

# روش های اپیدمیولوژی و پژوهش با نرم افزار Stata



دکتر علی احمدی  
حسین لشکر دوست

ویرایش جدید  
بهار ۱۴۰۳



زکات علم، انتشار دادن آن است.



در سه سطح مقدماتی، متوسطه و پیشرفته

## روش های اپیدمیولوژی و پژوهش با Stata

سرشناسه	: احمدی، علی، ۱۳۵۲ -
عنوان و نام پدیدآور	: اصول و روش های آماری در پژوهش های علوم پزشکی و اپیدمیولوژی با Stata /تالیف علی احمدی، حسین لشکردوست: [برای] دانشگاه علوم پزشکی شهرکرد، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی خراسان شمالی.
مشخصات نشر	: تهران: الیاس، ۱۳۹۲.
مشخصات ظاهری	: د، ۵۲۰ص:، مصور، جدول، نمودار؛ ۲۹ × ۲۲ س.م.
شابک	: ۹۷۸-۶۰۰-۵۷۲۷-۲۱-۰۰
وضعیت فهرست نویسی	: فیبا
موضوع	: نرم افزار استاتا
موضوع	: همه گیرشناسی -- برنامه های کامپیوتری
موضوع	: پزشکی -- تحقیق -- روش های آماری
موضوع	: آمار -- برنامه های کامپیوتری
سناسه افروده	: لشکردوست، حسین، ۱۳۶۱ -
سناسه افروده	: دانشگاه علوم پزشکی شهرکرد
سناسه افروده	: دانشگاه علوم پزشکی و خدمات بهداشتی درمانی خراسان شمالی
رده بندی کنگره	: ۱۳۹۲ الف۲الف/۳HAT۲
رده بندی دیویی	: ۰۰۵/۵۵
شماره کتابشناسی ملی	: ۳۳۲۶۷۱۵

۹۷۸-۹۶۴-۷۰۴۴-۲۹-۷



عنوان:	اصول و روش های آماری در پژوهش های علوم پزشکی و اپیدمیولوژی با Stata
تالیف:	علی احمدی - حسین لشکردوست
ناشر:	انتشارات الیاس
طرح جلد:	مهندس یاسر خسروشاه
چاپ و صحافی:	فرشویه، خیام
لیتوگرافی:	ترامسنج
ناظر چاپ:	امین رزاقی
چاپ اول:	پاییز ۱۳۹۲
تیراژ:	۱۰۰۰ جلد
بها:	۲۸۰۰۰ تومان

شابک: ۹۷۸-۶۰۰-۵۷۲۷-۲۱-۰۰

همراه با CD نرم افزار Stata و فایل های داده ای کتاب

فروشگاه و دفتر مرکز پخش: تهران، خیابان انقلاب، نبش ۱۲ فروردین، پلاک ۱۳۱۰، انتشارات الیاس، تلفن: ۶۶۴۰۵۰۸۴ و ۶۶۹۵۵۸۷۸

این کتاب در معاونت پژوهشی دانشگاه علوم پزشکی خراسان شمالی مورد داوری و به شماره ۹۲/۶۰/۷/۸۲۶ مورخ ۹۲/۷/۳ در کمیته انتشارات و نشر کتاب مورد تصویب قرار گرفت. کتاب در شورای انتشارات معاونت تحقیقات و فن آوری دانشگاه علوم پزشکی شهرکرد نیز تصویب و جهت چاپ مشترک با دانشگاه علوم پزشکی بجنورد مورد موافقت قرار گرفت. کلیه حقوق مادی و معنوی آن متعلق به مولفین می باشد. هر گونه چاپ مجدد با اجازه مولف خواهد بود.

### شناسنامه کتاب بالا قبلا توسط مولفین این کتاب منتشر گردید.

با توجه به شرایط پساپاندمی کووید-۱۹ و برای دستیابی فراگیران به اهداف یادگیری و کسب توانمندی های مورد انتظار، در راستای پاسخگویی اجتماعی به نیازهای جامعه و حمایت از علاقمندان استفاده از کتاب و به ویژه دانشجویان گرامی و برای توسعه آموزش های اپیدمیولوژی و ایفای مسئولیت حرفه ای، فایل PDF رنگی نسخه جدید کتاب به ارزش هر جلد پانصد هزار تومان، به صورت رایگان در دسترس علاقمندان قرار گرفته است.

کلیه حقوق مادی و معنوی این کتاب متعلق به مولفین است.



## اصول و روش های آماری

در پژوهش های علوم پزشکی و اپیدمیولوژی با

# STATA<sup>®</sup>

تالیف:

علی احمدی

حسین لشکردوست

با راهنمایی و مقدمه:

دکتر علی اکبر حقدوست

استاد اپیدمیولوژی دانشگاه علوم پزشکی کرمان

صفحه جلد کتاب منتشر شده در سال ۱۳۹۲

بنام خالق زیبایی ها

آن کس که پژوهش را استمرار نمی بخشد، از درک دانش بی نصیب است.

امام علی (ع)

## تقدیم به:

پیشگامان و استاتید ارجمند  
آمار زیستی و اپیدمیولوژی کشور

### بویژه:

استاد دکتر ابوالحسن ندیم

استاد دکتر سیدمحمدتقی آیت اللهی

استاد دکتر حمید سوری

استاد دکتر جعفر حسن زاده

استاد دکتر یدالله محرابی

و

محمد رضا و علیرضا

**فهرست مطالب**

۱.....	دیباچه به قلم استاد: دکتر حمید سوری
۲.....	مقدمه به قلم استاد: دکتر علی اکبر حقدوست
۳.....	پیشگفتار سال ۱۳۹۲ مولفین
۷.....	<b>پیشگفتار سال ۱۴۰۳</b> مولفین و تغییرات ایجاد شده در کتاب
<b>۹.....</b>	<b>بخش اول: اصول و مبانی کار با نرم افزار Stata</b>
۹.....	<b>فصل اول:</b> آشنایی با نرم افزار
۱۰.....	انواع نرم افزار و راه اندازی Stata
۱۴.....	آشنایی عمومی با محیط نرم افزار
۱۶.....	آشنایی اختصاصی با منوهای نرم افزار و ابزارها
۲۰.....	نگارش فرمان ها در نرم افزار
۲۲.....	ثبت و ذخیره کارهای انجام شده در نرم افزار
۲۳.....	کار عملی با نرم افزار
۲۵.....	تمرین ها
۲۶.....	<b>فصل دوم:</b> مدیریت داده ها در Stata
۲۷.....	مدیریت داده ها
۲۸.....	ساخت یک مجموعه از داده های جدید (ایجاد Dataset و ورود داده ها)
۳۲.....	سفارشی کردن داده ها و تغییر ویژگی های متغیرها
۳۸.....	تولید جدول مجازی برای داده ها و ساخت متغیرها و مدیریت آنها
۴۰.....	کار عملی با نرم افزار
۴۸.....	تمرین ها
<b>۵۰.....</b>	<b>بخش دوم: مبانی آمار توصیفی و کار با Stata</b>
۵۰.....	<b>فصل سوم:</b> مبانی آمار توصیفی و توصیف داده ها در Stata
۵۱.....	مبانی و اصول آمار توصیفی
۶۵.....	آماره های توصیفی و خلاصه در Stata
۶۶.....	رسم جدول ساده و متقاطع
۶۸.....	کار عملی با نرم افزار
۷۱.....	رسم حرفه ای جدول ها
۷۵.....	تمرین ها
۷۶.....	<b>فصل چهارم:</b> رسم نمودار در Stata
۷۷.....	نمودارها
۷۸.....	نمودارها در Stata

۷۹	گزینه های عمومی فرمان های مربوط به رسم نمودار .....
۸۱	کار عملی با نرم افزار و ترسیم انواع نمودارها .....
۸۴	نمودار پراکنش .....
۸۶	نمودار های بقا، منحنی کاپلان مایر، نمودار جدول عمر .....
۹۱	نمودار رادار .....
۹۳	ترسیم نقشه .....
۹۸	تمرین ها .....

### بخش سوم: آمار تحلیلی و کار با نرم افزار Stata ..... ۹۹

۹۹	<b>فصل پنجم: آزمون فرضیه ها و آنالیز دو متغیره و چند متغیره در Stata</b> .....
۱۰۰	تعاریف مربوط به آمار تحلیلی و آزمون های آماری .....
۱۰۳	توزیع نرمال و استاندارد .....
۱۰۹	استنباط آماری .....
۱۱۳	عدم قطعیت .....
۱۱۹	یادآوری آزمون های مهم و رایج .....
۱۲۵	راهنمای انجام آزمون های آماری رایج با نرم افزار Stata .....
۱۲۶	آزمون های مختلف تی (t-test) .....
۱۳۴	Chi square و آزمون های تعقیبی .....
۱۳۶	بررسی رابطه و انواع همبستگی ها .....
۱۳۸	همبستگی پارامتریک .....
۱۳۸	همبستگی ناپارامتریک .....
۱۴۴	همبستگی جزئی / تفکیکی .....
۱۴۸	آزمون های پارامتری برای مقایسه بیشتر از دو گروه (K گروه): آزمون های تحلیل واریانس .....
۱۵۲	کار عملی با نرم افزار و انجام تحلیل واریانس (ANOVA): .....
۱۵۴	آزمون های ناپارامتریک: کروسکال والیس، ویلکاکسون .....
۱۵۶	آنالیز واریانس دو طرفه MANOVA .....
۱۵۸	آزمون تحلیل واریانس یا کوواریانس تک متغیره ANCOVA .....
۱۶۲	تحلیل واریانس چند متغیره .....
۱۶۳	کار با نرم افزار و انجام تحلیل کامل MANOVA .....
۱۷۳	آزمون تحلیل واریانس اندازه های تکراری .....
۱۷۵	کار با نرم افزار .....
۱۸۴	معایب تحلیل واریانس تکراری و مدل Mixed .....
۱۸۸	تمرین ها .....

**بخش چهارم: مدل سازی آماری در اپیدمیولوژی با Stata ..... ۱۸۹**

**فصل ششم: همبستگی، رگرسیون و تحلیل های پیشرفته در اپیدمیولوژی با Stata ..... ۱۸۹**

۱۹۰ ..... مدل سازی

۱۹۵ ..... کار با نرم افزار در مدل سازی ها

۲۰۲ ..... همبستگی

۲۰۴ ..... رگرسیون خطی

۲۰۸ ..... کار با نرم افزار و ارزیابی مدل رگرسیون خطی در Stata

۲۱۶ ..... رگرسیون لجستیک ساده

۲۱۸ ..... اندازه های مهم در رگرسیون لجستیک

۲۲۱ ..... مفروضات مدل رگرسیون لجستیک

۲۲۲ ..... کار با نرم افزار

۲۲۷ ..... رویکرد های مدل سازی و تحلیل در رگرسیون لجستیک

۲۲۹ ..... نسبت حداکثر درستی ..... نسبت

۲۳۳ ..... ROC منحنی

۲۳۴ ..... مقایسه منحنی های ROC

۲۳۹ ..... رگرسیون لجستیک شرطی

۲۴۱ ..... کار با نرم افزار و انجام رگرسیون لجستیک شرطی

۲۴۹ ..... مقایسه رگرسیون لجستیک معمولی و شرطی

۲۵۰ ..... رگرسیون لجستیک رتبه ای

۲۵۷ ..... رگرسیون لجستیک چندوجهی

۲۶۰ ..... رگرسیون پواسون

۲۷۳ ..... آزمون فرض ها با رگرسیون پواسون

۲۷۴ ..... آزمون فرض اینترکشن در مدل پواسن

۲۷۵ ..... تحلیل داده های مطالعه کوهورت با رگرسیون پواسون

۲۷۷ ..... تحلیل بقا: (جدول عمر، روش کاپلان مایر و رگرسیون کاکس)

۲۸۲ ..... پیش فرضهای استفاده از روشهای آنالیز بقا

۲۸۵ ..... کار با نرم افزار

۲۹۱ ..... رگرسیون کاکس

۲۹۶ ..... مقایسه رگرسیون کاکس با رگرسیون پواسون

۲۹۸ ..... کنترل متغیرهای مخدوش کننده

۳۰۱ ..... تحلیل داده های وابسته به هم

۳۰۳ ..... مدل Robust Standard Error

۳۰۴ ..... مدل GEE



۳۰۸.....	Random Effect مدل
۳۱۰.....	مقایسه روش های تحلیل داده های به هم وابسته
۳۱۱.....	تحلیل مسیر Path analysis
۳۱۸.....	کار با نرم افزار و تحلیل مسیر
۳۲۹.....	تحلیل مدل چند سطحی
۳۳۰.....	مثال مفهومی تحلیل چند سطحی
۳۳۲.....	مدل چند سطحی برای متغیر پاسخ پیوسته
۳۳۹.....	افزافه کردن متغیرهای مستقل سطح دوم
۳۴۱.....	بر همکنش بین سطحی
۳۴۴.....	مدل چند سطحی برای متغیر پاسخ دو حالتی
۳۵۰.....	کار با نرم افزار
۳۵۴.....	تمرین ها

## بخش پنجم: روش های آماری و اپیدمیولوژی در تحقیقات علوم پزشکی ..... ۳۵۶

۳۵۶.....	فصل هفتم: مطالعات اپیدمیولوژیک و محاسبات اندازه های آنها در Stata
۳۵۷.....	مفاهیم پایه و علیت
۳۶۱.....	مطالعات اپیدمیولوژیک
۳۶۳.....	اندازه بیماری (پیامد) و اندازه اثر مواجهه
۳۶۴.....	درخت تصمیم گیری در برآورد رابطه در مطالعات
۳۶۶.....	کار با نرم افزار و محاسبه شاخص های اثر و رابطه در مطالعه مقطعی
۳۶۹.....	مطالعات همگروهی (کوهورت)
۳۷۱.....	شاخص های خطر نسبی
۳۷۳.....	کار با نرم افزار و محاسبه شاخص های اثر و رابطه در مطالعه کوهورت
۳۷۶.....	استانداردسازی مستقیم و غیرمستقیم میزان ها
۳۸۰.....	کار با نرم افزار و استانداردسازی میزان ها با Stata
۳۸۶.....	تحلیل رگرسیون کاکس برای مطالعات کوهورت
۳۸۷.....	مطالعات مورد - شاهد
۳۸۸.....	محاسبه شاخص های اثر و رابطه در مطالعه مورد شاهد
۳۹۰.....	تخمین خطر نسبی از نسبت شانس
۳۹۶.....	کار با نرم افزار و تحلیل مطالعات مورد شاهد
۴۰۱.....	مطالعات مداخله ای و کارآزمایی های بالینی
۴۰۲.....	انواع طرح های کارآزمایی بالینی
۴۰۵.....	طراحی تحقیقات بالینی در مطالعات درمان سرطان
۴۰۶.....	آنالیز داده ها در مطالعات مداخله ای و کارآزمایی های بالینی

۴۰۷	طرز محاسبه NNT در کارآزمایی بالینی
۴۰۸	اثر درمان (Treatment effect) و محاسبه با Stata
۴۱۳	خطاهای متداول در پژوهش های اپیدمیولوژیک
۴۱۳	شبیه سازی آماری و روش های باز نمونه گیری
۴۱۵	بوت استرپ
۴۱۶	جک نایف
۴۱۸	خطاهای پژوهش: تورش انتخاب، تورش اطلاعات
۴۲۹	انواع ارتباط بین متغیرها
۴۳۰	برهمکنش (اینترکشن)
۴۳۷	متغیرهای تعدیل کننده و میانجی (Moderation & Mediation)
۴۳۸	کار با نرم افزار و تعیین اثر مستقیم و غیرمستقیم متغیرها
۴۴۲	متغیر تعدیل کننده کیفی
۴۴۳	متغیر تعدیل کننده کمی
۴۴۵	ارزیابی عوامل موثر بر متغیر کیفی
۴۵۳	تحلیل حساسیت و کمی کردن خطاها
۴۵۳	کمی کردن خطای طبقه بندی نادرست
۴۵۵	کمی کردن تورش انتخاب
۴۵۷	کمی کردن مخدوش کننده کنترل نشده
۴۶۳	محاسبه ترکیب خطاهای پژوهش
۴۶۴	مطالعه ارزیابی تست های تشخیصی
۴۶۶	محاسبه و ارزیابی تست های تشخیصی با Stata
۴۶۷	محاسبه کاپا
۴۷۱	حدود اطمینان در آزمون های تشخیصی
۴۷۵	آزمون های چندتایی
۴۷۸	عدالت در سلامت و اندازه گیری نابرابری در سلامت
۴۸۰	کار با نرم افزار و محاسبه شاخص های نابرابری در سلامت
۴۸۲	ضریب جینی و شاخص تمرکز
۴۸۶	محاسبه شاخص های نابرابری مبتنی بر رگرسیون
۴۹۲	مدل DDD و کار با نرم افزار
۴۹۶	سری های زمانی
۴۹۹	عناصر سری زمانی
۵۰۰	کار با نرم افزار و تحلیل سری زمانی
۵۰۱	راهکارهای مدل سازی

۵۰۲	..... نمودارهای سری زمانی
۵۰۹	..... مدل ARIMA
۵۱۷	..... تمرین ها
۵۱۸	..... <b>فصل هشتم: حجم نمونه و توان در مطالعات</b>
۵۱۹	..... روش های نمونه گیری
۵۲۰	..... کار با نرم افزار و تعیین حجم نمونه و توان مطالعه
۵۲۲	..... حجم نمونه برای مقایسه میانگین یک گروه با مقدار مشخص
۵۲۴	..... مقایسه دو میانگین
۵۲۶	..... مقایسه دو نسبت
۵۲۸	..... کار با نرم افزار و محاسبه توان آزمون
۵۳۰	..... <b>فصل نهم: متا آنالیز</b>
۵۳۱	..... مقدمه ای بر متا آنالیز
۵۳۳	..... گزارش نتایج متاآنالیز
۵۳۳	..... متاآنالیز با Stata
۵۳۴	..... متاآنالیز متغیرهای پیوسته با نرم افزار
۵۳۶	..... متاآنالیز متغیرهای کیفی
۵۳۸	..... کار با نرم افزار و معرفی دستورات مهم
۵۴۶	..... دستورات متاآنالیز (meta)
۵۵۰	..... دستورات متاآنالیز (metan)
۵۵۳	..... دستورات متاآنالیز (metacum)
۵۶۱	..... متاآنالیز P- Values
۵۶۲	..... تمرین ها
۵۶۳	..... <b>فصل دهم: روایی و پایایی</b>
۵۶۴	..... مقدمه ای بر روان سنجی ابزارهای جمع آوری داده ها
۵۶۵	..... روایی
۵۷۳	..... پایایی
۵۷۴	..... انواع خطا و رابطه با پایایی و روایی
۵۷۵	..... انواع سنجش های پایایی
۵۷۶	..... محاسبه ICC با نرم افزار
۵۷۹	..... محاسبه آلفای کرونباخ با نرم افزار Stata
۵۸۱	..... تحلیل عامل
۵۸۲	..... کار با نرم افزار

۵۹۳.....	تحلیل کلاس پنهان (LCA) .....
۶۰۰.....	پیوست شماره یک: معرفی فایل داده های کار با نرم افزار Stata.....
۶۰۱.....	پیوست شماره دو: فلوجارت های استنباط آماری.....
۶۰۵.....	پیوست شماره سه: فرمول های مهم در مطالعات اپیدمیولوژیک.....
۶۱۴.....	ادامه پیوست سه: سنجش های اساسی در مطالعات اپیدمیولوژی .....
۶۱۷.....	پیوست شماره چهار: معرفی روش های جدید تحلیل در Stata .....
۶۱۸.....	پیوست شماره چهار: شکل های طراحی انواع مطالعات .....
۶۲۴.....	پیوست شماره پنج: جدول توزیع Z .....
۶۲۵.....	پیوست شماره شش: جدول توزیع t .....
۶۲۶.....	پیوست شماره هفت: جدول توزیع F .....
۶۲۷.....	پیوست شماره هشت: جدول عمر خلاصه شده.....
۶۲۸ .....	ایندکس (واژه نامه فارسی - انگلیسی) .....
۶۳۱ .....	منابع .....

## دیباچه

### جناب آقای دکتر احمدی

با سلام، تشکر و خسته نباشید. کتاب ارزشمندی نگارش کرده اید. به شما تبریک می گویم. کتاب اصول و روش های آماری در پژوهش های علوم پزشکی با استفاده از نرم افزار Stata کتابی مفید و کاربردی است که مطالعه آن برای پژوهشگران مختلف توصیه می شود. پیچیدگی های روش شناسی در عرصه تحقیق به خصوص پژوهش های کمی که مبنای اپیدمیولوژی دارند، ضرورت استفاده از ابزارهای پیشرفته ای چون Stata را در بسیاری از موارد اجتناب ناپذیر کرده است. امکان خودآموزی این ابزارها نیز محدود و به خصوص برای افراد کم تجربه تر، بسیار مشکل است. لذا در اختیار داشتن کتابی به شکل خودآموز، می تواند بسیار ارزنده باشد. محصولات به دست آمده از تحقیق که در حکم منبعی برای گزارش یافته های تحقیق، به کار می رود داده نامیده می شوند. در صورتی که این داده ها به خوبی تجزیه و تحلیل و گزارش نشوند عملاً محصول تحقیق اشتباه و برای کاربر، نتایج گمراه کننده خواهد بود. قبل از انتشار، اهمیت داده ها برای محقق در این است که با بررسی داده ها، صحت و سقم آنچه انجام شده مشخص می شود که این مسئله در چه قسمت هایی خوب پیش رفته و در چه قسمت هایی مشکل داشته است. این کار در نتیجه گیری و حمایت از یافته های معتبر و قابل گزارش استفاده می شود. پس از انتشار، اهمیت داده ها در این است که پژوهشگر و سایر افراد را قادر می سازد تا از آنها برای بررسی صحت آنالیزها استفاده کنند، به آزمون فرضیه ی جایگزین، از تقاضای مالکیت معنوی حمایت یا با ادعاهای نادرست مخالفت کند. از آنجا که داده اشکال گوناگونی دارد، مدیریت داده ها نیز دارای قواعدی است که محقق باید توانایی اداره کردن این گستره ی متفاوت از انواع داده را داشته باشد. مدیریت مؤثر داده ها با برنامه ریزی قبل از آغاز جمع آوری داده ها شروع می شود و در این راستا فرآوری داده ها و استفاده از ابزاری مناسب برای تجزیه و تحلیل داده ها بسیار با اهمیت است. گرچه برای تجزیه و تحلیل برخی از روش های آماری معمول، می توان از ابزارهای عمومی تر آماری استفاده کرد ولی این نرم افزار، برای پردازش برخی مدل ها و استفاده از داده های مگا کاربردهای ویژه ای دارد. نرم افزار Stata کاربرد جهانی دارد و در بسیاری از مراکز آموزشی و علمی جهان یادگیری آن ضروری است. داشتن مهارت لازم برای استفاده از این نرم افزار جزء مهارت هایی است که برای احراز بسیاری از موقعیت های شغلی علمی لازم است. کتاب حاضر علاوه بر ساختار مناسب، به شیوه ای آسان امکان آموزش حداکثری خواننده را فراهم کرده و با ارائه تمرین های خوب یک خودآزمایی قابل سنجش و مفید را برای مخاطب فراهم کرده است. گرچه این کتاب برای افرادی تدوین شده که باید اطلاعات پایه ای آمار، اپیدمیولوژی و روش تحقیق را بدانند اما با ارائه مثال های متنوع امکان بهره برداری محققان رسته های مختلف علوم پزشکی را نیز فراهم آورده است. کاربرد تلفیقی متن با تصاویر و شکل ها همراه داده های مختلف، ساختار کتاب را جذاب تر کرده است. به نظر می رسد که تمامی مراکز تحقیقاتی و گروه های آموزشی مرتبط، به نسخه ای از این کتاب نیاز دارند. شاید تلفیقی از مطالعه کتاب با کارگاه های آموزشی مرتبط بهره برداری از کتاب را چند برابر خواهد کرد. اینجانب ضمن تشکر از نویسنده محترم کتاب که تلاش قابل تحسینی را انجام داده اند مطالعه آن را به تمامی دانشجویان و محققان حوزه سلامت توصیه می کنم. مطالعه کتاب برای به خصوص دانشجویان رشته های اپیدمیولوژی و آمار و دانشجویان دکترای تخصصی پژوهش، می تواند در ارتقاء دانش ایشان و در نتیجه افزایش کیفیت نتایج تحقیق بسیار تاثیر گذار باشد.

دکتر حمید سوری – تابستان ۱۳۹۲

استاد اپیدمیولوژی – مدیر دپارتمان اپیدمیولوژی دانشگاه علوم پزشکی شهید بهشتی

رئیس مرکز تحقیقات پیشگیری از مصدومیت ها و ارتقاء ایمنی

مسئول دوره دکتری تخصصی پژوهش دانشگاه علوم پزشکی شهید بهشتی

استفاده از نرم افزارهای کاربردی در درک مفاهیم آماری؛ اجرای تحلیل های پیچیده و یافتن پاسخ سوالات کلیدی مبتنی بر نتایج تحقیق، یکی از ارکان اصلی عمده تحقیقات در علوم پزشکی شده است. این مهم آنقدر پر رنگ است که شاید بتوان به صورت قاطع گفت که در دنیای امروز دانستن آمار و توانایی کار با نرم افزارهای آماری یکی از اجزای بسیار مهم پژوهش و روش تحقیق است. به همین دلیل در برنامه آموزشی تقریباً تمامی رشته های تحصیلات تکمیلی این مباحث گنجانده شده است. پر واضح است که ضرورت فراگیری نرم افزارها برای رشته های خاصی مانند آمار حیاتی و اپیدمیولوژی بسیار پر رنگتر است. چرا که دیگران ممکن است حاشیه ای و در یک مقطع کوتاه از تحقیقات خود به این دانش و مهارت نیاز داشته باشند ولی این رشته ها به صورت روزمره با مباحث مذکور در ارتباط هستند و بایستی بتوانند با اشراف بسیار بالا حتی پاسخگوی سوالات دیگران نیز باشند. اما یک سوال کلیدی که عمدتاً ذهن متخصصین را به خود مشغول می کند انتخاب بهترین نرم افزارهایی است که بتوانند به شکل مناسب به نیازهای ایشان پاسخ دهد. در جامعه پزشکی و تحقیقات علوم پزشکی کشور، شاید رایجترین و شناخته شده ترین نرم افزار در این زمینه SPSS باشد و این اسم آنقدر جا افتاده است که با وجود تغییر اسم اصلی برنامه، هنوز به صورت عمومی همان نام قبلی را بکار می برند. استفاده از این برنامه ساده بوده و کاربر به راحتی می تواند با برنامه ارتباط برقرار نماید و از منوهای آن استفاده کند. اما شاید یک دلیل اصلی رایج شدن نرم افزار، وجود کتابها و سی دی های آموزشی بسیار زیاد در آن رابطه باشد. اما با وجود این مزایا، بسیاری از نیازهای آماری حرفه ای در تحقیقات علوم پزشکی توسط آن پاسخ داده نمی شود.

نرم افزار رایج دیگری که اختصاصاً برای تحقیقات علوم پزشکی و اپیدمیولوژی طراحی و دانشگاههای بسیار بزرگ دنیا آن را به عنوان نرم افزار اصلی خود برگزیده اند Stata می باشد. این نرم افزار بسیار گسترده و وسیع بوده و قابلیت های بسیار زیادی دارد. اما به اندازه SPSS در جامعه علمی ایران شناخته نشده است. یکی از رسالتهای اصلی مدرسین و متخصصان ایران باز کردن راه ورود این نرم افزار به دانشگاه ها و آشنا نمودن دانشجویان به این نرم افزار بسیار کاربردی و حرفه ای است. در این راستا قطعاً نوشتن کتابهایی به زبان ساده و کاربردی یکی از اهم وظایف است.

من از نگارنده محترم این کتاب که با زحمت زیاد این مجموعه مفید، جامع و عمیق را تدوین نموده اند کمال تشکر را دارم و مطمئن هستم در دسترس بودن این کتاب کمک شایانی به رشد دانش عمیق اپیدمیولوژی در کشور خواهد کرد. لذا به همه همکاران محترم توصیه اکید می کند که این کتاب را با دقت بخوانند. البته پر واضح است که به دلیل گستردگی بسیار زیاد نرم افزار قطعاً در یک کتاب نمی توان تمامی مفاهیم را وارد نمود ولی این را نیز باور دارم که قسمت عمده آموزش باید از طریق مطالعه و کنکاش شخصی حاصل شود و این مجموعه سرخط ها را به خوانندگان خواهد داد که ضمن آشنا شدن با مفاد اصلی نرم افزار؛ با شجاعت و اعتماد به نفس بیشتری با برنامه کار کنند و فرا بگیرند که چگونه سوالات خود را پاسخ دهند. بار دیگر از تلاشهای نگارنده محترم تشکر می کنم. امیدوارم که بتوانم در آینده نیز کتابهای بیشتری از ایشان مطالعه کنم.

دکتر علی اکبر حقدوست – اپیدمیولوژیست، دانشگاه علوم پزشکی کرمان

شهریور ماه ۱۳۹۲

## پیشگفتار مولفین

در دیباچه و مقدمه، اهمیت نرم افزار Stata به خوبی به قلم اساتید ارجمند و برجسته اپیدمیولوژی بیان گردید. در پیشگفتار به معرفی کتاب و اهداف آن پرداخته می شود.

اپیدمیولوژی عبارتست از مطالعه توزیع و بررسی عوامل تعیین کننده وقایع و پیشامدهای مرتبط با سلامت و بیماری در جمعیت های مشخص و بکارگیری این مطالعه برای حل مشکلات مرتبط با سلامت و ارتقا سطح سلامت افراد و پیشگیری از ابتلا و مرگ و میر در آنها. در بین علوم پایه موجود، اپیدمیولوژی دانشی نوپا و یکی از مهمترین و اساسی ترین علوم پایه در پزشکی است (۳-۱). بنا بر گفته استاد محترم جناب آقای دکتر اکبری - معاون سابق سلامت وزارت بهداشت درمان و آموزش پزشکی- اپیدمیولوژی اگر نباشد پایه های علوم پزشکی می لرزد. اپیدمیولوژی هم علم است و هم تکنیک. تکنیک های این علم، آن را از دیگر علوم متمایز و منحصر به فرد و خاص جلوه می دهد. این دانش و تکنیک از زمان بقراط در دنیای قدیم تا جان اسنو در دنیای جدید پیشرفت های چشمگیری داشته است. علیرغم همکاری های نزدیک متخصصین اپیدمیولوژی و پزشکان علاقمند به بهداشت عمومی با متخصصین آمار از جمله Cornfield, Mantel, Breslow و Cox و توسعه اپیدمیولوژی مدرن، همواره این همکاری ها و مشارکت ها و درک صحیح آمار در اپیدمیولوژی و بسط تکنیک ها و تبدیل تکنیک به علم و پاسخ به سوالات ضروری به نظر می رسد (۳، ۴). امروزه تحقیق و پژوهش در رشته های علوم پزشکی بدون توجه به اپیدمیولوژی و آمار امکان پذیر نیست. و لذا ناگزیر، پژوهشگران باید با این علوم به عنوان زیربنای تحقیقات، آشنایی لازم را داشته و با بهره گیری از تکنیک ها، به روش صحیح، به نیازها و سوالات پاسخ دهند (۴، ۵).

برای کار در حوزه اپیدمیولوژی نوین، دانستن تاریخچه، تعاریف، اهداف اختصاصی، چگونگی تحلیل داده های اپیدمیولوژیک و مباحث خاص و روش های اپیدمیولوژی و آماری از جمله:

- طبقه بندی انواع اپیدمیولوژی و حیطه های مورد گسترش و کاربردهای آن در سلامت و علوم پزشکی
  - مدل های اپیدمیولوژیک، مدل های علیت و <sup>۱</sup>DAG ها و کاربرد آنها در مطالعات مشاهده ای
  - مفاهیم پایه و اساسی از جمله تعاریف، کاربرد و تفسیر شاخص ها، میزان، نسبت و تناسب و روش های استاندارد سازی
  - شاخص های اندازه گیری اثر و رابطه، روایی و پایایی آنها، نحوه محاسبه، کاربرد و تفسیر آنها
  - انواع مطالعات، کاربرد و نقاط ضعف و قوت آنها، کاربرد آمار و مدل سازی در مطالعات اپیدمیولوژیک
  - تجزیه و تحلیل داده ها و تسلط بر نرم افزار مطلوب و مخصوصا تفسیر نتایج مطالعات و تهیه گزارش مناسب
  - شناخت خطاهای مختلف در انواع مطالعات و طرز تهیه یک پروپوزال مناسب، ضروری و مورد نیاز است (۳-۱، ۶، ۷).
- پرداختن به همه موارد بالا، از حوزه و اهداف این کتاب خارج است. اما تا جایی که مقدور بوده و مرتبط با مباحث آمار و روش های اپیدمیولوژی و کاربرد نرم افزار Stata باشد مطالب تشریح شده است. این کتاب در واقع با تجربه مولفین در تدوین کتاب راهنمای Stata برای پژوهشگران در سال ۱۳۸۶- به عنوان اولین کتاب فارسی راهنمای Stata در ایران تهیه گردیده است (۸). علاوه بر آن اسلایدهایی از کتاب مذکور تهیه و به صورت رایگان در وبلاگ اپیدمیولوژی گذاشته شد. بیش از ۱۰۰۰ بار اسلایدها دانلود و یا از طریق ایمیل درخواست شدند (۹). اینک پس از هفت سال از چاپ کتاب نخست، علیرغم همه دشواری ها و بعضا سختی ها، با برگزاری کارگاه های متعدد در دانشگاه ها و مراکز تحقیقاتی، کتاب حاضر به صورت کامل تر و با پوشش جامع تحلیل های آماری در اپیدمیولوژی به راهنمایی استاد ارجمند جناب آقای

دکتر حقدوست تهیه گردید. امید است این اثر نیز بتواند مفید واقع شود. منابع مورد استفاده که در تالیف، تدوین و گردآوری فصل های مختلف مورد استفاده قرار گرفته و در انتهای پیشگفتار نیز مشخص شده اند در پایان کتاب لیست گردیدند.

گرچه این کتاب برای راهنمایی پژوهشگران عمومی در حوزه های مختلف علوم پزشکی تدوین گردید، اما به عنوان یک کتاب مرجع و کمک درسی مطابق با سرفصل وزارت بهداشت، درمان و آموزش پزشکی برای دروس آمار و روش های اپیدمیولوژی در مقاطع مختلف رشته های علوم پزشکی و دروس تجزیه و تحلیل داده های بهداشتی و روشهای آماری در اپیدمیولوژی در مقطع کارشناسی ارشد اپیدمیولوژی و سایر رشته های تحصیلات تکمیلی علوم پایه پزشکی از جمله پرستاری، مامایی، آمار زیستی، آموزش بهداشت، بهداشت باروری، سالمندشناسی، فیزیولوژی، انگل شناسی، میکروبی شناسی، ویروس شناسی، پزشکی مولکولی، مهندسی بهداشت محیط، بهداشت حرفه ای، مدیریت خدمات بهداشتی درمانی، اقتصاد بهداشت و دارو، ایمونولوژی، پروتئومیکس، سیاستگذاری سلامت، دوره های مختلف عالی بهداشت عمومی (MPH)، پزشکی اجتماعی و دکترای اپیدمیولوژی تهیه و تنظیم گردیده است. کاربرد دیگر کتاب استفاده در برگزاری کارگاه های آموزشی مقدماتی، متوسطه و پیشرفته Stata، روش تحقیق و کاربرد آمار در پژوهش ها است. کتاب در نوع خود اولین مرجعی است که متناسب با دوره دکترای تخصصی پژوهشی (Ph.D by Research) در ایران، به جامعه پژوهشگران عرضه می شود.

کتاب مشتمل بر پنج بخش، ده فصل و هشت ضمیمه است. یک عدد سی دی حاوی نرم افزار Stata و مجموعه داده های مورد نیاز (سی و دو فایل داده) برای کار با نرم افزار ضمیمه کتاب می باشند. هر بخش حاوی یک فایل داده جداگانه است که برای کار با نرم افزار باید از آنها استفاده شود. این فایل ها در ضمیمه شماره یک کتاب معرفی شده اند. ذکر اهداف آموزشی، آموزش گام به گام نرم افزار و مفاهیم آماری از دیدگاه اپیدمیولوژی از مقدماتی تا پیشرفته، بیان خلاصه و نتیجه گیری در هر فصل و وجود تمرین در انتهای هر فصل، یکی دیگر از ویژگی های مناسب و کاربردی این کتاب است که کاربر می تواند از آنها استفاده نموده و ضمن ارزیابی خویش، مهارت کار با نرم افزار را افزایش دهد. وجود تحلیل های پیشرفته از جمله انواع مدل سازی های آماری، محاسبه شاخصهای اندازه اثر و ارتباط در مطالعات اپیدمیولوژیک، تحلیل حساسیت و کمی کردن خطاهای رایج در پژوهش ها، شبیه سازی آماری و معرفی روشهای بازنمونه گیری بوت استرپ، مونت کارلو و جک نایف، معرفی انواع طراحی کارآزمایی بالینی در تحقیقات علوم پزشکی و مخصوصا در بررسی سرطان ها و ضمیمه نمودن شکل های طراحی مطالعات و فرمول های آماری که به صورت جامع و یکجا و بی نظیر، جمع آوری شده اند و همچنین روشهای محاسبه شاخص های نابرابری در سلامت و پرداختن به اپیدمیولوژی اجتماعی و عدالت در سلامت نیز از نقاط قوت و موضوعات جدیدی است که برای اولین بار به فارسی به عنوان راهنمای پژوهشگران، در این کتاب در دسترس قرار می گیرند. در بخش اول کتاب به اصول و مبانی کار با نرم افزار Stata پرداخته شد. شناخت نرم افزار، نحوه نصب، پنجره ها، مدیریت داده ها، ساخت و تغییر در متغیرها، دستورات مورد نیاز، پنجره گفتگو و هر آنچه برای کار با این نرم افزار نیاز به دانستن است به صورت مفید و خلاصه توضیح داده شد. فرانس های این فصل در پایان کتاب آمده است (۸، ۱۲-۱۰).

در بخش دوم به اصول و مبانی آمار توصیفی و محاسبه شاخص های توصیفی مانند شاخص های تمایل مرکزی و شاخص های پراکندگی و انواع جداول و نمودارهای آماری پرداخته شده است. این بخش از کتاب های آماری گردآوری شده است. منابع مورد استفاده در پایان کتاب آمده است (۷، ۱۷-۱۳). در بخش سوم آمار تحلیلی و کار با نرم افزار و نحوه



استفاده از آزمون های آماری در پژوهش ها تشریح شد. در این بخش از آزمون های ساده مانند انواع آزمون های تی، کای اسکور، آنالیز واریانس یکطرفه و دو طرفه و آنالیز های پیشرفته تر از جمله آنالیز اندازه گیری های متواتر و تحلیل واریانس چند متغیره و تحلیل کوواریانس آمده است. مانند بخش دوم، در این بخش نیز برای گردآوری مطالب، از کتاب های آماری مرتبط و معتبر استفاده شده است (۴، ۱۴-۱۲، ۲۶-۱۸). **بخش چهارم** به مدل سازی و روش های آماری، اپیدمیولوژی و تحلیل های پیشرفته اختصاص یافته است. در این بخش انواع مدل های آماری از جمله رگرسیون خطی ساده و چندمتغیره، رگرسیون لجستیک معمولی، رگرسیون لجستیک شرطی، رگرسیون پواسون، تحلیل بقا و رگرسیون کاکس و تحلیل داده های وابسته به هم آمده است. در این بخش به سری های زمانی و کاربرد آنها در تحقیقات علوم پزشکی و همچنین به تحلیل چند سطحی نیز پرداخته شده است. همچنین از منابع معتبر و کاربردی منتشر شده و منتشر نشده و مشاوره با اساتید صاحب نظر استفاده گردید. منابع مورد استفاده در پایان کتاب آمده است (۶-۴، ۱۲-۱۰، ۲۳، ۳۲-۲۷). **خلاصه بخش های دوم تا پنجم** در قالب فلوجارت آزمون ها و انواع تحلیل و استنباط آماری که از کتاب مرجع اصول آمار زیستی تالیف پروفسور رازنر (Rosner) اقتباس شده است در ضمیمه شماره دو کتاب آمده است. استفاده از این فلوجارت ها در فرایند تصمیم گیری ها در تحلیل آماری کمک کننده است. شناخت اصولی و صحیح این فرایندها، به منزله درک صحیح آمار و روش های آن است. در **بخش پنجم** کتاب، روش های آماری در پژوهش های اپیدمیولوژیک آمده است. ابتدا به صورت خلاصه به مفاهیم و **مراحل هفت گانه** تحلیل داده های اپیدمیولوژیک پرداخته شد. سپس مطالعات و پژوهش های اپیدمیولوژیک معرفی و نحوه محاسبه شاخص های اثر و رابطه و کار با نرم افزار برای مطالعات مقطعی، کوهورت و مورد شاهد نیز ذکر گردید. در ادامه برای درک بیشتر مباحث تخصصی در تحلیل ها به مباحث انواع مطالعات کارآزمایی بالینی، طراحی مطالعات مربوط به ارزیابی داروها در سرطان، تست های تشخیصی، خطاهای تصادفی و منظم، تورش ها، اینترکشن یا اثر متقابل و مخدوش کننده ها نیز پرداخته شد. سپس نحوه کنترل و تطبیق بر اساس خطاها و کمی نمودن خطاهای اطلاعات، انتخاب و مخدوش شدگی کنترل نشده با نرم افزار محاسبه گردید.

به شبیه سازی آماری و روش های **بوت استرپ**، **مونت کارلو** و **جک نایف** نیز اشاره گردید. معرفی شاخص های عدالت در سلامت و اندازه گیری نابرابری در سلامت به عنوان یک مبحث مهم و اساسی در تحلیل های **اپیدمیولوژی اجتماعی**، در این بخش معرفی شده است. این مطالب توسط مولف به صورت سخنرانی جامع در هفتمین کنگره اپیدمیولوژی ایران در دانشگاه علوم پزشکی یاسوج ارائه شده بود. محاسبه حجم نمونه و توان آزمونها، انجام متآنالیز با نرم افزار Stata و تحلیل عامل و اندازه گیری روایی و پایایی نیز آخرین قسمت از این بخش است. به طور خلاصه و مفید نحوه محاسبه روایی و پایایی یک پرسشنامه و انجام متآنالیز و مثال های کاربردی و نحوه آماده ساختن داده ها برای متآنالیز توضیح داده شد. در این بخش از منابع معتبر و مشاوره با اساتید صاحب نظر و استفاده از منابع معتبر منتشر شده و منتشر نشده استفاده شده است. منابع مورد استفاده در پایان کتاب آمده است (۴-۱، ۶، ۷، ۱۱، ۱۳، ۱۶، ۲۹، ۴۳-۳۲).

کتاب توسط مولفین به استاد ارجمند جناب آقای دکتر سوری ارسال گردید. ایشان نیز درباره کتاب نکات ارزشمند و راهنمایی هایی را فرمودند که مورد استفاده قرار گرفت و دیباچه کتاب نیز به قلم استاد معظم، زیور طبع یافت که مایه خوشحالی و سپاسگزاری است. بنابر نظر مولفین و **تایید** اساتید اپیدمیولوژی و آمار زیستی مشهور، ممتاز و صاحب نظر، کتاب در نوع خود در زبان فارسی بی نظیر است. استفاده پژوهشگران و دانشجویان عزیز از این کتاب می تواند به توانمندسازی آنها بیانجامد. ان شاء الله

از راهنمایی ها و نقطه نظرات اساتید ارجمند گروه اپیدمیولوژی و آمار زیستی دانشکده بهداشت دانشگاه علوم پزشکی شهرکرد جناب آقایان دکتر سلیمان خیری و دکتر محمود مباحثی تقدیر و تشکر می شود. مجدداً از استاد بزرگوار جناب

آقای دکتر علی اکبر حقدوست استاد اپیدمیولوژی دانشگاه علوم پزشکی کرمان، بخاطر تشویق، مساعدت در تهیه منابع و راهنمایی در تدوین و نوشتن مقدمه کتاب صمیمانه تشکر می شود.

از اساتید ارجمند در **دپارتمان اپیدمیولوژی دانشگاه علوم پزشکی شهید بهشتی** جناب آقایان دکتر حمید سوری، دکتر یدالله محرابی، دکتر سهیلا خداکریم، دکتر کورش اعتماد، دکتر ایاد بهادری، دکتر بابک عشرتی، دکتر داود خلیلی، سرکار خانم دکتر پروین یآوری و **بویژه** از دکتر سید سعید هاشمی نظری بخاطر پاسخگویی به سوالات و راهنمایی صمیمانه سپاسگزاری می شود. از همکاران محترم و دانشجویان گرامی در دپارتمان اپیدمیولوژی، مخصوصا سرکار خانم دکتر مرضیه روحانی رصاف بخاطر همکاری در تدوین مطالب مربوط به کار با نرم افزار و تحلیل چند سطحی در انتهای بخش چهارم کتاب و سرکارخانم دکتر سیما مسعودی بخاطر مساعدت در تهیه منابع تشکر می شود.

مایه افتخار و خوشحالی است که در دوران کارشناسی ارشد اپیدمیولوژی، دروس آمار و روش های اپیدمیولوژی را از محضر اساتید بزرگوار جناب آقای دکتر سید محمد تقی آیت اللهی و دکتر جعفر حسن زاده در دانشگاه علوم پزشکی شیراز و آقای دکتر یدالله محرابی در دانشگاه علوم پزشکی شهید بهشتی بهره مند بودیم. بدینوسیله از ایشان نیز سپاسگزاری می شود.

از **معاون محترم پژوهشی دانشگاه علوم پزشکی خراسان شمالی** و همکاران ارجمند ایشان در حوزه معاونت پژوهشی آن دانشگاه نیز بخاطر داوری، قبول و تصویب کتاب، و همچنین از **معاونت محترم پژوهشی دانشگاه علوم پزشکی شهرکرد** بخاطر عقد تفاهم نامه همکاری با دانشگاه علوم پزشکی خراسان شمالی جهت چاپ کتاب سپاسگزاری می شود. سهم همکار محترم جناب آقای لشکردوست در تالیف کتاب بر اساس تفاهم نامه منعقد شده، یک سوم و معادل ۳۴٪ بوده است.

امید است این اثر به جامعه علمی کشور و پژوهشگران عزیز و دانش پژوهان اپیدمیولوژی کمک کننده باشد و در راستای فرمایشات و منویات مقام معظم رهبری مبنی بر استفاده از پژوهش در تصمیم گیری ها، بومی سازی دانش و توجه به تولید علم و رسیدن ایران اسلامی به جایگاه واقعی در دنیا، گام های موثری برداشته شود. ممکن است کتاب نقایصی داشته باشد. بی شک نظرات، پیشنهادات و انتقادات خوانندگان محترم در رفع نقایص موجود و بهبود کتاب در چاپ های آتی کمک کننده خواهد بود. نظرات خویش را به آدرس [aliamhadi2007@gmail.com](mailto:aliamhadi2007@gmail.com) و یا [doostaria@gmail.com](mailto:doostaria@gmail.com) ارسال فرمایید.

و ماتوفیقی الا بالله علیه توکلت والیه انیب

علی احمدی - دکترای تخصصی اپیدمیولوژی دانشگاه علوم پزشکی شهید بهشتی

**اپیدمیولوژیست دانشگاه علوم پزشکی شهرکرد**

حسین لشکردوست - عضو هیات علمی دانشگاه علوم پزشکی خراسان شمالی

**پاییز ۱۳۹۲**

## پیشگفتار نسخه جدید کتاب در سال ۱۴۰۳

خداوند متعال را سپاسگزاریم، بعد از ۲۰ سال کار با نرم افزار Stata و انتشار اولین کتاب مقدماتی آن در سال ۱۳۸۶ در ایران با عنوان "راهنمای کار با نرم افزار Stata" توسط انتشارات آصف (شاپک ۷-۳۹-۷۰۴۳-۹۶۴-۹۷۸) و سپس انتشار کتاب جامع "اصول و روش های آماری در پژوهش های علوم پزشکی و اپیدمیولوژی با Stata"، در سال ۱۳۹۲ توسط انتشارات الیاس (شاپک ۰-۲۱-۵۷۲۷-۶۰۰-۹۷۸)، اینک نسخه جدید کتاب در بهار ۱۴۰۳ با عنوان "روش های اپیدمیولوژی و پژوهش با Stata" و استفاده از نسخه ۱۷ و ۱۸ نرم افزار توسط مولفین تدوین و در اختیار علاقمندان قرار گرفته است.

### در نسخه جدید کتاب، موضوعات زیر اضافه و برای اولین بار در ایران در دسترس پژوهشگران است:

- استفاده از نسخه های ۱۷ و ۱۸ و بروزرسانی نرم افزار Stata و انجام کلیه تحلیل ها با آن
- بازننگری، بازنویسی و تکمیل دستورات مدیریت داده ها در فصل اول و دوم،
- پرداختن به عدم قطعیت، تکمیل نمودن دستورات برای آمار توصیفی و تحلیل اولیه داده ها و اجرای آنها با نرم افزار
- تکمیل دستورات برای جدول های یک طرفه و دوطرفه و ترسیم حرفه ای جدول، نقشه و نمودار رادار برای اولین بار
- ارایه گام های تحلیل داده ها و تمرکز بر ارزیابی کیفیت داده ها در سه مرحله Explore و Transform و Validate
- پرداختن به منوی Epidemiology & Related و ارایه دستورات برای تحلیل انواع مطالعات اپیدمیولوژیک (مقطعی، مورد- شاهد، کوهورت، تحلیل داده های همسان شده) با ذکر مثال و کار با نرم افزار استاتا
- ارایه نحوه محاسبه اثر درمان (Treatment Effect) در کارآزمایی های بالینی و کار با نرم افزار Stata
- ارایه مدل های DID و DDD برای برآورد اثرات مداخله در داده های پانل و طولی، تجربی و کارآزمایی بالینی
- ارایه محتوا پیرامون استانداردسازی مستقیم و غیرمستقیم و نحوه استفاده از نرم افزار برای تطبیق کردن میزان ها
- تحلیل مسیر و تهیه راهنمای کار با نرم افزار و اجرای مدل SEM و مدل تحلیل کلاس پنهان (LCA)
- پرداختن به روابط بین متغیرها و ترسیم دیاگرام متغیرهای تعدیل کننده و میانجی (Moderators & Mediation)
- تکمیل دستورات متا آنالیز و استفاده از منوهای برنامه استاتا با ذکر چند مثال و اجرای متاآنالیز توصیفی و تحلیلی
- ارایه محتوا پیرامون مدل سازی، ضرورت و گام های آن در مطالعات اپیدمیولوژیک و کار با نرم افزار stata و مدل سازی
- تکمیل دستورات انواع مدل های رگرسیونی ساده و چندگانه (خطی، لوجستیک، پواسون، مولتی نومیال، رتبه ای، کاکس)
- بازنویسی تحلیل های چند سطحی (Multilevel Analysis) و محاسبه شاخص های نابرابری و استفاده از DASP
- بازنویسی و تکمیل شاخص های ارزیابی تست های تشخیصی و کار با نرم افزار برای محاسبه آنها و ROC
- نحوه محاسبه توان (Power) و ارایه دستورات محاسبه حجم نمونه در مطالعات اپیدمیولوژیک با نرم افزار stata
- بازنویسی و تکمیل نحوه کار با نرم افزار و محاسبه شاخص های اعتبارسنجی ابزار جمع آوری داده ها و محاسبه روایی و پایایی ( , Kappa, ICC, Cronbach's alpha, Sen, Spe, Lr test, )
- ارایه خلاصه نحوه استفاده و مسیر رگرسیون بتا، لاسو، Zero inflated, رگرسیون توبیت (Tobit regression)

بعد از چاپ نسخه قبلی کتاب در سال ۱۳۹۲ و توزیع آن، با استقبال گسترده علاقمندان روبرو گردید. در چاپ مذکور بیش از هزار نسخه از کتاب توسط ناشر به فروش رفته و از سال ۱۳۹۴ کتاب در کشور نایاب گردید. با توجه به تقاضاهای متعدد علاقمندان، چندین بار از طرف ناشر درخواست برای تجدید چاپ کتاب شده بود، منتها با توجه به مشغله فراوان، عدم دسترسی به نسخه جدید نرم افزار استاتا به دلیل تحریم (Sanction)، طولانی شدن تدوین نسخه جدید کتاب و رخداد پاندمی کووید-۱۹، توفیق چاپ مجدد حاصل نشد. خوشبختانه بعد از چاپ نخست کتاب تا کنون، حداقل ۷ بار محتوی کتاب برای دروس کارشناسی ارشد رشته اپیدمیولوژی و به ویژه در درس روش های اپیدمیولوژی و درس تحلیل داده های سلامت در ترم های مختلف تدریس و اشکالات آن برطرف گردید.

با به بازار آمدن نسخه Stata ۱۸ و دسترسی به آخرین نسخه، مولفین تصمیم گرفتند بازنویسی کتاب را انجام و در راستای "**مسئولیت پذیری و پاسخگویی اجتماعی و حرفه ای به نیازهای جامعه**"، حمایت از علاقمندان استفاده از کتاب در جامعه و به ویژه دانشجویان گرامی و برای توسعه آموزش های اپیدمیولوژی مجازی، فایل PDF آن را به صورت **رایگان** و **تنها با خواندن سه صلوات برای شادی روح درگذشتگان مولفین**، برای دسترسی آسان و استفاده دانشجویان و پژوهشگران علاقمند به توسعه پژوهش های با کیفیت و دانش اپیدمیولوژی در ایران در دسترس قرار دهند.

در این نسخه از کتاب همه فصل های قبلی بازنویسی، ویرایش و مطالب جدیدتری به همه فصل ها اضافه شد. علاوه بر استفاده از فایل داده های نرم افزار، مثال هایی از کار با داده های واقعی در کتاب گنجانده شده است. فایل داده های مورد استفاده در این کتاب را از لینک <https://mhrc.skums.ac.ir/page-rcmh/fa/55/form/pId13470> و نسخه ۱۷ و ۱۸ نرم افزار را از سایت های ایرانی و خارجی می توان دانلود، نصب و استفاده نمود.

از استاد ارجمند جناب آقای دکتر عباسعلی کشتکار بخاطر معرفی کتاب به دانشجویان دکتری و استقبال و استفاده شرکت کنندگان در کارگاه های برگزار شده توسط ایشان و همچنین از خانم دکتر مرضیه روحانی رصاف، خانم جمیله معارفی و خانم فاطمه زینب کیانی بخاطر همکاری در تهیه برخی پیش نویس های متون، از **دانشگاه های علوم پزشکی شهرکرد** و **دانشگاه علوم پزشکی خراسان شمالی** بخاطر حمایت از چاپ قبلی کتاب و انتشارات الیاس تقدیر و تشکر می شود. کتاب در ۶۵۲ صفحه A4 و ۱۴۲۴۴۶ واژه بدون محاسبه شکل ها و خروجی های متعدد نرم افزار تدوین گردید. امید است که مورد رضایت خوانندگان قرار گیرد. بی شک نظرات، پیشنهادات و انتقادات خوانندگان محترم در رفع نقایص احتمالی و بهبود کتاب در چاپ های آتی کمک کننده خواهد بود. نظرات خویش را به آدرس [aliahmadi2007@gmail.com](mailto:aliahmadi2007@gmail.com) ارسال فرمایید.

و ماتوفیقی الا بالله علیه توکلت والیه انیب

دکتر علی احمدی – دانشیار اپیدمیولوژی عضو هیات علمی دانشگاه

حسین لشکردوست – عضو هیات علمی دانشگاه

## بخش اول :

# اصول و مبانی کار با نرم افزار Stata

## فصل اول: آشنایی با نرم افزار.....

راه اندازی Stata  
آشنایی عمومی با محیط نرم افزار، Menu ها و کادر محاوره ای یا پنجره گفتگو (Dialog box)  
معرفی پنجره های نرم افزار  
باز کردن فایل، ذخیره کردن فایل  
آشنایی با عملکرد برخی کلیدهای صفحه کلید در محیط Stata  
ثبت و ذخیره کارهای انجام شده در نرم افزار  
تمرین ها

## اهداف:

### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- آشنایی کامل با انواع مدل های نرم افزار Stata
- شناخت پنجره های نرم افزار Stata، اجرا و شناخت کارکرد ابزارهای نرم افزار
- انجام تنظیمات مربوط به نرم افزار متناسب با الگوی دلخواه
- شناخت عملکردهای صفحه کلید در نرم افزار و تعریف کلیدهای F برای عملکرد خاص
- یادگیری و کاربرد دستورات پایه در نرم افزار و استفاده از پنجره گفتگو (Dialog box)
- ذخیره سازی کارهای انجام شده با نرم افزار

## شروع کار با نرم افزار و انواع Stata :

نرم افزار Stata برگرفته از **Statistics and Data Science** به عنوان یکی از کامل ترین و از جمله مهم ترین نرم افزارهای مورد استفاده در تحقیقات علوم پزشکی است. این نرم افزار مورد توجه ویژه اپیدمیولوژیست ها و آمارشناسان و سایر محققان قرار گرفته است. این نرم افزار علاوه بر داشتن امکانات گسترده و قوی برای تحلیل داده ها، امکان برنامه نویسی را به کاربران میدهد. این نرم افزار آماری از سال ۱۹۸۴ تا کنون به ویرایش ۱۸ رسیده است. در این نسخه از کتاب، مثال ها با نسخه ۱۷ و ۱۸ انجام گردید. خرید نسخه اصلی این نرم افزار از آدرس [www.stata.com](http://www.stata.com) امکانپذیر است. بیش از ۱۷۰۰۰ صفحه محتوی به عنوان راهنمای نرم افزار توسط کمپانی Stata در دسترس استفاده کنندگان قرار دارد. قیمت نسخه اصلی این نرم افزار ۴۸۰۰ دلار است. در ایران از آدرس سایت دانلودلی میتوان نرم افزار را تهیه و نصب نمود.

کسب اطلاعات بیشتر از این نرم افزار و مقایسه ورژن های مختلف و آخرین تغییرات در وب سایت این نرم افزار در [www.Stata-journal.com](http://www.Stata-journal.com) آدرس اینترنتی [www.Stata.com](http://www.Stata.com) و ژورنال Stata به آدرس [www.Stata-journal.com](http://www.Stata-journal.com) قابل دسترس است. برنامه Stata مخفف (Abrivation) عبارتی نیست و لذا نوشتن آن با حروف بزرگ انگلیسی صحیح نیست. ویرایش های اولیه نرم افزار مبتنی بر دستور بوده و قابلیت استفاده با منو را نداشت. از ویرایش هشتم به بعد منوها به آن اضافه گردیده و قابلیت کار با منوها و استفاده از نرم افزار برای کاربران راحت تر گردید. نسخه های مختلفی از نرم افزار Stata وجود دارد. این نرم افزار به صورت متفاوت و در قالب مدل های زیر در بازار وجود دارد.

### ۱- مدل Stata/MP :

این مدل از نرم افزار معروف به Multiprocessor، پیشرفته ترین و پر سرعت ترین ویرایش Stata است. بر روی کامپیوترها و سرورهای بزرگ و ملی که دارای پردازنده های متعدد هستند استفاده می شود. قابلیت پشتیبانی از ۱۲۰۰۰۰ متغیر، ۶۵۵۳۲ متغیر مستقل و بی نهایت مشاهده را تا بیشتر از بیست بلیون داراست. متغیرهای اسمی را تا ۲۴۴ حرف و ماتریس ۱۱۰۰۰ در ۱۱۰۰۰ را پشتیبانی می کند.

### ۲- مدل Standard Edition Stata/SE :

نسخه استاندارد و حرفه ای Stata و پرسرعت ترین ویرایش نرم افزار است. این نسخه را می توان با حداکثر سرعت، بر روی کامپیوترهای معمولی نصب نمود. توانایی کار با حداکثر ۳۲۷۶۷ متغیر، ۱۰۹۹۸ متغیر مستقل را دارد و تنها چیزی که آن را محدود می کند محدودیت مشاهده، حافظه (رایانه) کامپیوتر می باشد. متغیرهای اسمی (string) را تا محدوده ۲۴۴ کاراکتر می پذیرد و ماتریس ها را تا ابعاد ۱۱۰۰۰×۱۱۰۰۰ پشتیبانی و آنالیز می کند.

### ۳- مدل Basic Edition Stata/BE :

این ویرایش نیز مانند دو مدل بالا است. تنها سرعت آن کمتر بوده و قابلیت پشتیبانی از ۲۰۴۸ متغیر، ۷۹۸ متغیر مستقل و بی نهایت مشاهده را دارد. تنها حافظه کامپیوتر می تواند محدودکننده قابلیت مشاهده آنها باشد. متغیرهای اسمی را تا حداکثر ۲۴۴ کاراکتر و ماتریس های ۸۰۰ در ۸۰۰ را پشتیبانی می کند. نسخه حرفه ای و کوچک (Small) دیگری از Stata به عنوان مدل چهارم وجود دارد که معمولاً ارزان تر و در دسترس تر است. سرعت مناسبی دارد اما سرعت آن از مدل های ذکر شده در بالا، کمتر است. توانایی پذیرش حداکثر ۹۹ متغیر و قابلیت پشتیبانی حداکثر ۱۰۰۰ مشاهده را دارد. متغیرهای string را تا ۲۰۰ کاراکتر می پذیرد و برای ماتریس تا اندازه ۴۰×۴۰ قابل استفاده است. معمولاً در کارگاه های آموزش Stata و کلاس های درسی از آن استفاده میشود.

### چه نسخه یا مدلی از نرم افزار Stata برای نصب مناسب است؟

ابتدا به کلید مجوز نرم افزار در دسترس خویش نگاه کنید. در صورت دارا بودن مجوز استفاده از Stata/SE، این نسخه را نصب نمائید. در صورت دارا بودن مجوز استفاده از ( Intercooled Stata:IC )، آنرا نصب نمائید و توجه داشته باشید که مجوز شما اجازه استفاده از Stata/SE را به شما نخواهد داد.

در صورت دارا بودن مجوز Small Stata، همین نسخه را نصب کنید و توجه داشته باشید که شما حق استفاده از دو نسخه دیگر را ندارید. در صورتیکه شما دارای مجوز کاربری - انحصاری (single-user) هستید، می توانید Stata را در هر دو کامپیوتر محل کار و منزل خویش، نصب کنید، اما امکان استفاده همزمان آن را روی دو کامپیوتر ندارید. در ایران معمولا نرم افزارها به صورت قفل شکسته و غیر قابل به روز رسانی در دسترس هستند. اینک در ایران، نسخه های ۱۷ و ۱۸ نرم افزار استاتا که غیرقابل بروز رسانی است در دسترس است.

### نصب Stata روی سیستم عامل Windows:

Stata در تمام نسخه های مدرن ( windows 95, 98, XP, windows 7 , windows 8 و بالاتر) قابل نصب و استفاده است. همچنین روی نسخه های Macintosh و لینوکس نیز قابل نصب و استفاده می باشد. بعضی از ویرایش های این نرم افزار از جمله ۱۱ و ۱۲ - بر خلاف ویرایش های قبلی که حتما ابتدا می بایست آنها را بر سیستم کامپیوتر نصب نمود - نیاز به نصب ندارند. برای استفاده از این نرم افزارها کافی است که منبع آن را در حافظه کامپیوتر یا فلش کپی نمود و از آنجا با دو بار کلیک کردن بر روی فایل اجرایی آن با پسوند (.exe) آن را اجرا و استفاده نمود. برای در دسترس قرار گرفتن نرم افزار میتوان با راست کلیک کردن بر آن و فرستادن آن به دسک تاپ همیشه آن را در دسترس داشت. یا با انتخاب گزینه Pin to Startt Menu برنامه را در لیست برنامه های قابل مشاهده به منوی Start اضافه و استفاده نمود.

### قبل از نصب نرم افزار به نکات زیر توجه باید کرد:

- ۱- از داشتن CD نصب نرم افزار Stata اطمینان داشته باشید. یعنی برنامه را در دسترس داشته باشید.
- ۲- از داشتن کلید مجوز استفاده از برنامه اطمینان حاصل کنید.
- ۳- معمولا در ایران نرم افزارهای Stata که در دسترس قرار دارند، کلید مجوز ندارند و به عبارتی قفل شکسته هستند. که وقتی از آنها استفاده می شود ممکن است در بعضی از دستورات کار نکنند. لذا باید به این مهم توجه داشت. اما خوشبختانه نرم افزارهای در دسترس در اکثریت موارد با نصب فایل کراک شده، پاسخگوی نیاز پژوهشگران و کاربران هست.
- ۴- نسخه مورد نظر خود را انتخاب کنید.
- ۵- فضای نصب نرم افزار Stata را انتخاب نمائید. پیشنهاد می شود برنامه را در درایو یا فضای C نصب کنید. C:\Stata 18 یا C:\Stata
- ۶- فضای کاری نرم افزاری خود را در کامپیوتر مشخص کنید. این فضا باید از محل نصب نرم افزار متفاوت باشد تا فایل های ایجاد شده با فایل های مربوط به خود نرم افزار Stata اشتباه و مخلوط نگردد. پس از انتخاب فضای کاری مورد نظر، برنامه Stata روی hard disk کپی کنید. برای اطمینان از نصب درست برنامه، پس از نصب با استفاده از دستور verinst که در قسمت دستورات باید آن را تایپ و تایید (Enter) نمود، میتوان به صحت یا عدم صحت نصب نرم افزار پی برد.

۷- وقتی دستور `verinst` نوشته شود، پیغام زیر در صفحه خروجی نتایج به نمایش در می آید.

Stata/MP 18.0 for Windows (64-bit x86-64)  
Revision 25 Apr 2023  
Copyright 1985-2023 StataCorp LLC

Total physical memory: 32.00 GB  
Available physical memory: 23.32 GB

Stata license: Single-user 1-core perpetual

در خروجی بالا، ورژن یا نسخه (نوع ویرایش) نرم افزار، تاریخ دریافت و کپی، ظرفیت حافظه مورد نیاز و در دسترس و مجوز نرم افزار نشان داده شده است. برای به روز رسانی نرم افزار از دستور `update all` و همچنین برای دریافت دستورات از فرمان `findit command` استفاده می شود. با این دستور می توان دستورات مورد نیاز را از اینترنت جستجو و نصب نمود. مثال `findit meta`

چنانچه محقق قصد دارد با مجموعه داده های بزرگ کار کند طبیعتاً به حافظه بیشتری نیاز دارد. نرم افزار Stata داده های شما را قبل از آنالیز، روی حافظه قرار می دهد. به طور قراردادی `Stata/SE` حدود ۱۰ مگا بایت، `Intercooled Stata` حدود ۱ مگا بایت و `Small Stata` حدود ۳۰۰ کیلو بایت حافظه را برای داده های خود اشغال می کند. حافظه نسخه `Small Stata` ثابت می باشد، بنابراین احتمال پاک شدن بقیه صفحه که در ظرفیت حافظه نمی گنجد وجود دارد. اما در نسخه های `Stata/SE` و `Intercooled Stata` باتوجه به حجم داده ها (dataset)، شما می توانید مقدار حافظه اختصاص یافته جهت داده ها را به دلخواه تعیین کنید. پس از آغاز اجرای Stata، امکان تغییر موقت مقدار حافظه کاری Stata وجود دارد. برای افزایش حافظه می توان از دستورات زیر استفاده کرد:

`Set mem 50m`

`set mem 200m`

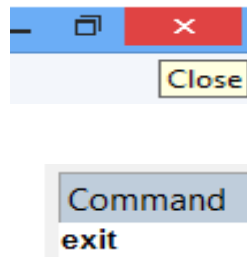
عدد بعد از فرمان `mem` مقدار فضای اختصاص یافته را نشان می دهد. مثلاً اختصاص ۲۰۰ مگابایت حافظه در فرمان ذکر شده مد نظر کاربر است. معمولاً حافظه نرم افزار بطور پیش فرض در 1m تنظیم شده است. که در برخی موارد کافی نیست. و لذا باید آن را تغییر داد. دستورات را می توان بصورت دستی (که مورد توجه کاربران پیشرفته می باشد) و یا به طریق `Point-and-Click` (نشان بده و کلیک کن) یا از طریق باز کردن کشویی (pull down) منوی Stata و انتخاب و احضار پنجره محاوره دستورات Stata، انجام داد. دستور یا فرمان `memory` وقتی نوشته شود حافظه مورد استفاده و در دسترس در کامپیوتر کاربر نمایش داده می شود. به عنوان مثال حافظه مورد استفاده سیستم مولف به شرح زیر نمایش داده می شود.

Memory usage	used	allocated
data (incl. buffers)	1,537	33,554,432
var. names, %fmts, ...	2,642	26,334
overhead	1,064,964	1,065,360
Stata matrices	0	0
ado-files	4,553	4,553
saved results	0	0
Mata matrices	0	0
Mata functions	0	0
set maxvar usage	1,391,728	1,391,728
other	3,566	3,566
total	2,465,606	36,045,973

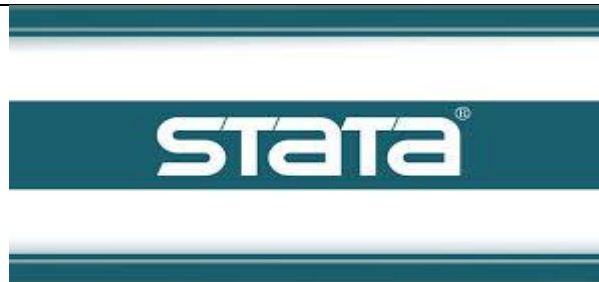


در صورتیکه Stata صحت نصب نرم افزار را اعلام کرد، برای خروج از نرم افزار، یکی از مراحل زیر را انجام دهید:

- ۱- در پنجره دستور **exit** را تایپ کنید و کلید **Enter** را فشار دهید. شما می‌توانید در صورت ذخیره کردن تغییرات بوجود آمده در داده‌های خود دستور **clear** و **exit** را تایپ کنید.
- ۲- از منوی **File**، گزینه **Exit** را انتخاب کنید.
- ۳- روی **Close box** (نشانه موجود در گوشه راست فوقانی پنجره Stata) کلیک کنید.



قبل از خارج شدن یا پس از خارج شدن در صورتی که از دستور **pwd** استفاده شود مسیر نصب نرم افزار نشان داده می‌شود. بنابراین برای اطلاع از این موضوع از دستور مذکور استفاده می‌شود.



## آشنایی عمومی با محیط Stata

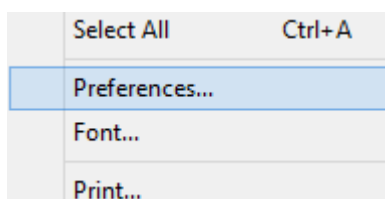
معرفی پنجره‌های Stata : صفحه‌ی اصلی Stata شامل چهار پنجره به شرح زیر است. محیط نرم افزار Stata و پنجره های آن در شکل زیر نمایش داده شده است.

- **Command** این پنجره به صورت یک نوار باریک در پایین صفحه قرار دارد. فرمان‌های Stata در این پنجره نوشته می‌شود.
- **Results** این پنجره بزرگترین پنجره‌ی محیط Stata است. خروجی فرمان‌های Stata در این پنجره نمایش داده می‌شود. در وسط صفحه بوده و نتایج را نشان می‌دهد.
- **History** این پنجره در سمت چپ، در بالای صفحه قرار دارد. هر فرمان پس از اجرا به این پنجره منتقل می‌شود. می‌توان برای اجرای مجدد یک فرمان به جای تایپ مجدد، آن را از پنجره‌ی History یا در نسخه‌های قبلی در پنجره Review فرا خواند. دستوراتی که اجرا نشده اند به رنگ قرمز مشاهده می‌شوند. (ممکن است علت آن اشتباه بودن دستور باشد).
- **Variables** این پنجره در سمت راست، در کنار صفحه قرار دارد و در آن نام متغیرهای مجموعه داده‌های فعال نوشته می‌شود.



کاربر می‌تواند با استفاده از نوار منوها در منوی Windows مجموعه پنجره‌های نرم افزار که ده پنجره هستند را مشاهده و کلیک نموده و پنجره‌های مورد درخواست را باز و یا بسته نگه دارد. کاربر می‌تواند شکل نرم افزار و رنگ آن را مطابق میل خود تنظیم کند. برای این کار می‌توان با راست کلیک کردن در صفحه وسط نرم افزار، تنظیمات مربوطه را انجام داد. شکل بالا نوع کلاسیک نرم افزار است.

وقتی در صفحه وسط نرم افزار (پنجره نتایج) راست کلیک کنید منوی کشویی زیر باز می شود. که کاربر می تواند ترجیحات، نوع و اندازه فونت و نوع چاپ را تنظیم نماید. رنگ قلم در پنجره های دستورات، مرور و متغیرها سیاه رنگ است. دستوراتی که اشتباه نوشته شود به رنگ قرمز نمایش داده می شوند. در قسمت نتایج رنگ فونت فرمان ها به رنگ سفید، عناوین جداول به رنگ سبز، نتایج به رنگ زرد و پیغام های خطا به رنگ قرمز نشان داده می شوند. به عنوان مثال با انتخاب Preferences در شکل زیر میتوان به زمینه دلخواه، نوع رنگ خروجی ها و فرمان ها را تنظیم نمود.



پس از تنظیم اگر فایل برنامه را ببینید و مجدداً آن را باز نمایید، برنامه مطابق تغییرات داده شده باز می شود. در زیر پنجره های نرم افزار به صورت جداگانه نمایش داده می شوند. پنجره نمایش متغیرها که در ورژن ۱۷ بر خلاف بقیه ورژن های قدیمی در سمت راست قرار دارد به صورت زیر است. با دکمه سمت چپ پنجره می توان آن را پنهان یا آشکار نمود.

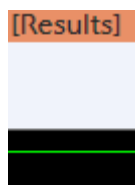
Variables	
Variable	Label
id	Subject identifier
time	Follow-up time in ...
d	Outcome (0=alive...

پنجره نمایش دستورات اجرا شده یا مرورگر در قسمت History یا Review به صورت زیر است. این پنجره در سمت چپ نرم افزار قرار دارد.

Review		
#	Command	_rc
1	use "C:\Users\ali\Desktop...	
2	des	

پنجره دستورات یا فرمان ها به صورت زیر است. دستور در پایین آن تایپ و با تایید (Enter) اجرا می شود.

```
Command
sum|
```

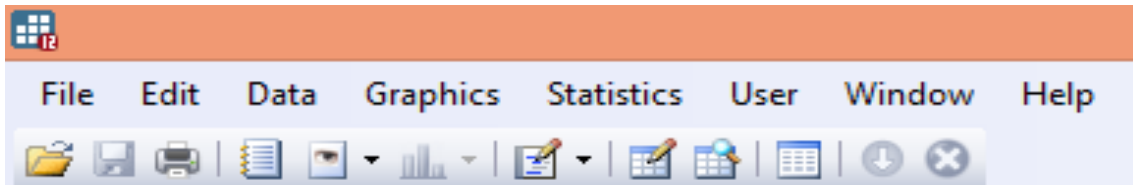


نتایج اجرای دستورات نیز در قسمت مرکزی نرم افزار نشان داده می شود.

## آشنایی اختصاصی با محیط نرم افزار، Menu ها و Dialog box

### نوار ابزار Stata



نوار ابزار Stata شامل کلیدهایی است که دسترسی سریع را به بخش‌های متداول تر Stata امکان‌پذیر می‌سازد. در صورتیکه شما عملکرد دکمه‌ای را به خاطر نمی‌آورید، نشانگر خود را روی آن نگه دارید تا توضیح عملکرد آن ظاهر شود. این ابزارها در شکل زیر نشان داده شده است.



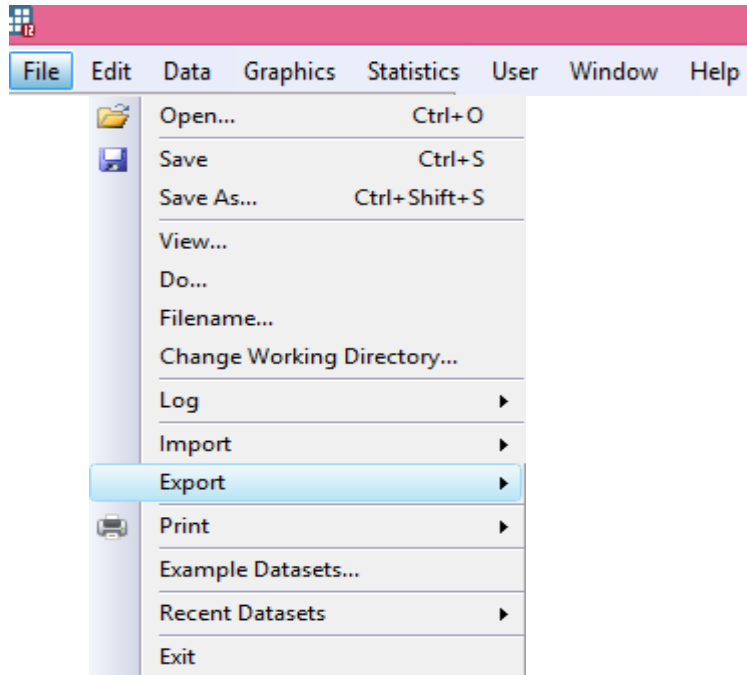
چنانچه نشانگر را به ترتیب روی ابزارها نگه دارید، ابزارهای زیر نمایش داده می‌شوند. این ابزارها معمولاً در همه نرم افزارها وجود دارند و کاربران کامپیوتر با آنها آشنایی دارند.



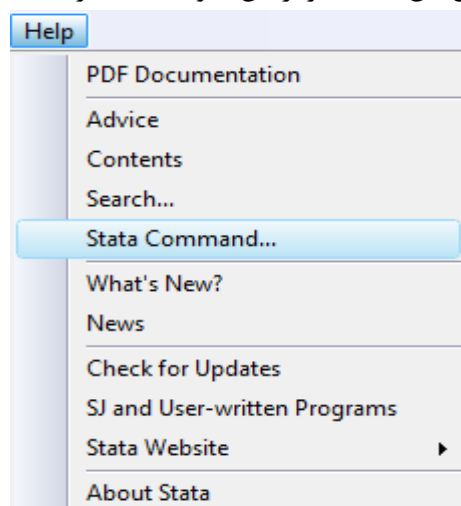
### توضیحات ابزارهای صفحه اصلی نرم افزار:

- **Open** : برای باز کردن فایل داده کاربرد دارد.
- **Save** : برای ذخیره کردن فایل استفاده می‌شود.
- **Print** : برای پرینت کردن و چاپ نمودن خروجی کاربرد دارد.
- **Log Begin/close/suspend** : شروع یک log جدید، رجوع به log موجود، بستن یا معلق نگه داشتن log فعلی. در واقع کارهای انجام شده را می‌توان ذخیره نمود.
- **New Viewer** : باز کردن یا جلو آوردن Viewer روی بقیه پنجره های Stata
- **Bring Graph window to Front** : آوردن پنجره نمودارها در جلوی سایر پنجره
- **New Do-file Editor** : باز کردن ویرایشگر Do-file. در فصل دوم به آن پرداخته می‌شود. این ابزار کارایی فوق العاده ای در ذخیره دستورات و اجرای مکرر آنها دارد.
- **Data Editor** : باز کردن ویراستار داده ها یا جلو آوردن آن روی سایر صفحات
- **Data Browser** : باز کردن یا جلو آوردن جستجو گر داده ها
- **Variable Manager** : مدیریت و تغییر مورد نیاز را برای متغیرها می‌توان انجام داد.
- **Clear-More-Condition** : برای مشاهده قسمت های بعدی نتایج، زمانی که خروجی نتایج ما طولانی است و در یک صفحه ظاهر نمی‌شود.
- **Stop** : برای توقف دستور فعلی در حال اجرا در Stata کاربرد دارد. منوهای Statistics, Data, Graphics در Stata امکان دستیابی و استفاده از روش Point-and-Click را در مورد بسیاری از دستورات Stata فراهم می‌کند. با این منوها و مخصوصاً منوی Statistics میتوان تحلیل های آماری

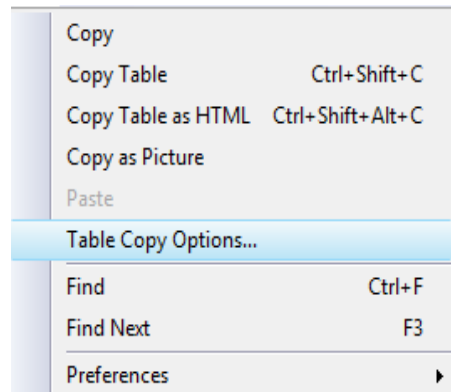
مورد نیاز را انجام داد. برای آشنایی با این منوها به تصاویر زیر خوب دقت فرمایید. در تصویر زیر وقتی منوی فایل را باز کنید می توانید اقدامات عمومی مانند باز و بسته کردن فایل، ذخیره کردن، انتقال داده ها، چاپ، تغییر محل ذخیره داده ها، مشاهده داده هایی که اخیرا باز شده اند و مثال های فرضی نرم افزار را مشاهده نمایید. در طی این کتاب به هر کدام از این موارد مبسوط توضیح داده می شود.



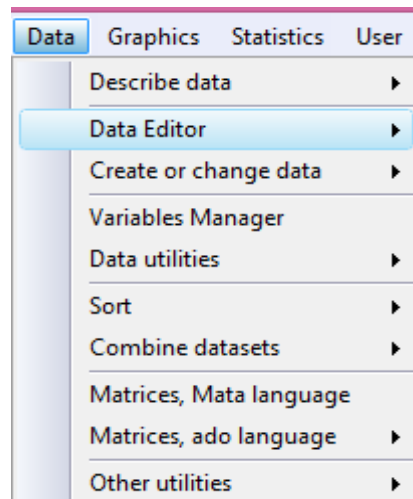
سیستم Help نرم افزار یکی از ویژگی های منحصر به فرد و با اهمیت نرم افزار است. با کلیک بر دکمه Help در منوی اصلی، فهرست گزینه های آن باز می شود و بر حسب مورد می توان از آن گزینه ها کمک گرفت. گزینه ها شامل فایل Pdf راهنمای نرم افزار، جستجو، یافتن فرمان مورد نظر، توصیه ها، محتویات و درباره نرم افزار و وب سایت آن است. در نگارش این کتاب نیز از این منو استفاده گردید.



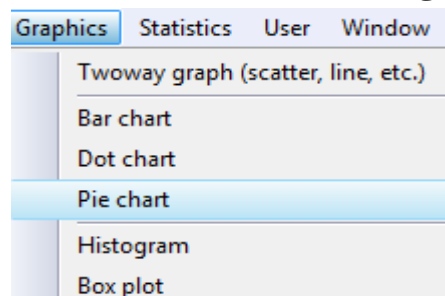
وقتی منوی دوم یعنی Edit را باز کنید. مجموعه کارهای عمومی از جمله کپی کردن را می توان انجام داد. این کارها و زیر منوی Edit در شکل زیر آمده است.



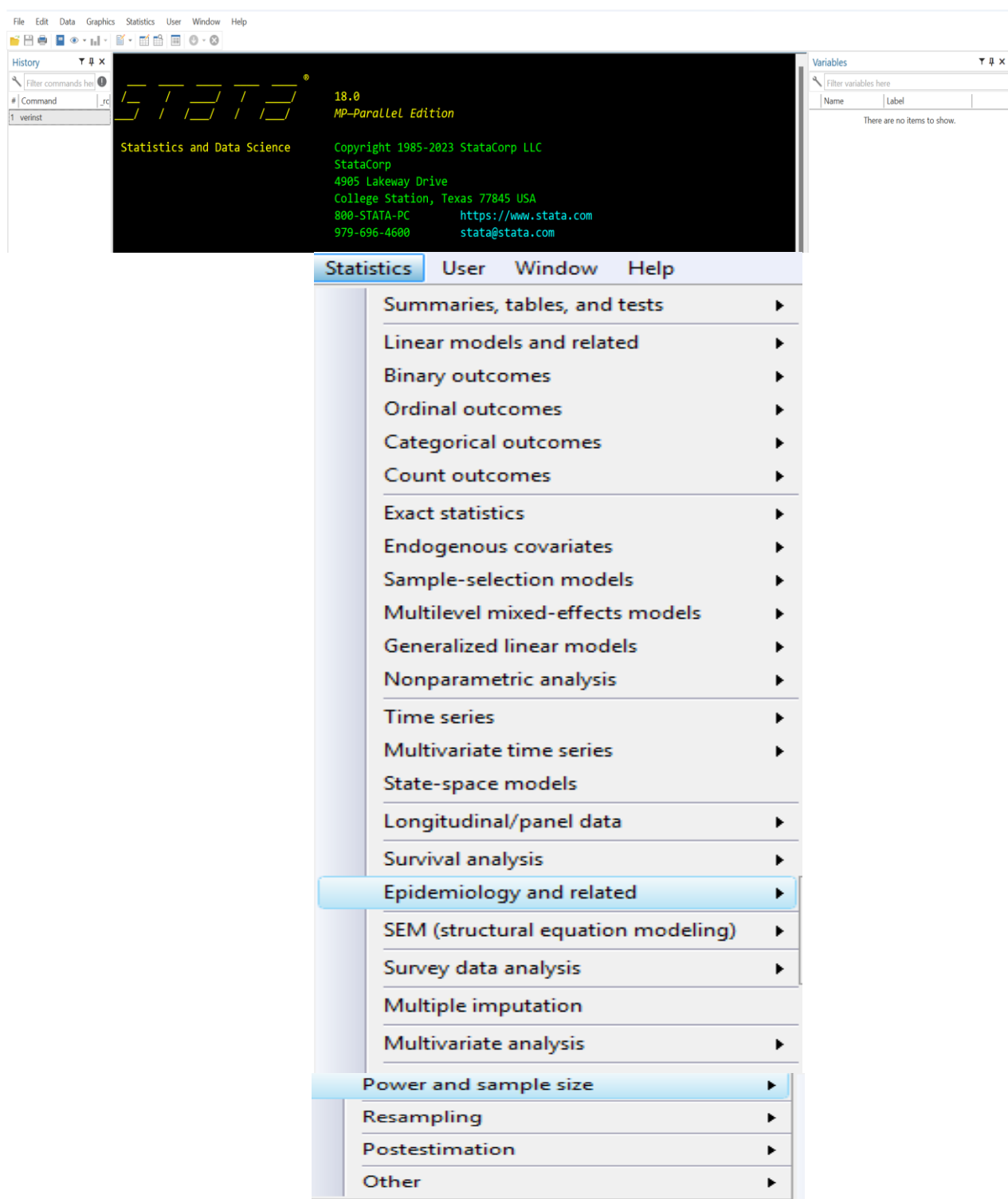
با باز کردن منوی Data می توان مدیریت داده ها و متغیرها را به نحو مورد نظر انجام داد. زیر منوی Data در شکل زیر آمده است. توصیف و مشاهده داده ها، ویرایش داده ها، ساخت متغیر داده های جدید، ویرایش متغیرها، مدیریت متغیرها و تنظیمات مربوط به داده ها و ترکیب کردن داده ها را می توان انجام داد.



همچنین می توان با باز نمودن منوی Graphics انواع مختلف نمودارها و از جمله منحنی ROC و منحنی های مربوط به تحلیل بقا از جمله منحنی کاپلان مایر را ترسیم نمود. این اقدامات در قسمت های مربوطه در فصول بعدی مبسوط توضیح داده می شوند.



مهمترین قسمت نرم افزار Stata برای آنالیز داده ها در شکل زیر نمایش داده شده است. که بسته به نوع آنالیز مورد نیاز کاربر می تواند از هر کدام از آنها استفاده نماید.



همانگونه که در تصویر بالا مشخص است یکی از گزینه های مهم **Epidemiology and related** است. این گزینه، زیر منوهای زیادی دارد که در بخش های مختلف کتاب، درباره کاربرد آنها، توضیحات مورد نیاز داده می شود. **در واقع این گزینه به جعبه ابزار اپیدمیولوژیست ها معروف است.**

## آشنایی با عملکرد برخی کلیدهای صفحه کلید در محیط Stata و نوشتن فرمان ها

در محیط Stata عملکرد برخی کلیدهای صفحه کلید، متفاوت است که در زیر به آن ها اشاره می شود:

خط فرمان را به طور کامل پاک می کند.	<b>Esc</b>
فرمان قبلی را از پنجره ی History به خط فرمان فرا می خواند.	<b>Pg Up</b>
فرمانی که در پنجره ی History بعد از فرمان فعلی قرار دارد، به خط فرمان فرا می خواند.	<b>Pg Dn</b>
اگر خروجی یک فرمان بیش از یک صفحه باشد، برای نمایش صفحات بعدی به کار می رود.	<b>Space</b>

**کلیدهای F** برخی از کلیدهای عملکردی دارای معانی مشخصی هستند. F1 راهنمای برنامه و F2 برای توصیف فایل ها است.

F3: ایجاد نوار جستجو در زیر پنجره نتایج برای جستجو کردن و تایپ عبارت مورد نظر در قسمت find و فشار دادن کلید Enter است. F7: معادل تایپ Save است و پس از کلیک روی آن می توانید نام فایل مربوطه را که مایل به ذخیره آن هستید وارد نمایید. توجه نمایید که معانی این کلیدها در نسخه ها و ویرایش های مختلف Stata (اعم از macintosh, windows) یکسان است. برای مشاهده تعاریف قراردادی کلیدها، عبارت دستور macro list را اگر وارد نمایید تعاریف زیر را برای کلیدهای F ملاحظه می کنید.

```
. macro list
S_FNDATE:      10 Aug 2013 23:47
S_FN:          C:\Users\ali\Desktop\data book stata\cox regression.DTA
S_level:       95
F1:            help advice;
F2:            describe;
F7:            save
F8:            use
S_ADO:         UPDATES;BASE;SITE;.;PERSONAL;PLUS;OLDPLACE
S_FLAVOR:      Intercooled
S_OS:          Windows
S_MACH:        PC
```

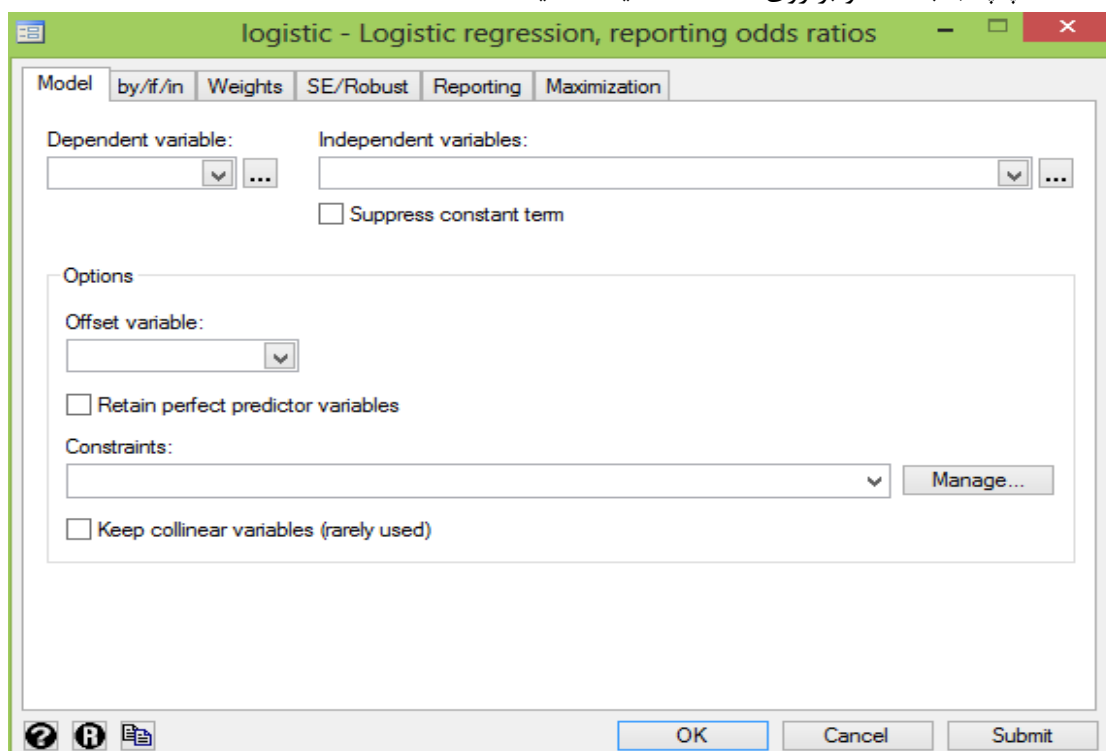
### نگارش فرمان ها یا دستورات در Stata

فرمان های Stata در خط فرمان یا دستور (پنجره ی Command) نوشته می شوند. در موقع نوشتن فرمان ها و نام متغیرها لازم است به این نکته توجه شود که نرم افزار Stata نسبت به کوچک و بزرگ بودن حروف حساس است. در طی این کتاب هر جا از حروف بزرگ در دستورات استفاده شده صحیح نبوده و تنها باید با حروف کوچک دستورات را بکار برد. برای اجرای برخی از فرمان های Stata لازم نیست نام فرمان به طور کامل نوشته شود و نوشتن چند حرف از حروف فرمان کفایت می کند، برای مثال می توان فرمان describe را صورت des نوشت. فرمان های قابل استفاده در نرم افزار را میتوان از طریق منوها و یا پنجره گفتگو و یا کادر محاوره ای که در این کتاب به عنوان کادر محاوره ای (**dialog box**) یا به اختصار db ذکر می شود بدست آورد و سپس دستور را اجرا نمود. یعنی پس از تایپ دستور db، کاری که می خواهید انجام دهید را تایپ کنید کادر محاوره ای نمایش داده می شود. به عنوان مثال چنانچه محققى بخواهد رگرسیون لجستیک انجام دهد و در قسمت فرمان تایپ نماید:

```
db logistic
```



کادر محاوره ای زیر باز می شود که بدون دانستن فرمان، میتوان با وارد کردن متغیرها و OK کردن، تحلیل را با استفاده از منو و یا کادر محاوره ای انجام داد. برای استفاده از راهنمای نرم افزار و به ویژه دستور مورد نیاز از گوشه سمت چپ جعبه گفتگو بر روی علامت **?** کلیک نمائید.



ساده ترین شکل نگارش فرمان های Stata به صورت زیر است:

[گزینه ها], نام متغیر(ها) نام فرمان

شکل کامل تر دستور بالا به شرح زیر است:

[by نام متغیر(ها)]:

[گزینه ها], [نام فایل using] [weight] [دامنه in] [شرط if] [exp] (نام متغیرها) نام فرمان

که در آن:

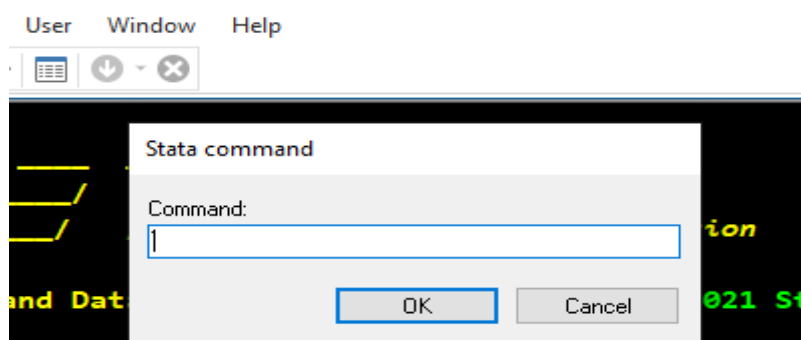
- by
- exp
- if
- in
- weight
- using
- گزینه ها

فرمان را برای زیرمجموعه هایی از مجموعه داده ها، به طور جداگانه اجرا می کند. عبارات جبری یا نویسه ای است که در برخی فرمان های Stata نوشته می شوند. فرمان را فقط برای مواردی که توسط شرط تعیین شده است، اجرا می کند. فرمان را فقط برای مواردی که در داخل دامنه ی معین قرار دارند، اجرا می کند. مقادیر مشاهدات را بر اساس متغیر تعیین شده، وزن دهی می کند.

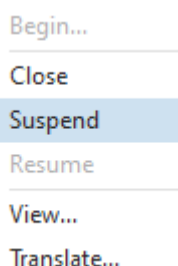
فقط در برخی فرمان های Stata به کار می رود و فایلی را که مجموعه داده ها در آن قرار دارد، معرفی می کند.

در هر یک از فرمان های Stata، برای کنترل بیشتر و اعمال تغییرات جزئی در عملکرد فرمان به کار می روند. توضیحات بیشتر درباره دستورات از طریق منوی Help برنامه در

دسترس است. دستور separator(10) را می توان در گزینه ها برای نمایش لیست ده تایی متغیر استفاده نمود. توضیحات بیشتر درباره دستورات از طریق منوی Help نرم افزار و تایپ دستور در جعبه زیر در برنامه در دسترس است.



**ثبت کارهای انجام شده:** تعریف و پیش فرض نسخه ۱۷ نرم افزار برای ذخیره کارهای انجام شده به صورت SMCL است و به صورت متنی یا (text) نیست. از منوی فایل، میتوان با انتخاب log مسیر ذخیره کارهای انجام شده را با انتخاب دکمه Begin معرفی نمود. سپس میتوان ذخیره را به صورت موقتی متوقف (suspend) یا مجدد با Resume شروع و یا کامل با close متوقف نمود. مسیر دستورات مذکور از قسمت فایل و با انتخاب log در زیر قابل مشاهده است:



• **ذخیره به صورت فایل متنی در نسخه ۱۶ نرم افزار:**

می توان فرمان ها و خروجی آن ها را در یک فایل متن ذخیره نمود، به این منظور باید پیش از نوشتن فرمان های مورد نظر، فرمان log using filename log, text نوشته و اجرا شود. و در پایان، فایل مورد نظر با استفاده از فرمان (log close) بسته شود. اگر کاربر مایل باشد فقط فرمان ها را در یک فایل متن ذخیره کند، باید به جای استفاده از واژه ی log، از cmdlog استفاده کند. در این صورت دیگر نیاز به استفاده از گزینه ی text نیست زیرا cmdlog به طور پیش فرض یک فایل متن می سازد.

• **باز کردن فایل، ذخیره کردن فایل**

اگر می خواهید فایل جدیدی باز کنید از قسمت File/ Open فایل مورد نظر خود را انتخاب و باز کنید.

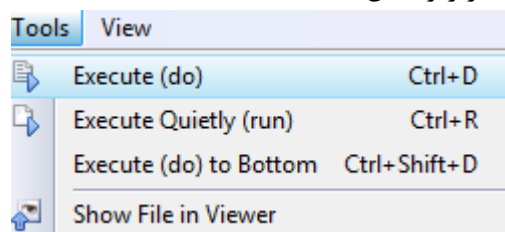
• **ذخیره نتایج**

در صورتی که بخواهید نتایج آنالیز را در Result window ذخیره کنید باید آن را در یک Log file ذخیره کنید.  
File/ Log/ Begin

در پنجره باز شده نام فایل خود و محلی که می خواهید آن را ذخیره کنید مشخص نمایید و بر روی OK کلیک کنید. در پنجره Results پیام باز شدن این فایل را می بینید. برای متوقف کردن موقت فایل Log، بستن آن، اضافه کردن به Log قبلی و مشاهده آن از همین مسیر استفاده کنید. فایل Log به دو فرمت smcl و log ذخیره می شود. در فرمت smcl (Stata Markup and Control Language) فونت و رنگ حفظ می شود ولی برای باز کردن مجدد این فرمت تنها توسط نرم افزار Stata امکان پذیر است ولی فرمت log قابل باز کردن و کپی نمودن در سایر نرم افزارها نظیر Word می باشد.

### • ذخیره دستورات

در صورتی که بخواهید دستورات خود را ذخیره کنید باید آن را در یک Do file ذخیره نمایید. برای این کار می توانید دستورات مورد نظر خود را در پنجره history یا Review Window انتخاب کنید. سپس با کلیک راست کردن و انتخاب گزینه Send to do file editor و save دستور اجرا می شود. دستورات شما در یک Do file با اسم مشخص که باید نامگذاری شود، ذخیره خواهد شد. می توانید با کلیک کردن بر روی New do file Editor در نوار ابزار، یک Do file جدید تعریف نمایید و یا Do file های قبلی را باز نمایید. می توانید دستورات خود را در Do file ویرایش نمایید. مهمترین کاربرد Do file امکان ثبت دستورات و اجرا نمودن مجدد آن ها است. برای اجرای دستورات در قسمت Tool از گزینه Do و یا Run استفاده کنید. Run و do هر دو دستور را اجرا می نمایند ولی do آن را به صورت Silent اجرا می کند. این دستورات در زیر نشان داده شد.



### • کار با نرم افزار:

با استفاده از مطالب پیشگفت کاربر گام به گام می تواند مراحل زیر را انجام دهد:

- ۱- از درستی نصب نرم افزار بر سیستم خویش اطمینان حاصل نماید. بدین منظور از دستور `verinst` استفاده می نماید. این دستور خروجی صحت نصب نرم افزار را نشان می دهد.
- ۲- برای کار عملی با نرم افزار لطفا دستور `sysuse cancer.dta` را اجرا نمایید. یک مجموعه فایل داده ای که در نرم افزار ذخیره شده است برای شما قابل مشاهده و قابل استفاده است. قبل از باز کردن برای ذخیره کردن دستورات و خروجی ها و ثبت کارهایی که قرار است بر روی این داده ها انجام شود فرمان های زیر بکار برده می شوند. برای اجرای فرمان اول، از مسیر فایل و منوی log مسیر ذخیره کارها را تعیین کنید.

```
log using "C:\Users\Tehran\Desktop\cancer.smcl"
. verinst
. memory
```

```
. db des  
. list  
. list groups  
. log close
```

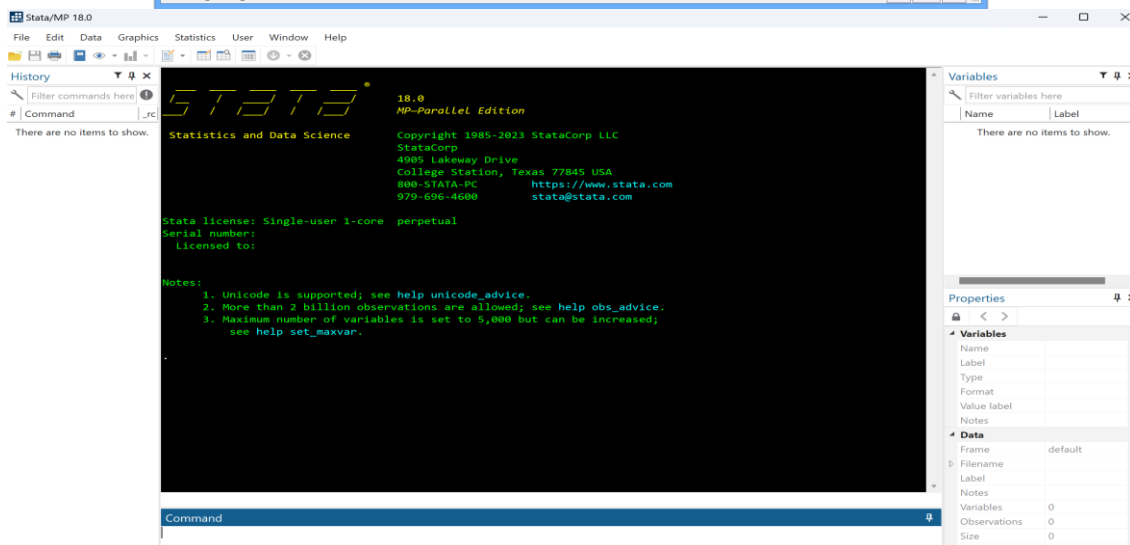
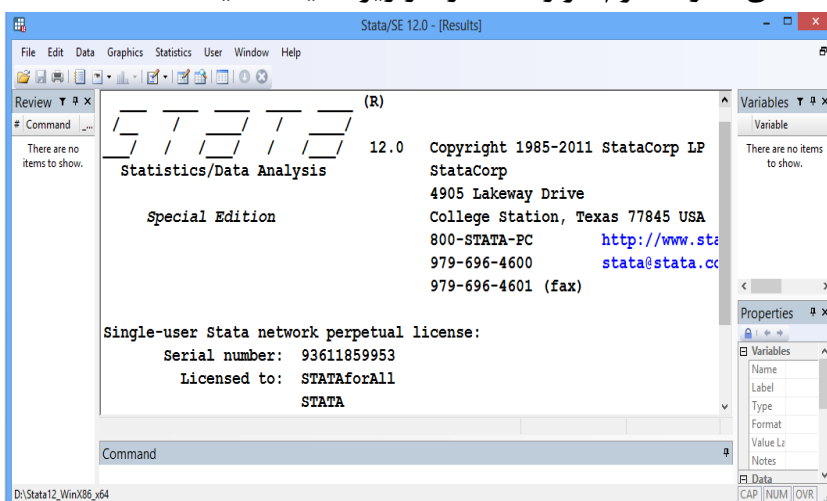
۳- با اجرای دستورات مذکور، مشخص است که این فایل حاوی ۸ متغیر و ۴۸ مشاهده است. صحت نصب نرم افزار و میزان حافظه در دسترس ارزیابی می شود. سپس کادر محاوره ای برای توصیف متغیرها باز می شود. پس از شناسایی متغیرها، آنها لیست می شوند. و ضمن آن لیستی از متغیر اول تهیه می گردد. در انتها برای ثبت کارهای انجام شده فایل بسته می شود.

### **خلاصه و نتیجه گیری:**

در این فصل نرم افزار Stata معرفی گردید. پنجره ها شرح داده شد. چگونگی راه اندازی و کار با نرم افزار بیان گردید. دستور بررسی صحت نصب و همچنین طرز خارج شدن از نرم افزار بیان گردید. تنظیمات مربوط به حافظه در دسترس برای کار با نرم افزار نیز بیان گردید. چگونگی ثبت و ذخیره کارهای انجام شده نیز گفته شد. طرز تهیه دو فایل (Do-File) و لیست کردن متغیرها نیز بیان گردید. در پایان نیز برای کار با نرم افزار گام های مورد نیاز با معرفی دستورات بیان شدند. زمانی که در نرم افزار یک فایل داده ای را باز کردید و بر روی آن فعالیتی انجام دادید، وقتی بخواهید فایل دیگری را باز نمایید نرم افزار خطا می دهد که فایل دیتای قبلی را از حافظه پاک نمایید. برای این کار باید از دستور **clear** استفاده نمایید.

تمرین ها :

- ۱- در برنامه Stata پنجره های اصلی را شناسایی نموده و نحوه باز کردن، بستن، جابجایی و تغییر اندازه آنها را تمرین نمایید.
- ۲- آیا میتوانید اندازه خط متن پنجره نتایج را تغییر دهید؟
- ۳- آیا می توانید رنگ نوشته های پنجره نتایج را تغییر دهید؟
- ۴- صفحه برنامه و خروجی برنامه را به رنگ ها و حالت های مختلف تغییر دهید.
- ۵- دکمه F8 را برای فرمان انجام رگرسیون لجستیک تنظیم و تعریف کنید.
- ۶- برای ذخیره سازی و ثبت فرمان ها و نتایج نرم افزار چگونه عمل می کنید؟
- ۷- شکل زیر کدام نوع از اشکال نرم افزار در بخش تنظیمات Preference است. در این شکل پنجره ها و منوها را بررسی و شناسایی نمایید.
- ۸- آیا می توان نام نرم افزار را STATA نوشت؟ در صورتیکه جواب خیر است چرا؟ نمای نسخه های ۱۲ و ۱۸ نرم افزار استاتا را در زیر مقایسه کنید.



## .....فصل دوم: مدیریت داده ها در Stata

مدیریت داده ها  
ساخت یک مجموعه از داده های جدید (ایجاد Dataset و ورود داده ها)  
تغییر نام متغیرها  
تعیین و تغییر ویژگی های متغیرها  
تبدیل یک متغیر اسمی به متغیر عددی  
کار عملی با نرم افزار  
تمرین دستورات

### • اهداف:

#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

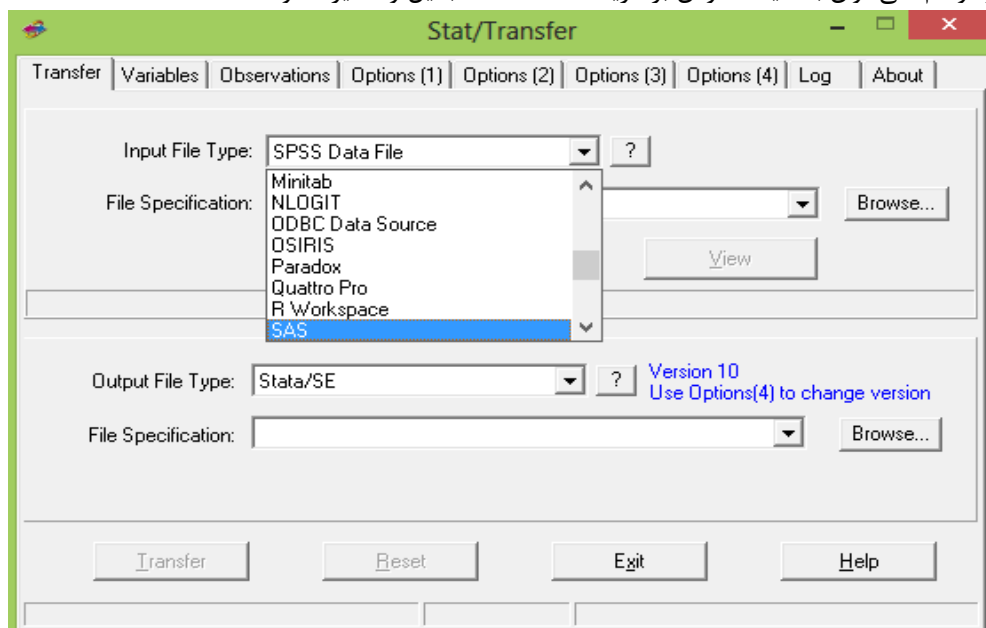
- یک مجموعه از داده های جدید ایجاد نماید. داده ها را وارد نرم افزار کند.
- متغیرها را بتواند تغییر نام دهد. برچسب به آنها الحاق نماید.
- متغیر جدید بسازد.
- متغیرها و داده ها را بتواند مدیریت و ویرایش نماید.
- ویرایش داده ها به عنوان گام مهم در تحلیل داده ها را به خوبی فرا گیرد.

مدیریت داده‌ها شامل انجام کارهایی مثل ساخت یک مجموعه داده‌ها، ویرایش داده‌ها به منظور تصحیح خطاها، اضافه کردن متغیر یا مشاهده‌ی جدید به مجموعه داده‌ها، ساخت متغیرهای جدید با استفاده از عبارتهای منطقی و جبری و... می‌شود. **اولین گام** برای تحلیل داده‌ها، وارد کردن داده‌ها به محیط Stata است. داده‌ها را به چند روش می‌توان به محیط Stata وارد کرد:

- با استفاده از فرمان edit، می‌توان داده‌ها را به طور مستقیم در محیط Stata وارد کرد.
- با استفاده از فرمان infile، داده‌هایی را که در یک فایل متن وارد شده‌اند، به محیط Stata فرا خواند.
- با استفاده از نرم‌افزار تبدیل‌کننده‌ی داده‌ها مانند نرم افزار **Stat Transfer** که به راحتی از اینترنت قابل دانلود کردن است، می‌توان داده‌هایی را که به فرمت صفحه‌گستر، بانک اطلاعاتی و... ساخته شده‌اند، به مجموعه داده‌های قابل استفاده برای Stata تبدیل کرد. شکل و لوگوی این نرم افزار و نحوه تبدیل داده‌ها در زیرآمده است.
- اخیراً از طریق منوی فایل و گزینه Import می‌توانید انواع فایل را بر اساس فرمت نرم افزارهای مختلف فراخوانی یا به فرمت‌های دلخواه از طریق Export تبدیل نمایید. در واقع کار نرم افزار Stat Transfer در استاتادغام شد.
- از طریق db infix و db insheet نیز می‌توان داده‌ها را وارد نمود. این روش‌ها رایج نیستند.

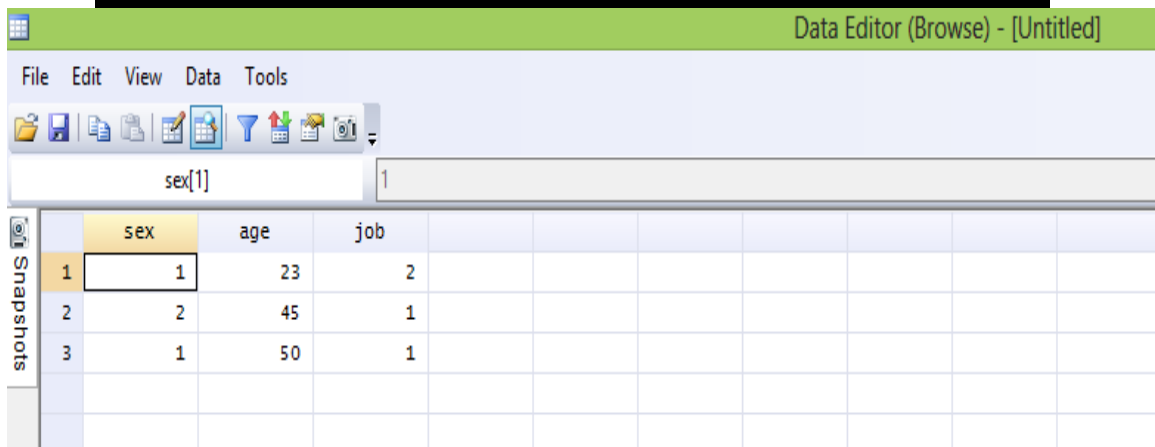


انواع مختلف فرمت داده‌ها، مربوط به نرم افزارهای مختلف آماری از جمله SAS و SPSS را می‌توان به فرمت Stata تبدیل نمود و برعکس. در شکل زیر که این نرم افزار اجرا شده است، می‌توان با دادن مسیر فایل ورودی (Input) که قرار است فرمت آن تغییر یابد، فایل مربوطه را به نرم افزار معرفی و سپس نوع فایل خروجی مورد نیاز (out put) را با آدرس محل ذخیره و نام، می‌توان با کلیک نمودن بر گزینه Transfer تبدیل و ذخیره نمود.



- از طریق صفحه کلید و تایپ دستورات و داده ها نیز میتوان داده ها را وارد نرم افزار نمود. به شکل زیر با دستور `input` و سپس وارد نمودن مقادیر داده ها برای متغیر های فرضی `sex`، `age` و `job` میتوان مقادیر را با یک فاصله و تایید (`Enter`) وارد نمود.
- در انتها با دستور `end` می توان خارج شد و بانک اطلاعاتی ساخته شده را به صورت زیر در ویراستار داده ها مشاهده نمود.

```
. input sex age job
      sex      age      job
1. 1 23 student
'student' cannot be read as a number
1. 1 23 2
2. 2 45 1
3. 1 50 1
4. end
```



متغیرها و دستورات بکار رفته در پنجره های نرم افزار به صورت زیر نمایش داده می شوند.

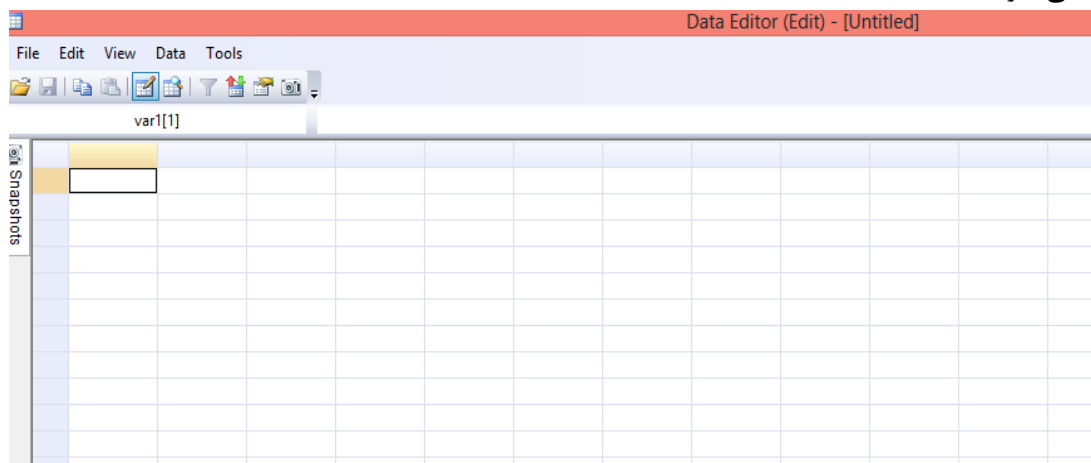
Variables		#	Command	_rc
sex		1	input sex age job	
age		2	1 23 student	
job		3	1 23 2	
		4	2 45 1	
		5	1 50 1	
		6	end	

ساخت یک مجموعه از داده های جدید یا بانک اطلاعاتی:

برای ساخت یک مجموعه داده های جدید در محیط Stata، استفاده از پنجره Editor یا آیکون ویرایشگر داده ها بهترین مسیر است. در Stata ساده ترین راه برای ساختن یک مجموعه داده های جدید، استفاده از پنجره ی



Editor است. در روش قبلی که توضیح داده شد وقت گیر و سخت است. با کلیک کردن روی گزینه Editor یا اجرای فرمان edit می توان به این پنجره وارد شد. وقتی این فرمان را تایپ کنید شکل زیر برای ورود داده ها مشاهده می شود.



پس از ورود به پنجره ی Editor می توان مقادیر مربوط به هر متغیر را در خانه های یک ستون وارد کرد. در صورتی که یکی از خانه ها خالی گذاشته شود، Stata به طور قراردادی مقدار مربوط به آن خانه را «مقدار گمشده» فرض می کند. اگر اولین مقداری که در یک ستون وارد می شود، عدد باشد Stata متغیر آن ستون را عددی و اگر حرفی باشد، متغیر آن ستون را نویسه ای فرض می کند. هر سطر به یک Case تعلق دارد و هر ستون نیز متعلق به یک متغیر می باشد. به طور مثال بیمار اول در سطر اول وارد می شود در حالی که سن همه بیماران در ستون اول وارد می شود. برای وارد نمودن مقدار متغیر، در خانه مورد نظر مقدار آن را تایپ کرده و سپس روی Enter یا Tab فشار دهید.

### تفاوت های Enter و Tab:

- Enter شما را به خانه بعدی در پایین خانه فعلی می برد و شما را در درون یک ستون حرکت می دهد.
  - Tab شما را به خانه سمت راست می برد و شما را ما بین ستونها حرکت می دهد.
- در بالای پنجره Data editor، مختصات متغیری که در حال وارد کردن آن هستید نشان داده می شود. به طور مثال اگر در حال وارد کردن مورد ۴ از متغیر ۳ هستید (یعنی ستون ۳ ردیف ۴) عبارت  $var3[4]=$  را مشاهده خواهید کرد و در مقابل آن مقدار متغیر، که توسط شما تایپ می شود قابل مشاهده است.

### طرز وارد کردن داده ها بصورت مورد به مورد:

- در بالاترین خانه موجود در اولین ستون خالی کلیک نمائید.
- مقدار را وارد نمائید.
- با Enter به خانه پایین وارد شوید.

### طرز وارد کردن داده ها بصورت متغیر به متغیر:

- در اولین خانه اولین ستون خالی کلیک کنید.
- مقدار مربوطه را وارد نمائید.
- برای ورود به خانه بعدی Tab را فشار دهید.

- پس از اتمام داده های مربوط به اولین مورد مشاهده ، روی خانه دوم در اولین ستون کلیک نمائید.
- مقادیر دومین مورد مشاهده را وارد نمائید، به همان ترتیب موارد بعدی را وارد نمائید.

کلید Tab یک کلید هوشمند است. به گونه ای تعداد متغیرها را می فهمد و بنابراین با خاتمه یافتن داده های مربوط به یک مورد مشاهده، بطور خودکار به اولین ستون خالی بر می گردد.

با دو بار کلیک کردن در هرستون متغیر **dialogue box** مربوط به **variable information** ظاهر می شود و می توانید نام متغیر مربوطه را وارد کنید.

نام هر متغیر باید بین ۱-۳۲ کاراکتر باشد، این کاراکترها می توانند:

- بصورت حروف باشند یا بصورت عددی باشند. یا بصورت خط **underscore** باشند.
- اما نام متغیر نمی تواند فضای خالی و سایر موارد را به عنوان کاراکتر بپذیرد.

توجه داشته باشید که Stata نسبت به حروف لاتین بزرگ یا کوچک حساس است و بنابراین حروف بزرگ را از کوچک در نامگذاری ها متمایز قلمداد می کند. Stata وجود یا اضافه کردن ستون یا سطر خالی را در حین وارد کردن داده ها نمی پذیرد، در صورت خالی گذاشتن سطر یا ستون، Stata، آن را به عنوان داده جا افتاده **missing** تلقی می کند.

### Missing data

- داده های وارد نشده در متغیرهای عددی به صورت . ثبت خواهد شد.
- داده های جا افتاده در متغیرهای متنی بصورت خانه خالی خواهند بود.

پس از بستن پنجره **Editor**، می توان مجموعه داده ها را با استفاده از فرمان زیر یا منوی **File→Save** در یک فایل روی دیسک ذخیره کرد: نام فایل **save**

افرادی که نگران تغییر در داده های خود هستند، باید بدانند که ویراستار (پنجره **Editor**) یک قابلیت خطرناک نیز دارد، بطوریکه به آسانی باعث تغییرات تصادفی می شود. لذا کاربران باید دقت نمایند، هر زمان که تنها قصد مشاهده داده های خود را دارند از جستجوگر داده ها (**Browse**) استفاده نمایند نه از ویراستار داده ها (**Edit**).

Stata فایل داده ها را با پسوند **dta** ذخیره می کند. اگر قبلاً فایلی به همین نام ایجاد شده باشد، برای این که فایل فعلی جایگزین فایل قبلی شود، فرمان **save** را باید به صورت زیر به کار برد:

**replace** , نام فایل **save**

تا قبل از خروج از Stata، برای مشاهده ی مجموعه داده ها، اصلاح مقادیر، اضافه کردن متغیر یا مقدار جدید به مجموعه داده های فعال، می توان مجدداً پنجره **Editor** را باز کرد.

در صورت خروج از Stata می توان مجموعه داده هایی را که قبلاً ذخیره شده است، با بکارگیری فرمان زیر یا منوی **File → Open** بازبازی کرد:

نام فایل **use**

برای خواندن یک فایل اطلاعات در برنامه از دستور **use** استفاده می شود برای استفاده از این دستور باید بعد از کلمه **use** آدرس فایل را نوشت مثلاً :

**use c:\data\cancer**

با نوشتن این دستور از زیر شاخه data فایل cancer توسط برنامه خوانده می شود. با تایپ دستور sysuse dir و دستور dir می توان لیست فایل ها و مستندات فایلی موجود در برنامه را مشاهده نمود. این لیست در زیر آمده است.

```
. sysuse dir
auto.dta      bpwide.dta      educ99gdp.dta  nlsw88.dta     surface.dta     voter.dta
auto16.dta    cancer.dta       gnp96.dta      nlswide1.dta   tsline1.dta     xtline1.dta
auto2.dta     census.dta       lifeexp.dta     pop2000.dta    tsline2.dta
autornd.dta   citytemp.dta     network1.dta    sandstone.dta  uslifeexp.dta
bplong.dta    citytemp4.dta    network1a.dta  sp500.dta      uslifeexp2.dta
```

```
. dir
<dir> 8/02/23 16:24 .
<dir> 8/02/23 16:10 ..
<dir> 8/02/23 16:10 ado
12.5k 4/24/23 10:28 auto.dta
<dir> 8/02/23 16:11 docs
0.0k 4/24/23 10:28 isstata.180
54.4M 4/24/23 9:22 mp-64.dll
0.2k 4/24/23 10:28 stata.html
79.6M 4/26/23 16:41 StataMP-64.exe
<dir> 8/02/23 16:11 utilities
```

برای باز نمودن یک فایل داده ای که به عنوان مثال در این نرم افزار وجود دارد می توانید از مسیر زیر اقدام نمایید: File/ Example Datasets / example dataset installed with Stata و فایل های داده ای Stata با پسوند dta را نشان میدهد.

این فایل ها به صورت پیش فرض و برای کار عملی با نرم افزار به عنوان مثال و فایل داده برای کار در نرم افزار وجود دارند. این مسیر را بخاطر بسپارید. در طی این کتاب از این داده ها نیز استفاده می شود.

نکته: فایلی با نام (auto.dta) را باز کرده و فرمان des را بنویسید تا مشخصات متغیرهای آن فایل برای شما نشان داده شود. یا با فرمان sysuse نام فایل مثال:

```
. sysuse auto.dta
(1978 Automobile Data)
```

### فراخوانی داده ها از یک فایل متنی (ASCII)

به منظور فراخوانی داده هایی که در یک فایل متنی ذخیره شده اند به محیط Stata، در صورتی که مقادیر متغیرهای مختلف مجموعه داده ها با حداقل یک فاصله خالی از هم جدا شده باشند، از فرمان infile استفاده می شود. اگر فایل داده ها شامل سه متغیر باشد، فرمان infile را باید به صورت زیر بکار برد:

```
infile x y z using نام فایل
```

اگر یکی از متغیرها نویسه ای باشد، قبل از نام متغیر، عبارت " تعدادحروف str " نوشته می شود.

نکته: وقتی که یک مجموعه داده ها در حافظه موجود است و برای فرا خواندن یک مجموعه داده های دیگر به حافظه، لازم است که مجموعه داده های فعال از حافظه پاک شود. به این منظور به دنبال فرمان infile use (یا هر فرمان دیگری که برای خواندن مجموعه داده ها به کار می رود) از گزینه ی clear استفاده می شود. در تجدید

نظر فعلی (ورژن جدید) نرم افزار این محدودیت بسته های قبلی حذف شده است. پس از ساختن مجموعه داده ها، می توان با استفاده از فرمان `describe`، مشخصات عمومی مجموعه داده ها (مثل تعداد مشاهدات، تعداد متغیرها، نام و نوع متغیرها) و با استفاده از فرمان `list`، مقادیر مجموعه داده ها را در پنجره ی خروجی Stata مشاهده کرد. چنانچه بخواهید در خروجی متغیرها با خطوط عمودی از همدیگر متمایز و قابل شناسایی باشند از فرمان `list, divider` استفاده میشود. `list, divider` همچنین می توان با فرمان `list, separator(0)` مشاهدات را بصورت ردیف های پشت سر هم و دلخواه نمایش داد. مثلاً نمایش مشاهدات بصورت ردیف های ده تایی با فرمان زیر انجام میشود.

`list, separator(10)`

### سفارشی کردن مجموعه داده ها (Customizing)

علاوه بر عملیات لازم برای ایجاد یک مجموعه داده های جدید، هرگونه تغییر و دستکاری در مجموعه داده ها که در نهایت منجر به گویاتر شدن خروجی فرمان ها یا نیل به اهداف تحقیق شود، در حیطه مدیریت داده ها قرار می گیرد. در این جا به چند مورد اشاره می شود:

### تغییر نام متغیرها

زمانی که داده ها در پنجره ی Editor وارد می شوند، Stata به طور قراردادی به متغیرها نام های `var1`، `var2` و... را منسوب می کند. بهتر است نام متغیرها تغییر داده شوند و از نام هایی که شناسایی متغیرها را آسان تر می کنند، استفاده شود. برای تغییر نام متغیر، فرمان `rename` به صورت زیر به کار می رود:

`rename var1` نام جدید

### برچسب گذاری متغیرها

در Stata می توان به مجموعه داده ها، متغیر یا مقادیر یک متغیر برچسبی مناسب، اضافه کرد. برای برچسب گذاری مجموعه داده ها از فرمان:

`label data` " برچسب مجموعه داده ها "

برای برچسب گذاری متغیرها از فرمان:

`label variable` " برچسب متغیر " نام متغیر

و برای برچسب گذاری مقادیر متغیرها از دو فرمان:

`label define` " برچسب اولین مقدار " اولین مقدار متغیر نام برچسب

... " برچسب دومین مقدار " دومین مقدار متغیر

`label values` نام برچسب نام متغیر

استفاده می شود.

### مرتب کردن مجموعه داده ها

فرمان `sort` مجموعه داده ها را بر اساس مقادیر متغیر مورد نظر، به طور صعودی مرتب می کند.

نام متغیر sort

فرمان sort مقادیر گم‌شده‌ی متغیر را در انتهای مجموعه داده‌ها قرار می‌دهد. به منظور مرتب کردن مجموعه داده‌ها به طور نزولی باید از فرمان gsort به صورت زیر استفاده شود:

نام متغیر [-] gsort

در صورت تمایل می‌توان ترتیب قرار گرفتن متغیرها در مجموعه داده‌ها را با استفاده از فرمان order تغییر داد.

order var1 var3 var2 var5 var4

order var5 , (first)

order var1 var2 var3, alphabetic

برای دیدن متغیرهای خاص در صفحه ویرایش داده می‌توان با دستور ed قبل از نام متغیرهایی که فقط آنها را می‌خواهید مشاهده نمایید استفاده نمود.

### تعیین حوزه‌ی عملکرد فرمان

می‌توان با استفاده از انتخابگرهای in و if حوزه‌ی عمل بسیاری از فرمان‌های Stata را تنها به بخشی از مجموعه داده‌ها محدود کرد، به این معنا که می‌توان از Stata خواست تا فرمان مورد نظر را فقط روی تعدادی از مشاهدات یا روی مشاهداتی که دارای مشخصه‌ی خاصی هستند، اعمال کند. انتخابگرهای in و if در انتهای فرمان قرار می‌گیرند. آن‌ها را می‌توان به شکل عمومی زیر به کار برد:

محدوده‌ی عمل in نام فرمان

شرط if نام فرمان

محدوده‌ی عمل برای انتخابگر in را می‌توان به چند شکل تعریف کرد:

شماره‌ی مشاهده‌ی مورد نظر in نام فرمان

شماره‌ی مشاهده‌ی انتهای محدوده / شماره‌ی مشاهده‌ی ابتدای محدوده in نام فرمان

l / شماره‌ی مشاهده‌ی ابتدای محدوده in نام فرمان

که l به معنای آخرین مشاهده در مجموعه داده‌ها است. به عنوان مثال برای نمایش مشاهدات شماره ۵ تا ۱۵ در یک فایل داده در پنجره ویراستاری یا editor می‌توان از دستور edit in 5/15 استفاده نمود.

به منظور تعریف شرط برای انتخابگر if، از عملگرهای مقایسه‌ای استفاده می‌شود.

این عملگرها عبارتند از:

< کوچکتر	> بزرگتر	= مساوی
<= کوچکتر یا مساوی	>= بزرگتر یا مساوی	!= نامساوی

در صورتی که تعداد شرط‌ها از یک شرط بیشتر باشد، می‌توان بنا به نیاز شرط‌ها را با استفاده از عملگر منطقی "&" و "|" در مقابل انتخابگر if نوشت.

گاهی مجموعه داده‌ها شامل یک متغیر رسته‌ای یا کیفی و یا طبقه‌ای (Categorical) است و تحلیل‌ها باید روی مقادیر مختلف متغیر رسته‌ای به طور جداگانه انجام شود. برای این کار ابتدا مجموعه داده‌ها برحسب متغیر

رسته‌ای مورد نظر مرتب می‌شود:

نام متغیر sort

سپس گزینه‌ی by را به یکی از دو صورت زیر به همراه فرمان مورد نظر به کار برد:

(نام متغیر رسته‌ای) by, ... نام فرمان

... نام فرمان : نام متغیر by

### تعریف متغیرهای جدید و مقادیر جدید برای متغیرها

گاهی لازم است با انجام عملیات ریاضی روی برخی از متغیرهای یک مجموعه داده‌ها، متغیر جدیدی ساخته شود یا مقادیر فعلی یک متغیر با مقادیر جدید، جایگزین شوند. فرمان‌های generate و egenate متغیر جدید و فرمان replace مقادیر جدید برای یک متغیر، را تعریف می‌کند.

تابعی ریاضی از متغیرهای مجموعه داده‌ها = نام متغیر جدید generate(e)

تابعی ریاضی از متغیرهای مجموعه داده‌ها = نام یکی از متغیرهای مجموعه داده‌ها replace

برای ساختن یک متغیر جدید که از لگاریتم گرفتن متغیر دیگر ساخته می‌شود از دستور زیر استفاده می‌شود.

`g= newVariable= ln(varname)`

تابع ریاضی از متغیرهای مجموعه داده‌ها را می‌توان با بکارگیری عملگرها و توابع Stata ساخت.

برخی از عملگرها و توابع Stata در زیر معرفی شده‌اند:

#### - عملگرهای جبری:

جمع +    تفریق -    ضرب \*    تقسیم /    توان ^

برای تغییر و تبدیل مقادیر یک متغیر از دستور db recode استفاده می‌شود. سپس در آنجا با انتخاب

نام متغیر و اعمال قوانین تغییرات را در مشاهدات انجام داد. قوانین در شکل زیر قابل مشاهده است.

rule	Example	Meaning
(# = #)	(3 = 1)	3 recoded to 1
(# # = #)	(2 . = 9)	2 and . recoded to 9
(#/# = #)	(1/5 = 4)	1 through 5 recoded to 4
(nonmissing = #)	(nonmiss = 8)	all other nonmissing to 8
(missing = #)	(miss = 9)	all other missings to 9

به عنوان مثال فایل ماشین‌ها را با دستور `sysuse auto.dta` باز نمائید. یک متغیر جدید به نام `price2`

از روی متغیر `price` بسازید. سپس متغیر جدید را به دسته ماشین‌های با کمتر از میانگین قیمت و

بالتر از میانگین کد بندی و جدول فراوانی بگیرید. دستورات زیر را ببینید.

```
gen price2 = price
sort price2
sum price2
recode price2 1/6165=1 6165.3/15906=2
tab price2
```

- برخی از توابع ریاضی:

abs(x)	the absolute value of x
exp(x)	the exponential function of x
int(x)	the integer obtained by truncating x towards zero
ln(x) or log(x)	the natural logarithm of x if $x > 0$
log10(x)	the log base 10 of x (for $x > 0$ )
logit(x)	the log of the odds for probability x: $\text{logit}(x) = \ln(x/(1-x))$
max(x1,x2,...,xn)	the maximum of x1, x2,..., xn, ignoring missing values
min(x1,x2,...,xn)	the minimum of x1, x2,..., xn, ignoring missing values
round(x)	x rounded to the nearest whole number
sqrt(x)	the square root of x if $x \geq 0$

در توابع بالا abs(x) قدر مطلق، exp(x) تابع نمایی، log(x) لگاریتم طبیعی، log10(x) لگاریتم بر مبنای ده و sqrt(x) ریشه‌ی دوم است.

**نکته:** برای به دست آوردن مقدار یک عبارت ریاضی، می‌توان از فرمان display که به عنوان ماشین حساب نرم افزار شناخته می‌شود به صورت زیر استفاده کرد:

عبارت ریاضی display

طرز ساخت متغیر مرکب یا ترکیبی مانند نمایه توده بدن (Body Mass Index: bmi)

$g \text{ bmi} = \text{weight} / (\text{hight}/100)^2$

حذف متغیر یا مشاهده از مجموعه داده‌ها

حذف یک یا چند متغیر یا مشاهده از مجموعه داده‌ها، به دو صورت زیر امکان پذیر است:  
 ۱- تعیین متغیرها یا مشاهداتی که باید حذف شوند، با استفاده از فرمان drop:

نام متغیر(ها) drop

محدوده in شرط drop if

drop if age <=50

۲- تعیین متغیرها یا مشاهداتی که باید نگه داشته شوند، با استفاده از فرمان keep:

نام متغیر(ها) keep

محدوده in شرط keep if

list if age >50

۳- برای حذف تمام متغیرها از دستور drop -all استفاده می‌شود.

**نکته:**

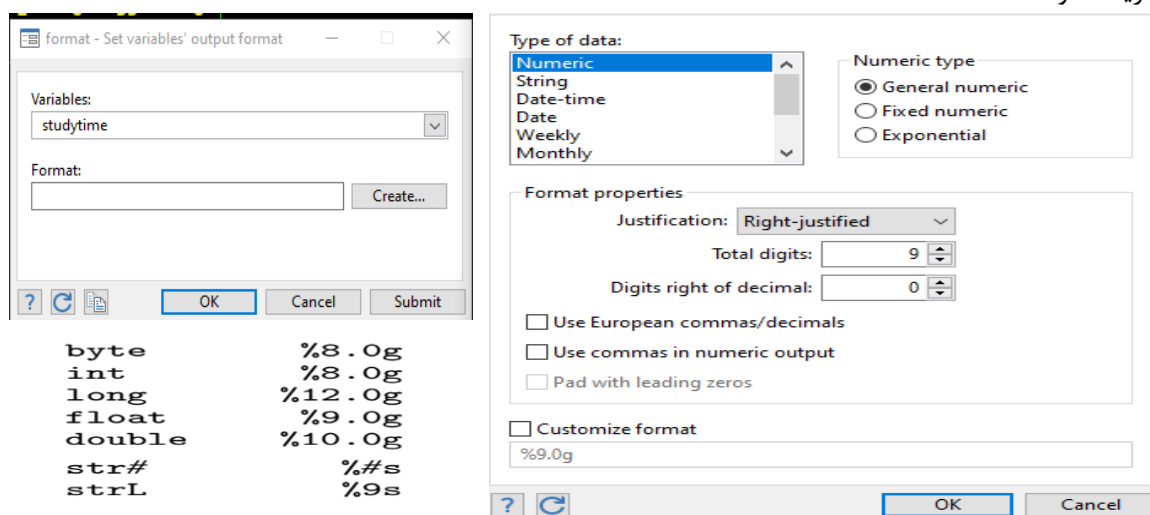
برای ساخت متغیرشناسایی یا کد دادن به بانک داده‌ها با دستور gen id=\_n می‌توان یک متغیر ایجاد نمود.

## تعیین و تغییر ویژگی های متغیرها

برای تعیین و تغییر ویژگی های متغیرها در Variables window روی متغیر مورد نظر خود کلیک راست کنید. در این حال امکان تغییرات فراهم می گردد. با استفاده از گزینه format می توانید فرمت متغیرها را تغییر دهید و آنها را به شکل عددی (با رقم های متفاوت اعشاری)، اسمی، ساعت، روز بنویسید. وقتی داده ها در متغیری مانند time، تعداد روز باشد و بخواهیم به فرمت تاریخ تبدیل شوند از فرمان زیر استفاده میشود. برای فرمت های دیگر از فرمان help dfm استفاده می شود.

`format time %d`

متغیرها را در نرم افزار استاتا به فرمت های مختلفی می توان تعریف نمود. در جدول زیر انواع فرمت ها گزارش شده است. با دستور db format می توان جعبه گفتگو درباره شکل متغیرها را باز نمود. سپس با انتخاب نام متغیر و فشردن دکمه یا زبانه Create، فرمت متغیر را بر حسب نوع متغیر (عددی، اسمی، زمان) و نحوه چیدمان آن را تعریف نمود.



```
byte          %8.0g
int           %8.0g
long         %12.0g
float        %9.0g
double       %10.0g
str#         %#s
strL        %9s
```

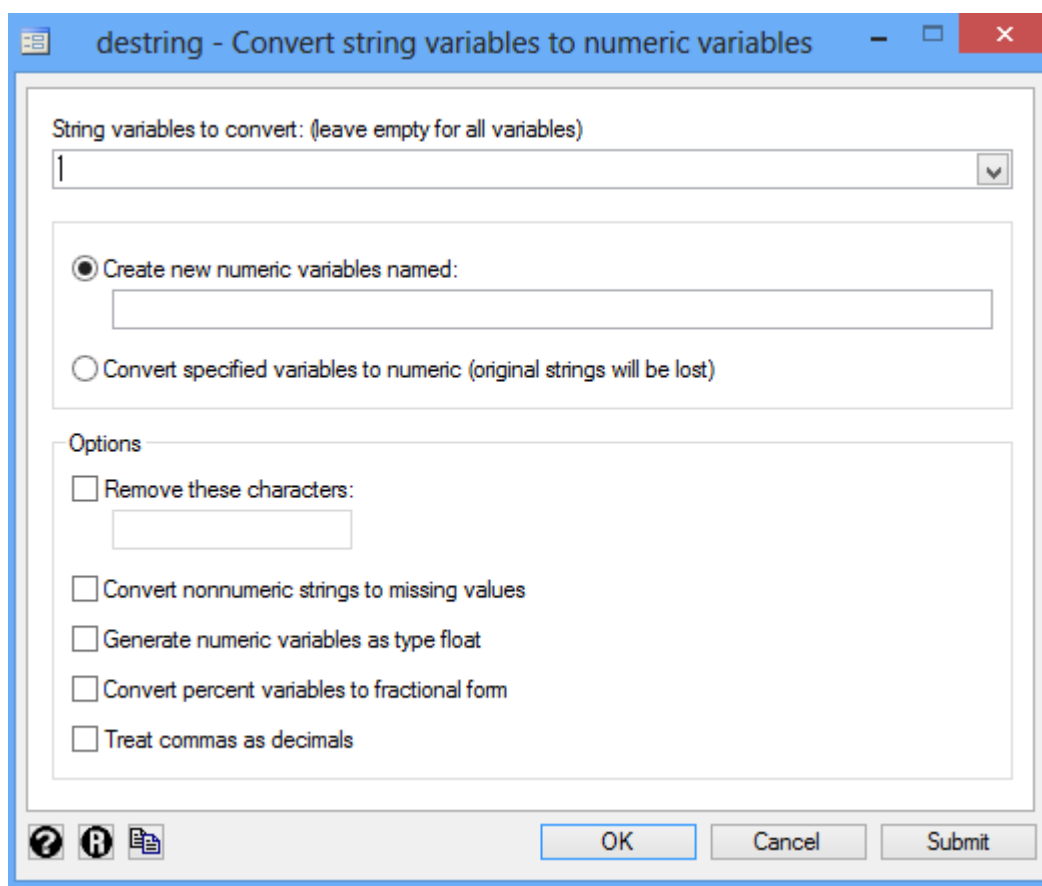
به عنوان مثال فرمت های متغیر عددی در شکل زیر قابل مشاهده است.

Numerical %fmt	Description	Example
right-justified		
%#.#g	general	%9.0g
%#.#f	fixed	%9.2f
%#.#e	exponential	%10.7e
%21x	hexadecimal	%21x
%16H	binary, hilo	%16H
%16L	binary, lohi	%16L
%8H	binary, hilo	%8H
%8L	binary, lohi	%8L
right-justified with commas		
%#.#gc	general	%9.0gc
%#.#fc	fixed	%9.2fc
right-justified with leading zeros		
%0#.#f	fixed	%09.2f
left-justified		
%-.#g	general	%-9.0g
%-.#f	fixed	%-9.2f
%-.#e	exponential	%-10.7e
left-justified with commas		
%-.#gc	general	%-9.0gc
%-.#fc	fixed	%-9.2fc

You may substitute comma (,) for period (.) in any of the above formats to make comma the decimal point. In %9,2fc, 1000.03 is 1.000,03. Or you can set dp comma.



همچنین با استفاده از مسیر زیر نیز می توان در متغیرها تغییر ایجاد کرد:  
 برای تغییر ویژگی های متغیرها می توانید صفحه Data Editor را باز نموده و روی نام متغیر دو بار کلیک نمایید. جعبه Variable Properties باز می شود. نام متغیر، Label، Format و Value label در اینجا قابل تغییر و تصحیح می باشد. برای تنظیم Value label روی دگمه Define/ modify کلیک نموده و روی Define کلیک کنید. با مشخص نمودن Value و مقادیر آن و Ok نمودن برای هر یک از مقادیر متغیر مورد نظر تعریف یا label مربوطه مشخص می گردد. برای مشاهده آنها در جدول متغیرها، روی نام متغیر کلیک راست کرده و در گزینه Assign value label to variable آن را انتخاب نمایید. از مسیر زیر نیز می توان داده های با فرمت string را با عددی تبدیل نمود. وقتی داده های یک فایل را مثلا از صفحه گسترده اکسل، کپی و در Stata آنها را paste کنید فرمت متغیرها و داده ها ممکن است توسط Stata نویسه ای یا string فرض شود و محاسبه های کمی بر روی آنها انجام نشود. لذا برای رفع این مشکل با دستور destring, replace و یا استفاده از جعبه گفتگوی زیر می توان تبدیل مورد نیاز را انجام داد.



با دستور collapse می توان میانبر را فراخوانی نمود و شاخص های میانگین، میانه و .. را برای متغیرها محاسبه نمود.

db collapse

**دستور expand):**

**تولید جدول مجازی برای داده ها**

در مبحث مدیریت داده ها و همچنین تحلیل، یکی از دستورات ویژه در نرم افزار، دستور expand است. با این دستور می توان برای جداول، داده تولید نمود. یا به عبارتی فایل اطلاعاتی تهیه کرد. به عنوان مثال چنانچه جدولی از نتایج یک طرح پژوهشی بررسی رابطه طبقه اجتماعی و ابتلا به سل را به شما به عنوان داور یک مجله بین المللی، جهت اعلام نظر درباره صحت رابطه گزارش شده بدهند و سردبیر مجله در گزارش مقادیر معنی داری رابطه این دو متغیر با همدیگر تردید نماید، شما می توانید برای جدول مزبور، داده تولید نموده و آن را آزمون و مقادیر معنی داری را بدست آورید. سپس نتایج را با مقادیری که نویسنده مقاله گزارش نمود مقایسه کنید و نتایج را به سردبیر مجله ارسال نمایید!!

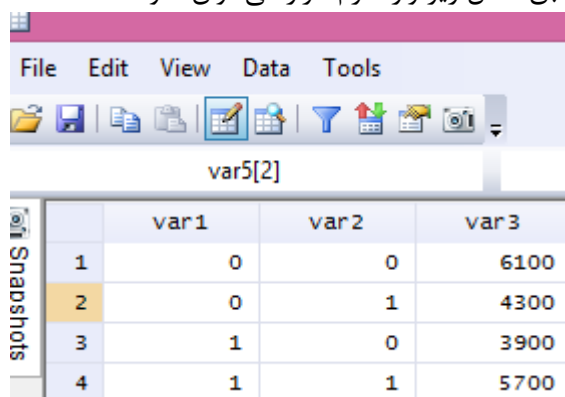
مثال زیر این داستان را نشان می دهد.

نسبت شانس طبقه اجتماعی و بیماری سل در جمعیت مورد بررسی با داده های جدول زیر 3.07 گزارش گردید. و ذکر گردید دامنه اطمینان یک را در برگرفته و نسبت شانس معنی دار نیست. و طبقه اجتماعی و بیماری سل ارتباط معنی داری ندارند. این نتایج را راستی آزمایی نمایید.

بیماری سل / طبقه اجتماعی	طبقه اجتماعی ضعیف	طبقه اجتماعی خوب
دارد	۵۷۰۰	۳۹۰۰
ندارد	۴۳۰۰	۶۱۰۰
جمع	۱۰۰۰۰	۱۰۰۰۰

**برای ساخت بانک اطلاعاتی برای جدول بالا و محاسبه نسبت شانس به ترتیب زیر عمل می شود:**

- ۱- ابتدا داده ها را باید وارد نرم افزار نمود. با دستور edit می توان داده ها را وارد پنجره ویراستار نمود. قبل از آن کاربر باید برای هر یک از متغیرهای سل و طبقه اجتماعی یک کد تعریف نماید. برای متغیر سل کد ۱ برای بیماران و کد ۰ برای کسانی که بیماری ندارند تعریف می شود. برای متغیر طبقه اجتماعی نیز کد ۱ برای طبقه اجتماعی ضعیف و کد ۰ برای طبقه اجتماعی خوب تعریف می شود.
- ۲- متغیرهای مذکور را طبق شکل زیر وارد نرم افزار می توان نمود.



- ۳- نام متغیرها بر اساس تعاریف گام یک ، ویرایش می گردد. بانک اطلاعاتی به صورت زیر نشان داده می شود.

```
rename var1 tb
rename var2 socialclass
```

	tb	socialclass	var3
1	0	0	6100
2	0	1	4300
3	1	0	3900
4	1	1	5700

نکته: البته برای تعداد زیادی سطر و ستون در یک جدول، می توان این بانک را ساخت. لیکن اینجا به عنوان مثال جدول دو در دو مد نظر است. این کار را برای جداول پیچیده نیز می توان انجام و شاخص هایی نظیر میزان توافق، پایایی و رابطه را از جداول محاسبه نمود.

۴- دستور expand var3 اجرا می شود. نتایج زیر حاصل می شود. مجدد پنجره ویراستار یا بانک اطلاعاتی را مشاهده نمایید. شکل زیر داده های این فایل را نشان می دهد.

	tb	socialclass	var3
1	0	0	6100
2	0	1	4300
3	1	0	3900

19990	1	1	5700
19991	1	1	5700
19992	1	1	5700
19993	1	1	5700
19994	1	1	5700
19995	1	1	5700
19996	1	1	5700
19997	1	1	5700
19998	1	1	5700
19999	1	1	5700
20000	1	1	5700

۵- بعد از آن می توان متغیر var3 را با دستور drop var3 حذف نمود. برای متغیرها برچسب ایجاد نمایید.

```
label define tb 0"no" 1"yes"
label values tb tb
label define socialclass 0"no" 1"yes"
label values socialclass socialclass
```

۶- اینک با دستور زیر جدول را تشکیل دهید. و مقدار نسبت شانس قابل محاسبه است.

. tab tb socialclass

tb	socialclass		Total
	no	yes	
no	6,100	4,300	10,400
yes	3,900	5,700	9,600
Total	10,000	10,000	20,000

mhodds tb socialclass

Odds Ratio: 2.07 chi2(1)=649.01 p= 0.0000

[95% Conf. Interval] 1.95 – 2.19

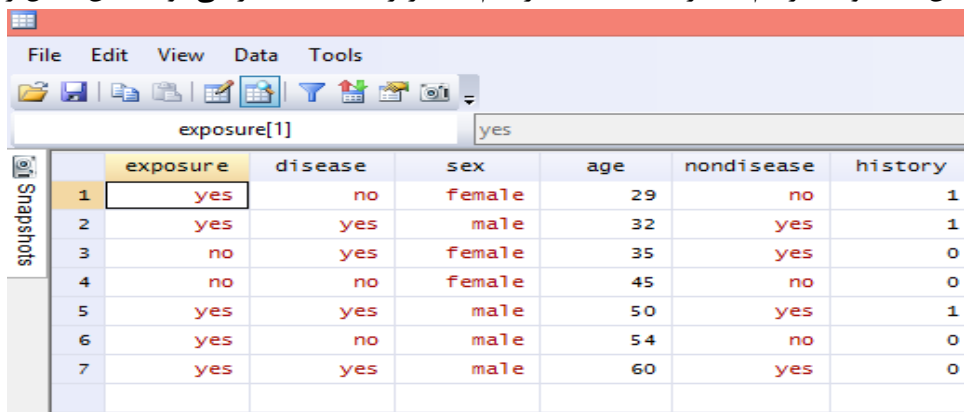
همانگونه که نتایج نشان می دهد نسبت شانس گزارش شده و مقدار معنی داری جدول صحت ندارد! نگران نباشید نحوه محاسبه، تعریف و تفسیر نسبت شانس و دستورات محاسبه نسبت شانس در فصول بعدی مفصل توضیح داده می شوند.

### کار عملی با نرم افزار و آموزش گام به گام مطالب پیشگفت :

۱- فایل داده های زیر را که حاوی شش متغیر مربوط به یک مثال فرضی است را به نرم افزار استتا وارد کنید.

Var6	Var5	Var4	Var3	Var2	Var1
1	35	female	Yes	No	1
1	32	male	Yes	Yes	2
0	45	female	no	No	3
0	54	male	No	Yes	2
1	50	male	Yes	Yes	3
0	29	female	No	Yes	1
0	60	male	yes	yes	1

۲- داده ها را در جای مشخص ذخیره و از نرم افزار خارج شوید. شکل زیر نمایش داده های وارد شده با فرمان edit و تغییر نام متغیرها با استفاده از گام های زیر است. شما نیز می توانید این شکل را بیاورید.



	exposure	disease	sex	age	nondisease	history
1	yes	no	female	29	no	1
2	yes	yes	male	32	yes	1
3	no	yes	female	35	yes	0
4	no	no	female	45	no	0
5	yes	yes	male	50	yes	1
6	yes	no	male	54	no	0
7	yes	yes	male	60	yes	0

۳- مجدد فایل داده های ذخیره شده را باز کنید. کلیه کارهایی که به شرح زیر قرار است روی این داده ها انجام دهید را ثبت و ذخیره نمایید. برای باز کردن مجدد فایل، در جایی که فایل قبلا ذخیره گردید میتوان با دوبار کلیک نمودن روی آن یا از مسیر فایل آن را باز نمود. برای ثبت کارهایی که قرار است انجام شود، از منوی فایل، با انتخاب log و سپس شروع و انتخاب مسیر ذخیره فعالیت ها، میتوان کارها را ثبت نمود.

```
"log using "C:\Users\Tehran\Desktop\ahmadi1400.smcl
```

۴- به داده ها بر چسب part 2 را الصاق نمایید. با استفاده از فرمان زیر این کار انجام می شود.

```
label data " part2"
```

۵- نام متغیرهای vra6 تا var1 را به ترتیب از چپ به راست به sex، age، disease و exposure و

location و history تغییر نام دهید. با استفاده از دستورات زیر میتوان تغییر نام متغیرها را انجام داد.

```
rename var1 location
```

```
rename var2 exposure
```

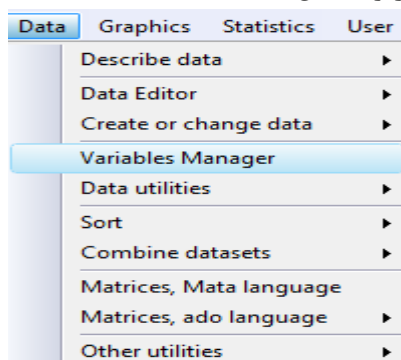
```
rename var3 disease
```

```
rename var4 sex
```

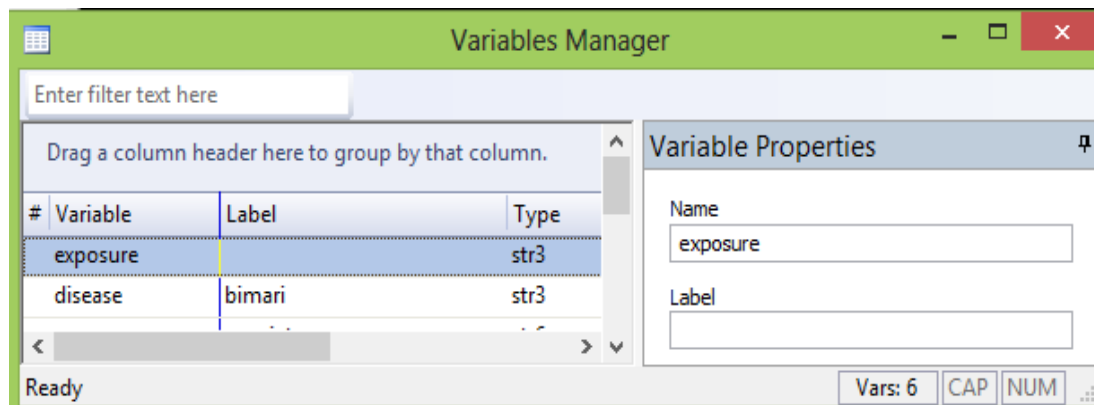
```
rename var5 age
```

```
rename var6 history
```

یا اینکه میتوان از طریق منوی data و انتخاب گزینه variable manager تغییرات مورد نظر را بر متغیرها اعمال کرد. این مسیر به صورت گرافیکی در زیر نمایش داده شده است.



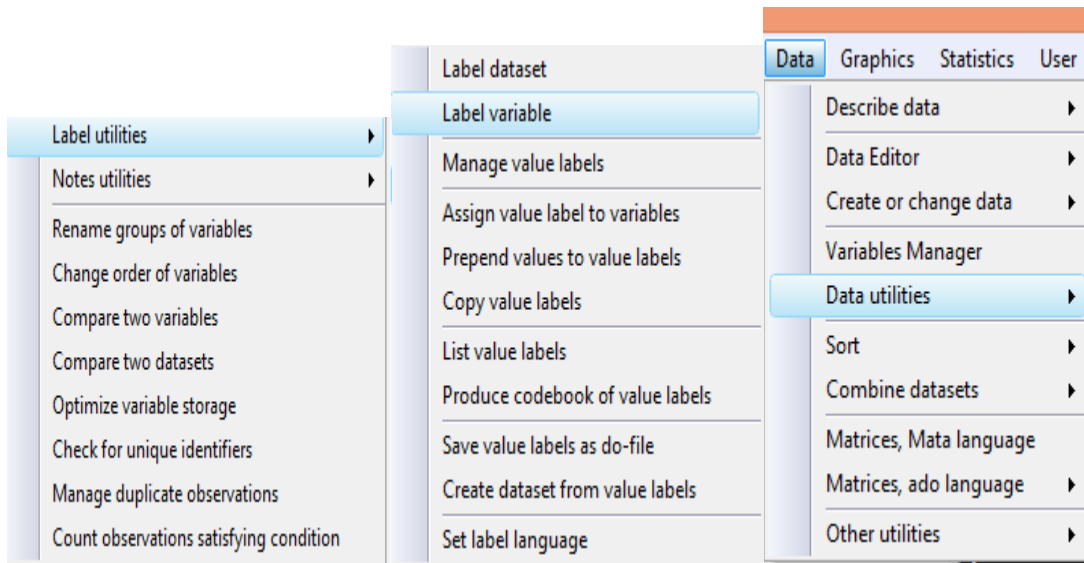
شکل زیر مدیریت متغیرها را نشان می دهد. در این شکل نیز می توان مشخصات متغیرها را مشاهده و تغییرات مورد نیاز را دنبال نمود.



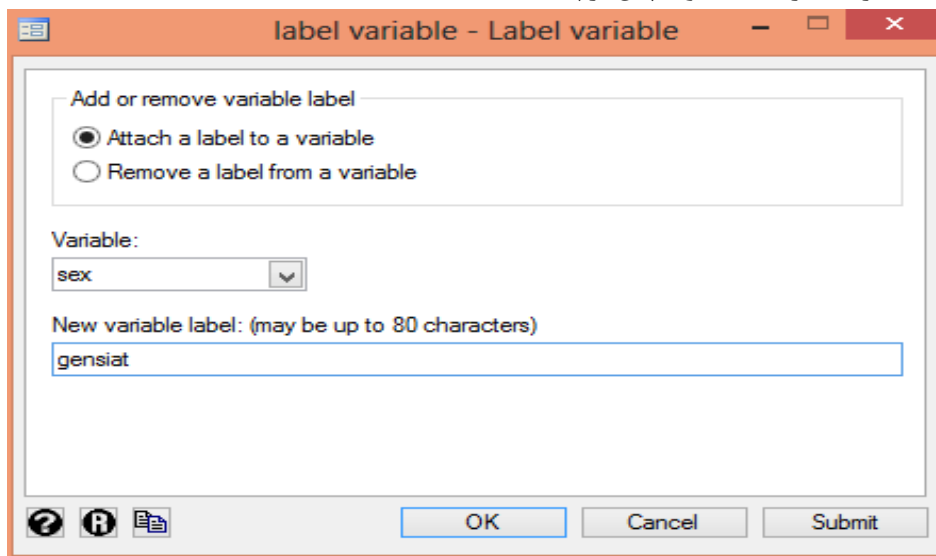
۶- به متغیر sex برچسب f برای زنان و m برای مردان و به متغیر disease نیز برچسب bimari و کد صفر و یک را برای داشتن یا نداشتن بیماری الصاق نمایید. برای این کار از دو راه میتوان اقدام کرد. راه اول از طریق اجرای فرمان های زیر است.

```
label variable sex "gensiat"
label define gensiat 1 "male" 2 "female"
label variable disease "bimari"
label define bimari 1 "yes" 0 "no"
```

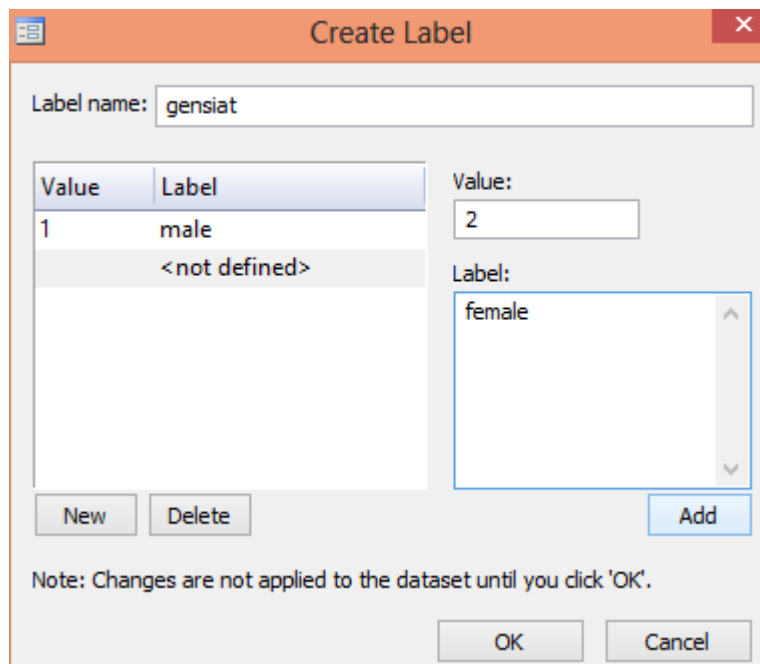
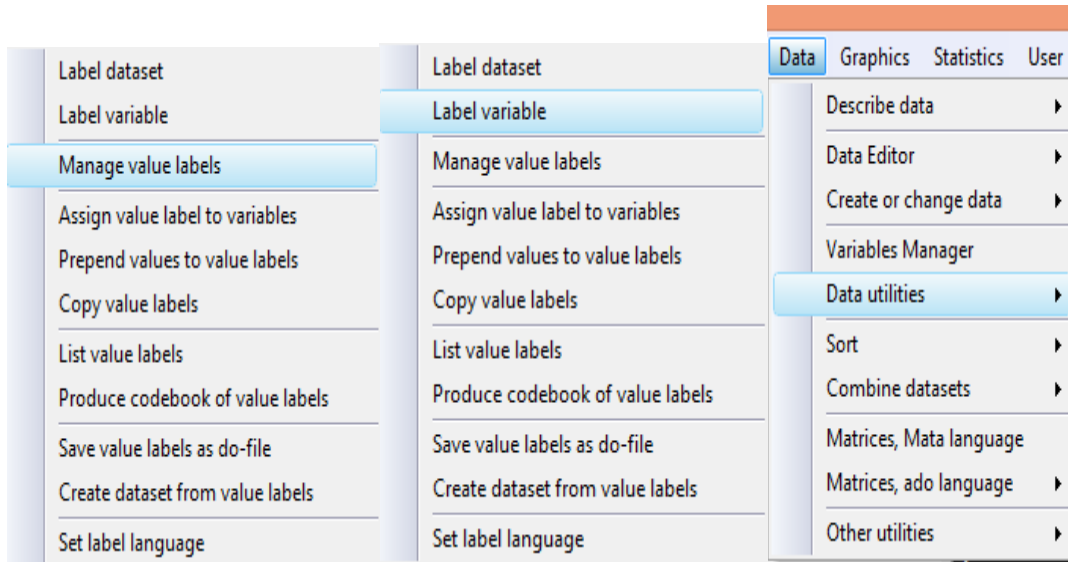
راه دوم از طریق منوی نرم افزار است که با استفاده از مسیرهای زیر در شکل نمایش داده می شوند.



سپس متغیری که قرار است به آن برچسب الصاق شود را انتخاب و برچسب مورد نظر در کادر پایین آن تایپ می شود. در اینجا متغیر sex را انتخاب و سپس برچسب gensiat اضافه شده است.



برای اضافه نمودن مقادیر عددی به برچسب نیز از مسیر زیر میتوان استفاده نمود. سپس در کادر محاوره ای باز شده مقادیر به برچسب اضافه می شوند.



۷- به متغیر history بر چسب no برای کد ۰ و yes برای کد ۱ بزنید. با دستورات زیر این کار انجام می شود. یا می توان با استفاده از منوی Data که در بالا گفته شد این کار را انجام داد.

```
label define history 0 "no" 1 "yes"
label values history history
```

۸- داده ها را بر حسب متغیر age مرتب نمایید.

```
sort age
```

۹- مشخصات عمومی مجموعه داده ها را توضیح دهید. از فرمان describe استفاده می شود. خروجی به شکل زیر نمایش داده می شود.

```
. des

Contains data from C:\Users\ali\Desktop\data book stata\part two.dta
  obs:          7                part2
  vars:         6                2 Aug 2013 01:01
  size:        140

-----+-----
variable name  storage  display  value  variable label
              type   format   label
-----+-----
exposure      str3    %9s
disease       str3    %9s      bimari
sex           str6    %9s      gensiat
age           byte   %8.0g
nondisease    str3    %9s      healther
history       float  %9.0g
```

۱۰- مقادیر مجموعه داده ها را در خروجی مشاهده نمایید.

list

۱۱- مشاهدات را به صورت سه تایی در خروجی به نمایش بگذارید.

list,separator(3)

	exposure	disease	sex	age	nondis~e	history
1.	yes	no	female	29	no	1
2.	yes	yes	male	32	yes	1
3.	no	yes	female	35	yes	0
4.	no	no	female	45	no	0
5.	yes	yes	male	50	yes	1
6.	yes	no	male	54	no	0
7.	yes	yes	male	60	yes	0

۱۲- متغیرها را بصورت تفکیک شده با خط عمودی نمایش دهید.

list ,divider

بعد از اجرای دستور بالا خروجی زیر که متغیرها از هم تفکیک شده اند نمایش داده می شود.

	exposure	disease	sex	age	nondis~e	history
1.	yes	no	female	29	no	1
2.	yes	yes	male	32	yes	1
3.	no	yes	female	35	yes	0
4.	no	no	female	45	no	0
5.	yes	yes	male	50	yes	1
6.	yes	no	male	54	no	0
7.	yes	yes	male	60	yes	0



۱۳- متغیر جدیدی بسازید بنام nonhistory که افراد بدون سابقه بیماری فقط در آن وجود داشته باشد.  
gen nonhistory=history

drop if history==1

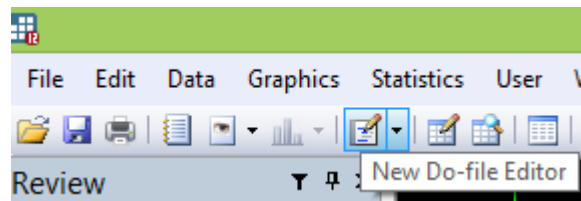
۱۴- متغیر location را حذف کنید.

drop location

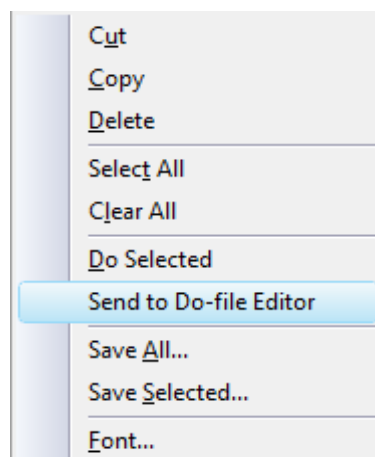
۱۵- برای مرتب کردن متغیرها می توان از دستور order استفاده نمود.

order var1 var2 var3  
order var1, last  
order var1, first  
order var1, before(var2)

برای انجام مراحل بالا می توان do-file درست نمود و کلید دستورات بکار رفته را در آنجا تایپ یا کپی نمود. سپس دستورات را از این محل فراخوان نموده و اجرا کرد. انجام مراحل بالا به شرح زیر است.



با انتخاب گزینه do-file editor در شکل بالا می توان دستورات بکار رفته را در آنجا تایپ و ذخیره نمود. سپس در مواقع مورد نیاز از آنها استفاده نمود. یا می توان در پنجره مرور، دستورات را انتخاب و سپس راست کلیک و send to do-file editor نمود. وقتی راست کلیک می شود شکل زیر نشان داده می شود.



فایل اجرایی تشکیل شده برای دستورات بالا به شکل زیر است.

```

1  rename var1 location
2  rename var2 exposure
3  rename var3 disease
4  rename var4 sex
5  rename var5 age
6  rename var6 history
7  label data " part2 "
8  label variable sex "gensiat"
9  label define gensiat 1 "male" 2 "female"
10 label variable disease "bimari"
11 label define bimari 1 "yes" 0 "no"
12 label define history 0 "no" 1 "yes"
13 label values history history
14 sort age
15 describe
16 list
17 list,separator(3)
18 list ,divider
19 gen nonhistory=history
20 drop if history==1
21 drop location

```

می توان فایل تشکیل شده را با نام دلخواه در مسیر مورد نظر ذخیره و در زمان های دیگر که نیاز به اجرای دستورات باشد، بدون تایپ مجدد، آنها را اجرا نمود.

### خلاصه و نتیجه گیری:

در این فصل طرز مدیریت داده ها و کار با آنها و تغییر متغیرها و خلاصه دستورات مورد نیاز توضیح داده شد. مهمترین دستورات و کاربرد آنها به شرح زیر است.

- use Use Stata dataset
- save Save datasets
- describe Describe data in memory or in file
- inspect Display simple summary of data's attributes
- codebook Describe data contents
- data types Quick reference for data types
- missing values Quick reference for missing values
- dates and times Date and time (%t) values and variables
- list List values of variables
- edit Browse or edit data with Data Editor
- varmanage Manage variable labels, formats, and other properties
- rename Rename variable
- format Set variables' output format
- label Manipulate labels
- generate Create or change contents of variable

- functions Functions
- egen Extensions to generate
- drop Eliminate variables or observations
- clear Clear memory

در قسمت مدیریت داده ها و ویرایش آنها، دستورات بسیار متعددی وجود دارد و کارهای زیادی قابل انجام است. نرم افزار Stata یکی از منحصر به فرد ترین نرم افزارها است که کاربران می توانند با فراگیری این دستورات کارایی کار با نرم افزار را بالا ببرند. همچنین می توان در نرم افزار برنامه نویسی نمود.

توصیه می شود هر گاه با نرم افزار Stata بر روی فایل داده ای قرار است کار انجام گیرد کاربر همیشه دستورات به کار برده شده را در یک Do-File ذخیره نماید.

و نکته مهم اینکه در فرایند تولید، ذخیره، ساختار، انتقال، دسترسی، بازیابی، کدبندی، برچسب گذاری، شناسایی، طبقه بندی و کنترل کیفی داده ها بایستی همواره دقت داشت که در فرایند مورد استفاده تغییری در داده ها ایجاد نشود و اصول علمی اخلاق در پژوهش را رعایت نمود. در شکل زیر فرایندهای مورد اشاره را ببینید.



**تمرین ها :**

- ۱- فایل های Stata مربوط به داده هایی که در کامپیوتر شما پسوند dta دارند را لیست نمایید.
- ۲- تفاوت بین دستور browse و edit چیست؟
- ۳- جدول زیر قسمتی از یک پروژه تحقیقاتی است که در دو گروه از بیماران با و بدون فشارخون بالا انجام گردید. متغیر ها از var1 تا var8 نامگذاری شدند. داده های جدول را در نرم افزار وارد کنید:

Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
1	55	0	160	74	خانه دار	35	65
2	40	1	180	103	برق کار	58	56
1	65	1	164	100	بازنشسته	54	60
2	25	0	160	58	دانشجو	51	55
1	69	1	169	58	بازنشسته	51	64
1	35	0	160	75	خیاط	49	50
1	47	0	165	58	ورزشکار	48	51
2	43	0	180	85	خانه دار	58	59
1	21	1	168	56	دانشجو	45	48
1	68	1	160	68	ازاد	57	58
2	40	1	170	63	معلم	50	55
2	34	0	164	60	داروساز	57	58
1	43	0	156	64	کارمند	47	39
2	26	1	180	73	ورزشکار	58	61
2	38	1	170	60	مهندس	50	51
2	43	0	160	75	خیاط	44	52
1	58	1	190	73	بازنشسته	59	36
1	62	0	160	54	خانه دار	43	58
2	60	0	180	75	بازنشسته	40	49
1	28	1	168	63	دانشجو	50	50
1	34	0	174	60	دانشجو	50	52
1	38	1	170	85	مهندس	45	53
2	43	0	163	50	خانه دار	48	48
1	47	0	158	65	خیاط	48	50
2	51	0	154	63	خانه دار	40	58
2	56	1	178	74	مکانیک	50	52
2	68	1	160	80	بازنشسته	50	53
2	59	0	165	60	خانه دار	48	49
1	30	1	170	63	حسابدار	43	48

- ۴- مجموعه داده هایی را که وارد نموده اید آنها را به نام sh\_ahmadi1400 ذخیره و برچسب بزنید.
  - ۵- متغیرها را به ترتیب زیر تغییر نام و نامگذاری کنید.
- Var5 : weight ، Var4 : height ، Var3 : sex ، Var2 : age ، Var1 : group  
 Var8:HDL2 و Var7 : HDL1 ، Var6 : job

- ۶- به مجموعه داده ها برچسب " information " و به متغیرها بر حسب های زیر را منسوب کنید.  
 HDL1: High-density lipoprotein (HDL) before intervention, HDL2: after intervention
- ۷- به مقادیر متغیر جنس، بر حسب های زیر را منسوب کنید: 0 : female و 1 : male
- ۸- مجموعه داده ها را بر حسب مقادیر متغیر age ابتدا به صورت صعودی و سپس به صورت نزولی مرتب کنید.
- ۹- متغیر نمایه توده بدنی را بسازید و مجموعه داده ها را بر حسب مقادیر متغیر BMI ابتدا به صورت نزولی و سپس به صورت صعودی مرتب کنید.
- ۱۰- در قسمت نوشتن دستورات، با دستور <http://www.stata-press.com/data/r13/census10> use فایل داده های سرشماری را فراخوانی و باز نمایش دهید. خروجی زیر با دستور des را اجرا و انواع فرمت متغیرها را بررسی کنید.

```
. use http://www.stata-press.com/data/r13/census10
(1980 Census data by state)

. describe

Contains data from http://www.stata-press.com/data/r13/census10.dta
Observations:      50      1980 Census data by state
Variables:         4        9 Apr 2013 08:05
```

Variable name	Storage type	Display format	Value label	Variable label
state	str14	%14s		State
region	int	%8.0g	cenreg	Census region
pop	long	%11.0g		Population
medage	float	%9.0g		Median age

با دستور `list in 1/8` می توان برای اولین متغیر (state)، هشت مشاهده مربوط به آن را به صورت زیر لیست نمود.

```
. list in 1/8
```

	state	region	pop	medage
1.	Alabama	South	3893888	29.3
2.	Alaska	West	401851	26.1
3.	Arizona	West	2718215	29.2
4.	Arkansas	South	2286435	30.6
5.	California	West	23667902	29.9
6.	Colorado	West	2889964	28.6
7.	Connecticut	NE	3107576	32
8.	Delaware	South	594338	29.8

همانگونه که قابل مشاهده است چیدمان نام ایالت ها بصورت راست چین و فرمت %14s است. چنانچه بخواهیم این چیدمان را تغییر دهیم از دستور `format state %-14s` استفاده می شود. چنانچه دستور مذکور اجرا شود خروجی شکل بالا نیز به شکل زیر خواهد بود.

```
. format state %-14s
. . list in 1/8
```

	state	region	pop	medage
1.	Alabama	South	3893888	29.3
2.	Alaska	West	401851	26.1
3.	Arizona	West	2718215	29.2
4.	Arkansas	South	2286435	30.6
5.	California	West	23667902	29.9
6.	Colorado	West	2889964	28.6
7.	Connecticut	NE	3107576	32
8.	Delaware	South	594338	29.8

## بخش دوم:

# مبانی آمار توصیفی و استفاده از Stata

## فصل سوم: آمار توصیفی و توصیف داده ها در Stata.....

مبانی و اصول آمار توصیفی

آماره های توصیفی و خلاصه

استانداردسازی و نرمالیزه کردن داده ها

میانگین

فراوانی

رسم جدول ساده و متقاطع

تمرین ها

### • اهداف:

در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند مبانی آمار توصیفی و انواع شاخص ها را بیان و استفاده نماید.
- کاربر تعریف متغیر و تقسیم بندی انواع متغیرها را بیان نماید.
- کاربر بتواند جدول ساده، توافقی و متقاطع برای متغیرها ترسیم نماید.
- میانگین، میانه و نما را بتواند محاسبه نماید.
- شاخص های انحراف معیار و واریانس را محاسبه نماید.
- کاربر بتواند با نرم افزار استاتا آماره های توصیفی را محاسبه و گزارش نماید.

## مبانی و اصول آمار توصیفی

آمار از سه بخش تشکیل شده است. آمار توصیفی، برآورد یا تخمین و آمار تحلیلی. برای ورود پیدا کردن به مبحث اصول آمار توصیفی و استفاده از نرم افزار Stata، ابتدا به خلاصه ای از مباحث پایه آماری که در فهم آمار توصیفی و استفاده از نرم افزار اهمیت دارند، پرداخته می شود. در شکل زیر "برآورد" به عنوان پلی میان آمار توصیفی و تحلیلی شناخته شده است. برخی آمار را در دو بخش توصیفی و تحلیلی می شناسند.



### آمار (Statistics) و اهمیت آن:

آمار از کلمه لاتین Status به معنی چگونه قرار گرفتن یا وضعیت مشتق شده است. امروزه در زمینه های گوناگون علمی، اقتصادی، ورزشی، پزشکی، فرهنگی، علوم انسانی و... مورد استفاده قرار می گیرد. آمار مجموعه ای از تکنیک ها و روش هایی است که در جمع آوری و تجزیه و تحلیل اطلاعات آماری مورد استفاده قرار می گیرد به عبارتی آمار علمی است که مشخصات جوامع را از نظر کمی و با در نظر گرفتن مشخص کننده های آن جامعه مورد بررسی قرار می دهد.

هدف نهایی آمار انجام استنباط در مورد یک جامعه بر اساس اطلاعات موجود در نمونه ایی از آن جامعه است. از آنجا که برای اظهار نظر در مورد یک جامعه نمی توان کلیه افراد آن جامعه را مورد بررسی قرار داد معمولاً نمونه ایی از آن جمعیت انتخاب شده و از اطلاعات موجود در نمونه برای استنباط در مورد جامعه استفاده می شود. برای اینکه بتوان نتایج حاصل از نمونه را به جمعیتی که نمونه از آن گرفته شده است تعمیم داد، نمونه انتخاب شده باید معرف و گویای جمعیتی باشد که از آن انتخاب شده است.

اولین گام در آمار، توصیف مجموعه ایی از اندازه ها می باشد که به صورت تشکیل توزیع های فراوانی و معیارهای توصیفی عددی انجام می گیرد. در مرحله بعدی مسئله این است که چگونه بر اساس اطلاعات موجود در نمونه می توان در باره جامعه استنباط نمود که برای این کار توزیع احتمال کمیت های نمونه ایی یعنی توزیع های نمونه ایی در نظر گرفته شده و سپس آزمونهای آماری انجام می شود.

### تحقیق Research:

در معنای لغوی یعنی بازجویی و بررسی، واری و واقعیت و درست نمودن می باشد. به بررسی و مطالعه منظم، کنترل شده و آزمایشی قضیه های فرضی درباره روابط احتمالی بین پدیده ها با دیدگاه انتقادی نیز تحقیق گفته می شود. لذا تحقیق مجموعه ای از فعالیت های منظم پژوهشی است که هدف از انجام آن، کشف حقیقت و پاسخ به یک سوال است. تحقیق می تواند تحقیق بنیادی، کاربردی و یا توسعه ای باشد. و یا با توجه به ماهیت از تقسیم بندی دیگر استفاده و به تحقیق کمی یا تحقیق کیفی آن را تقسیم بندی نمود.

## آمار زیستی **Biostatistics**:

به کاربرد آمار در علوم زیستی و اینکه اگر در تجزیه و تحلیل های آماری از داده های موجود در علوم زیستی و پزشکی استفاده شود به آن آمار زیستی گویند.

## آمار حیاتی **Vital statistics**:

اگر در تجزیه و تحلیل های آماری از داده هایی مانند تولد، مرگ، ازدواج، طلاق و مهاجرت استفاده شود آن را آمار حیاتی نامند.

## داده **data**:

اولین چیزهایی که در هنگام مواجهه با کلمه داده به ذهن متبادر می شود، عبارت است از:

- یک مرحله قبل از اطلاعات است.
  - مجموعه ای از حروف و اعداد و چیزهایی هستند که خام هستند و هیچ پردازشی روی آنها انجام نشده است.
  - داده بعد از پردازش تبدیل به اطلاعات می شود.
  - در واقع داده ها یک سری اطلاعات خام هستند.
  - آنچه از حواس پنجگانه به انسان منتقل می شود.
- واقعیت این است که مثل هر مفهوم دیگری در جهان علم برای داده (data) نیز تعاریف متفاوتی ارائه شده است. به صورت کلی می توان گفت که داده ها موارد حاصله از مشاهدات، بررسی ها و تحقیقات به عمل آمده هستند. مشاهده و بررسی و تحقیق می تواند از ساده ترین چیزها مثل اندازه گیری میزان درجه حرارت بدن ما تا پیچیده ترین و گسترده ترین پروژه های تحقیقاتی مثل پرتاب یک شاتل به فضا باشد. ماموریت داده ها، ایجاد نمودهای قابل فهم و تحلیل از رخدادها، پدیده ها و مفاهیم محیط پیرامونی ما می باشد. برای امکان پذیر کردن فرآیند ذخیره سازی و پردازش داده، داده ها باید به صورتی بیان شوند و بدست آیند که بتوان آنها را به سرعت ثبت و ضبط کرده و منتقل نمود. این کار در فضای تکنولوژی های اطلاعاتی امروزه از طریق نمایش دودویی (باینری یا همان صفر و یکی) و ذخیره سازی در حافظه کامپیوترها انجام می شود. داده ها معمولاً خود را به صورت ویژگی ها و صفات موجودیت های تحت بررسی نشان می دهند، مثل نمره دانش آموزان در درس های ریاضی، علوم و تاریخ که در این حالت موجودیت های ما دانش آموزان بوده که سه ویژگی از آنها یعنی نمره ریاضی، نمره علوم و نمره تاریخ مورد نظر قرار گرفته است. از نظر واژگان شناسی ریشه کلمه داده به عبارت لاتین datum (به معنای اطلاع) برمی گردد. در اصل کلمه data جمع واژه datum بوده هرچند که امروزه در هنگام استفاده عموماً وقتی می گوئیم data منظورمان حالت مفرد آن بوده و جمع آن را به شکل داده ها بیان می کنیم.

## اطلاعات **Information**:

داده دربردارنده مواردی است که مخاطب می تواند آنها را به صورت کامل یا حداقل تقریبی فهم و درک کرده و به وی برای انجام یک فعالیت یا اتخاذ یک تصمیم کمک می کند. این فعالیت می تواند یک اقدام فیزیکی (مثل تنظیم آینه خودرو بر اساس قد راننده)، الکترونیکی (کلیک بر روی یک آیکن) یا ذهنی (بررسی میزان فاصله تا خودروی مقابل در حین رانندگی) باشد. اطلاعات (information) از پردازش داده بدست می آید. مثلاً از اندازه



گیری و کنار هم قرار دادن دمای بدن بیمار در طی زمان (در اینجا دمای بدن بیمار داده است) متوجه می شویم که روند بیماری در حال شدت گرفتن یا بهبود بوده یا اینکه تغییر خاصی در وضعیت بیمار هنوز اتفاق نیفتاده است. در واقع خلاصه سازی داده ها و انجام عملیات جبری روی آنها برای استفاده در تصمیم گیری های مبتنی بر شواهد است.

### جمعیت Population:

جمعیت موجودات بزرگترین مجموعه ایی از موجودات یا اشیا است که در یک زمان معین، مطلوب و مورد نظر ما قرار می گیرد. اگر مقدار متغیری را بر هر یک از موجودات جمعیت اندازه گیری کنیم جمعیت مقادیر آن متغیر را بوجود آورده ایم. جمعیت می تواند محدود یا نامحدود باشد. **جمعیت محدود** از مقادیر ثابت و مشخصی تشکیل می شود. در مقابل **جمعیت نامحدود** از یک ردیف بی انتهای مقادیر تشکیل می گردد یا به عبارتی تعداد افراد آن غیر قابل شمارش است.

### جامعه آماری Statistical Population:

مجموعه ایی از افراد یا اشیا هستند که لاقول در یک صفت مشترک می باشند.

### نمونه Sample:

هر بخشی از جمعیت را نمونه می نامند. اگر جمعیت مورد نظر ما از وزن های همه دانش آموزان دبیرستانی تشکیل شده باشد اما برای تجزیه و تحلیل تنها مجموعه ایی از وزن بخشی از دانش آموزان را جمع آوری کنیم تنها قسمتی از وزن های دانش آموزان (یک نمونه) را خواهیم داشت. از یک جمعیت می توان نمونه های گوناگونی را انتخاب نمود.

### متغیر Variable:

هر ویژگی یا خصوصیتی که در یک جامعه آماری مورد بررسی و مطالعه قرار گیرد و این ویژگی بتواند از هر عضو این جامعه به عضو دیگر متفاوت و قابل تغییر باشد به آن متغیر گویند. هر گاه این تفاوتها یا تغییرات متغیر به عواملی وابسته باشد که خارج از کنترل انسان باشند به آن متغیر تصادفی گفته می شود.

### به طور کلی متغیرها را می توان به دو دسته تقسیم نمود: متغیرهای کمی و متغیرهای کیفی

**متغیر کمی:** (Quantitative variable) متغیری است که قابل اندازه گیری است یا حاصل سنجش آن توسط مقادیر عددی بیان می شود. متغیرهای کمی دو دسته هستند:

**الف- متغیر کمی پیوسته یا متصل (Continuous):** کمیت یا متغیری است که بتواند بین دو مقدار خود تمامی اعداد حقیقی ممکن را اختیار کند یا هیچ فاصله ایی بین هیچ یک از دو مقدار قابل تصور وجود نداشته باشد. مانند قد، وزن، سن.

**ب- متغیر کمی گسسته یا منفصل (Discrete):** کمیتی است که بتواند بین دو مقدار خود مجموعه شمارش پذیری از اعداد یا زیر مجموعه‌ایی از آن را اختیار کند به عبارتی بین مقادیر قابل تصور یک متغیر فاصله وجود داشته و این فاصله به علت عدم وجود مقادیر بین آن مقادیر خاص است. مانند تعداد افراد خانوار، تعداد دندانهای پوسیده. اغلب در عمل به علت ماهیت ابزارهای اندازه‌گیری متغیرهایی که ماهیتاً پیوسته هستند به صورت گسسته بیان می‌شوند باید توجه داشت که این نوع بیان متغیر، ماهیت آن را تغییر نمی‌دهد.

**متغیر کیفی (Qualitative variable):** ویژگی‌هایی وجود دارند که نمی‌توان آنها را در مفهوم قد، وزن و سن اندازه‌گیری نمود بلکه تنها می‌توان آنها را طبقه‌بندی نمود. این نوع متغیر تفاوت اعضاء جامعه مورد بررسی را بر مبنای کیفیتی خاص بیان می‌کند. مانند متغیر جنسیت که افراد را تنها بر مبنای جنسیتشان می‌توان در دو گروه زن و مرد قرار داد یا گروه‌های خونی و یا اینکه می‌توان گفت افراد از خصوصیتی خاص برخوردار هستند یا خیر مانند بیمار بودن یا نبودن.

### مقیاس‌های اندازه‌گیری Scales of Measurement:

مقیاس اندازه‌گیری یک صفت یا متغیر راهنمایی است برای شیوه نشان دادن اطلاعات و خلاصه نمودن آنها. مقیاس اندازه‌گیری و دقت اندازه‌گیری متغیر مورد مطالعه در تعیین روشهای آماری ممکن برای تحلیل داده‌ها مهم هستند. نحوه اندازه‌گیری صفات از مهمترین عواملی هستند که قبل از انجام یک پژوهش یا تحقیق آماری باید مورد بررسی قرار گرفته و معین شوند. کلیه متغیرها معمولاً با چهار مقیاس زیر سنجیده می‌شوند.

#### مقیاس اسمی (Nominal Scale):

ساده‌ترین سطح اندازه‌گیری است و زمانی به کار می‌رود که داده‌ها را می‌توان در گروه‌ها یا دسته‌های مختلف قرار داد و شامل مشاهدات بدون رتبه و دو یا چند حالتی است. مانند جنسیت، گروههای خونی و انواع مدارک تحصیلی. در این مقیاس هر فرد جامعه نسبت به صفت یا ویژگی خاص تحت عنوان و اسم خاصی قرار می‌گیرد یعنی یا این مقیاس تنها می‌توان شی یا فرد مورد مطالعه را به گروهی منتسب نمود و نام آن گروه را بر آن نهاد. عملگرهای جبری برای این مقیاس کاربردی ندارند.

#### مقیاس رتبه‌ای یا ترتیبی (Ordinal Scale):

متغیرهایی که با این مقیاس سنجیده می‌شوند نه تنها از خاصیت مقیاس اسمی برخوردار می‌باشند یعنی می‌توان آنها را در دسته‌ها یا گروه‌هایی طبقه‌بندی نمود بلکه بین دسته‌ها یا گروه‌های آن ترتیب ذاتی وجود دارد و دسته‌بندی باید با مراعات تقدم و تاخر صورت گیرد. در این مقیاس هر فرد در مطالعه نسبت به سایرین رتبه‌ای را به خود اختصاص می‌دهد.

معمولاً نمی‌توان فاصله دو رتبه متمایز یا نسبت اندازه بین آنها را مشخص نمود و با وجود ترتیب ذاتی بین دسته‌ها، دامنه دسته‌های مختلف یکسان نیست اما برای تجزیه و تحلیل آماری لازم است فرض شود که بین گروه‌ها فواصل مساوی وجود دارد.

مقیاس اسمی و رتبه‌ایی برای سنجش متغیرهای کیفی مورد استفاده قرار می‌گیرد و با این سنجش‌ها کیفیت شخص یا شی یا مشاهدات دسته‌بندی شده و در گروه‌های مختلف توصیف می‌شوند. داده‌های اسمی را به صورت نسبت و درصد ارائه می‌کنند و از نمودارهای ستونی برای نمایش اطلاعات آنها استفاده می‌کنند. داده‌های رتبه‌ای را نیز با درصد و نسبت بیان می‌کنند و گاهی نتایج این نوع از داده‌ها با آماره میانه نیز ارائه می‌گردد.

### مقیاس فاصله‌ای (Interval Scale):

در این مقیاس، اطلاعات حاصل اندازه‌گیری است که مبنای آن یک مبدا قراردادی است مانند درجه حرارت که صفر آن قراردادی است. با به کار بردن یک مبدا قراردادی نه تنها نتیجه سنجش را برحسب واحدهای ثابت و معین می‌توان اندازه‌گیری نمود بلکه فاصله دو شی یا فرد را نیز از نظر صفت مورد بررسی می‌توان معین نمود. اما چون فاقد صفر ذاتی است نمی‌توان نسبت اندازه خاصیت مورد مطالعه را در افراد مشخص نمود.

### مقیاس نسبتی (Ratio Scale):

در مقیاس نسبتی اطلاعات حاصل بر پایه مقیاسی سنجیده می‌شود که دارای صفر ذاتی است. در این جا نه تنها می‌توان فرد مورد مطالعه را به رده‌ایی منتسب نمود یا رتبه آن را معین نمود یا فاصله آن را در رابطه با افراد دیگر معین نمود بلکه می‌توان نسبت اندازه خاصیت مورد مطالعه را در افراد معین نمود. مانند سن و قد.

ارائه اندازه‌های کمیت پیوسته به دقت ابزار اندازه‌گیری بستگی دارد و در گزارشات نهایی تا نزدیکترین رقم کفایت می‌کند. مقیاس‌های فاصله‌ایی و نسبتی برای سنجش متغیرهایی که ماهیت کمی دارند استفاده می‌شود. نتایج غالباً به صورت جداول و نمودار ارائه شده و برای خلاصه نمودن داده‌ها از میانگین و انحراف معیار استفاده می‌شود.

### گروه بندی نتایج مشاهدات و بیان آن توسط جدول

بعد از جمع آوری مشاهدات به منظور درک بهتر داده‌ها، لازم است که حاصل مشاهدات را با توجه به بعضی خصوصیات صفت مورد مطالعه در گروه‌های کاملاً متمایز قرار داد. صفات کیفی مانند جنس، نژاد، وجود یا عدم وجود بیماری را می‌توان به آسانی گروه بندی نمود، اما در مواردی که انتخاب گروه به قضاوت فرد مربوط باشد گروه بندی ساده نخواهد بود. در مورد مشاهداتی که کمیت حاصل به صورت اندازه‌های گسسته بیان می‌شود انتخاب گروه نسبتاً آسان است زیرا تعداد گروه‌ها محدود و متمایز بوده و در انتساب افراد به یکی از گروه‌ها مشکلی بوجود نمی‌آید. در مشاهداتی که کمیت حاصل به صورت اندازه‌های پیوسته باشد باید برخی نکات را در گروه بندی در نظر گرفت. گروه بندی این اندازه‌ها به هر نحو که انجام شود بعضی از اطلاعات در گروه بندی از دست خواهد رفت. برای مثال اگر وزن بدو تولد کلیه نوزادان بدنیا آمده در یک بیمارستان را در چهار گروه کمتر از ۲۵۰۰ گرم، ۲۵۰۰-۳۰۰۰ گرم و ۳۰۰۰-۳۵۰۰ گرم و بیشتر از ۳۵۰۰ گرم گروه بندی کنیم تعداد نوزادانی که در هر گروه قرار می‌گیرند معلوم خواهد شد اما دیگر اطلاعاتی در مورد اینکه مثلاً چند نوزاد با وزن ۲۷۵۰ گرم وجود دارد را نخواهیم داشت. برای گروه بندی داده‌های پیوسته لازم است که تعداد گروه‌ها و عرض دامنه هر گروه تعیین شود. در مورد برخی از متغیرها فواصل معینی برای گروه‌ها وجود دارند که مورد پذیرش جوامع بین‌المللی بوده و برای گروه بندی این متغیرها بهتر است از گروه‌های استاندارد آنها استفاده شود. برای مثال اغلب مواقع سن

را در گروه های ۵ ساله یا ده ساله گروه بندی می کنند یا گروه های مشخصی برای مقادیر شاخص توده بدنی و همچنین تعیین گروههای فشارخون سیستولی و دیاستولی برای مشخص نمودن افرادی که در گروههای مختلف پرفشاری خون قرار می گیرند وجود دارد.

در صورتی که گروه بندی مشخصی برای متغیر مورد نظر وجود نداشته باشد یا بخواهیم که مقادیر صفت مورد نظر را بدون در نظر گرفتن گروههای تعیین شده، گروه بندی نمائیم، کتب مختلف آماری تعداد گروههای متفاوتی را پیشنهاد می کنند. استفاده از طبقات خیلی کم می تواند سبب از دست رفتن اطلاعات اساسی شود و اگر از تعداد گروههای بیشتری استفاده شود هدف اصلی از گروه بندی که همانا خلاصه نمودن داده هاست از بین می رود. در عمل شاید بهتر باشد که ابتدا تعداد طبقات را بیشتر در نظر گرفت که در صورت لزوم بتوان این طبقات را در هم ادغام نمود و جدول مناسبتری را بدست آورد. به طور کلی اگر در گروه بندی نتایج حاصل از مشاهدات از ۸ تا ۱۵ گروه استفاده شود اطلاعات قابل توجهی از بین نخواهد رفت. در اغلب مطالعات بهتر است از گروه های با فاصله های مساوی استفاده شود اما استفاده از فاصله های نامساوی نیز مرسوم است.

برای تعیین تعداد گروهها می توان از **قاعده یا دستور استورگس** استفاده نمود. در این قاعده تعداد گروهها از رابطه زیر بدست می آید.

$$K = 1 + 3.322(\log_{10} n)$$

در این رابطه  $K$  تعداد گروه ها و  $n$  اندازه نمونه می باشد. جواب این رابطه یک توصیه و رهنمود پیشنهادی است نه نهایی. یعنی برای مثال اگر اندازه نمونه ۵۰ نفر باشد جواب این رابطه ۶/۶۴ خواهد بود که بالطبع باید تعداد گروهها را ۵ یا ۶ انتخاب نمود.

بعد از تعیین تعداد گروهها، لازم است عرض دامنه هر گروه مشخص شود. برای تعیین عرض دامنه گروه ها از رابطه زیر استفاده می شود. در مواردی بجای  $h$  نیز از  $w$  استفاده می شود.

$$h = \frac{R}{K}$$

در این رابطه  $W$  یا  $h$  عرض دامنه گروه و  $R$  دامنه تغییرات یا تفاوت بین کمترین و بیشترین مقدار مشاهده شده ( $R = X_{\max} - X_{\min}$ ) می باشد. عرض دامنه گروه نیز معمولاً عدد صحیحی بدست نمی آید ولی می توان با گرد کردن عدد مناسبی را بدست آورد مگر اینکه دقت اندازه گیری به کار رفته زیاد بوده و یا اندازه مشاهدات به طور کلی کوچک باشند.

بعد از تعیین تعداد گروهها و عرض دامنه هر گروه، تعداد مشاهدات یا افراد هر گروه را که **فراوانی مطلق** آن گروه نامیده می شود، معین می کنند. جدولی که به این ترتیب بدست می آید **جدول توزیع فراوانی** نامیده می شود. برای تهیه جدول مذکور ابتدا دو ستون ایجاد می شود. در ستون اول گروه ها و در ستون دوم در مقابل هر گروه فراوانی مشاهداتی را که در آن گروه قرار می گیرند نوشته می شود. در واقع جدول توزیع فراوانی مجموعه ایی از طبقات از قبل تعیین شده همراه با شمارش تعداد مشاهداتی است که مقادیر آنها درون یک فاصله هر طبقه قرار می گیرند.

در تشکیل توزیع فراوانی، برای تعیین تعداد مشاهدات هر گروه، ابتدا مشاهدات را یکی یکی خوانده و بعد از مشخص نمودن گروه مربوطه، یک چوب خط در مقابل گروه رسم می شود.

برای راحتی شمارش چوب خط پنجم را به صورت مورب روی چهار چوب خط قبلی رسم و بعد از اتمام کار، تعداد چوب خط های هر گروه را شمرده و یادداشت می شوند.

مثال: ارقام زیر مربوط به یک مطالعه است که از اندازه گیری ذرات مواد شناور به میکروگرم به متر مکعب در نمونه های هوای ۵۷ شهر بدست آمده است

۶۸-۶۳-۴۲-۲۷-۳۰-۳۶-۲۸-۳۲-۷۹-۲۹-۲۲-۲۳-۲۴-۲۵-۴۴-۶۵-۴۳-۲۵-۷۴-۵۱-۳۶-۴۲-۲۸-۳۱-۲۸-۲۵-  
۴۵-۱۲-۵۷-۵۱-۱۲-۳۲-۴۹-۳۸-۴۲-۲۷-۳۱-۵۰-۳۸-۲۱-۱۶-۲۴-۶۹-۴۷-۲۳-۲۲-۴۳-۲۷-۴۹-۲۸-۲۳-  
۱۹-۴۶-۳۰-۴۳-۴۹-۱۲

برای تشکیل جدول توزیع فراوانی این داده ها همانطور که در این مشاهدات ملاحظه می شود کمترین مقدار مشاهده شده عدد ۱۲ و بیشترین مقدار ۷۹ است بنا براین دامنه تغییرات برابر است با:  $79-12=67$   
با توجه به قاعده استورگس تعداد تقریبی طبقات عبارت خواهد بود از:

$$K=1+ \sqrt[3]{322 \times (\log_{10} 57)} = 6/8$$

که می توان تعداد طبقات را ۷ انتخاب نمود. سپس عرض دامنه گروه یا فاصله طبقات را از هم با توجه به رابطه داده شده محاسبه می کنیم:

$$h = \frac{R}{K} = \frac{67}{7} = 9.57$$

که تقریباً برابر ۱۰ است با توجه به اینکه عرض دامنه گروه برابر ۱۰ و تعداد گروهها ۷

می باشد دامنه تغییرات  $10 \times 7 = 70$  می شود که سه واحد بیشتر از دامنه تغییرات محاسبه شده برای داده ها می باشد به این ترتیب اگر اولین گروه را از عدد ۱۰ شروع کنیم و فاصله ۱۰ تا را برای گروه بندی رعایت کنیم گروه بندی مناسبی به وجود خواهد آمد. جدول توزیع فراوانی مذکور به ترتیب زیر تشکیل می شود:

گروه ها	نمایش چوب خطی	$f_i$
۲۰-۱۰		۵
۳۰-۲۰		۱۹
۴۰-۳۰		۱۰
۵۰-۴۰		۱۳
۶۰-۵۰		۴
۷۰-۶۰		۴
۸۰-۷۰		۲
جمع		۵۷

هر گروه از داده ها در جدول دارای یک فراوانی است که به طور معمول آن را با  $f_i$  نمایش می دهند و آن را فراوانی مطلق آن گروه می نامند. اگر تعداد کل داده ها را با  $N$  نشان دهیم **فراوانی نسبی** هر گروه از داده ها برابر است با  $\frac{f_i}{N}$  که با ضرب نمودن آن در ۱۰۰، درصد فراوانی هر گروه بدست می آید. میتوان در هر جدول توزیع فراوانی فراوانی تجمعی را برای هر گروه محاسبه نمود که فراوانی تجمعی هر گروه از جمع کردن فراوانی همه گروههای

قبلی و فراوانی گروه فعلی به دست می آید. فراوانی تجمعی را معمولاً با  $F_i$  نشان می دهند برای فراوانی تجمعی نیز می توان فراوانی نسبی تجمعی و درصد آن را محاسبه نمود.

همچنین اغلب لازم است برای انجام محاسبات آماری در جداول توزیع فراوانی پیوسته، نماینده هر طبقه یا به عبارتی مقدار وسط طبقات محاسبه شود. برای این کار کرانه پایین هر طبقه را با کرانه بالای آن جمع و حاصل را بر دو تقسیم و نماینده آن طبقه بدست می آید. معمولاً آن را با  $x_i$  نشان می دهند. جدول زیر این محاسبه ها را نشان می دهد.

گروه ها		فراوانی مطلق $f_i$	فراوانی نسبی $f_i/N$	درصد فراوانی نسبی هر گروه $(f_i/N) \times 100$	فراوانی تجمعی $F_i$	نماینده طبقه $x_i$	فراوانی نسبی تجمعی $F_i/N$	درصد فراوانی نسبی تجمعی هر گروه $(F_i/N) \times 100$
۲۰-۱۰		۵	۰/۰۹	۹٪	۵	۱۵	۰/۰۹	۹٪
۳۰-۲۰		۱۹	۰/۳۳	۳۳٪	۲۴	۲۵	۰/۴۲	۴۲٪
۴۰-۳۰		۱۰	۰/۱۸	۱۸٪	۳۴	۳۵	۰/۶۰	۶۰٪
۵۰-۴۰		۱۳	۰/۲۳	۲۳٪	۴۷	۴۵	۰/۸۳	۸۳٪
۶۰-۵۰		۴	۰/۰۷	۷٪	۵۱	۵۵	۰/۹۰	۹۰٪
۷۰-۶۰		۴	۰/۰۷	۷٪	۵۵	۶۵	۰/۹۷	۹۷٪
۸۰-۷۰		۲	۰/۰۳	۳٪	۵۷	۷۵	۱	۱۰۰٪
جمع		۵۷	۱	۱۰۰				

### رعایت نکات زیر در تنظیم جدول های آماری برای گزارش نتایج مطلوب است:

- ۱- جدول باید حتی الامکان ساده باشد. بیان دو یا سه جدول ساده بر نمایش یک جدول بزرگ که شامل متغیرهای زیادی است ارجحیت دارد.
- ۲- جدول باید گویای محتوای خود باشد. لازم است کلیه علائم و نشانه هایی که در جدول به کار رفته اند در زیر جدول مشخص شده و کلیه سطر و ستون ها نامگذاری شده و واحد هر یک مشخص شود.
- ۳- باید برای سطرها و ستون ها، سطر و ستون جمع منظور گردد و با توجه به اهمیت جدول، اولین یا آخرین ستون یا سطر را به ستون یا سطر جمع اختصاص داد.
- ۴- باید در زیر جدول به ذکر منبع اطلاعات جدول اقدام شود که خواننده بتواند اعتماد خود را نسبت به محتوای جدول ارزیابی کند.
- ۵- باید برای هر جدول شماره ای در نظر گرفته شود که بتوان به آسانی با بیان شماره توجه خواننده را به جدول مورد نظر جلب نمود.
- ۶- جدول بایستی عنوان داشته باشد و در نوشتن عنوان جدول از جمله کوتاهی که شامل مفاهیم چه، کجا، چه وقت است استفاده شود.

**آمار توصیفی :** Descriptive Statistic

در مطالب پیشگفت، انواع متغیرها و مقیاسهای اندازه گیری متغیرها و گروه بندی داده ها و نحوه بیان داده ها توسط جدول ارایه شد. در این بخش آماره های رایج و متداولی را که برای خلاصه نمودن داده ها یا توزیع صفات یک مجموعه از داده ها مورد استفاده قرار می گیرد معرفی می شوند. در آمار توصیفی شاخص ها، نمودارها، نقشه ها و جداول بحث و ارایه می شوند که به این کار آمار توصیفی گفته می شود.

آماره ها یا شاخصهای توصیفی را می توان از داده های یک نمونه یا از داده های یک جمعیت محاسبه نمود. اگر چنانچه از داده های جمعیت برای محاسبه شاخص های توصیفی استفاده شود آنها را پارامتر می نامند. چنانچه از داده های نمونه برای محاسبه شاخص های توصیفی استفاده شود به آنها شاخص آماری یا آماره گفته می شود.



در مورد هر مجموعه از داده ها دو دسته شاخص توصیفی شامل شاخص های تمایل مرکزی و شاخص های پراکندگی و نوع توزیع آنها را می توان محاسبه نمود. در مجموع در آمار توصیفی حدود ۱۳ شاخص مطابق شکل بالا مورد محاسبه و استفاده قرار می گیرند.

### شاخص های تمایل مرکزی (Measures of Central Tendency):

این شاخص ها آماره هایی هستند که محل تمرکز توزیع را برای اندازه گیری های عددی و ترتیبی توصیف می کنند، به عبارتی جایگاه مرکز یک توزیع را بیان می کنند. یک توزیع Distribution شامل مقادیر مختلف اندازه گیری شده از یک خصوصیت همراه با فراوانی مربوط به هر اندازه گیری است. قبل از اینکه به بیان شاخصهای تمایل مرکزی بپردازیم ابتدا نماد  $\sum$  را معرفی می کنیم. این نماد که آن را زیگما می نامند برای نشان دادن جمع تعدادی داده مورد استفاده قرار می گیرد و دارای خواص زیر است.

$$1- \sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

$$2- \sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$$

$$3- \sum_{i=1}^n cx_i = c \sum_{i=1}^n x_i$$

### میانگین حسابی (Arithmetic):

#### الف: داده های خام

برای بدست آوردن میانگین حسابی، ابتدا همه داده ها را با هم جمع کرده و حاصل را بر مجموع تعداد داده ها تقسیم می کنند. اگر داده ها از طریق سرشماری بدست آمده باشند میانگین را با نماد  $\mu$  نشان می دهند و اگر داده ها مربوط به مقادیر صفت در یک نمونه باشند آن را با  $\bar{X}$  نمایش می دهند. برای محاسبه میانگین جمعیت از فرمول زیر استفاده می شود:

$$\mu = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum X_i}{N}$$

و برای محاسبه میانگین نمونه از فرمول زیر استفاده می شود.

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$$

#### ب- داده های گروه بندی شده:

برای محاسبه میانگین داده های گروه بندی شده، در مورد صفات کمی گسسته که در آن  $X_i$  مقدار صفت برای گروه  $i$  ام و  $N_i$  فراوانی مربوط به همان گروه می باشد میانگین از رابطه زیر بدست می آید:

$$\mu = \frac{N_1 X_1 + N_2 X_2 + \dots + N_k X_k}{N_1 + N_2 + \dots + N_k} = \frac{\sum N_i X_i}{N}$$

که  $N$  نشان دهنده تعداد کل مشاهدات است. در مورد صفات کمی پیوسته نیز از همین رابطه استفاده می شود ولی فرض بر این است که همه مقادیر دامنه گروه در نقطه وسط دامنه گروه قرار گرفته و خلاصه شده اند. در نتیجه از نماینده هر گروه یا نقطه وسط گروه برای محاسبه میانگین استفاده می شود به این ترتیب محاسبه



میانگین بطور تقریب امکان پذیر است. با توجه به نکته ای که قبلاً ذکر شد رابطه فوق برای محاسبه میانگین جمعیت می باشد. برای محاسبه میانگین نمونه از رابطه زیر استفاده می کنیم:

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

که  $f_i$  فراوانی مطلق هر گروه می باشد.

مثال ۱: میانگین داده های مربوط به مثال قبلی به صورت زیر محاسبه می شود.

گروه ها	$f_i$	$x_i$	$f_i x_i$
۲۰-۱۰	۵	۱۵	۷۵
۳۰-۲۰	۱۹	۲۵	۴۷۵
۴۰-۳۰	۱۰	۳۵	۳۵۰
۵۰-۴۰	۱۳	۴۵	۵۸۵
۶۰-۵۰	۴	۵۵	۲۲۰
۷۰-۶۰	۴	۶۵	۲۶۰
۸۰-۷۰	۲	۷۵	۱۵۰
جمع	۵۷		۲۱۱۵

$$\bar{X} = \frac{\sum f_i x_i}{n} = \frac{2115}{57} = 37.1$$

### میانگین هندسی (Geometric mean):

میانگین هندسی  $N$  عدد عبارتست از ریشه  $N$  ام حاصل ضرب اعداد. میانگین هندسی را  $\bar{X}_G$  نشان می دهند. از این میانگین زمانی استفاده می کنیم که داده های بدست آمده نسبت، درصد، نرخ رشد، PH باشد یا بخواهند جمعیت یک منطقه را مشخص نمایند. فرمول مورد استفاده به صورت زیر است.

$$\bar{X}_G = \sqrt[n]{x_1 x_2 \dots x_n}$$

مثال: میانگین هندسی سه عدد ۲ و ۴ و ۸ عبارتست از:

$$\bar{X}_G = \sqrt[3]{x_1 x_2 x_3} = \sqrt[3]{2 \times 4 \times 8} = 4$$

مثال: در سال ۱۹۵۵ جمعیت شهری ۲۵۰ هزار نفر و در سال ۱۹۶۵ برابر با ۴۹۰ هزار نفر بوده است. جمعیت این شهر در سال ۱۹۶۰ چند نفر بوده است؟

$$\bar{X}_G = \sqrt[2]{250000 \times 490000} = 500 \times 700 = 350000$$

جمعیت این شهر در سال ۱۹۶۰ برابر با ۳۵۰ هزار نفر بوده است.

### میانگین همساز یا هارمونیک (Harmonic Mean):

اگر مقیاس سنجش داده ها ترکیبی باشد (مانند کیلومتر در ساعت، متر در ثانیه و غیره) برای یافتن متوسط آنها از میانگین هارمونیک استفاده می شود. برای محاسبه این میانگین، تعداد اعداد بر مجموع معکوس آنها تقسیم می شود. فرمول مورد استفاده به صورت زیر است.

$$\bar{X}_H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N}} = \frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$$

مثال: خودرویی مسافتی را با سرعت ۱۲۰ کیلومتر در ساعت رفته و با سرعت ۱۰۰ کیلومتر در ساعت بازگشته است. سرعت متوسط این خودرو چقدر بوده است؟

$$\bar{X}_H = \frac{2}{\frac{1}{120} + \frac{1}{100}} = \frac{2}{\frac{11}{600}} = 109/09$$

خواص میانگین حسابی:

اگر  $Z_i = X_i + Y_i$  آنگاه  $\mu_Z = \mu_X + \mu_Y$

اگر  $Z_i = CX_i$  آنگاه  $\mu_Z = C \mu_X$  (C عددی است ثابت)

اگر  $Z_i = CX_i + A$  آنگاه  $\mu_Z = C\mu_X + A$  (C و A ثابت هستند).

**نما یا مد Mode:**

نمای یک مجموعه از مقادیر داده ها، مقداری است که در آن مجموعه، بیشتر از همه تکرار شده است یا به عبارتی داده یا داده هایی است که بیشترین فراوانی را دارند. نما شاخصی است که برای متغیرهای کیفی نیز مورد استفاده قرار می گیرد.

درمورد داده های کمی پیوسته که گروه بندی شده اند نما در طبقه ای قرار دارد که بیشترین فراوانی را دارد. بعد از تعیین گروهی که نما در آن قرار دارد نما را با استفاده از رابطه زیر محاسبه می کنیم.

$$Mod = L + \left( \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \times h \right)$$

که در این رابطه  $f_i$  فراوانی مطلق گروه نمایی،  $f_{i+1}$  فراوانی مطلق گروه ما بعد گروه نمایی و  $f_{i-1}$  فراوانی مطلق گروه ما قبل گروه نمایی، L کرانه پایین گروه نمایی و h عرض دامنه گروه نمایی می باشد.

مثال: در جدول زیر نما را بدست آورید.

گروه ها	$f_i$	$x_i$
۲۰-۱۰	۵	۱۵
۳۰-۲۰	۱۹	۲۵
۴۰-۳۰	۱۰	۳۵
۵۰-۴۰	۱۳	۴۵
۶۰-۵۰	۴	۵۵
۷۰-۶۰	۴	۶۵
۸۰-۷۰	۲	۷۵
جمع	۵۷	

با توجه به این جدول معلوم می گردد که گروه دوم دارای بیشترین فراوانی می باشد و به عبارتی نما در این گروه قرار می گیرد. پس:

$$Mod = L + \left( \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \times h \right) = 20 + \left( \frac{19 - 5}{(19 - 5) + (19 - 10)} \times 10 \right) = 20 + 6.08 = 26.08$$

### میانۀ Median:

میانۀ مقدار یا عددی است که ۵۰٪ داده ها بیشتر از آن و ۵۰٪ داده ها کمتر از آن هستند. در مورد داده های خام برای بدست آوردن میانۀ اعداد را به ترتیب صعودی یا نزولی مرتب کرده و سپس عدد وسط را پیدا می کنند. اگر تعداد داده ها فرد باشد عدد وسطی برابر با  $(N+1)/2$  داده ها خواهد بود. اگر تعداد داده ها زوج باشد متوسط دو عدد وسطی مقدار میانۀ خواهد بود. مثال: میانۀ اعداد زیر را بدست آورید.

۱۵ ۲۰ ۱۵ ۷ ۱ ۴ ۳ ۱ ۱۱ ۷ ۷

ابتدا اعداد را به ترتیب صعودی یا نزولی مرتب می کنیم.

۱ ۱ ۳ ۴ ۷ ۷ ۷ ۱۱ ۱۵ ۱۵ ۲۰

با توجه به این که ۱۱ عدد داریم  $(1+11)/2=6$  میانۀ برابر با عدد ششم است که برابر با ۷ می باشد. اگر داده ها گروه بندی شده و گسسته باشد برای پیدا کردن میانۀ ابتدا فراوانیهای تجمعی را محاسبه می کنیم تا بتوان موقعیت و ردیف اندازه صفات را مشخص نمود. سپس با توجه به ستون فراوانیهای تجمعی اندازه ایی را که در ردیف  $N/2$  قرار دارد به عنوان میانۀ انتخاب می شود. در مورد داده های کمی پیوسته محاسبه میانۀ تنها به صورت تقریبی امکان پذیر است. در این حالت ابتدا فراوانیهای تجمعی محاسبه می شود. سپس گروهی را که میانۀ در آن گروه قرار دارد مشخص می شود. بعد با استفاده از رابطه زیر میانۀ محاسبه می شود.

$$Med = L + \frac{\frac{N}{2} - F_{i-1}}{f_i} \times h$$

که در این رابطه  $L$  کران پایینی گروه میانۀ،  $F_{i-1}$  و  $f_i$  فراوانی مطلق گروه میانۀ و  $h$  عرض دامنه گروه میانۀ است. مثال: میانۀ اعداد جدول زیر را بدست آورید.

گروه ها	$f_i$	$x_i$	$F_i$
۲۰-۱۰	۵	۱۵	۵
۳۰-۲۰	۱۹	۲۵	۲۴
۴۰-۳۰	۱۰	۳۵	۳۴
۵۰-۴۰	۱۳	۴۵	۴۷
۶۰-۵۰	۴	۵۵	۵۱
۷۰-۶۰	۴	۶۵	۵۵
۸۰-۷۰	۲	۷۵	۵۷
جمع	۵۷		

$$\frac{N}{2} = \frac{57}{2} = 28.5$$

با نگاهی به ستون فراوانی تجمعی معلوم می شود که ۲۴ داده در دو گروه اول قرار دارند پس داده ۲۸ ام در گروه سوم قرار می گیرد. با جایگذاری اعداد مربوطه در فرمول بالا، مقدار میانه به صورت زیر محاسبه می شود.

$$Med = L + \frac{\frac{N}{2} - F_{i-1}}{f_i} \times h = 30 + \left( \frac{28.5 - 24}{10} \times 10 \right) = 34.5$$

### شاخص های پراکندگی:

این شاخص ها پراکندگی یا تغییر مشاهدات پیرامون میانگین را نشان می دهند. شاخص های مهم که در پزشکی بالینی و اپیدمیولوژی بیشتر استفاده می شود عبارتند از:

### الف) دامنه (ب) انحراف معیار (ج) واریانس (د) خطای معیار میانگین (ذ) ضریب تعیین

الف) دامنه: تفاوت حداکثر و حداقل مشاهدات است. از فرمول زیر محاسبه می شود. فقط از دو داده استفاده

$$R = X_{MAX} - X_{MIN} \quad \text{می شود.}$$

ب) انحراف معیار:

شاخصی است که نشان دهنده پراکندگی یا انتشار داده ها حول میانگین است. انحراف معیار در نمونه، جذر واریانس یا پراش می باشد.

انحراف معیار این امکان را فراهم می کند تا میزان تغییرپذیری منتظره در داده ها را برآورد کنیم. فرمول محاسبه آن به شرح زیر است.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 \quad \sqrt{S^2} = S = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

ج) واریانس، توان دوم انحراف معیار است. فرمول آن عبارت است از:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

### ویژگی های واریانس نمونه:

۱- واریانس عدد ثابت C برابر با صفر است.

۲- اگر مقدار ثابت  $\alpha$  رابه مشاهدات اضافه یا از آنها کم کنیم واریانس تغییر نمی کند.

۳- اگر مشاهدات در مقدار ثابت K ضرب یا برآن تقسیم شود واریانس جدید از ضرب یا تقسیم واریانس قدیم در  $K^2$  بدست می آید.

د) خطای معیار میانگین: این شاخص نقش مهمی بازی می کند. خطای معیار بزرگ دلالت بر بی دقت بودن برآورد است. کوچک بودن آن دلالت بر دقیق بودن برآورد است. با استفاده از افزایش حجم نمونه و کاهش تغییر پذیری در داده ها میتوان این شاخص را کاهش و عبارتی دقت برآورد را افزایش داد. یکی از کاربردهای مهم آن برای محاسبه حدود اطمینان است. کاربرد دیگر آن برآورد انحراف معیار جمعیت است. با تقسیم انحراف معیار بر جذر حجم نمونه بر اساس فرمول زیر محاسبه می شود.

$$Se \text{ (برآورد)} = \frac{Sd \text{ (صفت)}}{\sqrt{n}}$$

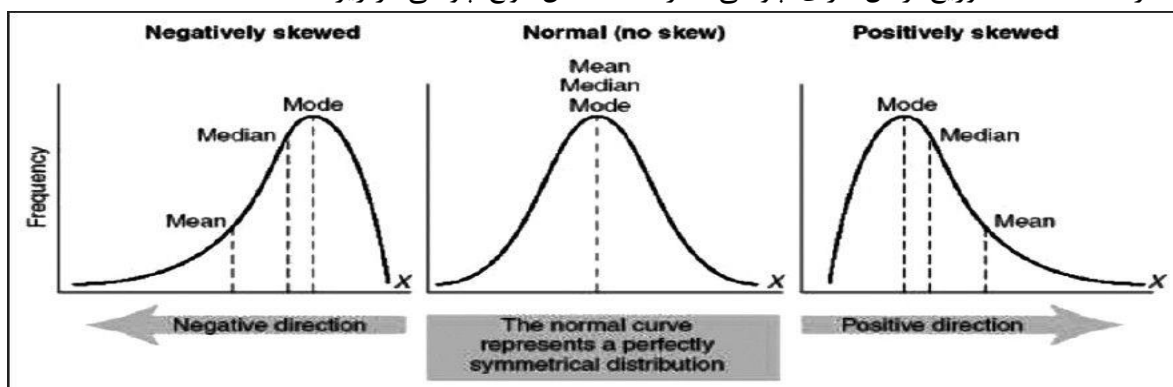
ذ) ضریب تغییر: نسبت انحراف معیار به میانگین است. به واحد اندازه گیری بستگی ندارد. برای مقایسه دو صفت از یک جامعه با واحدهای اندازه گیری متفاوت مورد استفاده قرار می گیرد. مجموعه مشاهداتی که دارای C.V کمتری است از سازگاری و همگنی بیشتری برخوردار هستند.

$$C.V = \frac{S}{\bar{x}} = \frac{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}}{\frac{1}{n} \sum x_i} = \frac{\sqrt{n \sum (x_i - \bar{x})^2}}{\sum x_i}$$

### کاربرد شاخص ها

برای استانداردسازی (standardization) داده ها، مقدار هر متغیر را از میانگین داده ها کم نموده و بر انحراف معیار تقسیم می کنند. برای نرمال کردن (normalization) داده ها مقادیر هر متغیر را از میانگین داده ها کم نموده و بر دامنه تقسیم می کنند. دامنه (range) تفاضل بیشترین و کمترین مقادیر مشاهدات است. در واقع با استانداردسازی داده ها واحد اندازه گیری حذف و داده ها به سمت مرکز هدایت می شوند.

**چولگی** یک معیار تقارن یا عدم تقارن در یک توزیع آماری است. در آمار، چولگی درجه ایی از عدم تقارن مشاهده شده در یک توزیع احتمالاتی را نشان می دهد که از **توزیع نرمال** متقارن (bell curve) در یک مجموعه داده منحرف شده است. توزیع نرمال دارای چولگی صفر است. شکل انواع چولگی در زیر آمده است.



### آماره های توصیفی و خلاصه در Stata:

آماره های توصیفی (مثل میانگین، میانه و انحراف استاندارد متغیرها) با استفاده از فرمان summarize، به دست می آیند. برای رسم جدول های توزیع فراوانی و انجام آزمون های استقلال و به دست آوردن آماره های بیان کننده ی میزان همبستگی از فرمان tabulate و برای رسم جدول هایی که خانه های آنها شامل آماره هایی مثل فراوانی، مجموع، میانگین یا میانه است، از فرمان table استفاده می شود.

### آماره های خلاصه برای متغیرها

برای به دست آوردن آماره های خلاصه (میانگین، انحراف استاندارد، می نیمم و ماکزیمم) برای یک متغیر از فرمان summarize استفاده می شود.

نام متغیر summarize

فرمان summarize مقادیر گم شده را در نظر نمی گیرد. اگر در فرمان summarize نام متغیر یا متغیرها ذکر نشود، Stata آماره های خلاصه را برای تمام متغیرهای مجموعه داده ها ارائه می دهد. در صورتی که فرمان summarize همراه گزینه ی detail به کار رود، خروجی شامل اطلاعات بیشتری، به شرح زیر خواهد بود:

- چندکها
- چهار مقدار از کوچکترین و چهار مقدار از بزرگترین مقادیر متغیر
- جمع وزن ها
- واریانس
- چولگی (جهت و مقدار نامتقارنی) و کشیدگی

### رسم جدول های فراوانی یک طرفه و دوطرفه

زمانی که مجموعه داده ها شامل متغیرهای رسته ای است، برای توصیف داده ها از جدول های فراوانی استفاده می شود. برای رسم یک جدول فراوانی یک طرفه فرمان tabulate به صورت زیر به کار می رود:

نام متغیر رسته ای tabulate

در صورتی که بعد از فرمان tabulate نام دو متغیر ذکر شود، جدول دوطرفه رسم خواهد شد. در این جدول سطوح متغیر اول در سطرها و سطوح متغیر دوم در ستون ها قرار می گیرد. برخی از گزینه های مفید فرمان tabulate، عبارتند از:

برای هر خانه ی جدول، درصد کل را نشان می دهد.	<b>cell</b>
برای هر خانه ی جدول، درصد ستونی را نشان می دهد.	<b>column</b>
برای هر خانه ی جدول، درصد سطری را نشان می دهد.	<b>row</b>
فراوانی های جدول را نشان نمی دهد.	<b>nofreq</b>
آزمون خی- دو یا کای اسکور پیرسون را برای آزمون استقلال متغیرهای سطر و ستون جدول انجام می دهد.	<b>chi2</b>
آزمون دقیق فیشر را برای آزمون فرض استقلال انجام می دهد.	<b>exact</b>
تمام آماره های بیانگر میزان همبستگی و آماره های لازم برای آزمون های استقلال را ارائه می دهد.	<b>all</b>
مقدار گم شده را در تحلیل ها لحاظ می کند.	<b>missing</b>

### جدول های چند طرفه و رسم چند جدول به طور همزمان

وقتی لازم است چند جدول فراوانی برای چند متغیر به طور همزمان رسم شوند، به جای این که برای هر جدول فرمان tabulate به طور جداگانه به کار رود، می توان از فرمان های tab1 و tab2 استفاده کرد. فرمان tab1 چند جدول یک طرفه و فرمان tab2 چند جدول دوطرفه را به طور همزمان رسم می کند.

نام متغیرها tab1

نام متغیرها tab2

به منظور رسم جدول سه طرفه با استفاده از فرمان `tabulate` می توان ابتدا مجموعه داده ها را برحسب یک متغیر مرتب و بعد با بکارگیری `by` به صورت زیر جدول سه طرفه را رسم کرد:

متغیر اول `sort`

متغیر سوم    متغیر دوم `tabulate` : متغیر اول `by`

از این روش برای تشکیل جدول های چندطرفه هم می توان استفاده کرد. به این ترتیب که ابتدا مجموعه داده ها برحسب همه ی متغیرهای موجود به جز دو متغیر، مرتب می شوند و سپس از عبارت `by` به ترتیبی که گفته شد، استفاده می شود.

متغیر n-ام , ... , متغیر اول `sort`

متغیر (n+2)ام    متغیر (n+1)ام `tabulate` : متغیر n-ام ... متغیر اول متغیر دوم `by`

اگر آزمون های آماری و مقدار درصدها مورد نیاز نباشد، راه ساده تر برای ساختن جدول های چندطرفه، استفاده از فرمان `table` است. دو گزینه ی مفید که همراه فرمان `table` به کار می روند عبارتند از:

با استفاده از این گزینه می توان سطوح حداکثر چهار متغیر را در سطرهای جدول قرار داد. **by** (نام متغیرها)

**contents** (کلمه کلیدی) این گزینه با توجه به کلمه ی کلیدی آن، تعیین کننده ی محتوای خانه های جدول است.

برخی از کلمات کلیدی که به همراه گزینه ی `contents` ( ) به کار می روند، عبارتند از:

<b>freq</b>	برای ارائه ی فراوانی
<b>mean</b>	نام متغیر برای ارائه ی میانگین متغیر مورد نظر
<b>sd</b>	نام متغیر برای ارائه ی انحراف استاندارد متغیر مورد نظر
<b>sum</b>	نام متغیر برای ارائه ی مجموع متغیر مورد نظر
<b>count</b>	نام متغیر برای ارائه ی تعداد مشاهدات غیرگمشده ی متغیر مورد نظر
<b>max</b>	نام متغیر برای ارائه ی حداکثر یا ماکزیمم مقادیر متغیر مورد نظر
<b>min</b>	نام متغیر برای ارائه ی حداقل یا مینیمم مقادیر متغیر مورد نظر
<b>median</b>	نام متغیر برای ارائه ی میانه ی متغیر مورد نظر

### جدولی از میانگین ها

برای ساختن جدول هایی که خانه های آن ها حاوی مقدار میانگین و انحراف استاندارد باشند، می توان از فرمان `tabulate` یا فرمان `table` به صورت زیر استفاده کرد:

### فرمان tabulate

tabulate (نام متغیر دیگر) summary, نام متغیر

tabulate means (نام متغیر) summary, نام متغیر اول نام متغیر دوم

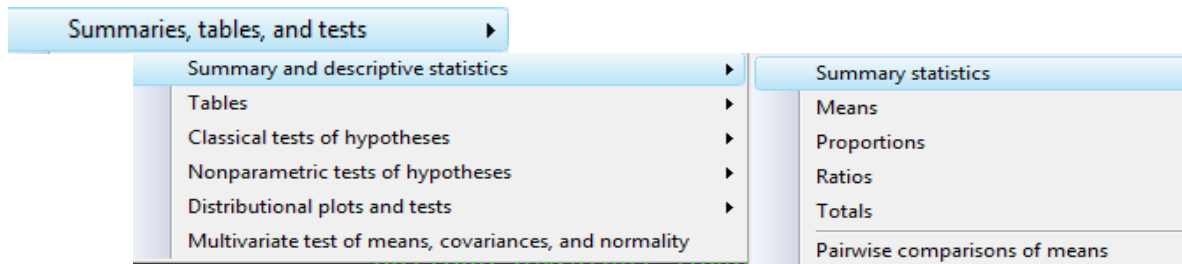
### فرمان table

table (نام متغیر دیگر mean) contents, نام متغیر

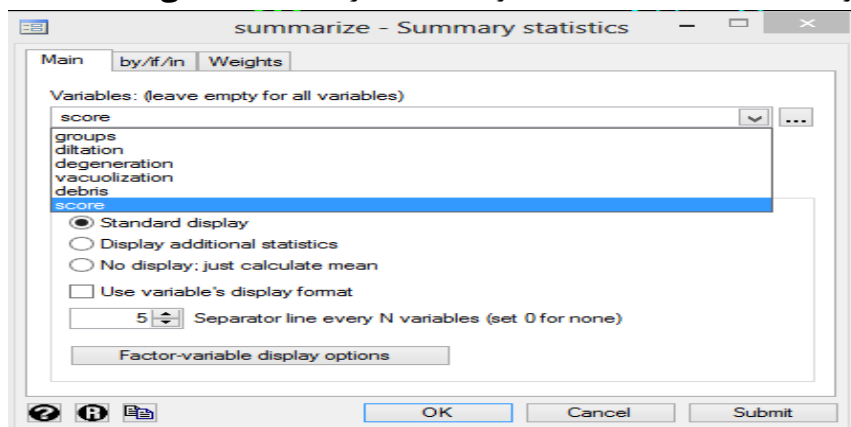
table (نام متغیر دیگر mean) contents, نام متغیر اول نام متغیر دوم

کار عملی با نرم افزار و محاسبه شاخص های توصیفی:

برای کار عملی از فایل part3 استفاده می شود. فایل part3 را باز نمایید. برای محاسبه شاخص های توصیفی متغیرهای این فایل از دو راه (منوها و تایپ دستورات) می توان استفاده کرد. اگر گزینه Summaries, tables, and test را از منوی Statistics باز نمایید زیر منوهای آن باز می شوند. با انتخاب هر کدام از گزینه ها محاسبات مورد نیاز را می توان انجام داد. این گزینه ها در تصویر زیر دیده می شوند.



در این فایل شاخص های توصیفی متغیر score که یک متغیر کمی است را اگر بخواهید محاسبه کنید از طریق مسیر تصویر بالا و انتخاب summary statistics و انتخاب متغیر score بدست می آیند.



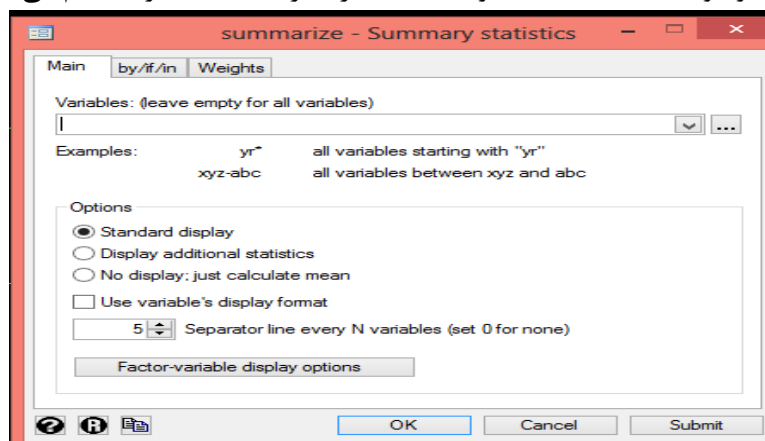
خروجی نرم افزار برای این دستور که شامل تعداد مشاهدات، میانگین، انحراف معیار، حداقل و حداکثر داده هاست در زیر آمده است.

```
. summarize score
```

Variable	Obs	Mean	Std. Dev.	Min	Max
score	32	33.78125	21.35376	1	73



برای تهیه گزارش خلاصه از متغیرهای موجود در فایل داده part three می توان به طریق زیر عمل کرد. ابتدا فرمان db sum را تایپ نمود. کادر محاوره ای مربوطه باز می شود. می توان با انتخاب متغیرها گزارش تهیه نمود. اگر متغیری انتخاب نشود و دکمه OK انتخاب شود خلاصه سازی برای همه متغیرها انجام می شود.



خروجی دستور بالا در زیر نمایش داده شده است.

Variable	Obs	Mean	Std. Dev.	Min	Max
groups	32	2.5	1.135924	1	4
diltation	32	7.875	6.173565	0	20
degeneration	32	10.34375	6.332759	0	25
vacuolizat~n	32	8.28125	5.832248	0	20
debris	32	7.28125	5.554622	0	20
score	32	33.78125	21.35376	1	73

اگر کاربر قصد محاسبه فراوانی تعداد نمونه ها در هر گروه را داشته باشد از دستور زیر استفاده می کند.  
table groups

وقتی دستور اجرا می شود فراوانی هر یک از گروه ها که چهار گروه با فراوانی ۸ است نمایش داده می شود.

اگر کاربر بخواهد میانگین را برای متغیر score بر حسب گروه محاسبه نماید از دستور زیر استفاده می کند.  
tabstat score, by(groups)

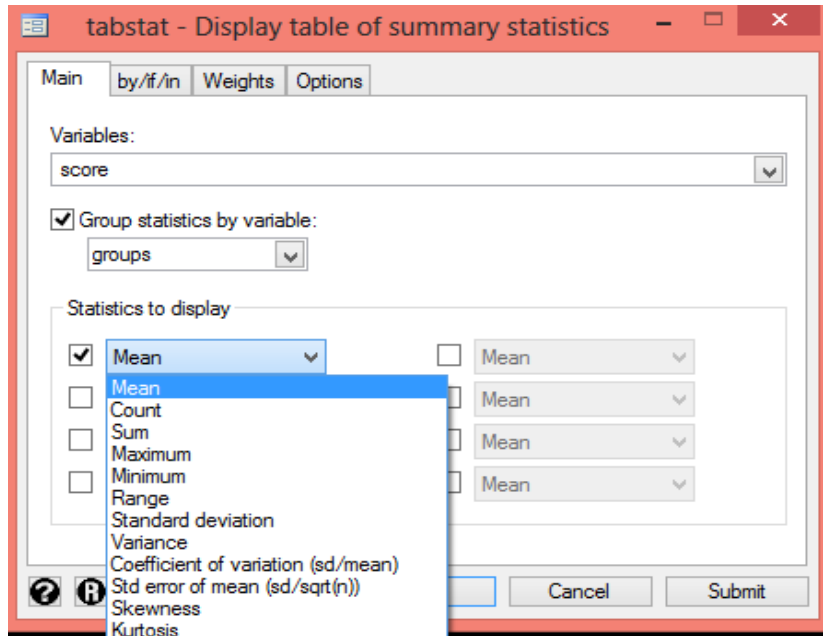
خروجی نتایج در تصویر زیر آمده است.

```
. tabstat score, by(groups)
Summary for variables: score
by categories of: groups
```

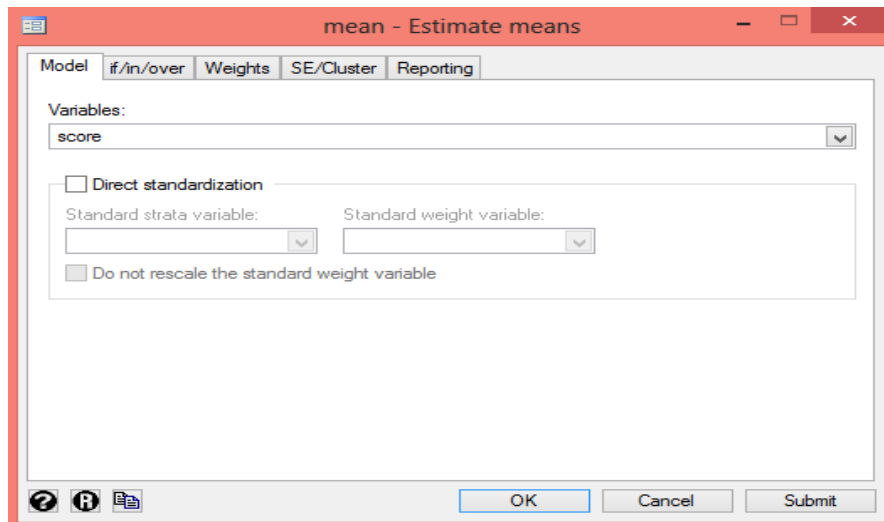
groups	mean
1	3.125
2	48.125
3	40.5
4	43.375
Total	33.78125

این محاسبات را از طریق منو نیز می توان انجام داد. اگر گزینه Summaries, tables, and test Statistics باز نمایید. زیر منوهای آن باز می شوند. دستور db tabstat اگر تایپ و اجرا شود (شکل زیر) می توان متغیر مورد نظر را در قسمت variable اضافه نمود. سپس می توان آماره های مختلف را برای آن و بر حسب گروه محاسبه نمود. کاربر می تواند با تایپ db mean و استفاده از کادر محاوره ای مربوطه محاسبات میانگین را

مانند بالا انجام دهد. وقتی کادر محاوره ای باز شد می توان با انتخاب هر کدام از گزینه های آن، محاسبات مورد نظر را انجام داد. مثلا وقتی در گزینه if/in/over متغیر مربوط به گروه را در قسمت مربوطه وارد نمایید محاسبات بر اساس گروه انجام میشوند. یا با انتخاب گزینه reporting می توان حدود اطمینان مورد نظر را تعیین و محاسبات را انجام داد.



تصویر استفاده از دستور db mean و کادر محاوره ای برای برآورد میانگین در زیر آمده است.



## ترسیم جداول حرفه ای در نرم افزار Stata

برای کار بیشتر با نرم افزار و ترسیم جدول برای نتایج پژوهش لطفا گام های زیر را اجرا نمایید.  
 -در نرم افزار دستور `webuse nhanes21` را هنگامی که به اینترنت متصل هستید اجرا نمایید. فایل داده ای باز می شود. این داده ها مربوط به دومین بررسی ملی بهداشت و تغذیه ( Second National Health and Nutrition Examination Survey ) هستند. این داده ها مشتمل بر داده های دموگرافیک، آنترپومتریک و بیولوژیک جمع اوری شده در امریکا هستند. توصیف داده ها را در زیر ببینید.

```
. describe age sex race height weight bmi highbp bpsystol bpdiastr tcresult tgresult hresult
```

Variable name	Storage type	Display format	Value label	Variable label
age	byte	%9.0g		Age (years)
sex	byte	%9.0g	sex	Sex
race	byte	%9.0g	race	Race
height	float	%9.0g		Height (cm)
weight	float	%9.0g		Weight (kg)
bmi	float	%9.0g		Body mass index (BMI)
highbp	byte	%8.0g	*	High blood pressure
bpsystol	int	%9.0g		Systolic blood pressure
bpdiastr	int	%9.0g		Diastolic blood pressure
tcresult	int	%9.0g		Serum cholesterol (mg/dL)
tgresult	int	%9.0g		Serum triglycerides (mg/dL)
hresult	int	%9.0g		High density lipids (mg/dL)

- اساسی ترین دستور در ترسیم جدول (ColVars) (RowVars) table است. با دستور `table (highbp)` جدولی برای فراوانی فشارخون ترسیم کنید.

`table (highbp) ()`

	Frequency
High blood pressure	
0	5,975
1	4,376
Total	10,351

به طور پیش فرض، جدول فراوانی هر دسته از `highbp` و فراوانی کل را نمایش می دهد. مجموعه دوم پرانتز خالی در این مثال ضروری نیست زیرا هیچ متغیر ستونی وجود ندارد. مثال زیر یک جدول برای متغیر ستون `highbp` ایجاد می کند. اولین مجموعه پرانتز خالی در این مثال ضروری است تا نرم افزار بداند که `highbp` یک متغیر ستونی است.

```
. . table () (highbp)
```

	High blood pressure		
	0	1	Total
Frequency	5,975	4,376	10,351

مثال زیر جدولی برای جنس به عنوان متغیر سطری و `highbp` به عنوان متغیر ستون ایجاد می کند. مجموع سطرها و ستون ها به طور پیش فرض گنجانده شده است.

. table (sex) (highbp)

	High blood pressure		
	0	1	Total
Sex			
Male	2,611	2,304	4,915
Female	3,364	2,072	5,436
Total	5,975	4,376	10,351

ما می توانیم مجموع سطرها و ستون ها را با درج گزینه nototals در دستور حذف کنیم.

. table (sex) (highbp), nototals

	High blood pressure	
	0	1
Sex		
Male	2,611	2,304
Female	3,364	2,072

ما همچنین می توانیم چند متغیر در سطر یا ستون یا هر دو را مشخص کنیم.

مثال زیر فراوانی را برای دسته های متغیر جنسیت که رده های درجه فشارخون بالا را دارند نمایش می دهد.

. table (highbp sex) (), nototals

	Frequency
High blood pressure	
0	
Sex	
Male	2,611
Female	3,364
1	
Sex	
Male	2,304
Female	2,072

ما می توانیم جدول را به شکل زیر نیز ترسیم کنیم.

. table (sex highbp) (), nototals

	Frequency
Sex	
Male	
High blood pressure	
0	2,611
1	2,304
Female	
High blood pressure	
0	3,364
1	2,072

یا به شکل های زیر جای متغیرهای ستونی را تغییر داد:

## روش های اپیدمیولوژی و پژوهش با Stata

. table () (highbp sex), nototals

	High blood pressure			
	0		1	
	Sex		Sex	
	Male	Female	Male	Female
Frequency	2,611	3,364	2,304	2,072

. table () (sex highbp), nototals

	Sex			
	Male		Female	
	High blood pressure		High blood pressure	
	0	1	0	1
Frequency	2,611	2,304	3,364	2,072

. table (sex) (highbp), statistic(frequency) statistic(percent) nototals

	High blood pressure	
	0	1
<b>Sex</b>		
<b>Male</b>		
Frequency	2,611	2,304
Percent	25.22	22.26
<b>Female</b>		
Frequency	3,364	2,072
Percent	32.50	20.02

دستورات زیر را اجرا نمائید. جدول های حرفه ای تری را ترسیم خواهید کرد.

table (sex) (highbp), statistic(frequency) statistic(percent) statistic(mean age) statistic(sd age) nototals  
table (sex) (highbp), statistic(frequency) statistic(percent) statistic(mean age) statistic(sd age)

	High blood pressure		
	0	1	Total
<b>Sex</b>			
<b>Male</b>			
Frequency	2,611	2,304	4,915
Percent	25.22	22.26	47.48
Mean	42.8625	52.59288	47.4238
Standard deviation	16.9688	15.88326	17.1683
<b>Female</b>			
Frequency	3,364	2,072	5,436
Percent	32.50	20.02	52.52
Mean	41.62366	57.61921	47.72057
Standard deviation	16.59921	13.25577	17.25716
<b>Total</b>			
Frequency	5,975	4,376	10,351
Percent	57.72	42.28	100.00
Mean	42.16502	54.97281	47.57965
Standard deviation	16.77157	14.90897	17.21483

با دستورات زیر می توان خروجی نرم افزار را به فرمت های Word,PDF تبدیل نمود.

Command	Purpose
<code>putdocx</code>	Create Word documents
<code>putpdf</code>	Create PDF files
<code>dyndoc</code>	Create HTML from Markdown

با دستور `putdocx` می توان فایل با پسوند `.docx` ایجاد نمود. اکنون فایل مثال ماشین ها در منوی فایل نرم افزار را باز کنید. اگر دیتاستی در نرم افزار شما در حال استفاده است با دستور `clear` یا `sysuse auto`, `replace` می توانید فایل `auto.dta` را باز نمائید. با دستور `putdocx begin` کار را برای ذخیره در فرمت `word` شروع نمائید. ما با برازش دو مدل رگرسیون شروع می کنیم. مدل اول عوامل تعیین کننده قیمت ماشین و فقط متغیر مسافت پیموده شده (`mpg`) خودرو را ارزیابی می کنیم. در مدل دوم علاوه بر آن، متغیر مدل ماشین " آیا خودرو از خارج از ایالات متحده وارد شده است (خارجی) یا خیر" را اضافه می کنیم. ما خیلی به خروجی رگرسیون و پیش فرض ها در اینجا اهمیت نمی دهیم و در بخش های دیگر کتاب به مدل های رگرسیونی پرداخته می شود، بنابراین این مدل ها را بی سر و صدا اجرا می دهیم. دستورات زیر را اجرا کنید:

```
. quietly regress price mpg
. estimates store model1
. quietly regress price mpg i.foreign
. estimates store model2
```

ما با سه مرحله می توانیم خروجی ها را به فرمت ورد (`word`) تبدیل کنیم. ایجاد یک سند در حافظه با دستور `putdocx begin`، نوشتن جدول با دستور `putdocx table tbl1 = etable` و ذخیره سند را با دستور `putdocx save mydoc` انجام می دهیم. ما باید نام جدول (`tbl1`) و نام سند خود (`mydoc`) مشخص کنیم.

## خلاصه و نتیجه گیری:

در این فصل تعاریف و مباحث پایه آماری و آمار توصیفی بیان شد. طرز محاسبه شاخص های توصیفی از جمله میانگین، میانه، نما، انحراف معیار و واریانس بیان گردید. با استفاده از دستورات `db` و `db tabstat` و `mean`، می توان پنجره گفتگو یا کادر محاوره ای مربوط به انجام آمار توصیفی و محاسبه میانگین را باز و با معرفی متغیرها و `Enter` کردن می توان خروجی را ملاحظه نمود. همچنین می توان انواع جداول یکطرفه و دو طرفه را ترسیم نمود. برای این کار نیز می توان از دستور `db table` استفاده نمود. مطالب این فصل یکی از مهمترین گام های اولیه در تحلیل داده هاست. بنابراین درک صحیح و عمیق این مطالب برای آماده سازی داده ها برای تحلیل مهم است.

تمرین ها :

- ۱- شاخص های میانگین، انحراف معیار و خطای معیار را تعریف کنید. تفاوت انحراف معیار و خطای معیار چیست؟
- ۲- فایل part two را باز نمایید و شاخص های خلاصه برای متغیرهای age را محاسبه نمایید.
- ۳- در فایل مذکور جدول یک طرفه برای محاسبه فراوانی متغیر exposure درست نمایید.
- ۴- جدول دوطرفه ای برای متغیر exposure و bimari درست نمایید. فراوانی ها را گزارش نمایید.
- ۵- میانگین، انحراف معیار و حدود اطمینان ۹۹٪ برای متغیر age را بنویسید. کدام جدول زیر پاسخ این سوال است؟

Mean estimation Number of obs = 7

	Mean	Std. Err.	[90% Conf. Interval]	
age	43.57143	4.476697	34.8724	52.27046

Mean estimation Number of obs = 7

	Mean	Std. Err.	[99% Conf. Interval]	
age	43.57143	4.476697	26.9744	60.16846

- ۶- در ارائه نتایج یک پژوهش، روش صحیح برای گزارش میانگین و انحراف معیار استفاده از Mean(SD) صحیح است یا استفاده از میانگین و خطای معیار Mean(SE)؟ چرا؟
- ۷- در فایل ahmadi1400 که قبلاً داده های آن را وارد نموده اید (تمرین فصل ۱)، شاخص های حداقل، حداکثر، میانه، نما، میانگین، انحراف معیار، واریانس و ضریب تغییرات را برای نمایه توده بدنی بیماران با و بدون فشارخون بالا محاسبه و گزارش کنید.

## .....فصل چهارم: رسم نمودار در Stata

نمودارها و کاربرد آنها  
گزینه های عمومی فرمان های مربوط به رسم نمودار  
نمودار دایره ای  
نمودار میله ای  
نمودار هیستوگرام  
نمودار خطی، نمودار پراکندگی  
نمودار کاپلان مایر و جدول عمر  
نمودار رادار  
تمرین ها

### • اهداف:

#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند انواع نمودار و کاربرد آنها را تعریف نماید.
- محاسبات مورد نیاز درباره آمار توصیفی و ترسیم نمودارهای مربوطه را در نرم افزار انجام دهد.
- کاربر بتواند نمودارهای مربوط به تحلیل بقا را ترسیم نماید.
- کاربر بتواند یک نمودار را ویرایش و به آن عنوان و توضیحات لازم را اضافه نماید.



## نمودارها:

### Graph

نمودار یکی از ابزارهای مهم آمار توصیفی است. نمودار وسیله ای است تصویری برای نمایش توزیع داده های آماری و نشان دادن موثر و تاثیرگذار داده ها. برای کشیدن نمودار از دستگاه مختصات استفاده می شود. در اغلب موارد اندازه صفت مورد مطالعه را روی محور طولها و فراوانی آن را روی محور عرضها مشخص می کنند. محورها باید طوری تقسیم بندی شود که کمترین و بیشترین فراوانی و یا اندازه صفت مورد مطالعه در آنها بگنجد. در ارائه یک نمودار باید کلیه نکاتی را که در مورد جدول شرح آن ذکر شد رعایت نمود.

### نمودار نرده ای یا ستونی Bar chart:

متداولترین نوع نمودار برای داده های کیفی (اسمی یا ترتیبی) نمودار نرده ای یا ستونی است. این نمودار تعداد یا درصدها را به صورت ستون نشان می دهد. در تهیه نمودار ستونی بهتر است قاعده ستون ها یا نرده ها و فواصل بین آنها را مساوی انتخاب نمود.

### نمودار دایره ای Pie chart:

از این نمودار برای نمایش داده های اسمی و رتبه ایی استفاده می شود. این نمودار را بر مبنای فراوانی نسبی داده ها رسم می کنند یعنی دایره را به نسبت فراوانی گروهها تقسیم می کنند. اگر داده یا گروهی دارای فراوانی  $\frac{f_i}{N}$  باشد نسبت  $360 \times \frac{f_i}{N}$  از دایره به آن گروه اختصاص داده می شود. از این نمودار در رشته های بهداشتی برای نشان دادن اطلاعاتی در مورد توزیع پزشکان به مناطق مختلف، تخصیص بودجه ملی به بهداشت عمومی استفاده می شود.

### هیستوگرام یا بافت نگار Histogram:

از هیستوگرام برای نمایش اطلاعات کمی پیوسته استفاده می شود. این نمودار در پزشکی کاربرد زیادی دارد و دارای دو محور افقی و عمودی است که روی محور افقی مقادیر یا اندازه های گروه های مورد نظر و روی محور عمودی تعداد یا درصد مقادیر مشاهدات را نشان می دهند. استفاده از تعداد یا درصد بستگی به هدف پژوهشگر دارد. از هیستوگرام درصدی زمانی استفاده می شود که بخواهند دو هیستوگرام با تعداد متفاوت را مورد مقایسه قرار دهند. در هیستوگرام سطح هر یک از مستطیل ها با درصد مشاهداتی که در آن حدود قرار دارند متناسب است و به عبارتی احتمال یا درصد انتخاب افراد آن گروه را بیان می دارد. برای بررسی وضعیت توزیع داده ها و تقارن آنها از نمودار هیستوگرام استفاده می شود. اگر در ترسیم هیستوگرام از فراوانی نسبی استفاده شود سطح کل زیر هیستوگرام برابر یک خواهد بود زیرا جمع فراوانی نسبی کلیه گروهها برابر یک است. به این ترتیب هر مقدار از سطح زیر هیستوگرام با میزان اطلاعاتی که داده ها در اختیار می گذارند برابر است. در ترسیم هیستوگرام بهتر است فواصل بین گروهها را مساوی در نظر گرفت در غیر این صورت بایستی ارتفاع هر یک از ستون های مستطیل را به نسبت مناسب اصلاح نمود تا سطح صحیحی را به خود اختصاص دهند.

### نمودار چند ضلعی فراوانی Polygon:

معمولاً برای نمایش داده های کمی گسسته مورد استفاده قرار می گیرد. در این نمودار مقادیر صفت را روی محور طولها و فراوانی متناظر با آن مقادیر را روی محور عرض ها مشخص می کنند. نموداری که از وصل کردن

نقاط به دست می آید چند گوش نامیده می شود. از این نمودار برای مقایسه دو توزیع روی یک نمودار استفاده می شود.

اگر در هیستوگرام وسط ستونها را به هم وصل کنیم یک چند گوش بدست می آید. این عمل نمایش چند هیستوگرام را روی یک دستگاه مختصات آسان می کند. اگر بخواهیم هیستوگرام را به نمودار چند گوش تبدیل کنیم باید از نقطه وسط آخرین ستون تا محور x ها ادامه دهیم تا فراوانی چند ضلعی کامل شود و معلوم گردد که مقادیر سایر گروهها قبل از اولین و آخرین گروه صفر بوده است. به این ترتیب در چنین نموداری از آنجا که بر مبنای هیستوگرام تهیه شده است سطح زیر این چند ضلعی برابر با سطح زیر هیستوگرام خواهد بود.

### نمودار توزیع تجمعی :

در این نمودار مقادیر صفت روی محور طولها و فراوانیهای تجمعی به صورت ساده یا نسبی روی محور عرضها در نظر گرفته می شود سپس نقاط متناظر مقادیر صفت و فراوانیهای تجمعی را مشخص نموده و از وصل نمودن این نقاط به هم در صفات پیوسته به صورت خطی یا تقریب منحنی و در صفات ناپیوسته به صورت پلکانی نمودار تجمعی حاصل می گردد. بر اساس این نمودار می توان صدک های مختلف مانند صدک ۲۵ (چارک اول)، صدک ۵۰ (میانه) و صدک ۷۵ (چارک سوم) را محاسبه نمود.

### نمودار رادار:

رادار چارت یا نمودار راداری یا نمودار عنکبوتی یا نمودار ستاره ای، نموداری است که با نمایش یک یا چند، چند ضلعی وضعیتی متغیرهای مختلف مورد نیاز را ترسیم می کند. با مشاهده نمودار راداری، میزان نزدیکی و شباهت متغیرهای هم جنس قابل درک خواهد بود. یکی از کاربردهای اصلی رادار چارت مقایسه وضعیت جاری کار با وضعیت تخمینی و یا پیش بینی می باشد که با مقایسه کلیه ابعاد مورد نظر به لحاظ واقعی و پیش بینی، عملکرد واقعی کار قابل رصد خواهد بود. ترسیم این نمودار مانند سایر نمودارهای پیشگفت در استاتا بر راحتی تعبیه نشده و باید دستورات آن دانلود و نصب یا نوشته شود که در این کتاب آمده است.

### نمودارها در Stata :

در Stata امکانات زیادی برای رسم نمودارها وجود دارد. در این بخش به معرفی فرمان های ترسیم نمودار پرداخته می شود. برخی نمودارها را می توان با استفاده از فرمانی که از ترکیب واژه ی graph با نام نمودار حاصل می شود، رسم نمود. در زیر به تعدادی از آنها اشاره می شود:

نمودار پراکندگی دو طرفه (نقطه ای)	<b>:twoway scatter</b>
نمودار پراکندگی دو طرفه (خطی)	<b>:twoway line</b>
ماتریس نمودار پراکندگی	<b>:matrix</b>
نمودار جعبه ای	<b>:box</b>
نمودار میله ای	<b>:bar</b>
نمودار دایره ای (کلوچه ای)	<b>:pie</b>

شکل عمومی فرمان graph به صورت زیر است:

گزینه‌ها , نام متغیر(ها) نام نمودار graph  
 برای هر یک از فرمان‌هایی که به منظور رسم نمودار به کار می‌روند، گزینه‌هایی وجود دارد که با استفاده از آن‌ها می‌توان شکل ظاهری نمودار را تغییر داد. برخی از این گزینه‌ها، به طور عمومی و برای همه انواع نمودارها قابل استفاده هستند، در حالی که برخی دیگر منحصراً برای یک نوع نمودار به کار می‌روند.

### گزینه‌های عمومی فرمان‌های مربوط به رسم نمودار

**title("عنوان")** این گزینه در بالای نمودار، عنوان مورد نظر را اضافه می‌کند.  
**subtitle("عنوان")** این گزینه در بالای نمودار، زیر عنوان مورد نظر را اضافه می‌کند.  
**caption("عنوان")** این گزینه در پایین نمودار، عنوان مورد نظر را اضافه می‌کند.  
**"عنوان" b1 و "عنوان" b2** با استفاده از این گزینه‌ها می‌توان تا حداکثر ۸ عنوان را به نمودار افزود. حروف **t1("عنوان") و t2("عنوان")** به ترتیب تعیین می‌کند که عنوان به ترتیب در پایین، بالا، چپ یا راست نمودار قرار گیرد و شماره‌های ۱ و ۲ ترتیب عنوان‌ها را بیان می‌کند.  
**l1("عنوان") و l2("عنوان")**  
**r1("عنوان") و r2("عنوان")**  
**by(نام متغیر رسته‌ای)** این گزینه موجب می‌شود که Stata برای هر یک از مقادیر متغیر رسته‌ای، نمودار جداگانه‌ای رسم کند.

**نکته:** قبل از بکارگیری گزینه‌ی by باید مجموعه داده‌ها برحسب متغیر رسته‌ای مورد نظر مرتب شده باشد.

### نمودارهای پراکندگی دوطرفه

برای رسم نمودارهای دوطرفه‌ی نقطه‌ای و خطی، از فرمان graph به صورت‌های زیر استفاده می‌شود:  
 نام متغیر محور عمودی نام متغیر محور افقی graph twoway scatter  
 نام متغیر محور عمودی نام متغیر محور افقی graph twoway line  
 برخی از گزینه‌هایی که می‌توان در نمودار پراکندگی دوطرفه به کار برد، عبارتند از:  
**xtitle("عنوان") و title("عنوان y")** به ترتیب به محور افقی و محور عمودی عنوان‌های مورد نظر را منتسب می‌کند.  
**xlabel(#دسته) و ylabel(#دسته)** (تعداد به ترتیب تعداد دسته‌ی اعداد روی محور افقی و محور عمودی نمودار را تعیین می‌کند.  
**xscale(range(...)) و yscale(range(...))** به ترتیب دامنه‌ی اعداد روی محور افقی و محور عمودی نمودار را تعیین می‌کند.  
**sort** قبل از رسم نمودار، مقادیر مربوط به محور افقی را به ترتیب صعودی مرتب می‌کند.

### ماتریس نمودار پراکندگی

برای رسم ماتریس نمودار پراکندگی کافی است از فرمان زیر استفاده شود:

نام متغیرها graph matrix

در این صورت نمودارهای پراکندگی متغیرها به صورت دو به دو در یک ماتریس نمودار پراکندگی رسم می شوند. با استفاده از گزینه ی half می توان فقط مثلث زیرین ماتریس پراکندگی را رسم کرد.

### نمودار جعبه ای

نمودار جعبه ای بیانگر اطلاعاتی درباره ی مرکز، پراکنش، تقارن توزیع و نقاط دورافتاده است. برای رسم یک نمودار جعبه ای از فرمان زیر استفاده می شود:

نام متغیر graph box

اگر مقیاس متغیرها یکسان باشد، برای مقایسه ی توزیع چند متغیر می توان فرمان زیر را به کار برد:

نام متغیرها graph box

خط افقی داخل جعبه، میانه ی داده ها و ارتفاع جعبه، دامنه ی میان چارکی است. نقاط دورافتاده در خارج از فاصله ۱/۵ برابر فاصله ی میان چارکی از چارک اول و سوم با دایره های کوچک نشان داده می شوند.

### نمودار دایره ای (کلوچه ای)

برای رسم یک نمودار دایره ای از فرمان زیر استفاده می شود:

نام متغیرها graph pie

که متغیرها همگی بیانگر تعداد برحسب واحد یکسان (نفر، تومان، ساعت و...) هستند. همچنین می توان برای یک متغیر رسته ای، نمودار دایره ای را به صورت زیر رسم کرد:

graph pie, over (نام متغیر رسته ای)

برخی از گزینه های این فرمان عبارتند از:

plabel(\_all sum) به هر یک از برش های نمودار، به ترتیب جمع یا درصد مربوط به آن برش را

plabel(\_all percent) منتسب می کند.

pie (explode, شماره برش) برشی را که شماره ی آن ذکر شده است، از نمودار دایره ای جدا می کند.

### نمودارهای میله ای

نمودار میله ای با استفاده از فرمان زیر رسم می شود:

graph bar (stat) نام متغیرها , over (نام متغیر رسته ای)

با استفاده از گزینه ی over می توان نمودار را برای مقادیر مختلف یک متغیر رسته ای رسم نمود. عبارت stat تعیین می کند که ارتفاع هر میله، مقدار کدام آماره از متغیر مورد نظر است. در حالت پیش فرض ارتفاع هر میله، بیانگر میانگین متغیر است.

### هیستوگرام

برای رسم هیستوگرام، از فرمان زیر استفاده می شود: histogram نام متغیر

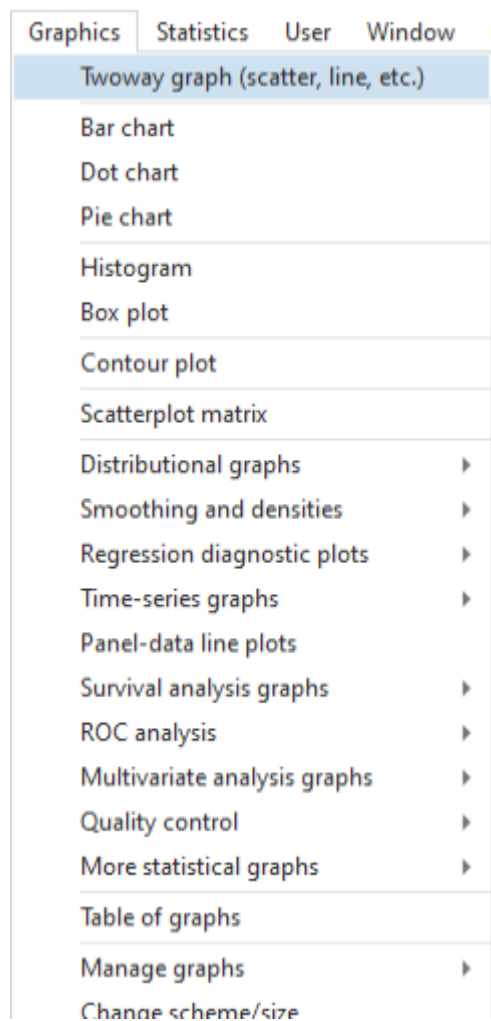
با اجرای این فرمان، به طور پیش فرض مقادیر فراوانی نسبی بر روی ستون های هیستوگرام قرار می گیرند.

برخی گزینه های فرمان histogram عبارتند از:

تعداد ستون های نمودار هیستوگرام را به تعداد مورد نظر، تغییر می دهد.	<b>bin(تعداد)</b>
روی محور عمودی، به جای فراوانی نسبی، فراوانی ها را قرار می دهد.	<b>freq</b>
روی محور عمودی، به جای فراوانی نسبی، درصدها را قرار می دهد.	<b>percent</b>
روی هیستوگرام یک منحنی نرمال با میانگین و انحراف استاندارد مربوط به مقادیر متغیر مورد نظر رسم می کند.	<b>normal</b>

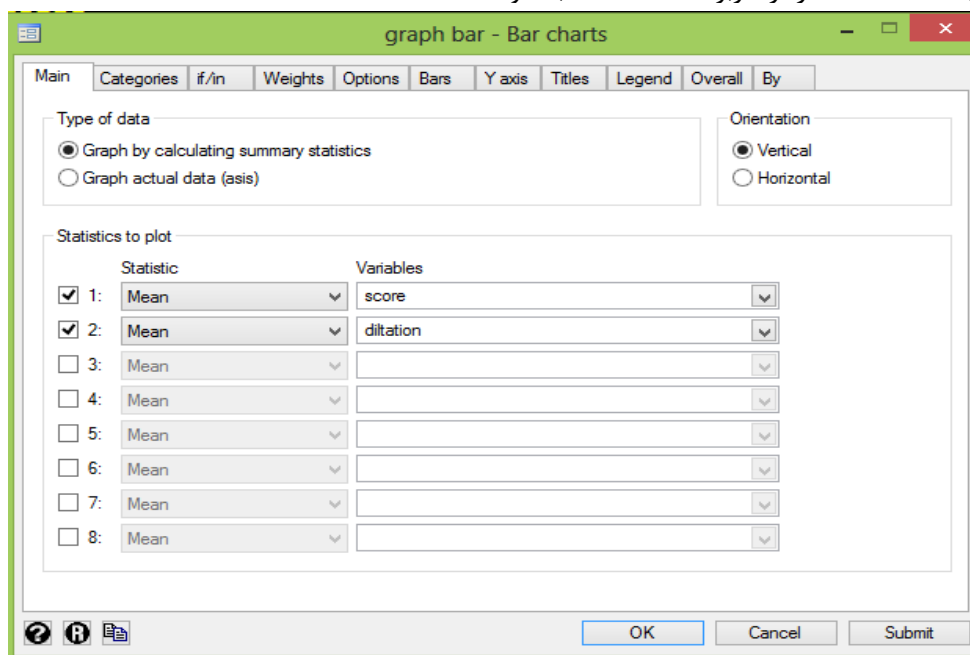
### کار عملی با نرم افزار و ترسیم انواع نمودارها:

در نرم افزار استاتا منوی Graphics می توان انواع نمودارها را انتخاب و ترسیم نمود. شکل زیر انواع نمودارهای قابل ترسیم را نمایش می دهد.



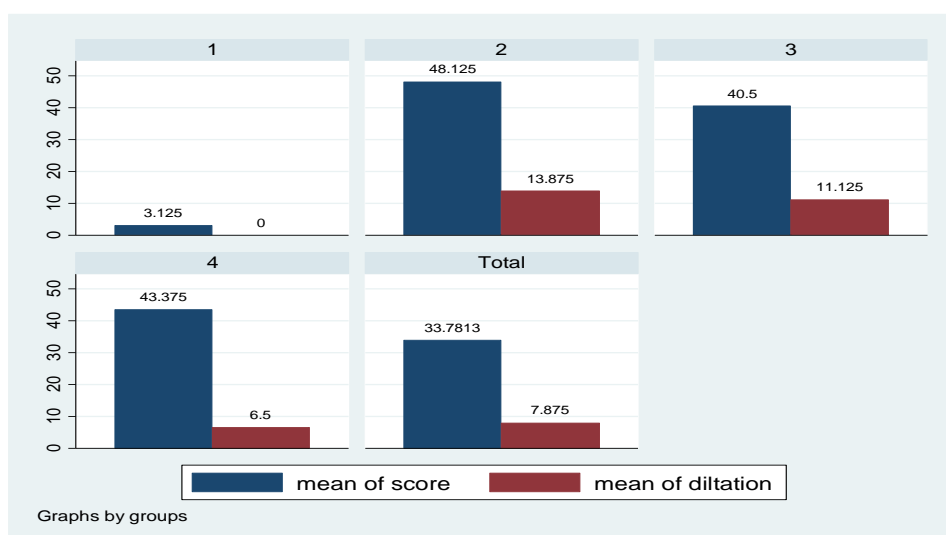
فایل part three را باز نمایید. نمودار ستونی برای متغیر score ترسیم نمایید. در این نمودار مقادیر میانگین ها نمایش داده شوند. نمودار بر حسب گروه ها ترسیم گردد. برای این کار از طریق زیر عمل می کنیم. از منوی graph گزینه bar- bar chart انتخاب می شود. میتوان با استفاده از تایپ **db graph** و انتخاب نوع

گراف به کادر محاوره ای رسید. در کادر محاوره ای که در زیر نمایش داده شده است، متغیر score را در قسمت variable و آماره مورد نظر را از قسمت statistic انتخاب نمود. سپس با انتخاب گزینه BY در بالای tab کادر محاوره ای و انتخاب نمودن متغیر گروه، نمودار را ترسیم نمود. همچنین برای نمایش مقادیر میانگین بر روی نمودار می توان گزینه bar labels را از سربرگ Bars انتخاب نمود.



با استفاده از دستور زیر نیز می توان به خروجی همانند استفاده از مسیر بالا رسید. نمودار ستونی ترسیم شده در زیر آمده است.

`graph bar (mean) score (mean) diltation, blabel(bar) by(groups, total)`

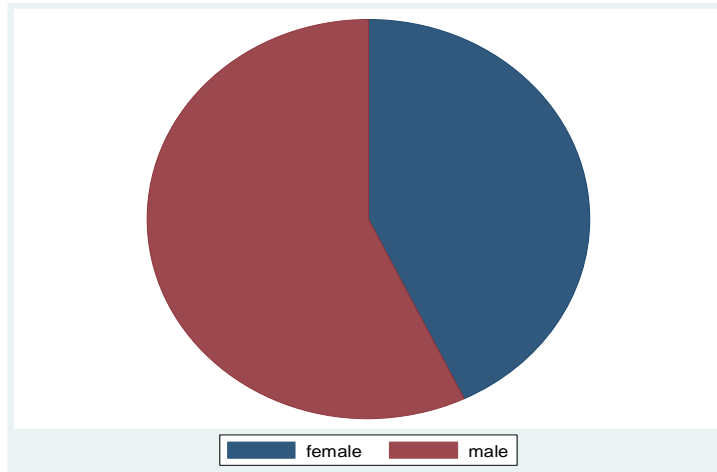


در فایل part two نمودار دایره ای برای متغیر sex ترسیم نمایید. برای رسم این نمودار از فرمان زیر استفاده می شود.

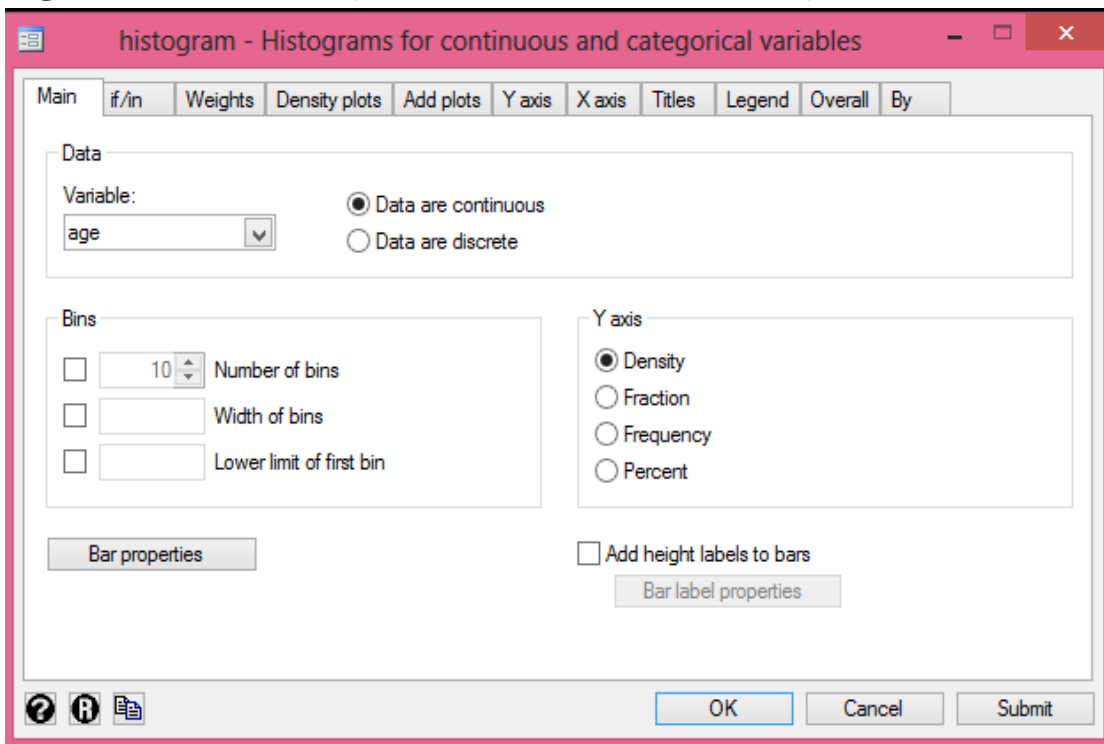
`graph pie,`

over(sex)

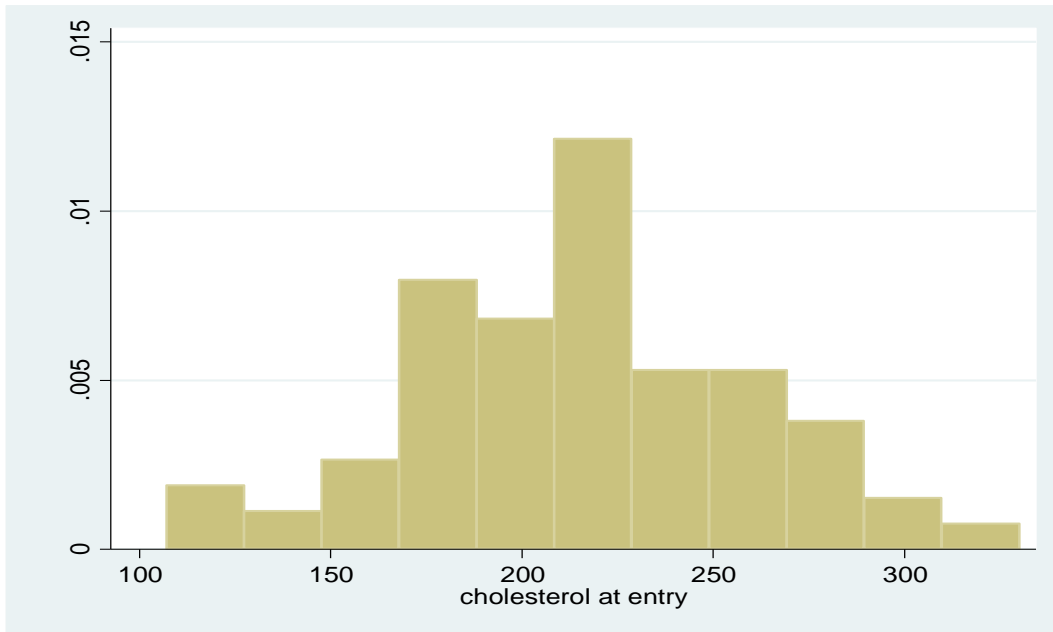
می توان از طریق دیگر نیز نمودار را ترسیم کرد. با انتخاب گزینه pie chart از منوی Graph و انتخاب متغیر sex در قسمت مربوطه نمودار را رسم کرد. خروجی در زیر آمده است.



در فایل مذکور نمودار هیستوگرام برای متغیر age را اگر کاربر بخواهد ترسیم نماید از مسیر زیر استفاده می کند.

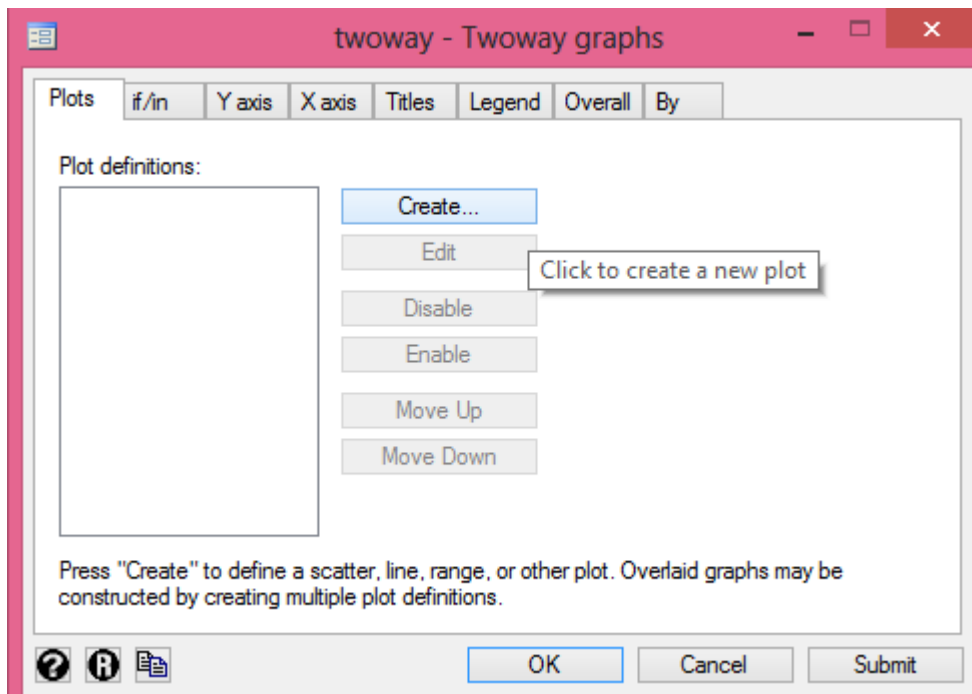


همچنین با تایپ دستور histogram cho نیز می توان نمودار را برای فایل part four رسم نمود. نمودار در زیر آمده است.

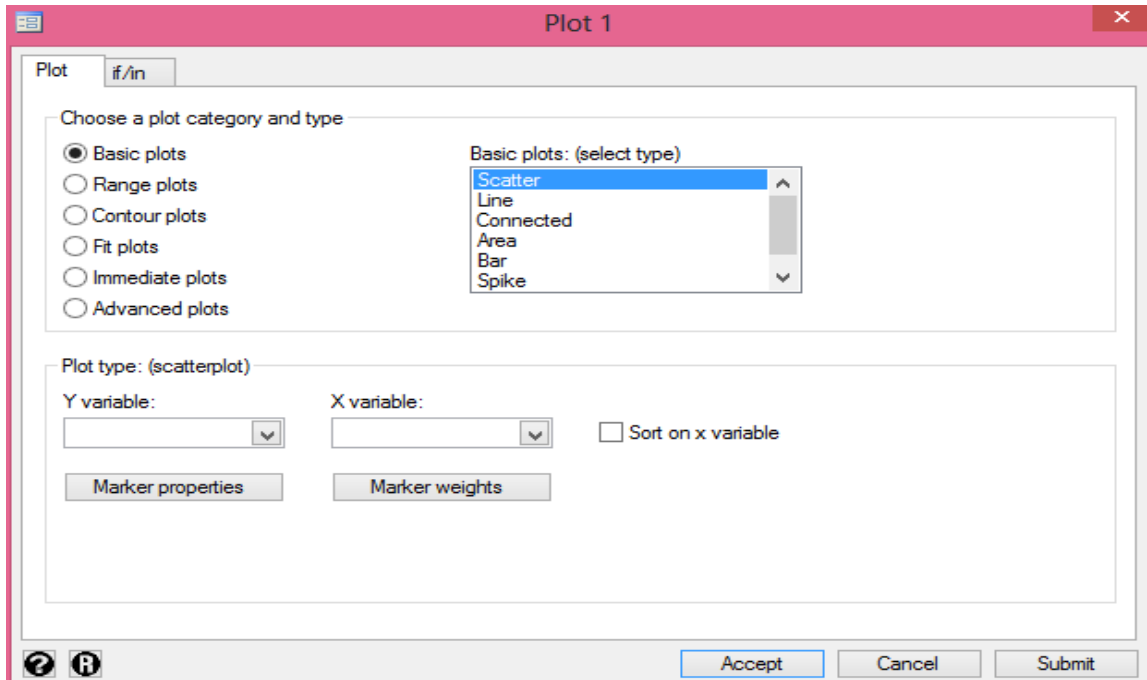


نمودار پراکنش:

برای رسم نمودار پراکنش یا Scatter plot با انتخاب گزینه twoway graphs(scatter, line,etc.) از منوی Graph و انتخاب creat می توان نوع نمودار را انتخاب و آن را رسم نمود. کادر محاوره ای مذکور در زیر نمایش داده شده است.



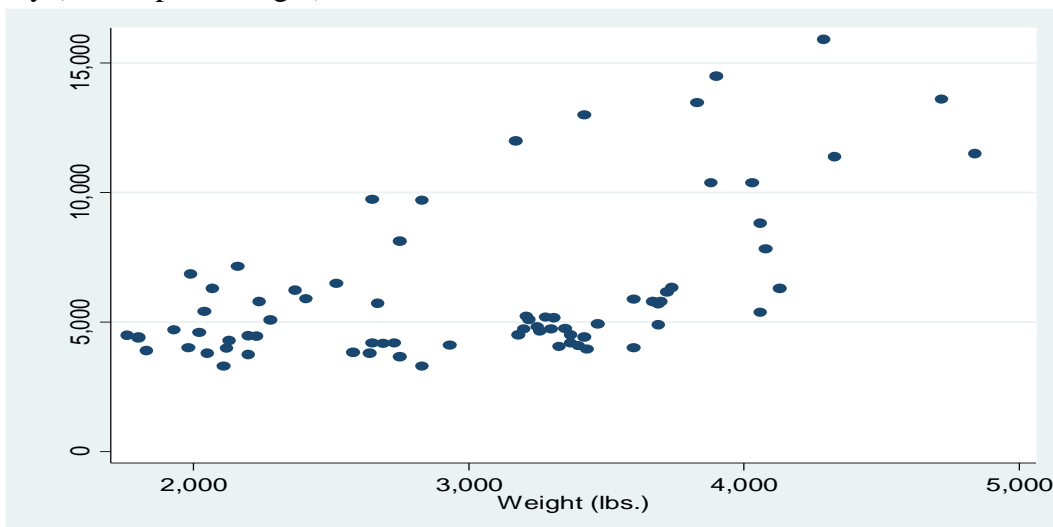




بعد از انتخاب نوع نمودار، متغیر وابسته و متغیر مستقل را می بایست در قسمت های مربوطه وارد نمود. چنانچه گزینه Basic plot انتخاب شود نمودار پراکنش ساده بین دو متغیر تعریف شده ترسیم می شود. چنانچه ترسیم خط برازش مورد نظر باشد باید گزینه fit plot و نوع آن مانند linear prediction را از گزینه edit انتخاب نمود. حدود اطمینان خط مربوطه را نیز میتوان ترسیم نمود.

برای این کار ابتدا از دستور sysuse auto.dta استفاده کنید. با استفاده از این دستور فایل مثال auto.dta که در نرم افزار وجود دارد باز می شود. از مسیر دیگری به این فایل می توان دست پیدا نمود. که در فصول گذشته به آن اشاره گردید و تاکید شد که این مسیر را بخاطر بسپارید. برای رسم نمودار پراکنش متغیر قیمت (price) در برابر متغیر وزن (weight) از دستور زیر نیز می توان استفاده نمود. نمودار در زیر آمده است.

twoway (scatter price weight)

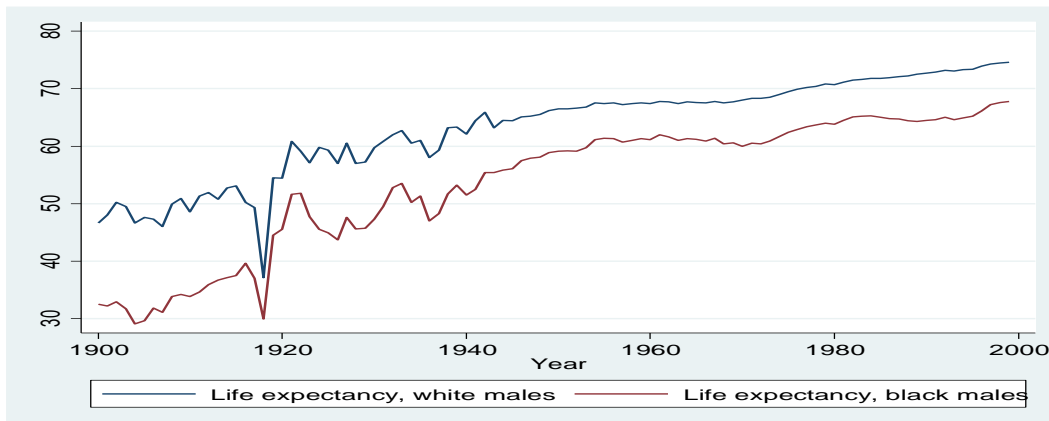


### نمودار های خطی و نقطه ای:

فایل `sysuse uslifeexp` که بعنوان مثال در داده های نرم افزار وجود دارد را باز نمایید. دستور زیر را اجرا نمایید.

```
twoway line le_wm year || line le_bm year
```

نموداری برای امید زندگی در مردان سفید پوست و سیاه پوست ترسیم می شود. نمودار به شکل زیر است.

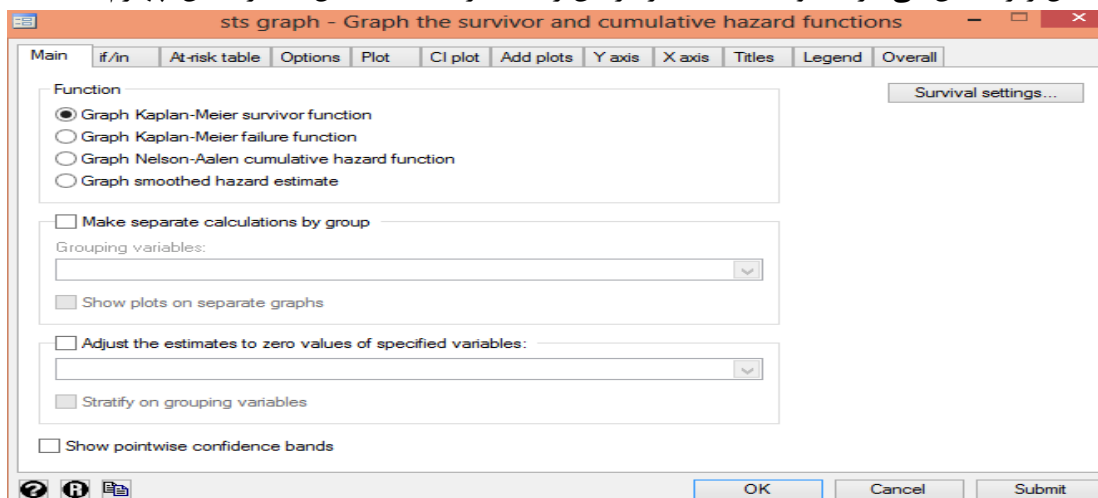


### نمودار کاپلان مایر:

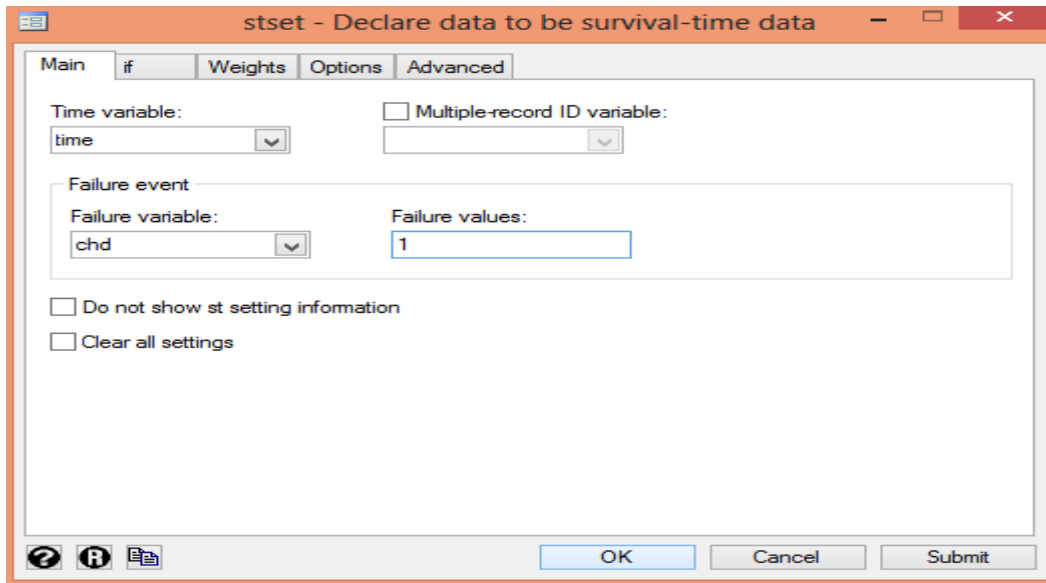
ابتدا فایل `part four` را باز نمایید. در این فایل برای شناسایی متغیرها از دستور `des` استفاده می شود. خروجی به صورت زیر نشان دادن می شود.

variable name	storage type	display format	value label	variable label
time	float	%9.0g		age at entry in years
chd	byte	%8.0g		death from chd
smok	byte	%8.0g		never/ex/1-14/15-24/25+
timein	float	%d		Date of entry (days)
timeout	float	%d		Date of exit (days)
timebth	float	%d		Date of birth (days)
chol	float	%9.0g		cholesterol at entry
all	byte	%8.0g		death from all causes
sbp	float	%9.0g		systolic bp at entry
smoking	float	%9.0g	smoking	

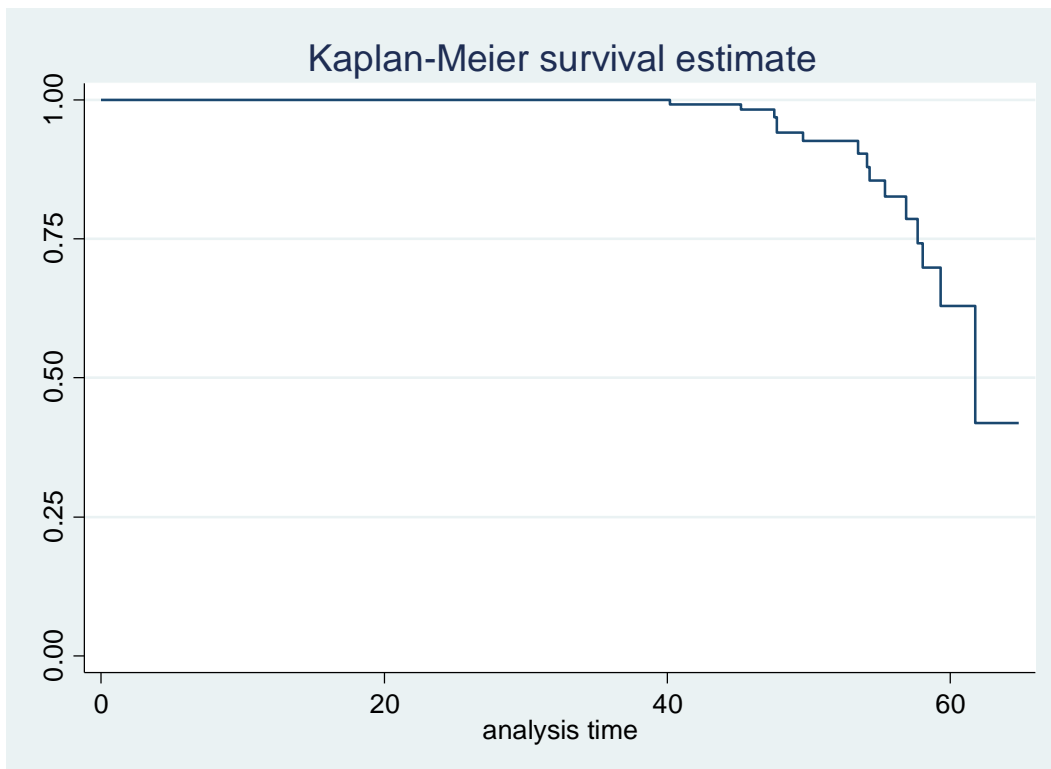
برای ترسیم منحنی بقا، از کادر گفتگو یا کادر محاوره ای `sts graph` استفاده می شود. وقتی این فرمان OK شود شکل زیر نمایان می گردد. توضیحات بیشتر در این زمینه، در قسمت تحلیل بقا در بخش چهارم آمده است.



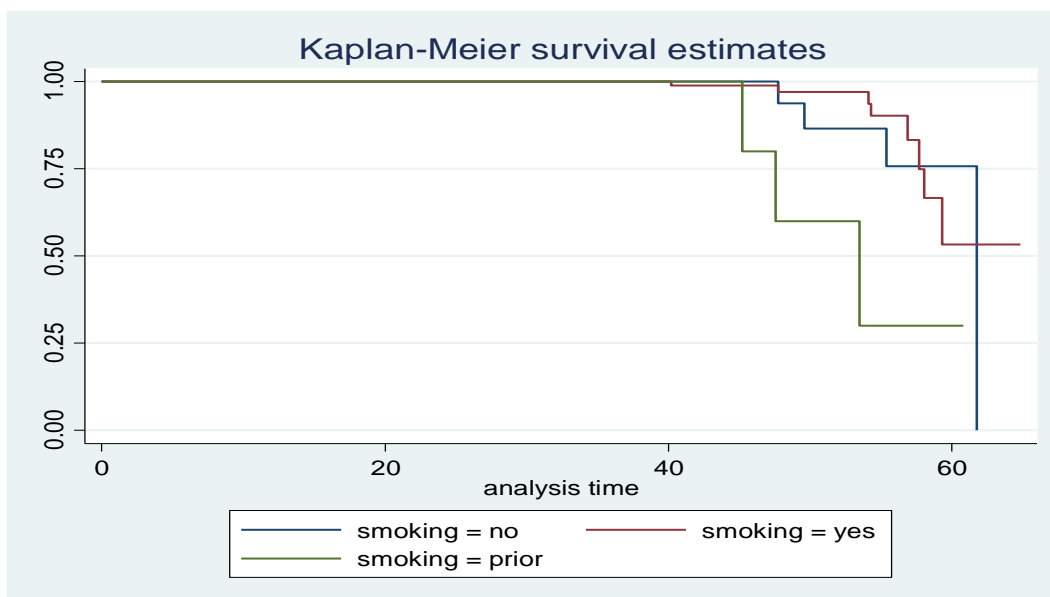
از کادر محاوره ای بازشده، گزینه Survival setting را انتخاب تا تنظیمات زمان ها و معرفی متغیرها انجام شود. در این کادر متغیر زمان (time) و وقوع رخداد (failure variable) برای نرم افزار تعریف و معرفی میشوند. و کادر محاوره ای مذکور بسته می شود. این کادر در تصویر زیر نمایش داده می شود.



اکنون با انتخاب ok دستور اجرا می شود. نمودار به شکل زیر نمایش داده می شود.



برای ترسیم این منحنی بر حسب گروه های سیگاری و غیر سیگاری تنظیمات مربوطه در کادر محاوره ای را باید انجام داد. برای این منظور می بایست گزینه `make separate calculation by group` نیز انتخاب شود. اکنون با انتخاب این گزینه، منحنی بر حسب متغیر (`smoking`) نمایانگر گروه های سیگاری و غیر سیگاری و آنهایی که قبلا سیگاری بودند ولی اکنون سیگار نمی کشند، ترسیم می شود. نمودار در زیر آمده است.



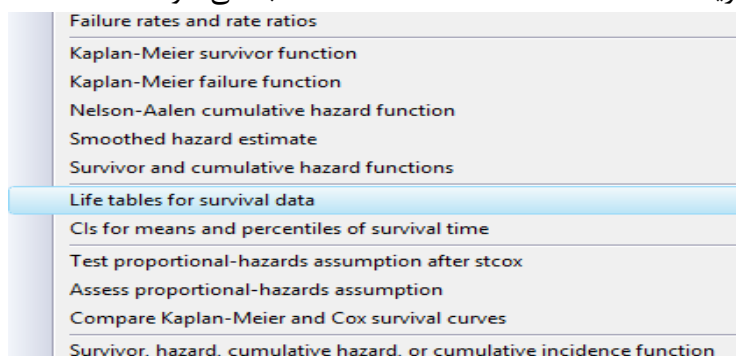
برای ترسیم نمودارهای مذکور می توان از دستورات زیر استفاده کرد.

```
db sts graph
stset time, failure(chd)
sts graph
sts graph, by(smoking)
```

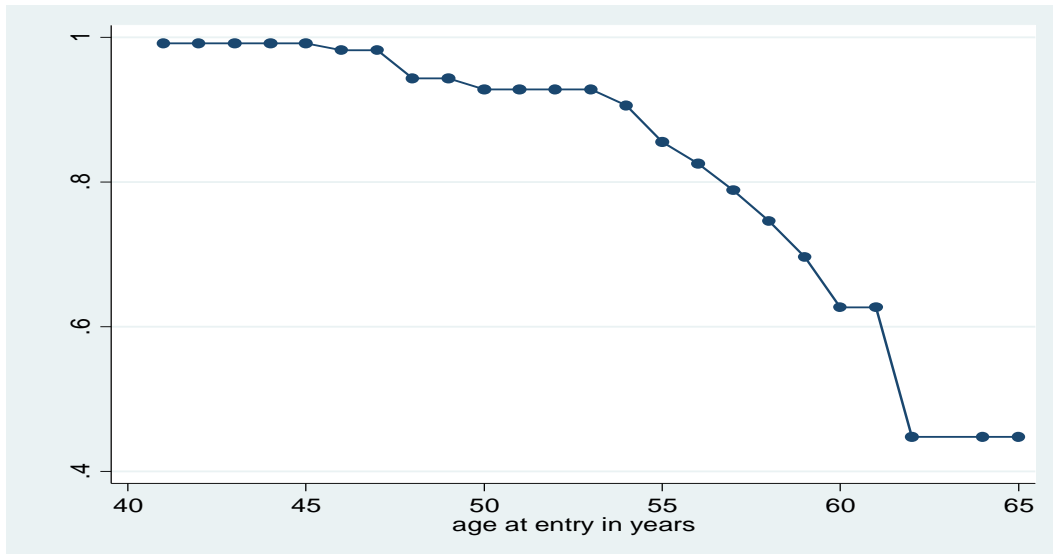
### نمودار برای جدول عمر (life tables):

برای ترسیم منحنی بقا بر اساس جدول عمر با استفاده از داده های فایل مذکور مانند ترسیم منحنی کاپلان مایر که گزینه `Kaplan meier survivor function` در شکل زیر را می بایست انتخاب کرد. در اینجا به طریق زیر عمل می کنیم.

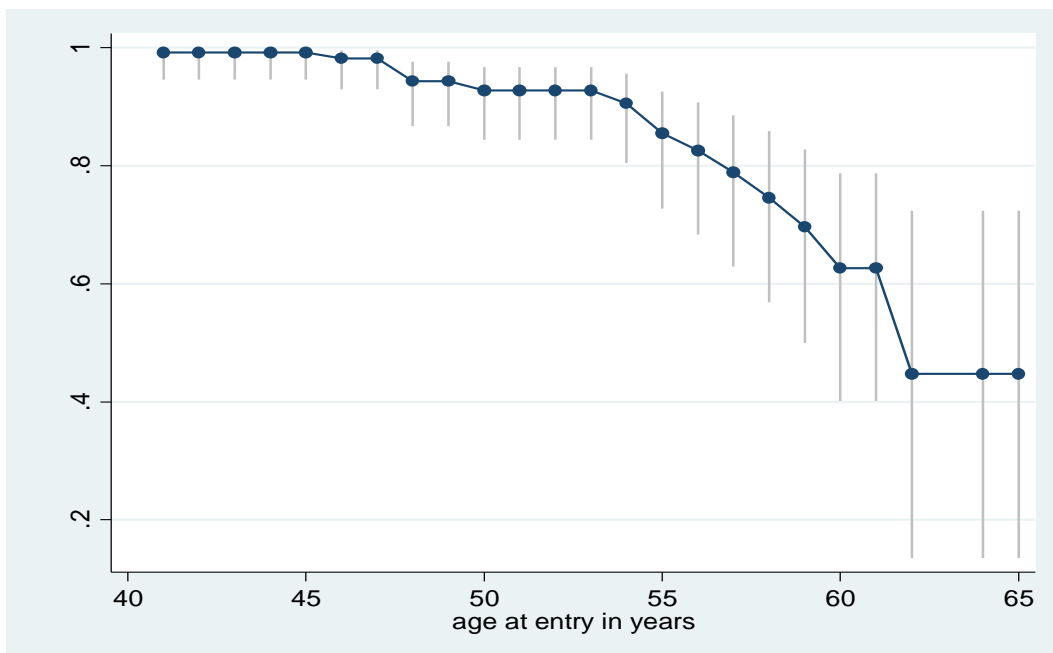
از منوی **Graphics** گزینه **Survival analysis graphs** را انتخاب و از کادر روبروی آن که در شکل زیر آمده است. گزینه `life tables for survival data` انتخاب می شود.



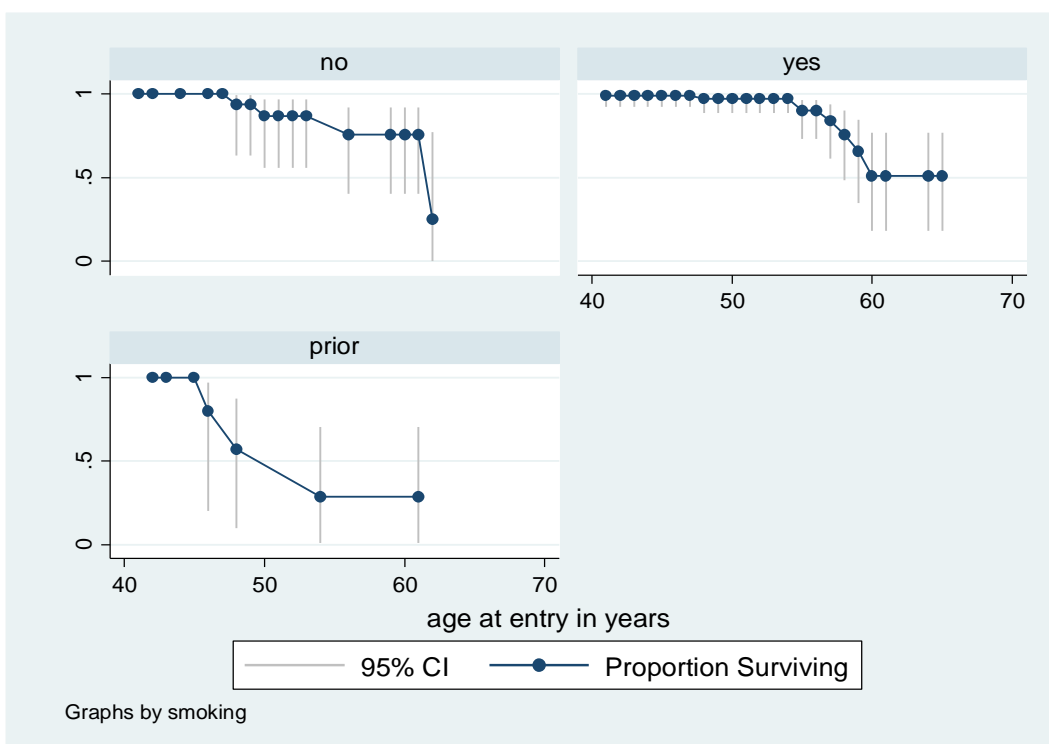
سپس همانند کادر محاوره ای برای ترسیم کاپلان مایر که در بالا توضیح داده شد، متغیرهای زمان و رخداد را تعریف و گزینه graph انتخاب می شود. می توان برای ترسیم جدول عمر گزینه table را نیز انتخاب نمود. همچنین ترسیم نمودار بر حسب گروه را نیز می توان انتخاب نمود تا نمودارها به تفکیک گروه ترسیم گردد. شکل زیر این نمودار را نمایش می دهد.



در ترسیم نمودار مذکور می توان با انتخاب حدود اطمینان مورد نظر و اینکه نمودار بر حسب گروه ترسیم گردد یا نه، نمودار دلخواه را ترسیم نمود. در تصاویر زیر این نمودارها نمایش داده می شوند.



نمودار بر حسب متغیر smoking :



با استفاده از دستورات زیر می توان نمودار مذکور را ترسیم نمود.

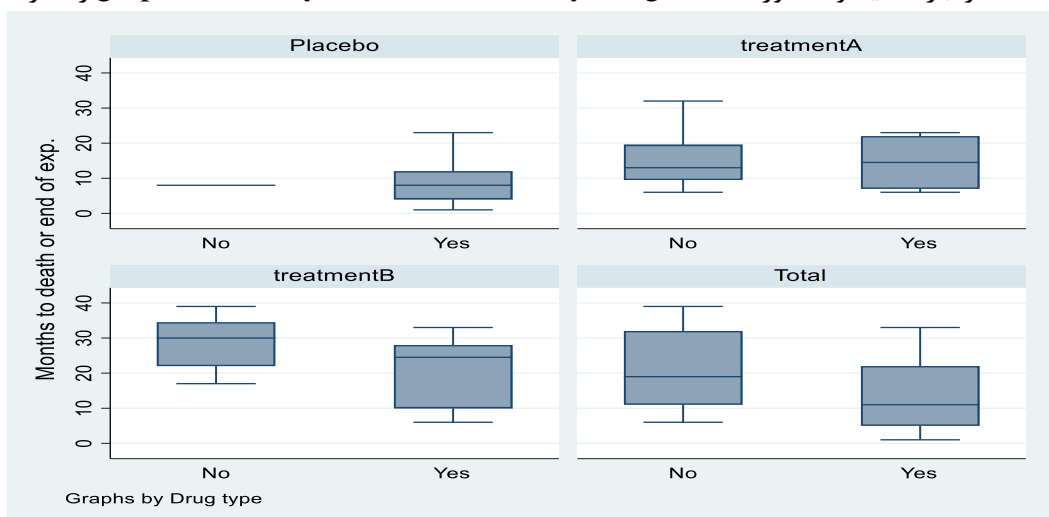
db ltable

ltable time chd, graph notable survival ci

ltable time chd, graph notable survival ci by(smoking)

یکی از نمودارهای مهم برای نمایش داده ها نمودار جعبه ای است. برای رسم نمودار جعبه فایل sysuse

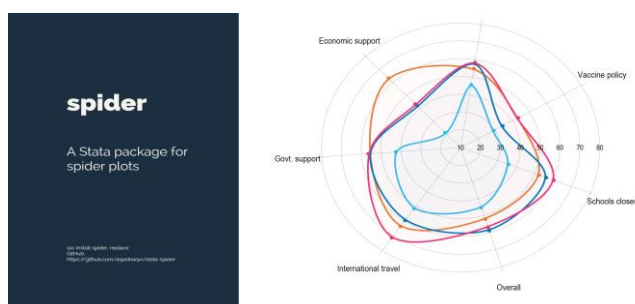
cancer.dta را باز نمائید و دستور graph box studytime, over(died) by(drug, total) را اجرا نمائید.



نمودار متغیر مدت زمان ورود به مطالعه تا مرگ (studytime) برای افراد فوت شده/ زنده مانده (yes/no) بر حسب سه گروه مورد مطالعه (پلاسبو/داروی A و داروی B) و کل افراد در بالا قابل مشاهده است.

## نمودار رادار

رادار چارت یا نمودار راداری یا نمودار عنکبوتی یا نمودار ستاره‌ای، نموداری است که با نمایش یک یا چند، چندضلعی وضعیت متغیرهای مختلف مورد نیاز را ترسیم می‌کند. با مشاهده نمودار راداری، میزان نزدیکی و شباهت متغیرهای هم جنس قابل درک خواهد بود. برای ترسیم این نمودار، منویی در نرم افزار استاتا تعبیه نشده است. شما باید پکیج ترسیم این نمودار را از آدرس <https://github.com/asjadnaqvi/stata-spider> دانلود و نصب نمایید.



با یک مثال رسم این نمودار را توضیح می‌دهیم. فرض کنید می‌خواهید نموداری برای مقایسه پنج نوع استراتژی برای کنترل کوید ۱۹ ترسیم کنید. در هر استراتژی می‌توانید وضعیت بروز کوید ۱۹ را مشاهده و مقایسه کنید. در شکل بالا استراتژی‌های واکسیناسیون، حمایت اقتصادی، کنترل حمل و نقل بین‌المللی، حمایت دولتی و بستن مدارس فرض شده است. این شکل به ما نشان می‌دهد که در بکارگیری هر استراتژی وضعیت بروز کوید ۱۹ در مقایسه با استراتژی‌های دیگر و بصورت کلی چه موقعیتی دارد. برای نصب دستورات ترسیم این نمودار باید **The SSC version (v1.21)** را نصب نمایید. دستورات زیر را اجرا و نصب نمایید.

```
ssc install spider, replace
net install spider, from("https://raw.githubusercontent.com/asjadnaqvi/stata-spider/main/installation/") replace
ssc install palettes, replace
ssc install colrspace, replace
ssc install schemepack, replace
set scheme white_tableau
graph set window fontface "Arial Narrow"
```

دستورات سینتکس نسخه ۲۱-۱ به صورت زیر است:

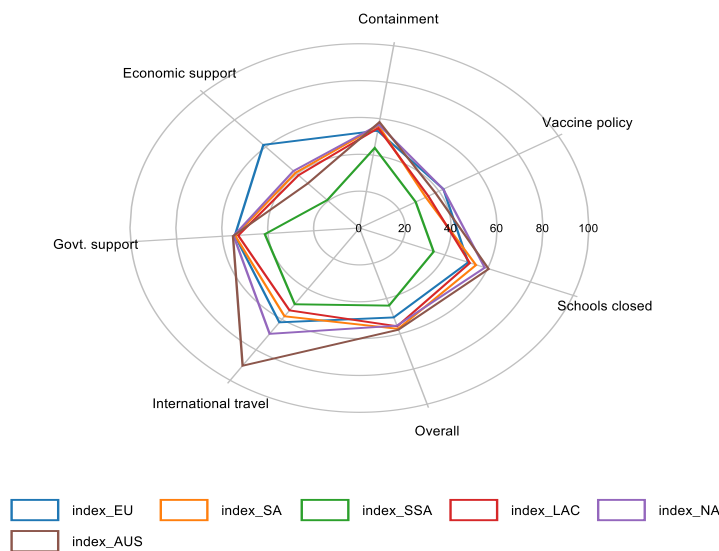
```
spider varlist [if] [in], over(cat var)
[ alpha(num 0-100) rotate(num) smooth(num 0-1) palette(str)
range(val1 val2) cuts(num) lwidth(str) msymbol(str)
raformat(fmt) ralabsize(str) ralabcolor(str) ralabangle(str)
msize(str) mlwidth(str) displacelab(num) displacspike(num)
ccolor(str) cwidth(str) scolor(str) swidth(str) slabsize(str)
nolegend legpositon(num) legpositon(num) legcolumns(num) legsize(num)
title(str) subtitle(str) note(str) scheme(str) xsize(num) ysize(num) name(str) ]
```

اکنون می توانید با دستور `spider numeric variables, over(category variable)` نمودار را ترسیم کنید.

به عنوان مثال فایل `radar_graph` را درباره استراتژی های کنترل کرونا را باز نمایید. دستور زیر را اجرا نمایید.  
`. tab policy`

GroupCode	Freq.	Percent	Cum.
Containment	766	14.29	14.29
Economic support	766	14.29	28.57
Govt. support	766	14.29	42.86
International travel	766	14.29	57.14
Overall	766	14.29	71.43
Schools closed	766	14.29	85.71
Vaccine policy	766	14.29	100.00
Total	5,362	100.00	

`spider index_EU index_SA index_SSA index_LAC index_NA index_AUS , over(policy) alpha(0) msym(none) ra(100)`



با دستور زیر نیز می توان نمودار را ارتقا داد یا نمودار صاف رسم نمود.

`spider index* , over(policy) msym(Oh) lw(0.4) ms(0.4) ra (0 100) cuts(5)`

`spider index* , over(policy) ra(10 80) cuts(8) smooth(0)`

`spider index* , over(policy) ra(10 80) cuts(8) smooth(0.5)`

`spider index* , over(policy) ra(10 80) cuts(8) smooth(1)`

`spider index* , over(policy) ra(0.1 0.7) cuts(4) raformat(%5.1f) ralabs(2)  
smooth(0) alpha(0)`

`spider index* , over(policy) msym(Oh) lw(0.4) ms(0.4) ra(0 100) ralabc(blue)  
ralabs(2) ralaba(45)`



## ترسیم نقشه با نرم افزار استاتا

برای ترسیم نمودار جغرافیایی یا نقشه باید پکیج `ssc install geoplot, replace` را نصب نمایید. `geoplot` نقشه ها را از فایل های `shape` و سایر مجموعه داده ها ترسیم می کند. روش کار به این صورت است که ابتدا یک یا چند فریم یا لایه حاوی داده های منبع را با استفاده از دستور `geoframe` ایجاد می کنیم و سپس `geoplot` را برای رسم داده های این لایه ها اعمال می کنیم. لایه های متعددی از عناصر مانند مناطق، مرزها، دریاچه ها، جاده ها، برجسب ها و نمادها را می توان آزادانه با هم ترکیب کرد. ظاهر عناصر (به عنوان مثال رنگ آنها) بسته به مقادیر متغیرها می تواند متفاوت باشد. برای نصب `geoplot` از دستور `SSC`، را تایپ کنید. با اجرای دستورات زیر

```
.ssc install palettes, replace
.ssc install colrspace, replace
.ssc install moremata, replace
```

سپس از مسیر زیر نصب پکیج را نهایی نمایید.

```
.net install geoplot, replace
from(https://raw.githubusercontent.com/benjann/geoplot/main/)
```

```
. . ssc install geoplot, replace
checking geoplot consistency and verifying not already installed...
installing into C:\Users\Administrator\ado\plus\...
installation complete.

. . ssc install palettes, replace
checking palettes consistency and verifying not already installed...
all files already exist and are up to date.

.
. . ssc install colrspace, replace
checking colrspace consistency and verifying not already installed...
all files already exist and are up to date.

.
. . ssc install moremata, replace
checking moremata consistency and verifying not already installed...
installing into C:\Users\Administrator\ado\plus\...
installation complete.

. . net install geoplot, replace from(https://raw.githubusercontent.com/benjann/geoplot/main/)
checking geoplot consistency and verifying not already installed...

the following files will be replaced:
C:\Users\Administrator\ado\plus\g\geoframe.ado
C:\Users\Administrator\ado\plus\g\geoplot.ado
C:\Users\Administrator\ado\plus\l\lgeoplot_source.sthlp
C:\Users\Administrator\ado\plus\_geoplot_symbol.ado
C:\Users\Administrator\ado\plus\_geoplot_pie.ado
C:\Users\Administrator\ado\plus\_geoplot_bar.ado
C:\Users\Administrator\ado\plus\_geoplot_layer.ado
C:\Users\Administrator\ado\plus\g\geoframe.sthlp
C:\Users\Administrator\ado\plus\l\lgeoplot.mlib
C:\Users\Administrator\ado\plus\g\geoplot.sthlp

installing into C:\Users\Administrator\ado\plus\...
installation complete.
```

مثالی برای ترسیم نقشه های اپیدمیولوژیک و پراکندگی و بکاربردن دستورات زیر:

```
local url http://fmwww.bc.edu/repec/bocode/i/
```

```
geoframe create regions `url'Italy-RegionsData.dta, id(id) coord(xcoord ycoord)  
shpfile(Italy-RegionsCoordinates.dta)
```

```
geoframe create country `url'Italy-OutlineCoordinates.dta
```

```
geoframe create capitals `url'Italy-Capitals.dta, coord(xcoord ycoord)
```

```
geoframe create lakes `url'Italy-Lakes.dta, feature(water)
```

```
geoframe create rivers `url'Italy-Rivers.dta, feature(water)
```

و سپس دستورات زیر:

```
geoplot (area regions) (line country, lwidth(medthick))
```

اکنون نقشه زیر مربوط به مناطق ایتالیا ترسیم می شود:



با دستور `geoplot (area regions) (area lakes) (line rivers)` نقشه زیر با لایه رودخانه ها و دریاچه ها رسم میشود:



سپس نقشه مناطق رنگ آمیزی شده بر اساس تعداد فالگیرها یا پیش بینی کننده ها (در هر میلیون جمعیت) را با

دستور زیر م ی‌توان ترسیم نمود:

`geoplot (area regions fortell) (line regions)`



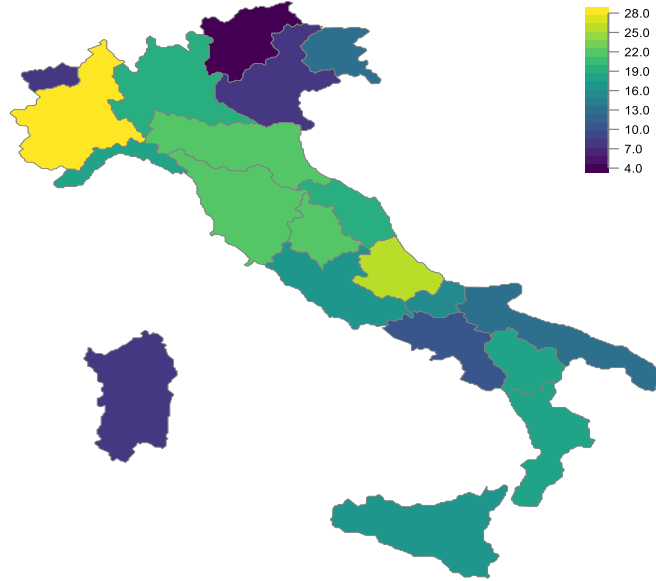
نقشه را با راهنمای متفاوتی نیز می توان با دستور زیر ترسیم نمود:

`geoplot (area regions fortell, label("@lb-@ub (N=@n)")) (line regions)`



در نسخه استاتای ۱۸ با دستور زیر می توان نقشه متفاوتی با راهنمای زیر ترسیم نمود:

geoplot (area regions fortell, levels(20) lcolor(gray)) , clegend(position(ne) height(30)) zlabel (4(3)28)



**نقشه کامل را با دستور زیر:**

geoplot (area regions) (point capitals i.size [w=pop98], color(Set1, opacity(50)) mlcolor(%0))  
(label capitals city if pop98>250000, color(black)) , legend compass sbar(length(300) units(km))

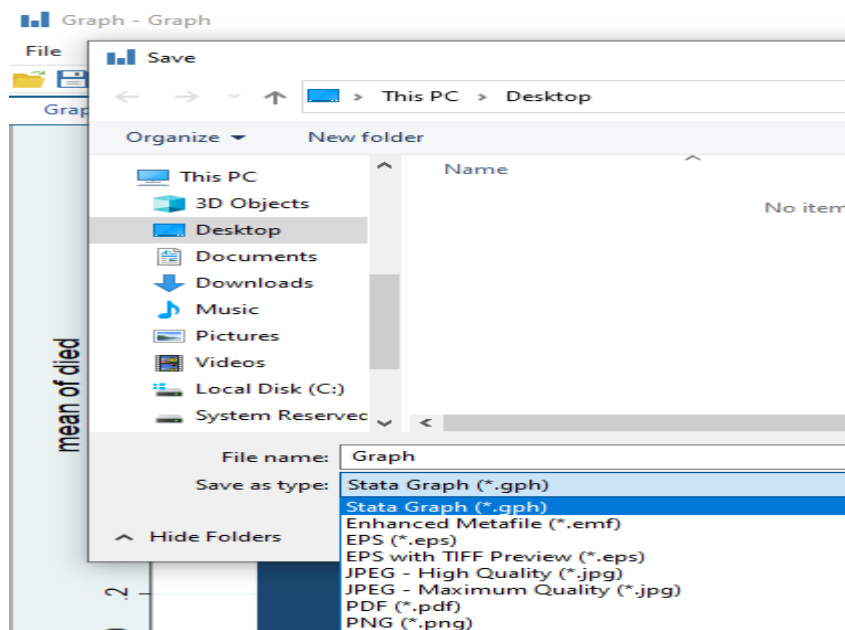


## خلاصه و نتیجه گیری:

در این فصل انواع نمودارها و نحوه ترسیم آنها با نرم افزار بیان گردید. نحوه ترسیم نقشه نیز ارائه شد، اما

Group	Option
Graph titles	title(text, size( )) subtitle(text, size( )) caption(text, size( )) note(text, size( ))
with by	place the above options inside the by( )
Axes	xtitle(text, size( )) ytitle(text, size( )) xlabel(numlist, labsize( ) angle( )) ylabel(numlist, labsize( ) angle( )) xscale(range(numlist) log) yscale(range(numlist) log)
Added lines	xline(#, lpattern( ) lcolor( )) yline(#, lpattern( ) lcolor( ))
Marker symbols	msymbol( ) msize( ) mcolor( ) mlabel( )
Connect style	connect( )
Legends	legend(label(# "text") label(# "text") ...) legend(order(# "text" # "text") ...)

نیاز به دقت بالا و دسترسی به نقاط طول و عرض جغرافیایی مکان هاست. برخی از دستورات تکمیلی نمودارها در بالا آمده است. از منوی فایل نمودار رسم شده می توان نمودار را با فرمت های مختلف که در شکل زیر مشخص است با انتخاب مسیر مناسب ذخیره نمود.



**تمرین ها:**

- ۱- فایل part four را باز نمایید. نمودار هیستوگرام برای متغیر chol و sbp ترسیم نمایید.
- ۲- نمودار دایره ای برای متغیر smoking رسم نمایید.
- ۳- نمودار کاپلان مایر را برای فایل مذکور با در نظر گرفتن متغیر time و متغیر all بعنوان متغیر رخدادهای پیامد ترسیم نمایید.
- ۴- نمودار جدول عمر این متغیرها را ترسیم نمایید.
- ۵- نمودار جدول عمر این متغیرها را بر حسب متغیر smok ترسیم نمایید.
- ۶- فایل ahmadi1400 که در فصل ۱ ساخته اید را باز نمائید. نموداری برای تعداد مردان و زنان و بر حسب گروه های مطالعه ترسیم کنید.
- ۷- برای کارهای انجام شده do-file تشکیل دهید. کارهای انجام شده را ذخیره نمایید.

## بخش سوم:

# آمار تحلیلی و کار با نرم افزار Stata

## فصل پنجم: آزمون فرضیه ها و آنالیز دو متغیره و چند متغیره در Stata

آمار تحلیلی، انواع فرضیه و خطاها  
آزمون های آماری  
بررسی فرض نرمال بودن داده ها  
مقایسه میانگین ها و انواع آزمون های t  
Chi square  
ANOVA  
آزمون های تعقیبی Post Hoc یا مقایسه های چندگانه  
آزمون های ناپارامتریک  
MANOVA  
ANCOVA  
MANCOVA  
Repeated Measure ANOVA

### • اهداف:

#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند داده ها را برای آنالیز تحلیلی آماده نماید.
- انواع آزمون های آماری را بتواند با کاربرد تشریح نماید.
- کاربر بتواند آزمونهای z، t، F، کای اسکور، آنالیز واریانس یکطرفه و دو طرفه، ANOVA، MANOVA، ANCOVA، MANCOVA، Repeated Measure را با نرم افزار انجام دهد.
- کاربر بتواند تفسیر مناسبی از آزمون های تحلیل دومتغیره و چندمتغیره ارائه دهد.
- کاربر نقش و نحوه کنترل متغیر مخدوش کننده در آنالیز ANCOVA تفسیر نماید.

## تعاریف مربوط به آمار تحلیلی و آزمون های آماری:

در فصل گذشته جمع آوری، تنظیم و خلاصه کردن داده ها و نمایش و ترسیم با جدول، نمودار، نقشه و شاخصها، آمار توصیفی (Descriptive Statistics) تعریف گردید. در این فصل به آمار تحلیلی پرداخته می شود. در ساده ترین شکل تحلیل، پژوهشگر به دنبال وجود ارتباط یا اختلاف بین دو متغیر است. در تحقیقات علوم پزشکی آزمون فرضیه (Hypothesis Testing) از اهمیت فراوانی برخوردار است. زیرا به محقق این امکان را می دهد که نتایج حاصل از نمونه تحت مطالعه را به جامعه مرجع تعمیم دهد. فرضیه پژوهش حدس یا تصویری است که انگیزه پژوهش قرار می گیرد. فرضیه تفکری است که از مشاهده پدیده های موجود در طبیعت ناشی می شود. مثلا مردان از زنان قد بلندتر هستند. هدف محقق از انجام آزمون فرضیه اثبات این مطلب است که نتایج یافته هایی که از مطالعه بدست آمده است از لحاظ آماری معنی دار هستند. آزمون فرضیه به قبول یا رد این گفته که نتایج مشاهده شده صرفا به دلیل شانس روی نداده اند بلکه به دلیل رابطه صحیح بین متغیرهای وابسته و مستقل روی داده اند، می پردازد. بنابراین آزمون فرضیه ادعایی درباره پارامتر یک یا چند جامعه می باشد و یک فرآیند تصمیم گیری است که این فرآیند مبتنی بر ذهنیاتی که از طریق مشاهده مستقیم و ساده اطلاعات بدست می آید نبوده، بلکه براساس عینیتی است که با محاسبه احتمال وقوع فرضیات مختلف حاصل می شود.

یک آزمون فرضیه روشی است که فرضیه ای که باید آزمون شود (فرضیه صفر) را بیان میکند. از اطلاعات نمونه استفاده کرده و یک قاعده تصمیم گیری را فرمول بندی می کند. بر مبنای نتیجه قاعده تصمیم گیری یک فرضیه به صورت آماری پذیرفته یا رد می شود. به طور کلی آزمون فرضیه ها و استنباط آماری به اثبات فرضیه نمی انجامند، بلکه تنها بر این نکته دلالت دارند که آیا فرضیه بوسیله داده های موجود تایید می شود یا نه. مثال: پژوهشگری رابطه بین مصرف سیگار و مرگ قلبی عروقی را مورد مطالعه قرار می دهد. یا مثال دیگر که با مشخص بودن نوع فرضیه ها در رابطه با رژیم غذایی و سطح ویتامین دی است در زیر می آید.

$H_0$ : رژیم غذایی جدید میانگین سطح ویتامین دی را تغییر نمی دهد

$H_a$ : رژیم غذایی جدید میانگین سطح ویتامین دی را کاهش می دهد.

**فرضیه صفر (Null Hypothesis):  $H_0$** 

فرضیه صفر وضع موجود را نشان میدهد. گاهی آن را فرضیه عدم تفاوت یا خنثی نیز می نامند. چون این فرضیه حاکی از موافقت یا عدم تفاوت با شرایط واقعی جمعیت مورد نظر را بیان می کند. فرضیه صفر همان فرضیه ای است که می خواهیم غلط بودن آن را به اثبات برسانیم (استفاده از برهان خلف).

در این مطالعه (رابطه بین مصرف سیگار و مرگ قلبی عروقی)، فرضیه صفر بیانگر آن است که تفاوتی بین افراد سیگاری و غیرسیگاری به عنوان عامل خطر بروز مرگ قلبی عروقی وجود ندارد و هر نوع تفاوت مشاهده شده، صرفا به دلیل ناشی از شانس یا تصادف (Random) است. فرضی که پژوهشگر می خواهد آن را آزمون کند عبارتست از اینکه آیا سیگاری ها نسبت به غیرسیگاری ها در معرض خطر بیشتری برای بروز مرگ های قلبی عروقی هستند. مثال یک: میانگین وزن نوزادان در طبقات محروم جامعه با مقدار استاندارد در جامعه یکسان است.

$$H_0 : \mu = 3000 \text{ gr}$$

$$H_0 : P_A = P_B$$

مثال دو: میزان بروز عوارض داروی A و داروی B یکسان است.



### فرضیه تحقیق (Alternative Hypothesis) $H_1$ :

این فرضیه درست برخلاف فرضیه  $H_0$  بیان می شود و ادعای محقق را نشان می دهد. فرضیه تحقیق بیانگر این است که بین افراد سیگاری و غیر سیگاری در رابطه با خطر بروز مرگ قلبی عروقی تفاوت وجود دارد و تفاوت های مشاهده شده به دلیل شانس نیست. اگر یافته های مطالعه از لحاظ آماری معنی دار تلقی گردند و فرضیه صفر صحیح نباشد، آنگاه می توان فرضیه صفر را رد و فرض مقابل را پذیرفت.

در مثال یک: میانگین وزن نوزادان در طبقات محروم جامعه کمتر از استاندارد است.

$$H_1 : \mu < 3000$$

در مثال دو : میزان عوارض داروی A و B یکسان نیست.

$$H_1 : P_A \neq P_B$$

### فرضیه یک طرفه (One-Tailed Hypothesis)

اینکه یک پارامتر جامعه کوچکتر یا بزرگتر از مقداری است را نشان می دهد.

$$(\mu < \mu_0), (\mu > \mu_0)$$

### فرضیه دو طرفه (Two-Tailed Hypothesis)

اینکه یک پارامتر جامعه، مخالف مقداری از پیش تعیین شده است را بیان می کند.

چون آزمون فرضیه بر مبنای داده های نمونه می باشد بنابراین ممکن است در تصمیم گیری دچار خطا شویم. برای تصمیم گیری در رد یا قبول فرضیه ها خطا ها را به خطای نوع اول و خطای نوع دوم تقسیم می کنند.

**خطای نوع اول**، رد فرضیه  $H_0$  وقتی آن درست می باشد تعریف می شود.

**خطای نوع دوم**، قبول فرضیه  $H_0$  وقتی آن غلط می باشد تعریف می شود.

ماکزیمم احتمال ارتکاب خطای نوع اول را با  $\alpha$  و سطح معنی داری نیز گفته می شود. احتمال ارتکاب خطای نوع دوم را با  $\beta$  نشان می دهند. نکته: بهترین آزمون فرضیه آزمون  $\alpha$  ثابت دارای  $\beta$  مینیمم باشد. برای کاهش

احتمالات هر دو خطای نوع اول و دوم ما نیاز به افزایش حجم نمونه داریم.

جدول زیر این خطاها را نشان می دهد.

نتیجه آزمون		شرایط واقعیت	
نوع تصمیم	قبول $H_0$	$H_0$ درست	$H_0$ غلط
	خطای نوع دوم	قضای صحیح	خطای نوع دوم
	خطای نوع اول	قضای صحیح	خطای نوع اول

مثال دیگر از خطاهای آماری: رابطه بین استفاده زیاد از تلفن همراه و کاهش شنوایی

رابطه ای وجود ندارد	واقعا رابطه وجود دارد	
رد فرضیه صفر به غلط: باز هم ادعای بی جا خطای نوع اول	آفرین: پس باید تلفنهای همراه را شکست	رابطه کشف می شود
آفرین: پس با خیال راحت از تلفن همراه استفاده کنیم	قبول فرضیه صفر به غلط: عدم توان یافتن حقیقت خطای نوع دوم	رابطه کشف نمی شود

مراحل انجام یک آزمون فرضیه:

- ۱- به صورت دقیق فرضیه ها بیان می شوند. البته ابتدا باید ماهیت دقیق داده ها و مقیاس اندازه گیری آنها درک گردد. مفروضات تعیین گردند. مثلا توزیع جمعیت، تساوی واریانس ها و استقلال نمونه ها.
- ۲- آماره آزمون مناسب ، با توجه به پارامتر مورد بررسی مشخص می شود. توزیع احتمال آن نیز مشخص می گردد.

برآورد نقطه ای - مقدار فرضی برای پارامتر

آماره آزمون =

انحراف معیار برآورد

۳- سطح معنی داری آزمون مشخص می شود. این سطح مساحتی از زیر منحنی توزیع شاخص آماری آزمون است که بالای مقادیر روی محور افقی است که ناحیه رد را تشکیل می دهد. داده های نمونه را جمع کرده و با توجه به آن آماره آزمون محاسبه می شود. احتمال خطای نوع اول سطح معنا داری یک آزمون نامیده می شود. معمولاً پژوهشگر سطح معنا دار معینی را انتخاب میکند. انتخاب سطوح معنا داری ۵٪ یا ۱٪ اغلب در بین پژوهشگران مرسوم و متداول است. استراتژی ما در برخورد با مسئله آزمون، کنترل خطای نوع اول است که خطای مهمی است. معمولاً محقق مایل است حداقل به ازای مقادیری که از دامنه مقادیر تعیین شده برای  $H_0$  (فرض صفر) دور است، احتمال خطای نوع دو در یک مقدار قابل اغماضی باقی بماند.

۴- توجه به فرضیه  $H_1$  و مقدار و توزیع احتمال آماره آزمون ناحیه رد و قبول فرضیه  $H_0$  مشخص میشود.

۵- اگر مقدار آزمون در ناحیه قبول  $H_0$  قرار گرفت فرضیه  $H_0$  را می پذیریم در غیر این صورت فرضیه  $H_0$  را رد می کنیم. مقدار آلفا در حقیقت احتمال رد کردن فرضیه صفر است. در صورتی که در واقع این فرضیه درست باشد(تصمیم گیری آماری). البته پس از آن تصمیم گیری بالینی، پزشکی و بهداشتی درمانی یا اداری انجام می شود. یعنی مقدار اختلاف و سطح معنی داری باید از نظر تفسیر کاربر نیز با اهمیت باشد. مقدار احتمال (P Value)، کمترین مقداری از احتمال خطای نوع اول (سطح آزمون) که ممکن است یافته آماره آزمون، موجب رد فرض صفر شود. P-value به شما می گوید که چقدر احتمال دارد که نتایج شما به دلیل شانس باشد، نه یک نتیجه واقعی.

### خطاهای متداول:

در تحقیقات علوم پزشکی برای پاسخ به سوالات و یا رد و قبول فرضیات، نیازمند اندازه گیری متغیرها هستیم. همیشه اندازه گیری دقیق و با حداقل خطا از دغدغه های اصلی محققان محسوب می گردد. در تقسیم بندی ساده خطاها به چهار زیر گروه اصلی تقسیم می شوند.

۱- خطای عدم توافق (Disagreement Error): زمانی بروز می کند که یک اندازه گیری توسط چند ابزار و یا چند فرد اندازه گیری کننده که با هم دقیقاً هم نظر نباشند، انجام شود.

۲- خطای تصادفی (Random Error): زمانی بروز می کند که مقدار اندازه گیری شده با مقدار واقعی تفاوت داشته باشد ولی همه اندازه گیری ها به صورت منظم به یک سمت متمایل نباشند.

۳- خطای منظم (Systematic Error) یا سوگیری (Bias)، زمانی دیده می شود که اندازه گیری ها به یک سمت متمایل باشند.

۴- خطای مخدوش کننده (Confounding Error): خطایی است که در مطالعات تحلیلی ایجاد می شود و روابط را به دلیل تاثیر متغیرهای مخدوش کننده کمتر یا بیشتر از مقدار واقعی نشان می دهد. در نظر گرفتن خطاها در سنجش متغیرها و پاسخ به سوالات و ارزیابی آزمونهای آماری، ضروری به نظر می رسد. در بخش پنجم کتاب نیز به این خطاها در مطالعات اپیدمیولوژیک و نحوه محاسبه آنها با نرم افزار پرداخته می شود.

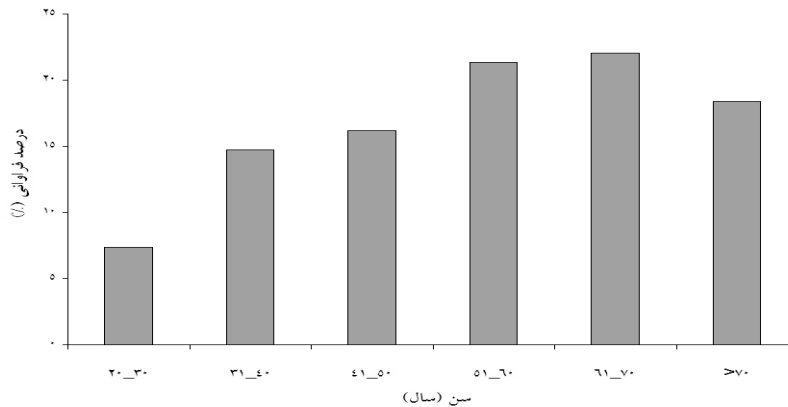
### تحلیل آماری پارامتریک و ناپارامتریک:

غالباً با توجه به توزیع داده ها و در صورت برقراری فرض نرمال بودن برای جامعه ای که نمونه از آن گرفته شده است و وجود نداشتن نقاط دورافتاده یا با بهره گیری از قضیه حد مرکزی و مستقل بودن مشاهدات، برای انجام آزمونهای آماری از روشهای کلاسیک استفاده می شود. با دانستن نوع فرضیه و سوال مطالعه، نوع مقیاس و تعداد متغیرها و شرایط زیر محقق نوع آزمون آماری را انتخاب می کند.

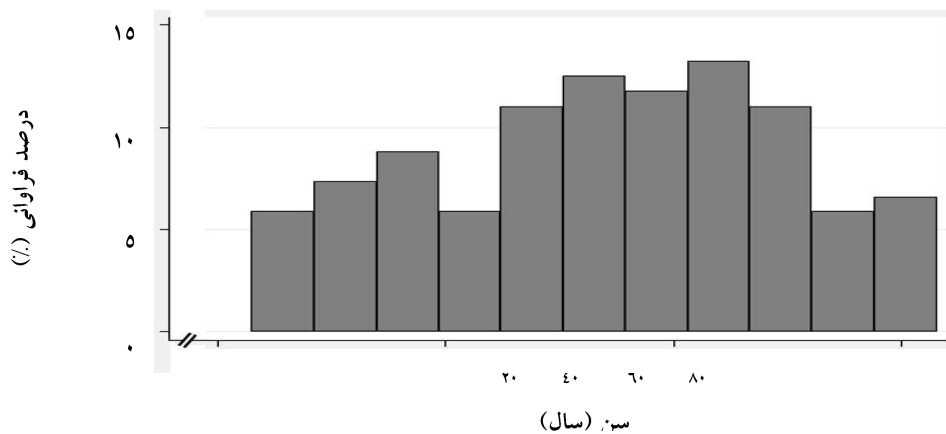
- مقایسه داده های دو گروه مستقل یا بیشتر (تعداد گروه ها) و یا مقایسه با عدد ثابت
- مقایسه دو یا بیشتر از دو گروه از داده های وابسته و تعداد بار اندازه گیری ها
- اندازه گیری ارتباط بین دو صفت
- وضعیت متغیر پیامد

برای تصمیم گیری نوع تحلیل و انتخاب آزمون مناسب از فلوجارت های ضمیمه شماره دو کتاب استفاده می شود. برای کسب اطلاع بیشتر به مراجع آماری و از جمله کتاب روش های آماری و شاخص های بهداشتی دکتر کاظم محمد یا به کتاب اصول و روش های آمار زیستی پرفسور دکتر آیت اللهی مراجعه نمایید. توزیع های زیادی در آمار مورد استفاده قرار می گیرد. در این بخش به توزیع های مهم پرداخته می شود.

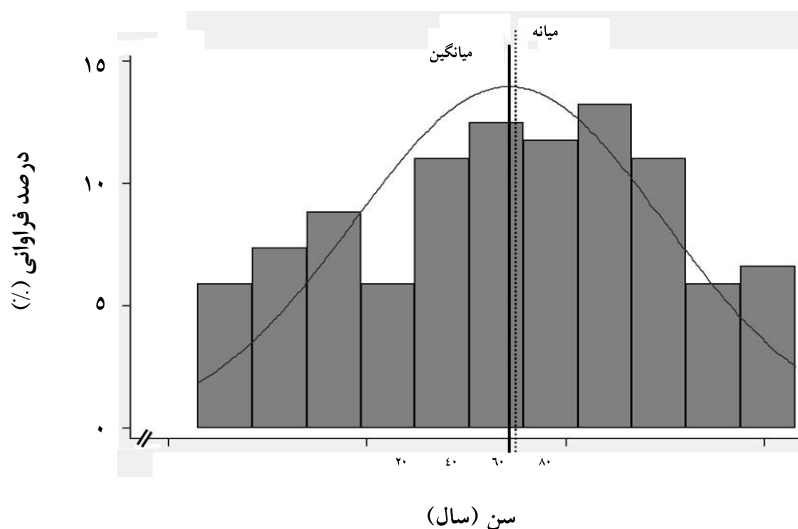
**توزیع نرمال و کاربردهای آن در آزمونهای آماری:** برای نمایش داده های کمی (مانند سن) معمولاً آنرا به دسته هایی تقسیم کرده و سپس فراوانی مطلق (frequency) یا درصد فراوانی (Percent) هر یک از دسته ها را با استفاده از نمودار ستونی (شکل زیر) نشان می دهند. اینگونه گزارشات آماری و اقدامات مشابه در فصول گذشته توضیح داده شد.



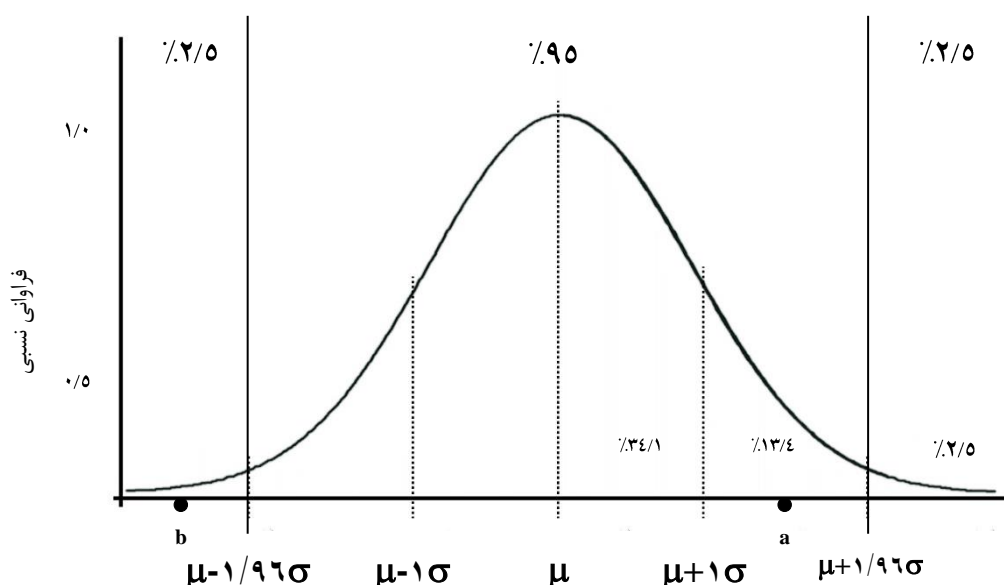
تبدیل متغیر سن به دسته های مختلف، باعث گردید تا توزیع فراوانی سن حالت گسسته یا منقطع پیدا کند، درحالیکه خود متغیر سن حالت کمی پیوسته دارد. برای برطرف کردن این مشکل، نمودار هیستوگرام توصیه شده است. در این نمودار، متغیر کمی پیوسته در محور X و فراوانی مربوطه در ستون Y نمایش داده می شود. در اینجا، اگرچه نمودار هیستوگرام، پیوستگی داده ها و توزیع آنها را بهتر نشان می دهد، اما همچنان حالت مضرس یا دنداندار دارد.



یکی از مهمترین توزیعهای فراوانی (معمولاً برای متغیرهای کمی پیوسته)، توزیع نرمال است. همانطور که در شکل زیر مشاهده می کنید، نمودار توزیع نرمال برای سن به نمودار هیستوگرام بالا اضافه شده است. توزیع نرمال سن، به خوبی توزیع پیوسته سن، مرکز توزیع و پراکندگی داده ها را نشان می دهد. توجه داشته باشید که توزیع نرمال سن، تنها برای ۱۳۵ نفر (۲۳ تا ۸۳ سال) ترسیم شده است.



توزیع نرمال از نوع پیوسته است زیرا هر مقداری اعم از عدد صحیح یا عدد اعشاری را می تواند قبول کند. منحنی این توزیع به صورت زنگوله ای است و نسبت به میانگین جمعیت که در وسط آن است حالت قرینه دارد. فاصله میانگین از هر یک از نقاط عطف (جایی که تحدب یا تقعر تغییر می کند) برابر با یک انحراف معیار توزیع است. چون منحنی توزیع نرمال تمامی حالات (۱۰۰٪ افراد) جمعیت را شامل می شود. سطح زیر منحنی با احتمال وجود همه مقادیر افراد (۱۰۰٪ افراد) در توزیع، برابر با یک است. بدلیل شکل قرینه نرمال، می توان منحنی را از میانگین به دو قسمت مساوی تقسیم کرد. شکل زیر توزیع نرمال را برای میانگین  $\mu$  و انحراف معیار  $\sigma$  نشان می دهد.



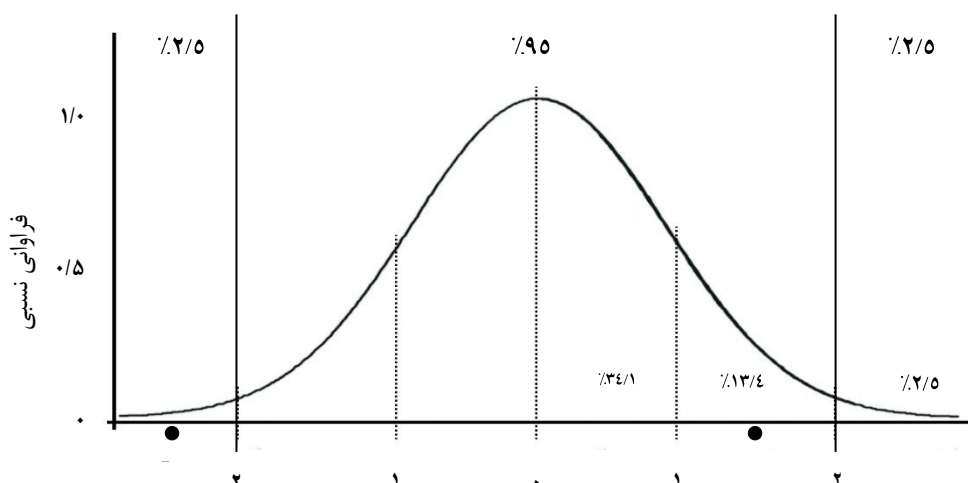
همانطور که در نمودار بالا مشاهده می شود ۹۵٪ داده ها در فاصله  $\mu \pm 2\sigma$  قرار دارند پس می توان چنین نتیجه گیری کرد که نقطه  $a$  در ۹۵٪ مرکز توزیع قرار دارد در حالی که نقطه  $b$  در ۵٪ انتهایی توزیع جای دارد. نکته : می توان با انتگرال گرفتن از معادله منحنی به ازای مقادیر  $a$  و  $b$  مشخص نمود که چه مقدار از داده ها بین فاصله  $a$  تا  $b$  قرار دارند. توجه داشته باشید که تنها متغیرهای منحنی فوق  $\mu$  و  $\sigma$  هستند پس برای هر جمعیت یک منحنی مخصوص به خودش وجود دارد.

### توزیع نرمال استاندارد (Z):

انتگرال گرفتن برای هر یک از نقاط و بدست آوردن سطح زیر منحنی برای آنها، کار دشواری است. پس اگر راهی وجود داشته باشد تا بتوان همه توزیع ها را به یک توزیع یکسان ( $\mu$  و  $\sigma$  یکسان) تبدیل کرد، می توان سطحهای زیر منحنی به ازای مقادیر مختلف را برای آن توزیع یکسان در جدولی تنظیم نمود و از آن به جای انتگرال گیری استفاده کرد. استفاده از تبدیل Z راه حل است. توزیعی که مقادیر Z را نشان می دهد، توزیعی نرمال، که همواره میانگین آن صفر و انحراف معیار آن برابر با یک است. برای تبدیل توزیع هر جمعیتی به توزیع Z کافی است که هر یک از داده های آنرا از میانگین جمعیت کم کرده و مقدار حاصل را بر انحراف معیار جمعیت تقسیم نمود.

$$Z = \frac{X - \mu}{\sigma}$$

شکل توزیع نرمال استاندارد همانند توزیع نرمال است با این تفاوت که میانگین و انحراف معیار آن ثابت و برابر ۱۰ است.



**مثال:** اگر توزیع طول قد جامعه ای با میانگین ۱۷۲/۵ سانتیمتر و انحراف معیار ۷/۵ سانتی متر نرمال باشد، طول

$$Z = \frac{178 - 172.5}{7.5} = 0.73$$

چه نسبتی از افراد بلندتر از ۱۷۸ سانتیمتر است؟  
در جدول Z مقدار ۰/۷۳ را پیدا می کنیم مطابق با جدول ۰/۲۳/۲۷ افراد مقادیر Z بیشتر از ۰/۷۳ دارند پس در جمعیت فوق هم ۰/۲۳/۲۷ افراد هم قدی بیشتر از ۱۷۸ سانتیمتر دارند.

### چگونه از توزیع Z برای تفسیر آماری استفاده می شود؟

بر اساس قضیه حد مرکزی توزیع میانگین نمونه ای توزیعی است نرمال با میانگین  $\mu$  (میانگین جمعیت کل) و انحراف معیار  $\sigma/\sqrt{n}$  پس می توان تبدیل را برای آن بدین صورت ترسیم کرد:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

گاهی اوقات نیاز است که مقدار یک کمیت (مثلاً قد افراد جامعه) را از روی تعداد محدودی نمونه پیش بینی کنید. مسلماً هرگز نمی توان با تعداد محدود دقیقاً مشخص کرد که میانگین قد افراد جامعه چقدر است اما با احتمالی مشخص می توان یک محدوده را برای آن پیشگویی کرد. این عمل برآورد فاصله ای (Interval Estimate) نام دارد.

**مثال ۱:** در مطالعه بررسی سن ابتلا به عفونت ادراری، تعداد ۱۳۵ بیمار مورد بررسی قرار گرفتند. بعد از جمع آوری داده ها، نتایج به صورت زیر بود:

$$\text{حجم نمونه } (n) = 135$$

$$\text{میانگین سن ابتلا } (\bar{x}) = 54/23 \text{ (سال)}$$

$$\text{انحراف معیار } (SD) = 15/56 \text{ (سال)}$$

در اینجا تنها نمونه ای از بیماران (یعنی ۱۳۵ نفر) مورد ارزیابی قرار گرفته اند. خیلی بعید است که میانگین سن ابتلا در جمعیت کل بیماران هم برابر با ۵۴/۲۳ سال باشد. دلیل آنهم وجود خطای نمونه گیری است. مقدار

میانگین واقعی (مربوط به جمعیت) را  $\mu$  می نامیم. این مقدار قابل اندازه گیری نیست. بنابراین، هدف ما یافتن فاصله ای است که به احتمال زیاد  $\mu$  را در بر بگیرد.

روشی که برای محاسبه این فاصله استفاده می شود، یافتن توزیع مقادیر  $\bar{x}$  های بدست آمده از تعداد زیاد نمونه های ۱۳۵ تایی ( $n=135$ ) است. به این توزیع، توزیع میانگین های نمونه ای (Sampling Distribution) گفته می شود.

در اینجا هم یک توزیع زنگوله ای شکل (توزیع نرمال) بوجود خواهد آمد که در آن، اکثر مقادیر  $\bar{x}$  در اطراف مقدار میانگین واقعی یعنی  $\mu$  قرار دارند، و تنها ۱٪ از ۲۰٪ (۵٪) از آنها دورتر از فاصله ۱/۹۶ خطای معیار از  $\mu$  قرار می گیرند. فرمول محاسبه خطای معیار میانگین (Standard Error of Mean – SE) به صورت زیر است :

$$SE = \sigma / \sqrt{n}$$

در این فرمول،  $n$  حجم نمونه (۱۳۵ نفر)، و  $\sigma$  انحراف معیار سن جمعیت بیماران مبتلا است.

**مثال ۲:** توزیع صفتی در جامعه تقریباً نرمال، میانگین و انحراف معیار آن به ترتیب برابر با ۴۰ و ۴ است. اگر از این جامعه نمونه تصادفی ۶۴ نفری انتخاب شود، احتمال اینکه میانگین نمونه از عدد ۲۱ بزرگتر باشد، چقدر است؟

$$\mu = 40$$

$$\sigma = 4$$

$$Z = (21 - 40) \div (4 \div \sqrt{64}) = 1 \div 0.5 = 2$$

مطابق با جدول  $Z$ ، چنین تفسیر می شود که تنها میانگین ۲/۲۸٪ از نمونه های ۶۴ تایی بزرگتر از ۲۱ می باشد.

مطابق با آنچه درباره خصوصیات توزیع نرمال گفته شد، برای هر میانگین بدست آمده از نمونه های ۱۳۵ تایی (مثل  $\bar{x} = 54/23$ )، ۹۵٪ احتمال دارد که در فاصله  $1/96 SE$  از میانگین  $\mu$  قرار دارد. به عبارتی ۹۵٪ محدوده اطمینان (Confidence Interval) برای  $\mu$  برابر است با

$$\bar{x} - 1/96 SE(\bar{x}) \text{ تا } \bar{x} + 1/96 SE(\bar{x})$$

و یا

$$\bar{x} - 1/96 \sigma / \sqrt{n} \text{ تا } \bar{x} + 1/96 \sigma / \sqrt{n}$$

**مثال ۳:** برای برآورد قد جمعیت مردان یک شهر، تعداد ۲۵ نفر بطور تصادفی انتخاب شدند. انحراف معیار جمعیت مردان برابر ۶/۵ سانتیمتر است. اگر میانگین قد افراد انتخاب شده ۱۷۴ سانتیمتر و انحراف معیار آن ۵ سانتیمتر باشد، به احتمال ۹۵٪ قد مردان آن شهر، بین چه مقادیری می باشد؟ (توزیع قد در جامعه نرمال است). این مسأله با مثالهای گذشته تفاوت دارد. در اینجا سطح زیر نمودار معلوم است پس می توان با استفاده از جدول مقادیر  $Z$  را پیدا کرد. ابتدا مقدار ۹۵٪ را به دو قسمت ۴۷/۵٪ بالایی و پایینی توزیع تقسیم می کنیم برای  $Z$  دو مقدار مساوی ۱/۹۶ بدست می آید. با جایگزینی هر یک از مقادیر  $Z$  در تساوی های زیر، حد بالا و پایین ۹۵٪ توزیع قد افراد بدست می آید.

$$\rightarrow Z_{0.975} = 1/96 \text{ محدوده اطمینان } 95\%$$

$$\bar{x} + 1/96 \sigma/\sqrt{n} = 174 + 1/96 \times (6/5 \div 5) = 174 + 2/54 = 176/54 \quad \text{حد بالا}$$

$$\bar{x} - 1/96 \sigma/\sqrt{n} = 174 - 1/96 \times (6/5 \div 5) = 174 - 2/54 = 171/46 \quad \text{حد پایین}$$

بنابراین، به احتمال ۰.۹۵، فاصله ۱۷۱/۴۶ تا ۱۷۶/۵۴ سانتیمتر، میانگین قد مردان را شامل می شود. مقدار  $\bar{x}$ ، برآورد نقطه ای (Point Estimate) میانگین جمعیت کل می باشد. این مقدار کاربرد چندانی ندارد، زیرا بدلیل خطای نمونه گیری با هر نمونه گیری تغییر می کند. فاصله اطمینان تا حدی این مشکل را برطرف می کند.

مشکل اصلی مربوط به فرمول  $\mu = \bar{x} \pm 1/96 \sigma/\sqrt{n}$  برای محاسبه محدوده اطمینان ۰.۹۵ میانگین جامعه، این است که معمولاً مقدار  $\sigma$  معلوم نیست. یکی از راه حلها جایگزین کردن SD به جای  $\sigma$  است. یعنی خواهیم داشت:

$$\bar{x} - 1/96 SD/\sqrt{n} \quad \text{تا} \quad \bar{x} + 1/96 SD/\sqrt{n}$$

چون SD تنها برآوردی از  $\sigma$  است، بهتر است بدلیل وجود خطای نمونه گیری در آن، مقدار فاصله اطمینان را بیشتر بگیریم. در واقع، در اینجا به جای توزیع Z، از توزیع دیگری به نام توزیع t استفاده می کنیم:

$$\bar{x} - t_{df, 0.975} SD/\sqrt{n} \quad \text{تا} \quad \bar{x} + t_{df, 0.975} SD/\sqrt{n}$$

مقدار t را از جدول مربوط به آن با در اختیار داشتن درجه آزادی (degree of Freedom – df) و سطح اطمینان مورد نظر بدست می آوریم. مقدار درجه آزادی در اینجا برابر با n-۱ است.

در مثال ۳، اگر انحراف معیار جمعیت را در اختیار نداشتیم، می توان محدوده اطمینان ۰.۹۵ را با استفاده از توزیع t و از روش زیر محاسبه کرد:

$$n = 25$$

$$\bar{x} = 174$$

$$SD = 5$$

$$df = n - 1 = 25 - 1 = 24$$

$$0.95 \text{ محدوده اطمینان} \rightarrow \alpha = 0.05 \rightarrow t_{24, 0.975} = 2.064$$

$$\bar{x} + t_{df, (1-\alpha/2)} SD/\sqrt{n} = 174 + 2.064 \times (5/\sqrt{25}) = 176.64$$

$$\bar{x} - t_{df, (1-\alpha/2)} SD/\sqrt{n} = 174 - 2.064 \times (5/\sqrt{25}) = 171.36$$

به عبارتی، به احتمال ۰.۹۵، فاصله ۱۷۱/۳۶ تا ۱۷۶/۶۴ سانتیمتر، میانگین قد مردان را شامل می شود.

توزیع t را توزیع Student t نیز می نامند و شامل توزیع میانگین هایی است با حجم نمونه های اندک. توزیع t مانند توزیع Z، نسبت به میانگین صفر قرینه است، ولی انحراف معیار آن بستگی به پارامتری به نام درجه آزادی دارد. پارامتر تعیین کننده درجه آزادی است. با افزایش درجه آزادی شکل آن به سمت توزیع نرمال می رود. در هر حال، اگر درجه آزادی (حجم نمونه) افزایش یابد، انحراف معیار این توزیع به سمت یک میل می کند و توزیع T عیناً برابر با توزیع Z می شود. حول میانگین متقارن است.

منحنی توزیع T نسبت به توزیع Z دارای ارتفاع کمتر، ولی دو دنباله آن بلندتر و کشیده تر است. مخصوصاً برای حجم نمونه پایین، انحراف معیار آن بزرگ است. اکثر پژوهشگران در اکثر مواقع حتی زمانیکه اندازه نمونه هم بزرگ است، توزیع T را به کار می برند. پس توزیع تی همواره پراکنده تر از توزیع نرمال استاندارد است.



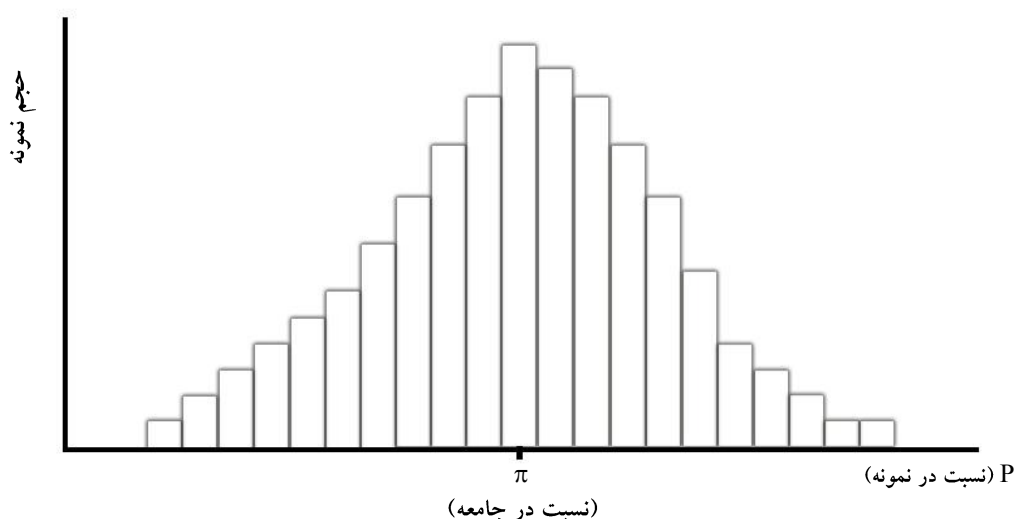
در آمار استنباطی واژه هایی کاربرد دارد که شناخت، فهم و درک این واژه ها و کاربرد آنها از اهمیت فوق العاده ای برخوردار است. **استنتاج آماری** فرایندی است که به ما امکان نتیجه گیری درباره یک جمعیت از یافته های مشاهده شده در یک نمونه را می دهد. روشهای آماری که در آن بکار می روند، بر این پیش فرض استوارند که نمونه تحت مطالعه بطور تصادفی از جمعیت مورد نظر (که بطور دقیق از قبل تعریف شده است) انتخاب شده باشد. یعنی، همه افراد جمعیت شانس یکسانی برای قرار گرفتن در نمونه مشاهده شده داشته باشند. معمولاً، انتخاب یک نمونه کاملاً تصادفی ممکن نیست و در این مواقع، همواره باید نگران تورش یا خطای انتخاب (Selection Bias) باشیم.

محدوده اطمینان (Confidence Interval : CI) تنها تاثیر خطای نمونه گیری (Sampling error) بر دقت برآورد بدست آمده را نشان می دهد و هیچ کنترلی بر خطاهای دیگر مانند تورش در زمان طراحی، اجرا و آنالیز داده ها ندارد. نسبت در جمعیت (Population Proportion:  $\pi$ ) و نسبت در نمونه (Sample Proportion: P) در هر نمونه تصادفی نسبتی را که بدست می آوریم تا حدودی با نسبت واقعی در کل جامعه متفاوت است. این تفاوتها اشاره به این موضوع دارند که همواره برآوردهای بدست آمده از نمونه، در معرض خطای نمونه گیری (Sampling Error) هستند.

اندازه نمونه (Sample Size:n) و خطای استاندارد (Standard Error:SE) است که ما معمولاً برای بیان کردن انحراف معیار یک توزیع نمونه ای از آن استفاده می کنیم. خطای استاندارد به ما نشان می دهد که برآورد نسبت واقعی ( $\pi$ ) که از روی نسبت بدست آمده از نمونه (P) محاسبه می شود، تاچه اندازه دقیق است. توجه داشته باشید که هر چقدر حجم نمونه (n) را افزایش دهیم، خطای استاندارد کمتر می شود و توزیع نمونه ای باریکتر می شود. این دقیقاً همان چیزی است که انتظار داریم، هر قدر تعداد نمونه افزایش یابد، برآورد پایاتری (Reliable) از  $\pi$  بدست آید. فرض کنید در یک بخش از مطالعه کشوری بررسی سلامت و بیماری، قصد داریم نسبت (proportion) مردانی را که در حال حاضر در یک شهر سیگار مصرف می کنند را تعیین کنیم. به دلایل عملی و مالی، مصاحبه با تک تک افراد آن شهر امکان پذیر نیست. به همین دلیل تصمیم میگیریم که نمونه ای تصادفی به تعداد ۳۰ مرد از این جامعه را انتخاب کنیم. در این نمونه، نسبت مردانی که در حال حاضر سیگار مصرف می کنند،  $23\% = 7 / 30$  است. معمولاً بررسی تمام افراد یک جامعه میسر نمی باشد؛ بنابراین، چاره ای نداریم بجز اینکه نمونه ای n نفری از این جمعیت را انتخاب کرده و مورد ارزیابی قرار دهیم و امیدوار باشیم که این نمونه نمایانگر کل جامعه ما باشد. بدون اندازه گیری تمام افراد جامعه، نمی توانیم درباره نسبت واقعی آن جامعه اطلاعی بدست آوریم. اگرچه نسبتی که از بررسی یک نمونه تصادفی بدست می آوریم، می تواند تا حدودی نسبت کل جامعه را برآورد کند. در مثال بالا، نسبت افراد سیگاری در نمونه تصادفی ۲۳٪ است. این مقدار برآوردی از نسبت واقعی اما نامعلوم مردان سیگاری در جمعیت آن شهر است.

### نسبت بدست آمده از یک نمونه تصادفی تا چه حد دقیق است ؟

در مثال بالا، نسبت افراد سیگاری در شهر، دقیقاً برابر با ۲۳٪ بدست آمده از نمونه گرفته شده نیست و این بدلیل خطای نمونه گیری است. سوالی که در اینجا مطرح می شود، این است که مقدار واقعی چقدر با ۲۳٪ تفاوت دارد؟ برای پاسخ به این سوال، نخست باید تاکید کرد که نمونه ۳۰ نفری انتخاب شده، فقط یکی از تعداد زیادی نمونه ۳۰ تایی است که می توانستیم بگیریم. فرض کنید که ما تمام نمونه های ۳۰ تایی ممکن را یک به یک انتخاب کرده و نسبت P را در آنها بدست آوریم. مقدار این نسبت (P) از نمونه ای به نمونه دیگر متفاوت است. حال اگر این نسبت ها را روی نموداری نمایش دهیم، نمودار زیر بدست می آید :



به این نمودار **توزیع نمونه ای نسبت (Sampling Distribution of P)** می گویند.

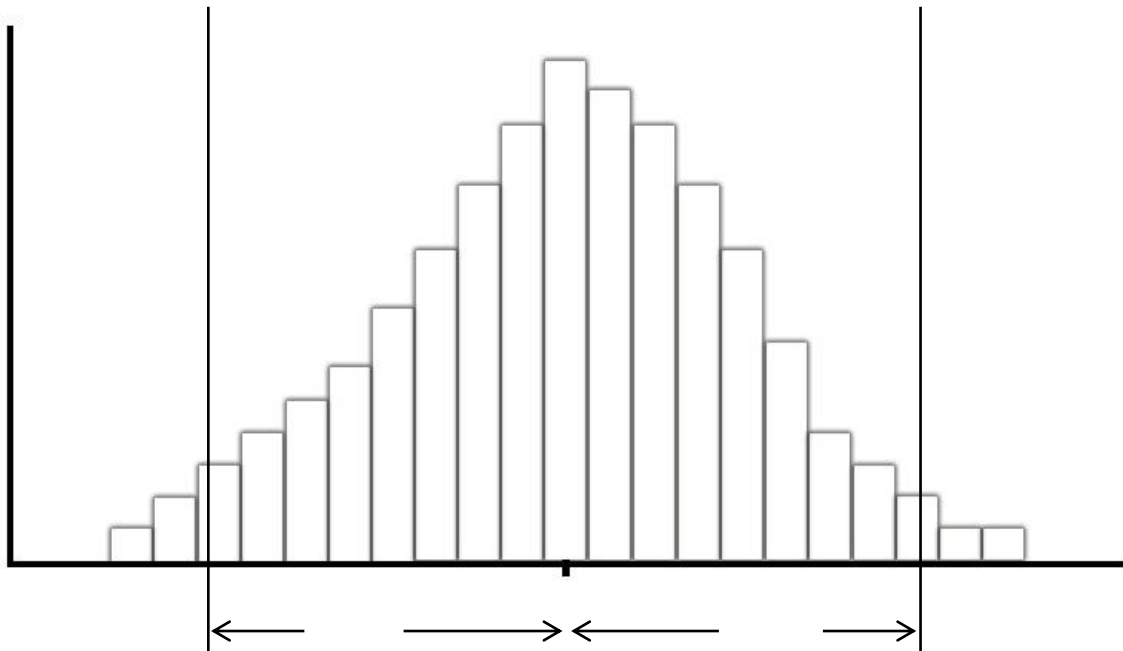
این نمودار نشان می دهد که اگرچه نسبت های مختلف بدست آمده از نمونه های مختلف متفاوت اند و با نسبت واقعی در جمعیت نیز تفاوت دارند، ولی همگی اطراف نسبت واقعی  $\pi$  (جمعیت) پراکنده شده اند. مقدار پراکندگی به ما نشان می دهد که چقدر این نسبت بدست آمده از نمونه، برای برآورد نسبت  $\pi$  واقعی، دقیق (Precise) است. اگر این پراکندگی خیلی زیاد باشد، یعنی خطای نمونه گیری ما زیاد است و نسبت بدست آمده از نمونه (P)، فاصله زیادی با مقدار حقیقی ( $\pi$ ) دارد. اگر این فاصله کم باشد، یعنی خطای نمونه گیری کم است و نسبت بدست آمده از نمونه خیلی به نسبت واقعی نزدیک خواهد بود. می توان پراکندگی یک توزیع را با کمیتی به نام **انحراف معیار** بیان کرد. بطور مشابهی، پراکندگی توزیع نمونه ای P، بدین صورت محاسبه می شود :

$$SE_{(P)} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

در واقع، وقتی که تحقیقی را انجام می دهیم، مقدار  $\pi$  را نداریم ( اگر داشتیم اصلاً نیازی به انجام تحقیق نبود)؛ بنابراین، نمی توانیم از فرمول بالا برای محاسبه  $SE(P)$  استفاده کنیم. بجای مقدار  $\pi$ ، مقدار بدست آمده از نمونه  $(p)$  را جایگزین می کنیم و این برآوردی نسبتاً خوب از خطای معیار است:

$$\text{Estimated } SE_{(P)} = \sqrt{\frac{P(1-P)}{n}}$$

یک بار دیگر به توزیع نمونه ای  $p$  نگاه کنیم؛ می توان با اثبات ریاضی نشان داد که ۹۵٪ از برآوردهای بدست آمده از نمونه، در محدوده دو خطای معیار ( $SE$ ) از مقدار واقعی  $\pi$  قرار می گیرند؛ و تنها ۲/۵٪ آنها در فاصله بیشتر از دوخطای معیار و ۲/۵٪ در فاصله کمتر از دو خطای معیار از مقدار واقعی جای دارند.



$$SE_{(P)} = \sqrt{\frac{0.23 \times 0.77}{30}} = 0.077$$

محدوده اطمینان ۹۵٪ برای نسبت واقعی مردان سیگاری در شهر برابر است با :

$$0.23 \pm (2 \times 0.077) = 0.076 \text{ تا } 0.384$$

در نتیجه بهترین برآوردی که اکنون (با این نمونه) برای نسبت افراد سیگاری در شهر می توان داشت، ۲۳٪ است، اما نسبت واقعی به راحتی می تواند هر عددی بین ۸٪ و ۳۸٪ باشد.

بنابراین محدوده اطمینان، محدوده ای است که می توان گفت به احتمال زیاد نسبت حقیقی جمعیت را دربر دارد، اما چندان هم نمی توان مطمئن بود. محدوده اطمینان بر پایه تکرار مطالعه بیان شده است. در نتیجه، اگر

مطالعه را ۱۰۰ مرتبه تکرار کنیم و برای ۱۰۰ نتیجه بدست آمده، محدوده اطمینان ۹۵٪ حساب کنیم؛ انتظار می رود که ۹۵ تا از آنها مقدار واقعی جمعیت را شامل شوند. پس در حقیقت، محدوده اطمینان بدست آمده از یک مطالعه، ممکن است مقدار واقعی جمعیت را دربر داشته باشد و یا شامل آن نشود. اگر بخواهیم که با اطمینان بیشتری بگوییم که محدوده ما شامل مقدار واقعی است، می توانیم محدوده اطمینان ۹۹٪ را محاسبه کنیم. این محدوده به سادگی با قرار دادن ۲/۵۸ به جای ۲ در فرمول بالا، قابل محاسبه است :

$$P \pm 2.58 \times \sqrt{\frac{p(1-p)}{n}}$$

در مثال ما، محدوده اطمینان ۹۹٪ برای نسبت مردان سیگاری شهر برابر است با ۰/۳۱ تا ۰/۴۲۹، و یا ۳٪ تا ۴۳٪.

### تا چه حد برآوردهای دیگر بدست آمده از نمونه دقیق هستند ؟

روشی که برای اندازه گیری و بیان خطای نمونه گیری بیان شد، فقط برای نسبت بدست آمده از نمونه (P) به کار نمی رود. ما میتوانیم محدوده اطمینان را برای برآوردهای دیگر بدست آمده از نمونه مانند میانگینها (Means)، ضرایب همبستگی (Correlation Coefficients)، ضرایب رگرسیون (Regression Slops)، میزانها (Rates)، خطر نسبی (Relative Risks)، نسبت شانس (Odds Ratios) و اندازه های دیگر نیز محاسبه کنیم. مفهوم پایه ای هر یک از آنها همانند چیزی است که در بالا درباره نسبتها بیان شد. در همه این موارد، محدوده اطمینان به ما نشان می دهد که چقدر مقدار برآورد شده از نمونه دقیق است. فرمول محاسبه و حدود اطمینان هر یک از شاخص های مذکور در ضمیمه شماره سه کتاب آمده است. فرض کنید در یک بخش از مطالعه کشوری بررسی سلامت و بیماری، قصد داریم میانگین قد مردان ساکن شهری را تعیین کنیم. یک نمونه ۳۰ تایی انتخاب کرده و میانگین قد آنها، ۱۶۰ سانتیمتر را بدست می آوریم. در اینجا هم احتمال اینکه میانگین قد کل مردان شهر دقیقاً برابر با ۱۶۰ سانتیمتر باشد، بسیار کم است. اگر تمام نمونه های ممکن ۳۰ تایی مردان شهر را بگیریم و میانگین قد آنها را محاسبه کنیم و نمودار آنرا رسم کنیم، توزیع نمونه ای میانگین (Sampling Distribution of Mean) بدست می آید. نکته قابل توجه این است که توزیع نمونه ای بدست آمده برای میانگین، شکلی شبیه به توزیع نمونه ای نسبت دارد. یعنی، توزیع "زنگوله ای شکل" است و اغلب برآوردهای بدست آمده از نمونه، در قسمت مرکزی و اطراف میانگین واقعی جمعیت تجمع یافته اند. بنابراین، به روش مشابه، محدوده اطمینان ۹۵٪ به صورت زیر قابل محاسبه است :

$$\pm 2 \times SE_{(\text{mean})} \text{ میانگین نمونه}$$

درمثال بالا، خطای معیار ۷ سانتیمتر بدست آمده است. محدوده اطمینان ۹۵٪ برابر است با :

$$160 \text{ cm} \pm 2 \times 7 \text{ cm} = 146 \text{ cm تا } 174 \text{ cm}$$

چگونه این محدوده اطمینان را تفسیر می کنیم؟ به احتمال ۹۵٪، میانگین واقعی قد مردان شهر یک مقداری بین ۱۴۶ تا ۱۷۴ سانتیمتر است.

بطور کلی، وقتی که توزیع نمونه ای یک برآورد خاص تقریباً "زنگوله ای شکل" باشد (به این توزیع، توزیع نرمال می گویند)، می توانیم محدوده اطمینان ۹۵٪ را به صورت زیر محاسبه کنیم:

$$\text{برآورد} \pm 2 \times SE_{\text{برآورد}} = \text{محدوده اطمینان } 95\%$$

و یا بطور دقیق تر :

$$\text{برآورد} \pm 1.96 \times SE_{\text{برآورد}} = \text{محدوده اطمینان } 95\%$$

۱/۹۶ مقدار دقیق آماری است که اغلب به جای آن از ۲ استفاده می شود. برای اینکه محدوده اطمینان ۹۹٪ محاسبه شود، باید به جای ۱/۹۶، از ۲/۵۸ در فرمول استفاده شود.

نکته : محدوده اطمینان ۹۵٪ **خطر قابل انتساب** با استفاده از فرمول زیر می شود. تعریف و توضیحات بیشتر درباره خطر قابل انتساب در بخش پنجم آمده است.

$$\text{خطر قابل انتساب} \pm 1.96 \times SE_{(\text{خطر قابل انتساب})}$$

حد بالا و پایین محدوده اطمینان محاسبه شده برای خطر قابل انتساب به یک فاصله از برآورد حاصل از نمونه قرار دارند، درحالیکه این موضوع درباره **خطر نسبی** صادق نیست. برخلاف **خطر قابل انتساب**، توزیع نمونه ای **خطر نسبی** دیگر متقارن (Symmetric) نیست. زیرا، کمترین مقدار ممکن آن صفر و بیشترین مقدار برای آن بی نهایت است.

یک راه رسیدن به توزیع متقارن، تبدیل لگاریتمی داده ها است. با این تبدیل لگاریتمی، یعنی در مقیاس لگاریتمی، حد بالا و پایین محدوده اطمینان در فاصله یکسانی از برآورد نمونه قرار خواهند داشت ولی با برگشت مقیاس داده ها به حالت اولیه (Antilogarithm)، این فواصل یکسان تا برآورد حاصل از نمونه، دیگر وجود نخواهند داشت. اکنون که با مفهوم حدود اطمینان آشنا شدیم، اشاره به عدم قطعیت و آشنایی با این موضوع نیز اهمیت دارد.

### مقدمه ای بر عدم قطعیت اندازه گیری

هر زمان که اندازه گیری انجام شود، نتیجه هرگز دقیق نخواهد بود. مهم نیست که چقدر دقیق یا علمی، هر اندازه گیری مستعد خطا و عدم قطعیت است. اندازه گیری بهترین تخمین کمی از یک مشاهده است که توسط یک یا چند عامل محدود کننده محدود می شود. این محدودیت ها در نتیجه مشاهده شده تردید ایجاد می کند و مفهوم عدم قطعیت را معرفی می کند.

### عدم قطعیت چیست؟

امروزه نتایج آزمون و کالیبراسیون‌هایی که در آزمایشگاه انجام می‌شود؛ بسیار حائز اهمیت است. به‌منظور حصول اطمینان از اعتبار و کیفیت نتایج به‌دست آمده در آزمایشگاه‌های کالیبراسیون و تأیید صلاحیت شده، می‌توان از مفاهیم و ابزارهایی همچون عدم قطعیت اندازه‌گیری و تضمین کیفیت نتایج استفاده کرد.

### اندازه‌گیری چیست؟

اندازه‌گیری راجع به خاصیت چیزی سخن می‌گوید. ممکن است این ویژگی سنگینی یک جسم، گرما و یا طول آن باشد. اندازه‌گیری، عددی را به آن خاصیت نسبت می‌دهد. اندازه‌گیریها همیشه با ابزاری مانند خط‌کش، فشارسنج، کرومومتر، ترازو و دماسنج و ... انجام می‌گیرند به‌طور طبیعی، نتیجه یک اندازه‌گیری از دو قسمت یعنی عدد و واحد اندازه‌گیری تشکیل شده است.

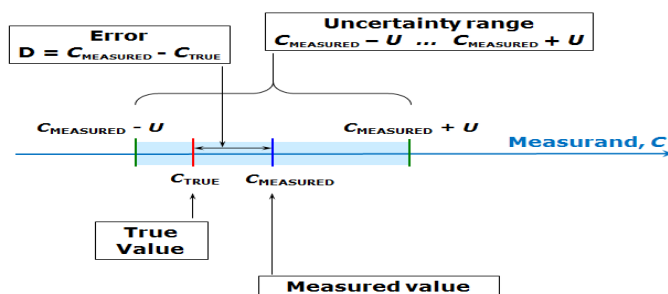
### چه چیزهایی یک اندازه‌گیری محسوب نمی‌شوند؟

بعضی فرآیندها ممکن است یک اندازه‌گیری به نظر آیند ولی در واقع چنین چیزی نیست. به‌عنوان نمونه، مقایسه دو رشته ریسمان جهت تعیین ریسمان بلندتر واقعاً یک اندازه‌گیری محسوب نمی‌شود. طبیعتاً شمارش هم اندازه‌گیری نیست. غالباً یک آزمون که به‌طور طبیعی منجر به جواب بله / خیر یا رد/قبول می‌شود اندازه‌گیری به‌شمار نمی‌آید. با وجود این ممکن است اندازه‌گیری‌ها قسمتی از فرآیندی باشد که به پاسخ یک تست منجر می‌شود.

### عدم قطعیت اندازه‌گیری چیست؟

عدم قطعیت به‌عنوان یک مفهوم عمومی، اصطلاحی است که عدم اطمینان انسان را در مورد برخی اشخاص یا اشیا منعکس می‌نماید، و بازه‌ی باز میان اطمینان کامل و عدم اطمینان محض را در بر می‌گیرد. در این بررسی عدم قطعیت یک عامل همراه با نتیجه اندازه‌گیری است که محدوده‌ی مقادیری را معین می‌کند که نتیجه اندازه‌گیری می‌تواند داشته باشد و مقدار آن نشان‌دهنده سطح اطمینانی است که مقدار واقعی مورد اندازه‌گیری شده در محدوده تعیین شده، قرار می‌گیرد. هدف از محاسبه عدم قطعیت در آزمایشگاه آزمون و کالیبراسیون ایجاد اطمینان از صحت و دقت نتایج آزمون‌کشش مطابق استاندارد ISO/IEC 17025 می‌باشد. نتایج حاصل از این بررسی نشان می‌دهد که عدم قطعیت اندازه‌گیری می‌تواند باعث کاهش خطای دستگاه و کاربران انجام آزمون و افزایش کارایی در آزمایشگاه گردد.

عدم قطعیت اندازه‌گیری اطلاعاتی راجع به کیفیت، کمیت مورد اندازه‌گیری بیان می‌کند در حقیقت، عدم قطعیت شک و گمانی است که درباره نتیجه هر اندازه‌گیری پدید می‌آید. عدم قطعیت نتیجه اندازه‌گیری از دانش ناقص ما از مقدار نتیجه آزمون و عوامل تأثیرگذار بر آن تشکیل می‌شود. به‌طور کلی کلمه عدم قطعیت با فرضیه عمومی تردید مرتبط است. عدم قطعیت اندازه‌گیری بیانگر تردید در خصوص اعتبار اندازه‌گیری نمی‌باشد، بلکه وجود دانش در خصوص عدم قطعیت بیانگر اطمینان بیشتر در مورد اعتبار نتیجه اندازه‌گیری است. همیشه در بسیاری از اندازه‌گیری‌ها حتی دقیق‌ترین آنها جای شک و تردید وجود دارد. در زبان محاوره‌ای این تردید به‌صورت عبارت کم و زیاد بیان می‌شود. مثلاً می‌گوییم این عصا تقریباً ۲ متر است با سانتی‌متر کم و زیاد.



### بیان کردن عدم قطعیت اندازه گیری:

چون تردید در هر اندازه گیری وجود دارد لذا پاسخ به سوالات زیر ضروری به نظر می رسد:

- میزان شک و تردید در اندازه گیری چقدر است؟
- این شک تا چه اندازه نادرست و بی اعتبار است؟

بنابراین برای تعیین عدم قطعیت اندازه گیری به دو عدد یکی فاصله و دیگری سطح اطمینان (یعنی میزان اطمینانی که مقدار واقعی اندازه گیری در داخل فاصله قرار می گیرد) نیازمندیم.

بعنوان مثال ممکن است بگوییم که طول چوبی در سطح اطمینان ۹۵٪ برابر ۲۰ سانتی متر، +۱ می باشد. این نتیجه را می توان به صورت مقابل نوشت: در سطح اطمینان ۹۵٪،  $cm\ 1 + cm\ 20$ . این عبارت بدین معنی است که با اطمینان ۹۵٪ طول چوب بین ۱۹ و ۲۱ سانتی متر خواهد بود.

### خطا در مقابل عدم قطعیت:

در اندازه گیری کمیت های فیزیکی خطا اجتناب ناپذیر است و یا بهتر است گفته شود خطا جزء جدا نشدنی اندازه گیری به شمار می آید. از این رو در اندازه گیری کمیتها به اندازه و نیز منابع خطا باید همواره توجه شود. به عبارت دیگر مطالعه و شناخت خطا ما را در یافتن روش های کاهش آن برای دستیابی به نتایج بهینه یاری می دهد. به این ترتیب از میزان خطا تا به اندازه ای می توان کاست که از اثر آن بر نتایج به دست آمده از اندازه گیری، بتوان چشم پوشی کرد.

نکته مهم آن است که اصطلاحات خطا و عدم قطعیت نباید با هم اشتباه شوند. خطا اختلاف بین مقدار اندازه گیری شده و مقدار واقعی کمیت مورد اندازه گیری است. در صورتی که عدم قطعیت تعیین شک و تردید راجع به نتیجه اندازه گیری است. با این حال سعی مان بر این است که تا حد امکان در هر زمانی خطاهای شناخته شده را تصحیح کنیم. در نتیجه خطا در آزمون یعنی تفاوت بین مقدار اندازه گیری شده با مقدار واقعی در صورتی که عدم قطعیت یعنی کمی نمودن تردید درباره اندازه گیری.

به عنوان نمونه این تصحیحات براساس گواهینامه های کالیبراسیون قابل انجام است. اما هر خطایی که معلوم نباشد منبع عدم قطعیت خواهد بود. به طور کلی سه نوع خطا در سیستم به شرح ذیل تعریف شده است:

### • خطای عمده:

این خطا عمدتاً مربوط به خطای انسانی است و در فرایند اندازه گیری این خطا اجتناب ناپذیر است.

### • خطای سیستماتیک:

میانگین نتایج حاصل از انجام تعداد نامحدودی اندازه گیری یک اندازه ده در شرایط تکرارپذیر منهی مقدار واقعی آن اندازه ده.

خطای سیستماتیک از روشهای ناقص، تجهیزات کالیبره نشده یا برخی از متغیرها که نتیجه ی اندازه گیری را هر بار با مقدار ثابتی تغییر میدهند، ناشی می شود. خطای سیستماتیک میتواند شناخته یا ناشناخته باشد. یک خطای سیستماتیک معلوم یا معین، بایاس نامیده میشود. دو مثال بارز از خطای سیستماتیک، خطای دستگاه اندازه گیری و خطای روش اندازه گیری است.

### خطای تصادفی:

نتیجه ی اندازه گیری منهای میانگین نتایجی که از انجام تعداد نامحدودی اندازه گیری یک اندازه ده در شرایط تکرارپذیر به دست می آید. خطای تصادفی از شرایط کنترل ناپذیر آزمایش و منابع ناشناخته ناشی میشود. این خطا در هر بار اندازه گیری، نتیجه ی اندازه گیری را با مقداری متفاوت تغییر میدهد. دیده شده است که نتایج اندازه گیری حتی پس از در نظر گرفتن تمام خطاهای سیستماتیک از یک خواننده به خواننده ی دیگر تغییر میکند.

### چرا عدم قطعیت اندازه گیری مهم است:

به منظور درک بهتر نتایج اندازه گیری و بدست آوردن اندازه گیری های با کیفیت بالا، ممکن است علاقه مند به محاسبه عدم قطعیت اندازه گیری باشیم. با وجود این دلایل خاص دیگری برای محاسبه و گزارش عدم قطعیت وجود دارد. اندازه گیری ها ممکن است قسمتی از یکی از موارد زیر باشند:

**کالیبراسیون:** عدم قطعیت اندازه گیری باید در گواهینامه کالیبراسیون گزارش شود.

**آزمون:** برای تعیین قبولی یا رد نتیجه آزمون می بایست عدم قطعیت اندازه گیری محاسبه و گزارش شود.

**ترانس:** ممکن است قبل از اینکه تصمیم گرفته شود که آیا ترانس مورد نظر بدست آمده نیازمند به دانستن عدم قطعیت باشیم.

### منابع خطا و عدم قطعیت

**خطاها و عدم قطعیت ها از کجا ناشی می شوند:** موارد بسیاری می تواند منجر به اشکال در یک اندازه گیری شود. این عیوب در اندازه گیری ممکن است قابل رویت یا غیرقابل دیدن باشند. از آنجایی که اندازه گیری ها واقعی هرگز تحت شرایط کاملی بدست نمی آیند لذا منابع ایجاد خطاها و عدم قطعیت ها اندازه گیری از موارد زیر ناشی می شوند:

**منابع خطا و عدم قطعیت، ابزار اندازه گیر:** ابزار اندازه گیری ممکن است از خطاهایی مانند اریبی (تمایل به یک طرف)، تغییرات ناشی از طول عمر، پوسیدگی، خوردگی، ساییش یا انواع دیگر انحراف، ناخوانا بودن، اغتشاش و سر و صدا (در وسایل الکتریکی) و بسیاری از مشکلات دیگر دچار خرابی شوند.

**منابع خطا و عدم قطعیت، جسم مورد اندازه گیری:** ممکن است جسم، پایدار و ثابت نباشد. تصور کنید می خواهید اندازه یک مکعب یخی را در هوای گرم اندازه گیری کنید.

**منابع خطا و عدم قطعیت، فرآیند اندازه گیری:** ممکن است انجام خود اندازه گیری با مشکلاتی همراه باشد. به عنوان نمونه مشکلات خاص مربوط به اندازه گیری ابعاد یک جسم خیلی کوچک را در نظر بگیرید.

**منابع خطا و عدم قطعیت، عدم قطعیت های ورودی:** معمولا عدم قطعیت مربوط به کالیبراسیون ابزار اندازه گیری با عدم قطعیت اندازه گیری های بدست آمده ترکیب خواهد شد. قابل ذکر است که عدم قطعیت ناشی از کالیبره نکردن ابزار بسیار مضر می باشد.

**منابع خطا و عدم قطعیت، مهارت اپراتور:** بعضی از اندازه گیری ها به مهارت و قضاوت اپراتور بستگی دارد. ممکن است در کار حساس تنظیم کردن و اندازه گیری، دقت شخصی نسبت به دیگری بیشتر بوده یا در خواندن اندازه ها با چشم، جزئیات را بهتر گزارش دهیم. مثلا استفاده از ابزاری مانند کرنومتر به زمان عکس العمل اپراتور



بستگی دارد. قابل ذکر است که این اشتباهات بزرگ و فاحش که از طرف اپراتور ناشی می شود ( خطای اپراتور) جزء منابع ایجاد عدم قطعیت اندازه گیری محسوب نخواهد شد.

**منابع خطا و عدم قطعیت، روش نمونه گیری:** اندازه گیری هایی که انجام می دهیم می بایست نماینده شایسته ای از فرآیند مورد ارزیابی باشد. به عنوان مثال اگر قصد اندازه گیری درجه حرارت محیط کار را داریم نباید از دماسنج نصب شده روی دیوار که در نزدیکی تهویه مطبوع قرار دارد استفاده کنیم. همچنین اگر برای اندازه گیری قصد انتخاب نمونه هایی از خط تولید را داریم نباید نمونه های انتخابی از ۱۰ محصولی که در صبح روز اول هفته تولید شده انتخاب گردند.

**منابع خطا و عدم قطعیت، محیط:** درجه حرارت، فشار هوا، رطوبت و شرایط بسیار دیگری می تواند بر ابزار اندازه گیری یا جسم مورد اندازه گیری تاثیر بگذارد.

**تذکرات کلی:** در مواردی که اندازه و اثر خطا معلوم باشد ( مثلا از گواهینامه کالیبراسیون ) می بایست تصحیحی برابر و با علامت مخالف بر روی نتیجه اندازه گیری اعمال کرد. به عنوان مثال اگر خطای دستگاه در گواهینامه کالیبراسیون ۰,۰۱ و نتیجه ۲ باشد با اعمال تصحیح یاد شده مقدار اندازه گیری شده برابر با ۱,۹۹ خواهد شد. اما در حالت کلی ، عدم قطعیت های ناشی از منابع یاد شده و از سایر منابع دیگر هر یک به صورت انفرادی سهمی از عدم قطعیت کلی را به خود اختصاص خواهند داد.

### فرآیند تخمین عدم قطعیت

تخمین عدم قطعیت در اصل ساده می باشد. پاراگرافهایی که در ادامه می آیند کارهای لازم جهت رسیدن به تخمین عدم قطعیت مرتبط با نتیجه اندازه گیری را به طور خلاصه توضیح می دهند. بخش های بعدی نیز راهنمایی های اضافی و کاربردی در شرایط مختلف را توضیح می دهند، این راهنمایی ها عمدتاً شامل چگونگی استفاده از داده های بدست آمده درون سازمانی و مطالعات صحنه گذاری متد مشترک، داده های QC داده های بدست آمده از آزمون مهارت (PT) و استفاده از اصول انتشار عدم قطعیت نرمال می باشند. مراحل انجام شده به ترتیب زیر می باشند:

### گام ۱: تعیین اندازه ده

یک توضیح روشن از آنچه که قرار است اندازه گیری شود را بنویسید. این توضیح میتواند شامل ارتباط بین کمیت اندازه ده و آن دسته از کمیت های ورودی (برای مثال، کمیت های اندازه گیری شونده، ثوابت، مقادیر استاندارد کالیبراسیون و غیره) باشند که اندازه ده به آنها وابسته است. هر جا که امکان داشته باشد، تصحیح های لازم برای اثرات سیستماتیک شناخته شده را هم لحاظ کنید. اطلاعات مربوط به مشخصات میبایست در روش اجرایی استاندارد (SOP) و یا دیگر تعریفهای متد مرتبط ارائه شده باشند.

### گام ۲: تعیین منابع عدم قطعیت

منابع احتمالی عدم قطعیت را فهرست کنید. این کار شناسایی منابعی است که میتوانند در عدم قطعیت روی پارامترهای موجود در رابطه ی تعیین شده گام ۱ مشارکت داشته باشند. هر چند منابع دیگری هم ممکن است وجود داشته باشند و در صورت نیاز حتی میبایست منابع احتمالی عدم قطعیت ناشی از فرضهای شیمیایی را نیز لحاظ کرد.

### گام ۳: تعیین کمی مولفه های عدم قطعیت

اندازه مولفه عدم قطعیت مرتبط با هر منبع احتمالی عدم قطعیت شناسایی شده را تخمین بزنید. اغلب، امکان تخمین یا تعیین یک سهم تک برای عدم قطعیت مرتبط با تعداد زیادی منابع جدا از هم، با استفاده از داده های مطالعات صحنه گذاری و داده های QC و غیره وجود دارد. استفاده از یک چنین داده هایی، کار لازم برای تخمین عدم قطعیت را به طور قابل ملاحظه ای کاهش میدهد، و چون در انجام این کار از داده های آزمایشگاهی واقعی استفاده می شود، لذا می تواند به تخمین های قابل اعتماد از عدم قطعیت نیز بینجامد. همچنین مهم است این نکته را در نظر بگیریم که آیا داده های موجود، توضیح کافی در مورد تمام منابع عدم قطعیت را ارائه می دهند یا نه، نکته دیگر آن که بعضی وقتها برای کسب اطمینان از این که آیا تمام منابع عدم قطعیت لحاظ شده اند یا نه به آزمایشها و مطالعات اضافی دیگری نیاز خواهیم داشت.

### گام ۴: محاسبه عدم قطعیت مرکب

اطلاعات بدست آمده از گام ۳، چه مرتبط با منابع تک یا مرتبط با اثرات ترکیبی چندین منبع باشند، در هر حالت شامل تعدادی از مولفه های با مقدار معین و سهم در عدم قطعیت کل خواهند بود. مقدار هر یک از سهم ها می بایست به شکل انحراف استاندارد بیان و سپس بر اساس قوانین خاصی با یکدیگر ترکیب شده تا منجر به عدم قطعیت استاندارد مرکب شوند. برای رسیدن به عدم قطعیت بسط یافته، فاکتور پوشش مناسب نیز می بایست اعمال شود.

فرمول محاسبه عدم قطعیت و حدود اطمینان

$$95\% \text{ confidence level} = \mu \pm 2 * u$$

در شکل زیر مثالی از نحوه محاسبه عدم قطعیت و رابطه بین خطای تصادفی، خطای منظم و عدم قطعیت برای ۴ حالت مختلف یک برآورد با نمونه شش نفری نمایش داده شد. در حالت چهارم که خطای معیار برآورد حداقل ممکن است و برآورد روایی و پایایی قابل قبولی دارد خطاهای دیگر و همچنین عدم قطعیت در حداقل ممکن است.

	A	B	C
5			
6	Readings (x <sub>i</sub> )	(x <sub>i</sub> - μ)	(x <sub>i</sub> - μ) <sup>2</sup>
7	15.33	0.04	0.0016
8	15.21	-0.08	0.0064
9	15.31	0.02	0.0004
10	15.25	-0.04	0.0016
11	15.35	0.06	0.0036
12			
13	Number of Reading (n)	5	
18	Uncertainty is calculated using the formula given below		
20	Uncertainty (u) = √ [Σ (x <sub>i</sub> - μ) <sup>2</sup> / (n * (n-1))]		
22	Uncertainty Formula	=SQRT(SUM(C7:C11)/(B13*(B13-1)))	
23	Uncertainty (u)	0.03	
24			

• Uncertainty = 0.03 seconds

68% of values fall within 1 standard deviation of the mean (-1s <= X <= 1s)

So Timing at 68% confidence level = μ ± 1 \* u

- Measurement at 68% confidence level = (15.29 ± 1 \* 0.03) seconds
- Measurement at 68% confidence level = **(15.29 ± 0.03) seconds**

Situation	Random effects	Systematic effects	Uncertainty
1.	Strong	Strong	High
2.	Strong	Weak (or absent)	Medium
3.	Weak	Strong	Medium
4.	Weak	Weak (or absent)	Low

## یادآوری آزمون های مهم و رایج:

فرایند چگونگی استفاده از این آزمون ها در ضمیمه شماره دو کتاب آمده است.

### ۱- آزمون اختلاف میانگین یک صفت در جامعه با یک عدد مشخص ( $\mu_0$ ) وقتی که انحراف معیار یا واریانس $\sigma^2$ آن صفت در جامعه معلوم باشد.

نمونه ها باید تصادفی و مستقل باشند. توزیع متغیر در جامعه نرمال باشد و یا حجم نمونه به اندازه کافی بزرگ باشد (استفاده از قضیه حد مرکزی). قبل از جمع آوری و مشاهده داده ها باید درباره انجام آزمون یک دامنه یا دو دامنه تصمیم گیری شود. چنانچه از قبل تصمیمی گرفته نشود باید از آزمون دو دامنه استفاده نمود.

**الف) آزمون دو دامنه** (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

**ب) آزمون یک دامنه** (وقتی جهت اختلاف یعنی بزرگتر یا کوچکتر بودن  $\mu$  از  $\mu_0$  مهم باشد)

ملاک آزمون	$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
فرضیه $H_0$	$\mu = \mu_0$
فرضیه $H_1$	$\mu > \mu_0$ یا $\mu < \mu_0$
شرایط رد فرضیه $H_0$	$Z > Z_{1-\alpha}$ یا $Z < Z_{\alpha}$

ملاک آزمون	$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
فرضیه $H_0$	$\mu = \mu_0$
فرضیه $H_1$	$\mu \neq \mu_0$
شرایط رد فرضیه $H_0$	$ Z  \geq Z_{1-\frac{\alpha}{2}}$

### ۲- آزمون اختلاف میانگین یک صفت در جامعه با یک عدد مشخص وقتی که انحراف معیار آن صفت در جامعه معلوم نباشد.

نمونه ها باید تصادفی و مستقل باشند. توزیع متغیر در جامعه نرمال باشد و یا حجم نمونه به اندازه کافی بزرگ باشد ( برقراری قضیه حد مرکزی). بجای انحراف معیار صفت در جامعه، از انحراف معیار نمونه استفاده می شود. بجای توزیع Z از توزیع t استفاده می شود.

**الف) آزمون دو دامنه** (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

**ب) آزمون یک دامنه** (وقتی جهت اختلاف یعنی بزرگتر یا کوچکتر بودن  $\mu$  از  $\mu_0$  مهم باشد)

ملاک آزمون	$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
فرضیه $H_0$	$\mu = \mu_0$
فرضیه $H_1$	$\mu \geq \mu_0$ یا $\mu < \mu_0$
شرایط رد فرضیه $H_0$	$t > t_{1-\alpha}$ یا $t < t_{\alpha}$

ملاک آزمون	$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	درجه آزادی = n-1
فرضیه $H_0$	$\mu = \mu_0$	
فرضیه $H_1$	$\mu \neq \mu_0$	
شرایط رد فرضیه $H_0$	$ t  \geq t_{1-\frac{\alpha}{2}}$	

### ۳-آزمون اختلاف نسبت یک صفت در جامعه ( $\bar{P}$ ) با یک نسبت مشخص ( $P_0$ )

نمونه ها باید تصادفی و مستقل باشند. حجم نمونه باید زیاد باشد. بطوریکه np و npq بزرگتر از ۵ باشد.

ب) آزمون یک دامنه (وقتی جهت اختلاف یعنی بزرگتر یا کوچکتر بودن  $\mu$  از  $\mu_0$  مهم باشد)

ملاک آزمون	$Z = \frac{\frac{x}{n} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$
فرضیه $H_0$	$\frac{x}{n} = P_0$
فرضیه $H_1$	$\frac{x}{n} < P_0$ یا $\frac{x}{n} > P_0$
شرایط رد فرضیه $H_0$	$Z < Z_{\alpha}$ یا $Z > Z_{1-\alpha}$

الف) آزمون دو دامنه (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

ملاک آزمون	$Z = \frac{\frac{x}{n} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$
فرضیه $H_0$	$\frac{x}{n} = P_0$
فرضیه $H_1$	$\frac{x}{n} \neq P_0$
شرایط رد فرضیه $H_0$	$ Z  \geq Z_{1-\frac{\alpha}{2}}$

### ۴-آزمون مساوی بودن واریانس دو جامعه

فرض کنید  $S_1$  و  $S_2$  برآورد واریانس صفت در دو نمونه مستقل از دو جامعه نرمال با واریانس های  $\sigma_1$  و  $\sigma_2$  باشد.

آزمون دو دامنه (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

ملاک آزمون	$F = \frac{S_1^2}{S_2^2}$
فرضیه $H_0$	$\sigma_1^2 = \sigma_2^2$
فرضیه $H_1$	$\sigma_1^2 \neq \sigma_2^2$
شرایط رد فرضیه $H_0$	$F > F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ یا $F < F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ درجه آزادی = $n_1 - 1, n_2 - 1$

۵- آزمون اختلاف میانگین یک صفت در دو جامعه نرمال ( $\mu_1$  و  $\mu_2$ ) وقتی واریانس دو جامعه مشخص بوده ( $\sigma_1^2$  و  $\sigma_2^2$ ) و میانگین دو نمونه مستقل از این دو جامعه برابر  $\bar{x}_1 - \bar{x}_2$  باشد.

آزمون دو دامنه (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	ملاک آزمون
$\mu_1 = \mu_2$	فرضیه $H_0$
$\mu_1 \neq \mu_2$	فرضیه $H_1$
$ z  > z_{1-\frac{\alpha}{2}}$	شرایط رد فرضیه $H_0$

۶- آزمون اختلاف میانگین یک صفت در دو جامعه نرمال ( $\mu_1$  و  $\mu_2$ ) وقتی واریانس دو جامعه مشخص نبوده و میانگین دو نمونه مستقل از این دو جامعه برابر  $\bar{x}_1$  و  $\bar{x}_2$  و واریانس این دو نمونه  $s_1^2$  و  $s_2^2$  باشد.

آزمون دو دامنه (وقتی فقط اختلاف مورد نظر باشد و جهت آن یعنی بزرگتر یا کوچکتر بود مهم نباشد):

$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$	$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	ملاک آزمون
$s_p^2 =$ برآورد ترکیبی واریانس در دو نمونه		
$\mu_1 = \mu_2$		فرضیه $H_0$
$\mu_1 \neq \mu_2$		فرضیه $H_1$
$ t  > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$	درجه آزادی = $n_1 + n_2 - 2$	شرایط رد فرضیه $H_0$

۷- آزمون اختلاف میانگین دو جامعه وقتی اطلاعات نتیجه مشاهدات دو تایی باشد. میانگین صفت در دو جامعه مستقل نباشد.

ب) اگر انحراف معیار اختلاف میانگینها ( $\bar{d}$ ) معلوم نباشد:  
(در آزمون نودامنه)

$t = \frac{\bar{d} - 0}{\frac{s}{\sqrt{n}}}$ $s = \frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}$	ملاک
$\bar{d} = 0$	فرضیه $H_0$
$\bar{d} \neq 0$	فرضیه $H_1$
$ t  > t_{1-\frac{\alpha}{2}}$	شرایط رد فرضیه $H_0$

الف) اگر انحراف معیار اختلاف میانگینها ( $\bar{d}$ ) معلوم باشد:  
(در آزمون نودامنه)

$z = \frac{\bar{d} - 0}{\frac{\sigma}{\sqrt{n}}}$	ملاک
$\bar{d} = 0$	فرضیه $H_0$
$\bar{d} \neq 0$	فرضیه $H_1$
$z > z_{1-\frac{\alpha}{2}}$	شرایط رد فرضیه $H_0$

۸- آزمون اختلاف نسبت در دو جامعه ( $P_1$  و  $P_2$ ) وقتی دو نمونه مستقل از آنها با نسبتهای  $\frac{x_1}{n_1}$  و  $\frac{x_2}{n_2}$  انتخاب شده باشند.

$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $P = \hat{P} = \frac{x_1 + x_2}{n_1 + n_2}$ <p><math>\hat{P}</math> = برآورد نسبت در دو جامعه</p>	ملاک
$P_1 = P_2$	فرضیه $H_0$
$P_1 \neq P_2$	فرضیه $H_1$
$z > z_{1-\frac{\alpha}{2}}$	شرایط رد فرضیه $H_0$

۹-آزمون تطابق نمونه با توزیع نظری با استفاده از ملاک کای دو

$\chi^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$	ملاک
m = تعداد پارامترهای مستقلی که توسط نمونه بر آورد شده n <sub>i</sub> = فراوانی مشاهده شده گروه i ام E <sub>i</sub> = فراوانی منتظره گروه i ام	
تطابق نمونه با توزیع نظری	فرضیه H <sub>0</sub>
عدم تطابق نمونه با توزیع نظری	فرضیه H <sub>1</sub>
$\chi^2 > \chi^2_{1-\alpha}(k-1-m)$ درجه آزادی = k-1-m	شرایط رد فرضیه H <sub>0</sub>

۱۰-آزمون روند:  $\chi^2_{-trend}$

در مواردی ممکن است جدولی وجود داشته باشد، که دارای دو سطر اما تعداد بیشتری ستون باشد. مثال زیر پژوهشی است که بر روی زنان باردار انجام شده است. هر زن باردار بر اساس رژیم غذایی، در یکی از گروههای تغذیه ای بد، متوسط و خوب قرار گرفته است که ستونهای جدول را تشکیل می دهد. در سطرهای جدول، ابتلا یا عدم ابتلا به پراکلامپسی مشخص شده است. هر خانه جدول، نشان دهنده نسبتی از افراد هر گروه تغذیه ای است که مبتلا به پراکلامپسی بوده اند:

مجموع	رژیم غذایی حین بارداری			پراکلامپسی
	خوب	متوسط	بد	
۳۰	۴ (%۷)	۱۱ (%۱۶)	۱۵ (%۲۹)	پراکلامپسی
۱۵۰	۵۷	۵۶	۳۷	بلی
۱۸۰	۶۱	۶۷	۵۲	خیر
	۳	۲	۱	ارزش عددی برای آزمون روند تغییرات (Score for Trend-Test)

در این جدول ۳×۲، هدف ارزیابی ارتباط بین رژیم غذایی و ابتلا به پراکلامپسی در زنان باردار است. از آنجا که پراکلامپسی یک پیامد است، هدف ما مقایسه نسبت افراد پراکلامپسی در ۳ رژیم غذایی مختلف می باشد. اگر تست معمول  $\chi^2$  را انجام دهیم، مقدار ۱۰/۰۵ حاصل می شود که با توجه به درجه آزادی ۲ مقدار  $p < ۰/۰۰۶$  حاصل می شود. بنابراین شواهد قوی برای ارتباط بین دو متغیر وجود دارد.

اما زمانی که متغیر ستونی ما، یک متغیر رتبه ای است، مانند مثال فوق، آزمون حساستری جهت تشخیص ارتباط (خطی) بین دو متغیر با استفاده از آزمون  $\chi^2$ -trend تست مورد نیاز است. بنابراین، استفاده از تست معمول  $\chi^2$  چندان مناسب نیست. برای محاسبه  $\chi^2$ -trend زمانیکه متغیر ما رتبه ای باشد، ابتدا به هر یک از ستونهای جدول، یک ارزش عددی تعلق می گیرد (Score for trend test). سپس، میانگین نمره  $\bar{x}_1$  و  $\bar{x}_2$  در افراد مبتلا به پراکلامپسی و افراد غیر مبتلا به صورت زیر محاسبه می گردد:

$$\bar{x}_1 = 1/63 = [(1 \times 15) + (2 \times 11) + (3 \times 4)]$$

$$\bar{x}_2 = 2/13 = \text{میانگین نمره در افراد غیرمبتلا به پراکلامپسی}$$

اگر هیچ تفاوتی بین ۳ نسبت وجود نداشته باشد، میانگین نمره  $\bar{x}_1$  و  $\bar{x}_2$  مساوی خواهد شد. درحالیکه، اگر یک روند صعودی یا نزولی در نسبتها وجود داشته باشد، میانگین نمره در افراد پراکلامپسی بیشتر یا کمتر خواهد شد.

بر این اساس، تست  $\chi^2$ -trend برای متغیرهای رتبه ای، از طریق مقایسه میانگین نمره ها (میانگین وزن داده شده)، بصورت زیر محاسبه میگردد:

$$\chi^2_{\text{-trend}} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2(1/n_1 + 1/n_2)}$$

$n_1$  و  $n_2$  مجموع تعداد بیماران مبتلا به پراکلامپسی و افراد سالم هستند.  $s$  انحراف معیار رژیم غذایی وزن داده شده برای کل افراد (پراکلامپسی+سالم) است. در مثال ما،  $s = 0.793$  است و  $\chi^2$ -trend بصورت زیر محاسبه می گردد:

$$\chi^2_{\text{-trend}} = \frac{(1.63 - 2.13)^2}{0.793^2(1/30 + 1/150)} = 9.94$$

این آزمون، بسیار به آزمون Z برای مقایسه میانگین دو گروه شباهت دارد.

$\chi^2$ -trend دارای توزیع  $\chi^2$  با درجه آزادی ۱ می باشد. در این مثال، مقدار محاسبه شده از مقدار جدول (۳/۸۴) بزرگتر است، بنابراین شواهد قوی به نفع وجود ارتباط خطی (یک روند خطی) بین متغیر رتبه ای رژیم غذایی و متغیر اسمی پراکلامپسی وجود دارد.



## راهنمای انجام آزمون های آماری رایج با نرم افزار Stata :

در این قسمت پس از ارائه روش ارزیابی و انجام آزمون های بررسی توزیع داده ها و برقراری فرض نرمالیتی، فرمان های لازم برای انجام آزمون های معروف آماری معرفی می شوند.

### آزمون های بررسی نرمال بودن داده ها و بررسی چولگی و کشیدگی داده ها

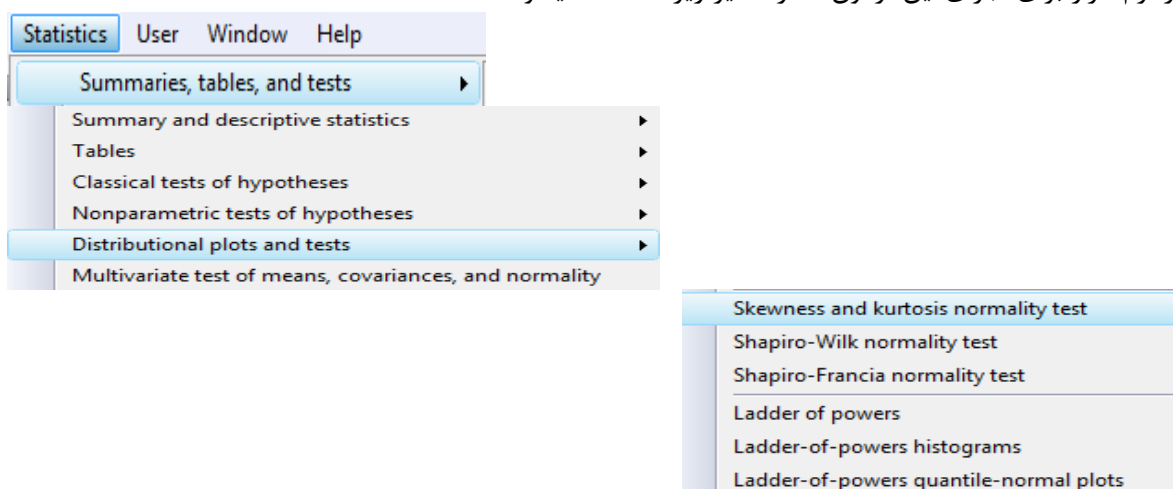
برای آزمون نرمال بودن یک جامعه، روش های مختلفی وجود دارد که در زیر به سه مورد اشاره می شود:

۱- آزمون چولگی-کشیدگی: نام متغیر (ها) `sktest`

۲- آزمون شاپیرو-ویلکز: نام متغیر(ها) `swilk`

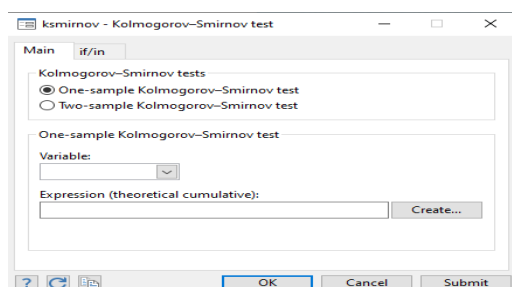
۳- آزمون شاپیرو-فرانسیا: نام متغیر(ها) `sfrancia`

در نرم افزار برای اجرای این آزمون ها از مسیر زیر استفاده میشود.



فرض صفر در ارزیابی توزیع داده ها این است که داده ها توزیع نرمال دارند. یا عبارتی بین توزیع متغیر مورد بررسی و توزیع نرمال اختلاف معنی داری نیست. چنانچه پس از ارزیابی مقدار پی آزمون کمتر از 0.05 باشد فرض صفر رد می شود و داده ها توزیع نرمال ندارند. بنابراین یکی از شرط های استفاده از آزمون های پارامتریک برقرار نیست. وقتی حجم نمونه بیشتر از ۳۰ باشد می توان با استفاده از قضیه حد مرکزی از روشهای پارامتریک استفاده نمود. می توان با دستور `db sktest` جعبه گفتگو را باز و با انتخاب نام متغیر ارزیابی مربوطه را انجام داد.

روش دیگر ارزیابی توزیع یک متغیر از منوی `statistics` در قسمت `Nonparametric analysis` می توان با انتخاب `Tests of hypotheses` و انتخاب `Kolmogorov-Smirnov test` مسیر ارزیابی توزیع نرمال یک متغیر را آزمون نمود.



## آزمون های مختلف تی (t-test)

### آزمون فرض برابری میانگین با مقدار ثابت:

برای آزمون این فرض که میانگین جامعه برابر یک مقدار ثابت هست یا نه، از دستور تابع `ttest` به صورت زیر استفاده می شود:

نام متغیر = مقدار ثابت `ttest`

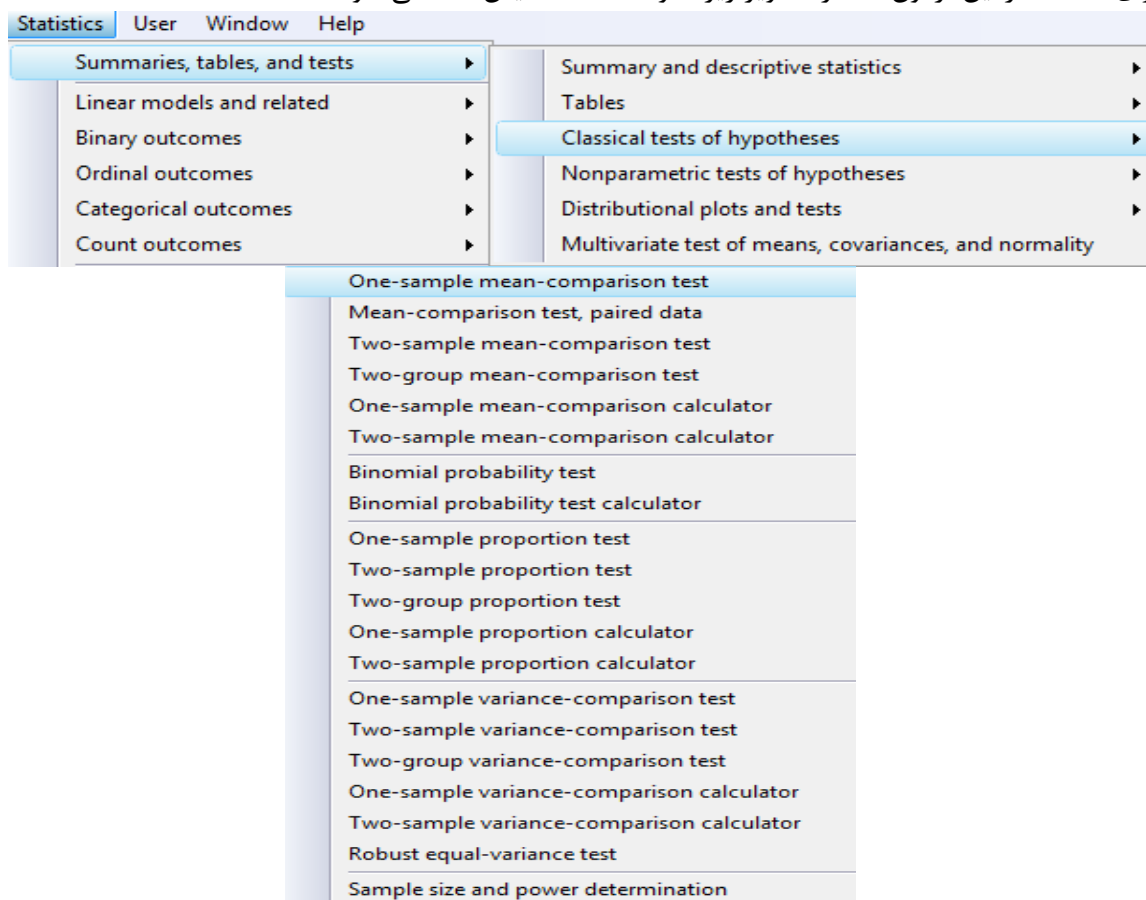
Stata در خروجی این فرمان علاوه بر سایر اطلاعات، بازه ی اطمینان ۹۵٪ برای میانگین جامعه را نیز ارائه می کند. به منظور تغییر سطح اطمینان این بازه می توان گزینه ی " (سطح اطمینان) `level` " را به کار برد:

`ttest` نام متغیر (سطح اطمینان) `level` , نام متغیر

**نکته:** برای ساختن بازه ی اطمینان برای میانگین متغیر از فرمان `ci` نیز می توان استفاده کرد، Stata به طور پیش فرض فاصله ی اطمینانی با سطح ۹۵٪ می سازد که با کمک گزینه ی ( ) `level` می توان سطح اطمینان بازه را تغییر داد. فرمان زیر بازه ی اطمینانی با سطح ۹۰٪ می سازد:

`ci` نام متغیر (90) `level` , نام متغیر

برای اجرای فرمان مذکور و یا استفاده از سایر تست های آماری می توان از مسیر منوی برنامه نیز استفاده کرد. برای استفاده از این آزمون ها، در تصاویر زیر نحوه استفاده نمایش داده می شود.



### آزمون فرض برابری میانگین‌ها برای نمونه‌ی زوج شده یا به هم وابسته و غیر مستقل:

در مواردی لازم است میانگین دو متغیر که هر دو برای یک مشاهده اندازه‌گیری شده‌اند، با هم مقایسه شوند. فرمان زیر برای انجام این آزمون به کار می‌رود:

نام متغیر اول = نام متغیر دوم ttest

### آزمون فرض برابری میانگین‌های دو جامعه مستقل:

این آزمون برای مقایسه‌ی میانگین دو جامعه‌ی مستقل از هم به کار می‌رود. برای انجام این آزمون در Stata از فرمان ttest به صورت زیر استفاده می‌شود:

(نام متغیر رسته‌ای) by, نام متغیر ttest

این فرمان به طور پیش‌فرض واریانس دو جامعه را برابر فرض می‌کند. در صورت برابر نبودن واریانس‌ها لازم است از گزینه‌ی unequal استفاده شود:

unequal (نام متغیر رسته‌ای) by, نام متغیر ttest

### آزمون فرض برابری واریانس با مقدار ثابت:

برای آزمون این فرض که واریانس جامعه برابر یک مقدار ثابت هست یا نه، از تابع sdtest به صورت زیر استفاده می‌شود:

نام متغیر = مقدار ثابت sdtest

### آزمون فرض برابری واریانس‌ها برای نمونه‌ی زوج شده:

در مواردی لازم است واریانس دو متغیر که هر دو برای یک مشاهده اندازه‌گیری شده‌اند، با هم مقایسه شوند. فرمان زیر برای انجام این آزمون به کار می‌رود:

نام متغیر اول = نام متغیر دوم sdtest

### آزمون فرض برابری واریانس‌های دو جامعه مستقل:

این آزمون برای مقایسه‌ی واریانس دو جامعه‌ی مستقل از هم به کار می‌رود. برای انجام این آزمون در Stata از فرمان sdtest به صورت زیر استفاده می‌شود:

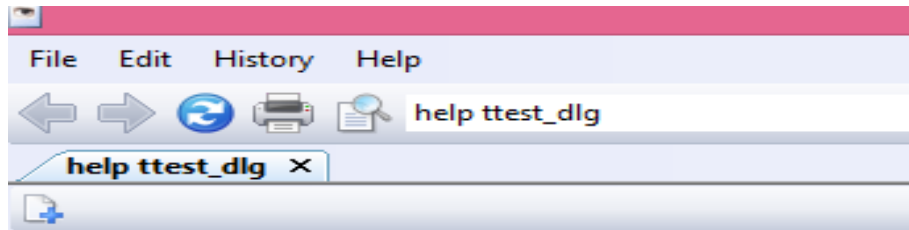
(نام متغیر رسته‌ای) by, نام متغیر sdtest

کار عملی با نرم افزار و انجام آزمون‌های آماری: فایل داده‌های پیوست بنام part five را باز نمایید. ابتدا متغیرهای موجود در آن را شناسایی نمایید. برای این کار از فرمان des استفاده می‌شود. در خروجی دستور ۳۴ متغیر و ۶۰ مشاهده وجود دارد.

### اولین آزمون: مقایسه میانگین با عدد ثابت: One Sample Mean – Comparison Test

برای انجام آزمون فرضیه " میانگین قد جامعه مورد بررسی برابر با ۱۷۰ سانتی متر است " از مسیر زیر استفاده می‌شود. می‌توان با فرمان db ttest راهنمای انجام آزمون‌های t را بدست آورد. سپس با انتخاب آزمون مورد نظر از ttest

dialogs کادر محاوره ای آن را باز نمود. شکل زیر راهنمای آزمون ها را نشان می دهد. این آزمون ها به منظور مقایسه میانگین در یک گروه با یک عدد مشخص و مقایسه میانگین ها در دو گروه و یا دو نمونه بکار می روند. توضیح هر آزمون در قسمت مربوطه با مثال آمده است.



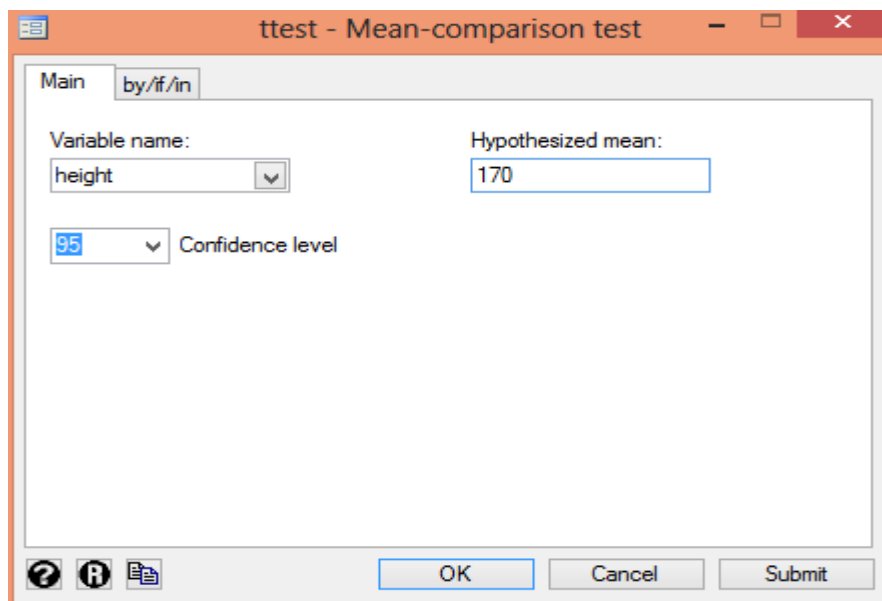
ttest dialogs

- One-sample mean-comparison test
- Two-sample mean-comparison test (unpaired)
- Two-sample mean-comparison test (paired)
- Two-group mean-comparison test

ttesti dialogs

- One-sample mean-comparison test calculator
- Two-sample mean-comparison test calculator

با انتخاب گزینه اول که برای مقایسه میانگین یک نمونه با یک عدد مشخص بکار می رود کادر محاوره ای زیر باز می شود. متغیر مورد آزمون (height) و عدد ۱۷۰ را بعنوان مقدار ثابت میانگین، برای گزینه های کادر محاوره ای معرفی می شوند.



نتایج آزمون در زیر نمایش داده شده است.

```
. ttest height == 170
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
height	60	160.9667	1.092855	8.465216	158.7799	163.1535

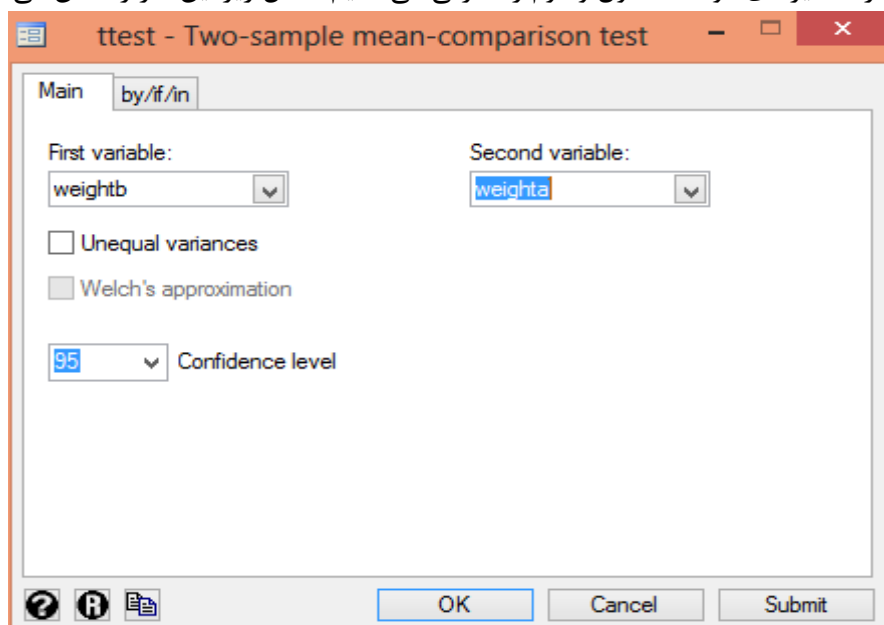
mean = mean(height) t = -8.2658  
 Ho: mean = 170 degrees of freedom = 59

Ha: mean < 170 Ha: mean != 170 Ha: mean > 170  
 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

ملاحظه می شود که میانگین قد در جامعه مورد بررسی با عدد ۱۷۰ اختلاف داشته و از نظر آماری با مقدار p برابر با 0.0000 معنی دار است. در اینجا فرضیه های مورد نظر (یکطرفه و دو طرفه) نیز ارائه شده اند. فرضیه مورد نظر در اینجا دو طرفه بوده است.

### دومین آزمون: مقایسه میانگین دو متغیر در دو جامعه Two-Sample Mean- Comparison Test

در فایل مذکور چنانچه فرض شود متغیر وزن افراد را در دو جامعه اندازه گیری و بنام های weighta و weightb ثبت شده باشند. برای انجام آزمون فرضیه " میانگین وزن دو جامعه مورد بررسی برابر است " از مسیر زیر استفاده می شود. ابتدا از کادر راهنمای آزمون t گزینه دوم انتخاب می شود. سپس با انتخاب آن کادر محاوره ای باز می شود. در این کادر، متغیرهای دو جامعه اول و دوم را معرفی می کنیم. شکل زیر این کار را نشان می دهد.



همانگونه که ملاحظه میگردد خروجی نرم افزار به عنوان نتایج آزمون در زیر آمده است.

```
. ttest weightb == weighta, unpaired
```

Two-sample t test with equal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
weightb	60	75.13333	1.605159	12.43351	71.92142	78.34525
weighta	60	74.95	1.582018	12.25426	71.78439	78.11561
combined	120	75.04167	1.122154	12.29258	72.81969	77.26364
diff		.1833333	2.253734		-4.279674	4.64634

```
diff = mean(weightb) - mean(weighta)          t = 0.0813
Ho: diff = 0                                degrees of freedom = 118
```

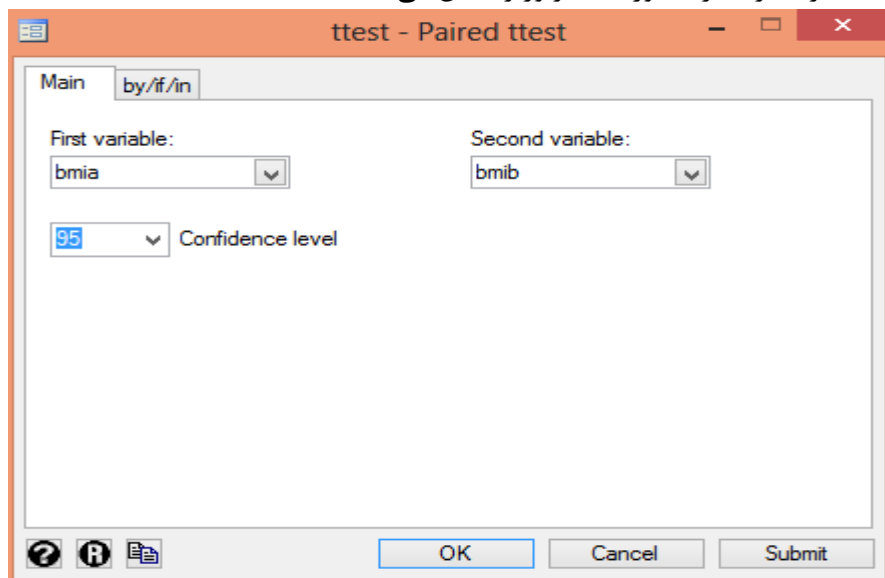
```
Ha: diff < 0                                Ha: diff != 0                                Ha: diff > 0
Pr(T < t) = 0.5323                          Pr(|T| > |t|) = 0.9353                          Pr(T > t) = 0.4677
```

نتایج نشان می دهد که وزن دو جامعه مورد بررسی از نظر آماری اختلاف معنی داری ندارد. مقدار دقیق p در بالا برای فرضیه مورد نظر آمده است.

**سومین آزمون: آزمون مقایسه میانگین دو گروه به هم وابسته یا آزمون میانگین قبل و بعد**

### Two-Sample Mean Comparison Test(Paired Test)

برای انجام این آزمون، گزینه سوم در کادر راهنمای آزمون t انتخاب و کادر محاوره ای مربوطه باز میگردد. به عنوان مثال در فایل مذکور برای مقایسه میانگین شاخص نمایه توده بدنی در یک گروه از افراد که مداخلات آموزشی و تغذیه ای انجام گردید آزمون انجام می شود. برای این کار در کادر محاوره ای باز شده متغیر bmia مربوط به قبل از مداخله و متغیر bmib مربوط به اندازه گیری نمایه توده بدنی بعد از مداخله معرفی می شوند. شکل زیر تعریف متغیرها در کادر محاوره ای مزبور را نشان می دهد.



نتایج آزمون مذکور در زیر آمده است.

```
. ttest bmia == bmib
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
bmia	60	29.03079	.6289056	4.871482	27.77235	30.28923
bmib	60	29.08959	.6290956	4.872954	27.83077	30.3484
diff	60	-.058797	.0806983	.625086	-.2202739	.1026798

```

mean(diff) = mean(bmia - bmib)          t = -0.7286
Ho: mean(diff) = 0                      degrees of freedom = 59

Ha: mean(diff) < 0                      Ha: mean(diff) != 0          Ha: mean(diff) > 0
Pr(T < t) = 0.2346                      Pr(|T| > |t|) = 0.4691        Pr(T > t) = 0.7654

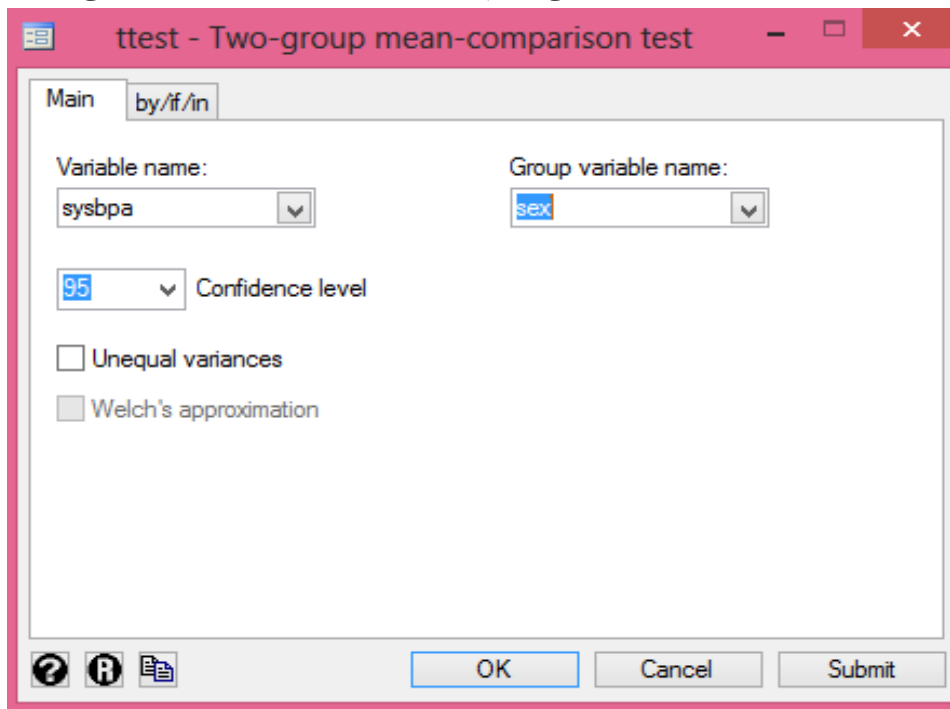
```

بر اساس یافته های بالا شاخص نمایه توده بدنی در قبل و بعد از مداخله انجام شده تغییری از نظر آماری نکرده و مقادیر p بر حسب فرضیه مورد نظر در بالا آمده است. نتایج نشان می دهد که میانگین مقدار دقیق شاخص نمایه توده بدنی، بعد از مداخله به 29.08959 رسید. مقدار معنی دار با  $p=0.4691$  گزارش گردید.

### چهارمین آزمون: آزمون مقایسه میانگین در دو گروه در یک جامعه

#### Two-Group Mean Comparison Test

برای آزمون مقایسه میانگین فشارخون سیستولیک (متغیر sysbpa) در دو گروه مردان و زنان (متغیر sex) آزمون t انجام می دهیم. چنانچه از راهنمای انجام آزمون t گزینه آخر two – group mean- comparison test انتخاب شود کادر محاوره ای زیر باز و با تعریف و معرفی متغیرها آزمون را می توان انجام داد. برای فرض برابری یا عدم برابری واریانس ها گزینه مربوطه را انتخاب می کنیم. شکل زیر کادر محاوره ای را نشان می دهد.



خروجی آزمون به صورت زیر است.

```
. ttest sysbpa, by (sex)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
male	17	120	4.112392	16.95582	111.2821	128.7179
female	43	116.2791	1.605405	10.52735	113.0392	119.5189
combined	60	117.3333	1.631551	12.63794	114.0686	120.5981
diff		3.72093	3.618953		-3.523192	10.96505

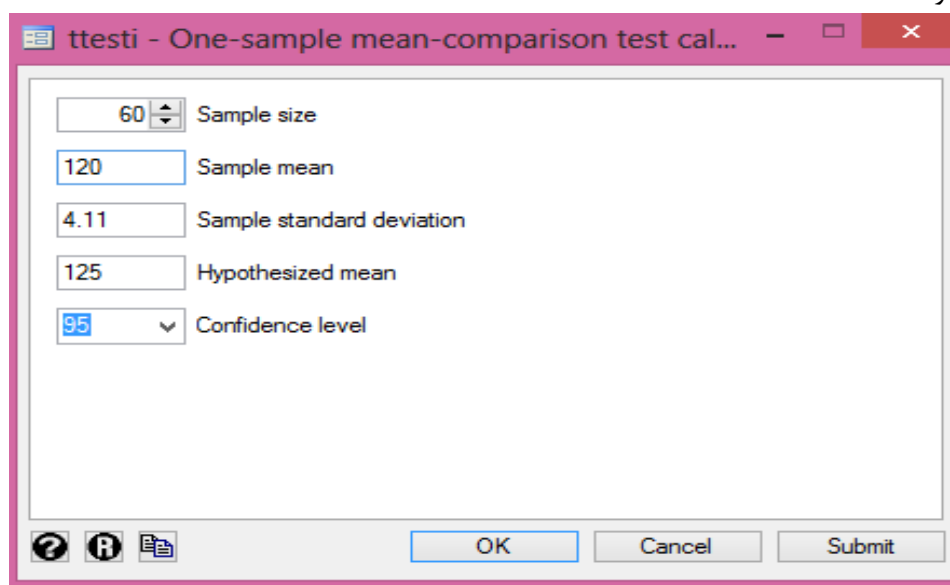
diff = mean(male) - mean(female) t = 1.0282  
Ho: diff = 0 degrees of freedom = 58

Ha: diff < 0 Pr(T < t) = 0.8459  
Ha: diff != 0 Pr(|T| > |t|) = 0.3081  
Ha: diff > 0 Pr(T > t) = 0.1541

ضمن ارائه جدول توصیفی و مشاهده تعداد در هر گروه، میانگین و انحراف معیار آنها، مقادیر اختلاف و حدود اطمینان ۹۵٪، فرضیه های مورد نظر، اختلاف معنی داری بین فشارخون مردان و زنان وجود ندارد. مقادیر p در بالا نوشته شده است. مقدار دقیق p برای آزمون فرضیه دو طرفه 0.3081 و معنی دار نبوده است.

### انجام آزمون های مقایسه میانگین با استفاده از کادر محاوره ای : ttesti dialogs

وقتی حجم نمونه، میانگین و انحراف معیار برای یک متغیر کمی وجود داشته باشد. مثلاً از پژوهش های قبلی بدست آمده و در دسترس باشد. اکنون هدف مقایسه میانگین مربوط به همان متغیر در پژوهش فعلی با پژوهش قبلی باشد از کادر راهنمای انجام آزمون t گزینه اول کادر محاوره ای ttesti dialogs انتخاب می شود. شکل زیر این انتخاب را نشان می دهد. اعداد مورد نظر وارد و OK انجام می شود. خروجی نرم افزار همانند آزمون های قبلی تفسیر می شود.



به عنوان مثال پژوهش قبلی در اندازه گیری میانگین فشارخون سیستولیک در مردان ۱۲۰ میلی متر جیوه با انحراف معیار ۴/۱۱ گزارش گردید. (خروجی آزمون چهارم در بالا)، اکنون محقق با انجام پژوهشی و بدست آوردن



میانگین فشارخون مردان برابر با ۱۲۵ میلی متر جیوه، بدون داشتن داده های قبلی و فقط استفاده از شاخص میانگین و انحراف معیار فشارخون که در گذشته انجام شده این آزمون را به صورت زیر انجام می دهد. کادر محاوره ای برای وارد کردن و تایپ اعداد در بالا آمده است.

```
. ttesti 60 120 4.11 125
```

One-sample t test

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
x	60	120	.5305987	4.11	118.9383	121.0617

mean = mean(x)

t = -9.4233

Ho: mean = 125

degrees of freedom = 59

Ha: mean < 125

Ha: mean != 125

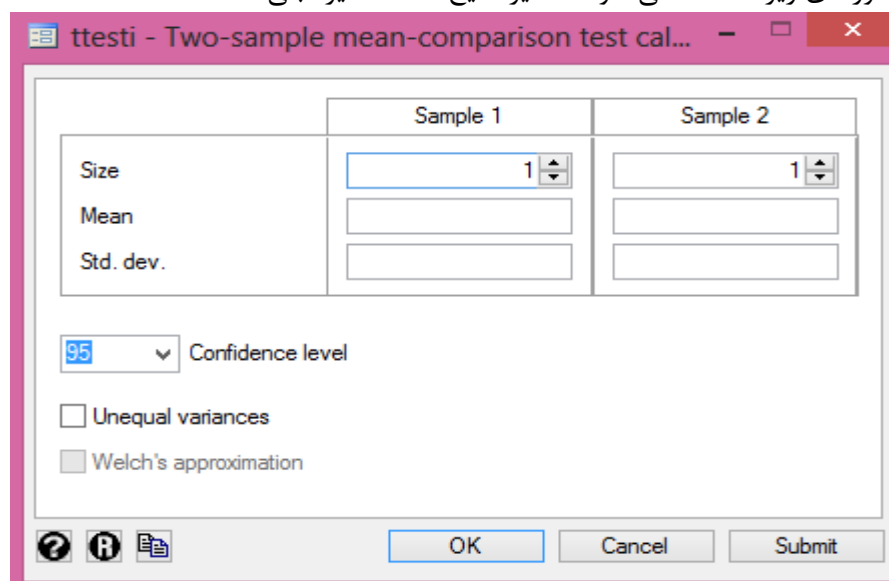
Ha: mean > 125

Pr(T < t) = 0.0000

Pr(|T| > |t|) = 0.0000

Pr(T > t) = 1.0000

نتایج نشان می دهد که بین میانگین فشارخون سیستولیک که محقق هم اکنون بدست آورد و میانگین فشارخون گزارش شده در قبل اختلاف آماری معنی داری وجود دارد. مقادیر دقیق پی در بالا گزارش شده است. همانند مسیر بالا، وقتی که میانگین و انحراف معیار در دو نمونه در دسترس باشد. و هدف آزمون میانگین در دو نمونه باشد از کادر محاوره ای زیر استفاده می شود. تفسیر نتایج مانند مسیر قبلی است.



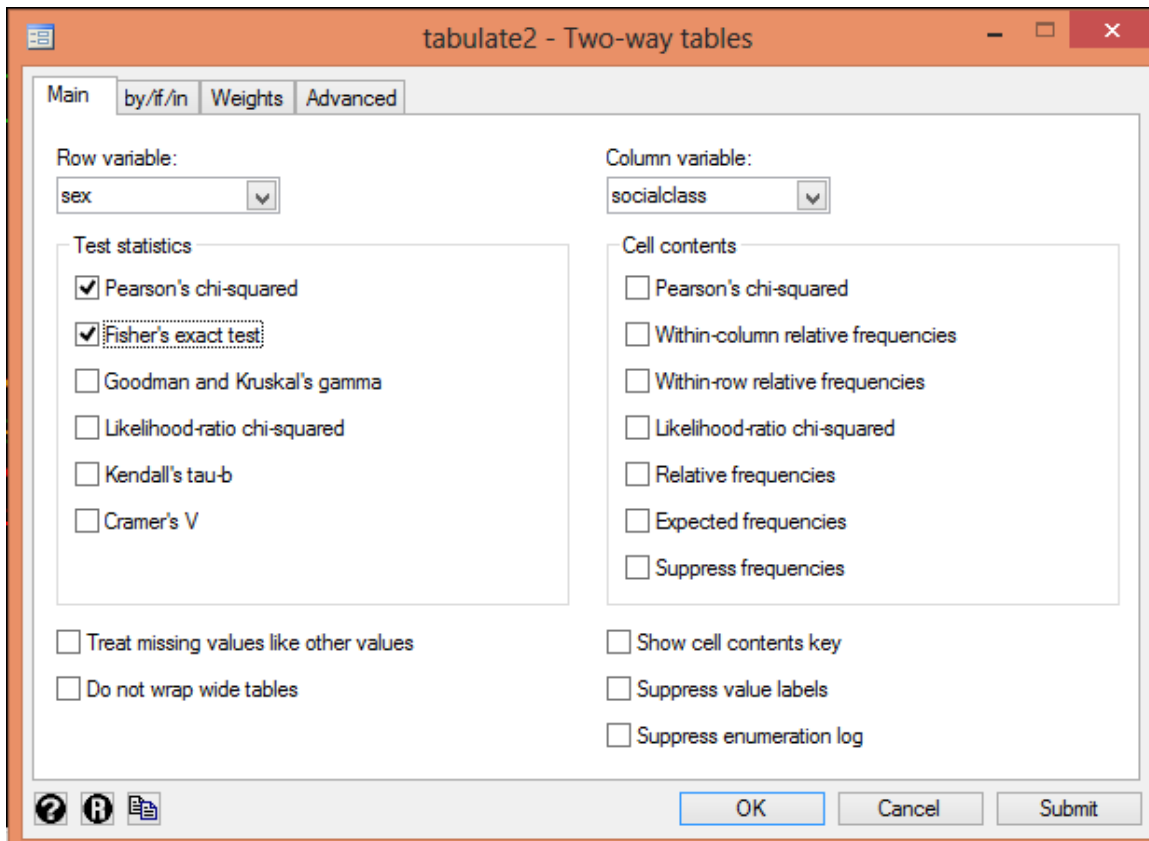
Do-file دستورات و ثبت کارهای مذکور، در زیر آمده است.

```
log using part5, text
ttest sysbpa, by(sex)
db ttest
ttest height == 170
ttest weightb == weightb, unpaired
ttest weightb == weighta, unpaired
ttest bmia == bmib
ttesti 60 120 4.11 125
ttesti 60 120 4.11 125
```

### آزمون کای دو و آزمون های تعقیبی: Chi square

وقتی محقق بخواهد فراوانی دو متغیر را به طور هم زمان در یک جدول نشان دهد از جدول دو طرفه استفاده می کند. این اقدام در فصل آمار توصیفی نمایش داده شد. چنانچه هدف بدست آوردن رابطه و انجام آزمون آماری کای دو (Chi square) باشد این اقدام از مسیر زیر انجام می شود.

به عنوان مثال در فایل part five وقتی هدف بررسی رابطه جنسیت و طبقه اجتماعی باشد. از کادر محاوره ای db tabulate2 استفاده می شود. کادر محاوره ای به شکل زیر باز و متغیرهای sex و socialclass به شکل زیر معرفی و گزینه های مورد نظر از جمله انجام کای اسکور و یا تست دقیق فیشر برای مواردی که فراوانی خانه ها کم است استفاده می شود. همچنین می توان فراوانی های مطلق و نسبی مشاهده شده و مورد انتظار برای سطر و ستون جدول را در کادر محاوره ای انتخاب و بدست آورد.



در کادر محاوره ای بالا می توان با معرفی دو متغیر، انواع شاخص های ارتباط را محاسبه نمود. خروجی نرم افزار در زیر آمده است.

```
. tabulate sex socialclass, chi2 exact
```

Enumerating sample-space combinations:

stage 3: enumerations = 1

stage 2: enumerations = 0

stage 1: enumerations = 0

sex	class			Total
	1	2	3	
male	6	7	4	17
female	14	17	12	43
Total	20	24	16	60

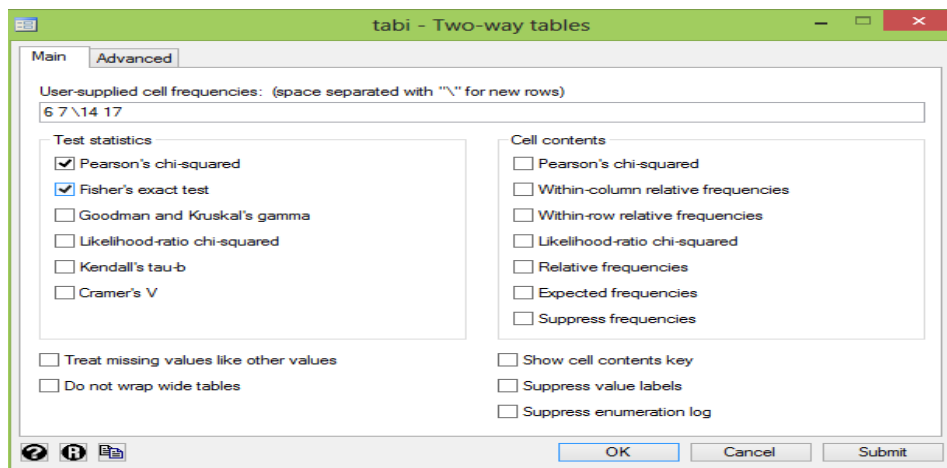
Pearson chi2(2) = 0.1231 Pr = 0.940

Fisher's exact = 1.000

نتایج نشان می دهد که بین طبقه اجتماعی و جنسیت ارتباط آماری معنی داری وجود ندارد. از قابلیت منحصر به فرد نرم افزار ایجاد جدول مجازی است. به عنوان مثال هنگامی که محقق جدولی از خلاصه داده های زیر را بدون داشتن فایل داده ها در اختیار دارد برای انجام آزمون بررسی رابطه در جدول به طریق زیر عمل می کند.

sex	class			Total
	1	2	3	
male	6	7	4	17
female	14	17	12	43
Total	20	24	16	60

ابتدا کادر محاوره ای db tabi را باز و سپس داده های جدول بالا را به صورت ردیف به ردیف و از چپ به راست وارد می کند. بین اعداد هر خانه فقط یک فاصله ایجاد شود. هر ردیف از ردیف بعدی با علامت \ جدا می شود. در کادر محاوره ای که در شکل زیر نمایش داده می شود. این اقدامات انجام شده است. وقتی دستور اجرا شود فایل داده برای جدول مذکور ایجاد می شود.



خروجی نرم افزار در زیر نمایش داده می شود.

. db tabi

. tabi 6 7 \14 17, chi2 exact

row	col		Total
	1	2	
1	6	7	13
2	14	17	31
Total	20	24	44

Pearson chi2(1) = 0.0036 Pr = 0.952  
 Fisher's exact = 1.000  
 1-sided Fisher's exact = 0.605

### بررسی رابطه و انواع همبستگی ها

آزمون کای اسکور معمولا برای رابطه هایی به کار می رود که هر دو متغیر ناپارامتری باشند. جز در موردی که یک متغیر اسمی و دیگری فاصله ای باشد. هر گاه در یک نمونه مورد مطالعه، هیچ رابطه سیستماتیکی بین دو متغیر وجود نداشته باشد می توان نتیجه گرفت که دو متغیر از یکدیگر مستقل هستند و به آن اصطلاحا استقلال آماری می گویند. هر چه مقدار آزمون کای دو بیشتر باشد احتمال وجود رابطه بین دو متغیر بیشتر است و برعکس. کای اسکور چگونگی و قوت رابطه را توضیح نمی دهد. بنابراین پس از محاسبه کای اسکور در صورت وجود رابطه بین متغیرها، باید با استفاده از شاخص های پیوند و همبستگی جهت رابطه و شدت آن را نیز تعیین کرد. در این آزمون با استفاده از تفاوت بین فراوانی های مشاهده شده و مورد انتظار به وجود یا عدم وجود رابطه بین دو متغیر پرداخته می شود.

برای استفاده از آزمون کای دو باید فراوانی های مورد انتظار از قبل مشخص شده باشند، که معمولا نیز برای انجام این کار، توزیع فراوانی های مشاهده شده را با یک توزیع نظری مثلا توزیع نرمال مقایسه می کنیم. تفاوت این دو فراوانی نشان از وجود رابطه و عدم تفاوت بین این دو فراوانی نشان از استقلال و عدم ارتباط بین دو متغیر مورد مطالعه دارد. با استفاده از فرمول زیر مقدار رابطه بدست می آید.

$$\chi^2 = R = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

درجه آزادی در این آزمون تابع حجم نمونه نیست ( بر خلاف آزمون های t و F ) بلکه تابع تعداد مقوله هاست. در یک جدول دو بعدی که تعداد سطر برابر k و تعداد ستون برابر l است درجه آزادی تعداد سطر منهای یک ضربدر تعداد ستون منهای یک است.

### انواع آزمون های کای اسکور استقلال:

- ۱- کای اسکور پیرسون: متداولترین آزمون مورد استفاده است.
- ۲- کای اسکور تصحیح یتس: تعدیل شده کای اسکور پیرسون است. زمانی بکار می رود که فراوانی مورد انتظار یک یا چند سلول کمتر از ۵ باشد. این آزمون تنها برای جداول دو در دو بکار می رود.
- ۳- کای اسکور نسبت درستنمایی: به آزمون درست نمایی یا آزمون جی معروف است. همانند پیرسون است.

اما این آزمون به حجم نمونه، توزیع حاشیه ای و ترتیب سطر و ستون حساسیت ندارد. بنابراین در برخی موارد نسبت به پیرسون ارجحیت دارد. تفسیر آن شبیه کای اسکور پیرسون است.

۴- آزمون دقیق فیشر: وقتی جدول دو در دو باشد و درجه آزادی یک و این تست بکار می رود. هنگامی که حجم نمونه کم باشد (معمولاً کمتر از ۵۰ مورد)، ادغام طبقات متغیرها در هم امکان پذیر نیست و از این آزمون استفاده می شود.

۵- آزمون همبستگی خطی - خطی (linear by linear association): برای جداول بزرگ و مستطیل شکل مانند جدول سه در چهار یا سه در دو و به منظور آزمون روند خطی بین دو متغیر استفاده میشود.

۶- کای اسکور مانتل هنزل: برای آزمون استقلال بین دو متغیر اسمی دو وجهی با شرط وجود یک یا چند متغیر کنترل یا لایه بکار می رود. ممکن است با نام آزمون کوکران و مانتل - هنزل نیز تعریف شود.

برای استفاده از این آزمون ها، داده ها باید تصادفی انتخاب و از همدیگر مستقل باشند. برای هر نوع متغیری اعم از اسمی، ترتیبی و فاصله ای قابل استفاده است. فرضیه در این آزمون فاقد جهت است. فراوانی مورد انتظار نباید در هیچ سلولی صفر باشد. نسبت به حجم نمونه حساس است. فراوانی مورد انتظار بیش از ۲۰٪ خانه ها نباید کمتر از ۵ باشد. اگر شرط فراوانی منتظره برقرار نبود باید طبقات تلفیق شده یا از آزمون بدیل (آزمون دقیق فیشر و تصحیح یتس) استفاده کرد.

### شاخص های همبستگی که به عنوان آزمون های تعقیبی کای اسکور استقلال بکار می روند:

#### الف) شاخص ها یا ضرایب برای همبستگی اسمی:

۱- ضریب توافق (contingency coefficient) که به ضریب توافق پیرسون معروف است. برای دو متغیر کیفی بکار می رود.

۲- ضریب فی (phi) وقتی بکار می رود که سطح سنجش دو متغیر اسمی باشد. یعنی هر دو متغیر دارای دو طبقه باشند.

۳- ضریب وی کرامر (Cramers V) گسترش یافته آزمون فی است و زمانی استفاده می شود که یکی از متغیرها در سطح اسمی و دیگری در سطح اسمی یا ترتیبی باشد. برای متغیرهای اسمی چندوجهی کاربرد دارد.

۴- ضریب لاندا (Lambda) این ضریب به شاخص گاتمن نیز معروف است. زمانی کاربرد دارد که هر دو متغیر مورد سنجش در سطح اسمی چندوجهی باشند.

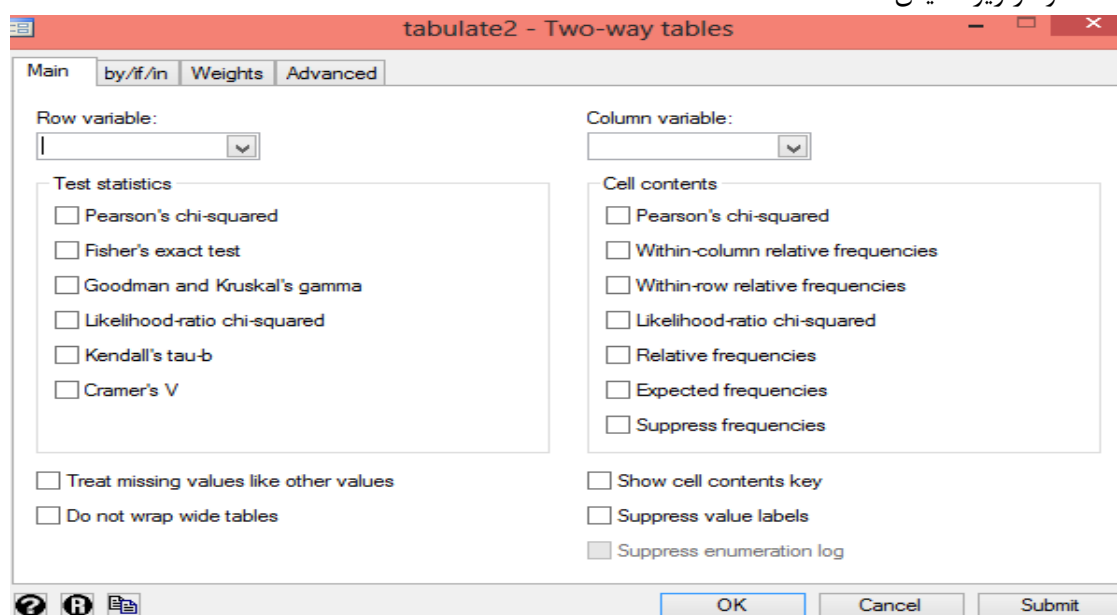
۵- ضریب تاوی گودمن و کروسکال (Goodman and kruskals gamma) زمانی کاربرد دارد که هر دو متغیر مورد سنجش در سطح اسمی چندوجهی باشند. این ضریب یک ضریب متقارن است و بنابراین وقتی متغیرهای مستقل و وابسته نیز مشخص نباشند کاربرد دارد. از داده های تمام جدول استفاده می کند. زمانی که تعداد طبقات متغیر وابسته بیشتر از متغیر مستقل باشد ممکن است بین ۰ تا ۱ نباشد.

#### ب) شاخص های همبستگی ترتیبی:

این ضرایب زمانی کاربرد دارند که متغیرها در سطح سنجش ترتیبی - ترتیبی، و یا ترتیبی - فاصله ای باشند. متغیرهای ترتیبی توزیع نامتقارن دارند. مقدار و جهت رابطه آنها از منهای یک تا مثبت یک تغییر

میکنند. این ضرایب عبارتند از:

- ۱- ضریب گاما که اغلب بنام ضریب تاوی گودمن و کروسکال (Goodman and kruskals gamma) مورد استفاده قرار می گیرد. و سنجش متغیرها در سطح ترتیبی باشد.
  - ۲- ضریب تاوی - بی کندال (kendalls tau-b) وقتی کاربرد دارد که متغیرها در مقیاس ترتیبی - ترتیبی باشند و جدول به شکل مربع بعنوان مثال سه در سه یا چهار در چهار باشد.
- این شاخص ها یا ضرایب با استفاده از کادر محاوره ای db tabi یا db tabulate2 به سهولت قابل دسترسی است و در زیر نمایش داده شده اند.



**برای اندازه گیری رابطه و همبستگی از شاخص های زیر استفاده می شود:**

**همبستگی پارامتری ( ضریب همبستگی پیرسون ) :**

برای محاسبه درجه و میزان ارتباط خطی بین دو متغیر در سطح فاصله ای و نسبتی بکار می رود. مقادیر آن بین منهای یک تا مثبت یک در نوسان است. علامت آن نشانگر جهت رابطه است. توزیع داده ها باید از توزیع نرمال تبعیت کند. برای اینکه از نحوه ارتباط خطی یا غیر خطی بین متغیرها اطلاع حاصل شود ابتدا باید نمودار پراکنش بین دو متغیر ترسیم گردد. این نمودار به کاربر کمک می کند.

**همبستگی ناپارامتری:**

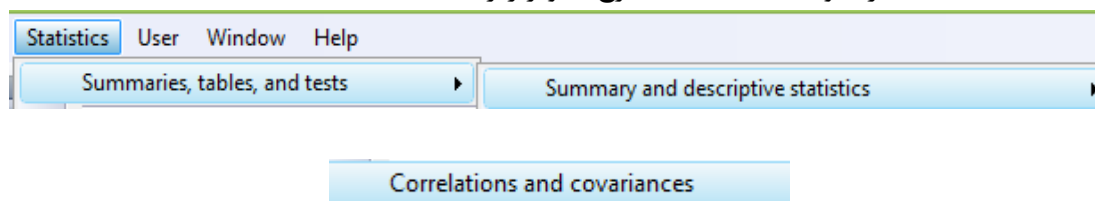
شامل دو ضریب اسپیرمن و تاوی - بی کندال است. که برای آزمون همبستگی بین دو متغیر با مقیاس ترتیبی به کار می رود. این ضرایب معادل ناپارامتری ضریب همبستگی پیرسون هستند. چنانچه تعداد موردها بسیار زیاد و طبقات هر متغیر محدود باشد بهتر است بجای اسپیرمن از تاوی - بی کندال استفاده شود.

ضریب تاوی - بی کندال یک ضریب متقارن است. و برای صفات مرتب شده یا رتبه داده شده بکار می رود. وقتی نسبت موردها به طبقات کم باشد یعنی تعداد افراد کم و تعداد متغیرها زیاد است ، ضریب اسپیرمن مناسب تر است. یا بعبارتی در حجم نمونه کم ضریب همبستگی اسپیرمن مطلوب تر است. در همبستگی اسپیرمن رتبه را بصورت نمره نشان میدهند در حالیکه در همبستگی تاوی - بی کندال از تعداد توافق ها و عدم توافق در رتبه

بندی برای محاسبه ضریب همبستگی استفاده می شود. بطور کلی مقدار ضریب اسپیرمن بیشتر از تاوی – کندال بوده و نزدیک به ضریب همبستگی پیرسون است. معمولا دامنه همبستگی 0.2 تا 0.35 را ضعیف، همبستگی 0.35 تا 0.65 را مناسب و متوسط ، همبستگی 0.65 تا 0.85 را مطلوب و قابل قبول و همبستگی بالاتر از 0.85 را بالا و نشان دهنده رابطه نزدیک بین دو متغیر می دانند.

### کار عملی با نرم افزار و محاسبه ضریب همبستگی:

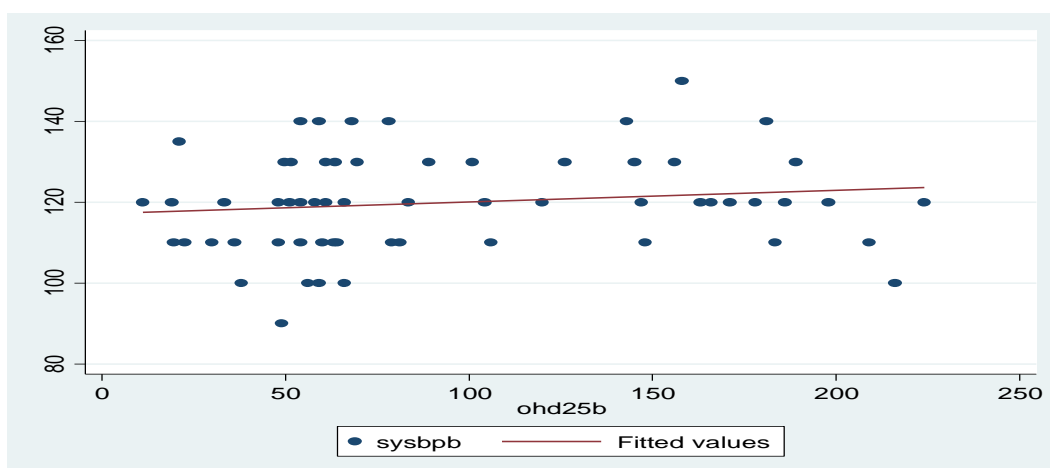
برای بدست آوردن رابطه همبستگی بین دو متغیر کمی از ضریب همبستگی پیرسون استفاده می شود. در فایل داده های part five برای بدست آوردن رابطه بین متغیر سطح ویتامین دی (ohd25b) با فشارخون سیستولیک (sysbpb) ، فشارخون دیاستولیک (diasbpb)، سن (age) و نمایه توده بدنی (bmib) و هموگلوبین گلیکوزیله (hba1cb) به طریق زیر عمل می شود. ابتدا گزینه correlation and covariances را از Summary and descriptive statistics از منوی statistics به شرح تصویر زیر انتخاب نمایید.



قبل از اجرای دستور ضریب همبستگی، نمودار پراکنش برای متغیرهای سطح ویتامین دی و فشارخون خون سیستولیک با خط برازش با دستورات زیر رسم می شود.

twoway (scatter sysbpb ohd25b)

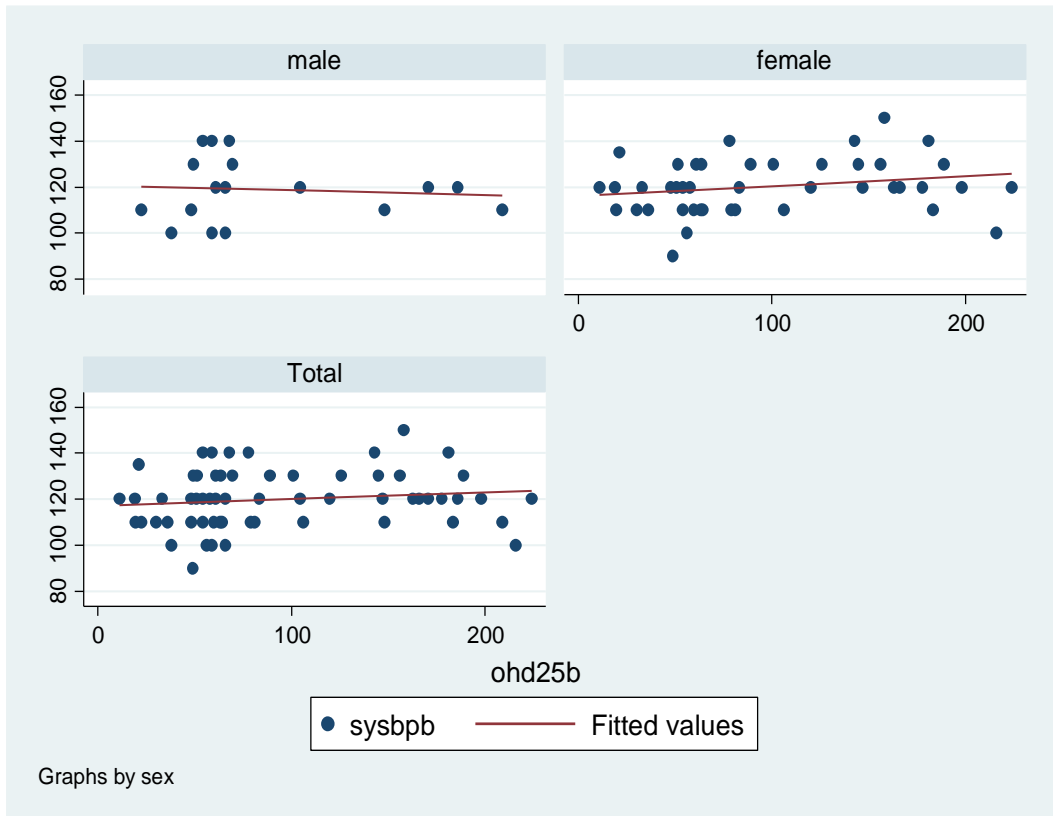
twoway (scatter sysbpb ohd25b) (lfit sysbpb ohd25b)



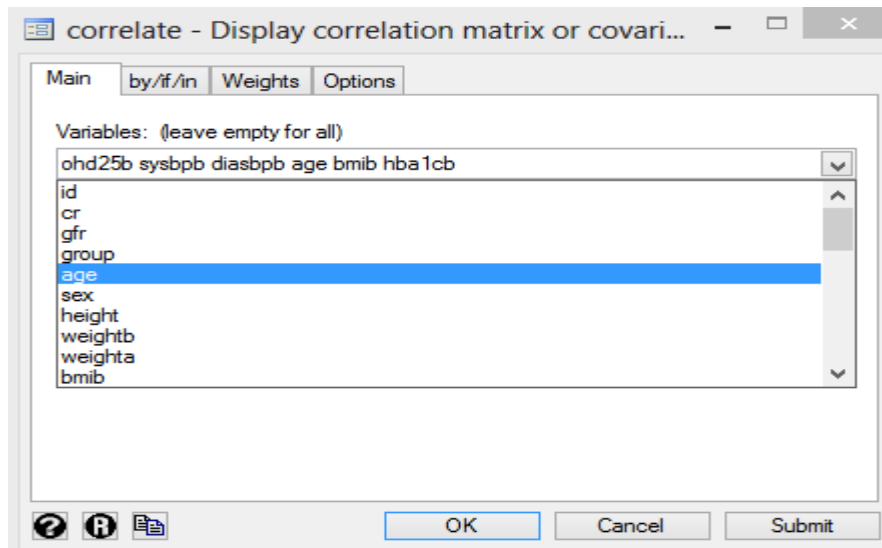
مجدد نمودار بر حسب جنسیت به صورت زیر ترسیم می شود.

twoway (scatter sysbpb ohd25b) (lfit sysbpb ohd25b), by(sex, total)

نمودار پراکنش فشارخون سیستولیک و سطح ویتامین دی بر حسب جنسیت:



سپس با دوبار کلیک نمودن بر correlation and covariances کادر محاوره ای زیر باز و متغیرهایی که قصد بررسی روابط بین آنها را دارید انتخاب می گردند.



نتایج در جدول زیر نمایش داده شده است. اگر نیاز به محاسبه میانگین و انحراف معیار و سایر شاخص های توصیفی این متغیرها باشد در گزینه option این شاخص ها باید انتخاب شوند. خروجی در زیر نشان داده شده است.



## روش های اپیدمیولوژی و پژوهش با Stata

```
. correlate ohd25b sysbpb diasbpb age bmib hba1cb, means
(obs=60)
```

Variable	Mean	Std. Dev.	Min	Max
ohd25b	94.87167	59.21349	11	224
sysbpb	119.9167	12.36766	90	150
diasbpb	80.41667	7.718109	60	95
age	55.01667	10.65084	34	76
bmib	29.08959	4.872954	20.04745	43.40947
hba1cb	7.64	.5015907	6.6	8.9

	ohd25b	sysbpb	diasbpb	age	bmib	hba1cb
ohd25b	1.0000					
sysbpb	0.1387	1.0000				
diasbpb	0.1685	0.8438	1.0000			
age	0.4009	0.4021	0.3783	1.0000		
bmib	0.0751	0.2090	0.2922	0.0370	1.0000	
hba1cb	-0.1154	-0.1101	-0.1182	-0.0959	0.0516	1.0000

در بررسی رابطه همبستگی بین دو متغیر کمی اندازه رابطه که با ضریب همبستگی بدست می آید، علامت ضریب همبستگی که جهت رابطه را تعیین میکند و مقدار معنی داری رابطه با اهمیت است. چنانچه قصد گزارش مقدار معنی داری روابط دارید از دستور زیر استفاده نمایید.

```
pwcorr ohd25b sysbpb diasbpb hba1cb age bmib, sig
```

با اجرای دستور مذکور خروجی زیر نمایش داده می شود.

```
. pwcorr ohd25b sysbpb diasbpb hba1cb age bmib, sig
```

	ohd25b	sysbpb	diasbpb	hba1cb	age	bmib
ohd25b	1.0000					
sysbpb	0.1387	1.0000				
	0.2905					
diasbpb	0.1685	0.8438	1.0000			
	0.1981	0.0000				
hba1cb	-0.1154	-0.1101	-0.1182	1.0000		
	0.3798	0.4023	0.3684			
age	0.4009	0.4021	0.3783	-0.0959	1.0000	
	0.0015	0.0014	0.0029	0.4659		
bmib	0.0751	0.2090	0.2922	0.0516	0.0370	1.0000
	0.5683	0.1090	0.0235	0.6953	0.7789	

خروجی بالا نشان می دهد که بین متغیر ohd25b و age ارتباط مستقیم معنا دار با ضریب 0.4 و مقدار پی 0.0015 وجود دارد. تفسیر بقیه ضرایب شبیه ارتباط این دو متغیر است. بین متغیر ohd25b و hba1cb ارتباط غیر مستقیم یا معکوس وجود دارد اما این ارتباط از نظر آماری معنی دار نیست.  $p = 0.3798$

### مقایسه و آزمون ضرایب همبستگی:

با دستور `corrcci` و دستور `corrcci` نیز میتوان ضریب همبستگی و حدود اطمینان آن را برای دو متغیر کمی محاسبه و گزارش نمود. به عنوان مثال محاسبه حدود اطمینان ضریب همبستگی بین دو متغیر سن و `ohd25b` به شکل زیر است.

```
. corrcci age ohd25b
      (obs=60)

      correlation and 95% limits
      age      ohd25b      0.401      0.164      0.594
```

این ضریب با روش فیشر و روش جیفری به شکل زیر است. برای آزمون دو ضریب همبستگی نیز می توان اقدام نمود. این دستورات به طور پیش فرض در نرم افزار تعبیه نشده اند و لذا باید با دستور `findit` آنها را دانلود و نصب نمود.

```
findit corrcci
findit corrcci
```

بعد از اجرای دستورات مذکور، پکیج های قابل نصب (`package installation`) دانلود و با نصب آنها می توان تحلیل همبستگی را انجام داد. با این دستورات می توان دو ضریب همبستگی را با هم مقایسه نمود.

```
. corrcci age ohd25b, fisher
      (obs=60)

      correlation and 95% limits
      age      ohd25b      0.401      0.150      0.585

. corrcci age ohd25b, jeffreys
      (obs=60)

      correlation and 95% limits
      age      ohd25b      0.401      0.154      0.579
```

همچنین با دستور `cortesti r1 n1 r2 n2` نیز می توان دو ضریب همبستگی  $r_1$  و  $r_2$  از دو نمونه مستقل با حجم نمونه  $n_1$  و  $n_2$  را آزمون کرد. فرض صفر برابری دو ضریب همبستگی است. این آزمون تقریبی است و در صورتی که حجم نمونه بالاتر از ۱۰ باشد کاربرد دارد. به عنوان مثال برای آزمون مقایسه ضریب همبستگی بین دو متغیر در نمونه ای با حجم ۲۹ که بدست آمد با ضریب همبستگی بین دو متغیر دیگر در نمونه دیگر با حجم ۳۵ و مقدار ضریب برابر با ۰.۶۴ نتایج زیر با نرم افزار بدست آمده است.

```
. cortesti 0.68 29 0.64 35

Test of equality of two correlation coefficients
drawn from the two different samples
-----

Coefficient 1 = 0.680 (n1 = 29)
Coefficient 2 = 0.640 (n2 = 35)

Ho: coefficient 1 = coefficient 2
      z = 0.269
Prob > |z| = 0.7882
```

همانگونه که نتیجه بالا نشان می دهد دو ضریب همبستگی با هم برابر هستند و از نظر آماری تفاوت معنی داری بین آن ها یافت نگردید.  $Z=0.269$  و  $P=0.7882$  علاوه بر آزمون های بالا، می توان با دستور `cortest rxy rXV` ضریب همبستگی را بین دو متغیر و متغیر سوم در یک نمونه مقایسه و آزمون نمود. به عنوان مثال در نظر بگیرید که در فایل مذکور ضریب همبستگی بین سن و قد ( $r_{XY}$ )، سن و وزن ( $r_{XV}$ ) و بین وزن و قد ( $r_{VY}$ ) و حجم نمونه ۶۰ نفر را می خواهید مقایسه و آزمون نمایید ابتدا مقادیر ضریب همبستگی بین متغیرهای بالا را محاسبه، سپس آزمون انجام می شود. نتیجه انجام این آزمون نیز در زیر آمده است.

```
. corr age height
(obs=60)
```

	age	height
age	1.0000	
height	-0.1017	1.0000

```
. corr age weightb
(obs=60)
```

	age	weightb
age	1.0000	
weightb	-0.0216	1.0000

```
. corr weightb height
(obs=60)
```

	weightb	height
weightb	1.0000	
height	0.2798	1.0000

```
. tab group
```

group	Freq.	Percent	Cum.
case	30	50.00	50.00
control	30	50.00	100.00
Total	60	100.00	

```
. cortesti -0.1017 -0.0216 0.2798 60
```

```
Comparison of correlation coefficients
drawn from the same sample (n = 60 obs)
```

```
-----
Coefficient r(x,y) = -0.102
Coefficient r(x,v) = -0.022
Coefficient r(v,y) = 0.280
```

```
t = -2.035 (df= 57)
```

```
Ha: r(x,y) < r(v,y) --- P < t = 0.023
```

```
Ha: r(x,y) = r(v,y) --- P = t = 0.046
```

```
Ha: r(x,y) > r(v,y) --- P > t = 0.977
```

### همبستگی تفکیکی / جزئی: Partial Correlation

به دلیل پیچیده بودن روابط بین متغیرها در تحلیل، اغلب اوقات برخی از روابط از دید کاربر تحلیل پنهان می ماند و ممکن است در تحلیل در نظر گرفته نشود. به عنوان مثال در پژوهش بالا ممکن است میزان سطح ویتامین دی با توجه به جنسیت یا سایر متغیرها متفاوت باشد. و یا این میزان بصورت مستقیم وابسته به جنسیت نباشد. و متغیرهای مخدوش کننده و کنترل در این رابطه نقش بازی کنند. برای رفع این مشکل و کنترل متغیرهای مخدوش کننده از این نوع همبستگی استفاده می شود. یعنی ضمن محاسبه ضریب همبستگی بین دو متغیر، اثر سایر متغیرهای وارد شده را نیز کنترل می کند.

هر چه مقدار این ضریب برای یک متغیر بیشتر باشد می توان گفت که نقش آن متغیر در مدل بیشتر است. برای استفاده از این ضریب، متغیرها باید متقارن و دارای رابطه خطی و در مقیاس کمی باشند. برای مدل های کوچک سه تا چهار متغیر مناسب است. برای کاربرد این ضریب در نرم افزار همانند بالا از مسیر Summary and descriptive statistics از منوی Partial correlations

انتخاب می شود.

در مثال بالا برای بررسی رابطه میزان ویتامین دی و فشارخون سیستولیک برای کنترل سایر متغیرها و بر حسب جنسیت از این ضریب به صورت زیر استفاده می شود.  
خروجی نرم افزار در زیر نمایانگر است.

```
. pcorr ohd25b age bmib sysbpb diasbpb hbalcb
(obs=60)
```

Partial and semipartial correlations of ohd25b with

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
age	0.3766	0.3692	0.1418	0.1363	0.0042
bmib	0.0663	0.0603	0.0044	0.0036	0.6273
sysbpb	-0.0717	-0.0653	0.0051	0.0043	0.5995
diasbpb	0.0520	0.0472	0.0027	0.0022	0.7037
hbalcb	-0.0895	-0.0816	0.0080	0.0067	0.5118

. by sex, sort : pcorr ohd25b age bmib sysbpb diasbpb hba1cb

-> sex = male  
(obs=17)

Partial and semipartial correlations of ohd25b with

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
age	0.5047	0.4610	0.2547	0.2125	0.0786
bmib	-0.0511	-0.0404	0.0026	0.0016	0.8683
sysbpb	-0.5454	-0.5131	0.2974	0.2633	0.0539
diasbpb	0.3283	0.2740	0.1078	0.0751	0.2735
hba1cb	-0.3320	-0.2776	0.1102	0.0770	0.2677

-> sex = female  
(obs=43)

Partial and semipartial correlations of ohd25b with

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
age	0.4757	0.4630	0.2263	0.2143	0.0022
bmib	0.0863	0.0742	0.0075	0.0055	0.6012
sysbpb	0.1004	0.0864	0.0101	0.0075	0.5431
diasbpb	-0.0413	-0.0354	0.0017	0.0013	0.8029
hba1cb	0.0047	0.0040	0.0000	0.0000	0.9775

آنچه از جداول خروجی بالا قابل تفسیر است این است که در بررسی رابطه سطح ویتامین دی (ohd25b) و فشارخون سیستولیک (sysbpb)، ضریب همبستگی 0.13 با مقدار  $p=0.13$  بدست آمد که معنی دار نیست. وقتی این ضریب بر حسب متغیر جنسیت بصورت جداگانه تحلیل می شود برای زنان 0.22 و برای مردان -0.08 بدست می آید. یعنی جهت رابطه برای مردان معکوس یا غیر مستقیم می شود. و عبارتی جنسیت نقش مخدوش کننده را دارد. در اینجا سطح ویتامین دی با فشارخون سیستولیک در مردان رابطه معکوس و در زنان رابطه مستقیم دارد. لذا برای کنترل سایر متغیرهایی که به نظر مخدوش کننده می رسند از ضریب همبستگی پارشیال استفاده شده و روابط مذکور مجدد به شرح زیر تغییر نموده است:

رابطه سطح ویتامین دی با فشارخون سیستولیک با کنترل سایر متغیرها به -0.07 رسیده است. یعنی اولاً جهت رابطه تغییر نموده و از 0.13 در ضریب همبستگی پیروسون به -0.07 رسید. وقتی بر حسب جنسیت مجدد تحلیل انجام میشود این ضریب از -0.07 برای کل افراد مورد بررسی در مردان به -0.54 و در زنان به 0.10 می رسد. یعنی جهت رابطه وقتی سایر متغیرها کنترل شوند فقط در مردان معکوس است. بنابراین در تفسیر ضرایب همبستگی توجه به این نکات دارای اهمیت است. مجدد به فایل part five برگردید. در این فایل متغیر group دو دسته افراد را معرفی می کند. یک گروه بنام case که ویتامین دی گرفتند و گروه دوم بنام control که

ویتامین دی دریافت نکردند.

. tab group

group	Freq.	Percent	Cum.
case	30	50.00	50.00
control	30	50.00	100.00
<b>Total</b>	<b>60</b>	<b>100.00</b>	

در بیماران گروه case یا مداخله که ویتامین دی دریافت نمودند می بایست سطح ویتامین دی قبل و بعد از مداخله متفاوت بوده ولی در گروه کنترل این تفاوت نباید وجود داشته باشد. برای آزمون این فرضیه از آزمون زوجی استفاده می شود. نتایج به صورت زیر است.

. by group, sort : ttest ohd25b == ohd25a

-> group = case

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
ohd25b	30	83.95667	9.513823	52.10935	64.49871	103.4146
ohd25a	30	164.1	10.51724	57.60529	142.5898	185.6102
diff	30	-80.14333	12.84701	70.36599	-106.4184	-53.86824

mean(diff) = mean(ohd25b - ohd25a) t = -6.2383  
Ho: mean(diff) = 0 degrees of freedom = 29

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0  
Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

ملاحظه می شود که در گروه case اختلاف معنی دار است. و پس از مداخله حدود 80.14 سطح ویتامین دی بالاتر رفت. (P=0.0000). در گروه کنترل یا شاهد نتایج به صورت زیر است.

-> group = control

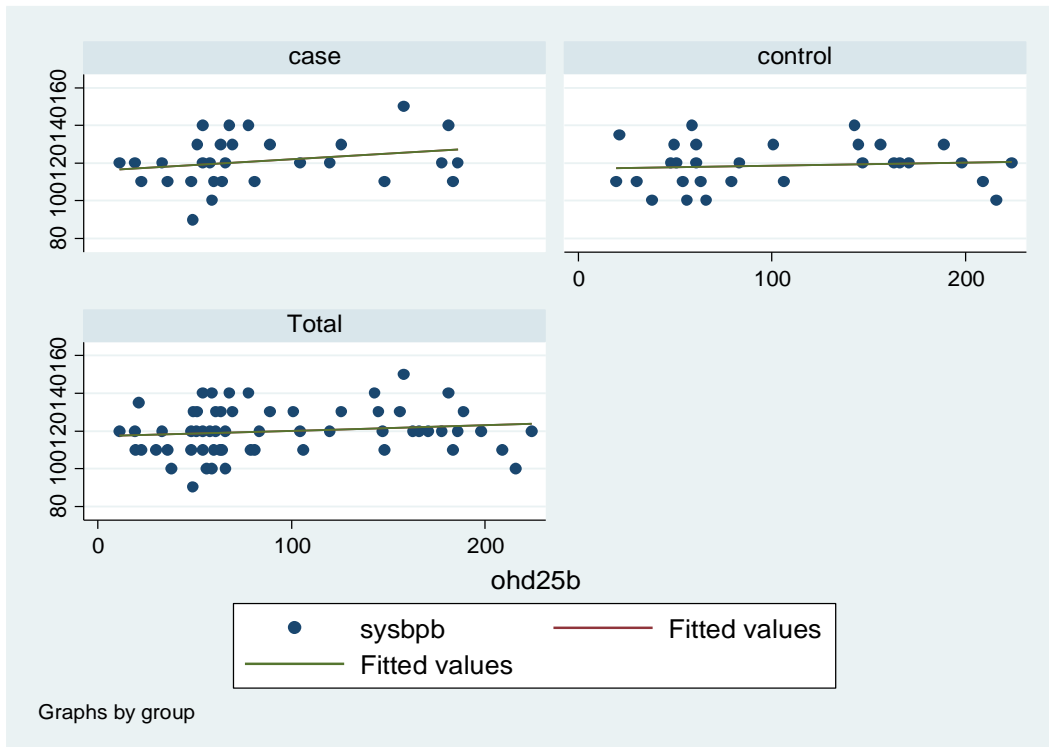
Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
ohd25b	30	105.7867	11.79195	64.58718	81.66942	129.9039
ohd25a	30	115.8283	17.24222	94.43955	80.56403	151.0926
diff	30	-10.04167	11.57736	63.41182	-33.72003	13.6367

mean(diff) = mean(ohd25b - ohd25a) t = -0.8674  
Ho: mean(diff) = 0 degrees of freedom = 29

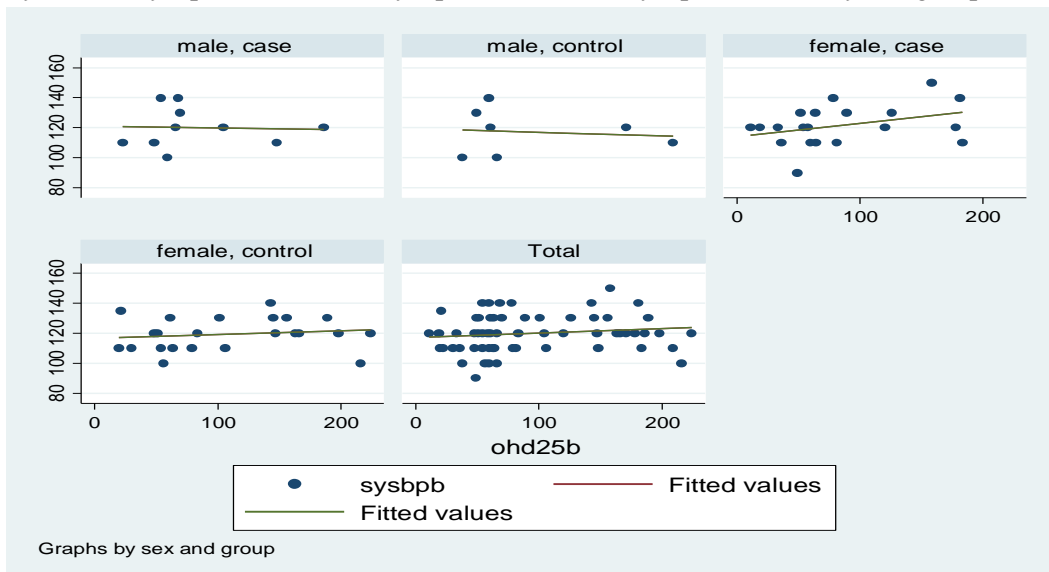
Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0  
Pr(T < t) = 0.1964 Pr(|T| > |t|) = 0.3929 Pr(T > t) = 0.8036

همانگونه که ملاحظه می شود در اینجا اختلافی بین سطح ویتامین دی قبل و بعد از مداخله نیست و نتیجه مورد انتظار هم هست. چون مداخله ای صورت نگرفت. در اینجا نیز رابطه سطح ویتامین دی با فشارخون سیستولیک در هر دو گروه مزبور بررسی می شود. ابتدا نمودار پراکنش دو متغیر به صورت زیر ترسیم می شود.



مجدد رابطه سطح ویتامین دی با فشارخون سیستولیک در هر دو گروه مزبور بر حسب جنسیت بررسی می شود. ابتدا نمودار پراکنش دو متغیر به صورت زیر ترسیم می شود.

`twoway (scatter sysbpb ohd25b) (lfit sysbpb ohd25b) (lfit sysbpb ohd25b), by(sex group, total)`



همبستگی بین دو متغیر سطح ویتامین دی و فشار خون در دو گروه در زیر آمده است.

by group, sort: pwcrr sysbpb ohd25b,sig  
-> group = control

	sysbpb	ohd25b
sysbpb	1.0000	
ohd25b	0.0898	1.0000
	0.6370	

-> group = case

	sysbpb	ohd25b
sysbpb	1.0000	
ohd25b	0.2380	1.0000
	0.2053	

### آزمون های پارامتری برای مقایسه بیشتر از دو گروه (K گروه) : آزمون های تحلیل واریانس

یکی از مهمترین فرضیه ها و هدف ها در اکثر پژوهش ها مقایسه تغییرات یک یا چند متغیر وابسته در بین یک یا چند گروه در قالب متغیر مستقل است. این آزمون ها عبارتند از :

- ۱- آزمون تحلیل واریانس یکطرفه و دو طرفه
- ۲- آزمون تحلیل واریانس یا کوواریانس تک متغیره (ANCOVA)
- ۳- آزمون تحلیل واریانس یا کوواریانس چند متغیره (MANOVA/ MANCOVA)

### آنالیز واریانس: ANOVA

#### آنالیز واریانس یک طرفه : One Way Analysis of Variance

آزمون t محدود به موقعیتی است که در آن تنها دو سطح از یک متغیر مستقل وجود دارد (مثلاً دو جنسیت: مرد و زن) و محقق قصد دارد دو میانگین را مقایسه نماید. اگر تعداد گروه های تجربی یا سطوح متغیر مستقل بیش از دو تا باشد آزمون t نامناسب خواهد بود. در این حالت آزمون مناسب آنالیز واریانس است. بنابراین این آزمون تعمیم یافته آزمون t است. که در آن واریانس کل جامعه به عوامل اولیه تجزیه می شود. همیشه نتیجه گیری در مورد میانگین های جامعه بر اساس پراکندگی میانگین های نمونه صورت می گیرد.

**پیش فرض ها :** مقیاس متغیر وابسته باید کمی و در سطح سنجش فاصله ای یا نسبی باشد. مقیاس متغیر مستقل باید کیفی و در سطح مقیاس اسمی یا ترتیبی باشد. توزیع داده های متغیر وابسته نرمال باشد. هر گروه باید یک نمونه تصادفی مستقل از جامعه آماری نرمال باشد. گروه ها از همدیگر مستقل باشند. واریانس گروه ها برابر باشد. نرم افزار Stata برابری واریانس ها را آزمون و گزارش می کند. در حجم نمونه های بالا واریانس مقاوم است.



آنالیز واریانس این مزیت را دارد که می‌تواند برای آنالیز موقعیت‌هایی به کار رود که در آن چند متغیر مستقل وجود دارند. در این موقعیت‌ها مانند آنالیز واریانس دو طرفه به ما می‌گوید که چگونه متغیرهای مستقل با یکدیگر تعامل یا برهمکنش (interaction) دارند و این تعاملات چه اثری روی متغیر وابسته دارند.

**یادآوری: فرض آزمون t:** میانگین دو نمونه با هم برابر است.

**فرض آزمون آنالیز واریانس:** میانگین همه گروه‌ها (سه گروه یا بیشتر) با هم برابر است.

بنابراین آنالیز واریانس این فرضیه را آزمون می‌کند که میانگین همه گروه‌ها با هم برابر هستند.

در آنالیز واریانس آماره F را داریم که مشابه آماره t است و این آماره مقدار واریانس سیستماتیک در داده‌ها را با مقدار واریانس غیر سیستماتیک unsystematic مقایسه می‌کند.

اگرچه آنالیز واریانس می‌تواند به ما بگوید که بین میانگین‌های گروه‌ها تفاوت وجود دارد یا حداقل یکی از میانگین‌ها با بقیه متفاوت است، اما نمی‌تواند بگوید که میانگین کدام گروه‌ها با هم تفاوت دارند. یعنی اگر نتیجه بگیریم که  $\bar{X}_1 = \bar{X}_2 = \bar{X}_3$  درست نیست این میانگین‌ها ممکن است به روش‌های زیادی با هم متفاوت باشند

الف: هر سه میانگین به هم تفاوت معنی‌دار داشته باشد ( $\bar{X}_1 \neq \bar{X}_2 \neq \bar{X}_3$ )

ب- میانگین گروه ۱ و ۲ یکسان هستند اما میانگین گروه ۳ با آنها تفاوت معنی‌داری دارد ( $\bar{X}_1 = \bar{X}_2 \neq \bar{X}_3$ )

ج- میانگین گروه ۲ و ۳ یکسان هستند اما میانگین گروه ۱ با آنها تفاوت معنی‌داری دارد ( $\bar{X}_1 \neq \bar{X}_2 = \bar{X}_3$ )

د- میانگین گروه ۱ و ۳ یکسان هستند اما میانگین گروه ۲ با آنها تفاوت معنی‌داری دارد ( $\bar{X}_1 = \bar{X}_3 \neq \bar{X}_2$ )

بنابراین آماره F تنها می‌تواند بگوید که مداخله ما اثر داشته یا بین سه گروه تفاوت وجود دارد یا ندارد. نمی‌تواند به ما بگوید که این اثر چه بوده یا این که کدام گروه‌ها با هم تفاوت معنی‌داری دارند. برای پیدا کردن محل اختلاف باید از آزمون‌های تعقیبی استفاده کرد.

آنالیز واریانس روش مقایسه نسبت واریانس سیستماتیک به واریانس غیر سیستماتیک در یک مطالعه تجربی است. نسبت این واریانس‌ها F-ratio می‌نامند. F-ratio نسبت واریانس توضیح داده شده توسط مدل را با خطا در مدل مقایسه می‌کند. اکنون تنها مدل برازش شده به داده‌ها با دستکاری تجربی سیستماتیک در بین گروه‌ها مرتبط است تا آنهایی که به طور طبیعی در متغیرها روی می‌دهند.

### گام‌ها و منطق انجام آنالیز واریانس

ساده‌ترین مدلی که می‌توان به یک مجموعه از داده‌ها برازش نمود میانگین کلی grand mean (میانگین متغیر پیامد) است. این مدل پایه بیان‌کننده عدم اثر یا عدم رابطه بین متغیر پیشگویی‌کننده و پیامد است.

می‌توان مدل متفاوتی را به داده‌های جمع‌آوری شده برازش نمود که فرضیات ما را بیان کند. اگر این مدل به داده‌ها خوب برازش گردد پس باید بهتر از استفاده از میانگین کلی باشد. گاهی اوقات یک مدل خطی (خط بهترین برازش) به داده‌ها برازش می‌گردد اما در پژوهش تجربی اغلب مدلی را بر اساس میانگین شرایط مختلف برازش می‌کنیم. به محض این که یک مدل به داده‌ها برازش شد می‌توان تغییرپذیری کل (total variability) یا  $SS_T$ ، خطا در مدل (SSR) یا تغییرپذیری بین گروهی (SSB) و بهبود به خاطر مدل (SSM)، تغییرپذیری درون گروهی (SSW) را محاسبه نمود

مدلی که برازش می‌کنیم بر مبنای میانگین‌های گروهی از شرایط تجربی مختلف است (زیرا اگر بدانیم کسی از گروه اول است می‌توانیم مقدار پیامد برای آن شخص را پیشگویی کنیم که میانگین گروه ۱ خواهد بود). اگر

میانگین های گروهی خیلی متفاوت باشند پس مدل در پیشگویی صحیح نمره فرد در مورد متغیر پیامد بهتر عمل خواهد کرد (زیرا گروه ها از یکدیگر به خوبی افتراق داده شده اند).

سپس می توانیم آزمون کنیم که مدل در مقایسه با میانگین کلی از طریق مقایسه واریانس توضیح داده شده توسط مدل (یعنی بهبود به خاطر استفاده از مدل برای پیشگویی پیامدها به جای استفاده از میانگین کلی) در مقابل خطا در مدل چقدر بهتر داده ها را برازش می کند. این شامل مقایسه  $SS_M$  و  $SS_R$  است. نتیجه F-ratio است.

### گام های انجام آنالیز واریانس :

#### گام اول: محاسبه مجموع مجذورات کلی (SST)

کل مقدار گوناگونی (variation) در داده ها را می توان با محاسبه تفاوت هر یک از داده ها از میانگین کلی بدست آورد سپس مجموع مجذور این تفاوتها به هم اضافه می شوند تا مجموع مجذورات کل یا SST را نتیجه بدهد.

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

$$s^2 = \frac{SS}{(N-1)}$$

توجه:

N: تعداد مشاهدات

$$SS = s^2(N - 1)$$

پس:

یعنی واریانس کلی گوناگونی ما بین همه نمرات است صرفنظر از وضعیت تجربی که نمرات از آن می آیند، واریانس کل مجذور فاصله بین نمره هر فرد و میانگین کل است. N-1 تعداد درجات آزادی مجموع مجذورات کل است درجه آزادی یکی کمتر از تعداد چیزهایی است که محاسبه می شوند.

#### گام دوم: محاسبه مجموع مجذورات مدل (SSM)

در گام اول اینکه مقدار کل گوناگونی در داده ها چقدر است محاسبه گردید. اکنون لازم است که بدانیم چقدر از این تغییر را مدل می تواند توجیه کند (توضیح دهد). در آنالیز واریانس مدل بر مبنای میانگین های گروهی از تجربه یا آزمایش ما است و بنابراین مدل مجموع های مجذورات به ما می گوید که چقدر از کل گوناگونی را می توان به وسیله این حقیقت توضیح داد که داده های مختلف از گروه های متفاوتی می آیند. به طور خلاصه مدل اثر دستکاری یا مداخله تجربی را بیان می کند. مدل مجموع مجذورات با گرفتن تفاوت بین مقادیر پیش بینی شده توسط مدل و میانگین کلی محاسبه می شود. در ANOVA مقادیر پیشگویی شده توسط مدل میانگین های گروهی هستند بنابراین برای هر فردی مقدار پیش بینی شده توسط مدل، میانگین گروهی است که آن فرد متعلق به آن است. بدین لحاظ بهبود به خاطر مدل، تفاوت بین میانگین کلی و میانگین هر گروه برای هر فرد خواهد بود. این تفاوت ها مجذور شده و با هم جمع می شوند. برای مدل مجموع مجذورات لازم است تفاوت بین مقدار پیشگویی شده هر فرد و میانگین کلی محاسبه شود سپس این تفاوت ها را مجذور کرده و با هم جمع می کنیم. می دانیم که مقدار پیش بینی شده برای هر فرد در گروه خاص میانگین آن گروه است. بنابراین ساده ترین روش برای محاسبه  $SS_M$  عبارتست از:

۱- محاسبه تفاوت بین میانگین هر گروه و میانگین کلی

۲- مجذور کردن این تفاوت ها

۳- ضرب کردن هر نتیجه در تعداد افراد گروه ( $n_i$ )

۴- اضافه کردن مقادیر هر گروه به هم

$$SS_M = \sum n_i (\bar{x}_i - \bar{x}_{grand})^2$$

درجه آزادی برای  $SS_M$  یکی کمتر از تعداد اقلام مورد استفاده برای محاسبه مجموع مجذورات است.

### گام سوم: محاسبه مجموع مجذورات باقیمانده ( $SS_R$ )

در گام های اول و دوم مشخص شد که چقدر واحد تغییر در داده های ما وجود دارد که مدل توانسته آن را توضیح دهد. مجموع مجذور نهایی مجموع مجذورات باقیمانده است ( $SS_R$ ) که به ما می گوید چه مقدار از تغییر نمی تواند توسط مدل توضیح داده شود. این عدد مقدار گوناگونی ناشی از عوامل خارجی مانند تفاوت های فردی در وزن، قد یا هر چیز دیگری است. ساده ترین کار برای محاسبه مجموع مجذورات باقیمانده وقتی که  $SS_M$  و  $SS_T$  را می دانیم این است که آنها را از هم کم کنیم ( $SS_R = SS_T - SS_M$ ). مجموع مجذورات باقیمانده تفاوت بین آن چیزی است که مدل پیشگویی می کند و آن چه که در واقعیت مشاهده شده است. برای یک فرد معین مدل میانگین گروهی را که فرد به آن تعلق دارد پیشگویی می کند. بنابراین  $SS_R$  با نگاه کردن به تفاوت های بین نمرات بدست آمده برای هر فرد و میانگین گروهی که به آن تعلق دارد بدست می آید. بنابراین مجموع مجذورات باقیمانده تفاوت های فردی بین افراد را در گروه تجربی بیان می کند. فاصله بین هر داده از میانگین گروهی را مجذور کرده و سپس با هم جمع می کنیم تا مجموع مجذورات باقیمانده بدست آید.

$$SS_R = \sum (x_i - \bar{x}_i)^2$$

or

$$SS_R = s_{group 1}^2 (n_1 - 1) + s_{group 2}^2 (n_2 - 1) + s_{group 3}^2 (n_3 - 1)$$

مجموع مجذورات برای هر گروه بیان کننده مجموع مجذور تفاوت های بین نمره هر فرد در آن گروه با میانگین گروهی است. بنا بر این مجموع مجذورات باقیمانده را می توان به صورت زیر نوشت:

$$SS_R = SS_{group 1} + SS_{group 2} + SS_{group 3}$$

درجه آزادی برای این قسمت برابر مجموع درجات آزادی برای هر یک از مجموع مجذورات خواهد بود.

### گام چهارم: محاسبه میانگین مجذورات

$SS_M$  به ما می گوید که چه مقدار از تغییرات (تفاوت ها) در مدل شرح داده می شود و  $SS_R$  به ما می گوید که چه مقدار از تغییرات یا تفاوت ها به خاطر عوامل خارجی است. به هر حال هر دوی این مقادیر مقادیر مجموع هستند، که به وسیله تعداد نمراتی جمع می شوند تحت تاثیر قرار می گیرد. برای حذف این تورش می توان متوسط جمع مجذورات را که میانگین مجذورات نامیده می شود ( $MS$ ) محاسبه می کنیم که به طور ساده مجموع مجذورات تقسیم بر تعداد درجات آزادی است.

$MS_M$  متوسط مقدار تغییر توصیف شده توسط مدل است (تغییر سیستماتیک)

$MS_R$  متوسط مقدار تغییر توصیف شده توسط متغیرهای خارجی است (تغییر غیرسیستماتیک)

### گام پنجم: محاسبه F-ratio

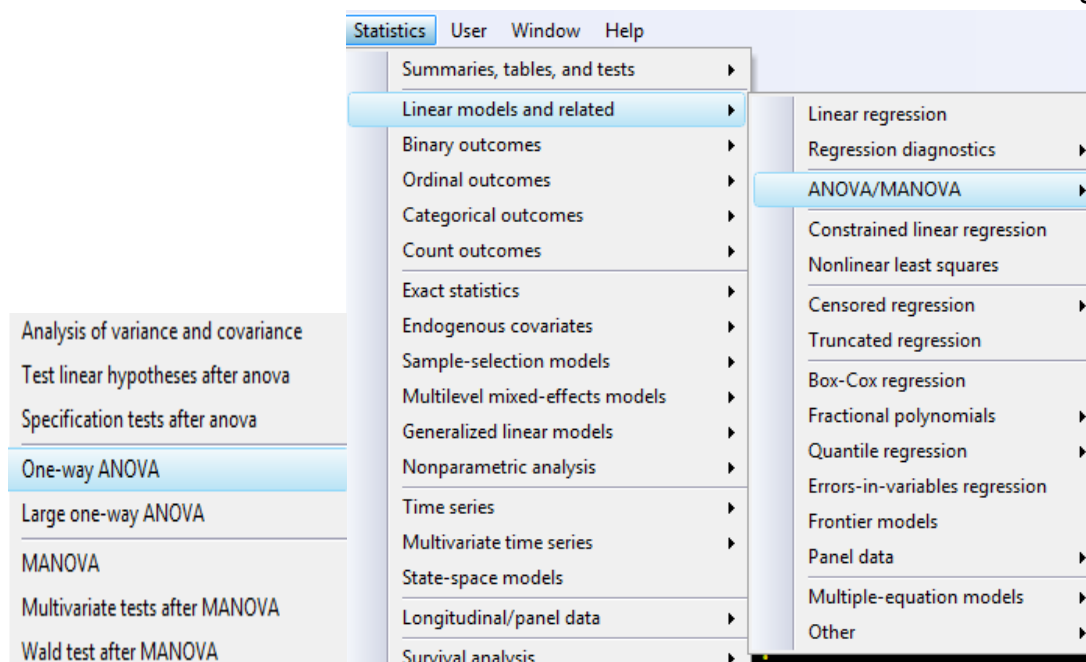
نسبت F اندازه نسبت تغییر توصیف شده توسط مدل و تغییر توصیف شده توسط عوامل غیر سیستماتیک است و از تقسیم میانگین مجزورات مدل به میانگین مجزورات باقیمانده بدست می آید.

$$F = \frac{MS_M}{MS_R}$$

در واقع F-ratio نسبت اثر تجربی به تفاوت های فردی یا نسبت واریانس سیستماتیک به واریانس غیر سیستماتیک یا نسبت واریانس بین گروه ها ناشی از متغیرهای مستقل به واریانس درون گروه ها ناشی از خطا است. اگر مقدار F کمتر از ۱ باشد به این معنی است که اثر معنی دار نیست زیرا در این حالت  $MS_R$  بزرگتر از  $MS_M$  است که به این معنی است که واریانس غیر سیستماتیک خیلی بیشتر از واریانس سیستماتیک است.

### کار عملی با نرم افزار و انجام تحلیل واریانس (ANOVA):

فایل داده های part five را باز نمایید. برای آزمون و مقایسه میانگین تری گلیسیرید قبل از مداخله (tgb) در بین طبقه اجتماعی افراد (socialclass) از این آزمون استفاده می شود. مسیر زیر انجام این آزمون با نرم افزار را نمایش می دهد.



یا با تایپ db oneway و انتخاب کادر محاوره ای زیر و تعریف متغیرهای مورد نظر تحلیل انجام می شود. همانطور که قبلا یادآوری شد آنالیز واریانس فقط می گوید که اختلاف وجود دارد یا نه. برای تعیین محل اختلاف در بین گروه ها باید آزمون تعقیبی انجام شود. آزمون های تعقیبی مورد نیاز در زیر گزینه Multiple-comparison test تعبیه شده اند که با انتخاب هر کدام محل اختلاف در بین گروه ها مشخص می گردد.

## مقایسه های چندگانه : Multiple- Comparison Test

آزمون بونفرونی (Bonferroni) : این آزمون بر اساس آماره  $t$  و با کنترل نرخ خطای کل، سطح معنی داری مشاهده شده را بر اساس واقعیتی که مقایسه های چندگانه انجام داده اند تعدیل می کند. برای آزمون تعداد جفت های کمتر از آن استفاده می شود.

آزمون سیداک (Sidak) : این آزمون مانند بونفرونی است و مرزهای نزدیکتری در مقایسه با آن ایجاد می کند. آزمون شفه (Scheffe) : اگر گروه ها حجم نابرابر داشته باشند، این آزمون مناسبترین است. اشکال عمده آن محافظه کار بودن آن است. این آزمون برای یکدسته از اطلاعات یکسان در مقایسه با آزمون توکی، فرض صفر را کمتر رد می کند. این آزمون به انحراف از پیش فرضهای نرمال بودن توزیع داده ها و همگونی واریانس ها حساس نیست و از مزیت آن به شمار می رود. آزمون های تعقیبی دیگری نیز وجود دارد منتها در بخش مقایسه های چندگانه بعد از آنالیز واریانس در نرم افزار Stata، سه آزمون بالا آمده است که از آنها می توان استفاده نمود. کادر محاوره ای زیر آزمون های تعقیبی و نحوه معرفی متغیرهای پاسخ و مستقل برای انجام تحلیل واریانس یکطرفه را نشان می دهد.

The screenshot shows the 'oneway - One-way analysis of variance' dialog box in Stata. It is divided into several sections:

- Main:** 'Response variable:' is set to 'tgb' and 'Factor variable:' is set to 'socialclass'.
- Multiple-comparison tests:** 'Bonferroni' and 'Scheffe' are checked, while 'Sidak' is unchecked.
- Output:** A group of options for suppressing output is shown, all of which are unchecked: 'Produce summary table', 'Suppress means', 'Suppress standard deviations', 'Suppress frequencies', 'Suppress number of obs.', 'Suppress the ANOVA table', 'Show numeric codes, not labels', and 'Do not break wide tables'.
- Other:** 'Treat missing values as categories' is unchecked.
- Buttons:** 'OK', 'Cancel', and 'Submit' are located at the bottom right.

نتایج در زیر آمده است.

```
. oneway tgb socialclass, bonferroni
```

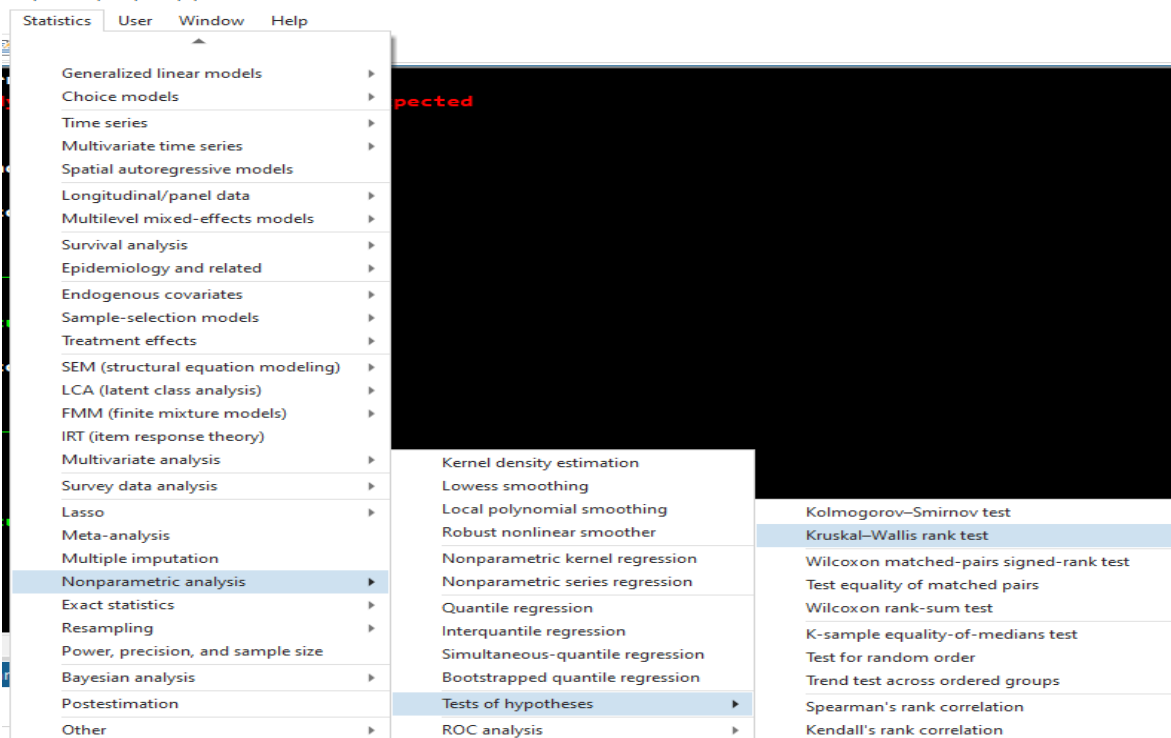
Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	28621.8375	2	14310.9187	0.55	0.5809
Within groups	1487298.35	57	26092.9534		
Total	1515920.18	59	25693.5624		

Bartlett's test for equal variances:  $\chi^2(2) = 34.7415$  Prob> $\chi^2 = 0.000$

نتایج بالا نشان می دهد که اختلافی بین تری گلیسیرید طبقات اجتماعی با  $F=0.55$  و مقدار معنی داری  $P=0.5809$  وجود ندارد. بنابراین نیازی به انجام آزمون تعقیبی نیست. در صورتی که پس از انجام تحلیل واریانس، مقدار پی در خروجی جدول معنی دار بود نشان از تفاوت بین میانگین گروه ها است. بنابراین باید با انتخاب آزمون تعقیبی مورد نظر، محل اختلاف در بین گروه ها را مشخص نمود.

### آزمون های ناپارامتریک:

پیش فرض های انجام آزمون برای مقایسه میانگین ها در بین دو گروه و بیشتر (مستقل و بهم وابسته) قبلا توضیح داده شد. چنانچه این پیش فرض ها برقرار نباشند از آزمون های ناپارامتریک استفاده می شود. معادل آزمون ناپارامتری برای مقایسه میانگین در دو گروه وابسته به هم آزمون من ویتنی است. معادل آزمون ناپارامتری برای مقایسه میانگین در دو گروه وابسته به هم آزمون ویلکاکسون است. معادل ناپارامتری آزمون آنالیز واریانس برای مقایسه میانگین، آزمون کروسکال والیس است. که از مسیر زیر قابل استفاده است.



در فایل part five : متغیر سطح ویتامین دی با اجرای دستور ارزیابی توزیع داد ه ها به صورت زیر  
by group, sort : swilk ohd25b

توزیع نرمال ندارد و حجم نمونه نیز برای هر گروه ۳۰ نفر و نسبتا کم است. بنابراین برای مقایسه میانگین آن در بین طبقات اجتماعی از آزمون کروسکال والیس به شکل زیر استفاده می شود.

```
. kwallis ohd25b, by(socialclass)
```

Kruskal-Wallis equality-of-populations rank test

social~s	Obs	Rank Sum
1	20	647.50
2	24	680.50
3	16	502.00

```
chi-squared = 0.633 with 2 d.f.  
probability = 0.7287
```

```
chi-squared with ties = 0.633 with 2 d.f.  
probability = 0.7286
```

نتایج این آزمون نشان می دهد که اختلافی بین میانگین سطح ویتامین دی در طبقات اجتماعی وجود ندارد.

```
chi-squared = 0.633 with 2 d.f.  
probability = 0.7287
```

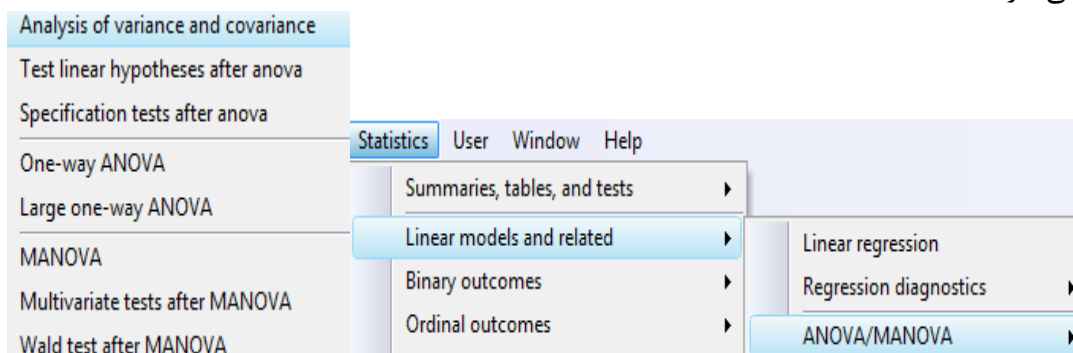
### سایر دستورات برای آزمون های ناپارامتریک:

- دستور انجام آزمون میانه یا آزمون رتبه ای علامت ویلکاکسون در یک گروه  
signrank Vname= مقدار فرضیه
- دستور انجام آزمون من ویتنی، معادل آزمون تی مستقل پارامتریک است.  
ranksum Vname, by(Variable)
- دستور انجام آزمون کروسکال والیس، معادل آزمون آنالیز واریانس پارامتریک است.  
kwallis Vname, by(Variable)
- دستور انجام آزمون ویلکاکسون، معادل آزمون تی زوجی است.  
signrank Variable1 = Variable2
- دستور انجام آزمون مک نمار (McNemar) برای شرایط پارامتریک و مطالعه مورد شاهد همسان شده  
mcci a b c d
- دستور انجام آزمون فریدمن: معادل آنالیز واریانس با اندازه های تکراری (درون گروهی است) که از آن برای مقایسه میانگین رتبه ها در بین k متغیر (گروه)
- دستور انجام همبستگی اسپیرمن  
friedman Variable 1 – Variable n  
spearman var1 var2

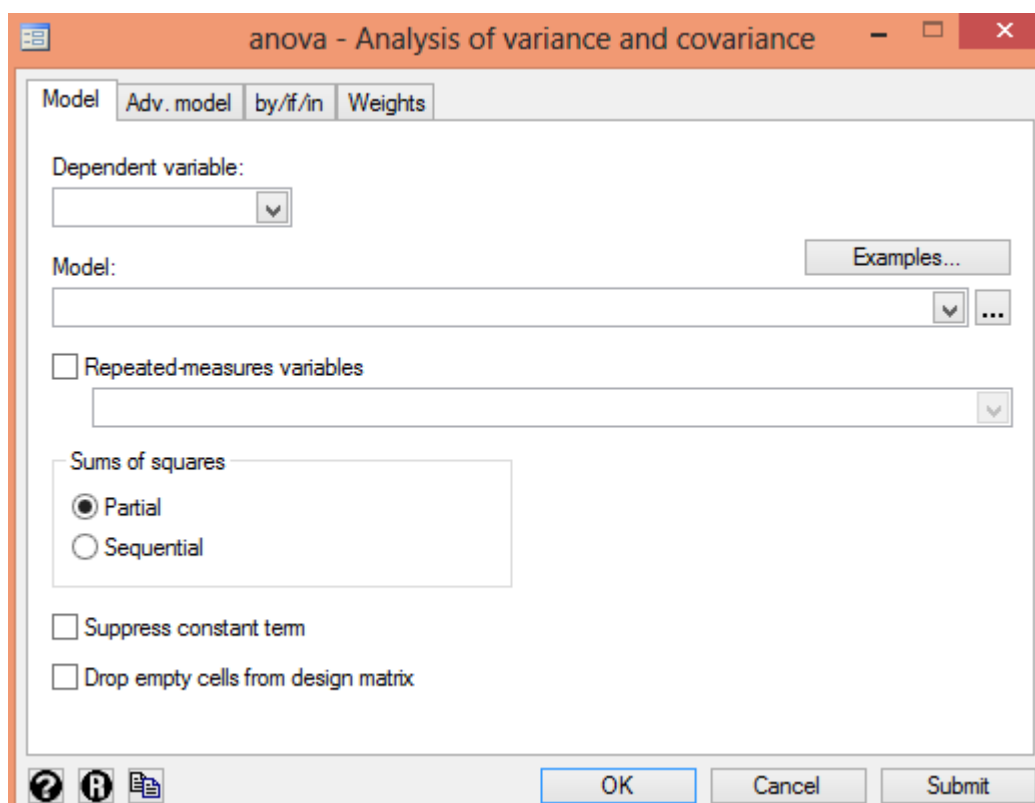
## Two Way ANOVA

## آنالیز واریانس دو طرفه

هنگامی که تعداد متغیر مستقل کیفی بیش از یکی باشد، (وقتی دو متغیر مستقل باشد) از آنالیز واریانس دو طرفه و اگر بیشتر از دو متغیر باشد از آنالیز واریانس چند طرفه استفاده می شود. برای انجام این تحلیل از مسیر زیر اقدام می شود.



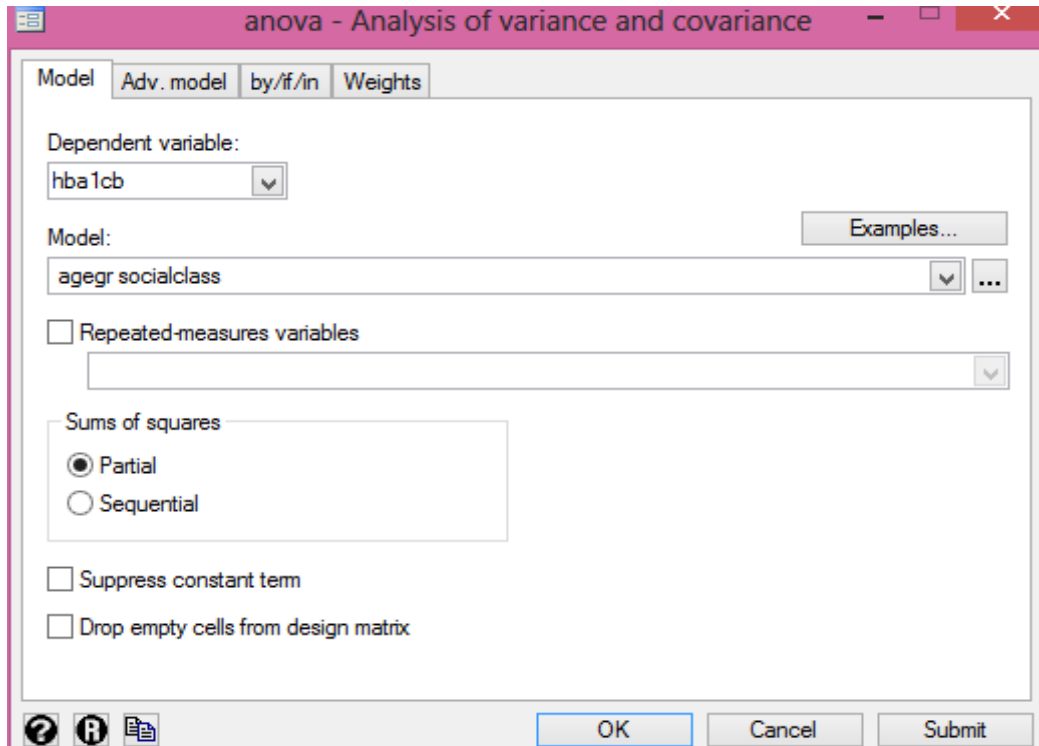
یا می توان با دستور `db anova` کادر محاوره ای مربوطه را باز نمود. از کادر محاوره ای آنالیز واریانس دو طرفه، می توان برای آنالیز واریانس یکطرفه نیز استفاده نمود اما عکس این استفاده مقدور نیست. زیرا در گزینه متغیر مستقل در کادر آنالیز واریانس یکطرفه، فقط یک متغیر می توان تعریف و معرفی نمود. کادر مزبور در زیر آمده است.



به عنوان مثال در فایل `part five` برای مقایسه میزان هموگلوبین گلیکوزیله در بین متغیرهای وضعیت طبقه



اجتماعی و گروه های سنی از آنالیز واریانس دو طرفه استفاده می شود. ابتدا متغیرهای مورد آزمون همانند زیر وارد کادر محاوره ای می شوند.



نتایج به صورت زیر نمایش داده می شود.

```
. anova hba1cb agegr socialclass
```

Source	Partial SS	df	MS	F	Prob > F
Model	2.78605462	5	.557210923	2.50	0.0419
agegr	1.6971378	3	.5657126	2.53	0.0665
socialclass	.848397787	2	.424198893	1.90	0.1595
Residual	12.0579437	54	.223295254		
Total	14.8439984	59	.251593192		

همانگونه که ملاحظه می شود در شرایط یکسان سنی (از نظر متغیر گروه های سنی) میانگین متغیر هموگلوبین گلیکوزیله در طبقات اجتماعی یکسان است ( $p=0.1595$ ). همچنین در شرایط طبقه اجتماعی یکسان، میانگین متغیر هموگلوبین گلیکوزیله در گروه های سنی نیز یکسان است ( $p=0.06$ ). این آنالیز وقتی بصورت یکطرفه و در سطح گروه های سنی تحلیل می شود معنا دار است. وقتی طبقه اجتماعی نیز وارد می شود اثر گروه سنی خنثی و غیر معنی دار می شود.

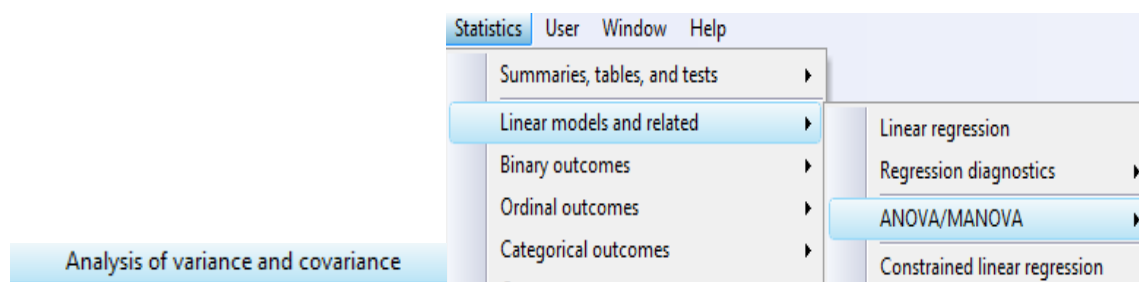
## آزمون تحلیل واریانس یا کوواریانس تک متغیره (ANCOVA)

این آزمون تعمیم یافته آزمون آنالیز واریانس یکطرفه است. در این آزمون بیش از یک متغیر مستقل وارد مدل شده و محقق قصد دارد ببیند کدام متغیر مستقل بیشترین اثر را بر متغیر وابسته دارد. این روش بر پایه روش مدل عمومی خطی است که در آن فرض بر این است که عامل ها (Factors) و کوواریت ها (Covariates) یک رابطه خطی با متغیر وابسته دارند. همانند آنالیز واریانس دو طرفه است منتها در اینجا یکی از متغیرهای مستقل حتما باید کمی باشد. در واقع م میتوان با این آزمون نقش مخدوش کننده یک متغیر کمی را کنترل نمود. هدف در انجام این آزمون نه تنها تاثیر جداگانه دو یا چند متغیر مستقل کیفی بر یک متغیر وابسته کمی بررسی می شود، بلکه آزمون تاثیر همزمان این دو متغیر بر متغیر وابسته نیز مد نظر است. در اینجا هر متغیر مستقل دو نوع اثر بر متغیر وابسته دارد:

۱- اثر اصلی (Main Effect) یا مستقیم : اثری که یک متغیر مستقل به تنهایی بر متغیر وابسته دارد. اینترکشن سایر متغیرها در نظر گرفته نمی شود.

۲- اثر متقابل یا تعاملی (Interaction) : اثری که یک متغیر مستقل در ترکیب با سایر متغیرهای مستقل دارد. بعبارت دیگر در اینجا نقش متغیر مستقل کمی که وارد مدل شده نیز کنترل می شود. یعنی اینترکشن متغیرهای مستقل و وابسته را با در نظر گرفتن متغیر ثالث مورد بررسی قرار می گیرد.

برای استفاده از این آزمون در نرم افزار، مسیر زیر قابل اجرا است.



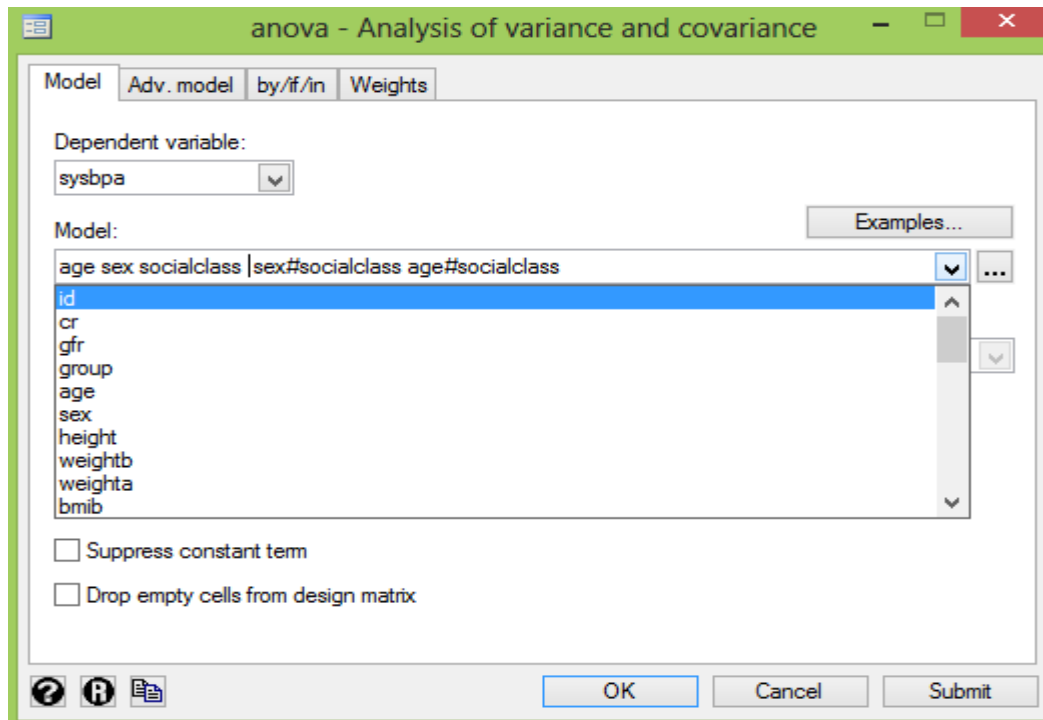
یا با دستور db anova می توان آن را اجرا نمود.

در فایل part five برای بررسی اثر طبقه اجتماعی، جنسیت و سن بر فشار خون سیستمولیک (متغیر وابسته) از این آزمون استفاده می شود. در اینجا اینترکشن جنسیت با طبقه اجتماعی و اینترکشن سن با طبقه اجتماعی نیز مورد بررسی قرار می گیرد.

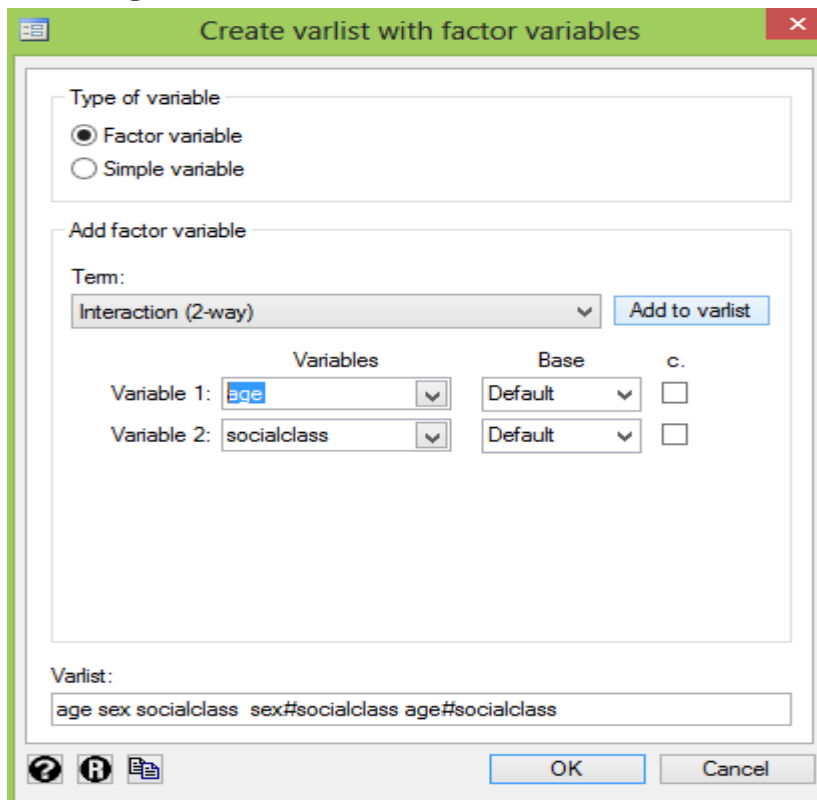
دستور مورد نیاز برای انجام این کار در زیر آمده است.

```
anova sysbpa age sex socialclass agegroup sex#socialclass age#socialclass
```

در کادر محاوره ای متغیرها به صورت زیر وارد می شوند.



برای وارد نمودن اینترکشن متغیرها از زبانه کنار گزینه Model در پایین Examples استفاده می شود. این عمل در شکل زیر نمایش داده می شود. لیست متغیرهای مدل پس از Add کردن در قسمت Varlist نشان داده میشود. این لیست پس از ok کردن به کادر محاوره ای در قسمت Model منتقل می شود.



خروجی نرم افزار به شکل زیر نمایش داده می شود.

```
. anova sysbpa age sex socialclass sex#socialclass age#socialclass
```

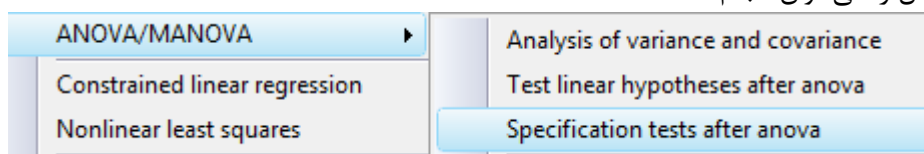
```
Number of obs =      60      R-squared      = 0.9143
Root MSE      = 8.56602      Adj R-squared = 0.5406
```

Source	Partial SS	df	MS	F	Prob > F
Model	8616.19048	48	179.503968	2.45	0.0547
age	5056.47162	30	168.549054	2.30	0.0728
sex	415.704825	1	415.704825	5.67	0.0365
socialclass	526.400772	2	263.200386	3.59	0.0632
sex#socialclass	1476.19048	2	738.095238	10.06	0.0033
age#socialclass	1807.63485	13	139.048835	1.90	0.1477
Residual	807.142857	11	73.3766234		
Total	9423.33333	59	159.717514		

همانگونه که ملاحظه میگردد اثر اصلی هر کدام از متغیرهای وارد شده به مدل به تفکیک نشان داده شده است. اثر sex با مقدار  $p=0.0365$  معنا دار است. اثر متغیر سن معنا دار نشده است. در بررسی تاثیر همزمان جنسیت و طبقه اجتماعی بر متغیر فشارخون مشخص است که جنسیت با طبقه اجتماعی نیز اینترکشن معناداری دارد. عبارتی فشارخون مردان و زنان در طبقات اجتماعی مختلف معنی دار بوده و با هم متفاوت است.

### ارزیابی مدل ANCOVA :

مستقل بودن مشاهدات، داشتن رابطه خطی بین متغیرهای مستقل و وابسته، توزیع نرمال متغیر وابسته، یکسان بودن واریانس متغیر وابسته برای کلیه متغیرهای مستقل، عدم رابطه خطی قوی بین متغیرهای مستقل و رابطه خطی بین متغیر همراه (Covariate) و متغیر وابسته باید بررسی گردد. اینترکشن بین متغیر مستقل اصلی و متغیر همراه وجود نداشته باشد. یا به عبارتی جهت و قدرت رابطه بین متغیر همراه و متغیر وابسته (یکسان بودن شیب رگرسیون) در هر یک از گروه ها مشابه باشد. با استفاده از دستورات زیر یا از طریق منو و انتخاب گزینه های زیر ارزیابی مدل را می توان انجام داد.



در بالای جدول خروجی نرم افزار مقدار RSquared خام و تعدیل شده نشان داده شده است. مقدار این ضریب نشان می دهد که متغیرهای وارد شده به مدل توانسته اند ۵۴ درصد از واریانس متغیر وابسته فشارخون سیستمولیک را تبیین کنند. حدود ۴۶ درصد دیگر از واریانس مربوط به متغیرهایی است که در مدل وارد نشده یا مورد بررسی قرار نگرفتند.

بعد از انجام تحلیل، آزمون برای فرض هم خطی بین متغیرها با فرض صفر ثابت بودن واریانس ها انجام میشود.

بعبارتی در مدل واریانس توزیع متغیر وابسته باید برای تمام مقادیر متغیر مستقل یکسان باشد. این فرض با دستور زیر انجام و نتیجه آن در زیر آمده است. خروجی نشان می دهد که فرض برقرار است.  $P=0.3669$

estat hottest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of sysbpa

chi2(1) = 0.81

Prob > chi2 = 0.3669

فرض خطی بودن روابط بین متغیرهای مستقل و وابسته نیز ارزیابی می شود که برقرار نیست. بنابراین در بررسی مجدد روابط بین متغیرها باید تجدید نظر گردد.

. test age sex socialclass# sex age# socialclass

Source	Partial SS	df	MS	F	Prob > F
age sex sex#socialclass					
age#socialclass	8505.56548	46	184.903597	2.52	0.0497
Residual	807.142857	11	73.3766234		

برای بررسی و اندازه گیری قدرت رابطه خطی بین متغیرهای مستقل از دستور estat vif استفاده می شود. این دستور برای محاسبه شاخص تولرانس (Tolerance) استفاده می شود. این شاخص سهم واریانس تبیین نشده متغیرهای مستقل (1-R Squared) را اندازه گیری می کند. که معمولاً مقادیر بالاتر از ۱۰ برای vif مناسب نیست. در اینجا میانگین این شاخص حدود ۱۱ گزارش شده است.

Mean VIF	11.12
----------	-------

برای وارد نمودن اینترکشن بین متغیرها و ساختن مدل مناسب از شکل زیر نیز می توان کمک گرفت.

Model Description

```

a          one factor
a b        two factors
a b a#b    two factors plus interaction
a##b      two factors plus interaction
a b c      three factors
a b c a#b a#c b#c  three factors plus two-way interactions
a b c a#b a#c b#c a#b#c  three factors plus all interactions
a##b##c   three factors plus all interactions
    
```

## آزمون تحلیل واریانس یا کوواریانس چند متغیره (MANOVA/ MANCOVA)

آزمون MANOVA و MANCOVA به دستورهای GLM Multivariate معروف هستند. به عبارتی MANOVA نوعی تحلیل رگرسیون و تحلیل واریانس از چندین متغیر وابسته بر اساس یک یا چند متغیر عامل و کووریت است. فرض مورد آزمون این است که آیا اثرات متغیرهای عامل بر میانگین های گروه بندی های متفاوت از یک توزیع مشترک متغیرهای وابسته متفاوت است یا خیر؟ یا به عبارت دیگر وقتی هدف مقایسه میانگین یک متغیر عددی در زیر گروه های مختلفی باشد که توسط چند متغیر ایجاد گردد سراغ آنالیز واریانس چند متغیره می رویم.

در این روش تفاوت در میانگین های طبقات مختلف متغیرهای مستقل بر حسب متغیرهای وابسته فاصله ای آزمون می شود. علاوه بر بررسی و تعیین اثرات اصلی و جداگانه هر عامل، تعاملات و اینترکشن بین عامل ها و بر متغیرهای وابسته را بررسی و آزمون می شوند.

متغیرهای مستقل می تواند کیفی ( اسمی / ترتیبی ) و کمی ( فاصله ای / نسبتی ) باشند. از متغیرهای کووریت کمی (بمعنای متغیرهای کنترل برای متغیرهای مستقل استفاده می شود. کاربرد این آزمون عبارتست از:

۱- مقایسه گروه هایی که توسط متغیرهای مستقل و بر حسب تفاوت های گروهی در مجموعه ای از متغیرهای وابسته فاصله ای تشکیل میشوند.

۲- استفاده از عدم تفاوت میانگین های طبقات مختلف متغیرهای مستقل در متغیرهای وابسته به عنوان معیاری برای کاهش تعداد متغیرهای مستقل به مجموعه ای کوچک تر با هدف مدل سازی بهتر.

۳- شناسایی متغیرهای مستقلی که نقش بیشتری در تفکیک مجموعه متغیرهای وابسته از همدیگر دارند.

## پیش فرض ها:

کلید پیش فرض ها برای انجام آزمون پارامتریک یا کلاسیک که در آزمون های بالا توضیح داده شد باید وجود داشته باشند. متغیرهای وابسته باید در سطح کمی باشند. حداقل دو متغیر وابسته ( فاصله ای / نسبتی ) باید باشند. متغیرهای مستقل می توانند هم کمی ( فاصله ای / نسبتی ) و هم کیفی ( اسمی / ترتیبی ) باشند. متغیرهای مستقل عامل باید مقیاس کیفی داشته باشند. متغیرهای مستقل کووریت باید در مقیاس کمی باشند. عامل ها و کووریت ها باید رابطه خطی با متغیر وابسته داشته باشند. یکسان بودن واریانس ها و کوواریانس ها، همخطی بین متغیرهای وابسته و یکسان بودن ارتباط بین متغیر وابسته و کووریت ها در گروه های مختلف یکسان باشد. چون بین متغیرهای وابسته همبستگی وجود دارد و ترکیب خطی از متغیرهای مستقل درست می گردد و پراکندگی آنها با هم هم پوشانی دارد نباید از ANOVA جداگانه برای هر متغیر وابسته و متغیرهای مستقل برای تحلیل استفاده نمود. در صورت استفاده خطای نوع یک رخ می دهد. نحوه اندازه گیری خطای نوع یک در زیر آمده است.  $a$  سطح معنی داری (  $0/05$  ) و  $t$  تعداد مقایسه های چندگانه است.

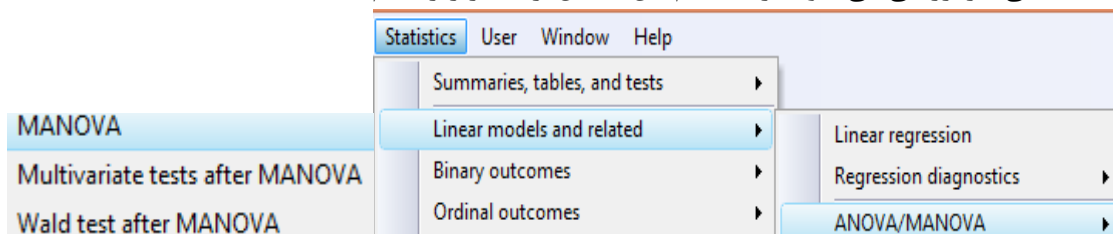
$$1 - (1 - a)^t$$

برای پیشگیری از این خطا باید از MANOVA استفاده نمود که توان بالاتری نسبت به ANOVA دارد. اولین گام در ارزیابی و انجام تحلیل، بررسی متغیرهای وابسته و تفاوت های آنها است. همانند استفاده از آزمون F در آنالیز واریانس، اما در اینجا از چهار شاخص دیگر استفاده می شود که در مباحث آنالیز به آنها پرداخته می شود.

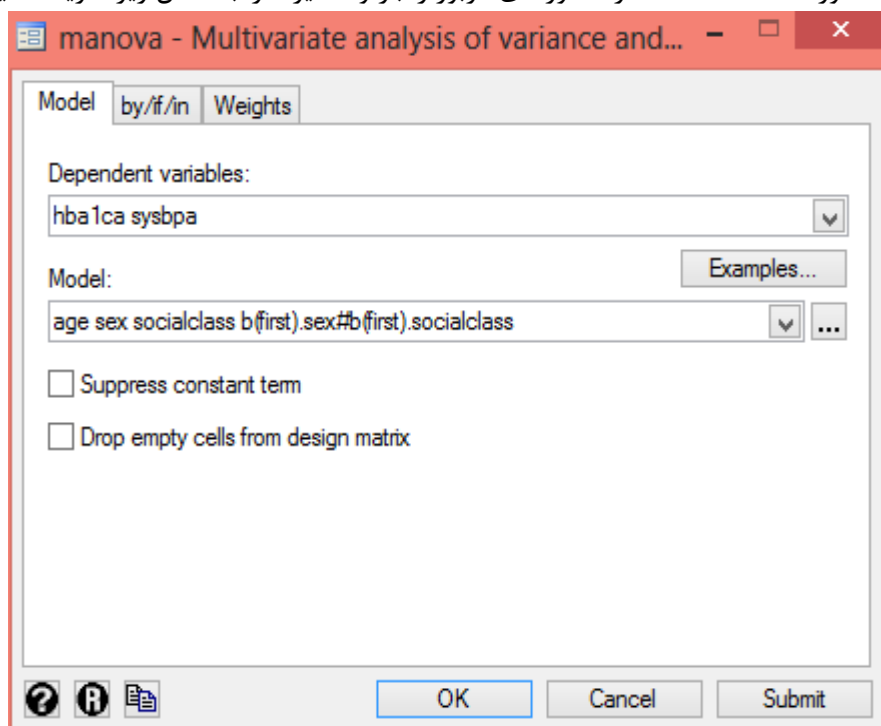
## MANOVA

## کار با نرم افزار و تحلیل

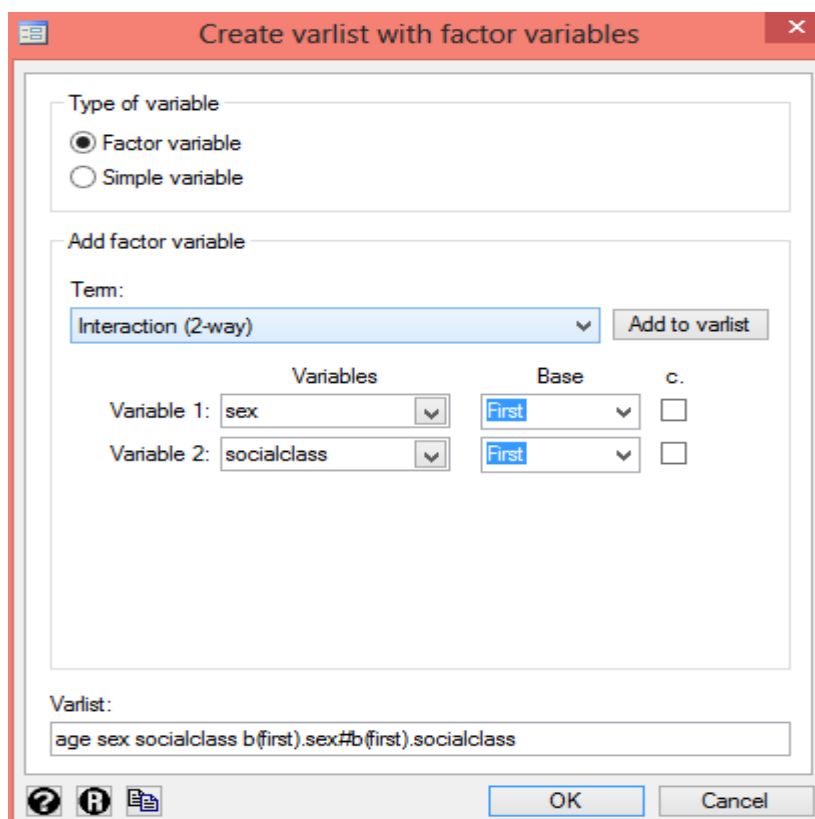
فایل part five را باز نمایید. متغیر های وابسته فشار خون سیستولیک (sysbpa) و هموگلوبین گلیکوزیله (hba1ca) می باشند. متغیرهای مستقل سن (age)، طبقه اجتماعی (socialclass) هستند. اینترکشن جنسیت با طبقه اجتماعی نیز ارزیابی می شود. برای انجام این تحلیل از مسیر زیر اقدام کنید



یا با استفاده از دستور db manova کادر محاوره ای مزبور را باز و متغیرها را به شکل زیر تعریف کنید.



برای تعریف متغیرهایی که با هم اینترکشن دارند همانند چگونگی معرفی اینترکشن در استفاده از مدل ANCOVA که قبلا توضیح داده شد استفاده می شود. از کادر محاوره ای بالا در سمت راست گزینه Model سربرگ معرفی اینترکشن بین متغیرها باز و متغیرها به شکل زیر تعریف می شوند.



خروجی نرم افزار به شکل زیر نمایش داده می شود. در جدول خروجی زیر نتیجه آزمون های چند متغیره بر اساس چهار آزمون، جهت معنی داری اثر هر متغیر مستقل در مدل، بصورت دقیق، تقریبی و بالای مقدار F گزارش شده است.

از میان چهار آزمون زیر، **لانداى ويلکز W (Wilks lambda)** معروفیت بیشتری نسبت به سایر آزمون ها داشته و با معیار نسبت احتمال مرتبط است. مقدار بین صفر تا یک دارد. مقادیر نزدیک به صفر نشان می دهند که میانگین های گروه ها از همدیگر متفاوت است. مقادیر نزدیک به ۱ نیز نشان از عدم تفاوت میانگین ها دارند. مقدار برابر با ۱ نشان می دهد که همه میانگین ها یکی هستند. اما در عمل آزمون اثر پیلای P (Pillais trace) قدرتمند تر از سایر آزمون هاست.

آزمون اثر هتلینگ L (LowLay Hotelling s trace) مقدار آن بر حسب مجموع مقادیر و ارزش های ویژه محاسبه میشود. چهارمین آزمون، بزرگترین ریشه روی (Roys largest root) نام دارد. برای تفسیر معنی داری این چهار آزمون باید از مقدار p استفاده نمود. هر گاه سطح معنی داری کمتر از 0.05 بود بر آن اساس می توان نتیجه گرفت که اثر مورد نظر معنی دار است. و در مدل نقش دارد. بر اساس جدول زیر مقدار p برای هیچ یک اثرات اصلی معنی دار نشده است. تنها برای اینترکشن جنسیت و طبقه اجتماعی مقدار آزمون Hotelling s LowLay trace معنی دار شده است و نشان از اثر اینترکشن دارد. بعبارتی تاثیر همزمان این دو متغیر بر متغیرهای وابسته است.



```
. manova hbalca sysbpa = age sex socialclass b(first).sex#b(first).socialclass
```

Number of obs = 60

W = Wilks' lambda      L = Lawley-Hotelling trace  
P = Pillai's trace      R = Roy's largest root

Source	Statistic	df	F(df1, df2) =	F	Prob>F	
Model	W	0.0995	35	70.0	46.0	1.43 0.1005 e
	P	1.3586		70.0	48.0	1.45 0.0860 a
	L	4.4460		70.0	44.0	1.40 0.1179 a
	R	2.8047		35.0	24.0	1.92 0.0486 u
Residual		24				
age	W	0.1434	30	60.0	46.0	1.26 0.2098 e
	P	1.2389		60.0	48.0	1.30 0.1727 a
	L	3.3082		60.0	44.0	1.21 0.2526 a
	R	1.9188		30.0	24.0	1.54 0.1425 u
sex	W	0.9347	1	2.0	23.0	0.80 0.4601 e
	P	0.0653		2.0	23.0	0.80 0.4601 e
	L	0.0698		2.0	23.0	0.80 0.4601 e
	R	0.0698		2.0	23.0	0.80 0.4601 e
socialclass	W	0.9081	2	4.0	46.0	0.57 0.6873 e
	P	0.0924		4.0	48.0	0.58 0.6779 a
	L	0.1006		4.0	44.0	0.55 0.6975 a
	R	0.0950		2.0	24.0	1.14 0.3366 u
sex#socialclass	W	0.7561	2	4.0	46.0	1.73 0.1607 e
	P	0.2462		4.0	48.0	1.68 0.1690 a
	L	0.3194		4.0	44.0	1.76 0.1547 a
	R	0.3095		2.0	24.0	3.71 0.0393 u
Residual		24				
Total		59				

e = exact, a = approximate, u = upper bound on F

### کار با نرم افزار و انجام تحلیل کامل با One-Way Multivariate Analysis of Variance :

وقتی متغیر پاسخ یا وابسته به صورت پیوسته (کمی) بوده و مشتمل بر دو متغیر یا بیشتر بوده و متغیرهای مستقل نیز در سطح کیفی باشند از این نوع آنالیز استفاده می شود. داده های فایل `manova difficult` را باز نمایید. در این فایل ۳۳ نفر مشاهده برای ۳ متغیر پاسخ وجود دارند. متغیرهای پاسخ عبارتند از `useful`، `importance` و `difficulty`. گروه مورد مطالعه سه سطح دارد. یک سطح بعنوان مداخله درمانی و دو سطح به

عنوان کنترل. سطح یک در متغیر گروه درمان (treatment group) را می گیرد. سطح ۲ در متغیر گروه ، کنترل برای سطح یک متغیر درمان است. سطح سه گروه نیز کنترل برای سطح دو متغیر گروه است. جدول زیر توصیف این داده ها را نشان می دهد.

. summarize difficulty useful importance

Variable	Obs	Mean	Std. Dev.	Min	Max
difficulty	33	5.715152	2.017598	2.4	10.25
useful	33	16.3303	3.292461	11.9	24.3
importance	33	6.475758	3.985131	.2	18.8

برای متغیر گروه نیز سطوح مورد نظر به صورت زیر تعریف شده اند.

. tabulate group

group	Freq.	Percent	Cum.
treatment	11	33.33	33.33
control_1	11	33.33	66.67
control_2	11	33.33	100.00
Total	33	100.00	

میانگین متغیرهای پاسخ بر حسب سه سطح از گروه و کل مشاهدات در زیر آمده است.

. tabstat difficulty useful importance, by(group)

Summary statistics: mean  
by categories of: group

group	diffic~y	useful	import~e
treatment	6.190909	18.11818	8.681818
control_1	5.581818	15.52727	5.109091
control_2	5.372727	15.34545	5.636364
Total	5.715152	16.3303	6.475758

ماتریس مقادیر ضریب همبستگی بین متغیرها و مقدار معنی داری آنها در زیر آمده است.

```
. pwcorr useful difficulty importance, sig
```

	useful	difficulty	importance
useful	1.0000		
difficulty	0.0978	1.0000	
importance	-0.3411	0.1978	1.0000
	0.0520	0.2698	

### برای انجام تحلیل بر روی داده های بالا چند نوع تحلیل می توان انجام داد:

- با توجه به اینکه سه متغیر وابسته کمی و یک متغیر کیفی مستقل وجود دارد، بهترین نوع تحلیل استفاده از MANOVA است.

- انجام Discriminant Function Analysis که نتایجی مشابه تحلیل بالا می دهد.

- داده ها تغییر فرمت داده شوند و به صورت Long تعریف و با استفاده Multilevel model تحلیل شوند. انواع فرمت داده ها و نحوه تبدیل آنها به یکدیگر در ذیل توضیحات آنالیز واریانس با داده های تکراری آمده است.

- انجام آنالیز های جداگانه ANOVA برای هر متغیر پاسخ. در این صورت توان آزمون پایین می آید و احتمال خطا بالا می رود. بنابراین تحلیل MANOVA در اینجا مطلوب بوده و این نوع تحلیل به شرح زیر انجام می شود.

### انجام تحلیل MANOVA:

خروجی نرم افزار نتایج این تحلیل را به صورت زیر نشان می دهد.

```
. manova difficulty useful importance = group
```

```
Number of obs = 33
```

```
W = Wilks' lambda      L = Lawley-Hotelling trace
```

```
P = Pillai's trace     R = Roy's largest root
```

Source	Statistic	df	F(df1, df2) =	F	Prob>F	
group	W	0.5258	2	6.0	56.0	3.54 0.0049 e
	P	0.4767		6.0	58.0	3.02 0.0122 a
	L	0.8972		6.0	54.0	4.04 0.0021 a
	R	0.8920		3.0	29.0	8.62 0.0003 u
Residual		30				
Total		32				

e = exact, a = approximate, u = upper bound on F

نتایج نشان می دهد که به بطور کلی نتایج معنی دار است. میانگین اندازه گیری شده بین سطوح مختلف گروه تفاوت دارند. توضیحات مربوط به چهار نوع آزمون که در جدول بالا به اختصار با حروف W و P و L و R نشان

داده شده است در مطالب گذشته آمده است. برای پیدا نمودن محل اختلاف بین گروه ها آزمون های تعقیبی را باید انجام داد.

برای این کار از فرمان زیر استفاده و گروه ها به شرح زیر مرتب می شوند.

```
. manovatest, showorder
```

Order of columns in the design matrix

1: (group==1)

2: (group==2)

3: (group==3)

4: \_cons

برای مقایسه گروه درمان با گروه کنترل، گروه یک با میانگین دو گروه دیگر مقایسه می گردد. فرضیه مورد آزمون در اینجا این است که میانگین گروه کنترل (گروه ۲ و ۳) با میانگین گروه درمان (گروه ۱) برابر است. خروجی بالا نشان می دهد که ۴ عنصر اصلی در ماتریکس مورد نظر وجود دارد. که آنها را باید تنظیم نمود. برای این کار (مقایسه گروه ها) مجدد باید ماتریسی درست نمود با نام C1 و آزمون برای آنها انجام داد. تنظیمات این کار با دستورات زیر و به صورت زیر انجام می شود.

```
. matrix c1=(2,-1,-1,0)
```

```
. manovatest, test(c1)
```

Test constraint

```
(1) 2*1.group - 2.group - 3.group = 0
```

	W = Wilks' lambda	L = Lawley-Hotelling trace	P = Pillai's trace	R = Roy's largest root		
Source	Statistic	df	F(df1, df2) =	F	Prob>F	
manovatest	W 0.5290	1	3.0 28.0	8.31	0.0004	e
	P 0.4710		3.0 28.0	8.31	0.0004	e
	L 0.8904		3.0 28.0	8.31	0.0004	e
	R 0.8904		3.0 28.0	8.31	0.0004	e
Residual		30				

e = exact, a = approximate, u = upper bound on F

نتایج بالا نشان می دهد که گروه یک از نظر آماری با میانگین گروه ۲ و ۳ اختلاف معنی داری دارد. اکنون مقایسه گروه ۱ کنترل (گروه ۲) با گروه ۲ کنترل (گروه ۳) باید انجام شود. مجدد باید ماتریسی درست نمود و آن را C2 نامگذاری نمود. نتایج این کار نیز در زیر آمده است.



```
. margins group, predict(equation(useful))
```

Adjusted predictions Number of obs = 33

Expression : Linear prediction, predict(equation(useful))

group	Delta-method					
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
1	18.11818	.9438243	19.20	0.000	16.26832	19.96804
2	15.52727	.9438243	16.45	0.000	13.67741	17.37713
3	15.34545	.9438243	16.26	0.000	13.49559	17.19532

```
. margins group, predict(equation(importance))
```

Adjusted predictions Number of obs = 33

Expression : Linear prediction, predict(equation(importance))

group	Delta-method					
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
1	8.681818	1.136676	7.64	0.000	6.453973	10.90966
2	5.109091	1.136676	4.49	0.000	2.881246	7.336936
3	5.636364	1.136676	4.96	0.000	3.408519	7.864208

در خروجی هر یک از سه جدول بالا نشان داده شده است که میانگین پیش بینی شده برای گروه های ۲ و ۳ خیلی شبیه به هم هستند. میانگین پیش بینی شده برای گروه ۱ بالاتر از میانگین پیش بینی شده برای گروه های ۲ و ۳ است.

اکنون اختلاف میانگین برای هر یک از متغیرهای وابسته در هر یک از گروه های کنترل (گروه ۲ و گروه ۳) و مقایسه با گروه درمان (گروه ۱) باید محاسبه و گزارش گردد.

با توجه به متغیر *difficulty*، تفاوت بین میانگین گروه ۱ کنترل در مقابل میانگین در گروه درمان، تقریباً با حدود اطمینان مورد نظر (6.19 - 5.58) 0.61- است.

تفاوت بین میانگین گروه ۲ کنترل با گروه درمان نیز (6.19 - 5.37) 0.82- می باشد.

در جدول زیر این تفاوت ها نشان داده شده است.

```
. margins, dydx(group) predict(equation(difficulty))
```

Conditional marginal effects                                Number of obs    =            33

Expression    : Linear prediction, predict(equation(difficulty))

dy/dx w.r.t. : 2.group 3.group

group	Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]
2	-.6090908	.8748585	-0.70	0.486	-2.323782    1.1056
3	-.8181818	.8748585	-0.94	0.350	-2.532873    .8965094

Note: dy/dx for factor levels is the discrete change from the base level.

اختلاف های مذکور برای متغیر useful در جدول زیر نشان داده شده است.

```
. margins, dydx(group) predict(equation(useful))
```

Conditional marginal effects                                Number of obs    =            33

Expression    : Linear prediction, predict(equation(useful))

dy/dx w.r.t. : 2.group 3.group

group	Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]
2	-2.590909	1.334769	-1.94	0.052	-5.207008    .0251907
3	-2.772727	1.334769	-2.08	0.038	-5.388827    -.1566278

Note: dy/dx for factor levels is the discrete change from the base level.

```
. margins, dydx(group) predict(equation(importance))
```

group	Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]
2	-3.572727	1.607503	-2.22	0.026	-6.723375    -.4220792
3	-3.045454	1.607503	-1.89	0.058	-6.196103    .1051936

سر انجام، با انجام ANOVA برای هر یک از متغیرهای وابسته به صورت جداگانه، تحلیل نهایی انجام می شود. این تحلیل ها در زیر آمده است.

```
. foreach vname in difficulty useful importance {
2.
. anova `vname' group
3.
. }
```

Number of obs = 33 R-squared = 0.0305  
 Root MSE = 2.05173 Adj R-squared = -0.0341

Source	Partial SS	df	MS	F	Prob > F
Model	3.97515121	2	1.9875756	0.47	0.6282
group	3.97515121	2	1.9875756	0.47	0.6282
Residual	126.287277	30	4.20957589		
Total	130.262428	32	4.07070087		

Number of obs = 33 R-squared = 0.1526  
 Root MSE = 3.13031 Adj R-squared = 0.0961

Source	Partial SS	df	MS	F	Prob > F
Model	52.9242378	2	26.4621189	2.70	0.0835
group	52.9242378	2	26.4621189	2.70	0.0835
Residual	293.965442	30	9.79884808		
Total	346.88968	32	10.8403025		

Number of obs = 33 R-squared = 0.1610  
 Root MSE = 3.76993 Adj R-squared = 0.1051

Source	Partial SS	df	MS	F	Prob > F
Model	81.8296936	2	40.9148468	2.88	0.0718
group	81.8296936	2	40.9148468	2.88	0.0718
Residual	426.370896	30	14.2123632		
Total	508.20059	32	15.8812684		

هیچ یک از آنالیز واریانس های انجام شده در بالا از نظر آماری در سطح آلفای 0.05 معنی دار نشده است. به عنوان مثال برای متغیر **difficulty** ، مقدار F-ratio کمتر از ۱ شده است.

**برای ارزیابی پیش فرض ها در MANOVA اقدامات زیر باید انجام شود:**

۱- ابتدا اینکه توزیع متغیر وابسته در هر یک از سطوح مختلف متغیر مستقل توزیعی نرمال داشته باشد. برای این ارزیابی از partial test و دستور زیر استفاده می شود. خروجی در زیر نشان می دهد که توزیع نرمال نبوده است.



```
. mvtest normality difficult useful importance
```

```
Test for multivariate normality
```

```
Doornik-Hansen          chi2(6) =    13.371    Prob>chi2 =    0.0375
```

۲- دوم اینکه فرض ماتریس کوواریانس جمعیت باید همگن یا عبارتی (a.k.a. sphericity) برقرار باشد. این فرض می گوید که واریانس و کوواریانس همه متغیرهای وابسته باید در بین گروه ها بر حسب متغیرهای مستقل برابر باشند. برای ارزیابی این فرض از دستور زیر استفاده می شود. خروجی در زیر آمده و همگنی واریانس ها را نشان می دهد.

```
. mvtest covariance difficult useful importance, by(group)
```

```
Test of equality of covariance matrices across 3 samples
```

```
Modified LR chi2 =    12.02242
Box F(12,4361.5) =      0.85      Prob > F =    0.5938
Box chi2(12) =      10.29      Prob > chi2 =    0.5909
```

### نکته:

- حجم نمونه کم، توان آزمون را پایین می آورد. اما چنانچه متغیرهای وابسته نرمال باشند آنالیز واریانس چند متغیره آزمون با توان بالا تر از آنالیزهای واریانس جداگانه است.

- بعد از معنی دار شدن MANOVA حداقل پنج نوع تحلیل تعقیبی وجود دارد. این تحلیل ها عبارتند از:

- Multiple univariate ANOVAs
- Stepdown Analysis
- Dependent Variable Contribution
- Multivariate Contrasts

## آزمون تحلیل واریانس اندازه های تکراری :

### Repeated Measures Analysis of Variance

#### دو نوع اندازه گیری مکرر وجود دارد:

- ۱- هنگامی که روش درمان یا موضوع مورد پژوهش ثابت بوده اما هر فرد چند بار مورد مشاهده قرار می گیرد. یعنی یک فرد در دفعات مکرر درمان را گرفته و آن درمان مورد سنجش قرار می گیرد.
  - ۲- هنگامی که روش درمان یا موضوع مورد پژوهش متغیر است و هر فرد نمونه به چند درمان پاسخ می دهد.
- تحلیل واریانس ناپارامتری برای اندازه گیری های مکرر، فریدمن است که دستور انجام آن در آزمون های ناپارامتریک بیان گردید. در این آزمون گروه های مورد بررسی از هم مستقل نیستند. تعمیم یافته آزمون t زوجی است. اندازه های تکراری یا مکرر به طراحی آماری گفته می شود که در آن هر یک از آزمودنی ها یا مشاهدات در

معرض بیش از یک متغیر مستقل قرار گرفته می شود.

با این روش می توان مقادیر متغیرهای وابسته چندگانه را که در دوره های زمانی متعدد مورد سنجش و اندازه گیری قرار می گیرند را با توجه به متغیرهای مستقل کیفی و کمی و همچنین دوره های زمانی که در آن سنجش شده اند مدل سازی نمود.

در این طرح تغییراتی را که ناشی از همبستگی اندازه گیری هاست، از تغییرات کل جدا شده و تغییرات خطا را با کم کردن تغییرات ناشی از تفاوت های فردی کاهش می دهند. در این طرح ابتدا واریانس کل به واریانس بین آزمودنی ها و واریانس درون آزمودنی ها تقسیم می شود.

واریانس درون آزمودنی ها نیز به واریانس بین متغیرهای آزمایشی و واریانس باقی مانده تقسیم می شود. درجات آزادی نیز به درجه آزادی بین مشاهدات و درون مشاهدات تقسیم می شود. دو نوع تحلیل در این طراحی انجام می شود. یکی اینکه مقایسه روند زمانی انجام می شود. مثلاً اینکه آیا تاثیر درمان در طول زمان ثابت است؟ و دیگری مقایسه بین گروه ها انجام می شود. مثلاً اینکه آیا تاثیر زمان در گروه های مختلف یکسان است؟

### پیش فرض ها:

برقراری پیش فرض هایی که در تحلیل واریانس یکطرفه ذکر گردید و شرایط زیر:

۱- متغیر وابسته توزیع نرمال داشته باشد. یکسان بودن واریانس ها (Homoscedasticity) و فرض کرویت (Sphericity) برقرار باشد. دو شرط در برقراری فرض کرویت مهم است. یکی بین اندازه ها باید همبستگی مشابه و یکسانی وجود داشته باشد و دیگری واریانس متغیر پاسخ (وابسته) در هر نقطه زمانی مورد بررسی یکسان باشد. عبارتی اینکه ضرایب همبستگی بین اندازه های متواتر همه با هم تقریباً مساوی اند یا خیر مورد آزمون قرار می گیرد. اگر این اندازه های متواتر یکسان بودند شرط کرویت برقرار است. یا به عبارتی ماتریس های کوواریانس مشاهده شده متغیرهای وابسته در بین گروه های مختلف با یکدیگر برابر باشند.

Box s test فرضیه برابری ماتریس های کوواریانس مشاهده شده متغیرهای وابسته در بین گروه های مختلف را ارزیابی می کند. این فرض در ارزیابی مدل باید برقرار و مقدار  $p$  آن کمتر از 0.05 نباشد. آزمون Mauchly test برای ارزیابی فرض کرویت (Sphericity) و کرویت بودن ماتریس کوواریانس متغیر وابسته انجام می شود. وقتی این آزمون معنی دار شد و فرض برقرار نبود باید تصحیح بر مبنای درجه آزادی و کاهش آن صورت گیرد و از آزمون های محافظه کار گرینهاوس - گیسر، هیون فلت و حد پایین استفاده نمود. نرم افزار بطور خودکار با دستور بکار رفته این آزمون ها را انجام می دهد.

۲- عامل ها و متغیرهای کمکی با متغیرهای وابسته رابطه خطی داشته باشند.

۳- متغیرهای مستقل عامل باید در سطح کیفی باشند.

۴- متغیرهای مستقل کووریت باید در سطح کمی باشند.

## کار با نرم افزار و انجام تحلیل واریانس اندازه گیری های مکرر:

فایل داده های بنام RM2 را باز نمایید. در این فایل داده هایی با فرمت اندازه های تکراری وجود دارد. در اندازه های تکراری فرمت داده ها به دو صورت است. یکی ممکن است به صورت گسترده یا Wide باشند. بدین صورت که هر نمونه (id) یا subject در طی زمان های مختلف و متعدد (y1 تا y4) مورد سنجش و اندازه گیری برای متغیرهای پاسخ مورد نظر قرار می گیرد. شکل زیر این فرمت از داده ها را نشان می دهد.

```
id y1 y2 y3 y4
1 3.5 4.5 7.5 7.5
2 6.5 5.5 8.5 8.5
```

در فرمت دیگر که به صورت طولانی یا long است یک نمونه یا subject برای یک دوره زمانی مورد سنجش و اندازه گیری برای متغیر پاسخ مورد نظر (Y) قرار می گیرد. این فرمت از داده ها بنام شخص - دوره مد نظر است. در نرم افزارهای SPSS و SAS داده ها به صورت Wide با ANOVA مورد تحلیل قرار می گیرند. در Stata داده ها به فرمت long و با ANOVA و Mixed models مورد تحلیل قرار می گیرند. شکل زیر این فرمت از داده ها را نشان می دهد.

```
id time y
1 1 3.5
1 2 4.5
1 3 7.5
1 4 7.5
2 1 6.5
2 2 5.5
2 3 8.5
2 4 8.5
```

در این فایل ۸ فرد (id) یا Subject در طی چهار نقطه زمانی مورد اندازه گیری برای متغیر پاسخ قرار می گیرند. داده ها در فرمت Wide هستند. Y1 اندازه متغیر پاسخ در زمان اول، Y2 متغیر پاسخ در زمان دوم و به همین ترتیب متغیرهای Y3 و Y4. مشاهدات شامل دو گروه و مشتمل بر ۴ نفر هستند که تحت درمان هستند. یعنی در هر زمان می توان ۸ نفر یا دو گروه چهار نفره داشت. متغیر پاسخ trt نام داشته و مورد اندازه گیری در زمان های مختلف (چهار بار) قرار می گیرد. شکل زیر این داده ها را در نرم افزار Stata نشان می دهد.

	id	trt	y1	y2	y3	y4
1	1	1	3.5	4.5	7.5	7.5
2	2	1	6.5	5.5	8.5	8.5
3	3	1	3.5	4.5	7.5	9.5
4	4	1	3.5	3.5	6.5	8.5
5	5	2	1	2	5	10
6	6	2	2	3	6	10
7	7	2	2	4	5	9
8	8	2	2	3	6	11

شاخص های توصیفی متغیر پاسخ اندازه گیری شده در طی چهار بار در زیر آمده است.

. summarize y1-y4

Variable	Obs	Mean	Std. Dev.	Min	Max
y1	8	3	1.690309	1	6.5
y2	8	3.75	1.101946	2	5.5
y3	8	6.5	1.253566	5	8.5
y4	8	9.25	1.101946	7.5	11

شاخص های خلاصه (تعداد افراد، میانگین، انحراف معیار و واریانس) برای کل افراد و هر یک از گروه ها در طی چهار زمان مختلف در زیر آمده است.

. tabstat y1-y4, by(trt) stat(n mean sd var)

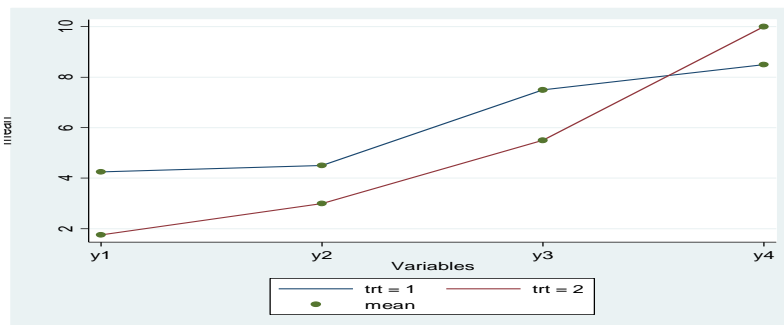
Summary statistics: N, mean, sd, variance  
by categories of: trt

trt	y1	y2	y3	y4
1	4	4	4	4
	4.25	4.5	7.5	8.5
	1.5	.8164966	.8164966	.8164966
	2.25	.6666667	.6666667	.6666667
2	4	4	4	4
	1.75	3	5.5	10
	.5	.8164966	.5773503	.8164966
	.25	.6666667	.3333333	.6666667
Total	8	8	8	8
	3	3.75	6.5	9.25
	1.690309	1.101946	1.253566	1.101946
	2.857143	1.214286	1.571429	1.214286

با استفاده از دستور **profileplot y1-y4, by(trt)**، نمودار متغیر اندازه گیری شده بر حسب دو گروه ترسیم می گردد.

نکته: ممکن است دستور بالا در نرم افزار سیستم شما نصب نباشد! بنابراین این دستور را نمی توانید اجرا نمایید! برای نصب این دستور بر نرم افزار سیستم خویش با استفاده از دستور زیر این دستور را از اینترنت دانلود و نصب نمایید. **findit profileplot**

اکنون دستور ترسیم نمودار را اجرا نمایید. نمودار زیر مشاهده می گردد.



این نمودار میانگین متغیر اندازه گیری شده در طی زمان های مختلف (چهار بار اندازه گیری) بر حسب دو گروه

درمانی را نمایش می دهد. در طی زمان چهارم ملاحظه می شود که جهت رابطه معکوس شده اما معنی دار نیست. این موضوع در نمودار بالا به صورت قطع نمودن دو خط نمودار در طی اندازه گیری در زمان چهارم نمایش داده شد. زمانی که خطوط رسم شده در اینگونه نمودارها تقریباً دو خط موازی باشد می توان انتظار داشت که تداخل یا اینترکشن بین متغیرهای مذکور معنی دار نباشد و الگوی تغییرات متغیر عددی وابسته در زمان های مختلف یکسان باشد. اما وقتی خطوط یکدیگر را قطع نمایند دلالت بر وجود تداخل معنی دار دارد. برای آشنایی با ضرایب همبستگی، جهت و مقدار معنی داری مقادیر اندازه گیری شده در طی چهار بار ماتریس این اندازه ها در زیر آمده است.

```
. pwcorr y1-y4, sig
```

	y1	y2	y3	y4
y1	1.0000			
y2	0.8820 0.0038	1.0000		
y3	0.9102 0.0017	0.8273 0.0113	1.0000	
y4	-0.5752 0.1358	-0.6471 0.0829	-0.5171 0.1894	1.0000

ماتریس همبستگی کوواریانس های این مقادیر اندازه گیری نیز در زیر آمده است.

```
. correlate y1-y4, cov  
(obs=8)
```

	y1	y2	y3	y4
y1	2.85714			
y2	1.64286	1.21429		
y3	1.92857	1.14286	1.57143	
y4	-1.07143	-.785714	-.714286	1.21429

در انجام آنالیز واریانس برای داده های تکراری فرض این است که ساختار کوواریانس در بین افراد (within-subject covariance structure) باید متقارن باشد. این فرض در اینجا رعایت نشده است.

**تغییر فرمت داده ها از Wide به Long:**

برای این کار از دستور زیر استفاده می شود.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. reshape long y, i(id) j(time)
(note: j = 1 2 3 4)
```

Data	wide	->	long
Number of obs.	8	->	32
Number of variables	6	->	4
j variable (4 values)		->	time
xij variables:	y1 y2 ... y4	->	y

همانگونه که ملاحظه می شود با دستور reshape این کار صورت گرفته و تعداد مشاهدات از ۸ نفر به ۳۲ نفر، تعداد متغیرها از ۶ تا به ۴ تغییر نموده است. پس در داده های long تعداد مشاهدات بیشتر از فرمت داده ها بصورت wide است. J متغیر زمان (۴ بار اندازه گیری) و i نشان دهنده شماره فرد است. اکنون داده ها به صورت زیر که فقط ۹ مورد از آنها آمده است در نرم افزار تغییر یافته اند. شکل زیر را با شکل قبلی مقایسه نمایید.

		id	time	trt	y
1	1	1	1	1	3.5
2	1	1	2	1	4.5
3	1	1	3	1	7.5
4	1	1	4	1	7.5
5	2	2	1	1	6.5
6	2	2	2	1	5.5
7	2	2	3	1	8.5
8	2	2	4	1	8.5
9	3	3	1	1	3.5

با استفاده از دستور db reshape نیز می توان کادر محاوره ای تبدیل فرمت داده ها به یکدیگر را به صورت زیر در دسترس داشت.

## روش های اپیدمیولوژی و پژوهش با Stata

Long format	id	year	sex	inc	ue
	1	80	0	5000	0
	1	81	0	5500	1
	1	82	0	6000	0
	2	80	1	2000	1
	2	81	1	2200	0
	2	82	1	3300	0
	3	80	0	3000	0
	3	81	0	2000	0
	3	82	0	1000	1

Wide format	id	sex	inc80	inc81	inc82	ue80	ue81	ue82
	1	0	5000	5500	6000	0	1	0
	2	1	2000	2200	3300	1	0	0
	3	0	3000	2000	1000	0	0	1

Variable edit field	Value
ID variable(s) - the i() option	id
Subobservation identifier variable - the j() option	year
Variables that differ on subobservations	inc ue

برای نمایش لیست کلیه متغیرها و داده ها با استفاده از دستور `list` ، متغیرها به صورت زیر (صفحه بعد) نمایش داده می شوند. در این داده ها دو نوع اثر ( Between-Subject and Within-Subject Effects ) وجود دارد. که به آن مدل اثرات مخلوط ( Mixed Effects Model ) گفته می شود. برای آنالیز داده های `long` به شرح داده های موجود، باید مقادیر خطای هر دو نوع اثر را محاسبه نمود. بر اساس قانون کلی یک خطا برای اثرات میان فردی و خطای جداگانه برای هر یک از افراد وجود دارد.

```
. list, sep(4)
```

	id	time	trt	y
1.	1	1	1	3.5
2.	1	2	1	4.5
3.	1	3	1	7.5
4.	1	4	1	7.5
5.	2	1	1	6.5
6.	2	2	1	5.5
7.	2	3	1	8.5
8.	2	4	1	8.5
9.	3	1	1	3.5
10.	3	2	1	4.5
11.	3	3	1	7.5
12.	3	4	1	9.5
13.	4	1	1	3.5
14.	4	2	1	3.5
15.	4	3	1	6.5
16.	4	4	1	8.5
17.	5	1	2	1
18.	5	2	2	2
19.	5	3	2	5
20.	5	4	2	10
21.	6	1	2	2
22.	6	2	2	3
23.	6	3	2	6
24.	6	4	2	10
25.	7	1	2	2
26.	7	2	2	4
27.	7	3	2	5
28.	7	4	2	9
29.	8	1	2	2
30.	8	2	2	3
31.	8	3	2	6
32.	8	4	2	11

اکنون برای انجام آنالیز واریانس داده های تکراری زمینه مناسب است. برای انجام این آنالیز پیش فرض متقارن

بودن ماتریس کوواریانس درون افراد مورد نیاز است. یا بعبارتی واریانس در هر زمان برای مقادیر اندازه گیری شده باید برابر باشد. اگر این شرط برقرار نباشد مقدار پی ویو گزارش شده بوسیله آزمون آنالیز واریانس با داده های تکراری مبتنی بر احتمالات صحیح نیست. در نرم افزار برای این نقص حربه ای بنام ((repeated(time)) در دستور اجرای ANOVA وجود دارد که بکار می رود. نرم افزار برای این کار مقدار پی ویو را بصورت محافظه کارانه گزارش می کند. `anova y trt / id|trt time trt#time, repeated(time)`

	Number of obs =	32	R-squared =	0.9624	
	Root MSE =	.712	Adj R-squared =	0.9352	
Source	Partial SS	df	MS	F	Prob > F
Model	233.375	13	17.9519231	35.41	0.0000
trt	10.125	1	10.125	6.48	0.0438
id trt	9.375	6	1.5625		
time	194.5	3	64.8333333	127.89	0.0000
trt#time	19.375	3	6.4583333	12.74	0.0001
Residual	9.125	18	.506944444		
Total	242.5	31	7.82258065		

Between-subjects error term: id|trt  
 Levels: 8 (6 df)  
 Lowest b.s.e. variable: id  
 Covariance pooled over: trt (for repeated variable)

Repeated variable: time

Huynh-Feldt epsilon = 0.9432  
 Greenhouse-Geisser epsilon = 0.5841  
 Box's conservative epsilon = 0.3333

Source	df	F	Prob > F			
			Regular	H-F	G-G	Box
time	3	127.89	0.0000	0.0000	0.0000	0.0000
trt#time	3	12.74	0.0001	0.0002	0.0019	0.0118
Residual	18					

در جدول بالا نتایج به این صورت تفسیر می گردد:

اثرات درمان و همچنین زمان و اینترکشن زمان (time) با درمان (Treatment) با توجه به مقادیر p معنی دار هستند. خروجی پایین جدول بوسیله سه آزمون محافظه کار F به شرح زیر گزارش گردید.

- ۱- Huynh-Feldt
- ۲- Greenhouse-Geisser
- ۳- Box's conservative F



این مقادیر شاخص هایی هستند که اگر پیش فرض متقارن بودن برقرار نبود می بایست مورد توجه قرار گیرند. در اینجا می توان ماتریس کوواریانس درون افراد را به صورت کلی با دستور `matrix list e(Srep)` اجرا نمود. با توجه به وجود اینترکشن زمان با متغیر درمان و معنی دار بودن آن باید این اینترکشن را آزمون و آن را با اثرات زمان بر درمان توضیح داد. برای آزمون اثرات زمان بر درمان از آزمون ساده اثرات ( `Tests of Simple Effects` ) دستور زیر استفاده می شود.

### contrast time& trt, effect

Margins : asbalanced

	df	F	P>F
time@trt			
1	3	35.96	0.0000
2	3	104.67	0.0000
Joint	6	70.32	0.0000
Residual	18		

	Contrast	Std. Err.	t	P> t	[95% Conf. Interval]
time@trt					
(2 vs base)					
1	.25	.5034602	0.50	0.626	-.8077307 1.307731
(2 vs base)					
2	1.25	.5034602	2.48	0.023	.1922693 2.307731
(3 vs base)					
1	3.25	.5034602	6.46	0.000	2.192269 4.307731
(3 vs base)					
2	3.75	.5034602	7.45	0.000	2.692269 4.807731
(4 vs base)					
1	4.25	.5034602	8.44	0.000	3.192269 5.307731
(4 vs base)					
2	8.25	.5034602	16.39	0.000	7.192269 9.307731

درجه آزادی اثرات ساده زمان برای هر یک از سطوح درمان سه، و برای کل شش است. برای آزمون اثرات ساده از دستور **contrast** استفاده گردید. مقادیر معنی داری برای اثرات مختلف در سطوح مختلف نشان داده شده است. هر یک از آزمون های اثرات ساده مشتمل بر پی گیری در چهار نقطه زمانی برای مقایسه های جفتی است. برای انجام مقایسه های جفتی متعدد از دستور زیر استفاده می شود.

```
margins time, at(trt=1) pwcompare(effects) noestimcheck
margins time, at(trt=2) pwcompare(effects) noestimcheck
```

نتایج این دستورات در زیر آمده است.

	Delta-method		Unadjusted		Unadjusted	
	Contrast	Std. Err.	z	P> z	[95% Conf. Interval]	
time						
2 vs 1	.25	.5034602	0.50	0.619	-.736764	1.236764
3 vs 1	3.25	.5034602	6.46	0.000	2.263236	4.236764
4 vs 1	4.25	.5034602	8.44	0.000	3.263236	5.236764
3 vs 2	3	.5034602	5.96	0.000	2.013236	3.986764
4 vs 2	4	.5034602	7.95	0.000	3.013236	4.986764
4 vs 3	1	.5034602	1.99	0.047	.013236	1.986764

Expression : Linear prediction, predict()  
at : trt = 2

	Delta-method		Unadjusted		Unadjusted	
	Contrast	Std. Err.	z	P> z	[95% Conf. Interval]	
time						
2 vs 1	1.25	.5034602	2.48	0.013	.263236	2.236764
3 vs 1	3.75	.5034602	7.45	0.000	2.763236	4.736764
4 vs 1	8.25	.5034602	16.39	0.000	7.263236	9.236764
3 vs 2	2.5	.5034602	4.97	0.000	1.513236	3.486764
4 vs 2	7	.5034602	13.90	0.000	6.013236	7.986764
4 vs 3	4.5	.5034602	8.94	0.000	3.513236	5.486764

در اینجا می توان برای هر کدام از درمان ها ( گروه های درمانی ۱ و ۲ ) بر اساس جداول بالا نتایج مقایسه های جفتی در زمان های مختلف را مشاهده نمود. همانگونه که مشاهده می شود:

در گروه درمانی اول برای زمان دوم در مقایسه با زمان اول مقدار اثر درمان معنی دار نگردید (P=0.619). بقیه اثرات درمانی در زمان های مختلف معنی دار گردید. در گروه درمانی دوم در زمان های مختلف اثرات درمان دوم معنی دار شده است.

اثرات درمان دوم در زمان های اول و دوم در مقایسه با زمان های دوم و سوم و چهارم بیشتر از درمان گروه اول بوده ولیکن در گروه دوم در مقایسه زمان چهارم با سوم این اثر کمتر از زمان قبلی شده است.

برای آزمون اثرات هر یک از درمان نیاز به محاسبه خطا کلی است. اکنون با دستور `anova y trt##time` این خطا ها و خطای باقی مانده ها محاسبه میشوند. مجدد با دستور زیر اثرات هر یک از درمان ها در زمان های مختلف تعیین می گردد. نتایج در زیر نشان می دهد که اثرات در زمان های چهارگانه به شرح زیر معنی دار بوده است.

Contrasts of marginal linear predictions

Margins : asbalanced

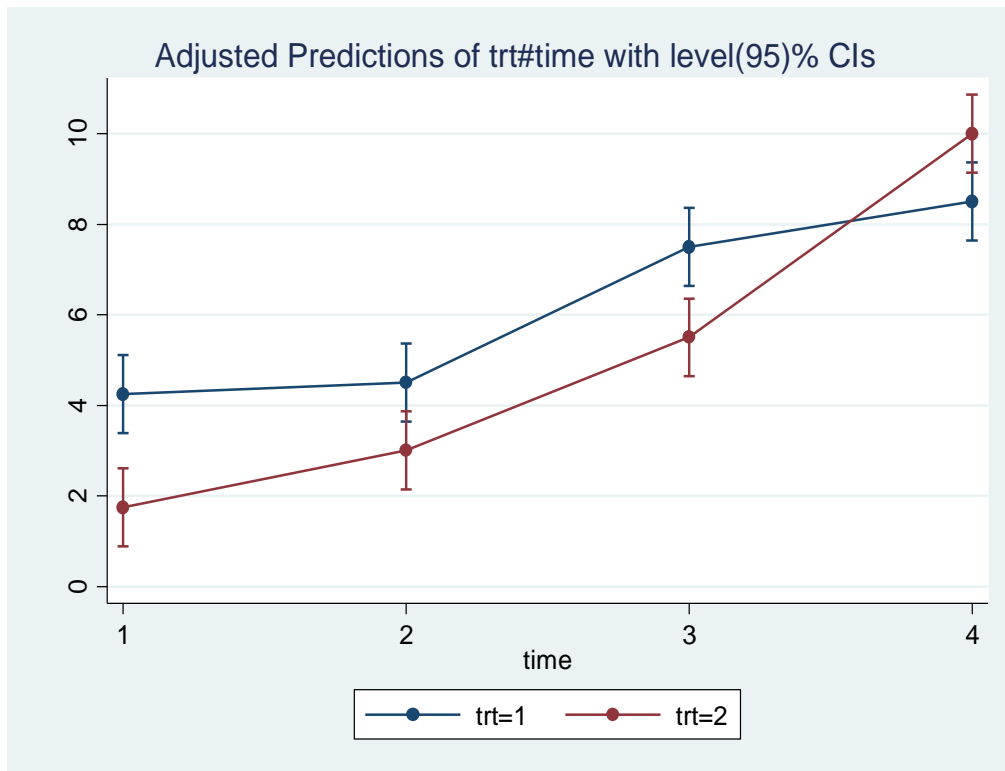
	df	F	P>F
trt@time			
1	1	16.22	0.0005
2	1	5.84	0.0237
3	1	10.38	0.0036
4	1	5.84	0.0237
Joint	4	9.57	0.0001
Residual	24		

	Contrast	Std. Err.	t	P> t	[95% Conf. Interval]	
trt@time						
(2 vs base)						
1	-2.5	.6208194	-4.03	0.000	-3.781308	-1.218692
(2 vs base)						
2	-1.5	.6208194	-2.42	0.024	-2.781308	-.2186918
(2 vs base)						
3	-2	.6208194	-3.22	0.004	-3.281308	-.7186918
(2 vs base)						
4	1.5	.6208194	2.42	0.024	.2186918	2.781308

در مقایسه درمان دوم با اول در زمان های مختلف ( زمان های چهارگانه)، جدول بالا نشان می دهد که همه اثرات درمان اول در سه زمان اول بر درمان نوع دوم ارجح است. منتها درمان دوم در زمان چهارم بر درمان اول ارجح تر است.

ترسیم نمودار برای نمایش اینترکشن بسیار مناسب است. از دستور margins trt#time برای محاسبه اینترکشن درمان و زمان و از دستور marginsplot, x(time) برای نمایش گرافیکی اینترکشن زمان با درمان استفاده می شود. نتایج نشان میدهد که در همه زمان ها درمان با زمان اینترکشن دارد. این نمودار در زیر آمده است.

	Delta-method				
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
trt#time					
1 1	4.25	.4389856	9.68	0.000	3.389604 5.110396
1 2	4.5	.4389856	10.25	0.000	3.639604 5.360396
1 3	7.5	.4389856	17.08	0.000	6.639604 8.360396
1 4	8.5	.4389856	19.36	0.000	7.639604 9.360396
2 1	1.75	.4389856	3.99	0.000	.8896041 2.610396
2 2	3	.4389856	6.83	0.000	2.139604 3.860396
2 3	5.5	.4389856	12.53	0.000	4.639604 6.360396
2 4	10	.4389856	22.78	0.000	9.139604 10.8604



### معایب روش آنالیز واریانس با اندازه های تکراری:

- باید مقادیر خطا برای هر یک از اثرات بصورت صحیح محاسبه گردد.
- برقرار بودن فرض متقارن بودن یا عدم تغییر پذیری ساختار کوواریانس ها
- وقتی مشاهدات درون افراد نابرابر باشند اجازه انجام ممکن نیست.

راه دیگر انجام آنالیز واریانس برای اندازه های تکراری استفاده از Repeated Measures Mixed Model است. در این مدل نیازی به محاسبه خطاها بصورت جداگانه نیست. فقط نیاز به نام متغیرها در اندازه های تکراری است. با استفاده از دستور زیر این آنالیز امکان پذیر است.

```
xtmixed y trt##time || id:, var reml
```

از گزینه reml در دستور بالا برای قابل مقایسه بودن این روش با آنالیز واریانس استفاده می شود. نتایج اثرات درمان ها در زمان های مختلف و اینترکشن آنها با زمان در زیر آمده است.

## روش های اپیدمیولوژی و پژوهش با Stata

```

Mixed-effects REML regression      Number of obs   =      32
Group variable: id                 Number of groups =      8

                                   Obs per group: min =      4
                                   avg       =      4.0
                                   max       =      4

                                   Wald chi2(7)    =    428.37
Log restricted-likelihood = -34.824379  Prob > chi2     =    0.0000
    
```

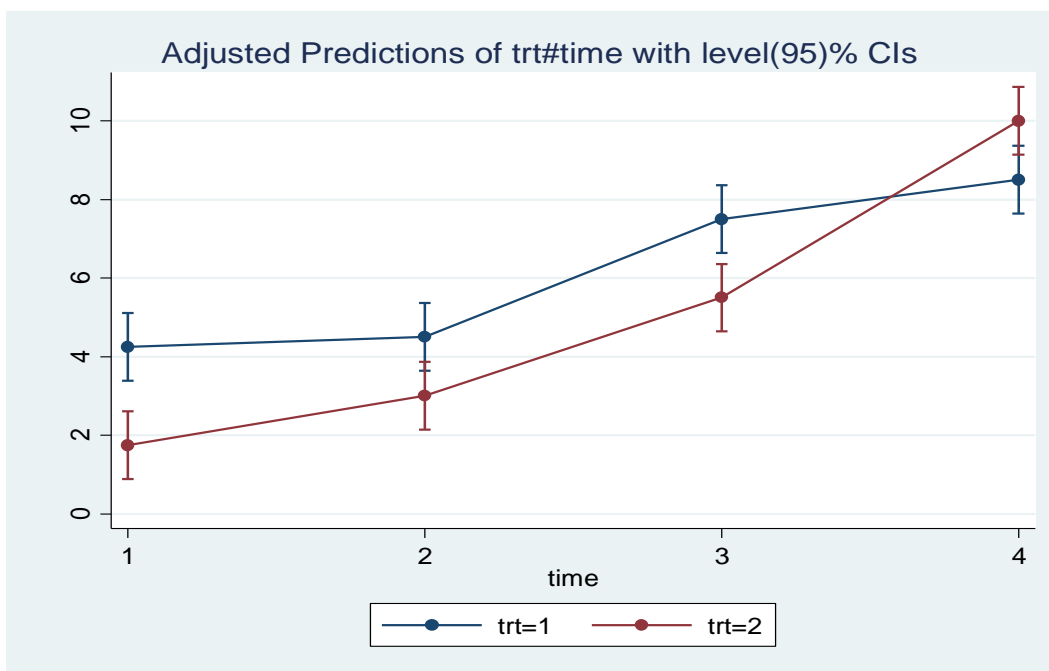
y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
2.trt	-2.5	.6208193	-4.03	0.000	-3.716783	-1.283217
time						
2	.25	.5034603	0.50	0.619	-.736764	1.236764
3	3.25	.5034603	6.46	0.000	2.263236	4.236764
4	4.25	.5034603	8.44	0.000	3.263236	5.236764
trt#time						
2 2	1	.7120004	1.40	0.160	-.3954951	2.395495
2 3	.5	.7120004	0.70	0.483	-.8954951	1.895495
2 4	4	.7120004	5.62	0.000	2.604505	5.395495
_cons	4.25	.4389855	9.68	0.000	3.389604	5.110396

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Identity				
var(_cons)	.2638887	.2294499	.0480071	1.450562
var(Residual)	.5069445	.1689815	.2637707	.9743036

LR test vs. linear regression: chibar2(01) = 3.30 Prob >= chibar2 = 0.0346

برای محاسبه اثرات اینترکشن و نمایش گرافیکی اثرات و اینترکشن در این روش همانند روش آنالیز واریانس که قبلاً توضیح و ترسیم گردید نمودار و جدول زیر نمایش داده می شود. همانگونه که ملاحظه می گردد تفاوتی بین دو روش به نظر نمی رسد.

trt#time	Delta-method				
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
1 1	4.25	.4389855	9.68	0.000	3.389604 5.110396
1 2	4.5	.4389855	10.25	0.000	3.639604 5.360396
1 3	7.5	.4389855	17.08	0.000	6.639604 8.360396
1 4	8.5	.4389855	19.36	0.000	7.639604 9.360396
2 1	1.75	.4389855	3.99	0.000	.8896042 2.610396
2 2	3	.4389855	6.83	0.000	2.139604 3.860396
2 3	5.5	.4389855	12.53	0.000	4.639604 6.360396
2 4	10	.4389855	22.78	0.000	9.139604 10.8604



آزمون اثرات به صورت ساده نیز طبق دستورات قبلی آنالیز واریانس به شرح زیر انجام می شود.

```
contrast time@trt, effect
margins time, at(trt=1) pwcompare(effects)
margins time, at(trt=2) pwcompare(effects)
contrast trt@time, effect
```

برای انجام آزمون های تعقیبی از دستور **contrast p.time@trt, effect** استفاده میشود. محل اختلاف در زیر مشخص است.

	Contrast	Std. Err.	z	P> z	[95% Conf. Interval]	
y						
time@trt						
(linear)						
1	1.760904	.1780001	9.89	0.000	1.41203	2.109777
(linear)						
2	3.046643	.1780001	17.12	0.000	2.697769	3.395516
(quadratic)						
1	.1875	.1780001	1.05	0.292	-.1613738	.5363738
(quadratic)						
2	.8125	.1780001	4.56	0.000	.4636262	1.161374
(cubic)						
1	-.5310661	.1780001	-2.98	0.003	-.8799399	-.1821924
(cubic)						
2	.0838525	.1780001	0.47	0.638	-.2650212	.4327263

### مزایای Mixed Model:

- محاسبه اتوماتیک و صحیح خطای معیار برای هر یک از اثرات
- وقتی در داده ها درون افراد مشاهدات کم شده باشد یا عدم تعادل، محاسبات انجام می شود.

- نیازی به یکسان بودن فاصله های زمانی مورد بررسی نیست.
- با ساختار کوواریانس متفاوت در مشاهدات ، محاسبات انجام می شود.
- با متغیر کیفی یا پیوسته ، مداخله در گروه ها را می توان اندازه گیری نمود.

### معایب:

Xtmixed محاسبات و نتایج را مبتنی بر کای اسکور (chi-square) گزارش می کند. مقدار p برای حجم نمونه بزرگ مناسب است در حجم نمونه های کم ممکن است با خطا همراه باشد.

### خلاصه و نتیجه گیری:

در این فصل ابتدا به مفاهیم و تعاریف پایه در آزمون فرضیه ها پرداخته شد. انواع خطا و شیوه گام به گام آزمون فرض تشریح گردید. انواع آزمونهای آماری بیان و موارد کاربرد و تفسیر آنها نیز ذکر گردید. آزمون های مقایسه میانگین (انواع آزمون t)، آزمون کای اسکور، انواع آزمون های تحلیل واریانس یکطرفه و دو طرفه و طرز کار با نرم افزار برای انواع آزمون ها بیان گردید. آزمون های تعقیبی برای انجام مقایسه میانگین ها (مقایسه های چندگانه) از جمله آزمون بونفرونی، سیداک و شفه نیز بیان گردید. آزمون های ناپارامتریک رایج نیز بیان و دستورات انجام آزمون های میانه، من ویتنی، ویلکاکسون، فریدمن، کروسکال والیس بیان گردیدند. نحوه ارزیابی توزیع متغیرها نیز بیان گردید. در این فصل انواع تحلیل دو و چند متغیره از جمله آنالیز واریانس تک متغیره و چند متغیره از جمله ANOVA، MANOVA، ANCOVA و Repeated Measure و طرز استفاده و تفسیر آنها با نرم افزار Stata بیان شد. برای هر کدام از آزمون ها و کار با نرم افزار Stata یک مثال ذکر گردید. استفاده از برنامه Help نرم افزار برای استفاده از آزمون های مذکور کمک کننده است.

**تمرین ها :**

- ۱- پیش فرض های استفاده از آنالیزهای دو متغیره چیست؟
- ۲- انواع آزمون های پارامتریک در آنالیزهای دو متغیره را نام ببرید؟
- ۳- معادل ناپارامتریک آزمون های پارامتریک برای مقایسه میانگین در دو گروه مستقل و وابسته چیست؟
- ۴- پیش فرض های انجام آنالیز واریانس یکطرفه و دو طرفه چیست؟
- ۵- وقتی سه متغیر وابسته کمی داشته باشید و دو متغیر مستقل کیفی و کمی، نوع آنالیز مربوطه چیست؟ پیش فرض ها چگونه است؟
- ۶- برای ارزیابی بعد از انجام MANOVA چه اقداماتی باید انجام داد؟
- ۷- در آنالیز واریانس با اندازه گیری های مکرر پیش فرض ها چیست؟
- ۸- داده های Long و داده های Wide چگونه ترکیب و فرمتی دارند؟ چگونه قابل تبدیل به یکدیگر هستند؟
- ۹- در فایل داده های پیوست، فایل part five و فایل manova را باز نمایید و یکبار آنالیزهای ANOVA و MANOVA را مانند آنچه در بالا و طی فصل گذشته توضیح داده شد اجرا نمایید.
- ۱۰- انواع و کاربرد آزمون های ناپارامتریک و دستورات انجام آنها با نرم افزار Stata را بیان فرمایید.
- ۱۱- فایل **ahmadi1400** که در فصل ۱ ساخته اید را باز نمایید. به سوالات زیر پاسخ دهید:
  - ۱-۱۱- آزمون فرضیه برای برابری نمایه توده بدنی در دو گروه بیماران
  - ۲-۱۱- آزمون فرضیه برای برابری متغیر HDL قبل و بعد از مداخله در هر یک از گروه ها
  - ۳-۱۱- آزمون فرضیه مقایسه وزن در دو گروه بیماران و آزمون برابری میانگین وزن برابر ۶۵ کیلوگرم
  - ۴-۱۱- از متغیر شغل افراد، یک متغیر سه طبقه ای از شغل ها به نام Job\_cat ایجاد نمایید. آزمون برابری وزن بین گروه های شغلی را انجام دهید.
  - ۵-۱۱- متغیر اختلاف بین متغیر HDL قبل و بعد از مداخله ایجاد و با نام diff\_HDL ذخیره نمایید. اکنون میانگین این متغیر را بین گروه های شغلی با و بدون کنترل BMI مقایسه کنید.
  - ۶-۱۱- از اختلاف متغیر diff\_HDL یک متغیر دو طبقه ای با تعریف مقادیر نرمال و غیرنرمال ایجاد نمایید. آیا جنسیت بیماران و گروه های درمانی با این متغیر رابطه ای دارد؟



## بخش چهارم:

# مدل سازی آماری در تحقیقات علوم پزشکی و

## روش های اپیدمیولوژی با Stata

### فصل ششم: همبستگی، رگرسیون و تحلیل های پیشرفته در اپیدمیولوژی با Stata

مروری بر همبستگی و مدل سازی های آماری در اپیدمیولوژی  
رگرسیون خطی ساده، رگرسیون چندگانه  
محاسبه اثر درمان در یک کارآزمایی بالینی و انجام آنالیز واریانس و رگرسیون خطی  
رگرسیون لجستیک ساده و چندگانه، رگرسیون لجستیک شرطی  
رگرسیون چندوجهی، رگرسیون رتبه ای، رگرسیون پواسون و رگرسیون کاکس  
روش های تحلیل داده های وابسته به هم: GEE و Random Effect  
تحلیل چند سطحی (Multi-Level)  
کار با نرم افزار  
تمرین ها

#### • اهداف:

#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند مفهوم رگرسیون را درک و پیش فرض های تحلیل های رگرسیونی را بداند.
- کاربر بتواند انواع تحلیل های رگرسیونی و مدل سازی آماری را انجام دهد.
- کاربر بتواند نتایج و خروجی های مورد نیاز حاصل از نتایج تحلیل های رگرسیونی انجام شده را تهیه و ارائه نماید.
- کاربر بتواند اثر درمان را در مطالعات محاسبه نماید.
- کاربر بتواند تفسیر مناسب از نتایج تحلیل های رگرسیونی انجام شده ارائه نماید.
- مدل سازی را تعریف و انواع مدل های آماری در اپیدمیولوژی را تشریح نماید.

## Modelling

## مدل سازی

تبدیل و بیان یک مفهوم آماری، به زبان ریاضی، نوعی از مدل سازی است. مدل سازی نوعی تکنیک و روش آماری محسوب می گردد. مدل سازی یکی از روش های بیان خلاصه واقعیت ها است. فرایند ایجاد و انتخاب مدل ها را مدل سازی نامیده اند. مدل ها، انواع گوناگون (مانند مدل های فیزیکی، ریاضی، عددی، هواشناسی، نرم افزاری و...) و کاربردهای حیاتی متنوع و فراوانی در همه زمینه های علوم و فن آوری دارند. هرچه مفاهیم زبان ریاضی استفاده شده در مدل سازی ساده تر باشند، مدل سازی ارزش بیشتری دارد. در مبحث کاربرد آمار در تحقیقات علوم پزشکی با به کار گیری داده های موجود و وارد نمودن متغیرهای مختلف در مدل و کنترل همزمان متغیرهای مخدوش کننده و بررسی نقش اینترکشن و پیش بینی مقادیر اطلاعات با توجه به داده های موجود، نوعی از مدل سازی در تحقیقات علوم پزشکی محسوب می شود. به عبارت دیگر مدل سازی یکی از روش های کنترل متغیرهای مخدوش کننده می باشد. برای برخورد و کنترل خطای مخدوش شدگی در مرحله طراحی مطالعه از روش های محدود سازی، همسان سازی و تصادفی سازی استفاده می شود. این روش ها در انتهای فصل پنجم توضیح داده می شوند. دو روش مهم در برخورد با خطای مخدوش کنندگی در مرحله تجزیه و تحلیل، روش طبقه بندی شده (Stratified Analysis) و روش مدل سازی آماری (Statistical Modeling) است. در روش طبقه بندی شده که به آن تطبیق کردن (Adjustment) نیز گفته می شود رابطه میان متغیرهای مواجهه و پیامد مورد نظر به صورت لایه بندی شده در طبقه های مختلف متغیر مخدوش کننده به طور جداگانه محاسبه و در نهایت میانگین رابطه مشاهده شده در طبقه های مختلف، با توجه به وزن هر طبقه از لحاظ تعداد نمونه موجود در آن محاسبه می شود. این روش دارای محدودیت هایی است از جمله با افزایش تعداد متغیرهای مخدوش کننده و تعداد طبقات متغیرهای مخدوش کننده، اندازه کلی نمونه در طبقه ها شکسته شده و به دلیل تعداد کم نمونه ها در هر طبقه، توان کمی برای کشف روابط وجود خواهد داشت. لذا برای رفع این محدودیت و کنترل همزمان چندین متغیر مخدوش کننده از مدل سازی آماری و رگرسیونی استفاده می شود. بکار گیری مدل های رگرسیونی برای پیش بینی و تعیین عوامل تعیین کننده و موثر بر بیماریها از رایجترین استفاده مدل سازی آماری در اپیدمیولوژی است. قبل از پرداختن به انواع مدل های آماری در اپیدمیولوژی و بلاخص مدل های رگرسیونی، ابتدا به گام های مدل سازی پرداخته می شود. سپس یادآوری مفهوم همبستگی بین متغیرها ارائه می شود.

مدل سازی نمایش دادن بخشی از واقعیت است. وقتی آگاهانه ویژگی ها و خصوصیت های برخی متغیرهای مهم را برای نشان دادن یک ساختار یا یک فرآیند در یک سیستم بکار می بریم در واقع نوعی مدل سازی است. مدل سازی یک مفهوم انتزاعی و انتخابی از واقعیت است. ممکن است ارائه همه واقعیت در مدل سازی مقدور نبوده یا امکان پذیر نباشد. پدیده ها و پیامدهای مرتبط با سلامتی و بیماریها در جامعه غالباً با ریسک فاکتورها و رخداد های چند علیتی و چندوجهی و از جمله با عوامل اجتماعی، فرهنگی، سیاسی و اقتصادی مرتبط می باشند، بنابراین مدل سازی این پدیده ها می تواند به درک بهتر چند علیتی بودن آنها کمک نماید. برای آشکار نمودن و ساده سازی روابط بین اجزای تشکیل دهنده یک رخداد مرتبط با سلامتی و تعیین ساختار پدیده، از مدل سازی استفاده می شود.

این مدل‌ها احتمالاً بیانگر همه واقعیت‌های مرتبط با آن‌ها نباشند و البته ضمن ساده بودن، باید تا حد امکان مؤلفه‌های مهم تأثیرگذار در پدیده‌ها، هنگام مدل‌سازی‌ها در نظر گرفته شود. در علوم غیرپزشکی نیز از مدل‌ها به‌وفور استفاده می‌شود. ماکت کره زمین (کره جغرافیایی) یا نقشه‌های راهنما، انواعی از مدل‌ها یا الگویی از واقعیت جغرافیایی جهان موجود هستند. این مدل‌ها تصور بهتر همه مناطق جغرافیایی جهان را ممکن و برای استفاده در آموزش جغرافیا و یا دسترسی به مناطق مختلف و تبیین مناطق همسایه کاربرد دارند. مدل‌سازی فقط جنبه‌های خاصی از واقعیت را به انتخاب و هنر مدل‌ساز ارائه می‌دهد. به‌عنوان مثال در علم اقتصاد یک مدل‌ساز برای نشان دادن تعادل "مقدار کالا" و "قیمت" از دو منحنی عرضه و تقاضا استفاده می‌کند. این مدل بسیاری از امور واقعی مؤثر بر تعادل بازار را در نظر نمی‌گیرد و فقط سه فاکتور قیمت، تقاضا و عرضه را در نظر دارد. در مدل‌سازی درک جهان خارج، از طریق حواس پنج‌گانه، تعقل و تدبیر انجام می‌شود. مدل‌ها نمایش مجددی از واقعیت‌ها هستند که می‌توانند گاهی به‌طور دقیق واقعیت‌ها را بیان دارند و یا به علل اشتباهات مدل‌ساز و یا ورودی‌های اشتباه و ناقص، پاسخ‌های غیردقیق یا دور از واقع ارائه نمایند. اعتبار یک مدل بستگی به میزان واقع‌گرایی آن دارد. مدل‌سازی دارای توالی و مستلزم برقراری ارتباط بین جهان واقعیت، مدل، مدل‌ساز و کار تیمی با تخصص‌های موردنیاز است. سپس تفسیر واقعیت هستی بر اساس درک ما، مدل‌سازی می‌شود.

### در فرآیند و تعریف مدل‌سازی سه نکته زیر مورد توافق متخصصین این فیلد است:

- ۱- واقعیتی از طریق عقل، تفکر و حواس به فرد یا افراد مدل‌ساز (تیم/گروه) منتقل می‌شود.
  - ۲- مدل‌ساز بر اساس تفکرات و دانش خویش روابط بین متغیرها را برقرار و اطلاعات را تحلیل و تفسیر می‌کند.
  - ۳- فرد، تفسیر (درک) خود را به‌عنوان واقعیت یا مدل ارائه می‌دهد.
- بنابراین می‌توان گفت "مدل‌سازی فرآیندی است که در شبکه عصبی مغز با تفکر آغاز و با درک واقعیات، فرموله کردن تجربیات، پردازش و نمایش مجدد اطلاعات جهان خارج ادامه می‌یابد. نتیجه این فرآیند مدل نامیده می‌شود".

از واژه مدل در سه نقش اسم، صفت و فعل استفاده می‌شود. در نقش اسم مدل بیانگر یک مفهوم ذهنی و یا شکل ذهنی است. مدلی که یک معمار از یک ساختمان بیمارستانی در مقیاس کوچک‌تر یا یک فیزیکدان از یک اتم یا یک ویروس‌شناس از یک ویروس کووید-۱۹، مدلی در اندازه بزرگ‌تر می‌سازد مثالی از مدل‌ها در نقش اسم هستند.

در نقش صفت، مدل به صفات عالی و مطلوب و نمونه‌ای شاخص و الگو اشاره دارد. خانه بهداشت مدل و یا دانشجوی مدل که بیانگر الگوی کامل و شاخصی برای ساختن یک خانه بهداشت ایده آل و یا آموزش و تربیت یک دانشجوی نمونه یا یک انسان سالم و یا یک بیمارنا واجد شاخص‌های خاص مثالی از مدل‌ها در نقش صفت است.

در نقش فعل، مدل نمایش دادن بخشی از واقعیت و فرآیند ایجاد یک پدیده یا رخداد است. مدل آماری عوامل تعیین کننده سکتة قلبی در ایران یا مدل ریاضی سیر پیشرفت سرطان در یک بیمار مبتلا به سرطان معده یا مدل عوامل تعیین کننده طغیان اسهال خونی در یک شهر یا مدل عوامل تعیین کننده و پیش بینی کننده کووید-۱۹ در ایران نمونه هایی از مدل ها در نقش فعل است.

مدل ها انواع مختلفی از جمله فیزیکی (شمایلی و قیاسی) و نمادین (کلامی، ریاضی و آماری) دارند و در علوم مختلف کاربرد متعددی دارند. یکی از انواع آن ها که در علوم پزشکی و سلامت بسیار پر کاربرد و با اهمیت هستند مدل های ریاضی و آماری است. این مدل ها با استفاده از حروف، اعداد، شبیه سازی ها، توابع، عمل گراها و روابط ریاضی، جهان واقعی و متغیرهای مرتبط و تعیین کننده یک حالت یا پیامد سلامتی و بیماری را در مراحل مختلف پیش آگهی، تشخیص، درمان و پیشگیری را به تصویر می کشند یا پیش بینی می کنند.

این مدل ها بسیار انتزاعی و قابل انعطاف بوده و بیشترین میزان کاربرد را در سیاست گذاری و تصمیم گیری مبتنی بر شواهد در نظام سلامت دارند. اما در انتهای مقدمه و تعاریف مدل سازی باید جمله معروف و منتسب به آمارشناس انگلیسی ([George E. P. Box](#)) که در شکل زیر نشان داده شده است را نیز همواره در استفاده از مدل ها و ترجمان دانش آن ها مورد نظر داشت. "همه مدل ها اشتباه هستند، اما برخی از آنها مفید هستند" یعنی هر مدلی اشتباه است. زیرا متغیرهای زیادی را در نظر نمی گیرند و واقعیت را ساده سازی می کنند.

برخی از مدل ها، به ویژه در علوم "سخت" یا زمینه های خاصی از علوم طبیعی فقط کمی اشتباه دارند. آن ها مواردی مانند اصطکاک یا اثر گرانشی اجسام کوچک را نادیده می گیرند. مدل های دیگر بسیار اشتباه هستند - آن ها چیزهای بزرگ تر را نادیده می گیرند. در علوم اجتماعی، ما خیلی نادیده می گیریم. - ساده سازی واقعیت می تواند بسیار مفید باشد. آن ها می توانند در توضیح، پیش بینی و درک جهان و همه اجزای مختلف آن به ما کمک کنند. این فقط در آمار درست نیست! نقشه ها نوعی مدل هستند. آن ها اشتباه می کنند اما نقشه های خوب بسیار مفید هستند.



تصویر آمارشناس انگلیسی ([George E. P. Box](#))

## ضرورت مدل سازی ها

- ۱) آشکار نمودن روابط بین اجزای سازنده مدل.
- ۲) درک بیشتر از اجزای مدل و متغیرها.
- ۳) امکان تجزیه و تحلیل ریاضی آن.
- ۴) امکان پذیر نمودن اجرای فرایندهای واقعی و شبیه سازی آنها.
- ۵) قابلیت استفاده در آموزش و پیشگیری.
- ۶) قابلیت استفاده در تصمیم سازی ها و تصمیم گیری ها.
- ۷) مدل های بیماری نقش مهمی در درک و مدیریت پویایی انتقال عوامل مختلف بیماری را دارند. ما می توانیم از آنها برای توصیف الگوهای مکانی و زمانی شیوع بیماری ها و همچنین کشف یا درک بهتر عواملی که بر بروز عفونت تأثیر می گذارند استفاده کنیم. در درک اینکه چه روش های درمانی و مداخله ای می توانند مؤثرتر باشند یا این روش ها چقدر مقرون به صرفه هستند و هنگام تلاش برای کنترل، حذف و ریشه کنی بیماری چه فاکتورهای خاصی باید در نظر گرفته شود استفاده از مدل ها ضروری است. از مدل سازی با حداقل هزینه های مورد نیاز می توان برای هدایت سیاست اجرای راه حل های عملی در دنیای واقعی استفاده کرد.

## فرآیند مدل سازی در سلامت و بیماری ها

مدل سازی در یک فرآیند ۸ مرحله ای قابل انجام است:

- ۱- **تعریف مسئله یا مشکل مرتبط با سلامتی:** تعیین سؤال پژوهش، درک، توضیح، تبیین، تعیین وضعیت موجود، تشریح و شناخت واقعی و درک مسئله، انجام یک بررسی متون و مرور سیستماتیک درباره مشکل مرتبط با سلامتی. موضوع اپیدمی کووید-۱۹ در ایران مانند سایر کشورهای دنیا به عنوان یک مشکل اولویت دار نیازمند مدل سازی و پیش بینی است. اطلاعات ما درباره اپیدمی کووید-۱۹ ناقص و با گذشت زمان تکمیل می گردد.
- ۲- **جمع آوری داده ها:** انجام یک پژوهش اولیه یا ثانویه و جمع آوری پارامترها و فاکتورهایی که بر مسئله مورد بررسی تأثیرگذار هستند. انجام مطالعات اپیدمیولوژیک و استفاده از مشاهده، مصاحبه، پرسشنامه، نمونه گیری و به طور کلی اجرای گام های یک پژوهش آکادمیک. در واقع نیاز به فراهمی درون داده های مورد نیاز و معتبر برای مدل سازی از جمله ضرورت های مدل سازی کووید-۱۹ است. نادانسته های ما درباره این بیماری نوپدید و عدم قطعیت های آن نیز در مدل سازی اثرگذار بوده و به آن نیز باید توجه نمود.
- ۳- **ساخت و توسعه مدل:** نیاز به مهارت و فنون اپیدمیولوژی، آمار زیستی، ریاضی، فناوری اطلاعات و نرم افزارها و آمیخته نمودن آنها با تخیل، هنر، هنرمندی و تجربه، تعیین روابط متقابل بین متغیرها. دیدگاه ها و راه های مختلفی برای ساخت مدل ها وجود دارد. برای ساخت یک مدل، فرموله کردن مدل و در واقع تعیین اجزا متشکله یک مدل، ورودی ها و خروجی ها، متغیرهای تصمیم و قابل کنترل (متغیرهایی که مبنای تصمیم گیری قرار

می گیرند و با داشتن مقادیر مختلف و متعدد از آنها، مقدار منطقی و با نظر تیم مدل ساز تعیین می شوند.)، متغیرهای غیرقابل کنترل و پارامتر (متغیرهای مربوط به سطح جامعه و اینکه اطلاع چندانی از توزیع آنها در دست نیست)، اندازه های عملکرد، متغیرهای نتیجه، متغیرهای مستقل و وابسته، مخدوش کننده ها، متغیرهای برهمکنش کننده، متغیرهای کولایدر و متغیرهای ترکیبی مرتبط با موضوع مدل سازی را باید شناخت و سپس به مدل سازی اقدام نمود. در جهان واقعی تنها یک راه صحیح برای ساخت یک مدل وجود ندارد. در یک سال گذشته از شروع پاندمی کووید-۱۹ تاکنون بیش از یک صد مدل ارائه شده و مقالات متعددی در این زمینه منتشر شده است. به عنوان مثال در [این مقاله](#) حداقل ۳ مدل ( Gompertz, Von Bertalanffy, and Least Squared ) (Error (LSE) در سال ۱۳۹۸ توسط نگارنده این متن برای پیش بینی کووید-۱۹ در ایران معرفی و استفاده شده است. یا مدل های دیگر را [در این لینک](#) و [این لینک](#) ببینید. دو مدل بیشتر از بقیه مدل ها به دلیل دینامیک بودن و در دسترس قرار داشتن پیش بینی ها و خروجی ها یا کار با مدل و مدل سازی کووید-۱۹ مورد علاقه بیشتر پژوهشگران و مدیران در نظام سلامت قرار گرفته است. مدل اول مربوط به موسسه سنجش و ارزیابی سلامت در آمریکا (IHME) است که خروجی های آن برای همه کشورها در دسترس بوده و می توان پیش بینی ها را از وبسایت موسسه و [این لینک](#) مشاهده نمود. در گزارش سالیانه اپیدمی کووید-۱۹ در ایران که توسط تهیه کنندگان این کتاب تهیه شده بود، خروجی های این مدل قبلاً گزارش و در [این لینک](#) در دسترس است. مدل دوم مربوط به دانشگاه آکسفورد و معروف به CommoModel است که از [این لینک](#) قابل استفاده است. در مدل دوم مدل ساز می تواند ورودی ها مورد نیاز مدل و از جمله متغیرهای ابتلا و مرگومیر کووید-۱۹ به همراه مداخلات انجام گرفته را وارد اپلیکیشن یا نرم افزار R نموده و با کالیبره کردن و اعتبارسنجی به خروجی ها دست پیدا نماید. در واقع مدل آکسفورد از نظر آموزشی و سیاست گذاری کاربرد بیشتری از مدل اول دارد. راهنمای مدل سازی با مدل آکسفورد برای کووید-۱۹ توسط نگارنده این کتاب تهیه و از [این لینک](#) قابل دانلود است.

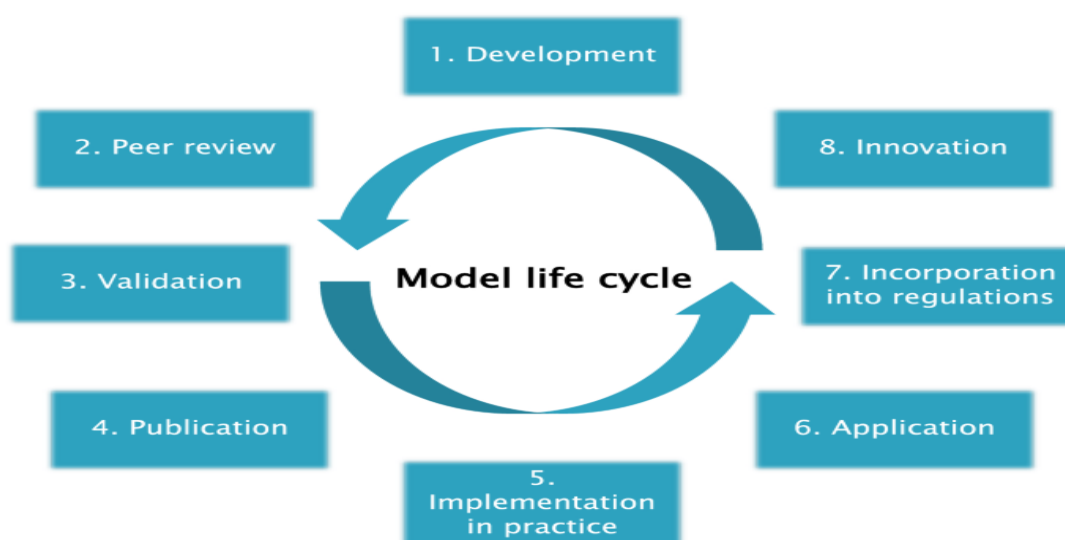
**۴- ارزیابی اعتبار و بررسی صحت عملکرد مدل:** این مرحله از مدل سازی و به ویژه در به کارگیری در سلامتی و جهان واقعی بسیار مهم است. مشاوره با متخصصین علوم بالینی مرتبط با موضوع مدل سازی در سلامت و همچنین متخصصین علوم پایه مرتبط، احتمال پذیرش مدل ساخته شده را افزایش می دهد. بنابراین ارزیابی پیش فرض های مدل سازی و ممیزی و صحت عملکرد نیاز به تأیید دارد. بازنگری داده ها، محاسبه شاخص های ارزیابی مدل، صحت، دقت، روایی و پایایی، بازنگری پیچیدگی های مدل، ساده بودن مدل و تعداد کمتر متغیر، نیاز به زمان کمتر برای اجرا و استفاده دارد.

**۵- حل مدل:** پیاده سازی تصمیمات یا استراتژی های توصیه شده توسط مدل در این گام مورد توجه است. بهینه کردن مدل، استفاده از نرم افزارهای مورد نیاز و گزارش الگوریتم ها و دستورات بکار رفته، سناریوسازی، تحلیل حساسیت و تعیین پارامترهای حساس و در صورت نیاز اصلاح و تغییر در مدل و نهایی کردن مدل سازی.

**۶-ارائه نتایج مطالعه:** در اختیار قرار دادن نتایج مدل سازی برای کاربران و سفارش دهندگان. با توجه به اینکه کاربران با مفاهیم مدل سازی آشنایی دقیقی ندارند و ممکن است به سختی مدل را متوجه یا قبول کنند توصیه می شود از ابتدا درگیر مراحل مدل سازی گردند.

ارائه نتایج بر اساس سطوح پیشگیری در نظام سلامت بسیار کمک کننده است. کدام یک از متغیرهای تعیین کننده مدل را می توان در آموزش به مردم، غربالگری، پیش آگهی، تشخیص، درمان و کنترل بیماری بکار بست؟

**۷ و ۸- اجرا و به کارگیری مدل و خلق ابتکار و نوآوری:** به کارگیری مدل در عمل با انجام شش مرحله گذشته بسیار مهم است و می توان با مدل سازی به نوآوری و خلاقیت های بیشتر دست پیدا نمود و مدل ها را توسعه و چرخه مدل سازی را گسترش داد. این چرخه در شکل زیر قابل مشاهده است. در این چرخه مراحل اصلی مدل سازی در ۸ گام نمایش داده شده است.



### آشنایی با مدل سازی با نرم افزار Stata

از منوی **statistics** گزینه های زیر را مشاهده نمایید. مسیر انتخاب مدل در زیر معرفی شده است.

Choice models	Setup and utilities	Declare data to be choice model data
Time series	Conditional logit (McFadden's choice) model	Tabulate choice sets
Multivariate time series	Mixed logit model	Summarize variables by chosen alternatives
Spatial autoregressive models	Panel mixed logit model	Tabulate chosen alternatives
Longitudinal/panel data	Multinomial probit model	Display reasons for sample exclusion
Multilevel mixed-effects models	Nested logit model	
Survival analysis	Rank-ordered logit model	
Epidemiology and related	Rank-ordered probit model	

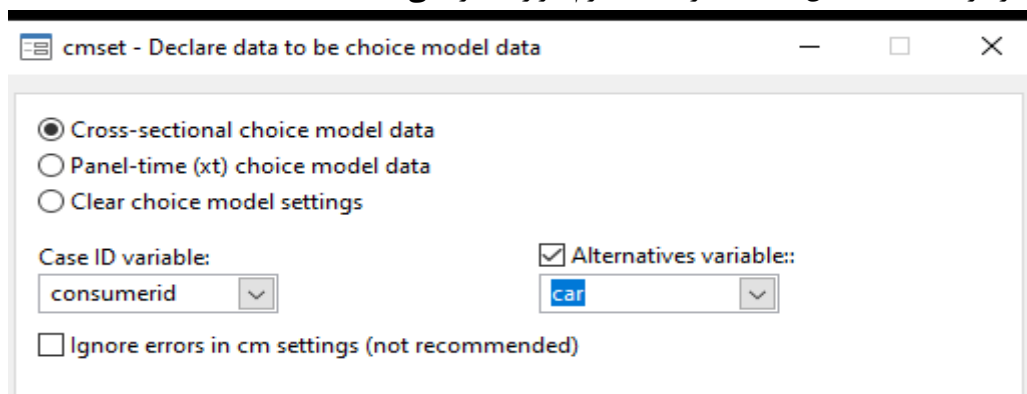
۵ مرحله را باید قبل از انتخاب مدل ها برای مدل سازی انجام دهید. این مراحل عبارتند از معرفی داده ها به نرم افزار، تهیه جدولی از متغیرها، خلاصه کردن آمار بر حسب متغیر جایگزین، تهیه جدول برای متغیر جایگزین و نمایش دلایل برای خارج کردن نمونه.

برای کار با نرم افزار فایل carchoice را باز نمائید. ابتدا ساختار متغیرها را به نرم افزار معرفی نمائید. متغیرهای این فایل را با دستور `list consumerid car purchase if consumerid <= 4, sepby(consumerid) abbr(10)` ببینید. در این فایل ساختار داده ها به صورت Cross-sectional یا مقطعی است.

	consumerid	car	purchase
1.	1	American	1
2.	1	Japanese	0
3.	1	European	0
4.	1	Korean	0
5.	2	American	1
6.	2	Japanese	0
7.	2	European	0
8.	2	Korean	0
9.	3	American	0
10.	3	Japanese	1
11.	3	European	0
12.	4	American	1
13.	4	Japanese	0
14.	4	European	0

متغیر consumerid به عنوان مصرف کننده، متغیر شناسه مورد است. ماشین متغیر جایگزین ها را تعریف می کند. این داده های ساختگی، نشان دهنده افرادی است که خودرو را با انتخاب خود بر اساس ملیت سازنده، آمریکایی، ژاپنی، اروپایی یا کره ای خریداری کرده اند.

جعبه زیر (db cmset) این مسیر را نشان می دهد. با دستور `cmset consumerid car` شروع کنید. این دستور فایل داده را برای انتخاب مدل مناسب در حافظه نرم افزار ذخیره می کند.



بسته به اینکه داده ها به صورت مقطعی یا به صورت پانل و در چند نوبت اندازه گیری شده باشند یکی از گزینه ها را انتخاب کنید. اگر مشاهدات شماره گذاری نشده بودند با دستور `gen id=_n` مشاهدات را شماره گذاری و متغیر ID را بسازید. مشاهدات یکسان و تکراری یک شماره باید اختصاص داده شود. این داده ها را با نام `carchoice_cmset` ذخیره کنید.

```
. cmset consumerid car
note: alternatives are unbalanced across choice sets; choice sets of different sizes found.
```

```
Case ID variable: consumerid
Alternatives variable: car
```

```
. cmset
note: alternatives are unbalanced across choice sets; choice sets of different sizes found.
```

```
Case ID variable: consumerid
Alternatives variable: car
```

دستور `cmchoiceset` را اجرا کنید.



Tabulation of choice-set possibilities

Choice set	Freq.	Percent	Cum.
1 2 3	380	42.94	42.94
1 2 3 4	505	57.06	100.00
Total	885	100.00	

ممکن است داده هایی که باز می کنید نرم افزار پیام خطا بدهد. فایل `carchoice_errors` را باز کنید.

```
. cmsset consumerid car
at least one choice set has more than one instance of the same alternative
r(459);
```

باید از دستور `cmsset consumerid car, force` استفاده کنید. بعد از آن خطا رفع می شود.

```
. cmsset consumerid car, force
note: at least one choice set has more than one instance of the same alternative.
```

Case ID variable: consumerid  
Alternatives variable: car

دستور `cmsample, generate(flag)` را اجرا کنید تا محل خطا را با ستاره نمایش دهد.

```
. cmsample, generate(flag)
```

Reason for exclusion	Freq.	Percent	Cum.
observations included	3,153	99.78	99.78
repeated alternatives within case*	7	0.22	100.00
Total	3,160	100.00	

\* indicates an error

با دستور `list consumerid car flag if flag != 0, sepby(consumerid) abbr(10)` مشاهدات دارای خطا را لیست کنید.

consumerid	car	flag
397.	111 American	repeated alternatives within case*
398.	111 Japanese	repeated alternatives within case*
399.	111 Japanese	repeated alternatives within case*
1035.	290 American	repeated alternatives within case*
1036.	290 Japanese	repeated alternatives within case*
1037.	290 Japanese	repeated alternatives within case*
1038.	290 Korean	repeated alternatives within case*

برای برخی از مدل ها و برآورد پارامترها مانند مدل `cmrologit` نیاز به متغیر جایگزین نیست و می توانید دستور `cmsset consumerid, noalternatives` را اجرا نمایید.

```
. cmsset consumerid, noalternatives
note: alternatives are unbalanced across choice sets; choice sets of different sizes found.
```

Case ID variable: consumerid  
Alternatives variable: <none>

ساختار داده های پانل در زیر معرفی می گردد. فایل `transport` را باز کنید. سه متغیر در این فایل

وجود دارد. متغیر t زمان است. در ۳ زمان تعریف گردید. متغیر alt نیز متغیر جایگزین است.  
دستور list id t alt if id == 1, sepby(t) اجرا کنید.

```
. list id t alt if id == 1, sepby(t)
```

	id	t	alt
1.	1	1	Car
2.	1	1	Public
3.	1	1	Bicycle
4.	1	1	Walk
5.	1	2	Car
6.	1	2	Public
7.	1	2	Bicycle
8.	1	2	Walk
9.	1	3	Car
10.	1	3	Public
11.	1	3	Bicycle
12.	1	3	Walk

```
. cmset id t alt
```

note: case identifier \_caseid generated from id and t.

note: panel by alternatives identifier \_panelaltid generated from id and alt.

```

      Panel data: Panels id and time t
      Case ID variable: _caseid
      Alternatives variable: alt
Panel by alternatives variable: _panelaltid (strongly balanced)
      Time variable: t, 1 to 3
      Delta: 1 unit

```

Note: Data have been xtset.

دستورات زیر را اجرا کنید.

```
sort id t alt
```

```
list id t alt _caseid _panelaltid if inlist(id, 1, 2), sepby(t) abbr('')
```

با دستور xtset داده ها را به حافظه نرم افزار معرفی کنید.

```
. xtset
```

```

Panel variable: _panelaltid (strongly balanced)
Time variable: t, 1 to 3
Delta: 1 unit

```

```
. cmset id t, noalternatives
```

note: case identifier \_caseid generated from id and t.

```

      Panel data: Panels id and time t
      Case ID variable: _caseid
Alternatives variable: <none>
      Time variable: t, 1 to 3
      Delta: 1 unit

```

## محاسبه شاخص های آماری خلاصه برای مدل سازی

فایل `carchoice` را باز کنید. دستور `cmset consumerid car` را اجرا کنید.

```
. cmset consumerid car
note: alternatives are unbalanced across choice sets; choice sets of different sizes found.
```

```
Case ID variable: consumerid
Alternatives variable: car
```

این داده های ساختگی نشان دهنده افرادی است که خودرو را با انتخاب خود بر اساس ملیت سازنده، آمریکایی، ژاپنی، اروپایی یا کره ای خریداری کرده اند. آمارهای خلاصه بر اساس ملیت خودرو محاسبه می شود. دوم، `cmsummarize` را تایپ می کنیم که به طور پیش فرض میانگین ها را محاسبه می کند. با تعیین متغیر درآمد، درآمد ناشی از ملیت خودروی خریداری شده محاسبه می شود.

```
. cmsummarize income, choice(purchase)
```

```
Statistics by chosen alternatives (purchase = 1)
```

```
income is constant within case
```

```
Summary for variables: income
Group variable: _chosen_alternative (purchase = 1)
```

_chosen_alternative	Mean
American	40.52394
Japanese	43.15127
European	45.80462
Korean	35.585
Total	42.05429

با دستور `cmsummarize gender income dealers, choice(purchase) statistics(N min mean max)` می توان جزئیات بیشتری را محاسبه و گزارش نمود.

```
. cmtab, choice(purchase)
```

```
Tabulation of chosen alternatives (purchase = 1)
```

Nationality of car	Freq.	Percent	Cum.
American	384	43.39	43.39
Japanese	326	36.84	80.23
European	135	15.25	95.48
Korean	40	4.52	100.00
Total	885	100.00	

بیشتر افراد خودرو آمریکایی خریداری کردند. با دستور `cmtab gender, choice(purchase)` بر اساس جنسیت نیز گزارش می شود. مردان بیشتر از زنان خودرو خرید نمودند و این رابطه معنی دار است.

```
Tabulation for chosen alternatives (purchase = 1)
```

```
gender is constant within case
```

Nationality of car	Gender: 0 = Female, 1 = Male		Total
	Female	Male	
American	96	280	376
Japanese	110	206	316
European	22	108	130
Korean	8	32	40
Total	236	626	862

دستور را اجرا نمائید. `cmtab gender, choice(purchase) row chi2 nokey transpose`

## روش های اپیدمیولوژی و پژوهش با Stata

```
. cmtab gender, choice(purchase) row chi2 nokey transpose
```

```
Tabulation for chosen alternatives (purchase = 1)
```

```
gender is constant within case
```

Gender: 0 = Female, 1 = Male	Nationality of car				Total
	American	Japanese	European	Korean	
Female	96 40.68	110 46.61	22 9.32	8 3.39	236 100.00
Male	280 44.73	206 32.91	108 17.25	32 5.11	626 100.00
Total	376 43.62	316 36.66	130 15.08	40 4.64	862 100.00

```
Pearson chi2(3) = 17.6654 Pr = 0.001
```

در این مرحله داده ها برای مدل سازی آماده هستند. فایل **transportation** را باز نمایید.

```
. list in 1/12, sepby(t)
```

	id	t	alt	choice	trcost	trtime	age	income	parttime
1.	1	1	Car	1	4.14	0.13	3.0	3	Full-time
2.	1	1	Public	0	4.74	0.42	3.0	3	Full-time
3.	1	1	Bicycle	0	2.76	0.36	3.0	3	Full-time
4.	1	1	Walk	0	0.92	0.13	3.0	3	Full-time
5.	1	2	Car	1	8.00	0.14	3.2	5	Full-time
6.	1	2	Public	0	3.14	0.12	3.2	5	Full-time
7.	1	2	Bicycle	0	2.56	0.18	3.2	5	Full-time
8.	1	2	Walk	0	0.64	0.39	3.2	5	Full-time
9.	1	3	Car	1	1.76	0.18	3.4	5	Part-time
10.	1	3	Public	0	2.25	0.50	3.4	5	Part-time
11.	1	3	Bicycle	0	0.92	1.05	3.4	5	Part-time
12.	1	3	Walk	0	0.58	0.59	3.4	5	Part-time

دستور `cmset id t alt` را اجرا کنید.

مدل سازی را با دستور

```
cmxtmixlogit choice trcost, random(trtime) casevars(age income)
```

سرفصله خروجی نشان می دهد که مدل ما با استفاده از داده های ۵۰۰ پانل مناسب بوده است. هر پانل در مثال ما یک فرد است. هر گروه/شخص در هر ۳ دوره زمانی، انتخابی را انجام داد، بنابراین ۱۵۰۰ مورد وجود دارد. از آنجا که هر مورد دارای ۴ جایگزین است، در کل ۶۰۰۰ مشاهده وجود دارد.

برآورد ضریب ثابت روی متغیر `trcost` برابر ۰.۸۴ است. میانگین برآورد تصادفی برای متغیر `trtime` برابر با ضرایب ۱.۵۱ - است. انحراف استاندارد تخمینی (`sd(trtime)`) برابر ۱/۹۵، و ناهمگونی قابل توجهی در این ضرایب نسبت به افراد جامعه وجود دارد. هر متغیر موردی دارای ضرایب متفاوتی برای هر گزینه است. این ضرایب مشابه `logit` چند جمله ای معمولی تفسیر می شوند. به عنوان مثال، ضرایب برآورد شده جایگزینی دوچرخه و عموم، در مورد سن مثبت است وقتی خودرو جایگزین اصلی است به این معنا که افزایش سن باعث می شود که افراد از وسایل نقلیه عمومی استفاده کنند یا از دوچرخه بجای ماشین شخصی استفاده کنند. به طور مشابه، افزایش درآمد باعث می شود کمتر کسی از وسایل نقلیه عمومی یا دوچرخه به جای ماشین استفاده کند.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. cmxtmixlogit choice trcost, random(trtime) casevars(age income)
```

Fitting fixed parameter model:

Fitting full model:

```
Iteration 0: log simulated-likelihood = -1025.707 (not concave)
Iteration 1: log simulated-likelihood = -1014.2513
Iteration 2: log simulated-likelihood = -1006.2212
Iteration 3: log simulated-likelihood = -1005.9904
Iteration 4: log simulated-likelihood = -1005.9899
Iteration 5: log simulated-likelihood = -1005.9899
```

```
Mixed logit choice model      Number of obs      =      6,000
                              Number of cases      =      1,500
Panel variable: id           Number of panels   =       500

Time variable: t             Cases per panel:  min =        3
                              avg =       3.0
                              max =        3

Alternatives variable: alt    Alts per case:   min =        4
                              avg =       4.0
                              max =        4

Integration sequence:        Hammersley
Integration points:          594      Wald chi2(8)      =      432.68
Log simulated-likelihood = -1005.9899  Prob > chi2      =      0.0000
```

choice	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
<b>alt</b>						
trcost	-.8388216	.0438587	-19.13	0.000	-.9247829	-.7528602
trtime	-1.508756	.2641554	-5.71	0.000	-2.026492	-.9910212
/Normal sd(trtime)	1.945596	.2594145			1.498161	2.526661
<b>Car</b> (base alternative)						
<b>Public</b>						
age	.1538915	.0672638	2.29	0.022	.0220569	.2857261
income	-.3815444	.0347459	-10.98	0.000	-.4496451	-.3134437
_cons	-.5756547	.3515763	-1.64	0.102	-1.264732	.1134222
<b>Bicycle</b>						
age	.20638	.0847655	2.43	0.015	.0402426	.3725174
income	-.5225054	.0463235	-11.28	0.000	-.6132978	-.4317131
_cons	-1.137393	.4461318	-2.55	0.011	-2.011795	-.2629909
<b>Walk</b>						
age	.3097417	.1069941	2.89	0.004	.1000372	.5194463
income	-.9016697	.0686042	-13.14	0.000	-1.036132	-.7672078
_cons	-.4183279	.5607111	-0.75	0.456	-1.517302	.6806458

```
. margins
```

```
Predictive margins      Number of obs = 6,000
Model VCE: OIM
```

```
Expression: Pr(alt), predict()
```

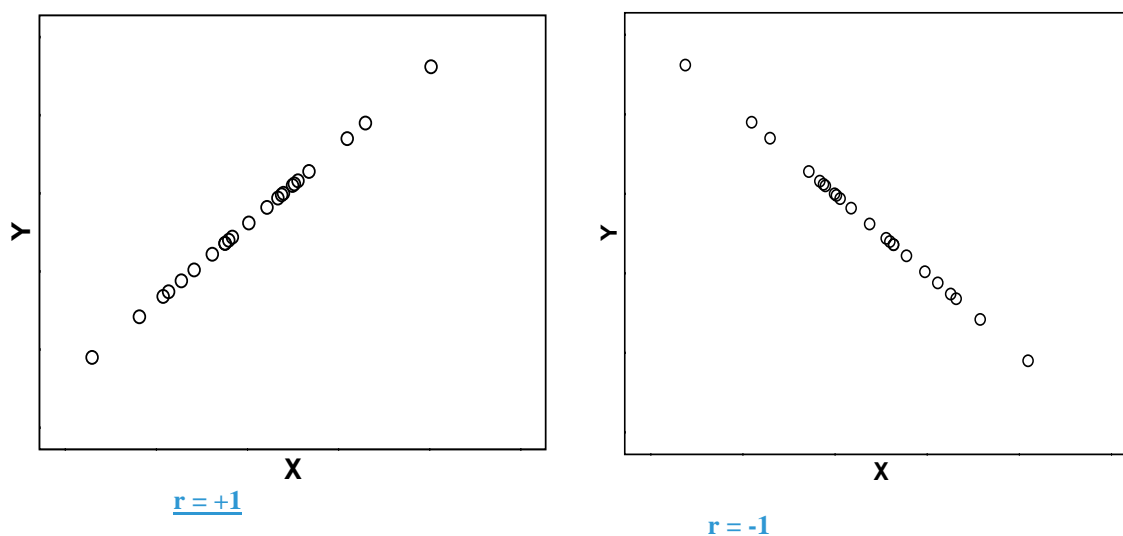
	Delta-method		z	P> z	[95% conf. interval]	
	Margin	std. err.				
_outcome						
Car	.6546572	.009617	68.07	0.000	.6358082	.6735062
Public	.1699625	.008265	20.56	0.000	.1537635	.1861615
Bicycle	.0951967	.0069522	13.69	0.000	.0815706	.1088227
Walk	.0801836	.0065606	12.22	0.000	.0673252	.0930421

در فصل های بعد مدل سازی های دیگری نیز ارایه می شود.

## همبستگی Correlation

همانگونه که در فصول گذشته توضیح داده شد برای بررسی ارتباط بین متغیرها از شاخص های مختلفی برای بررسی و محاسبه ضریب همبستگی استفاده می شود. ضریب مهم در بررسی روابط بین متغیرهای کمی که شرایط پارامتریک را داشته باشند، ضریب همبستگی پیرسون است. متغیرهای مورد بررسی باید در سطح مقیاس فاصله ای و یا نسبتی باشند. رابطه بین آنها خطی باشد. با نمودار پراکنش این رابطه قابل مشاهده است. شکل توزیع دو متغیر مشابه باشد. وقتی شرایط پارامتریک وجود نداشته باشد از ضریب همبستگی اسپیرمن و ضریب همبستگی تاوی بی کندال استفاده می شود. ضرایب مذکور فقط رابطه و همبستگی متغیرها را گزارش می کنند. از منهای یک تا یک مثبت را به خود می گیرند. چنانچه در بررسی روابط بین متغیرها نیاز به کنترل متغیرهای مخدوش کننده باشد از ضریب همبستگی جزئی یا Partial استفاده می شود. این ضرایب در فصول گذشته با مثال کاربردی و رسم نمودار پراکنش برای درک چگونگی جهت روابط توضیح داده شد.

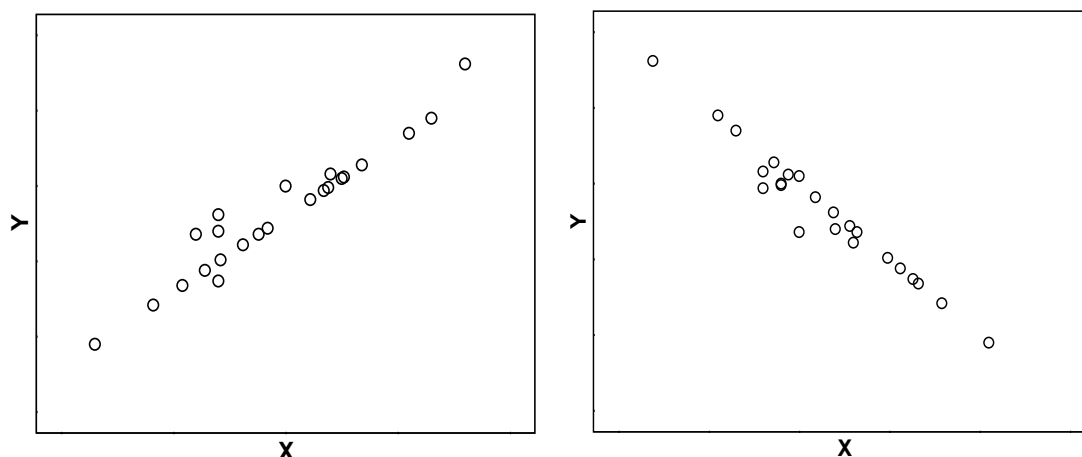
خلاصه اینکه در تجزیه و تحلیل همبستگی به اندازه گیری قدرت رابطه بین متغیرها پرداخته می شود. و شناخت متغیرهای مستقل و وابسته مد نظر نیست. ضریب همبستگی ( $r$ ) دارای ویژگی های زیر می باشد: مقدار آن بین  $-1$  و  $+1$  قرار دارد. اگر  $r = +1$  یا  $r = -1$  باشد، ارتباط کاملاً خطی است و تمام نقاط در نمودار پراکنش دقیقاً روی خط مستقیم قرار می گیرند. همبستگی مثبت کامل ( $r = +1$ ) و همبستگی منفی کامل ( $r = -1$ ) به ترتیب در شکل های زیر نمایش داده شده است.



فرمول ضریب همبستگی به صورت زیر تعریف می شود :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

این فرمول ها نسبت به  $X$ ,  $Y$  متقارن هستند. بنابراین همبستگی  $Y$ ,  $X$  مثل همبستگی  $X$ ,  $Y$  است. اگر  $r$  عددی بین  $0$  و  $+1$  یا بین  $0$  و  $-1$  باشد نقاط در اطراف خط، پراکندگی دارند. شکل های زیر به ترتیب  $r > 0$  و  $r < 0$  را نشان می دهد. هر چقدر  $r$  به یک نزدیک تر باشد پراکندگی کمتر است.



اگر  $r = 0$  باشد هیچ ارتباط خطی بین  $X$ ,  $Y$  وجود ندارد. این ممکن است بدین معنا باشد که اصلاً هیچ ارتباطی بین این دو متغیر وجود ندارد، ولی هم چنین ممکن است  $r = 0$  باشد ولی یک ارتباط غیر خطی (منحنی) بین  $X$  و  $Y$  موجود باشد. مجذور این ضریب یعنی  $r^2$ ، عبارت است از نسبتی از تغییرات متغیر  $Y$  که از طریق ارتباط خطی آن با متغیر  $X$ ، تبیین می شود. برای مثال، ضریب همبستگی  $DBP$  و  $BMI$ ،  $0.49$  است بنابراین حدود یک چهارم ( $0.49^2 = 0.24$ ) از واریانس  $DBP$  از طریق  $BMI$  تبیین می شود. برای محاسبه ضریب همبستگی علاوه بر دستورات و مسیرهای قبلی از دستور  $corr$  نیز استفاده می شود. در خروجی این دستور که قبل از دو متغیر کمی بکار می رود حدود اطمینان  $95\%$  نیز گزارش می شود.

### فرضیاتی که مبنای روش های کارکرد ضریب همبستگی هستند عبارتند از:

ضریب همبستگی برای هر نوع اطلاعاتی از نمونه گیری انجام شده از جمعیت قابل محاسبه است. نمونه باید نماینده جمعیت مورد مطالعه باشد. بنابراین برای اعتبار آزمون فرضیه ی مربوط، باید فرض شود که دو متغیر مشاهده شده نمونه های تصادفی از افراد هستند و این که اطلاعات حداقل برای یکی از متغیرها دارای توزیع نرمال در جامعه باشد. اگر اطلاعات دارای توزیع نرمال نباشند برای نرمال شدن می توان یکی از متغیرها یا هر دو را تبدیل کرد (برای مثال گرفتن لگاریتم). و اینکه رابطه دو متغیر خطی و تغییر پذیری واریانس آنها مشابه باشد. اما چنانچه هدف در بررسی روابط بین متغیرها، پیش بینی یا برآورد مقدار یک متغیر از روی متغیر یا متغیرهای دیگر باشد نیاز به انجام تحلیل های رگرسیونی است. انتخاب مدل رگرسیونی مناسب همانند آنالیز واریانس در فصل قبل، بر اساس جنس و ماهیت متغیر وابسته است. وقتی ما مجموعه ای از اطلاعات برای دو متغیر پیوسته داشته باشیم ممکن است سؤالات دیگری مطرح شود که دیگر نتوان با ضریب همبستگی پاسخ داد. بنابراین به مدل های رگرسیونی نیاز است. انواع مدل های رگرسیونی در زیر مورد بحث قرار می گیرند.

## Linear Regression

## رگرسیون خطی

وقتی محقق دنبال بررسی و تاثیر چندین متغیر مستقل بر یک متغیر عددی یا کمی با توزیع نرمال است از رگرسیون خطی استفاده می نماید. به عبارت دیگر متغیر مستقل کمی بر اساس یک یا چند متغیر مستقل دیگر پیش بینی می شود. بعضی وقتها ما می خواهیم ارتباط بین دو متغیر را خلاصه کنیم و بدین طریق بتوانیم مقدار یک متغیر برای یک فرد را براساس مقدار متغیر دیگر فرد (که برای ما مشخص است) پیش بینی کنیم که انجام این عمل به کمک همبستگی امکان پذیر نیست. چرا که ضریب همبستگی فقط قدرت ارتباط را تنها از طریق یک عدد نشان می دهد بنابراین ما به تکنیک دیگری به نام رگرسیون نیاز داریم. در رگرسیون وقتی رابطه یک متغیر مستقل با متغیر وابسته سنجیده شود رگرسیون خطی ساده گفته می شود. معادله آن به صورت زیر است.  $X$  متغیر مستقل و  $y$  متغیر وابسته است.

$$y = \alpha + \beta_1 x_1$$

وقتی پیش بینی متغیر وابسته بر اساس بیش از یک متغیر مستقل انجام شود رگرسیون خطی چندگانه نامیده می شود. معادله آن به صورت زیر نوشته می شود.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

برای مثال وقتی رابطه فشارخون دیاستول (DBP) را به عنوان متغیر پاسخ و نمایه توده بدنی (BMI) را به عنوان متغیر مستقل بررسی کنید، این ارتباط را از طریق یک خط مستقیم که از این داده ها می گذرد می توان نشان داد. هر خط مستقیمی که روی یک نمودار کشیده می شود با معادله  $y = a + bx$  نمایش داده می شود.  $Y$  نشان دهنده ی مقدار متغیر پاسخ و  $X$  نشان دهنده ی متغیر کمکی یا مستقل می باشد. این معادله به ما نشان می دهد که  $x$  ,  $y$  چگونه به هم مرتبط می باشند.  $a$ ، عرض از مبدأ، نقطه ای است که خط محور  $y$  ها را قطع می کند و مقدار  $y$  به ازای  $x=0$  می باشد.  $b$ ، شیب خط، افزایش (یا کاهش) در مقدار  $y$  به ازای هر واحد افزایش در  $x$  می باشد. برآورد حداقل مربعات شیب رگرسیونی عبارت است از :

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$b$  نسبت به  $x$  ,  $y$  متقارن نیست. بنابراین رگرسیون  $y$  روی  $x$  در حالت کلی با رگرسیون  $x$  روی  $y$  یکسان نیست. وقتی که مقدار عرض از مبدأ ( $a$ ) از یک معادله خطی به دست می آید. با علم به این که می دانیم خط رگرسیون از نقطه ی  $x = \bar{x}$  و  $y = \bar{y}$  می گذرد:  $a = \bar{y} - b\bar{x}$  برازش خط مستقیمی به این اطلاعات که بهترین برآورد  $y$  را برای هر مقدار  $x$  بدهد مهمترین نکته در این معادله است. می توان یک خط به صورت چشمی کشید ولی یک تکنیک بهتر (objective) استفاده از روش حداقل مربعات است. این انتخاب  $a$  ,  $b$  به طریقی است که فاصله ی عمودی نقاط از خط رگرسیون کمترین مقدار باشد. هر یک از این فواصل تفاضل مقدار مشاهده شده فرد از مقداری که از طریق خط رگرسیون به دست می آید، می باشد. برای دقت بیشتر ما  $a$  ,  $b$  را به طریقی



انتخاب می کنیم که مجموع مجذورات این فواصل عمودی کمترین مقدار شود. خط برآورد شده، خطی است که مجموع مربعات باقی مانده ها را می نیمم می کند. به نظر عجیب می رسد که از این روش به عنوان معیاری برای بهترین خط رگرسیون استفاده شود. البته می توان از روش های دیگر مثل مجموع قدر مطلق مقادیر باقی مانده ها استفاده نمود. اما این مسئله بدین خاطر است که این روش ها به ریاضیات در سطح بالاتری نیاز داشته که برای این کتاب مناسب نیست. به خصوص اگر ارتباط در جامعه واقعاً یک خط راست باشد برآورد پارامترهایی که ما از رگرسیون به دست می آوریم نا اریب هستند و دارای کمترین واریانس در بین تمام برآورد کننده های خطی نا اریب ممکن می باشند. علاوه بر آن اگر انحراف ها از خط مستقیم به صورت نرمال توزیع شده باشند، آن گاه برآوردهای حداقل مربعات هم چنین برآوردهای ماکزیمم درستنمایی نزدیک می باشند. خطی که به این روش به دست می آید خط رگرسیون نام دارد. خط رگرسیون متوسط مقدار  $y$  برای یک مقدار داده شده  $x$  را برآورد می کند. این خط از نقطه ای که به وسیله  $x$  میانگین  $x$  ها و میانگین  $y$  ها تعریف می شود  $(\bar{x}, \bar{y})$  می گذرد. شیب،  $(b)$  که ضریب رگرسیون نامیده می شود، متوسط افزایش (یا کاهش) در  $y$  به ازای یک واحد افزایش  $x$  را بیان می کند. معادلات برای برآوردهای حداقل مربعات  $a$  ,  $b$  در ادامه آمده است. وقتی داده های DBP و BMI آنالیز شوند مقادیر زیر برای عرض از مبدأ و شیب (ضرایب رگرسیونی) به دست می آید.

$$b = 1.046 \quad a = 64.5$$

بنابراین خط رگرسیونی حاصله عبارت است از :

$$DBP = 64.5 + 1.046 \times BMI$$

که در notation اخیر  $y$ ، DBP و  $x$  با BMI است.

این معادله ما را قادر به محاسبه ی مقادیر برآورد شده DBP برای هر مقدار نمایه توده بدن (BMI) می سازد. برای مثال برای یک BMI برابر با  $20 \text{ (kg/m}^2\text{)}$  مقدار برآورد شده DBP عبارت است از :

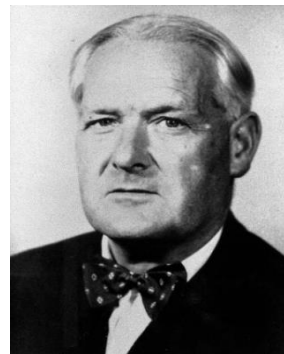
$$64.5 + (1.046 * 20) = 85.5 \text{ mmHg}$$

## همبستگی ها، ارتباطات، علت و معلول (Cause and Effect)

داشتن یک همبستگی خطی مستقیم، دلیلی بر وجود اثر علی و معلولی بین متغیرها نمی باشد. برای ممانعت از چنین مسائلی، از کلمات «ارتباط» و «همبستگی» استفاده می شود. بررسی رابطه علیت به چیزی فراتر از تجزیه و تحلیل اطلاعات بستگی دارد. برای درک بیشتر موضوع به معیارهای علیتی دال و هیل مراجعه شود.

این معیارها و موضوع علیت در **بخش پنجم** آمده است. در جعبه زیر این معیارها اشاره شده است.

**رهنمودهای علیتی Bradford Hill برای قضاوت درباره علیتی بودن یک ارتباط:**



- ۱- رابطه زمانی
- ۲- قدرت ارتباط
- ۳- رابطه دوز- پاسخ
- ۴- تکرارپذیری یافته ها
- ۵- همخوانی زیستی
- ۶- بررسی سایر توضیحات جایگزین
- ۷- توقف مواجهه
- ۸- سازگاری و همخوانی با سایر دانسته ها
- ۹- ویژگی و اختصاصی بودن رابطه

### آزمون فرضیه برای ضرایب رگرسیون و همبستگی

مثل بقیه ی قسمت های آنالیز آماری، معمولاً براساس اطلاعات نمونه ای، درباره ی ارتباط واقعی (جامعه) بین دو متغیر استنباط انجام می شود. برای مثال، اگرچه آنالیز بالا یک ارتباط نسبتاً خوبی را بین BMI و DBP پیشنهاد می کند ولی تا چه حد این یافته شانس است؟ به عبارت دیگر، چقدر احتمال دارد با بدست آوردن این نتایج، در واقعیت این دو متغیر به یکدیگر مرتبط نباشند؟ بررسی این که آیا هیچ کدام از همبستگی های خطی هستند یا این که (همه ی ارتباطات ظاهر شده فقط به خاطر شانس می باشد) از طریق آزمون فرضیه صورت می گیرد.

#### آزمون فرضیه ی ضریب رگرسیون $b = 0$ :

برآورد  $b$  بر خطای معیار تقسیم و نسبت حاصله را با توزیع  $t$  با  $(n-2)$  درجه آزادی مقایسه می کنند.

$$t = \frac{b}{se(b)}$$

$Se(b)$  خطای معیارشیب می باشد و به طور مستقیم از برنامه های کامپیوتری در رگرسیون خطی به دست می آید. نرم افزار Stata این مقدار را گزارش می کند.

#### آزمون فرضیه ی ضریب همبستگی $r = 0$ :

تحت فرض صفر که هیچ همبستگی در جامعه موجود نمی باشد می توان نشان داد که :

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

این معادله نیز دارای توزیع t با (n-2) درجه آزادی است که می توان آن را با جدول t، با همان روش معمول بررسی کرد.

مثال : برای اطلاعات DBP و BMI، خطای معیار b، 0.398 به دست آمده است.

در آزمون فرضیه ی صفر  $b = 0$  :

$$t = \frac{(1.046 - 0)}{0.398} = 2.63$$

$$t = 0.49 \sqrt{\frac{24-2}{1-0.49^2}} = 2.64$$

و آزمون کردن فرضیه ی صفر  $r = 0$  :

تفاوت بین این دو آزمون به خاطر گرد کردن خطا است. وقتی با هر دو مقدار  $t = 2.63$  یا  $t = 2.64$  به جدول توزیع t با 22 درجه آزادی مراجعه می شود مقدار  $P < 0.015$  به دست می آید که گواه نسبتاً خوبی از این است که رگرسیون و همبستگی واقعی مساوی صفر نیست. فواصل اطمینان برای شیب را می توان به روش معمول، با استفاده از توزیع t با (n-2) درجه آزادی به عنوان ضریب خطای استاندارد محاسبه کرد. در این مقدار برای مقادیر بزرگ n تقریباً برابر 1.96 می باشد).

در مثال بالا مقدار t با 22 درجه آزادی برابر 2.073 است و فاصله اطمینان 95% عبارت است از :

$$1.046 \pm (2.073 \times 0.398)$$

$$(0.22, 1.87) \text{ mmHg} \left( \frac{\text{kg}}{\text{m}^2} \right)$$

در مثال بالا، در شروع، بین متغیر وابسته یا پاسخ (DBP) و متغیر مستقل یا کمکی (BMI) تفاوت قائل شدیم. در بسیاری از موقعیت ها مشخص نیست که کدام یک متغیر پاسخ و کدام یک متغیر کمکی است. این تفاوت برای رگرسیون خیلی مهم تر از همبستگی است. برای مثال اگر BMI را به عنوان متغیر پاسخ و DBP را به عنوان متغیر کمکی در نظر بگیرید، باز هم همین ضریب همبستگی خواهد بود. اما خط رگرسیونی در این حالت کاملاً متفاوت خواهد شد. این مطلب تأکید می کند که رگرسیون خطی تنها برای موقعیت هایی مناسب است که این دو متغیر به طور روشنی از هم متمایز باشند. رگرسیون و همبستگی تنها به ما اجازه ی بررسی همبستگی خطی دو متغیر کمی را می دهد. حتی وقتی همبستگی قوی دیده می شود حتماً بدین معنا نیست که این ارتباط علی و معلولی است. این تمایز بین همبستگی و علی بودن باید همیشه در تفسیر نتایج آنالیز در مطالعات مشاهده ای در نظر گرفته شود. روش هایی که در این جا بررسی شد برای بررسی ارتباط خطی بود و به طور کلی برای آنالیز ارتباط های غیر خطی مناسب نمی باشند. برای ارتباط های غیر خطی مثبت یا منفی یک

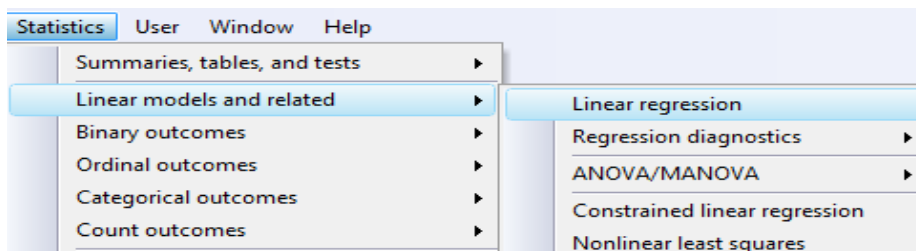
کمیت به نام ضریب همبستگی اسپیرمن یا رتبه ای می تواند مورد استفاده قرار گیرد. تحت شرایط خاصی از رگرسیون برای پیش بینی  $y$  در یک مقدار مشخص  $x$  استفاده می شود. توجه کنید که خط رگرسیون نباید برای پیش بینی ها، برای مقادیری از  $x$  که خارج از دامنه ی مقادیر اطلاعات مشاهده ای است مورد استفاده قرار گیرد. وقتی هیچ مدرکی در مورد این رابطه خارج از اطلاعات مشاهده ای ندارید چنین استنباطی بی مورد است.

### فرضیاتی که مبنای روش های آنالیز رگرسیونی هستند عبارتند از:

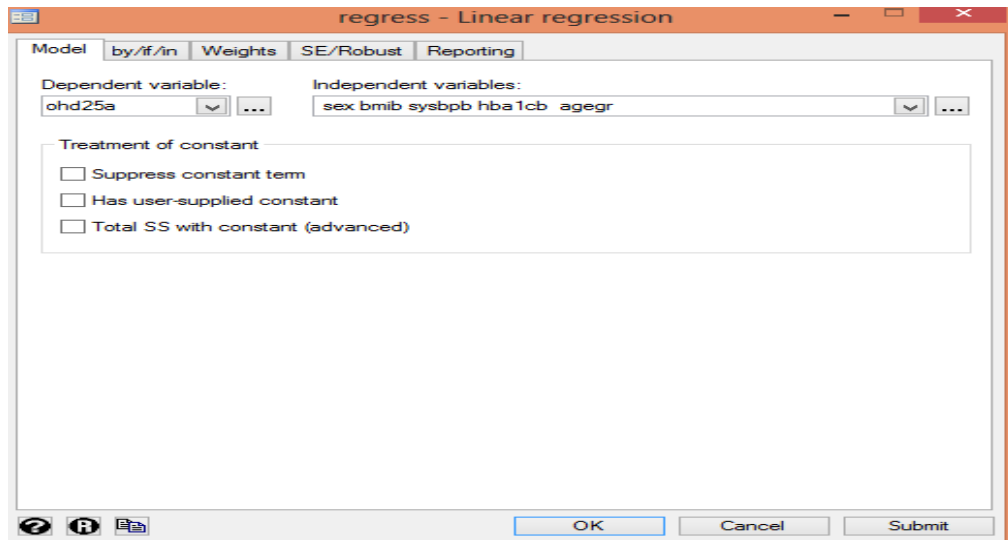
اولا نمونه گرفته شده باید نماینده و گویای جامعه مورد مطالعه باشد که قرار است درباره آن استنباط شود. مقادیر متغیر پاسخ برای هر مقدار متغیر کمکی باید دارای توزیع نرمال باشد. متغیر مستقل با متغیر وابسته ارتباط خطی داشته باشد. تغییر پذیری متغیر پاسخ که از طریق واریانس یا انحراف معیار ارزیابی می شود برای هر مقدار از متغیر کمکی باید یکسان باشد اگر دو فرضیه ی بالا برقرار نباشد بعضی اوقات مسئله از طریق تبدیل مناسب روی متغیر پاسخ حل می شود. متأسفانه پارامترهای رگرسیون  $(b, a)$ ، ضریب همبستگی  $(r)$ ، چیز معنی داری در سطح جامعه را اندازه گیری نمی کنند. برای مثال مقدار مشاهده شده ی  $r$  به سادگی با افزودن اطلاعات جدید دارای مقادیر  $x$  دور افتاده ای باشند افزایش پیدا کند.

### کار با نرم افزار و انجام رگرسیون خطی:

فایل Part five را باز نمایید. متغیرهای این فایل قبلا توضیح داده شدند. در این پژوهش فرضی می خواهید ارتباط بین سطح ویتامین دی (ohd25a) را با جنسیت (sex)، گروه های سنی (agegr)، نمایه توده بدنی (bmib) و فشارخون سیستولیک (sysbpb) و هموگلوبین گلیکوزیله (hba1cb) بررسی نمایید. متغیر وابسته کمی است. متغیرهای مستقل نمایه توده بدنی و فشارخون سیستولیک کمی، متغیرهای جنسیت و گروه های سنی نیز کیفی هستند. برای انجام رگرسیون یا با تایپ دستور db regress کادر محاوره ای رگرسیون خطی را باز نمایید و یا از طریق زیر عمل نمایید.



در کادر محاوره ای زیر متغیرهای مورد نظر در مدل تعریف و معرفی می شوند.



خروجی نرم افزار در زیر نمایش داده می شود. برای متغیرهایی که به صورت کیفی هستند در تجزیه و تحلیل بهتر است به نرم افزار معرفی گردند و قبل از نام متغیر عبارت  $I$  تایپ گردد. به این صورت نرم افزار این متغیر را به صورت متغیر صوری یا ظاهری (dummy) می شناسد و تحلیل را انجام می دهد. در نرم افزار وقتی متغیر دو یا بیشتر از دو حالت داشته باشد به صورت پیش فرض نرم افزار طبقه اول را به عنوان مرجع گرفته و بقیه طبقات را با آن مقایسه می کند. یا عبارتی مقادیر بزرگتر را نسبت به مقادیر کوچکتر می سنجد. در اینجا متغیر جنسیت دو طبقه ۱ برای مردان و ۲ برای زنان دارد. بنابراین زنان را با مردان می سنجد. متغیر گروه سنی نیز ۴ طبقه دارد که طبقه اول به عنوان رفرانس در نظر گرفته شده و بقیه طبقات با آن سنجیده می شوند.

```
. regress ohd25a i.sex bmib sysbpb hba1cb i.agegr
```

Source	SS	df	MS			
Model	42143.6756	7	6020.52509	Number of obs =	60	
Residual	347687.34	52	6686.29499	F( 7, 52) =	0.90	
Total	389831.015	59	6607.30534	Prob > F =	0.5133	
				R-squared =	0.1081	
				Adj R-squared =	-0.0120	
				Root MSE =	81.77	

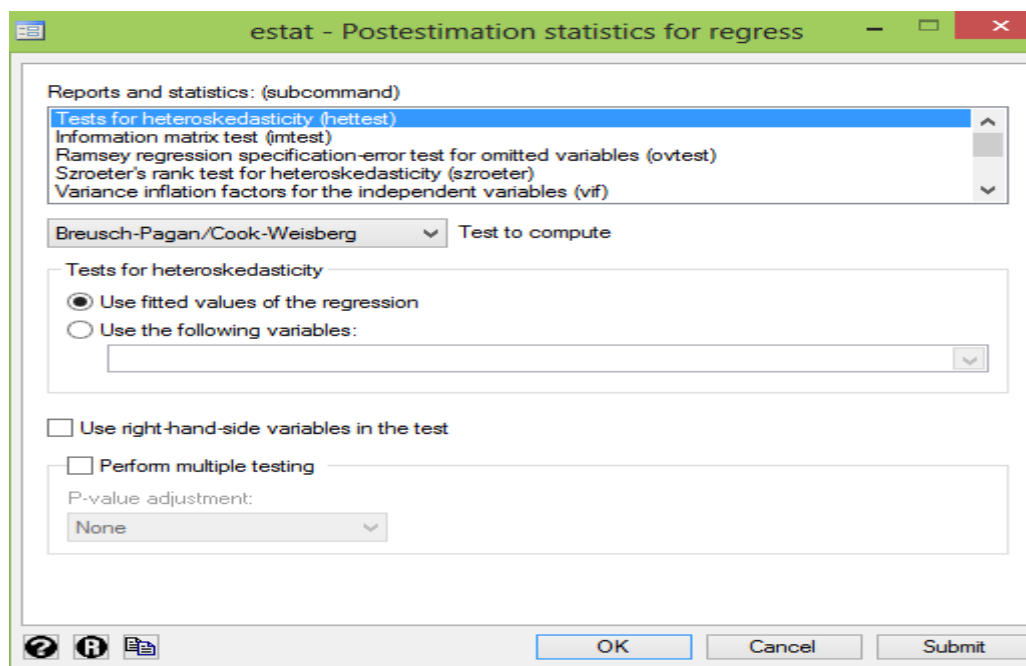
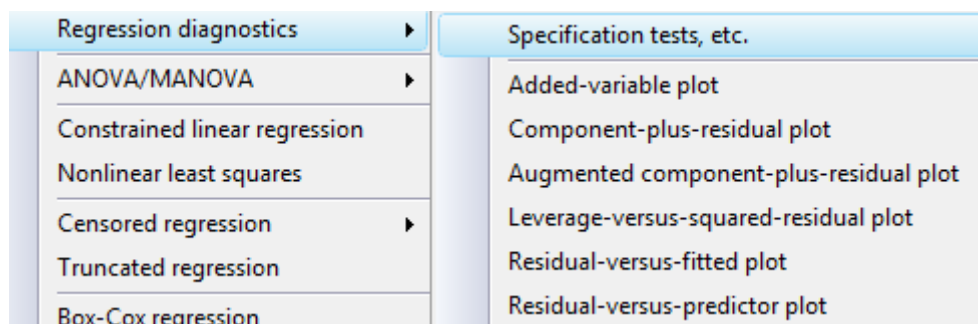
  

ohd25a	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
2.sex	7.483384	25.8329	0.29	0.773	-44.35412	59.32089
bmib	1.351909	2.334749	0.58	0.565	-3.333107	6.036925
sysbpb	-.6312408	.9827057	-0.64	0.523	-2.603184	1.340702
hba1cb	-18.419	22.94979	-0.80	0.426	-64.47112	27.63311
agegr						
2	82.4312	40.98379	2.01	0.049	.1912112	164.6712
3	69.29794	45.69863	1.52	0.135	-22.40307	160.999
4	48.89014	54.36273	0.90	0.373	-60.19667	157.9769
_cons	243.4478	219.3338	1.11	0.272	-196.6775	683.5732

همانگونه که خروجی بالا نشان می دهد اولاً نتایج به صورت کلی معنی دار نیست. ( $F=0.9$  و  $P=0.5133$ ). به عبارت دیگر رابطه ی مناسبی بین متغیرهای مستقل و وابسته مشاهده نشده است. در گروه های سنی سطح ویتامین دی متفاوت است. عدد 82.43 بیان گر این نکته است که در شرایط یکسان و ثابت بودن بقیه متغیرها، میانگین سطح ویتامین دی در گروه سنی دو حدود 82.43 میکرو از میانگین بقیه گروه های سنی بیشتر می باشد. در خروجی مذکور متغیر `-cons` عرض از مبدا یا مقدار ثابت است. برای نوشتن معادله رگرسیون این مقدار استفاده می شود. ضریب تعیین ( $R^2$ ) پایین و منفی است. ممکن است پیش فرض ها برقرار نباشند. لذا باید ارزیابی انجام گیرد.

### ارزیابی مدل رگرسیون خطی در Stata:

در منوی آنالیز رگرسیون مسیر زیر برای ارزیابی مدل رگرسیون برازش شده استفاده می شود. یا با دستور `estat` کادر محاوره ای مربوطه باز و ارزیابی انجام می شود. این مسیر گزینه پس از تخمین است.



در هریک از سطوح متغیرهای مستقل، واریانس مقادیر باقی مانده باید ثابت باشد. در ارزیابی بالا اولین گزینه میکند. نتایج این ارزیابی در زیر آمده است. این شرط تقریباً برقرار است.

```
chi2(1)      =      3.54
Prob > chi2   =      0.0598
```

می توان با آزمون ماتریس اطلاعات چولگی، کشیدگی و برابری واریانس ها را نیز ارزیابی نمود. این ارزیابی با گزینه دوم انجام می شود. آماره هایی که برای ارزیابی بکار می روند در شکل زیر که بعنوان قسمت اول کادر محاوره ای بالا می باشد آمده اند.

Reports and statistics: (subcommand)

```
Tests for heteroskedasticity (hettest)
Information matrix test (imtest)
Ramsey regression specification-error test for omitted variables (ovtest)
Sroeter's rank test for heteroskedasticity (szroeter)
Variance inflation factors for the independent variables (vif)
```

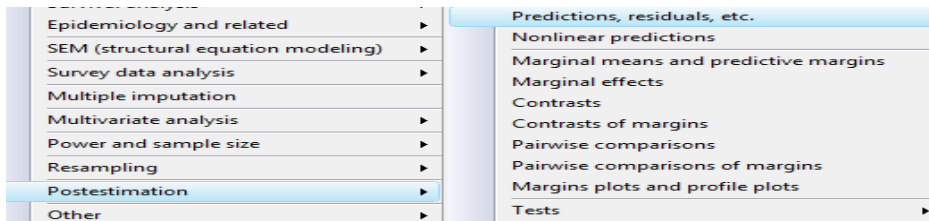
نتایج ارزیابی دوم در زیر آمده و شروط تقریباً برقرار نیست.

```
. estat imtest
```

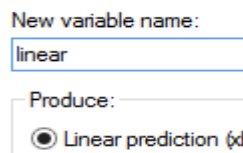
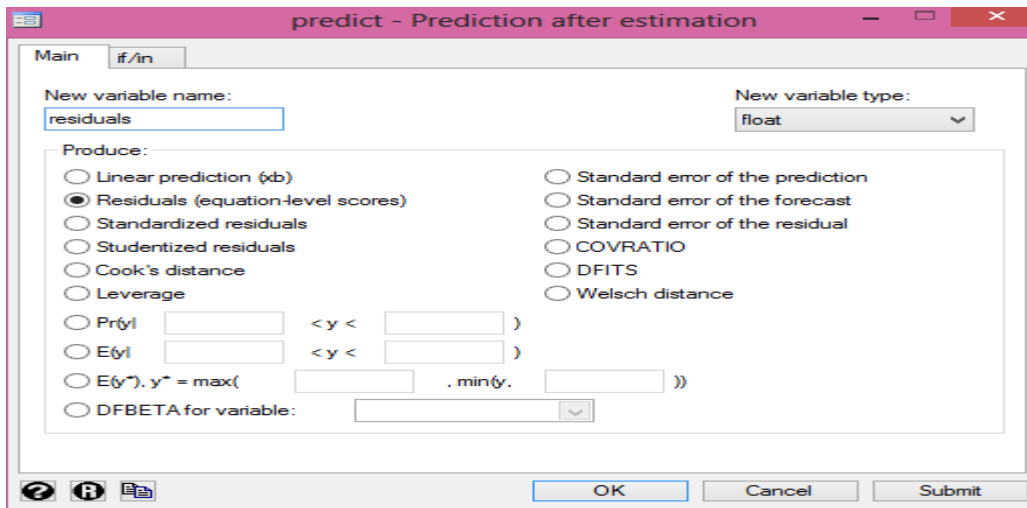
Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	2.78	1	0.0957
Skewness	6.86	1	0.0088
Kurtosis	0.87	1	0.3502
Total	10.50	3	0.0147

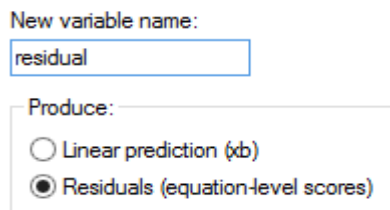
در بررسی روند رگرسیون انجام شده ابتدا نمودار پراکنش برای بررسی وضعیت داده ها باید ترسیم شود. داده های پرت شناسایی شوند. رابطه ها باید خطی باشند. متغیر وابسته توزیع نرمال داشته باشد. واریانس توزیع متغیر وابسته برای مقادیر متغیرهای مستقل یکسان باشد یا بعبارتی صفر نباشد. Homogeneity of Variance یا (homoscedasticity) برقرار باشد. یعنی واریانس مقادیر خطا ثابت باشد. و اینکه رابطه خطی قوی بین متغیرهای مستقل نباشد. این رابطه بوسیله شاخص تولرانس یا VIF سنجیده می شود. توضیحات این شاخص در قسمت تحلیل واریانس نیز ذکر گردید. در ادامه همین مبحث درباره این شاخص توضیحات لازم آمده است. این شاخص با دستور `estat vif` یا پنجمین گزینه در کادر محاوره ای ارزیابی مدل رگرسیون اجرا می شود. برای رسم منحنی مقادیر باقی مانده در مقابل مقادیر برازش شده از دستور `rvfplot` و برای رسم منحنی مقادیر باقی مانده و مقادیر پیش بینی شده برای هر متغیری مانند `sex` از دستور `rvpplot sex` استفاده می شود. همچنین می توان از مسیر زیر ارزیابی پس از تخمین مدل سازی رگرسیون را انجام داد. از منوی اصلی گزینه `postestimation` انتخاب گردد.



بعد از باز شدن کادر محاوره ای به شکل زیر ، نامگذاری متغیرها با دو گام ، انجام می شود.



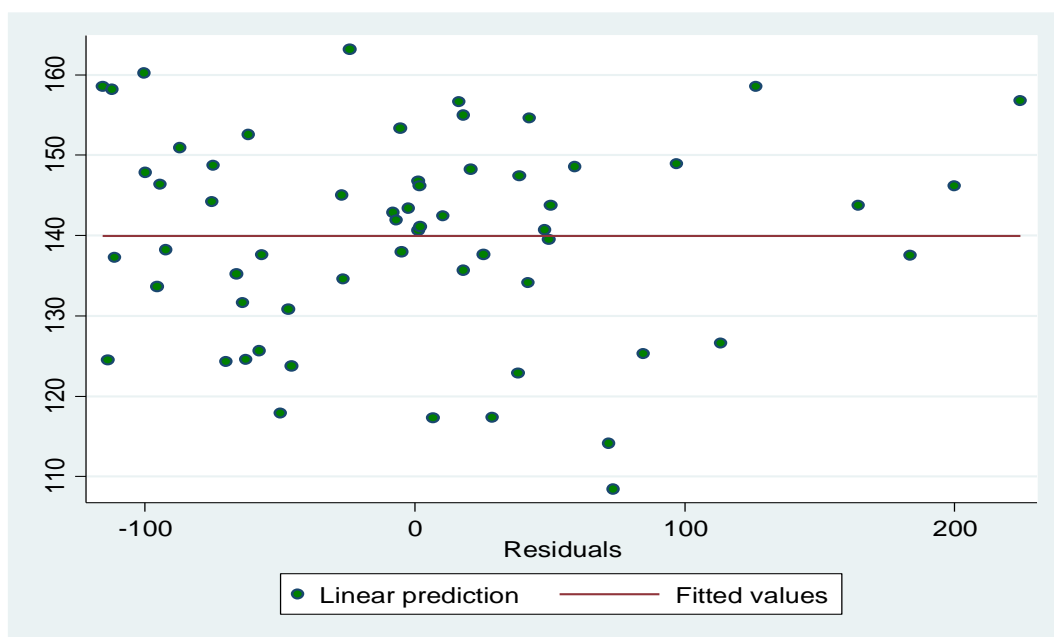
و یکبار هم متغیر دیگری نامگذاری می شود با انتخاب گزینه residuals به صورت زیر:



این دستورات دو متغیر با نام های داده شده ایجاد می کند که در پنجره متغیرها وجود دارند. اکنون می توان با نمودار پراکنش دو متغیر ساخته شده و ترسیم خط برازش آن را به صورت زیر مشاهده نمود.

twoway (scatter linear residuals, sort mfcolor(green)) (lfit linear residuals)





اگر معادله رگرسیون بین متغیرهای مورد بررسی کارایی مناسبی داشته باشد باید مجموع مربعات تبیین شده نسبت بزرگی از مجموع مربعات کل را تشکیل دهد. این نسبت همان ضریب تعیین است. برای مناسب بودن مدل در خروجی بالا مقدار R خام (برای نمونه آماری) و استاندارد شده (تعدیل شده برای جامعه آماری) پایین می باشد. بنابراین مدل نامناسب است.

مقدار مجذور R که ضریب تعیین نام دارد نشان می دهد که چند درصد از پراکندگی مشاهده شده یا واریانس سطح ویتامین دی توسط متغیرهای وارد شده به مدل توجیه می شود. مقدار ضریب هر متغیر نشان می دهد به ازای یک واحد افزایش در متغیر مستقل مربوطه (مثلا ۱ واحد افزایش به شاخص نمایه توده بدنی)، 1.36 واحد متغیر وابسته یا سطح ویتامین دی افزایش پیدا می کند.

### چگونگی ورود متغیرها به مدل رگرسیون :

در تحلیل می توان از دستورات مذکور (وارد کردن همزمان کلیه متغیرها) استفاده نمود یا با استفاده از دستورات زیر نوع مدل ورود متغیرهای رگرسیونی را تعریف نمود.

- Enter method: regress y x1 x2 x3 .... روش همزمان وارد کردن
- Stepwise met (روش پیشرونده) روش گام به گام وارد کردن
- Forward selection:
  - sw regress y x1 x2 x3 ....., pe(0.05) pr(0.1) forward
  - sw regress y x1 x2 x3 ....., pe(0.05)
- Backward elimination: روش حذف پس روند
  - sw regress y x1 x2 x3 ....., pe(0.05) pr(0.1)
  - Sw regress y x1 x2 x3 ....., pr(0.1)

## خلاصه و نتیجه گیری برای چک کردن مدل رگرسیون خطی:

- برای چک نمودن نرمال بودن مقادیر باقی مانده (Checking normality of Residuals) از دستورات زیر استفاده می شود.

- Histogram; **histogram r, normal**
- Kernel Density; **kdensity r, normal**
- Normality test; Shapiro-Wilk test; **swilk r**

- برای چک کردن هموژن بودن واریانس مقادیر باقی مانده (Checking homogeneity of Residual Variance) با استفاده از نمودار مقادیر باقی مانده در مقابل مقادیر پیش بینی شده از دستورات زیر استفاده می شود.

- **rvfplot, yline(0)**
- **estat imtest**
- **estat hettest**

- برای چک نمودن چند همخطی نبودن بین متغیرهای مستقل (Multicollinearity) از دستور زیر استفاده می شود.

- **Vairiance Inflation Factor; vif**

وقتی هم خطی کامل بین متغیرها وجود داشته باشد واریانس ضرایب برآورد شده به سمت بی نهایت میل نموده و در این صورت مقدار آماره تی نیز به سمت صفر میل خواهد نمود. لذا ممکن است علیرغم معنی دار بودن ضرایب کلی رگرسیون (آماره F) ضرایب تکی متغیرهای وارد شده به مدل از لحاظ آماری معنی دار نگردند. لذا از دستور بالا برای ارزیابی هم خطی بین متغیرها استفاده می شود. و عامل افزایش دهنده واریانس ضرایب برآورد شده محاسبه می شود. در این خصوص قاعده مشخصی وجود ندارد. اما مطالعات تجربی نشان داده است که اگر مقدار این شاخص (VIF) برای متغیر مستقلی بزرگتر از ۱۰ یا ۱۱ باشد نشانگر هم خطی آن متغیر با سایر متغیرهای مستقل است. یا به عبارتی آن متغیر دارای ترکیب خطی با سایر متغیرهاست. میانگین  $vif$  بالاتر از ۱ هم خطی بین متغیرها را نشان می دهد. هر چه مقدار عامل افزایش واریانس از عدد ۲ بزرگتر باشد میزان همخطی بیشتر است. هر چقدر مقدار تولرانس بیشتر باشد مقدار عامل تورم یا افزایش واریانس کمتر است و برعکس. برای کاهش یا حذف همخطی از تبدیل مقادیر متغیرهای مستقل به نمره استاندارد، استفاده از تحلیل عاملی برای متغیرهای مستقل و کاهش آنها، استفاده از لگاریتم، وارد کردن متغیر مجازی، استفاده از تابع اولین تفاضل و

یا کنار گذاشتن برخی متغیرهای دارای همبستگی بالا با متغیرهای دیگر می توان استفاده نمود. بسته به مهارت تحلیل کننده از هر یک از این روش ها برای حذف همخطی استفاده می شود.

- برای چک نمودن رابطه خطی (Linearity) ابتدا از نمودار پراکنش استفاده می شود.

برای رگرسیون ساده استفاده از : `scatter y x`

- برای رگرسیون چندگانه استفاده از : استاندارد شده مقادیر باقی مانده در مقابل هر متغیر مستقل

- `scatter r x1, Scatter r x2 , .....`
- `acprplot x1, lowess Isopts(bwidth(1))`

برای ارزیابی مناسب بودن حجم نمونه در آنالیز رگرسیونی از قانون زیر استفاده می شود.

Good and Hardin's Rule of Thumb:

- **n**: sample size
- **m**: number of observations are needed to precisely define a straight line (usually 5)
- **p**: number of predictor variables

با در نظر گرفتن تعداد متغیرهای پیش بینی کننده یا مستقل (P) و تعداد مشاهداتی که برای تعریف دقیق یک خط راست مورد نیاز است که معمولاً آن را ۵ می گیرند (m)، حجم نمونه مورد نیاز (n) با فرمول زیر محاسبه می شود. یا اینکه به ازای هر ۳۰ پاسخگو (آزمودنی) حداقل یک متغیر مستقل لازم است. نسبت تعداد متغیرهای مستقل به حجم نمونه کمتر از یک سی ام باشد.

$$p = \frac{\log n}{\log m}$$

وقتی شرط نرمال بودن برقرار نباشد باید متغیر را تبدیل و تغییر داد مثلاً با لگاریتم گرفتن و... و یا انجام رگرسیون غیرخطی مثلاً لوجستیک. وقتی شرط یکسان بودن مقادیر واریانس باقی مانده ها برقرار نباشد از مدل واریانس Robust نیز استفاده می شود. از دستورات زیر برای تحلیل می توان استفاده کرد.

- `regress y x1 x2 x3..., robust`
- `regress y x1 x2 x3..., vce(robust)`
- `rreg y x1 x2 x3 ...`

Logistic Regression

رگرسیون لوجستیک:

رگرسیون لوجستیک حالت خاصی از تحلیل های رگرسیون است که در مواردی که متغیر پاسخ، کیفی باشد مورد استفاده برای مدل سازی قرار می گیرد. به عبارت دیگر وقتی متغیر پاسخ دو گزینه ای یا دو حالتی (Dichotomous) باشد از رگرسیون دو حالتی یا دو جمله ای (Bionomial) استفاده می شود. وقتی متغیر پاسخ بیش از دو طبقه داشته باشد از رگرسیون چند جمله ای یا چندحالتی (Multinomial) استفاده می شود. در مبحث رگرسیون خطی برای پیش بینی متغیر وابسته از تابع خطی به صورت معادله زیر استفاده گردید.

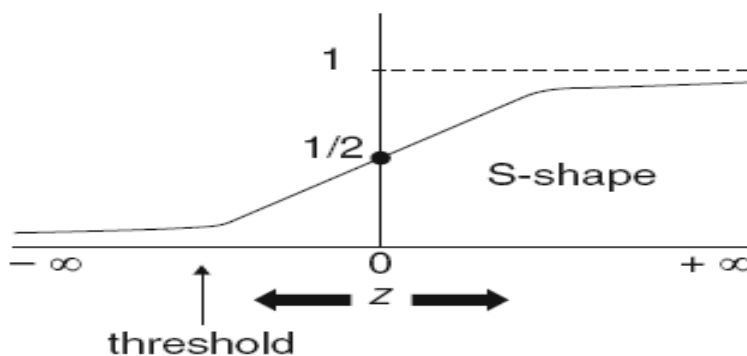
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

در اینجا از تابع لوجستیک استفاده می شود. این تابع به صورت زیر تعریف می شود.

$$Z = P(D) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

$$Z = P(D) = \frac{1}{1 + e^{-\sum \beta_i X_i}}$$

یا از تابع  $Z = P(D) = \frac{1}{1 + e^{-\sum \beta_i X_i}}$  که غیر خطی (nonlinear) و شکل آن به صورت "S shaped" بوده و در زیر نشان داده شده، استفاده می شود. همانگونه که در شکل مشخص است  $Z$  مقادیر بین ۰ تا ۱ و  $X$  مقادیر منفی بینهایت تا مثبت بی نهایت را می گیرد.



در این مدل، برای متقارن کردن  $Z$  به عنوان پیامد، ابتدا متغیر وابسته را به نسبت و سپس به یک متغیر لوجیت

$$\text{Logit } P(D) = \ln \left[ \frac{P(D)}{1 - P(D)} \right]$$

( لگاریتم طبیعی شانس های وقوع یا عدم وقوع) تبدیل می کنند. معادله به صورت خواهد شد. سپس لوجیت مقادیر  $-\infty$  تا  $+\infty$  را به خود می گیرد. تابع لوجستیک به صورت زیر است.

$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

مقادیر بتا در رگرسیون لجستیک برابر لگاریتم نسبت های شانس است. اگر آنتی لگاریتم این مقادیر گرفته شود معادل نسبت های شانس خواهند بود. البته نرم افزار Stata در قسمت نحوه گزارش نتایج گزینه ای دارد که با انتخاب آن گزینه مقادیر نسبت شانس گزارش خواهند شد. برای تبدیل تابع لوجیت به تابع خطی ، اگر تابع

$$\text{لوجیت } P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \text{ با شد، و تابع } 1 - P(X) \text{ به صورت زیر :}$$

$$\begin{aligned} 1 - P(X) &= 1 - \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \\ &= \frac{e^{-(\alpha + \sum \beta_i X_i)}}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \end{aligned}$$

آنگاه با محاسبه ساده و تقسیم دو فرمول بالا بر همدیگر خواهیم داشت:

$$\begin{aligned} \frac{P(X)}{1 - P(X)} &= \frac{\frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}}{\frac{e^{-(\alpha + \sum \beta_i X_i)}}{1 + e^{-(\alpha + \sum \beta_i X_i)}}} \\ &= e^{(\alpha + \sum \beta_i X_i)} \end{aligned}$$

اگر از طرفین لگاریتم گرفته شود. تابع خطی زیر بدست می آید.

$$\begin{aligned} \ln_e \left[ \frac{P(X)}{1 - P(X)} \right] &= \ln_e \left[ e^{(\alpha + \sum \beta_i X_i)} \right] \\ &= \underbrace{(\alpha + \sum \beta_i X_i)}_{\text{linear sum}} \end{aligned}$$

نهایتاً با توجه به محاسبات انجام شده معادله خطی زیر برای استفاده بدست می آید که در مدل رگرسیون لجستیک برای تفسیر ضرایب و مقدار نسبت های شانس بکار می رود.

$$\log \text{it } P(X) = \ln \text{ odds}(X) = \alpha + \sum \beta_i X_i$$

این مدل را به صورت زیر نیز می توان نوشت.

$$\ln \left[ \frac{P(y|x)}{1 - P(y|x)} \right] = \alpha + \beta x$$

اگر متغیر های مستقل بیش از یکی باشند رگرسیون چندگانه و اگر تنها یک متغیر مستقل باشد رگرسیون لجستیک ساده (**Simple logistic regression**) نامیده می شود. معادله رگرسیون لجستیک چندگانه (**Multiple logistic regression**) نیز به صورت زیر است.

$$\ln \left( \frac{P}{1-P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

رگرسیون لجستیک یک مدل ریاضی است که می تواند برای توصیف رابطه چندین متغیر  $X$  با یک متغیر وابسته دو حالتی یا چند حالتی (متغیری که فقط دارای دو یا چند وضعیت متفاوت است) به عنوان  $Y$  مورد استفاده قرار گیرد. منظور از متغیر دو حالتی (**binary**)، متغیری است که فقط دارای دو جواب باشد. مانند مردن یا زنده ماندن، کمربند ایمنی بستن یا نبستن، حاضر بودن یا غایب بودن و بیمار بودن یا بیمار نبودن. اغلب برای این متغیرها از کدهای صفر و یک استفاده می شود، کد یک را برای حالت مثبت بودن (موفقیت) آن خاصیت و کد صفر برای منفی بودن (شکست) آن به کار می رود. متغیر های مستقل می توانند کمی (سن)، کیفی اسمی (بلی یا خیر) یا رتبه ای (طبقه اجتماعی ضعیف، متوسط، بالا) باشند. معمولاً کاربرد این مدل برای هنگامی که متغیرهای مخدوش کننده زیادی وجود دارند بوده و می توان با این نوع تحلیل اثر سایر متغیرها را حذف و کنترل نمود.

#### اندازه های مهم در رگرسیون لجستیک:

نسبت شانس یا بخت یک اندازه اساسی و مهم در تحلیل رگرسیون لجستیک است. **شانس عبارتست از نسبت احتمال موفقیت به عدم موفقیت.** اگر شانس رخداد یک پدیده در دو گروه بر هم تقسیم شود آن را نسبت شانس (OR) می نامند. مفهوم OR مساوی ۳ برای رابطه چاقی و سرطان کولون این است که شانس سرطان کولون در افراد چاق ۳ برابر افراد غیرچاق است. نسبت شانس (Odds Ratio) یکی از شاخص های مهم و قابل محاسبه در این نوع تحلیل و مدل سازی به شمار می آید. که می توان آن را به صورت برآورد خام و یا تطبیق شده و کنترل سایر متغیرهای دیگر برآورد نمود. نسبت شانس را در همه مطالعات می توان محاسبه نمود. چنانچه احتمال وقوع یک بیماری ۲۰ درصد باشد. یا بعبارتی میزان شیوع آن ۲۰ درصد باشد شانس وقوع آن با تقسیم ۲۰ درصد بر ۸۰ درصد (شانس عدم وقوع بیماری) معادل ۰/۲۵ می شود. یا چنانچه فراوانی شیوع یک بیماری ۵۰ درصد باشد میزان شانس وقوع آن ۱ خواهد بود. **وقتی میزان شانس به احتمال مربوطه نزدیک می شود که وقوع یا رخداد پدیده نادرتر و فراوانی کمتر شود.** برای هنگامی که احتمال بیماری ۰/۱ باشد شانس ۰/۱۱ یا یک نهم می شود. وقتی احتمال بیماری یک دوم یا نصف است شانس بیماری ۱ می گردد. فرمول روابط شانس و احتمال برای تبدیل آنها به یکدیگر و مقادیر متفاوت آنها در زیر آمده است.

$$P(\text{probability}) = \text{odds} / 1 + \text{odds} \quad \text{و} \quad \text{Odds} = p / 1 - p$$

Probability $p$	Corresponding odds $o$
.0	.00
.1	.11
.2	.25
.3	.42
.4	.67
.5	1.00
.6	1.50
.7	2.33
.8	4.00
.9	9.00
1.0	$\infty$

وقتی مقادیر احتمال کوچک باشد معمولا شانس (Odds= $p/1-p$ ) تقریبا با احتمال ( $p$ ) برابر است. زیرا  $1-p$  تقریبا ۱ است. اغلب اپیدمیولوژیست ها برای پدیده های نادر، بجای شانس از خطر یا risk استفاده می کنند. و مقادیر شانس و احتمال را یکسان در نظر می گیرند. برای مقادیر  $p$  کمتر 0.1 مقادیر شانس در جدول زیر نشان داده شده است.

Probability $p$	Corresponding odds $o$
.1	.11...
.01	.0101...
.001	.001001...
.0001	.00010001...

برای محاسبه نسبت شانس یا OR با نرم افزار، فایل داده های مربوط به اتومبیل ها با نام use auto,clear را باز نمایید. برای بررسی اثر متغیرهای وزن و mpg بر foreign رگرسیون لجستیک انجام و OR مربوط به متغیرها را ملاحظه نمایید. سپس دستورات زیر را اجرا نمایید.

```
logistic foreign mpg weight          predict double p
replace mpg = mpg+1                  predict double q
gen double or = (q/(1-q))/(p/(1-p))  summarize or
```

خروجی دستورات بالا در زیر نشان داده شده است.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. logistic foreign mpg weight
```

```
Logistic regression                Number of obs   =       74
                                   LR chi2(2)         =      35.72
                                   Prob > chi2         =      0.0000
Log likelihood = -27.175156        Pseudo R2       =      0.3966
```

foreign	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
mpg	.8448578	.0776572	-1.83	0.067	.7055753 1.011635
weight	.9961009	.0010077	-3.86	0.000	.9941279 .9980779

```
. predict double p
(option pr assumed; Pr(foreign))
```

```
. replace mpg = mpg+1
(74 real changes made)
```

```
. predict double q
(option pr assumed; Pr(foreign))
```

```
. gen double or = (q/(1-q))/(p/(1-p))
```

```
. summarize or
```

Variable	Obs	Mean	Std. Dev.	Min	Max
or	74	.8448578	1.09e-15	.8448578	.8448578

همانگونه که ملاحظه می شود نسبت شانس محاسبه شده با نسبت شانس مدل لجستیک در بالا یکسان است. برای مقایسه نسبت خطر با نسبت شانس با استفاده از دستورات زیر محاسبات به صورت زیر انجام می شود.

```
. gen double rr=q/p
```

```
. summarize or rr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
or	74	.8448578	1.09e-15	.8448578	.8448578
rr	74	.8879504	.0457727	.8449534	.9816245

تقریباً برابر بودن مقادیر نسبت شانس و نسبت خطر در جدول بالا نشان داده شد. نسبت خطر مقدار اندکی از نسبت شانس بیشتر است.



مفروضات مدل رگرسیون لجستیک:

- متغیر وابسته حتما در سطح سنجش اسمی دو حالت باشد. اگر متغیر وابسته دو حالت ( Binary or Dichotomous) باشد از رگرسیون لجستیک (Binary Logistic)، اگر متغیر وابسته ترتیبی (Ordinal) باشد از رگرسیون رتبه ای یا Ordered Logistic و اگر متغیر وابسته اسمی چندوجهی (Multinomial) باشد از رگرسیون چندوجهی (Multinomial Logistic) استفاده می شود.

- متغیرهای مستقل می توانند کمی یا فاصله ای باشند. وقتی متغیرهای مستقل کیفی باشند باید آنها را به عنوان متغیر dummy به نرم افزار معرفی نمود.

- همانند رگرسیون خطی از مقادیر  $\beta$  به عنوان شاخص اثر گزارش می شود. چنانچه در مقیاس لگاریتمی ( Ln Scale ) بخواهید تفسیر کنید به این صورت است که ۱ واحد تغییر در متغیر مستقل (X)،  $e^\beta$  بار تغییر در شانس بیماری ایجاد می شود. اگر مقیاس نسبت شانس باشد تفسیر می شود ۱ واحد تغییر در متغیر X ،  $\beta$  بار تغییر در شانس بیماری (Odds of Disease) ایجاد می شود.

- وجود رابطه خطی (Linear relationship) میان متغیرهای پیش بین و لوجیت متغیر وابسته ( Logit of Outcome ) و چند هم خطی نبودن متغیرهای مستقل (Collinearity)

مثال کاربردی و تفسیر ضرایب رگرسیون لجستیک:

چنانچه در یک مطالعه مورد شاهد سه متغیر مستقل  $X_1$  تا  $X_3$  به ترتیب  $X_1$ : smoker (Binary) و  $X_2$ : age (Continuous) و  $X_3$ : HTN (Binary) به عنوان مواجهه برای بروز سکته قلبی در نظر گرفته شوند. نتایج مدل رگرسیون لجستیک با ضرایب زیر گزارش می گردد:

Case-control Printout	
Variable	Coefficient
Constant	$-4.50 = \hat{\alpha}$
$X_1$	$0.70 = \hat{\beta}_1$
$X_2$	$0.05 = \hat{\beta}_2$
$X_3$	$0.42 = \hat{\beta}_3$

در تفسیر عوامل موثر بر سکته قلبی با ضرایب رگرسیونی بالا به صورت زیر عمل می شود:

خطر سکته قلبی در کسانی که سیگار می کشند (متغیر  $X_1$ )، ( $e^{0.70} = 2.01$ ) بار بیشتر از کسانی است که سیگار نمی کشند. خطر سکته قلبی با افزایش ۱ سال به سن، ( $e^{0.05} = 1.05$ ) بار یعنی حدود ۵٪ خطر افزایش می یابد. خطر سکته قلبی در کسانی فشارخون بالا دارند نیز، ( $e^{0.42} = 1.52$ ) بار بیشتر از کسانی است که فشارخون بالا ندارند.

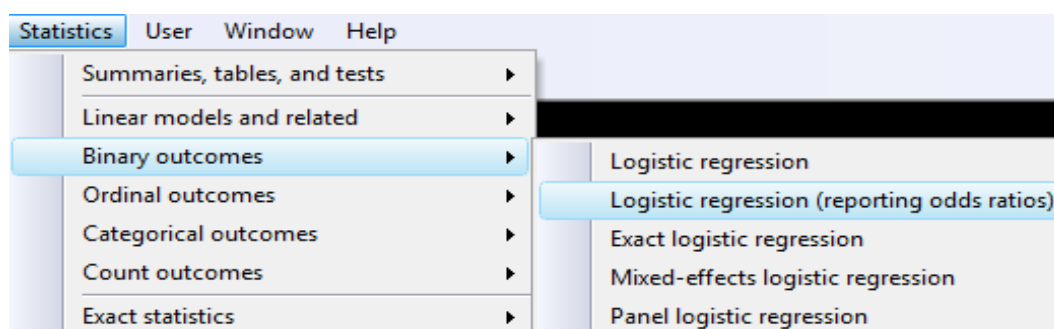
## کار با نرم افزار و انجام تحلیل رگرسیون لجستیک:

برای استفاده از رگرسیون لجستیک، فایل LOGISTIC را باز نمایید. متغیر MI بعنوان سکت قلبی و آن را به عنوان متغیر وابسته در نظر بگیرید.

متغیرهای دیگر از جمله استعمال سیگار (smk)، سن (age)، نوار قلب (ecg)، مقدار کلسترول (chl)، مقدار کاتکولامین (cat) و هپارین (hpt)، فشارخون سیستولیک (sbp) فشارخون دیاستولیک (dbp) نیز بعنوان متغیرهای مستقل هستند.

اینترکشن کلسترول با کاتکولامین با متغیر CC و اینترکشن هپارین با کاتکولامین با متغیر ch بکار رفته است.

برای آنالیز عوامل موثر بر سکت قلبی از دستور db logistic یا db logit استفاده و کادر محاوره ای مزبور را باز نمایید. یا از طریق منوی برنامه به صورت زیر اقدام نمایید.



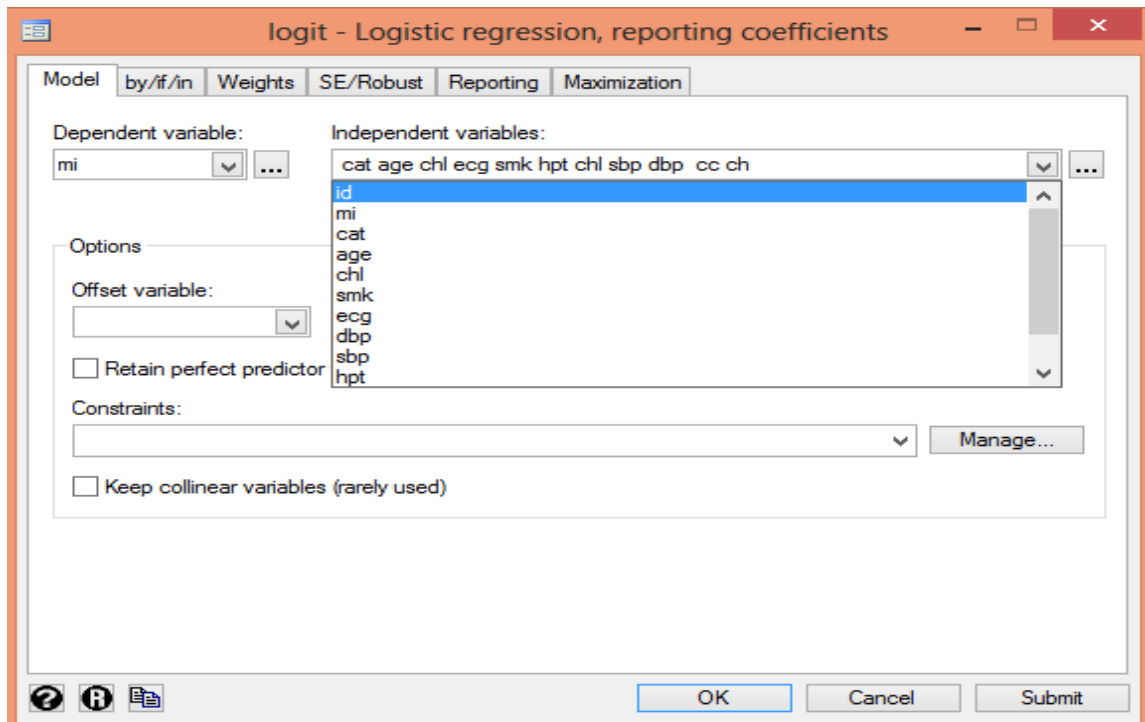
ابتدا با در نظر گرفتن اینترکشن ها متغیرها را با دستور زیر

```
logit mi cat age chl ecg smk hpt chl sbp dbp cc ch
```

یا با دستور db logit به صورت زیر در کادر محاوره ای (صفحه بعد) تعریف و مدل را اجرا (OK) می کنیم.

**نکته:**

وقتی دستور logit D X1 X2 X3.. بکار می رود مقادیر  $\beta$  در مقیاس ln scale گزارش می شوند. وقتی از دستور logistic D X1 X2 X3 استفاده گردد مقادیر ضرایب در مقیاس نسبت شانس گزارش می شوند.



نتایج این تحلیل در زیر آمده است. ۵ بار شبیه سازی انجام شده تا کمترین log lr بدست آمده است. ضرایب هر متغیر، آماره والد و مقدار معنی داری هر متغیر در جدول آمده است. تعداد مشاهدات یا حجم نمونه ۶۰۹ نفر بوده است. متغیر chl بخاطر داشتن collinearity حذف گردید. مدل راند شده برای متغیر های وارد شده معنی دار می باشد.  $Lr=92.64$  و مقدار  $p=0.0000$ .

می توان در کادر محاوره ای بالا در قسمت reporting این نوع گزارش را برای ضرایب مدل انتخاب نمود. یا در ادامه دستور مذکور عبارت or, را برای گزارش نسبت شانس ها می توان اضافه نمود. برای متغیر شاخص یا indicator و انجام رگرسیون رتبه ای یا چندوجهی در ابتدای دستور xi: به طریق زیر می آید.

xi: logistic D X1 X2 i.X3..

نتایج تحلیل در زیر آمده است.

## روش های اپیدمیولوژی و پژوهش با Stata

```

note: chl omitted because of collinearity
Iteration 0:   log likelihood = -219.27915
Iteration 1:   log likelihood = -183.63421
Iteration 2:   log likelihood = -173.80939
Iteration 3:   log likelihood = -172.96882
Iteration 4:   log likelihood = -172.96063
Iteration 5:   log likelihood = -172.96062
    
```

```

Logistic regression               Number of obs   =           609
                                LR chi2(10)         =           92.64
                                Prob > chi2         =           0.0000
Log likelihood = -172.96062      Pseudo R2       =           0.2112
    
```

mi	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cat	-12.75108	3.125892	-4.08	0.000	-18.87772	-6.624448
age	.0427362	.0175756	2.43	0.015	.0082886	.0771838
chl	-.005606	.0042128	-1.33	0.183	-.0138628	.0026509
ecg	.3794937	.3302582	1.15	0.251	-.2678004	1.026788
smk	.8008935	.3296377	2.43	0.015	.1548155	1.446971
hpt	1.013943	.4433979	2.29	0.022	.1448992	1.882987
chl	0	(omitted)				
sbp	-.0093386	.0092933	-1.00	0.315	-.0275532	.008876
dbp	.0159687	.0163001	0.98	0.327	-.015979	.0479163
cc	.0697787	.0144656	4.82	0.000	.0414266	.0981308
ch	-2.267078	.7507192	-3.02	0.003	-3.738461	-.7956954
_cons	-4.576811	1.783906	-2.57	0.010	-8.073203	-1.08042

گزارش مقادیر ضرایب یا نسبت شانس برای متغیرهای مدل به صورت زیر می آید.

mi	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
cat	2.90e-06	9.06e-06	-4.08	0.000	6.33e-09	.0013275
age	1.043663	.018343	2.43	0.015	1.008323	1.080241
chl	.9944097	.0041892	-1.33	0.183	.9862328	1.002654
ecg	1.461544	.482687	1.15	0.251	.7650605	2.792083
smk	2.22753	.7342779	2.43	0.015	1.167443	4.250223
hpt	2.756449	1.222204	2.29	0.022	1.155923	6.57311
chl	1	(omitted)				
sbp	.9907048	.0092069	-1.00	0.315	.9728229	1.008915
dbp	1.016097	.0165625	0.98	0.327	.984148	1.049083
cc	1.072271	.0155111	4.82	0.000	1.042297	1.103107
ch	.1036145	.0777854	-3.02	0.003	.0237907	.4512673
_cons	.0102876	.0183522	-2.57	0.010	.0003118	.339453

متغیرهای cat، age، smk، hpt و مقادیر اینترکشن بین متغیرها (CC و Ch) در مدل بالا با توجه به مقادیر پی ویو و حدود اطمینان نسبت های شانس آنها، معنی دار شده است. برای نمایش ماتریس واریانس کوواریانس پارامترها و محاسبه حدود اطمینان بعد از راند کردن مدل، با دستور vce نمایش داده می شوند. برای اینکه متوجه بشوید مدل با اینترکشن و بدون اینترکشن چگونه کار می کند، وقتی مدل را بصورت کامل اجرا نمودید از

دستور `lrtest` استفاده نمایید. سپس برای ذخیره نمودن مدل از دستور `estimates store full` استفاده و مجدد مدل را کاهش و ترم های اینترکشن را وارد مدل نکنید. مجدد بعد از راند شدن مدل بدون وارد نمودن اینترکشن ها از دستور `lrtest full` استفاده نمایید. این دستورات متعاقبا در قسمت رویکردهای مدل سازی بیشتر توضیح داده می شوند.

### برای ارزیابی مدل رگرسیون لجستیک به صورت زیر عمل می شود:

۱- برای ارزیابی توزیع دو جمله ای متغیر وابسته از دستورات زیر استفاده می شود. مقدار احتمال قابل پیش بینی در برابر متغیر وابسته نوشته می شود و با آن مقدار احتمال آزمون انجام می شود. با توجه به مقدار  $p$  می توان اظهار نظر نمود که توزیع متغیر وابسته از توزیع دو جمله ای تبعیت می کند یا خیر.

`bitest mi==predicted probability`

`bitest mi==0.12`

`bitest mi==0.12, detail`

`. bitest mi==0.12, detail`

Variable	N	Observed k	Expected k	Assumed p	Observed p
mi	609	71	73.08	0.12000	0.11658
Pr(k >= 71)	= 0.620742	(one-sided test)			
Pr(k <= 71)	= 0.427893	(one-sided test)			
Pr(k <= 71 or k >= 75)	= 0.851633	(two-sided test)			
Pr(k == 71)	= 0.048635	(observed)			
Pr(k == 74)	= 0.049100				
Pr(k == 75)	= 0.047761	(opposite extreme)			

۲- برای آزمون رابطه خطی داشتن متغیر کمی مستقل با لجیت متغیر وابسته با دستورات زیر عمل می شود.

- `logit D X1`
- `predict p, pr`
- `predict log, xb`
- `scatter p X1`
- `scatter log X1`

پس از بکارگیری دستورات بالا برای متغیر وابسته و سن به صورت زیر تحلیل ارزیابی انجام می شود.

logit mi age

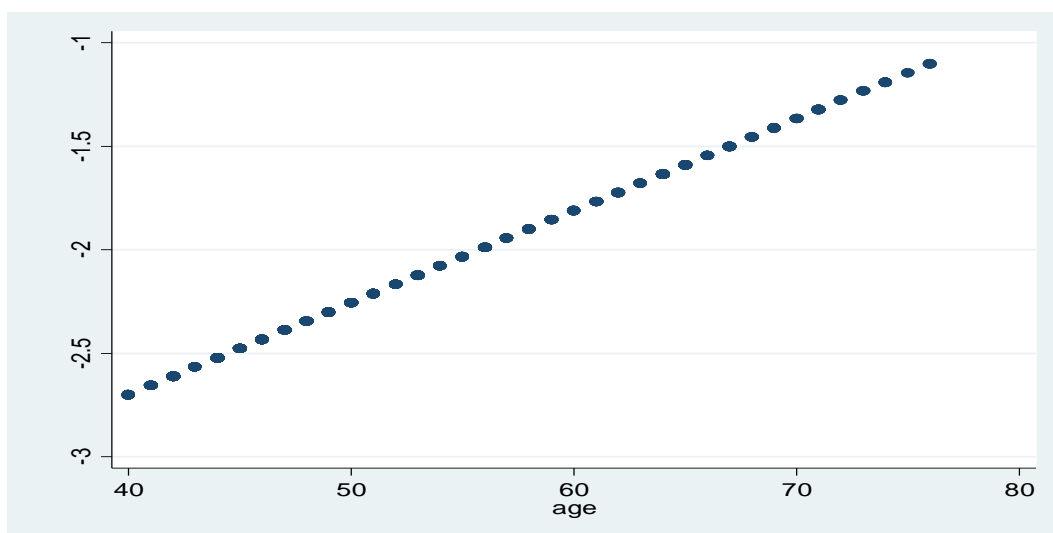
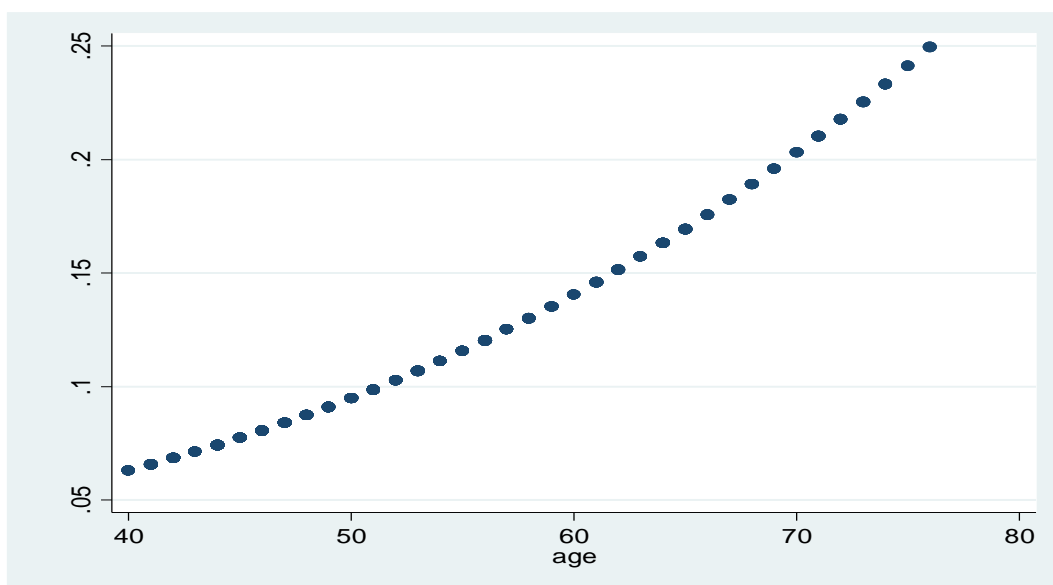
predict p, pr

predict log, xb

scatter p age

scatter log age

خروجی زیر نمودارهای مربوطه را که نشان از رابطه خطی دارد را نمایش می دهد.



رویکرد های مدل سازی و تحلیل در رگرسیون لجستیک:

۱. Assessing relationship: رویکرد ارزیابی رابطه بین متغیرهای مستقل و وابسته و در نظر گرفتن متغیرهای محدود کننده و اینترکشن بین متغیرها که در مثال رابطه بین سگته قلبی و متغیرهای مستقل در بالا به آن پرداخته شد.

۲. Model Making/ Modeling: رویکرد مدل سازی

۳. Mixed approach: رویکرد مشتمل بر انتظارات بالا یعنی در نظر گرفتن رویکرد اول و دوم بطور همزمان

رویکرد ساختن بهترین مدل، برای تعیین متغیرهای موثر بر متغیر وابسته کاربرد دارد. برای برآزش مدل از Maximum Likelihood Method استفاده می شود. این روش اندازه ای است که برای پیدا نمودن حداقل انحراف و فاصله از مقادیر مشاهده شده و مقادیر مورد انتظار استفاده می کند. در رگرسیون خطی از Least Square Method و حداقل نمودن مربع مقادیر باقی مانده استفاده می شد. Likelihood احتمالی است که توسط داده ها به شرط داده های مشاهده شده، پارامتر را برآورد می کند. باید دید داده ها از چه مدلی تبعیت می کنند آنگاه با استفاده از آن مدل پارامترها را برآورد نمود. روش برآورد پارامترهای مدل likelihood نام دارد. به عنوان مثال سه نمونه خون از ۱۰ نمونه خون گرفته شده برای آزمون تشخیص بیماری بروسلوز یا تب مالت تهیه گردید. احتمال (likelihood) بدست آوردن ۳ نمونه مثبت در بین ۱۰ نمونه، چنانچه شیوع واقعی را با این نسبت (p) داشته باشیم به صورت زیر محاسبه می گردد. این نسبت از توزیع دو جمله ای تبعیت و عبارتست از:

$$P_n(X) = C_n^x P^x (1-P)^{n-x}$$

$$C_n^x = \frac{n!}{x!(n-x)!}$$

$$P_{10}(3) = L(P) = C_{10}^3 P^3 (1-P)^{10-3}$$

$$C_{10}^3 = \frac{10!}{3!(10-3)!}$$

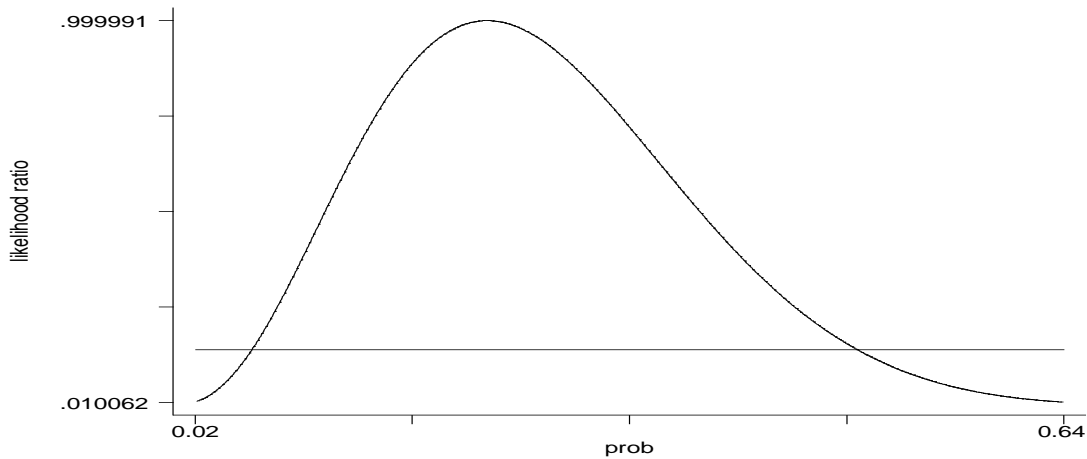
شیوع در سه نمونه :

برای مقدار واقعی P=0.30 مقدار log likelihood می شود -0.628

برای مقدار واقعی P=0.20 مقدار log likelihood می شود -1.60

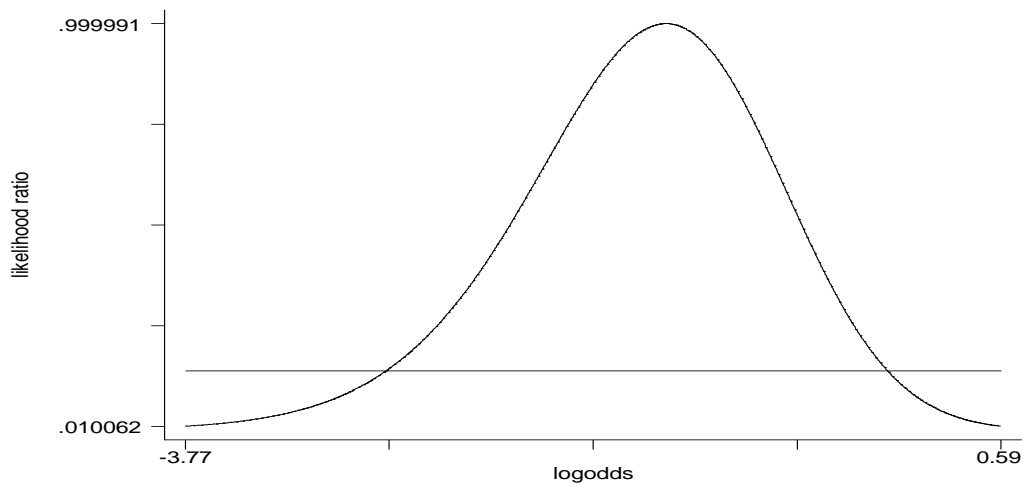
برای مقدار واقعی P=0.35 مقدار log likelihood می شود -1.38

بنابراین حداکثر مقدار احتمال با  $\log(L)$  or  $\log$  Likelihood برای مقدار  $p=0.30$  بدست آمده است. مقادیر likelihood برای ۳ نمونه مثبت در بین ۱۰ نمونه خون گرفته شده و شکل تابع تشکیل شده با دستور زیر ترسیم می شود.



```
cut-point .1465
Most likely value for prob    0.23077
Likelihood based limits for prob  0.06303  0.49513
```

اگر در دستور بالا بجای احتمال از  $\log$ odds یا odds استفاده شود شکل نمودار به صورت زیر خواهد شد.



```
. blik 3 10, param(logodds)

cut-point .1465
Most likely value for logodds -1.20397
Likelihood based limits for logodds -2.69903 -0.01948
```

دستورات بالا ممکن است در نرم افزار شما اجرا نشوند، ابتدا آنها را نصب یا دانلود نمایید.



نسبت حداکثر درستنمایی و کار با نرم افزار ( محاسبه Maximum Likelihood Method )

برای محاسبه حداکثر درستنمایی ( LR ) در نرم افزار علاوه بر دستورات قبلی، از دستورات زیر نیز استفاده می شود. به عنوان مثال چنانچه در فایل مذکور به رابطه بین سگته قلبی و سیگار کشیدن بپردازیم و بخواهیم حداکثر نسبت L را محاسبه کنیم از دستورات زیر استفاده می شود.

logit mi smk

```
Logistic regression                Number of obs   =       609
                                   LR chi2(1)        =         5.75
                                   Prob > chi2         =       0.0165
Log likelihood = -216.40647        Pseudo R2       =       0.0131
```

mi	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smk	.6706383	.2919299	2.30	0.022	.0984661	1.242811
_cons	-2.489797	.2523917	-9.86	0.000	-2.984475	-1.995118

predict pi, pr

replace pi = 1 - pi if mi == 0

gen lnpi=ln(pi)

total lnpi

```
Total estimation                Number of obs   =       609
```

	Total	Std. Err.	[95% Conf. Interval]	
lnpi	-216.4065	15.93818	-247.707	-185.1059

برای آزمون معنی داری متغیرها در مدل های رگرسیون لجستیک از آزمون والد ( Wald Statistics ) یا z با فرمول زیر استفاده می شود. به عنوان مثال برای آزمون معنی داری متغیر sbp این آزمون استفاده می شود. نتیجه معنی دار نیست. قبلا در مثال ابتدای بحث به این آزمون پرداخته شد.

- $Wald = [\beta / s.e.\beta]^2$

- $Z = [\beta / s.e.\beta]$

. test sbp

( 1) [mi]sbp = 0

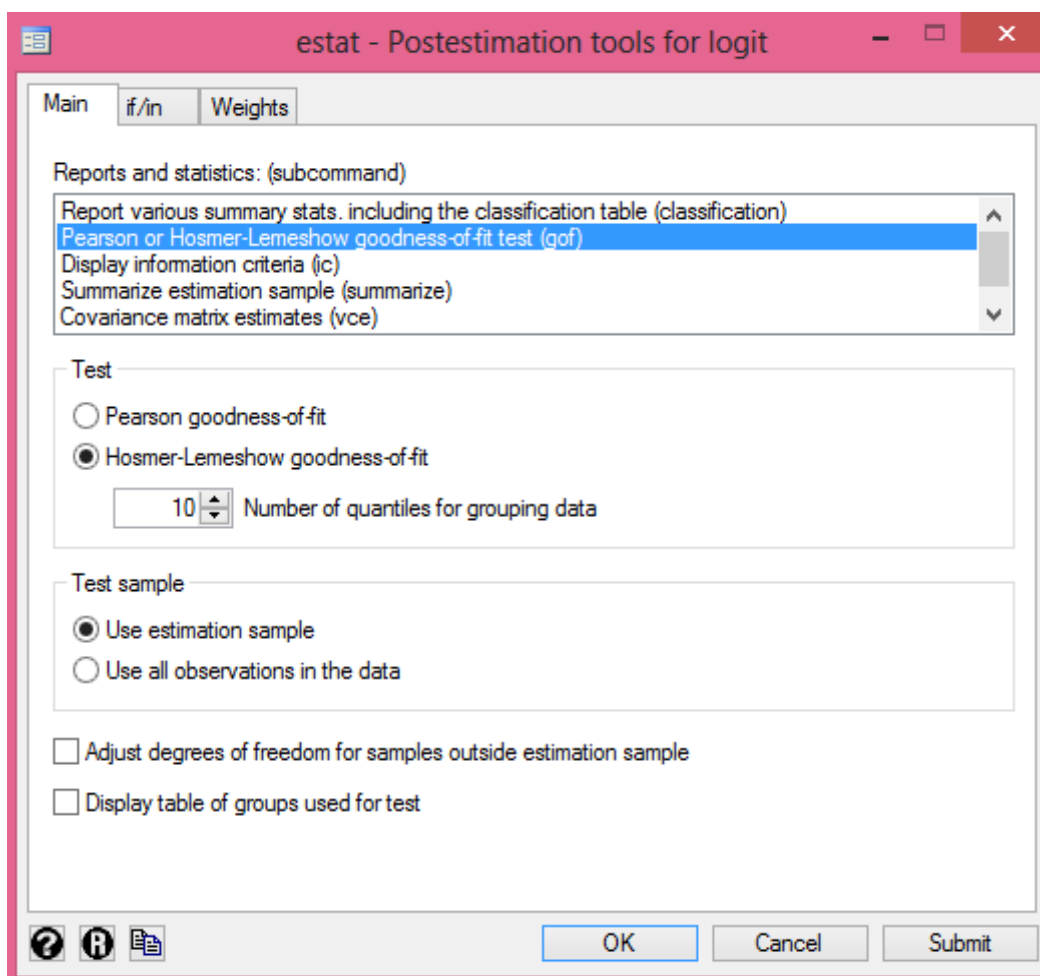
```
chi2( 1) = 1.01
Prob > chi2 = 0.3150
```

برآوردهای بعد از مدل سازی رگرسیون لجستیک با نرم افزار:

در برآوردهای بعد از مدل سازی، بلافاصله بعد از راند کردن مدل لجستیک، گزینه **Postestimation** را از **Binary outcomes** از منوی statistics انتخاب و شکل زیر ظاهر می گردد.

- Goodness-of-fit after logistic/logit/probit
- Classification statistics after logistic/logit/probit/ivprobit
- ROC curve after logistic/logit/probit/ivprobit
- Sensitivity/specificity plot

با انتخاب گزینه اول و باز شدن کادر محاوره ای زیر آزمون **Hosmer-Lemeshow** انتخاب و برازش مدل ارزیابی می شود.



```
. estat gof, group(10)
```

Logistic model for mi, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

```
number of observations =      609
number of groups      =       10
Hosmer-Lemeshow chi2(8) =       8.08
Prob > chi2           =      0.4256
```

وقتی نتیجه آزمون مذکور که بر مبنای تعداد موارد مشاهده شده و مورد انتظار کای اسکور کار می کند، معنی دار نشد نشان می دهد که برازش مدل مناسب است.

با انتخاب گزینه دوم **Classification statistics after logistic/logit/probit/ivprobit** از منوی مذکور می توان آماره های مربوط به طبقه بندی افراد مورد نظر و قدرت مدل در تفکیک افراد در طبقات متغیر وابسته و پیش بینی پذیری مدل را بدست آورد. و مدل را از نظر این شاخص ها و به صورت کلی اعتبار مدل را ارزیابی نمود. نتایج زیر درصد حساسیت، ویژگی، ارزش اخباری مثبت و منفی و درصد منفی و مثبت واقعی و کاذب را نشان داده است. به فرمول های این شاخص ها در بخش پنجم و در قسمت آزمون های تشخیصی پرداخته شده است. ارزیابی مذکور نشان داده است که مدل بیش از ۹۰ درصد در طبقه بندی افراد خوب کار می کند و اعتبار مدل مطلوب است.

```
. estat classification
```

Logistic model for mi

Classified	True		Total
	D	~D	
+	18	5	23
-	53	533	586
Total	71	538	609

```
Classified + if predicted Pr(D) >= .5
True D defined as mi != 0
```

Sensitivity	Pr( +  D)	25.35%
Specificity	Pr( - ~D)	99.07%
Positive predictive value	Pr( D  +)	78.26%
Negative predictive value	Pr(~D  -)	90.96%
False + rate for true ~D	Pr( + ~D)	0.93%
False - rate for true D	Pr( -  D)	74.65%
False + rate for classified +	Pr(~D  +)	21.74%
False - rate for classified -	Pr( D  -)	9.04%
Correctly classified		90.48%

### خلاصه دستورات ارزیابی مدل رگرسیون لجستیک:

پس از راند کردن مدل، دستورات خلاصه برای ارزیابی مدل به شرح زیر است.

lift

lsens

lstat

lstat, cutoff(0.6)

lroc

### منحنی ROC:

در این منحنی شاخص حساسیت (درصد مثبت حقیقی) را در محور  $y$  ها و شاخص ویژگی  $- 1$  (درصد مثبت کاذب) را در محور  $x$  ها قرار می دهند. این منحنی برای مدل مذکور با استفاده از گزینه زیر انجام می شود.

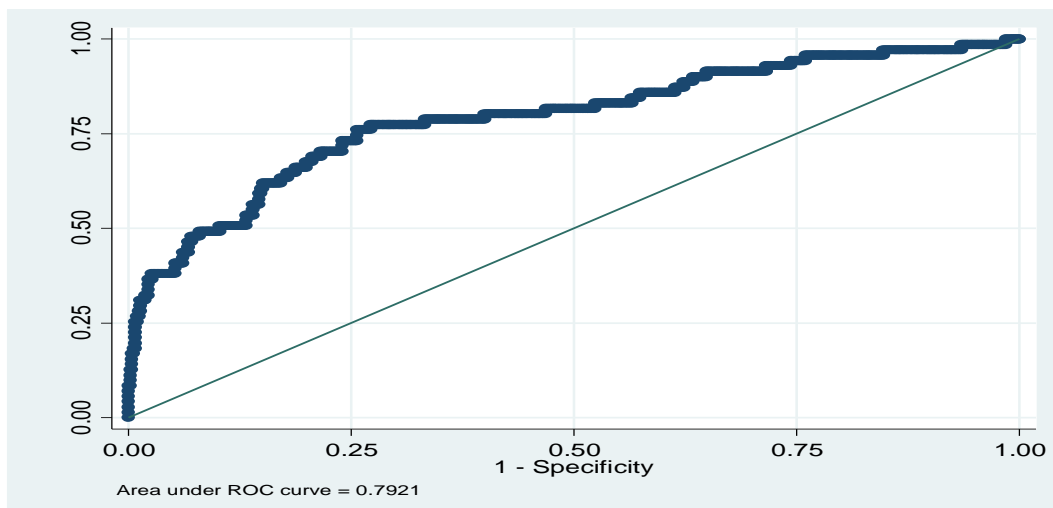
#### ROC curve after logistic/logit/probit/ivprobit

نتایج و نمودار در زیر آمده است. سطح زیر منحنی حدود ۸۰ درصد (0.7921) گزارش شده است. این مقدار از سطح زیر منحنی برای مدل نیز مطلوب است.

```
. lroc
```

```
Logistic model for mi
```

```
number of observations =      609
area under ROC curve   =    0.7921
```



برای ترسیم منحنی ROC از مسیر زیر نیز می توان منحنی را رسم نمود. توضیحات مرتبط با این منحنی در قسمت آزمون های تشخیصی در بخش پنجم کتاب شرح داده شده است.

Epidemiology and related	ROC analysis	ROC regression models
SEM (structural equation modeling)	Tables for epidemiologists	Parametric ROC analysis without covariates
Survey data analysis	Other	Nonparametric ROC analysis without covariates
Multiple imputation		Test equality of two or more ROC areas
Multivariate analysis		Test equality of ROC area against gold standard

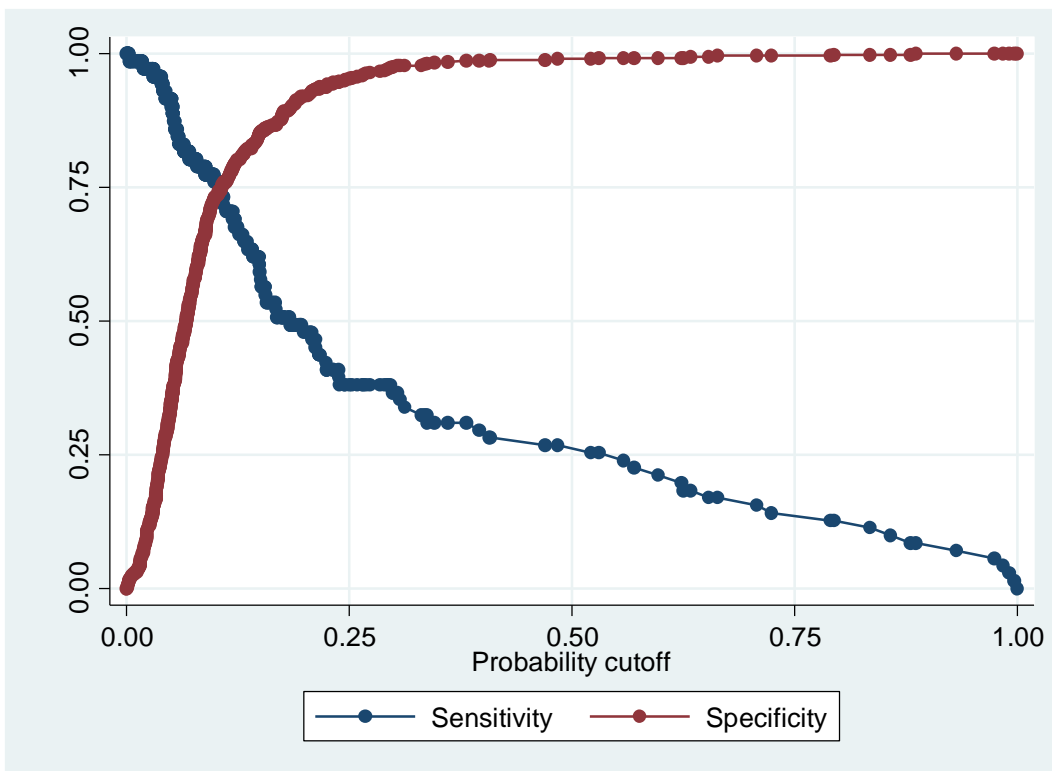
بر طبق یک قاعده عمومی مقادیر سطح زیر منحنی ROC به صورت زیر تقسیم بندی می شود:

۱- مقدار غیرقابل قبول  $ROC=0.5$

۲- مقدار قابل قبول  $ROC \geq 0.7$

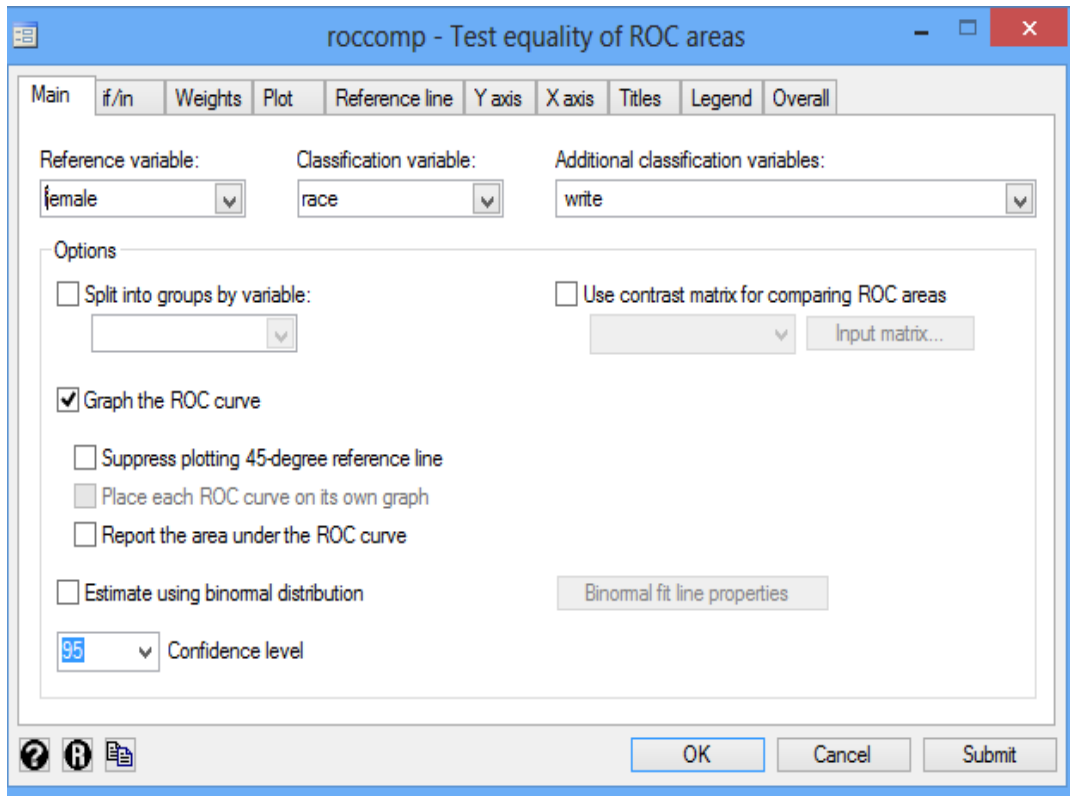
۳- مقدار مطلوب  $ROC \geq 0.8$  و مقدار بالاتر از 0.9 معمولا غیرمعمول است.

با استفاده از گزینه سوم منوی مذکور یعنی **Sensitivity/specificity plot** یا با استفاده از دستور **Isens** می توان نمودار حساسیت و ویژگی و نقطه برش را در مدل مذکور تعیین و آن را به صورت زیر ترسیم نمود.



مقایسه منحنی های ROC :

ابتدا منحنی ROC برای هر کدام از مدل های رگرسیونی ترسیم می شود. این منحنی در مطالب پیشگفت با دستور logit یا logistic و استفاده از دستور Iroc قابل انجام بود. اما با دستور Iroc نمی توان مقایسه دو منحنی را انجام داد. بنابراین از دستور roccomp استفاده می شود. از مسیر دستور db roccomp نیز می توان با انتخاب کادر محاوره ای زیر و معرفی متغیرها مقایسه دو یا چند منحنی را انجام داد.



برای کار با نرم افزار و استفاده از این دستورات فایل comparisonROC را باز و متغیرها را مانند کادر محاوره ای بالا معرفی نمایید. سپس مراحل زیر را انجام دهید. متغیر hon را بعنوان متغیر پیامد دو حالتی و متغیر female را بعنوان متغیر مستقل در نظر گرفته و مدل لوجستیک را اجرا نمایید. قبل از اجرای مدل، متغیر پیامد را بررسی نمایید. برای تبدیل نمودن آن به یک متغیر دو حالتی از دستور زیر استفاده می شود.

```
generate hon=write>=60
```

نتیجه خروجی راند کردن مدل نشان می دهد به صورت کلی مدل معنی دار است اما متغیر female با ضریب و مقدار P و حدود اطمینان ارائه شده معنی دار نیست. برای گزارش مقدار سطح زیر منحنی ROC از دستور زیر استفاده می شود. در صورتیکه عبارت nograph در دستور نوشته نشود منحنی نیز ترسیم می شود. در اینجا هدف عدم نمایش منحنی است.

```
. lroc, nograph
```

```
Logistic model for hon
```

```
number of observations =      200
area under ROC curve   =    0.5785
```

با استفاده از دستور `predict xb1, xb` پیش بینی ترکیب خطی برای مدل مذکور (مدل ۱) انجام می شود. مدل دوم همانند مدل یک اجرا می شود. در مدل دوم متغیر `read` اضافه شده است. نتیجه مدل دوم به صورت زیر است.

```
. logit hon female read, nolog
```

```
Logistic regression                Number of obs   =      200
                                   LR chi2(2)         =     60.40
                                   Prob > chi2        =     0.0000
Log likelihood = -85.44372          Pseudo R2      =     0.2612
```

hon	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
female	1.120926	.4081043	2.75	0.006	.3210558 1.920795
read	.1443657	.0233338	6.19	0.000	.0986322 .1900991
_cons	-9.603364	1.426412	-6.73	0.000	-12.39908 -6.807647

خروجی بالا نشان می دهد که متغیر `read` در مدل معنی دار است. سطح زیر منحنی آن در خروجی زیر آمده است. برای این مدل نیز با دستور `predict xb2, xb` پیش بینی ترکیب خطی مدل دوم انجام می شود.

```
. lroc, nograph
```

```
Logistic model for hon
```

```
number of observations =      200
area under ROC curve   =    0.8330
```

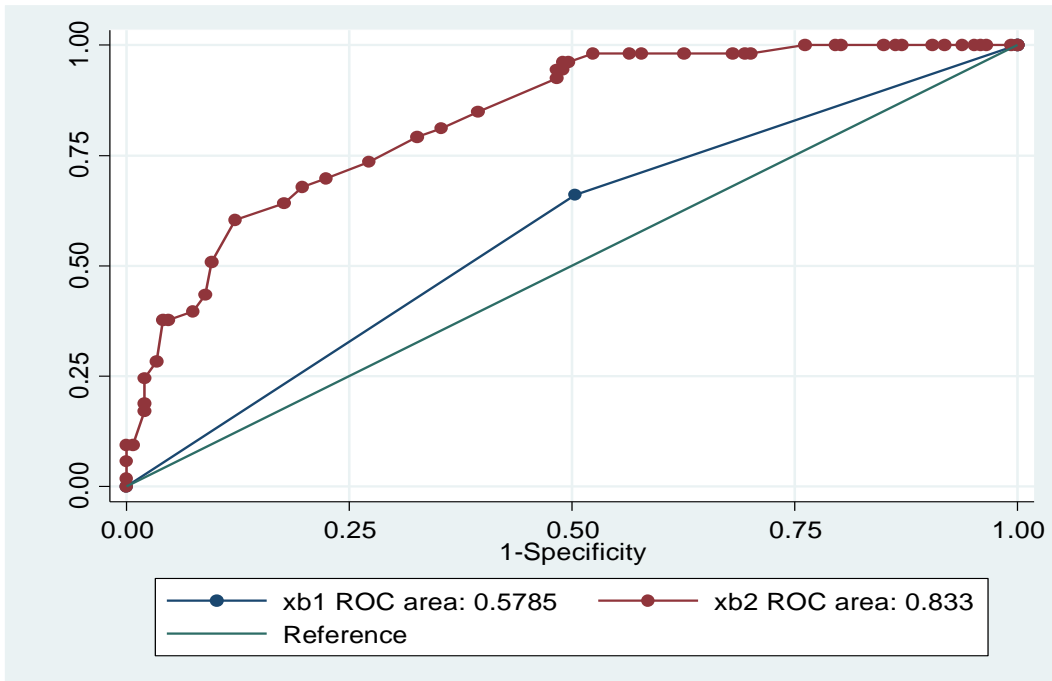
اینک دو منحنی ROC با سطح زیر منحنی 0.5785 و 0.8330 وجود دارد. در صورتیکه محقق قصد مقایسه این دو منحنی را داشته باشد به صورت زیر آزمون انجام میشود.

```
. roccomp hon xb1 xb2, graph summary
```

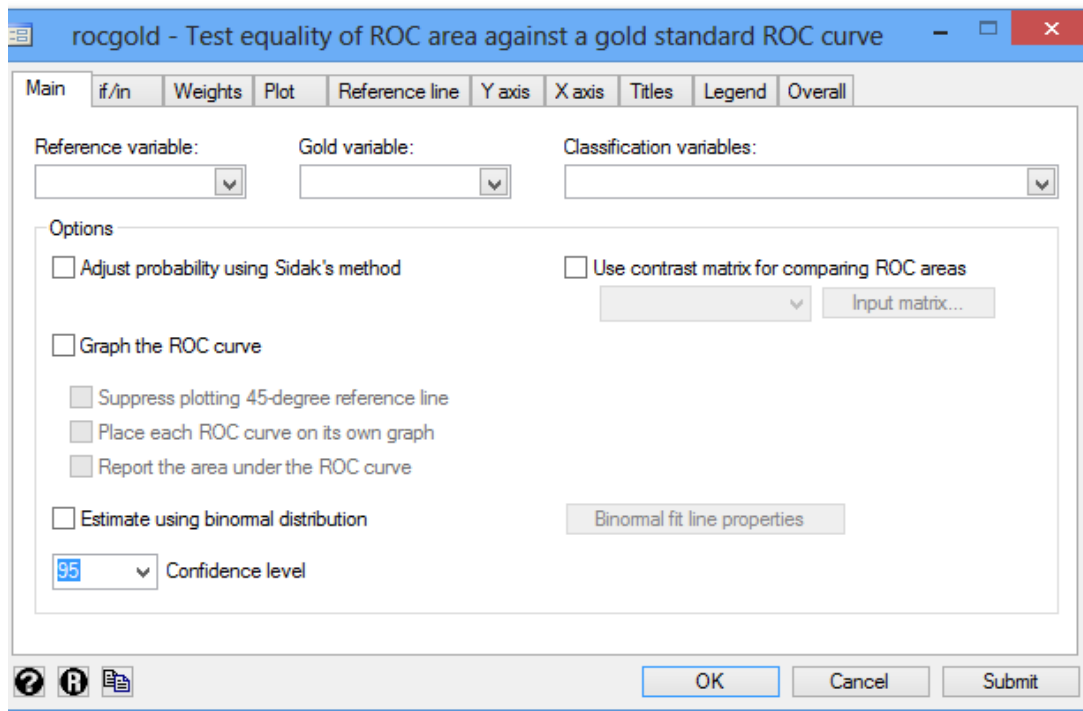
	Obs	ROC Area	Std. Err.	-Asymptotic Normal- [95% Conf. Interval]	
xb1	200	0.5785	0.0388	0.50242	0.65456
xb2	200	0.8330	0.0301	0.77397	0.89205

```
Ho: area(xb1) = area(xb2)
chi2(1) = 34.91      Prob>chi2 = 0.0000
```

خروجی بالا تعداد مشاهدات هر مدل، سطح زیر منحنی و خطای معیار و حدود اطمینان را نشان می دهد. فرض صفر مساوی بودن سطح زیر دو منحنی است. که در آزمون انجام شده، نتیجه نشان می دهد که تفاوت آماری معنی داری بین دو منحنی وجود دارد ( $P=0.0000$ ). مقایسه دو منحنی در شکل زیر آمده است.



**نکته:** برای مقایسه منحنی با استاندارد طلایی با دستور `db rocgold` و کادر زیر مقایسه را می توان انجام داد.





در ابتدای بحث بالا (مدل ها و رویکردهای تحلیل رگرسیون لجستیک) در رویکردهای مدل سازی به رویکرد (Mixed approach) که مشتمل بر انتظارات و در نظر گرفتن رویکرد اول و دوم، بطور همزمان بود پرداخته شد. یعنی مدلی درست نماییم با حداقل اطلاعات و متغیر، اما برازش مدل مطلوب باشد. برای انتخاب بهترین مدل از روش (Likelihood Ratio Test) استفاده می شود. این روش بر مبنای اختلاف بین دو مدل کامل و کاهش یافته ( $-2 \log \text{Likelihood}$ ) و با توزیع کای دو کار می کند. مدل کامل وقتی است که همه متغیرها و اینترکشن آنها در مدل وارد می شود. در مدل کاهش یافته اینترکشن وارد نمی شود. فرمول های محاسبه کارکرد مدل ها در زیر آمده است.

$$G = \chi^2 = D_{null} - D_k$$

$$= -2LL_{null} - (-2LL_k)$$

$$G = \chi^2 = -2 \ln \left( \frac{L_{null}}{L_k} \right)$$

برای استفاده از مباحث مذکور و انتخاب بهترین مدل مناسب به صورت زیر استفاده می شود. ابتدا مدل کامل به صورت زیر اجرا میشود.

logit mi cat age chl ecg smk hpt chl sbp dbp cc ch, or

Logistic regression	Number of obs	=	609
	LR chi2 (10)	=	92.64
	Prob > chi2	=	0.0000
Log likelihood = -172.96062	Pseudo R2	=	0.2112

mi	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
cat	2.90e-06	9.06e-06	-4.08	0.000	6.33e-09 .0013275
age	1.043663	.018343	2.43	0.015	1.008323 1.080241
chl	.9944097	.0041892	-1.33	0.183	.9862328 1.002654
ecg	1.461544	.482687	1.15	0.251	.7650605 2.792083
smk	2.22753	.7342779	2.43	0.015	1.167443 4.250223
hpt	2.756449	1.222204	2.29	0.022	1.155923 6.57311
chl	1	(omitted)			
sbp	.9907048	.0092069	-1.00	0.315	.9728229 1.008915
dbp	1.016097	.0165625	0.98	0.327	.984148 1.049083
cc	1.072271	.0155111	4.82	0.000	1.042297 1.103107
ch	.1036145	.0777854	-3.02	0.003	.0237907 .4512673
_cons	.0102876	.0183522	-2.57	0.010	.0003118 .339453

سپس مدل کامل با نام `fullmodel` با دستور `estimates store fullmodel` ذخیره می شود. مجدد مدل کاهش یافته یعنی بدون اینترکشن با دستور `logit mi cat age chl ecg smk hpt chl sbp dbp`, اجرا می شود. نتایج زیر حاصل می شود.

```

Logistic regression                               Number of obs   =       609
                                                    LR chi2(8)      =       39.96
                                                    Prob > chi2     =       0.0000
Log likelihood = -199.30006                       Pseudo R2      =       0.0911
    
```

mi	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
cat	2.117545	.7988092	1.99	0.047	1.010953 4.435418
age	1.038504	.0166483	2.36	0.018	1.006381 1.071652
chl	1.008712	.003325	2.63	0.009	1.002216 1.01525
ecg	1.496944	.4432864	1.36	0.173	.8378021 2.674668
smk	2.380889	.7314647	2.82	0.005	1.303857 4.347585
hpt	1.703615	.6537628	1.39	0.165	.8030127 3.614267
chl	1	(omitted)			
sbp	.9894843	.0079882	-1.31	0.190	.9739509 1.005265
dbp	1.012479	.0144625	0.87	0.385	.9845265 1.041226
_cons	.0011538	.0018227	-4.28	0.000	.0000522 .0255174

این مدل نیز با نام `reducemodel` و دستور `estimates store reducemodel` نیز ذخیره می شود. اکنون برای مقایسه دو مدل ساخته شده از دستور `lrtest reducemodel fullmodel, stats` استفاده و آماره  $(-2 \log \text{Likelihood})$  محاسبه می شود.

نتایج به صورت زیر اختلاف معنی دار بین دو مدل را نشان می دهد.

```

Likelihood-ratio test                               LR chi2(2) =       52.68
(Assumption: reducemodel nested in fullmodel)     Prob > chi2 =       0.0000
    
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
<u>reducemodel</u>	609	-219.2791	-199.3001	9	416.6001	456.3065
<u>fullmodel</u>	609	-219.2791	-172.9606	11	367.9212	416.4512

Note: N=Obs used in calculating BIC; see [R] BIC note

وقتی بین دو مدل تفاوت معنی دار بود مطلوب استفاده از مدل کامل و در نظر گرفتن اینترکشن بین متغیرها در مدل است. در نتایج بالا می توان گفت که اختلاف بین دو مدل ۵۲/۶۸ بوده و این مقدار اختلاف معنی دار است. لذا از مدل کامل استفاده می شود.

در واقع برای مقایسه مدل ها از شاخص های `AIC`, `BIC` استفاده می کنیم و مدلی که `Akaike (AIC)` Information Criterion یا `Bayesian information criterion (BIC)` کمتری داشته باشد مناسب تر است.

## رگرسیون لجستیک شرطی: Conditional Logistic Regression

دو نوع رویکرد شرطی و غیر شرطی ( unconditional estimation approach and the conditional estimation approach) برای برآورد Maximum likelihood estimation در رگرسیون لجستیک وجود دارد. رویکرد اول رگرسیون لجستیک معمولی است که در مبحث قبلی تشریح گردید. رویکرد دوم، برآورد بوسیله رگرسیون لجستیک شرطی است. رگرسیون لجستیک شرطی بعنوان رگرسیون لجستیک اثرات ثابت شرطی و رگرسیون لجستیک مورد-شاهدی همسان شده نیز معروف است. الگوریتم برآورد در هر ۳ مدل یکسان است. اختلاف آنها بر این اساس است که داده‌ها چطور ساختار یافته‌اند و تحلیل برای چه هدفی بکار می‌رود. مدل‌های شرطی در همه گونه‌هایشان نوعی مدل پانل (panel) هستند، یعنی یک مورد یا فرد شامل مجموعه اطلاعاتی در یک دوره زمانی است.

مدل پانل زیر را در نظر بگیرید :

$$y_{ij} = \beta_0 + \beta_i x_{ij} + \gamma_i + \varepsilon_{ij}$$

در مدل اثرات ثابت،  $\gamma$  شامل مشخصات ثابت شده‌ای از یک فرد یا مورد است. مشکل اساسی در تحلیل مدل‌های پانل چگونگی تمایز بین تغییرات درون فردی از تغییرات بین فردی است. مدل‌های اثرات ثابت شده شرطی، تغییرات بین فردی را نادیده می‌گیرند. با این کار خطای استاندارد زیاد می‌شود. مدل اثرات تصادفی هم تغییرات بین فردی و هم تغییرات درون فردی را در نظر می‌گیرد. وقتی تغییرات درون فردی قابل توجهی وجود داشته باشد، آنگاه مدل اثرات ثابت مناسب‌ترین است. در واقع این مدل برای آشکارسازی تغییرات درون فردی مفید است. جنسیت، پیشینه اخلاقی، مذهب و کشور محل تولد تغییرات درون فردی خیلی کمی دارند. وقتی واضح است که توزیعی به داده‌ها مناسب نیست، یک مدل اثرات ثابت شرطی ممکن است برآوردهایی با اریبی کمتر بدست آورد. وقتی تعداد زیادی پانل وجود دارد، ترجیح داده می‌شود که از مدل‌های اثرات ثابت شرطی استفاده شود.

در مطالعات مورد شاهد برای حذف یا کنترل متغیرهای مخدوش کننده روش همسان سازی استفاده می‌شود. نوع آنالیز این داده نیز متفاوت از وقتی است که مطالعه مورد شاهد معمولی انجام می‌شود و همسان سازی صورت نمی‌گیرد. بنابراین در روش مطالعه همسان سازی شده از مدل رگرسیون لجستیک شرطی استفاده می‌شود. همسان سازی در مطالعات کوهورت رابطه مواجهه - مخدوشگر از بین می‌رود. همسان سازی در مطالعات مورد-شاهدی رابطه بیماری - مخدوشگر را از بین می‌برد.

اگر در مطالعه همسان شده، مخصوصاً در مطالعه همسان شده مورد-شاهدی، از یک تحلیل همسان شده استفاده نشود، اریبی یا تورش انتخاب رخ می‌دهد و همچنین برآوردهای نامعتبری از OR بدست می‌آید. وقتی داده‌ها از یک مطالعه مورد-شاهدی همسان شده باشند، معمولاً رگرسیون لجستیک شرطی برای تحلیل به کار می‌رود. در مطالعه‌های کوهورت همسان شده، رگرسیون کاکس یا پواسن برای تحلیل به کار می‌رود.

پس اگر در یک مطالعه داده‌ها همسان نشوند، رابطه بین متغیر مواجهه و بیماری با استفاده از رگرسیون معمولی (غیر شرطی) بدست می‌آید. در صورت همسان شدن رگرسیون لجستیک شرطی به کار می‌رود. رگرسیون لجستیک شرطی مشابه CMH chi-square است. در آن روی جفت‌های همسان‌شده طبقه‌بندی صورت می‌گیرد و سپس روش برآورد درست‌نمایی ماکزیمم شرطی به کار می‌رود. چون تعداد متغیرهای ظاهری (dummy) زیاد است (یکی کمتر از تعداد مجموعه‌های همسان‌شده)، به کار بردن روش درست‌نمایی ماکزیمم شرطی ضرورت دارد. در واقع تعداد پارامترها نسبت به تعداد مشاهدات بزرگ است. در مطالعات همسان‌شده اگر از روش رگرسیون لجستیک معمولی استفاده شود، OR فرابرازش یا بیش برآورد (overestimate) می‌شود.

به عنوان مثال ۴۰ جفت از افراد مبتلا به سکته قلبی حاد برای تسکین درد قلب هر یک مداخله درمانی مورد نظر (مداخله) یا درمان استاندارد (عدم مداخله) را دریافت کردند. رگرسیون لجستیک شرطی بصورت زیر برای این افراد فرمول‌بندی می‌شود. طبقات (جفت‌های همسان‌شده) با استفاده از متغیرهای نشانگر تعریف می‌شوند.

$$\text{logit } P(X) = \beta_0 + \beta_1 E + \sum_{i=1}^{39} \gamma_i V_i$$

$$V_i = \begin{cases} 1 & \text{for subject } i \\ 0 & \text{otherwise} \end{cases}$$

وقتی داده‌ها به صورت فردی همسان سازی شده‌اند از آنالیز رگرسیون شرطی و چنانچه داده‌ها به صورت گروهی همسان سازی گردند از آنالیزهای معمول استفاده می‌شود. از رگرسیون لجستیک شرطی برای تحلیل داده‌های همسان‌شده به صورت گروهی استفاده نمی‌شود. ساختار داده‌ها در مدل رگرسیون لجستیک شرطی به صورت زیر است. در درون گروه‌ها بر روی تعداد متغیر پیامد مساوی با یک شرط گذاشته می‌شود.

```

group 1:
  obs. 1  outcome=1  x1 = ...  x2 = ...
  obs. 2  outcome=0  x1 = ...  x2 = ...
group 2:
  obs. 3  outcome=1  x1 = ...  x2 = ...
  obs. 4  outcome=0  x1 = ...  x2 = ...
  obs. 5  outcome=0  x1 = ...  x2 = ...
group 3:
  ...
.
.
Group k:
  ...
    
```

وقتی مدل کامل از رگرسیون لجستیک شرطی مد نظر باشد و متغیرهای همسان‌شده، متغیرهای مخدوش کننده و اینترکشن را در مدل وارد شوند مدل به صورت کلی با فرمول زیر نوشته می‌شود.

$$\begin{aligned} \text{logit } P(\mathbf{X}) = & \alpha + \beta E \\ & + \underbrace{\sum \gamma_{1i} V_{1i}}_{\text{matching}} + \underbrace{\sum \gamma_{2j} V_{2j}}_{\text{confounders}} \\ & + E \underbrace{\sum \delta_k W_k}_{\text{interaction}} \end{aligned}$$

اگر تعداد جفت‌های همسان شده ۱۰۰ باشد، ۹۹ متغیر dummy یا ظاهری وجود دارد.

- مجموعه همسان شده ۹۹ ام :

$$V_{1,99} = 1, \quad V_{1,1} = V_{1,2} = \dots = V_{1,98} = 0$$

- مجموعه همسان شده ۱۰۰ ام :

$$V_{1,1} = V_{1,2} = \dots = V_{1,99} = 0$$

قابل ذکر است که در روش رگرسیون لجستیک شرطی پاسخ‌ها برای یک فرد (آزمودنی) خاص فرض می‌شود که مستقل است. به عبارت دیگر پاسخ‌ها می‌توانند مستقل باشند اگر بوسیله افراد شرطی شوند.

برآورد پارامترهای عرض از مبدا یا متغیرهای ظاهری بدست نمی‌آید. در هر صورت این مشکل بزرگی نیست زیرا پارامتر مطلوب ضریب متغیر مستقل است. وقتی همه متغیرهای پیشگو در همه طبقه های جور شده همسان باشند آنگاه مدل رگرسیون لجستیک شرطی اجرا نخواهد شد و همه پارامترها صفر برآورد می‌شوند.

وقتی تغییرات درون فردی برای حداقل یک متغیر مستقل وجود داشت آنگاه مدل اجرا خواهد شد و پارامترها تنها برای آن متغیرها برآورد می‌شود. یکی از مزایای تحلیل همسان شده اینست که متغیرهای مخدوش کننده ای که قابل کنترل نیستند را کنترل می‌کند.

یکی از معایب تحلیل همسان شده نیز این است که نمی‌توان اثرات مجزای عامل‌های مستقل از زمان را مدل‌بندی نمود. در مدل رگرسیون لجستیک شرطی افراد بصورت اثر ثابت با پارامتر  $\gamma$  مدل‌بندی می‌شوند.

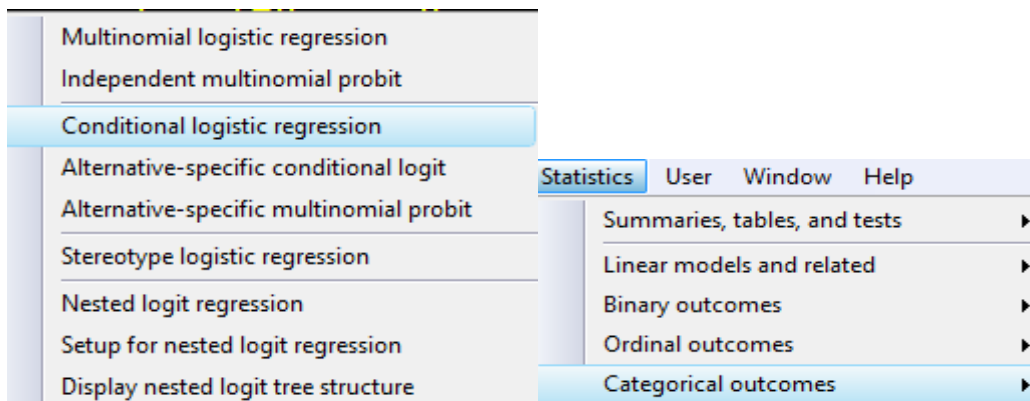
### کار با نرم افزار و انجام رگرسیون لجستیک شرطی:

فایل داده های Matched data را باز نمایید. اولاً داده ها به طرز زیر وارد محیط نرم افزار می شوند. یعنی متغیر اول بنام match برای معرفی جفت های همسان شده است. در این مطالعه به ازای یک مورد (در ستون دوم متغیر mi مورد با کد ۱ تعریف گردید و شاهد با کد ۰ تعریف شد)، ۲ نفر شاهد به ازای یک مورد انتخاب و همسان سازی صورت گرفت. بیماران و شاهدها با متغیر mi نشان داده شده اند. متغیرهای سیگار کشیدن،

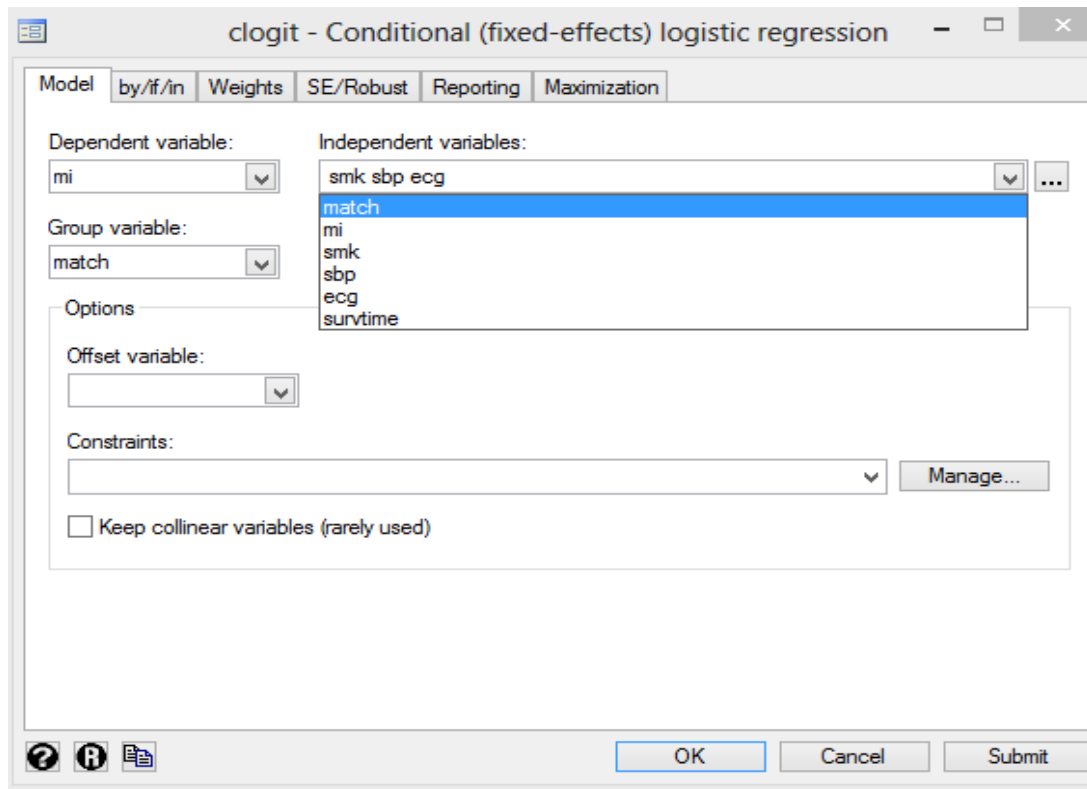
وضعیت فرد و نوار قلب مشکل دار با کد ۱ و بدون مشکل با کد ۰ به عنوان مواجهه تعریف شده است. فشارخون سیستمولیک نیز به صورت عددی وارد گردید.

	match	mi	smk	sbp	ecg	survtime
1	1	1	0	160	1	1
2	1	0	0	140	0	2
3	1	0	0	120	0	2
4	2	1	0	160	1	1
5	2	0	0	140	0	2
6	2	0	0	120	0	2
7	3	1	0	160	0	1
8	3	0	0	140	0	2
9	3	0	0	120	0	2

**نکته:** در این داده ها شماره ردیف وجود ندارد. اولاً متغیری با دستور `gen id=_n` درست می شود و شماره ردیف برای افراد نشان داده می شود. این دستور از جمله دستورات مهم مدیریت داده ها به شمار می رود. برای انجام آنالیز رگرسیون شرطی با دستور `db clogit` یا از مسیر زیر کادر محاوره ای باز می شود.



در کادر محاوره ای به صورت زیر متغیرها معرفی می شوند. متغیری که جفت های همسان شده را معرفی می کند در جعبه مربوط به `Group variable`. متغیر وابسته و متغیرهای مستقل نیز در قسمت مربوطه تعریف می شوند. چنانچه در خروجی نیاز به گزارش نسبت های شانس باشد در قسمت `reporting` گزینه مربوطه انتخاب می شود. چنانچه نتایج بصورت لگاریتمی نیاز باشد نیز گزینه مربوطه انتخاب می شود. متغیرهایی که برای آنها همسان سازی صورت گرفت وارد مدل نخواهند شد. بنابراین اثر آنها را نمی توان بررسی کرد.



خروجی زیر تحلیل مورد نظر را نشان می دهد.

```
. clogit mi smk sbp ecg, group(match) or
```

```
Iteration 0: log likelihood = -32.362396
Iteration 1: log likelihood = -31.750966
Iteration 2: log likelihood = -31.745464
Iteration 3: log likelihood = -31.745464
```

```
Conditional (fixed-effects) logistic regression   Number of obs   =       117
                                                    LR chi2(3)      =       22.20
                                                    Prob > chi2     =       0.0001
Log likelihood = -31.745464                    Pseudo R2       =       0.2591
```

	mi	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	smk	2.073127	1.163557	1.30	0.194	.6900538 6.22829
	sbp	1.046699	.0159589	2.99	0.003	1.015883 1.07845
	ecg	4.949382	4.223875	1.87	0.061	.92923 26.36203

نتایج نشان می دهد مدل معنی دار شده است. متغیر فشارخون سیستمولیک نیز بعنوان متغیر اثر گذار معنی دار شده است. نسبت شانس و حدود اطمینان آن نیز گزارش شده است. بقیه متغیرها معنی دار نشده اند. تفسیر مقادیر نسبت شانس مانند رگرسیون لجستیک معمولی است. پس از انجام تحلیل برای ارزیابی برازش مدل از دستور fitstat استفاده می شود و مدل ارزیابی می شود. ممکن است این دستور در سیستم شما نصب نباشد. لذا

ابتدا باید آنرا با دستور `findit fitstat` پیدا و نصب نمایید. پس از اجرای این دستور خروجی زیر نمایش داده می شود. می توان مدل ها را با همدیگر مقایسه نمود. در اینجا فقط دو مدل ابتدایی ( بدون حضور متغیرهای مستقل) و مدل کامل با هم مقایسه شده است. می توان اینترکشن ها را وارد نمود و مدل ها را مقایسه نمود. برای این کار نیز قبلا توضیحات و دستور لازم در قسمت رگرسیون لجستیک معمولی داده شد.

`. fitstat`

Measures of Fit for clogit of mi

Log-Lik Intercept Only:	-42.846	Log-Lik Full Model:	-31.745
D(36) :	63.491	LR(3) :	22.201
		Prob > LR:	0.000
McFadden's R2:	0.259	McFadden's Adj R2:	0.189
Maximum Likelihood R2:	0.434	Cragg & Uhler's R2:	0.488
Count R2:	0.688		
AIC:	1.782	AIC*n:	69.491
BIC:	-68.397	BIC' :	-11.210

آزمون McFadden's R2 که با فرمول زیر محاسبه می شود برای ارزیابی برازش مدل بکار می رود.

$$R^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$

شبه مربع R است. یعنی مدل تحقیق چقدر از برازش مناسب برخوردار است. این آزمون میزان تغییرات در مدل LR را نشان می دهد. یا اینکه متغیرهای مستقل چقدر قادر به پیش بینی متغیر وابسته هستند. کوچک بودن نسبت log likelihoods نشان می دهد که مدل کامل (full model) از مدل intercept model بهتر است. اگر دو مدل با استفاده از داده های یکسان با همدیگر مقایسه گردند McFadden's برای مدلی که likelihood بالاتری دارد بیشتر است. آماره های AIC و BIC برای تطبیق نمودن LR با تعداد متغیرها و حجم نمونه مطالعه بکار می روند.

LL بعنوان log-likelihood و K تعداد متغیرهای مستقل یا پیش بینی کننده است. 2k برای جبران تعداد متغیرها در مدل بکار می رود. حجم نمونه بالا بر  $-2LL$  اثر می گذارد. فرمول های آنها به صورت زیر است.

$$\begin{aligned} & \text{ESTAT-IC} \\ \text{AIC} &= -2*LL + 2*k = -2(LL-k) \end{aligned}$$

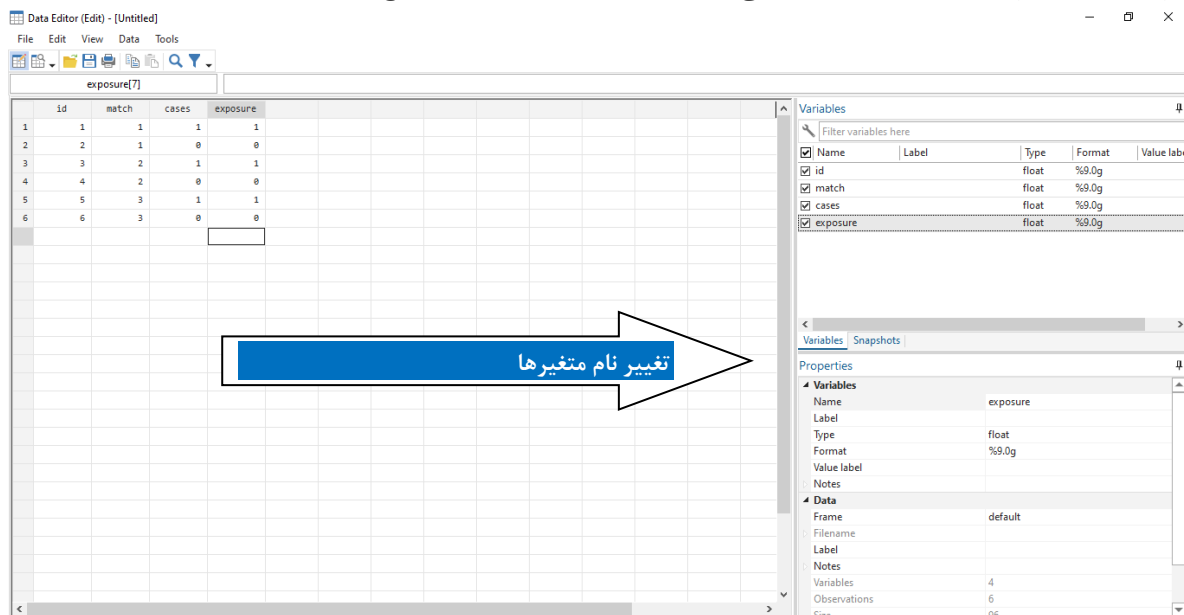
$$\begin{aligned} & \text{GLM} \\ \text{AIC} &= -2*LL + 2*k - 2(LL - k) \\ & \text{-----} = \text{-----} \\ & \quad \quad \quad n \quad \quad \quad n \end{aligned}$$



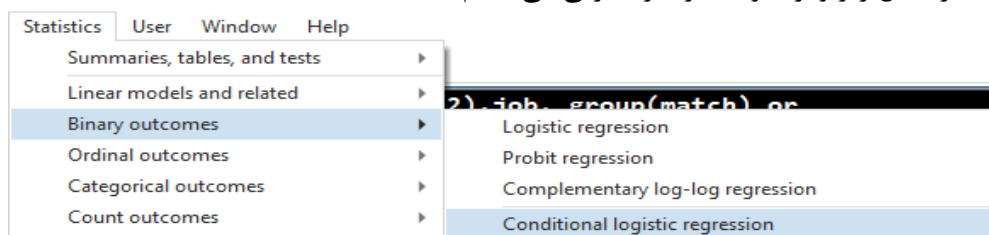
طرز ساخت بانک اطلاعاتی برای داده های همسان شده فردی و تحلیل رگرسیون لجستیک شرطی

**مثال:** در یک مطالعه مورد- شاهد به ازای هر مورد یک شاهد انتخاب و از نظر سن همسان شدند. برای وارد کردن داده ها به نرم افزار به صورت زیر عمل می شود:

- ۱- ابتدا پنجره دیتا ادیتور را باز می کنیم و داده ها را وارد می کنیم. به صورت پیش فرض نرم افزار به متغیرها به ترتیب var1 تا به تعداد متغیرها شماره می دهد. مثلا در دیتای شکل زیر تا var4.
- ۲- سپس در همان پنجره از قسمت راست می توان متغیرها را نامگذاری و اصلاح نمود.



- ۳- در قسمت match شماره های یکسان با یکدیگر از نظر متغیر سن همسان شدند. وقتی نمونه ها را از نظر یک متغیر با یکدیگر همسان سازی نمائید دیگر نمی توان اثر متغیر همسان شده را بین دو گروه مورد و شاهد مقایسه و تحلیل نمود. زیرا نمونه ها شبیه به هم انتخاب شدند.
- ۴- در قسمت cases نیز چنانچه فرد وارد شده به مطالعه در گروه مورد باشد عدد ۱ و چنانچه در گروه کنترل یا شاهد باشد عدد صفر ثبت می گردد. به عبارتی مطالعه مورد- شاهد با داشتن و نداشتن پیامد شروع می شود و در اینجا برای افرادی که پیامد را دارند عدد ۱ و افرادی که پیامد را ندارند عدد ۰ ثبت می گردد.
- ۵- در متغیر بعدی مواجهه اگر به صورت دو حالت مانند سیگار کشیدن یا نکشیدن باشد بصورت ۱ و ۰ ثبت می شود. به همین شکل متغیرهای دیگر مواجهه و داده ها را می توان تعریف و وارد نمود.
- ۶- اکنون بانک داده برای تحلیل آماده است. با دستور db clogit جعبه معرفی متغیرها را باز می کنیم. یا از مسیر شکل زیر و از منو متغیرها را معرفی می کنیم.



۷- فایل داده ها یا بانک اطلاعات ساخته شده را در زیر ملاحظه نمایید.

Id	match	cases	exposure	job	weight
1	1	1	1	1	45
2	1	0	0	2	48
3	2	1	1	0	45
4	2	0	0	0	25
5	2	1	1	1	45
6	2	0	1	2	36
7	1	1	1	1	58
8	1	0	0	2	48
9	2	1	1	0	41
10	2	0	1	1	52
11	1	1	0	2	48
12	2	0	0	0	78
13	1	1	0	0	52
14	1	0	1	1	25
15	2	1	1	2	45
16	2	0	1	2	65
17	2	1	1	1	32
18	1	0	0	0	58
19	1	1	1	1	56
20	1	0	0	2	45
21	1	1	1	0	25
22	2	0	0	1	85
23	1	1	1	1	56
24	2	0	0	1	56
25	1	1	0	2	89
26	1	0	0	2	25
27	2	1	1	2	58
28	2	0	0	0	45
29	1	1	0	0	45
30	2	0	0	0	54
31	1	1	0	2	58
32	1	0	0	2	55
33	2	1	0	2	62
34	2	0	1	2	62
35	1	1	0	1	32
36	2	0	1	2	32
37	1	1	0	1	32
38	1	0	0	2	45
39	2	1	0	0	45
40	2	0	0	2	45
41	1	1	1	1	45
42	2	0	0	2	44
43	1	1	1	0	66
44	1	0	0	0	22
45	2	1	1	0	21
46	2	0	0	0	28
47	2	1	0	0	84
48	2	0	0	0	45

**دیکشنری متغیرها:** همیشه برای فایل داده و پژوهش باید دیکشنری متغیرها و توصیف آن بر اساس کدها و تعریف های متغیرها و نحوه اندازه گیری و واحد آنها را قبل از تحلیل تهیه نمایید.

**ID:** شماره شناسایی هر نمونه وارد شده به مطالعه است. ۴۸ نمونه یا فرد وارد مطالعه گردید.

**Match:** نمونه های با کدهای همسان شده است. به عبارتی به هر جفت نمونه همسان شده، عدد یکسان داده می شود.

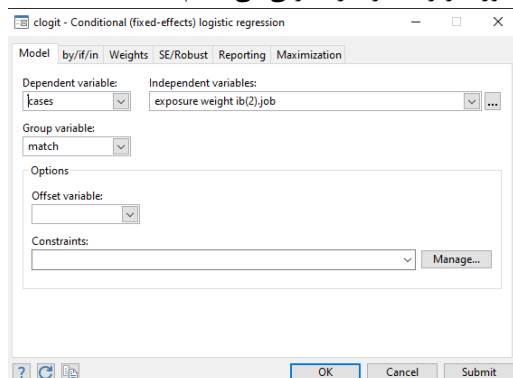
**Cases:** داشتن یا نداشتن پیامد است. برای گروه مورد عدد ۱ و برای گروه شاهد عدد ۰ ثبت شد. افرادی که بر اساس آزمایش قندخون ناشتا به عنوان بیمار دیابتی نوع دو تعریف شدند به عنوان گروه مورد تعریف می گردند. کنترل ها نباید به دیابت نوع دو مبتلا باشند.

**Exposure:** استعمال یا عدم استعمال سیگار با کدهای ۱ و صفر است.

**Job:** در سه طبقه شغلی ۰ و ۱ و ۲ تعریف شد.

**Weight:** وزن افراد بر حسب کیلوگرم بصورت کمی ثبت شد.

با دستور db clogit جعبه معرفی متغیرها را باز و به صورت زیر متغیرها را معرفی می کنیم.



خروجی را میتوان مشاهده نمود. در واقع در تاریخچه می توان دستور اجرا شده زیر را مشاهده نمود:

clogit cases exposure weight ib(2).job, group(match) or  
نتایج و ادامه تحلیل را در صفحه بعد ملاحظه نمایید.

خروجی تحلیل رگرسیون شرطی انجام شده:

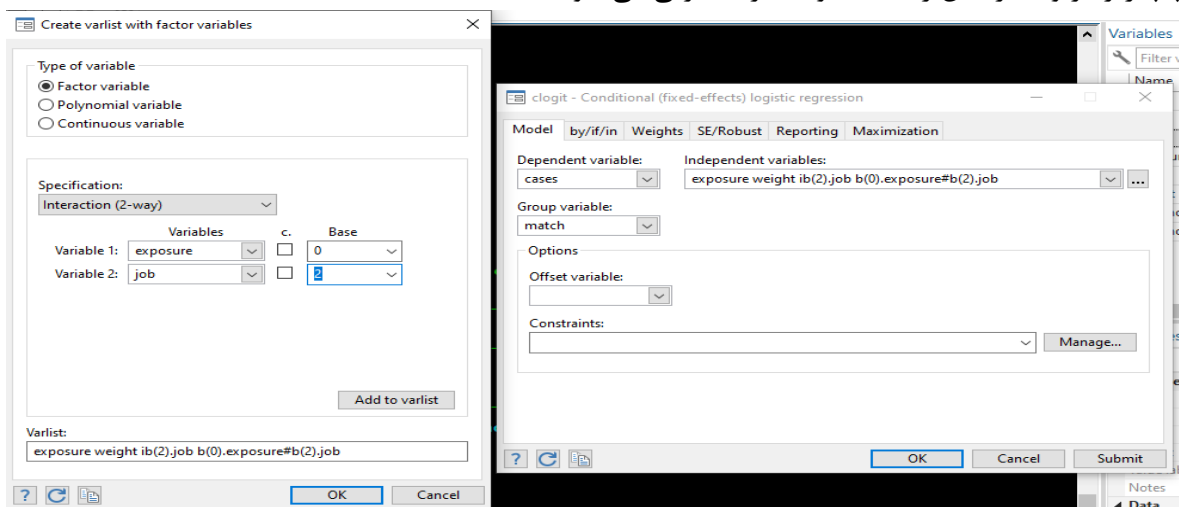
. clogit cases exposure weight ib(2).job, group(match) or  
note: multiple positive outcomes within groups encountered.

Iteration 0: log likelihood = -23.365757  
Iteration 1: log likelihood = -22.930828  
Iteration 2: log likelihood = -22.929941  
Iteration 3: log likelihood = -22.929941

Conditional (fixed-effects) logistic regression      Number of obs =      48  
LR chi2(4) =      12.09  
Prob > chi2 =      0.0167  
Pseudo R2 =      0.2087  
Log likelihood = -22.929941

cases	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
exposure	8.528337	7.331882	2.49	0.013	1.581537	45.98852
weight	1.029615	.021483	1.40	0.162	.9883581	1.072593
job						
0	4.584914	3.803451	1.84	0.066	.9020077	23.30516
1	2.705518	2.345352	1.15	0.251	.4947242	14.79578

نسبت شانس برای مواجهه سیگار ۸/۵۲ ، برای وزن ۱/۰۲ و برای شغل در طبقه صفر برابر ۴/۵ و در طبقه یک برابر ۲/۷ و طبقه شغلی دو را به عنوان مرجع یا گروه مقایسه گرفته است. برای معرفی یک طبقه از یک متغیر کیفی چند حالت به عنوان مرجع از `ib` و داخل پرانتز طبقه مرجع معرفی می شود. در اینجا از `ib(2).job` استفاده گردید. حدود اطمینان ۹۵٪ برای نسبت های شانس محاسبه شده را مشاهده و به ترتیب برای مواجهه با سیگار معنی دار، برای وزن غیرمعنی دار و در واقع حدود اطمینان از ۰/۹۸ تا ۱/۰۷ مقدار یک را دربرگرفته و معنی دار نیست. برای شغل نیز در هیچ یک از طبقات در مقایسه با طبقه دو شغلی معنی دار نشدند. در کنار حدود اطمینان می توان `P-Value` را نیز مشاهده و درباره معنی داری یا عدم معنی داری آماری متغیرهای مستقل نتیجه گیری نمود. با دستور `estimates store model1` مدل اجرا شده را ذخیره می کنیم. سپس مدل دیگری را اجرا می کنیم و اینترکشن سیگار و شغل را لحاظ می کنیم. مسیر معرفی اینترکشن در زیر آمده است. زبانه سه نقطه کنار متغیرهای مستقل در شکل سمت راست را باز می کنیم. سپس منوی سمت چپ زیر باز و اینترکشن را انتخاب و متغیرها معرفی می شود.



خروجی مدل دوم به صورت زیر است.

```
. clogit cases exposure weight ib(2).job b(0).exposure#b(2).job, group(match) or
note: 1.exposure#2.job omitted because of collinearity.
note: multiple positive outcomes within groups encountered.
```

```
Iteration 0: log likelihood = -27.787237
Iteration 1: log likelihood = -21.106483
Iteration 2: log likelihood = -20.714761
Iteration 3: log likelihood = -20.67653
Iteration 4: log likelihood = -20.673308
Iteration 5: log likelihood = -20.672681
Iteration 6: log likelihood = -20.672564
Iteration 7: log likelihood = -20.672553
Iteration 8: log likelihood = -20.67255
Iteration 9: log likelihood = -20.67255
```

```
Conditional (fixed-effects) logistic regression      Number of obs =    48
LR chi2(6) = 16.61
Prob > chi2 = 0.0108
Pseudo R2 = 0.2866
Log likelihood = -20.67255
```

cases	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
exposure	3.505651	4.598671	0.96	0.339	.2680234	45.85267
weight	1.037567	.0236138	1.62	0.105	.9923021	1.084897
job						
0	2.006243	1.986937	0.70	0.482	.287987	13.97636
1	3.106847	3.782012	0.93	0.352	.2858506	33.76762
exposure#job						
1 0	5.46e+07	1.50e+11	0.01	0.995	0	.
1 1	1.10277	2.002095	0.05	0.957	.0314135	38.7127
1 2	1	(omitted)				

با دستور `estimates store model2` مدل اجرا شده را ذخیره می کنیم. سپس دو مدل را با دستور زیر مقایسه و مدل مناسب را انتخاب می کنیم.

```
. lrtest model1 model2, stats
```

```
Likelihood-ratio test
Assumption: model1 nested within model2
```

```
LR chi2(2) = 4.51
Prob > chi2 = 0.1046
```

```
Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
<u>model1</u>	48	-28.97731	-22.92994	4	53.85988	61.34469
<u>model2</u>	48	-28.97731	-20.67255	6	53.3451	64.57231

Note: BIC uses N = number of observations. See [\[R\] BIC note](#).

نتایج مدل سازی نشان می دهد که در مدل دوم اینترکشنی بین سیگار و شغل وجود نداشت. با توجه به شاخص BIC کمتر در مدل اول، می توانیم مدل مذکور را به عنوان مناسب ترین مدل انتخاب کنید.

## مقایسه رگرسیون لجستیک معمولی با رگرسیون لجستیک شرطی

تقریباً نتایج تحلیل رگرسیون لجستیک شرطی با رگرسیون لجستیک معمولی، هنگامی که فقط یک گروه وجود داشته باشد شبیه به هم هستند. به عنوان مثال در فایل اتومبیل ها use auto,clear اگر این فایل را باز و دستورات زیر را اجرا نمایید به نتایج یکسان خواهید رسید. در اولی مقدار ثابت \_cons\_ نیز گزارش شده ولی در دومی گزارش نشد. اگر به آخر دستورات OR اضافه شود خروجی با OR گزارش میشود.

```
gen grp=1
```

```
logit foreign mpg weight
```

```
clogit foreign mpg weight ,group(grp)
```

```
. logit foreign mpg weight
```

```
Iteration 0:  log likelihood = -45.03321
Iteration 1:  log likelihood = -29.238536
Iteration 2:  log likelihood = -27.244139
Iteration 3:  log likelihood = -27.175277
Iteration 4:  log likelihood = -27.175156
Iteration 5:  log likelihood = -27.175156
```

```
Logistic regression                Number of obs   =          74
                                   LR chi2(2)       =          35.72
                                   Prob > chi2      =          0.0000
Log likelihood = -27.175156        Pseudo R2      =          0.3966
```

foreign	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
mpg	-.1685869	.0919175	-1.83	0.067	-.3487418 .011568
weight	-.0039067	.0010116	-3.86	0.000	-.0058894 -.001924
_cons	13.87695	4.603836	3.01	0.003	4.853602 22.90031

```
.
```

```
. clogit foreign mpg weight ,group(grp)
```

```
note: multiple positive outcomes within groups encountered.
```

```
Iteration 0:  log likelihood = -25.175561
Iteration 1:  log likelihood = -25.17194
Iteration 2:  log likelihood = -25.171939
```

```
Conditional (fixed-effects) logistic regression  Number of obs   =          74
                                                  LR chi2(2)     =          35.14
                                                  Prob > chi2    =          0.0000
Log likelihood = -25.171939                    Pseudo R2      =          0.4111
```

foreign	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
mpg	-.1650079	.0908013	-1.82	0.069	-.3429752 .0129594
weight	-.0038276	.0009968	-3.84	0.000	-.0057812 -.0018739

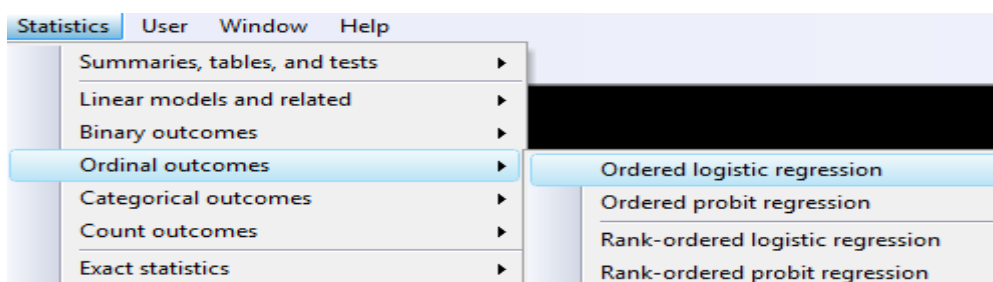
## رگرسیون لوجستیک رتبه ای: Ordinal Logistic Regression

اگر متغیر وابسته ترتیبی (Ordinal) باشد از رگرسیون رتبه ای (Ordered Logistic) استفاده می شود. در فایل polytomus logistic متغیر grade سرطان را بعنوان متغیر وابسته ترتیبی در نظر بگیرید. این متغیر سه طبقه قابل تشخیص (۲)، متوسط (۱) و ضعیف (۰) به شرح زیر دارد.

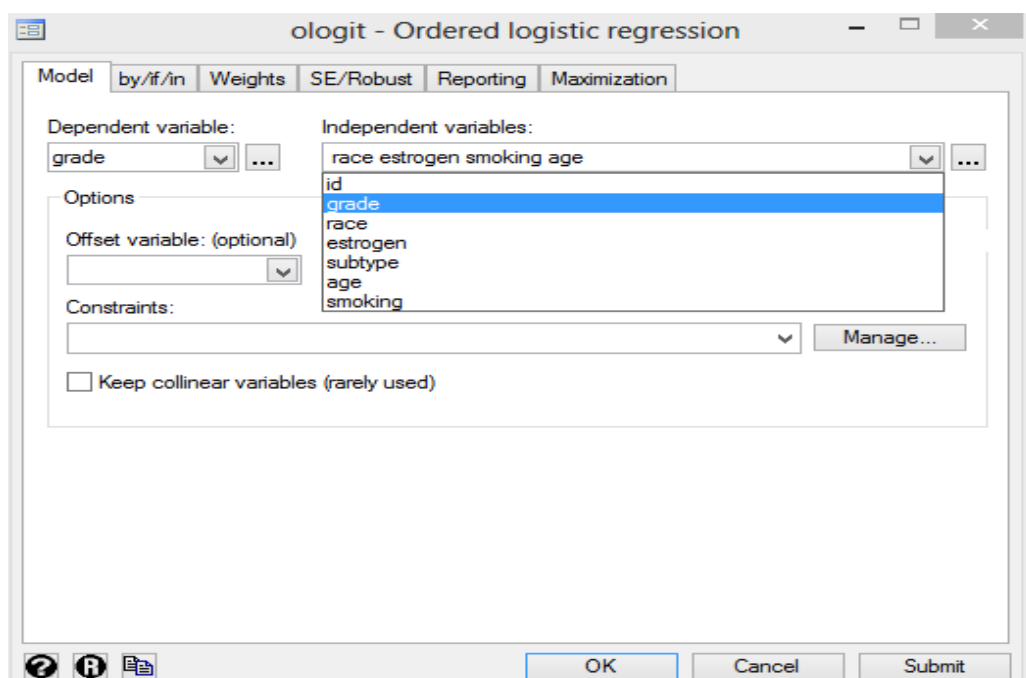
```
. tab grade
```

grade	Freq.	Percent	Cum.
0	130	45.14	45.14
1	105	36.46	81.60
2	53	18.40	100.00
Total	288	100.00	

برای اجرای این تحلیل از مسیر زیر اقدام می شود.



یا با تایپ دستور db ologit کادر محاوره ای زیر باز و متغیرها به شکل زیر تعریف می شوند.



خروجی در جدول زیر نمایش داده می شود.

```
. ologit grade race estrogen smoking age, or
```

```
Iteration 0: log likelihood = -297.46248
Iteration 1: log likelihood = -286.67187
Iteration 2: log likelihood = -286.62328
Iteration 3: log likelihood = -286.62326
```

```
Ordered logistic regression          Number of obs   =      286
LR chi2(4)                          =      21.68
Prob > chi2                          =      0.0002
Pseudo R2                            =      0.0364

Log likelihood = -286.62326
```

grade	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
race	1.5559	.4257398	1.62	0.106	.9100562 2.660084
estrogen	.4634796	.1159627	-3.07	0.002	.283831 .7568355
smoking	1.358202	.5189829	0.80	0.423	.6422612 2.872217
age	1.370221	.3303399	1.31	0.191	.8542376 2.197874
/cut1	-.2782189	.2723221			-.8119603 .2555226
/cut2	1.517787	.2896328			.950117 2.085457

در جدول بالا نیز هر متغیر مستقل یک نسبت شانس برای رخداد دارد. به ازای تعداد سطوح متغیر وابسته منهای یک، نیز ضریب  $\alpha$  وجود دارد به عنوان مثال در بالا دو ضریب cut1 و cut2 به این منظور آمده است. که استفاده ای معمولاً ندارند. فقط برای نوشتن مدل ریاضی رگرسیون کاربرد دارند.

نرم افزار برای وقتی که هیچ متغیر مستقلی به مدل وارد نشود یعنی فرض صفر، Iteration 0 را نیز انجام و log likelihood = -297.46248 بدست می آورد. مدل معنی دار است. متغیر استروژن نیز معنی دار شده است. به ازای ۱ واحد افزایش در مصرف استروژن، در لگاریتم متغیر وابسته (گرید ۲)، 0.464 افزایش می یابد. یعنی با مصرف ۱ واحد بیشتر استروژن، خطر برای سرطان قابل تشخیص log0.464 افزایش می یابد. به عبارت دیگر نسبت شانس ۰/۴۶ در نتیجه مدل نشان می دهد که استروژن اثر محافظت کننده برای سرطان دارد.

در مدل رگرسیون لوجستیک رتبه ای لازم است که فرض تناسب نسبت های شانس رعایت گردد. این فرض با آزمون اسکور انجام می شود. نرم افزار در راند کردن مدل این فرض را ارزیابی می کند.

### تحلیل رگرسیون لوجستیک رتبه ای از نگاه دیگر:

در فایل اتومبیل ها که به عنوان مثال در نرم افزار وجود دارد. متغیر rep78 به عنوان متغیر وابسته ترتیبی به شرح ۵ طبقه زیر از ضعیف تا عالی و متغیر وزن ماشین ها که به عنوان متغیر کمی تعریف شده است، به شرح زیر توصیف و تحلیل می شود.

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
<b>Total</b>	<b>69</b>	<b>100.00</b>	

برای انجام رگرسیون لوجستیک رتبه ای برای متغیرهای مذکور، دستورات زیر را اجرا نمایید.

keep if rep78 !=.

ologit rep78 mpg weight

predict double (p1 p2 p3 p4 p5)

replace mpg=mpg+1

predict double (q1 q2 q3 q4 q5)

gen double o5 =(q5/(q1+q2+q3+q4)) / (p5/(p1+p2+p3+p4))

gen double o4=((q5+q4)/(q1+q2+q3))/((p5+p4)/(p1+p2+p3))

gen double o3=((q5+q4+q3)/(q1+q2))/((p5+p4+p3)/(p1+p2))

gen double o2=((q5+q4+q3+q2)/(q1))/((p5+p4+p3+p2)/(p1))

summarize o5 o4 o3 o2

display exp(\_b[mpg])

خروجی دستورات بالا به عنوان نتایج تحلیل رویکرد مذکور در زیر آمده است. ابتدا نتایج را نگاه کنید. درباره دستورات مذکور و نتایج، توضیحات لازم پس از نتایج آمده است.



```
. keep if rep78 !=.
(5 observations deleted)
```

```
. ologit rep78 mpg weight
```

```
Iteration 0: log likelihood = -93.692061
Iteration 1: log likelihood = -86.794936
Iteration 2: log likelihood = -86.513907
Iteration 3: log likelihood = -86.513267
Iteration 4: log likelihood = -86.513267
```

```
Ordered logistic regression                Number of obs =          69
LR chi2(2) =                             14.36
Prob > chi2 =                             0.0008
Log likelihood = -86.513267              Pseudo R2 =             0.0766
```

rep78	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mpg	.1170693	.0712216	1.64	0.100	-.0225224	.2566611
weight	-.0003287	.0004849	-0.68	0.498	-.001279	.0006217
/cut1	-2.183186	2.982933			-8.029628	3.663255
/cut2	-.4252723	2.918105			-6.144652	5.294108
/cut3	1.911175	2.89813			-3.769055	7.591405
/cut4	3.52733	2.940878			-2.236685	9.291346

```
. predict double (p1 p2 p3 p4 p5)
(option pr assumed; predicted probabilities)
```

```
. replace mpg=mpg+1
(69 real changes made)
```

```
. predict double (q1 q2 q3 q4 q5)
(option pr assumed; predicted probabilities)
```

```
. gen double o5 = (q5 / (q1+q2+q3+q4)) / (p5 / (p1+p2+p3+p4))
```

```
. gen double o4 = ((q5+q4) / (q1+q2+q3)) / ((p5+p4) / (p1+p2+p3))
```

```
. gen double o3 = ((q5+q4+q3) / (q1+q2)) / ((p5+p4+p3) / (p1+p2))
```

```
. gen double o2 = ((q5+q4+q3+q2) / (q1)) / ((p5+p4+p3+p2) / (p1))
```

```
. summarize o5 o4 o3 o2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
o5	69	1.124197	1.22e-15	1.124197	1.124197
o4	69	1.124197	4.21e-16	1.124197	1.124197
o3	69	1.124197	3.29e-16	1.124197	1.124197
o2	69	1.124197	4.62e-16	1.124197	1.124197

```
. display exp(_b[mpg])
1.1241973
```

### توضیحات دستورات و خروجی بالا:

- ابتدا برآورد مدل رگرسیون لوجستیک رتبه ای برای متغیر وابسته و مستقل انجام گردید.
- از p1 تا p5 به عنوان مقادیر احتمال قابل پیش بینی برای متغیر وابسته که ۵ طبقه دارد برآورد گردید. البته جمع مقادیر احتمالات مذکور ۱ است.
- مقدار ۱ واحد به متغیر mpg اضافه گردید و مجدد پیش بینی مقادیر احتمال p1 تا p5 انجام گردید.
- متغیر o به عنوان نسبت شانس ساخته شد. متغیر o5 برابر با تعریف زیر ساخته شد.

$$o5 = (q5 / (q1+q2+q3+q4)) / (p5 / (p1+p2+p3+p4))$$

صورت کسر شانس متغیر وابسته در طبقه ۵ است. وقتی ۱ واحد به متغیر mpg اضافه می شود مخرج کسر شانس همان طبقه وقتی مقادیری به متغیر mpg اضافه نشده است. بنابراین o5 نسبت شانس است.

-متغیر o4 و o3 و o2 نیز به ترتیب مذکور ساخته می شوند. مثلا o4 برای طبقه مساوی و بیشتر از ۴ ساخته می شود و از تعریف یا فرمول زیر استفاده می شود.

$$o4 = ((q5+q4) / (q1+q2+q3)) / ((p5+p4) / (p1+p2+p3))$$

سپس مقادیر o ساخته شده بصورت خلاصه گزارش شده اند. همانگونه که ملاحظه می شود این مقادیر با هم برابر و یکسان هستند. و این مقادیر با  $\exp(-b[mpg])$  خروجی دستور ologit یکسان هستند.  $\exp()$  این مقادیر حساب می شوند. زیرا ologit مقادیر نسبت شانس را گزارش نمی کند. برای محاسبه حدود اطمینان برای نسبت شانس ها به طریق زیر عمل می شود.

ologit rep78 mpg weight

rep78	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
mpg	.1170693	.0712216	1.64	0.100	-.0225224 .2566611
weight	-.0003287	.0004849	-0.68	0.498	-.001279 .0006217
/cut1	-2.066117	3.048036			-8.040157 3.907923
/cut2	-.308203	2.984741			-6.158188 5.541782
/cut3	2.028244	2.96575			-3.784519 7.841008
/cut4	3.6444	3.00898			-2.253093 9.541893

نسبت شانس برای متغیر mpg با  $\exp(0.1170693)$  برابر با 1.124 بدست می آید. و برای متغیر weight با  $\exp(-.0003287)$  برابر 0.99967 بدست می آید. دستورات محاسبه exp به صورت زیر است.

نکته : برای این محاسبات از دستور `display` به شکل زیر استفاده می شود.

```
. display exp(_b[mpg])
1.1241973

. display exp(_b[ weight])
.99967138
```

در خروجی قبلی حدود اطمینان ضرایب مدل نیز وجود داشت با `exp` کردن حدود پایین و بالای ضرایب نسبت شانس بدست می آید. این محاسبات در زیر آمده است. بنابراین نسبت شانس و حدود اطمینان نسبت های شانس برای دو متغیر به شرح زیر است.

`mpg` برای نسبت شانس متغیر  $1.12(0.799 - 1.29)$

`weight` برای نسبت شانس متغیر  $0.99967138 (0.99 - 1.0006219)$

```
. display exp([-0.225224])
.79833737

. display exp([0.2566611])
1.292607

. display exp([-0.001279])
.99872182

. display exp([0.0006217])
1.0006219
```

راه آسان تر برای محاسبه حدود اطمینان استفاده از دستورات زیر است. خروجی در زیر آمده است.

```
. lincom mpg,or
```

```
( 1) [rep78]mpg = 0
```

rep78	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	1.124197	.0800671	1.64	0.100	.9777293 1.292607

```
. lincom weight,or
```

```
( 1) [rep78]weight = 0
```

rep78	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	.9996714	.0004847	-0.68	0.498	.9987218 1.000622

بعد از محاسبات بالا مقادیر احتمالات پیش بینی شده نیز محاسبه می گردد. قبلا در ابتدای بحث دستورات ذکر گردید. اگر مقادیر شانس مد نظر باشد به طریق زیر اقدام و نتایج به شرح زیر است.

```
. gen odds5=p5/(p1+p2+p3+p4)
.
. gen odds4=(p5+p4)/(p1+p2+p3)
.
. gen odds3=(p5+p4+p3)/(p1+p2)
.
. gen odds2=(p5+p4+p3+p2)/(p1)
. summarize o5 o4 o3 o2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
o5	69	1.124197	1.22e-15	1.124197	1.124197
o4	69	1.124197	4.21e-16	1.124197	1.124197
o3	69	1.124197	3.29e-16	1.124197	1.124197
o2	69	1.124197	4.62e-16	1.124197	1.124197

خروجی بالا مقادیر متغیرهای ساخته شده را نشان می دهد.

## Multinomial Logistic Regression

## رگرسیون لوجستیک چند وجهی:

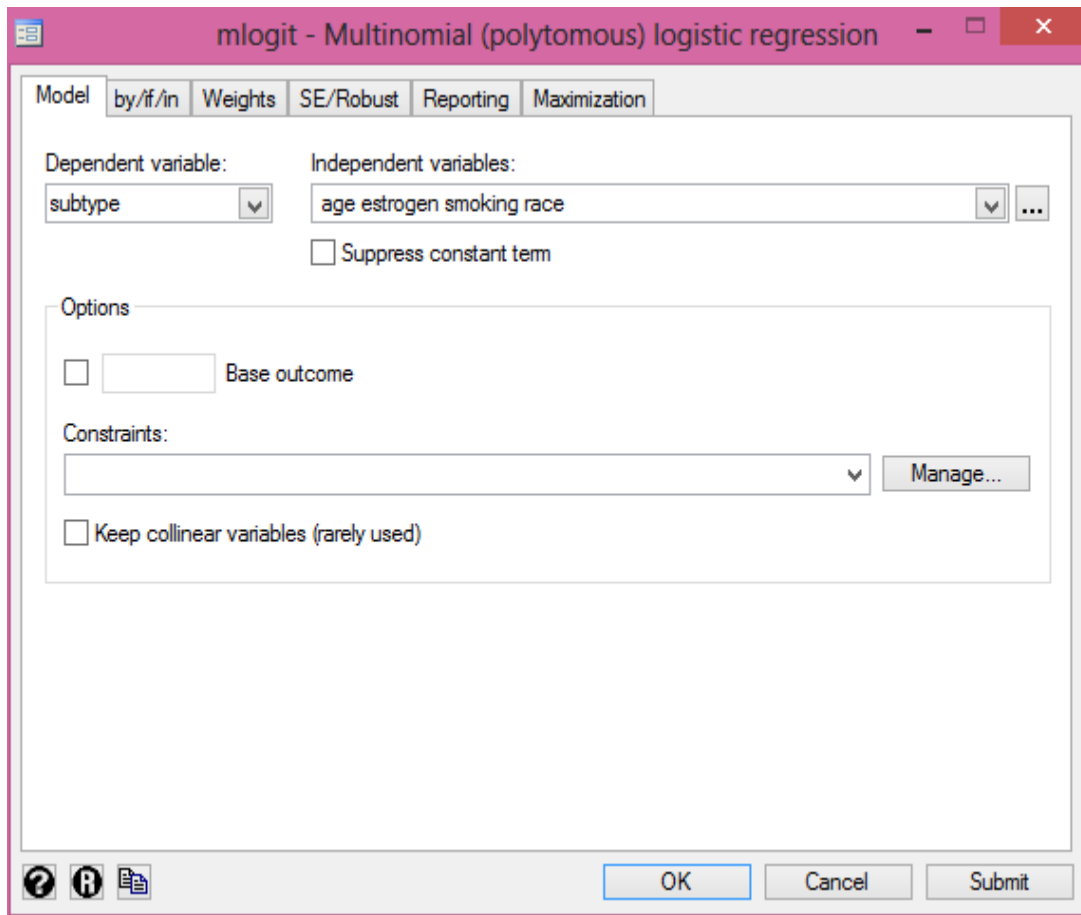
اگر متغیر وابسته، اسمی چندوجهی (Multinomial) باشد، از رگرسیون لوجستیک چندوجهی استفاده می شود. توضیحات مبنای انجام رگرسیون لوجستیک در بالا بصورت مفصل آمده است. اما آنچه در اینجا قابل بحث است این است که نوع آنالیز برای وقتی که متغیر وابسته چندسطحی باشد از رگرسیون لوجستیک معمولی با متغیر پاسخ دو طبقه ای، متفاوت است. اگر متغیر وابسته دو طبقه یا سطح داشته باشد شانس و نسبت های خطر با هم یکسان هستند اما اگر متغیر وابسته بیش از دو طبقه باشد این مقادیر با هم متفاوت خواهد شد. جدول مثال زیر این اختلاف را نشان می دهد.

Category	Probability	Odds	Rel. Risk rel. to category 3
1	.3	.429	1.5
2	.5	1.000	2.5
3 (base)	.2	.250	1.0

برای استفاده از این نوع مدل در نرم افزار، لطفا فایل `polytomus logistic` را باز نمایید. از منوی اصلی نرم افزار از گزینه `statistics` منوی زیر را انتخاب نمایید یا کادر محاوره ای مربوطه را با دستور `mlogit` باز نمایید.

Categorical outcomes	Multinomial logistic regression
Count outcomes	Independent multinomial probit
Exact statistics	Conditional logistic regression
Endogenous covariates	Alternative-specific conditional logit
Sample-selection models	Alternative-specific multinomial probit
Multilevel mixed-effects models	Stereotype logistic regression
Generalized linear models	Nested logit regression
Nonparametric analysis	Setup for nested logit regression
Time series	Display nested logit tree structure

کادر محاوره ای در زیر آمده است. متغیر وابسته چند سطحی `subtype` است. این متغیر انواع سرطان های آدنو کارسینوما، آدنواسکوموس و سایر را نشان می دهد. متغیر های مستقل سن، استروژن، سیگار کشیدن و نژاد است که برای متغیرهای مربوطه به صورت زیر در کادر محاوره ای تعریف می شوند.



در متغیر پیامد نرم افزار کد صفر یا کمتر را به عنوان رفرانس می گیرد. تعداد برآوردها برای ضرایب متغیرهای مستقل و  $\alpha$  (آلفا) به اندازه یک منهای سطوح متغیر وابسته است. یعنی در اینجا که متغیر وابسته سه سطح دارد، هر یک از متغیرهای مستقل دو ضریب دارند.

```
. tab subtype
```

subtype	Freq.	Percent	Cum.
0	186	64.58	64.58
1	45	15.63	80.21
2	57	19.79	100.00
<b>Total</b>	<b>288</b>	<b>100.00</b>	

خروجی در زیر نشان داده شده است. همچنین می توان در گزینه reporting، با انتخاب relative risk ratio ضرایب تنظیم گردند.

```
. mlogit subtype age estrogen smoking race, rrr
```

```
Iteration 0: log likelihood = -256.31174
Iteration 1: log likelihood = -247.58596
Iteration 2: log likelihood = -246.97421
Iteration 3: log likelihood = -246.9652
Iteration 4: log likelihood = -246.96519
```

```
Multinomial logistic regression      Number of obs =      286
LR chi2(8) =      18.69
Prob > chi2 =      0.0166
Pseudo R2 =      0.0365
Log likelihood = -246.96519
```

subtype	RRR	Std. Err.	z	P> z	[95% Conf. Interval]	
0	(base outcome)					
1						
age	2.659473	1.095112	2.38	0.018	1.18656	5.960758
estrogen	.4827819	.1823929	-1.93	0.054	.2302347	1.012351
smoking	2.43463	1.280137	1.69	0.091	.8686962	6.823354
race	.8064241	.3328374	-0.52	0.602	.3591245	1.810848
_cons	.1694265	.0757618	-3.97	0.000	.0705266	.4070146
2						
age	1.331885	.437618	0.87	0.383	.6994988	2.535983
estrogen	.942988	.3229943	-0.17	0.864	.4818932	1.845277
smoking	.1664359	.1742136	-1.71	0.087	.0213928	1.294869
race	1.127685	.4242584	0.32	0.749	.5394465	2.357368
_cons	.2818773	.1065329	-3.35	0.001	.1343868	.59124

در جدول بالا مدل لوجستیک ساخته شده با LR=18.69 و P=0.0166 معنی دار است. تعداد مشاهدات ۲۸۶ نفر بوده است. برای هر متغیر مستقل، یک ردیف در جدول خروجی بالا نتایج آمده است. در ردیف اول طبقه دوم با طبقه اول و در ردیف دوم طبقه سوم با طبقه اول مقایسه شده است. یعنی طبقه اول بعنوان رفرانس گرفته شده است. می توان گروه رفرانس را با انتخاب base outcome در گزینه option کادر محاوره ای تغییر داد. خطر افزایش سرطان آدنوکارسینوما با یکسال افزایش در متغیر سن، ( ln 2.65 ) و برای سایر سرطان ها با یکسال افزایش در سن ، ( ln 1.33 ) بار بیشتر می شود. وقتی که متغیرهای سیگار کشیدن و نژاد و استروژن درمانی کنترل شده باشند. معادله را بصورت زیر می توان نوشت.

$$\log(P(\text{subtype}=1)/P(\text{subtype}=0))=0.170+2.65\text{age}+0.49\text{estrogen}+0.17\text{smoking} +0.8 \text{ race}$$

$$\log(P(\text{subtype}=2)/P(\text{brand}=0)) = 0.29+1.33\text{age}+0.94\text{estrogen}+0.17 \text{ smoking} +1.13 \text{ race}$$

**Poisson Regression**      رگرسیون پواسون:

آنچه تا کنون برای تحلیل و مدل سازی داده ها گفته شد متفاوت از تحلیل رگرسیون پواسون است. جدول زیر ساختار داده های متغیر وابسته که به صورت کمی پیوسته و دوحالته باشند و نوع آنالیز مورد نیاز را نشان می دهد. این آنالیزها قبلا مفصل بحث شدند.

Type of Data	Analytic Model(s)
Continuous	t-test ANOVA Linear Regression
Binary	Logistic Regression

وقتی متغیر وابسته به صورت شمارشی باشد مدل تحلیل مورد استفاده برای داده ها رگرسیون پواسون است. برای محاسبه نسبت میزان های خام و میزان های استاندارد شده و کنترل متغیرهای مخدوش کننده احتمالی در مطالعات اپیدمیولوژیک و مخصوصا مطالعه کوهورت از تحلیل رگرسیون پواسون برای مدل سازی استفاده می شود. ساختار داده ها نه تنها شمارشی است بلکه ممکن است در پریودهای زمانی مختلف این مقادیر اندازه گیری و به صورت تعداد ماه های تحت نظر یا شخص - زمان پی گیری شده باشند. این آنالیز از توزیع پواسون که احتمال وقوع رخداد متغیر پیوسته را محاسبه می کند، گرفته شده است. میانگین و واریانس در توزیع پواسن با هم برابر است. مدل عمومی رگرسیون پواسون به صورت زیر است. هنگامی که رخداد متغیر وابسته نادر است و مدت پی گیری افراد در مطالعه متفاوت باشد، استفاده از این آنالیز مناسب است. به عنوان مثال متغیر Y را بعنوان متغیر وابسته و تعداد موارد بیماری سل در نظر بگیرید.  $\lambda$  نیز میزان رخداد موارد بیماری در واحد زمان است. V تعداد واحدهای زمانی است. مثلا شخص - سال.

$$\Pr(Y = y) = \frac{e^{-\lambda V} (\lambda V)^y}{y!}$$

$$\text{Log}(\# \text{ of counts}) = B_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

If  $\ln(\lambda) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$  ,  
 then  $\lambda = e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}$

$$\text{RR (male vs female)} = \frac{e^{\beta_0 + \beta_1 \text{Age} + \beta_2}}{e^{\beta_0 + \beta_1 \text{Age}}}$$

$$= e^{\beta_2}$$

$$\Rightarrow \beta_2 = \ln(\text{RR}_{\text{males vs females}})$$



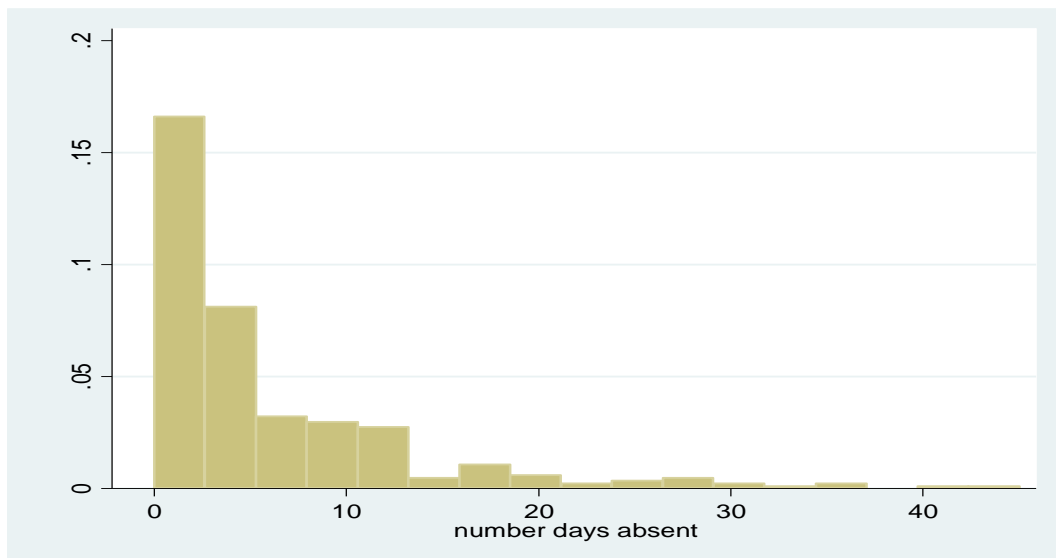
در معادله بالا دو متغیر مستقل سن و جنسیت وارد مدل شده و ضرایب آنها به عنوان نسبت میزان بروز متغیر وابسته گزارش می شود.

### کار با نرم افزار و انجام تحلیل رگرسیون پواسون:

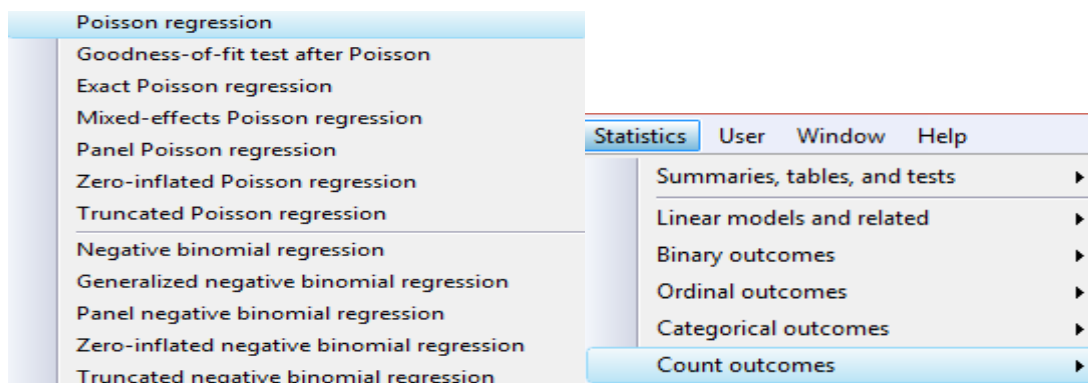
فایل داده های مربوطه بنام poisson regression را باز نمایید. در این فایل متغیری بنام روزهای غیبت از مدرسه در طی یکسال بنام daysabs مربوط به ۳۱۶ دانش آموز وجود دارد. این متغیر به صورت شمارشی روزهای غیبت از مدرسه را نشان می دهد. متغیر جنسیت با gender و متغیر school نوع مدرسه است که با کد ۱ و ۲ مشخص شده است. متغیرهای دیگری نیز در فایل وجود دارد که در زیر آمده است. در تحلیل بررسی اثر دو متغیر نمره استاندارد شده دانش آموز در درس زبان (langnce) و درس ریاضی (mathnce) مد نظر است.

id	float	%9.0g		
gender	float	%9.0g	g1	
ethnic	float	%10.0g	e1	ethnicity
school	float	%9.0g		school 1 or 2
mathpr	float	%9.0g		ctbs math pct rank
langpr	float	%9.0g		ctbs lang pct rank
mathnce	float	%9.0g		ctbs math nce
langnce	float	%9.0g		ctbs lang nce
biling	float	%12.0g	b1	bilingual status
daysabs	float	%9.0g		number days absent

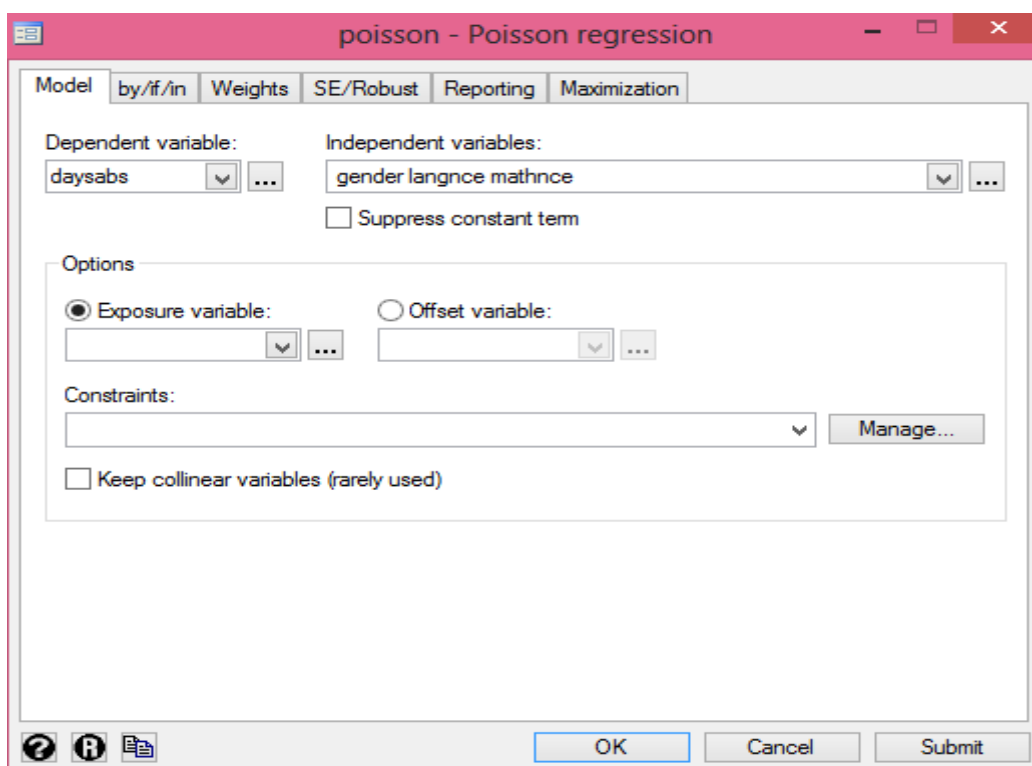
هدف بررسی عوامل موثر بر غیبت دانش آموزان در مدرسه است. هیستوگرام متغیر وابسته یا روزهای غیبت در مدرسه در زیر آمده است. به نظر داده ها به طور قابل توجهی چولگی به راست دارند. و رگرسیون ( OLS regression) برای تحلیل نامناسب است. اما با توجه به اینکه داده ها شمارشی هستند انجام رگرسیون پواسون شاید مناسب باشد.



از نظر تئوری در توزیع پواسون میانگین و واریانس با هم برابر است. با دستور summarize daysabs, detail فرض مورد بررسی قرار می گیرد. وقتی دستور اجرا می شود نتایج نشان می دهد که واریانس و میانگین متغیر مذکور با هم برابر نیست. این مقادیر برای میانگین 5.810127 با انحراف معیار 7.449003 و برای واریانس 55.4876 است. واریانس نزدیک به ده برابر بیشتر از میانگین است. قبل از اینکه سراغ آنالیز دیگر بروید آنالیز رگرسیون پواسون را با توجه به شمارشی بودن متغیر وابسته با در نظر گرفتن اینکه این نوع تحلیل نامناسب است انجام دهید. برای انجام تحلیل رگرسیون پواسون از مسیر زیر اقدام نمایید.



یا با استفاده از دستور db poisson کادر محاوره ای مربوطه را باز و متغیرها را به صورت زیر معرفی نمایید.



خروجی تحلیل در زیر آمده است.

Iteration 1: log likelihood = -1547.9709

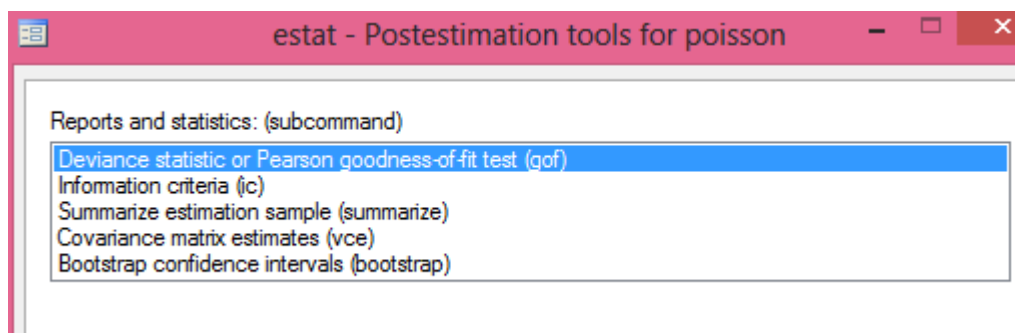
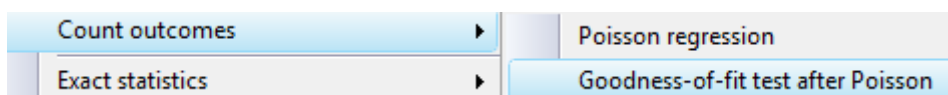
Poisson regression

Number of obs = 316  
 LR chi2(3) = 175.27  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.0536

Log likelihood = -1547.9709

daysabs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gender	-.4009209	.0484122	-8.28	0.000	-.495807 - .3060348
langnce	-.0121521	.0018348	-6.62	0.000	-.0157483 - .0085559
mathnce	-.0035232	.0018213	-1.93	0.053	-.007093 .0000466
_cons	3.088587	.1017365	30.36	0.000	2.889187 3.287987

برازش مدل پس از اجراء با استفاده از مسیر زیر ارزیابی می شود.



مقدار معنی داری آزمون زیر نشان می دهد که مدل مناسب نیست. این دلیل بر دلیل بالا که واریانس و میانگین یکسان نبودند نیز افزوده شده و لذا نشان می دهد که تحلیل رگرسیون پواسون مناسب نیست.

. poisgof

```
Deviance goodness-of-fit = 2234.546
Prob > chi2(312) = 0.0000

Pearson goodness-of-fit = 2774.414
Prob > chi2(312) = 0.0000
```

در اینگونه موارد، از توزیع منفی دوجمله ای (negative binomial analysis) برای آنالیز نیز استفاده می شود. وقتی که با دستور nbreg daysabs gender mathnce langnce این تحلیل انجام می شود نتایج به صورت زیر است. likelihood ratio test در بالای خروجی، overdispersion یا بیش پراکندگی پارامتر آلفا را نشان می دهد. وقتی بیش پراکندگی پارامتر آلفا در توزیع داده ها صفر باشد توزیع منفی دو جمله ای با توزیع پواسون برابر و یکسان است. در اینجا با توجه به معنی داری آلفا، و تفاوت با صفر، توزیع پواسون نیز برقرار نیست.

```
Negative binomial regression      Number of obs   =       316
LR chi2(3)                       =       20.74
Dispersion   = mean              Prob > chi2     =       0.0001
Log likelihood = -880.87312      Pseudo R2      =       0.0116
```

daysabs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gender	-.4311844	.1396656	-3.09	0.002	-.704924	-.1574448
mathnce	-.001601	.00485	-0.33	0.741	-.0111067	.0079048
langnce	-.0143475	.0055815	-2.57	0.010	-.0252871	-.003408
_cons	3.147254	.3211669	9.80	0.000	2.517778	3.776729
/lnalpha	.2533877	.0955362			.0661402	.4406351
alpha	1.288383	.1230871			1.068377	1.553694

Likelihood-ratio test of alpha=0: chibar2(01) = 1334.20 Prob>=chibar2 = 0.000

در خروجی بالا نیز نتایج نشان می دهد که هر دوی متغیر های gender و langnce معنی دار هستند ولی mathnce معنی دار نیست. برای جنسیت کد ۱ و ۲ وجود دارد. و این نشان می دهد غیبت دانش آموزان دختر در مدرسه نسبت به پسران بیشتر و معنی دار است. ضریب معنی دار برای langnce نشان می دهد که دانش آموزانی که توانایی بالاتری در درس زبان دارند، غیبت کمتری دارند نسبت به دانش آموزان که توانایی کمتری در درس زبان دارند. چنانچه کاربر ضرایب را در مقیاس نسبت میزان بروز بخواد گزارش نماید در قسمت reporting با انتخاب گزینه مربوطه نتایج به صورت زیر نمایش داده می شود.

daysabs	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
gender	.669703	.0324218	-8.28	0.000	.6090792	.736361
langnce	.9879214	.0018127	-6.62	0.000	.984375	.9914806
mathnce	.996483	.0018149	-1.93	0.053	.9929321	1.000047

نتایج نشان می دهد که دانش آموزان دختر نسبت به پسران و دانش آموزانی که نمره زبان بهتری دارند نسبت به دانش آموزانی که نمراتشان در این درس کمتر است در مدرسه غیبت کمتری دارند. نمره ریاضی معنی دار نیست و بر غیبت در مدرسه اثری ندارد.

**رویکرد تحلیل دیگر در رگرسیون پواسون،** بررسی میزان بروز یک رخداد یا واقعه در واحد زمان و استفاده از شخص – سال است. وقتی در دوره های مطالعه بتوان فرض گذاشت که میزان مورد نظر، با مرور زمان تغییر زیادی نکند. یا در طی دوره های زمانی مشخص، میزان ثابت باشد از این روش تحلیل استفاده می شود. در رگرسیون پواسون برای تحلیل میزان (Rate)، با فرض عدم وجود اینترکشن یا multiplicativity از مسیر زیر اقدام می شود.

فایل POISSON REGRESSION RATE را باز نمایید. این فایل یک نمونه ده درصدی از داده های مطالعه کوهورت Whitehall است. این مطالعه به منظور بررسی عوامل خطر مرگ های ناشی از بیماری ایسکمیک قلب انجام شده است. ابتدا تحلیل اولیه ، برای محاسبه میزان ها با روش معمول انجام می گیرد. سپس به تحلیل

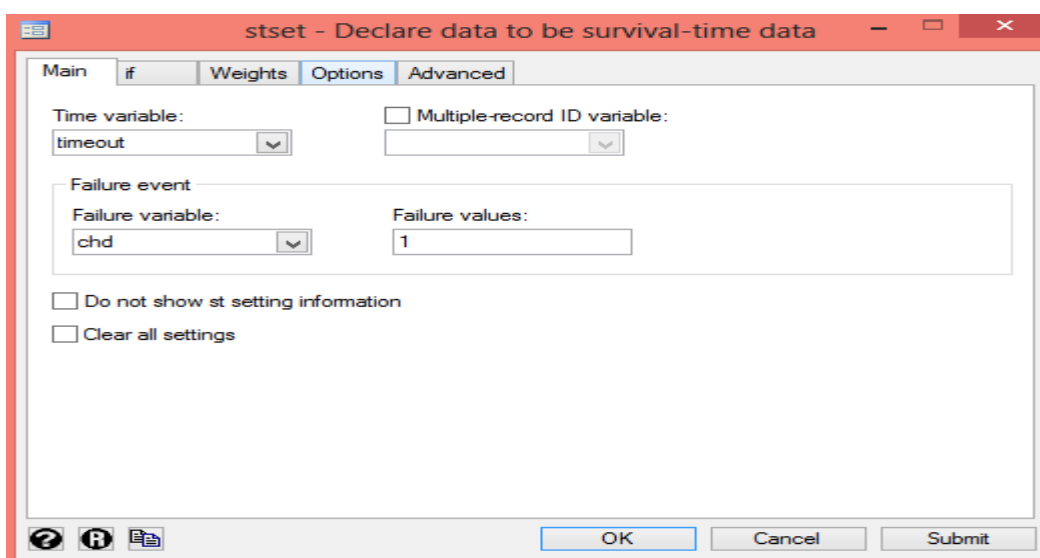
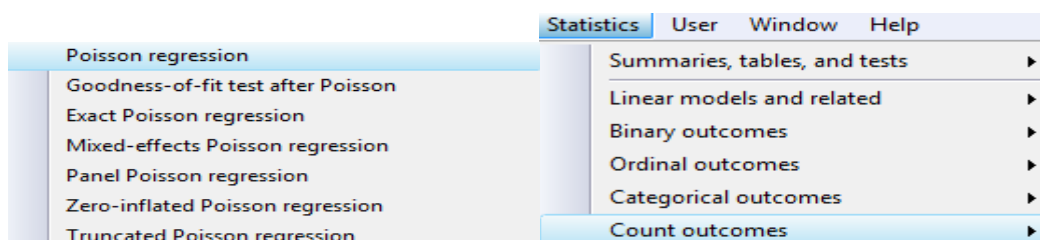
رگرسیون پواسون پرداخته می شود. به این منظور برای معرفی داده ها به نرم افزار و بیان مقدار پی گیری افراد، یعنی زمان ورود و خروج افراد به مطالعه و رخداد واقعه برای آنها کادر محاوره ای مزبور با دستور db stset می گردد. یا با دستور زیر متغیرها به نرم افزار معرفی و زمان تنظیم می گردد.

stset timeout, fail(chd) id (id) origin(timebth) enter(timein) scale(365.25)

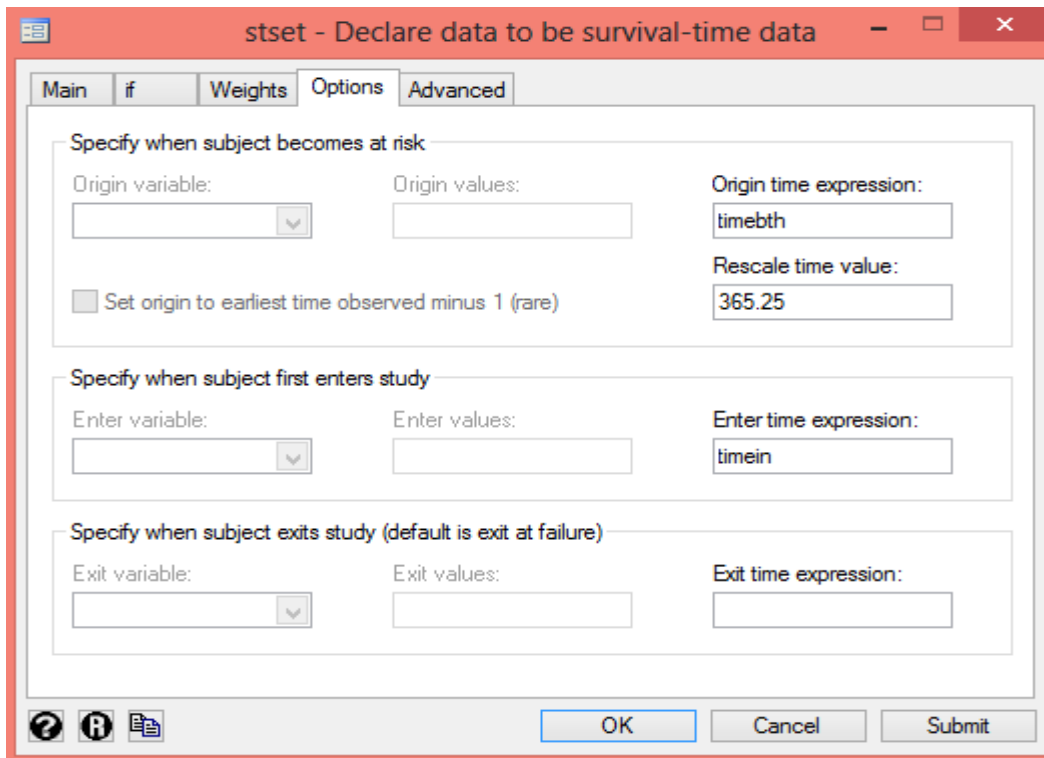
### در دستور بالا:

بعد از دستور stset متغیر زمان یا تاریخ خروج افراد از مطالعه می آید. بعد از آن در پرانتز fail متغیر رخداد واقعه یا متغیری که فرد به دلیل آن از مطالعه خارج می شود که در اینجا مرگ ناشی از بیماری ایسکمیک قلب (chd) است وارد می گردد. متغیر id نیز شماره شناسایی هر فرد است که در پرانتز تایپ می گردد. در داخل پرانتز origin نیز متغیر تاریخ یا زمان در معرض خطر بودن فرد در مطالعه بسته به اینکه زمان مواجهه چه موقع در نظر گرفته شود وارد می شود. اگر زمان مواجهه از زمان تولد افراد باشد متغیر تاریخ تولد افراد وارد می شود.

اگر مواجهه از زمان ورود فرد به مطالعه باشد نیز متغیر مربوطه وارد می شود. در اینجا زمان تولد افراد timebth مد نظر بوده و وارد شده است. در پرانتز enter نیز زمان ورود فرد به مطالعه که در اینجا متغیر timein است وارد می شود. این زمان برای بعضی از مطالعات با متغیر origin یکسان است. و در انتهای دستور نیز مقیاس زمان مطالعه با توجه به سال کیسه و تاریخ نرم افزار Stata که از ۱ ژوئن ۱۹۶۰ تنظیم شده ، عدد ۳۶۵/۲۵ وارد می شود و با این زمان شخص سال تبدیل می گردد. از مسیر زیر نیز می توان این تنظیمات را انجام داد.



از سربرگ option در کادر محاوره ای بالا بقیه متغیر به شکل زیر تعریف می شوند.



اکنون با توجه به تنظیمات بالا که نتیجه آن در زیر آمده است، می توان تحلیل مورد نیاز را انجام داد.

```
. stset timeout, failure(chd==1) enter(time timein) origin(time timebth) id(id) scale(
> 365.25)
```

```

      id:  id
failure event:  chd == 1
obs. time interval:  (timeout[_n-1], timeout]
enter on or after:  time timein
exit on or before:  failure
t for analysis:  (time-origin)/365.25
      origin:  time timebth
```

---

```
1677 total obs.
      0 exclusions
```

---

```
1677 obs. remaining, representing
1677 subjects
154 failures in single failure-per-subject data
27605.37 total analysis time at risk, at risk from t = 0
          earliest observed entry t = 40.052
          last observed exit t = 85.76587
```

خروجی بالا متغیرهای معرفی شده و مدت زمان در معرض خطر بودن افراد که 27605.37 سال است و تعداد موارد رخداد مرگ ناشی از ایسکمیک قلبی که ۱۵۴ مورد بوده و کمترین زمان و بیشترین زمان برای سن افراد پی گیری شده در مطالعه را نشان می دهد. به عبارتی نشان می دهد که فردی که ۶۰ سال سن دارد ۶۰ سال در معرض خطر بوده و فردی که ۲۰ سال سن دارد ۲۰ سال در معرض خطر بوده است. بر سن افراد به نوعی تطبیق انجام می شود. چنانچه در تنظیمات بالا در پراتنز origin بجای تاریخ تولد افراد، متغیر زمان ورود به مطالعه قرار گیرد خروجی مذکور به شرح زیر متفاوت می گردد. به عبارتی در تنظیمات بالا متغیر زمان نیز کنترل می گردد ولی در خروجی زیر ، اثر سن افراد کنترل نمی شود.

```
27605.37 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 19.38123
```

اولین گام در تحلیل یک مطالعه کوهورت و از جمله مطالعه مذکور، اندازه گیری میزان بروز است.

در اینجا محاسبه بروز مرگ ناشی از بیماری ایسکمیک قلب مد نظر است. برای این محاسبه با دستور `strate` خروجی زیر بدست می آید.

D	Y	Rate	Lower	Upper
154	2.8e+04	0.0055786	0.0047636	0.0065331

میزان بروز مرگ ناشی از بیماری ایسکمیک قلب و حدود اطمینان آن در جدول آمده است. چنانچه کاربر بخواهد این میزان را در هزار نفر شخص سال و بر اساس متغیر `grade` محاسبه نماید با دستور `strate grade`, per(1000) نتیجه به شرح زیر می باشد.

grade	D	Y	Rate	Lower	Upper
1	90	20.3398	4.4248	3.5989	5.4403
2	64	7.2656	8.8087	6.8946	11.2541

از تقسیم دو میزان، نسبت خطر یا نسبت میزان (Rate Ratio) بدست می آید. برای محاسبه نسبت دو میزان در متغیر `grade` نتیجه به شکل زیر است. این کار با دستور `stmh grade` انجام می شود.

نکته : دستور `stmh` همیشه بعد از `stset` بکار می رود.

Maximum likelihood estimate of the rate ratio  
comparing grade==2 vs. grade==1

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
1.991	18.44	0.0000	1.445	2.743

خروجی بالا نشان می دهد که خطر مرگ ناشی از بیماری ایسکمیک قلب در گرید دو نسبت به گرید یک حدود ۲ برابر یا (1.991) بار بیشتر است. این مقدار نیز معنی دار است. حدود اطمینان آن نیز گزارش گردید. ممکن است سن در این تحلیل متغیر مخدوش کننده باشد. اما فرض گردید که در طی زمان ، میزان های بروز مرگ برای سنین مختلف یکسان است. اما فرض بهتر این است که این میزان در طی سنین مختلف ثابت باشد. لذا برای کنترل اثر سن، متغیر سن (currage) را با توجه به دستور زیر ساخته و آن را به گروه های سنی مختلف تقسیم و سپس میزان ها محاسبه می گردد.

```
. stsplot curage, at(40, 50(5) 80,90)
(5243 observations (episodes) created)
```

خروجی زیر مفصل است. اولاً نسبت میزان های بروز مرگ ایسکمیک قلبی در گروه های سنی مختلف برای گرید دو به یک در جدول زیر آمده است. مثلاً نسبت بروز در گروه سنی ۵۰ ساله ها و برای گرید دو نسبت به گرید اول برابر با 2.35 می باشد. این نسبت به صورت کلی با کنترل متغیر سن، 1.391 بدست آمد. همچنین نتایج نشان می دهد که نسبت های میزان مرگ در سنین مختلف با یکدیگر متفاوت هستند و بنابراین بین سن و گرید اینترکشن وجود دارد. چنانچه اینترکشن را در نظر نگیریم و این نسبت میزان که برای سن تطبیق شده است را با نسبت میزان خام که قبلاً 1.99 بدست آمد مقایسه شود مشخص است که میزان تطبیق شده کمتر از میزان خام است.



```
. stmh grade , by( curage)
```

```

failure _d: chd == 1
analysis time _t: (timeout-origin)/365.25
origin: time timebth
enter on or after: time timein
id: id

```

Maximum likelihood estimate of the rate ratio  
 comparing grade==2 vs. grade==1  
 by curage

RR estimate, and lower and upper 95% confidence limits

curage	RR	Lower	Upper
40	.	.	.
50	2.35	0.59	9.39
55	1.56	0.55	4.43
60	2.18	1.15	4.12
65	1.84	1.00	3.40
70	0.61	0.28	1.34
75	0.62	0.22	1.75
80	1.66	0.17	15.96

Overall estimate controlling for curage

RR	chi2	P>chi2	[95% Conf. Interval]	
1.391	4.26	0.0391	1.015	1.905

```

Approx test for unequal RRs (effect modification): chi2(7) = 14.32
Pr>chi2 = 0.0458

```

چنانچه از متغیر `agecat` شش گروه سنی با دستور زیر درست شود. سپس نسبت میزان ها در هر گروه سنی محاسبه گردد. نتیجه به صورت زیر خواهد بود.

```
egen agecat=cut(agein), at(40,45,50,55,60,65,70)
```

```
. stmh grade, by( agecat)
```

```

failure _d: chd == 1
analysis time _t: (timeout-origin)/365.25
              origin: time timebth
enter on or after: time timein
              id: id

```

Maximum likelihood estimate of the rate ratio  
 comparing grade==2 vs. grade==1  
 by agecat

RR estimate, and lower and upper 95% confidence limits

agecat	RR	Lower	Upper
40	2.29	0.61	8.62
45	1.10	0.31	3.85
50	2.33	1.18	4.61
55	1.48	0.85	2.59
60	0.67	0.33	1.35
65	0.67	0.18	2.51

Overall estimate controlling for agecat

RR	chi2	P>chi2	[95% Conf. Interval]	
1.286	2.35	0.1249	0.932	1.774

Approx test for unequal RRs (effect modification): chi2(5) = 8.46  
 Pr>chi2 = 0.1326

نسبت های میزان در گروه های سنی مختلف در جدول بالا آمده است. نسبت میزان کلی با کنترل متغیر سن در قالب گروه های سنی ساخته شده حدود 1.286 است. این نسبت معنی دار نیست و نشان می دهد با توجه به گروه های سنی ساخته شده نسبت میزان مرگ ایسکمیک قلبی یکسان است. و بنابراین اینترکشن هم وجود ندارد. در مقایسه این نسبت تطبیق شده یک کاسه یا کلی (1.286) با نسبت خام (1.99)، مشخص است که سن می تواند بعنوان یک مخدوش کننده مد نظر باشد. چنانچه اختلاف مقدار تطبیق شده با خام بیش از 10% بود مخدوش کننده وجود دارد.

این اختلاف به صورت زیر محاسبه می شود.

$$\text{display } (1.99 - 1.286) / 1.286 = 0.4575$$

برای برازش مدل رگرسیون پواسون به داده های فایل مذکور و برآورد اثر خام سن از دستورات زیر استفاده می شود. ابتدا باید متغیر زمان پی گیری افراد را ساخته و سپس تحلیل انجام شود. در اینجا نیازی به دستور St و تنظیم زمان نیست. خروجی به شرح زیر است.

```
. gen y=( timeout- timein)/365.25

. xi: poisson chd i.grade, e(y)
i.grade          _Igrade_1-2          (naturally coded; _Igrade_1 omitted)

Iteration 0:    log likelihood =  -619.5724
Iteration 1:    log likelihood =  -619.57209
Iteration 2:    log likelihood =  -619.57209

Poisson regression                               Number of obs   =       1677
                                                LR chi2(1)      =       16.76
                                                Prob > chi2     =       0.0000
Log likelihood = -619.57209                    Pseudo R2      =       0.0133
```

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Igrade_2	.6885037	.1635118	4.21	0.000	.3680264	1.008981
_cons	-5.420525	.1054093	-51.42	0.000	-5.627123	-5.213926
ln(y)	1	(exposure)				

در مقایسه خروجی بالا با نتایجی که با دستور strate بدست آمد وقتی ضریب جدول بالا در مقیاس لگاریتم طبیعی محاسبه شود با نتایج قبلی دستور strate یکسان است. ضریب \_cons نیز لگاریتم میزان در گروه رفرانس را نشان می دهد. چنانچه کاربر بخواهد نسبت های میزان را گزارش دهد در انتهای دستور IIR نیز اضافه می شود. خروجی به شکل زیر خواهد بود.

```
. xi: poisson chd i.grade, e(y) irr
i.grade          _Igrade_1-2          (naturally coded; _Igrade_1 omitted)

Iteration 0:    log likelihood =  -619.5724
Iteration 1:    log likelihood =  -619.57209
Iteration 2:    log likelihood =  -619.57209

Poisson regression                               Number of obs   =       1677
                                                LR chi2(1)      =       16.76
                                                Prob > chi2     =       0.0000
Log likelihood = -619.57209                    Pseudo R2      =       0.0133
```

chd	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
_Igrade_2	1.990735	.3255086	4.21	0.000	1.44488	2.742805
_cons	.0044248	.0004664	-51.42	0.000	.0035989	.0054403
ln(y)	1	(exposure)				

خروجی بالا نشان می دهد که خطر مرگ ناشی از بیماری ایسکمیک قلب در گرید دو نسبت به گرید یک حدود ۲ برابر یا (1.991) بار بیشتر است. نتیجه مذکور با نتیجه دستور `strate grade` یکسان است.

برای تطبیق تاثیر `grade` و مرگ ناشی از ایسکمیک قلبی بر روی متغیر سن و برازش مدل پواسون و کنترل سن کنونی افراد (`current age`)، مجدداً مانند تحلیل قبلی باید متغیر سن به سطوح و طبقات مختلف ساخته شود. و این پیش فرض که می گوید `rate` در سطوح مختلف سن، ثابت است را بررسی نمود. یا اینکه در معرفی متغیرهای زمان و تنظیمات مربوطه، که در ابتدای مبحث توضیح داده شد، `origin` تاریخ تولد افراد (`timebth`) باشد. با این روش نیز اثر سن به عنوان متغیر مخدوش کننده کنترل می شود. اما وقتی متغیر سن گروه بندی یا `split` می شود زمان ها نیز متفاوت می شود. و لذا باید مجدد متغیر زمان ساخته و تحلیل انجام شود. این مراحل به صورت زیر انجام می شود.

```
. stsplot currage, at(40,50(5) 80,90)
(5243 observations (episodes) created)

. replace y=( _t- _t0)
(6893 real changes made)

. xi: poisson _d i.grade i.currage, e(y) irr
i.grade          _Igrade_1-2      (naturally coded; _Igrade_1 omitted)
i.currage        _Icurrage_40-80  (naturally coded; _Icurrage_40 omitted)

Iteration 0:    log likelihood = -832.49262
Iteration 1:    log likelihood = -831.85702
Iteration 2:    log likelihood = -831.8421
Iteration 3:    log likelihood = -831.8421

Poisson regression              Number of obs   =       6920
                                LR chi2(8)        =       97.51
                                Prob > chi2         =       0.0000
                                Pseudo R2          =       0.0554

Log likelihood = -831.8421
```

_d	IRR	Std. Err.	z	P> z	[95% Conf. Interval]
_Igrade_2	1.412587	.237513	2.05	0.040	1.016 1.963978
_Icurrage_50	6.31751	6.659313	1.75	0.080	.8003703 49.86558
_Icurrage_55	8.681497	8.933833	2.10	0.036	1.155186 65.24349
_Icurrage_60	18.47156	18.71842	2.88	0.004	2.5347 134.6111
_Icurrage_65	27.23922	27.58815	3.26	0.001	3.741878 198.2895
_Icurrage_70	33.67959	34.30024	3.45	0.001	4.575927 247.8874
_Icurrage_75	45.16834	46.74968	3.68	0.000	5.940625 343.4283
_Icurrage_80	88.93968	99.73993	4.00	0.000	9.874831 801.0534
_cons	.0003001	.0003002	-8.11	0.000	.0000422 .0021326
ln(y)	1	(exposure)			

نسبت بروز تطبیق شده برای گرید دو به یک 1.41 است. این عدد خیلی نزدیک به تحلیل مانتل هنزل با دستور `stmh` است که در ابتدای مبحث انجام و 1.39 بدست آمد. این مدل نیز با فرض عدم وجود اینترکشن گرید و سن انجام شده است و بر خلاف نتایج دستور `stmh`، در اینجا آزمون اینترکشن انجام نمی شود. در خروجی بالا ستون دوم نسبت های میزان بروز تطبیق شده گزارش شده است. این نسبت ها برای گروه های سنی مختلف نسبت به گروه سنی ۴۰ سال محاسبه شده اند.

### آزمون فرض ها با رگرسیون پواسون:

آماره Z یا آزمون والد که قبلا درباره آن توضیح داده شد نسبت های معنی دار در مدل بالا را نشان می دهد. فرض مورد آزمون یا فرض صفر در اینجا این است که نسبت میزان برابر ۱ (RR=1) است. یا `log rate ratio` برابر صفر (log RR=0) است. در خروجی زیر `log rate ratio` برای متغیر `grade` برابر با 0.3454228 گزارش شده است.

_d	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Igrade_2	.3454228	.1681405	2.05	0.040	.0158736	.674972

همه نسبت های بروز در گروه های سنی معنی دار هستند. حدود اطمینان نسبت های بروز نیز گزارش گردید. برای ارزیابی مدل و معنی دار بودن متغیر در مدل سه روش وجود دارد. روش آزمون والد، روش آزمون اسکور و `likelihood ratio`. توضیحات این سه روش در قسمت های گذشته داده شد.

برای استفاده از روش سوم در اینجا به صورت زیر عمل می شود. استفاده از دستور `quietly` در ابتدای دستور یعنی اینکه نرم افزار محاسبات را در حافظه انجام دهد و خروجی ندهد.

```
. quietly xi: poisson _d i.grade i.currage, e(y) irr
. estimates store fullmodel
. quietly xi: poisson _d i.currage, e(y) irr
. estimates store reducemodel
. lrtest reducemodel fullmodel, stats
```

```
Likelihood-ratio test
(Assumption: reducemodel nested in fullmodel)
LR chi2(1) = 4.13
Prob > chi2 = 0.0421
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
<u>reducemodel</u>	6920	-880.5963	-833.9068	8	1683.814	1738.551
<u>fullmodel</u>	6920	-880.5963	-831.8421	9	1681.684	1743.264

Note: N=Obs used in calculating BIC; see [R] BIC note

نتیجه شبیه آزمون والد است. نتیجه نشان می دهد که اثر grade کنترل شده برای متغیر currage نیز معنی دار است. یکبار هم تحلیل بر مبنای اثر currage و کنترل grade انجام می شود. در این صورت نیز مدل معنی دار است. نتیجه نشان می دهد که اثر currage کنترل شده برای متغیر grade شدیداً معنی دار است. فرض صفر در اینجا این است که نسبت میزان بروز برای همه هفت گروه سنی (در مقیاس log rate ratio) صفر می باشد.

```
. quietly xi: poisson _d i.grade i.currage, e(y) irr
. estimates store fullmodel
. quietly xi: poisson _d i.grade , e(y) irr
. estimates store reducemodel
. lrtest reducemodel fullmodel, stats
```

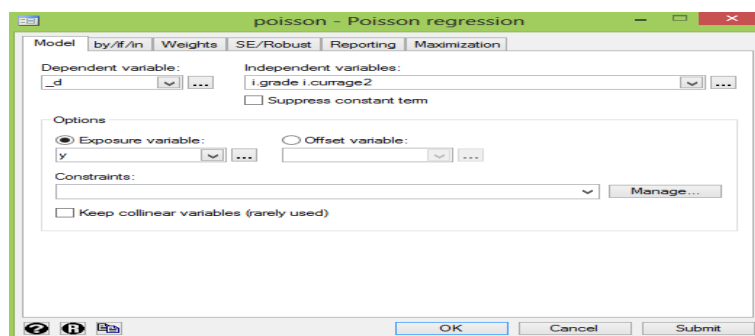
Likelihood-ratio test LR chi2(7) = 80.75  
 (Assumption: reducemodel nested in fullmodel) Prob > chi2 = 0.0000

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
<u>reducemodel</u>	6920	-880.5963	-872.2156	2	1748.431	1762.116
<u>fullmodel</u>	6920	-880.5963	-831.8421	9	1681.684	1743.264

Note: N=Obs used in calculating BIC; see [R] BIC note

### ارزیابی فرض عدم وجود اینترکشن در مدل رگرسیون پواسون:

همه تحلیل های بالا بر مبنای فرض عدم وجود اینترکشن انجام گردید. نرم افزار برای آزمون این فرض خروجی نمی دهد. مجدد یادآوری می شود در ابتدای مبحث، وقتی آزمون مانند مانتل هنزل انجام گردید و اثر سن در رابطه گرید و مرگ های ایسکمیک قلبی کنترل شد، نتیجه نشان داد که اینترکشن وجود دارد و آنجا برای ادامه تحلیل ذکر گردید که فعلاً اینترکشن را در نظر نگیرید. اما در تحلیل رگرسیون پواسون برای ارزیابی اینترکشن، همانند رگرسیون لجستیک عمل می شود. یکبار اینترکشن دو متغیر و یکبار نیز بدون اینترکشن متغیرها وارد می شوند. سپس دو مدل مقایسه می شوند. برای انجام این کار به صورت زیر عمل می شود.





برای معرفی ترم اینترکشن همانند رگرسیون لجستیک، گزینه کنار متغیر های مستقل  
محاوره ای بالا انتخاب می شود و به صورت زیر اینترکشن متغیرها تعریف و add می شود.

Add factor variable

Term:  
Interaction (2-way) Add to varlist

Variables Base c. Time operators

Variable 1: grade First  0 Lag

Variable 2: currage2 First  0 Lag

نتیجه تحلیل بالا با دستورات در زیر آمده است. نتیجه نشان می دهد که با  $p=0.0640$  اینترکشن بسیار ضعیفی  
بین grade و age وجود دارد.

### تحلیل داده های مطالعه کوهورت با رگرسیون پواسون:

فایل داده های cohort1 را باز نمایید. در این فایل متغیر شخص - سال پی گیری شده با متغیر (persontime)  
در یک مطالعه همگروهی در فایل مذکور در دسترس است. متغیر رخداد واقعه با کد صفر و یک، وقوع پیامد  
مرگ ناشی از سکته قلبی (deathMI) را نشان می دهد. متغیرهای مستقل نیز در این داده ها مانند فشارخون و  
سیگار و کلسترول وجود دارند. برای محاسبه نسبت میزان های بروز و تحلیل پواسون از دستور زیر استفاده می شود.

```
. poisson _d i.grade#i.currage2, exposure(y) irr

Iteration 0: log likelihood = -829.8583
Iteration 1: log likelihood = -829.68092
Iteration 2: log likelihood = -829.68055
Iteration 3: log likelihood = -829.68055

Poisson regression                               Number of obs =      6920
LR chi2(11) =                                     =      101.83
Prob > chi2 =                                     =       0.0000
Pseudo R2 =                                       =       0.0578

Log likelihood = -829.68055
```

_d	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
grade#						
currage2						
1 55	2.667453	1.333726	1.96	0.050	1.001143	7.107184
1 60	4.921002	2.277981	3.44	0.001	1.986204	12.19223
1 65	7.623793	3.529131	4.39	0.000	3.077098	18.88865
1 70	14.96203	6.964446	5.81	0.000	6.008662	37.25661
1 75	21.81954	11.26756	5.97	0.000	7.930252	60.03495
2 50	3.377171	2.179954	1.89	0.059	.9530261	11.96744
2 55	4.158737	2.51824	2.35	0.019	1.269209	13.62666
2 60	10.70629	5.083967	4.99	0.000	4.221229	27.15431
2 65	14.06353	6.546216	5.68	0.000	5.647829	35.01927
2 70	9.157862	4.826617	4.20	0.000	3.259654	25.72863
2 75	17.29743	9.116545	5.41	0.000	6.156855	48.59641
_cons	.0009496	.0003877	-17.05	0.000	.0004266	.0021138
ln(y)	1	(exposure)				

```
. estimates store fullmodel
. quietly xi: poisson _d i.grade i.currage2, exposure(y) irr
. estimates store reducemodel
. lrtest reducemodel fullmodel, stats

Likelihood-ratio test                               LR chi2(5) =      10.43
(Assumption: reducemodel nested in fullmodel)       Prob > chi2 =     0.0640
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
reducemodel	6920	-880.5963	-834.8937	7	1683.787	1731.683
fullmodel	6920	-880.5963	-829.6806	12	1683.361	1765.467

دو مدل تفاوت معنی داری را نشان نمی دهند.  $P = 0/06$ . نسبت های میزان بروز در سنین مختلف در جدول بالا گزارش شده است. اینترکشن بین گرید و سن به نظر معنی دار نیست. بنابراین از مدل کاهش یافته استفاده می شود. تحلیل را می توان به شکل قبل و با این تفاوت که متغیر شخص زمان در قسمت Exposure variable قرار گیرد نیز انجام داد. شکل زیر این نوع داده ها را نشان می دهد.

	deathMI	sbp	chol	smok	PersonTime
1	.	120	273	4	2.517455
2	.	120	273	4	5
3	.	120	273	4	5
4	.	120	273	4	5
5	0	120	273	4	1.86377

poisson - Poisson regression

Model by/if/in Weights SE/Robust Reporting Maximization

Dependent variable: deathMI

Independent variables: sbp chol smok

Suppress constant term

Options

Exposure variable: PersonTime

Offset variable:

Constraints: Manage...

Keep collinear variables (rarely used)

OK Cancel Submit

خروجی زیر نتایج این تحلیل را نشان می دهد.



```
. poisson deathMI sbp chol smok, exposure(PersonTime) irr
```

```
Iteration 0: log likelihood = -493.07339
Iteration 1: log likelihood = -493.07261
Iteration 2: log likelihood = -493.07261
```

```
Poisson regression          Number of obs   =      1417
                             LR chi2(3)             =      46.10
                             Prob > chi2            =      0.0000
                             Pseudo R2              =      0.0447
Log likelihood = -493.07261
```

deathMI	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
sbp	1.01493	.0035728	4.21	0.000	1.007952	1.021957
chol	1.005666	.0015906	3.57	0.000	1.002553	1.008788
smok	1.332668	.0901574	4.25	0.000	1.167177	1.521624
_cons	.0007178	.0004607	-11.28	0.000	.000204	.0025255
ln(Person~e)	1	(exposure)				

در خروجی بالا نسبت میزان بروز یا خطر سکتة قلبی در کسانی که سیگار مصرف می کردند نسبت به کسانی که سیگار مصرف ننمودند 1.33 بار بیشتر است. این میزان خطر معنی دار می باشد. در این جا اثر چند متغیر مستقل به عنوان عوامل خطر سکتة قلبی نیز وارد مدل شده و به عبارتی کنترل گردیدند. و اثر عوامل خطر به صورت تطبیق شده گزارش می شود.

### تحلیل بقا: (جدول عمر، روش کاپلان مایر و رگرسیون کاکس)

در رگرسیون پوآسون، با فرض یکسان بودن rate در دوره های زمانی مختلف، یا تقسیم دوره زمانی مطالعه به دوره های کوتاه تر و فرض ثابت بودن میزان، تحلیل قابل انجام بود. در این نوع تحلیل اثر زمان قابل اندازه گیری بوده، اما اندازه گیری زمان بقای افراد در این روش مقدور نبود. با توجه به تغییر قابل توجه میزان ها در طی زمان، مثلا سن افراد، و اینکه در مواقعی که نتوان فرض نمود میزان در طی مدت زمان های مختلف یکسان است یا اینکه هدف اندازه گیری زمان بقای افراد از مدت پی گیری است یا اندازه گیری زمان تا رخداد واقعه (time to Event) باشد، رگرسیون پوآسون مناسب نیست و لذا تحلیل رگرسیون کاکس مورد نیاز می باشد. این تحلیل نیز رویکردی دیگر برای مطالعات کوهورت است. جمع آوری داده ها و در نظر گرفتن افراد حذف شده یا سانسور شده نیز در این مطالعات رخ می دهند که نیازمند توجه هستند.

### برای انجام تحلیل بقا گام های زیر نیاز است:

- ۱- سوال و فرضیه پژوهش چیست؟
- ۲- محل پژوهش و متغیرهای مورد نیاز مثلا متغیرهای پیش بین یا مستقل و پیامد چیست؟
- ۳- نمونه مورد نیاز پژوهش چیست و چگونه باید انتخاب و پی گیری شود؟

۴- متغیرها چگونه اندازه گیری و ثبت گردند؟ با داده های سانسور شده چگونه برخورد می شود؟ با خطر های رقیب چگونه برخورد می شود؟

۵- تعیین و تعریف متغیری که زمان بقای افراد یا مشاهدات را اندازه گیری می نماید. زمان به عنوان یک متغیر مهم می بایست در جمع آوری متغیرها مد نظر قرار گیرد. مثلاً تاریخ تولد دقیق افراد یا تاریخ دقیق ورود و خروج افراد در مطالعه و وقوع پیامد.

۶- تحلیل داده ها و استفاده از روش کاپلان مایر یا جدول عمر و یا رگرسیون کاکس و ارزیابی پیش فرضها و مدل

با توجه به نکات بالا به نظر می رسد زمان بقا یک متغیر کلیدی در این نوع تحلیل است. زمان بقا یک متغیر تصادفی غیرمنفی است که فاصله زمانی بین شروع و وقوع یک پدیده خاصی را اندازه گیری می کند. پس متغیر اصلی، زمان بقا (Survival Time) است. این متغیر در واحد زمان از صفر تا بی نهایت را می تواند به خود بگیرد. این رخداد در رایج ترین حالت مطالعات آنالیز بقا، همان مرگ است اما به واقع میتواند هر رخداد قابل مشاهده یا اندازه گیری دیگر مثل ابتلا به یک بیماری مشخص، پاسخ به یک درمان مشخص، عود بیماری و... باشد. متغیر مهم دوم در آنالیز بقا، وقوع یا عدم وقوع آن رخداد مشخص است که در ساده ترین و رایج ترین حالت به شکل یک متغیر دو وضعیتی (وقوع یا عدم وقوع) است. در مثالهای این فصل رخداد، همان شکل رایج آن یعنی مرگ در نظر گرفته شده و وقوع رخداد را شکست (Failure) و عدم وقوع رخداد آن موفقیت (Success) نامیده می شود. این متغیر معمولاً بجای توزیع نرمال، توزیع چولیده دارد. سابقه آنالیز بقا شاید به زمانی در قرن ۱۷ میلادی برسد که شخصی به نام Graunt اولین فهرست هفتگی مرگ و میر را در لندن منتشر نمود و به دنبال آن Healey اولین جدول عمر را منتشر کرد. بدیهی است که از ابتدا این مفهوم جهت بررسی بقای انسان یا به عبارتی سنجش طول عمر انسان بکار می رفت. اما در طی جنگ جهانی دوم محققان از روشهای آنالیز بقا برای بررسی طول عمر و پایایی ابزارآلات جنگی و صنعتی نیز استفاده کردند. با گسترش این مفهوم در حال حاضر آنالیز بقا در زمینه های مختلف پزشکی، اپیدمیولوژی، بهداشت محیط، تجارت، نجوم، صنعت و... بکار برده میشود. مثال های زیر کاربرد این نوع تحلیل را نشان می دهند.

- بررسی زمان بوجود آمدن رتینوپاتی دیابتی از زمان تشخیص دیابت در بیماران دیابتی نوع ۲
- بررسی زمان پاسخ به درمان در یک پروتکل درمانی مشخص،
- بررسی عود بیماری در بیماران پس از درمان،
- بررسی زمان آزاد شدن زندانیان در یک زندان،
- بررسی مدت زمان دوام اولین ازدواج در زوجهای یک شهرستان،
- بررسی طول عمر وسایل الکترونیک تولید شده توسط یک کارخانه

## جداول عمر (Life Tables)

یک روش آنالیز اطلاعات بسیار معمول و رایج در این نوع مطالعات استفاده از جدول عمر است. فرض کنید که گروهی از مردم (یک همگروه (Cohort)، در یک دوره زمانی مشخص پیگیری و تا وقوع رخداد خاصی مانند مرگ در آنها اندازه گیری شود. اگر در انتهای این دوره زمانی مجموع تعداد رخداد، هدف اندازه گیری باشد به سادگی محاسبه میشود. اما ممکن است علاوه بر آن، هدف بررسی الگوی وقوع رخداد در طول مدت زمان پیگیری نیز باشد. به عبارت دیگر ممکن است اندازه گیری طول زمان قبل از وقوع رخداد مد نظر باشد. تا اینکه هدف فقط رخ دادن واقعه باشد یا خیر. به عنوان مثال، اگر بر روی بیماری سکت قلبی مطالعه کنید و تعدادی بیمار مبتلا را برای مدت پنج سال پیگیری نمایید، طبعاً تنها صرف بررسی وقوع مرگ در این بیماران، کافی نیست و مهمتر از آن این است که بدانید هر کدام از بیماران بعد از چه مدت فوت کرده‌اند. در چنین شرایطی، داده‌ها می‌تواند با استفاده از جداول عمر نمایش داده شده و آنالیز شود.

احتمال شکست در یک جمعیت  $N(f)$  نفره که همگی به مدت مشخصی پیگیری شده و در طول این مدت  $D$  نفر مواجه با شکست میشوند به سادگی از طریق فرمول زیر قابل محاسبه است:

$$\frac{D}{N(f)}$$

پایان یافتن زود هنگام پیگیری به هر دلیل، سانسور شدن نامیده میشود. سانسور از مفاهیم بسیار مهم در آنالیز بقا است و انواع (چپ، راست و ...) و علل مختلفی دارد که رایج ترین آنها در جدول زیر آمده است.

علت سانسور	مثال
عدم امکان پیگیری	مهاجرت، عدم تمایل به ادامه مطالعه، عدم مراجعه جهت پیگیری
رسیدن به رخداد جایگزین	مرگ بر اثر سکت قلبی در مطالعه آنالیز بقای سرطان معده
خروج افراد از مطالعه	بروز عوارض جانبی درمان، ملاحظات اخلاقی ادامه مطالعه
زمانهای متفاوت ورود به مطالعه	افراد در زمانهای متفاوتی به مطالعه وارد میشوند

چگونگی برخورد با موارد سانسور شده در این تحلیل مهم است. آسان ترین روش این است که بطور کلی این موارد از مطالعه حذف شوند. اما هنگام محاسبه احتمال شکست، منطقی نیست که سانسور نادیده گرفته شود. زیرا این موارد نیز به هر حال شانس برای شکست (هر چند کمتر) داشته‌اند. این مهم که آنها در طول حضور در مطالعه مواجه با شکست نشده‌اند، اطلاعاتی هر چند ناکافی درباره احتمال شکست در آنها به ما می‌دهد که می‌تواند مورد استفاده قرار بگیرد. مشکل این است که ما نمی‌دانیم چه نسبتی از موارد سانسور شده در صورت پیگیری کامل مواجه با شکست می‌شده‌اند. بنابراین، ما نمی‌توانیم هیچ اصلاح و تعدیلی روی تعداد موارد مرگ مشاهده شده انجام دهیم، اما میتوان مخرج کسر را اصلاح کرد. به این شکل که اگر دوره پیگیری در نظر گرفته شده طولانی باشد، میتوان بصورت ساده فرض کرد که بطور متوسط هر بیمار سانسور شده به مدت نصف دوره

مورد نظر پیگیری شده، بدون اینکه با شکست مواجه شده باشد. پس در حالت کلی، اگر یک کوهورت N نفره را در نظر بگیرید که باید به مدت مشخصی پیگیری می شده‌اند و در این مدت در D نفر شکست رخ داده و L نفر سانسور شده‌اند، در این صورت احتمال شکست در مدت پیگیری مورد نظر بصورت زیر محاسبه خواهد شد:

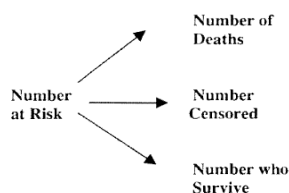
$$\frac{D}{N - 0.5L}$$

در واقع اینطور میتوان گفت که اثر سانسور این بود که تعداد جمعیت موثر کوهورت که به عنوان اندازه موثر کوهورت (Effective Cohort Size) نامیده میشود را به مقدار  $L/5$  کاهش داد. از این روش نسبتاً خام برای لحاظ کردن سانسور، زمانی میتوان استفاده کرد که L (یعنی تعداد موارد سانسور شده) در مقایسه با N (یعنی کل جمعیت کوهورت) کوچک باشد. استفاده از یک دوره زمانی ثابت به عنوان دوره پیگیری با سه اشکال اساسی همراه است.

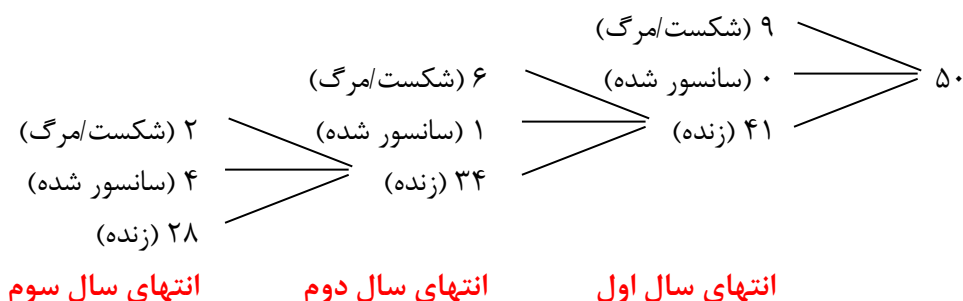
اول اینکه تنها زمانی میتوان دو کوهورت را با هم مقایسه کرد که هر دو، مدت زمان پیگیری یکسان داشته باشند. مثلاً مرگ و میر تجمعی ۵ ساله نمیتواند با مرگ و میر تجمعی ۱۰ ساله مقایسه شود. دوم اینکه اطلاعات مربوط به زمان مواجهه فرد با شکست نادیده گرفته می شود. در واقع ممکن است تغییرات مهمی در احتمال شکست در مقاطع مختلف در طول مدت پیگیری بوجود آمده که دانستن آن میتواند مهم باشد، بخصوص در مطالعات پیگیری بالینی.

سوم اینکه روش گفته شده برای لحاظ کردن موارد سانسور شده نسبتاً خام است چون زمان وقوع سانسور را نادیده می گیرد. راه چاره برای هر سه شکل این است که به جای استفاده از یک فاصله زمانی پیگیری طولانی، از چند فاصله زمانی کوتاهتر متوالی استفاده شود. سپس با استفاده از محاسبات آماری، از نتایج بررسی کوهورت در طول هر یک از این فواصل کوتاه میتوان برای بررسی کوهورت در کل طول زمان پیگیری استفاده کرد. این روش به عنوان **روش جدول عمر** نامیده میشود.

در این روش به جای محاسبه منفرد احتمال شکست برای یک فاصله زمانی طولانی، یک سری از محاسبات خواهیم داشت که هر کدام مربوط به یک فاصله کوتاهتر خواهد بود. این حالت میتواند در قالب یک نمودار درختی تصور شود که در آن شاخه بالای درخت معرف شکست (یا مرگ)، شاخه پایین معرف موفقیت (یا زنده بودن) و شاخه وسطی درخت معرف مشاهدات سانسور شده است.



در مثال زیر فاصله زمانی پیگیری ۵ ساله، به فواصل یکساله تقسیم شده است. به نمودار زیر که شامل سه سال اول می باشد، توجه کنید. در این نمودار تعداد بیماران فرضی نمایش داده شده است.



موارد موفقیت در این مثال افراد زنده در انتهای یکسال، جمعیت شروع شاخه‌های درخت برای سال بعد هستند. در مثال فوق، در ابتدای سال اول ۵۰ نفر مطالعه را شروع کردند. ۹ نفر مردند و هیچکس سانسور نشد. در نهایت ۴۱ بیمار باقی ماندند تا سال دوم را شروع کنند و همین طور الی آخر.

احتمال شکست برای هر دوره زمانی محاسبه می شود. احتمال شکست در هر فاصله زمانی، با لحاظ کردن موارد سانسور شده در همان فاصله زمانی محاسبه می شود. احتمال موفقیت در هر فاصله زمانی هم شاخص مهمی است که به سادگی برابر است با: احتمال شکست در آن فاصله زمانی - ۱ = احتمال موفقیت

مطابق با قانون احتمالات، احتمال وقوع توأم دو پیامد مستقل برابر است با حاصلضرب احتمال هر یک از آنها. پس احتمال بقا در دو سال پیاپی، برابر خواهد بود با حاصلضرب احتمال بقا در سال اول و احتمال بقا در سال دوم. برای محاسبه احتمال هر یک از این پیامدها تنها لازم است از ابتدای درخت به سمت پایین حرکت کرده و مقادیر شاخه‌ها را تا رسیدن به شاخه مربوط به پیامد مورد نظر در هم ضرب کنیم.

به این ترتیب احتمال هر یک از پیامدها در مثال بالا به شرح زیر است:

$$\text{احتمال شکست در سال اول} = 0/18$$

$$\text{احتمال شکست در سال دوم} = 0/148 \times 0/82 = 0/121$$

$$\text{احتمال شکست در سال سوم} = 0/063 \times 0/852 \times 0/82 = 0/044$$

$$\text{احتمال عدم شکست (احتمال موفقیت) تا انتهای سال سوم} = 0/937 \times 0/852 \times 0/82 = 0/655$$

این احتمالات به ترتیب ۰/۱۸، ۰/۱۲۱، ۰/۰۴۴ و ۰/۶۵۵ محاسبه شدند. بدیهی است که مجموع آنها برابر ۱ میشود، چون مجموع آنها تمامی پیامدهای ممکن را تشکیل میدهند. احتمال وقوع شکست در طول پیگیری برابر است با  $0/345 = 0/044 + 0/121 + 0/18$ . روش محاسبه ساده‌تر این احتمال این است که احتمال زنده بودن (عدم شکست) در انتهای سال سوم را از عدد ۱ کم شود:  $0/655 = 1 - 0/345$ .

در کوهورت‌هایی که افراد در آن معاینه سالیانه میشوند، داده‌ها اغلب به شکل یک جدول عمر (جدول عمر کوهورت (Cohort Life Table)) ارائه میشود که در آن تعداد موارد شکست و سانسور که در هر سال رخ داده است نمایش داده میشود. بنابراین جدول عمر، تنها یک شیوه ساده برای نمایش احتمالات شکست و بقا است. این روش برای مطالعه اینکه آیا احتمال شکست در طول زمان پیگیری تغییر میکند و برای محاسبه مقادیری مانند احتمال بقای پنج ساله مفید است. نمودار یا نقشه‌ای که احتمال تجمعی بقا را نمایش میدهد، منحنی بقا (Survival Curve) نامیده میشود. بهتر است که منحنی بقا از مقدار احتمال یک شروع شود که بیانگر احتمال زنده ماندن برای حداقل صفر سال است. این نمودار در قسمت کار با نرم افزار نمایش داده می شود.

در محاسبات جدول عمر، موارد شکست را بطور کلی در فرمول  $D/(N - 0.5L)$  قرار میدهند و بدین ترتیب احتمال وقوع شکست در طی دوره مورد بررسی را محاسبه میکنند. اگر زمان وقوع شکست در هر فرد معلوم باشد، به نحو بهتری میتوان عمل کرد. بدین ترتیب که دوره‌های زمانی مورد بررسی را به نحوی انتخاب میکنیم که منطبق با وقوع یک شکست باشد. به عبارت دیگر لازم نیست دوره پیگیری را به فواصل ثابت، مثلا یکساله، تقسیم کنیم بلکه احتمال بقا بعد از وقوع هر شکست محاسبه خواهد شد. این روش محاسبه احتمال تجمعی بقا کاپلان-مایر (Kaplan-Meier) نامیده میشود. در اینجا نیز رسم نمودار احتمال تجمعی بقا مفید است اما اینجا نمودار به جای یک منحنی به شکل یک نمودار پله‌ای خواهد شد چون بقای تجمعی دقیقا در زمان وقوع شکست پایین می‌افتد. تفسیر و تحلیل نمودارهای بقا که با استفاده از روش کاپلان-مایر رسم میشود تقریبا به روش مشابه با نمودارهایی است که با روش جدول عمر رسم میشود. منحنی‌های بقا توصیف خوبی از بررسی بقا در یک کوهورت ارائه می‌دهند. در بسیاری از مواقع، کوهورت ممکن است به چندین زیرگروه تقسیم شده و مقایسه بقا در این زیرگروه‌ها مورد توجه باشد. در این شرایط می‌توان احتمال بقای تجمعی را برای هر زیرگروه بطور جداگانه محاسبه کرد و دو منحنی را برای مقایسه روی یک نمودار نمایش داد. این مقایسه چشمی منحنی‌های بقا بسیار مفید است اما روشهای آماری دقیق‌تری نیز برای مقایسه وجود دارد مانند:

- log rank test
- Wilcoxon
- Tarone-Ware
- Peto
- Flemington-Harrington

### پیش‌فرضهای استفاده از روشهای آنالیز بقا

**مطابق پیش‌فرض اول** که مختص جدول عمر است، باید وقوع رخدادها در طول هر فاصله زمانی مورد بررسی تقریبا یکسان باشد. به بیان دیگر، اگر احتمال وقوع رخداد در طول یک فاصله زمانی به سرعت تغییر کند و شما از روش جدول عمر برای آنالیز بقا استفاده کنید، در واقع شما مقدار متوسط احتمال وقوع رخداد را برای تمام

طول آن فاصله زمانی بصورت ثابت در نظر گرفته‌اید که این مطابق واقع نیست و مطالعه را دچار تورش می‌کند. راه حل منطقی برای رفع این خطا دقت در انتخاب فواصل زمانی است. به بیانی دیگر هنگامی که در روش جدول عمر، طول مدت پیگیری را به فواصل کوتاهتر تقسیم می‌کنیم، باید توجه کنیم که فواصل به گونه‌ای انتخاب شوند که وقوع رخدادها در طول یک فاصله زمانی تقریباً مشابه باشد. البته بدیهی است که در روش کاپلان مایر این خطا پیش نخواهد آمد چون در این روش، عملاً طول فواصل زمانی مورد بررسی بر اساس وقوع رخداد تعیین میشود و میزان بقا پس از وقوع هر رخداد محاسبه خواهد شد؛ لذا چنین خطایی، بی‌معنی خواهد بود. جدای از این پیش‌فرض که مختص روش جدول عمر بود، دو پیش‌فرض زیر برای هر دو روش لازم است.

**پیش‌فرض دوم** بیان میکند که احتمال وقوع رخداد در موارد سانسور شده باید مشابه احتمال وقوع رخداد در مواردی باشد که در مطالعه باقی‌مانده‌اند. به بیان دیگر لازم است سانسور مستقل از بقا باشد. اجازه بدهید با یک مثال منظور را روشن‌تر کنیم. فرض کنید یک مطالعه بقا در حال انجام است که در آن رخداد مورد نظر، مرگ به علت سکنه قلبی است. در طول مدت پیگیری اگر متوجه شویم که افراد سانسور شده بیشتر افراد مسن و یا افراد سیگاری هستند، چه خطایی به نظرتان محتمل است؟ یک فرض منطقی به ما میگوید که افراد سانسور شده در این مطالعه که اکثراً سیگاری و پیر بوده‌اند بیشتر از بقیه افراد کوهورت، در خطر سکنه قلبی بوده‌اند. پس ما در حقیقت با حذف این افراد از پیگیری بیشتر، احتمال وقوع رخداد را کم کرده‌ایم و در بقای محاسبه شده دچار کم‌شماری (Underestimation) خواهیم شد. البته این خطا میتواند در شرایطی نتایج را به سمت بیش‌شماری (Overestimation) هم هدایت کند که از توضیح آن در اینجا میگذریم. مطابق **پیش‌فرض سوم** که آن را عدم وجود Secular trend می‌نامند، در طول مدت پیگیری نباید تغییرات واضحی در خصوصیات کوهورت که مرتبط با رخداد مورد مطالعه است و یا مواجهه‌های مرتبط پیش بیاید. باز هم با یک مثال موضوع روشن‌تر میشود. فرض کنید مطالعه‌ای به بررسی بقا در بیماران دیابتی که بیماری آنها بین سالهای ۱۹۱۵ تا ۱۹۳۵ تشخیص داده شده است، پرداخته است. اگر بدانید که جایی در میانه این بازه زمانی، انسولین کشف شده و مورد استفاده قرار گرفته است، آیا به نظرتان هنوز منطقی است که این بیماران در یک گروه واحد مورد مطالعه قرار بگیرند؟ بدیهی است که بیماران دیابتی پس از کشف انسولین، بطور قابل توجهی طول عمر بیشتری یافته‌اند. در واقع یک تغییر فاحش در مواجهه در این کوهورت که مرتبط با موضوع مورد مطالعه است باعث شده است که اگر محقق این کوهورت را در قالب یک گروه واحد بررسی کند، دچار تورش واضحی شود. روش دیگر آنالیز روش رگرسیون کاکس است. در این روش اثر متغیرهای مخدوش‌کننده مهم و متغیرهای موثر بر زمان بقای افراد را می‌توان کنترل و محاسبه نمود. برای متغیرهایی می‌توان وارد مدل کاکس نمود که فرض **proportional hazard ratio** برای آنها برقرار باشد. عبارتی نسبت خطر در طول زمان ثابت باشد.

در جدول عمر هم گروهی (Cohort life table) مرگ و میر واقعی یک گروه از افراد تولد اولین نفر تا مرگ آخرین نفر ثبت می‌شود. جمع‌آوری داده‌های مورد نیاز برای ساختن چنین جدولی مستلزم تعقیب یک نسل از تولد تا مرگ است و لذا ساختن جدول عمر هم گروهی کاری دشوار است. جدول عمر دوره زمانی (Period life table) بر اساس اطلاعات مرگ به دست آمده در طی یک سال یا یک دوره زمانی مشخص ساخته می‌شود. در چنین جدولی فرض بر آن است که افرادی که در این دوره زمانی به دنیا آمده‌اند در سال‌های آینده از احتمال

مرگی معادل احتمال مرگ همان گروه سنی در این دوره زمانی برخوردارند. یعنی فرض بر آن است که احتمال مرگ کودکی که در این دوره زمانی به دنیا آمده است در سن ۲۵ سالگی با احتمال مرگ فرد ۲۵ ساله در همین دوره زمانی برابر است. در قیاس با جدول عمر هم گروهی دست یابی به اطاعات مورد نیاز برای جدول عمر دوره زمانی با سهولت بیشتری انجام می شود و عملی تر است. لذا در تمامی مواردی که به جدول عمر اشاره می شود، مقصور جدول عمر دوره زمانی است.

تقسیم بندی دیگر جداول عمر بر اساس کامل یا خلاصه بودن آن است. جدول عمر کامل جدولی است که میزان های مرگ در آن برای تمامی سال های عمر محاسبه شده است. در جدول عمر خلاصه شده، که معمول تر است، میزان های مرگ برای گروه های سنی محاسبه می شود. گروه های سنی در این جدول عبارتند از: صفر تا ۱۱ ماه، ۱ تا ۴ سال و پس از آن گروه های سنی ۵ ساله تا گروه سنی بالاتر از ۸۵.

**نمونه جدول عمر در ضمیمه آمده است. ستون های جدول عمر خلاصه شده عبارتند از:**

**ستون اول (X):** سال اول هر گروه سنی

**ستون دوم (n):** فاصله هر گروه سنی

**ستون سوم ( $nP_x$ ):** جمعیت گروه سنی در وسط سال مورد نظر

**ستون چهارم ( $nD_x$ ):** تعداد مرگ رخ داده در گروه سنی در سال مورد نظر

**ستون پنجم ( $nM_x$ ):** میزان مرگ اختصاصی سن در سال مورد نظر که از تقسیم ستون چهارم به ستون سوم به دست می آید.

**ستون ششم (a):** کسری از فاصله گروه سنی که افراد فوت شده آن فاصله، زندگی کرده اند. این کسر برای تمامی گروه های سنی به استثنای گروه اول ۰/۵ در نظر گرفته می شود. این بدان معناست که افرادی که در هر گروه سنی فوت کرده اند به طور متوسط ۲/۵ سال از فاصله ۵ ساله آن گروه سنی را زنده بوده اند. در گروه سنی اول به واسطه آن که احتمال مرگ زیر یک سال در ماه های اول زندگی بیشتر است، مقدار a کمتر از ۰/۵ منظور می شود. این مقدار در جوامع دارای مرگ و میر بالای اطفال برابر با ۰/۳ و در جوامع با مرگ و میر پایین ۰/۱ منظور می شود. برای گروه سنی دوم نیز که احتمال مرگ در سنین پایین بیشتر است، مقدار ۰/۴ برای a در نظر گرفته می شود.

**ستون هفتم ( $nq_x$ ):** احتمال وقوع مرگ در فاصله سن X و X + n است. این احتمال از رابطه زیر محاسبه می شود:

$$nq_x = n \cdot nM_x / [1 + n \cdot (1 - na_x) \cdot nM_x]$$

هدف از به کارگیری این رابطه محاسبه جمعیت اول فاصله گروه سنی به کمک جمعیت وسط سال و کسر a است.



**ستون هشتم ( $nP_x$ ):** احتمال بقا در فاصله سن  $x$  و  $x+n$  است و از رابطه زیر به دست می آید:

$$nP_x = 1 - nq_x$$

**ستون نهم ( $I_x$ ):** تعداد افرادی هستند که در سن دقیق  $x$  زنده هستند.  $L_0$  معمولاً عددی چون ۱، ۱۰۰۰، یا ۱۰۰۰۰۰۰ اختیار می شود و با فرض این که یک جمعیت برای مثال ۱۰۰۰۰۰ نفری در سال های مختلف عمر در معرض احتمال مرگی مشابه آنچه که بر اساس این جدول محاسبه شده است قرار می گیرند؛ تعداد افراد زنده در سن  $x$  محاسبه می شود.  $I_x$  یک عدد فرضی است و هیچ ارتباطی با تعداد واقعی افراد  $x$  ساله زنده در جمعیت واقعی ندارد. با توجه به این تعریف  $I_x$  از رابطه زیر محاسبه می شود:

$$I_x = I_{x-n} \cdot nP_{x-n}$$

بر اساس اطلاعات این ستون منحنی بقا برای جامعه ترسیم می شود. منحنی بقا نشان می دهد که چه بخشی از یک هم گروه فرضی در هر سن زنده هستند.

**ستون دهم ( $nd_x$ ):** تعداد افرادی که در فاصله سن  $x$  تا  $x+n$  فوت می کنند. این مقدار از رابطه زیر محاسبه می شود:

$$nd_x = I_x \cdot nq_x$$

با توجه به آن که تمامی افراد در آخرین گروه سنی فوت می کنند،  $d_{x+n} = I_x$

**ستون یازدهم ( $nL_x$ ):** تعداد کل سال هایی که افراد در فاصله  $x$  تا  $x+n$  سال زندگی کرده اند. افرادی که تمام این فاصله را زنده بوده اند  $(I_{x+n})$  تمام فاصله را زندگی کرده اند  $(n \cdot I_{x+n})$  و افرادی که در این فاصله فوت کرده اند  $(nd_x)$  کسری از این فاصله را زندگی کرده اند  $(n \cdot na_{x+n} \cdot nd_x)$ . بنابراین مقدار این ستون از رابطه زیر به دست می آید:

$$nL_x = n(I_{x+n} + na_{x+n} \cdot nd_x)$$

برای آخرین گروه سنی مقدار این ستون از رابطه زیر به دست می آید:

$$L_{85+} = d_{85+} / M_{85+}$$

مقدار این ستون برای محاسبه شاخص های کلی تندرستی جامعه مورد استفاده قرار می گیرد.

**ستون دوازدهم ( $T_x$ ):** کل سال هایی که افراد بعد از سن  $x$  زندگی کرده اند. برای آخرین گروه سنی

$$T_{85+} = L_{85+}$$

و برای سایر گروه های سنی

$$T_x = T_{x+n} + nL_x$$

**ستون سیزدهم ( $e_x$ ):** امید زندگی در سن  $x$  است و از رابطه زیر به دست می آید:

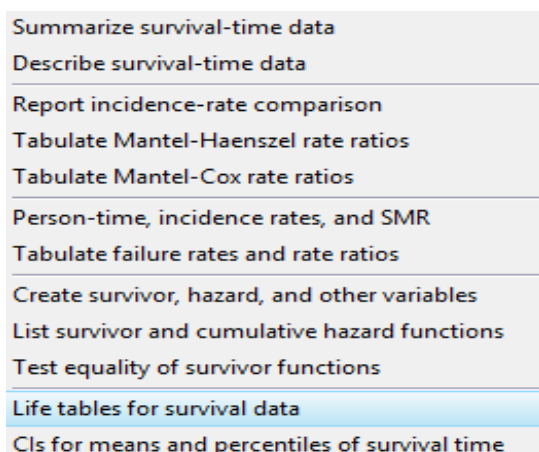
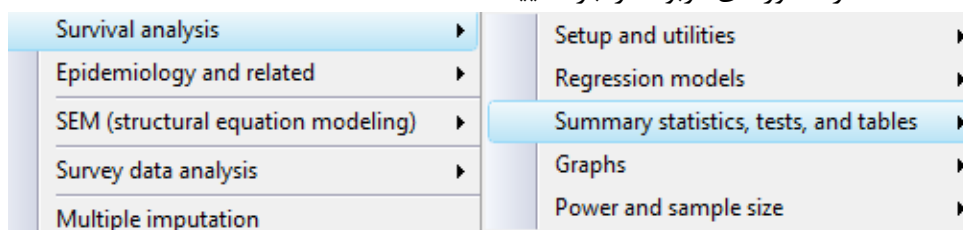
$$e_x = T_x / I_x$$

**کار با نرم افزار و ترسیم نمودار و جدول عمر:**

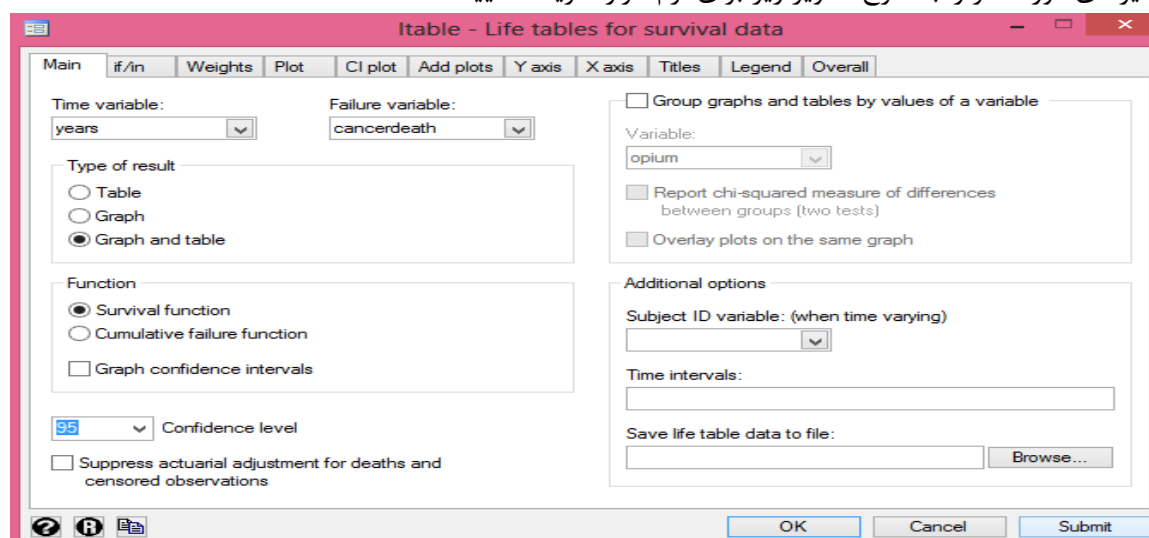
فایل lifeTable را باز نمایید. توضیح و تعریف متغیرها در قسمت label یا برچسب متغیرها داده شده است. این فایل مربوط به یک مطالعه کوهورت فرضی با حجم نمونه ۲۰۰ نفر برای وقوع مرگ ناشی از سرطان مری طراحی گردید. برای تحلیل بقا با چند متغیر آن از جمله متغیر death یا مرگ از همه علل، متغیر years بعنوان متغیر سال های پی گیری افراد برای رخداد واقعه مرگ ناشی از سرطان مری با نام cancerdeath که به عنوان متغیر پیامد مد نظر است و متغیر opium بعنوان متغیر استعمال تریاک و متغیر نوشیدن چای داغ بنام teahot و گروه

قومیت بنام ethgp تحلیل انجام می شود.

اولا جدول عمر را برای این مطالعه و نمودار مربوطه را ترسیم نمایید. ثانيا جدول را برحسب دو گروه از افراد پی گیری شده بر حسب مصرف تریاک و بروز مرگ ناشی از سرطان مری نیز ترسیم نمایید. برای این تحلیل از مسیر زیر اقدام می شود. گزینه Survival analysis را از منوی اصلی statistics به شرح زیر انتخاب یا با دستور db ltable محاوره ای مربوطه را باز نمایید.



متغیرهای مورد نظر را به شرح تصویر زیر برای نرم افزار تعریف نمایید.

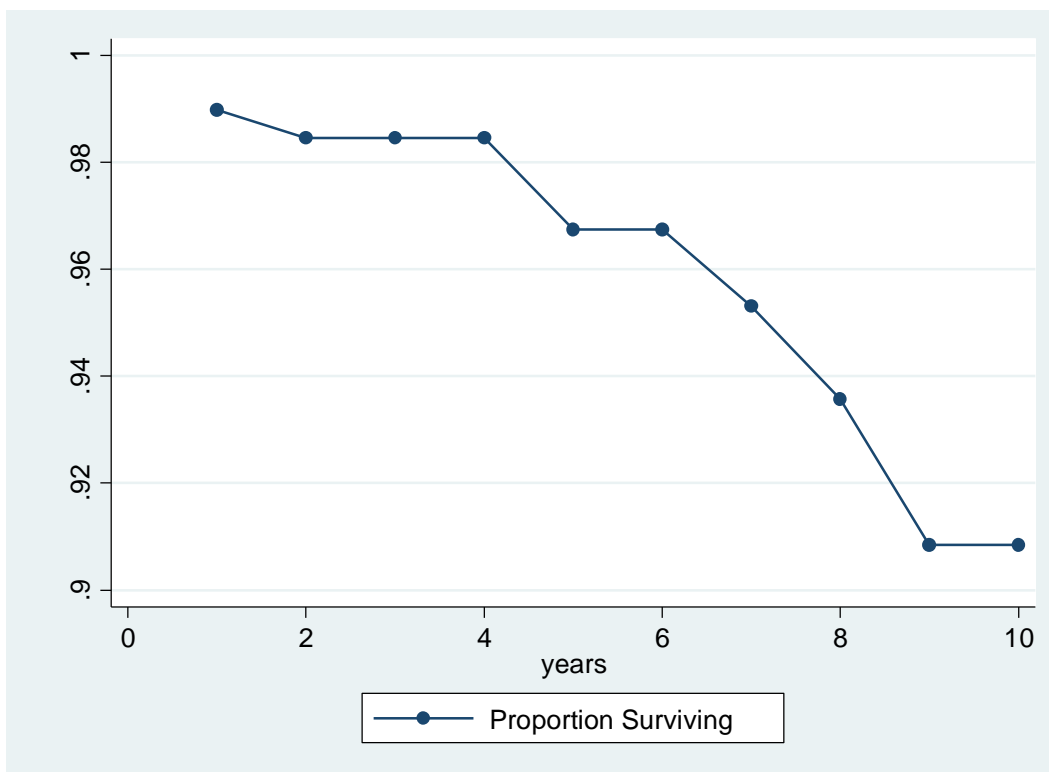


می توان در کادر بالا توابع مختلف بقا را انتخاب و یا اینکه جدول به تنهایی یا همراه با نمودار و رسم نمودار بر حسب متغیر کیفی و ذخیره نمودار در مسیر دلخواه نتایج را تنظیم نمود. خروجی به شرح زیر است.

. ltable years cancerdeath, graph survival legend(on)

Interval	Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf. Int.]
0 1	200	2	6	0.9898	0.0071	0.9600 0.9975
1 2	192	1	5	0.9846	0.0088	0.9531 0.9950
2 3	186	0	3	0.9846	0.0088	0.9531 0.9950
3 4	183	0	8	0.9846	0.0088	0.9531 0.9950
4 5	175	3	6	0.9675	0.0131	0.9289 0.9853
5 6	166	0	21	0.9675	0.0131	0.9289 0.9853
6 7	145	2	20	0.9531	0.0164	0.9078 0.9765
7 8	123	2	26	0.9358	0.0201	0.8822 0.9655
8 9	95	2	53	0.9085	0.0273	0.8376 0.9493
9 10	40	0	40	0.9085	0.0273	0.8376 0.9493

جدول بالا نشان می دهد که در شروع مطالعه ۲۰۰ نفر وارد گردیده و تا پایان مطالعه حدود ۱۲ نفر فوت نموده است. ستون تعداد مرگ ها، تعداد از دست رفته ها یا سانسور شده ها، ستون های احتمال بقا و خطای معیار آن و حدود اطمینان بقا نیز نشان داده شده است. نمودار بقای این افراد در زیر آمده است.



برای رسم نمودار و جدول عمر بر حسب متغیر مصرف تریاک نتایج به شرح زیر است.

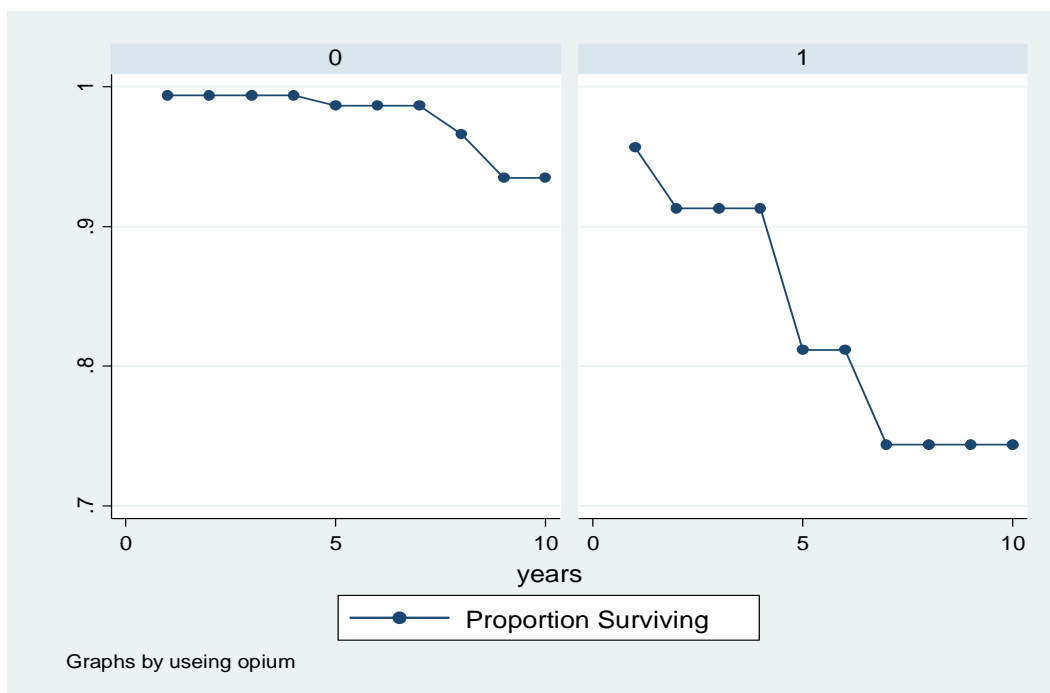
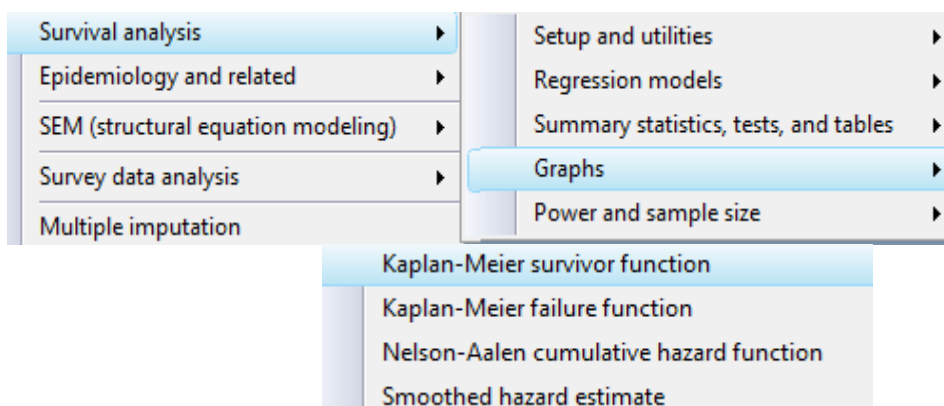


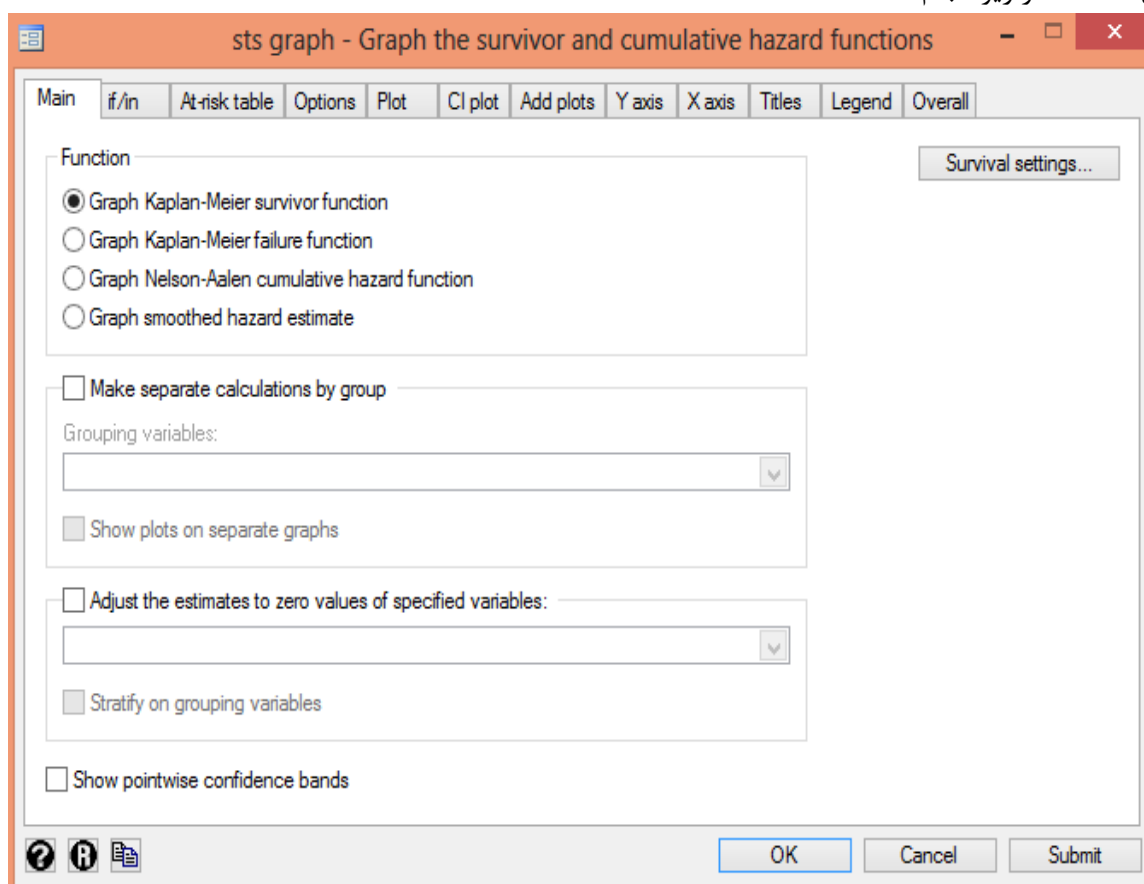
table years cancerdeath, graph survival by(opium) byopts(legend(on))

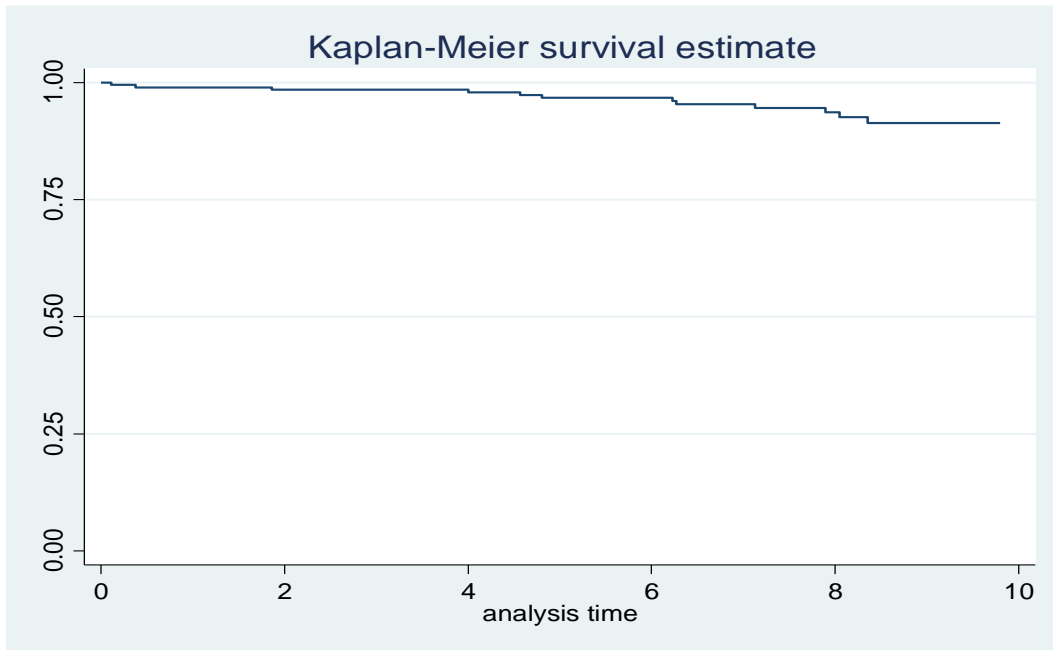
Interval	Beg.	Total	Deaths	Lost	Survival	Std. Error	[95% Conf. Int.]
<b>opium = 0</b>							
0	1	159	1	3	0.9937	0.0063	0.9558 0.9991
1	2	155	0	3	0.9937	0.0063	0.9558 0.9991
2	3	152	0	2	0.9937	0.0063	0.9558 0.9991
3	4	150	0	6	0.9937	0.0063	0.9558 0.9991
4	5	144	1	6	0.9866	0.0094	0.9474 0.9966
5	6	137	0	18	0.9866	0.0094	0.9474 0.9966
6	7	119	0	14	0.9866	0.0094	0.9474 0.9966
7	8	105	2	19	0.9659	0.0171	0.9097 0.9874
8	9	84	2	45	0.9345	0.0274	0.8535 0.9715
9	10	37	0	37	0.9345	0.0274	0.8535 0.9715
<b>opium = 1</b>							
0	1	23	1	0	0.9565	0.0425	0.7293 0.9938
1	2	22	1	0	0.9130	0.0588	0.6949 0.9775
2	3	21	0	1	0.9130	0.0588	0.6949 0.9775
3	4	20	0	2	0.9130	0.0588	0.6949 0.9775
4	5	18	2	0	0.8116	0.0854	0.5707 0.9253
5	6	16	0	2	0.8116	0.0854	0.5707 0.9253
6	7	14	1	4	0.7440	0.1016	0.4813 0.8873
7	8	9	0	4	0.7440	0.1016	0.4813 0.8873
8	9	5	0	3	0.7440	0.1016	0.4813 0.8873
9	10	2	0	2	0.7440	0.1016	0.4813 0.8873

این جدول نیز همانند جدول بالاست. منتها برحسب دو گروه مصرف کننده یا عدم مصرف تریاک تهیه گردید. در روش کاپلان مایر نیز به طریق زیر از منوی **statistics** منوهای زیر انتخاب می شوند.

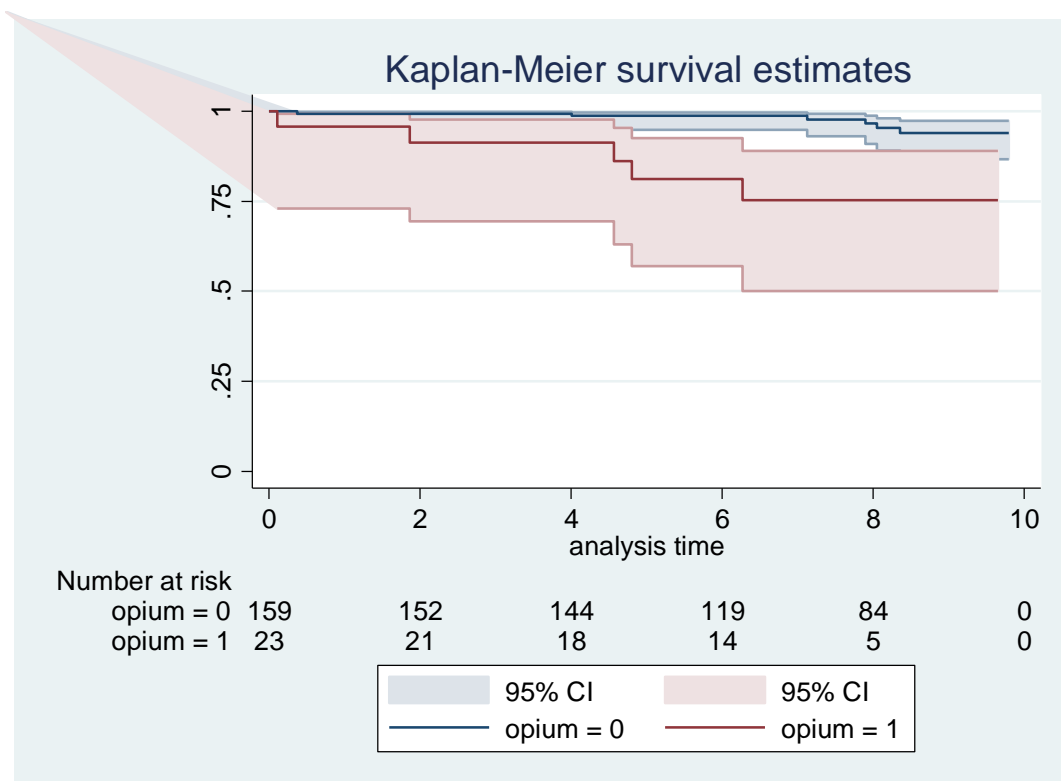


یا می توان با دستور `db stsggraph` کادر محاوره ای مربوطه را باز نمود. در روش کاپلان مایر نیاز به معرفی زمان ورود افراد به مطالعه، زمان خروج افراد از مطالعه و تنظیمات مربوط به زمان است. این تنظیمات قبلا در مبحث رگرسیون پوآسون مفصل توضیح داده شد. در اینجا نیز از دستور `stset` استفاده می شود. سپس برای مقایسه نمودارهای مربوطه و استفاده از آزمون لوگ رنگ از دستور `sts` استفاده می شود. این اقدامات در زیر انجام شده است.





نمودار بالا را می توان بر حسب متغیر کیفی مصرف تریاک و حدود اطمینان آن به همراه جدول به شرح زیر رسم نمود.



با استفاده از دستور زیر نیز می توان دو نمودار یا بقا در دو گروه را آزمون نمود.

. sts test opium

```
failure _d: cancerdeath == 1
analysis time _t: (timeout-origin)/365.25
origin: time timein
```

**Log-rank test for equality of survivor functions**

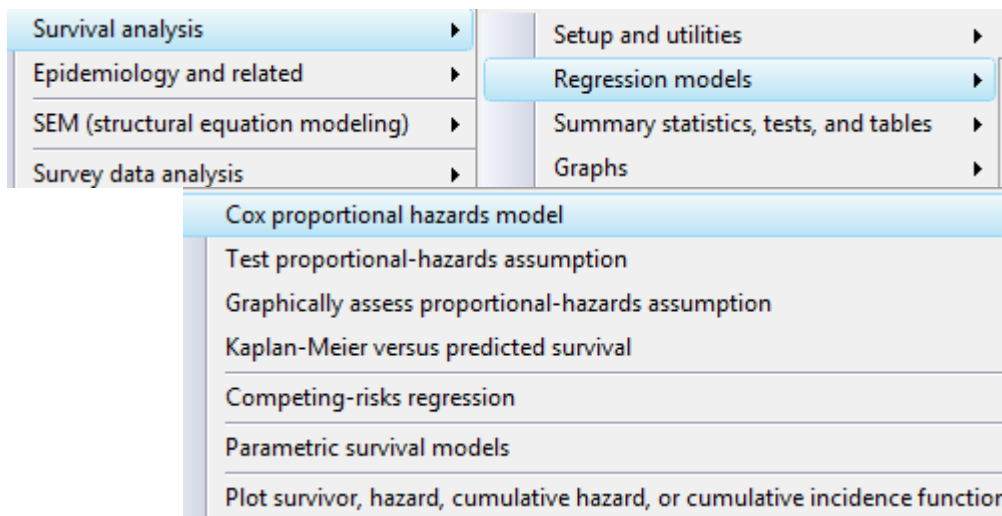
opium	Events observed	Events expected
0	6	9.93
1	5	1.07
Total	11	11.00

chi2(1) = 16.01  
Pr>chi2 = 0.0001

نتایج نشان می دهد بین دو گروه از نظر تابع بقا تفاوت معنی داری وجود دارد.  $P=0.0001$

**رگرسیون کاکس:**

برای مدل سازی و استفاده از رگرسیون کاکس با انتخاب گزینه های زیر از منوی statistics برنامه، از مسیر زیر اقدام می شود.



یا با دستور db stcox می توان کادر محاوره ای را باز و متغیرها را وارد نمود. برای انجام این تحلیل فایل COX regression را باز نمایید. این فایل نمونه ای از یک مطالعه کارآزمایی بالینی است. حجم نمونه ۱۹۱ نفر است. بیماران مبتلا به سیروز صفراوی (biliary cirrhosis) که بیماری مزمن و کشنده است به دو گروه درمانی مداخله مورد نظر با حجم نمونه ۹۷ نفر و گروه پلاسبو با حجم نمونه ۹۴ تقسیم تصادفی شدند. در طی مطالعه به ترتیب ۴۹ و ۴۷ نفر از بازوهای مطالعه فوت نمودند. ۷ مورد سانسور شده وجود داشت. توصیف گروههای مطالعه در زیر آمده است.

Treatment (1=placebo, 2=active)	Freq.	Percent	Cum.
1	90	48.91	48.91
2	94	51.09	100.00

متغیر d نیز به عنوان پیامد مطالعه مد نظر است. توصیف این متغیر در زیر آمده است.

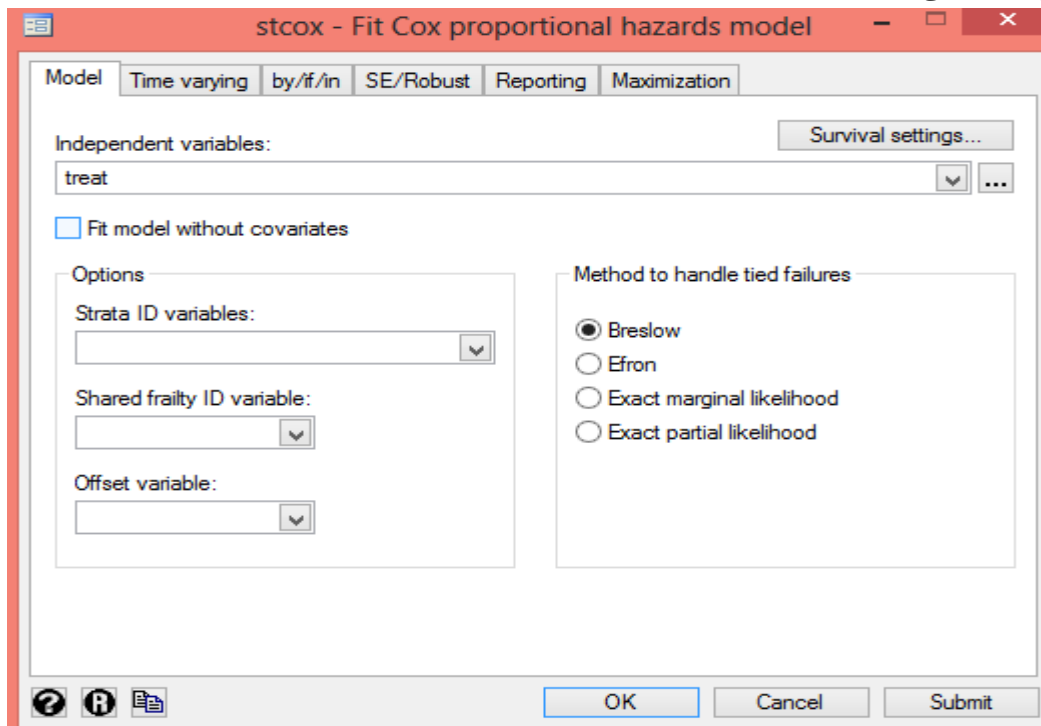
Outcome (0=alive, 1=dead)	Freq.	Percent	Cum.
0	88	47.83	47.83
1	96	52.17	100.00

جدول دو در دوی زیر، طبقه بندی افراد گروه ها و پیامد رخ داده را نشان می دهد.

. tab d treat

Outcome (0=alive, 1=dead)	Treatment (1=placebo, 2=active)		Total
	1	2	
0	41	47	88
1	49	47	96
<b>Total</b>	<b>90</b>	<b>94</b>	<b>184</b>

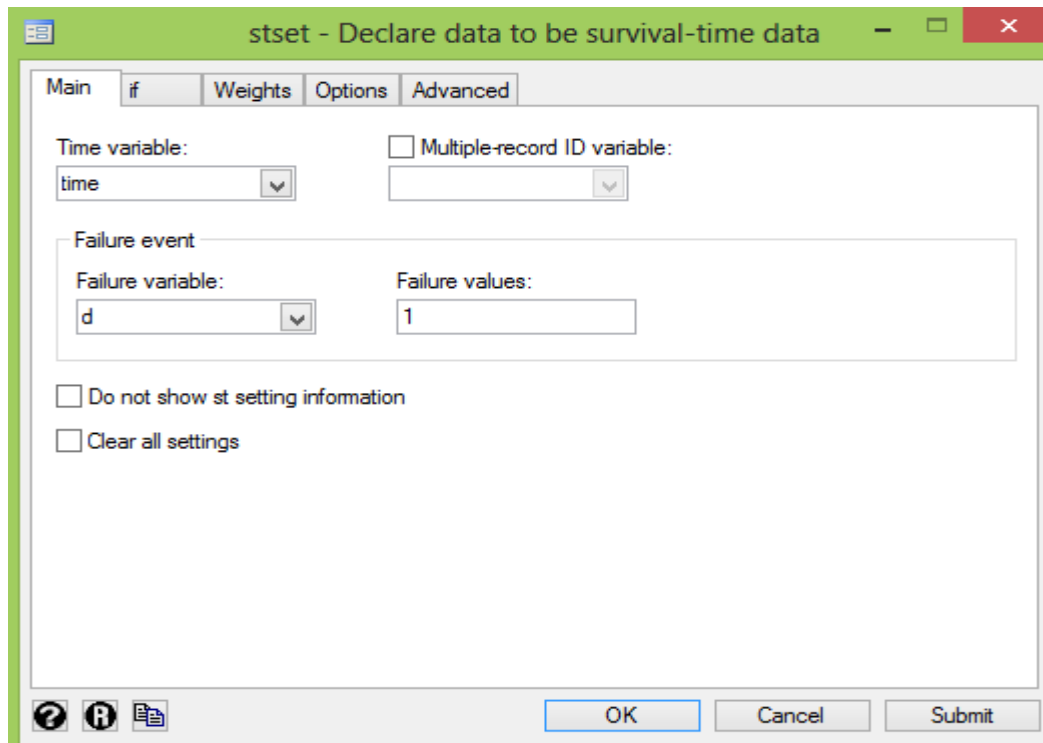
برای انجام تحلیل کاکس باید متغیر زمان پی گیری افراد (time) و متغیر رخداد واقعه (d) را به نرم افزار معرفی نمود. این تعاریف در قسمت survival setting از کادر محاوره ای زیر که متغیرهای مستقل و مخدوش کننده احتمالی تعریف می شوند قابل تعریف است.



کادر محاوره ای زیر تنظیمات زمان و معرفی متغیرها را که با دستور `stset time, failure(d==1) id(id)` نیز



قابل انجام است را نشان می دهد.



نتیجه تحلیل رگرسیون کاکس در زیر آمده است.

### stcox treat

Cox regression -- Breslow method for ties

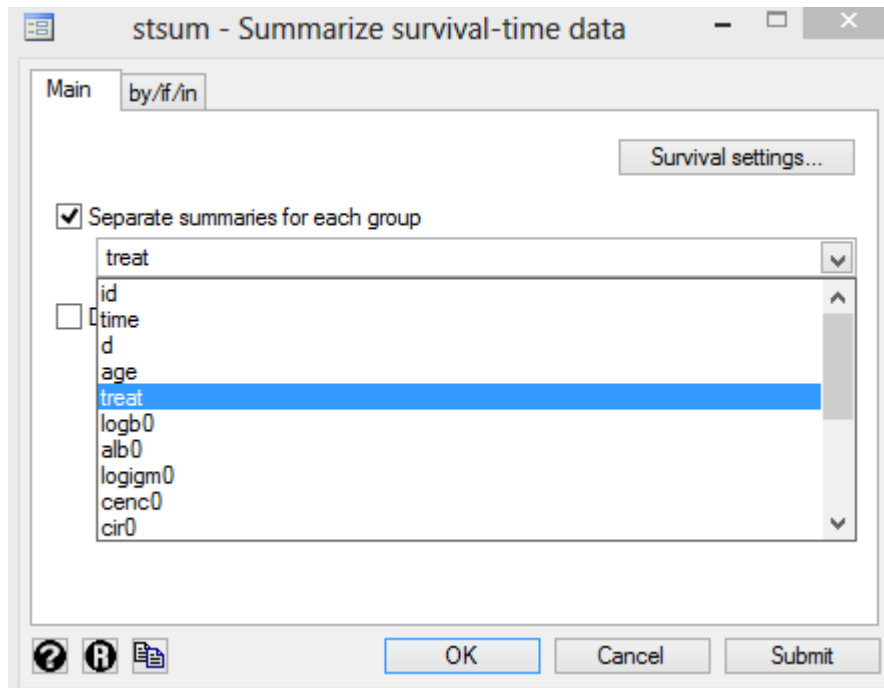
```

No. of subjects =          184          Number of obs =          184
No. of failures =           96
Time at risk   = 747.0390152
Log likelihood = -432.77204
LR chi2(1)     =           0.57
Prob > chi2    =           0.4509
    
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
treat	.8564564	.1760527	-0.75	0.451	.5724437 1.281379

نتایج نشان داده است که درمان مورد مداخله موثر نبوده و تفاوتی با پلاسبو نداشت.  $P=0.451$

با دستور `db stsum` نیز می توان کادر محاوره ای مربوطه را به صورت زیر باز و متغیر را معرفی و شاخص های توصیفی و زمان بقا را گزارش نمود.



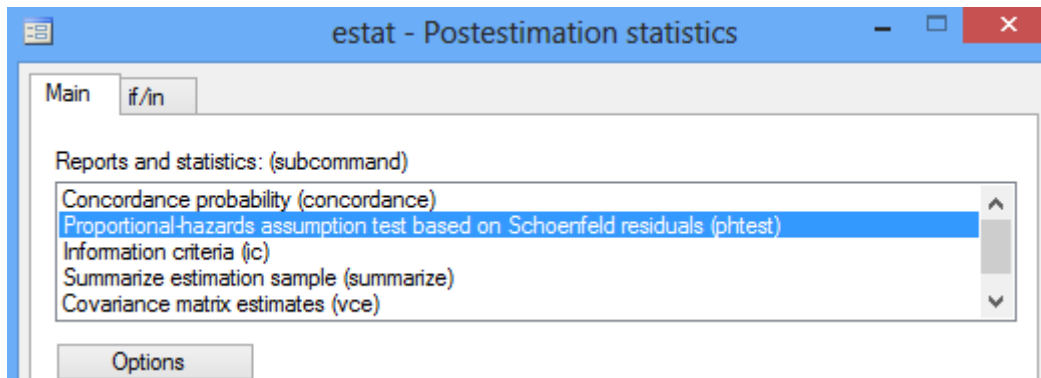
خروجی دستور بالا در زیر آمده است. میانه زمان بقا در دو گروه و کل افراد مورد بررسی ، چارک اول و سوم یا بقای ۲۵٪ و ۷۵٪ نیز گزارش شده است. میانه زمان بقا در گروه اول کمتر از گروه دوم است.

```
. stsum, by(treat)
```

```
failure _d: d == 1
analysis time _t: time
```

treat	time at risk	incidence rate	no. of subjects	Survival time		
				25%	50%	75%
1	355.0006853	.1380279	90	2.803559	5.275838	7.78371
2	392.0383299	.1198862	94	2.182067	6.047913	.
total	747.0390152	.1285073	184	2.592745	5.407255	9.259412

برای ارزیابی فرض تناسب خطر با دستور db estat و کادر زیر ارزیابی فرض انجام می شود.



نتیجه زیر نشان می دهد که فرض برقرار است.  $P=0.1437$

. estat phtest

Test of proportional-hazards assumption

Time: Time

	chi2	df	Prob>chi2
global test	2.14	1	0.1437

در تحلیل بالا متغیرهای مخدوش کننده وارد نشده است. مجدد تحلیل با ورود سایر متغیرهای مستقل از جمله مقدار بیلی روبین (**logb0**) که بر درمان تاثیر گذار هست تحلیل انجام می گیرد. نتایج و دستورات به شرح زیر است.

Cox regression -- Breslow method for ties

No. of subjects =	184	Number of obs =	184
No. of failures =	96		
Time at risk =	747.0390152		
Log likelihood =	-389.61311	LR chi2(2) =	86.89
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
treat	.6524091	.1380407	-2.02	0.044	.4309408 .9876942
logb0	10.61886	2.728634	9.19	0.000	6.417295 17.57128

اکنون نتایج بالا نشان می دهد که اثر درمان موثر بوده است. همچنین متغیر بیلی روبین نیز معنی دار شده است. یعنی متغیر بیلی روبین باعث تغییر اثر درمان شده است. و در مدل قبلی به عنوان متغیر مخدوش کننده اثر درمان مد نظر می باشد. چنانچه متغیر طبقه اجتماعی افراد نیز وارد شود مدل به صورت زیر خواهد شد. و

نشان می دهد که اثر درمان موثر نیست. و متغیرهای بیلی روبین و طبقه اجتماعی دوم در مدل معنی دار هستند.

xi: stcox treat logb0 i.socialclass

Cox regression -- Breslow method for ties

No. of subjects =	184	Number of obs =	184
No. of failures =	96		
Time at risk =	747.0390152		
Log likelihood =	-386.77699	LR chi2(4) =	92.56
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
treat	1.220113	.4093223	0.59	0.553	.6321784 2.354834
logb0	11.75412	3.042274	9.52	0.000	7.077449 19.52106
_Isocialcla_2	.3523194	.1525017	-2.41	0.016	.1508319 .8229619
_Isocialcla_3	.6109178	.1789852	-1.68	0.093	.3440328 1.08484

### مقایسه رگرسیون کاکس با رگرسیون پواسون:

برای مدل سازی بهتر و درک بهتر متغیرهای مخدوش کننده و اثر سن، مقایسه رگرسیون کاکس با رگرسیون پواسون انجام می گیرد. فایل coxregression2 را باز نمایید. ابتدا تنظیمات مربوط به زمان به صورت زیر انجام می شود.

```
stset timeout, id(id) failure(death) origin(time timein) scale(365.25)
2204.539 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 9.798768
```

با دستور زیر متغیر سیگار کشیدن (smokenum) و داده های گمشده در آنها توصیف می شود.

. tab smokenum, miss

No. of cigarettes per day	Freq.	Percent	Cum.
non-smok	140	44.03	44.03
ex-smoke	68	21.38	65.41
1-9 cigs	29	9.12	74.53
10-19 ci	30	9.43	83.96
20-29 ci	26	8.18	92.14
30+ cigs	24	7.55	99.69
.	1	0.31	100.00
Total	318	100.00	

جدول بالا نشان می دهد که این متغیر در ۶ طبقه وجود دارد. ۱ مورد نیز داده گمشده وجود دارد. از متغیر بالا متغیر جدیدی با ۳ طبقه ساخته می شود. دو طبقه اول مجدد ساخته می شوند و طبقات ۳ تا ۶ به عنوان یک طبقه ساخته می شوند. دستورات زیر ساخت این متغیر را با سه طبقه نشان می دهد.

```
gen smok= smokenum
```

```
recode smok 2/max=2
```

```
. tab smok
```

smok	Freq.	Percent	Cum.
0	140	44.16	44.16
1	68	21.45	65.62
2	109	34.38	100.00
<b>Total</b>	<b>317</b>	<b>100.00</b>	

میزان های بروز اختصاصی مرگ بر حسب متغیر سیگار کشیدن که جدید ساخته شد با دستور زیر به شرح زیر برآورد می شوند.

```
strate smok, per(1000)
```

smok	D	Y	Rate	Lower	Upper
0	30	0.9844	30.475	21.308	43.587
1	18	0.4616	38.995	24.569	61.893
2	40	0.7490	53.403	39.173	72.804

نتایج بالا نشان می دهد که به صورت روشن میزان ها در حال افزایش از طبقه ۰ به ۲ هستند. برای برآورد نسبت خطر مرگ در گروه های سیگاری نسبت به غیرسیگاری ها، تحلیل با رگرسیون کاکس به صورت زیر انجام می شود.

```
xi: stcox i.smok
```

Cox regression -- Breslow method for ties

```
No. of subjects =          317          Number of obs =          317
No. of failures =           88
Time at risk    = 2195.017112
Log likelihood  = -479.70522
LR chi2(2)     =           5.44
Prob > chi2    =           0.0658
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_ismok_1	1.294995	.386338	0.87	0.386	.7216519 2.323852
_ismok_2	1.749532	.4226447	2.32	0.021	1.089663 2.808999

نتایج نشان می دهد که برآورد برای گروه دوم نسبت به گروه رفرانس معنی دار است. نسبت خطر گروه اول به گروه صفر معنی دار نیست. حدود اطمینان آنها نیز آمده است. آزمون Likelihood ratio در بالای جدول نشان می دهد که اثر سیگار کشیدن در مقابل عدم کشیدن سیگار با  $p=0.0658$  در مرز بینا بینی معنی داری قرار دارد. این نتیجه نشان می دهد یا این متغیر معنی دار نیست و یا اینکه خوب تعریف و طبقه بندی نشده است. برای مقایسه نتایج کاکس با رگرسیون پواسون به صورت زیر اقدام می شود.

```
Poisson regression
Log likelihood = -282.9754
Number of obs = 317
LR chi2(2) = 5.49
Prob > chi2 = 0.0644
Pseudo R2 = 0.0096
```

death	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ismok_1	1.279578	.3814964	0.83	0.408	.713325	2.295335
_Ismok_2	1.752358	.4232347	2.32	0.020	1.091536	2.813245
_cons	.0304752	.005564	-19.12	0.000	.0213078	.0435867
ln(years)	1	(exposure)				

نتایج این تحلیل دقیقاً مشابه تحلیل رگرسیون کاکس است که در خروجی قبلی نشان داده شد. در اینجا ممکن است سن اثر مخدوش کننده داشته باشد و باید کنترل شود. دو راه برای کنترل اثر سن کنونی افراد وجود دارد یا داده ها را بر روی متغیر جدید گروه بندی شده سن بازسازی شود و سپس تحلیل انجام گیرد یا اینکه برای origin متغیر تاریخ تولد افراد تعریف گردد. برای این کار مجدد تنظیمات زیر برای زمان انجام می گیرد.  
 tset timeout, id(id) failure( death) enter( timein) origin( timebth) scale(365.25)

### راه های کنترل اثر مخدوش کننده:

دو راه برای کنترل اثر سن به عنوان متغیر مخدوش کننده وجود دارد.

### راه اول: از طریق رگرسیون کاکس:

با وارد کردن متغیر سن تولد افراد در قسمت Origin مربوط به تنظیمات برنامه ، نتیجه در زیر آمده است.

Cox regression -- no ties

```
No. of subjects = 317
No. of failures = 88
Time at risk = 2195.017112
Log likelihood = -430.44829
Number of obs = 757
LR chi2(2) = 5.68
Prob > chi2 = 0.0585
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ismok_1	1.277505	.3814011	0.82	0.412	.7115987	2.293455
_Ismok_2	1.773063	.4299107	2.36	0.018	1.10239	2.851761

## راه دوم از طریق رگرسیون پوآسون:

داده ها بر اساس گروه سنی جدید بازسازی و با توجه به اینکه در گروه های سنی جدید زمان ها فرق دارد باید متغیر زمان در معرض خطر بودن و پی گیری افراد ابتدا حذف و متغیر پیامد جدید بنام `_d` و متغیر زمان در معرض خطر بودن بنام `y` جدید ساخته گردند. این اقدامات و نتایج به شکل زیر است.

```
stsplit ageband, at(60(5) 80)
tab ageband _d
drop y
gen y=_t - _t0
gen d=_d
xi: poisson d i.smok i.ageband, e(y) irr
```

نتایج زیر نشان می دهد که نسبت میزان بروز برای متغیر سیگار کشیدن همانند گذشته تغییری ننموده است و بنابراین مخدوش کننده نیست.

Poisson regression	Number of obs	=	757
	LR chi2(5)	=	11.82
	Prob > chi2	=	0.0373
Log likelihood = -351.893	Pseudo R2	=	0.0165

d	IRR	Std. Err.	z	P> z	[95% Conf. Interval]
_Ismok_1	1.264397	.3771704	0.79	0.432	.7046438 2.268806
_Ismok_2	1.785577	.4326039	2.39	0.017	1.110586 2.870814
_Iageband_65	1.583683	.5365641	1.36	0.175	.8152196 3.076535
_Iageband_70	1.252102	.4559221	0.62	0.537	.6133319 2.556135
_Iageband_75	2.772364	1.184875	2.39	0.017	1.199657 6.406834
_cons	.0208765	.0069991	-11.54	0.000	.0108214 .0402746
ln(y)	1	(exposure)			

برای ارزیابی فرض ثابت بودن نسب خطر در طی زمان (RR در هر زمانی ثابت است.) که معادل فرض عدم اینترکشن اثر مورد نظر با زمان در رگرسیون پوآسون است. به چند طریق عمل می شود. یکی اینکه بعد از راند کردن مدل رگرسیون کاکس با دستور `estat phtest` این کار صورت گیرد. گزینه های زیر از منوی تحلیل بقا انتخاب می شود.

Regression models	Cox proportional hazards model
Summary statistics, tests, and tables	Test proportional-hazards assumption
Graphs	Graphically assess proportional-hazards assumption

. estat phtest

Test of proportional-hazards assumption

Time: Time

	chi2	df	Prob>chi2
global test	0.20	2	0.9031

آزمون انجام شده نشان می دهد که فرض تناسب خطرها برقرار است.

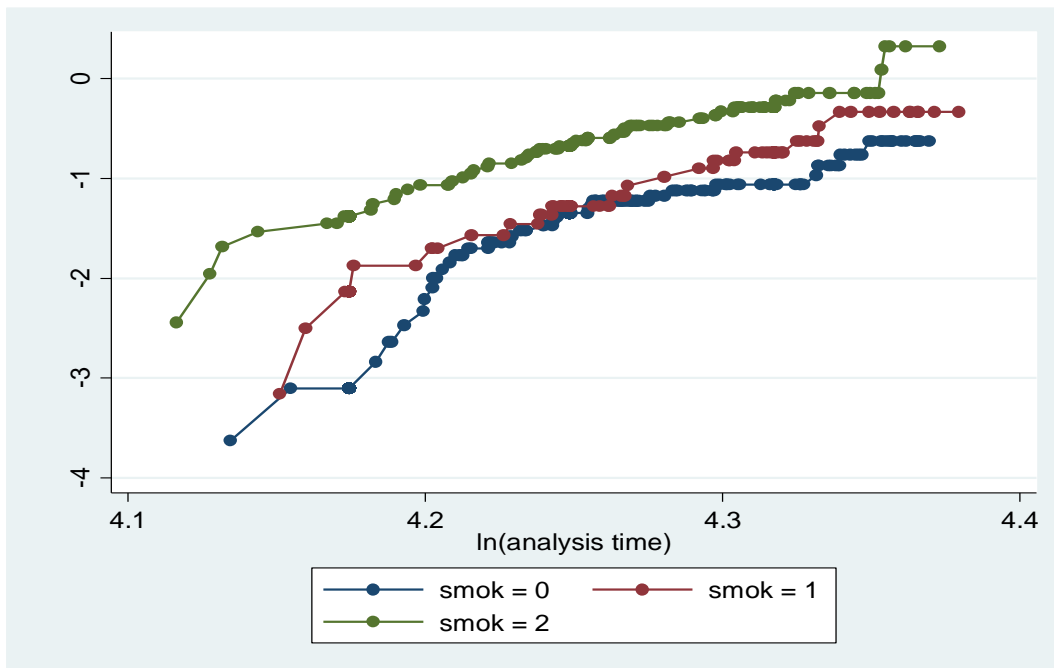
روش دیگر از طریق گرافیکی است. که با انتخاب گزینه سوم در تصویر بالا به صورت زیر انجام می شود.

. sthplot, by(smok)

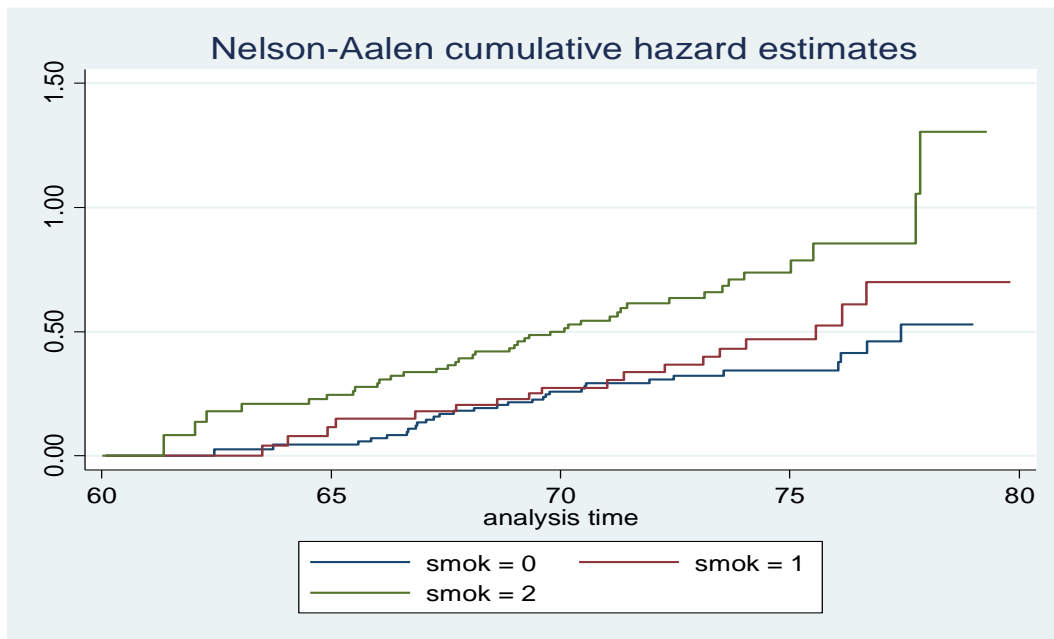
```

failure _d: death == 1
analysis time_t: (timeout-origin)/365.25
              origin: time timein
enter on or after: time timein
              id: id
    
```

نمودار را می توان در مقیاس لگاریتمی یا با نمودار نلسون ان (sts graph if smok!=., by(smok) na نیز به صورت زیر رسم نمود. لذا متغیرهایی باید وارد شوند که این فرض (موازی بودن تابع خطر) را داشته باشند.







بنابراین در مدل رگرسیون کاکس با فرض PH، می توان چند متغیر را وارد مدل نمود و اثر آنها را بررسی کرد. کاکس یک مدل نیمه پارامتریک و Robust است. بر خلاف روش کاپلان مایر که می توان دو سطح از یک متغیر را با هم مقایسه نمود در اینجا همه سطوح متغیر را می توان مقایسه کرد.

### تحلیل داده های وابسته به هم:

تحلیل هایی که در قسمت های گذشته معرفی و بکار رفته بوند ، با فرض استقلال مشاهدات انجام گردیدند. وقتی شرط استقلال برقرار نباشد از روشهای دیگر تحلیل استفاده می شود. در این قسمت از کتاب مقدمه ای بر این روش ها گفته می شود.

### این روش ها عبارتند از:

Robust Standards Errors

Generalised Estimating Equations (GEE)

Random Effect(Multi-Level) Models

لذا در داده هایی که چندین بار اندازه گیری برای پیامد صورت می گیرد و افراد در زمان های مختلف پی گیری می شوند، نمونه ها یا افراد استقلال نداشته و بهم وابسته هستند. نمونه ها می تواند فرد در خانوار، چشم در یک فرد یا اعضای داخل یک خانوار باشند. بنابراین در تحلیل، این وابستگی ها باید لحاظ گردد. در روش های تحلیل بالا همبستگی در نظر گرفته می شود. اگر این مهم در نظر گرفته نشود ممکن است خطای نوع اول رخ دهد. زیرا مقدار p کوچک می شود و حدود اطمینان باریک تر میگردد و واریانس کوچکتر می شود. و در واقع در بررسی

رابطه ارتباطی وجود ندارد و بخاطر نوع تحلیل نادرست ذکر می شود که ارتباط وجود دارد. لذا باید در هر کلاستر یا خوشه که تحلیل انجام می شود یک شاخص خلاصه محاسبه و با کلاسترهای دیگر مقایسه گردد. از آنجاییکه اثر هر فرد را دقیق نمیتوان اندازه گیری نمود و پیچیده است این نگاه نیز با مشکل روبرو می شود.

### کار با نرم افزار و انجام تحلیل داده های وابسته به یکدیگر:

برای مدل های تحلیل داده های وابسته به هم فایل GEE را باز نمایید. این مربوط به یک مطالعه است بر روی ۷۰ نفر بیمار مبتلا به سل ریوی در زیمباوه، ۴۲ نفر از این افراد HIV مثبت و ۲۸ نفر دیگر منفی هستند.

هدف از این مطالعه تعیین این است که افراد مبتلا به سل و HIV مثبت آیا بیماری سل را به دیگر اعضای خانواده بیشتر یا کمتر منتقل می کنند؟ یا عبارتی فرد شاخص مبتلا به سل (فرد ایندکس) اگر مثبت باشد انتقال سل در خانوار بیشتر است یا کمتر؟ و آیا وابسته به افراد هر خانوار می باشد یا نه؟ جدول زیر توصیف اولیه این متغیرها را نشان می دهد. متغیر MANTOUX وضعیت بیماران سلی را طبقه بندی و نشان می دهد. متغیر HIV نیز وضعیت ایدز در بیماران را مشخص نموده است.

Mantoux test status	HIV status of index case		Total
	Negative	Positive	
Negative	36 28.57	87 48.07	123 40.07
Positive	90 71.43	94 51.93	184 59.93
Total	126 100.00	181 100.00	307 100.00

جدول نشان می دهد که شیوع تست مثبت سل در کلیه افراد حدود ۶۰ درصد (59.93) است. همچنین در این بیماران حدود ۵۲ درصد HIV مثبت هستند.

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
0.432184	11.72	0.0006	0.263575	0.708651

نتایج نشان می دهد فردی که HIV مثبت است و شانس اینکه بیماری سل را به دیگران منتقل نماید کمتر از فردی است که HIV منفی باشد. نسبت شانس مذکور و حدود اطمینان آن در جدول بالا آمده است. نسبت شانس مذکور معنی دار است. چرا؟ چنانچه اثر خوشه ای بودن افراد بیمار را در نظر نگیریم بدست آمدن مقدار

پی ویو در جدول بالا شواهد قوی را برای این ارتباط نشان می دهد. و این رابطه ناشی از شانس و خطای تصادفی نیست. نتیجه گرفتن از این رابطه که ممکن است ناشی از در نظر نگرفتن اثر خوشه باشد نیز غلط است. چه توجیهی برای این رابطه قابل تصور است؟

### روش اول تحلیل: Robust Standards Errors

در این روش بجای تغییرات مشاهده شده که در روش های معمول استفاده می شود، چون فرض استقلال برقرار نیست از تغییرات مشاهده شده در کل داده ها استفاده می شود. یعنی بر مبنای تفاوت تغییرات واریانس در مشاهدات و واریانس باقی مانده های مدل کار می کند. در هر کلاستر واریانس محاسبه می شود سپس واریانس کل کلاسترها با هم جمع و محاسبه می شوند. وقتی تحلیل رگرسیون لجستیک برای این داده ها بدون در نظر گرفتن کلاستر و همبستگی داده ها انجام شود نتایج زیر حاصل می شود.

```

Logistic regression                Number of obs   =       307
                                   LR chi2(1)       =       11.98
                                   Prob > chi2        =       0.0005
Log likelihood = -200.70621        Pseudo R2       =       0.0290
    
```

	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Ihiv_2	.4321839	.1067603	-3.40	0.001	.2663179 .7013532
_cons	2.5	.4930066	4.65	0.000	1.69856 3.679588

در اینجا نیز نتایج همانند روش کلاسیک که در ابتدای مبحث تحلیل گردید بوده و نسبت شانس معنی دار است. اما در روش محاسبه واریانس Robust با استفاده از دستور زیر اجرا می شود. یعنی روش محاسبه واریانس به مدل کلاسیک لجستیک اضافه شده است. نتایج زیر نشان می دهد تفاوت خیلی زیادی بین دو روش مشاهده نمی شود. در این روش نسبت شانس و واریانس (خطای معیار) دقیق تر محاسبه شده است اما تفاوت قابل ملاحظه ای وجود ندارد.

```

. xi: logistic mantoux i.hiv, robust
i.hiv          _Ihiv_1-2      (naturally coded; _Ihiv_1 omitted)
    
```

```

Logistic regression                Number of obs   =       307
                                   Wald chi2(1)       =       11.50
                                   Prob > chi2        =       0.0007
Log pseudolikelihood = -200.70621        Pseudo R2       =       0.0290
    
```

	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]
_Ihiv_2	.4321839	.1069346	-3.39	0.001	.2661075 .7019079
_cons	2.5	.4938115	4.64	0.000	1.697488 3.681911

در اجرای روش بالا می توان به نرم افزار دستور داد که ضمن محاسبه بالا اثر درون هر خانوار بعنوان خوشه نیز در نظر گرفته شود. این تحلیل و نتیجه آن در زیر می آید.

در خروجی این تحلیل که در زیر آمده است مشخص است که خطای معیار بزرگتر شده است. یعنی از 0.106 در روش بالا به 0.143 رسیده است. و این بخاطر در نظر گرفتن اثر خوشه است. چون خطای معیار تحت تاثیر قرار گرفت بزرگتر گردید. حدود اطمینان نیز پهن شده است. اما همچنان نسبت شانس معنی دار است.

```
. xi: logistic mantoux i.hiv, robust cluster( id)
i.hiv          _Ihiv_1-2          (naturally coded; _Ihiv_1 omitted)

Logistic regression              Number of obs   =       307
                                Wald chi2(1)    =        6.37
                                Prob > chi2     =       0.0116
Log pseudolikelihood = -200.70621 Pseudo R2      =       0.0290

                                (Std. Err. adjusted for 70 clusters in id)
```

mantoux	Robust		z	P> z	[95% Conf. Interval]	
	Odds Ratio	Std. Err.				
_Ihiv_2	.4321839	.1436037	-2.52	0.012	.2253389	.828898
_cons	2.5	.6662136	3.44	0.001	1.482884	4.21476

خلاصه اینکه روش محاسبه خطای معیار Robust با استفاده از رویکرد حداکثر درست نمایی (ML) برآورد نقطه ای پارامتر مورد نظر را بدون در نظر گرفتن همبستگی در داده ها محاسبه می کند. برای برآورد خطای معیار، بجای استفاده از Log likelihood، از روش تقریبی quadratic استفاده می کند. خطای معیار در این روش در سطح خوشه ها محاسبه می شود. سپس ارتباط بین خوشه ها با توجه به این خطای معیار محاسبه می شود. در این روش برای مقایسه مدل ها از آزمون likelihood ratio نمی توان استفاده کرد. در استفاد از این مدل پیشنهاد می شود وقتی حداقل ۳۰ خوشه وجود دارد و توزیع نرمال باشد از این روش استفاده شود.

### روش دوم تحلیل: Generalised Estimating Equations (GEE)

روش GEE نیز مانند روش قبل با خطای معیار Robust کار می کند و مانند آن است با این تفاوت که در این روش همبستگی بین داده ها نیز در نظر گرفته می شود. از روش های مختلفی برای برآورد همبستگی بین داده ها استفاده می شود. روش اول استفاده از Exchangeable در محاسبه همبستگی است. دو مشاهده در این روش در یک خوشه همبستگی یکسان دارند اما بین خوشه ها همبستگی یکسان نیست. معمولاً این روش استفاده می شود. روش دیگر در هنگامی که اندازه گیری های تکراری در طول زمان در افراد مشابه، برای اندازه گیری پیامد وجود دارد، استفاده از روش autocorrelation است. وقتی اندازه ها در طولانی مدت یا زمان طولانی انجام شود همبستگی قوی تری بدست می آید. همبستگی در زمان کوتاه، کمتر بدست می آید. فاصله

زمانی بین دو اندازه گیری وقتی طولانی باشد همبستگی کوچکتر بدست می آید. انواع دیگر از روشهای محاسبه همبستگی نیز وجود دارد که برای اطلاع بیشتر به کتاب آماری مراجعه شود.

برای استفاده از این روش ابتدا با فرض استقلال داده ها با دستور "corr (ind)" مدل تحلیل انجام می شود. در این روش با رویکرد دانستن متوسط جمعیت (pa) تحلیل انجام می شود. یعنی برای هر فرد در یک کلاستر خاص یکسان است اما هر فرد در یک کلاستر نسبت به کلاستر دیگر متفاوت است.

روش تحلیل با دستور xi: xtlogit mantoux i.hiv, pa corr(ind) i(id) انجام و نتیجه زیر حاصل می شود.

Iteration 1: tolerance = 3.622e-16

GEE population-averaged model		Number of obs	=	307
Group variable:	id	Number of groups	=	70
Link:	logit	Obs per group: min	=	1
Family:	binomial	avg	=	4.4
Correlation:	independent	max	=	13
		Wald chi2(1)	=	11.53
Scale parameter:	1	Prob > chi2	=	0.0007
Pearson chi2(307):	307.00	Deviance	=	401.41
Dispersion (Pearson):	1	Dispersion	=	1.307532

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_Ihiv_2	-.8389041	.2470252	-3.40	0.001	-1.323065 - .3547436
_cons	.9162907	.1972027	4.65	0.000	.5297806 1.302801

نتایج بالا در مدل تحلیل GEE نشان می دهد که برآورد نقطه ای و خطای معیار بدست آمده با روش استاندارد و استفاده از Likelihood یعنی همان تحلیل با رگرسیون لجستیک که قبلا انجام شده است یکسان است. برای اینکه به نرم افزار دستور داده شد که اولاً با فرض استقلال داده ها کار نماید و ثانياً اینکه از روش خطای معیار Robust استفاده نکند.

چنانچه به نرم افزار گفته شود با روش خطای معیار Robust تحلیل را با ماتریس همبستگی و شرط استقلال انجام دهد نتایج زیر بدست می آید. در اینجا نیز برآوردهای بدست آمده مشابه رویکرد اول (خطای معیار Robust) است. و همبستگی درون خانوارها نیز در نظر گرفته شد. که در اینجا با توجه به اینکه به نرم افزار گفته شد با فرض استقلال محاسبه را انجام دهد تطبیق بر اساس همبستگی درون خانوار انجام گردید.

```
. xi: xtlogit mantoux i.hiv, pa corr(ind) i( id) robust
i.hiv          _Ihiv_1-2          (naturally coded; _Ihiv_1 omitted)

Iteration 1: tolerance = 3.622e-16

GEE population-averaged model
Group variable:          id          Number of obs      =      307
Link:                   logit        Number of groups   =      70
Family:                 binomial     Obs per group: min =      1
Correlation:           independent   avg =              4.4
Scale parameter:       1            max =              13
Wald chi2(1)          =      6.37
Prob > chi2           =      0.0116

Pearson chi2(307):     307.00    Deviance           =      401.41
Dispersion (Pearson): 1            Dispersion         =      1.307532

(Std. Err. adjusted for clustering on id)
```

mantoux	Semirobust			z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.					
_Ihiv_2	-.8389041	.3322744	-2.52	0.012	-1.49015	-.1876581	
_cons	.9162907	.2664854	3.44	0.001	.3939889	1.438593	

چنانچه برای محاسبه همبستگی درون خانوارها بجای فرض استقلال، دنبال استفاده از روش Exchangeble باشید از دستور زیر استفاده نمایید ، نتایج به صورت زیر خواهد بود.

```
. xi: xtlogit mantoux i.hiv, pa corr(exch) i( id) robust
i.hiv          _Ihiv_1-2          (naturally coded; _Ihiv_1 omitted)

Iteration 1: tolerance = .06827002
Iteration 2: tolerance = .00201465
Iteration 3: tolerance = .00003962
Iteration 4: tolerance = 7.362e-07

GEE population-averaged model
Group variable:          id          Number of obs      =      307
Link:                   logit        Number of groups   =      70
Family:                 binomial     Obs per group: min =      1
Correlation:           exchangeable   avg =              4.4
Scale parameter:       1            max =              13
Wald chi2(1)          =      8.75
Prob > chi2           =      0.0031

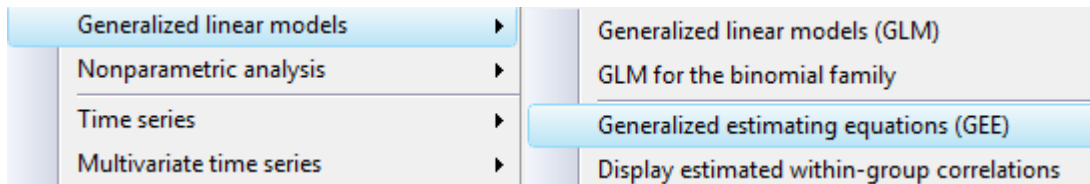
(Std. Err. adjusted for clustering on id)
```

mantoux	Semirobust			z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.					
_Ihiv_2	-.9684832	.3273525	-2.96	0.003	-1.610082	-.3268841	
_cons	1.011113	.2617372	3.86	0.000	.4981178	1.524109	

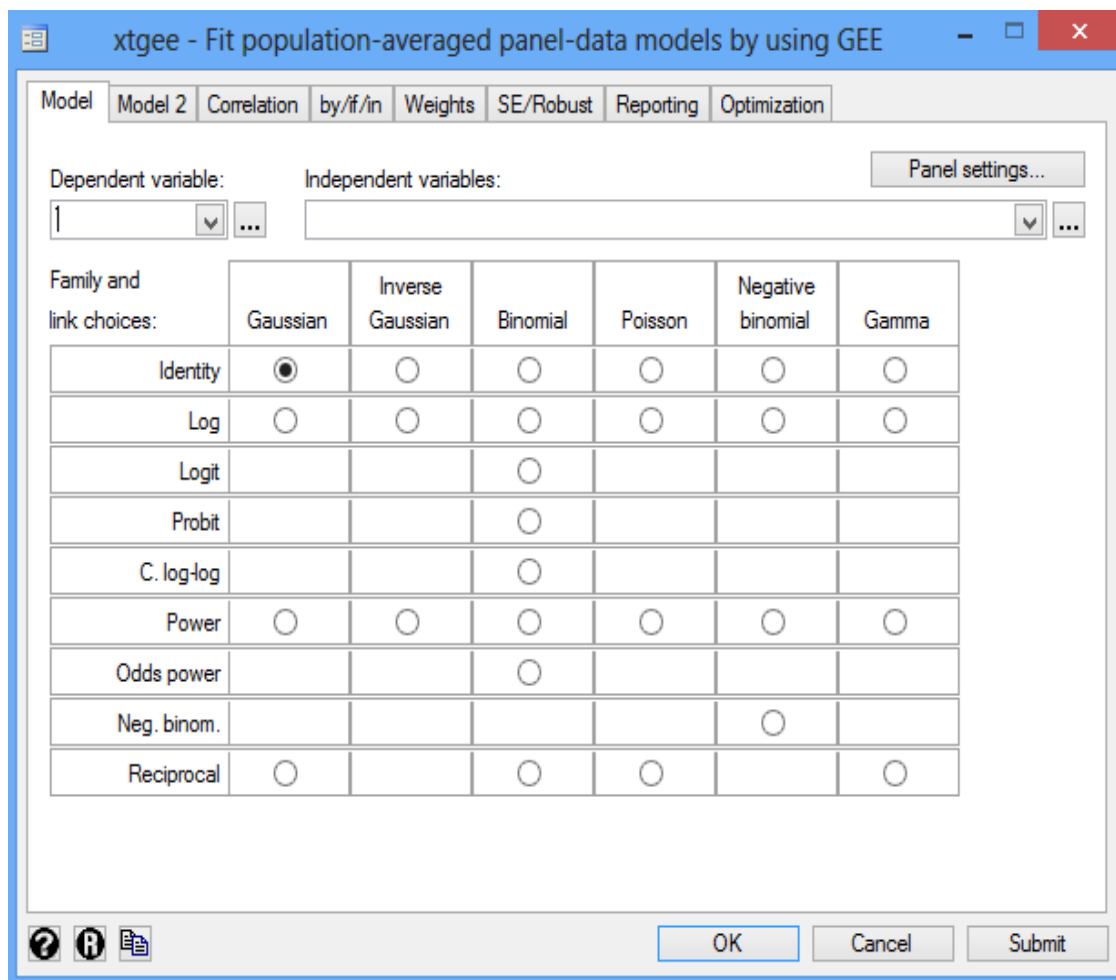
در تحلیل بالا نیز برآوردهای انجام شده مشابه تحلیل قبلی است. اما برآورد نقطه ای log odds ratio به -0.9685 تغییر پیدا نمود. در مدل قبلی این برآورد -0.8389 بود. بخاطر اینکه این نوع تحلیل همبستگی درون

خانوارها یا خوشه ها را در نظر گرفت. در روش GEE محاسبه واریانس بر اساس مدل احتمال نیست. فقط ماتریس همبستگی برای مدل تعریف می شود و همبستگی ها در هر کلاستر محاسبه و پارامترها برآورد می شوند. این تحلیل با همبستگی بین کلاسترها کاری ندارد. اما داخل کلاسترها را در نظر می گیرد.

روش GEE وقتی حجم نمونه بالا باشد و تعداد خوشه ها نیز زیاد باشد روشی دقیق و پایدار است. این روش را از مسیر زیر نیز می توان انجام داد.



با معرفی متغیرهای وابسته و مستقل و انتخاب نوع مدل در کادر زیر تحلیل مانند بالا قابل انجام است.



## روش سوم تحلیل : Random Effect(Multi-Level) Models

این روش بر اساس احتمال کار می کند. و از روش ماتریس همبستگی استفاده نمی کند. در این مدل اثر در هر کلاستر برای هر فرد متفاوت است. احتمال برای هر فرد در یک کلاستر، نسبت به کلاسترهای دیگر ممکن است تغییر کند. افراد داخل یک کلاستر یکسان هستند. عبارتی فرد 1 داخل کلاستر 2 با فرد 2 داخل کلاستر 1 یکسان است. و فرد 2 در کلاستر 1 با فرد 2 در کلاستر 2 متفاوت هستند. پارامتری که برای متغیر برآورد می شود متوسط برای همه افراد است. در خلاصه می توان گفت effect یک نفر در هر کلاستر باعث اختلاف بین کلاسترها می شود. همچنین متغیرهای زمینه ای که برای افراد اندازه گیری نشد می تواند متفاوت باشند. اگر همه متغیرها اندازه گیری و وارد مدل می شدند واریانس صفر بود. اما آن متغیرهایی که اندازه گیری نشدند و تفاوت هایی که افراد در کلاسترها با هم دارند را می توان از علل اختلاف کلاسترها دانست. در این مدل فرض می شود که داده ها توزیع احتمال داشته و از توزیع نرمال تبعیت می کنند. برای اجرای این تحلیل در مثال گذشته دستور زیر مدل را راند می کند.

```
xi: xtlogit mantoux i.hiv, re i( id)
```

```
Random-effects logistic regression      Number of obs      =      307
Group variable: id                    Number of groups   =       70

Random effects u_i ~ Gaussian          Obs per group: min =        1
                                         avg =       4.4
                                         max =       13

Wald chi2(1)                          =       8.54
Prob > chi2                            =      0.0035

Log likelihood = -194.25722
```

mantoux	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ihiv_2	-1.1508	.3938605	-2.92	0.003	-1.922753	-.378848
_cons	1.200259	.3166709	3.79	0.000	.5795955	1.820923
/lnsig2u	-.0324611	.4965861			-1.005752	.9408298
sigma_u	.9839005	.2442956			.6047888	1.600658
rho	.2273547	.0872326			.1000563	.4378191

```
Likelihood-ratio test of rho=0: chibar2(01) = 12.90 Prob >= chibar2 = 0.000
```

در خروجی بالا مقدار rho برابر با 0.22 گزارش گردید. این مقدار همبستگی درون خانوار را نشان می دهد. اگر این مقدار و مقدار برآورد sigma u صفر باشند یعنی اینکه کلاستر درون خانوارها وجود ندارد. این مقدار نیز آزمون گردید (فرض rh=0) یعنی واریانس درون کلاسترها صفر است.



وقتی آزمون انجام گردید نتایج در پایین جدول مذکور نشان می دهد که فرض رد می شود و واریانس درون کلاسترها صفر نیست. برآورد  $\sigma^2 u$  نیز 0.98 گزارش گردید. اگر این مقدار صفر باشد یعنی واریانس بین کلاسترها وجود ندارد. برآورد نسبت شانس با مدل قبلی نیز تفاوت پیدا کرده است اما همچنان معنی دار است. نتایج نشان می دهد که کلاستر (Clustering) در دورن خانوارهای وارد شده به مدل وجود دارد. در این مدل از توزیع نرمال و توزیع دوجمله ای استفاده می شود تا برآوردها به صورت تقریبی محاسبه گردد و روش پیچیده ای است. محاسبات ممکن است دقیق نباشند لذا باید این مدل را چک نمود. چنانچه  $\rho$  بیشتر از 0.25 بود یا تعداد افراد داخل هر کلاستر بیشتر از ۲۰ بود، باید این مدل با دستورات زیر چک گردد. این چک کردن به منظور تصمیم گیری درباره پایدار بودن مدل و برآوردها انجام می شود. اگر تفاوت نسبی بیشتر از 0.01 باشد نشان از تردید در برآوردها وجود دارد. در مدل مذکور پایایی مدل نیاز به ارزیابی ندارد. با فرض مقدار  $\rho$  برابر با ۰/۴۹، پایایی مدل به صورت زیر ارزیابی می شود. وقتی مدل ارزیابی می شود نتایج زیر بدست می آید:

. quadchk, nooutput

Refitting model intpoints() = 8  
Refitting model intpoints() = 16

#### Quadrature check

	Fitted quadrature 12 points	Comparison quadrature 8 points	Comparison quadrature 16 points	
Log likelihood	-194.25722	-194.25726 - .00003849 1.982e-07	-194.25722 -3.751e-09 1.931e-11	Difference Relative difference
mantoux: _Ihiv_2	-1.1508004	-1.1507835 .00001687 -.00001466	-1.1508001 2.900e-07 -2.520e-07	Difference Relative difference
mantoux: _cons	1.2002591	1.200242 -.00001711 -.00001426	1.2002588 -3.547e-07 -2.955e-07	Difference Relative difference
lnsig2u: _cons	-.03246107	-.03248382 -.00002275 .00070087	-.03246251 -1.438e-06 .00004429	Difference Relative difference

با نگاه به تفاوت های نسبی (relative difference) اگر این مقادیر بیشتر از 0.01 باشند برآوردها پایدار نیستند. این مشکل در مدل ناشی از ترکیب توزیع دوجمله ای و توزیع نرمال است. وقتی پیامد کمی پیوسته باشد این مشکل مرتفع می شود.

**خلاصه مقایسه نتایج روش های مختلف تحلیل داده های وابسته به هم در فابل مورد بحث :**

ردیف	نوع مدل تحلیل	نسبت شانس برای HIV و CI95%	خطای معیار برای log odds ratio	P Value
۱	رگرسیون لجستیک معمولی	0.43 (0.27-0.7)	0.25	0.001
۲	رگرسیون لجستیک با روش خطای معیار Robust برای خوشه ها	0.43 (0.23-0.83)	0.33	0.01
۳	رگرسیون لجستیک با GEE و روش Exchangeable برای ماتریس همبستگی و خطای معیار Robust	0.38 (0.2 -0.72)	0.33	0.003
۴	روش رگرسیون لجستیک Random Effect	0.32 (0.15-0.69)	0.39	0.003

در **روش اول** تحلیل افراد در هر خانوار با افراد خانوارهای دیگر یکسان در نظر گرفته شده است. نوع تحلیل برای اینگونه داده ها صحیح نیست. در **روش دوم** برای محاسبه خطای معیار ، کلاسترها در نظر گرفته شده است. اما در برآورد نسبت شانس، کلاسترها در نظر گرفته نمی شود.

برآورد نسبت شانس در این روش مانند روش اول است. اما در اینجا خطای معیار بزرگتر شده است. این رویکرد صحیح است اما بهینه و مناسب نیست. در **روش تحلیل سوم** و استفاده از GEE، روش قبلی ارتقا و بهبود پیدا می کند و کلاسترها نیز در محاسبه نسبت شانس لحاظ شدند. برآورد نسبت شانس در این روش با روش قبلی متفاوت شده اما خطای معیار دو روش شبیه هم است. این رویکرد تحلیل نیز صحیح بوده اما از نظر رویکرد آماری شاید راضی کننده نباشد.

در **روش چهارم** که رویکرد متفاوتی با روشهای قبلی دارد. برآورد نسبت شانس و خطای معیار هر دو با در نظر گرفتن خوشه ها محاسبه شده است. نسبت شانس در این روش با روش سوم متفاوت بوده است. این رویکرد نیز صحیح و معتبر است. اما در این روش نیاز به چک کردن مدل و برآوردها می باشد تا پایایی آنها مورد تایید قرار گیرد.

نسبت شانس که با مدل GEE برآورد می شود اغلب نسبت شانس متوسط جمعیت خوانده می شود. نسبت شانس که با روش اثر تصادفی (Random Effect) برآورد می شود اغلب نسبت شانس اختصاصی خوشه ها تلقی می شود. وقتی توزیع متغیر پیامد نرمال باشد دو روش پارامترها را یکسان برآورد می کنند.

بر اساس شواهد مختلف اختلاف کمی بین رویکرد تحلیل GEE و Random Effect وجود دارد. و به طور شفاف، برتری آنها بر یکدیگر پیشنهاد نشده است. اگر میان خوشه ها تغییرات باشد تفاوت بین دو روش مذکور بیشتر می شود. روش اثر تصادفی اغلب بوسیله آمارشناسان بیشتر توصیه می شود. بخاطر اینکه تماما مبتنی بر مدل احتمالات است.

## Path Analysis

## تحلیل مسیر

مطالب این قسمت از پایان نامه دفاع شده خانم جمیله معارفی در رشته اپیدمیولوژی به راهنمایی نگارنده این کتاب برگرفته شده است. در پایان نامه مذکور رابطه بین شادکامی، عوامل خطر و بیماری های غیرواگیر در مطالعه کوهورت پرشین شهرکرد از تحلیل مسیر استفاده گردید. متغیرهای دموگرافیک، عوامل خطر بیماری های غیرواگیر، بیماری های غیرواگیر شایع، شادکامی، کیفیت زندگی، سلامت عمومی و استرس با پرسشنامه های معتبر اندازه گیری و جمع آوری شدند.

تحلیل مسیر، تعمیم یافته روش رگرسیون چندمتغیره در رابطه با تدوین مدل های علیتی است و یکی از روش های پیشرفته آماری است که به کمک آن می توان علاوه بر تأثیرات مستقیم، تأثیرات غیرمستقیم و اثر کل هر یک از متغیرهای مستقل بر متغیر وابسته را نیز شناسایی کرد. تحلیل مسیر با ارائه مدلی ساختاری از روابط علیتی احتمالی بین متغیرها، بیان موجهی از همبستگی بین متغیرها نشان می دهد.

## تاریخچه تحلیل مسیر

تحلیل مسیر برای اولین بار طی سال های ۱۹۱۸-۱۹۲۱ از سوی Sewall Wright، جهت مطالعه اثرات مستقیم و غیرمستقیم متغیرهای علت بر متغیرهای معلول و همچنین تجزیه همبستگی بین دو متغیر در یک مدل علی مطرح شد. در حقیقت وی به صورت ناخودآگاه مبانی تحلیل مسیر امروزی را برای حل مسائل وراثت و ژنتیک به کار برد. سپس Henry Nilz به نقد مقاله تحلیل مسیر Sewall Wright پرداخت. چکیده انتقاد وی به این صورت بود که تحلیل مسیر نیازمند یک دانش قبلی در زمینه وجود رابطه علی بین متغیرهاست و آمار با آزمون کردن آن تنها تأیید یا رد کننده این روابط است. در حقیقت Nilz معتقد بود که روش تحلیل مسیری که Wright معرفی کرده به خودی خود قادر به کشف علیت بین متغیرها نیست بلکه تنها می تواند یک مدل علی مبتنی بر دانش قبلی و نظری را آزمون کند. اولین کاربرد تحلیل مسیر پس از Wright توسط Burkese در سال ۱۹۲۸ در یک مطالعه به بررسی تأثیر هوش والدین و محیط بر هوش کودکان انجام شد. تحلیل مسیر در اواسط قرن بیستم به ویژه بیست سال پس از Burkese مورد توجه جامعه شناسانی مانند Gold, Herbert Blak, Dancan, Berger و اقتصاددانان و اپیدمیولوژیست هایی چون Rao (در سال ۱۹۶۴) قرار گرفت. با ترکیب تحلیل مسیر و تحلیل عاملی در دهه ۱۹۷۰، مدل یابی معادلات ساختاری به عنوان ابزار قوی در تحلیل های علت و معلولی پدید آمد. استفاده از SEM در سال های اخیر به ویژه در علوم اجتماعی، آموزش و پرورش، تجارت و علوم زیستی افزایش یافته است اما تاکنون استفاده از این روش و بحث در مورد آن در پژوهش های پزشکی و اپیدمیولوژی محدود بوده است.

## مفروضات مدل تحلیل مسیر

- ✓ ارتباط بین متغیرهای مستقل با وابسته، خطی و جمع پذیر است.
- ✓ باقیمانده ها با یکدیگر و با هیچ یک از متغیرهای موجود در مدل همبستگی ندارند.
- ✓ در مدل، مسیرهای ممکن به صورت پیکان های یک طرفه هستند.
- ✓ متغیرهای مستقل بدون خطا اندازه گیری می شوند.
- ✓ هم خطی چندگانه بین متغیرها اندک است (هم خطی چندگانه زمانی بروز می یابد که بین حداقل

دو متغیر مستقل همبستگی بالایی (بیشتر از ۰/۸) وجود داشته باشد که باعث عدم دقت و بی ثباتی در برآورد ضرایب رگرسیون می شود).

- ✓ حجم نمونه به اندازه کافی باشد. Kline در سال ۱۹۸۹ در مقاله‌ای توصیه می کند که به ازای هر پارامتر (نه متغیر) در مدل، حداقل ۱۰ مورد به حجم نمونه اضافه شود. در نظر گرفتن نسبت ۲۰ نمونه برای هر پارامتر بسیار مطلوب است.
- ✓ متغیرهای مدل پیوسته (فاصله‌ای یا نسبتی) باشند.

### انواع متغیرها

- متغیر آشکار یا مشاهده شده
- متغیر مشاهده شده (گویه) متغیری است که مستقیماً قابل مشاهده و اندازه گیری است.
- متغیر پنهان
- بسیاری از مفاهیم و معیارهای موجود در علوم اجتماعی، پزشکی و اپیدمیولوژی به طور واضح و مستقیم قابل رؤیت و اندازه گیری نیستند. نگرش‌ها، استعداد، میزان آگاهی و بعضی از بیماری‌های روان پزشکی مانند استرس، افسردگی، اسکیزوفرنی و... از این مقوله هستند که در ادبیات آماری به متغیرهای پنهان (مکنون-سازه) موسوم هستند. چنین مفاهیمی را فقط می توان به طور غیرمستقیم و به وسیله متغیرهای مشاهده شده (از طریق پرسشنامه) که بیانگر آن متغیر پنهان هستند ارزیابی نمود.

### مولفه های مدل تحلیل مسیر

در تحلیل مسیر یک متغیر در مدل علی ممکن است نسبت به بعضی از متغیرها مستقل و نسبت به بعضی دیگر وابسته باشد. برای جلوگیری از ابهام در مورد تعیین نوع متغیرها به جای مستقل و وابسته در روش تحلیل مسیر از دو اصطلاح به نام های «متغیر درون‌زا<sup>۲</sup>» و «متغیر برون‌زا<sup>۳</sup>» استفاده می شود.

### متغیر درون‌زا

- متغیرهایی که هم مستقل اند و هم وابسته در واقع یکجا علت اند و یکجا معلول.
- برای هر متغیر درون‌زا در SEM یک معادله ساختاری وجود دارد.
- این متغیرها حداقل یک مسیر را از متغیر دیگر دریافت می کنند.

### متغیر مستقل یا برون‌زا

- همیشه مستقل اند و به عنوان علت در نظر گرفته می شوند.
- متغیرهایی هستند که در سمت راست معادلات ساختاری قرار می گیرند.
- این متغیرها حداقل یک مسیر به متغیر دیگر وارد می کنند (متغیر مستقل).
- این متغیرها ممکن است با یکدیگر همبستگی داشته باشند که از طریق پیکان‌های دوطرفه به هم متصل می شوند.

<sup>۲</sup> endogenous  
<sup>۳</sup> exogenous

## جمله خطا<sup>۴</sup>

متغیرهای تحلیل مسیر همواره می توانند تنها بخشی از واریانس متغیر وابسته را تبیین کنند به همین دلیل آنچه به عنوان اثر یا عوامل ناشناخته باقی می ماند به وسیله  $e$  که به کمیت خطا معروف است، نشان داده می شود.

## نمودار مسیر

روش تحلیل مسیر بر پایه مجموعه ای از تحلیل های رگرسیون چندگانه و بر اساس فرض ارتباط بین متغیرهای مستقل و وابسته استوار است. این روش بر استفاده ابتکاری از نمودار تصویری که به نمودار مسیر معروف است تأکید زیادی دارد. به طور کلی یکی از ساده ترین روش های برقراری ارتباط با یک مدل معادلات ساختاری ترسیم نموداری آن با استفاده از نمادهای ترسیمی خاصی است که آن را نمودار مسیر می نامند که پیوندهای علی احتمالی بین متغیرها، از طریق رسم پیکان های جهت دار از سوی متغیرهای علت به سمت متغیرهای معلول مشخص می شود. رایج ترین نمادهای ترسیمی به کار رفته به صورت زیر است:



متغیرهای مشاهده شده در مربع یا مستطیل قرار می گیرند.



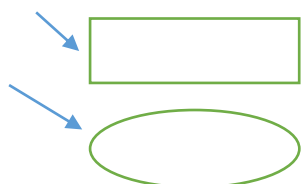
متغیرهای پنهان در دایره یا بیضی قرار می گیرند.

مسیر یک جهت (جهت علیت) متغیر انتهایی پیکان در مدل (متغیر وابسته) توسط متغیر ابتدای پیکان (متغیر مستقل) تبیین می شود.



ارتباط دو جهت همبستگی بین متغیرهای مستقل را نشان می دهد.

هر متغیر وابسته (مشاهده شده و یا پنهان) یک جمله خطا دارد. جملات خطا در واقع بیانگر وجود متغیرهایی است که در مدل لحاظ نشده اند.



خطای اندازه گیری در متغیرهای مشاهده شده

باقیمانده یا خطا در متغیر پنهان

## ضریب مسیر و برآورد آنها

$\beta_{ij}$  (ضریب مسیر) عبارت است از اثر مستقیم  $j$  امین متغیر علت بر  $i$  امین متغیر معلول که مقدار تغییر در متغیر وابسته به ازاء یک واحد تغییر در متغیر مستقل نشان می دهد. در نمودار مسیر می توان رابطه متغیرهای علت و معلول را با استفاده از معادلات رگرسیونی نشان داد. اگر تمام متغیرهای مدل به متغیرهای استاندارد شده تبدیل شوند و برای هر متغیر معلول یک معادله رگرسیونی بر علت هایش نوشته شود آنگاه ضرایب مسیر همان ضرایب رگرسیونی خواهند بود (در این حالت معادلات رگرسیونی بدون عرض از مبدأ هستند).

اگر رابطه رگرسیونی چندگانه به صورت زیر باشد:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_rX_r + e$$

<sup>۴</sup> Error term

$b_i$  ها ضرایب رگرسیونی هستند.

برای مقادیر استاندارد شده  $X_i$  ها و  $Y$ ، ثابت می شود ضریب مسیر از متغیر علت  $X_i$  به متغیر معلول  $Y$  ضریب رگرسیونی استاندارد شده است. پس

$$P_{YX_i} = b_i \frac{S_{X_i}}{S_Y}$$

$S_{X_i}$  انحراف استاندارد متغیر  $X_i$  و  $S_Y$  انحراف استاندارد متغیر  $Y$ .

### ضرایب مسیر و تجزیه همبستگی

اصل اساسی موردنظر در تحلیل مسیر این است که هر ضریب همبستگی بین دو متغیر را می توان به مجموعه ای از مسیرها تجزیه کرد. در واقع یکی از مزایای بارز تحلیل مسیر، تجزیه همبستگی است که آثار مستقیم و غیرمستقیم یک متغیر را بر روی متغیر دیگر اندازه گرفته و باهم مقایسه می کند. بنابراین ضریب همبستگی بین دو متغیر علت و معلول به تعداد کل مسیرهایی که از سوی متغیر علت به سمت معلول وجود دارد، تجزیه می شود. بنابراین پس از محاسبه اثرات مستقیم یک متغیر بر متغیر دیگر، می توان مجموع اثرات غیرمستقیم متغیر فوق بر متغیر وابسته را از طریق کم کردن ضریب بتای استاندارد شده از ضریب همبستگی بدست آورد. این عمل بر مبنای تجزیه همبستگی امکان پذیر است. تحلیل مسیر برای تبیین دقیق تر روابط علی متغیرها به تجزیه همبستگی بین متغیرها می پردازد. آنچه ضرایب همبستگی نشان می دهند اثرات مستقیم و غیرمستقیم متغیرهاست به عبارت دیگر:

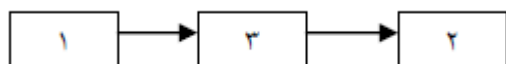
$$\text{اثرات غیرمستقیم} + \text{اثرات مستقیم} = \text{همبستگی}$$

اثرات مستقیم، همان ضرایب بتا برای مسیرهای مختلف است که با مقایسه ارقام این ضرایب توسط محقق اهمیت هر مسیر مشخص می شود. برای محاسبه اثرات غیرمستقیم، ضریب بتای هر مسیر تا رسیدن به متغیر وابسته در هم ضرب می شود و چون هر متغیر ممکن است از طریق مسیرهای متعدد بر متغیر وابسته اثر بگذارد، بنابراین برای محاسبه اثرات غیرمستقیم کل یک متغیر بر متغیر دیگر، اثرات مسیرهای غیرمستقیم آن دو متغیر با همدیگر جمع می شود.

در تحلیل مسیر این امکان وجود دارد که ضریب همبستگی بین متغیرها را به چهار اثر تجزیه کرد: اثر مستقیم: تأثیر مستقیم یک متغیر بر متغیر دیگر بدون حضور هیچ متغیر دیگری در بین آنها (همان ضریب مسیر است).

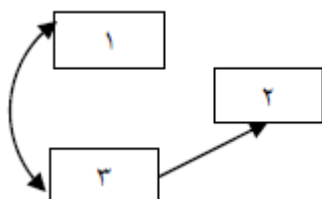


اثر غیرمستقیم: ارتباط دو متغیر با یکدیگر از طریق یک متغیر مداخله گر سوم.

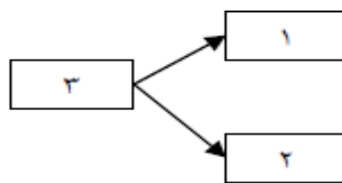


اثر علی کل یک متغیر بر متغیر دیگر: جمع اثر مستقیم با مجموع آثار غیرمستقیم آن که آثار غیرمستقیم از طریق حاصل ضرب ضرایب هر مسیر محاسبه می شود.

اثر تحلیل نشده: به دلیل همبستگی بین دو متغیر برون زای ۱ و ۳ و ارتباط متغیر ۳ با متغیر ۲، تبیین تغییرپذیری بین ۱ و ۲ توسط مدل امکان پذیر نیست به این نوع اثر، اثر تحلیل نشده می گویند.



اثر کاذب: رابطه بین ۱ و ۲ وقتی کاذب<sup>۵</sup> است که متغیر ۳ علت هر دو متغیر ۱ و ۲ باشد.



اثر غیر علی: مجموع اثر کاذب و تحلیل نشده.

ضریب همبستگی: اثر علی + اثر غیر علی.

در تحقیقات اجتماعی معمولاً شناسایی کلیه عوامل تأثیرگذار بر متغیر وابسته برای محقق امکان پذیر نیست بنابراین متغیرهای تحلیل مسیر همواره می توانند تنها بخشی از واریانس متغیر وابسته را تبیین کنند به همین دلیل آنچه به عنوان اثر یا عوامل ناشناخته باقی می ماند به وسیله  $e$  که به کمیت خطا معروف است، نشان داده می شود. مقدار  $e$  بیانگر میزان واریانس متغیر وابسته است که متغیرهای مستقل قبلی در مدل قادر به تبیین آن نبودند. پس  $e$  واریانس تبیین نشده متغیر وابسته است. چون  $e$  مقدار واریانس خطا نام دارد آن را  $e^2$  نیز می نامند در صورتی که مقدار  $e^2$  از عدد ۱ کم شود مقدار واریانس تبیین شده (که آن را با  $R^2$  نشان می دهند) به این صورت بدست می آید:

$$R^2 = 1 - e^2$$

در تحلیل مسیر از ضریب تعیین ( $R^2$ ) بدست آمده از طریق رگرسیون، مناسب بودن مدل بدست می آید. این فن از ضرایب رگرسیون استاندارد شده جزئی به عنوان ضرایب مسیر استفاده کرده و اثرات مستقیم هر متغیر را تعیین می کند. به عبارت دیگر این ضریب سهم یا وزن متغیر مستقل را در تبیین واریانس متغیر وابسته نشان می دهد. طبیعی است که هرچه مقدار  $e^2$  کم باشد در نتیجه مقدار  $R^2$  زیاد است و مدل از برازندگی بیشتری برخوردار است بنابراین واریانس متغیرهای وابسته توسط متغیرهای مستقل موجود در مدل تبیین شده است و به ورود متغیرهای مستقل بیشتری در مدل نیاز نیست.

## روش کلی کار در الگوی معادلات ساختاری

## مشخص کردن الگویی بر پایه یک نظریه

الگو یا مدل یک عبارت آماری درباره روابط بین متغیرهاست. ممکن است یک نمودار مسیر ترسیم شود و روابط علی بین متغیرها نشان داده شود و بعد از مشخص شدن متغیرهای پنهان، متغیرهای مشاهده شده مناسبی را انتخاب و به آنها وصل شود.

## ارائه تخمین برای مدل پیشنهادی

در هر مدل معادلات ساختاری پارامترهای نامعلوم طوری برآورد می‌شوند که مدل قابلیت شبیه‌سازی از ماتریس همبستگی یا کواریانس نمونه مورد تحلیل را داشته باشد. با مقایسه دو ماتریس کواریانس حاصل از مدل و ماتریس کواریانس حاصل از نمونه، یک ماتریس باقیمانده تولید می‌شود. هدف این است که مدلی ارائه شود تا ماتریس باقیمانده به حداقل ممکن برسد.

## ارزیابی تناسب یا برازش مدل

هنگامی که ماتریس کواریانس حاصل از مدل با ماتریس کواریانس داده‌های مشاهده شده (نمونه) معادل باشند، ماتریس باقیمانده نزدیک به صفر می‌شود و مدل با داده‌های مشاهده شده تناسب دارد. این تناسب را از طریق معیارهای برازندگی مدل از قبیل آزمون خی دو و تقریب جذر میانگین مربعات خطا (RMSEA) و ... که در ادامه توضیح داده می‌شوند سنجیده می‌شوند.

## اصلاح مدل

اگر ماتریس واریانس-کواریانس برآورد شده از مدل، نتواند به خوبی ماتریس واریانس-کواریانس نمونه‌ای را برآورد کند (شاخص‌های نیکویی برازش نامناسب باشند)، مدل باید تصحیح گردد. تصحیح مدل با وارد کردن متغیرهای جدید، حذف متغیرهای فعلی و تغییر در فرضیات مدل انجام می‌شود. خلاصه شاخص‌های برازندگی مدل در معادلات ساختار یافته در جدول زیر گزارش شد.

آزمون	توضیحات و معادل فارسی	مدل ایده ال	مدل قابل قبول
RMSEA	ریشه میانگین مجذورات خطای تقریب Root Mean Square Error of Approximation	بین صفر تا ۰/۰۵	بین ۰/۰۵ تا ۰/۰۸
SRMR	ریشه میانگین مجذور باقیمانده استاندارد شده Root Mean Square Residual Standardized	بین صفر تا ۰/۰۵	بین ۰/۰۵ تا ۰/۰۸
CFI	برازندگی تطبیقی Comparative Fit Index	بین ۰/۹۵ تا ۱	بین ۰/۹ تا ۰/۹۵
AIC	ملاک اطلاعات Akaike	هرچه کوچک‌تر برازش بهتر	هرچه کوچک‌تر برازش بهتر

از آنجایی که موضوع برازش مدل مفهومی و شاخص‌هایی که به بهترین نحو توانایی تفسیر برازش مدل را داشته باشند بسیار متنوع و پیچیده است و پیوسته در حال توسعه و تکامل می‌باشند، هنوز درباره حتی یک آزمون بهینه نیز توافق همگانی وجود ندارد. به طور کلی چندین شاخص برای سنجش برازش مدل مورد استفاده قرار می‌گیرند که برای تأیید مدل استفاده از سه تا پنج شاخص کافی است.

## شاخص‌های برازش مطلق

شاخص‌های اندازه‌گیری مطلق از اصلی‌ترین معیارها برای صحت و سقم برازش داده‌ها بر اساس مدل یا فرضیات پیشنهاد شده می‌باشند که بر پایه تفاوت میان ماتریس واریانس-کواریانس مشاهده شده و تدوین شده توسط مدل، اندازه‌گیری می‌شوند.



شاخص هایی که در این دسته بندی قرار می گیرند عبارت اند از:

شاخص<sup>۶</sup> RMR (ریشه میانگین مجذور باقیمانده) و<sup>۷</sup> SRMR (ریشه میانگین مجذور باقیمانده استاندارد شده)

ریشه میانگین مجذور باقیمانده شاخصی برای واریانس باقیمانده در برازش هر پارامتر به داده های نمونه و یا به عنوان میانگین باقیمانده های گزارش شده در مدل، بیان می شود که فرمول آن به صورت زیر است:

$$RMR = \sqrt{\left[ \frac{1}{K} \sum_{ij} (S_{ij} - \sigma_{ij})^2 \right]}$$

$$SRMR = \sqrt{\left\{ 2 \sum_{ij} [(S_{ij} - \sigma_{ij}) / S_{ii} S_{jj}]^2 \right\} / p(p + 1)}$$

K تعداد مقادیر مجزای ماتریس S است و  $S_{ij}$  عناصر ماتریس کواریانس نمونه و  $\sigma_{ij}$  عناصر ماتریس کواریانس به دست آمده از مدل و p تعداد متغیرهای مشاهده شده است. دامنه RMR بر طبق وزن هر متغیر مشاهده پذیر محاسبه شده و برای پرسشنامه ای حاوی گویه هایی با طیف های متفاوت (لیکرت ۵ تایی و ۷ تایی) استناد به این شاخص اعتبار چندانی ندارد ولی این مشکل با SRMR برطرف شده و تفسیر در این گونه موارد بر اساس شاخص SRMR اعتبار بالایی دارد. هر چه قدر این دو معیار کوچک تر باشند یعنی به صفر نزدیک تر باشند، مدل از نیکویی برازش خوبی برخوردار است برای مدل هایی با برازش خوب زیر ۰/۰۵ گزارش شده است و تا ۰/۰۸ هم مدل قابل قبول است.

معیار<sup>۸</sup> RMSEA (ریشه میانگین مجذورات تقریب)

ریشه میانگین مجذورات تقریب همان آزمون انحراف هر درجه آزادی است. این شاخص یکی از اصلی ترین شاخص های برازش مدل در تحلیل خروجی نرم افزار لیزرل بوده که میزان نامناسب بودن مدل برازش شده را در مقابل مدل اشباع شده برآورد می کند فرمول آن چنین است:

$$RMSEA = \sqrt{[(\chi^2/df) - 1 / (N - 1)]}$$

N حجم نمونه است. اگر  $\chi^2$  کوچک تر از درجه آزادی (df) باشد، RMSEA برابر با صفر قرار داده می شود (۱۱۶). اگر این مقدار کمتر از ۰/۰۵ باشد نشان دهنده برازندگی خوب برای مدل است و از ۰/۰۵ تا ۰/۰۸ و در بعضی منابع تا ۰/۱ نشان دهنده خطای معقول برای تقریب در جامعه است یعنی مدل قابل قبول است و اگر RMSEA بیشتر از ۰/۱ باشد مدل برازش ضعیفی دارد.

1. Root Mean Square Residual
2. Root Mean Square Residuals Standardized
- Root Mean Square Error of Approximation<sup>۸</sup>

### شاخص برازندگی تطبیقی CFI<sup>۹</sup>

این شاخص از طریق مقایسه یک مدل مستقل که در آن هیچ رابطه‌ای بین متغیرها وجود ندارد با مدل پیشنهادی موردنظر، به آزمودن مدل پیشنهادی می‌پردازد. این شاخص از نظر معنا مانند NFI است ولی برای حجم نمونه جریمه می‌دهد یعنی به ازای هر پارامتر برآورد شده یک جریمه معادل یک را می‌پردازد. اگر  $d = \chi^2 - df$  باشد، شاخص برازندگی تطبیقی به این صورت خواهد بود:

$$CFI = \frac{d \text{ مدل پیشنهادی} - d \text{ مدل (صفر) مستقل}}{d \text{ مدل (صفر) مستقل}}$$

این شاخص نیز همانند شاخص‌های قبل تفسیر می‌شود. مقادیر بالای ۰/۰۹ قابل قبول و مقادیر بالای ۰/۹۵ نشان‌دهنده برازش خوب مدل است.

### ملاک اطلاعات AIC<sup>۱۰</sup>

هرچه این ملاک عدد کوچک‌تری داشته باشد برازش بهتری را نشان می‌دهد. اندازه آن استاندارد نیست و برای یک مدل معین تفسیر نمی‌شود. برای دو مدل که از یک مجموعه داده بدست آمده، مدل با AIC کمتر ترجیح داده می‌شود فرمول آن به صورت زیر است:

$$AIC = \chi^2 + k(k - 1) - 2df$$

در این فرمول  $k$  تعداد متغیرهای موجود در مدل،  $df$  درجه آزادی مدل و مقدار  $k(k - 1) - 2df$  معرف تعداد پارامترهای آزاد موجود در مدل است قدر مطلق آن بی‌معنا و اندازه نسبی آن مهم است. به‌طور کلی این آزمون‌ها به این سؤال پاسخ می‌دهند که چه میزان مدل موردنظر برازنده داده‌های پژوهش است.

### طرز کار با نرم افزار stata برای مدلسازی معادلات ساختاری / تحلیل مسیر

از منوی statistics در شکل زیر SEM را انتخاب و مسیر ابتدایی پیش رو را اجرا نمایید.

The screenshot shows the Stata software interface. The 'Statistics' menu is open, listing various statistical models. The command window displays the following statistics:

```

Number of obs =      69
LR chi2(2)      =    14.36
Prob > chi2     =    0.0008
Pseudo R2      =    0.0766
    
```

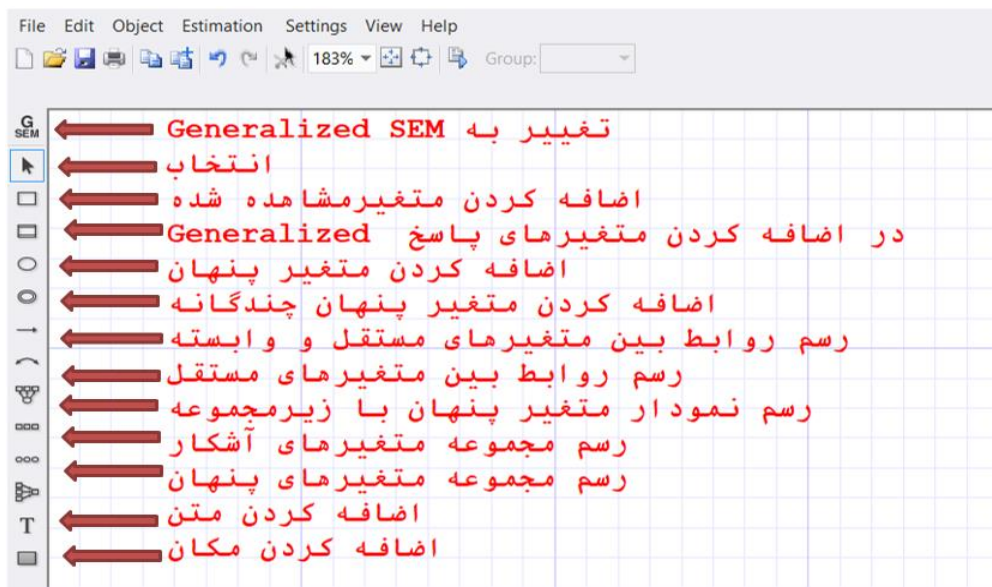
Below this, a table of parameter estimates is shown:

rr.	z	P> z	[95% conf. interval]	
16	1.64	0.100	-.0225224	.2566611
49	-0.68	0.498	-.001279	.0006217
17			-8.019659	3.419148
			131551	5.046868

<sup>۹</sup> Comparative Fit Index  
<sup>۱۰</sup> Akaike

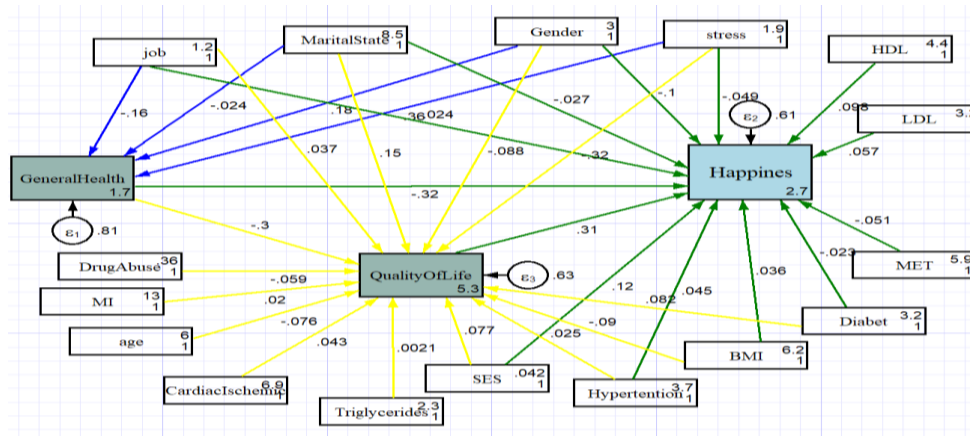
برای انجام این کار، پس از وارد کردن فایل داده‌ها به نرم‌افزار، از گزینه Statistic: SEM(structural equation modeling) در نوار ابزار سمت چپ شکل زیر (Model building and estimation): Model building and estimation وارد محیط رسم مدل می‌شویم.

از مسیر بالا SEM (structural equation modeling) را انتخاب کنید. شکل زیر ظاهر می‌گردد. در این صفحه باید مسیر متغیرها را تعریف و ترسیم کرد. در نوار ابزار سمت چپ شکل زیر ابزارهای ترسیم مسیرها را باید انتخاب و مسیرها را رسم نمود. کارکرد هر یک از ابزارها در شکل زیر توضیح داده شد.



### رسم مدل با استفاده از تحلیل مسیر

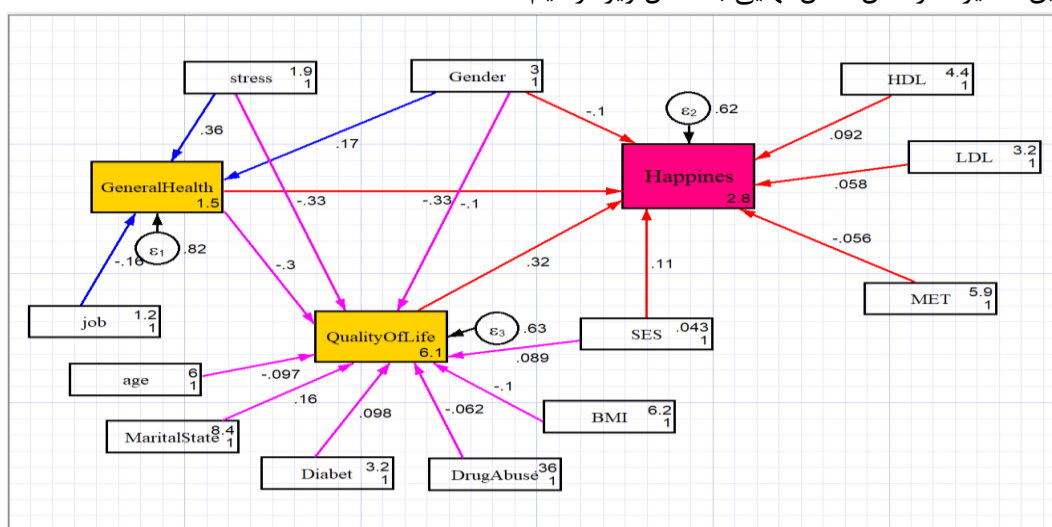
با رعایت مفروضات تحلیل مسیر با انتخاب نمونه‌ای با حجم بالا، وارد کردن متغیرهای کمی و کیفی دوحالته و بررسی عدم همبستگی بین متغیرهای مستقل مدل اولیه برای عوامل موثر و مرتبط با شادکامی طراحی شد. با انجام آنالیز اولیه، متغیرهایی که مقدار احتمال آنها کمتر از ۰/۲ بود وارد مدل فرضی پیشنهادی شدند. در این صفحه با استفاده از ابزارهایی که در سمت چپ صفحه قرار دارند می‌توان متغیر آشکار (□)، متغیر پنهان (○)، روابط یک طرفه (→) و روابط دوطرفه (↔) و... را رسم نمود. مدل پیشنهادی اولیه را در شکل زیر می‌توانید ببینید.



پس از رسم مدل پیشنهادی بایستی روابط بین متغیرها در مدل سنجیده شود. برای این کار در صفحه SEM Builder از مسیر Estimation: estimate اقدام می‌کنیم. پنجره‌ای که باز می‌شود از Tab Model به ما امکان محاسبه likelihood با یا بدون داده‌های missing را می‌دهد. از Tab Reporting نیز با انتخاب گزینه Display standardized coefficient and values می‌توان ضرایب مسیر را محاسبه کرد.

پس از طی این مسیر، نرم افزار در مدل ضرایب مسیر و در صفحه اول stata، ضریب مسیر (coef)، انحراف استاندارد (SE)، مقدار احتمال ( $P > |Z|$ ) و دامنه اطمینان (95% conf. interval) هر مسیر منتهی به یک متغیر وابسته میانی یا نهایی را نشان می‌دهد.

تجزیه و تحلیل مسیر یک رویکرد تست مدل سازی است، نه یک مدل سازنده. در حالت ایده آل، اگر یک متغیر خارجی گنجانده شده باشد که ضریب مسیر آن کم و غیر قابل توجه باشد، نشان می‌دهد که مدل باید بدون آن متغیر تجدید نظر شود. در این مدل متغیرهایی که مقدار احتمال آنها بیشتر از ۰/۲ بود از مدل حذف شده و بنابراین مسیر آنها نیز از مدل حذف شدند. بدین ترتیب با حذف متغیرهای شغل، فشارخون، وضعیت تاهل، استرس، دیابت و BMI که مسیر آنها به متغیر شادکامی وصل می‌شد، متغیرهای شغل، فشارخون، تری‌گلیسرید، سکت قلبی و سابقه بیماری قلبی عروقی که مسیر آنها به متغیر کیفیت زندگی می‌رسید و متغیر وضعیت تاهل که مسیر آن به متغیر سلامت عمومی وصل می‌شد، مدل دوم طراحی شد. در مدل دوم مسیر متغیر فشارخون به شادکامی و مسیر متغیر شغل که به کیفیت زندگی منتهی می‌شدند مقدار احتمال قابل قبولی نداشتند. در نهایت با حذف این متغیرها از مدل، مدل نهایی به شکل زیر ترسیم شد.

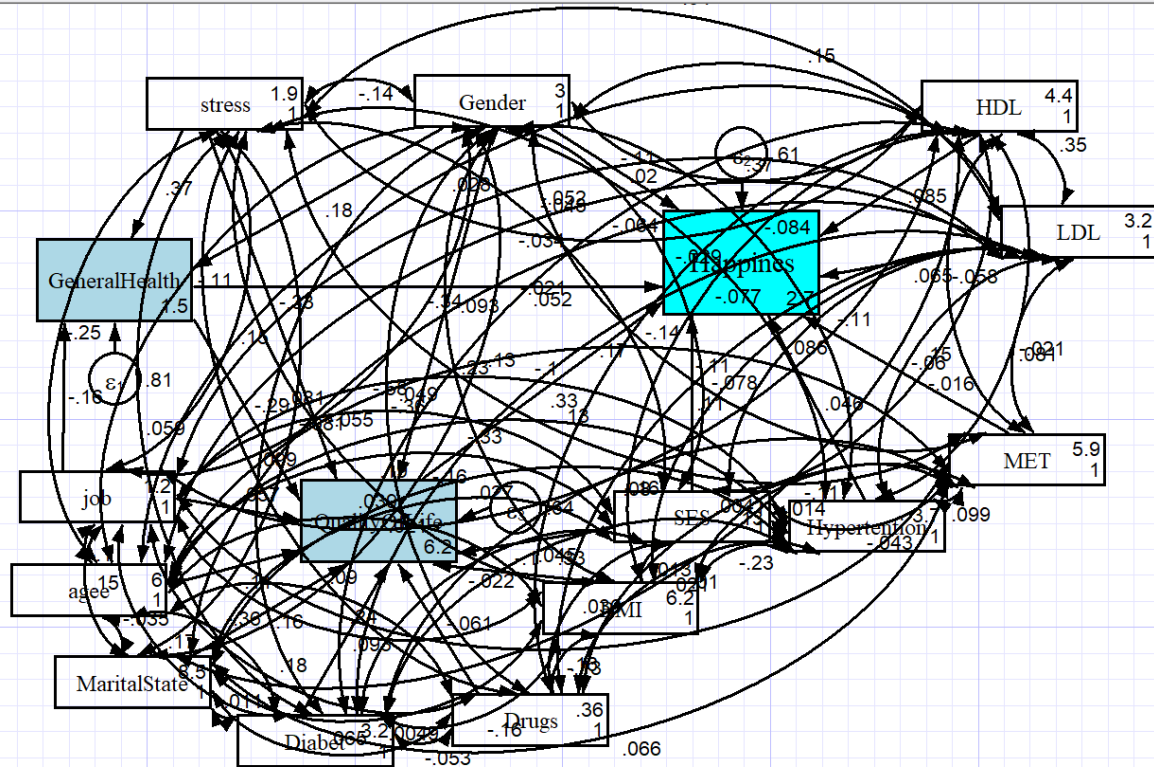


همانطور که قابل مشاهده است مدل نهایی نسبت به مدل پیشنهادی اول واضح‌تر و قابل فهم‌تر می‌باشد. در اینجا نیز در ابتدا وضعیت ضرایب مسیر را در مدل بررسی می‌کنیم. مطابق با خروجی جدول ضرایب مسیر مربوط به مدل نهایی، تمامی ضرایب مقدار احتمال آماری قابل قبولی دارند.

همانطور که در شکل بالا مشاهده می‌شود در این مدل فقط روابط بین متغیرهای مستقل و وابسته نمایش داده شده است در حالیکه ممکن است بین متغیرهای مستقل نیز روابطی وجود داشته باشد که بر روابط بین متغیرهای مستقل و وابسته تاثیرگذار باشد. برای ترسیم روابط بین متغیرهای مستقل، گزینه estimate در پنجره SEM را انتخاب می‌کنیم و از tab SE/Robust گزینه Sattora-Bentler estimator را انتخاب می‌کنیم.

با انتخاب این گزینه تمام روابط بین متغیرهای مستقل در مدل ترسیم شده (شکل زیر) و جدول ضرایب مسیر نیز در صفحه نتایج قابل مشاهده خواهد بود. همانگونه که در شکل زیر دیده می شود مدل با رسم تمام روابط بین متغیرهای مستقل نامفهوم خواهد شد. از طرفی ضرایب مسیر و مقدار احتمال متغیرهای مستقل در جدول نتایج تغییری نسبت به مدل قبل نداشتند که نشان دهنده عدم تاثیر روابط بین متغیرهای مستقل در این مدل می باشد. در نتیجه با این تفاسیر شکل قبلی به عنوان مدل انتخاب شد.

**شکل مدل دوم همراه با نمایش رابطه بین متغیرهای مستقل**



با به توان دو رساندن ضرایب مسیر جدول خروجی، مقدار  $R^2$  به دست می آید که همان ضریب تعیین اصلاح شده می باشد. بدین ترتیب، مطابق نتایج ضرایب مسیر، ۹٪ تغییرات شادکامی در این مدل توسط متغیر سلامت عمومی و ۹٪ با کیفیت زندگی تبیین می شود. ۹ درصد از تغییرات سلامت عمومی تحت تاثیر استرس قرار دارد. برای متغیر کیفیت زندگی ۴٪ از تغییرات آن با سلامت عمومی و ۹٪ با استرس تبیین می شود. با جمع مقادیر  $R^2$  مقدار کل  $R^2$  مدل برابر ۰/۴۹ خواهد بود که نشان از دقت قابل قبول مدل می باشد.

در این مرحله و با تایید ضرایب مسیر، بایستی برازندگی مدل را آزمون کرد. برای آزمون برازندگی مدل در تحلیل مسیر در Stata از دستور `estat gof, stats (all)` استفاده می شود. با این دستور تمامی آزمون های برازندگی مدل انجام می شود.

برای تایید برازندگی یک مدل در تحلیل مسیر، حداقل بایستی سه آزمون را انجام داد. آزمون های RMSEA، CFI و SRMR نتایجی در محدوده ایده آل برای این مدل داشتند. با توجه به شاخص های برازش الگوی تحلیل مسیر، مشاهده می شود که مقدار شاخص  $CFI=0/98$  است که چون بین ۰/۹۵ تا ۱ است مقدار ایده آلی است و با

توجه به شاخص  $SRMR=0/017$  و شاخص  $RMSEA=0/031$  است (مقادیر کمتر از  $0/05$  نشان دهنده ایده آل بودن مدل است) مدل پیشنهادی از برازش ایده آلی برخوردار است.

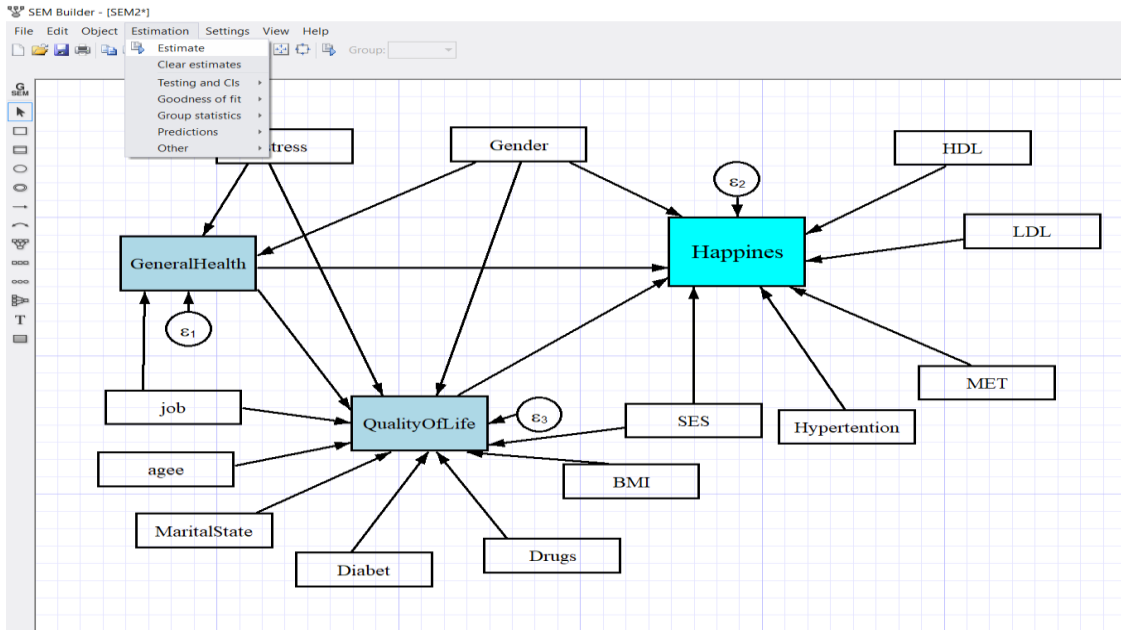
جدول نتایج شاخص های برازندگی مدل نهایی

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(20)	32.471	model vs. saturated
p > chi2	0.039	
chi2_bs(39)	752.592	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.032	Root mean squared error of approximation
90% CI, lower bound	0.007	
upper bound	0.051	
pclose	0.938	Probability RMSEA <= 0.05
Information criteria		
AIC	46288.337	Akaike's information criterion
BIC	46398.837	Bayesian information criterion
Baseline comparison		
CFI	0.983	Comparative fit index
TLI	0.966	Tucker-Lewis index
Size of residuals		
SRMR	0.018	Standardized root mean squared residual
CD	0.394	Coefficient of determination

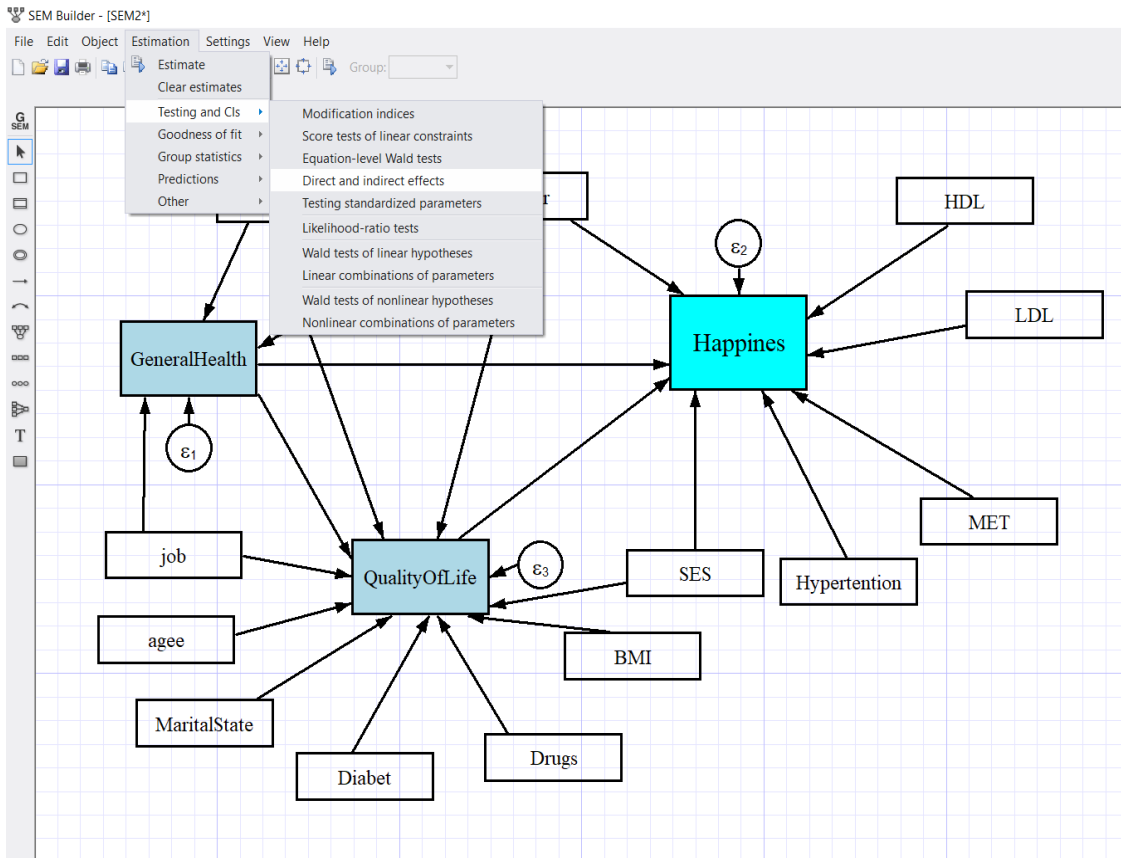
باتوجه به نتایج ضرایب مسیر و مطابق با برازندگی ایده آل مدل، این مدل به عنوان مدل نهایی تایید شد. با تایید مدل، مسیرهای مستقیم و غیرمستقیم تاثیر متغیرهای مستقل بر متغیرهای وابسته را محاسبه می کنیم. برای انجام این کار، از گزینه Estimation در پنجره SEM در stata، گزینه Testing and CLs و سپس Direct and indirect effects را انتخاب می کنیم. با انتخاب این گزینه، مسیرهای تاثیر مستقیم، غیرمستقیم و تاثیر کلی هر یک از متغیرهای مستقل بر متغیرهای وابسته در جداول مجزا نمایش داده می شود.

در شکل زیر نحوه محاسبه و تعیین ضرایب مسیر در نرم افزار stata نشان داده می شود.

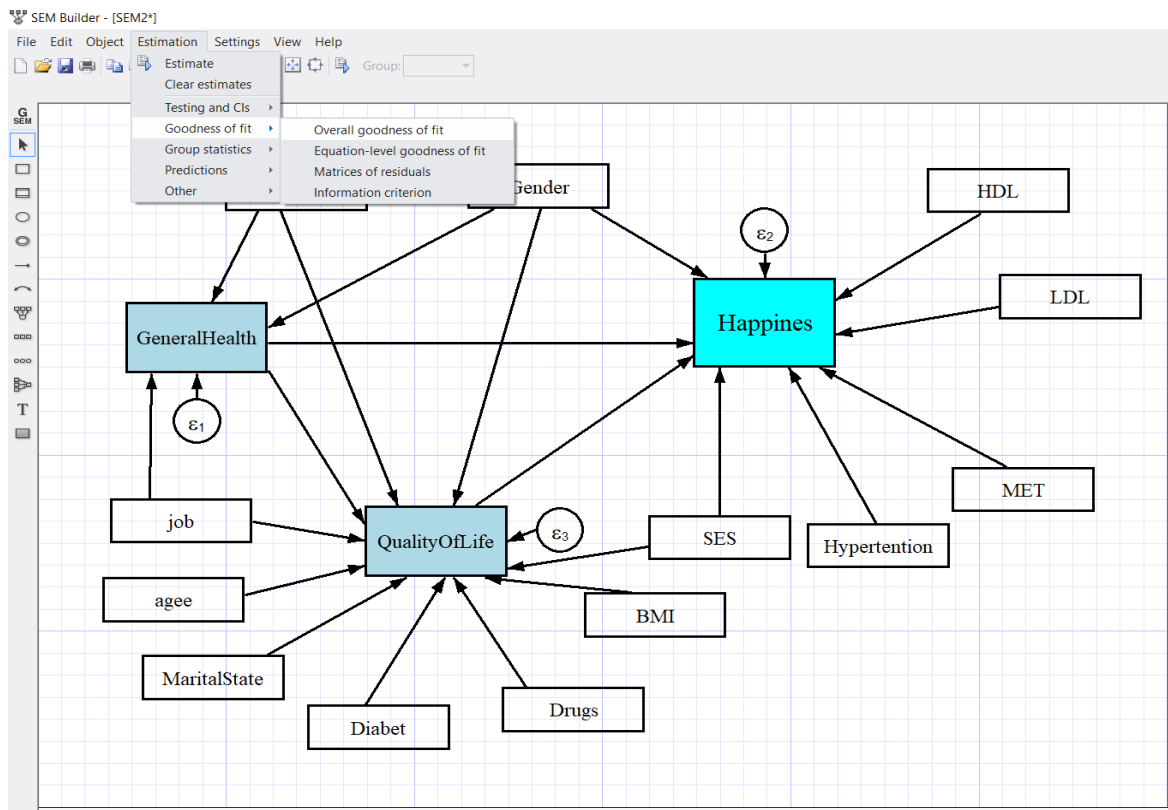
## روش های اپیدمیولوژی و پژوهش با Stata



در شکل زیر مسیر محاسبه و تعیین اثرات مستقیم و غیرمستقیم هر مسیر در نرم افزار آمده است.



در شکل زیر مسیر محاسبه و نحوه تعیین برازندگی مدل در نرم افزار آمده است.



### کار با نرم افزار و انجام تحلیل مسیر

فایل webuse sem\_sm2, clear را باز نمائید. توصیف فایل در زیر آمده است. این فایل با نام SEM ذخیره کنید.  
 . ssd describe

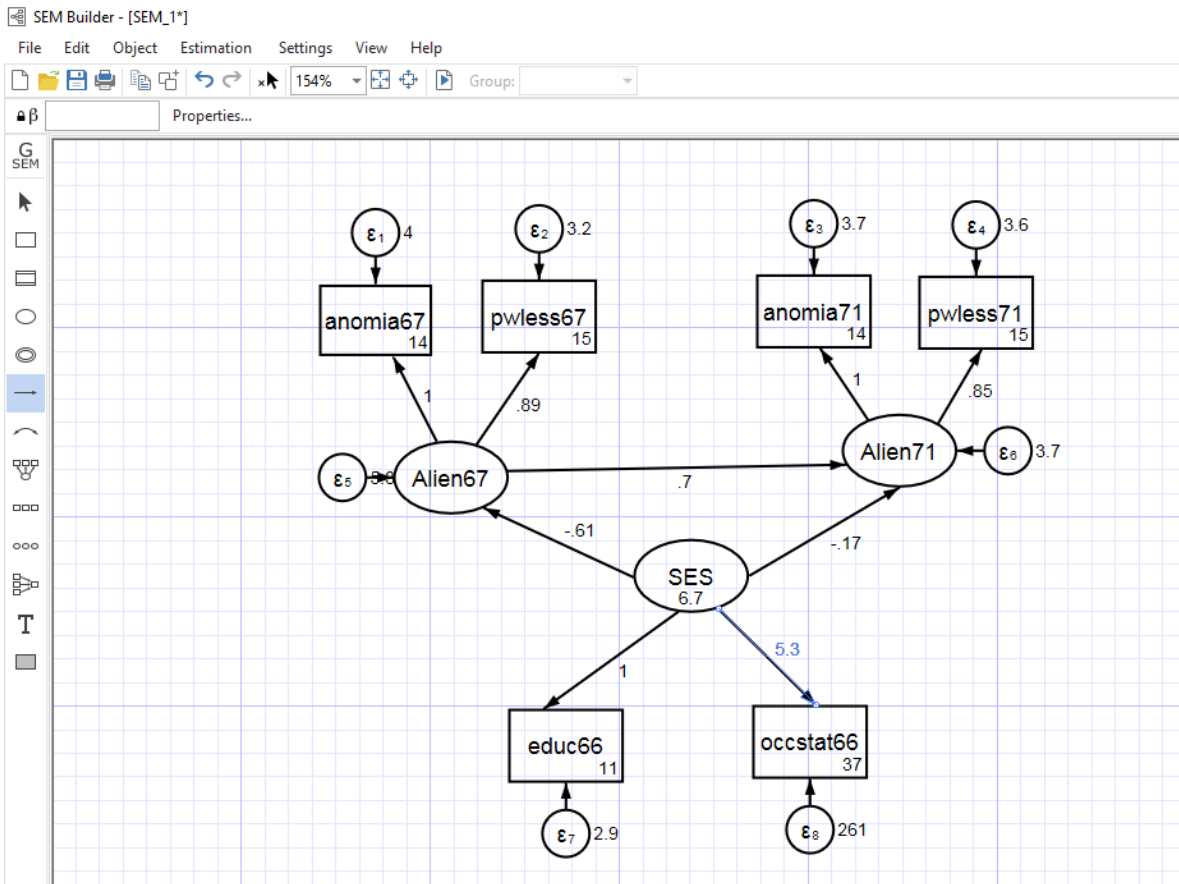
```
Summary statistics data from https://www.stata-press.com/data/r17/sem_sm2.dta
Observations:      932      Structural model with measurem..
Variables:         13      25 May 2020 11:45
                    (_dta has notes)
```

Variable name	Variable label
educ66	Education, 1966
occstat66	Occupational status, 1966
anomia66	Anomia, 1966
pwless66	Powerlessness, 1966
socdist66	Latin American social distance, 1966
occstat67	Occupational status, 1967
anomia67	Anomia, 1967
pwless67	Powerlessness, 1967
socdist67	Latin American social distance, 1967
occstat71	Occupational status, 1971
anomia71	Anomia, 1971
pwless71	Powerlessness, 1971
socdist71	Latin American social distance, 1971

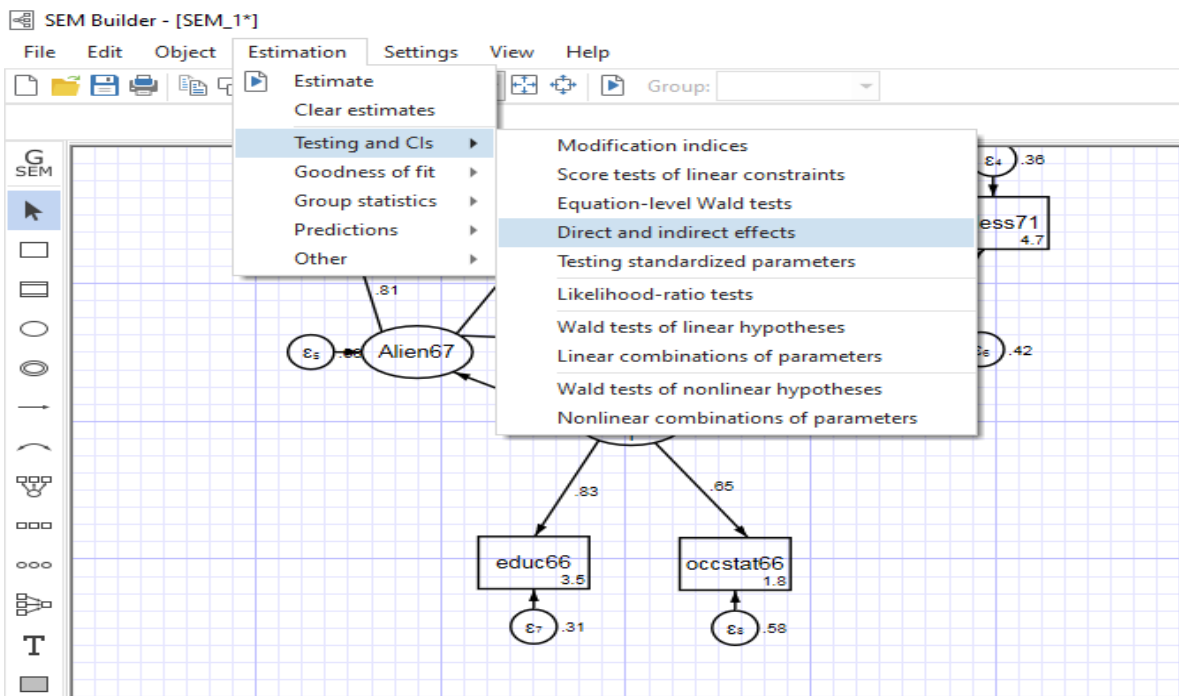
متغیرهای نهفته درون زا: educ66 occstat66 anomia67 pwless67 anomia71 pwless71 و متغیرهای پنهان: Alien67 و Alien71 و متغیرهای برون زا (Exogenous variables) و متغیر پنهان SES است.



دیگرام رابطه بین متغیرها را به صورت شکل زیر ترسیم کنید. این فایل با نام SEM1 ذخیره شده است.



estimation را اجرا و سپس از آنجا CI و testing را انتخاب و سپس اثرات مستقیم و غیرمستقیم را اجرا می کنیم.



## روش های اپیدمیولوژی و پژوهش با Stata

اکنون خروجی نتایج به صورت زیر است:

```
. sem (Alien67 -> anomia67, ) (Alien67 -> pwless67, ) (Alien67 -> Alien71, ) (Alien71 -> anomia71, ) (Alien71
> -> pwless71, ) (SES -> Alien67, ) (SES -> Alien71, ) (SES -> educ66, ) (SES -> occstat66, ), latent(Alien67
> Alien71 SES ) nocapslatent
```

Endogenous variables

```
Measurement: anomia67 pwless67 anomia71 pwless71 educ66 occstat66
Latent: Alien67 Alien71
```

Exogenous variables

```
Latent: SES
```

Fitting target model:

```
Iteration 0: log likelihood = -15249.988
Iteration 1: log likelihood = -15246.584
Iteration 2: log likelihood = -15246.469
Iteration 3: log likelihood = -15246.469
```

Structural equation model

Number of obs = 932

Estimation method: ml

Log likelihood = -15246.469

- ( 1) [anomia67]Alien67 = 1
- ( 2) [anomia71]Alien71 = 1
- ( 3) [educ66]SES = 1

		OIM						
		Coefficient	std. err.	z	P> z	[95% conf. interval]		
Structural Alien67	SES	-.6140404	.0562407	-10.92	0.000	-.7242701	-.5038107	
	<hr/>							
Alien71	Alien67	.7046342	.0533512	13.21	0.000	.6000678	.8092007	
	SES	-.1744153	.0542489	-3.22	0.001	-.2807413	-.0680894	
<hr/>								
Measurement anomia67	Alien67 _cons	1 13.61	(constrained) .1126205	120.85	0.000	13.38927	13.83073	
	<hr/>							
pwless67	Alien67 _cons	.8884887 14.67	.0431565 .1001798	20.59 146.44	0.000 0.000	.8039034 14.47365	.9730739 14.86635	
	<hr/>							
anomia71	Alien71 _cons	1 14.13	(constrained) .1158943	121.92	0.000	13.90285	14.35715	
	<hr/>							
pwless71	Alien71 _cons	.8486022 14.9	.0415205 .1034537	20.44 144.03	0.000 0.000	.7672235 14.69723	.9299808 15.10277	
	<hr/>							
educ66	SES _cons	1 10.9	(constrained) .1014894	107.40	0.000	10.70108	11.09892	
	<hr/>							
occstat66	SES _cons	5.331259 37.49	.4307503 .6947112	12.38 53.96	0.000 0.000	4.487004 36.12839	6.175514 38.85161	
	<hr/>							
var(e.anomia67)		4.009921	.3582978			3.365724	4.777416	
var(e.pwless67)		3.187468	.283374			2.677762	3.794197	
var(e.anomia71)		3.695593	.3911512			3.003245	4.54755	
var(e.pwless71)		3.621531	.3037908			3.072483	4.268693	
var(e.educ66)		2.943819	.5002527			2.109908	4.107319	
var(e.occstat66)		260.63	18.24572			227.2139	298.9605	
var(e.Alien67)		5.301416	.483144			4.434225	6.338201	
var(e.Alien71)		3.737286	.3881546			3.048951	4.581019	
var(SES)		6.65587	.6409484			5.511067	8.038482	

LR test of model vs. saturated: chi2(6) = 71.62

Prob > chi2 = 0.0000

## روش های اپیدمیولوژی و پژوهش با Stata

. estat mindices

Modification indices

	MI	df	P>MI	EPC	Standard EPC
<b>Measurement</b>					
<b>anomia67</b>					
anomia71	51.977	1	0.00	.3906425	.4019984
pwless71	32.517	1	0.00	-.2969297	-.2727609
educ66	5.627	1	0.02	.0935048	.0842631
<b>pwless67</b>					
anomia71	41.618	1	0.00	-.3106995	-.3594367
pwless71	23.622	1	0.00	.2249714	.2323233
educ66	6.441	1	0.01	-.0889042	-.0900664
<b>anomia71</b>					
anomia67	58.768	1	0.00	.429437	.4173061
pwless67	38.142	1	0.00	-.3873066	-.3347904
<b>pwless71</b>					
anomia67	46.188	1	0.00	-.3308484	-.3601641
pwless67	27.760	1	0.00	.2871709	.2780833
<b>educ66</b>					
anomia67	4.415	1	0.04	.1055965	.1171781
pwless67	6.816	1	0.01	-.1469371	-.1450411
<b>cov(e.anomia67,e.anomia71)</b>					
	63.786	1	0.00	1.951578	.5069627
<b>cov(e.anomia67,e.pwless71)</b>					
	49.892	1	0.00	-1.506704	-.3953794
<b>cov(e.anomia67,e.educ66)</b>					
	6.063	1	0.01	.5527612	.1608845
<b>cov(e.pwless67,e.anomia71)</b>					
	49.876	1	0.00	-1.534199	-.4470094
<b>cov(e.pwless67,e.pwless71)</b>					
	37.357	1	0.00	1.159123	.341162
<b>cov(e.pwless67,e.educ66)</b>					
	7.752	1	0.01	-.5557802	-.1814365

## روش های اپیدمیولوژی و پژوهش با Stata

. estat teffects

Direct effects

		OIM				
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Measurement anomia67	Alien67	1	(constrained)			
	SES	0	(no path)			
pwless67	Alien67	.8884887	.0431565	20.59	0.000	.8039034 .9730739
	SES	0	(no path)			
anomia71	Alien67	0	(no path)			
	Alien71	1	(constrained)			
pwless71	Alien67	0	(no path)			
	Alien71	.8486022	.0415205	20.44	0.000	.7672235 .9299808
educ66	SES	1	(constrained)			
	occstat66	5.331259	.4307503	12.38	0.000	4.487004 6.175514
Structural Alien67	SES	-.6140404	.0562407	-10.92	0.000	-.7242701 -.5038107
	Alien71	.7046342	.0533512	13.21	0.000	.6000678 .8092007
SES	SES	-.1744153	.0542489	-3.22	0.001	-.2807413 -.0680894

Indirect effects

		OIM				
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Measurement anomia67	Alien67	0	(no path)			
	SES	-.6140404	.0562407	-10.92	0.000	-.7242701 -.5038107
pwless67	Alien67	0	(no path)			
	SES	-.5455679	.0500916	-10.89	0.000	-.6437457 -.4473901
anomia71	Alien67	.7046342	.0533512	13.21	0.000	.6000678 .8092007
	Alien71	0	(no path)			
pwless71	Alien67	-.6070892	.0582633	-10.42	0.000	-.7212832 -.4928952
	SES	.5979541	.0472921	12.64	0.000	.5052633 .690645
educ66	SES	0	(no path)			
	occstat66	0	(no path)			
Structural Alien67	SES	0	(no path)			
	Alien71	0	(no path)			
SES	SES	-.4326739	.047684	-9.07	0.000	-.5261327 -.339215

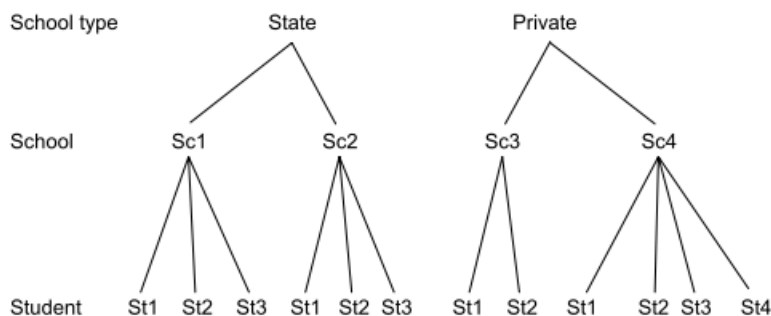
Total effects

		OIM				
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Measurement anomia67	Alien67	1	(constrained)			
	SES	-.6140404	.0562407	-10.92	0.000	-.7242701 -.5038107
pwless67	Alien67	.8884887	.0431565	20.59	0.000	.8039034 .9730739
	SES	-.5455679	.0500916	-10.89	0.000	-.6437457 -.4473901
anomia71	Alien67	.7046342	.0533512	13.21	0.000	.6000678 .8092007
	Alien71	1	(constrained)			
pwless71	Alien67	-.6070892	.0582633	-10.42	0.000	-.7212832 -.4928952
	SES	.5979541	.0472921	12.64	0.000	.5052633 .690645
educ66	SES	.8486022	.0415205	20.44	0.000	.7672235 .9299808
	occstat66	5.331259	.4307503	12.38	0.000	4.487004 6.175514
Structural Alien67	SES	-.6140404	.0562407	-10.92	0.000	-.7242701 -.5038107
	Alien71	.7046342	.0533512	13.21	0.000	.6000678 .8092007
SES	SES	-.6070892	.0582633	-10.42	0.000	-.7212832 -.4928952

دکتر مرضیه روحانی رصاف: استادیار اپیدمیولوژی، با همکاری نویسندگان کتاب

### تحلیل مدل چند سطحی: Multi-Level Model

روش تحلیل چند سطحی یک روش پیشرفته و نسبتاً جدید است. در سال ۱۹۹۵ تنها ۲۲ مقاله با این نوع تحلیل در پایگاه اطلاعات Pubmed منتشر گردیده بود. این رقم در سال ۲۰۰۰ میلادی به ۸۶ مقاله و در سال ۲۰۰۴ به ۱۷۰ مقاله رسید. افزایش سیر مقالات منتشر شده با این روش و تقریباً کم بودن تعداد مقالات در مقایسه با استفاده از سایر روش های تحلیل آماری نشان از این دارد که این تحلیل رایج نیست و محققان کمتر از آن استفاده می کنند. شاید عدم آشنایی پژوهشگران با این تحلیل مهمترین دلیل این امر باشد. این تحلیل همان رگرسیون پیشرفته است. در رگرسیون بین متغیرهای مستقل رابطه قوی نباید باشد و اگر رابطه وجود داشت استفاده از رگرسیون باعث برآورد بیشتر از حد واریانس شده و یک محدودیت تلقی می شود. اما این محدودیت با استفاده از روش چندسطحی با قرار گرفتن افراد که معمولاً شبیه به هم هستند در یک کلاستر، برطرف می گردد. وقتی واحدی در سطح پائین تر از یک یا بیشتر از یک واحد بالاتر قرار گیرد ساختار سلسله ای یا سلسله مراتبی پیش می آید. در واقع، وقتی داده ها در بیش از یک طبقه، گروه بندی و سطح بندی شده باشند (مثل کشور، منطقه) از مدل چند سطحی استفاده می شود. مثال بارز آن مردم ساکن در نواحی مختلف (مثلاً بیست گانه) شهرداری تهران و سپس بوجود آمدن هر چند ناحیه در یک منطقه و بوجود آوردن مناطق مختلف و سپس شهرداری کل است. یا مثال دیگر دانش آموزان در مدارس مختلف است که این مدارس را در سطح بالاتری دوباره می توان دسته بندی و ناحیه ها را به وجود آورد. این مثال در قالب شکل زیر در سه سطح دیده می شود.



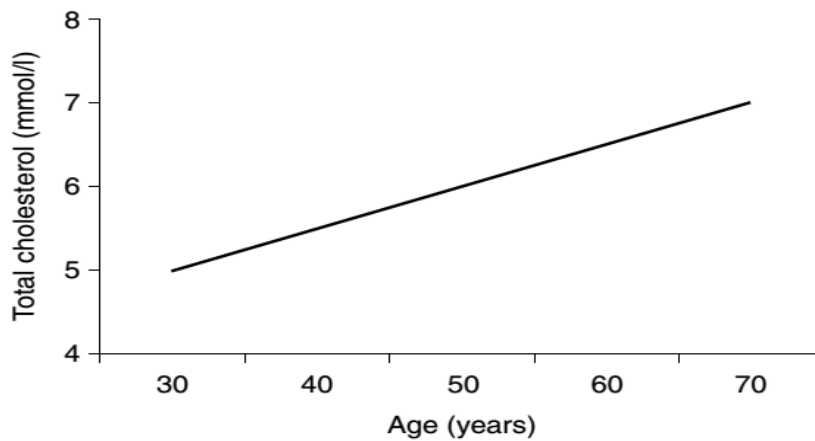
در مدل چند سطحی تخمین میانگین هر گروه و تاثیر این متغیرها در سطوح بالاتر که در هر گروه متفاوت است بر متغیر پاسخ فردی بررسی می گردد. در مدل ساده بدون در نظر گرفتن سطوح میانگین، واریانس بین گروه ها نادیده گرفته می شود اما در مدل چند سطحی اینکه چقدر از واریانس متغیر وابسته مربوط به فاکتورهای سطح است و چقدر از آن مربوط به فاکتورهای فردی است مشخص می شود. بنابراین در این نوع تحلیل محقق دقت و صحت لازم را در تحلیل داده ها مد نظر قرار می دهد. شناخت فرم داده ها و تحلیل مورد نیاز مهم و ضروری است. متغیر وابسته کمی یا کیفی دو حالت می تواند باشد. متغیرهای مستقل نیز کمی و کیفی می توانند باشند. تحلیل چند سطحی می تواند در سطوح مختلفی انجام گیرد. تحلیل دو سطحی، تحلیل سه سطحی و بیشتر. تحلیل چندسطحی همان رگرسیون است اما با توسعه و رویکرد بهتر که مثال زیر این توسعه مدل را نشان می دهد.

مثال مفهومی تحلیل چند سطحی:

برای بررسی رابطه رگرسیونی بین کلسترول و سن، تحلیل را به چند طریق می توان انجام داد. یکی اینکه کلسترول را به عنوان پیامد در نظر گرفت. متغیر کلسترول را به دو طبقه تقسیم و تحلیل رگرسیون لجستیک ساده انجام داد. دیگری اینکه متغیر کلسترول را به سه حالت اسمی یا رتبه ای تقسیم و رگرسیون لجستیک چندوجهی یا رتبه ای را انجام داد. یا اینکه می توان کلسترول را به عنوان متغیر کمی پیوسته در نظر گرفت و تحلیل رگرسیون خطی ساده را انجام داد. معادله زیر رابطه بین کلسترول و سن را نشان می دهد. این رابطه در نمودار زیر نشان داده شده است.

$$\text{Total cholesterol} = \beta_0 + \beta_1 \times \text{age} + \varepsilon$$

where total cholesterol = outcome variable;  $\beta_0$  = intercept;  $\beta_1$  = regression coefficient for age; age = independent variable, and  $\varepsilon$  = error/residual.

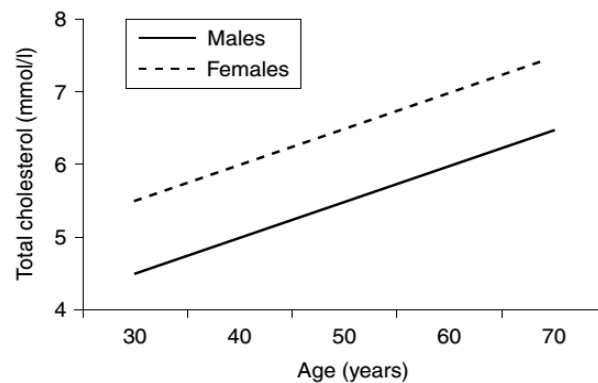


وقتی متغیر مستقل سن، صفر باشد عرض از مبدا ( $\beta_0$ ) intercept برابر با کلسترول می شود. ضریب رگرسیون برای سن ( $\beta_1$ ) است. این ضریب اشاره به اختلاف در کلسترول در سنین مختلف دارد. یعنی با افزایش ۱ سال به سن، مقدار ( $\beta_1$ ) به کلسترول فرد اضافه می شود.

چنانچه محقق بخواهد تحلیل را بر حسب جنسیت انجام دهد معادله به صورت زیر خواهد شد.

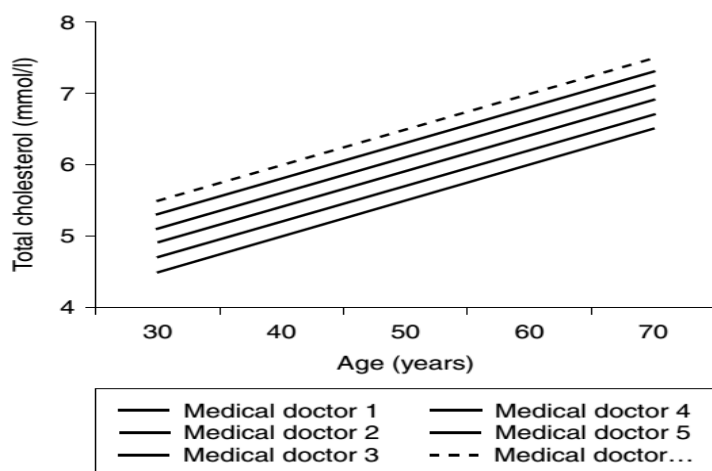
$$\text{Total cholesterol} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{gender} + \varepsilon$$

شکل زیر این تحلیل را بر حسب جنسیت نشان می دهد.



چنانچه محقق بخواهد متغیر دیگری مانند درمان های پزشکی یا نسخ پزشکی را وارد مدل نماید و این متغیر بیش از دو طبقه داشته باشد و بخواهد بر اساس آن مدل را اجرا نماید با توجه به عدم کمی بودن متغیر درمان های پزشکی و بیشتر از دو طبقه بودن باید تبدیل به متغیر مجازی یا ظاهری (Dummy) گردد تا بتوان اثر آن را در تحلیل در نظر گرفت و تصحیح مورد نظر را انجام داد. معادله زیر این مدل را نشان می دهد. متغیر مجازی به تعداد طبقات متغیر منهای یک ، ساخته و وارد مدل می شود. یک طبقه را به عنوان مرجع باید در نظر گرفت.

$$\begin{aligned} \text{Total cholesterol} = & \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{dummyMD}_1 \\ & + \beta_3 \times \text{dummyMD}_2 + \dots + \beta_m \\ & \times \text{dummyMD}_{m-1} + \varepsilon \end{aligned}$$



با توجه به تعداد طبقات متغیردرمان های پزشکی که ۱۲ طبقه دارد باید ۱۱ ضریب برای مدل رگرسیونی (علاوه بر ضرایب سایر متغیرهای مستقل) برآورد گردد. وقتی تعداد سطوح متغیر زیاد باشد کارایی و قدرت مدل رگرسیون به شدت آسیب می بیند. زیرا مطلوب محقق هم نیست که مقادیر مختلف کلسترول برای هر یک از درمان های پزشکی وجود داشته باشد.

**بنابراین مدل تحلیل مناسب تحلیل چند سطحی است.** در این مدل تحلیل ضرایب به صورت جداگانه برای این متغیر برآورد نمی شوند بلکه واریانس ضرایب برآورد می شود و بجای وارد کردن ۱۱ ضریب در مدل، می توان یک متغیر بنام واریانس ضرایب را وارد مدل نمود. برآورد واریانس مذکور با فرض اینکه ضرایب تصادفی هستند محاسبه می شوند. در تحلیل چند سطحی مشاهدات بیماران هستند. این بیماران در درون هر یک از درمان های پزشکی خوشه یا کلاستر نامیده می شوند. بیماران درون هر خوشه از درمان های پزشکی ، با همدیگر مرتبط و وابسته به هم هستند. بنابراین تصحیح باید بر اساس متغیر درمان پزشکی انجام شود. خوشه ای شدن در ساختار داده ها در دو سطح وجود دارد. بیماران مشاهده شده در سطح پایین و درمان های پزشکی در سطح بالاتر وجود دارند. توزیع نرمال داشتن ضرایب مختلف برای متغیر سطح دوم با اهمیت است. فرضیات مدل رگرسیون خطی که در قسمت تحلیل و مدل سازی رگرسیون توضیح داده شد در اینجا نیز باید برقرار باشند. بزرگترین مزیت تحلیل چند سطحی این است که می توان متناسب با متغیر پیامد نوع تحلیل رگرسیونی را انتخاب و مدل تحلیل چندسطحی را اجرا نمود.

**تحلیل چند سطحی با نام های دیگری از جمله** Hierarchical Models, Cotextual Models, Fixed Effect / Random Effect و Marginal Model. شناخته می شود. در قسمت قبلی در تحلیل داده های وابسته به هم، به مدل های Fixed Effect / Random Effect پرداخته شد. در این قسمت اختصاصاً به کاربرد این تحلیل در مدل های کمی و دوحالته پرداخته می شود. پیش فرض اصلی در استفاده از این نوع تحلیل وجود همبستگی قابل توجه در درون هر سطح یا طبقه است. برای ارزیابی این فرض نیز از ضریب همبستگی درون طبقه یا ICC استفاده می شود. این ضریب نسبت واریانس میان گروه ها یا سطح ها به واریانس کل است. که در بخش های قبلی نحوه محاسبه آن توضیح داده شد. داده ها شامل یک سری متغیرهای مربوط به طبقات، یک ستون متغیر پاسخ و یک سری متغیرهای مستقل می باشد. شکل داده ها در جدول زیر آمده است. شناخت ساختار این جدول و سطوح متغیرها پژوهشگر را قادر به این تحلیل می کند.

متغیرهای مستقل		متغیر پاسخ	سطوح یا طبقات		
جنس	نمره قبلی	نمره دانش آموز	نوع مدرسه	مدرسه	دانش آموز
زن	۴۶	۶۵	دولتی	۱	۱
زن	۴۵	۶۱	دولتی	۱	۲
مرد	۷۰	۸۱	خصوصی	۲	۱
زن	۵۰	۶۸	خصوصی	۲	۲
مرد	۴۶	۳۷	خصوصی	۲	۳

پس از شناخت ساختار داده ها، در ادامه تحلیل چند سطحی با دو نوع متغیر پاسخ پیوسته و دوحالته و کار با نرم افزار Stata پرداخته می شود:

#### الف) مدل چند سطحی برای متغیر پاسخ پیوسته:

برای کار با نرم افزار از فایل داده های multilevel1 استفاده می شود. در این فایل تاثیر مدارس مختلف و همچنین دیگر متغیرهای پیشگویی کننده در سطح مدرسه بر نمره امتحان دانش آموزان و تعیین درصد مشارکت واریانس مربوط به مدارس بررسی می شود. متغیر وابسته نمره کل می باشد. آنالیز داده ها شامل متغیر سطح فردی (دانش آموزان) و سه متغیر در سطح مدرسه است. جدول متغیرها در زیر مشاهده می شود.

Variable name	Description and codes
caseid	Anonymised student identifier
schoolid	Anonymised school identifier
score	Point score calculated from awards in Standard grades taken at age 16. Scores range from 0 to 75, with a higher score indicating a higher attainment

نمونه در کوهورت سالهای ۱۹۸۴-۸۶-۸۸-۹۰-۹۶ و ۹۸ است که متغیر cohort90 از کم کردن هر مقدار(هر سال) از سال ۹۰ محاسبه شده است مثلاً ۶- برابر سال ۸۴ است. طبقه اجتماعی ۱ برای طبقه بالاتر تا ۳ که کارگر و ۴ که نامشخص است طبقه بندی شده است، سه متغیر سطح مدرسه شامل نوع مدرسه (دولتی و خصوصی)،



طبقه شهری ۱ و روستایی صفر مدرسه، مذهب مدرسه شامل کاتولیک یا بدون هیچ فرقه ای. این داده ها ۳۳۹۸۸ دانش آموز است که در ۵۰۸ مدرسه طبقه بندی شدند. سایر متغیرها در جدول زیر معرفی شده اند.

cohort90	The sample includes the following cohorts: 1984, 1986, 1988, 1990, 1996 and 1998. The cohort90 variable is calculated by subtracting 1990 from each value. Thus values range from -6 (corresponding to 1984) to 8 (1998), with 1990 coded as zero
female	Sex of student (1 = female, 0 = male)
sclass	Social class, defined as the higher class of mother or father (1 = managerial and professional, 2 = intermediate, 3 = working, 4 = unclassified)
sctype	School type, distinguishing independent schools from state-funded schools (1 = independent, 0 = state-funded)
schurban	Urban-rural classification of school (1 = urban, 0 = town or rural)
schdenom	School denomination (1 = Roman Catholic, 0 = non-denominational)

از دستور describe برای نشان دادن خلاصه داده ها استفاده کنید.

. describe

Contains data from F:\book\Statistical Methodes with STATA\data\multilevel1.dta

Observations: 33,988  
Variables: 9 12 Feb 2010 14:49

Variable name	Storage type	Display format	Value label	Variable label
caseid	float	%9.0g		Case ID
schoolid	int	%9.0g		School ID
score	byte	%9.0g		Score
cohort90	byte	%9.0g		Cohort
female	byte	%9.0g		Female
sclass	byte	%9.0g		Social class
sctype	byte	%9.0g		School type
schurban	byte	%9.0g		School urban-rural classification
schdenom	byte	%9.0g		School denomination

در گام اول مدل را با ساده ترین مدل چند سطحی شروع می کنیم. ابتدا تاثیر سطح مدرسه بر نمره بدون متغیر مستقل بررسی می شود. به این مدل null گویند. در مدل چند سطحی باید گروههای موجود را به عنوان یک نمونه تصادفی در نظر گرفت پس تفاوت مقدار ثابت در هر گروه از میانگین کل مقداری خطا خواهد داشت.

$$score_{ij} = \beta_0 + u_{0j} + e_{ij}$$

Score نمره دانش آموز  $i$  در مدرسه  $j$  در مدل null برابر با مجموع  $B_0$  میانگین کل مدارس،  $u_{0j}$  تفاوت بین میانگین مدرسه  $j$  با همه مدارس (واریانس بین گروهی که تاثیر سطح دو را بر نمره نشان می دهد) و  $e_{ij}$  باقیمانده سطح فردی خواهد بود. در همین مدل ساده می توان به درصد تغییرات متغیر وابسته متاثر از متغیر سطح اول و دوم پی برد اگر مقدار  $u_{0j}$  بطور معنادار صفر نباشد یعنی متغیر سطح دوم بر متغیر وابسته تاثیر داشته و با کاهش در مقدار آن در مدل های پیچیده تر می توان به برازندگی بهتر مدل رسید.

دستور برای برازش این مدل mixed است و بعد به همراه آن متغیر پاسخ و سپس متغیرهای مستقل ثابت سطح فردی است. با دستور db mixed می توان متغیرها را معرفی نمود.

متغیر تصادفی سطح دو (مدرسه) بعد از دو خط راست نوشته می شود. Mle تخمین بیشترین درستنمایی را میدهد. variance واریانس هر ضریب تصادفی در مدل را می دهد. اگر نوشته نشود انحراف استاندارد را خواهد داد. nostderr از محاسبه خطای استاندارد برای پارامترهای تصادفی جلوگیری می کند. دستور در زیر آمده است. خروجی این دستور نیز آمده است.

```
mixed score || schoolid:, mle variance nostderr
. mixed score || schoolid:, mle variance nostderr
```

Performing EM optimization ...

Performing gradient-based optimization:

Iteration 0: log likelihood = -143269.53

Iteration 1: log likelihood = -143269.53

Mixed-effects ML regression	Number of obs	=	33,988
Group variable: schoolid	Number of groups	=	508
	Obs per group:		
	min	=	1
	avg	=	66.9
	max	=	190
	Wald chi2(0)	=	.
	Prob > chi2	=	.
Log likelihood = -143269.53			

score	Coefficient	Std. err.	z	P> z	[95% conf. interval]
_cons	30.6006	.3694317	82.83	0.000	29.87652 31.32467

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
schoolid: Identity			
var(_cons)	61.02457	.	. .
var(Residual)	258.3572	.	. .

LR test vs. linear model:  $\chi^2(0) = 3749.78$  Prob >=  $\chi^2 = 0.0000$

مدل با استفاده از EM (expectation maximisation) تا ۲۰ بار چرخش و شبیه سازی، مدل را برازش می کند. میانگین کل نمره ۳۰،۶۰ است و میانگین برای مدرسه  $j$   $U_{0j} + 30,60$  که تفاوت بین میانگین مقدار ثابت برای همه گروهها و گروه  $j$  است.

واریانس بین مدارس (سطح دو)  $\text{var}(\_cons)$  ۶۱،۰۲ است. واریانس بین دانش آموزان (سطح یک)  $\text{var}(\text{Residual})$  ۲۵۸،۳۶ است پس واریانس کل ۳۱۹،۳۸ است. ۶۱،۰۲ تقسیم بر واریانس کل ۰،۱۹ می شود که نشان دهنده این است که ۱۹٪ واریانس نمره مربوط به تفاوت در بین مدارس است.

برای معنا داری تاثیر سطح مدرسه تست نسبت درستنمایی انجام می شود خط آخر نتایج آماره LR مدل چند سطحی null را با مدل یک سطحی null مقایسه می کند در واقع LR برابر با دو برابر تفاوت مقدار log likelihood در دو مدل با یک درجه آزادی چون تنها یک پارامتر بین دو مدل متفاوت است و مقایسه آن با توزیع کای اسکوتر با یک درجه آزادی که ۳،۸۴ است.

چون این مقدار از ۳،۸۴ بسیار بیشتر است پس نتیجه می گیریم که تاثیر سطح مدرسه معنادار است. برای مدل تک سطحی null می توان دستور زیر را اجرا کرد:

```
mixed score, mle variance nostderr
```

$$LR = 2(-143269.53 - -145144.42) = 3750$$

در گام دوم می توانیم متغیرهای مستقل سطح اول را وارد کنیم که مدل مقادیر ثابت تصادفی نامیده می شود.

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + e_{ij}$$

در این مدل ضریب متغیر مستقل  $B_1$  است یعنی شیب خط از گروهی به گروه دیگر تغییر نمی کند به عبارت دیگر فرض می شود که تاثیر خصوصیات فردی بر پیامد تحت تاثیر گروهی که به آن تعلق دارند موثر نیست لذا به آن متغیر ثابت گفته می شود. دستور و نتایج آن در زیر آمده است.

mixed score cohort90 || schoolid:, mle variance nostderr

. mixed score cohort90 || schoolid:, mle variance nostderr

Performing EM optimization ...

Performing gradient-based optimization:

Iteration 0: log likelihood = -140456.79

Iteration 1: log likelihood = -140456.79

Mixed-effects ML regression	Number of obs	=	33,988
Group variable: schoolid	Number of groups	=	508
	Obs per group:		
	min	=	1
	avg	=	66.9
	max	=	190
	Wald chi2(1)	=	6120.93
Log likelihood = -140456.79	Prob > chi2	=	0.0000

score	Coefficient	Std. err.	z	P> z	[95% conf. interval]
cohort90	1.214954	.0155293	78.24	0.000	1.184518 1.245391
_cons	30.55915	.3225441	94.74	0.000	29.92698 31.19133

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
schoolid: Identity			
var(_cons)	45.98856	.	.
var(Residual)	219.2879	.	.

LR test vs. linear model:  $\text{chibar2}(01) = 3158.04$  Prob >=  $\text{chibar2} = 0.0000$

$$\text{scôre}_{ij} = 30.559 + 1.215 \text{cohort90}_{ij}$$

در اینجا تاثیر کوهورت با شیب ۱,۲۱ برای همه مدارس یکسان فرض شده است. با مقایسه نتایج در دو مدل می بینیم که متغیر کوهورت مقدار واریانس در هر دو سطح فردی و گروهی را کم کرده است. واریانس بین مدارس از ۶۱,۰۲ به ۴۵,۹۹ و واریانس داخل مدارس از ۲۵۸,۳۶ تا ۲۱۹,۲۹ رسیده است. کاهش در واریانس داخل مدارس مورد انتظار بود چون متغیر کوهورت یک متغیر سطح فردی است اما کاهش بیشتر در واریانس سطح مدرسه نشان دهنده این است که توزیع دانش آموزان بوسیله متغیر کوهورت از مدرسه ای به مدرسه دیگر تفاوت دارد. بعد از محاسبه تاثیر کوهورت نسبت واریانس تبیین نشده به خاطر تفاوت بین مدارس کاهش کوچکی دارد ۴۵,۹۹ تقسیم بر واریانس کل ۲۱۹,۲۹+۴۵,۹۹ برابر ۱۷٪ است.

برای برازش مدلی که هم  $\text{intercept}$  و هم شیب متغیر  $\text{cohort90}$  را در بین سطوح مدارس بطور تصادفی متفاوت در نظر بگیرد مدل زیر را برازش می دهیم که به مدل ضرایب تصادفی معروف است.

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij}$$

توجه کنید که ulj به مدل اضافه شده است و ضریب برای هر سطح مقدار خاصی می گیرد. در دستور برای برازش این مدل باید متغیر cohort90 را بعد از سطح دو بیاورید تا متغیر مورد نظر در سطوح مختلف مدرسه بطور تصادفی در نظر گرفته شود. چون پیش فرض نرم افزار استقلال شیب خط و مقدار ثابت است دستور covariance(unstructured) تعیین میکند که این دو بهم وابسته هستند.

```
. xtmixed score cohort90 || schoolid: cohort90, covariance(unstructured)mle variance
> nostderr
```

Performing EM optimization:

Performing gradient-based optimization:

```
Iteration 0: log likelihood = -140343.09
Iteration 1: log likelihood = -140343.09
```

```
Mixed-effects ML regression
Group variable: schoolid
Number of obs = 33988
Number of groups = 508
Obs per group: min = 1
                avg = 66.9
                max = 190
```

```
Log likelihood = -140343.09
Wald chi2(1) = 2376.07
Prob > chi2 = 0.0000
```

score	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
cohort90	1.233902	.0253135	48.74	0.000	1.184289 1.283516
_cons	30.60963	.313448	97.65	0.000	29.99529 31.22398

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
schoolid: Unstructured			
var(cohort90)	.1605836	.	.
var(_cons)	42.85854	.	.
cov(cohort90,_cons)	-1.024181	.	.
var(Residual)	215.7394	.	.

```
LR test vs. linear regression: chi2(3) = 3385.44 Prob > chi2 = 0.0000
```

در این مدل نیز تست LR مدل حاضر را با مدل تک سطحی مقایسه می کند. برای تفاوت تاثیر کوهورت در میان مدارس مانند مثال قبل باید ۲ برابر تفاوت لوگ درستنمایی در دو مدل را بدست بیاوریم و با مقدار کای اسکوئر مقایسه کنیم. مقدار لوگ درستنمایی برای مدل مقدار ثابت تصادفی (مدل قبل) ۱۴۰۴۵۷- بود. فرض صفر این است که دو پارامتر اضافی (هم ضریب ثابت و هم شیب خط) هر دو همزمان برابر صفر باشد.

$$LR = 2(-140343 - -140457) = 228 \text{ on } 2 \text{ d.f.}$$

معناداری قوی نشان دهنده این است که تاثیر کوهورت در میان مدارس فرق می کند. تاثیر کوهورت برای مدرسه j برابر  $1.234 + \hat{u}_{1j}$  و واریانس بین مدارس ۰,۱۶۱ تخمین زده شده است. ۹۵٪ فاصله برای شیب خط  $1.234 \pm 1.96\sqrt{0.161}$ ، ۰,۴۴۸ تا ۰,۲۰۲ پس با فرض نرمال انتظار داریم که ۹۵٪ مدارس یک شیب خطی بین این دو عدد داشته باشند. واریانس مقدار ثابت ۴۲,۸۵۸ وقتی کوهورت برابر صفر باشد یعنی برای سال

۱۹۹۰ تفسیر می شود. کوواریانس منفی یعنی مدارس با مقدار ثابت بالا (نمره متوسط بالادرسال ۹۰) تمایل به داشتن شیب خط کندتر دارند و مدارس با شیب کم افزایش قابل توجهی در نمره با تغییر کوهورت خواهند داشت. مدل بعدی اضافه کردن ضریب تصادفی جنس است. تفاوت نمره در دو جنس متفاوت است این موضوع را می توان با دستور زیر اجرا کرد و با دیدن مقدار معناداری  $p$  برای ضریب جنس آنرا تأیید کرد

regress score female

حالا آیا این تفاوت در میان مدارس مختلف یکسان خواهد بود؟ ابتدا با اضافه کردن تاثیر ثابت برای جنس شروع می شود. این مدل با مدل ضریب تصادفی مقایسه خواهد شد

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij}$$

```
. xtmixed score cohort90 female || schoolid: cohort90, covariance(unstructured) mle var
> iance nostderr
```

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -140272.07

Iteration 1: log likelihood = -140272.07

Mixed-effects ML regression

Group variable: schoolid

Number of obs = 33988

Number of groups = 508

Obs per group: min = 1

avg = 66.9

max = 190

Log likelihood = -140272.07

Wald chi2(2) = 2517.61

Prob > chi2 = 0.0000

روش های اپیدمیولوژی و پژوهش با Stata

score	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cohort90	1.227326	.0253264	48.46	0.000	1.177687	1.276965
female	1.944526	.1629805	11.93	0.000	1.62509	2.263962
_cons	29.58487	.3240554	91.30	0.000	28.94974	30.22001

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
schoolid: Unstructured			
var(cohort90)	.1612602	.	.
var(_cons)	42.57498	.	.
cov(cohort90, _cons)	-1.030571	.	.
var(Residual)	214.8374	.	.

LR test vs. linear regression: chi2(3) = 3403.21 Prob > chi2 = 0.0000

و برای اضافه کردن ضریب تصادفی برای جنس مدل زیر را برازش می دهیم که در سطوح مختلف شیب متفاوت خواهد داشت.

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + u_{0j} + u_{1j} \text{cohort90}_{ij} + u_{2j} \text{female}_{ij} + e_{ij}$$

```
. xtmixed score cohort90 female|| schoolid: cohort90 female, covariance(unstructured)
> mle variance nostderr
```

Performing EM optimization:

Performing gradient-based optimization:

```
Iteration 0: log likelihood = -140272
Iteration 1: log likelihood = -140269.46
Iteration 2: log likelihood = -140269.45
Iteration 3: log likelihood = -140269.45
```

```
Mixed-effects ML regression      Number of obs      =      33988
Group variable: schoolid        Number of groups   =         508

Obs per group: min =           1
                  avg =          66.9
                  max =          190
```

```
Wald chi2(2) = 2524.31
Log likelihood = -140269.45      Prob > chi2 = 0.0000
```

score	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cohort90	1.22777	.0253452	48.44	0.000	1.178094	1.277446
female	1.93145	.1738955	11.11	0.000	1.590621	2.272278
_cons	29.58908	.317659	93.15	0.000	28.96648	30.21169

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
<b>schoolid: Unstructured</b>			
var(cohort90)	.1617015	.	.
var(female)	1.37019	.	.
var(_cons)	40.55858	.	.
cov(cohort90, female)	-.0530665	.	.
cov(cohort90, _cons)	-1.008308	.	.
cov(female, _cons)	1.535505	.	.
var(Residual)	214.5159	.	.

LR test vs. linear regression: chi2(6) = 3408.45 Prob > chi2 = 0.0000

تست  $LR$  این مدل را با مدل قبل که جنس یک متغیر ثابت بود مقایسه می کند و فرض صفر این است که همه سه متغیر برابر صفر می باشد.  $LR = 2(-140269.45 - -140272.07) = 5.24$  on 3 d.f. که در سطح ۵٪ معنادار نیست (کای اسکوئر با ۳ درجه آزادی ۷,۸۲) پس نمی توانیم فرض صفر را رد کنیم پس نتیجه گیری می شود که تاثیر جنس برای همه سطوح مدارس برابر است پس به مدل با ضریب ثابت برای جنس برمیگردیم. برای تمرین؛ طبقه اجتماعی را بعد از تبدیل آن به متغیر ظاهری (*Dommy*) در مدل اضافه کنید.

### ب) اضافه کردن متغیرهای مستقل سطح دوم:

مزیت استفاده آنالیز چند سطحی ارزیابی تاثیرات متغیرهای مستقل سطح دو بر پیامد سطح ۱ و مقدار واریانس که آنها می توانند تبیین نمایند می باشد. ممکن است نمره دانش آموزان تحت تاثیر مدرسه ای که در آن تدریس می کردند باشد. در این داده ها متغیرهای سطح مدرسه که پیشگویی کننده نمره دانش آموزان است شامل *schdenom* و *schurban* و *schtype* می باشد.

قبل از اضافه کردن طبقه اجتماعی به مدل، با دستورات زیر متغیر را به ۳ متغیر ظاهری تبدیل می کنیم و طبقه ۳ را رفرنس قرار می دهیم تا بتوانیم تاثیر هر طبقه را محاسبه کنیم

```
generate sclass1 = sclass==1
generate sclass2 = sclass==2
generate sclass4 = sclass==4
```

معادله مدل به صورت زیر است:

$$\text{score}_{ij} = \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + \beta_3 \text{sclass1}_{ij} + \beta_4 \text{sclass2}_{ij} + \beta_5 \text{sclass4}_{ij} + \beta_6 \text{schtype}_j + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij}$$

اکنون دستور زیر را اجرا نمایید.

```
mixed score cohort90 female sclass1 sclass2 sclass4 schtype || schoolid: cohort90,
covariance(unstructured) mle variance nostderr
```

## روش های اپیدمیولوژی و پژوهش با Stata

```
. mixed score cohort90 female sclass1 sclass2 sclass4 schtype || schoolid: cohort90, covariance(unstructured)
> mle variance nostderr
```

Performing EM optimization ...

Performing gradient-based optimization:

Iteration 0: log likelihood = -138333.44

Iteration 1: log likelihood = -138333.44

```
Mixed-effects ML regression      Number of obs   =   33,988
Group variable: schoolid        Number of groups =    508
                                Obs per group:
                                min =         1
                                avg =        66.9
                                max =        190
                                Wald chi2(6)   =   6997.32
                                Prob > chi2    =    0.0000

Log likelihood = -138333.44
```

score	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
cohort90	1.184027	.0242101	48.91	0.000	1.136576	1.231478
female	1.963768	.154259	12.73	0.000	1.661426	2.26611
sclass1	11.03064	.2069528	53.30	0.000	10.62502	11.43626
sclass2	5.856441	.2041423	28.69	0.000	5.45633	6.256553
sclass4	-3.750241	.2845443	-13.18	0.000	-4.307938	-3.192544
schtype	4.247496	.8168863	5.20	0.000	2.646428	5.848564
_cons	24.27927	.2787147	87.11	0.000	23.733	24.82554

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]	
schoolid: Unstructured				
var(cohort90)	.1481224	.	.	.
var(_cons)	20.56986	.	.	.
cov(cohort90,_cons)	-.4585435	.	.	.
var(Residual)				
	192.9941	.	.	.

LR test vs. linear model: chi2(3) = 1681.11      Prob > chi2 = 0.0000

فردی در یک مدرسه خصوصی است (کد ۱) انتظار می رود تا نمره ای که دارد ۴,۲۵ برابر بیشتر از فردی در مدرسه دولتی (با کد صفر) باشد( در همان کوهورت و با همان جنس و طبقه اجتماعی) و آن بطور قوی معنادار است چون ضریب آن ۵ برابر بیشتر از خطای استاندارد آن است و فاصله اطمینان یک را در بر نگرفته است. همچنین یک کاهش کمی در واریانس سطح مدرسه دیده شد.

بعد از محاسبه برای نوع مدرسه واریانس بین مدارس برای سال ۹۰ ( واریانس مقدار ثابت ) از ۲۲,۵ یعنی مدلی که تنها طبقه اجتماعی و جنس را دارد به ۲۰,۶ در این مدل کاهش یافت. اما هنوز مقدار زیادی از واریانس تبیین نشده بین مدارس باقیمانده است پس می توان متغیر *schurban* همچنین *schdenom* را هم اضافه کرد.

$$\begin{aligned} \text{score}_{ij} = & \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + \beta_3 \text{sclass1}_{ij} + \beta_4 \text{sclass2}_{ij} + \beta_5 \text{sclass4}_{ij} \\ & + \beta_6 \text{schtype}_j + \beta_7 \text{schurban}_j + \beta_8 \text{schdenom}_j \\ & + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij} \end{aligned}$$



```
. mixed score cohort90 female sclass1 sclass2 sclass4 schtype schurban schdenom || schoolid: cohort90, covaria
> nce(unstructured) mle variance nostderr
```

Performing EM optimization ...

Performing gradient-based optimization:  
Iteration 0: log likelihood = -138328.92  
Iteration 1: log likelihood = -138328.92

```
Mixed-effects ML regression          Number of obs   =   33,988
Group variable: schoolid             Number of groups =    508
                                      Obs per group:
                                      min =         1
                                      avg =        66.9
                                      max =        190
                                      Wald chi2(8)    =   7019.84
Log likelihood = -138328.92          Prob > chi2     =    0.0000
```

score	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
cohort90	1.182018	.0242188	48.81	0.000	1.13455	1.229486
female	1.966649	.1542535	12.75	0.000	1.664318	2.26898
sclass1	11.0335	.2069607	53.31	0.000	10.62786	11.43913
sclass2	5.847271	.2041904	28.64	0.000	5.447065	6.247477
sclass4	-3.740548	.284578	-13.14	0.000	-4.298311	-3.182786
schtype	4.398108	.8108271	5.42	0.000	2.808916	5.9873
schurban	-1.462191	.4849347	-3.02	0.003	-2.412646	-.511737
schdenom	.1698495	.6015188	0.28	0.778	-1.009106	1.348805
_cons	25.2484	.4293585	58.80	0.000	24.40687	26.08993

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
schoolid: Unstructured			
var(cohort90)	.1481931	.	.
var(_cons)	19.96586	.	.
cov(cohort90,_cons)	-.4598982	.	.
var(Residual)	193.0126	.	.

LR test vs. linear model: chi2(3) = 1647.69 Prob > chi2 = 0.0000

دانش آموزی در یک مدرسه شهری ۱،۴۶ برابر کمتر از دانش آموزی در یک مدرسه روستایی نمره میگیرد این نمره کنترل شده برای تاثیر دیگر متغیرهای وارد شده در مدل است. متغیر schdenom هم تاثیر معناداری نداشته چون فاصله اطمینان یک را در بر گرفته است پس می توان آنرا از مدل حذف کرد. واریانس بین مدارس در سال ۹۰ از ۲۰،۶ به ۱۹،۹ رسیده است.

### ج: بر همکنش بین سطحی:

تا به حال متغیرها را بدون در نظر گرفتن اینترکشن و اثر متقابل در مدل وارد کردیم اما در عمل رابطه بین متغیر وابسته و مستقل  $X_1$  ممکن است به مقدار متغیر دیگر  $X_2$  بستگی داشته باشد یعنی تاثیر برهمکنش بین  $X_1$  و  $X_2$  در یک مدل چند سطحی ممکن است این دو در یک سطح یا در سطوح مختلف باشد. در حالت دوم به آن cross-level interaction گویند. برای روشن شدن این موضوع ما اینترکشن بین کوهورت (سطح ۱) و نوع مدرسه (سطح دو) را امتحان می کنیم و بررسی می کنیم که آیا اینترکشن مدرسه و کوهورت تفاوت در نمره بین مدارس را در یک روند توجیه می کند به عبارت دیگر واریانس قسمت تصادفی شیب برای کوهورت را کاهش می دهد.

اول باید متغیر اینتراکشن را ایجاد و در مدل وارد کنیم. برای این کار از دستور زیر استفاده می شود.

```
generate cohort90Xschtype = cohort90*schtype
```

معادله مدل به صورت زیر معرفی می شود.

$$\begin{aligned} \text{score}_{ij} = & \beta_0 + \beta_1 \text{cohort90}_{ij} + \beta_2 \text{female}_{ij} + \beta_3 \text{sclass1}_{ij} + \beta_4 \text{sclass2}_{ij} + \beta_5 \text{sclass4}_{ij} \\ & + \beta_6 \text{schtype}_j + \beta_7 \text{schurban}_j + \beta_8 \text{cohort90Xschtype}_{ij} \\ & + u_{0j} + u_{1j} \text{cohort90}_{ij} + e_{ij} \end{aligned}$$

### نکته :

وقتی داده ها در بیش از یک طبقه یا سطح دسته بندی شوند و داده ها از همدیگر مستقل نباشند برای تعیین اثر سطوح از مدل چند سطحی استفاده می شود. در این مدل برآورد میانگین هر گروه و تاثیر متغیرها در سطوح بالاتر که در هر گروه متفاوت هستند بر متغیر پاسخ فردی بررسی می گردد. در این مدل سهم مقدار واریانسی که مربوط به فاکتورهای سطح و فاکتورهای فردی است به تفکیک مشخص می شوند.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. mixed score cohort90 female sclass1 sclass2 sclass4 schtype schurban schdenom cohort90*schtype || schoolid:
> cohort90, covariance(unstructured) mle variance nostderr
```

Performing EM optimization ...

```
Performing gradient-based optimization:
Iteration 0: log likelihood = -138312.51
Iteration 1: log likelihood = -138312.51
```

```
Mixed-effects ML regression          Number of obs   =   33,988
Group variable: schoolid             Number of groups =    508
Obs per group:
    min =          1
    avg =         66.9
    max =         190
Wald chi2(9)                         =   7155.78
Prob > chi2                           =    0.0000

Log likelihood = -138312.51
```

score	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
cohort90	1.213586	.0244189	49.70	0.000	1.165725	1.261446
female	1.970232	.1541852	12.78	0.000	1.668035	2.27243
sclass1	11.01977	.2068929	53.26	0.000	10.61427	11.42527
sclass2	5.830855	.2041144	28.57	0.000	5.430798	6.230912
sclass4	-3.743213	.2844503	-13.16	0.000	-4.300726	-3.185701
schtype	5.302542	.8325247	6.37	0.000	3.670824	6.934261
schurban	-1.420995	.4915901	-2.89	0.004	-2.384494	-.4574965
schdenom	.1170688	.6092813	0.19	0.848	-1.077101	1.311238
cohort90Xschtype	-.5990071	.103776	-5.77	0.000	-.8024043	-.3956098
_cons	25.17844	.434224	57.98	0.000	24.32737	26.0295

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
schoolid: Unstructured			
var(cohort90)	.1379715	.	.
var(_cons)	20.42063	.	.
cov(cohort90,_cons)	-.3958606	.	.
var(Residual)	192.8525	.	.

LR test vs. linear model: chi2(3) = 1648.39                      Prob > chi2 = 0.0000

ضریب متغیر اینتراکشن معنادار است یعنی تاثیر کوهورت برای مدرسه خصوصی و دولتی متفاوت است یا می توانیم بگوئیم که تفاوت بین مدارس دولتی و خصوصی در کوهورت متفاوت خواهد بود. با اینحال تفاوت بین مدارس در کوهورت ۹۰ تنها از ۰,۱۴۸ به ۰,۱۳۸ رسیده است. برای فهم تاثیر برهمکنش مدل زیر را برای دو متغیر cohort90 و schtype در نظر بگیرید

$$1.214 \text{ cohort90} + 5.291 \text{ schtype} - 0.599 \text{ cohort90Xschtype}$$

برای schtype با کد صفر این معادله به 1.214 cohort90 کاهش می یابد یعنی در state school ما انتظار داریم با افزایش یک سال ۱,۲۱۴ برابر نمره اضافه شود و برای schtype با کد یک

$$1.214 \text{ cohort90} + 5.291 - 0.599 \text{ cohort90} = (1.214 - 0.599) \text{ cohort90} + 5.291 \\ = 0.615 \text{ cohort90} + 5.291$$

یعنی در independet school با کد یک با افزایش یک سال ۰,۶۱۵ برابر به نمره اضافه می شود و تفاوت مورد انتظار بین دو نوع مدرسه در سال ۹۰ یعنی وقتی کوهورت صفر است برابر ۵,۲۹۱ است. نتیجه نهایی این است که میانگین نمره در independet school بالاتر از state schools است اما independet school افزایش کوچکتری در نمره با کوهورت دارند (چون ضریب کوچکتری دارد).

د: آنالیز چند سطحی برای متغیر پاسخ دوتایی:

در قسمت قبل بر روی داده هایی کار شد که متغیر پاسخ پیوسته بود. در این قسمت داده های سلامت زنان بنگلادشی در فایل داده های multilevel2 آنالیز می شود. متغیر پاسخ antemed یک متغیر دوتایی است اینکه آیا خانم حداقل یکبار مراقبت های قبل از بارداری را دریافت کرده است یا نه. داده ها در دو سطح هستند ۵۳۶۶ زن در سطح یک که در بین ۳۶۱ جامعه در comm در سطح دو جای گرفتند. سن در زمان تولد و تحصیلات در سطح اول و منطقه سکونت شهری و روستایی در سطح دو متغیرهای پیش بینی کننده هستند. این متغیرها در زیر نمایش داده شده اند.

Variable name	Description and codes
comm	Community identifier
womid	Woman identifier
antemed	Received antenatal care at least once from a medically-trained provider, e.g. doctor, nurse or midwife (1 = yes, 0 = no)
bord	Birth order of child (ranges from 1 to 13)
mage	Mother's age at the child's birth (in years)
urban	Type of region of residence at survey (1 = urban, 0 = rural)
meduc	Mother's level of education at survey (1 = none, 2 = primary, 3 = secondary or higher)
islam	Mother's religion (1 = Islam, 0 = other)
wealth	Household wealth index in quintiles (1 = poorest to 5 = richest)

اول یک مدل null که تنها یک مقدار ثابت به همراه خطاهای تصادفی سطح دو که از توزیع نرمال پیروی می کند برازش می دهیم.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + u_{0j}$$

دستور برای برازش متغیر وابسته دوتایی xtmelogit است. از دستور زیر استفاده می شود.

xtmelogit antemed || comm:, variance

## روش های اپیدمیولوژی و پژوهش با Stata

```
. xtmelogit antemed || comm:, variance
```

Refining starting values:

```
Iteration 0: log likelihood = -3321.6208
Iteration 1: log likelihood = -3313.2849
Iteration 2: log likelihood = -3313.2818
```

Performing gradient-based optimization:

```
Iteration 0: log likelihood = -3313.2818
Iteration 1: log likelihood = -3313.2817
```

```
Mixed-effects logistic regression      Number of obs      =      5366
Group variable: comm                  Number of groups   =      361

Obs per group: min =          3
                  avg =      14.9
                  max =          25

Integration points =      7            Wald chi2(0)      =          .
Log likelihood = -3313.2817          Prob > chi2      =          .
```

antemed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	.1486212	.0727516	2.04	0.041	.0060307 .2912118

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
comm: Identity var(_cons)	1.502371	.1591921	1.220628 1.849145

LR test vs. logistic regression: chibar2(01) = 808.64 Prob>=chibar2 = 0.0000

می توان بیان داشت که لوگ شانس دریافت مراقبت در یک جامعه متوسط (وقتی  $u_{0j}=0$ )  $B_0=0,149$  تخمین زده می شود. مقدار ثابت برای جامعه  $z$  برابر  $0,149+u_{0j}$  است که واریانس  $u_{0j}$  برابر  $1,502$  است. آماره LR برای تست فرض صفر که واریانس  $u_{0j}$  برابر صفر است نشان می دهد که واریانس بین جوامع بطور معنادار غیر صفر است.

قدم بعد وارد کردن یک متغیر مستقل مثل سن مادر به مدل است. معادله به صورت زیر خواهد بود.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 \text{magec}_{ij} + u_{0j}$$

```
xtmelogit antemed magec || comm:, variance
```

```
. xtmelogit antemed magec || comm:, variance
```

Refining starting values:

```
Iteration 0: log likelihood = -3302.4674
Iteration 1: log likelihood = -3294.2858
Iteration 2: log likelihood = -3294.28
```

Performing gradient-based optimization:

```
Iteration 0: log likelihood = -3294.28
Iteration 1: log likelihood = -3294.2799
```

```
Mixed-effects logistic regression      Number of obs      =      5366
Group variable: comm                  Number of groups   =      361
```

```
Obs per group: min =      3
                  avg =     14.9
                  max =     25
```

```
Integration points =      7              Wald chi2(1)      =     37.44
Log likelihood = -3294.2799             Prob > chi2       =     0.0000
```

antemed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
magec	-.0323352	.0052842	-6.12	0.000	-.0426921 -.0219783
_cons	.1451359	.072752	1.99	0.046	.0025447 .2877272

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
comm: Identity			
var(_cons)	1.5001	.1592894	1.218245 1.847165

LR test vs. logistic regression: chibar2(01) = 796.58 Prob>=chibar2 = 0.0000

با توجه به اینکه تغییر کمی در تخمین واریانس بین جوامع دیده شد پیشنهاد می شود که توزیع سن مادر در میان جوامع مشابه است.

این معادله رابطه بین لگاریتم شانس دریافت مراقبت و سن مادر را نشان می دهد:

$$\log\left(\frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}}\right) = 0.145 - 0.032\text{magec}_{ij}$$

در این قسمت یک مدل تک سطحی را با یک مدل چند سطحی مقایسه می کنیم انتظار می رود وقتی متغیر بصورت تصادفی اضافه میشود اندازه ضرایب زیاد شود

در این مدل تحصیلات رتبه ای مادر به دو متغیر ظاهری (دامی) meduc1 برای ابتدایی و meduc2 برای طبقه بالاتر تبدیل شده است و متغیر wealth خانوار هم که بصورت چارک است را بصورت زیر به متغیر پیوسته تبدیل می کنیم. با دستور summarize میانگین متغیر wealth را بدست آورده سپس متغیر جدید wealthc را می سازیم. اعداد حول میانگین خواهد بود

summarize wealth

generate wealthc = wealth - 3.008

مدل زیر را برازش می دهیم که متغیر سن را بصورت درجه دوم هم وارد می کنیم

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 \text{magec}_{ij} + \beta_2 \text{magecsq}_{ij} + \beta_3 \text{meduc2}_{ij} + \beta_4 \text{meduc3}_{ij} + \beta_5 \text{wealthc}_{ij} + u_{0j}$$

. xtmelogit antemed magec magecsq meduc2 meduc3 wealthc || comm:, variance

Refining starting values:

Iteration 0: log likelihood = -2994.175

Iteration 1: log likelihood = -2987.3033

Iteration 2: log likelihood = -2987.2587

Performing gradient-based optimization:

Iteration 0: log likelihood = -2987.2587

Iteration 1: log likelihood = -2987.2586

Mixed-effects logistic regression

Group variable: comm

Number of obs = 5366

Number of groups = 361

Obs per group: min = 3

avg = 14.9

max = 25

Integration points = 7

Log likelihood = -2987.2586

Wald chi2(5) = 591.90

Prob > chi2 = 0.0000

antemed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
magec	-.0002284	.0066466	-0.03	0.973	-.0132555	.0127987
magecsq	-.0010115	.0006909	-1.46	0.143	-.0023656	.0003426
meduc2	.5509808	.0863579	6.38	0.000	.3817224	.7202392
meduc3	1.315807	.1002754	13.12	0.000	1.119271	1.512343
wealthc	.3977915	.0302409	13.15	0.000	.3385204	.4570627
_cons	-.4560695	.0842313	-5.41	0.000	-.6211598	-.2909791

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
comm: Identity				
var(_cons)	.8886033	.1105534	.6963182	1.133987

LR test vs. logistic regression: chibar2(01) = 323.64 Prob>=chibar2 = 0.0000

اضافه کردن meduc2, meduc3 و wealthc واریانس بین جوامع را کاهش داد یعنی توزیع یک یا بیشتر آنها در میان جوامع متفاوت است. انتظار می رود که چون بعضی جوامع نسبت بالاتری از زنان تحصیل کرده یا خانواده های ثروتمندتر دارند این تفاوت را ایجاد می کند.

حالا مدل یک سطحی را ایجاد می کنیم (همان متغیرهای مستقل اما بدون تاثیر تصادفی جامعه) مدل ساده را با دستور logit برآزش می دهیم.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 \text{magec}_{ij} + \beta_2 \text{magecsq}_{ij} + \beta_3 \text{meduc2}_{ij} + \beta_4 \text{meduc3}_{ij} + \beta_5 \text{wealthc}_{ij}$$

. logit antemed magec magecsq meduc2 meduc3 wealthc

Iteration 0: log likelihood = -3717.6012  
 Iteration 1: log likelihood = -3149.4921  
 Iteration 2: log likelihood = -3149.0774  
 Iteration 3: log likelihood = -3149.0774

Logistic regression	Number of obs	=	5366
	LR chi2(5)	=	1137.05
	Prob > chi2	=	0.0000
Log likelihood = -3149.0774	Pseudo R2	=	0.1529

antemed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
magec	-.0046213	.0059102	-0.78	0.434	-.0162051	.0069625
magecsq	-.0013503	.0006307	-2.14	0.032	-.0025864	-.0001143
meduc2	.3968979	.0757902	5.24	0.000	.2483518	.5454439
meduc3	1.072526	.0849994	12.62	0.000	.9059303	1.239122
wealthc	.4819516	.0239126	20.15	0.000	.4350838	.5288194
_cons	-.3686811	.0593236	-6.21	0.000	-.4849531	-.252409

در مقایسه نتایج دو مدل مشاهده می شود که ضرایب تحصیلات در مدلی که تاثیر تصادفی اضافه شده (مدل اول) نسبت به مدل ساده دوم بیشتر است. برعکس ضریب wealthc وقتی مدل با تاثیر تصادفی وجود دارد کاهش می یابد. انتظار می رود که کنترل برای ویژگیهای غیر قابل مشاهده جامعه در مدل چند سطحی تاثیر wealthc را کم خواهد کرد. در این مدل تاثیر سن و تحصیلات و ثروت فرض شده تا در هر جامعه یکسان باشد. می توان شیب تصادفی برای این متغیرها در نظر گرفت به طوریکه برای هر جامعه متفاوت باشد .



## روش های اپیدمیولوژی و پژوهش با Stata

```
. xtlogit antemed magec magecsq meduc2 meduc3 wealthc || comm: wealthc, covariance(u
> nstructured) mle variance
```

Refining starting values:

```
Iteration 0: log likelihood = -3088.6724 (not concave)
Iteration 1: log likelihood = -2993.8962
Iteration 2: log likelihood = -2984.5034
```

Performing gradient-based optimization:

```
Iteration 0: log likelihood = -2984.5034
Iteration 1: log likelihood = -2981.8795
Iteration 2: log likelihood = -2981.8289
Iteration 3: log likelihood = -2981.8286
```

```
Mixed-effects logistic regression          Number of obs    =    5366
Group variable: comm                      Number of groups =     361

Obs per group: min =          3
                  avg =         14.9
                  max =         25
```

```
Integration points = 7                    Wald chi2(5)     =    528.82
Log likelihood = -2981.8286                Prob > chi2      =     0.0000
```

antemed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
magec	-.0001365	.006661	-0.02	0.984	-.0131919	.0129189
magecsq	-.0010593	.0006936	-1.53	0.127	-.0024188	.0003002
meduc2	.5491784	.0870629	6.31	0.000	.3785383	.7198185
meduc3	1.311499	.0999473	13.12	0.000	1.115606	1.507392
wealthc	.4069847	.0326996	12.45	0.000	.3428948	.4710747
_cons	-.4568416	.0844117	-5.41	0.000	-.6222855	-.2913977

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
comm: Unstructured				
var(wealthc)	.0249109	.0238408	.0038172	.1625653
var(_cons)	.8652962	.1103503	.6739256	1.111009
cov(wealthc,_cons)	-.1159484	.0380808	-.1905854	-.0413115

LR test vs. logistic regression: chi2(3) = 334.50 Prob > chi2 = 0.0000

تست LR دو برابر تفاوت مقدار دو لوگ درست‌نمایی بین مدل حاضر با مدلی که wealth بدون شیب تصادفی در نظر گرفته شده بود.

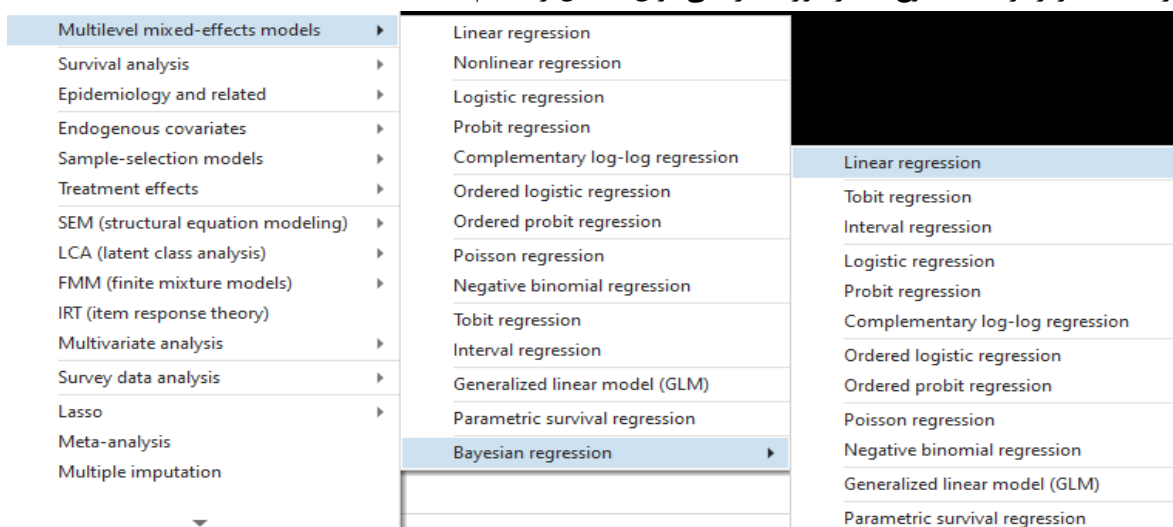
$LR = 2(-2981.8286 - -2987.2586) = 10.86$  on 2 d.f.

کای اسکوتر با یک درجه آزادی ۵,۹۹ که نشان دهنده تأثیر تفاوت معنادار wealth در جوامع (سطح دو) است. تأثیر ثروت بر لگاریتم شانس دریافت مراقبت در جامعه  $z = 0.407 + 0.05u$  است و واریانس بین جوامع در تأثیر ثروت ۰,۲۵ تخمین زده شده است. چون ثروت حول میانگین اندازه گیری شده است واریانس مقدار ثابت برابر ۰,۸۶۵

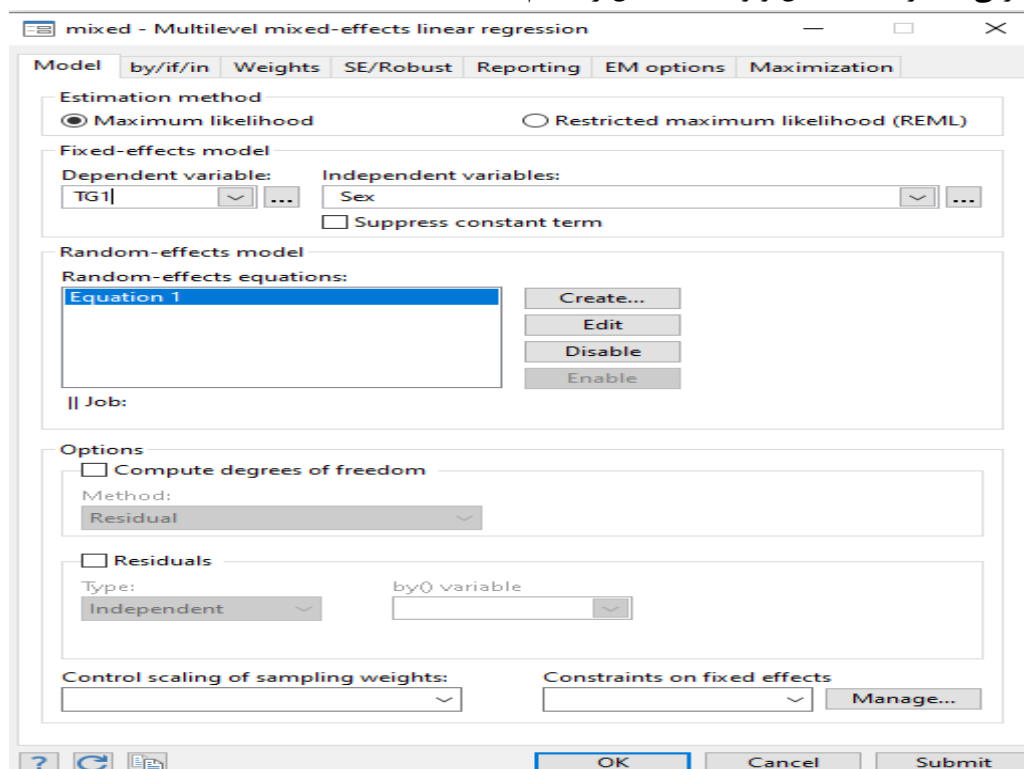
به عنوان واریانس بین گروهی در لگاریتم شانس دریافت خدمت در میانگین wealth تفسیر می شود. می توان مانند وقتی که متغیر پاسخ پیوسته بود متغیرها را بصورت گروهی یا فردی اضافه کنید و تفاوت ها را ببینید و آنها را تفسیر کنید.

### کار با نرم افزار: تحلیل چند سطحی از طریق منوی نرم افزار

برای انجام تحلیل چند سطحی (Multilevel Mixed-Effect Models) از مسیر Statistics و انتخاب یکی از گزینه های زیر بر حسب نوع آنالیز مورد نیاز، می توان تحلیل را انجام داد.



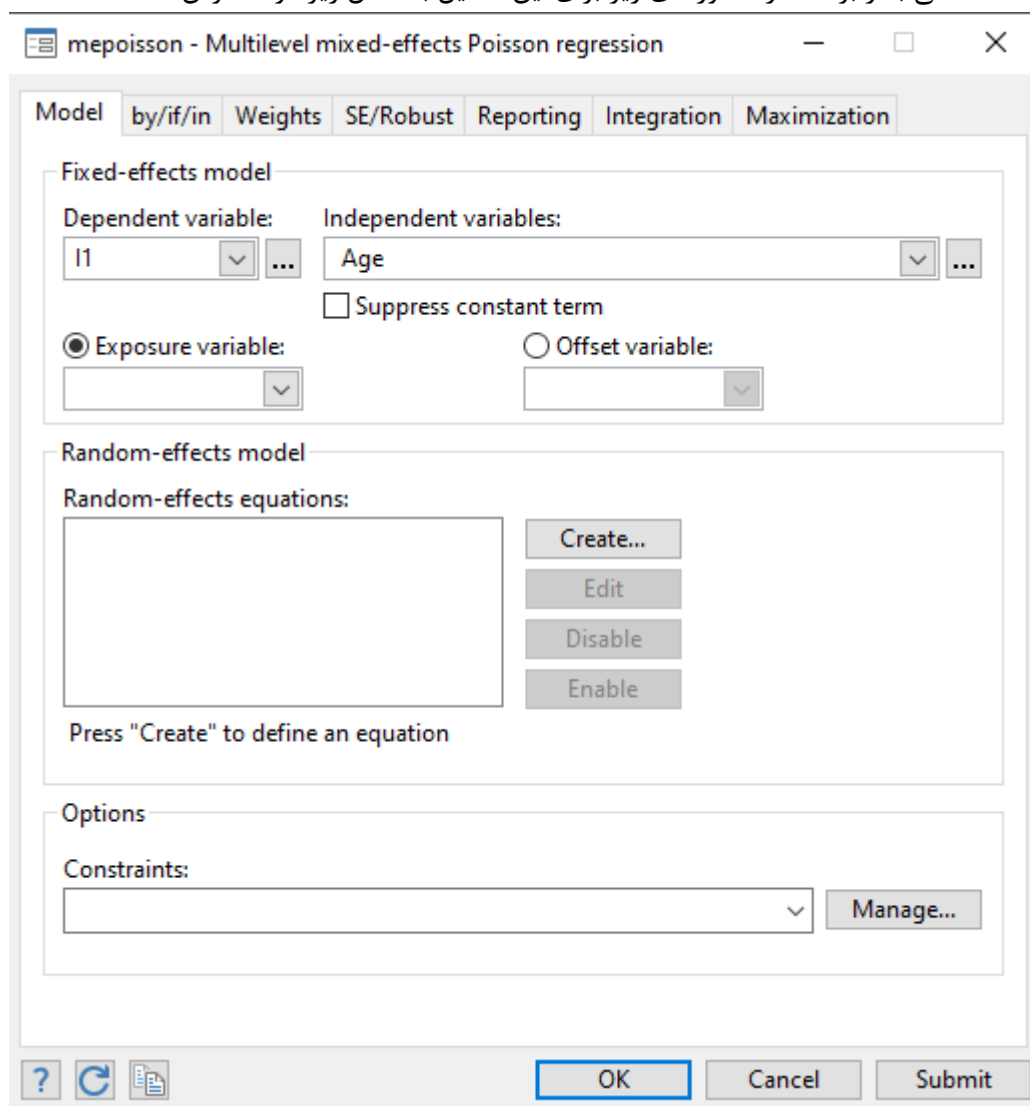
یا با دستور db mixed می توان کادر محاوره ای برای انجام تحلیل اثرات مختلط رگرسیون خطی را به شکل زیر باز و با معرفی متغیرهای مستقل و وابسته تحلیل را انجام داد.



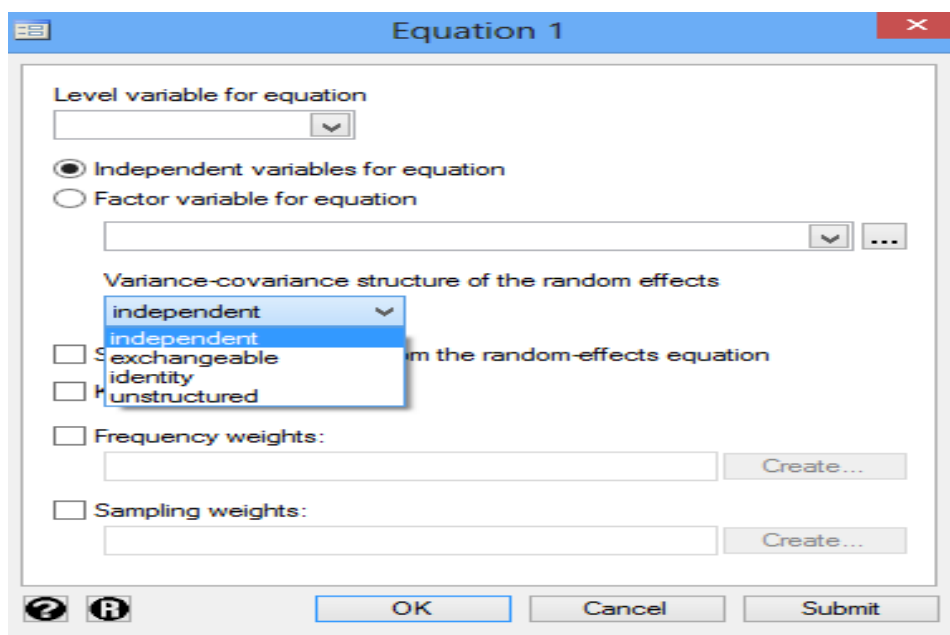
خروجی این دستور نیز مانند رگرسیون خطی تفسیر می شود. منتها در این جا وابسته بودن داده ها مد نظر قرار می گیرد. در گزینه های بالای کادر محاوره ای نیز می توان نوع خروجی مورد نظر و نوع انحراف معیار قابل استفاده را مشخص نمود و تحلیل را انجام داد. با دستور `db melogit` می توان برای متغیرهای دو حالتی تحلیل رگرسیون لجستیک و اثرات مختلط را محاسبه نمود. کادر محاوره ای به شکل زیر برای این دستور باز می شود. با معرفی متغیرهای مستقل و وابسته و تعریف سطوح متغیرها و انتخاب نوع مدل ، تحلیل انجام می شود.

تفسیر نتایج این تحلیل نیز مانند رگرسیون لجستیک است. نوع مدل اثرات تصادفی را نیز می توان به نرم افزار معرفی و تحلیل را بر اساس مدل انجام داد.

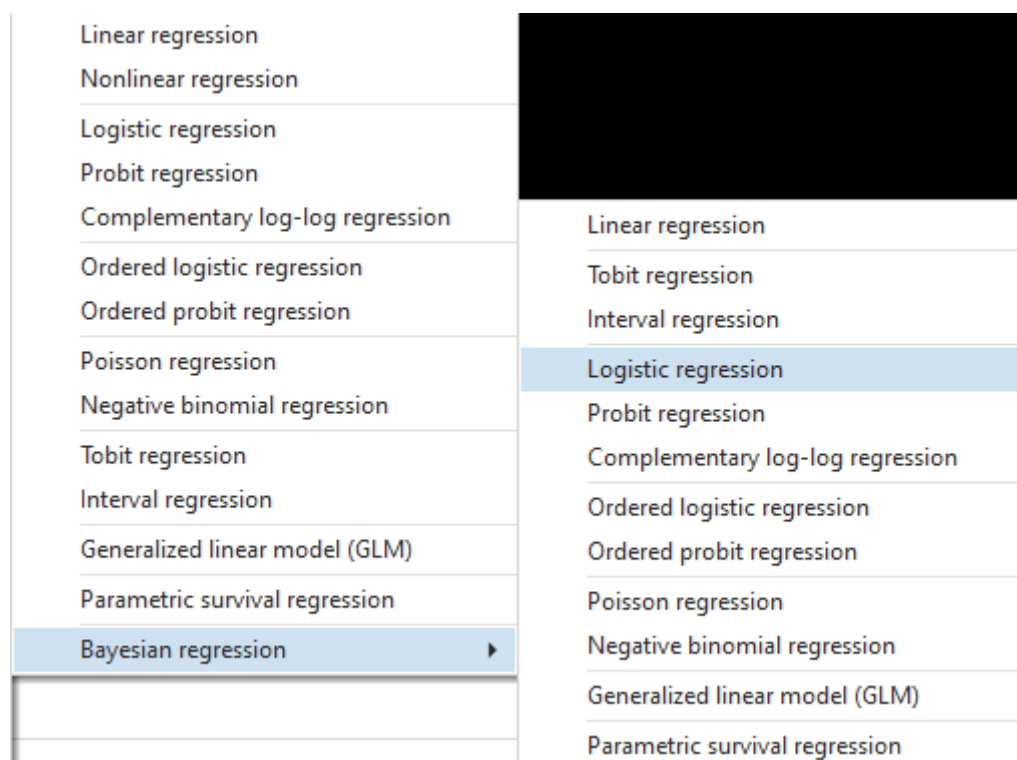
همچنین با دستور db mepoisson نیز می توان تحلیل رگرسیون پواسون را برای داده های وابسته به هم و تحلیل چند سطحی بکار برد. کادر محاوره ای زیر برای این تحلیل به شکل زیر در دسترس است.



در کادر بالا نیز مانند تحلیل رگرسیون پواسون که در قسمت های قبل توضیح داده شد متغیرها را می توان وارد کادر محاوره ای نمود و تحلیل را انجام داد. تفسیر این مدل تحلیل نیز مانند رگرسیون پواسون است. منتها در اینجا اثرات مختلط مدل و برهمکنش و وابسته بودن داده ها مد نظر قرار می گیرد. سطوح متغیر را با انتخاب گزینه Equation به شکل زیر و همچنین انتخاب نوع ماتریس کوواریانس واریانس، برای نرم افزار معرفی و تعریف نمود.



در نرم افزار استاتا تحلیل های چندسطحی برای انواع مدل های زیر را می توان انجام داد.



تمرین ها :

الف) فایلی بنام **part six regression** را باز نمایید. به سوالات زیر پاسخ دهید.

- ۱- با استفاده از مدل رگرسیون چندگانه رابطه بین سن، جنس، وزن، قد، BMI، مصرف سیگار، وضعیت فعالیت فیزیکی، فشار خون سیستولیک و دیاستولیک را با تراکم استخوانی ناحیه فمور را تحلیل نمایید.
- ۲- پیش فرضهای مدل رگرسیون خطی مربوطه را ارزیابی نمایید.
- ۳- مدل را از نظر پدیده های **influence** و **collinearity** ارزیابی نمایید.
- ۴- آیا نیاز به تغییراتی در متغیرها وجود دارد؟
- ۵- اگر هدف از تحلیل رگرسیون خطی، ارزیابی روابط بین متغیرها و تعیین اثر مخدوش کنندگی احتمالی برخی از آنها باشد، مدل چند متغیره مربوطه را با متغیرهای فوق تشکیل دهید و نتایج را دسته بندی نمایید.
- ۶- در مورد **interaction** بین جنس و مصرف سیگار در رابطه آن با تراکم استخوانی (منطقه گردن فمور) اظهار نظر نمایید، چه اقداماتی نیاز است که انجام دهید؟ تفسیر شما چیست؟

ب) فایلی بنام **part exam** را باز و به سوالات زیر پاسخ دهید. در این فایل متغیر **cvd** وضعیت مرگ و میر به دنبال بیماری های قلبی عروقی را نشان می دهد. متغیرهای **timebth**، **timeout**، **timein** به ترتیب زمان ورود، خروج و زمان تولد را نشان می دهد.

- ۷- بروز مرگ و میر در این مطالعه در ۱۰۰۰۰ نفرسال چقدر است؟
- ۸- بروز مرگ و میر بیماری قلبی عروقی در این مطالعه در ۱۰۰۰۰ نفر بر حسب گروه کلسترول (**cholgrp**) چقدر است؟
- ۹- استفاده از مدل **COX** اثر متغیر کلسترول گروه بندی شده بر روی بقا و اثر کلسترول گروه بندی کنترل شده برای **smok** بر روی بقا را محاسبه نمایید
- ۱۰- در خصوص تاثیر کلسترول و سیگار و نقش مخدوش کنندگی سیگار توضیح دهید.
- ۱۱- پیش فرض های استفاده از مدل فوق چیست؟
- ۱۲- بررسی نمایید آیا سن افراد در زمان وقوع پیامد، در مدل بالا مخدوش کننده است؟

ج) فایل داده های **tb.dta** را باز نمایید. این فایل مربوط به داده های تعدادی خانوار می باشد که در هر خانوار یک نفر بیمار مبتلا به سل وجود دارد. می خواهیم اثر وجود **cavity** در ریه و اثر درجه **intimacy** را بر روی انتقال **tb** به افراد خانوار بررسی نماییم.

۱۳- مشخص نمایید میزان متوسط خطر ابتلا برای یک فرد در جامعه که در خانواده اش یک فرد با سل ریوی وجود دارد و این فرد **cavity** در ریه دارد و این دو نفر از یک بستر استفاده مینمایند ( **share a common bed**) در مقایسه با خانواد های که فرد اندکس در ریه **cavity** ندارد و درجه **Intimacy** آنها استفاده از **living room** است جقدر است ؟

۱۴- همین مقدار اثر را برای یک نفر در یک خانواده خاص برآورد نمایید.

۱۵- در خصوص تفاوت برآورد اثر در سوال بالا اظهار نظر فرمایید.

۱۶- آیا بین **intimacy** و وجود **cavity** در ریه اثر متقابل ضربی وجود دارد ؟ - بررسی نمایید.

۱۷- تحلیل چند سطحی در چه مواردی کاربرد دارد؟

۱۸- مترادف های دیگر تحلیل چند سطحی را نام ببرید.

۱۹- ساختار داده ها و پیش فرض انجام تحلیل چند سطحی چیست و چگونه بررسی می شود؟

۲۰- دستورات انجام تحلیل چند سطحی با نرم افزار **Stata** را به صورت خلاصه در زیر بنویسید.

## بخش پنجم:

# روش های آماری و اپیدمیولوژیک با Stata

### فصل هفتم: مطالعات اپیدمیولوژیک و محاسبه اندازه های اثر و ارتباط در آنها با Stata

مفاهیم علیت و اندازه اثر  
مبانی و اندازه های اساسی در مطالعات اپیدمیولوژیک  
مطالعات مقطعی  
مطالعات همگروهی  
مطالعات مورد - شاهد  
تحلیل حساسیت و کمی کردن خطاهای پژوهش  
شبیه سازی آماری: بوت استرپ، مونت کارلو و جک نایف  
مطالعه ارزیابی تست های تشخیصی  
مطالعات مداخله ای و کارآزمایی های بالینی  
انواع طراحی مطالعات بالینی در تحقیقات درمان سرطان ها  
اندازه گیری نابرابری در سلامت و معرفی شاخص ها  
سری های زمانی  
متاآنالیز  
روایی و پایایی  
تحلیل کلاس پنهان

### اهداف:

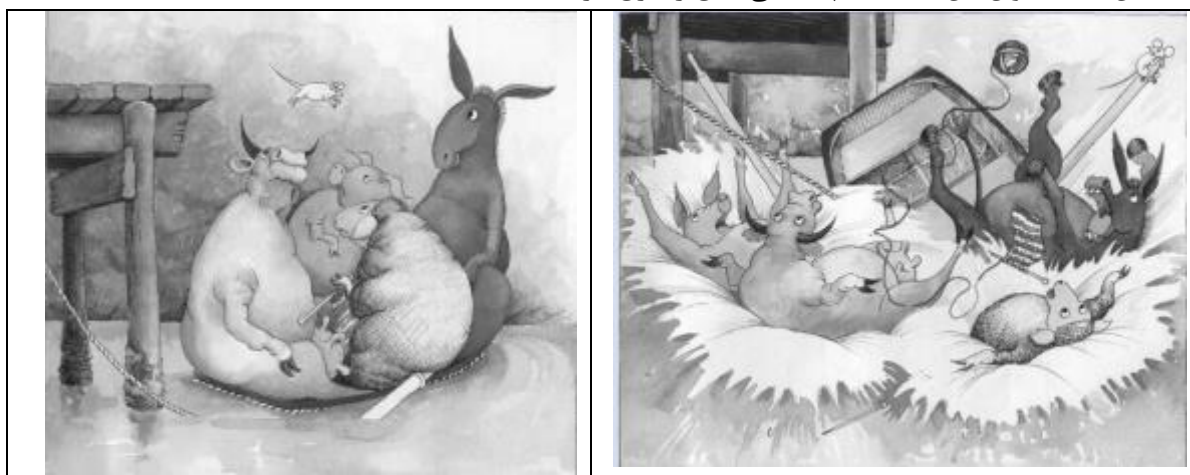
#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند با مفاهیم علیت آشنایی پیدا نماید و ملاحظات گزارش روابط علیتی را بیان کند.
- کاربر بتواند انواع مطالعات اپیدمیولوژیک و نقاط قوت و ضعف آنها را دسته بندی نماید.
- کاربر بتواند اندازه های قابل محاسبه در انواع مطالعات و خطاهای پژوهش را بیان نماید.
- کاربر بتواند روشهای آماری در اپیدمیولوژی را فرا گرفته و با نرم افزار بتواند اندازه های اساسی در مطالعات را محاسبه نماید. با تحلیل حساسیت بتواند خطاهای پژوهش را به صورت کمی محاسبه و تفسیر نماید.
- کاربر بتواند انواع روش های آماری که در فصل های قبل توضیح داده شد، را در انواع پژوهش های اپیدمیولوژیک استفاده و تفسیر نماید.



## مفاهیم پایه در علیت (Causation):

داستان کوتاه زیر را بخوانید و به این سوال که "چه کسی قایق را غرق کرد؟" پاسخ دهید. پنج حیوان در کنار دریا در بندری زندگی می کردند، یک گاو، یک الاغ، یک گوسفند، یک خوک و یک موش. یک روز خوب آنها تصمیم گرفتند که در دریا قایقرانی کنند. ابتدا گاو سوار قایق شد، کمی تکان خورد اما او از پشت و با حالت خوابیده راحت در قایق نشست. سپس خر با دقت وارد شد و در قسمت جلو نشست تا تعادل قایق را حفظ کند. بعد از آن خوک در قایق رفت و زیر چتر قایق نشست، پس از ورود این سه حیوان اندکی آب به قایق وارد شد. سپس گوسفند به درون قایق پرید و مقابل خوک نشست. قایق هنوز شناور است. سرانجام موش کوچولو داخل قایق پرید و - فاجعه! قایق واژگون شد و حیوانات باید تا ساحل شنا کنند. شکل زیرگویای این داستان است. سوال این است که چه کسی قایق را غرق کرد؟

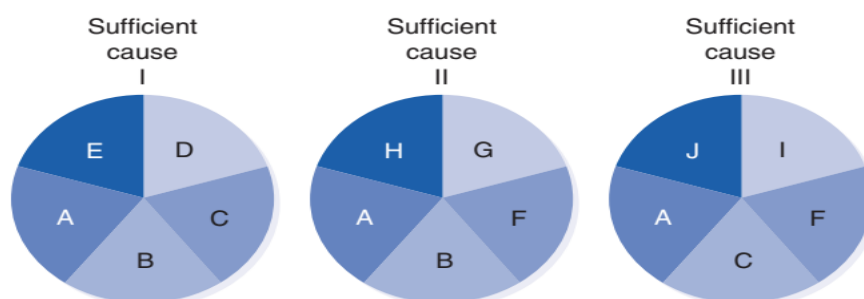


برای بررسی رابطه علت و معلولی و از جمله تعیین علل درد شکم حاد اهالی یک روستا و مراجعه آنها به بیمارستان یا اینکه بررسی بکنیم چرا قایق غرق شد و اظهار نظر درباره علت یا علل هر پدیده ای ابتدا بایستی درباره واژه‌های مرتبط و اصول علمی تعیین علل و علیت آگاهی داشته باشیم. در زیر به این مفاهیم و موارد پرداخته می شود. طبق فرهنگ لغت آمریکایی، علت به مفهوم شرایط، دلیل یا دلایل و واقعه ای است که مسئول یک عمل بوده یا نتیجه ای به همراه دارد. به عبارتی ایجاد یک اثر یا نتیجه حتما ناشی از عللی است. به عنوان مثال، دلیل حادثه رانندگی در یک بزرگراه شرایط یخبندان جاده است یا دلیل سرقت از یک انبار آهن آلات، نبود نگهبان شب برای انبار است. از نظر سوسر (Susser)، علل واقعی یک پدیده یا واقعه دارای سه ویژگی اساسی هستند: ارتباط، ترتیب زمانی و جهت.

منظور از ارتباط این است که "عامل علیتی (X) باید همراه با اثر (Y) قلمداد و رخ دهد. به عبارت دیگر، باید "وابستگی یا همبستگی آماری" بین عامل و معلول وجود داشته باشد. اگر ارتباط آماری بین دو متغیر X و Y پیدا نشود، علیت را می توان رد کرد. متخصصان اپیدمیولوژی با مقایسه مطلق یا نسبی بین دو یا چند گروه، ارتباطات را کمی می کنند.

منظور از ترتیب زمانی به این معناست که علت باید مقدم بر معلول باشد. یک علت ممکن است مدت زمان طولانی یا در کوتاه مدت اثر کند و سبب ایجاد واقعه یا اثر گردد. اولی را علت دور (*distant cause*) می نامند و دومی را علت نزدیک (*proximate cause*) می نامند. درباره بیماریهای مزمن مانند سرطان، علت دور اغلب با دهه ها مقدم بر تشخیص سرطان هستند و یا علت تقریبی با ماهها یا سالها مقدم بر آن هستند. به عنوان مثال، در معرض مواجهه قرار گرفتن رحم با دی اتیل استیل بسترول (DES)، علت دور آدنوکارسینوما واژن در نظر گرفته می شود. زیرا این در معرض قرار گرفتن به طور متوسط ۲۰ سال قبل از تشخیص سرطان در زنان مبتلا رخ می دهد. در مقابل، درمان جایگزینی با هورمون به عنوان یکی از دلایل اصلی سرطان پستان و به عنوان علت نزدیک در نظر گرفته می شود. زیرا فقط در معرض قرار گرفتن فعلی با این هورمون و نه در معرض خطر قبلی، خطر بیشتری برای ابتلا به سرطان پستان دارد. در سال ۱۹۷۰ علیت توسط اپیدمیولوژیست سرشناسی بنام راتمن ارایه و مفاهیم اپیدمیولوژیک در علیت (Causation) معرفی گردید. بر اساس نظریات راتمن، علت (Cause) یک واقعه یا رخداد (event)، حالت یا مشخصاتی است که قبل از وقوع آن واقعه رخ می دهد و بدون آن واقعه یا به صورت کلی در آن زمان اتفاق نمی افتد و یا بعدا رخ می دهد. به عبارت دیگر برای وقوع یک بیماری یا رخداد، علت ضروری و مقدم بر رویداد است. البته با فرض اینکه تمام شرایط دیگر ثابت و بدون تغییر باشند.

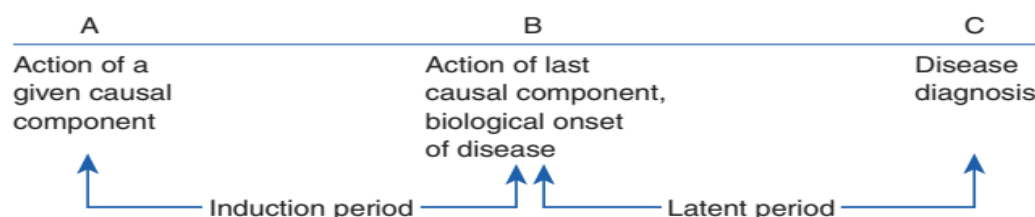
به مجموعه ای از حداقل شرایطی است که رخ دادن بیماری را به صورت غیرقابل اجتناب، تضمین می کند علت کافی یا علت تامه (Sufficient Cause) گفته می شود. اگر یک دایره را به چند قاچ تقسیم کنید هر یک از قاچ ها بعنوان یک عنصر از علت کافی تلقی می شوند. عناصر هر علت کافی را علل جزئی (component cause) می نامند. یک بیماری ممکن است دارای چند علت کافی برای رخداد باشد. به عنوان مثال در ایجاد سرطان معده یا شکم درد حاد ممکن است چند مسیر علل کافی داشته باشیم: میکروب هلیکوباکتر پیلوری، استعمال سیگار، استعمال تریاک، دریافت رژیم غذایی پرچرب، دریافت زیاد نمک، دریافت بیش از اندازه غذاهای نمک سود، سرخ شده و پرچرب، زمینه مساعد ژنتیکی، سن، جنسیت و برخی عوامل دیگر. نمای شماتیک مسیرهای مختلف ایجاد یک بیماری و مدل های چند علت کافی برای رخداد یک واقعه یا رخداد در شکل زیر قابل مشاهده است.



### نمای شماتیک چند علت کافی برای رخداد یک واقعه یا رخداد

دو علت جزئی در یک مجموعه علت کافی مشابه و یکسان اثر خود را اعمال می کنند تا بیماری مورد نظر رخ بدهد. در اصطلاح این دو جز علت کافی با یکدیگر در ایجاد بیماری اثر متقابل یا اینترکشن داشته اند. قدرت اثر

متقابل دو علت جزئی به شیوع فاکتورهای دیگر در مسیرهای علیتی کافی بستگی دارد. برای یک علت جزئی خاص در یک مجموعه علل کافی باید مفهوم دوره القا که به فاصله زمانی بین تاثیر یک علت (A) و آغاز بیماری (B) است توجه نمود. یک بیماری ممکن است دوره القا کوتاه یا طولانی داشته باشد. مفهوم دوره القا را تنها می توان در ارتباط با یک علت جزئی خاص که در یک علت کافی خاص عمل می کند درک و بیان کرد. زیرا هر علت جزئی در هر سازوکار علیتی می تواند در زمانی متفاوت از سایر علل جزئی اثر خود را اعمال کند. بنابراین هر یک از علل جزئی دوره القای منحصر به فردی دارد. و علل جزئی که به عنوان آخرین مورد عمل می کند دوره القای معادل صفر خواهد داشت. بنابراین دوره القا یک زوج "علت-اثر" خاص را مشخص می کند نه اینکه فقط اثر را مشخص نماید. تقریباً دوره القا برای علل جزئی یکسان برابر است. اما دوره نهفتگی که بعد از دوره القا است متفاوت است. فاصله زمانی بین وقوع بیماری برگشت ناپذیر (B) و تشخیص آن بیماری (C) دوره نهفتگی نام دارد. این دوره را با بهبود و ارتقای روش های تشخیص بیماری می توان کوتاه نمود اما دوره القا تغییری نمی کند. چون شروع بیماری به منزله پایان دوره القا است هر عاملی که به عنوان تسریع کننده در یک فرایند علیتی عمل نماید می تواند دوره القای سایر عوامل را سرعت بدهد و خود یک علت محسوب شود و برعکس آن عامل نیز می تواند یک اقدام پیشگیرانه محسوب گردد. در شکل زیر مفاهیم دوره های القا و نهفتگی نمایش داده شد.



### شکل مفاهیم دوره های القا و نهفتگی در شروع تغییرات بیولوژیکی و ایجاد بیماری

پس بطور خلاصه میتوان گفت: علت (Cause) یک واقعه یا رخداد (event)، حالت یا مشخصاتی است که قبل از وقوع بیماری رخ می دهد و بدون آن بیماری (پیامد یا رخداد) یا به صورت کلی در آن زمان اتفاق نمی افتد و یا بعداً رخ می دهد. به عبارت دیگر برای وقوع یک بیماری یا رخ داد، علت ضروری و مقدم بر رویداد است. البته با فرض اینکه تمام شرایط دیگر ثابت و بدون تغییر باشند. به مجموعه ای از حداقل شرایطی است که رخ دادن بیماری را به صورت غیرقابل اجتناب، تضمین می کند علت کافی یا علت تامه (Sufficient Cause) گفته می شود. عناصر هر علت کافی را علل جزئی (component cause) می نامند. علیت و همبستگی دو مفهوم جدا از هم می باشند. روابط علیتی directed هستند ولی همبستگی ها undirected می باشند.

علیت بر خلاف همبستگی نمونه ای قابل مشاهده نمی باشد. با این وجود همبستگی مشاهده شده نتیجه روابط علیتی می باشند. در واقع اگر  $Y$  علت  $X$  باشد در عمل بین این دو متغیر همبستگی مشاهده می شود. ولی نکته ای که می بایست به آن توجه شود، متغیرهای دیگر مثل متغیرهای مخدوش کننده است که توزیع  $Y$  را تحت تاثیر قرار می دهند و رابطه بین  $X$  و  $Y$  دیگر تنها تحت تاثیر رابطه علیتی بین  $X$  و  $Y$  نخواهد بود.

بنابراین در تفسیر مطالعات مشاهده ای، دقت و احتیاط در تفسیر نتایج لازم است.

## گام های تحلیل داده ها در مطالعات اپیدمیولوژی

۱- توجه به داده های خام (Raw Data) : ارزیابی کیفیت، صحت، اعتبار، پایایی و کامل بودن آنها. این فرایند در طی ویرایش داده ها (Data Edithing) و بررسی خطاها و اصلاح آنها انجام می گیرد. در این مرحله بکار بردن نهایت دقت برای تهیه یک بانک اطلاعاتی مناسب بسیار مهم و حساس است. ارزیابی کیفیت داده ها در **سه مرحله** explore و transform و validate انجام می شود. در **مرحله اول** به بررسی کیفیت داده ها با شناخت فیزیک داده ها، توزیع داده ها، دیکشنری متغیرها، ترسیم گراف ها، آمار توصیفی، راستی آزمایی و صحت داده ها، تاریخچه دیتا و پروتکل های جمع آوری و استفاده داده ها بررسی می گردد.

در **مرحله دوم** یا transform و تغییر شکل داده ها و ایجاد متغیرهای جدید به ساختار داده ها و داده های گمشده، شناسایی داده های پرت و پیدا کردن دلیل آنها پرداخته می شود. و در **مرحله سوم** تهیه داده ها برای الگوریتم و تحلیل، استاندارد کردن، نرمال کردن و تهیه جداول فراوانی و نمودارهای گرافیکی مد نظر است.

۲- برآورد و انجام تبدیل های مورد نیاز بر روی داده ها مبتنی بر سوال پژوهش و پیش فرض ها. انجام تحلیل توصیفی (در توصیف داده ها P-Values، حدود اطمینان و سایر آماره هایی که برای استنباط مد نظر هستند محاسبه نمی شود). مانند جداول و نمودارها و گروه بندی داده ها بر حسب فاکتورهای مواجهه، بیماری، مخدوش کننده ها و اثرات متقابل و خلاصه سازی داده ها (Data Description and Summarization). در این مرحله تحلیل گر باید سعی نماید اولاً با داده ها به خوبی آشنا شده و متغیرها را نیز بتواند توصیف نماید. پیدا کردن دید کلی از بانک اطلاعاتی مورد تحلیل در این مرحله لازم است.

۳- ترسیم جداول متقاطع برای داده ها (Data Tabulation): در پژوهش های مختلف میانگین ها، میانه ها و سایر اندازه ها برای متغیر های کمی به صورت خلاصه گزارش می شوند. در پژوهش های اپیدمیولوژیک مقادیر خلاصه با جداول توافقی (Contingency Tables) گزارش می شوند.

۴- انتخاب گروه بندی و طبقه بندی مناسب : ویرایش داده ها و خلاصه نمودن آنها در طبقه ها و جداول توافقی بر اساس نیاز تصمیم گیرندگان گام مهمی است. به عنوان مثال طبقه بندی بر حسب قومیت، مذهب و سایر متغیرهای اسمی یا طبقه بندی متغیر کمی فشارخون سیستمولیک بر حسب پرفشاری خون و فشارخون طبیعی. در این مرحله به متغیرهای مخدوش کننده و برهمکنش توجه می شود.

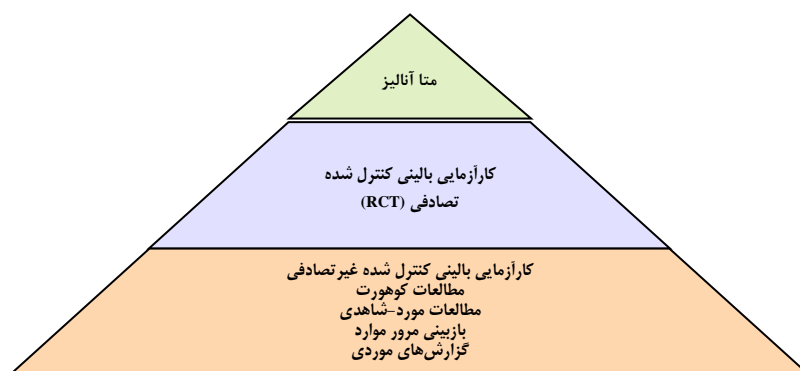
۵- طبقه بندی مشاهدات و توجه به شخص - زمان برای بیماران و غیربیماران، مواجهه یافتگان و غیرمواجهه یافتگان و توجه به داده های گمشده و مدیریت آنها. در این مرحله محاسبه اندازه های اساسی با توجه به ماهیت داده ها و نوع مطالعه مهم است.

۶- استفاده از روش های آزمون فرضیه ها و برآوردها و محاسبه آماره ها و مقادیر معنی داری آزمون ها ( P-Values ) و حدود اطمینان آماره ها، مانند آزمون مانتل هنزل، کای اسکور، والد. در این مرحله تحلیل های اولیه انجام می شود.

۷- توجه خاص به تحلیل های دو متغیره و مدل سازی و انجام تحلیل های چند متغیره و کنترل متغیره های مخدوش کننده و بررسی اینترکشن بین متغیره ها و همچنین کمی نمودن خطاهای متداول در پژوهش. بخش عمده ای از هفت مرحله بالا به عنوان مراحل تجزیه و تحلیل داده های اپیدمیولوژیک در فصل های قبل کتاب همراه با کار با نرم افزار Stata بیان گردید. در این بخش به تحلیل مطالعات اپیدمیولوژیک پرداخته شده است.

### مطالعات اپیدمیولوژیک

معمولا در ارزیابی مطالعات اپیدمیولوژیک، هرم شواهد به صورت زیر ارائه می شود که نحوه ارزیابی سطوح مختلف شواهد توسط برخی مؤسسات مثل انستیتو ملی تعالی بالینی (NICE) را نشان می دهد.

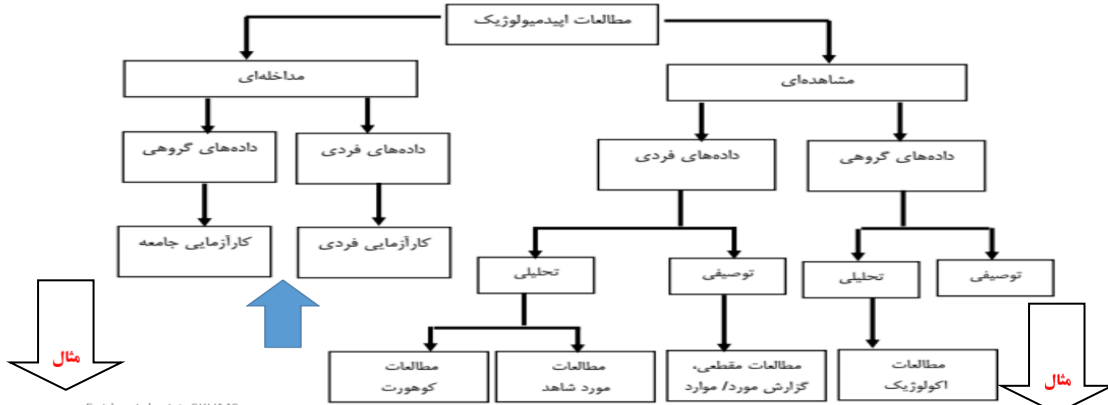


هرم شواهد مذکور را نیز می توان در قالب جدول زیر طبقه بندی نمود.

سطح	شواهد
Ia	شواهد حاصل از مرور منظم و متاآنالیز کارآزمایی های بالینی کنترل شده تصادفی
Ib	شواهد حاصل از حداقل یک کارآزمایی بالینی کنترل شده تصادفی
IIa	شواهد حاصل از حداقل یک مطالعه کنترل شده غیر تصادفی
IIb	شواهد حاصل از حداقل یکی از سایر انواع مطالعات شبه-تجربی
III	شواهد حاصل از مطالعات توصیفی غیرتجربی مثل مطالعات مقایسه ای، مطالعات همبستگی و مطالعات موردشاهدی
IV	شواهد حاصل از گزارش ها یا نظریات کمیته های خبرگان و/ یا تجربه بالینی افراد صلاحیت دار

روش طراحی انواع مطالعات اپیدمیولوژیک در ضمیمه شماره ۴ کتاب به صورت شکل آمده است.

**مطالعات اپیدمیولوژیک** بطور کلی به دو دسته مشاهده ای (Observational یا non-experimental) و مداخله ای (Interventional یا Experimental) تقسیم می شوند. این تقسیم بندی در دیاگرام زیر آمده است.



**طراحی یک پژوهش بالینی برای مطالعه اثر هورمون درمانی پس از یائسگی برای جلوگیری از CHD**

مطالعه	مداخله	نوع مطالعه
	عدم مداخله	۱- مطالعات مشاهده‌ای:
مطالعه هم‌گروهی / کوهورت	پیشگیری یک هم‌گروه در طول زمان معین و اندازه گیری پیامد	محققین یک گروه از خانم‌های یائسه را برای چندسال از لحاظ بروز حمله قلبی، در دو گروه مصرف‌کننده، هورمون و آنها که هورمون مصرف نمی‌کنند به طور سالانه مورد معاینه و بررسی قرار می‌دهند.
مطالعه مقطعی	بررسی یک نمونه در یک دوره زمانی کوتاه مدت	در یک مقطع زمانی خانم‌های یائسه از لحاظ شیوع سکنه قلبی و مصرف یا عدم مصرف هورمون مورد بررسی و طبقه‌بندی قرار می‌گیرند.
مطالعه مورد-شاهدی	بررسی دو گروه بر مبنای وقوع یا عدم وقوع پیامد و اندازه گیری مواجهه‌ها	محققین یک گروه از خانم‌های یائسه را که دچار حمله قلبی شده‌اند (موردها) با یک گروه سالم (بدون سکنه قلبی) از خانم‌های یائسه (شاهدها) از لحاظ مصرف یا عدم مصرف هورمون مقایسه می‌کنند.
	مداخله	۲- مطالعات تجربی:
کارآزمایی تصادفی و کور شده	بررسی دو گروه که به تصادف هورمون یا دارونما مصرف می‌کنند و از نوع درمان اعمال شده بی‌اطلاع هستند.	محققین به طور تصادفی به افراد مورد مطالعه هورمون یا دارونما تجویز می‌کنند و پس از پیشگیری هر دو گروه طی مدت چند سال میزان بروز حمله قلبی را بین دو گروه مقایسه می‌نمایند.

کاربرد انواع مختلف مطالعات مشاهده ای

اکولوژیک	مقطعی	مورد-شاهدی	همگروهی	
+++	-	++++	-	بررسی بیماری نادر
++	-	-	++++	بررسی مواجهه نادر
+	++	-	++++	بررسی چندین پیامد
++	++	++++	+++	بررسی چندین مواجهه
+	-	+	++++	بررسی زمان بین مواجهه تا پیامد
-	-	+	++++	اندازه گیری مستقیم بروز
-	-	+++	*+++	بررسی پیامدهای با دوره کمون طولانی

\*البته در مطالعه همگروهی تاریخی

نقاط قوت و ضعف انواع مختلف مطالعات مشاهده ای

همگروهی	مورد-شاهدی	مقطعی	اکولوژیک	
				احتمال وجود
کم	زیاد	متوسط	مورد ندارد	تورش انتخاب
کم	زیاد	زیاد	مورد ندارد	تورش اندازه گیری
زیاد	کم	مورد ندارد	مورد ندارد	خروج افراد از مطالعه
کم	متوسط	متوسط	زیاد	مخدوش کننده
زیاد	متوسط	متوسط	کم	مدت زمان لازم
زیاد	متوسط	متوسط	کم	هزینه بالا

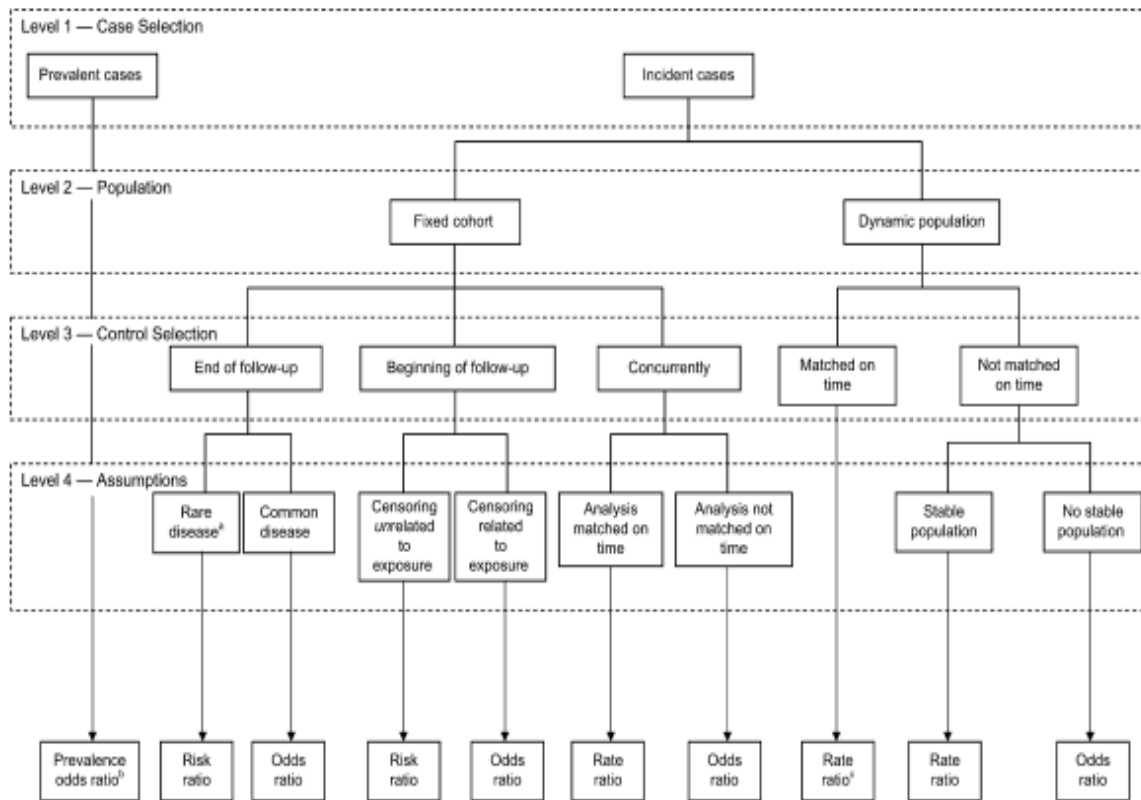
اندازه بیماری (پیامد) و اندازه اثر مواجهه

انواع مختلف مطالعات تحلیلی، اندازه های مختلفی از وقوع بیماری و یا اثر مواجهه را بررسی و گزارش می کنند. در جدول زیر مهمترین آنها بیان شده است.

نوع مطالعه	اندازه وقوع بیماری (پیامد)	اندازه اثر مواجهه
اکولوژیک	میزان، خطر، شیوع	ضریب همبستگی یا رگرسیون
مقطعی	شیوع	نسبت شیوع، اختلاف شیوع، نسبت شانس (Odds Ratio)
همگروهی	میزان، خطر، Odds، میانگین یا میانه	.Risk Ratio، .Rate Ratio، نسبت شانس، .Rate Difference، Risk Difference، اثربخشی واکسن، اختلاف بین میانگینها یا میانه ها
مورد-شاهدی	قابل اندازه گیری نیست <sup>۱</sup>	نسبت شانس، اثربخشی واکسن
مداخله ای	میزان، ریسک، Odds، میانگین یا میانه	.Risk Ratio، .Rate Ratio، نسبت شانس، .Rate Difference، Risk Difference، اثربخشی واکسن، اختلاف بین میانگینها یا میانه ها

<sup>۱</sup> مگر اینکه نسبت نمونه گیری (Sampling Fraction) هم برای گروه مورد و هم برای گروه شاهد معلوم باشد؛ یعنی نسبتی که موردها و شاهدها از جمعیت انتخاب شده اند، معلوم باشد.

## درخت تصمیم گیری در برآورد اندازه های رابطه از نسبت شانس



با توجه به چگونگی انتخاب مورد ها و شاهد ها ، جمعیت مورد بررسی (همگروه ثابت یا دینامیک) و داشتن پیش فرض درباره نادر بودن بیماری در جمعیت و یا ارتباط سانسور شده ها با مواجهه ، اندازه های رابطه برآورد می گردند .

هنگامی که **موردها** از موارد شیوع (موجود) انتخاب گردند نسبت شانس برآوردی از **نسبت شانس شیوع** خواهد بود.

هنگامی که **موردها** از موارد بروز (جدید) انتخاب گردند و جمعیت مورد بررسی **کوهورت ثابت** باشد اگر کنترل ها در **انتهای پی گیری** انتخاب گردند و فرض نادر بودن بیماری برقرار باشد نسبت شانس برآوردی از **Risk Ratio** است. اگر

بیماری شایع باشد (فرض نادر بودن برقرار نباشد) **نسبت شانس (Odds Ratio)** برآورد می شود. اگر کنترل ها در

**شروع پیگیری** انتخاب گردند و سانسور شده ها با مواجهه مرتبط نباشند برآورد **Risk Ratio** است. اگر سانسور شده ها

با مواجهه مرتبط باشند برآورد **نسبت شانس** است. چنانچه کنترل ها **همزمان** با موردها انتخاب شوند و همسان سازی بر

زمان انجام شود برآورد **نسبت میزان (Rate Ratio)** خواهد بود.

اگر همسان سازی بر زمان انجام نشود برآورد **نسبت شانس** است. هنگامی که موردها از موارد بروز انتخاب شوند و جمعیت

مورد بررسی دینامیک باشد ، اگر بر زمان همسان سازی انجام شود برآورد **Rate Ratio** خواهد بود ولی اگر همسان سازی

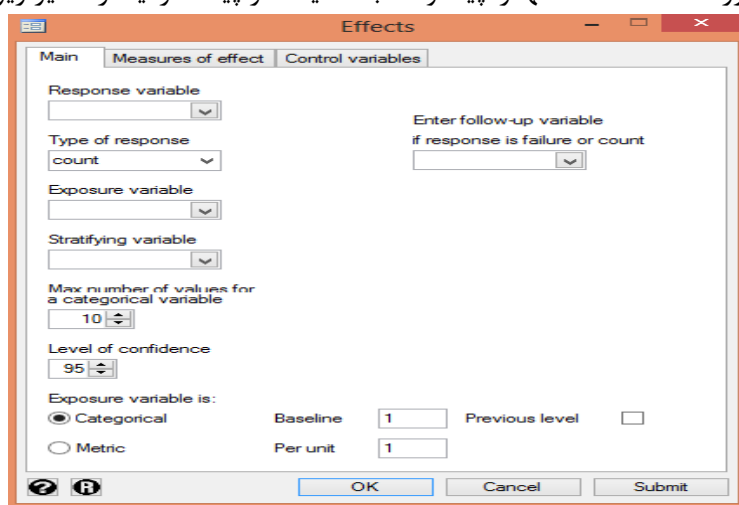
بر زمان انجام نشود و جمعیت پایدار باشد برآورد **Rate Ratio** خواهد بود.

اگر جمعیت پایدار نباشد برآورد **نسبت شانس** است. **در ادامه روش محاسبه این شاخص ها با نرم افزار خواهد آمد.**



### محاسبه شاخص های اندازه اثر و رابطه در مطالعات اپیدمیولوژیک:

در کاربرد آمار در اپیدمیولوژی، در اغلب موارد تعریف شاخص اثر، محاسبه شاخص اثر و حدود اطمینان آن، چگونگی محاسبه حدود اطمینان برای شاخص اثر مبتنی بر Bootstrap و چگونگی محاسبه این شاخص با نرم افزار در همه مطالعات اپیدمیولوژیک لازم و ضروری است. در طی این بخش به این سوالات پاسخ داده می شود. شاخص های اثر در مطالعات اپیدمیولوژیک را اکثراً با دستور db effect و کادر محاوره ای زیر و تعریف متغیرها می توان محاسبه نمود. این دستور را باید از اینترنت دانلود و نصب نمود. حدود ۶۲۸ پکیج برای نصب وجود دارد. می توانید با دستور findit effect آنها را پیدا و نصب نمایید. اگر پیدا نکردید از مسیر زیر استفاده کنید!!



Classical tests of hypotheses

در منوی

Summaries, tables, and tests

از مسیر Statistics و از قسمت

#### Effect size based on mean comparison

منوی را باز نمایید. در اینجا می توان متغیرها را معرفی و اندازه اثر را مقایسه نمود. در مطالعات مشاهده ای اطلاعات درباره رویدادهایی که نسبت به آنها کنترلی وجود ندارد، جمع آوری می شود. راه مفید تقسیم بندی مطالعات مشاهده ای، دسته بندی آنها به انواعی است که در آنها اطلاعات یا از گروهها و جمعیتها (داده های گروه بندی شده – Aggregated Data) و یا از تک تک افراد (Individual Data) کسب می شود. سپس اگر پیامد مورد بررسی به تنهایی و بدون توجه به مواجهه بررسی شود، مطالعه از نوع توصیفی است و اگر عوامل مرتبط با پیامد هم ارزیابی شوند، نوع مطالعه تحلیلی است. یا به عبارتی وقتی فرضیه ای مورد آزمون قرار گیرد مطالعه تحلیلی محسوب می شود. این راه مناسبی برای تقسیم بندی مطالعات است. اما گاهی آنها با هم همپوشانی دارند. بعضی از مطالعات مقطعی (Cross Sectional) یا همگروهی (Cohort) ممکن است هم اجزاء توصیفی داشته باشند و هم اجزاء تحلیلی. در مطالعات اکولوژیک، مهم است که بدانید نتیجه گیری باید تنها در سطح جمعیتها صورت گیرد و نه سطح فردی. این برداشت غلط از یافته های یک مطالعه اکولوژیک، که در سطح فردی ممکن است تفسیر انجام شود Ecological Fallacy نامیده می شود. مطالعات مقطعی معمولاً به آسانی قابل اجرا هستند، زمان کمی می برند و نسبتاً ارزان هستند. بنابراین، این نوع مطالعات به وفور، برای برنامه ریزی مبتنی بر شواهد، اجرا می شوند. اشکال عمده آنها این است که به سختی روابط علیتی در آنها قابل تفسیر است. دلیل آن هم این است که نمی توان ترتیب علیت و پیامد را از نظر زمانی مشخص نمود. یا نمی توان مشخص کرد که آیا پیامد بدنبال مواجهه بوقوع پیوسته است و یا اینکه این مواجهه است که بعد از پیامد بوجود آمده است؟!

کار با نرم افزار و محاسبه شاخص های اثر و رابطه در مطالعه مقطعی:

فایل epidemiology را باز نمایید. با فرض اینکه داده های استعمال سیگار بعنوان مواجهه در یک مطالعه مقطعی برای موارد فوت (all) جمع آوری گردیده است. یک متغیر جدید بنام smoking از متغیر smok بسازید. به این متغیر کد صفر و یک بدهید. یعنی کدهای ۱ و ۲ را صفر و کدهای ۳ تا ۵ را به ۱ تغییر دهید. سپس به متغیر ساخته شده برچسب no را مقادیر ۰ و مقادیر ۱ را با برچسب ۱ تعریف نمایید. خطر، اختلاف خطر و خطر منتسب برای سیگار کشیدن و احتمال مرگ (death) برای بیماری قلبی عروقی (chd) و کل مرگ ها (all) را محاسبه نمایید.

```
gen smoking= smok
recode smoking 1/2=0 3/5=1
label define smokingg 0"no" 1"yes"
label values smoking smoking
```

در خروجی زیر ۴۰۳ نفر و حدود ۲۴ درصد فوت نموده است. با دستور `tab all`

death from all causes	Freq.	Percent	Cum.
0	1,274	75.97	75.97
1	403	24.03	100.00
Total	1,677	100.00	

از بین افرادی که فوت نمودند، ۲۳۵ نفر یا حدود ۳۳ درصد سیگاری بود. نتایج آزمون کای اسکور نشان می دهد این دو متغیر با یکدیگر رابطه معنی دار آماری دارند.

```
. tabulate smoking all, chi2
```

smoking	death from all causes		Total
	0	1	
0	795	168	963
1	479	235	714
Total	1,274	403	1,677

Pearson chi2(1) = 53.7321 Pr = 0.000

نسبت مرگ ها در سیگاری ها و در کل جمعیت مورد بررسی در زیر محاسبه شد.

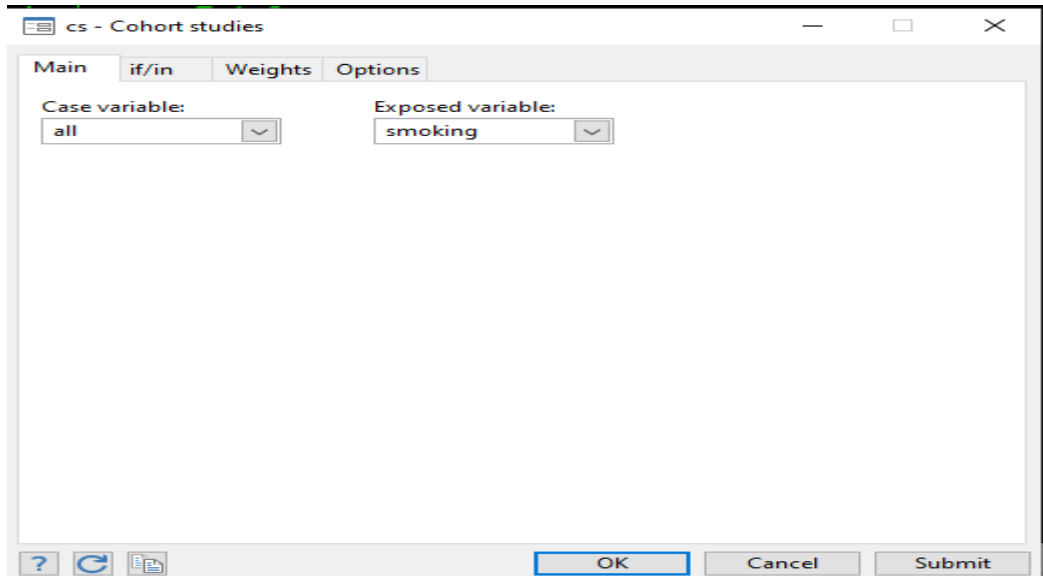
```
. tabulate smoking all, cell chi2 row
```

Key
frequency
row percentage
cell percentage

smoking	death from all causes		Total
	0	1	
0	795	168	963
	82.55	17.45	100.00
	47.41	10.02	57.42
1	479	235	714
	67.09	32.91	100.00
	28.56	14.01	42.58
Total	1,274	403	1,677
	75.97	24.03	100.00
	75.97	24.03	100.00

Pearson chi2(1) = 53.7321 Pr = 0.000

مقادیر خطر، اختلاف خطر و خطر منتسب مواجهه و خطر منتسب در جمعیت و حدود اطمینان آنها در زیر آمده است. این محاسبات را با کادر محاوره ای و دستور db cs نیز می توان انجام داد. با دستور db cs می توان کادر محاوره ای زیر را باز و شاخص های مورد نظر را محاسبه نمود.



در بین کل افراد نیز ۷۱۴ نفر سیگاری بوده است. اختلاف خطر مرگ در افراد سیگاری و غیر سیگاری حدود ۱۵ درصد بود. نسبت خطر  $1/89$  و نسبت مرگ های منتسب به سیگار به عنوان مواجهه ۴۶٪ و در کل جمعیت ۲۷٪ گزارش گردید. به عبارت دیگر ۴۶ درصد از مرگ ها در کل افراد سیگاری به سیگار کشیدن منتسب است و ۲۷ درصد از کل مرگ ها در کل جامعه به سیگار کشیدن منتسب است.

### نکته:

کسر قابل انتساب به جامعه تابع شیوع مواجهه در جامعه است. لذا زمانی قابلیت استفاده و استناد دارد که نسبت افراد مواجهه یافته و مواجهه نیافته در مطالعه با نسبت افراد مواجهه یافته و مواجهه نیافته در جامعه برابر باشد در غیر این صورت این شاخص کاربرد ندارد. بنابراین در این مطالعه مقطعی این شاخص ها با احتیاط باید تفسیر گردند. در مطالعه مقطعی شیوع برآورد می گردد و تقدم و تاخر علت و معلول مشخص نیست.

## روش های اپیدمیولوژی و پژوهش با Stata

. tab all smoking, col

Key
<i>frequency</i>
<i>column percentage</i>

death from all causes	smoking		Total
	0	1	
0	795 82.55	479 67.09	1,274 75.97
1	168 17.45	235 32.91	403 24.03
Total	963 100.00	714 100.00	1,677 100.00

. db cs

. cs all smoking

	smoking		Total
	Exposed	Unexposed	
Cases	235	168	403
Noncases	479	795	1274
Total	714	963	1677
Risk	.3291317	.1744548	.2403101
	Point estimate		[95% conf. interval]
Risk difference	.1546768		.112695    .1966586
Risk ratio	1.88663		1.587309    2.242394
Attr. frac. ex.	.4699543		.3700029    .554048
Attr. frac. pop	.2740428		

chi2(1) = 53.73 Pr>chi2 = 0.0000

در این مطالعه مقطعی نسبت شانس مرگ برای متغیرهای مستقل به صورت زیر محاسبه می شود:

. logistic all smoking sbp chol agein

Logistic regression

Number of obs = 1,677

LR chi2(4) = 318.15

Prob > chi2 = 0.0000

Log likelihood = -765.68268

Pseudo R2 = 0.1720

all	Odds ratio	Std. err.	z	P> z	[95% conf. interval]
smoking	2.356227	.302545	6.67	0.000	1.83198    3.030493
sbp	1.015865	.0030283	5.28	0.000	1.009947    1.021818
chol	1.003649	.0013861	2.64	0.008	1.000936    1.00637
agein	1.141149	.0119755	12.58	0.000	1.117917    1.164864
_cons	9.74e-06	7.29e-06	-15.41	0.000	2.25e-06    .0000422

نسبت شانس مرگ در افراد سیگاری نسبت به افراد غیرسیگاری ۲/۳۵ برابر است.

## مطالعات همگروهی (کوهورت):

یک مطالعه همگروهی (مطالعه طولی یا مطالعه پیگیری)، مطالعه مشاهده ای است که در آن گروهی از افراد مواجهه یافته (exposed) به همراه گروهی از افراد مواجهه نیافته (non-exposed) به یک عامل بالقوه بیماری، در طول یک دوره زمانی پیگیری می شوند و بروز (incidence) بیماری در دو گروه با یکدیگر مقایسه می گردد. برای آشنایی با نحوه طراحی و اجرای مطالعات کوهورت، به گزارش مطالعه کوهورت پرشین شهرکرد در این لینک مراجعه نمایید.

به منظور بررسی رابطه بین مصرف سیگار و سکته قلبی ۳۰۰۰۰ فرد سیگاری و ۳۰۰۰۰ فرد غیر سیگاری وارد یک مطالعه هم گروهی و به مدت ۵ سال پی گیری شدند. هیچکدام از افراد در ابتدای ورود به مطالعه مبتلا به سکته قلبی نبودند. تعداد افراد مبتلا پس از ۵ سال به قرار زیر است:

مبتلا به سکته قلبی	غیر مبتلا به سکته قلبی	
۳۹	۲۹,۹۶۱	افراد سیگاری
۳	۲۹,۹۹۷	افراد غیر سیگاری

مصرف سیگار در این مطالعه به عنوان **ریسک فاکتور (Risk Factor)** یا **مواجهه (Exposure)** شناخته شده و سکته قلبی به عنوان **پیامد (Outcome)** می باشد. پیامد پس از مواجهه رخ داده است، بنابراین می توان در مورد رابطه مصرف سیگار با افزایش خطر ابتلا به سکته قلبی بحث کرد. رابطه مشاهده شده زمانی بیشتر اهمیت دارد که این دو گروه (افراد سیگاری و غیر سیگاری) از لحاظ دیگر عوامل افزایش دهنده خطر ابتلا به سکته قلبی نیز یکسان باشند.

استفاده از اصطلاح مطالعات گذشته نگر و آینده نگر تا حدی گمراه کننده است. مطالعات مورد-شاهدی گذشته نگر هستند چون از نظر جهت زمانی به عقب بر می گردند و از بیماری به سمت مواجهه نگاه می کنند، در حالیکه مطالعات همگروهی مطالعات آینده نگری هستند که جهت آنها از مواجهه به سمت بیماری است. اما از لحاظ زمانی، مطالعات مورد-شاهدی می توانند داده هایشان را به صورت آینده نگر، نیز جمع آوری کنند و مطالعات همگروهی می توانند از داده های تاریخی (داده های مربوط به گذشته) و یا نمونه های سرمی یا بافتی ذخیره شده برای تعریف مواجهه استفاده کنند.

بنابراین، به طور مثال در پیگیری ۱۰۰۰۰۰ کارمند دفتری، افرادی که مبتلا به یک سرطان نادر هستند، می توانند به عنوان گروه مورد و آنهایی که مبتلا نشده اند، به عنوان گروه شاهد در یک مطالعه دیگر وارد شوند. سپس مواجهه مورد نظر در این زیرگروه (به جای تمام صد هزار نفر) بررسی خواهد شد و از این طریق می توان هزینه بسیاری را صرفه جویی کرد. اگرچه در این نوع مطالعه افراد در طول زمان وارد مطالعه می شوند، اما یک مطالعه مورد-شاهدی است که **مطالعه مورد-شاهدی لانه گزیده (nested case control study)** نامیده می شود. در این نوع از مطالعات هر زمان که یک مورد در بین کوهورت مورد نظر رخ داد از جمعیت در معرض خطر یک نفر یا بیشتر به عنوان شاهد یا کنترل انتخاب می شود.

ایده آل است که گروههای مواجهه یافته و مواجهه نیافته از تمامی جنبه ها به جز خود مواجهه شبیه هم باشند. اما در واقعیت این کار چندان عملی نیست. به طور مثال، در بررسی رابطه بین مصرف سیگار و سرطان معده، افراد سیگاری ممکن است دارای دیگر ریسک فاکتورهای سرطان معده نیز باشد. برای مثال، ممکن است رژیم های غذایی متفاوتی داشته باشند و یا نسبت به افراد غیر سیگاری، محل کار متفاوتی داشته باشند. یک متغیر مخدوش کننده (Confounding Variables) متغیری است که با مواجهه مورد نظر رابطه دارد و به طور مستقل خطر ابتلا به پیامد مورد نظر را تحت تاثیر قرار می دهد. بنابراین در اینجا، رژیم غذایی، قومیت، محل کار و مصرف زیاد تریاک، می توانند مخدوش کننده باشند در صورتی که هم با مصرف سیگار و هم با سرطان معده رابطه داشته باشند. مصرف زیاد تریاک که اغلب با مصرف سیگار همراه است تنها در صورتی مخدوش کننده خواهد بود که خطر ابتلا به سرطان معده را تحت تاثیر قرار دهد. همچنین، مواجهه با الکل و مشروبات مخدوش کننده نخواهد بود، در صورتی که تعداد برابری از افراد سیگاری و غیر سیگاری با آن مواجهه داشته باشند.

برای مثال، داده های فرضی بررسی رابطه الکل و سکته قلبی به صورت زیر بیان شده اند:

عدم ابتلا به سکته قلبی	مبتلا به سکته قلبی	
۵۲(۴۲٪)	۷۱(۵۸٪)	افرادی که الکل مصرف می کنند
۴۸(۶۲٪)	۲۹(۳۸٪)	افرادی که الکل مصرف نمی کنند

همانطور که نشان داده شده است، خطر سکته قلبی در گروه مصرف کنندگان الکل بیشتر است. اما معلوم است افرادی که الکل مصرف می کنند، تمایل بیشتری به مصرف سیگار دارند و مصرف سیگار با سکته قلبی رابطه دارد؛ بنابراین، زمانی که روابط بالا طبقه بندی می شوند:

در افراد غیر سیگاری

عدم ابتلا به سکته قلبی	مبتلا به سکته قلبی	
۱۶(۵۰٪)	۸(۵۰٪)	افرادی که الکل مصرف می کنند
۲۲(۵۰٪)	۲۲(۵۰٪)	افرادی که الکل مصرف نمی کنند

در افراد سیگاری

عدم ابتلا به سکته قلبی	مبتلا به سکته قلبی	
۳۶(۳۶٪)	۶۳(۶۴٪)	افرادی که الکل مصرف می کنند
۴(۳۶٪)	۷(۶۴٪)	افرادی که الکل مصرف نمی کنند

بنابراین، در اینجا شواهدی برای اینکه مصرف الکل خطر سکته قلبی را در بین افراد سیگاری و غیر سیگاری افزایش می دهد، وجود ندارد.

در مطالعات کوهورت تحلیل داده ها معمولاً با محاسبه میزان بروز (تعداد پیامد تقسیم بر شخص-زمان در معرض خطر : Incidence Rate) در گروههای مواجهه یافته و مواجهه نیافته و مقایسه آنها با یکدیگر صورت می گیرد. در این حالت، خطر نسبی محاسبه شده از نوع Rate Ratio است که در زیر آورده شده است:

تعداد موارد بیماری	شخص- زمان در معرض خطر
a	$Pyr_1$
c	$Pyr_0$

میزان بروز در گروه مواجهه  $(R_1) = a / Pyr_1$   
 میزان بروز در گروه مواجهه نیافته  $(R_0) = c / Pyr_0$   
 نسبت میزان بروز و تفاوت میزان به صورت زیر بدست می آید.

$$\text{Rate Ratio} = R_1 / R_0$$

$$\text{Rate Difference} = R_1 - R_0$$

این جداول ۲×۲ می توانند گسترش یابند و مواجهه های بیش از دو حالت را نیز شامل شوند. اگر از سطوح مختلف مواجهه در مطالعه همگروهی استفاده شود، می توان روند (Trend) بروز بیماری برای مقادیر مختلف مواجهه را نیز بررسی کرد. زمانی که در مورد متغیرهای مخدوش کننده اطلاعاتی در دسترس باشد، تطبیق مقدار خطر نسبی با توجه به متغیرهای مخدوش کننده لازم است.

### خطر نسبی (Relative Risk):

شاخص های مختلفی برای بیان خطر نسبی وجود دارند. این شاخصها مقدار بزرگی رابطه بین مواجهه و بیماری را نشان می دهند و بیان می کنند که احتمال ایجاد بیماری در گروه مواجهه یافته چند برابر گروه مواجهه نیافته است.

سه نوع شاخص نسبی (Relative Measures) قابل محاسبه است :

خطر (بروز تجمعی) در گروه مواجهه یافته	= نسبت خطر (Risk Ratio)
خطر (بروز تجمعی) در گروه مواجهه نیافته	
میزان بروز در گروه مواجهه یافته	= نسبت بروز (Rate Ratio)
میزان بروز در گروه مواجهه نیافته	
شانس بیماری در گروه مواجهه یافته	= نسبت شانس (Odds Ratio)
شانس بیماری در گروه مواجهه نیافته	

در مجموع، به این شاخصها، شاخصهای خطر نسبی (Relative Risk) گفته می شود. خطر نسبی به عنوان شاخصی برای نشان دادن شدت ارتباط (aetiological strength) بکار برده می شود. مقدار ۱ نشان می دهد که بروز بیماری در گروه های مواجهه یافته و بدون مواجهه یکسان است و بنابراین، ارتباطی بین مواجهه و بیماری بر اساس این داده ها مشاهده نمی شود. مقدار بزرگتر از ۱ نشاندهنده وجود یک رابطه مستقیم یا افزایش خطر ابتلا در بین مواجهه

یافتگان با آن عامل است. همچنین، خطر نسبی کمتر از ۱ به معنی وجود رابطه معکوس یا کاهش خطر ابتلا در مواجهه یافتگان است، یعنی مواجهه اثر محافظت کننده دارد. در عمل، در شرایطی که احتمال وقوع پیامد مورد نظر اندک باشد (مانند بیشتر سرطانها، نقایص مادرزادی و...)، سه شاخص اندازه گیری اثر، برآوردهای یکسانی از خطر نسبی را نشان می دهند. اما در مواقعی که پیامد مورد بررسی، امری شایع باشد (مانند اکثر بیماریهای عفونی)، برآوردهای حاصل از این سه شاخص بطور قابل ملاحظه ای با هم تفاوت دارند. هر سه شاخص مربوط به خطر نسبی (نسبت خطر، نسبت بروز و نسبت شانس) در **مطالعات همگروهی (Cohort Studies)** که میزان بروز بیماری در گروه مواجهه یافته و بدون مواجهه معلوم است، قابل محاسبه هستند. در **مطالعات مورد-شاهدی (Case-control studies)**، انتخاب افراد براساس وضعیت بیماری آنها انجام می شود (افراد مبتلا به بیماری مورد نظر = case و افراد سالم = control) و نه بر اساس وضعیت مواجهه؛ بنابراین، امکان محاسبه بروز بیماری در گروه مواجهه و بدون مواجهه وجود ندارد. علی رغم این محدودیت، امکان محاسبه **شانس مواجهه (Odds of Exposure)** در گروه مورد و **شانس مواجهه** در گروه شاهد و بدست آوردن **نسبت شانس مواجهه (Odds Ratio of Exposure)** وجود دارد. می توان نشان داد که نسبت شانس مواجهه برابر است با نسبت شانس بیماری. به همین دلیل است که از نسبت شانس مواجهه برای برآورد خطر نسبی در مطالعات مورد-شاهدی استفاده می شود. اطلاع از خطر نسبی به تنهایی تمامی ابعاد رابطه بین مواجهه و بیماری را نشان نمی دهد. در جدول زیر خطر نسبی (که با **نسبت بروز** نشان داده شده) بیماری A و B در مواجهه با یک عامل خطر واقعی نشان داده شده است. اگرچه در هر دو حالت نسبت بروز برابر با ۲ است، اما میزان بروز بیماری A از ۵ به ۱۰ در صد هزار نفر-سال در معرض خطر افزایش می یابد، درحالیکه میزان بروز بیماری B از ۴۰ به ۸۰ در صد هزار نفر-سال در معرض خطر زیاد می شود. مشخص است که مقدار اثر مطلق مواجهه یا **خطر مطلق (Absolute Risk)** برای این دو بیماری متفاوت است.

بیماری B	بیماری A	
۸۰ در صد هزار نفر-سال	۱۰ در صد هزار نفر-سال	میزان بروز در گروه مواجهه یافته
۴۰ در صد هزار نفر-سال	۵ در صد هزار نفر-سال	میزان بروز در گروه مواجهه نیافته
۲/۰	۲/۰	نسبت بروز
۴۰ در صد هزار نفر-سال	۵ در صد هزار نفر-سال	اختلاف بروز

**خطر قابل انتساب (attributable risk):** خطر منتسب شاخصی است برای بیان تاثیر یک مواجهه و با مقیاس جبری (absolute scale) نشان می دهد که فراوانی بیماری در گروه مواجهه چقدر بیشتر از گروه مواجهه نیافته است، با این پیش فرض که رابطه بین مواجهه و بیماری یک رابطه علیتی است. این شاخص را می توان از اختلاف بروز تجمعی (risk difference) یا میزان بروز (rate difference) دو گروه مواجهه یافته و بدون مواجهه بدست آورد. خطر مطلق (risk diff.) = خطر در گروه مواجهه یافته - خطر در گروه بدون مواجهه

بروز مطلق (rate diff.) = میزان بروز در گروه مواجهه یافته - میزان بروز در گروه بدون مواجهه



خطر قابل انتساب حداکثر انتظار ما را در کاهش ابتلا به بیماری، در صورتی که قادر به حذف کامل عامل مواجهه باشیم، نشان می دهد. در مثال بالا، شدت همبستگی عامل مواجهه (به عنوان یک فاکتور علیتی) با بیماری با نسبت بروز بیان شده و برای هر دو بیماری A و B یکسان است. اما از نگاه بهداشتی، اهمیت مواجهه در ارتباط با بیماری B بسیار بیشتر است زیرا با حذف مواجهه موارد بیشتری از بیماری B، نسبت به بیماری A، قابل پیشگیری خواهند بود. برخلاف خطر نسبی، نمی توان خطر قابل انتساب را به جوامع دیگر تعمیم داد، چرا که کاملاً وابسته به مقدار بروز در گروه مواجهه نیافته است که در جوامع مختلف تغییر می کند.

### کار با نرم افزار و محاسبه شاخص های اثر و رابطه در مطالعه کوهورت:

محاسبات مربوط به میزان و نسبت میزان در فصل گذشته با نرم افزار محاسبه گردید. برای این محاسبات مطالعه باید کوهورت باشد. چنانچه در فایل epidemiology داده ها را به صورت کوهورت در نظر بگیرید. ابتدا باید تنظیمات مربوط به زمان با دستور زیر که قبلاً نیز مفصل بحث گردید انجام شود.

```
stset timeout, id(id) failure(chd==1) enter(timein) origin(timebth) scale(365.25)
```

```
. stset timeout, id(id) failure(chd==1) enter(timein) origin(timebth) scale(365.25)
```

Survival-time data settings

```

ID variable: id
Failure event: chd==1
Observed time interval: (timeout[_n-1], timeout]
Enter on or after: time timein
Exit on or before: failure
Time for analysis: (time-origin)/365.25
Origin: time timebth

```

```

1,677 total observations
0 exclusions

```

```

1,677 observations remaining, representing
1,677 subjects
154 failures in single-failure-per-subject data
27,605.371 total analysis time at risk and under observation
At risk from t = 0
Earliest observed entry t = 40.052
Last observed exit t = 85.76587

```

میزان مرگ قلبی عروقی (chd) در هزار نفر را برای کسانی که سیگار (smoking) می کشند و کسانی که سیگار نمی کشند را محاسبه نمایید. متغیر سیگار به عنوان یک متغیر دوحالتی در تمرین قبلی ساخته شد.

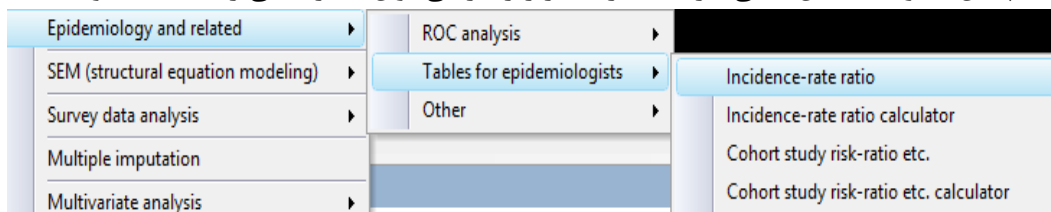
```
strate smoking, graph per(1000)
```

smoking	D	Y	Rate	Lower	Upper
no	66	16.3626	4.0336	3.1689	5.1341
yes	88	11.2427	7.8273	6.3514	9.6460

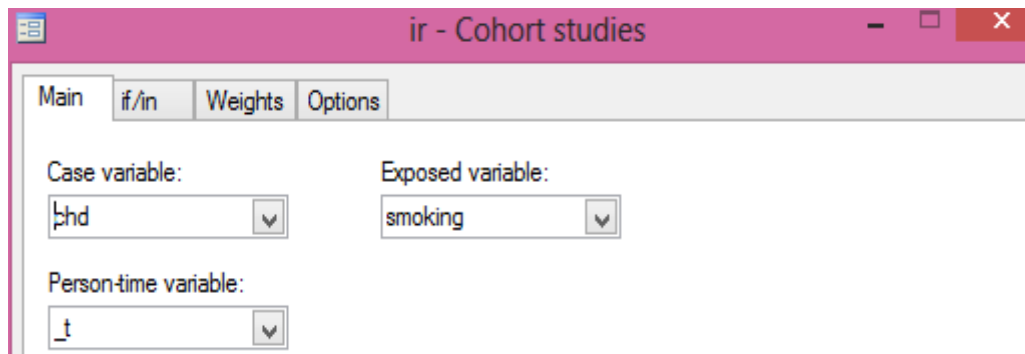
خطر نسبی یا نسبت دو میزان را محاسبه نمایید. با دستور `stmh smoking` می توان خطر نسبی را محاسبه کرد. نتیجه در زیر آمده است.

RR	chi2	P>chi2	[95% Conf. Interval]	
1.941	17.19	0.0000	1.410	2.670

خطر نسبی در جدول بالا ۱/۹۴۱ گزارش شده است. نتیجه نشان می دهد که خطر بیماری قلبی در کسانی که سیگار می کشند نسبت به کسانی که سیگار نمی کشند ۱/۹۴۱ بار بیشتر است. این مقدار با توجه به مقدار معنی داری و همچنین حدود اطمینان معنی دار است. از مسیر زیر نیز می توان خطر نسبی را محاسبه نمود.



یا با دستور `db ir` نیز می توان کادر محاوره ای مزبور را باز نمود. در فایل مذکور متغیر `event` را `chd` و متغیر مواجهه را `smoking` و متغیر شخص زمان را `_t` قرار دهید.



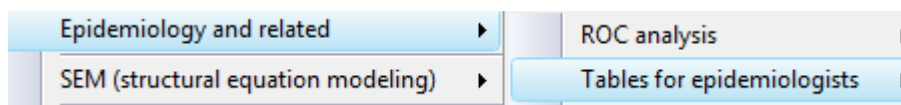
خروجی به صورت زیر نمایش داده می شود.

`. ir chd smoking _t`

	smoking		Total
	Exposed	Unexposed	
death from chd _t	88 48985.08	66 65992.12	154 114977.2
Incidence rate	.0017965	.0010001	.0013394
	Point estimate		[95% Conf. Interval]
Inc. rate diff.	.0007963		.0003501 .0012426
Inc. rate ratio	1.796251		1.290665 2.510391 (exact)
Attr. frac. ex.	.4432849		.2252058 .6016557 (exact)
Attr. frac. pop	.2533057		
	(midp) Pr(k>=88) =		0.0002 (exact)
	(midp) 2*Pr(k>=88) =		0.0003 (exact)

بر اساس نتایج بالا خطر ابتلا به بیماری قلبی عروقی در افراد سیگاری  $1/79$  برابر افراد غیر سیگاری است. حدود اطمینان این نسبت نیز معنی دار و نشان داده شده است. خطر قابل انتساب مواجهه حدود ۴۵ درصد است. یعنی تقریباً ۴۵ درصد از موارد وقوع بیماری قلبی عروقی در افراد سیگاری به دلیل مصرف سیگار بوده است. خطر قابل انتساب به جمعیت حدود ۲۵ درصد است. یعنی ۲۵٪ از مرگ های قلبی عروقی در جمعیت به سیگار کشیدن منتسب است. این محاسبات را با داشتن اعداد متغیر مواجهه و عدم مواجهه و شخص زمان در معرض خطر به صورت زیر با دستور db irri نیز می توان انجام داد.

کادر محاوره ای بالا با دستور db csi قابل دسترسی است. کادرهای محاوره ای فوق الذکر همه از مسیر زیر قابل دسترسی هستند.



## استانداردسازی مستقیم و غیرمستقیم میزان ها با نرم افزار Stata

استاندارد کردن یکی از روش های تطبیق یا Adjustment متغیر سن برای مقایسه دوجمعیت با ساختار سنی متفاوت است. یعنی حذف اثر یک عامل مخدوش کننده احتمالی که با روش های مختلفی این کار انجام می شود. اگر میزان های مرگ دو جمعیت را بخواهیم مقایسه کنیم، میزان های خام مرگ یک عدد را به ما می دهد، میزان های اختصاصی هم یک عدد، که البته مقایسه میزان های اختصاصی بهتر است اما به دلیل تعدد میزان های اختصاصی، مقایسه آن ها مشکل است.

در جدول زیر میزان های خام جمعیت اول تقریباً ۳ برابر جمعیت دوم است. وقتی که میزان های خام با هم تفاوت داشته باشند در واقع علت مهم میتواند تفاوت ساختار هرم سنی و جنسی دو جمعیت باشد و برای فراهم شدن امکان مقایسه، باید استانداردسازی صورت گیرد و در واقع با یک ترازوی واحد شاخص ها مقایسه بشود تا واقعیت را ببینیم. اگر میانگین وزنی میزان های اختصاصی را حساب کنیم، به عبارتی استانداردسازی انجام می شود که برای این کار ما باید هر دو جمعیت را با یک ترازو (یک جمعیت استاندارد) میزان کنیم و مورد انتظارها را به دست آوریم. جمعیت استاندارد ترجیحاً باید جمعیتی باشد که دو جمعیت اول و دوم که می خواهند مقایسه شوند را در بر گیرد یا مجموع دو تا جمعیت را می توانیم به عنوان جمعیت استاندارد بگیریم و نباید از جمعیت ثالث استفاده شود. حالا بسته به این که کدام اطلاعات را در جمعیت استاندارد داشته باشیم از روش های استاندارد مستقیم و غیر مستقیم استفاده می کنیم؛ اگر فقط جمعیت استاندارد را داشتیم از روش مستقیم و اگر میزان های اختصاصی جمعیت استاندارد را داشتیم از روش غیر مستقیم استفاده می شود. جدول زیر را مشاهده کنید.

گروه سنی	میزان مرگ اختصاصی سنی A	میزان مرگ اختصاصی سنی B	جمعیت استاندارد	مرگ های مورد انتظار A	مرگ های مورد انتظار B
کمتر از ۱۵	۱/۱۹	۱/۵۹	۲۳۹۶۱۰۰۰	۲۸۵۱۳/۵۹	۳۸۰۹۷/۹
۱۵-۲۴	۰/۷۱	۰/۹	۱۵۴۲۰۰۰۰	۱۰۹۴۸۲	۱۳۸۷۸
۲۵-۴۴	۱/۵۶	۱/۱۳	۲۱۳۵۳۰۰۰	۳۳۳۱۰/۶۸	۷۴۱۸۸۱
۴۵-۶۴	۷/۱۲	۶/۰۲	۱۹۶۰۹۰۰۰	۱۳۹۶۱۶/۰۸	۱۱۸۰۴۶
بالای ۶۵	۳۹/۱۱	۳۹	۱۰۶۸۵۰۰۰	۴۱۷۸۹۰/۳۵	۴۱۶۷۱
CDR	۸/۹۲	۲/۶۷	۹۱۰۲۸۰۰۰	۶۳۰۲۷۸/۹	۶۱۰۸۶۶/۰۶

### روش مستقیم استاندارد سازی

در مثال بالا ما فقط جمعیت استاندارد را داریم و بنابراین با حاصلضرب میزان های مرگ اختصاصی سنی هر جمعیت در جمعیت استاندارد، مرگ های مورد انتظار آن دو جمعیت را به دست می آوریم و سپس با تقسیم جمع مرگ های مورد انتظار هر جمعیت بر جمعیت استاندارد، می توانیم مرگ های دو جمعیت را مقایسه کنیم.

$\text{جمعیت دوم} = \frac{610866/06}{91028000} = 6/71$	$\text{جمعیت اول} = \frac{630278/9}{91028000} = 6/92$
--	---

روش غیر مستقیم استاندارد سازی

هنگامی که میزان های اختصاصی در جمعیت های مورد مقایسه وجود ندارد و فراوانی مرگ های مشاهده شده در دسترس است و میزان های اختصاصی جمعیت استاندارد در دسترس و وجود دارد، از روش استاندارد سازی غیرمستقیم استفاده می شود که روش دقیقی نیست ولی می تواند جایگزین روش مستقیم شود. جدول زیر را مشاهده نمائید.

مرگ های مورد انتظار برای جمعیت B	مرگ های مورد انتظار برای جمعیت A	میزان های اختصاصی مرگ در جمعیت استاندارد	جمعیت B	جمعیت A	گروه سنی
۴۹۷۹۹/۷۶	۱۵۳۲۲۹	۱/۳۴	۳۷۱۶۴	۱۱۴۳۵۰	۱۵ >
۱۱۶۲۰/۸۸	۴۶۵۵۰/۲۲	۰/۵۸	۲۰۰۳۶	۸۰۲۵۹	۲۴-۱۵
۴۵۷۷۰/۲	۱۸۶۸۱۶	۱/۴	۳۲۶۹۳	۱۳۳۳۴۰	۴۴-۲۵
۱۰۶۷۲۱/۵۸	۱۰۱۸۶۶۳۸	۷/۱۴	۱۳۹۴۷	۱۴۲۶۷۰	۶۴-۴۵
۱۰۲۸۳۲/۲۷	۴۵۶۳۲۳۷/۶۸	۴۹/۵۱	۲۰۷۷	۹۲۱۶۸	بالای ۶۵
۲۱۶۷۴۴/۶۹	۵۹۶۸۴۹۶/۷		۱۰۵۹۱۷	۵۶۲۷۸۷	جمع

$$SMR = \frac{\text{مرگ های مشاهده شده}}{\text{مرگ های مورد انتظار}} = \frac{5022}{5968496/7} = 0/0008$$

$$SMR_{B \text{ جمعیت}} = \frac{285}{316743/69} = 0/0009$$

اگر نسبت مرگ استاندارد شده (SMR) برای جمعیت ها را در میزان مرگ و میر خام جمعیت استاندارد ضرب کنیم. میزان های استاندارد شده غیر مستقیم مرگ و میر دو تا جمعیت را محاسبه می شود.

$$SDR_{A \text{ جمعیت}} = SMR_{A \text{ جمعیت}} \times CDR_{\text{جمعیت استاندارد}} = 6/84$$

$$SDR_{B \text{ جمعیت}} = SMR_{B \text{ جمعیت}} \times CDR_{\text{جمعیت استاندارد}} = 7/32$$

## روش های اپیدمیولوژی و پژوهش با Stata

**مثال:** استاندارد کردن مستقیم و غیر مستقیم را در جمعیت های زیر بکار ببرید. مقایسه و تفسیر کنید.

میزان مرگ اختصاصی B	موارد مرگ (B)	جمعیت B	میزان مرگ اختصاصی A	موارد مرگ (A)	جمعیت A	گروه سنی
۰/۰۰۲	۱۰	۵۰۰۰	۰/۰۰۲	۲	۱۰۰۰	۴۹-۴۰
۰/۰۰۴	۴	۱۰۰۰	۰/۰۰۴	۲۰	۵۰۰۰	۵۹-۵۰
۰/۰۰۲	۱۴	۶۰۰۰	۰/۰۰۳	۲۲	۶۰۰۰	جمع

میزان های اختصاصی	موارد مرگ	جمعیت استاندارد	گروه سنی
۰/۰۰۱	۳۰	۳۰۰۰۰	۴۹-۴۰
۰/۰۰۳	۱۲۰	۴۰۰۰۰	۵۹-۵۰
۰/۰۰۲	۱۵۰	۷۰۰۰۰	جمع

میزان های خام مرگ در جمعیت A بیشتر از این میزان (CDR = ۰/۰۰۲) در جمعیت B است.

**الف) استاندارد سازی مستقیم:**

جمعیت B	جمعیت A	گروه سنی
$30000 \times 0/002 = 60$	$30000 \times 0/002 = 60$	۴۹-۴۰
۱۶۰	۱۶۰	۵۹-۵۰
۲۲۰	۲۲۰	جمع

$$A \text{ جمعیت} = \frac{220}{70000} = 0/003$$

و برای جمعیت B ؟

**ب) استاندارد سازی غیر مستقیم:**

جمعیت B	جمعیت A	گروه سنی
مرگ های مورد انتظار	مرگ های مورد انتظار	
۵	$1000 \times 0/001 = 1$	۴۹-۴۰
۳	$5000 \times 0/003 = 15$	۵۹-۵۰
۸	۱۶	جمع

$$SMR_B = \frac{14}{8} = 1/75 \quad SMR_A = \frac{\text{مرگ های مشاهده شده}}{\text{مرگ های مورد انتظار}} = \frac{22}{16} = 1/375$$

$$SDR_A = SMR_A \times CDR_{\text{standard}} = 0/00275$$

$$SDR_A = 1/375 \times 0/002 = 0/00275$$

$$SDR_B = 0/0035$$

سوال: آیا آزمون آماری وجود دارد که روش های استاندارد کردن مستقیم و غیر مستقیم را بشود در دو جمعیت با هم مقایسه کرد؟ بله. نوع آزمون مورد استفاده را پیدا و گزارش کنید!

روش استاندارد کردن (ASCAR): Age Standardized Cancer Ratio

روش استاندارد کردن

هنگامی که در مرکز ثبت سرطان فقط تعداد موارد سرطان بر حسب سن و جنس و نوع سرطان ثبت می شود، می توانیم برای مقایسه ثبت آن ها با جمعیت استاندارد (درصد مورد انتظار سرطان در گروه های سنی) که توسط آقای Parkin پیشنهاد شده است، استاندارد سازی را انجام دهیم. مثال: در مرکز ثبت تومور استان چهارمحال و بختیاری در طی سال های ۸۰-۱۳۷۶ مطالعه ای در مورد ثبت لوکمیا انجام گردید که به شرح زیر است.

نسبت مورد انتظار (۳×۴)	انتظار می رود چند درصد سرطان ها در این گروه سنی باشند(۴)	نسبت سرطان لوکمیا به کل سرطان ها(۳)	تعداد کل سرطان ها به جز سرطان های پوست	تعداد لوکمیا ثبت شده	گروه های سنی
۳/۰۷۵	۵	۰/۶۱۵	۲۳۴	۱۴۴	۱۴-۰
۲/۴۴	۵	۰/۴۸۸	۲۱۵	۱۰۵	۲۴-۱۵
۱/۱۷	۵	۰/۲۳۴	۱۵۴	۳۶	۳۴-۲۵
۱/۶۸	۱۰	۰/۱۶۸	۲۷۳	۴۶	۴۴-۳۵
۲/۶۲	۲۰	۰/۱۳۱	۲۹۸	۳۹	۵۴-۴۵
۲/۳۷۵	۲۵	۰/۰۹۵	۲۰۱	۱۹	۶۴-۵۵
۱	۲۰	۰/۰۵۰	۱۸۰	۹	۷۴-۶۵
۰	۱۰	۰	۶۱	۰	۷۵>
٪ ۱۳/۱۹	۱۰۰	٪ ۲۴/۵ یا ۰/۲۴۵	۱۶۱۹	۳۹۷	جمع

اگر نسبت مورد انتظار ما کمتر از نسبت مشاهده شده باشد، ثبت سرطان خون بیشتر از مورد انتظار است. که از دلایل آن: یا سرطان خون (لوکمیا) ثبت مناسبی دارد یا ممکن است بروز سرطان لوکمیا در جامعه ما زیاد باشد و یا ممکن سرطان های دیگر خوب ثبت نشوند. بنابراین ASCAR سرنخ هایی را به ما می دهد درباره ی این که ثبت هر سرطان در جامعه چگونه است؟ و این که کدام سرطان بیشتر و ثبت آن ها چگونه است؟

ASCAR را باید برای تمام سرطان ها محاسبه کنیم و اگر دیدیم اکثریت سرطان ها کمتر از مورد انتظار ثبت می شوند و سرطان خون بیشتر از مورد انتظار ثبت می شود، بروز این سرطان در جامعه زیاد است یا این که سرطان لوکمیا ثبت مناسبی دارد.

حدود اطمینان برای استاندارد سازی مستقیم

$$ASR \pm z_{\alpha/2} \cdot SE (ASR)$$

$$z_{\alpha/2}(CI\ 95\%)=1/96 / z_{\alpha/2}(CI\ 99\%)=2/58$$

$$Var (ASR) = \frac{\sum [a_i w_i^2 (100000 - a_i) / n_i]}{(\sum w_i)^2}$$

$$SE = (ASR) = \sqrt{Var (ASR)}$$

$a_i$ : میزان اختصاصی سنی در گروه سنی

$w_i$ : جمعیت استاندارد در گروه سنی  $i$

$n_i$ : شخص زمان، تعداد افراد در گروه سنی

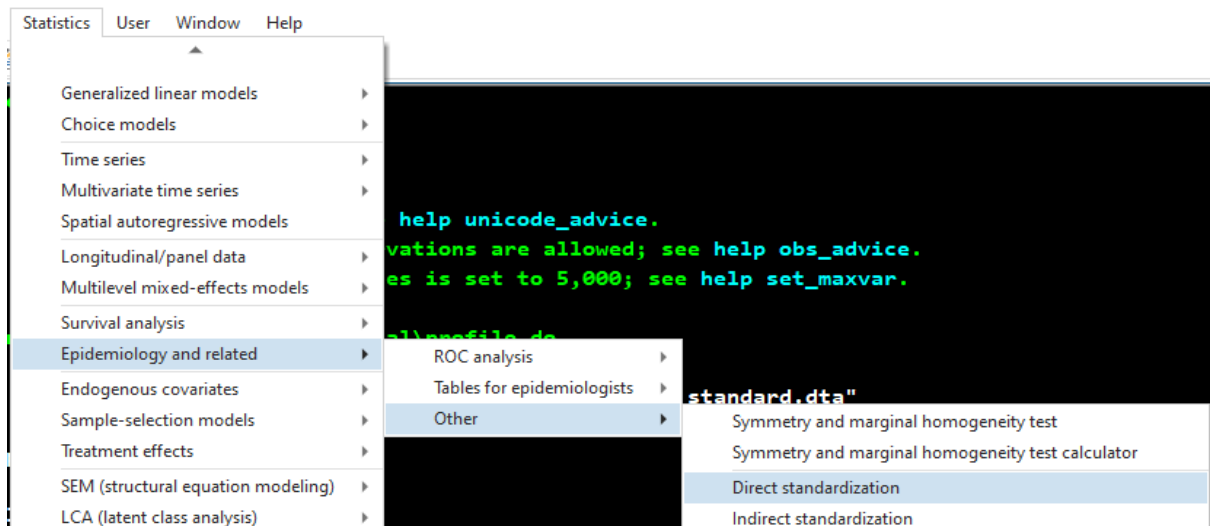
$$= \left( ASR_1 / ASR_2 \right) 1 \pm \left( Z_{\alpha/2} / x \right)$$

نسبت میزان استاندارد کردن

$$x = \frac{(ASR_1 - ASR_2)}{\sqrt{SEASR_1^2 + SEASR_2^2}}$$

استاندارد سازی میزان ها با نرم افزار Stata

مسیر زیر را مشاهده نمایید. بسته به داده های موجود یکی از روش های استانداردسازی مستقیم یا غیرمستقیم را انتخاب کنید.



مثال: برای انجام استاندارد سازی میزان بروز سرطان در کشور A به دو فایل نیاز دارید:

**الف)** فایل اول میزان بروز سنی و جنسی و جمعیت کشور A که ۴ متغیر جنسیت، گروه سنی، تعداد موارد سرطان و تعداد یا درصد جمعیت هر گروه سنی و جنسی.



**ب)** فایل دوم ساختار جمعیت استاندارد با ۳ متغیر جنسیت، گروه سنی و تعداد یا درصد جمعیت هر گروه سنی و جنسی در جمعیت استاندارد که معمولاً جمعیت استاندارد سازمان جهانی بهداشت است. توجه داشته باشید نام متغیرهای دو فایل بایستی یکسان باشد. اکنون از مسیر بالا یا با دستور `db dstdize` کار محاوره ای را باز نمائید و متغیرها را وارد کنید. متغیر تعداد موارد یا درصد را در گزینه `characteristic` و متغیر جمعیت را در گزینه `population` و متغیر گروه سنی را در گزینه `strata variables` و متغیر جنسیت را در گزینه `grouping variable` قرار دهید. سپس جمعیت استاندارد را نیز از مسیر خواسته معرفی کنید. برای این مثال از داده های مطالعه دکتر راتمن در سال ۱۹۸۶ درباره میزان های مرگ در سوئد و پاناما در سال ۱۹۶۲ استفاده می شود. سپس میزان های مرگ دو کشور مقایسه می گردد. فایل `Rotman1986` را باز کنید.

```
. des

Contains data from C:\Users\Tehran\Desktop\stata_book1400\Rotman1986.dta
Observations:      6      1962 Mortality, Sweden & Panama
Variables:         4      3 Oct 2021 21:12
```

Variable name	Storage type	Display format	Value label	Variable label
nation	str6	%9s		Nation
age_category	byte	%9.0g	age_lbl	Age category
population	float	%10.0gc		Population in age category
deaths	float	%9.0gc		Deaths in age category

```
. list, sepby(nation) abbrev(12) divider
```

	nation	age_category	population	deaths
1.	Sweden	0-29	3145000	3,523
2.	Sweden	30-59	3057000	10,928
3.	Sweden	60+	1294000	59,104
4.	Panama	0-29	741,000	3,904
5.	Panama	30-59	275,000	1,421
6.	Panama	60+	59,000	2,456

جدول جمع جمعیت و مرگ ها را بر حسب دو کشور طبق دستور زیر محاسبه می کنیم.

```
. collapse (sum) pop deaths, by(nation)
. list, abbrev(10) divider
```

	nation	population	deaths
1.	Panama	1075000	7,781
2.	Sweden	7496000	73,555

با دستورات زیر میزان خام مرگ محاسبه می شود.

```
generate crude = deaths/pop
list, abbrev(10) divider
```

	nation	population	deaths	crude
1.	Panama	1075000	7,781	.0072381
2.	Sweden	7496000	73,555	.0098126

در مقایسه میزان خام مرگ در دو کشور، مشخص است که این میزان خام و تعداد مرگ ها در سوئد بیشتر از پاناما است. با مراجعه به داده های گروه بندی شده جمعیت بر حسب سن، مشخص است که درصد بالایی از جمعیت سوئد سالمند است. ۱۲۹۴۰۰۰ نفر در گروه سنی ۶۰ سال و بالاتر در مقایسه با ۵۹۰۰۰ نفر در همین گروه سنی در پاناما. بنابراین با توجه به ساختار سنی متفاوت در دو کشور برای مقایسه مرگها، باید استاندارد سازی سن انجام شود. برای این مقایسه میزان های مرگ در این دو جمعیت نیاز به جمعیت استاندارد است. فایل 1982.dta را باز کنید. این فایل درصد توزیع جمعیت استاندارد در گروه های سنی را نشان می دهد.

. list, abbrev(12) divider

	age_category	population
1.	0-29	.35
2.	30-59	.35
3.	60+	.3

توزیع سنی جمعیت بالا را می توان به عنوان وزن برای محاسبه میزان های دو کشور به کار گرفت. با دستور زیر میزان های خام مرگ، تعداد مرگ، تعداد جمعیت بر حسب گروه های سنی در دو کشور را در زیر ببینید.

list, sepby(age\_category) abbrev(12)

	nation	age_category	population	deaths	crude
1.	Sweden	0-29	3145000	3,523	.0011202
2.	Sweden	30-59	3057000	10,928	.0035747
3.	Sweden	60+	1294000	59,104	.0456754
4.	Panama	0-29	741,000	3,904	.0052686
5.	Panama	30-59	275,000	1,421	.0051673
6.	Panama	60+	59,000	2,456	.0416271

## روش استانداردسازی مستقیم

طبق جعبه گفتگوی زیر (db dstdize) متغیرهای فایل mortality را به نرم افزار معرفی کنید. از قسمت Browse نیز جمعیت استاندارد (1962.dta) را انتخاب و معرفی کنید. دستور را اجرا نمایید.

نتایج استانداردسازی مستقیم میزان های مرگ را در دو کشور سوئد و پاناما را در زیر ببینید.

## روش های اپیدمیولوژی و پژوهش با Stata

### Direct standardization

```
-> nation = Panama
```

Stratum	Pop.	Cases	Unadjusted		Std. pop. dist.	s*P
			Pop. dist.	Stratum rate		
0-29	741,000	3,904	0.689	0.0053	0.350	0.0018
30-59	275,000	1,421	0.256	0.0052	0.350	0.0018
60+	59,000	2,456	0.055	0.0416	0.300	0.0125

Total: 1,075,000 7,781  
 Note: s\*P is Stratum rate multiplied by Std. pop. dist.

Adjusted cases = 17,351.2  
 Crude rate = 0.0072  
 Adjusted rate = 0.0161  
 95% conf. interval: [0.0156, 0.0166]

```
-> nation = Sweden
```

Stratum	Pop.	Cases	Unadjusted		Std. pop. dist.	s*P
			Pop. dist.	Stratum rate		
0-29	3,145,000	3,523	0.420	0.0011	0.350	0.0004
30-59	3,057,000	10,928	0.408	0.0036	0.350	0.0013
60+	1,294,000	59,104	0.173	0.0457	0.300	0.0137

Total: 7,496,000 73,555  
 Note: s\*P is Stratum rate multiplied by Std. pop. dist.

Adjusted cases = 115,032.5  
 Crude rate = 0.0098  
 Adjusted rate = 0.0153  
 95% conf. interval: [0.0152, 0.0155]

### Summary of study populations

nation	N	Crude rate	Adjusted rate	[95% conf. interval]	
Panama	1,075,000	0.007238	0.016141	0.015645	0.016637
Sweden	7,496,000	0.009813	0.015346	0.015235	0.015457

نتایج نشان می دهد میزان های مرگ استاندارد شده در سوئد کمتر از پاناما است. این نتیجه برخلاف میزان های خام مرگ در دو کشور است. میزان خام مرگ در سوئد بیشتر از پاناما بود. بنابراین برای مقایسه میزان های مرگ در دو کشور با ساختار سنی متفاوت باید از استانداردسازی استفاده نمود.

## روش استانداردسازی غیر مستقیم

وقتی میزان های اختصاصی در دو جمعیت نامعلوم یا قابل اعتماد نباشند از استانداردسازی غیرمستقیم استفاده می شود. برای مقایسه مرگ و میر در کالیفرنیا و ماین (Maine) از داده های ۱۹۸۹ مطالعه kahn استفاده می کنیم. تعداد جمعیت در گروه های سنی ویژه در دو ایالت در دسترس است. میزان های مرگ اختصاصی دو ایالت در دسترس نیست. بنابراین استانداردسازی مستقیم امکان پذیر نیست و باید استانداردسازی غیرمستقیم را انجام داد. فایل جمعیت استاندارد آمریکا با نام popkahn را باز نمائید. در این فایل میزان مرگ (rate)، تعداد مرگ و جمعیت گروه های سنی (pop) موجود است.

```
. list age pop deaths rate, sep(4)
```

	age	population	deaths	rate
1.	<15	57,900,000	103,062	.00178
2.	15-24	35,441,000	45,261	.00128
3.	25-34	24,907,000	39,193	.00157
4.	35-44	23,088,000	72,617	.00315
5.	45-54	23,220,000	169,517	.0073
6.	55-64	18,590,000	308,373	.01659
7.	65-74	12,436,000	445,531	.03583
8.	75+	7,630,000	736,758	.09656

به داده های دو ایالت مراجعه کنید. فایل kahn را باز کنید.

```
. use https://www.stata-press.com/data/r17/kahn
```

```
. list, sep(4)
```

	state	age	populat-n	death	st	death_~e
1.	California	<15	5,524,000	166,285	1	.0016
2.	California	15-24	3,558,000	166,285	1	.0013
3.	California	25-34	2,677,000	166,285	1	.0015
4.	California	35-44	2,359,000	166,285	1	.0028
5.	California	45-54	2,330,000	166,285	1	.0067
6.	California	55-64	1,704,000	166,285	1	.0154
7.	California	65-74	1,105,000	166,285	1	.0328
8.	California	75+	696,000	166,285	1	.0917
9.	Maine	<15	286,000	11,051	2	.0019
10.	Maine	15-24	168,000	.	2	.0011
11.	Maine	25-34	110,000	.	2	.0014
12.	Maine	35-44	109,000	.	2	.0029
13.	Maine	45-54	110,000	.	2	.0069
14.	Maine	55-64	94,000	.	2	.0173
15.	Maine	65-74	69,000	.	2	.039
16.	Maine	75+	46,000	.	2	.1041

دستور زیر را اجرا نمایید.

. istdize death pop age using [https://www.stata-press.com/data/r17/popkahn,by\(state\) pop\(deaths pop\) print](https://www.stata-press.com/data/r17/popkahn,by(state) pop(deaths pop) print)  
 استفاده `print` در دستور بالا، جدول خلاصه برای جمعیت استاندارد را به ما می دهد. متغیرها را به نرم افزار معرفی نمایید.

```
Standard population

      Stratum      Rate
-----
      <15          0.00178
      15-24         0.00128
      25-34         0.00157
      35-44         0.00315
      45-54         0.00730
      55-64         0.01659
      65-74         0.03583
      75+           0.09656

      Crude rate = 0.00945

Indirect standardization

-> state = California

      Stratum      Standard
                    population
                    rate      Observed
                    population      Expected
                    cases
-----
      <15          0.0018      5,524,000      9,832.72
      15-24         0.0013      3,558,000      4,543.85
      25-34         0.0016      2,677,000      4,212.46
      35-44         0.0031      2,359,000      7,419.59
      45-54         0.0073      2,330,000     17,010.10
      55-64         0.0166      1,704,000     28,266.14
      65-74         0.0358      1,105,000     39,587.63
      75+           0.0966           696,000     67,206.23

Total:                                19,953,000      178,078.73

      Observed cases = 166,285
      SMR (Obs/Exp) = 0.93
      SMR exact 95% conf. interval: [0.9293, 0.9383]

      Crude rate = 0.0083
      Adjusted rate = 0.0088
      95% conf. interval: [0.0088, 0.0089]

-> state = Maine

      Stratum      Standard
                    population
                    rate      Observed
                    population      Expected
                    cases
-----
      <15          0.0018      286,000      509.08
      15-24         0.0013      168,000      214.55
      25-34         0.0016      110,000      173.09
      35-44         0.0031      109,000      342.83
      45-54         0.0073      110,000      803.05
      55-64         0.0166      94,000      1,559.28
      65-74         0.0358      69,000      2,471.99
      75+           0.0966           46,000      4,441.79

Total:                                992,000      10,515.67

      Observed cases = 11,051
      SMR (Obs/Exp) = 1.05
      SMR exact 95% conf. interval: [1.0314, 1.0707]

      Crude rate = 0.0111
      Adjusted rate = 0.0099
      95% conf. interval: [0.0097, 0.0101]

Summary of study populations, reporting rates

      state      Observed
                    cases      Crude rate      Adjusted
                    rate      [95% conf. interval]
-----
California      166,285      0.008334      0.008824      0.008782      0.008866
Maine           11,051      0.011140      0.009931      0.009747      0.010118

Summary of study populations, reporting SMRs

      state      Observed
                    cases      Expected
                    cases      SMR      Exact
                    [95% conf. interval]
-----
California      166,285      178,078.73      0.934      0.929290      0.938271
Maine           11,051      10,515.67      1.051      1.031405      1.070688
```

## رویکرد رگرسیون کاکس در تحلیل مطالعات کوهورت

در تحلیل داده های مطالعه مقطعی، قبلا از رگرسیون لجستیک استفاده گردید و نسبت شانس مرگ برای متغیرهای سیگار کشیدن، فشارخون، کلسترول و سن گزارش شد. در اینجا نسبت مخاطره (Hazard Ratio) برای متغیرهای مذکور قابل محاسبه است. خروجی زیر را ملاحظه نمایید.

```
. stcox smoking sbp chol agein

      Failure _d: chd==1
      Analysis time _t: (timeout-origin)/365.25
                Origin: time timebth
      Enter on or after: time timein
                ID variable: id

Iteration 0:  log likelihood = -1013.083
Iteration 1:  log likelihood = -984.78125
Iteration 2:  log likelihood = -984.49673
Iteration 3:  log likelihood = -984.49666
Refining estimates:
Iteration 0:  log likelihood = -984.49666

Cox regression with no ties

No. of subjects =      1,677                Number of obs =  1,677
No. of failures =       154
Time at risk   = 27,605.3707

LR chi2(4)      = 57.17
Prob > chi2     = 0.0000

Log likelihood = -984.49666
```

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
smoking	1.918062	.314852	3.97	0.000	1.390391	2.645991
sbp	1.016183	.0035337	4.62	0.000	1.009281	1.023133
chol	1.007659	.0016548	4.65	0.000	1.004421	1.010908
agein	1.025482	.0194589	1.33	0.185	.9880434	1.064338

در مقایسه نتایج بالا با نتایج مدل رگرسیون لجستیک در استفاده از داده های مقطعی قابل ذکر است که در مطالعه مقطعی همه متغیرهای وارد شده به مدل لجستیک معنی دار بودند. اما در اینجا همانگونه که مشخص است سن معنی دار نشده است و بقیه اندازه های اثر نیز تعدیل شدند. نسبت شانس مرگ برای سیگار از ۲/۳۵ در مطالعه مقطعی به ۱/۹۱ در مطالعه کوهورت تغییر پیدا نمود. پیش فرض تناسب نسبت مخاطره نیز چک و برقرار بود.

```
. estat phtest
```

Test of proportional-hazards assumption

Time function: Analysis time

	chi2	df	Prob>chi2
Global test	4.39	4	0.3558

با دستور `stcoxkm, by(smoking)` می توان منحنی کاپلان مایر و بقا در افراد سیگاری و غیرسیگاری را ترسیم نمود. همچنین می توان با دستور `stsum, by(smoking)` شاخص های توصیفی بقا از جمله میانه بقا را محاسبه نمود.

## Case Control Study

## مطالعه مورد شاهد:

مطالعه مورد-شاهدی (Case-Control Studies) یک مطالعه تحلیلی است که در آن یک فرضیه مورد آزمون قرار می‌گیرد. این فرضیه بیانگر این است که آیا یک مواجهه (که به عنوان علت احتمالی یک بیماری خاص) با بیماری یا پیامد مورد نظر همراهی دارد. یک مطالعه مورد-شاهدی با تعیین یک گروه از موردها (Cases) (افراد) که پیامد، بیماری یا وضعیت مورد نظر را دارا می‌باشند) و یک گروه از شاهدها (Controls) (افراد) که پیامد یا بیماری مورد نظر را ندارند) آغاز می‌شود. سپس نسبت مواجهه با فاکتور مورد نظر در هر دو گروه تعیین و مقایسه می‌شود. اگر نسبت مواجهه به فاکتورها در موردها بیشتر از شاهدها بود، مواجهه ممکن است یک ریسک فاکتور برای بیمار باشد. اگر این نسبت کمتر بود مواجهه ممکن است یک فاکتور محافظت کننده برای بیماری باشد. افراد در مطالعه مورد-شاهدی بر پایه پیامد وارد می‌شوند، به عنوان مثال بر پایه داشتن یا نداشتن بیماری مورد نظر تقسیم می‌شوند. هدف از انتخاب گروه کنترل دادن اطلاعات درباره وضعیت مواجهه در کوهورتی است که موردها از آن حاصل می‌شوند. مطالعات مورد-شاهدی از دو جهت کارآمد هستند: گذشته نگر بودن و عدم نیاز به پی‌گیری طولانی برای وقوع موردها (بیماری) و کاربرد داشتن در بررسی بیماریهای با دوره کمون طولانی و استفاده از گروه کنترل به عنوان نمونه‌ای از تمامی کسانی که پیامد مورد نظر را ندارند و در نتیجه نیاز به نمونه‌های بسیار کمتری جهت مشخص کردن وضعیت مواجهه است. اینکه چه افرادی وارد گروه مورد شوند بایستی به دقت مورد توجه قرار گیرد. مطالعه ممکن است بر پایه بیمارستان (Hospital based) یا بر پایه جمعیت (Population Based) باشد. در مطالعه بر پایه بیمارستان موردها از تمامی بیمارانی که همه معیارهای تعریف مورد (Case) را داشته و در یک بیمارستان خاص حضور دارند گرفته می‌شوند. در مطالعه بر پایه جمعیت موردها از یک جمعیت تعریف شده در یک دوره زمانی ثابت گرفته می‌شوند. در کل انتخاب شاهدها مشکل‌ترین قسمت طراحی یک مطالعه مورد-شاهدی است. جهت انتخاب نمونه‌ای از شاهدها که معرف جمعیتی باشد که موردها از آن حاصل شده‌اند، می‌توان از یک جمعیت تعریف شده کوهورت استفاده کرد. زمانی که شاهدها از یک جمعیت تعریف شده کوهورت انتخاب می‌شوند چهارچوب نمونه‌گیری (Sampling Frame) شامل افراد در ریسک در زمانی که موردها در طول دوره پیگیری ظاهر می‌شوند (Density Sampling) و یا جمعیت کوهورت پایه (Baseline Cohort) می‌باشد، که اولی مربوط به طراحی مورد-شاهدی لانه‌ای (Nested Case Control) و دومی مربوط به طراحی مورد-کوهورت (Case Cohort) است.

مطالعه مورد-شاهدی لانه‌ای (Nested Case Control Study) راهی است برای امتحان فرضیه‌هایی که هنگام انجام یک مطالعه کوهورت در مورد بیماری مورد نظر (نه مواجهه مورد نظر) ایجاد می‌شوند. در این مطالعه پس از پی‌گیری نمونه‌ها افرادی که بیمار می‌شوند به عنوان موارد بیماری انتخاب می‌شوند و یک نمونه از افرادی که در پی‌گیری بیمار نشده‌اند به عنوان کنترل در نظر گرفته می‌شوند. در واقع در طول دوره پی‌گیری زمانی که موارد ایجاد می‌شوند شاهدها نیز تقریباً در همان دوره زمانی از افرادی که سالم هستند انتخاب می‌گردند (density sampling). در این روش، اگر فرد مبتلا پس از بهبودی دوباره در معرض خطر بیماری باشد و پیگیری شود، ممکن است در ادامه مطالعه، هم به عنوان مورد و هم به عنوان شاهد دوباره مورد ارزیابی قرار گیرد. در این مطالعه گروه مورد و شاهد به طور خودکار همسان شده‌اند. این راهی است برای ضمانت این موضوع که شاهدها معرف جمعیتی

هستند که موارد را تولید کرده‌اند که در واقع همان جمعیت کوهورت می‌باشند. نسبت شانس (OR) محاسبه شده در این نوع مطالعه مورد-شاهدی، برآورد مناسبی از RH (Relative Hazard) است. در طراحی case cohort design، گروه مورد از مواردی که در طول دوره پی‌گیری شناسایی می‌شوند و گروه شاهد نمونه‌ای از کل جمعیت کوهورت در زمان شروع پیگیری (baseline) می‌باشند. در این حالت این امکان وجود دارد که یک نفر هم به عنوان شاهد و هم به عنوان مورد در مطالعه وارد شود. نتیجه این خواهد بود که OR بدست آمده برآورد مناسبی از RR می‌باشد. در این طراحی همچنین این امکان وجود دارد که موارد از شاهد های کوهورت پایه (baseline) جدا شوند. بنابراین، چهارچوب نمونه‌گیری شاهد ها به وسیله افرادی که در طول دوره پی‌گیری سالم باقی مانده‌اند شکل می‌گیرد. در این حالت، OR بدست آمده درست همان مقداری خواهد بود که از روش مرسوم انجام مطالعه مورد-شاهدی بدست می‌آید.

### تحلیل مطالعات مورد-شاهد:

با استفاده از داده های فرضی که در جداول زیر نشان داده شده است، می توان در مقایسه ۲۱۰ نفر (گروه مورد) که به سکت قلبی مبتلا شدند با ۱۹۷۹۰ فرد سالم (گروه شاهد)، یک تحلیل مورد-شاهدی انجام داد.

مبتلا به سکت قلبی	غیر مبتلا به سکت قلبی
فشار خون بالا (a) ۱۸۰	۹۸۲۰ (b)
فشار خون طبیعی (c) ۳۰	۹۹۷۰ (d)
(a + c) ۲۱۰	(b + d) ۱۹۷۹۰

نسبت مواجهه در گروه مورد و نسبت مواجهه در گروه شاهد به صورت زیر محاسبه می شود.

$$P_{exp} = a / (a+c) \text{ (در گروه مورد)}$$

$$P_{exp} = b / (b+d) \text{ (در گروه شاهد)}$$

بنابراین، شانس (Odds) مواجهه در گروه مورد و شاهد برابر خواهد بود با

$$Odds_{exp} = P_{exp} / (1 - P_{exp}) = [ a / (a+c) ] / [ 1 - ( a / (a+c) ) ] = a / c \text{ (در گروه مورد)}$$

$$Odds_{exp} = P_{exp} / (1 - P_{exp}) = [ b / (b+d) ] / [ 1 - ( b / (b+d) ) ] = b / d \text{ (در گروه شاهد)}$$

با تقسیم شانس مواجهه در دو گروه مورد و شاهد، نسبت شانس مواجهه ( $OR_{exp}$ ) برابر خواهد بود:

$$OR_{exp} = Odds_{exp} / Odds_{sep} = (a/c) / (b/d) = ad / bc$$

مراحل محاسباتی بالا را دوباره تکرار می کنیم، اما در اینجا نسبت بیماری ( $P_d$ ) را در گروه مواجهه یافته و عدم مواجهه محاسبه می کنیم :

$$P_d = a / (a+b) \text{ (در گروه مواجهه یافته)}$$

$$P_d = c / (c+d) \text{ (در گروه عدم مواجهه)}$$



بنابراین، شانس (Odds) بیماری در گروه مواجهه و عدم مواجهه برابر خواهد بود با

$$\text{Odds}_d = P_d / (1 - P_d) = [ a / (a+b) ] / [ 1 - (a / (a+b)) ] = a / b$$

(در گروه مواجهه یافته)

$$\text{Odds}_d = P_d / (1 - P_d) = [ c / (c+d) ] / [ 1 - (c / (c+d)) ] = c / d$$

(در گروه عدم مواجهه)

با تقسیم شانس بیماری در دو گروه مواجهه و عدم مواجهه، نسبت شانس بیماری ( $OR_d$ ) برابر خواهد بود:

$$OR_d = \text{Odds}_d / \text{Odds}_d = (a/b) / (c/d) = ad / bc$$

همانطور که مشاهده می کنید، نسبت شانس مواجهه ( $OR_{exp}$ ) با نسبت شانس بیماری ( $OR_d$ ) برابر است. و این خاصیت قرینگی  $OR$  یکی از مهمترین خصوصیت‌های این شاخص اندازه گیری رابطه مواجهه و پیامد است. در مثال ذکر شده، نسبت شانس برابر است با :

$$OR_{exp} = (a \div c) \div (b \div d) = (180 \div 30) \div (982 \div 997) = 6/0.9$$

$$OR_d = (a \div b) \div (c \div d) = (180 \div 982) \div (30 \div 997) = 6/0.9$$

در مثال فوق که محاسبات برای داده های آینده نگر انجام شده بود، تمام موردها و شاهد ها جهت برآورد نسبت شانس استفاده شدند. اما در مطالعات مورد-شاهدی، معمولاً کسری از شاهدها ( و کسری از موردها) مورد نمونه گیری قرار می گیرند. اگر تعداد کل موردها کم باشد (همانطور که در جدول بالا نشان داده شده است)، پژوهشگر می تواند از تمام موردها و یک نمونه (کسری) از شاهدها استفاده کند.

به عنوان مثال، اگر ۱۰۰٪ موردها و تنها ۱۰٪ شاهدها مورد مطالعه قرار گیرند، توزیع داده ها به صورت زیر خواهد بود :

مبتلا به سکته قلبی	غیر مبتلا به سکته قلبی		
(a) ۱۸۰	(b) ۹۸۲	فشار خون بالا	
(c) ۳۰	(d) ۹۹۷	فشار خون طبیعی	
(a + c) ۲۱۰	(b + d) ۱۹۷۹		

$$OR = (a \div c) \div (b \div d) = (180 \div 30) \div (982 \div 997) = 6/0.9$$

این مثال تاییدکننده این موضوع است که نسبت نمونه گیری (از جمعیت کل) در موردها و شاهدها لازم نیست حتماً برابر باشد. با این وجود، جهت بدست آوردن نمونه عاری از سوگرایی از جمعیت موردها و شاهدها، نسبت نمونه گیری باید مستقل از مواجهه باشد.

این واقعیت که نسبت شانس مواجهه با نسبت شانس بیماری یکسان است، این اجازه را میدهد که یک تفسیر آینده نگر از نسبت شانس در مطالعه مورد-شاهدی داشته باشیم.

در مطالعه مورد-شاهدی تک به تک همسان شده ((Individual Matched Case-Control)) که در آن نسبت گروه مورد به شاهد ۱ به ۱ است، دیگر با تک افراد سروکار نداریم، بلکه این جفتها هستند که تعیین کننده وضعیت مواجهه و پیامد می باشند.

بنابراین، برای برآورد عاری از سوگرایی نسبت شانس در این نوع مطالعه، تعداد جفت هایی که در آنها موردها مواجهه داشته و شاهدها مواجهه نداشته اند [case (+), control (-)] بر تعداد جفت هایی که در آنها شاهدها مواجهه داشته و موردها مواجهه نداشته اند [case(-), control(+)] تقسیم می کنیم:

شاهد های مواجهه نیافته	شاهد های مواجهه یافته	
b	A	مورد های مواجهه یافته
d	C	مورد های مواجهه نیافته

$$OR_{match} = b / c$$

باید توجه داشت که آنالیز داده های همسان شده، دارای پیچیدگیهای خاصی است و با روشهای آماری خاصی انجام می شود. برای کنترل مخدوش کننده ها از روشهای دیگری همچون محدود سازی، طبقه بندی و لوجستیک رگرسیون بر روی داده های همسان و غیر همسان نیز استفاده می شود. تحلیل این نوع مطالعات در قسمت رگرسیون لوجستیک توضیح داده شد.

نسبت شانس که اغلب برای کنترل مخدوش کننده ها به کار می رود، نسبت شانس تطبیق داده شده نامیده می شود. در داده های ناهمسان، استفاده از روشهای طبقه بندی و لوجستیک رگرسیون امکان پذیر است. در داده های همسان شده انجام آنالیز طبقه بندی شده امکان پذیر نیست و روش انتخابی لوجستیک رگرسیون شرطی (Conditional Logistic Regression) می باشد.

آنالیز مطالعه با (Group Matching Case-control) مانند استفاده از روشهایی که برای داده های همسان نشده بیان شد، قابل انجام است. اما باید به این نکته توجه داشت که همیشه آنالیز را به تفکیک براساس متغیرهایی که با آنها همسان سازی گروهی انجام شده است، انجام داد. برای مثال، اگر مطالعه از لحاظ سن به صورت گروهی تنها در دو گروه سنی همسان شده است. آنالیز بر اساس سن در این دو گروه طبقه بند شده و برای هر گروه آنالیز بطور جداگانه انجام می گیرد.

### تخمین خطر نسبی (Relative Risk) از نسبت شانس در مطالعه مورد-شاهدی :

در مطالعه مورد-شاهدی استفاده از نسبت شانس برای تخمین خطر نسبی بر اساس این فرض صورت می گیرد که میزان بروز بیماری مورد مطالعه کم باشد، یعنی بیماری نادر باشد. با توجه به فرمول زیر، مشاهده می کنید که رابطه بین نسبت شانس و خطر نسبی به بروز پیامد وابسته است. اگر بروز پیامد، کمتر از ۱٪ باشد، خطر نسبی به نسبت شانس خیلی نزدیک می شود. حتی برای پیامدهای نسبتاً شایع (با فراوانی ۱-۵٪) نسبت شانس و خطر نسبی تقریباً برابر خواهند بود.

$$Relative Risk = OR / 1 - [q_{(-)} - (OR \times q_{(-)})]$$

$$q_{(-)} = \text{بروز در گروه مواجهه نیافته}$$

نتایج مطالعات مختلف نشان داده است هنگامی که پیامد مورد نظر شایعتر است (با فراوانی بیشتر از ۱۰ تا ۲۰٪)، مقدار نسبت شانس که از مطالعه مورد-شاهدی بدست می آید با خطر نسبی به طور واضحی تفاوت دارد. بنابراین، در تفسیر نسبت شانس باید دقت بسیاری داشت. زمانی که گروه شاهد نمونه ای از جمعیت کل است (مانند آنچه

درباره مطالعه Case-Cohort در بالا گفته شد)، نسبت شانس برآورد مستقیمی از خطر نسبی بدون در نظر گرفتن فراوانی پیامد مورد نظر است. در مطالعه Case-Cohort، نسبت شانس به این صورت محاسبه می‌شود:

جمعیت کل	شاهدها	موردها		
a + b	b	A	بله	
c + d	d	C	خیر	مواجهه

شانس مواجهه در موردها

= نسبت شانس

شانس مواجهه در کل جمعیت

$$OR = [a / c] / [(a + b) / (c + d)] = [a / (a + b)] / [(c / (c + d))] = q_{(+)} / q_{(-)} = \text{Relative Risk}$$

$q_{(+)}$  = احتمال بیماری در گروه مواجهه

$q_{(-)}$  = احتمال بیماری در گروه عدم مواجهه

### محاسبه خطر قابل انتساب (Attributable Risk) در مطالعات مورد-شاهدی

خطر قابل انتساب به مواجهه ( $AR_{exp}$ ) را در مطالعات مورد-شاهدی زمانی می‌توان محاسبه نمود که نسبت شانس، برآورد مناسبی از خطر نسبی باشد (یعنی بیماری نادر باشد).

$$AR_{exp} = [(OR - 1) / OR] \times 100$$

خطر نسبی (Relative Risk) و نسبت شانس برآوردی از مقدار همراهی مواجهه و پیامد می‌باشند. یک مقدار خطر نسبی برابر با ۳ به این معنای است که مواجهه با فاکتور مورد با افزایش ۳ برابری ریسک بیماری در گروه مواجهه همراه است. خطر نسبی ۰/۵ به معنای این است که مواجهه همراه با نصف شدن ریسک بیماری است. خطر نسبی برابر با ۱، هیچ همراهی را بین مواجهه و پیامد نشان نمی‌دهد. برای نسبت شانس هم، تفسیر فوق برقرار است اما آنچه درباره آن صحبت می‌شود، شانس (Odds) بیماری است.

**در نتایج اگر همراهی یافت شود ( $OR \neq 1$  یا  $RR$ )، پژوهشگر باید نکات زیر را در نظر داشته باشد:**

- آیا این همراهی ناشی از شانس و تصادف است یا خیر؟
- آیا سوگرایی انتخاب در انتخاب موردها و شاهدها وجود دارد؟
- آیا سوگرایی اطلاعات در جمع‌آوری داده‌های مربوط به مواجهه و یا پیامد وجود دارد؟
- آیا فاکتورهای مخدوش کننده مورد توجه قرار گرفته اند و آیا اثرات آنها بطور کامل کنترل شده است یعنی آیا اثر مخدوش کنندگی باقیمانده است؟
- آیا توالی زمانی بین مواجهه و پیامد وجود داشته و حفظ شده است؟ احتمال اینکه مواجهه، در واقع پیامد باشد (Reverse Causality) چقدر است؟

مثال: اگر همراهی بین رژیم غذایی و سرطان معده در یک مطالعه مورد-شاهدی یافت شود، احتمال دارد که رژیم غذایی بدنبال بروز علائم زودرس سرطان تغییر کرده باشد. در یک مطالعه مورد-شاهدی که جهت بررسی رابطه مصرف قهوه و سرطان پانکراس انجام شده بود، نتایج زیر حاصل شد:

مجموع	شاهد	مورد		
۸۹۰	(b)۴۴۰	(a)۴۵۰	بله	مصرف قهوه
۷۱۰	(d)۴۱۰	(c)۳۰۰	خیر	
	۸۵۰	۷۵۰	مجموع	

OR خام محاسبه شده برای رابطه بین مصرف قهوه و سرطان پانکراس برابر است با: ۱/۴۰

این نتیجه نشان می دهد که خطر سرطان پانکراس با مصرف قهوه مرتبط است. توجیه های مختلفی برای این نتیجه مشاهده شده وجود دارد:

۱. شانس: ارتباط مشاهده شده بین مصرف قهوه و سرطان پانکراس ناشی از شانس باشد. مقدار آزمون  $X^2$  برای این جدول به صورت زیر محاسبه می شود:

$$X^2 = \frac{[(ad-bc)-N/2]^2 \times N}{[(a+b)(c+d)(a+c)(b+d)]} = ۱۰/۶۲$$

مقدار آزمون و  $P\text{-value} = ۰/۰۰۱$  نشان می دهد که شانس، یک توجیه غیرمحتمل برای این یافته است. حدود اطمینان نسبت شانس را نیز می توان محاسبه و ارزیابی را انجام داد.

۲. تورش: نسبت شانس مشاهده شده ۱/۴۰ بدلیل تورش می باشد و بیانگر مقدار واقعی نیست.

۳. مخدوش کنندگی: مقدار ۱/۴۰ مشاهده شده برای نسبت شانس، مقداری واقعی است، اما بدلیل اثر سایر متغیرها می باشد. برای مثال، ممکن است افرادی که قهوه مصرف می کنند، بیشتر هم سیگار بکشند و این سیگار است که منجر به افزایش خطر سرطان پانکراس شده و نه مصرف قهوه. این حالت، مخدوش کنندگی نام دارد. با استفاده از روش مانتل هنزل که در ادامه بیان می شود، اثر مخدوش کنندگی اصلاح می شود.

۴. رابطه علیتی: مصرف قهوه شانس سرطان پانکراس را افزایش می دهد و مقدار OR ۱/۴۰ برآورد شده، نتیجه درستی است.

**Mantel-Haenszel Odds Ratio (OR<sub>MH</sub>) (نسبت شانس مانتل - هنزل):**

جهت محاسبه OR<sub>MH</sub> اولین قدم ساخت جدول متقاطع پیامد و مواجهه به تفکیک طبقه های مختلف عامل مخدوش کننده است:

		طبقه ۱		طبقه ۲	
		بیماری		بیماری	
مواجهه	بله	a <sub>1</sub>	b <sub>1</sub>	a <sub>2</sub>	b <sub>2</sub>
	خیر	c <sub>1</sub>	d <sub>1</sub>	c <sub>2</sub>	d <sub>2</sub>
		n <sub>1</sub>		n <sub>2</sub>	

سپس OR<sub>MH</sub> از طریق فرمول زیر محاسبه می گردد (I تعداد طبقات متغیر مخدوش کننده است):

$$OR_{MH} = \frac{\sum_{i=1}^I (a_i d_i / n_i)}{\sum_{i=1}^I (b_i c_i / n_i)}$$

و برای وقتی که تنها دو طبقه برای مخدوش کننده وجود دارد :

$$OR_{MH} = \frac{a_1 d_1 / n_1 + a_2 d_2 / n_2}{b_1 c_1 / n_1 + b_2 c_2 / n_2}$$

چنانچه داده های مثال رابطه سرطان و مصرف قهوه بر حسب متغیر سیگار گروه بندی شود نسبت شانس برابر است با :

$$OR_{MH} = \frac{470 / (50 \times 220) + 1130 / (400 \times 190)}{470 / (100 \times 100) + 1130 / (340 \times 200)} = 1/11$$

بنابراین، OR اصلاح شده برای مصرف سیگار ۱/۱ می باشد. این مقدار را با OR=۱/۴۰ (که معمولا Crude OR خوانده می شود) بدست آمده از جدول کلی اولیه مقایسه کنید. پس از اصلاح OR برای مصرف سیگار، این عدد در مقایسه با OR خام به مقدار ۱ نزدیکتر شده است که در واقع بیانگر آنست که ارتباط بین مصرف قهوه و سرطان پانکراس چندان هم که تصور می شد، قوی نیست.

روش MH می تواند برای اصلاح اثر مخدوش کنندگی برای شاخصهای Relative Risk و Relative Ratio از طریق محاسبه میانگین وزن داده شده RR طبقه های مختلف بکار رود. به این روشها، روشهای آنالیز طبقه ای (Stratified Analysis) هم گفته می شود ، زیرا رابطه بین مواجهه و پیامد در طبقه های (لایه ها - سطوح) مختلف متغیر مخدوش کننده بررسی می شود. در قسمت مدل سازی به این تکنیک اشاره شده بود.

**Mantel-Haenszel Chi-square test ( $X^2_{MH}$ )**

آزمون  $X^2_{MH}$  جهت تعیین اینکه آیا  $OR_{MH}$  (یا  $RR_{MH}$ ) از نظر آماری معنی دار است یا نه (یعنی با مقدار ۱ تفاوت آماری معنی دار دارد) بکار می رود. این آزمون، مشابه آزمون  $X^2$  معمولی است. فرضیه صفر در واقع عدم ارتباط بین مواجهه و پیامد در هیچ یک از طبقه ها است. جهت انجام این آزمون، باید موارد زیر را برای هر یک از جداول داشته باشیم:

۱. مقدار مشاهده شده  $a$

۲. مقدار قابل انتظار  $a$ ، با در نظر گرفتن برقرار بودن فرضیه صفر یعنی عدم وجود رابطه :

$$E_a = (a+b)(a+c)/n$$

۳. واریانس  $a$ ، با در نظر گرفتن برقرار بودن فرضیه صفر یعنی عدم وجود رابطه :

$$V_a = (a+b)(c+d)(a+c)(b+d) / [n^2(n-1)]$$

در این مثال اعداد بصورت زیر جایگذاری و محاسبه انجام می شود:

$$V_a = (a+b)(c+d)(a+c)(b+d) / [n^2(n-1)]$$

	a	$E_a=(a+b)(a+c)/n$	
سیگاری	۴۰۰	۳۹۲/۹	۶۳/۷
غیر سیگاری	۵۰	۴۷/۹	۲۲/۲
	۴۵۰	۴۴۰/۸	۸۵/۹

(تعداد سطوح متغیر مخدوش کننده)  $I = 2$

اگر فرضیه صفر درست باشد، انتظار داریم که اختلاف بین مشاهدات و مقادیر قابل انتظار اندک باشد. ما آزمون می کنیم که آیا این اختلاف بزرگتر از مقداری است که بتواند تنها بدلیل شانس بوجود آمده باشد. در اینجا از اختلاف بین  $\sum a$  و  $\sum E_a$  استفاده می کنیم:

$$\chi^2_{MH} = \frac{(\sum_{i=1}^I a_i - \sum_{i=1}^I E a_i)^2}{\sum_{i=1}^I V_i}$$

و با استفاده از جدول توزیع  $X^2$  با درجه آزادی ۱، مقدار P-value را بدست می آوریم. در این مثال،  $X^2$  محاسبه شده برابر ۰/۸۸ و درجه آزادی برابر با ۱ می باشد. نتیجه مقدار  $P\text{-value} > ۰/۳۴$  خواهد بود که در واقع اینگونه تفسیر می گردد که پس از اصلاح اثر مصرف سیگار، شواهدی جهت وجود ارتباط بین مصرف قهوه و سرطان پانکراس وجود ندارد. بنابراین  $OR = ۱/۱$  اختلاف معنی داری با عدد ۱ ندارد.

## مقایسه نسبتها برای داده های زوجی :

تاکنون آزمونهای مختلفی آماری برای داده های دسته بندی شده ارائه شده است. این روشها تنها برای مشاهدات کیفی (دسته بندی شده) که داده ها حالت زوجی ندارند کاربرد دارند. به عبارتی زمانیکه مشاهدات یک گروه با مشاهدات گروه دیگر جفت نشده یا وابسته نباشند، می توان از آنها استفاده کرد. برای مثال، موارد سرطان پانکراس کاملا مستقل از گروه کنترل نمونه گیری شده اند، و هیچ ارتباط یا جورسازی صورت نگرفته است. مثالهایی از گروههای وابسته عبارتند از : مطالعه مورد-شاهدی جور شده (Matched)، اندازه گیریهای "قبل و بعد"، و مقایسه دو مشاهده گر در ارتباط با یک موضوع. این داده ها به روشهای خاص آنالیز که این جفت شدن ها را مد نظر دارد، نیاز دارند. برای ارزیابی اینکه آیا تفاوت مشاهده شده بین تعداد این گروهها، بیش از مقداری است که ما بر اساس شانس و تصادف انتظار داشتیم، از آزمون مک نمار ( $\chi^2$  McNemar's) استفاده می کنیم. اگر ما تعداد زوجهای (مورد مواجهه یافته و شاهد مواجهه نیافته) را با b نشان دهیم و تعداد زوجهای (مورد مواجهه نیافته و شاهد مواجهه یافته) را با c نشان دهیم،  $\chi^2$  McNemar's بصورت زیر محاسبه می گردد:

$$\chi^2_{McN} = \frac{(b-c)^2}{(b+c)}$$

توزیع آن همان توزیع  $\chi^2$  با درجه آزادی ۱ است.

**خلاصه و نتیجه گیری:**

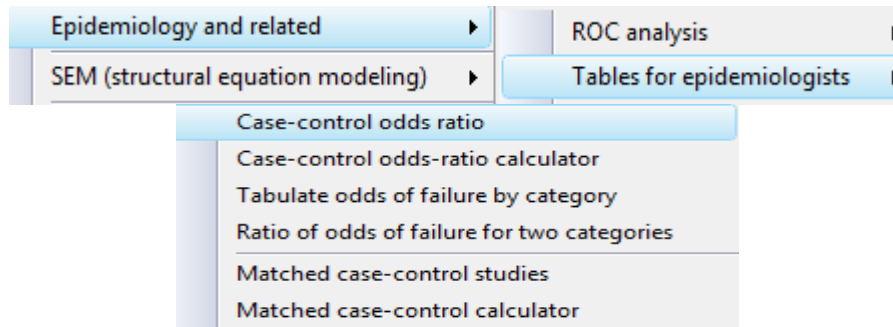
در این قسمت انواع مطالعات اپیدمیولوژیک مقطعی، کوهورت و مورد شاهد مورد بحث قرار گرفت. شاخص های اثر و رابطه در این مطالعات بیان و نحوه محاسبه آنها ذکر گردید. به طور ویژه مفاهیم پایه بحث گردید. خطر (Risk) و شیوع (Prevalence) هر کدام یک Porportion هستند. واحد زمانی ندارند. بر خلاف خطر، میزان (Rate) به زمان بستگی ندارد و در طولانی مدت مستقل است. میزان در واقع متوسط احتمال بیماری در طول زمان است. با افزایش فاصله زمانی، خطر گرایش به بالا رفتن دارد. اگر اثر مواجهه در طول زمان تغییر کند خطر رویکرد مناسبی برای اندازه گیری بیماری نیست. خطر مقادیر صفر تا ۱ می گیرد. از جنس احتمال است. میزان مقادیر صفر تا بی نهایت را می گیرد. نسبت شانس نیز مانند میزان است. به مخدوش شدگی و نحوه کنترل آن با طبقه بندی و استفاده از آزمون مانتل هنزل اشاره گردید.

کار با نرم افزار و محاسبه شاخص های اثر و رابطه در مطالعه مورد شاهد:

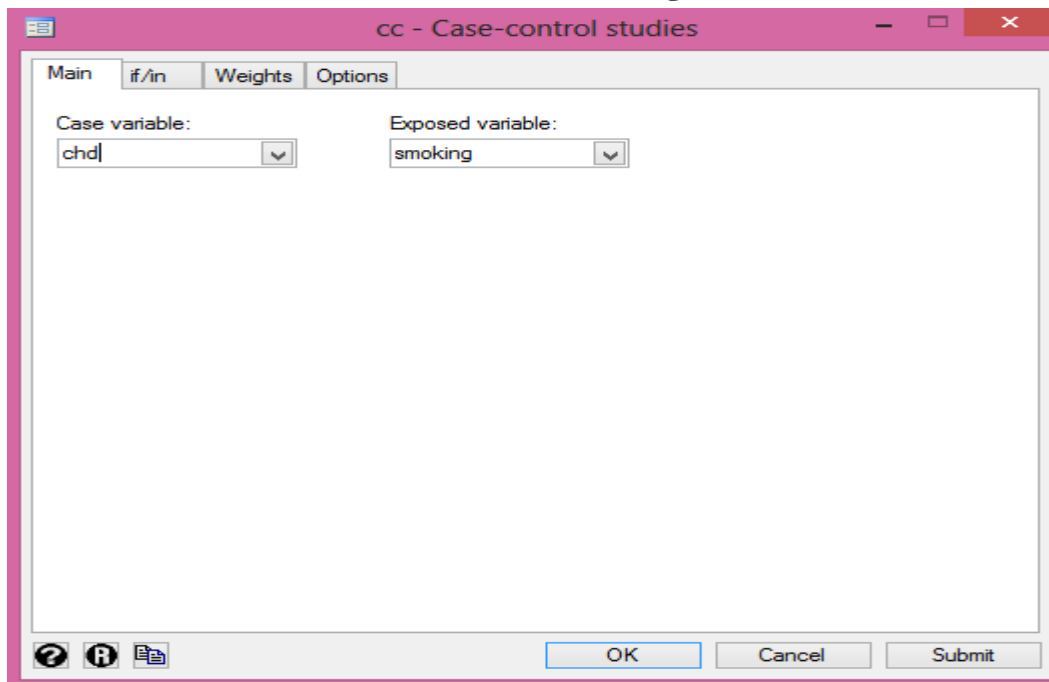
در فایل epidemiology چنانچه متغیر بیماری قلبی عروقی را بعنوان افراد مطالعه و بیماران و کنترل ها را با کد صفر و یک در نظر بگیرید و آن را مطالعه مورد شاهد فرض نمایید. به کد صفر برچسب شاهد و به کد ۱ برچسب کنترل بزنید. متغیر سیگار کشیدن را نیز بعنوان متغیر مواجهه در نظر بگیرید. ۱۵۴ مورد را با ۱۵۲۳ شاهد مقایسه کنید. فرض کنید با توجه به حجم نمونه پایین در گروه مورد که بر پایه یک مطالعه کوهورت انتخاب شدند همه افراد غیر بیمار قلبی عروقی به عنوان شاهد (۱۵۲۳ نفر) در نظر گرفته شدند.

death from chd	Freq.	Percent	Cum.
0	1,523	90.82	90.82
1	154	9.18	100.00
<b>Total</b>	<b>1,677</b>	<b>100.00</b>	

محاسبات این نوع مطالعه به صورت زیر از مسیر زیر قابل دسترسی است.

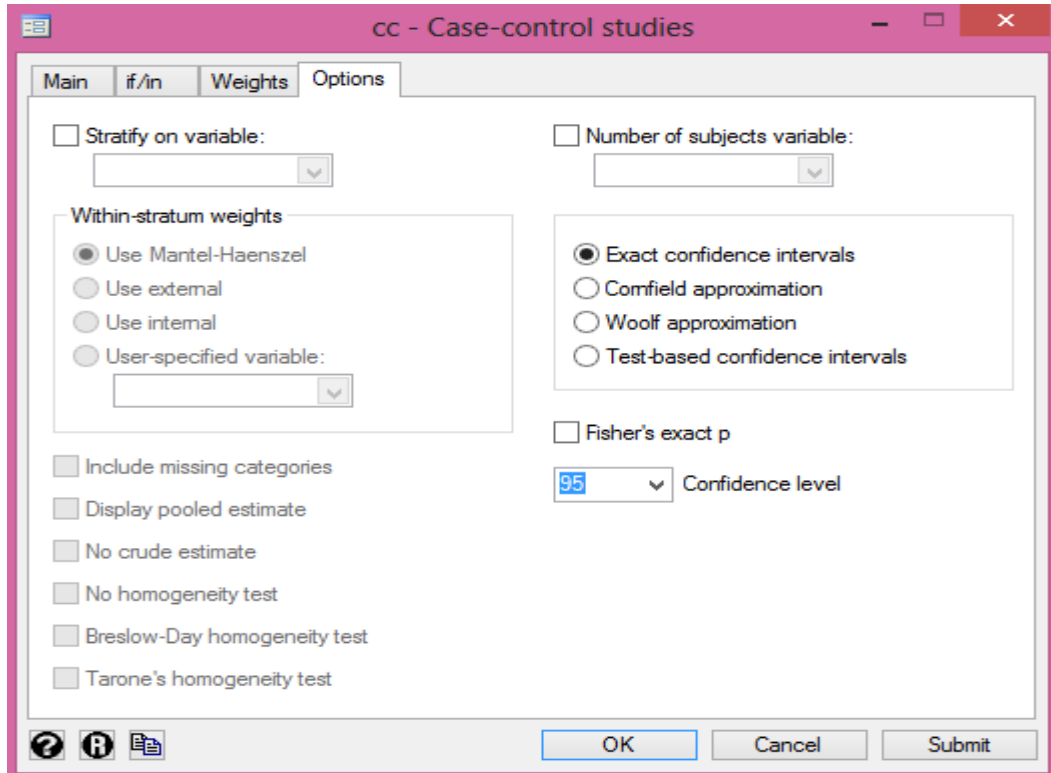


یا با دستور db cc کادر محاوره ای زیر باز می شود.





از کادر محاوره ای بالا در قسمت option می توان متغیر سوم را نیز وارد نمود. بر حسب متغیر سوم، طبقه بندی یا تطبیق را می توان انجام داد. همچنین می توان هر یک از روش های محاسبه نسبت شانس تطبیق شده از جمله روش ولف را انتخاب نمود. این روش ها در شکل زیر نمایش داده می شوند.



کادرهای مزبور را می توان با دستور db cc نیز بدست آورد. خروجی تحلیل بالا در زیر آمده است.

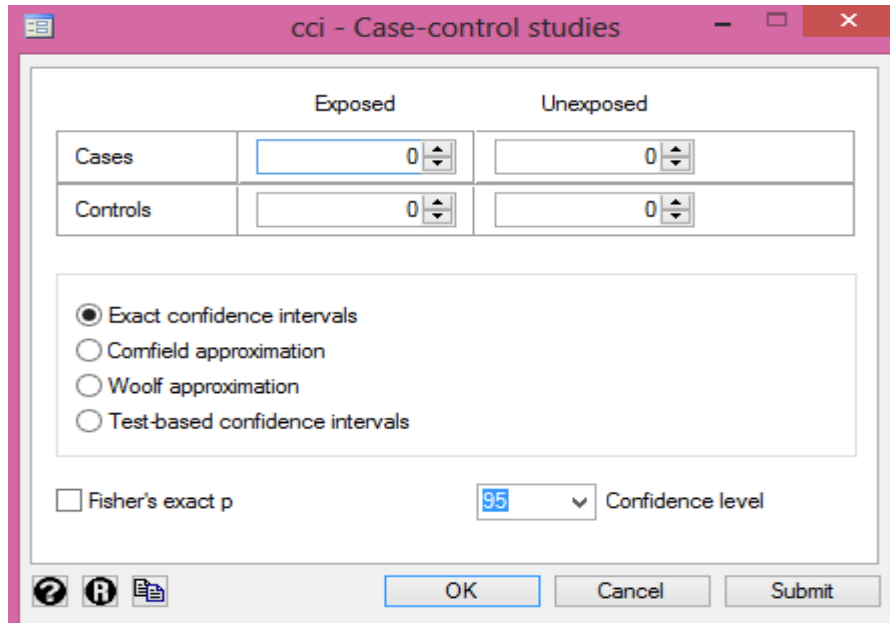
. cc chd smoking

	Exposed	Unexposed	Total	Proportion Exposed
Cases	88	66	154	0.5714
Controls	626	897	1523	0.4110
Total	714	963	1677	0.4258
	Point estimate		[95% Conf. Interval]	
Odds ratio	1.910543		1.34884	2.713874 (exact)
Attr. frac. ex.	.4765886		.2586222	.631523 (exact)
Attr. frac. pop	.2723364			

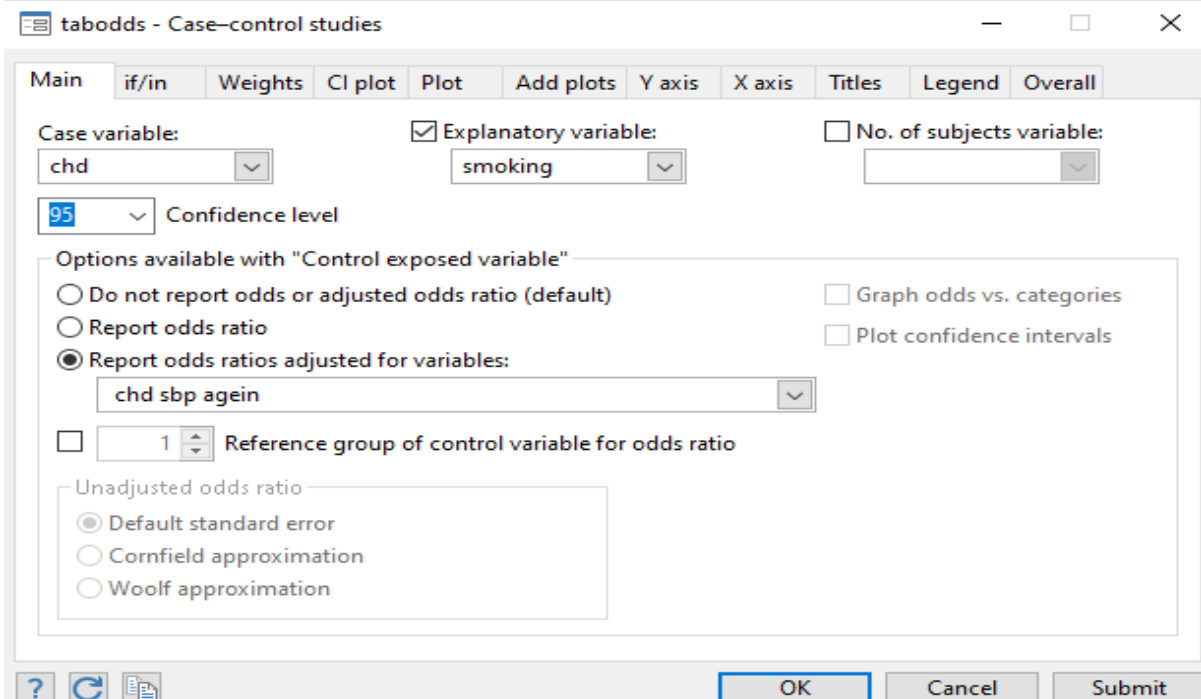
chi2(1) = 14.72 Pr>chi2 = 0.0001

نسبت شانس ابتلا به بیماری قلبی عروقی در کسانی که سیگار می کشند حدود دو برابر (1.91) افرادی است که سیگار نمی کشند. این رابطه از نظر آماری معنی دار است. حدود اطمینان آن نیز گزارش گردید.

همچنین در صورت داشتن ارقام موردها و شاهد‌ها و سابقه مواجهه آنها با دستور db cci می توان محاسبات بالا را انجام داد. کادر محاوره ای مزبور در زیر آمده است.



برای محاسبه شانس، تست هموجینیتی و برابر بودن نسبت های شانس و تست ترند از دستور db tabodds و یا کادر محاوره ای زیر می توان محاسبات را انجام داد.



می توان نسبت شانس را برای برخی متغیرها تطبیق و گزارش نمود. نام متغیرهای مورد نظر را باید در قسمت مربوطه اضافه نمود.

محاسبات مربوط به شانس و آزمون مانتل هنزل را با دستور db mhodds می توان انجام داد.

می توان با دستور db mcc تحلیل را برای مطالعات مورد شاهد همسان شده نیز انجام داد. فرض کنید در فایل مذکور موردها بیماران سرطان کبد با متغیر death و کنترل ها با نام متغیر smoking باشند. و از نظر متغیرهای سن و جنس همسان شده و مواجهه نیز الکل باشد. برای انجام تحلیل به صورت زیر عمل می شود.

. mcc death smoking

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	235	168	403
Unexposed	479	795	1274
Total	714	963	1677

McNemar's chi2(1) = 149.49 Prob > chi2 = 0.0000  
 Exact McNemar significance probability = 0.0000

Proportion with factor

Cases	.2403101		
Controls	.4257603	[95% Conf. Interval]	
difference	-.1854502	-.2144187	-.1564817
ratio	.5644258	.5143324	.619398
rel. diff.	-.3229491	-.3824941	-.2634041
odds ratio	.3507307	.2924309	.418974 (exact)

خروجی بالا نتیجه تحلیل مطالعه مورد شاهد همسان شده است. که نسبت شانس معنی داری را گزارش نموده است. در منوی epidemiology and related از مسیر زیر می توان محاسبات مربوط به انواع مطالعات اپیدمیولوژیک را انجام داد.

## مطالعات مداخله ای و کارآزمایی های بالینی:

از جمله مشخصات مهم مطالعات مداخله ای و کارآزمایی های بالینی می توان به آینده نگر بودن، داشتن مداخله و گروه کنترل و تصادفی سازی در کارآزمایی بالینی اشاره نمود. یک نوع از مطالعات مداخله ای مطالعات تجربی است که در حیوانات و محیط آزمایشگاه انجام می شود. کارآزمایی های بالینی در انسان انجام می شوند. در مطالعات فارماکوپیدمیولوژیک، معمولاً دارای چند فاز هستند و به صورت مرحله ای اجرا می شوند. در فاز صفر مطالعات میکرو دوز در انسان انجام می شود. جذب، توزیع و متابولیسم دارو یا به عبارتی سرنوشت دارو در بدن انسان (Pharmacokinetic) و مکانیسم اثر دارو (Pharmacodynamic) در بدن انسان مطالعه می شود. این فاز حدود ۱۰ تا ۱۵ نفر حجم نمونه دارد. در بعضی از تقسیم بندی ها فاز صفر جزء مطالعات پیش بالینی و به منظور بررسی دوز دارو و پاسخ در حیوانات انجام می شود.

در فاز یک کارآزمایی های بالینی اثرات فارماکولوژیک سمی دارو (حاد و وابسته به دوز) با حجم نمونه حدود ۲۰ تا ۵۰ نفر در افراد داوطلب انجام می شود. حداکثر دوز قابل تحمل (MTD) در این فاز تعیین می شود. در فاز دو اثرات توکسیک فارماکولوژیک دارو و تا حدودی اثربخشی داروی جدید و تعیین دوز بهینه ارزیابی می شود. حجم نمونه حدود ۲۰ تا ۳۰۰ نفر است. در فاز سه عمدتاً کارایی دارو و به میزان کمتر اثرات توکسیک آن با حجم نمونه ۳۰۰ تا ۳۰۰۰ نفر بررسی می شود. به فاز های مذکور فازهای قبل از بازار (Premarketing) نیز گفته می شود. فاز چهارم بعد از ورود دارو به بازار است. در واقع اثربخشی و اثرات توکسیک در جمعیت های بزرگ تر بررسی می شود و به مدت طولانی تر مراقبت انجام می شود. در این فاز گروه کنترل وجود ندارد. در مطالعات مداخله ای محقق عمده "مواجهه" را به گروهی از افراد و یا جوامع تحمیل می کند. همانطور که قبلاً گفته شد، بدلیل اخلاقی، نمی توانیم مردم را با عوامل مخاطره آمیز که خطر بیماری یا مرگ را افزایش می دهند، مواجهه دهیم، اما می توانیم مواجهه (مثلاً سیگار کشیدن) را کاهش دهیم یا از روش درمانی دیگری استفاده کنیم. ارجح ترین نوع مطالعه مداخله ای، کارآزمایی شاهددار تصادفی است، که در آن مداخله (یا مواجهه) بطور تصادفی به گروهها (Community Trial) و یا به افراد (Individual Field Trial یا Clinical Trial) اختصاص می یابد. همه مطالعات مداخله ای، از نوع تحلیلی هستند زیرا اثر مواجهه یا مداخله را بررسی می کنند. برای بررسی اثر مواجهه، پیامد(های) مورد نظر ارزیابی و در گروههای مختلف مواجهه با یکدیگر مقایسه می شوند. کورسازی (Blinding) یک راه مهم برای کاهش تورش اندازه گیری (Measurement Bias) است. در این روش آزمایش شوندگان و محققان نمی دانند که فرد به چه گروهی تعلق دارد. کورسازی محقق و بیمار نسبت به گروه مداخله یا کنترل (Double Blinding)، برای ارزیابی صحیح عوارض جانبی و پیامدهای نامطلوب امری ضروری است. اطلاع از اینکه بیمار در چه گروهی قرار دارد، می تواند گزارش عوارض جانبی توسط بیمار را تحت تاثیر قرار دهد. در بعضی مطالعات تنها یکی از دو گروه محقق و بیماران کورسازی (Single Blinding) می شود.

در بعضی موارد بیمار نمی تواند کورسازی شود. برای مثال، وقتی که هدف ما بررسی اثر خوابیدن در معرض یک پشه مالاریا در بروز پاراسیتامیا مالاریا باشد، دیگر کورسازی بیمار امکان پذیر نیست. اما این امکان وجود دارد که پاتولوژیست کورسازی شود. در عوض، در یک مطالعه مقایسه عود فتق (Hernia) در دو روش جراحی (تنها برش جراحی کمی متفاوت است)، امکان کورسازی آزمایش شوندگان وجود دارد درحالیکه امکان کورسازی دو سویه

ممکن نیست. تصادفی سازی به عنوان قلب کارآزمایی های بالینی شناخته می شود. تصادفی سازی معیار مهمی است که باعث تعادل گروه های مطالعه و اجتناب از سوگیری می شود. تصادفی سازی ممکن است به صورت ثابت و یا تطبیقی انجام شود. روشهای تصادفی سازی رایج تصادفی سازی ساده ، بلوک بندی شده و طبقه ای نام دارند. گروه کنترل Placebo concurrent در یک کارآزمایی استاندارد، معیار اصلی برای ارزیابی اثر داروست.

در شکل زیر مراحل مجوز گرفتن یک دارو را مشاهده نمائید.

**THE DRUG DISCOVERY, DEVELOPMENT AND APPROVAL PROCESS**

It takes 10-15 years on average for an experimental drug to travel from the lab to U.S. patients. Only five in 5,000 compounds that enter preclinical testing make it to human testing. One of these five tested in people is approved.

Discovery/ Preclinical testing		Clinical Trials			FDA	Phase IV
Years	6.5	Phase I	Phase II	Phase III	1.5	
Test Population	Laboratory and animal studies	20 to 100 healthy volunteers	100 to 500 patient volunteers	1,000 to 5,000 patient volunteers	Review process/ approval	Additional post-marketing testing required by FDA
Purpose	Assess safety, biological activity and formulations	Determine safety and dosage	Evaluate effectiveness, look for side effects	Confirm effectiveness, monitor adverse reactions from long-term use		
Success Rate	5,000 compounds evaluated	5 enter trials			1 approved	

انواع مختلف کارآزمایی های بالینی تصادفی شده شاهددار:

**(Randomized Controlled Clinical Trials)**

- Community Randomized Trial (CRT) : این نوع مطالعه با نام Cluster Randomized Trial هم مشهور است. در اینجا جوامع به جای افراد بطور تصادفی تقسیم می شوند. یک جامعه ممکن است یک روستا، یک گروه از روستاها، یک مدرسه و یا یک گروه از افراد تحت پوشش یک کلینیک بهداشتی باشد. هر جا که ممکن باشد، باید بیش از یک جامعه در هر گروه ( مداخله یا کنترل) قرار گیرد. تعداد خوشه های (Cluster) توصیه شده برای هر گروه حداقل شش خوشه می باشد. بدلیل وجود تشابه بین افراد درون یک جامعه نسبت به افراد جوامع دیگر، محاسبه حجم نمونه باید با توجه به اثر این همبستگی درون گروهی انجام شود. این اثر، design effect نامیده می شود. آنالیز داده ها باید با توجه به این موضوع که واحد تصادفی شده، جوامع هستند و نه تک تک افراد، انجام شود. در ارزیابی برنامه های اجتماعی و آموزش سبک زندگی وقتی واحد مطالعه بجای فرد، شهر، مدرسه یا واحدهای اجتماعی باشد از این روش استفاده می شود. هر چقدر همبستگی درون خوشه ها بزرگتر باشد مقدار واریانس کل خوشه ها بزرگتر از واریانس نمونه تصادفی خواهد بود.
- کارآزماییهای چند عاملی (Multi Factorial Trials) : در این مطالعات دو مداخله یا بیشتر مقایسه می شوند. مداخله ها یا به تنهایی و یا به صورت ترکیبی از چند مداخله طراحی و اعمال شده و با گروه کنترل

مقایسه می شوند. این طرح اجازه می دهد که اثر هم افزایی (Synergy) بین مداخله های مختلف بررسی شود.

- **Crossover Trials**: در این مطالعه، هر آزمایش شونده، کنترل خودش است و حداقل دو درمان (برای مثال، یک داروی جدید (درمان A) و یک داروی استاندارد (درمان B)) را در زمانهای مختلف در طول مطالعه دریافت می کند. ترتیب اینکه هر فرد (یا هر خوشه) کدام درمان را اول دریافت کند (A سپس B و یا برعکس)، بطور تصادفی مشخص می گردد. بین دو مداخله باید یک دوره پاک سازی یا استراحت (Wash Out Period) وجود داشته باشد، تا اثر مداخله قبلی (Carry Over Effect) به حداقل برسد. اگرچه این طرح فقط در جایی کاربرد دارد که مداخله اثر طولانی نداشته باشد، اما این امکان را فراهم می آورد که هر فرد، کنترل خودش باشد یعنی بیشترین تشابه بین گروه ها را باعث می شود. پراکندگی درون افراد حذف می شود. در بیماریهای مزمن و پایدار بیشتر کاربرد دارد.
- **Field Trials**: کارآزمایی هایی هستند که آزمایش شوندگان آنها بیماران نیستند، بلکه اعضاء یک جمعیت و یا یک زیرگروه از جمعیت عمومی هستند. در مواقع بسیاری بویژه در کشور های در حال توسعه، بیماران انتخاب شده از بیمارستان، گروههای خاصی هستند که یافته های حاصل از بررسی آنها قابل تعمیم به جمعیت عمومی نمی باشد. یافته های بدست آمده از Field Trial قابلیت تعمیم پذیری بیشتری دارند اگرچه بررسی یافته های بیوشیمیایی در اینگونه مطالعات مشکل تر است.
- **Effectiveness Studies**: این مطالعات تاثیر (Impact) برنامه های بهداشتی (مانند واکسیناسیون) را در شرایط واقعی می سنجند و نه در شرایط ایده آل و کنترل شده. این مطالعات معمولاً تصادفی شده و دو سو کور هستند؛ و بیشتر برای بررسی مداخلاتی انجام می گیرند که در مطالعات قبلی (شاید کوچکتر) اثربخشی و بی خطر بودن آنها نشان داده شده است.
- **Equivalence Trials**: در مطالعات قبلی، سعی بر این بود که نشان دهند یک مداخله بهتر از مداخله دیگر است (مثلاً درمان مالاریا با داروی جدید لاپدپ، بهتر از روشهای درمانی رایج است). اما در این نوع از مطالعات، هدف نشان دادن تساوی دو مداخله است. برای مثال، ما حدس می زنیم که درمان با داروی لاپدپ به تنهایی به همان اندازه موثر است که اگر به صورت ترکیبی با داروی زود اثر آرتسانات ترکیب شده باشد. این ترکیب درمانی میتواند از پیدایش مقاومت دارویی نسبت به لاپدپ جلوگیری کند. فرضیه ما در چنین مطالعاتی این است که دو مداخله دارویی (لاپدپ در مقایسه با لاپدپ+آرتسانات) به یک اندازه موثر هستند.
- **Titration Disign**: در فاز یک بر روی افراد سالم برای بدست آوردن Safety دارو انجام می شوند. در فاز دو برای تعیین دوز - پاسخ و کارایی دارو استفاده می شوند. در طراحی استاندارد این نوع مطالعه، مداخله بر اساس دوزهای از پیش تعیین شده انجام می شود. ابتدا مداخله با پلاسبو شروع و داده های پایه اندازه گیری می شود. سپس همه بیماران با دوز مشابه (حداقل دوز) شروع می کنند. اگر فرد معیارهای از پیش تعیین شده را داشت در فرد همان دوز ادامه می یابد و اگر معیار وجود نداشت دوزها افزایش می یابند. طراحی دیگری نیز وجود دارد بنام (Forced Dose Escalation) که در سطوح دوز پایین دارو کارایی

لازم را ندارد و اگر عوارض جانبی دارو مشاهده نشد دوز دارو افزایش می یابید. یعنی بر اساس معیار عوارض جانبی مداخله صورت می گیرد.

- **Enrichment** : بعضی از بیماران می توانند به دارو پاسخ دهند. یعنی پاسخ افراد به داروهای مختلف متفاوت است. به عبارتی بیماریها به داروها عکس العمل های متفاوتی نشان می دهند. در این طراحی افراد بدون کورسازی وارد می شوند. آنهایی که به دارو پاسخ می دهند وارد فاز دوم می شوند. کسانی که پاسخ نمی دهند خارج می شوند. سپس مطالعه بر روی افراد پاسخ دهنده با کورسازی دو طرفه و تصادفی سازی و مقایسه با گروه پلاسبو مورد بررسی قرار می گیرند. مهمترین مشکل این طرح تورش انتخاب است. یعنی خارج کردن افرادی که به دارو پاسخ نمی دهند. این طراحی از یک برنامه غربالگری استفاده می کند. تصادفی سازی نسبتا کاملی ندارد.
- **Blinder reader designs** : این طرح برای ارزشیابی فرآورده های دارویی و تصویربرداری پزشکی مورد استفاده قرار می گیرد. قبل از تصویر برداری یکسری دارو داده می شود که رادیولوژیست بر اساس آنها ارزشیابی عکسها را انجام میدهد و برای نامطلع کردن رادیولوژیست ها از داروی تجویز شده این مطالعه انجام می شود.
- **Zelen Design** : در طراحی زلن افراد واجد شرایط قبل از اعلام رضایت نامه شرکت در مطالعه ، مورد تصادفی سازی قرار می گیرند. بازوهای مطالعه تشکیل می گردد. سپس از افرادی که قرار است در گروه مداخله یا درمان قرار گیرند رضایت نامه اخذ می شود.
- **Venberg Design** : در این طرح افراد بر حسب حجم نمونه که واجد شرایط هستند معمولا به دو گروه تمایل یا ترجیح و گروه کارآزمایی بالینی تقسیم می شوند. این تقسیم ممکن است تصادفی باشد. سپس گروه کارآزمایی بالینی مورد تصادفی سازی قرار می گیرند و بازوهای مطالعه شکل می گیرد. در گروه ترجیح نیز بازوهای مطالعه تشکیل می شود.
- کارآزمایی های چند مرکزی : در این نوع طراحی یک کارآزمایی با پروتکل یکسان و مناسب در چند مرکز اجرا می شود. برای اندازه گیری برآورد کلی محاسبه اینترکشن کمی و کیفی در نتایج این مراکز مهم و شرط اساسی است. مقدار رابطه و جهت رابطه باید تعیین گردد.

### طراحی های متداول کارآزمایی های بالینی بر حسب هدف مطالعه

نشان دادن بهتر بودن اثربخشی درمان جدید نسبت به دارونما /درمان استاندارد با معنی داری آماری و بالینی کاهش مطلق رویدادهایی که از طریق یک درمان به دست می آید، شبیه به درمان دیگر با تفاوت در یک محدوده از پیش تعریف شده است. اثربخشی مداخله دو رژیم دارویی در درمان یکسان است.	برتری هم تراز / هم ارزی
نشان دادن کمتر بودن میانگین اثربخشی درمان جدید از درمان استاندارد در یک محدوده از پیش تعیین شده و اما تفاوت معنی دار بالینی کمتری ندارد. اثربخشی داروی جدید نسبت به داروی موجود پایین تر نیست.	نا برتری / بدتر نبودن
هم ارزی یا هم تراز، برابری درمان جدید با دارونما /درمان استاندارد بر اساس پیامدهای بالینی مانند کاهش مرگ و میر یا سکنه مغزی	هم ارزی بالینی
هم ارزی یا برابری درمان جدید با دارونما /درمان استاندارد بر اساس پارامترهای فارماکوکینتیک مانند غلظت خون یا میزان اشباع گیرنده های دارو	هم ارزی زیستی



## انواع دیگر مطالعات مداخله ای :

انواع دیگری از مطالعات مداخله ای وجود دارد؛ هر چند که پایه و اساس آنها یکی است (دو یا چند گروه نسبتاً مشابه، که بطور تصادفی مداخله یا دارونما را دریافت می کنند)، اما هیچکدام شواهد محکمی را که یک RCT واقعی تولید می کند، بوجود نمی آورند. دو نوع اصلی برای این نوع از مطالعات بیان شده است :

### • Before/after studies

در این مطالعه میزان پیامد قبل از مداخله با میزان پیامد بعد از مداخله مقایسه می شود. در این طرح مطالعه، از کنترل‌های تاریخی استفاده می شود. بر این مطالعه همواره این ایراد وارد است که تغییرات مشاهده شده هیچ ارتباطی با مداخله انجام شده نداشته باشند و تنها بدلیل تغییر زمان (Secular Trends) بوجود آمده باشند. برای مثال، آموزش مسائل جنسی در مدارس ممکن است با کاهش بروز عفونت جدید همراه باشد؛ ممکن است این حالت حتی بدون گذاشتن دوره آموزشی هم اتفاق بیفتد. استفاده از این نوع مطالعه در ارزیابی خدمات بهداشتی بسیار متداول است. روشهایی برای نشان دادن اثر واقعی مداخله وجود دارد :

۱. پایش دقیق تغییرات دیگر در کنار مداخله انجام گرفته در جمعیت مطالعه.
۲. پایش داده های که نباید بوسیله مداخله تحت تاثیر قرار گیرند و نشان دادن این که آنها تغییر نکرده اند.
۳. پایش تغییرات پیامد مورد نظر در جمعیت عمومی و مقایسه بزرگی تغییرات پیامد مورد نظر در گروه مداخله با تغییر مشاهده شده در جمعیت عمومی.
۴. انتخاب جمعیتی که هر مداخله را به طور تصادفی دریافت کند (قابلیت اجرای اندکی دارد).

### • Non-randomized Controlled Trials

در این مطالعه آزمایش شوندهایی که مداخله را دریافت می کنند و آنهایی که در گروه کنترل هستند، به طور تصادفی انتخاب نمی شوند. برای مثال، افرادی که داوطلب دریافت مداخله بودند با آنهایی که داوطلب نیستند، باید از انجام آن پرهیز کرد.

## انواع طراحی های اپیدمیولوژی بالینی در تحقیقات درمانی سرطان ها :

در فاز یک کارآزمایی های بالینی ، افراد سالم و داوطلب وارد مطالعه می شوند ، اما در مطالعات مربوط به سرطان ها افراد بیمار و مبتلا به سرطان که در انتهای مرحله سرطان هستند وارد مطالعه می شوند. در این تحقیقات بالینی هدف پیدا کردن دارویی است با دوز مناسب که موثر و کارا باشد. دوز درمانی معمولاً با یک دهم تا یک سوم دوز کشنده حیوان (موش) شروع می شود.

**طراحی نوع A:** در این طراحی ، درمان بیماران مبتلا به سرطان با گروه سه نفره از بیماران و با دوز استاندارد و از قبل معلوم شده یک دارو شروع و سپس در گروه های بعدی با توجه به سمیت دارو ، دوز ها تشدید و یا کاهش می یابند. اگر توکسیسیته یا سمیت یا DLT (Dose Limiting Toxicity) در بیماران مشاهده نشد ، در گروه بعدی بیماران ( سه نفر ) دوز درمانی افزایش می یابد. چنانچه هر گونه مسمومیت یا عارضه و سمیتی مشاهده گردد دوز دارو افزایش نمی یابد و درمان با سه نفر دیگر ، با همان دوز قبلی ادامه می یابد. اگر یکی از شش نفر سمیت دارو را نشان دهد دوز دارو افزایش نمی یابد و طرح متوقف می شود. به عبارت دیگر توقف طرح بر اساس نتیجه سمیت در شش نفر آخری است. بیشترین دوز هنگامی است که حداقل یک ششم از بیماران ، مبتلا به مسمومیت بشوند. تعیین حداکثر دوز مجاز با احتیاط انجام می شود.

**طراحی B:** برای رفع مشکلات طراحی نوع A و صرفه جویی در وقت و حجم نمونه ، این طراحی با مشاهده و درمان یک نفر بیمار سرطانی شروع می شود. اگر مسمومیت در بیمار مشاهده نگردد ، دوز داروی درمانی در نفر بعدی افزایش نمی یابد. سپس اگر در این نفر مسمومیت مشاهده نشد دوز دارو در نفر بعدی افزایش می یابد و اگر مسمومیت مشاهده شد دوز دارو کاهش می یابد. اگر در بیمار بعدی مسمومیت مشاهده شود طرح متوقف می شود. این طراحی 1Up , 1Down نامیده می شود.

**طراحی C:** با مشاهده و درمان یک نفر بیمار با دوز خاصی طراحی درمان شروع می شود. اگر سمیت در دو نفر مشاهده نگردد دوز افزایش می یابد . اگر سمیت مشاهده شد دوز کاهش می یابد. غالباً طراحی نوع B و C ادغام و با نام B شناخته می شود. اگر مسمومیت در دو بیمار مشاهده نگردد ، دوز داروی درمانی در نفر بعدی افزایش می یابد. اگر مسمومیت مشاهده شد دوز دارو کاهش می یابد. اگر دو بیمار پشت سر هم مسمومیت مشاهده نگردد نفر سوم با دوز بالاتر درمان می شود. اگر مسمومیت در دو بیمار مشاهده شود دوز را کاهش و اگر بیمار سوم نیز مسمومیت نشان دهد طرح متوقف می شود.

**طراحی D:** با یک گروه سه نفره از بیماران طراحی درمان با دوز نزدیک به حداکثر دوز مجاز شروع می شود. اگر مسمومیت و سمیت در بیشتر از یک نفر نبود ، دوز در گروه بعدی بیماران افزایش می یابد. اگر یک نفر مسمومیت را نشان داد دوز درمانی در گروه بعدی بیماران کمتر می شود. اگر دو یا سه نفر مسمومیت را نشان دهد دوز درمانی در گروه بعدی بیماران کاهش می یابد.

**طراحی E:** طراحی با درمان دو نفر بیمار شروع می شود. اگر مسمومیت مشاهده نشد دوز داروی مورد استفاده افزایش می یابد. اگر یک مورد مسمومیت مشاهده شود دوز تغییر نمی کند. اگر در هر دو نفر مسمومیت مشاهده شود دوز کاهش می یابد. طراحی های دو مرحله ای نیز برای درمان سرطان ها وجود دارد. دوره درمان با این طراحی ها کاهش می یابد. حجم نمونه مورد استفاده نیز کمتر است. برای اطلاع بیشتر از این طراحی ها به کتاب های تخصصی کارآزمایی های بالینی مراجعه شود.

## آنالیز داده ها در مطالعات مداخله ای و کارآزمایی های بالینی:

### اندازه اثر (Measures of Effect)

اندازه اثر اصلی از یک مطالعه مداخله ای، خطرنسبی (Risk Ratio) و نسبت میزان (Rate Ratio) است. خطر منتسب (Attributable Risk) و خطر قابل انتساب به جامعه (Population Attributable Risk) نیز قابل محاسبه است.

در بررسی اثرات مداخله بر یک پیامد (بیماری)، مقدار ارزش کلینیکی این اثرات نیز اهمیت پیدا می کند. معمولاً، ارزش کلینیکی یک مداخله را با شاخص NNT (Number Needed to Treat) بیان می کنند. NNT عبارت است تعداد افرادی که باید مورد مداخله قرار گیرند، تا از یک مورد پیامد مورد نظر پیشگیری شود. محاسبه این شاخص برای کارآزمایی بالینی، برای تصمیم گیری بالینی بهتر درباره یک بیمار کمک کننده است. روش محاسبه NNT به این صورت است:

$$NNT = 1 / ARR$$

$$ARR = ARC - ART$$

**ARR : Absolute Risk Reduction** – مقدار خطر کاهش یافته در اثر مداخله است. با کم کردن خطر قابل انتساب گروه کنترل از گروه مداخله بدست می آید.

**ARC : Attributable Risk in Control group** – خطر قابل انتساب در گروه کنترل است

**ART : Attributable Risk in Treatment group** – خطر قابل انتساب در گروه مداخله است.

شاخص دیگر نیز NNH است. تعداد بیمارانی که باید درمان شوند تا یک مورد عارض جانبی در آنها رخ دهد تعریف می شود. بر خلاف NNT، این شاخص هر چه بیشتر باشد بهتر است. یعنی عوارض جانبی شیوع کمتری داشته باشد.

$$NNH = 1 / \text{Absolut Risk Increase}$$

### طرز محاسبه NNT با نرم افزار

با دستورات زیر می توان مقدار کاهش خطر مطلق و تعداد مورد نیاز برای درمان را برای تفاوت بین نسبت بهبود بیماران در گروه مداخله و نسبت بهبود بیماران در گروه کنترل محاسبه نمود.

`bcii #a #b #c #d [, options]`

`bcib #a #b #c #d #e #f #g #h [, options]`

	Exposed	Unexposed	Total
Cases	<i>a</i>	<i>b</i>	<i>a + b</i>
Noncases	<i>c</i>	<i>d</i>	<i>c + d</i>
Total	<i>a + c</i>	<i>b + d</i>	<i>a + b + c + d</i>

برای محاسبه NNT سه مثال زیر را در نرم افزار اجرا نمایید.

`. bcii 82 62 143 194`

```
Risk of improvement for control (p0): 0.242
Risk of improvement for intervention (p1): 0.364
Risk difference (p1 - p0): 0.122
Newcombe Method 10 95% CI: 0.040-0.203
Number needed to treat (Improvement): 8.179
Bender's 95% CI: 4.925-24.958
```

`. bcii 82 62 143 194, leve(90)`

```
Risk of improvement for control (p0): 0.242
Risk of improvement for intervention (p1): 0.364
Risk difference (p1 - p0): 0.122
Newcombe Method 10 90% CI: 0.053-0.190
Number needed to treat (Improvement): 8.179
Bender's 90% CI: 5.256-18.771
```

`. bcib 82 62 143 194 5 12 220 243`

```
Risk of improvement for control (p0): 0.242
Risk of improvement for intervention(p1): 0.364
Risk of deterioration for control (p0det): 0.047
Risk of deterioration for intervention (p1det): 0.022
Risk difference (p1 - p1det - p0 + p0det): 0.147
Newcombe Method 10 95% CI: 0.057-0.235
Number needed to treat (Benefit): 6.798
Bender's 95% CI: 4.258-17.618
```

## رویکرد آنالیز (Methods of Analysis) :

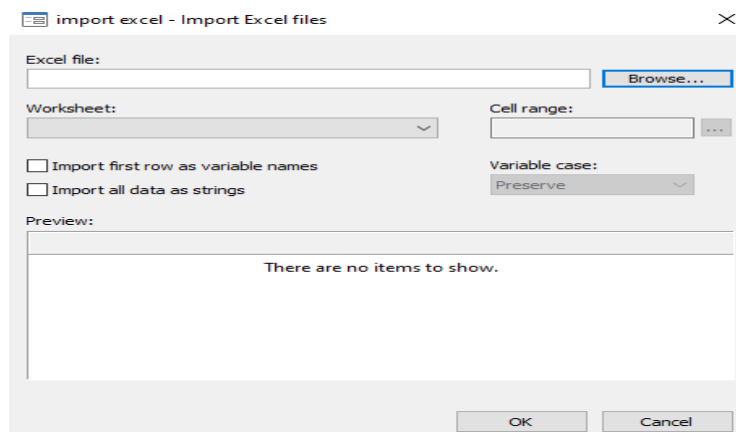
مهم است که کارآزمایی بالینی پروتکل داشته باشد و بر اساس آن، یک طرح جامع برای آنالیز داده ها در نظر گرفته و سپس آنالیز انجام شود. باید تعریف دقیقی از پیامد و مواجهه داشته باشید. آنالیز داده ها باید هدفمند و کاملاً براساس فرضیه های مشخص انجام گیرد. برای روش آنالیز، دو انتخاب وجود دارد :

**Per Protocol و Intention to Treat**

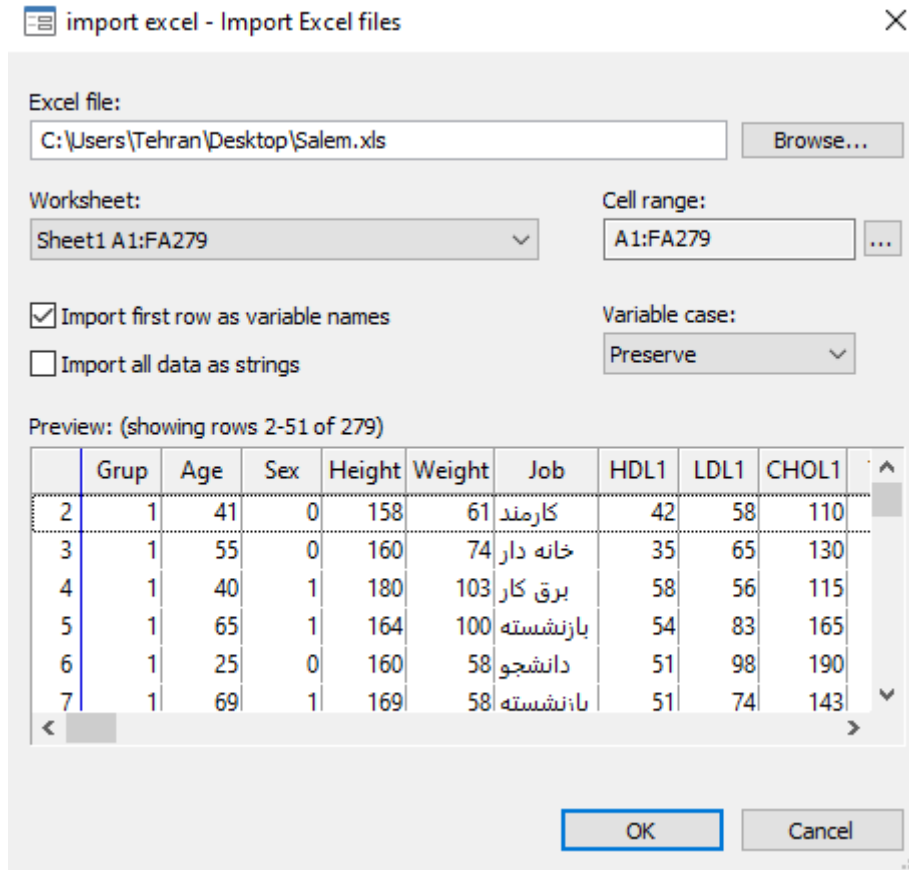
در روش Intention to Treat گروه ها تصادفی شده باقی می ماندند؛ اثر تصادفی کردن، یعنی یکسان بودن گروه ها برقرار است. و بر اساس آن گروه های مطالعه با هم مقایسه و تحلیل انجام می شوند. ولی در روش دوم بر اساس نوع مداخله ای که بیماران دریافت نمودند تحلیل انجام می شود. و اثر تصادفی سازی ممکن است از بین برود. در کارآزماییهای بزرگ و طولانی مدت، روش آنالیز های بینابینی (Interim analysis) نیز مورد استفاده قرار می گیرد. در مثال کارآزمایی واکسن پنوموکوک، حجم نمونه ۵۰۰۰۰ نفر بر اساس میزان مرگ قابل انتظار ناشی از بیماری پنوموکوک و مقدار اثربخشی واکسن مورد مطالعه محاسبه شده است. اما در محلی مانند آفریقا، تشخیص دقیق علت مرگ بسیار مشکل است بویژه اگر مرگ ناشی از بیماری پنوموکوک مدنظر باشد. همچنین، مقدار اثربخشی واکسن ممکن است با آنچه ما از یافته های مطالعات کوچکتر بدست آورده ایم، متفاوت باشد. بنابراین، داده های گردآوری شده بعد از ۲/۵ سال (بستگی به سرعت ورود افراد به مطالعه دارد)، مورد آنالیز قرار می گیرند. اگر اثربخشی واکسن پنوموکوک در این مرحله، بدون هرگونه تردیدی، بیشتر از حد انتظار باشد، دلیلی برای ادامه مطالعه و بالابردن حجم نمونه وجود ندارد. اگر اثربخشی مشاهده شده کمتر از چیزی باشد که انتظار داشتیم، باید گرفتن نمونه های بیشتر مدنظر قرار گیرد. البته Interim Analysis باید توسط گروهی غیر از محققین پژوهش انجام گیرد. این نوع تحلیل در کارآزمایی های متوالی بیشترین استفاده را دارد. نکات مهم و پیچیده ای در مورد این روش آنالیز وجود دارد. در یک کارآزمایی بالینی هر چند وقت یک بار (یعنی با چه فاصله زمانی) این آنالیز را باید انجام داد؟ آیا آنالیز مکرر داده ها، خدشه ای بر پیش فرضهای آماری آزمونهای مورد استفاده وارد نمی کند؟ توقف کارآزمایی وقتی که دلیل قانع کننده ای به نفع یا بر له اثربخشی یک مداخله وجود ندارد، کاملاً اشتباه است.

**مثال کاربردی و تعیین اثر درمان (Treatment Effect) در یک کارآزمایی بالینی با نرم افزار**

فایل salem به صورت یک فایل اکسل را از آدرس داده های کتاب دانلود و در دستکتاپ ذخیره نمائید. از منوی فایل و قسمت import فایل مذکور را به نرم افزار فراخوان نمائید. شکل زیر این مسیر را نمایش می دهد.



وقتی داده های فایل اکسل را فراخوانی نمودید اگر در ردیف اول فایل اکسل نام متغیرها وجود دارد باید تیک مربوطه را بزنید. در شکل زیر این مسیر نمایش داده شد.



در واقع دستور زیر را اجرا نمودید:

**import excel "C:\Users\Tehran\Desktop\Salem.xls", sheet("Sheet1") firstrow**

اکنون فایل مذکور که اکسل فراخوانی شده است برای تحلیل با استاتا آماده است. این فایل مربوط به داده های یک کارآزمایی بالینی سه سو کور است. ۱۰۵ بیمار دچار یبوست در متغیری به نام Grup در سه گروه پلاسبو و آموزش اصلاح سبک زندگی، درمان نوع A و درمان نوع B به صورت موازی وارد مطالعه گردید و هیچ یک از بیماران، درمان گر و تحلیل گر از گروه ها اطلاع ندارند.

تصادفی سازی در سه گروه انجام و در هر گروه ۳۵ بیمار قرار گرفت. هدف از این مطالعه تحلیل و مقایسه سه گروه برای شدت یبوست (متغیر H) و تعداد بار اجابت مزاج به عنوان پیامد اولیه (متغیر I) و متغیرهای قند و لیپید به عنوان پیامدهای ثانویه تعریف شده است.

مقایسه ها بصورت قبل و بعد در هر گروه و همچنین به صورت بین گروهی انجام می شود. شاخص های خلاصه متغیرهای پایه (سن، جنسیت، وزن، قد و ...) در همه بیماران را با دستور sum مشاهده و با دستور زیر این شاخص ها را بر حسب گروه ملاحظه نمائید.

by Grup, sort : summarize Age Height Weight HDL1 LDL1 CHOL1 TG1 FBS1 TSH H1 I1

## روش های اپیدمیولوژی و پژوهش با Stata

```
. sum Age Height Weight HDL1 LDL1 CHOL1 TG1 FBS1 TSH H1 I1
```

Variable	Obs	Mean	Std. dev.	Min	Max
Age	105	45.48571	12.84457	20	69
Height	105	168.219	9.284997	154	190
Weight	105	67.17143	11.67547	45	115
HDL1	105	53.20952	8.085713	35	82
LDL1	105	62.58095	18.58648	36	110
CHOL1	105	123.1048	35.25646	72	195
TG1	105	96.51429	18.53692	63	140
FBS1	105	92.31429	13.14806	67	118
TSH	105	2.505524	.5061658	1.7	3.5
H1	105	1.819048	.5847575	1	3
I1	105	3.161905	.7736503	1	4

با انجام آنالیز واریانس تفاوتی بین متغیرهای پایه بین سه گروه وجود ندارد. فقط متغیر HDL1 بین سه گروه معنی دار است. با دستور زیر این مقایسه را برای بقیه متغیرها انجام دهید.

```
. oneway HDL1 Grup, tabulate
```

Grup	Summary of HDL 1		
	Mean	Std. dev.	Freq.
1	50.571429	8.2543597	35
2	53.428571	8.4341244	35
3	55.628571	6.8902562	35
Total	53.209524	8.085713	105

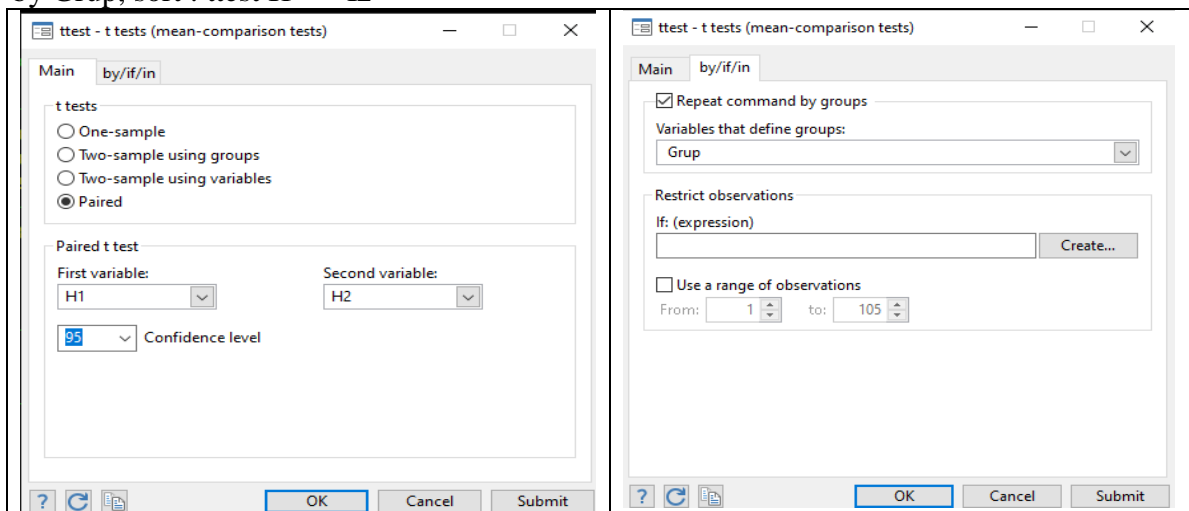
Source	Analysis of variance			F	Prob > F
	SS	df	MS		
Between groups	450.07619	2	225.038095	3.62	0.0304
Within groups	6349.31429	102	62.2481793		
Total	6799.39048	104	65.3787546		

Bartlett's equal-variances test:  $\chi^2(2) = 1.5813$  Prob> $\chi^2 = 0.454$

مقایسه اختلاف مقادیر متغیرهای پیامد را با آزمون تی زوجی بر حسب گروه انجام دهید. مسیر معرفی متغیرها و انجام تحلیل در زیر آمده است. دستور انجام آزمون تی زوجی برای متغیرهای پیامد اولیه:

```
by Grup, sort : ttest H1 == H2
```

```
by Grup, sort : ttest I1 == I2
```



نتایج نشان داد که مقادیر پیامد در هر یک از گروه ها اختلاف آماری معنی داری داشته و به عبارتی در هر سه گروه مداخله های انجام شده موثر بوده است. اکنون نوبت به محاسبه مقادیر اختلاف تغییرات ایجاد شده در هر یک

از گروه ها و در واقع مقایسه مقادیر اختلاف بین سه گروه مورد مداخله است. این کار را با آنالیز واریانس انجام می دهیم. دستورات محاسبه اختلاف مقادیر قبل و بعد و ساخت متغیر جدید H و I در زیر آمده است.

```
. gen diff_H = H2- H1
. gen diff_I = I2- I1
. sum diff_H diff_I
```

Variable	Obs	Mean	Std. dev.	Min	Max
diff_H	105	2.038095	1.184222	0	4
diff_I	105	-1.07619	.816833	-3	1

اکنون با آنالیز واریانس متغیر پیامد H را بین سه گروه مقایسه می کنیم. نتایج به صورت زیر است:

```
. oneway diff_H Grup, bonferroni tabulate
```

Grup	Summary of diff_H		
	Mean	Std. dev.	Freq.
1	2.8	.90097986	35
2	2.5142857	.81786769	35
3	.8	.63245553	35
Total	2.0380952	1.1842217	105

Source	Analysis of variance			F	Prob > F
	SS	df	MS		
Between groups	81.9047619	2	40.952381	65.33	0.0000
Within groups	63.9428571	102	.626890756		
Total	145.847619	104	1.40238095		

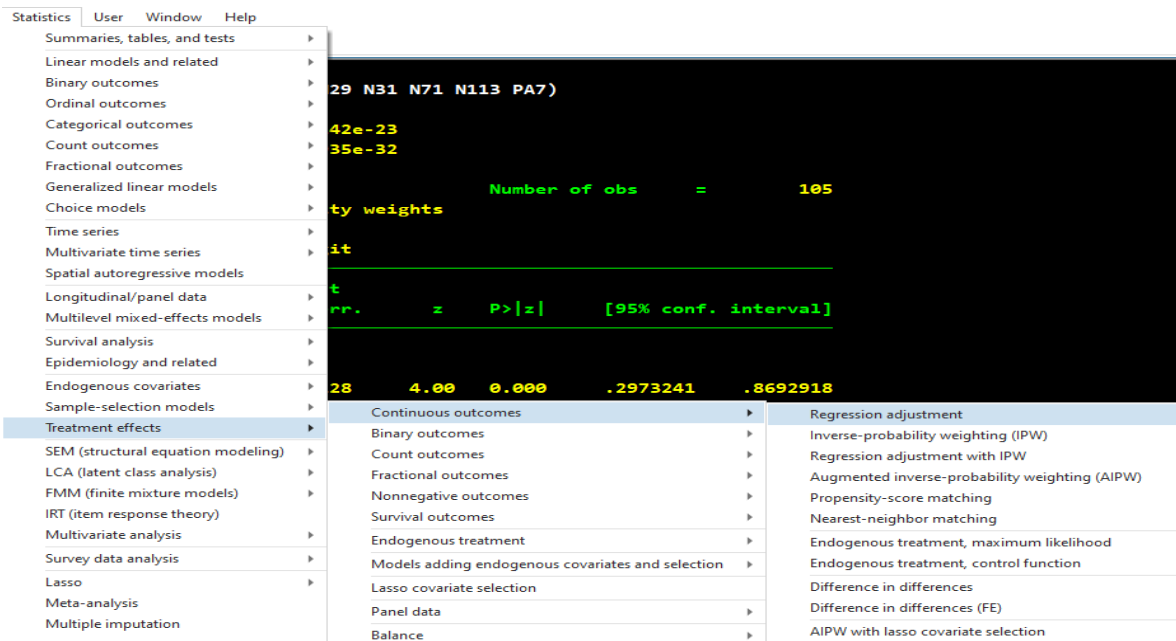
Bartlett's equal-variances test:  $\chi^2(2) = 4.2286$  Prob> $\chi^2 = 0.121$

Comparison of diff_H by Grup (Bonferroni)		
Row Mean- Col Mean	1	2
2	-.285714 0.403	
3	-2 0.000	-1.71429 0.000

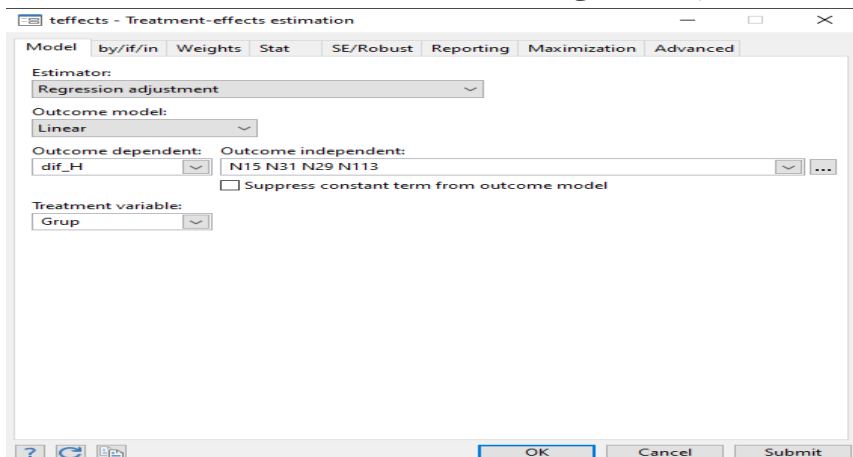
در خروجی بالا نتایج مقایسه گروه ها نشان می دهد بین گروه ۲ و ۱ اختلاف آماری معنی داری ( $P=0.403$ ) نیست. بین گروه ۳ و ۱ اختلاف آماری معنی داری ( $P=0.0001$ ) وجود دارد. بین گروه ۲ و ۳ نیز اختلاف آماری معنی داری ( $0.0001$ ) وجود دارد. گروه ۳ افرادی بودند که فقط آموزش های اصلاح شیوه زندگی را دریافت داشتند. گروه ۱ و ۲ دریافت کننده درمان دارویی و درمان گیاهی بودند. شما آنالیز واریانس را برای متغیر I انجام دهید و نتایج را گزارش کنید.

در این فایل داده های مربوط به تغذیه (متغیرهای N1 تا N113) و فعالیت فیزیکی (متغیرهای PA1 تا PA7) افراد نیز جمع اوری شده است. با مقایسه این متغیرها در سه گروه مقدار نشستن افراد بین سه گروه با متغیر (PA7) و دریافت برخی مواد غذایی از جمله موز، لبنیات و حبوبات متفاوت بوده که با آنالیز واریانس هر یک از این متغیرها می توان نتایج را دید. Dofile دستورات با نام salem را می توانید دانلود و نتایج را مقایسه کنید.

اکنون برای تعیین اثر درمان (Treatment Effect) از مسیر زیر اقدام کنید.



متغیرها را طبق شکل زیر به نرم افزار معرفی کنید.



در واقع با اجرای دستور `teffects ra (dif_H N15 N31 N29 N113) (Grup)` آنالیز انجام می شود. خروجی زیر نشان میدهد مقدار estimate potential-outcome mean برای مداخله گروه ۱ از بقیه گروه ها موثرتر و بین گروه ۱ و ۲ اختلاف آماری معنی داری در estimate average treatment effect on the treated وجود ندارد. بین مداخله گروه ۳ و ۱ نیز متوسط اثر درمان تفاوت آماری معنی دار وجود داشت.

		Number of obs = 105				
		Estimator : regression adjustment				
		Outcome model : linear				
		Treatment model: none				
	dif_H	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]
<b>ATE</b>						
Grup						
(2 vs 1)		-.3496814	.2314591	-1.51	0.131	-.803333 1.039701
(3 vs 1)		-2.214593	.1803854	-12.28	0.000	-2.568142 -1.861044
<b>POmean</b>						
Grup						
1		2.996106	.1403248	21.35	0.000	2.721074 3.271137



## خطاهای متداول در پژوهش های اپیدمیولوژیک

در تحلیل نتایج مطالعات اپیدمیولوژیک توجه به ماهیت رابطه و اثر بدست آمده بسیار مهم و اساسی است. رابطه ی بین مواجهه و بیماری به طور کلی تحت چهار عنوان زیر، قابل بررسی است:

۱. شانس (خطای تصادفی)

۲. تورش یا سوگرایی

۳. مخدوش شدگی

۴. ارتباط و علیت (شامل علیت معکوس نیز می شود)

قبل از پرداختن به خطاها و محاسبه آنها با نرم افزار و انجام تحلیل حساسیت و تطبیق رابطه ها بر اساس خطاها ، به شبیه سازی آماری (Statistical Simulation) که در کمی نمودن خطاها و کاهش نقش خطاها در بررسی روابط مورد استفاده قرار می گیرد پرداخته می شود. در قسمت های مدل سازی آماری نیز به شبیه سازی های آماری اشاره شده بود. لیکن در این قسمت با توضیحات مختصر به این مبحث پرداخته می شود.

## شبیه سازی آماری: Statistical Simulation

شبیه سازی (Simulation) شاخه ای از علم آمار است که به کمک انجام آزمایش های مصنوعی یا تکراری که بسیار به وضعیت واقعیت شبیه می باشند به سوالات مربوط به دنیای واقعی پاسخ می دهند. یکی از دلایل مهم انجام شبیه سازی آماری در هنگامی است که محقق قادر نیست آزمایش واقعی یا پژوهش واقعی را انجام دهد، محدودیت زمانی و مالی و عدم ناتوانی محقق می باشند. شبیه سازی آماری زیر مجموعه ای از شبیه سازی کامپیوتری است. مدل های کامپیوتری در گروه های مدل های تصادفی (Stochastic) یا مدل های قطعی (Deterministic)، مدل های ایستا یا پویا (Static and Dynamic) ، و مدل های پیوسته یا گسسته تقسیم می شوند. یکی از مفاهیم مهم در شبیه سازی آماری ، روش های مونت کارلو (Monte Carlo) است. این روش ها یک کلاس از الگوریتم های محاسباتی هستند که بر مبنای تکرار نمونه های تصادفی هستند. روش مونت کارلو در دهه ۱۹۴۰ توسط جان ون نیومن ، استانیسلاو اولام و نیکولاس متروپلیس ابداع گردید. این افراد بر روی پروژه های بمب اتمی در آزمایشگاه ملی لوس آلاموس آمریکا کار می کردند. انتخاب این اسم به دلیل تجلیل از قمارخانه مشهوری به نام مونت کارلو بود که عمومی اولام در آنجا اغلب پول خود را از دست می داد!

روش های مونت کارلو متفاوت اند. اما عموماً شامل گام های زیر است:

ابتدا دامنه ورودی های ممکن تعریف می شود. سپس ورودی ها طبق یک توزیع احتمال روی دامنه به طور تصادفی تولید می شوند. بر روی ورودی ها ، اعمال ریاضی و قطعی انجام و نهایتاً نتایج جمع می گردد. بنابراین برای تعریف مونت کارلو تعریف توافق شده ای وجود ندارد اما شبیه سازی مونت کارلو با استفاده از نمونه گیری های تکرار شونده ، یا باز نمونه گیری (Resampling) خصوصیات برخی پدیده ها را تعیین می کند.

هر روش شبیه سازی نیازمند داشتن اعداد تصادفی است. روش مونت کارلو همیشه نیازمند اعداد تصادفی واقعی نیست. امروزه به هر روشی برای تولید اعداد تصادفی ، روش مونت کارلو نیز گفته می شود. مهمترین گام در تولید اعداد تصادفی ، تولید اعداد تصادفی از توزیع یکنواخت است. تولید اعداد تصادفی از توزیع یکنواخت پایه تولید اعداد تصادفی از دیگر توزیع آماری است. توزیع های مهم آماری در مباحث قبلی مورد اشاره قرار گرفتند. روش های آماری

متفاوتی برای تولید اعداد تصادفی در توزیع های مختلف وجود دارد که پرداختن به آنها از حوصله این کتاب خارج است. نرم افزار Stata در مدل های تحلیل آماری که اجرا می کند این روش ها را با توجه به توزیع داده ها مورد استفاده قرار می دهد. و نتایج را بر اساس توزیع های مختلف گزارش می کند. در تحلیل داده ها بهترین حالت زمانی است که توزیع داده ها معلوم باشد و نمونه به اندازه کافی بزرگ باشد. اگر توزیع معلوم باشد و داده ها کم باشند از شبیه سازی استفاده می شود. اگر توزیع جامعه نامعلوم است ولی نمونه به اندازه کافی بزرگ باشد از تحلیل های مجانبی ( قضیه حد مرکزی) استفاده می شود. نامطلوب ترین حالت زمانی است که هم توزیع جامعه نامعلوم باشد و هم حجم نمونه کم باشد.

### بازنمونه گیری

وقتی توزیع جامعه نامعلوم و نمونه نیز کوچک باشد از **بازنمونه گیری (Resampling)** استفاده می شود. از روش های معمول بازنمونه گیری، می توان به بوت استرپ (Bootstrap)، جک نایف (Jackknife) و آزمون های جایگشتی (Permutation Test) اشاره نمود. واژه بوت استرپ (بند پوتین) از ماجرای یک کاپیتان بازنشسته ارتش در اواخر قرن ۱۸ که می خواست خودش را از ته دریاچه بدون هیچ کمکی بالا بکشد و از بند پوتین خود استفاده می نماید، گرفته شده است.

این روش در آمار نوعی شبیه سازی محسوب می شود. به اندازه دلخواه نمونه با جایگذاری از نمونه اصلی برداشته می شود و از میانگین این نمونه ها ، به عنوان برآوردی بهتر استفاده می شود. وقتی توزیع داده ها نامعلوم باشد از توزیع تجربی نمونه هایی که در دسترس هستند استفاده و شبیه سازی انجام می شود. بوت استرپ یک روش ساده ولی قدرتمند مونت کارلو آماری است که به برآورد توزیع نمونه ای یک آماره مانند میانگین می پردازد. محاسبات این روش طولانی است اما نرم افزار Stata این محاسبات را انجام و نتایج را گزارش می کند. در نرم افزار Stata روش های بازنمونه گیری از منوی اصلی آنالیز زیر قابل استفاده است.

Resampling	▶	Bootstrap estimation
Postestimation	▶	Report bootstrap results
Other	▶	Jackknife estimation
		Permutation tests
		Draw bootstrap sample
		Draw random sample

روش جک نایف شبیه روش بوت استرپ است و در برآورد اریبی و واریانس یک آماره مورد استفاده قرار می گیرد. ایده برآورد واریانس در این روش بر مبنای محاسبه مجدد و منظم آماره مورد نظر است. در هر مرحله یک یا چند داده از داده های اصلی حذف می شوند و از مجموعه داده جدید ، آماره مورد نظر محاسبه می شود. در نهایت نیز برآوردی از اریبی و واریانس آن آماره محاسبه می شود. در روش بوت استرپ به دلیل انتخابی تصادفی داده ها در هر بار اجرای کامل نتایج عددی متفاوتی را نشان میدهد. اما در روش جک نایف در هر اجرا نتایج یکسانی دارد. بوت استرپ علاوه بر برآورد واریانس برآوردگر ، کل توزیع آن را نیز بدست می آورد. اما جک نایف که برای کاهش واریانس استفاده می شود فقط برآورد واریانس نقطه ای برآورد شده را ارائه می کند. توصیه می شود برای کاهش خطا و اریبی ، از روش جک نایف و برای برآورد توزیع آماره از روش بوت استرپ استفاده شود. معمولاً بعد از

بکارگیری بوت استرپ ، برای کاهش خطا از روش جک نایف استفاده می شود. روش جک نایف برای میانه و صدک ها قابل استفاده نیست. برای میانگین ، تناسب ، نسبت ، نسبت شانس و ضرایب رگرسیون کاربرد دارد.

### کار با نرم افزار و انجام بازنمونه گیری (Resampling):

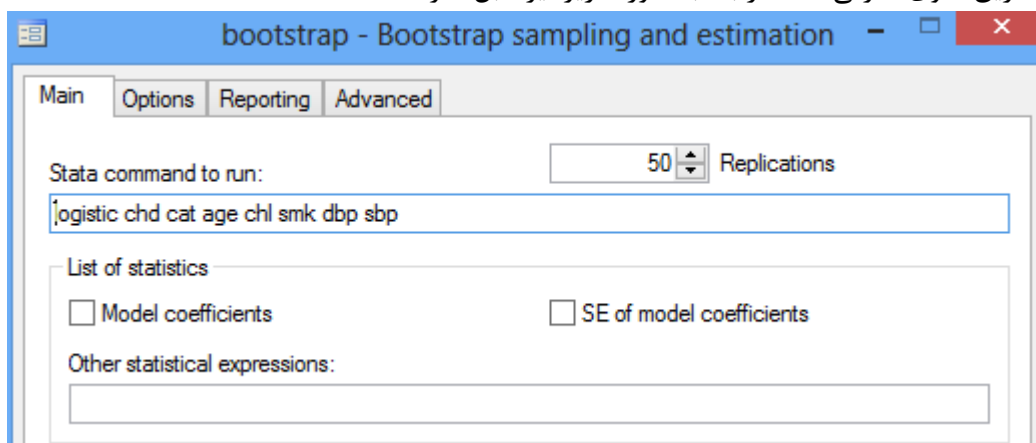
برای استفاده از روش های باز نمونه گیری که در بالا توضیح داده شدند فایل logistic را باز نمایید. ابتدا برای بررسی عوامل موثر بر بیماری قلبی عروقی (متغیر وابسته chd)، مدل رگرسیون لجستیک با متغیر های مستقل cat، age ، chl ، smk ، dbp ، sbp که قبلا در قسمت رگرسیون لجستیک این متغیرها توضیح داده شدند اجرا نمایید. نتیجه مدل و نسبت شانس برای متغیرهای مستقل به شرح زیر است. بجز متغیرهای فشارخون سیستولیک و دیاستولیک بقیه متغیرها معنی دار هستند.

chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
cat	2.401576	.8945696	2.35	0.019	1.157257	4.983824
age	1.040158	.0166359	2.46	0.014	1.008058	1.07328
chl	1.008605	.0033129	2.61	0.009	1.002133	1.015119
smk	2.269423	.689771	2.70	0.007	1.250836	4.117472
dbp	1.020613	.0137544	1.51	0.130	.9940082	1.047931
sbp	.9933866	.007814	-0.84	0.399	.9781889	1.00882

حجم نمونه در مطالعه مذکور ۶۰۹ نفر و با توجه به تعداد زیاد متغیر به نظر کافی نیست. چنانچه محقق قصد استفاده از روشهای بازنمونه گیری را داشته باشد و حجم نمونه بیشتری را مد نظر قرار دهد این مدل با روش های بازنمونه گیری به صورت زیر انجام و اجرا می شود.

### الف: روش بوت استرپ (Bootstrap)

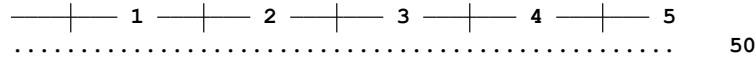
در روش بوت استرپ چنانچه مقدار تکرار بازنمونه گیری ۵۰ بار فرض گردد مدل با دستور زیر اجرا می گردد. این دستور از طریق منوی معرفی شده در بالا به صورت زیر نیز قابل اجرا است.



## روش های اپیدمیولوژی و پژوهش با Stata

```
. bootstrap, reps(50) : logistic chd cat age chl smk dbp sbp
(running logistic on estimation sample)
```

Bootstrap replications (50)



```
Logistic regression          Number of obs    =    609
                             Replications        =     50
                             Wald chi2(6)         =    27.94
                             Prob > chi2         =    0.0001
                             Pseudo R2          =    0.0828

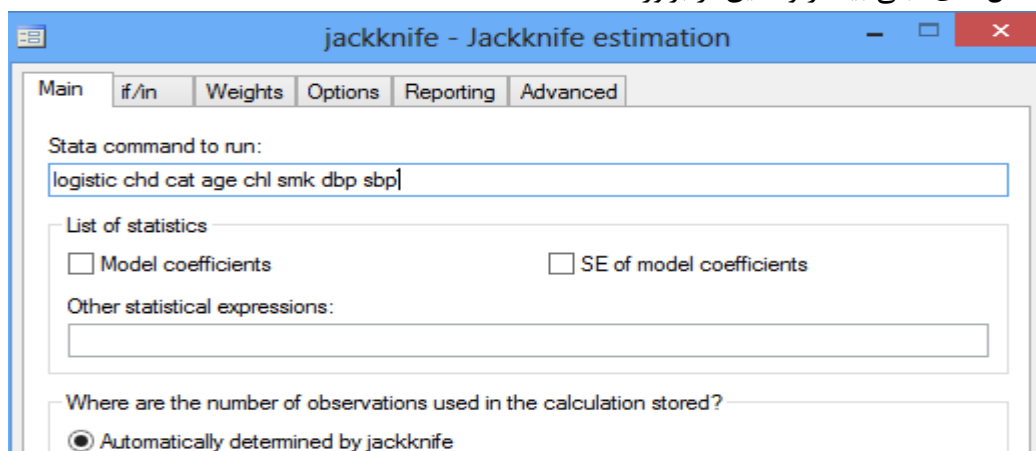
Log likelihood = -201.12197
```

chd	Observed Odds Ratio	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
cat	2.401576	.9384727	2.24	0.025	1.116527	5.165632
age	1.040158	.0180163	2.27	0.023	1.005439	1.076075
chl	1.008605	.0032774	2.64	0.008	1.002202	1.015049
smk	2.269423	.7889153	2.36	0.018	1.14819	4.485564
dbp	1.020613	.0166936	1.25	0.212	.9884134	1.053862
sbp	.9933866	.007713	-0.85	0.393	.9783838	1.00862
_cons	.0004256	.000677	-4.88	0.000	.0000188	.0096129

همانگونه که نتایج مدل بالا نشان می دهد مقادیر نسبت های شانس با اجرا کردن مدل از طریق بوت استرپ تغییر قابل ملاحظه ای بجز در انحراف معیار ضرایب مدل به وجود نیامده است و مانند مدل رگرسیون لجستیک ، کلیه متغیرهای مستقل بجز دو متغیر فشارخون سیستولیک و دیاستولیک ، معنی دار هستند. بنابراین اطمینان به مدل راند شده بیشتر شده و محقق با اطمینان بهتر می تواند به نسبت های شانس گزارش شده تکیه نماید. و نقش شانس را ناچیز بداند.

### ب: روش جک نایف (Jackknife)

در اجرای روش جک نایف برای مدل اجرا شده در بالا همانگونه که دستور و نتایج زیر نشان می دهند تغییرات قابل توجهی در مدل بوجود آمده است. در این روش برآورد واریانس برای برآوردها به صورت دقیق تر انجام می شود. در این مدل متغیر اول یعنی cat نسبت شانس غیرمعنی داری پیدا نموده است. خطای معیار نسبت شانس در این مدل نسبت به مدل های قبلی بیشتر و دقیق تر برآورد شده است.



## روش های اپیدمیولوژی و پژوهش با Stata

```
. jackknife : logistic chd cat age chl smk dbp sbp
      (running logistic on estimation sample)
```

```
Jackknife replications (609)
```

	1	2	3	4	5
.....					50
.....					100
.....					150
.....					200
.....					250
.....					300
.....					350
.....					400
.....					450
.....					500
.....					550
.....					600

Logistic regression	Number of obs	=	609
	Replications	=	609
	F( 6, 608)	=	5.58
	Prob > F	=	0.0000
Log likelihood = -201.12197	Pseudo R2	=	0.0828

	Odds Ratio	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
chd						
cat	2.401576	1.112678	1.89	0.059	.9668035	5.965604
age	1.040158	.0181092	2.26	0.024	1.005195	1.076337
chl	1.008605	.0035136	2.46	0.014	1.001728	1.015529
smk	2.269423	.6902355	2.69	0.007	1.248849	4.124025
dbp	1.020613	.0160962	1.29	0.196	.9894869	1.052719
sbp	.9933866	.0084346	-0.78	0.435	.9769595	1.01009
_cons	.0004256	.0006407	-5.16	0.000	.0000221	.0081815

### ج: روش آزمون جایگشتی (Permutation Test)

در این روش نیز مدل به صورت زیر اجرا می شود. منته در این روش نسبت ها نسبت به مدل های قبلی با احتیاط و محافظه کارانه گزارش شده است. برای هر متغیر یکصد بار نمونه با جایگذاری انجام گردید. سپس بر اساس وزن و حجم نمونه، تصحیح نسبت انجام می شود.

```
. permute chd_b, reps(100) : logistic chd cat age chl smk dbp sbp
      (running logistic on estimation sample)
```

```
Permutation replications (100)
```

	1	2	3	4	5
.....					50
.....					100

```
Monte Carlo permutation results           Number of obs = 609
```

```
command: logistic chd cat age chl smk dbp sbp
permute var: chd
```

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf. Interval]	
chd							
cat	.8761252	3	100	0.0300	0.0171	.00623	.0851761
age	.0393722	0	100	0.0000	0.0000	0	.0362167
chl	.0085683	3	100	0.0300	0.0171	.00623	.0851761
smk	.8195258	0	100	0.0000	0.0000	0	.0362167
dbp	.0204037	18	100	0.1800	0.0384	.1103112	.2694771
sbp	-.0066354	46	100	0.4600	0.0498	.3598434	.5625884
_cons	-7.76192	0	100	0.0000	0.0000	0	.0362167

از منوی بازنمونه گیری، می توان از داده های مطالعه در دست تحلیل، نمونه های تصادفی مختلفی به حجم دلخواه را تهیه نمود و در هر کدام از نمونه ها تحلیل را انجام و مقایسه ها را انجام داد. یا اینکه می توان با روش بوت استرپ نمونه مورد نظر را تهیه نمود و طبقه بندی و اثر خوشه را با متغیر مزبور در مدل در نظر گرفت.

**الف ( خطاهای تصادفی (Random Error) در مطالعات اپیدمیولوژیک**

وقتی اتفاق می دهد که تفاوت بین مقدار مشاهده شده و مقدار حقیقی تنها ناشی از شانس باشد. این خطا باعث ناپایداری در نتایج و رقیق شدن (Diluted) ارتباط در مطالعه می شود. خطای تصادفی در هنگام برآورد مقدار یک اثر (مانند OR، RR و...) رخ می دهد و احتمال وقوع آن برای مقادیر بالاتر و پایین تر از مقدار واقعی برابر است. این خطا شامل دو خطای نمونه گیری و خطای اندازه گیری می باشد. **خطای نمونه گیری** علت اصلی خطای تصادفی در مطالعات اپیدمیولوژی است. خطای نمونه گیری قابل حذف شدن نیست، اما می توان با روشهای آماری مقدار آنرا اندازه گیری کرد. یکی از اهداف طراحی مناسب مطالعه، کاهش این خطا تا سطحی قابل قبول (با توجه به محدودیتها مختلف موجود) است. وجود خطای نمونه گیری به این معنی است که هر زمان که یک فرضیه صفر آزمون می شود (به عنوان مثال  $H_0: OR=1$ )، میزان احتمال مشخصی وجود دارد که فرضیه صفر رد شود، در حالیکه در واقع  $H_0$  درست است (**خطای نوع ۱** که مقدار آن با حرف  $\alpha$  مشخص می شود)؛ و یا  $H_0$  رد نشود در حالیکه در واقع درست نباشد (**خطای نوع ۲** که مقدار آن با حرف  $\beta$  مشخص می شود). به طور کلی، معمولا با افزایش حجم نمونه، خطای تصادفی کنترل می شود. همچنین با داشتن مقدار اثری که پژوهشگر دنبال یافتن آن است و میزان وقوع بیماری در گروه غیر مواجهه، پژوهشگر از راههای زیر می تواند خطای نمونه گیری را کاهش دهد و بنابراین احتمال خطاهای ۱ و ۲ را کم کند:

**الف.** در مطالعات همگروهی: افزایش تعداد نمونه و یا طول دوره پیگیری.

**ب.** در مطالعات مقطعی و یا همگروهی: بررسی جمعیت پرخطر که در آن موارد بیشتری از بیماری وجود دارد.

**ج.** در مطالعات مورد-شاهدی: بررسی جمعیتی که شیوع مواجهه مورد مطالعه در آنها بیشتر باشد.

همواره بایستی در زمان طراحی مطالعه، با در نظر گرفتن احتمال مشخص برای خطاهای نوع ۱ و ۲، حداقل حجم نمونه لازم برای تشخیص مقدار اثر مورد نظر را محاسبه نمود. این حجم نمونه محاسبه شده باید در مقاله ای که در پایان کار نوشته می شود، بیان شود. خطای تصادفی در اندازه گیری می تواند به یک خطای منظم (تورش) در رابطه ی بین بیماری و مواجهه منجر شود.

**ب) تورش یا سوگرایی (Bias) در مطالعات**

خطای منظم در برآورد مقدار یک اثر، **تورش** یا **سوگرایی** خوانده می شود. این خطا باعث بی اعتباری نتایج مطالعه و ایجاد رابطه جعلی در پژوهش می شود. خطای منظم در برآورد یک اثر (مانند OR، RR و...) وقتی رخ می دهد که احتمال مقدار مشاهده شده به یک سمت بیشتر باشد، یعنی یا بیشتر از مقدار واقعی باشد و یا کمتر از آن. به عبارت دیگر، در هر بار اندازه گیری، نتایج همیشه بیشتر یا کمتر از حد واقعی است. و این بستگی به ماهیت خود خطای منظم دارد.

تورش نتیجه مجموعه ای از نقص ها در طراحی و یا اجرای یک مطالعه اپیدمیولوژی ناقص است. تاثیر تورش در زمان آنالیز داده های مطالعه قابل ارزیابی نیست، در حالیکه اگر متغیر **مخدوش کننده** اندازه گیری شده باشد، اثر مخدوش کنندگی قابل محاسبه است. برای اینکه تفاوت بین تورش و مخدوش کنندگی بهتر مشخص شود، به دو تفسیر از یک مطالعه اپیدمیولوژیک که رابطه بین یک بیماری و یک مواجهه را گزارش می کند، توجه کنید:

- **اگر رابطه مشاهده شده بدلیل تورش باشد:** به این معنی است که در جمعیت تحت مطالعه، بیماری ربطی به مواجهه نداشته است؛ و این یافته کاملاً اشتباه است.
- **اگر رابطه مشاهده شده بدلیل مخدوش کنندگی باشد:** به این معنی است که اگرچه بیماری با مواجهه مورد نظر رابطه دارد، اما برخی عوامل دیگر (که با مواجهه مورد مطالعه ارتباط داشته و نیز بطور مستقل با بیماری هم ارتباط دارند) می توانند این رابطه را توجیه کنند.

**مثال ۱.** یک مطالعه مورد-شاهدی را در نظر بگیرید که رابطه ای بین نوشیدن قهوه و سرطان مثانه کشف کرده است. یک توجیه برای وجود این رابطه این است که بیماران مبتلا به سرطان مثانه، خوردن قهوه خود را بهتر از گروه شاهد به یاد می آورند، این یک تورش است. توجیه دیگر ممکن است اینگونه باشد که استعمال دخانیات با نوشیدن قهوه همراه است و همچنین نشان داده شده که دخانیات، خطر ابتلا به سرطان مثانه را افزایش می دهد، در این حالت، استعمال دخانیات یک عامل مخدوش کننده است.

### دو نوع اصلی تورش در مطالعات اپیدمیولوژیک وجود دارد:

#### تورش انتخاب (Selection Bias) و تورش اطلاعات (Information Bias)

#### تورش انتخاب:

هنگامی روی می دهد که بیماران مورد بررسی در مطالعه، نماینده بیماران؛ و یا شاهد های مورد بررسی، نماینده مناسبی از گروه غیر بیمار (غیر مواجهه) در جامعه نباشد. مهمترین و شایعترین انواع تورش انتخاب، شامل موارد زیر است:

#### ۱- تورش در روش نمونه گیری (Bias in Sampling Method):

ناشی از طراحی نادرست روش اجرای مطالعه و انتخاب افراد است. برای مثال، در یک مطالعه مورد-شاهدی اگر روشی برای انتخاب گروه مورد و گروه شاهد انتخاب شود که آنها نمایانگر افراد بیمار و غیر بیمار جمعیت هدف نباشند، این سوگرایی رخ می دهد. اگر موردها (یا شاهدها) بر اساس معیاری به مطالعه وارد و یا از آن خارج شوند که در ارتباط با عامل مواجهه مورد مطالعه باشد، برآورد خطر نسبی با سوگرایی همراه خواهد بود. این یک مشکل رایج در مطالعات مورد-شاهدی است که معمولاً در اثر انتخاب نادرست گروه شاهد بوجود می آید.

**مثال ۲.** فرض کنید در یک مطالعه مورد-شاهدی بررسی علل سرطان ریه، گروه شاهد از بین افراد مبتلا به بیماری تنفسی غیربدخیم انتخاب شوند. سیگار کشیدن یکی از علل ابتلا به برونشیت و دیگر بیماریهای تنفسی غیربدخیم است. بنابراین، شیوع سیگار کشیدن در گروه شاهد بیشتر از شیوع آن در جمعیتی خواهد بود که موارد سرطان ریه از آن گرفته شده اند. در نتیجه، رابطه بین سیگار کشیدن و سرطان ریه کمتر از مقدار واقعی برآورد خواهد شد.

## ۲- تورش عدم پاسخ و خروج از مطالعه (Non-response bias and Loss to Follow-up):

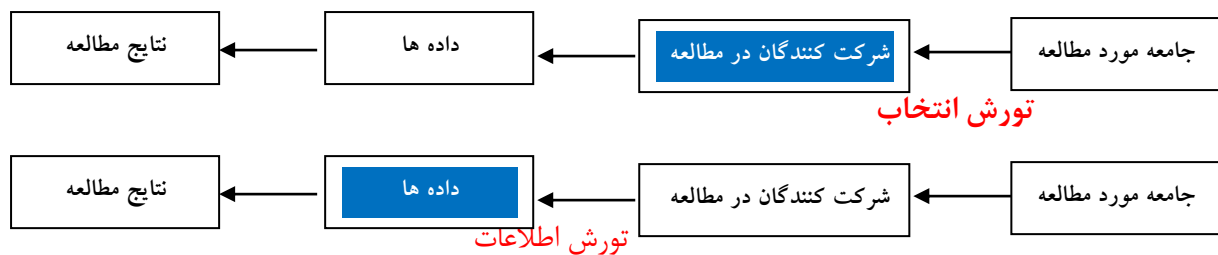
بطور کلی، در مطالعات اپیدمیولوژیک احتمال بیمار بودن (بدحال بودن) افرادی که از مطالعه خارج می شوند، بیش از افرادی است که در مطالعه باقی می ماند. در نتیجه، در اغلب مطالعات شیوع بیماری کمتر از مقدار واقعی برآورد می شود. بطور مشابهی، بعضی از انواع مواجهه مانند (مصرف تریاک، هروئین و الکل) همواره با میزان پاسخگویی اندکی همراه هستند. این یکی از دلایلی است که مطالعات بررسی این مواجهه ها همواره برآوردی کمتر از مقدار واقعی را در جوامع نشان می دهند. تورش عدم پاسخ لزوماً همواره بر ارتباط بین مواجهه و بیماری اثر نمی گذارد. فقط اگر میزان عدم پاسخدهی مرتبط با مواجهه در افراد بیمار و غیر بیمار متفاوت باشد، یا میزان پاسخدهی مرتبط با بیماری در دو گروه مواجهه و غیرمواجهه، تفاوت داشته باشد (Differential)، نتایج دچار سوگرایی می شوند.

### تورش اطلاعات:

وقتی اتفاق می افتد که اطلاعات جمع آوری شده در رابطه با مواجهه یا بیماری یا هر دو، به صورت منظم (و نه تصادفی)، اشتباه باشد. به عبارت دیگر، این سوگرایی در یک مطالعه زمانی رخ می دهد که افراد مورد مطالعه، بر اساس وضعیت بیماری، وضعیت مواجهه و یا هر دو آنها، به غلط دسته بندی (Misclassification) شده باشند.

- **دسته بندی نادرست غیر افتراقی (Non-differential Misclassification)** زمانی رخ می دهد که خطا در دسته بندی افراد بر اساس وضعیت مواجهه در گروه مورد و شاهد یکسان باشد. همچنین این تورش زمانی اتفاق می افتد که خطا در دسته بندی افراد بر اساس وضعیت بیماری در گروه مواجهه یافته و مواجهه نیافته یکسان باشد. این نوع اشتباه در دسته بندی، تقریباً همیشه منجر به برآورد کمتر از مقدار واقعی رابطه بین مواجهه و بیماری می شود. یعنی، جهت آن به سمت فرضیه صفر ( $OR=1$  یا  $RR=1$ ) است.
- **دسته بندی نادرست افتراقی (Differential Misclassification)** زمانی رخ می دهد که اشتباه در دسته بندی وضعیت بیماری، با وضعیت مواجهه ارتباط داشته باشد و یا بر عکس. برای مثال، در یک مطالعه مورد-شاهدی، یادآوری افراد گروه مورد از مواجهه قبلی شان با عوامل خطر ممکن است با یادآوری افراد گروه شاهد متفاوت باشد. این دسته بندی نادرست افتراقی ممکن است باعث سوگرایی برآوردها به سمت و یا برخلاف فرضیه صفر شود. پس می تواند باعث ایجاد روابط غیر واقعی گردد.

### • محل وقوع تورش انتخاب، تورش اطلاعات یا تورش طبقه بندی نادرست در پژوهش





**مثال ۳.** در یک مطالعه همگروهی بررسی رابطه بین مصرف چای داغ و سرطان مری نتایج به صورت خلاصه شده است. همانطور که مشاهده می کنید، کاهش اختلاف بروز (Rate Difference) سرطان مری از ۴۰۰۰۰۰ در سال به ۲۰۰۰۰۰ در سال و نسبت بروز (Rate ratio) از ۵ به ۱/۷ بدلیل دسته بندی نادرست غیر افتراقی گروهی از مصرف کنندگان چای داغ و گروه شاهد بوده است.

نسبت بروز (Rate ratio)	اختلاف بروز ( $1.5 \times$ )	میزان بروز ( $1.5 \times$ )	
<b>واقعیت (عدم اشتباه در دسته بندی)</b>			
۵/۰	۴۰	۵۰	۱۰۰۰۰۰۰ نفر مصرف کننده چای داغ
		۱۰	۵۰۰۰۰۰ نفر شاهد
<b>نصف مصرف کنندگان چای به اشتباه در گروه شاهد جای گرفته باشند</b>			
۱/۷	۲۰	۵۰	۵۰۰۰۰۰ نفر مصرف کننده چای داغ
		۳۰	۱۰۰۰۰۰۰ نفر شاهد (۵۰٪ در واقع چای داغ می خوردند)
<b>نصف مصرف کنندگان چای به اشتباه در گروه شاهد جای گرفته باشند و یک سوم شاهد ها در گروه مصرف کننده چای داغ طبقه بندی شده باشند.</b>			
۱/۲	۶	۴۰	۶۶۷۶۶۶ نفر مصرف کننده چای داغ (۲۵٪ در واقع چای داغ نمی خوردند)
		۳۴	۳۳۳۸۳۳ نفر شاهد (۶۰٪ در واقع چای داغ می خوردند)
<sup>۱</sup> در اینجا یک دسته بندی نادرست غیر افتراقی اتفاق افتاده است.			

تنها موارد معدودی وجود دارد که در آنها اثر دسته بندی نادرست غیر افتراقی در جهتتی خلاف فرضیه صفر است. برای مثال، اگر مواجهه بیش از ۲ حالت داشته باشد، اما دسته بندی نادرست فقط در دو حالت آن اتفاق افتاده باشد. همانطور که در جدول زیر مشاهده می کنید، بطور یکسان هم در گروه مورد و هم در گروه شاهد ۴۰٪ از افراد دارای مواجهه شدید، به اشتباه در گروه مواجهه خفیف جای گرفته اند. نتیجه این اشتباه، افزایش OR گروه مواجهه خفیف از ۲ به ۳/۱ است. یعنی در مسیر خلاف فرضیه صفر OR تغییر کرده است.

توجه داشته باشید که بطور مشابهی، دسته بندی نادرست غیر افتراقی افراد بر اساس متغیر مخدوش کننده، باعث می شود که اثر مخدوش کنندگی کمتر از میزان واقعی برآورد شود؛ یعنی، اثر مخدوش کننده باقی مانده (Residual Confounding) است. و این باعث خواهد شد که نتوان با روشهای آماری، اثر دیگر متغیرها (مخدوش کننده های احتمالی) را بر رابطه مشاهده شده بین مواجهه و بیماری مورد نظر بررسی کرد.

مثال تورش دسته بندی یا طبقه بندی نادرست غیرافتراقی:

مواجهه نیافته	مواجهه خفیف	مواجهه شدید	
واقعیت (عدم اشتباه در دسته بندی)			
مورد	۲۰۰	۶۰۰	۱۰۰
شاهد	۱۰۰	۱۰۰	۱۰۰
	OR=۲	OR=۶	
۴۰٪ مواجهه شدید ← مواجهه خفیف <sup>۱</sup>			
مورد	۴۴۰	۳۶۰	۱۰۰
شاهد	۱۴۰	۶۰	۱۰۰
	OR=۳/۱	OR=۶	

در اینجا یک دسته بندی نادرست غیر افتراقی اتفاق افتاده است.

**تورش یادآوری (Recall Bias):** این تورش افتراقی، بویژه در مطالعات مورد-شاهدی به عنوان یک مشکل عمده مطرح است، زیرا ممکن است موردها (بدلیل بیماری که دارند و انگیزه بیشتر برای یافتن دلیلی برای آن) وضعیت مواجهه را بهتر از افراد گروه شاهد بخاطر آورند. برای مثال، در بررسی ارتباط بدشکلیهای مادرزادی و استفاده از اسپری اسپرم کش، مادران کودکان مبتلا نسبت به مادران گروه شاهد انگیزه بیشتری برای یادآوری استفاده از این ماده را دارند.

**سوگرایی گزارش دهی (Reporting Bias):** نوعی تورش غیر افتراقی است که در آن افراد به دلیل نگرش ها و اعتقادات خود، تمایل به گزارش غیر واقعی از مواجهه یا عدم مواجهه خود یا نمونه های مورد پژوهش دارند. مانند انکار سیگار کشیدن، نوشیدن الکل یا رابطه جنسی مشکوک در جوامعی که این رفتارها به عنوان ضد ارزش شناخته می شود؛ یا:

پزشکان و پرستاران وقتی که در مورد وجود یک مشکل مثل اپیدمی وبا و عامل احتمالی آن مثل سبزی آلوده خبرهایی شایع شده باشد، موارد مبتلا و مواجهه مربوط به آنرا بیشتر گزارش می کنند.

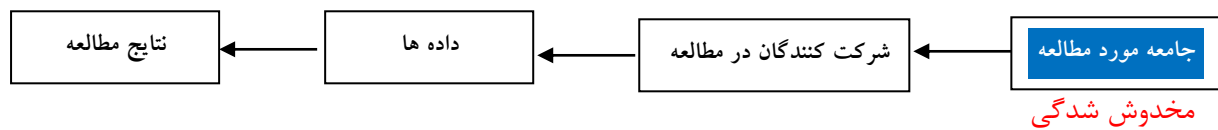
**سوگرایی مشاهده گر (Observer Bias):** مشاهده گرهایی که از وضعیت مواجهه افراد آگاهی دارند، ممکن است به صورت خودآگاه یا ناخودآگاه، پیامد مورد نظر را طوری اندازه گیری کنند که به نفع فرضیه مورد مطالعه باشد. برای مثال، در یک کارآزمایی بالینی بررسی تاثیر درمانهای غیر دارویی، ممکن است مشاهده گر فشار خون افراد گروه مداخله را کمتر و فشار خون افراد گروه شاهد را بیشتر از حد واقعی اندازه گیری و گزارش کند.

تورش می تواند در روش نمونه گیری یا در جمع آوری اطلاعاتی راجع به بیماری یا تماس با عامل خطر در یک مطالعه بوجود آید. همچنین در مطالعات مروری و متاآنالیز تمایل پژوهشگر به استفاده از نتایج مطالعاتی که با

نظریات خودش هماهنگی دارد و حذف نتایج مطالعات دیگر می تواند به نوعی تورش بنام تورش انتشار یا گزارش (Publication Bias) بیانجامد.

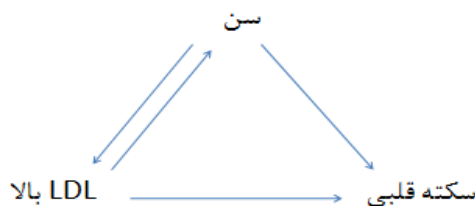
### ج) مخدوش شدگی (Confounding)

در بعضی از منابع معتبر مخدوش شدگی جزء تورش ها طبقه بندی می شود. و در بعضی از متون دیگر، جداگانه در نظر گرفته می شود. در قسمت قبل محل وقوع خطاهای انتخاب و اطلاعات در دیاگرام نمایش داده شد. در دیاگرام زیر محل وقوع مخدوش شدگی در جامعه مورد مطالعه که در بین همبستگی بین متغیرهای مورد بررسی در جامعه وجود دارد، نمایش داده می شود.

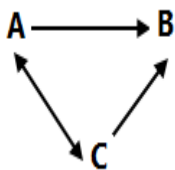


فرض کنید در یک مطالعه می خواهیم رابطه بین مقدار LDL سرم خون به عنوان متغیر مواجهه و خطر ابتلای به سکته قلبی (متغیر پیامد) را بررسی کنیم. متغیری به نام سن (متغیر سوم) به عنوان متغیر مخدوش کننده شناخته می شود زمانی که سه شرط زیر را داشته باشد:

- ۱- سن عامل خطر شناخته شده سکته قلبی باشد. به عبارت دیگر متغیر سوم به صورت مستقل با متغیر پیامد رابطه علیتی داشته باشد.
- ۲- سن با متغیر مستقل (مقدار LDL) رابطه داشته باشد. به عبارت دیگر متغیر سوم با مواجهه نیز ارتباط یک طرفه یا دو طرفه داشته باشد.
- ۳- در مسیر علیتی مواجهه و بیماری، سن به عنوان متغیر واسطه نقش نداشته باشد.

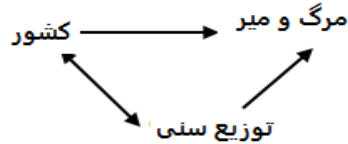


ارتباط متغیر سوم با متغیرها Exposure و Outcome به شرطی که ارتباط با A و B دارای جهت های مشخصی باشد. یعنی ارتباط متغیر مخدوش کننده (متغیر سوم) با پیامد یک رابطه است ولی رابطه آن با Exposure می تواند یک طرفه یا دوطرفه باشد. این رابطه می تواند به خطا موجب ایجاد رابطه غیر واقعی بین متغیر مواجهه و پیامد شود و یا برعکس رابطه واقعی بین این دو متغیر را مخفی کنند.



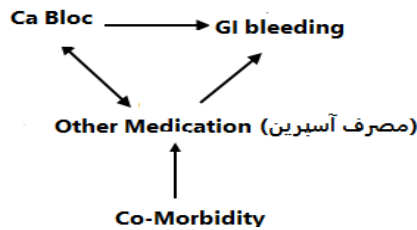
چند مثال برای انواع مخدوش کننده ها:

۱- ارتباط کشور (منطقه جغرافیایی) با مرگ و میر و متغیر سوم سن

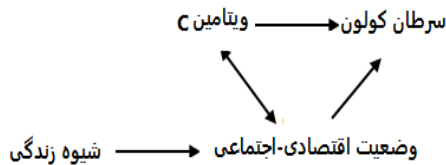


۲- رابطه بین Sexual activity به عنوان یک مواجهه برای ابتلا به ایدز و متغیر جنسیت

۳- رابطه بین مصرف داروهای بلاکر کلسیم با GI bleeding و متغیر سوم مثل مصرف آسپرین که این متغیر حاصل یک متغیر دیگر است. کسی که داروهای دیگر مصرف می کند، به خاطر این است که بیماری دیگری هم دارد. یعنی کسی که داروی ضد فشارخون می خورد ممکن است بیماری قلبی هم داشته و آسپرین هم مصرف نماید یا برعکس.



۴- وضعیت اقتصادی اجتماعی با ویتامین C ممکن است ارتباط یک طرفه یا دوطرفه داشته باشد. (توزیع متفاوت دارد) کسانی که وضع مناسبی دارند ویتامین بیشتری را مصرف می کنند (در ارتباط با یکدیگرند)



ارزیابی مخدوش کننده ها: در یک مطالعه مواجهه، جنسیت و پیامد، مالاریا است. (مطالعه مورد-شاهد)

		Outcome		
		Case	Control	
Exposure	Male	۸۸	۶۸	$OR = \frac{88 \times 82}{68 \times 62} = 1/71$
	Female	۶۲	۸۲	
		۱۵۰	۱۵۰	

جنسیت مرد یک ریسک فاکتور برای مالاریا است. اما ممکن است این رابطه به دلیل مخدوش کننده ها باشد. برای مثال شغل در این جا می تواند یک مخدوش کننده باشد. برای این که وضعیت شغل را ارزیابی کنیم باید ارتباط اش را با A و B ارزیابی کنیم.

ابتدا رابطه جنسیت و شغل را بررسی می کنیم (Confounder V.S Exposure)

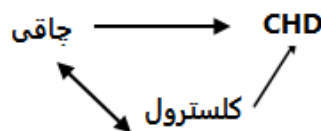
		شغل	
		خارج منزل	داخل منزل
Exposure	Male	۶۸	۱۳
	Female	۶۲	۱۳۱
		۸۱	۲۱۹

در این جا نیز مردها نسبت به زن ها بیشتر بیرون از منزل کار می کنند و بین جنسیت و محل کار  $OR=7/8$  ارتباط وجود دارد. رابطه شغل به عنوان Confounder را در مقابل Outcome بررسی می کنیم. در این جا هم پیامد با شغل با  $OR=5/3$  ارتباط دارد.

		Case	Control
		شغل	خارج منزل
داخل منزل	۸۷		۱۳
		۱۵۰	۳۱

$$OR = \frac{63 \times 13}{18 \times 87} = 5/3$$

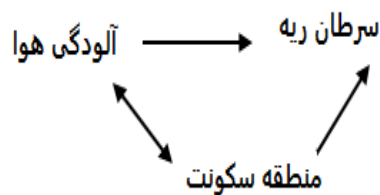
پس هر دو شرط برقرار است. شغل (مخدوش کننده) هم با پیامد و هم با جنسیت (مواجهه) ارتباط دارد. شرط سوم: متغیر مخدوش کننده (C) به عنوان یک واسطه (Intermediate) بین A و B عمل نکند. در این جا ارزیابی آماری وجود ندارد و ارزیابی بیولوژیک است. چون در این زنجیره علیتی بین A و B قرار ندارد. لذا به عنوان متغیر واسطه تلقی نشده و بنابراین شرط سوم هم برقرار است. و لذا شغل یک متغیر مخدوش کننده است. رابطه بین چاقی و CHD:



مقدار کلسترول خون با CHD رابطه علیتی دارد و کلسترول با چاقی هم رابطه دارد. آیا کلسترول در این رابطه نقش Confounder دارد؟

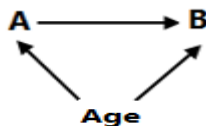
چون یکی از مسیرهای ایجاد بیماری CHD بالا بودن کلسترول است و لذا در این جا نقش واسطه ندارد و مخدوش کننده نیست. اگر در مطالعه ای متغیری نقش واسطه ای داشته باشد و اشتباها به عنوان مخدوش کننده در نظر گرفته شود چه اتفاقی می افتد؟ چون ما یکی از مسیرها ی علیتی بین A و B را در نظر نگرفتیم. بنابراین اثر متغیروابسته و مستقل ضعیف تر نشان داده می شود.

مثلا اگر در مثال بالا تنها کلسترول بالا در رابطه چاقی و CHD عامل علیتی باشد و آن را به عنوان مخدوش کننده بگیریم در این جا نیز رابطه به غلط نشان داده نمی شود. گاهی وقت ها رابطه بین Exposure و Confounder خیلی قوی است ، در ضمن انجام این مطالعه چه مشکلی ممکن است پیدا شود؟  
مثال:



در این جا اگر برای هر نوع محل سکونتی ارتباط را جدا آنالیز کنیم چون در بررسی مواجهه در بیماران یا همه مواجهه با آلودگی دارند و یا همه ندارند و دیگر این جا آلودگی متغیر نیست که بررسی شود. به ارتباط قوی که بین Exposure و Confounder وجود دارد و این مشکل ایجاد می شود. برای این که اثر متغیر مخدوش کننده را از بین ببریم یکی از راه ها این است که بیائیم برای متغیر مخدوش کننده را در Level های مختلف جداگانه آنالیز کنیم.

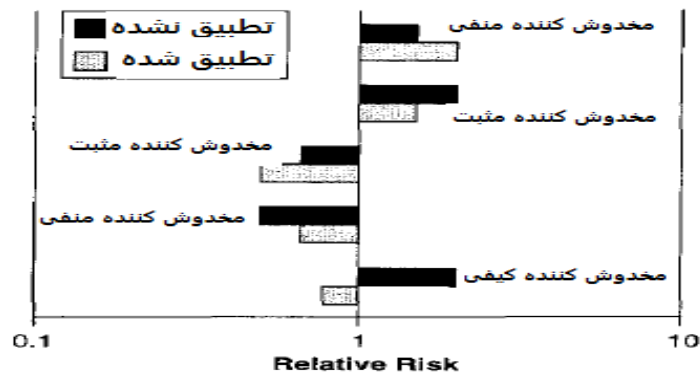
به عنوان مثال



که متغیر Age را در گروه های ریز ۰-۵ سال ، ۶-۱۰ سال، ۱۱-۱۵ سال، و بالای ۱۵ سال اگر بررسی کنیم ، به حذف مخدوش کننده کمک می کنند. ولی آیا همه افراد بالای ۱۵ سال در این جا یکسان است؟ خیر چون هنوز هم اثر سن به طور کامل در سنین بالای ۱۵ سال حذف نشده است و اصطلاحا به آن Confounder Residual می گویند.

انواع اثرات مخدوش کننده ها

نوع اثر	در هر مثال چه اتفاقی افتاده ؟	تطبیق شده	RR تطبیق نشده	ردیف
مثبت	قبل از تطبیق متغیر مخدوش کننده به عنوان عامل مثبت تلقی می شد که بعد از تطبیق اثرش کمتر شد	۱	۳/۵	-۱
مثبت	قبل از تطبیق متغیر مخدوش کننده به عنوان عامل مثبت تلقی می شد که بعد از تطبیق اثرش کمتر شد	۲/۱	۳/۵	-۲
مثبت	قبل از تطبیق اثر ۰/۳ (قوی تر) از ۰/۷ بود و متغیر مخدوش کننده اثر مثبت داشت.	۰/۷	۱/۳	-۳
منفی	متغیر مخدوش کننده قبل از تطبیق رابطه را ضعیف نشان می داد و اثر منفی داشت، بعد از تطبیق RR بیشتر شده است	۳/۲	۱/۵	-۴
کیفی	قبل از تطبیق ریسک فاکتور و بعد از تطبیق محافظت کننده شد. در این جا کیفیت رابطه تغییر کرده است.	۰/۷	۲	-۵



ارتباط مخدوش کننده با بیماری	ارتباط مخدوش کننده و مواجهه	نوع اثر مخدوش کننده	Diagram
مستقیم	مستقیم	مثبت	<pre>           A → B          /  \         /    \        C      CHD         \    /          ←   سیگار           قهوه             </pre>
معکوس	مستقیم	منفی	<pre>           سرطان کولون          /  \         /    \        ←     سطح اقتصادی-اجتماعی         \    /          سیگار             </pre>
معکوس	معکوس	مثبت	<pre>           سیگار → CHD          /  \         /    \        ←     ورزش             </pre>
مستقیم	معکوس	منفی	<pre>           چاقی → CHD          /  \         /    \        ←     ورزش             </pre>

**برای جلوگیری از مخدوش کنندگی چه باید کرد؟**

می توان با طراحی مناسب مطالعه احتمال وجود مخدوش کنندگی را کاهش داد. **تصادفی سازی**، (Randomization) مطمئن ترین راهی است که با آن می توان مخدوش کننده های شناخته شده و ناشناخته را در هنگام مقایسه دو روش درمانی حذف کرد. البته این روش برای مطالعات مشاهده ای کاربرد ندارد. بسیار اهمیت دارد که داده های تمامی افرادی که بطور تصادفی تقسیم شده اند، در پایان مورد آنالیز قرار گرفته و گزارش شوند (روش Intention to Treat). استفاده از داده های تمامی افرادی که تصادفی شده اند بسیار حائز اهمیت است. اگر افرادی باشند که به نوع درمان خود پایبند نبوده (مثلاً باید دارونما می خورده و دارو اصلی را مصرف کرده و برعکس) و آنالیز هم بر این اساس که چه کسی واقعا چه دارویی گرفته انجام شود (Per-Protocol)، دیگر تصادفی کردن کارآیی خود را نداشته و احتمال مخدوش شدن و سوگرایی یافته ها وجود دارد. در انواع مطالعات بخصوص مطالعه مورد-شاهدی، گروه هایی که باید با هم مقایسه شوند (مانند گروه مورد و شاهد)، اغلب از نظر عوامل مخدوش کننده بالقوه، با یکدیگر **جور** (Match) می شوند. باید به این نکته توجه داشت که **جورسازی** علی رغم افزایش **دقت** (Precision) مطالعه، به خودی خود مخدوش کنندگی را حذف نمی کند. بلکه باید با در نظر گرفتن فاکتور جورسازی شده در آنالیز داده ها، تاثیر آنرا حذف کرد. اگر این کار صورت نگیرد، نتایج مطالعه مخدوش خواهند بود. با محدود کردن گروه مورد مطالعه به تنها یک سطح از متغیر مخدوش کننده، می توان از مخدوش کنندگی جلوگیری کرد. وقتی با انتخاب افراد غیرسیگاری، اثر مخدوش کنندگی سیگار در مطالعه کنار گذاشته می شود به این روش کنترل اثر مخدوش کنندگی در زمان طراحی مطالعه، **محدودسازی** (Restriction) گفته می شود. این روش باعث کاهش تعمیم پذیری مطالعه می شود. علی رغم آنچه در مورد روشهای کنترل مخدوش کننده در زمان طراحی مطالعه گفته شد، معمولاً کنترل اثرات مخدوش کننده ها در زمان آنالیز داده ها نیز انجام می شود، اما به شرطی که اطلاعات مربوط به آن عوامل مخدوش کننده جمع آوری شده باشند. چندین روش برای انجام این کار وجود دارد. **روش لایه سازی** (Stratification) یکی از روشهایی است که به طور گسترده مورد استفاده قرار می گیرد که شامل روش استاندارد کردن (Standardization) و روش Mantel-Haenzel است. استفاده از روشهای پیچیده آماری برای کنترل اثرات مخدوش کنندگی چندین عامل، امروزه بسیار رایج شده است که تحت عنوان روش های رگرسیون چندگانه (Multiple Regression Technique) یا مدل سازی آماری شناخته می شوند. این روش ها در فصول قبل توضیح داده شدند.

**د) ارتباط و علیت (Association and Causation)**

وقتی دو متغیر A و B با هم اتفاق بیافتند، و این اتفاق فراتر از شانس باشد، می گوییم بین این دو متغیر ارتباط وجود دارد. درجه قدرت ارتباط بین این دو متغیر را با همبستگی (Correlation) نشان می دهیم، که می توان آن را با شاخص هایی چون ضریب همبستگی، نسبت شانس، خطر نسبی و ... نشان داد. اگر معلوم شود که یک مواجهه با یک بیماری همبستگی دارد، سوال بعدی این خواهد بود که "آیا این همبستگی علیتی است یا خیر؟" پس از مشخص شدن این ارتباط و همبستگی، اگر یکی از این متغیرها (A) مطلقاً به ایجاد متغیر دیگری (B) باشد، می گوییم بین این دو متغیر، احتمالاً رابطه علیتی (Causality) برقرار است. این تفسیر نیاز به تامل کافی، دقت و



احتیاط دارد. ترسیم نمودارهای علیتی و در نظر گرفتن نقش متغیرهای کولایدر، متغیرهای ابزاری و بررسی روابط بین متغیرها در مطالعات مشاهده ای کمک کننده است.

## انواع ارتباط بین متغیرها:

۱- رابطه جعلی Spurious Association

۲- رابطه علیتی مستقیم Direct Association

۳- رابطه علیتی غیر مستقیم Indirect Association

**رابطه جعلی:** این رابطه در نتیجه مقایسه نامناسب ایجاد می شود. رابطه جعلی بین مواجهه و بیماری، واقعی نیست. مثلاً مشاهده مرگ و میر بیشتر نوزادانی که در بیمارستان به دنیا می آیند نسبت به نوزادانی که در خانه به دنیا می آیند. که نتیجه گیری کنیم "تولد نوزاد در خانه برای سلامتی نوزاد مناسب تر است!" در صورتی که ممکن است در نتیجه یک مقایسه نامناسب ایجاد شده باشد؛ مانند: انجام زایمان های با خطر بیشتر در بیمارستان، و در نتیجه میزان مرگ و میر بیشتر نسبت به زایمان های انجام شده در خانه.

**رابطه علیتی مستقیم:** عامل بیماری به طور مستقیم و بدون واسطه باعث ایجاد بیماری می شود. مانند نقش پلی مورفیزم ژن F508 در ایجاد بیماری فیروز سیستیک.

**رابطه علیتی غیر مستقیم:** در این نوع رابطه عامل ایجاد بیماری، از طریق یک واسطه باعث ایجاد بیماری می شود. مانند نقش کلسترول بالا بر سگته قلبی؛ عوامل واسطی بین این دو، از جمله فاکتور های خونی و ضخامت سرخرگ ها جهت ایجاد سگته قلبی لازم است.

**انواع رابطه علیتی:** وقتی یک رابطه علیتی باشد، این رابطه علیتی به یکی از چهار طریق زیر عمل می کند:

- ۱- **رابطه علیتی لازم و کافی است:** وجود عامل، برای ایجاد بیماری هم لازم است و هم کفایت می کند. مانند مسمومیت با بوتولیسم در مرگ ناشی از بوتولیسم؛ یا توالی خاص ژنتیکی در بیماریهایی با منشا صرفاً ژنتیکی
- ۲- **رابطه علیتی لازم است ولی کافی نیست:** حضور عامل برای ایجاد بیماری ضروری است ولی کفایت ندارد و علاوه بر آن، عوامل دیگری نیز بایستی حضور داشته باشند. مثلاً، علاوه بر باسیل سل، در ایجاد بیماری توبرکلوزیس، تغذیه نامناسب و ضعف سیستم ایمنی نیز جهت ایجاد بیماری، لازم است.
- ۳- **رابطه علیتی لازم نیست ولی کافی است:** عامل به تنهایی قادر به ایجاد بیماری است ولی آن بیماری را عوامل دیگری نیز به تنهایی می توانند ایجاد کنند. مانند مواجهه با تشعشعات رادیواکتیو و یا بنزن در ایجاد لوسمی
- ۴- **رابطه علیتی نه لازم و نه کافی است:** در این حالت، وجود یک عامل به تنهایی برای ایجاد بیماری، نه لازم است و نه کافی. این حالت رابطه علیتی، بسیار پیچیده است. بیشتر بیماری های مزمن دارای چنین مدل رابطه ی علیتی هستند.

### معیارهایی جهت شناسایی علیتی بودن یک رابطه واقعی :

- ۱- **رابطه زمانی یا تقدم زمانی:** بدیهی است عامل مظنون، بایستی قبل از ایجاد بیماری در بدن حضور داشته باشد، تا رابطه علیتی برقرار شود. رابطه زمانی مهمترین معیار علیتی است.
- ۲- **قدرت رابطه:** که با اندازه هایی مانند نسبت شانس، خطر نسبی و سایر سنجه های رابطه مشخص می شود.
- ۳- **توجیه پذیری بیولوژیکی:** در سطح زیست شناختی، مطالعات بیولوژیک، چه در حال و چه در آینده، بایستی تایید کننده این همبستگی باشد. بعد از تقدم زمانی مهمترین معیار است.
- ۴- **تکرارپذیری یافته ها و همخوانی با مطالعات دیگر:** اگر رابطه علیتی باشد، انتظار خواهیم داشت که در مطالعات مشابه با جمعیت های متفاوت نیز، نتیجه یکسانی به دست آید.
- ۵- **رابطه مقدار-پاسخ:** انتظار می رود در صورتیکه رابطه علیتی بین یک عامل و بیماری وجود دارد، رابطه ای نیز بین مقدار مواجهه بدن با عامل خطر، و افزایش احتمال ابتلا به بیماری وجود داشته باشد. مانند افزایش تعداد سیگار مصرفی در روز و افزایش خطر ابتلا به سرطان ریه.
- ۶- **شواهد تجربی، شباهت، پیوستگی و عدم تناقض رابطه مشاهده شده با سیر طبیعی و بیولوژی بیماری و هم راستا بودن با سایر مطالعات و اختصاصی بودن رابطه از معیارهای دیگر هستند.** وجود یا عدم وجود معیارهای مذکور که به نام معیارهای دال و هیل معروف هستند به طور قطعی ارتباط علیتی را تایید و یا رد نمی کنند.

### برهمکنش (اینترکشن):

واژه **اثر متقابل** (یا Effect Modification) در اپیدمیولوژی به شرایطی اطلاق می شود که دو یا چند متغیر اثر یکدیگر بر روی پیامد را تحت تاثیر قرار دهند. باید توجه داشت که این پدیده با اثر مخدوش کنندگی متفاوت است. برای متغیرهای دو حالتی، اثر متقابل به این معناست که اثر مواجهه بر روی پیامد با توجه به وجود یا عدم وجود متغیر سوم (Effect Modifier) متفاوت است. در صورتیکه که وجود این متغیر باعث افزایش قدرت رابطه شود، اثر متغیر فوق Synergistic است و اثر متقابل مثبت وجود دارد. ولی در صورتیکه وجود متغیر سوم باعث کاهش قدرت رابطه شود، اثر متغیر Antagonistic است و اثر متقابل موجود از نوع منفی است. به همین صورت در مورد متغیرهای پیوسته وجود اثر متقابل به این معناست که اثر مواجهه بر روی پیامد وابسته به سطوح مختلف متغیر سوم متفاوت است. باید توجه داشت که برای بررسی اثر متقابل، وجود حداقل سه متغیر الزامی است: مواجهه (E)، پیامد (O) و متغیر سوم (Z) که Effect Modifier بودن آن مورد بررسی قرار می گیرد. اثر متقابل کاملاً دوطرفه (Reciprocal) است، به این معنی که اگر فاکتور Z رابطه بین E و O را تحت تاثیر قرار می دهد، فاکتور E نیز اثری مشابه بر رابطه Z و O دارد.

### وجود اثر متقابل را به دو طریق می توان ارزیابی کرد:

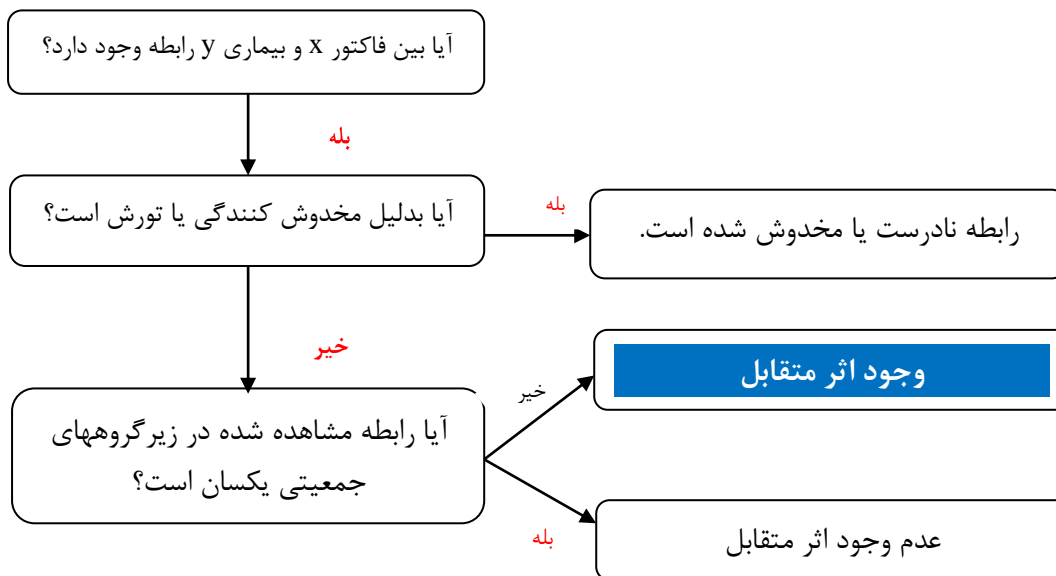
- از طریق بررسی هموزن یا هتروژن بودن اثرات: اثر متقابل زمانی وجود دارد که اثر مواجهه بر پیامد در لایه های مختلف متغیر سوم هموزن نیست.

- از طریق مقایسه بین اثرات توأم قابل انتظار و مشاهده شده مواجهه (E) و متغیر سوم (Z): اثر متقابل زمانی حضور دارد که اثر توأم مشاهده شده متفاوت از اثر مورد انتظار با فرض استقلال دو متغیر (مواجهه (E) و متغیر سوم (Z)) است.

با توجه به معیاری که برای بررسی اثر یک رابطه استفاده می شود، وجود دو نوع اثر متقابل قابل بررسی است، اثر متقابل جمع (Additive Interaction) و اثر متقابل ضربی (Multiplicative Interaction).

### بررسی وجود اثر متقابل از طریق بررسی هتروژن بودن اثرات:

بدنبال مشاهده وجود یک رابطه آماری و اطمینان از اینکه رابطه موجود بدلیل Confounding، Bias و شانس نیست، سوال کلیدی در یافتن اثر متقابل این است که آیا رابطه شدت رابطه مشاهده شده در سطوح مختلف متغیر سوم یکسان است؟ مثبت بودن پاسخ این سوال، بیانگر عدم وجود اثر متقابل است. به عنوان مثال، در بررسی ارتباط دیابت با بیماری عروق کرونر، دیابت ریسک فاکتور مهم تری برای بیماری عروق کرونر در مردان نسبت به زنان است. یعنی جنسیت در این رابطه به عنوان Effect Modifier عمل می کند. به عبارتی، اثر متقابل وجود دارد. برای بررسی اثر مخدوش کنندگی و اثر متقابل، پیشنهاد می شود مسیر زیر دنبال گردد:



### بررسی وجود اثر متقابل جمع:

اثر متقابل جمع زمانی وجود دارد که خطر قابل انتساب (Attributable Risk - AR) در گروه مواجهه مثبت (E+) با توجه به وجود یا عدم وجود متغیر سوم (Z) متفاوت است. برای ارزیابی وجود اثر متقابل جمع، AR برای گروه مواجهه مثبت را در سطوح مختلف متغیر سوم محاسبه کرده، تفاوت AR در سطوح مختلف متغیر سوم به معنای وجود اثر متقابل جمع است.

به مثالهای زیر توجه کنید:

A

خطر قابل انتساب (در ۱۰۰۰ نفر)	میزان بروز (در ۱۰۰۰ نفر)	مواجهه (E)	متغیر سوم (Z)
۰	۱۰	بلی	-
۱۰	۲۰	خیر	
۰	۳۰	بلی	+
۱۰	۴۰	خیر	

B

خطر قابل انتساب (در ۱۰۰۰ نفر)	بروز	میزان (در ۱۰۰۰ نفر)	مواجهه (E)	متغیر سوم (Z)
۰		۵	خیر	-
۵		۱۰	بلی	
۰		۱۰	خیر	+
۲۰		۳۰	بلی	

همانطور که در جدول A دیده می شود، AR در سطوح مختلف متغیر سوم یکسان است که بیان کننده عدم وجود اثر متقابل جمعی است. در صورتیکه در جدول B، AR در حضور متغیر سوم (Z+) بیشتر از عدم حضور آن (Z-) است. این یافته، بیانگر هتروژن بودن اثر مواجهه (AR) با توجه به متغیر سوم یعنی وجود اثر متقابل جمعی است.

بررسی وجود اثر متقابل ضربی:

اثر متقابل ضربی زمانی وجود دارد که تفاوت خطر نسبی یک پیامد بین گروههای مواجهه مثبت و منفی، با توجه به وجود یا عدم وجود متغیر سوم متفاوت باشد. برای ارزیابی وجود اثر متقابل ضربی، RR برای گروه مواجهه مثبت را در سطوح مختلف متغیر سوم محاسبه کرده، متفاوت بودن RR در سطوح مختلف متغیر سوم به معنای وجود اثر متقابل ضربی است. به مثالهای زیر توجه کنید:

A

خطر نسبی (RR) (در ۱۰۰۰ نفر)	بروز	میزان (در ۱۰۰۰ نفر)	مواجهه (E)	متغیر سوم (Z)
۱		۱۰	خیر	-
۲		۲۰	بلی	
۱		۲۵	خیر	+
۲		۵۰	بلی	

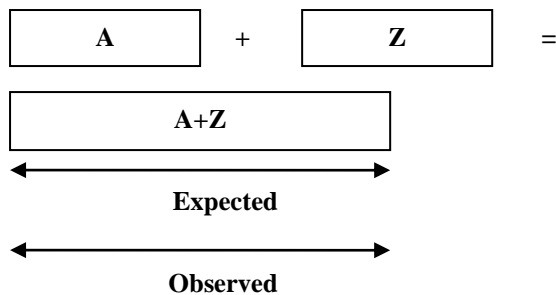
متغیر سوم (Z)	مواجهه (E)	میزان (در ۱۰۰۰ نفر)	بروز خطر نسبی (RR) (در ۱۰۰۰ نفر)
-	خیر	۵	۱
-	بلی	۱۰	۲
+	خیر	۲۵	۱
+	بلی	۱۲۵	۵

همانطور که در جدول A دیده می شود، RR در سطوح مختلف متغیر سوم یکسان است و این به معنی عدم وجود اثر متقابل ضربی است. در صورتیکه در جدول B، RR در گروه مواجهه یافته با Z، بیشتر از گروه مواجهه نیافته است، این یافته بیانگر هتروژن بودن اثر مواجهه (RR) با توجه به متغیر سوم است. به عبارتی، در این حالت اثر متقابل ضربی وجود دارد.

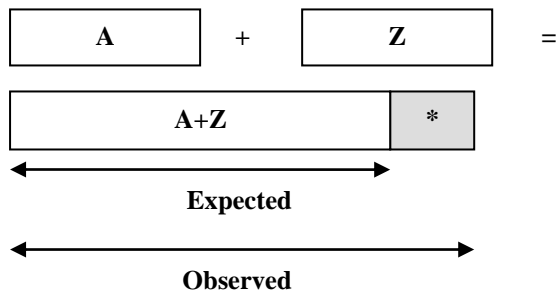
**مقایسه اثرات توأم قابل انتظار و مشاهده شده:** روش دیگر بررسی وجود اثر متقابل مقایسه اثرات توأم مشاهده شده با اثرات قابل انتظار است. اثر متقابل زمانی وجود دارد که این دو مقدار با هم متفاوت باشند. مقادیر قابل انتظار با فرض مستقل بودن دو متغیر مواجهه و متغیر سوم بدست می آیند. همانند روش قبل بررسی وجود اثر متقابل در هر دو مدل (جمععی یا ضربی) بر یک اساس صورت می گیرد و تنها تفاوت دو مدل، شاخص مورد استفاده یعنی AR یا RR است.

**شکل شماتیک برای نشان دادن اثر متقابل براساس مقایسه اثر مشاهده شده با مقدار قابل انتظار**

A : وقتی که اثر متقابل وجود نداشته باشد، اثر مشاهده شده با مقدار قابل انتظار برابر است:

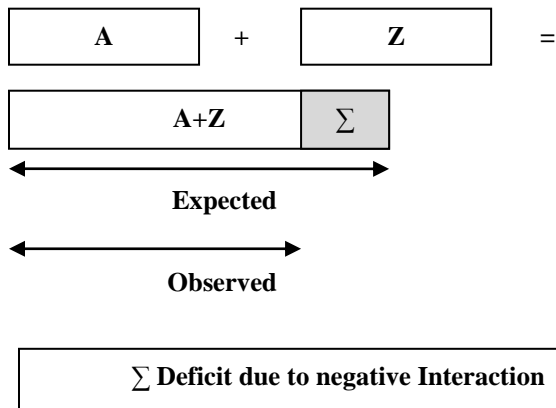


B : وقتی که اثر متقابل مثبت وجود داشته باشد، اثر مشاهده شده بیشتر از مقدار قابل انتظار است:



\* Excess due to positive Interaction

C : وقتی که اثر متقابل منفی وجود داشته باشد، اثر مشاهده شده کمتر از مقدار قابل انتظار است:



**بررسی اثر متقابل جمعی:** در این روش مقدار قابل انتظار اثر توام مواجهه و متغیر سوم از جمع جبری اثرات مستقل (در این مدل خطر قابل انتساب) برای مواجهه مثبت بدست می آید. متفاوت بودن دو مقدار مشاهده شده و قابل انتظار بیانگر وجود اثر متقابل جمعی است.

A

خطر قابل انتساب مشاهده شده (در ۱۰۰۰ نفر)		میزان بروز مشاهده شده (در ۱۰۰۰ نفر)		متغیر سوم (Z)
مواجهه (E)		مواجهه (E)		
+	-	+	-	
۲۰	۰	۳۰	۱۰	-
۳۰	۱۰	۴۰	۲۰	+

مقدار قابل انتظار AR :  $30 = 10 + 20$

مقدار مشاهده شده AR : ۳۰

B

خطر قابل انتساب مشاهده شده (در ۱۰۰۰ نفر)		میزان بروز مشاهده شده (در ۱۰۰۰ نفر)		متغیر سوم (Z)
مواجهه (E)		مواجهه (E)		
+	-	+	-	
۲۰	۰	۳۰	۱۰	-
۵۰	۱۰	۶۰	۲۰	+

مقدار قابل انتظار AR :  $30 = 10 + 20$

مقدار مشاهده شده AR : ۵۰

قدم اول محاسبه میزان بروز در هر گروه است. گروهی که با هیچ یک از فاکتورها مواجهه نداشته اند (E-Z-) را مرجع قرار داده و AR مشاهده شده را برای سه گروه دیگر محاسبه می کنیم. مقدار منتظره نیز از جمع جبری ARهای مستقل دو فاکتور بدست می آید:

$$\text{Expected AR}_{E+Z+} = \text{OBS AR}_{E+Z-} + \text{OBS AR}_{E-Z+}$$

همانطور که در جدول A دیده می شود، AR قابل انتظار (۲۰+۱۰=۳۰) برابر AR مشاهده شده است که نشان دهنده عدم وجود اثر متقابل جمعی است. اما این دو مقدار در جدول B برابر نیستند. مقدار مشاهده شده بیشتر از مقدار منتظره است، این موضوع بیانگر اثر متقابل جمعی مثبت است.

### بررسی اثر متقابل ضربی

محاسبه اثر متقابل ضربی مشابه روش اثر متقابل جمعی است با این تفاوت که در بررسی اثر متقابل ضربی، اثر توأم از ضرب دو خطر نسبی E و Z بدست می آید.

$$\text{Expected RR}_{E+Z+} = \text{OBS RR}_{E+Z-} \times \text{OBS RR}_{E-Z+}$$

به مثالهای زیر توجه کنید:

A

خطر نسبی مشاهده شده (RR) (در ۱۰۰۰ نفر)		میزان بروز مشاهده شده (در ۱۰۰۰ نفر)		متغیر سوم (Z)
مواجهه (E)		مواجهه (E)		
+	-	+	-	
۳	۱	۳۰	۱۰	-
۶	۲	۶۰	۲۰	+

مقدار قابل انتظار RR :  $۶=۲ \times ۳$  و مقدار مشاهده شده RR : ۶

B

خطر نسبی مشاهده شده (RR) (در ۱۰۰۰ نفر)		میزان بروز مشاهده شده (در ۱۰۰۰ نفر)		متغیر سوم (Z)
مواجهه (E)		مواجهه (E)		
+	-	+	-	
۳	۱	۳۰	۱۰	-
۹	۲	۹۰	۲۰	+

مقدار قابل انتظار RR :  $۶=۲ \times ۳$  و مقدار مشاهده شده RR : ۹

همانند مثال قبل، در قسمت A مقادیر قابل انتظار و مشاهده شده برابرند که بیان کننده عدم وجود اثر متقابل ضربی است. در صورتیکه مقدار RR مشاهده شده در قسمت B بیش از مقدار قابل انتظار و نشان دهنده اثر متقابل ضربی مثبت است.

### اثر متقابل کمی (Quantitative) و کیفی (Qualitative) :

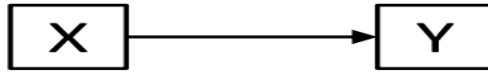
در مواردی که بین مواجهه با فاکتور E و پیامد O رابطه وجود دارد، ولی قدرت این رابطه در سطوح مختلف متغیر سوم Z متفاوت است، اثر متقابل کمی وجود دارد. با تغییر مقیاس متغیر، می توان اثر متقابل کمی را برداشت. زمانی که مواجهه با فاکتور E و پیامد O در سطوح مختلف متغیر سوم Z، رابطه برعکس دارند یا زمانی که در یک سطح متغیر سوم (Z) بین مواجهه E و پیامد ارتباط وجود دارد و در سطح دیگر متغیر سوم، ارتباطی بین E و O وجود ندارد، اثر متقابل کیفی وجود دارد.

لازم به ذکر است که در حضور اثر متقابل کیفی، اثر متقابل در هر دو مدل ضربی و جمعی وجود دارد. این نوع اثر متقابل، مستقل از مقیاس اندازه گیری است و با تغییر مقیاس مرتفع نمی شود. آنچه در بالا به آن اشاره شده، استفاده از روش طبقه بندی برای بررسی اثر متقابل و مخدوش کنندگی برای تنها یک متغیر سوم (Z) بود. در حالتی که تعداد متغیرها عوامل خارجی زیاد باشد و یا سطوح مختلفی برای یک متغیر خارجی وجود داشته باشد، استفاده از روش طبقه بندی مشکل و بعضاً غیر قابل استفاده خواهد بود. روشهای جایگزینی برای روش طبقه بندی مانند روش لوجیستیک چندگانه وجود دارد، که در مبحث مدل‌های رگرسیونی آمده است.

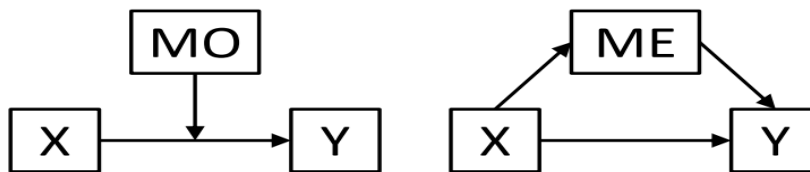


متغیرهای تعدیل کننده (Moderation) و میانجی (Mediation) در نرم افزار stata

ساده ترین رابطه بین دو متغیر به صورت زیر است:

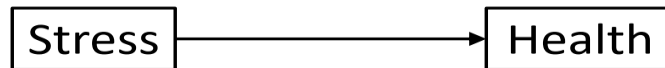


در رابطه بین دو متغیر، وقتی متغیر سومی حضور داشته باشد رابطه مذکور با Moderators و Mediators قابل توجیه است. متغیر سومی که رابطه بین دو متغیر X و Y را تعدیل نماید، متغیر تعدیل کننده (MO) نام دارد. متغیر سومی که رابطه بین X و Y را توضیح دهد متغیر میانجی (ME) نام دارد. در شکل زیر این دو رویکرد نمایش داده شد.



متغیر تعدیل کننده

مثال در شکل زیر رابطه دو متغیر استرس با سلامتی ترسیم شد.



آیا رابطه بالا در حضور متغیرهای مختلف و موارد زیر یکسان است:

همه گروهها/مکانها/زمانها؟

همه انواع استرس؟

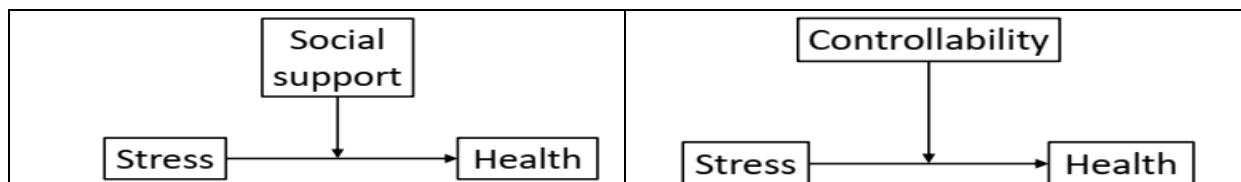
آیا عوامل حفاظتی و پیشگیری کننده وجود دارد؟

آیا عوامل خطر وجود دارد؟

پاسخ این است که می توان گفت حداقل دو متغیر مهم حمایت اجتماعی (Social support) و کنترل پذیری

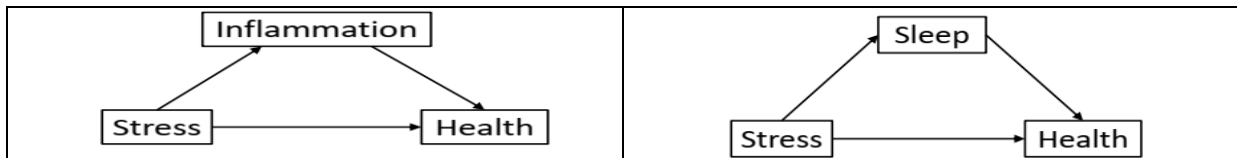
(Controllability) در این رابطه موثر بوده و باید نقش آنها را در رابطه بین استرس و سلامتی که در شکل زیر

آمده است در نظر گرفت.



### متغیر میانجی

چرا استرس بر سلامتی تأثیر می گذارد؟  
 مکانیسم علی ارتباط استرس و سلامتی چیست؟  
 توضیحات احتمالی چیست؟ نقش خواب و التهاب در رابطه مذکور چیست؟



با توضیحات بالا، می توان گفت متغیر تعدیل کننده یا Moderator، متغیر سومی است که رابطه بین دو متغیر X و Y را تغییر می دهد. متغیر میانجی یا Mediators، متغیر سومی است که رابطه بین دو متغیر X و Y را توضیح (explains) و توجیه می کند.

### کار با نرم افزار و بررسی متغیرهای ME و MO

فایل nhanes را باز کنید. با دستور `compact, codebook` توصیف متغیرها را ببینید.

Variable	Obs	Unique	Mean	Min	Max	Label
sex	10351	2	.4748333	0	1	Sex (0=Female; 1=Male)
race	10351	3	1.143561	1	3	Race (1=White; 2=Black; 3=Hispanic)
age	10351	55	47.57965	20	74	Age (years)
heartatk	10349	2	.0574935	0	1	Heart attack (1=Yes)
weight	10351	745	71.89752	30.84	175.88	Weight (kg)
sbp	10351	108	130.8817	65	300	Systolic blood pressure (mm/Hg)
diast	10351	68	81.715	35	150	Diastolic blood pressure (mm/Hg)
cholesterol	10351	316	217.6697	80	828	Serum cholesterol (mg/dL)
triglycerids	5050	423	143.8958	16	2238	Serum triglycerides (mg/dL)

در فایل مذکور دیاگرام رابطه بین سن و فشارخون سیستول و معادله رگرسیون خطی آن را در زیر ببینید.



نتیجه مدل رگرسیونی در زیر آمده است.

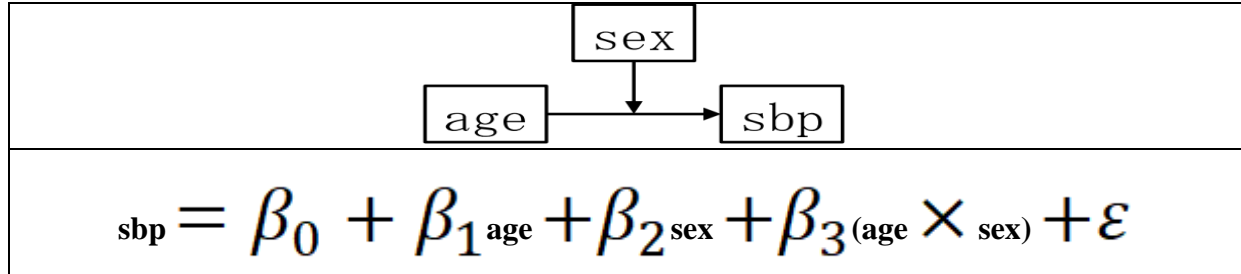
`. regress sbp age`

Source	SS	df	MS	Number of obs	=	10,351
Model	1304200.02	1	1304200.02	F(1, 10349)	=	3116.79
Residual	4330470.01	10,349	418.443328	Prob > F	=	0.0000
Total	5634670.03	10,350	544.412563	R-squared	=	0.2315
				Adj R-squared	=	0.2314
				Root MSE	=	20.456

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]
age	.6520775	.0116801	55.83	0.000	.6291823 .6749727
_cons	99.85603	.5909867	168.96	0.000	98.69758 101.0145

## روش های اپیدمیولوژی و پژوهش با Stata

جنسیت به عنوان یک متغیر تعدیل کننده و تغییر دهنده در رابطه سن با فشارخون سیستمولیک را ارزیابی کنید. معادله رگرسیونی مربوطه را بنویسید.



متغیر جنسیت یک متغیر کیفی است. با اضافه کردن `i.sex` در دستور زیر معادله رگرسیونی رابطه بین سن و فشارخون برای مردان و در نظر گرفتن زیان به عنوان گروه مرجع `regress sbp age i.sex c.age#sex` به شرح زیر است.

```
. regress sbp age i.sex
```

Source	SS	df	MS	Number of obs	=	10,351
Model	1345777.75	2	672888.876	F(2, 10348)	=	1623.51
Residual	4288892.27	10,348	414.465817	Prob > F	=	0.0000
				R-squared	=	0.2388
				Adj R-squared	=	0.2387
Total	5634670.03	10,350	544.412563	Root MSE	=	20.358

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]
age	.6530798	.0116249	56.18	0.000	.6302929 .6758668
sex					
Male	4.013619	.4007284	10.02	0.000	3.228114 4.799124
_cons	97.90254	.6196663	157.99	0.000	96.68788 99.11721

در معادله رگرسیونی بالا اینترکشن سن و جنسیت را وارد کنید. نتیجه عبارت است از:

```
. regress sbp c.age i.sex c.age#sex
```

Source	SS	df	MS	Number of obs	=	10,351
Model	1437147	3	479049	F(3, 10347)	=	1180.87
Residual	4197523.03	10,347	405.675367	Prob > F	=	0.0000
				R-squared	=	0.2551
				Adj R-squared	=	0.2548
Total	5634670.03	10,350	544.412563	Root MSE	=	20.141

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]
age	.8163546	.0158314	51.57	0.000	.7853219 .8473873
sex					
Male	20.45813	1.165263	17.56	0.000	18.17398 22.74227
sex#c.age					
Male	-.3457346	.0230373	-15.01	0.000	-.3908923 -.300577
_cons	90.11098	.8033587	112.17	0.000	88.53624 91.68571

## روش های اپیدمیولوژی و پژوهش با Stata

در دستور معادله بالا، می توان ## را در دستور regress sbp c.age##sex اضافه و اثر اصلی و اینترکشن دو متغیر را همزمان محاسبه نمود.

```
. regress sbp c.age##sex
```

Source	SS	df	MS	Number of obs	=	10,351
Model	1437147	3	479049	F(3, 10347)	=	1180.87
Residual	4197523.03	10,347	405.675367	Prob > F	=	0.0000
				R-squared	=	0.2551
				Adj R-squared	=	0.2548
Total	5634670.03	10,350	544.412563	Root MSE	=	20.141

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]
age	.8163546	.0158314	51.57	0.000	.7853219 .8473873
sex					
Male	20.45813	1.165263	17.56	0.000	18.17398 22.74227
sex#c.age					
Male	-.3457346	.0230373	-15.01	0.000	-.3908923 -.300577
_cons	90.11098	.8033587	112.17	0.000	88.53624 91.68571

در نرم افزار می توان نمودار رابطه سن و فشارخون برای مردان و زنان را در بر حسب گروه های سنی با دستورات زیر محاسبه و گزارش نمود. شکل زیر را ببینید. marginsplot و margins sex, at (age=(20(10)60)) vsquish

```
. margins sex, at (age=(20(10)60)) vsquish
```

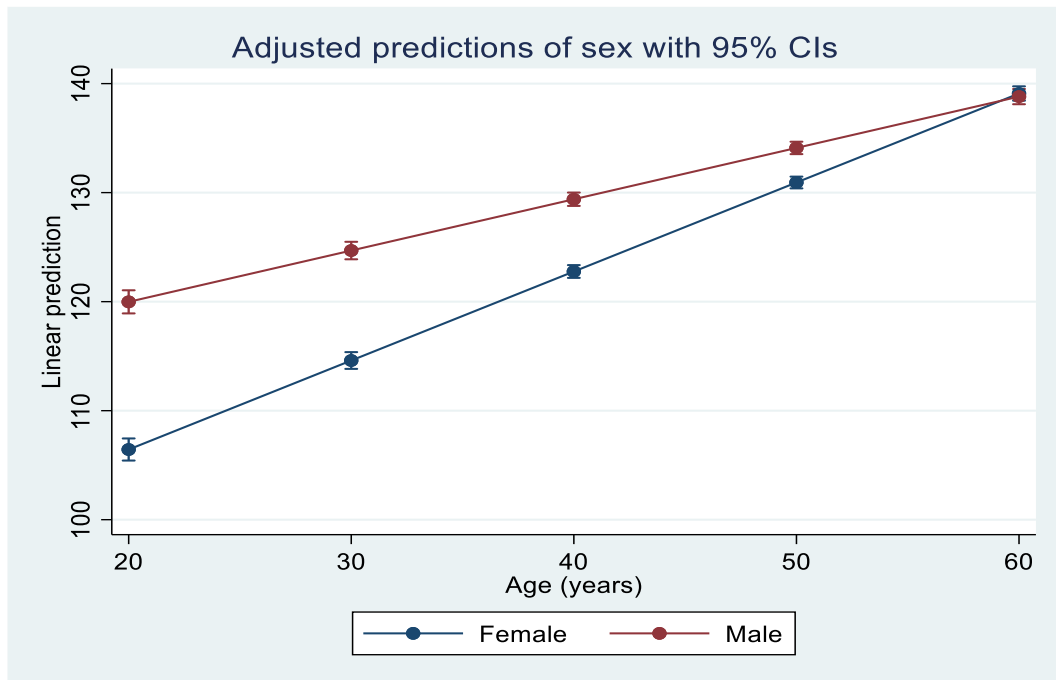
Adjusted predictions  
Model VCE: OLS

Number of obs = 10,351

Expression: Linear prediction, predict()

- 1.\_at: age = 20
- 2.\_at: age = 30
- 3.\_at: age = 40
- 4.\_at: age = 50
- 5.\_at: age = 60

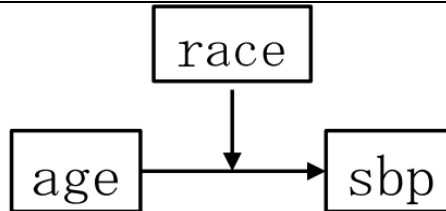
	Delta-method			P> t	[95% conf. interval]	
	Margin	std. err.	t			
_at#sex						
1#Female	106.4381	.5169356	205.90	0.000	105.4248	107.4514
1#Male	119.9815	.5414609	221.59	0.000	118.9201	121.0429
2#Female	114.6016	.3915755	292.67	0.000	113.8341	115.3692
2#Male	124.6877	.4093516	304.60	0.000	123.8853	125.4901
3#Female	122.7652	.2992777	410.20	0.000	122.1785	123.3518
3#Male	129.3939	.3130088	413.39	0.000	128.7803	130.0075
4#Female	130.9287	.2755536	475.15	0.000	130.3886	131.4688
4#Male	134.1001	.2905118	461.60	0.000	133.5306	134.6696
5#Female	139.0923	.3352899	414.84	0.000	138.435	139.7495
5#Male	138.8063	.3561412	389.75	0.000	138.1082	139.5044



تا کنون نحوه آزمایش تعدیل کننده متغیر دو حالت (باینری) جنسیت را با مثال نشان دادیم. این ساده ترین مورد است. در مورد تعدیل کننده دسته ای یا کیفی با ۳ سطح و متغیر پیوسته در ادامه توضیح داده می شود.

متغیر تعدیل کننده دسته بندی شده (کیفی) چند حالتی و کار با نرم افزار

در شکل زیر در رابطه بین سن و فشارخون سیستولیک نقش متغیر نژاد به عنوان یک متغیر کیفی سه حالتی و تعدیل کننده، رابطه را ارزیابی و معادله را با اینترکشن سن و نژاد بنویسید.



$$sbp = \beta_0 + \beta_1 age + \beta_2 race1 + \beta_3 race2 + \beta_4 (age \times race1) + \beta_5 (age \times race2) + \varepsilon$$

با دستور `regress sbp age c.age## race` معادله بالا حل و اجرا می شود.

```
. regress sbp age c.age## race
note: age omitted because of collinearity.
```

Source	SS	df	MS	Number of obs	=	10,351
Model	1333586.56	5	266717.312	F(5, 10345)	=	641.51
Residual	4301083.47	10,345	415.764472	Prob > F	=	0.0000
				R-squared	=	0.2367
				Adj R-squared	=	0.2363
Total	5634670.03	10,350	544.412563	Root MSE	=	20.39

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.6344876	.0124758	50.86	0.000	.6100325	.6589426
age	0	(omitted)				
race						
Black	-1.638109	1.85557	-0.88	0.377	-5.275385	1.999166
Hispanic	-13.2599	4.001067	-3.31	0.001	-21.10276	-5.417033
race#c.age						
Black	.1313702	.0376089	3.49	0.000	.0576495	.2050908
Hispanic	.3134428	.0843473	3.72	0.000	.1481058	.4787799
_cons	100.2206	.6342408	158.02	0.000	98.97741	101.4639

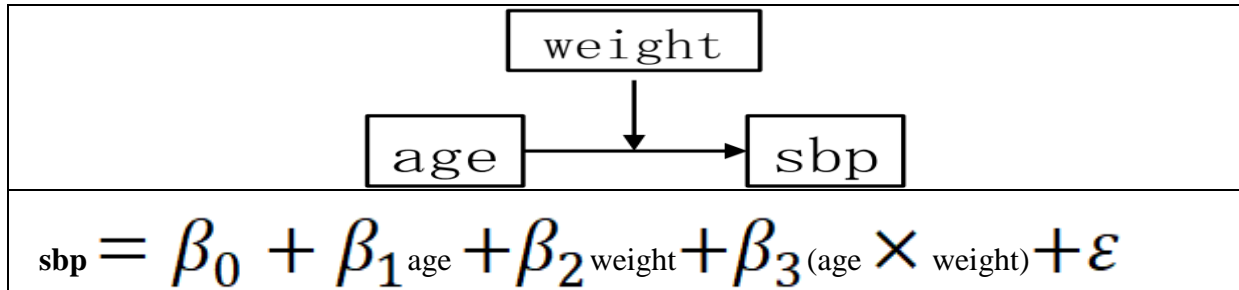
با دستور `testparm c.age# race` می توان اینترکشن را بین نژاد و سن را ارزیابی نمود. نتیجه نشان می دهد که اینترکشن وجود دارد.

- ( 1) 2.race#c.age = 0
- ( 2) 3.race#c.age = 0

```
F( 2, 10345) = 12.40
Prob > F = 0.0000
```

متغیر تعدیل کننده کمی پیوسته

در شکل زیر در رابطه بین سن و فشارخون سیستمولیک نقش متغیر وزن به عنوان یک متغیر کمی پیوسته و تعدیل کننده، رابطه را ارزیابی و معادله را با اینترکشن سن و وزن بنویسید.



معادله بالا را با دستور regress sbp c.age## c.weight می توان اجرا نمود:

```
. regress sbp c.age## c.weight
```

Source	SS	df	MS	Number of obs	=	10,351
Model	1716436.54	3	572145.514	F(3, 10347)	=	1510.88
Residual	3918233.49	10,347	378.683047	Prob > F	=	0.0000
				R-squared	=	0.3046
				Adj R-squared	=	0.3044
Total	5634670.03	10,350	544.412563	Root MSE	=	19.46

sbp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.8898576	.0536198	16.60	0.000	.7847525	.9949627
weight	.5733109	.0368295	15.57	0.000	.5011179	.6455039
c.age#c.weight	-.003581	.0007458	-4.80	0.000	-.0050429	-.0021191
_cons	59.60983	2.64211	22.56	0.000	54.43079	64.78888

```
. sum age weight
```

Variable	Obs	Mean	Std. dev.	Min	Max
age	10,351	47.57965	17.21483	20	74
weight	10,351	71.89752	15.35642	30.84	175.88

-1SD: 71-15=56 , Mean: 71, +1SD: 71+15=86

## روش های اپیدمیولوژی و پژوهش با Stata

با دستور ( age=(20(10)60) weight=86 ) at (age=(20(10)60) weight=71) at (age=(20(10)60) weight=56 ) quietly margins, at

مقادیر پیش بینی فشارخون بر حسب گروه های سنی و وزن در زیر آمده است.

Adjusted predictions  
Model VCE: OLS

Number of obs = 10,351

Expression: Linear prediction, predict()

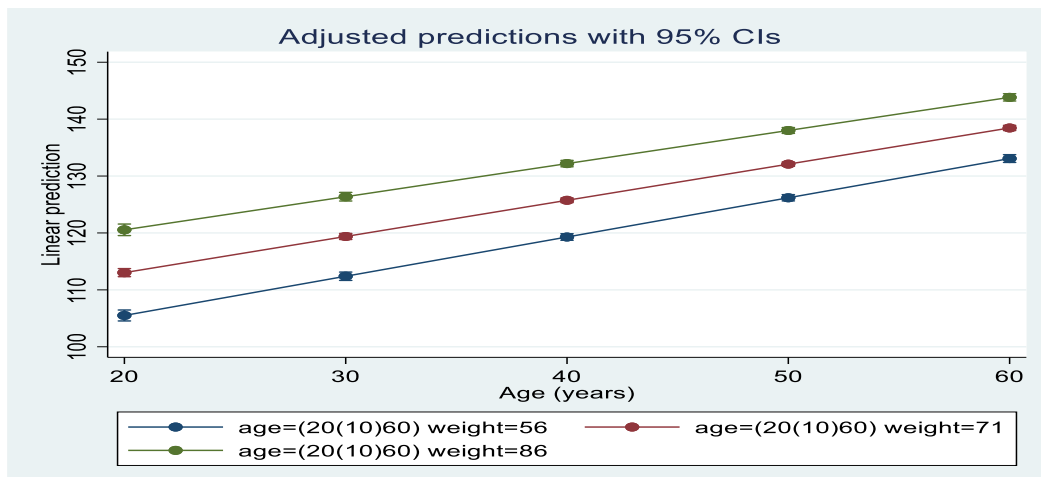
```

1._at: age = 20
      weight = 56
2._at: age = 30
      weight = 56
3._at: age = 40
      weight = 56
4._at: age = 50
      weight = 56
5._at: age = 60
      weight = 56
6._at: age = 20
      weight = 71
7._at: age = 30
      weight = 71
8._at: age = 40
      weight = 71
9._at: age = 50
      weight = 71
10._at: age = 60
      weight = 71
11._at: age = 20
      weight = 86
12._at: age = 30
      weight = 86
13._at: age = 40
      weight = 86
14._at: age = 50
      weight = 86
15._at: age = 60
      weight = 86
    
```

_at	Delta-method				
	Margin	std. err.	t	P> t	[95% conf. interval]
1	105.5016	.4901742	215.23	0.000	104.5408 106.4625
2	112.3948	.3728086	301.48	0.000	111.6641 113.1256
3	119.288	.2920397	408.47	0.000	118.7156 119.8605
4	126.1812	.281378	448.44	0.000	125.6297 126.7328
5	133.0744	.3473404	383.12	0.000	132.3936 133.7553
6	113.027	.3619141	312.30	0.000	112.3176 113.7364
7	119.383	.2739441	435.79	0.000	118.8461 119.92
8	125.7391	.2094939	600.20	0.000	125.3284 126.1497
9	132.0951	.1936719	682.06	0.000	131.7155 132.4748
10	138.4512	.2364503	585.54	0.000	137.9877 138.9147
11	120.5524	.5166914	233.32	0.000	119.5395 121.5652
12	126.3712	.3858926	327.48	0.000	125.6148 127.1277
13	132.1901	.2878989	459.15	0.000	131.6258 132.7545
14	138.009	.2624654	525.82	0.000	137.4946 138.5235
15	143.8279	.3269862	439.86	0.000	143.187 144.4689

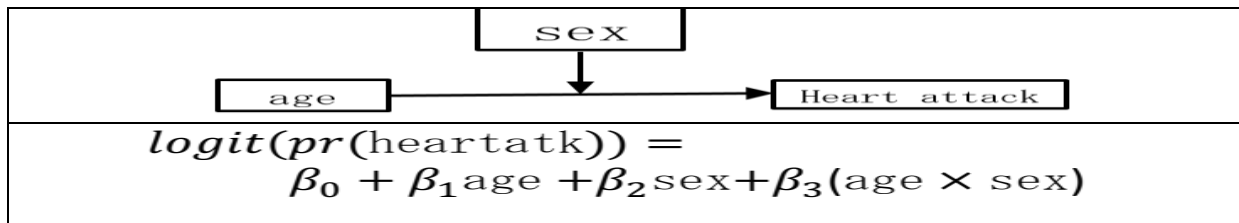


نمودار پیش بینی روند فشارخون بر حسب گروه های سنی با دستور `marginsplot` به صورت زیر است.



پیامد دوحالته و بررسی نقش متغیر تعدیل کننده

در شکل زیر در رابطه بین سن و حمله قلبی به عنوان پیامد (متغیر کیفی دوحالته) و جنسیت به عنوان متغیر تعدیل کننده، رابطه را ارزیابی و معادله را با اینترکشن سن و جنسیت بنویسید.



با دستور `logit heartatk c.age## sex,or` مدل رگرسیون لجستیک اجرا می شود:

```
Iteration 0: log likelihood = -2276.928
Iteration 1: log likelihood = -2064.07
Iteration 2: log likelihood = -1836.4926
Iteration 3: log likelihood = -1829.8036
Iteration 4: log likelihood = -1829.7339
Iteration 5: log likelihood = -1829.7338
```

Logistic regression

```
Number of obs = 10,349
LR chi2(3) = 894.39
Prob > chi2 = 0.0000
Pseudo R2 = 0.1964
```

Log likelihood = -1829.7338

heartatk	Odds ratio	Std. err.	z	P> z	[95% conf. interval]
age	1.081366	.0085734	9.87	0.000	1.064692 1.098301
sex					
Male	1.062546	.6620164	0.10	0.922	.3133349 3.603185
sex#c.age					
Male	1.019639	.0099367	2.00	0.046	1.000348 1.039301
_cons	.0003549	.0001799	-15.67	0.000	.0001314 .0009585

اگر دستور بالا `logit heartatk c.age## sex, or nolog` نوشته شود لگاریتم های بالای جدول نوشته نمیشوند.

با دستور `quietly margins sex, at (age=(20(10)60))` پیش بینی حمله قلبی برای زنان و مردان محاسبه می شود.

```
. margins sex, at (age=(20(10)60))
```

Adjusted predictions  
Model VCE: OIM

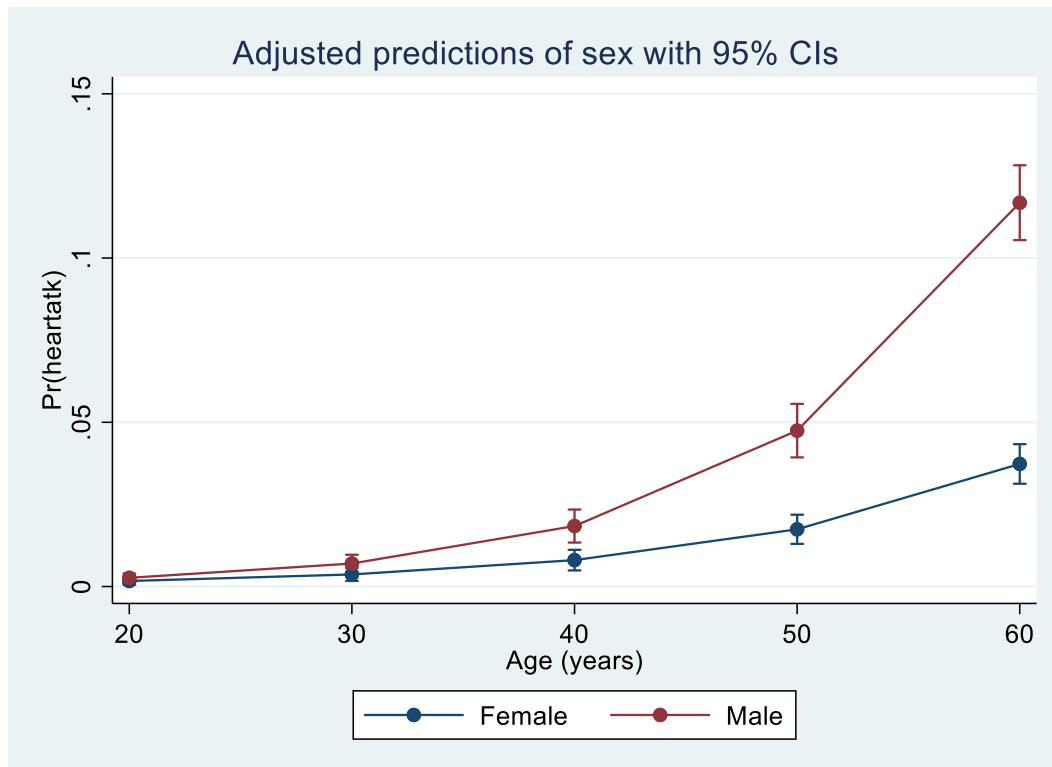
Number of obs = 10,349

Expression: Pr(heartatk), predict()

```
1._at: age = 20
2._at: age = 30
3._at: age = 40
4._at: age = 50
5._at: age = 60
```

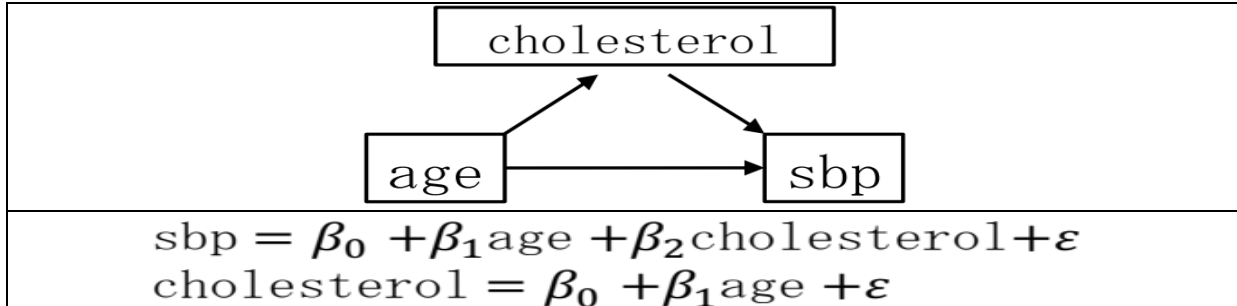
	Delta-method			z	P> z	[95% conf. interval]	
	Margin	std. err.					
_at#sex							
1#Female	.0016936	.0005941	2.85	0.004	.0005292	.002858	
1#Male	.0026525	.0006632	4.00	0.000	.0013527	.0039524	
2#Female	.0036954	.0010121	3.65	0.000	.0017116	.0056792	
2#Male	.0070138	.0013628	5.15	0.000	.0043426	.0096849	
3#Female	.0080443	.0016009	5.02	0.000	.0049066	.0111821	
3#Male	.0184131	.0025664	7.17	0.000	.0133831	.0234431	
4#Female	.0174218	.0022645	7.69	0.000	.0129833	.0218602	
4#Male	.0474543	.0041561	11.42	0.000	.0393084	.0556002	
5#Female	.0373194	.003074	12.14	0.000	.0312945	.0433443	
5#Male	.1168469	.0058088	20.12	0.000	.1054619	.1282318	

با دستور `marginsplot` نمودار تطبیق شده حمله قلبی برای جنسیت ترسیم می شود.

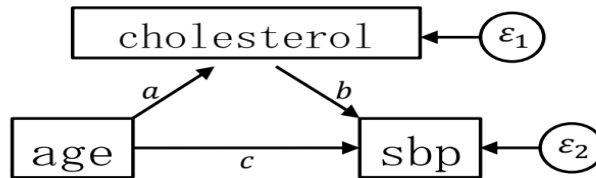


مثال متغیر میانجی و توضیح دهنده در روابط بین متغیرها با نرم افزار

دیاگرام و معادله رابطه سن و فشارخون و متغیر سوم کلسترول به عنوان متغیر میانجی را ترسیم نمائید.



مدل سازی معادلات ساختاری (SEM) یک چارچوب تجزیه و تحلیل آماری چند متغیره است که امکان برآورد همزمان یک سیستم معادلات، از جمله متغیرهای اندازه گیری شده با خطا و متغیرهای نهفته را فراهم می آورد. با استفاده از این مدل می توانید سهم مستقیم، غیرمستقیم و کل متغیرها بر فشارخون را محاسبه نمائید. دیاگرام زیر را ببینید.

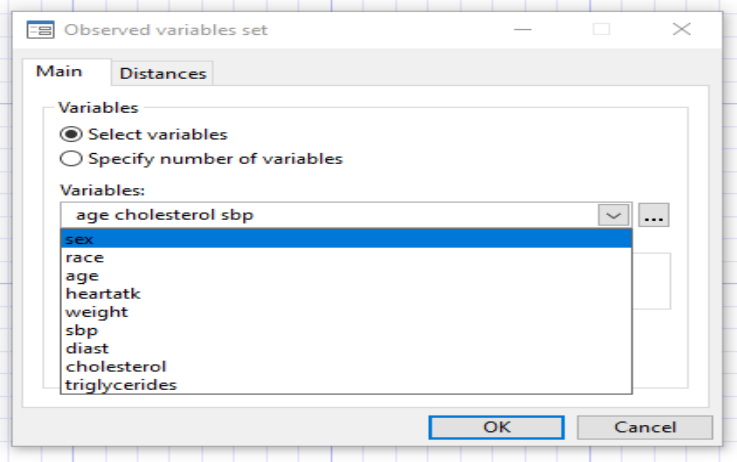


Direct Effect:  $c$   
 Indirect Effect:  $a \times b$   
 Total Effect:  $a \times b + c$   
 Percent of effect being mediated:  $\frac{\text{Indirect}}{\text{Total}}$

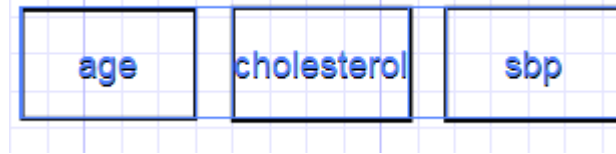
برای ترسیم دیاگرام و محاسبات بالا در نرم افزار از منوی statistics باید

انتخاب نمائید. SEM (structural equation modeling) ▶ Model building and estimation

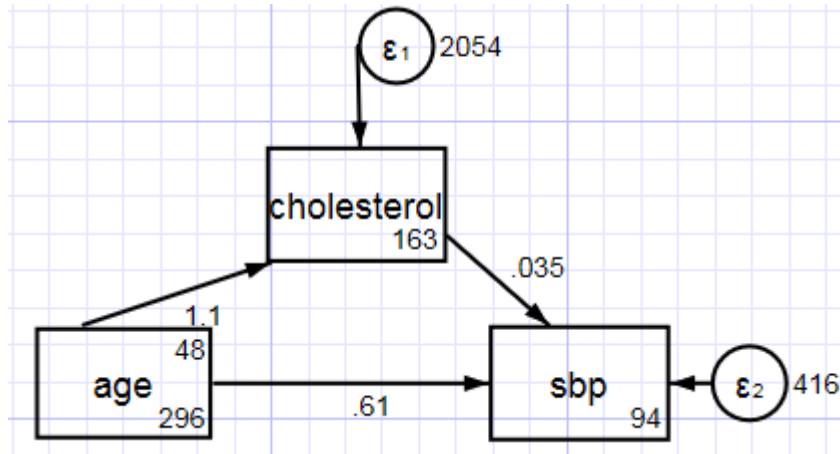
از صفحه ظاهر شده ابزار سمت چپ را انتخاب و بر روی صفحه شطرنجی کلیک نموده و متغیرها را در پنجره زیر معرفی کنید.



پس از اجرای دستور، اکنون مستطیل های مربوط به سه متغیر به شکل زیر ترسیم و متغیرها نامگذاری می شوند.



با انتخاب در سمت چپ، چیدمان مستطیل ها تنظیم می گردد. با انتخاب روابط یا مسیر فلش دار بین متغیرها نیز ترسیم می شود. سپس با انتخاب منوی Estimation محاسبات انجام می شود.



نتایج مدل به صورت زیر در خروجی نرم افزار قابل مشاهده است:

```
sem (age -> cholesterol, ) (age -> sbp, ) (cholesterol -> sbp, ), nocapslatent
. sem (age -> cholesterol, ) (age -> sbp, ) (cholesterol -> sbp, ), nocapslatent
```

Endogenous variables

Observed: cholesterol sbp

Exogenous variables

Observed: age

Fitting target model:

Iteration 0: log likelihood = -144204.81

Iteration 1: log likelihood = -144204.81

Structural equation model

Number of obs = 10,351

Estimation method: ml

Log likelihood = -144204.81

	OIM					[95% conf. interval]	
	Coefficient	std. err.	z	P> z			
<b>Structural</b>							
<b>cholesterol</b>							
age	1.139262	.0258793	44.02	0.000	1.08854	1.189985	
_cons	163.464	1.309436	124.84	0.000	160.8975	166.0304	
<b>sbp</b>							
cholesterol	.0352269	.0044221	7.97	0.000	.0265596	.0438941	
age	.6119448	.0126865	48.24	0.000	.5870797	.63681	
_cons	94.09771	.9325231	100.91	0.000	92.26999	95.92542	
var(e.cholesterol)	2054.233	28.55443			1999.023	2110.968	
var(e.sbp)	415.8133	5.779925			404.6378	427.2975	

LR test of model vs. saturated: chi2(0) = 0.00

Prob > chi2 = .

در خروجی بالا ضریب اثر متغیر سن بر کلسترول (a)، ضریب اثر کلسترول بر فشارخون (b)، و ضریب اثر مستقیم سن بر کلسترول (c) مشخص شده است.

با دستور estat teffects, compact در خروجی زیر سهم اثرات مستقیم، غیرمستقیم و کل گزارش گردید.

Direct Effect: *c*

Indirect Effect:  $a \times b$

Total Effect:  $a \times b + c$

Percent of effect being mediated:

$$\frac{\text{Indirect}}{\text{Total}}$$

## روش های اپیدمیولوژی و پژوهش با Stata

Direct effects

	OIM		z	P> z	[95% conf. interval]	
	Coefficient	std. err.				
Structural cholesterol age	1.139262	.0258793	44.02	0.000	1.08854	1.189985
sbp cholesterol age	.0352269	.0044221	7.97	0.000	.0265596	.0438941
	.6119448	.0126865	48.24	0.000	.5870797	.63681

**a=1.13**

**b=0.35**

**C=0.611**

Indirect effects

	OIM		z	P> z	[95% conf. interval]	
	Coefficient	std. err.				
sbp age	.0401327	.0051198	7.84	0.000	.030098	.0501673

**a × b=0.04**

Total effects

	OIM		z	P> z	[95% conf. interval]	
	Coefficient	std. err.				
Structural cholesterol age	1.139262	.0258793	44.02	0.000	1.08854	1.189985
sbp cholesterol age	.0352269	.0044221	7.97	0.000	.0265596	.0438941
	.6520775	.0116789	55.83	0.000	.6291872	.6749678

**a × b + c=0.65**

. estat summarize

Estimation sample sem                          Number of obs =    10,351

Variable	Mean	Std. dev.	Min	Max
cholesterol	217.6697	49.38694	80	828
sbp	130.8817	23.33265	65	300
age	47.57965	17.21483	20	74

Direct Effect: **C** اثر مستقیم سن روی فشارخون بالا

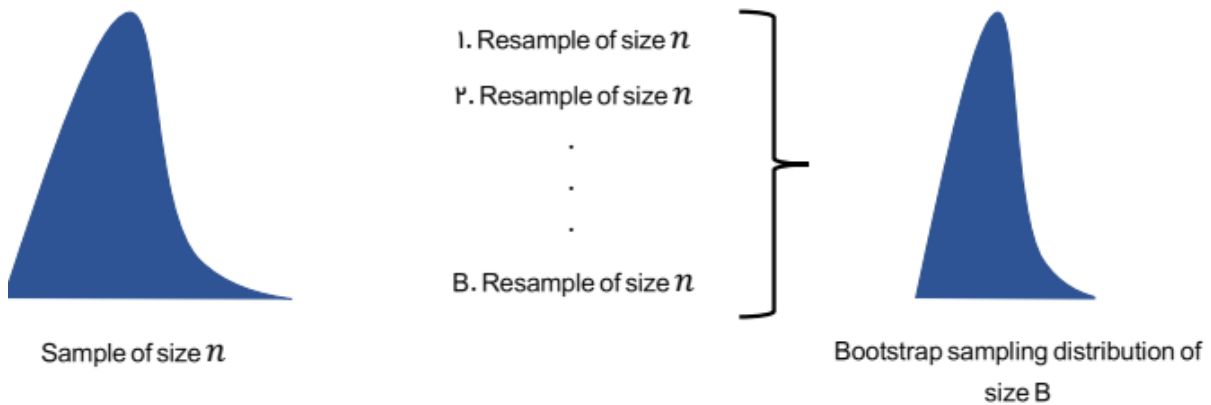
Indirect Effect:  $a \times b$  حاصل ضرب اثر سن روی کلسترول و اثر کلسترول روی فشارخون

Total Effect:  $a \times b + c$

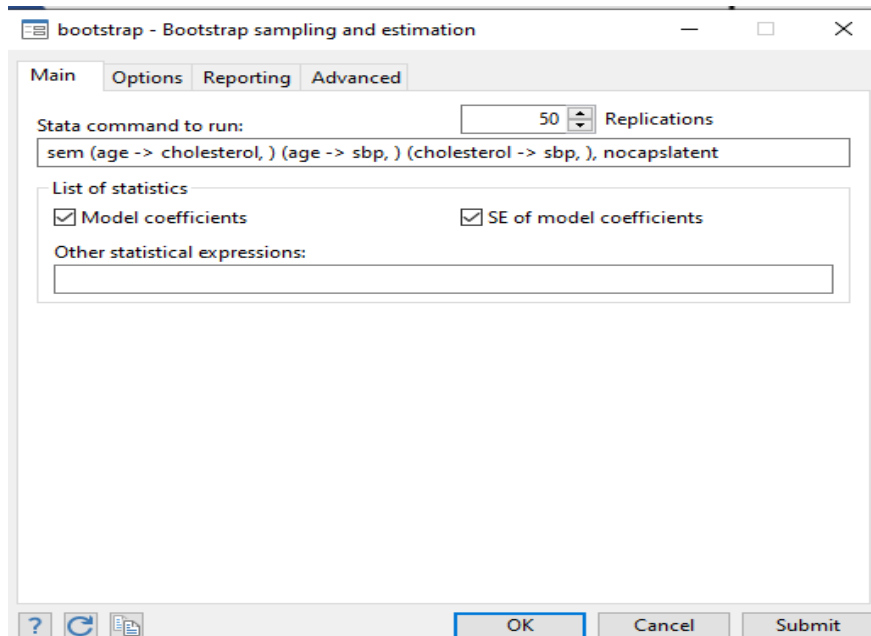
Percent of effect being mediated: نسبت اثر غیرمستقیم به کل اثرات سهم مربوط به متغیر میانجی است.

## بوت استرپ Bootstrap Mediation برای متغیر میانجیگر

از خطاهای استاندارد روش Sobel برای آزمایش اثر متغیر میانجیگر استفاده می شود (Sobel, 1987). این روش فرض می کند که  $a$  و  $b$  مستقل هستند و این روش توزیع نرمال دارد. Bootstrapping یک رویکرد غیر پارامتریک است که پیش فرضی برای توزیع آن وجود ندارد. این روش از روش Sobel قوی تر است، اما از نظر محاسباتی تمرکز بیشتری دارد. در شکل زیر توزیع نمونه گیری های مجدد نمایش داده شد.



می توان در روش بوت استرپ دستوراتی که برای برآورد پارامترهای sem اجرا شد را با بازنمونه گیری انجام و اجرا نمود. با دستور db bootstrap منوی مربوطه را باز نمائید. اکنون دستورات اجرا شده برای مدل sem که در بالا وجود دارند را در قسمت بوت استرپ معرفی نمائید.



`bootstrap, reps(50) : sem (age -> cholesterol, ) (age -> sbp, ) (cholesterol -> sbp, ), nocapslatent  
estat teffects`

نتیجه بوت استرپ در زیر آمده است.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. bootstrap, reps(50) : sem (age -> cholesterol, ) (age -> sbp, ) (cholesterol -> sbp, ), nocapslatent
(running sem on estimation sample)
```

```
Bootstrap replications (50)
-----|-----|-----|-----|-----|-----|
1     2     3     4     5
.....
50
```

```
Structural equation model           Number of obs = 10,351
Estimation method: ml             Replications = 50
Log likelihood = -144204.81
```

	Observed coefficient	Bootstrap std. err.	z	P> z	Normal-based [95% conf. interval]	
Structural						
cholesterol						
age	1.139262	.0260248	43.78	0.000	1.088255	1.19027
_cons	163.464	1.274622	128.25	0.000	160.9658	165.9622
sbp						
cholesterol	.0352269	.0047993	7.34	0.000	.0258204	.0446334
age	.6119448	.0134292	45.57	0.000	.585624	.6382657
_cons	94.09771	.97639	96.37	0.000	92.18402	96.01139
var(e.cholesterol)	2054.233	58.06606			1943.521	2171.252
var(e.sbp)	415.8133	8.377646			399.7134	432.5617

```
. estat teffects
```

```
Direct effects
```

	Observed coefficient	Bootstrap std. err.	z	P> z	Normal-based [95% conf. interval]	
Structural						
cholesterol						
age	1.139262	.0260248	43.78	0.000	1.088255	1.19027
sbp						
cholesterol	.0352269	.0047993	7.34	0.000	.0258204	.0446334
age	.6119448	.0134292	45.57	0.000	.585624	.6382657

```
Indirect effects
```

	Observed coefficient	Bootstrap std. err.	z	P> z	Normal-based [95% conf. interval]	
Structural						
cholesterol						
age	0 (no path)					
sbp						
cholesterol	0 (no path)					
age	.0401327	.0056401	7.12	0.000	.0290783	.051187

```
Total effects
```

	Observed coefficient	Bootstrap std. err.	z	P> z	Normal-based [95% conf. interval]	
Structural						
cholesterol						
age	1.139262	.0260248	43.78	0.000	1.088255	1.19027
sbp						
cholesterol	.0352269	.0047993	7.34	0.000	.0258204	.0446334
age	.6520775	.0126245	51.65	0.000	.6273339	.6768211

```
. estat bootstrap
```

```
Structural equation model           Number of obs = 10,351
Replications = 50
```

	Observed coefficient	Bias	Bootstrap std. err.	[95% conf. interval]	
cholesterol					
age	1.1392622	.0009573	.02602481	1.087978	1.176215 (BC)
_cons	163.46399	-.0399995	1.2746217	160.2863	165.7456 (BC)
sbp					
cholesterol	.03522689	.0010769	.00479934	.0242194	.0449506 (BC)
age	.61194483	-.0009091	.01342924	.5831704	.6340507 (BC)
_cons	94.097705	-.1861799	.97638995	92.33478	95.84893 (BC)
/					
var(e.chol~)	2054.2331	-1.51874	58.066057	1976.933	2208.514 (BC)
var(e.sbp)	415.81331	-.368945	8.3776459	392.7501	428.1553 (BC)



## کمی کردن خطاهای پژوهش با رویکرد تحلیل حساسیت با نرم افزار Stata :

در قسمت های قبلی درباره تقسیم بندی و تعریف خطاهای پژوهش، توضیحات مشروح ذکر گردید. برای کمی نمودن خطاهای پژوهش از دو رویکرد مبتنی بر مدل های قطعی (Deterministic Sensitivity Analysis) و تصادفی یا مبتنی بر احتمالات (Probabilistic Sensitivity Analysis) و انجام تحلیل حساسیت استفاده می شود. بدین منظور از نتایج حاصل از اطلاعات مطالعه مورد - شاهد گرینلند (Greenland) و همکاران در سال ۱۹۹۴ درخصوص بررسی رابطه مواجهه شغلی با رزین و مرگ ناشی از سرطان ریه استفاده می شود. دستور cci برای محاسبه نسبت شانس در یک جدول دو در دو، حاصل از مطالعه مذکور به شرح زیر آمده است.

. cci 45 94 257 945, woolf

	Exposed	Unexposed	Proportion	
			Total	Exposed
Cases	45	94	139	0.3237
Controls	257	945	1202	0.2138
Total	302	1039	1341	0.2252
	Point estimate		[95% Conf. Interval]	
Odds ratio	1.760286		1.202457	2.576898 (Woolf)
Attr. frac. ex.	.4319106		.1683693	.6119365 (Woolf)
Attr. frac. pop	.1398272			

chi2(1) = 8.63 Pr>chi2 = 0.0033

بر اساس نتایج بالا همبستگی مثبت و معنی داری بین مواجهه شغلی با رزین و مرگ ناشی از سرطان ریه با نسبت شانس (OR=1.76) بدست آمده است.

وقتی بر اساس سن تطبیق انجام می شود تغییری در رابطه مذکور ایجاد نمی شود. چنانچه فرض شود خطای اندازه گیری در بدست آوردن اطلاعات مواجهه بین گروه های مطالعه یا خطای انتخاب در گروه های مطالعه رخ داده باشد این خطاها را به صورت زیر می توان کمی و محاسبه نمود. دستور کلی کمی کردن خطاها در Stata به صورت زیر است. گزینه های این دستور در زیر توضیح داده شد.

`episens var_case var_exposed [var_time] [if] [in] [weight] [, options]`

`episensi #a #b #c #d [, options]`

### ۱- رویکرد تحلیل حساسیت در مدل های قطعی: Deterministic Sensitivity Analysis

#### الف: تورش طبقه بندی نادرست مواجهه Exposure Misclassification

شاخص های حساسیت و ویژگی در شناسایی و طبقه بندی موردها و شاهدها این اجازه را به پژوهشگر می دهند تا به تطبیق داده های مشاهده شده و طبقه بندی خطاها و کمی نمودن آنها بر اساس تحلیل حساسیت بپردازد. ابتدا با آنالیز حساسیت، با فرض عدم خطای طبقه بندی غیراتراقی برای اندازه گیری مواجهه و با استفاده از مدل ثابت و مقادیر یکسان حساسیت و ویژگی برای افراد بیمار و غیر بیمار تحلیل انجام می شود. برای این تحلیل از نتایج

مطالعه گرینلد و همکاران که در بالا مطرح گردید استفاده می شود. این تحلیل با دستور `episensi` انجام می شود. دو پکیج از این دستور در اینترنت وجود دارد باید آنها را در نرم افزار نصب نمائید. بعد از معرفی ارقام جدول دو در دو، نام مطالعه در پرانتز `st` می آید. بعد از آن مقادیر حساسیت و ویژگی برای بیماران و غیر بیماران و مواجهه یافته ها و مواجهه نیافته ذکر می شود. همانگونه که در دستور زیر آمده است این مقدار برای همه گروه ها یکسان و  $1/9$  فرض گردید. نتیجه تحلیل در زیر آمده است.

```
. episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.9)) dsenc(c(.9)) dspnc(c(.9))
```

```
Se|Cases : Constant(.9)
Sp|Cases : Constant(.9)
Se|No-Cases: Constant(.9)
Sp|No-Cases: Constant(.9)
```

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for misclassification of the exposure

External adjusted Odds Ratio = 2.34

Percent bias = -25%

بر اساس نتایج بالا، نسبت شانس تطبیق شده برای طبقه بندی نادرست مواجهه  $2/34$  می باشد. این نسبت در مقابل External adjusted Odds Ratio آمده است. نسبت شانس قبل از تطبیق با حدود اطمینان  $95\%$  نیز گزارش شده است. این نسبت همان نسبت شانس معنی دار مطالعه گرینلد است. درصد تورش یا خطای مطالعه نیز منفی  $25\%$  گزارش گردید. این درصد به صورت زیر محاسبه می شود و نشان می دهد نسبت شانس افزایش یافته است و به عبارت دیگر قبل از تطبیق  $25\%$  خطا وجود داشت که پس از تطبیق علامت آن منفی است.

```
. display (1.76-2.34) / 2.34 * 100
-24.786325
```

دستورات زیر را اجرا نمائید. این دستورات با سناریوهای متفاوت مقادیر حساسیت و ویژگی است.

```
. episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.9)) dsenc(c(.9)) dspnc(c(.9))"
```

```
. episensi 45 94 257 945, st(cc) dpunexp(c(.55)) dorce(c(2.45)) drrcd(c(5))"
```

```
. episensi 45 94 257 945, st(cc) dpexp(uni(.4 .7)) dpunexp(uni(.4 .7)) drrcd(log-n(2.159 .280))"
```

```
. episensi 45 94 257 945, st(cc) nodots dpunexp(uni(.4 .7)) drrcd(log-n(2.159 .280)) dorce(log-normal(0 .639))"
```

خروجی زیر مربوط به اولین دستور است. آن را ملاحظه نمائید:

```
. findit episensi
```

```
. . episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.9)) dsenc(c(.9)) dspnc(c(.9))"
```

```
Se|Cases : Constant(.9)
Sp|Cases : Constant(.9)
Se|No-Cases: Constant(.9)
Sp|No-Cases: Constant(.9)
```

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for misclassification of the exposure

External adjusted Odds Ratio = 2.34

Percent bias = -25%

اگر بر اساس سناریوی دیگری حساسیت در بیماران  $90\%$  درصد و در غیربیماران  $80\%$  درصد بکار رفته باشد دستور و نتیجه تحلیل به صورت زیر خواهد بود.

```
. episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.8)) dsenc(c(.8)) dspnc(c(.8))
```

```
Se|Cases : Constant(.9)
Sp|Cases : Constant(.8)
Se|No-Cases: Constant(.8)
Sp|No-Cases: Constant(.8)
Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]
```

```
Deterministic sensitivity analysis for misclassification of the exposure
External adjusted Odds Ratio = 9.11
Percent bias = -81%
```

در این خروجی نیز نسبت شانس تطبیق یافته برای تورش طبقه بندی نادرست ۹/۱۱ بدست آمده است. درصد خطا نیز به منفی ۸۱٪ رسیده است.

بنابراین در سناریوهای مختلف درصد خطا با تغییر نمودن حساسیت و ویژگی، تغییر می کند. در دستور بکار رفته برای خطای طبقه بندی نادرست مواجهه گزینه های دستور به صورت زیر است.

<u>dseca</u> (pdf_options)	define the sensitivity among the cases
<u>dspca</u> (pdf_options)	define the specificity among the cases
<u>dsenc</u> (pdf_options)	define the sensitivity among the noncases
<u>dspnc</u> (pdf_options)	define the specificity among the noncases
<u>corrrens</u> (#)	set the correlation between case and noncase sensitivities to #
<u>corrspec</u> (#)	set the correlation between case and noncase specificities to #

### ب: تورش انتخاب Selection Bias

با توجه به فقدان اطلاعات کافی درباره شغل افراد، مطالعه گرینلند نشان می دهد که احتمال خطای انتخاب در انتخاب بیماران و غیربیماران به عنوان گروه های مورد و شاهد در مطالعه به ترتیب ۰/۷ و ۰/۶ است. اگر انتخاب افراد بیمار و غیر بیمار با طبقه بندی مواجهه یافته و مواجهه نیافته ها در مطالعه مرتبط باشند خطای انتخاب رخ می دهد. این خطا را می توان محاسبه و کمی نمود.

همچنین می توان نسبت شانس تطبیق یافته برای این خطا را بدست آورد. دستور و نتایج این تحلیل در زیر آمده است. در این دستور نیز مانند دستور قبلی احتمال انتخاب بیمار در بین مواجهه یافته ها و مواجهه نیافته ها و احتمال انتخاب غیر بیمار در مواجهه یافته ها و مواجهه نیافته ها در دستور به نرم افزار به صورت زیر داده می شوند. گزینه های دستور کمی نمودن تورش انتخاب به صورت زیر است.

```

dpscex(pdf_options)      define the selection probability among cases
                             exposed
dpscun(pdf_options)      define the selection probability among cases
                             unexposed
dpsnex(pdf_options)      define the selection probability among noncases
                             exposed
dpsnun(pdf_options)      define the selection probability among noncases
                             unexposed
dsbfactor(pdf_options)   define the selection-bias factor

. episensi 45 94 257 945, st(cc) dpscex(c(.7)) dpscun(c(.7)) dpsnex(c(.6)) dpsnun(c(.
> 6))

Pr Case Selection Exposed: Constant(.7)
Pr Case Selection No-Exposed: Constant(.7)
Pr No-Case Selection Exposed: Constant(.6)
Pr No-Case Selection No-Exposed: Constant(.6)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
  External adjusted Odds Ratio = 1.76
  Percent bias = 0%

اگر احتمالات انتخاب بیمار و غیربیمار با توجه به وضعیت مواجهه متفاوت باشد فاکتور خطا بیشتر از یک خواهد شد.
اگر احتمال انتخاب بیمار در افراد مواجهه یافته ۰/۹ و در افراد مواجهه نیافته ۰/۵ باشد. و احتمال انتخاب غیر بیمار در
افراد مواجهه یافته و مواجهه نیافته به ترتیب ۰/۵ و ۰/۷ باشد فاکتور خطای انتخاب معادل ۲/۵ خواهد بود. محاسبات
در زیر آمده است.

. display (0.9 / 0.5) / (0.5 / 0.7)
2.52

نسبت شانس تطبیق یافته برای خطای انتخاب برابر است با : ۰/۷

. display 1.76 / 2.52
.6984127

خروجی این تحلیل در زیر آمده است.

. episensi 45 94 257 945, st(cc) dpscex(c(.9)) dpscun(c(.5)) dpsnex(c(.5)) dpsnun(c(.
> 7))

Pr Case Selection Exposed: Constant(.9)
Pr Case Selection No-Exposed: Constant(.5)
Pr No-Case Selection Exposed: Constant(.5)
Pr No-Case Selection No-Exposed: Constant(.7)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
  External adjusted Odds Ratio = 0.70
  Percent bias = 152%

```

در خروجی بالا همانگونه که ملاحظه می شود نسبت شانس تطبیق یافته برای خطا ۰.۷. گزارش شده است. طرز محاسبه این نسبت شانس قبلاً ذکر گردید. درصد خطا در این خروجی ۱۵۲ درصد گزارش شده است. در سناریوی دیگری چنانچه احتمال انتخاب بیمار در مواجهه یافته ها و مواجهه نیافته ها به ترتیب ۰.۵ و ۰.۹ باشد و احتمال انتخاب غیربیمار در مواجهه یافته ها و مواجهه نیافته ها به ترتیب ۰.۷ و ۰.۵ باشد فاکتور یا عامل خطای انتخاب برابر با ۰.۴ خواهد بود. نسبت شانس تطبیق یافته و مقدار خطا در زیر آمده است.

```
. display (0.5/0.9) / (0.7 /0.5)
.3968254

. display 1.76 / 0.3968254
4.4352
. episensi 45 94 257 945, st(cc) dpsex(c(.5)) dpscun(c(.9)) dpsnex(c(.7)) dpsnun(c(.
> 5))

Pr Case Selection Exposed: Constant(.5)
Pr Case Selection No-Exposed: Constant(.9)
Pr No-Case Selection Exposed: Constant(.7)
Pr No-Case Selection No-Exposed: Constant(.5)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
External adjusted Odds Ratio = 4.44
Percent bias = -60%
```

### ج: مخدوش کننده های کنترل نشده یا اندازه گیری نشده: Uncontrolled Confounders

در مطالعه بررسی رابطه مواجهه شغلی با رزین و مرگ ناشی از سرطان ریه چنانچه اطلاعاتی پیرامون سیگار کشیدن افراد در دسترس تحلیل کننده نباشد و محقق بخواهد خطای ناشی از متغیر مخدوش کننده بالقوه ای مانند سیگار کشیدن که اطلاعات آن در مطالعه جمع آوری نشده را در نظر بگیرد، تحلیل به صورت زیر انجام می شود. ابتدا محقق به مطالعات قبلی برمی گردد و با فرض اینکه رابطه سیگار کشیدن و مرگ سرطان ریه خطر نسبی معادل ۵ داشته باشد و شیوع سیگار کشیدن در میان مواجهه یافته های شغلی و غیر مواجهه یافتگان شغلی به ترتیب ۰.۷ و ۰.۵ باشد کمی نمودن خطای مخدوش شدگی با دستور زیر محاسبه می شود.

```
. episensi 45 94 257 945, dpexp(c(.7)) dpunexp(c(.5)) drrcd(c(5))
```

در دستور بالا گزینه ها به صورت زیر تعریف می شوند.

<code>dpexp(pdf_options)</code>	define the prevalence of the confounder among the exposed
<code>dpunexp(pdf_options)</code>	define the prevalence of the confounder among the unexposed
<code>drrcd(pdf_options)</code>	define the confounder-disease relative risk
<code>dorce(pdf_options)</code>	define the confounder-exposure odds ratio
<code>corrprev(#)</code>	set the correlation between exposure-specific confounder prevalences to #

نتیجه این دستور در زیر آمده است.

Pr(c=1|e=1): Constant(.7)  
 Pr(c=1|e=0): Constant(.5)  
 RR\_cd : Constant(5)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for unmeasured confounding  
 External adjusted Odds Ratio = 1.39  
 Percent bias = 27%

نسبت شانس برای مواجهه با رزین و مرگ سرطان ریه که برای متغیر مخدوش کننده سیگار تطبیق شده است برابر با ۱/۳۹ گزارش شده است. این مقدار نسبت شانس کمتر از نسبت شانس مشاهده شده (قبل از تطبیق) است. زیرا در محاسبات بالا فرض گردیده است که رابطه متغیر مخدوش کننده (سیگار) و پیامد (مرگ ناشی از سرطان ریه) همبستگی مثبت و بیشتر از ۱ دارد. در بالا این مقدار ۵ در نظر گرفته شد. همچنین رابطه متغیر مخدوش کننده با متغیر مواجهه (رزین) نیز ۰/۷ و بیشتر از ۰/۵ فرض گردید. درصد خطای ناشی از این مطالعه برای متغیر مخدوش کننده اندازه گیری نشده برابر با ۲۷٪ محاسبه و گزارش شده است.

## ۲- رویکرد تحلیل حساسیت مبتنی بر احتمالات: Probabilistic Sensitivity Analysis

محدودیت مهم در در تحلیل حساسیت مبتنی بر مدل های قطعی، تهدید خطا برای پارامتر در مدل است. وقتی محقق مقدار پارامتر را ثابت فرض نماید و یا از آن آگاه باشد ممکن است در برآورد پارامتر دچار خطا گردد. لذا در این روش بر خلاف روش قبلی از توزیع های مبتنی بر احتمالات استفاده می شود. و با توجه به این توزیع ها مقدار خطای پارامتر برآورد و کمی می گردد.

در این رویکرد نیز برای طبقه بندی نادرست مواجهه، تورش انتخاب و مخدوش کننده کنترل نشده، مقدار نسبت شانس تطبیق می گردد و همچنین مقادیر این خطاها کمی و گزارش می گردد. این محاسبات برای هر کدام از خطاها در زیر با مثال ارائه می شوند.

### الف: طبقه بندی نادرست مواجهه: Exposure Misclassification

ابتدا با فرض طبقه بندی نادرست غیرافتراقی مواجهه، با در نظر گرفتن تابعی از ویژگی و حساسیت در میان بیماران و غیربیماران و استفاده از توزیع ذوزنقه ای (Trapezoidal) با حداقل ۰/۷۵ و حداکثر ۱ و فاصله میان این مقادیر با احتمال برابر با ۰/۸۵ تا ۰/۹۵. برای بیماران، غیربیماران، مواجهه یافتگان و مواجهه نیافتگان، محاسبات انجام می شود.

دستور محاسبه و نتایج آن در زیر آمده است.

## روش های اپیدمیولوژی و پژوهش با Stata

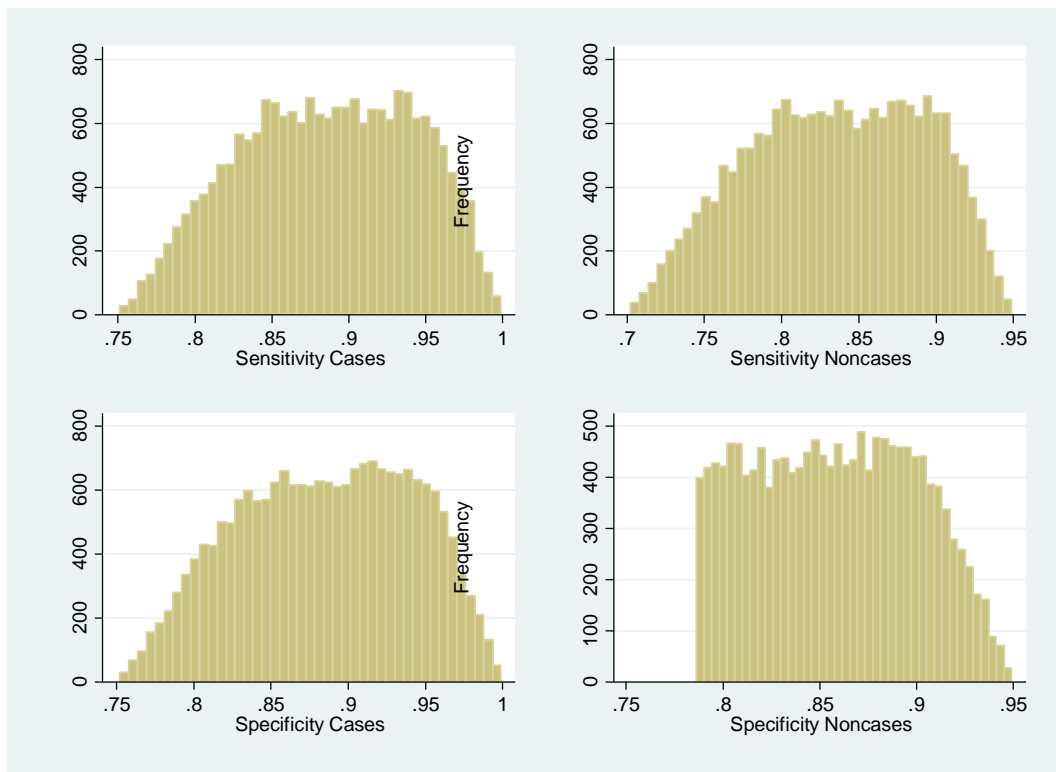
```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dseca(trap(.75 .85 .95 1)) dspca
> (trap(.75 .85 .95 1)) dsenc(trap(.75 .85 .95 1)) dspnc(trap(.75 .85 .95 1)) seed(12
> 3)
```

```
Se|Cases : Trapezoidal(.75,.85,.95,1)
Sp|Cases : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.75,.85,.95,1)
Sp|No-Cases: Trapezoidal(.75,.85,.95,1)
```

Probabilistic sensitivity analysis for misclassification of the exposure

	Percentiles			Ratio
	2.5	50	97.5	97.5/2.5
Conventional	1.20	1.76	2.58	2.14
Systematic error	1.87	2.47	14.71	7.86
Systematic and random error	1.49	2.57	15.00	10.07

خروجی بالا نشان می دهد نسبت شانس بر اساس تطبیق انجام شده برای مدل ، با صدک های ۲/۵ ، میانه و ۹۷/۵ با شبیه سازی با ۲۰۰۰۰ بار انجام و محاسبه گزارش شده است. نسبت شانس تطبیق یافته برای خطای منظم ۱/۸۷ تا ۱۴/۷۱ ، نسبت شانس تطبیق یافته برای خطای تصادفی ۱/۲ تا ۲/۵۸ گزارش گردید. نسبت شانس با در نظر گرفتن کلیه خطا های مذکور نیز تطبیق و از ۱/۴۹ تا ۱۵ گزارش شده است. هیستوگرام توزیع های فرض شده با ۲۰۰۰۰ بار شبیه سازی و شکل آنها برای مقادیر حساسیت و ویژگی در بیماران و غیربیماران در زیر آمده است.



گزینه های انجام شبیه سازی در دستورات بالا به شرح زیر تعریف و استفاده می گردند.

### Simulations

<code>reps(#)</code>	specify the number of replications to be performed
<code>nodots</code>	suppress the replication dots
<code>seed(#)</code>	set the random-number seed to #
<code>ndraw(#)</code>	number of observations drawn at each replication
<code>saving(filename)</code>	save results to <i>filename</i>
<code>grprior</code>	histogram of the priors
<code>grarrsys</code>	histogram of the adjusted relative risk (systematic error)
<code>grarrtot</code>	histogram of the adjusted relative risk (systematic error plus random error)

### ب: تورش انتخاب : Selection Bias

در محاسبه خطای انتخاب با رویکرد قبلی نشان داده شد که حساسیت نسبت شانس برای رابطه مواجهه با رزین و مرگ ناشی از سرطان ریه در سناریوهای مختلف برای کمی کردن خطای انتخاب متفاوت می باشد. چنانچه رابطه کوچکی بین مواجهه و پیامد وجود داشته باشد کمی کردن خطای انتخاب کمک کننده نیست. در اینجا با پیش فرض استفاده از توزیع احتمالات برای انتخاب هر یک از بیماران و غیربیماران و دانستن توزیع پیشین آنها می توان فاکتور خطای انتخاب را محاسبه و کمی نمود. با فرض تبعیت پارامترها از توزیع لگاریتمی نرمال با میانگین صفر و انحراف معیار ۰.۲۱ و ۹۵٪ حدود اطمینان برای احتمال پیشین عامل خطا ۰.۷ تا ۱/۵ می گردد. با دستور زیر می توان نسبت شانس را برای این خطا تطبیق و خطای انتخاب را برآورد نمود.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dsbfactor(log-n(0 0.21)) seed(123)
> grprior
```

نتایج این دستور در زیر آمده است.

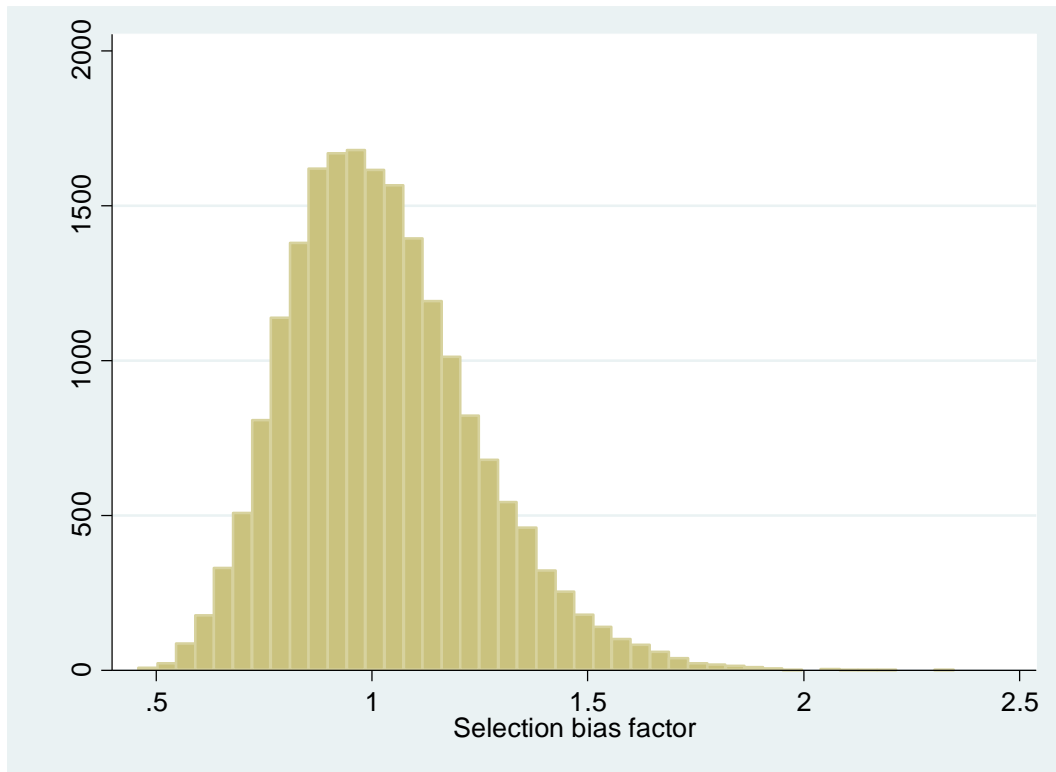
```
selection bias factor: Log-Normal(0.00,0.21)
```

#### Probabilistic sensitivity analysis for selection bias

	Percentiles			Ratio
	2.5	50	97.5	97.5/2.5
Conventional	1.20	1.76	2.58	2.14
Systematic error	1.16	1.76	2.66	2.29
Systematic and random error	1.01	1.75	3.08	3.05

نمودار هیستوگرام خطای انتخاب نیز با ۲۰۰۰۰ شبیه سازی انجام شده و استفاده از توزیع لگاریتم نرمال در زیر آمده است.





### ج: مخدوش کننده کنترل نشده: Uncontrolled Confounder

چنانچه دو توزیع یکسان یا یکنواخت و مستقل برای متغیر مخدوش کننده بالقوه ( سیگار کشیدن) در میان مواجهه یافتگان و مواجهه نیافتگان در نظر گرفته شود و مقادیر آن  $.4$  و  $.7$  فرض گردد و این توزیع ها با احتمالات پیشین مستقل باشند و نسبت شانس خطر نیز از توزیع لگاریتمی نرمال تبعیت نماید. میانگین این توزیع  $2/159$  و انحراف معیار  $.28$  محاسبات به صورت زیر و با دستور زیر انجام می شوند.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dpexp(uni(.4 .7)) dpunexp(uni(.4 .7))
> 7)) drrcd(log-n(2.159 .280)) seed(123) grarrtot gprior
```

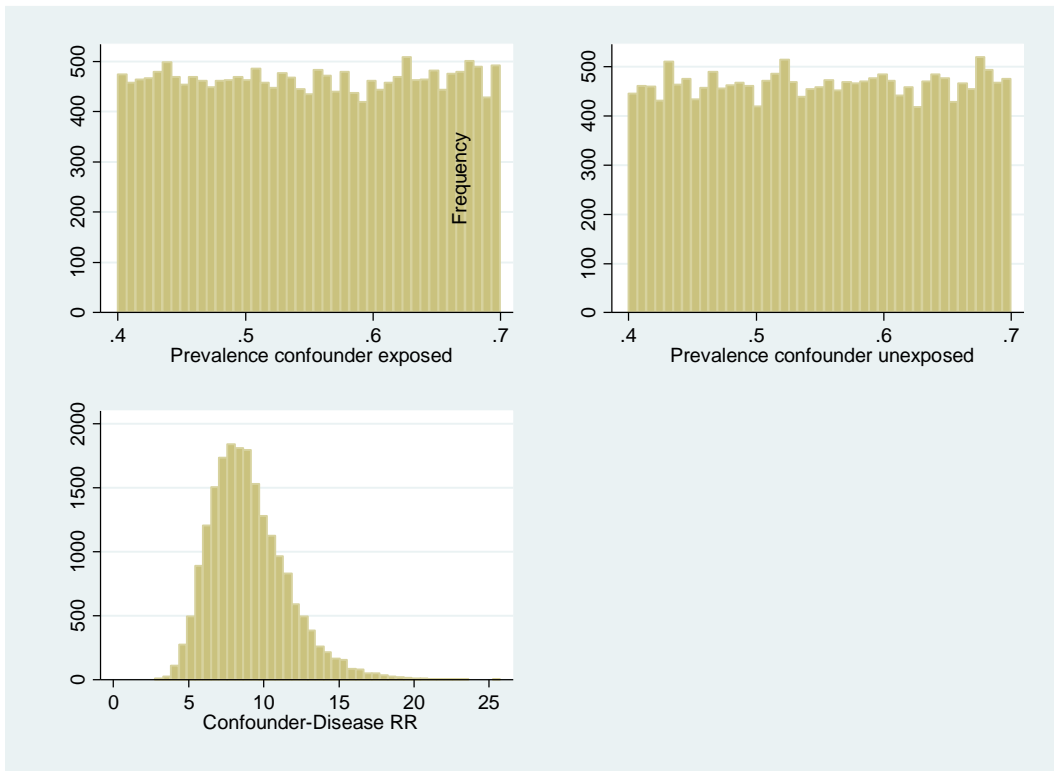
```
Pr(c=1|e=1): Uniform(.4,.7)
Pr(c=1|e=0): Uniform(.4,.7)
RR_cd      : Log-Normal(2.16,0.28)
```

#### Probabilistic sensitivity analysis for unmeasured confounding

	2.5	Percentiles		Ratio
		50	97.5	97.5/2.5
Conventional	1.20	1.76	2.58	2.14
Systematic error	1.25	1.76	2.49	2.00
Systematic and random error	1.05	1.76	2.96	2.83

در خروجی بالا نیز مقدار نسبت شانس تطبیق یافته در صدک های مختلف برای خطاهای مختلف محاسبه و گزارش شده است.

نمودار هیستوگرام شبیه سازی برای تطبیق متغیر مخدوش کننده و شیوع متغیر مخدوش کننده در گروه های مواجهه یافته و مواجهه نیافته در زیر آمده است.



### ۳- محاسبه و تحلیل کمی ترکیب خطاها : Combined Analysis of Biases

برای تطبیق کردن نسبت شانس برای خطاهای چندگانه (خطای انتخاب ، اطلاعات و مخدوش شدگی کنترل نشده) که تا کنون در مباحث قبلی به آنها پرداخته شد می توان با دستور زیر محاسبه را برای ترکیب خطاها انجام داد. نتیجه نشان می دهد که نسبت شانس تطبیق یافته برای همه خطاهای مذکور از ۱/۳۴ تا ۵۷/۵ قابل گزارش است.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dseca(trap(.75 .85 .95 1)) dspca(
> trap(.75 .85 .95 1)) dsenc(trap(.7 .8 .9 .95)) dspnc(trap(.7 .8 .9 .95)) corrsens(.
> 8) corrspec(.8) dsbfactor(log-n(0 0.21)) dpunexp(uni(.4 .7)) dorce(log-n(0 0.639))
> drrcd(log-n(2.159 .280)) seed(123) combined
```

```
Se|Cases : Trapezoidal(.75,.85,.95,1)
Sp|Cases : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.7,.8,.9,.95)
Sp|No-Cases: Trapezoidal(.7,.8,.9,.95)
Corr Se|Cases and Se|No-Cases : .8
Corr Sp|Cases and Sp|No-Cases : .8
selection bias factor: Log-Normal(0.00,0.21)
Pr(c=1|e=0): Uniform(.4,.7)
RR_cd : Log-Normal(2.16,0.28)
OR_ce : Log-Normal(0.00,0.64)
```

Probabilistic sensitivity analysis - Combined corrections  
 Misclassification of the exposure  
 Selection bias  
 Unmeasured confounding

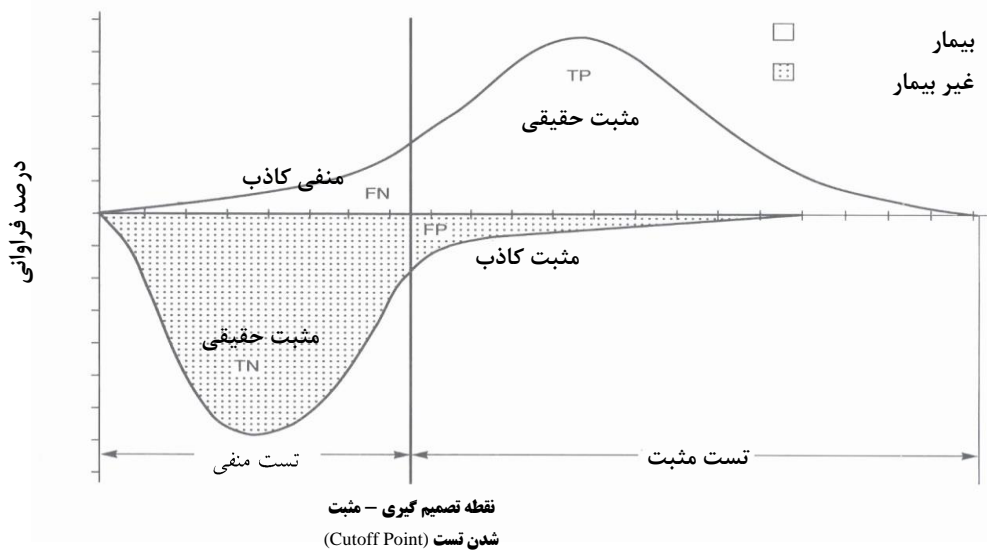
	Percentiles			Ratio
	2.5	50	97.5	97.5/2.5
Conventional	1.20	1.76	2.58	2.14
Systematic error	1.47	3.87	56.17	38.14
Systematic and random error	1.34	3.92	57.56	42.98

### خلاصه و نتیجه گیری:

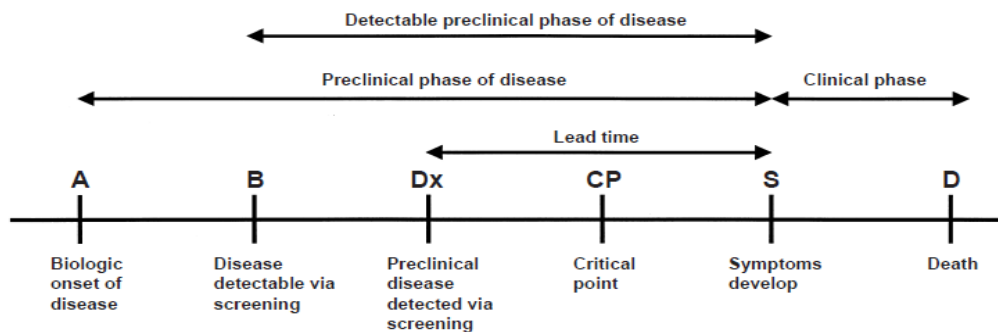
در این قسمت ابتدا به مبحث شبیه سازی آماری ، بوت استرپ ، روش های مونت کارلو و جک نایف اشاره گردید. سپس انواع خطاهای متداول در پژوهش های علوم پزشکی و اپیدمیولوژی طبقه بندی گردیدند. در انتها با استفاده از نرم افزار Stata و استفاده از دو رویکرد مبتنی بر قطعیت و تصادفی و با استفاده از اطلاعات مطالعه بررسی رابطه مواجهه با رزین و خطر مرگ سرطان ریه و متغیر سیگار کشیدن به عنوان متغیر مخدوش کننده ، نسبت شانس برای این رابطه با تطبیق بر اساس انواع خطاهای اطلاعات ، انتخاب و مخدوش شدگی کنترل نشده محاسبه گردید. سپس مقدار هر یک از خطاها به تفکیک و ترکیبی محاسبه و به درصد هر کدام از خطاها به صورت کمی گزارش گردیدند.

مطالعه ارزیابی تست های تشخیصی (آزمون های تشخیصی):

در الگوی جدید تصمیم گیری بالینی از پزشکی مبتنی بر شواهد که شامل درهم آمیختن بهترین شواهد تحقیقی، تجربه بالینی، ارزش ها و ارجحیت بیمار است استفاده می شود. اصول تصمیم گیری بالینی در تست های تشخیصی یکی از مباحث کاربردی و شایع در پزشکی مبتنی بر شواهد است. اطلاعات حاصل از آزمونهای تشخیصی بر اساس مقیاسهای اسمی، رتبه ای یا فاصله ای بیان می شود. در ساده ترین حالت جواب آزمایش دو تایی است: طبیعی یا غیر طبیعی. اکثر تستها مقادیر عددی (Quantitative Tests) می دهند، نه فقط یک جواب ساده مثبت یا منفی. بنابراین باید تعیین شود که تا چه مقداری طبیعی است و از چه مقدار غیر طبیعی است. به عبارتی، باید مرزی بین طبیعی و غیر طبیعی در نظر گرفت. این مرز در شکل زیر آمده است.



ساده ترین حالت تست تشخیصی بصورتی است، که در آن، نتیجه تست مثبت یا منفی است (دو حالت). نتیجه تست (بیمار بودن یا نبودن) با استاندارد طلایی (Gold Standard) یا مطمئن ترین روش یا روشهای تشخیصی موجود مقایسه می شود. یکی از مهمترین معیارها برای اینکه یک بیماری کاندید غربالگری باشد این است که آن بیماری باید یک دوره پیش بالینی قابل تشخیص (Detectable Preclinical Phase (DPCP) داشته باشد تا آزمون غربالگری را بتوان برای آن بکار برد. شکل زیر دوره بیماری را نمایش می دهد. شکل زیر سیر طبیعی یک بیماری است. دوره پیش بالینی را ملاحظه نمائید.



هر بیماری از یک نقطه که تغییر بیولوژیک نام دارد شروع میشود (نقطه A). بعد از تغییرات بیولوژیک در مرحله ای علایم بیماری بروز خواهد کرد (نقطه S). ممکن است در نهایت بیماری به مرگ یا بهبودی منجر شود. به فاصله بین شروع تغییرات بیولوژیک تا بروز علایم مرحله پیش بالینی می گویند. با توجه به عدم اطلاع از بیماری توسط فرد و پزشکان و همچنین دانش فعلی ما، معمولا دورتر از نقطه تغییرات بیولوژی و تقریبا در نقطه (B) می توان بیماری را تشخیص داد. البته با فرض اینکه بیمار تحت مراقبت بوده و در صورتی است که به هر دلیل به بیماری شک کرده باشیم و بیمار تحت بررسی دقیق باشد و آزمونهای دقیق تشخیصی نیز موجود باشد. در غیر این صورت و به صورت عادی و معمول، بیماری در زمان بروز علایم (نقطه S) تشخیص داده خواهد شد. معمولا " زمانی که علایم ظاهر میشود، بیماری از نقطه بحرانی (نقطه CP) قابل درمان گذشته و عوارضی را ایجاد کرده است و یا غیر قابل درمان شده است (این روند در مورد برخی بیماریها مشاهده نمیشود). پس اگر بتوانیم قبل از نقطه بحرانی (CP)، بیماری را تشخیص داده و مداخله نماییم برای ما مفید و خوشایند خواهد بود. بنابراین با غربالگری باید تلاش نمود در نقطه ای بین نقطه A و نقطه بحرانی بیماری (نقطه CP)، بیماری تشخیص داده شود. آزمون و تستی که برای غربالگری بکار می رود باید اعتبار مناسب داشته باشد. آزمون باید Validity و پایایی مناسبی داشته باشد. Validity خود دارای چندین جزء شامل حساسیت، ویژگی، ارزش اخباری مثبت و منفی است. جدول زیر مربوط به محاسبه شاخص های آزمون های تشخیصی است.

### شکل کلی جدول دو در دو برای تجزیه و تحلیل تست های تشخیصی:

		Gold Standard		total
		+	-	
Test	+	a	b	m <sub>1</sub>
	-	c	d	m <sub>0</sub>
total		n <sub>1</sub>	n <sub>0</sub>	N

در این جدول شاخص های زیر قابل محاسبه است:

$$\text{Sensitivity} = a/(a+c)$$

$$\text{Specificity} = d/(b+d)$$

$$\text{Accuracy} = (a+d)/(a+b+c+d)$$

$$\text{Positive Predictive Value (PPV)} = a/(a+b)$$

$$\text{Negative Predictive Value (NPV)} = d/(c+d)$$

- (۱) حساسیت<sup>۱</sup>: احتمال شناسایی درست کل بیماران
- (۲) ویژگی<sup>۲</sup>: احتمال شناسایی درست کل افراد سالم
- (۳) ارزش اخباری مثبت<sup>۳</sup>: احتمال وجود بیماری در یک نتیجه مثبت تست تشخیصی.
- (۴) ارزش اخباری منفی<sup>۴</sup>: احتمال عدم وجود بیماری به دنبال یک نتیجه منفی تست تشخیصی.

1. Sensitivity (Se)

2. Specificity (Sp)

3. Positive predictive value (PPV)

4. Negative predictive value (NPV)

۵) **صحت:** عبارت است از نسبت نتایجی که درست است به تمام نتایج (درست و غلط). در مقایسه یک تست تشخیصی با آزمون طلایی، صحت با درصد توافق برابر می شود. حدود اطمینان صحت برابر است با:

$$CI_{(Accuracy)} = Accuracy \pm Z_{\alpha/2} \sqrt{\frac{Accuracy(1 - accuracy)}{N}}$$

### ارزیابی تست های تشخیصی و محاسبه شاخص ها با Stata

فایل **diagnosis** را باز نمائید. اعتبار سنجی آزمون بکار رفته برای شناسایی قند خون بالا در مقابل آزمون طلایی و استاندارد تشخیص دیابت نوع دو را محاسبه کنید. سطح زیر منحنی راک را محاسبه کنید. از نصب بودن پکیج **diagt** یا با دستور **findit** آنرا نصب کنید.

. diagt Gold\_Standard test

Category 2	Category 1		Total
	Pos.	Neg.	
Abnormal	9	6	15
Normal	10	5	15
Total	19	11	30

True abnormal diagnosis defined as Gold\_Standard = 1

[95% Confidence Interval]

		50%	31%	68.7%
Prevalence	Pr(A)			
Sensitivity	Pr(+ A)	60%	32.3%	83.7%
Specificity	Pr(- N)	33.3%	11.8%	61.6%
ROC area	(Sens. + Spec.)/2	.467	.289	.645
Likelihood ratio (+)	Pr(+ A)/Pr(+ N)	.9	.521	1.55
Likelihood ratio (-)	Pr(- A)/Pr(- N)	1.2	.466	3.09
Odds ratio	LR(+)/LR(-)	.75	.176	3.2
Positive predictive value	Pr(A +)	47.4%	24.4%	71.1%
Negative predictive value	Pr(N -)	45.5%	16.7%	76.6%

خروجی بالا شاخص های اعتبارسنجی آزمون غربالگری اجرا شده با سه میزان شیوع را گزارش کرده است. سطح زیر منحنی در زیر گزارش شد.

. roctab Gold\_Standard test

Obs	ROC area	Std. err.	Asymptotic normal [95% conf. interval]	
30	0.4667	0.0909	0.28860	0.64473

۶- **نسبت درست‌نمایی<sup>۱</sup>:** برای محاسبه احتمال بیماری پس از یک آزمون مثبت یا منفی به کار می رود. LR نسبت دو احتمال است: احتمال نتیجه تست (+ یا -) در میان بیماران، نسبت به احتمال آن نتیجه در غیر بیماران. LR بزرگتر از یک نشان دهنده آن است که احتمال نتیجه تست (+ یا -) در میان بیماران بیشتر از نتیجه این تست در غیر بیماران است و بر عکس. نسبت درست‌نمایی مثبت<sup>۲</sup>: عبارت است از نسبت میزان مثبت حقیقی (حساسیت) به

$$LR(+) = \frac{Se}{FPR} \quad \text{میزان مثبت کاذب (ویژگی-۱):}$$

<sup>۱</sup>. Likelihood ratio (LR)

<sup>۲</sup>. LR+

نسبت درست‌نمایی منفی<sup>۱</sup>: عبارت است نسبت میزان منفی کاذب به منفی حقیقی:

$$LR(-) = \frac{FNR}{Sp}$$

به ارزش اخباری احتمال پسین هم می گویند. این احتمال را می توان از حاصلضرب احتمال پیشین (یا همان شیوع بیماری) در نسبت درست‌نمایی به دست آورد.

۷- **نسبت شانس**<sup>۲</sup>: عبارت است از شانس نتیجه مثبت نسبت به نتیجه منفی، در میان بیماران تقسیم بر شانس نتیجه مثبت به نتیجه منفی، در میان افرادی که بیمار نیستند. اگر OR بزرگتر از ۱ باشد، نشان دهنده این است که شانس نتیجه مثبت تست مورد بررسی برای بیماران بیشتر از افراد غیر بیمار است.

۸- **آماره کاپا**<sup>۳</sup>: به منظور تعیین درجه توافق تست های مورد بررسی، که فراتر از توافق شانسی بوده است، به کار می رود. آماره کاپا از فرمول زیر به دست می آید:

$$Kappa = \frac{P - P_0}{100 - P_0}$$

که در آن:  $P_0$  در صد توافق قابل انتظار است که بر مبنای شانس حاصل می شود،  $P$  در صد توافق مشاهده شده می باشد. درجه بندی مقادیر کاپا بر اساس شدت توافق بر شرح زیر است:

کاپای ۰/۷۵ و بیشتر: درجه توافق مشاهده شده عالی؛ کاپای ۰/۴۰-۰/۷۵: درجه توافق متوسط تا خوب؛ کاپای ۰/۴۰ و کمتر حاکی از درجه توافق ضعیف دارد.

۹- **درصد توافق**<sup>۴</sup>: عبارت است از نسبت مجموع نتایجی که هر دو تست، مثبت یا منفی هستند، به کل دفعاتی که هر دو تست انجام شده اند.

برای محاسبه کاپا از دستور db kappa استفاده و در جعبه گفتگو متغیرها را معرفی کنید. بعد از معرفی متغیر ارزیابی کننده اول و دوم با kap rata ratb, wgt(mine) می توانید محاسبه را انجام دهید.

دستور کلی محاسبه کاپا . kap varname1 varname2 varname3 [...] [if] [in] [weight]

. db kappa

. kappa rating judge

Two-outcomes, multiple raters:

Kappa	Z	Prob>Z
-0.0150	-0.38	0.6490

برای اجرای دستورات محاسبه کاپا از ۴ مسیر زیر در نرم افزار می توانید اقدام کنید.

کاپا برای دو مشاهده گر مستقل (kap: two unique raters)

Statistics > Epidemiology and related > Other > Interrater agreement, two unique raters

کاپا وزن دهی شده (kapwgt)

Statistics > Epidemiology and related > Other > Define weights for the above (kap)

1. LR-

2. Odds ratio (OR)

3. Kappa statistic

4. Concordance (Overall percent Agreement)

کاپا برای مشاهده گران وابسته (kap: nonunique raters)

Statistics > Epidemiology and related > Other > Interrater agreement, nonunique raters

کاپا و درصد توافق مشاهده گران (kappa)

Statistics > Epidemiology and related > Other > Interrater agreement, nonunique raters with frequencies

فایل kappa را باز نمائید. دستورات زیر را برای محاسبه توافق بین دو نفر پاتولوژیست برای خوانش لام های سرطان اجرا کنید.

```
. kap rada radb
```

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

```
. kap rada radb, tab
```

Radiologist A's assessment	Radiologist B's assessment				Total
	Normal	Benign	Suspect	Cancer	
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspect	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

```
. kap rada radb, wgt(w)
```

Ratings weighted by:

1.0000	0.6667	0.3333	0.0000
0.6667	1.0000	0.6667	0.3333
0.3333	0.6667	1.0000	0.6667
0.0000	0.3333	0.6667	1.0000

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
86.67%	69.11%	0.5684	0.0788	7.22	0.0000

```
. kap rada radb, wgt(w) absolute
```

Ratings weighted by:

1.0000	0.6667	0.3333	0.0000
0.6667	1.0000	0.6667	0.3333
0.3333	0.6667	1.0000	0.6667
0.0000	0.3333	0.6667	1.0000

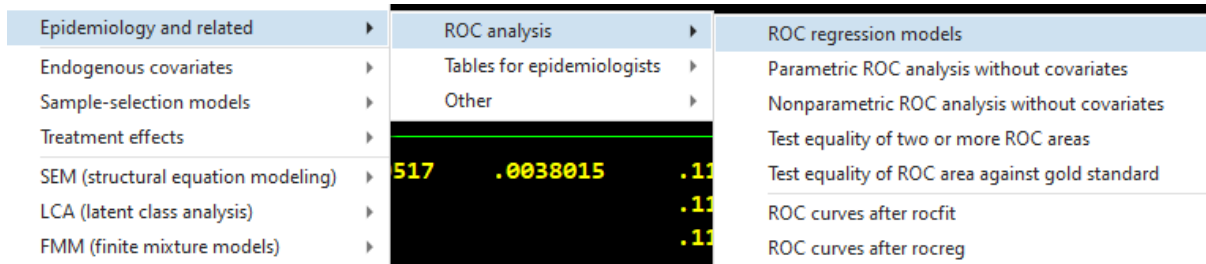
Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
86.67%	69.11%	0.5684	0.0788	7.22	0.0000

۱۰- تجزیه و تحلیل منحنی ROC<sup>۱</sup>: این منحنی با رسم میزان مثبت حقیقی (حساسیت) در برابر میزان مثبت کاذب (ویژگی-۱) ساخته می شود. می توان درستی یک تست تشخیصی را به صورت سطح زیر منحنی ROC توصیف کرد. هر چه سطح زیر منحنی بزرگتر باشد، تست تشخیصی، تست مناسبتری است. مقدار سطح زیر منحنی بین ۰ و ۱ متغیر است. توصیف سطح زیر منحنی: "چقدر احتمال وجود دارد که بیمار نتیجه مثبت داشته باشد". یک منحنی ROC نقشه نقطه چینی است که از حساسیت و ویژگی-۱ تشکیل شده است. به طوری که هر نقطه از منحنی با یک نقطه مرزی متناظر است. در صورتی که از نظر بالینی حساسیت و ویژگی ارزش یکسانی داشته باشند،

<sup>۱</sup>. ROC curve analysis



از نظر آماری، بهترین نقطه مرزی، نقطه ای در منحنی است که به بالا و گوشه چپ منحنی نزدیکتر باشد. تجزیه و تحلیل این منحنی، با ارزش ترین روش برای توصیف و مقایسه تست های تشخیصی می باشد. برای تحلیل منحنی راک از مسیر `Statistics > Epidemiology and related > ROC analysis` استفاده کنید. با انتخاب شرایط پارامتریک و ناپارامتریک و با یا بدون سایر متغیرها و یا مقایسه دو منحنی می توان تحلیل را انجام داد. مسیر منوها را در زیر ببینید.



فایل `hanley_ROC` را باز نمائید. در این فایل هانلی و مک نیل (۱۹۸۲) داده های یک مطالعه را ارائه نمودند. در آن از یک داور خواسته شد تا با استفاده از مقیاس پنج نقطه ای، یک نمونه تصادفی از ۱۰۹ تصویر توموگرافی از بیماران با مشکلات عصبی را طبقه بندی کند. مقیاس رتبه بندی (`rating`) به شرح زیر بود: ۱ قطعاً طبیعی است، ۲ احتمالاً طبیعی است، ۳ مشکوک است، ۴ احتمالاً غیر طبیعی است و ۵ قطعاً غیر طبیعی است. وضعیت واقعی بیماری برای ۵۱ نفر و برای ۵۸ بیمار غیرواقعی بود.

True disease status of subject	Freq.	Percent	Cum.
0	58	53.21	53.21
1	51	46.79	100.00
Total	109	100.00	

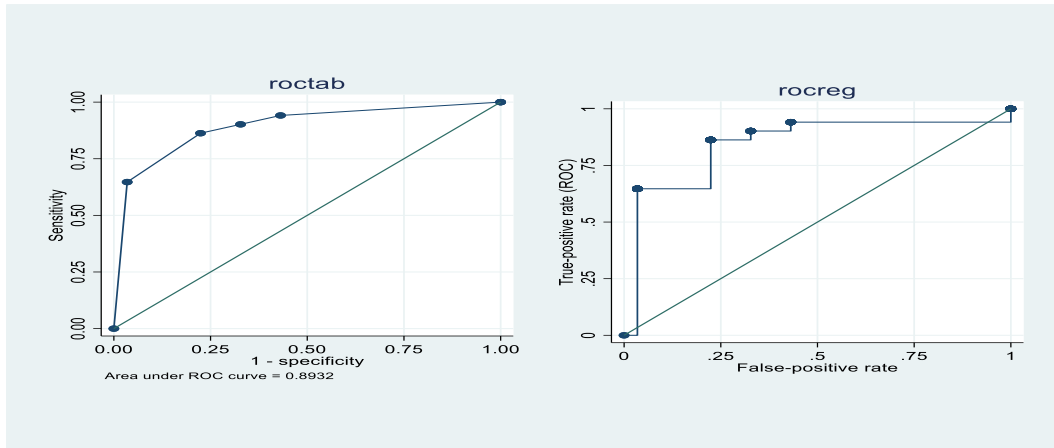
`. list disease rating in 1/9`

	disease	rating
1.	1	5
2.	0	1
3.	1	5
4.	0	4
5.	0	1
6.	0	3
7.	1	5
8.	0	5
9.	0	1

تحلیل منحنی با روش ناپارامتریک به صورت زیر است:

```
. roctab disease rating, graph aspectratio(1) name(a) nodraw title("roctab")
. quietly rocreg disease rating
. rocregplot, name(b) nodraw legend(off) title("rocreg")
. graph combine a b
```

منحنی های زیر را ببینید.



هر دو دستور `roctab` و `rocreg` میزان مثبت کاذب و مقادیر ROC را محاسبه می کنند. سبک اتصال خط پلکان نمودار سمت راست بر ماهیت تجربی تخمین های آن تأکید می کند. به طور پیش فرض، `roctab` مساحت زیر منحنی ROC (AUC) را با استفاده از تقریب ذوزنقه ای به برآورد میزان مثبت کاذب و میزان مثبت واقعی محاسبه می کند. دستور `roctab disease rating, summary` را اجرا نمائید.

`. roctab disease rating, summary`

Obs	ROC area	Std. err.	Asymptotic normal [95% conf. interval]	
109	0.8932	0.0307	0.83295	0.95339

سطح زیر منحنی را با دستور `(rocreg disease rating, tiecorrected bseed (29092))` با بوت استرپ برآورد نمائید.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. rocreg disease rating, tiecorrected bseed(29092)
(running rocregstat on estimation sample)

Bootstrap replications (1,000)
-----|-----|-----|-----|-----|-----|
      1       2       3       4       5
.....
      50
.....
     100
.....
     150
.....
     200
.....
     250
.....
     300
.....
     350
.....
     400
.....
     450
.....
     500
.....
     550
.....
     600
.....
     650
.....
     700
.....
     750
.....
     800
.....
     850
.....
     900
.....
     950
.....
    1,000

Bootstrap results                                Number of obs =   109
                                                Replications = 1,000

Nonparametric ROC estimation

Control standardization: empirical, corrected for ties
ROC method               : empirical

Area under the ROC curve

Status   : disease
Classifier: rating

-----+-----
      AUC      | Observed      | Bias      | Bootstrap      | [95% conf. interval]
                | coefficient    |           | std. err.      |
-----+-----+-----+-----+-----+-----+
                | .8931711      | .0010376  | .0309808       | .8324498 .9538923 (N)
                |               |           |                 | .8223829 .9475383 (P)
                |               |           |                 | .8084577 .9435818 (BC)
```

### محاسبه حدود اطمینان در آزمون های تشخیصی:

$$Var(Se) = \frac{Se(1-Se)}{n_1} = \frac{a \times c}{n_1^3} \quad \text{فرمول واریانس برای حساسیت:}$$

$$Var(Sp) = \frac{b \times d}{n_0^3} \quad \text{فرمول واریانس برای ویژگی:}$$

$$A \pm Z_{1-\frac{\alpha}{2}} \sqrt{Var(A)} \quad \text{قاعده کلی برای فاصله اطمینان:}$$

مثلا فاصله اطمینان ۹۵٪ برای حساسیت عبارتست از:  $Se \pm 1/96 \times \sqrt{Var(Se)}$  از فرمول زیر نیز می توان فاصله اطمینان را برای حساسیت محاسبه نمود:

$$CI_{(Sensitivity)} = Sensitivity \pm Z_{1-\alpha/2} \sqrt{\frac{Sensitivity(1-Sensitivity)}{N}}$$

از فرمول زیر نیز می توان فاصله اطمینان را برای ویژگی محاسبه نمود:

$$CI_{(Specificity)} = Specificity \pm Z_{\alpha/2} \sqrt{\frac{Specificity(1-Specificity)}{N}}$$

با فرمول های زیر می توان ارزش اخباری مثبت و فاصله اطمینان را برای آن محاسبه نمود. ارزش اخباری مثبت از فرمول  $PPV = \frac{TP}{TP + FP}$  محاسبه می شود. نسبت تعداد بیماران دارای تست مثبت به کل افراد دارای تست مثبت است.

$$PPV = \frac{Sensitivity \times prevalence}{(sensitivity \times prevalence) + (1 - specificity)(1 - prevalence)}$$

$$CI_{(PPV)} = PPV \pm Z_{\alpha/2} \sqrt{\frac{PPV(1-PPV)}{N}}$$

فرمول برآورد واریانس برای نسبت درست‌نمایی: چون توزیع LR دارای چولگی می باشد (مستقر نیست)، اما لگاریتم LR از توزیع طبیعی پیروی می کند، برآورد واریانس لگاریتم LR(+) شامل:

$$Var(\ln(LR(+))) = \frac{1-Se}{a} + \frac{Sp}{b}$$

فرمول حدود اطمینان برای LR(+):

$$\exp\left(\ln\left(\frac{Se}{1-SP}\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-Se}{a} + \frac{Sp}{b}}\right)$$

فرمول واریانس برای لگاریتم LR(-):

$$Var(\ln(LR(-))) = \frac{Se}{c} + \frac{1-Sp}{d}$$

فرمول حدود اطمینان برای LR(-):

$$\exp\left(\ln\left(\frac{1-Se}{SP}\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{Se}{c} + \frac{1-Sp}{d}}\right)$$

ارزش اخباری به مقدار حساسیت، ویژگی و شیوع بیماری وابسته است. ارتباط بین ارزش اخباری مثبت با حساسیت، ویژگی و شیوع بیماری در فرمول زیر گزارش شد.

$$PPV = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} = \frac{1}{1 + \frac{\text{False positives}}{\text{True positives}}}$$

$$PPV = \frac{1}{1 + \frac{(1 - specificity)(1 - prevalence)}{sensitivity \times prevalence}}$$

PPV با نسبت مثبت کاذب به مثبت واقعی، نسبت عکس دارد. هرچه این نسبت کمتر باشد PPV بزرگتر خواهد بود. اگر بیماری نادر و شیوع کمی داشته باشد جزء (1-prevalence) به ۱ نزدیکتر خواهد شد. در این صورت حتی اگر حساسیت برابر ۱۰۰ درصد فرض شود، بازهم نسبت مثبت کاذب به مثبت واقعی کمتر از نسبت  $\frac{1 - Specificity}{prevalence}$  نخواهد شد. لذا با شیوع کم، حتی اگر مثبت کاذب خیلی کم باشد (مثلاً ۱٪)، بازهم ارزش اخباری مثبت کم خواهد بود (مثبت کاذب بالا) و تاثیر ویژگی بر ارزش اخباری بسیار بیشتر از حساسیت است. البته این در شرایطی است که شیوع

بیماری بالا نباشد. وقتی شیوع بالا نیست، تعداد بیشتری از افراد در طرف راست جدول قرار میگیرند که این طرف روی ویژگی تاثیر دارد. بنابراین ارزش اخباری یک تست تحت تاثیر آن جمعیتی است که تست در آنها انجام میشود. با فرمول های زیر می توان ارزش اخباری منفی و فاصله اطمینان را برای آن محاسبه نمود.

$$NPV = \frac{(1 - Prevalence)(Specificity)}{(1 - prevalence)(specificity) + (prevalence)(1 - sensitivity)}$$

$$CI_{(NPV)} = NPV \pm Z_{\alpha/2} \sqrt{\frac{NPV(1-NPV)}{N}}$$

فرمول کلی نسبت شانس:

$$OR = \frac{\frac{Se}{1-Se}}{\frac{1-Sp}{Sp}} = \frac{Se \times Sp}{FNR \times FPR} = \frac{a \times d}{b \times c}$$

برآورد واریانس برای نسبت شانس:

$$Var(OR) = OR \left( \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)$$

$$Var(\log(OR)) = \left( \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)$$

مانند نسبت درستنمایی، توزیع نسبت شانس هم چوله می باشد:

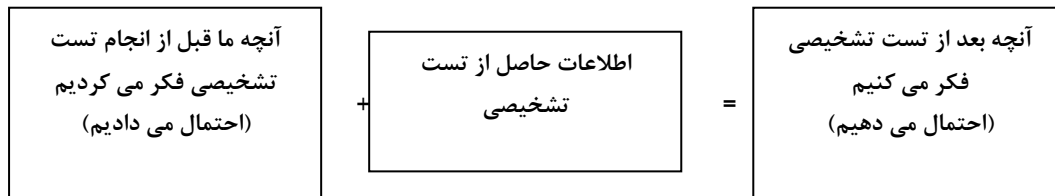
#### چند نکته مهم در تفسیر آزمون های تشخیصی:

- تستهایی که حساسیت بالا دارند، منفی شان ارزش دارد و با اطمینان بالا بیماری را رد می کنند.
  - تستهایی که ویژگی بالایی دارند، مثبت شان ارزش بالایی دارد و در صورت مثبت شدن احتمال وجود بیماری بالاست.
  - در بیماران با شدت بیماری بیشتر یا کمتر یا در مراحل مختلف بیماری، حساسیت یا ویژگی یک آزمون ممکن است بالاتر یا پایین تر باشد.
  - از شاخص یودن و استفاده همزمان از حساسیت و ویژگی و هر چه به ۱ نزدیک تر باشد مطلوب تر است.
- Youden's Statistic = (sensitivity + specificity) - 1** حدود اطمینان این شاخص از رابطه زیر قابل محاسبه است.

$$CI_{(Youden)} = Youdens \pm Z_{\alpha/2} \sqrt{\frac{Sensitivity(1-Sensitivity)}{TP} + \frac{Specificity(1-Specificity)}{TN}}$$

- ارزش اخباری Predictive Value اگرچه تحت تاثیر ویژگی و حساسیت آزمون می باشد، تحت تاثیر میزان شیوع بیماری نیز می باشد. هر چه شیوع بیماری در جمعیتی به صفر نزدیکتر شود، ارزش اخباری مثبت آزمون به صفر نزدیکتر می شود. هر چه شیوع بیماری در جمعیتی به صددرصد نزدیکتر شود، ارزش اخباری منفی آزمون به صفر نزدیکتر می شود.

- مهمترین علت انجام تستهای تشخیصی در واقع تصحیح احتمال ابتلا بیمار به بیماریهای مورد نظر (در لیست تشخیصهای افتراقی) بر اساس نتیجه تست تشخیصی است. نتیجه تست تشخیصی، می تواند احتمال وجود بیماری را تغییر دهد. بدیهی است:



یکی از مهمترین روشهایی که برای تعیین اینکه یک تست تشخیصی چقدر احتمال یک بیماری را کم یا زیاد می کند، Likelihood Ratio (LR) است. چگونه pretest probability را تعیین کنیم؟  
برای این کار دو راه داریم:

۱. ارزیابی اطلاعات کلینیکال و یا تجربه بالینی
۲. بر اساس آمار موجود از شیوع بیماری در جمعیت مورد نظر.

$$LR = \frac{LR \text{ positive}}{LR \text{ negative}}$$

$$LR \text{ positive} = \frac{\text{احتمال مثبت بودن نتیجه تست تشخیصی در یک فرد بیمار}}{\text{احتمال منفی بودن نتیجه تست تشخیصی در یک فرد غیر بیمار}} = \frac{a / a+c}{b / b+d}$$

$$LR \text{ negative} = \frac{\text{احتمال منفی بودن نتیجه تست تشخیصی در یک فرد بیمار}}{\text{احتمال منفی بودن نتیجه تست تشخیصی در یک فرد غیر بیمار}} = \frac{c / a+c}{d / b+d}$$

با دانستن حساسیت و ویژگی هم می توان LR را محاسبه کرد.

$$LR + = \text{sensitivity} / (1 - \text{specificity})$$

$$LR - = (1 - \text{sensitivity}) / \text{specificity}$$

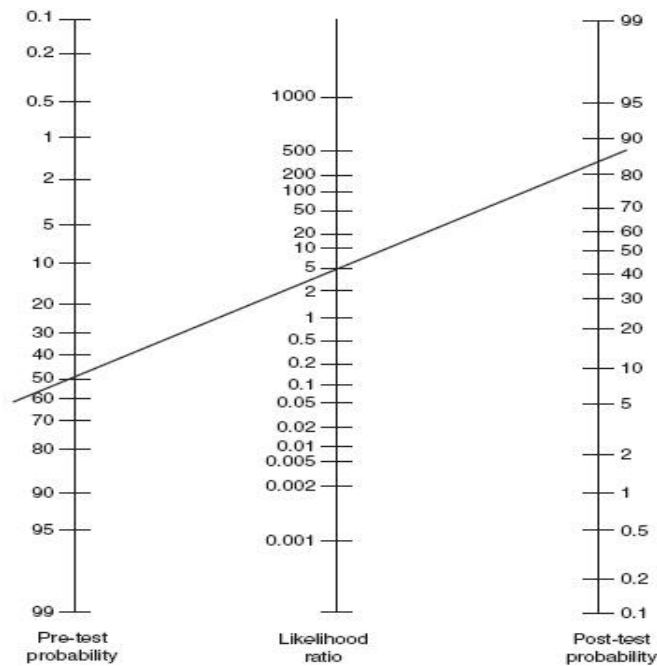
برای محاسبه احتمال وجود بیماری بعد از نتیجه تست تشخیصی از فرمولهای زیر استفاده می شود.

$$\text{Posttest Odds} = \text{Pretest Odds} \times LR$$

$$\text{Odds} = \text{Probability} / (1 - \text{Probability})$$

$$\text{Probability} = \text{Odds} / (1 + \text{Odds})$$

البته راه بسیار ساده تری هم برای محاسبه این فرمولها وجود دارد و آن استفاده از نانوگرام (شکل زیر) است.



در صورتی که چندین تست تشخیصی را بخواهید انجام بدهید، باید از فرمول زیر استفاده کنید  

$$\text{Post test Odds} = \text{Pretest Odds} \times \text{LR}_1 \times \text{LR}_2 \times \text{LR}_3 \times \dots \times \text{LR}_n$$
 مزیت استفاده از likelihood ratio در تشخیص کلینیکال بقرار زیر است:

- در این روش می توان در تستهایی که بیش از دو جواب دارند (نه فقط مثبت یا منفی) به راحتی استفاده کرد.
- این روش مشخص می کند که تست مورد نظر دقیقاً به چه میزان احتمال بیماری را زیاد یا کم کرده است.
- به راحتی در تستهای متوالی کاربرد دارد. مفهوم ابتلا به بیماری با در صد بیان می شود که ملموس تر است.

### آزمون های چند تایی: تستهای تشخیصی متعدد را می توان به دو صورت انجام داد

۱- موازی (Parallel): بدین معنی است که پزشک چند تست تشخیصی را همزمان درخواست می کند و جواب مثبت در هر کدام از تستها، شاهدهی مبنی بر وجود بیماری است. خصوصیات تستهای موازی در زیر آمده است.

باعث صرفه جوئی در وقت پزشک می شود. باعث افزایش حساسیت در تشخیص بیماری می شوند. باعث کاهش ویژگی و PPV در تشخیص بیماری می شوند.

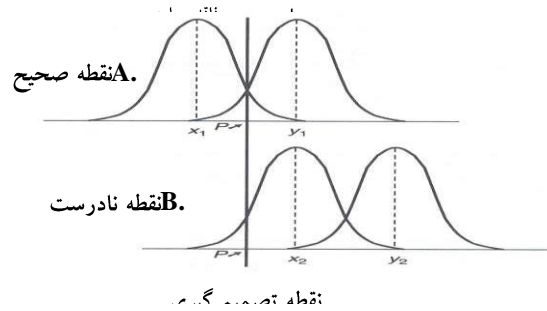
۲- متوالی (Serial): در این صورت هر تست بعد از در اختیار داشتن جواب تست قبلی درخواست می شود، چون تصمیم گیری در مورد تست بعد به تست قبلی ارتباط دارد. در تستهای متوالی، تمام تستها باید جواب مثبت بدهند، تا تشخیص بیماری گذاشته شود.

### خصوصیات تستهای متوالی:

وقت بیشتری می برند. می توان تستهای گرانتر و پر خطر تر را بعد از تستهای ساده تر و ارزانتر انجام داد. ، باعث افزایش ویژگی و PPV در تشخیص بیماری می شوند. باعث کاهش حساسیت و NPV در تشخیص بیماری می گردند.

### تعیین نقطه تصمیم گیری یا مثبت شدن تست تشخیصی (Cutoff Point)

همانطور که گفته شد، نتیجه اکثر تستها مقادیر عددی پیوسته است، نه فقط یک جواب ساده مثبت یا منفی. بنابراین، باید تعیین شود که تا چه مقداری طبیعی است و از چه مقدار غیر طبیعی است. این نقطه تصمیم گیری (Cutoff Point) یک تصمیم اختیاری است که در تعیین آن، برقراری توازن بین حساسیت و ویژگی لازم است. این توضیحات در تصویر زیر نمایان است.

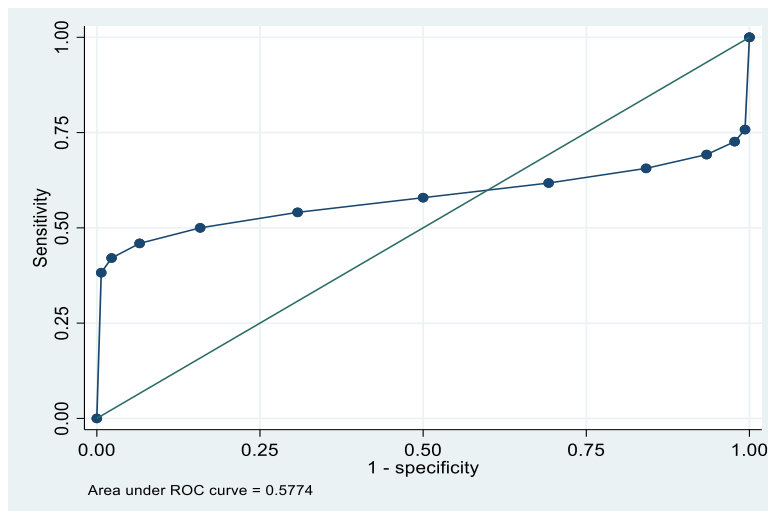


### کار با نرم افزار و ارزیابی تست های تشخیصی

فایل `roc1400` با نام `roc1400` را اجرا نمائید. این فایل با نام `roc1400` در `http://www.stata-press.com/data/r13/lorenz` قابل استفاده است. دستور `roctab disease class [fweight=pop], graph summary` در فایل دیتای کتاب قابل استفاده است. دستور `roctab disease class [fweight=pop], graph summary` سطح زیر منحنی محاسبه می شود.

`roctab disease class [fweight=pop], graph summary`

Obs	ROC area	Std. err.	Asymptotic normal [95% conf. interval]	
884	0.5774	0.0215	0.53517	0.61959





برای اکثر تستهای تشخیصی، نقطه تصمیم گیری جایی تعیین می شود که تست ویژگی و حساسیت بالایی داشته باشد برای این کار می توان از منحنی ROC (Receiver Operating Characteristic) استفاده کرد. منحنی ROC، تغییرات حساسیت و ویژگی آزمون را بسته به محل انتخاب نقطه تصمیم گیری نشان می دهد. تغییر محل نقطه تصمیم گیری باعث می شود که حساسیت افزایش یابد، اما ویژگی کاهش یابد یا بالعکس. می توان با استفاده از تغییر نقطه تصمیم گیری، و ترسیم منحنی، بهترین کارایی را برای یک تست تشخیصی بدست آورد. کار با نرم افزار و محاسبه شاخص های حساسیت و ویژگی در بخش چهارم کتاب و در قسمت رگرسیون لوجستیک و ترسیم منحنی ROC و مقایسه منحنی های ROC توضیح داده شد.

## خلاصه و نتیجه گیری:

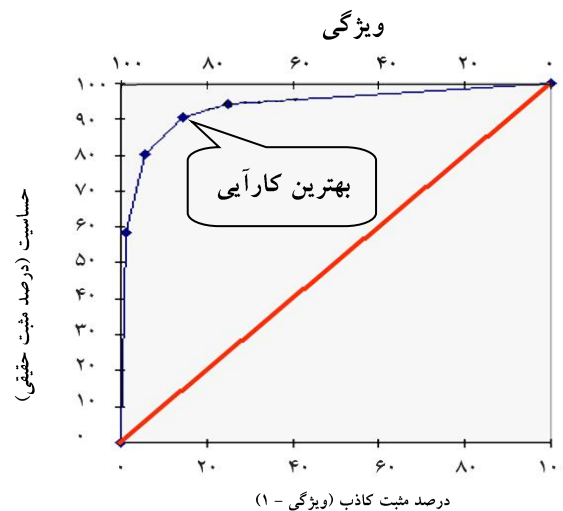
در این بخش به انواع مطالعات اپیدمیولوژیک و مفاهیم مرتبط و شاخص های اثر و ارتباط پرداخته شد. محاسبه شاخص ها با مثال کاربردی با نرم افزار انجام گردید. درک و تفسیر این شاخص ها مهمترین هدف در این بخش است. همچنین در این بخش کاربرد فرمول های رایج و نحوه حدود اطمینان آنها بحث گردید. فرمول های زیادی در اپیدمیولوژی کاربرد دارد، علاوه بر توضیح فرمول های مهم در متن، سایر فرمول های رایج و عمومی نیز در ضمیمه شماره سه کتاب آمده است. به ارزیابی و کاربرد آزمون های تشخیصی و منحنی ROC (Receiver Operating Characteristic) پرداخته شد. سطح زیر منحنی را محاسبه و این منحنی در زیر ترسیم شده است. دستورات مهم در اندازه گیری شاخص های اعتبار عبارتند از:

.diagt diagvar testvar [weight] [if exp] [in range] [, prev(#) level(#)

.diagti #a #b #c #d, prev(#) level(#)

در دستور بالا، شیوع باید بین صفر تا ۱۰۰٪ و حدود اطمینان مورد نظر را وارد نمود. اعداد a تا d نیز اعداد

سلول های جدول دو در دوی برای ارزیابی آزمون تشخیصی است.



## عدالت در سلامت و اندازه گیری نابرابری در سلامت:

سازمان جهانی بهداشت (WHO) از سال ۱۹۹۵ توجه دولت ها و NGO ها را به مقوله برابری (Equality) و عدالت (Equity) در سلامت معطوف نمود. به هر گونه اختلاف، تفاوت و گوناگونی غیرضروری، قابل اجتناب و غیر منصفانه و غیرعدالانه در شاخص های اندازه گیری سلامت و عوامل خطر آن در جوامع مختلف، نابرابری در سلامت گفته میشود.

نابرابری در سلامت اغلب در گروههای اجتماعی اقتصادی، و بویژه در جوامع با درآمد کم و مابین لایه های یک جامعه مشاهده میشود. یکی از مباحث مهم در اپیدمیولوژی اجتماعی پرداختن به عدالت در سلامت و اندازه گیری نابرابری در سلامت است.

شاخصهای بهداشتی هر زیرگروه، مانند میزان ها یا فراوانی ها را با جدول یا نمودار نمایش و مورد ارزیابی قرار میدهند. این شاخص ها در فصل های ابتدایی کتاب مورد بحث قرار گرفتند. این کار به فهم کلی از موقعیت زیرگروه های جمعیتی کمک خواهد کرد. جداول و نمودار های ساده اختلاف و نابرابری های بهداشتی را نشان می دهند اما همه جنبه های نابرابری را نمایان نمی کنند.

از جمله مشکلات استفاده از اندازه هائی که متغیر های دو حالت را می سنجند این است که در صورتی که زیرگروههای اجتماعی بیش از دو حالت باشند، اطلاعات مربوط به دیگر گروه ها نادیده گرفته می شود. برای رفع این مشکل در اندازه گیری نابرابری ها، از سنجه هایی استفاده میشود که اطلاعات تمام گروه ها را در بر گرفته و همزمان ضمن مقایسه همه ی زیرگروه ها، آنها را بصورت یک اندازه خلاصه برای تمام زیرگروه ها ارائه داد.

این رویکرد با استفاده از اندازه های نابرابری مبتنی بر رگرسیون (Slope Index of Inequalities-SII) و (Relative Index of Inequality-RII) بدست می آید.

برای سنجش نابرابری، ابتدا باید مشخص شود که دو گروه و یا بیشتر از دو گروه را می خواهیم مقایسه کنیم. اگر بخواهیم دو گروه را با یکدیگر مقایسه کنیم از اندازه های مطلق و نسبی ساده استفاده میشود.

اما اگر زیرگروه هایی که می خواهیم استفاده کنیم بیشتر از دو مورد باشد و دارای یک ترتیب طبیعی باشند (مثل تحصیلات و درآمد) در این صورت اندازه های (Absolute concentration Index- ACI) و یا

(Slope Index of Inequalities- SII) و (Relative Index of Inequality –RII) استفاده میشوند. برای

محاسبه SII، ابتدا در محور X ها طبقات اقتصادی اجتماعی جامعه را مثلا بر حسب میزان تحصیلات یا میزان محرومیت یا میزان درآمد به ترتیب از پایین ترین تا بالاترین گروه مرتب می کنیم و در محور Y ها پیامد یا متغیر وابسته مربوط به سلامت (مثلا شیوع استعمال دخانیات، شیوع چاقی یا هر شاخص دیگری) قرار میگیرد.

سپس جمعیت حاضر در هر گروه را بر اساس نسبت آن از کل جمعیت مشخص کرده و بر مبنای نقطه میانی دامنه خود در میزان توزیع کلی جمعیت، مقدار آن را مشخص می کنیم، سپس فراوانی نسبی عامل خطر را در هر گروه بر روی نقطه میانی مشخص شده هر گروه مشخص می کنیم. به این ترتیب در نقطه میانی توزیع هر گروه نسبت عامل خطر مورد مطالعه مشخص می شود سپس یک خط رگرسیون بر روی این نقاط رسم کرده و معادله خط رگرسیونی رسم شده را به صورت  $\bar{y}_i = \beta_0 + \beta_1 \bar{R}_j$  محاسبه می کنیم.

در این حالت  $\bar{y}_i \bar{y}_i$  وضعیت فراوانی نسبی یا میزان عامل خطر یا شاخص سلامت مورد مطالعه،  $\bar{R}_j \bar{R}_j$  مقدار میانگین مشخص شده هر گروه  $\beta_0 \beta_0$  فراوانی نسبی عامل خطر برای اولین فرد در اولین گروه خواهد بود. در این حالت  $\beta_i \beta_i$  اختلاف فراوانی نسبی عامل خطر در اولین فرد در اولین گروه و آخرین فرد در بالاترین گروه خواهد بود.

در معادله خط رگرسیونی رسم شده  $\beta_i \beta_i$  در فرمول بالا، معادل SII است که یکی از اندازه های نابرابری مبتنی بر رگرسیون است.

این اندازه ها از تمام اطلاعات موجود در همه گروه های اجتماعی استفاده می کنند و نسبت به جهت شیب اجتماعی حساس هستند. این اندازه ها تغییرات در توزیع جمعیت ابتدایی و نهایی را در همه گروه های اجتماعی در طول زمان منظور می کنند. در هنگام مقایسه گروه های اجتماعی که ترتیب مرتبی ندارند و یا هنگامی که بیشتر از دو گروه را با همدیگر مقایسه کنیم، اندازه هایی مثل واریانس بین گروهی، شاخص انتروپی، ضریب Henri Theil توصیه می شود.

برای رهایی از مشکلات مربوط به اندازه های مطلق (Absolute) با تقسیم SII بر میانگین فراوانی نسبی عامل خطر در کل جمعیت، یک اندازه نسبی (Relative) به دست می آید که در کنار شاخص قبلی خاصیت های شاخص نسبی (RII)، برای بیان نابرابری را دارد.

تفسیر SII به این صورت است که این مقدار نشان دهنده تفاوت نسبت عامل خطر در اولین و آخرین فرد گروه است. SII یک شاخص مطلق است. اگرچه استفاده از این اندازه برای تعیین نابرابری نیازمند این پیش فرض است که ارتباط بین گروه ها باید خطی باشد و استفاده از آن را تا حدودی محدود می کند. اما دو ویژگی بارز و مهم باعث می شود تا این شاخص مورد استفاده قرار گیرد. یکی اینکه چنانچه در طول زمان تغییرات جمعیتی در بین گروه های اجتماعی حاصل شود، این اندازه نسبت به آن ها حساس بوده و آن را نشان می دهد. و دیگری اینکه بیان و درک آنها ساده و قابل فهم است. لذا این فواید استفاده از آن را در سنجش نابرابری در سلامت امکانپذیر میکند.

شاخص تمرکز (Concentration Index-CI) یکی دیگر از شاخص هایی است که برای سنجش نابرابریهای اقتصادی اجتماعی در سلامت بکار می رود.

این شاخص شباهت های زیادی با ضریب جینی (Gini) دارد اما در محاسبه با آن تفاوت دارد. این شاخص بعنوان یکی از روش های اپیدمیولوژی اجتماعی برای تعیین نابرابریها در زیرگروههای اقتصادی اجتماعی جامعه کاربرد دارد. برای تعیین شاخص تمرکز (CI)، ابتدا باید (Concentration Curve - CC) رسم شود. برای رسم منحنی تمرکز (CC) به متغیر وابسته (پیامد سلامت) و متغیری که وضعیت و رتبه اقتصادی، اجتماعی فرد را در جامعه مورد مطالعه نشان دهد نیاز است.

برای رسم منحنی، درصد تجمعی وضعیت متغیر سلامت (فراوانی یا میزان عامل خطر) را در محور  $y$ ها و نسبت اقتصادی اجتماعی افراد از فقیرترین تا غنی ترین گروه ها، در محور  $x$ ها قرار میگیرد.

منحنی ترسیم شده بیانگر وضعیت نابرابری عامل خطر در گروه های اقتصادی و اجتماعی مختلف خواهد بود. برای محاسبه داده های گروه بندی شده از فرمول (Relative Concentration Index- RCI) استفاده میشود.

$$RCI = \frac{2}{\mu} \left[ \sum_{j=1}^J P_j \mu_j R_j \right] - 1$$

در این فرمول  $P_j P_j$  نسبت گروه جمعیتی،  $\mu_j \mu_j$  فراوانی نسبی یا میزان عامل خطر در کل جمعیت،  $R_j R_j$  برابر رتبه نسبی (Relative Rank) گروه زدر بین گروه های اقتصادی و اجتماعی است.  $R_j R_j$  به صورت زیر محاسبه می شود:

$$R_j = \sum_{j=1}^J P_j - 1/2 P_j$$

که  $P_j P_j$  سهم تجمعی جمعیت تا حد نهایی گروه و  $P_j P_j$  نشانگر رتبه نسبی تا حد میانی این گروه به صورت تجمعی می باشد.

اگر چه در معرفی و محاسبه شاخص تمرکز، به RCI بیشتر توجه می شود، اما منحنی تمرکز نیز قابل محاسبه است. منحنی تمرکز بر اساس متد های Covariance و Convenient Regression محاسبه میشود. در تفسیر این منحنی چنانچه منحنی در پایین خط Equity باشد عامل خطر مورد مطالعه بیشتر در بین افراد غنی تمرکز دارد و اگر منحنی در بالای خط Equity قرار گیرد عامل خطر مورد مطالعه بیشتر در بین افراد فقیر تمرکز دارد.

شاخص تمرکز به دو برابر ناحیه بین منحنی تمرکز و خط برابری گفته میشود. مقدار CI را می توان به دو صورت Relative CI (RCI) و Absolute CI (ACI) ارائه داد. معمولاً RCI کاربرد بیشتری دارد. مقادیر RCI بین منهای 1 تا مثبت 1 بوده و اگر CC بالای خط برابری باشد مقادیر CI بین صفر تا حداکثر منهای یک را در بر می گیرد و نشانه توزیع عامل خطر در بین فقرا است. اگر CC پایین خط برابری باشد مقادیر CI بین صفر تا حداکثر یک مثبت را در بر می گیرد و نشانه توزیع عامل خطر در بین اغنیاء است. این روش کاربرد بسیاری در اپیدمیولوژی اجتماعی دارد که در ایران کمتر از آن استفاده میشود. ترویج و استفاده از این شاخص معتبر، توصیه می شود.

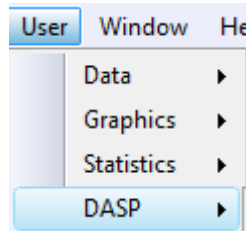
#### کار با نرم افزار و محاسبه مهمترین شاخص های نابرابری در سلامت:

برای استفاده از نرم افزار Stata و محاسبه شاخص های نابرابری باید برنامه نویسی نمود. اغلب کاربران آشنایی کافی به زبان برنامه نویسی در نرم افزار Stata را ندارند. به این منظور بسته برنامه ای به نام (DASP) که مخفف Distributive Analysis Package Stata است را می توان از اینترنت دانلود و بر نرم افزار Stata نصب و از آن استفاده نمود. برای تنظیمات این برنامه پس از دانلود بسته برنامه مذکور از دستورات زیر استفاده می شود.

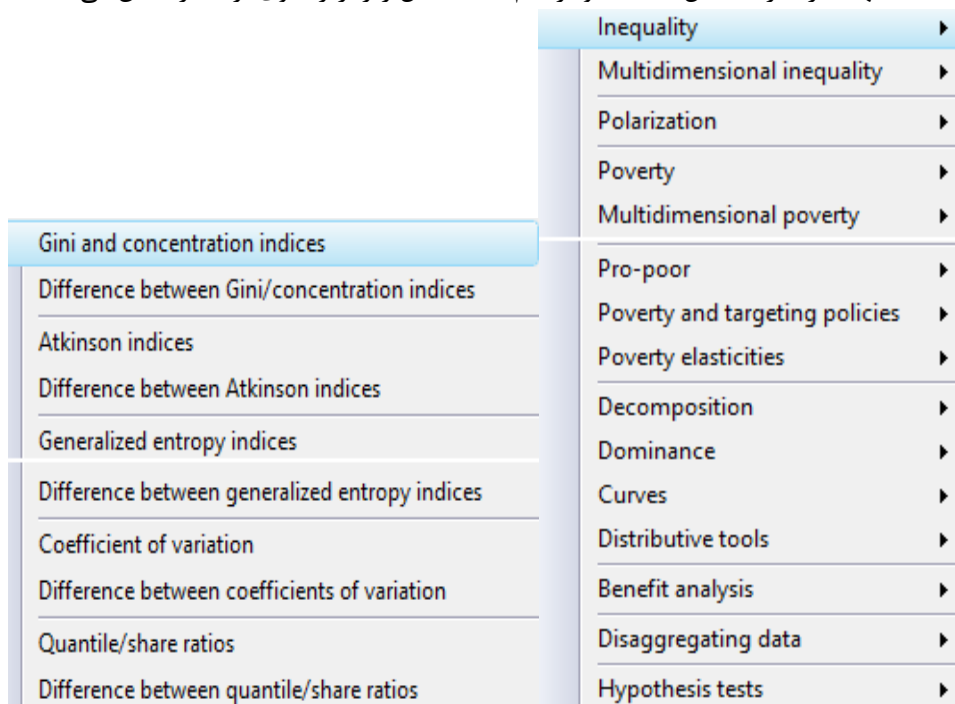
ادرس وب سایت: <http://dasp.ecn.ulaval.ca>

```
set more off
net from c:/dasp
net install dasp_p1, force
net install dasp_p2, force
net install dasp_p3, force
net install dasp_p4, force
```

شکل زیر اضافه شدن این برنامه (DASP) در زیر منوی اصلی User نرم افزار Stata را نشان می دهد.



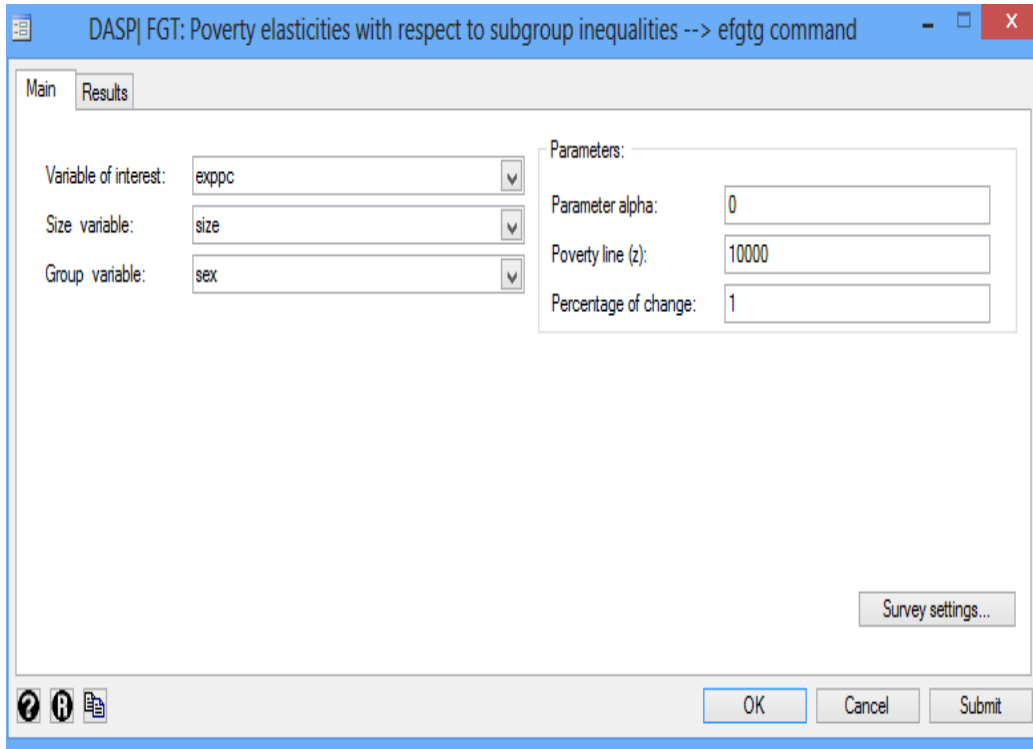
در منوی این برنامه اقدامات زیادی را برای اندازه گیری فقر، نابرابری و شاخص های مرتبط با نابرابری در سلامت و منحنی های مربوطه که در بالا به آنها اشاره گردید قابل محاسبه و ترسیم است. شکل زیر، زیر منوی برنامه را نشان می دهد.



فایل inequality را باز نمایید. برای اندازه گیری فقر از شاخص Foster- Greer- Thorbecke یا (FGT) استفاده می شود. با دستور db efgtg و باز شدن کادر محاوره ای زیر، می توان متغیرها را به نرم افزار معرفی و شاخص های اندازه گیری فقر در جامعه را محاسبه و گزارش نمود. در کادر محاوره ای هر یک از شاخص های مذکور وارد کردن متغیرهای زیر مهم است:

- Variable of interest : متغیر یا متغیرهایی که شرایط زندگی استاندارد و وضعیت موجود را در جامعه مورد بررسی نشان می دهند ، وارد این کادر می شوند. مانند درآمد سرانه، دریافت کالری، قد نرمال برای سن در کودکان، ثروت و دارایی افراد
- Size variable : این متغیر اشاره به اقلیت ها و قومیت ها و اندازه فیزیکی و فراوانی مشاهدات مرتبط است.
- طبقه بندی شاخص های مورد درخواست را می توان بر حسب متغیرهای گروه بندی شده و طبقه ای مانند جنسیت انجام داد. بنابراین در قسمت مربوطه باید متغیر کیفی وارد شود.

- وزن نمونه مورد بررسی نیز مهم است که باید بر اساس عکس مقدار احتمال نمونه برداری انجام شده در جامعه وارد شود.
- کادر محاوره ای زیر که در اکثریت تحلیل ها وجود دارد، این متغیرها را نشان می دهد.



خروجی دستور محاسبه شاخص فقر (FGT) و ضریب جینی در جمعیت مورد بررسی در زیر آمده است. در نتایج زیر مقادیر شاخص های ضریب جینی و شاخص FGT برای اندازه گیری فقر در جامعه گزارش شده است.

```
. efgtg exppc, hgroup(sex) hsize(size) alpha(0) pline(10000) prc(1)
```

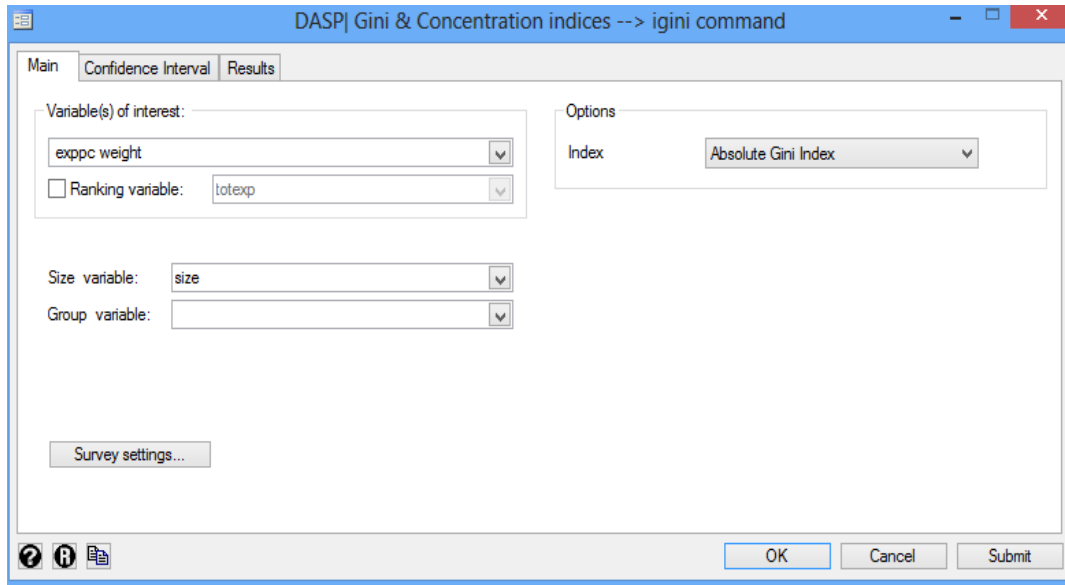
Poverty and inequality indices

Indices	Estimate
FGT	0.007173
Gini	0.493527

Marginal Impact & Elasticities By Groups

Group	Population Share	Marginal impact on ineq.	Marginal impact on pov.	Elasticity
1: Group_1	0.950787	0.004554	0.002624	39.646763
2: Group_2	0.049213	0.000316	0.000177	38.463089
Within	.	0.004871	0.002801	39.569851
Between	.	0.000049	0.000005	6.631795
Population	1.000000	0.004935	0.002862	39.893036

هر چه ضریب جینی به صفر نزدیکتر باشد یعنی نابرابری در جامعه کمتر است. یا فقر کمتر است. با دستور db igini نیز می توان ضریب جینی و شاخص تمرکز و منحنی آن را به صورت زیر محاسبه نمود. کادر محاوره ای زیر باز و متغیرها را می توان معرفی نمود.



نتایج در زیر آمده است.

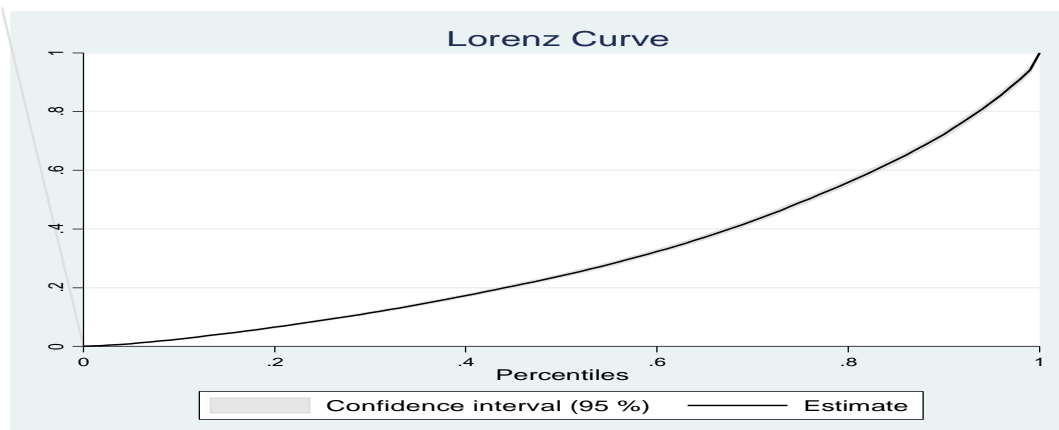
```
. igini exppc weight, hsize(size) dec(2) type(abs)
```

```
Index          : Gini index
Household size : size
```

Variable	Estimate	STE	LB	UB
1: A_GINI_exppc	40091.16	807.96	38507.37	41674.95
2: A_GINI_weight	54.47	1.16	52.20	56.75

### منحنی لورنز:

برای رسم منحنی لورنز با دستور db clorenzs می توان متغیرها را معرفی و منحنی را به شکل زیر ترسیم نمود.





برای محاسبه شاخص تمرکز و رسم منحنی تمرکز یا Concentration با دستور نیز می توان متغیرها را در کادر  
مجاوره ای زیر معرفی و منحنی را به شکل زیر ملاحظه نمود.

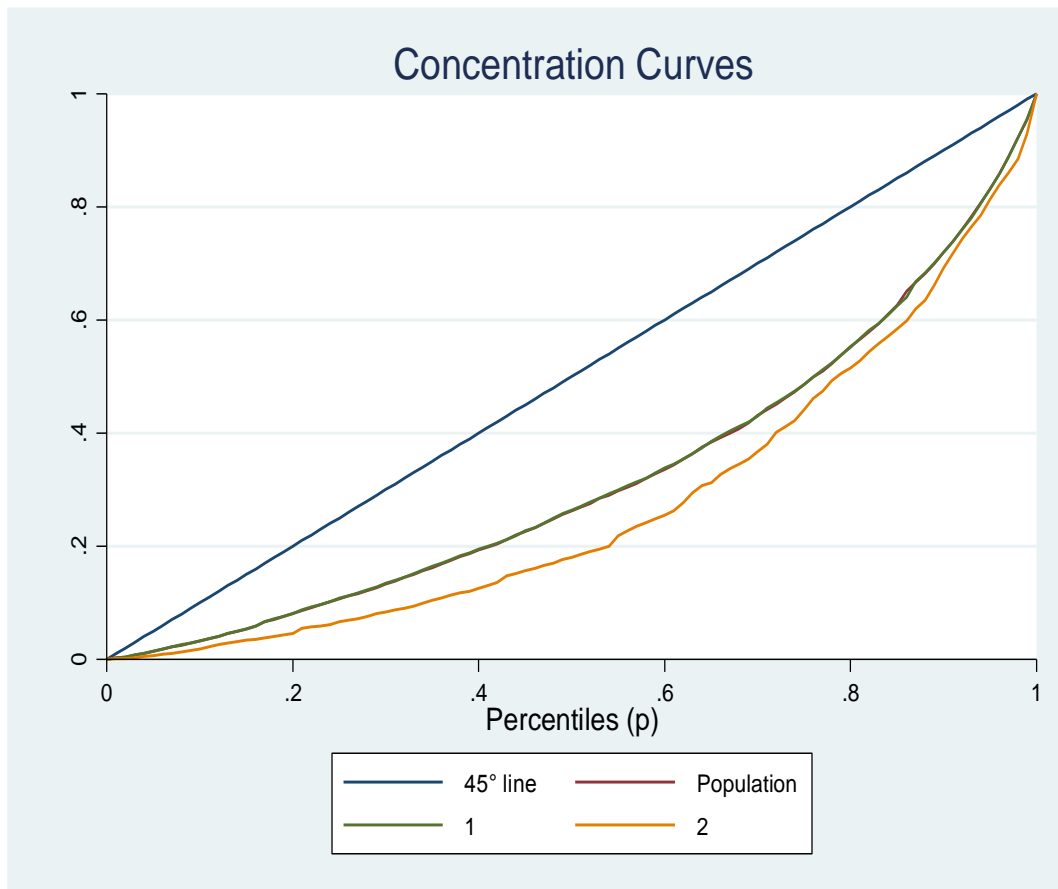
```
. igini weight, rank(totexp) hsize(size) hgroup(sex) dec(2)
```

```

Index          : Concentration index
Ranking variable : totexp
Household size  : size
Group variable  : sex
    
```

> -	Group	Estimate	STE	LB	UB
> -					
1: 1		-0.02	0.01	-0.03	-0.0
> 0					
2: 2		-0.06	0.02	-0.09	-0.0
> 3					
> -					
Population		-0.02	0.01	-0.03	-0.0
> 0					

نتایج بالا ضریب تمرکز، خطای معیار و حدود اطمینان آن را بر حسب متغیر گروه بندی شده (جنسیت) و در کل  
جامعه را نشان می دهد.



منحنی بالا بر حسب دو گروه مرد و زن ترسیم شده است. نحوه تفسیر منحنی مذکور در ابتدای مبحث شاخص های اندازه گیری نابرابری بیان شده است. لطفاً با مراجعه به آن مبحث منحنی بالا را تفسیر نمایید. برای محاسبه شاخص اتکینسون با دستور `db iatkinson` متغیرها را در کادر محاوره ای باز شده به نرم افزار معرفی می شوند. این شاخص به صورت زیر برای مردان، زنان و کل جامعه محاسبه شده است.

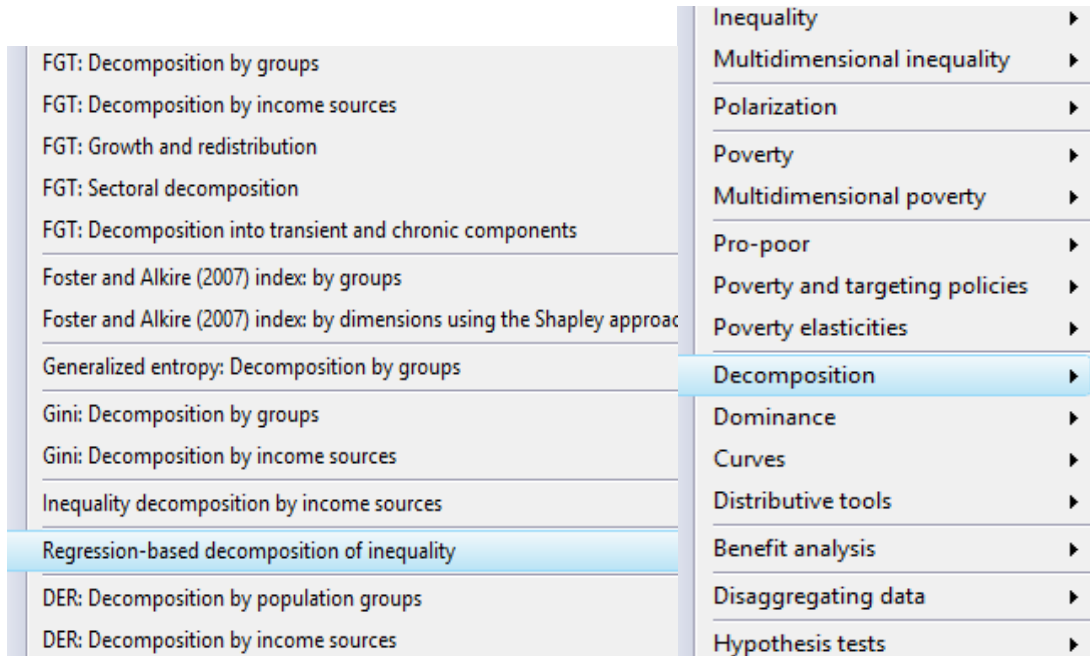
```
. iatkinson psu, epsilon(0.5) hsize(size) hgroup(sex)
```

```
Index          : Atkinson index
Parameter epsilon : .5
Household size  : size
Group variable  : sex
```

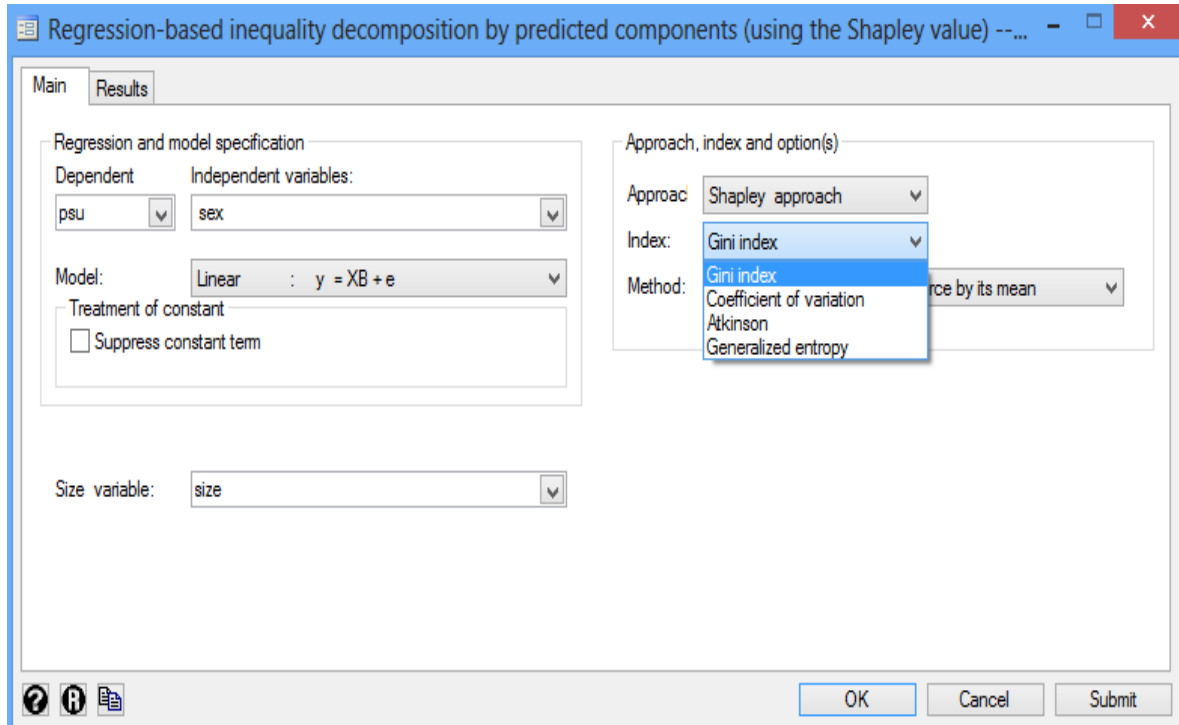
Group	Estimate	STE	LB	UB
1: 1	0.116561	0.002440	0.111777	0.121345
2: 2	0.078634	0.006517	0.065859	0.091408
Population	0.115148	0.002350	0.110542	0.119755

محاسبه شاخص های نابرابری مبتنی بر رگرسیون:

برای محاسبه نابرابری در بین گروه ها و استفاده از رگرسیون از منوی Decomposition گزینه های زیر انتخاب و اقدام می شود.



به این منظور با محاسبه نابرابری در بین زنان و مردان به عنوان متغیر مستقل یا پیش بین و متغیر psu به عنوان متغیر وابسته و معرفی متغیرهای زیر در کادر محاوره ای، نتایج زیر حاصل می شود.



نتایج تحلیل رگرسیونی انجام شده:

Regression-based inequality decomposition by predicted income components (using the SH > apley value).

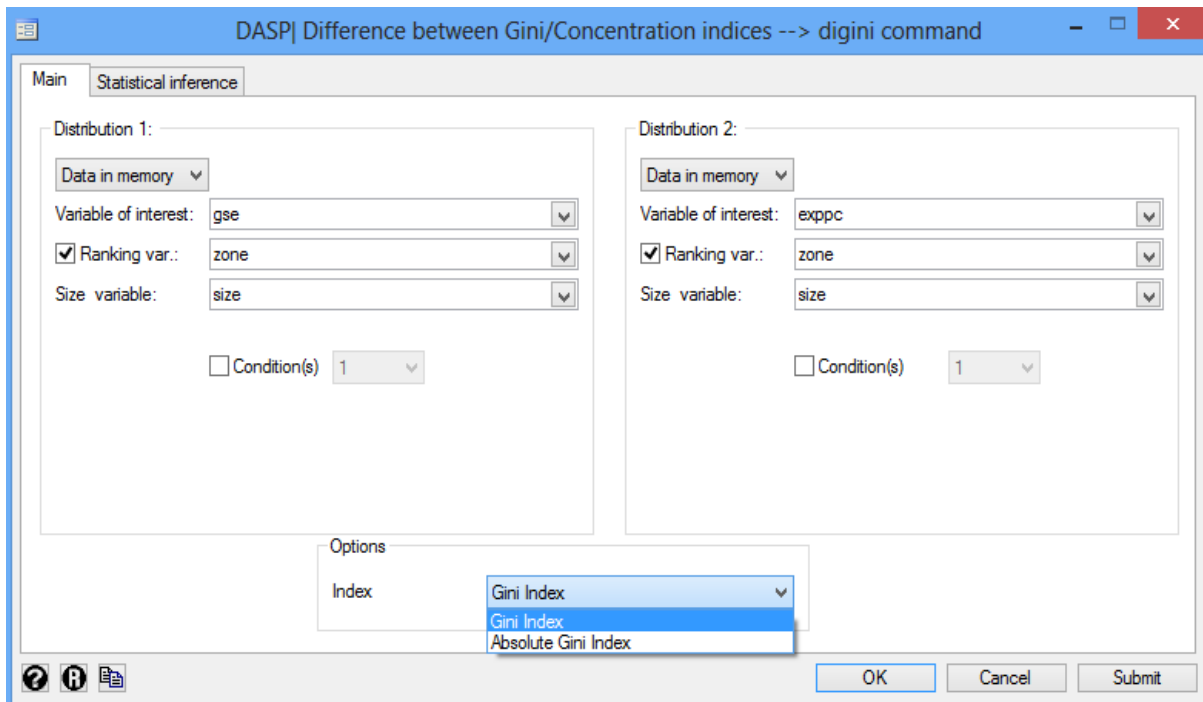
Execution time : 1.22 second(s)  
 Inequality index : Gini index  
 Estimated inequality : 0.338923  
 Household size : size

Sources	Income Share	Absolute Contribution	Relative Contribution
1: _p_cons	0.7210	-0.0000	-0.0000
2: _p_sex	0.2790	0.0124	0.0367
3: _p_resi	0.0000	0.3265	0.9633
Total	0.0000	0.3389	1.0000

Marginal contributions:

Source	level_1	level_2
1: _p_cons	0.000000	-0.000000
2: _p_sex	0.006222	0.006222

برای مقایسه میان ضریب جینی یا شاخص تمرکز در بین دو متغیر با معرفی متغیرها در کادر محاوره ای زیر می توان مقایسه را انجام داد.



## روش های اپیدمیولوژی و پژوهش با Stata

نتیجه مقایسه ضریب جینی برای دو متغیر در بین جامعه به صورت زیر حاصل می شود.

```
. digini gse exppc, rank1(zone) hsize1(size) rank2(zone) hsize2(size)
```

Index	Estimate	Std. Err.	t	P> t	[95% Conf. Interval]	
CONC_Dis1	-.0633351	.0025775	-24.5723	0.0000	-.0683876	-.0582826
CONC_Dis2	.1817702	.0064701	28.0939	0.0000	.1690873	.1944531
diff.	.2451053	.0077261	31.7243	0.0000	.2299603	.2602503

نتایج نشان می دهد که بین دو متغیر مورد بررسی، ضریب جینی و شاخص تمرکز معنی دار است. مقدار اختلاف و خطای معیار و مقدار معنی داری در جدول بالا آمده است. فایل مورد استفاده صرفاً برای آموزش و بکارگیری دستورات و محاسبه شاخص ها و ترسیم منحنی ها قابل استفاده می باشد. برای محاسبه دقیق و مقایسه شاخص ها و منحنی ها نیاز به داده های واقعی می باشد.

### کار با نرم افزار stata

۱- فایل bkf94.dta را باز نمائید. متغیرهای فایل را توصیف نمائید.

```
. des
```

```
Contains data from C:\Users\Tehran\Downloads\bkf94.dta
Observations:      8,625
Variables:          10                2 Nov 2006 15:50
```

Variable name	Storage type	Display format	Value label	Variable label
weight	float	%9.0g		
size	byte	%8.0g		
strata	byte	%8.0g		
psu	int	%8.0g		
gse	byte	%8.0g		
sex	byte	%8.0g		
zone	byte	%8.0g		
exppc	double	%10.0g		
totexp	double	%10.0g		
expeq	double	%10.0g		

۲- با دستور زیر متغیرهای `strata`, `psu`, and `weight` را به نرم افزار معرفی نمائید.

```
. svyset psu [pweight=weight], strata(strata) vce(linearized) singleunit(missing)
```

```
Sampling weights: weight
                  VCE: linearized
                  Single unit: missing
                  Strata 1: strata
Sampling unit 1: psu
FPC 1: <zero>
```

با دستور `db ifgt` می توانید متغیرها را معرفی کنید.

۳- با دستور `svydes` طبقه های نمونه گیری، واحدها و تعداد افراد را لیست کنید.

## روش های اپیدمیولوژی و پژوهش با Stata

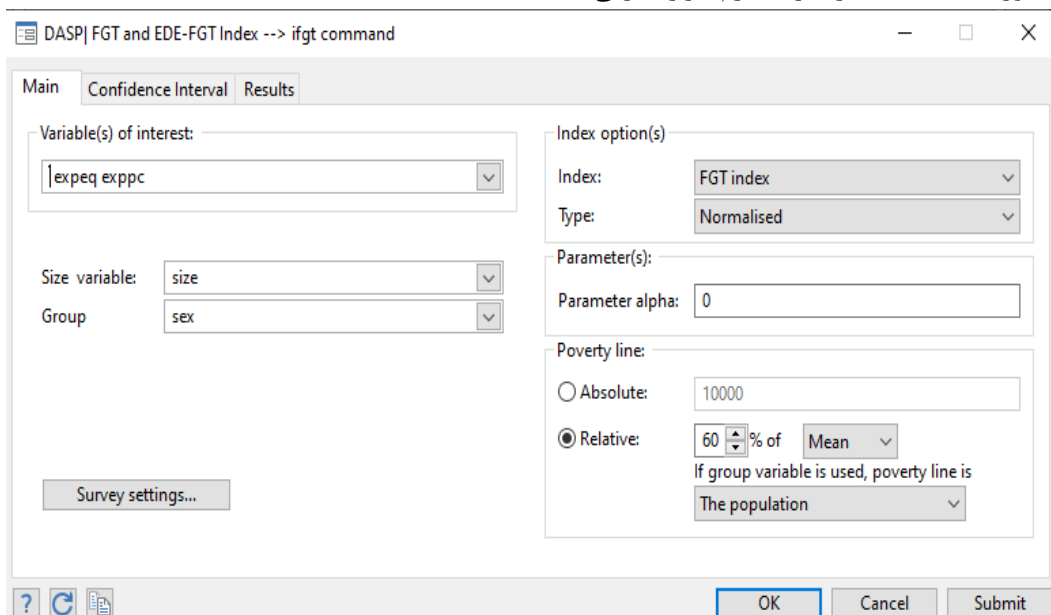
```
. svydes
```

```
Survey: Describing stage 1 sampling units
```

```
Sampling weights: weight
                   VCE: linearized
                   Single unit: missing
                   Strata 1: strata
Sampling unit 1: psu
                   FPC 1: <zero>
```

Stratum	# units	# obs	Number of obs per unit		
			Min	Mean	Max
1	42	838	19	20.0	20
2	37	733	17	19.8	20
3	98	1,959	19	20.0	20
4	55	1,093	19	19.9	20
5	66	1,286	13	19.5	20
6	41	779	1	19.0	20
7	97	1,937	19	20.0	20
7	436	8,625	1	19.8	20

۴- با دستور `db ifgt` متغیرها را به نرم افزار معرفی نمائید.



۵- برای متغیرهای `exped` و `expcc` شاخص FGT (Foster-Greer-Thorbecke indices) را برای اندازه گیری فقر در جامعه را محاسبه کنید.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. ifgt expeq exppc, alpha(0) pline(10000)
```

```
Poverty index : FGT index
Sampling weight : weight
Parameter alpha : 0.00
```

Variable	Estimate	STE	LB	UB	Pov. line
expeq	0.001216	0.000336	0.000556	0.001876	10000.00
exppc	0.002789	0.000598	0.001613	0.003965	10000.00

```
. ifgt expeq exppc, alpha(0) hsize(size) hgroup(strata) pline(10000)
```

```
Poverty index : FGT index
Household size : size
Sampling weight : weight
Group variable : strata
Parameter alpha : 0.00
```

Group	Estimate	STE	LB	UB	Pov. line
1: Group_1	0.002457	0.001874	-0.001225	0.006140	10000.00
2: Group_2	0.000000	0.000000	0.000000	0.000000	10000.00
3: Group_3	0.003449	0.002963	-0.002376	0.009273	10000.00
4: Group_4	0.000000	0.000000	0.000000	0.000000	10000.00
5: Group_5	0.020360	0.007104	0.006398	0.034323	10000.00
6: Group_6	0.000000	0.000000	0.000000	0.000000	10000.00
7: Group_7	0.000000	0.000000	0.000000	0.000000	10000.00
Population	0.002337	0.000861	0.000645	0.004029	10000.00

### خروجی محاسبه شاخص اندازه گیری فقر در زیر گروه ها

```
. ifgt expeq exppc, alpha(0) hsize(size) opl(mean) prop(60)
```

```
Poverty index : FGT index
Household size : size
Sampling weight : weight
Parameter alpha : 0.00
```

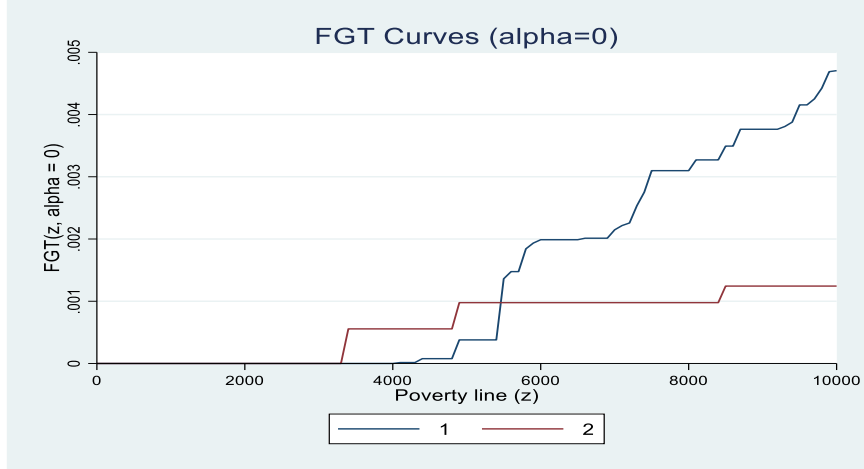
Variable	Estimate	STE	LB	UB	Pov. line
expeq	0.456067	0.012966	0.430582	0.481552	56686.21
exppc	0.484310	0.013377	0.458018	0.510602	43670.00

```
. ifgt expeq exppc, alpha(0) hsize(size) hgroup(sex) opl(mean) prop(60)
```

```
Poverty index : FGT index
Household size : size
Sampling weight : weight
Group variable : sex
Parameter alpha : 0.00
```

Group	Estimate	STE	LB	UB	Pov. line
1: Group_1	0.463948	0.013651	0.437117	0.490778	56686.21
2: Group_2	0.287582	0.027983	0.232582	0.342582	56686.21
Population	0.456067	0.012966	0.430582	0.481552	56686.21

cfgt exppc, alpha(0) hsize(size) hgroup(zone) min(0) max(10000)



### خلاصه و نتیجه گیری:

اندازه گیری نابرابری در سلامت یکی از تکنیک های مهم در اپیدمیولوژی اجتماعی است. به طور کلی روش های اندازه گیری نابرابری در سلامت به صورت زیر تقسیم و با نرم افزار می توان مهمترین آنها را محاسبه نمود. روش محاسبه ضریب تمرکز، ضریب جینی، شاخص اتکینسون، منحنی لورنز و منحنی تمرکز با نرم افزار و بیان دستورات و منوها تشریح گردید. دسته بندی شاخص های اندازه گیری نابرابری در سلامت در زیر آمده است.

۱- اندازه گیری نابرابری در کل جمعیت :

- اختلاف هر فرد با میانگین افراد جامعه (Individual-Mean Differences: (IMD)
- اختلاف میان افراد با همدیگر در جامعه (Inter-Individual Differences (IID)

۲- اندازه گیری نابرابری در گروه های اجتماعی و زیر گروه های جامعه :

- روش مقایسه های دو به دو
- روش مقایسه ای مبتنی بر رگرسیون :
- ( Slope Index of Inequality : SII)
- ( Relative Index of Inequality :RII )

۳- اندازه گیری بار نابرابری و متوسط عدم تناسب (Measures of Average Disproportionality)

- Concentration Index : CI
- Gini Index : GI

هر یک از شاخص های مذکور در ابتدای مبحث کار با نرم افزار تشریح شدند. دسته سوم شاخص ها کاربرد بیشتری دارند.



Difference – in- differences (ID) and DDD models

مدل های DID و DDD

وقتی داده های مطالعه مقطعی در چند نوبت تکرار شود و واحدهای مطالعه در زمان های مختلف تغییر نمایند از این نوع تحلیل و مدل سازی می توان استفاده کرد. در واقع داده ها با روی هم ریختن چند مطالعه مقطعی و به صورت پانلی و طولی هستند. برای برآورد متوسط اثرات مداخله یا درمان در گروه درمان شده بوسیله مقایسه میانگین پیامد در گروه کنترل و گروه مداخله انجام می شود. مثال: مطالعه ای برای بررسی اثر فرایند جدید پذیرش بیمارستانی بر رضایتمندی بیماران به صورت ماهیانه انجام می شود و داده ها جمع آوری می شوند. قبل از اجرای فرایند جدید نیز رضایت مندی بیماران در چند بیمارستان ارزیابی گردید.

کار با نرم افزار

فایل hospddd را باز کنید. متغیر satis رضایتمندی بیماران، متغیر درمان یا مداخله با نام procedure نامگذاری شدند.

Variable name	Storage type	Display format	Value label	Variable label
hospital	byte	%9.0g		Hospital ID
frequency	byte	%9.0g	size	Hospital visit frequency
month	byte	%8.0g	mnth	Month
procedure	byte	%9.0g	pol	Admission procedure
satis	float	%9.0g		Patient satisfaction score

. tab frequency procedure

Hospital visit frequency	Admission procedure		Total
	Old	New	
Low	1,612	444	2,056
Medium	1,392	392	1,784
High	1,240	272	1,512
Very high	1,592	424	2,016
Total	5,836	1,532	7,368

didregress (satis) (procedure), group(hospital) time(month)

دستور اجرای مدل

. didregress (satis) (procedure), group(hospital) time(month)

Number of groups and treatment time

Time variable: month  
Control: procedure = 0  
Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression  
Data type: Repeated cross-sectional

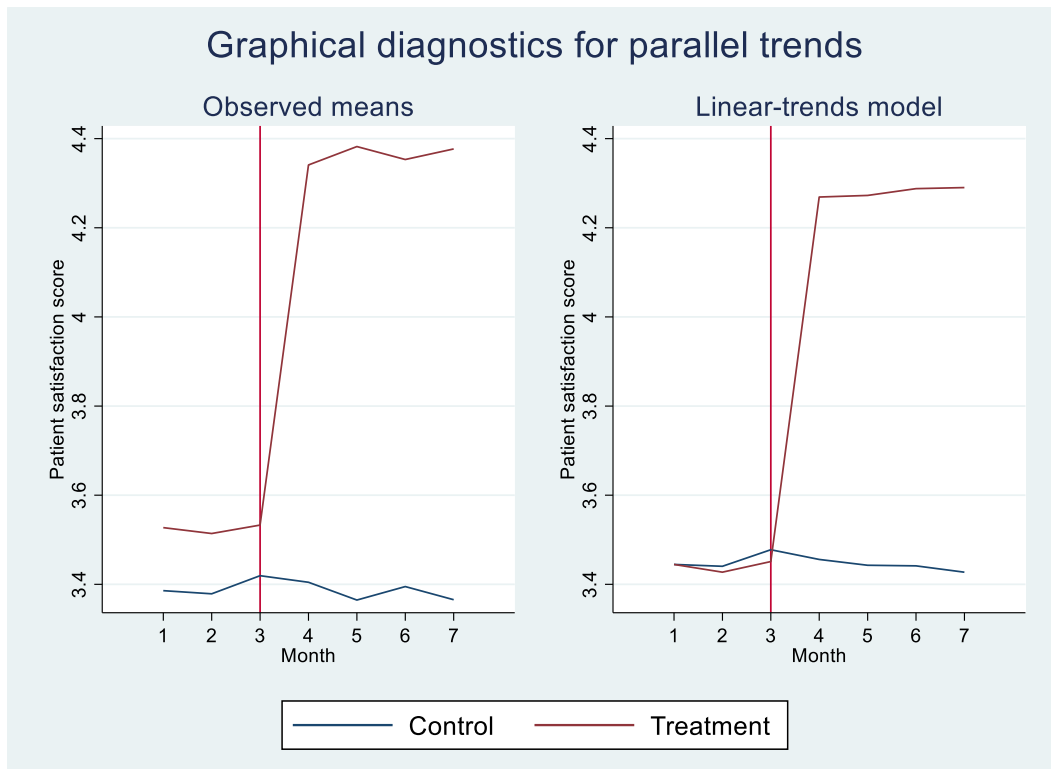
Number of obs = 7,368

(Std. err. adjusted for 46 clusters in hospital)

	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
ATET procedure (New vs Old)	.8479879	.0321121	26.41	0.000	.7833108 .912665

Note: ATET estimate adjusted for group effects and time effects.

نتایج مدل نشان می دهد: جدول اول اطلاعاتی در مورد گروه های کنترل و درمان (مداخله) و زمان مداخله ارائه می دهد. بخش اول به ما می گوید که ۲۸ بیمارستان به روش فرایند قبلی بیماران را پذیرش و ادامه دادند. ۱۸ بیمارستان فرایند پذیرش بیماران در بیمارستان را به فرایند جدید تغییر و اجرا نمودند. بخش دوم به ما می گوید که تمام بیمارستان هایی که روش جدید را اجرا کردند در دوره زمانی چهارم این کار را انجام دادند. اگر بعداً برخی بیمارستانها این سیاست را اتخاذ می کردند، حداقل و حداکثر زمان اولین درمان متفاوت بود. جدول دوم تخمین ATET (**average treatment effect on the treated**) یا متوسط اثر درمان در گروه مداخله را نشان می دهد. رضایت بیماران نسبت به بیمارستان هایی که فرایند روش جدید را اجرا نکرده بودند، ۰,۸۵ امتیاز افزایش یافته بود. یکی از مفروضات این مدل این است که مسیرهای متغیر رضایت مندی (satis) برای گروه های کنترل و درمان یا مداخله قبل از اجرای روش جدید با یکدیگر موازی باشند. یک بررسی بصری از این مسیرها را می توان با ترسیم ابزارهای نتیجه در طول زمان برای هر دو گروه یا با تجسم نتایج مدل روندهای خطی به دست آورد. ما می توانیم هر دو این بررسی های تشخیصی را با استفاده از دستور estat trendplot انجام دهیم.



قبل از اجرای سیاست جدید پذیرش بیمار، بیمارستان های کنترل و درمان مسیر موازی را دنبال می کردند. علاوه بر آن، ما می توانیم این فرض را با استفاده از آزمون روندهای موازی با دستور **estat ptrends** ارزیابی کنیم.

Parallel-trends test (pretreatment time period)  
H0: Linear trends are parallel

F(1, 45) = 0.55  
Prob > F = 0.4615

شواهد کافی برای رد فرضیه صفر روندهای موازی وجود ندارد. این آزمون و تجزیه و تحلیل گرافیکی از فرض روندهای موازی پشتیبانی می کند. آزمایش دیگری که ممکن است بخواهیم انجام دهیم این است که ببینیم آیا

گروه های کنترل یا درمان در پیش بینی درمان رفتار خود را تغییر می دهند یا خیر. این با استفاده از آزمون علیت گرنجر با استفاده از دستور `estat granger` ارزیابی می شود.

Granger causality test  
H0: No effect in anticipation of treatment

F(2, 45) = 0.33  
Prob > F = 0.7239

در این مثال، داده های تعداد کافی بیمارستان (۴۶) را تا استنباط های موثقی در مورد اثر درمانی خود بررسی شدند. داشته باشیم. اگر فقط اطلاعات ۱۵ بیمارستان وجود داشت، ممکن است روش های جایگزین را در نظر بگیرید. برای استفاده از خطاهای استاندارد اصلاح شده با سوگیری از Bell و McCaffrey (۲۰۰۲) استفاده می شود. دستور زیر را اجرا کنید.

`didregress (satis) (procedure), group(hospital) time(month) vce(hc2)`  
. didregress (satis) (procedure), group(hospital) time(month) vce(hc2)

Computing degrees-of-freedom:  
procedure .....

Number of groups and treatment time

Time variable: month  
Control: procedure = 0  
Treatment: procedure = 1

	Control	Treatment
Group hospital	28	18
Time Minimum	1	4
Maximum	1	4

Difference-in-differences regression

Number of obs = 7,368  
No. of clusters = 46

Data type: Repeated cross-sectional

satis	Robust HC2		t	P> t	[95% conf. interval]	
	Coefficient	std. err.				
ATET procedure (New vs Old)	.8479879	.0325552	26.05	0.000	.7819941	.9139816

Note: ATET estimate adjusted for group effects and time effects.

برای استفاده از روش تجمیع پیشنهاد شده توسط دونالد و لانگ (۲۰۰۷)، دستور زیر را اجرا کنید.

`didregress (satis) (procedure), group(hospital) time(month) aggregate(dlang, varying)`

می توانید با استفاده از روش بوت استرپ `wild-cluster` برای بدست آوردن مقادیر p و فاصله اطمینان استفاده کنید تا نتایج قابل تکرار باشد. دستور زیر را اجرا کنید.

`didregress (satis) (procedure), group(hospital) time(month) wildbootstrap(rseed(111))`

## روش های اپیدمیولوژی و پژوهش با Stata

```
. didregress (satis) (procedure), group(hospital) time(month) aggregate(dlang, varying)
Donald-Lang varying-coefficient regression (322)
```

```
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
      1       2       3       4       5
..... 50
..... 100
..... 150
..... 200
..... 250
..... 300
.....
```

Number of groups and treatment time

```
Time variable: month
Control:       procedure = 0
Treatment:    procedure = 1
```

	Control	Treatment
Group hospital	28	18
Time Minimum	1	4
Maximum	1	4

```
Difference-in-differences regression                               Number of obs = 322
Data type: Repeated cross-sectional
Aggregation: Donald-Lang, varying coefficients
```

	satis	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ATET procedure (New vs Old)		.8500467	.0255727	33.24	0.000	.7997311 .9003623

Note: ATET estimate adjusted for group effects and time effects.

**didregress (satis) (procedure), group(hospital) time(month) wildbootstrap(rseed(111))**

```
. didregress (satis) (procedure), group(hospital) time(month) wildbootstrap(rseed(111))
computing 1000 replications
```

```
Finding p-value
..... 50%
..... 100%
Confidence interval lower bound
.....
Confidence interval upper bound
.....
```

Number of groups and treatment time

```
Time variable: month
Control:       procedure = 0
Treatment:    procedure = 1
```

	Control	Treatment
Group hospital	28	18
Time Minimum	1	4
Maximum	1	4

```
DID with wild-cluster bootstrap inference                               Number of obs = 7,368
                                                                    No. of clusters = 46
                                                                    Replications = 1,000
```

```
Data type: Repeated cross-sectional
Error weight: rademacher
```

	satis	Coefficient	t	P> t	[95% conf. interval]
ATET procedure (New vs Old)		.8479879	26.41	0.000	.7806237 .9157614

Note: ATET estimate adjusted for group effects and time effects.

در این قسمت برای آشنایی با تجزیه و تحلیل سری های زمانی مطالب زیر بحث می شود. به منظور آشنایی با مطالب دیگر در این زمینه به کتاب های تخصصی سری های زمانی مراجعه شود. تاکید می گردد که این قسمت صرفاً برای آشنایی مقدماتی کاربران نرم افزار تدوین گردیده و برای درک عمیق به منابع دیگر مراجعه شود. یک سری زمانی مجموعه ای از مشاهدات درباره یک متغیر است که در فواصل زمانی مختلف اندازه گیری گردیده و بر حسب زمان مرتب شده اند. بنابراین یک سری زمانی از مشاهده یک پدیده در طول زمان بدست می آید. یک سری زمانی، به عنوان دنباله ای از مشاهدات بر روی یک متغیر در نظر گرفته می شود. متغیر در نقاط گسسته ای از زمان که معمولاً فاصله های مساوی دارند، اندازه گیری می شود. بررسی رفتار گذشته سری و پیش بینی مقادیر آینده رخداد متغیر یا عبارتی کشف و شناسایی مدل های احتمالی داده ها و پیش بینی مقادیر این اندازه ها در آینده از جمله اهداف مهم این نوع تحلیل است. مثال: تعداد موارد آنفلوآنزای ماهیانه در طی ماه های مختلف سال، شاخص خشکسالی در طی ماه ها و سال های متعدد، میزان بیکاران فصلی، میزان درجه حرارت روزانه، تعداد موارد سرطان دستگاه گوارش در طی بیست سال، وقوع موارد تصادفا رانندگی در طی سال های متوالی.

سری های زمانی را به طور کلی به دو دسته تقسیم می شوند.

۱- سری زمانی پیوسته : هرگاه مشاهدات به طور پیوسته در طی زمان ایجاد شده باشند.

۲- سری زمانی گسسته : هرگاه مشاهدات فقط در زمانهای معینی که معمولاً به فواصل مساوی از یکدیگر قرار دارند اندازه گیری شده باشند.

در تحلیل یک مجموعه سری زمانی رسم نمودار سری از جمله اصول اولیه تحلیل می باشد. نمودار اطلاعات مفیدی در مورد طبیعت داده ها به تحلیل کننده می دهد. در ترسیم نمودار سری زمانی، مشاهدات یک سری زمانی را در برابر زمان رخداد آنها قرار داده می شود.

### راهکارهای الگو سازی یا مدل سازی:

پیدا کردن الگوهای مناسب برای سری های زمانی یک اقدام مهم بشمار می رود. الگوی چند مرحله ای که بوسیله باکس و جنکینس در سال ۱۹۷۶ ارائه گردید بعنوان استراتژی مهم در مدل سازی مطرح است. این الگو مشتمل بر موارد زیر است:

(۱) تشخیص یا شناسایی الگو : دسته ای از الگوهای سری های زمانی را که برای سری زمانی مشاهده شده مناسب است انتخاب می گردد. نمودار زمانی سری مورد توجه قرار می گیرد. آماره های مرتبط از داده ها محاسبه می شوند. الگوی انتخاب شده آزمایشی است و به تجدید نظر در مرحله تجزیه و تحلیل بستگی دارد. الگویی که بکار برده می شود باید کمترین تعداد پارامتر که بطور قابل قبولی داده ها را مشخص می کند داشته باشد. جمله معروف انیشتاین "هر چیز بایستی تا جایی که ممکن است ساده شود ولی نه ساده تر" مد نظر قرار گیرد.

۲) برازش الگو : پیدا کردن بهترین برآوردهای ممکن پارامترهای نامعلوم الگوی داده شده با استفاده از روش حداکثر درست نمایی یا کمترین مربعات.

۳) تشخیص درستی الگو: تجزیه و تحلیل کیفیت الگویی که تشخیص و برآورد گردید مد نظر است. یا به عبارتی برازش مدل به داده ها تا چه اندازه مناسب است؟ آیا فرض های مدل برقرار است؟ اگر فرض ها صادق باشند آن گاه می توان مقادیر آینده را پیش بینی نمود. در غیر این صورت به مرحله تشخیص بازگشته و الگوی دیگری انتخاب می شود و بدین ترتیب سه مرحله را آنقدر دور باید زد تا مدل مناسب و قابل قبول انتخاب و برازش گردد. محاسبات پیچیده بوده و این مراحل با نرم افزا انجام می شود.

**همبستگی بین مشاهدات سری زمانی:** در سری های زمانی ویژگی مهم این است که معمولا مشاهدات متوالی مستقل نیستند و داده ها به هم وابسته هستند. لذا تحلیل اساسی در سری زمانی، بررسی وابستگی داده ها به همدیگر است که در تحلیل برای این وابستگی از تابع خود همبستگی و تابع خود همبستگی جزئی استفاده می شود. چون داده ها به هم وابسته هستند نمی توان از تحلیل های رگرسیونی استفاده نمود.

خود همبستگی در تاخیر  $K$ : عبارت است از همبستگی بین مشاهداتی که  $K$  واحد زمانی با یکدیگر فاصله دارند. تابع خود همبستگی نظری که آن را با  $\rho_k$  نشان می دهیم، عبارتست از: (autocorrelation)

$$\rho_k = \frac{\text{cov}(x_t, x_{t+k})}{\text{var}(x_t)}$$

همبستگی بین  $x_t$  و  $x_{t+k}$  بعد از حذف اثر متغیرهای  $x_{t+1}, x_{t+2}, \dots, x_{t+k-1}$  را ضریب خود همبستگی جزئی می نامند. کوواریانس و همبستگی هر دو وابستگی (خطی) بین متغیرهای تصادفی را اندازه گیری می کند. با فرمول زیر نیز میتوان این ضریب را نشان داد.

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

- $r_k$  = autocorrelation coefficient for a k period lag.
- $\bar{y}$  mean of the time series.
- $y_t$  = Value of the time series at period t.
- $y_{t-k}$  = Value of time series k periods before period t.

**مدل سازی سری های زمانی به روش باکس - جنکینز (ARIMA):** مدل سازی یک سری زمانی بطور کلی مشتمل بر تشخیص مدل آزمایشی، تخمین پارامترهای مدل ( برازش مدل ) و بررسی مناسب بودن مدل است. این مراحل در بالا توضیح داده شد. برای تشخیص مدل آزمایشی رسم acf و pacf نمونه ای ، مرحله اول تشخیص مدل آزمایشی است. بعد از آن آزمون وجود روند قطعی در مدل انجام می شود.

مدل ARIMA سه پارامتر دارد.  $P$ ،  $d$  و  $q$ . پارامتر اول همان ضریب خود همبستگی است. اگر صفر باشد یعنی ارتباطی وجود ندارد. پارامتر دوم که اکثراً  $d$  در نظر گرفته می شود، همان تفاضل گیری است که ممکن است چندبار انجام شود. و پارامتر سوم میانگین متحرک است. که  $AC$  بررسی می شود اگر  $q$  بیشتر از صفر شد یعنی ارتباط وجود دارد.

ARIMA model specification

ARIMA(p,d,q) specification:

Autoregressive order (p)

Integrated (difference) order (d)

Moving-average order (q)

بنابراین با آزمون و خطا با کاهش و افزایش پارامترهای مدل، مدل مناسب پیدا می شود. برای بررسی مناسب بودن مدل از بررسی باقی مانده ها (اختلاف مقدار واقعی و مقدار پیش بینی شده) استفاده میشود. مثلاً اگر مدل مناسب باشد باقی مانده ها باید توزیع نرمال داشته باشند و همچنین ضریب خود همبستگی و ضریب خود همبستگی جزئی باقی مانده ها در هیچکدام از تاخیر ها معنی دار نباشد. بعد از پیدا کردن یک مدل مناسب برای سری زمانی میتوان، مقادیر آینده را از روی آن پیش بینی نمود. برای مدل سازی روندها از روش های غیر تصادفی بودن در مقابل روندهای تصادفی، برآورد یک میانگین ثابت و روش های رگرسیونی خطی، مرتبه دوم، میانگین های فصلی و کسینوسی استفاده می شود.

### ایستایی داده ها:

برای انجام یک استنباط آماری درباره فرایند های تصادفی باید فرض های ساده و معقولی را در نظر داشت. یکی از مهمترین فرضها ایستایی داده هاست. مفهوم اساسی ایستایی این است که قوانین احتمالی حاکم بر فرایند با زمان تغییر نمی کند. یا به عبارتی فرایند در تعادل آماری است.

هر داده ای که ترند داشته باشد ایستا نیست. چون میانگین در طول زمان تغییر می کند. برای تحلیل باید داده های نایستا را مثلاً با تفاضلی کردن و یا تبدیلات لگاریتمی و توانی به ایستا تبدیل نمود. داده هایی که در طول زمان میانگین و واریانس ثابتی داشته باشند و میانگین و واریانس آنها تغییر نکند داده های ایستا نامیده می شود. ایستایی داده ها از جمله گام های اولیه است که در تحلیل بررسی می شود. تفاضلی کردن زیاد نیز ممکن است همبستگی غیرلازمی را در الگو ایجاد کند و الگوی ساده را پیچیده نماید.

برای رفع بی نظمی داده ها و نوسانات داده ها یا هموار سازی (Smoothing) داده ها از تکنیک های مختلفی استفاده می شود. یکی از روش های مورد استفاده رایج استفاده از میانگین متحرک یا Moving Averag است. در نرم افزار با استفاده از دستورات زیر می توان هموارسازی داده ها را انجام داد.

- `tssmooth ma ma1= varname, win(1)`
- `tssmooth ma ma2= varname, win(2)`
- `tssmooth ma ma3= varname, win(3)`
- `tsline ma3`
- `tsline varname ma3`

در این روش داده هایی که نزدیک به هم هستند را یک میانگین می گیرند تا بی نظمی از بین برود. از دادن وزن هم اجتناب می شود. فرمول محاسبات به شکل زیر است.

$$F_{t+1} = \frac{1}{n}(D_t + D_{t-1} + \dots + D_{t+1-n})$$

$$F_{t+1} = \frac{1}{n} \sum_{i=t+1-n}^t D_i$$

$F_t$  is average of  $n$  previous observations or actuals  $D_t$

- Note that the  $n$  past observations are equally weighted.

روش دوم برای هموارسازی داده ها روش نمایی است. این روش به داده های قبلی وزن کمتر و به داده های جدید وزن بیشتر می دهد. مقدار آلفا اختیاری است ولی دستور مربوطه بهترین مقدار را در نظر می گیرد. بهترین مقدار برای آلفا مقداری است که کمترین مقدار Root mean squared error را دهد. این روش داده ها را پیشگویی نیز می کند. برای استفاده از این روش دستورات به شرح زیر می باشد. همانند دستورات روش قبلی است منتها بجای ma در اینجا از exp استفاده می شود.

- tssmooth exp sm2= varname
- tssmooth exp sm3= varname, p(.7)
- tsline varname sm2

روش ها و مدل های دیگری برای هموارسازی داده ها وجود دارد. روشهایی که برای داده های دارای ترند استفاده می شود عبارتند از:

- Moving averages –
- Holt's linear exponential smoothing –
- Simple regression –
- Exponential models –
- Time series decomposition –
- Autoregressive integrated moving average (ARIMA) –

عناصر یک سری زمانی: برای هر سری زمانی عناصر زیر وجود دارد.

۱. روند یا تغییرات دراز مدت در سطح داده ها (trend)
۲. تغییرات دوره ای (cyclical variation)
۳. تغییرات فصلی (seasonal variation)
۴. تغییرات نامنظم (irregular variation)

### ۱ – روند (T) :

روند یا تغییرات دراز مدت به تغییرات متغیر مورد مطالعه در یک دوره زمانی طولانی بدون در نظر گرفتن تغییرات دوره ای، فصلی و نامنظم تعریف می شود. به عبارت دیگر روند عبارتست از ترند افزایشی و یا کاهششی و ایستا و یا حرکات رو به بالا و پایین یک سری زمانی که نشان دهنده کاهش یا افزایش بلند مدت یک سری زمانی است.



## ۲- تغییرات دوره ای یا سیکل (C) :

تغییرات دوره ای عبارت است از تکرار حرکات رو به بالا و پایین حول سطوح روند. این نوع تغییرات دارای دوره نوسان بیشتر از یک سال می باشد. نوسانات دوره ای ممکن است دقیقا از طرح های مشابهی بعد از فواصل زمانی مساوی پیروی کنند ولی همیشه معمول نیست. یک دوره کامل را که معمولا ۷ تا ۹ سال طول می کشد اصطلاحا یک "دوره" می نامند. یکی از معمولی ترین نوسانات سیکلی داده های سری زمانی، سیکل تجاری است. سیکل تجاری وقوع مکرر دوره های رونق و رکود است.

## ۳- تغییرات فصلی (S) :

تغییرات فصلی تغییراتی هستند که در دوره های تناوبی کوتاه مانند فصل، پیش می آیند. این تغییرات مربوط به عواملی هستند که به طریقی منظم و چرخه ای روی یک دوره کمتر از یک سال حرکت می کنند. در واقع تغییرات فصلی رفتار دوره ای متغیر را نشان می دهد، یعنی رفتاری که معمولا هر سال در همان فصل تقریبا با همان شدت روی می دهد. مثلا اپیدمی مالاریا در یک فصل خاص در طی سال های متوالی.

## ۴- تغییرات نامنظم (I) :

تغییرات نامنظم عبارت است از حرکات پراکنده در یک سری زمانی که از الگوی منظم و مشخصی پیروی نمی کنند. به عبارتی الگویی که مطابق با سه الگوی قبلی نباشد. در واقع این حرکات بیان می کنند که پس از محاسبه روند، تغییرات دوره ای و تغییرات فصلی چه چیز دیگری در سری زمانی بجا می ماند. نوسانات نامنظم معمولا ناشی از وقایع غیر معمولی هستند که قابل پیش بینی نیستند. مانند زمین لرزه، طوفان، جنگ، تصادفات

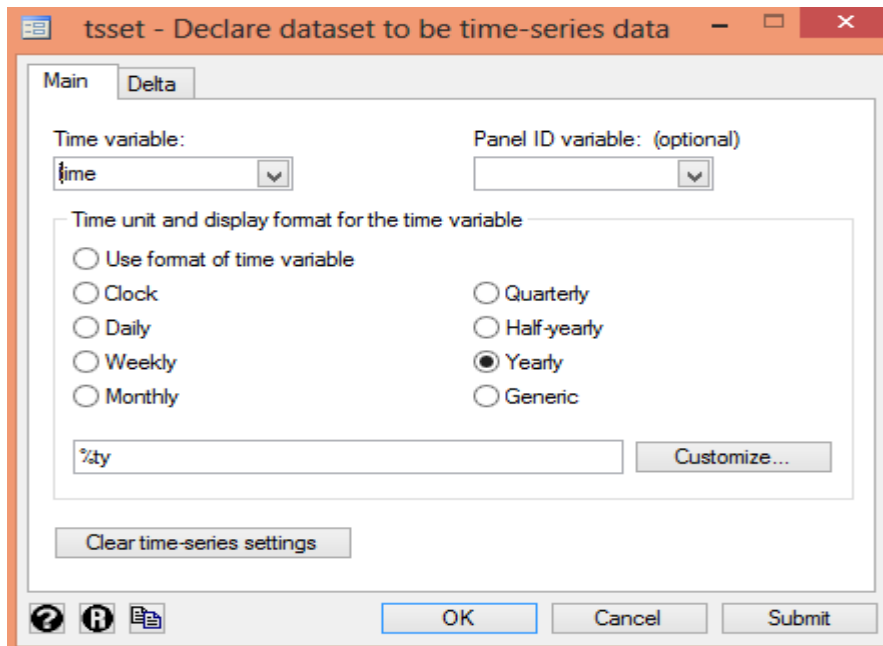
## کار با نرم افزار و انجام تحلیل سری های زمانی:

در مدیریت داده ها و عملگرهای سری زمانی ابتدا باید مطمئن بود متغیر مورد بررسی در سری زمانی به فرمت Stata باشد و سپس باید داده ها را در این متغیر سورت نمود و بعد از آن داده ها را با دستور `tsset` به نرم افزار معرفی نمود. سپس می توان آنالیز تک متغیره سری زمانی (Univariate time series) و آنالیز چند متغیره سری زمانی (Multivariate time series) را با نرم افزار انجام داد. برای انجام این تحلیل از منوی `statistics` از مسیر زیر اقدام می شود.

Time series	Setup and utilities	Declare dataset to be time-series data
Multivariate time series	ARIMA and ARMAX models	Report time-series aspects of dataset
Spatial autoregressive models	ARCH/GARCH	Fill in gaps in time variable
Longitudinal/panel data	ARFIMA models	Add observations to time-series dataset
Multilevel mixed-effects models	Unobserved-components model	
Survival analysis	Markov-switching model	
Epidemiology and related	Threshold regression model	
Endogenous covariates	Prais-Winsten regression	
Sample-selection models	Regression with Newey-West std. errors	
Treatment effects	State-space models	
SEM (structural equation modeling)	Forecasting	
LCA (latent class analysis)	Postestimation	
FMM (finite mixture models)	Rolling-window and recursive estimation	
IRT (item response theory)	Smoothers/univariate forecasters	
Multivariate analysis	Filters for cyclical components	
Survey data analysis	Tests	
Lasso	Graphs	

با تایپ دستور `db tsset` نیز می توان کادر محاوره ای زیر را باز و متغیر زمان را به نرم افزار معرفی نمود. در فایل `skin diseases` که بیماریهای پوستی را به صورت فرضی و سالیانه در قاره آفریقا نشان می دهد با این دستور به برنامه معرفی می شود.

نتایج در زیر نشان داده شده است. داده ها فرضی است و لذا ممکن است مدل نامناسب باشد.



```
. sort time
. tsset time, yearly
    time variable:  time, 1 to 0144
      delta: 1 year
```

راهکارهای مدل سازی: در مدیریت داده ها در سری های زمانی با دستورات زیر کار شروع می شود.

<code>tsset</code>	Declare data to be time-series data
<code>tsfill</code>	Fill in gaps in time variable
<code>tsappend</code>	Add observations to a time-series dataset
<code>tsreport</code>	Report time-series aspects of a dataset or estimation sample
<code>tsrevar</code>	Time-series operator programming command
<code>haver</code>	Load data from Haver Analytics database
<code>rolling</code>	Rolling-window and recursive estimation
	دستورات در تحلیل تک متغیره ( Univariate time series ) شامل موارد زیر است.
<code>Arima</code>	ARIMA, ARMAX, and other dynamic regression models
<code>Arima postestimation</code>	Postestimation tools for arima

arch Autoregressive conditional heteroskedasticity (ARCH) family of estimators

arch postestimation Postestimation tools for arch

newey Regression with Newey–West standard errors

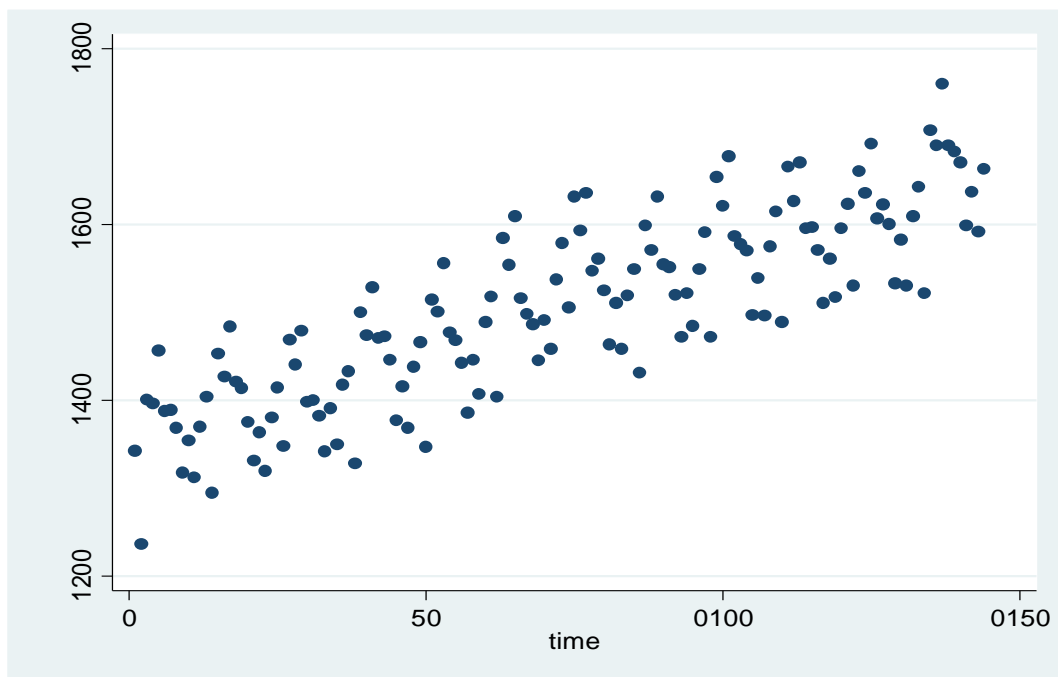
newey postestimation Postestimation tools for newey

prais Prais–Winsten and Cochrane–Orcutt regression

### نمودارهای سری زمانی

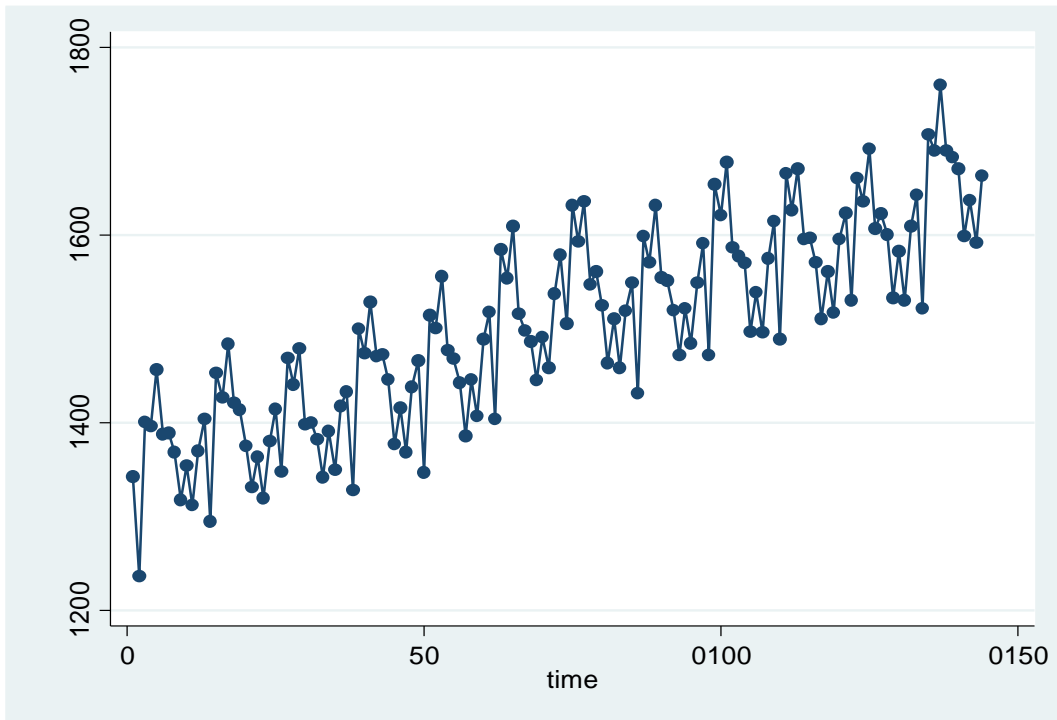
پس تنظیم متغیرها اولین قدم در بررسی داده های زمانی، ترسیم روند تغییرات مشاهدات در طول زمان است. لذا با استفاده از نمودار نقاط پراکنده، پراکندگی تعداد موارد بیماری پوستی، در طول زمان ترسیم می شود.

scatter total time

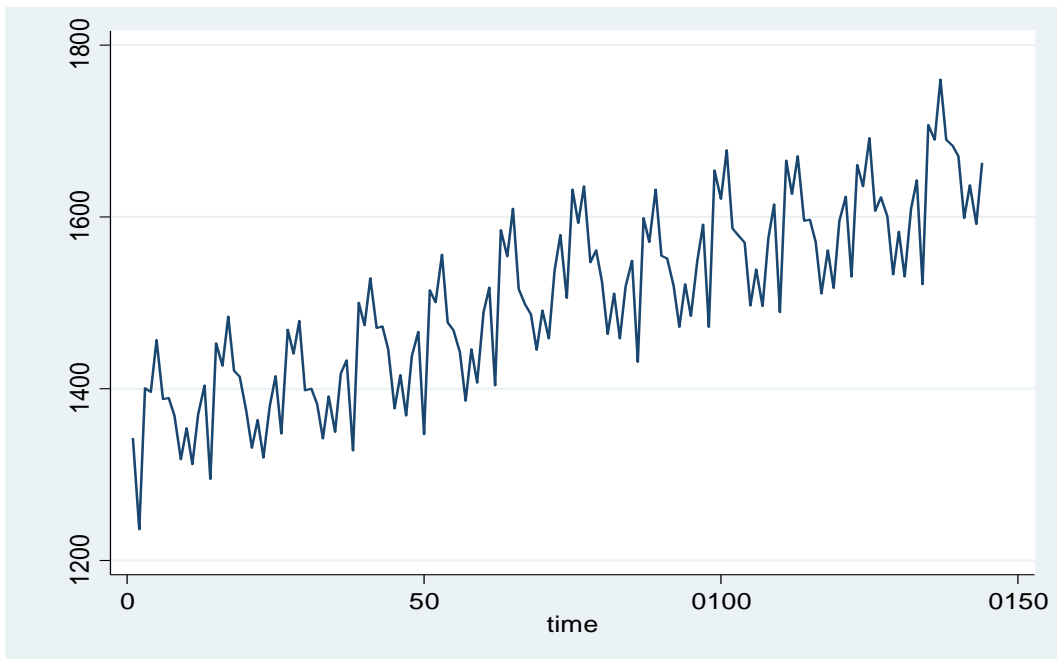


برای اتصال نقاط در نمودار بالا و ترسیم نمودار به صورت زیر، از دستور زیر استفاده کنید:

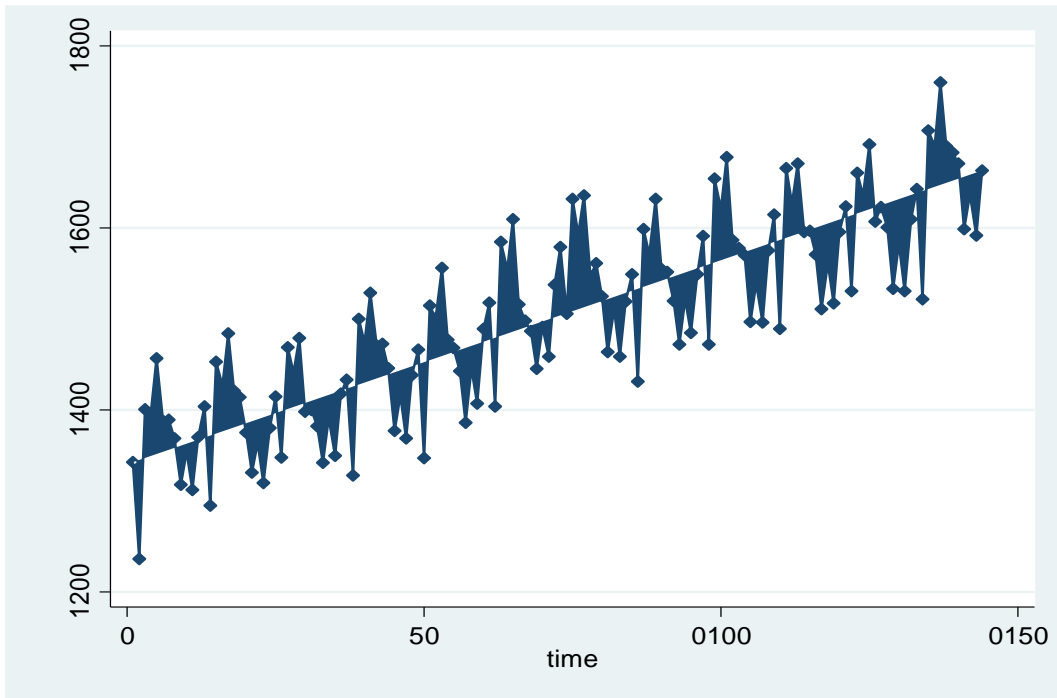
scatter total time, connect(l)



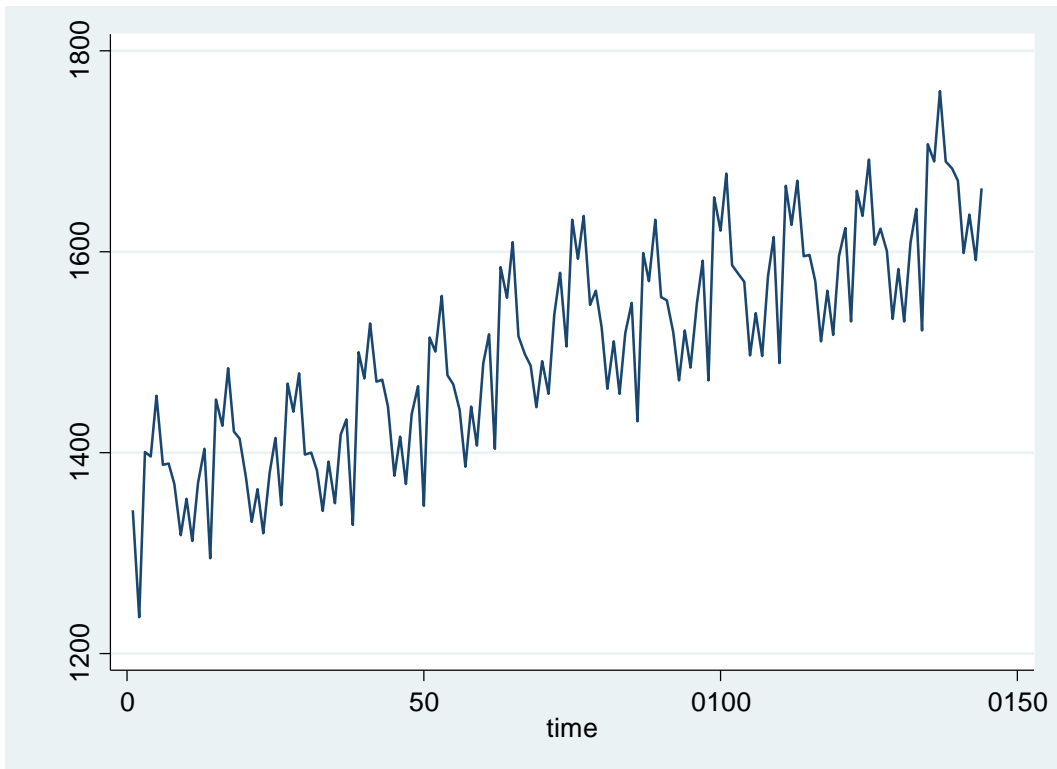
سپس برای ترسیم نمودار واضح تر و بیان ترند زمانی و لگ ها از دستور زیر استفاده نمایید.  
`scatter total time, connect(l) msymbol(i)`



اگر به جای `l` در پرانتز آخر `X` یا `D` بگذارید، نمودارها به شکل زیر ترسیم می شوند.



دستور اختصاصی برای ترسیم روند برای داده های سری زمانی، دستور `tsline` است. دستور زیر را اجرا کنید:



این دستور `tsline skindisease` ، قابلیت بسیاری برای جداکردن زمانهای خاص در نمودار، خارج کردن یک زمان مشخص از محدوده زمانی و ترسیم خطوط راهنماهای متعدد را دارد.

بررسی LAG : در آنالیز زمانی، تعیین میزان تاثیر LAG (زمان های تاخیر) های مختلف در میزان خودوابستگی (Autocorrelation) مشاهدات بسیار مهم است. برای شناسایی الگوی موجود در مدل این تحلیل انجام می شود. برای مثال، در داده های بیماری های پوستی، برای سنجش ارتباط بین تعداد بیماری در این سال با تعداد بیماری پوستی مشاهده شده در سال قبل از ضریب خود همبستگی استفاده می شود. قرار دادن تابع ضریب خودهمبستگی در مقابل زمان تاخیر (time lag) نمودار Correlogram نامیده می شود. زمان تاخیری در محور xها و ضریب خود همبستگی در محور yها قرار می گیرد. این نمودار و الگوی آن برای استفاده از نوع تحلیل مفید می باشد. با دستور زیر این ضریب محاسبه می شود.

corrgram skindisease, lags(1)

خروجی دستور، به صورت زیر خواهد بود:

. corrgram skindisease, lags(1)

LAG	AC	PAC	Q	Prob>Q	-1	0	1	-1	0	1
					[Autocorrelation]			[Partial Autocor]		
1	0.7488	0.7611	82.427	0.0000		—			—	

AC : Autocorrelation

PAC : Partial Autocorrelation

Q : Statistics Prob>Q : P-value

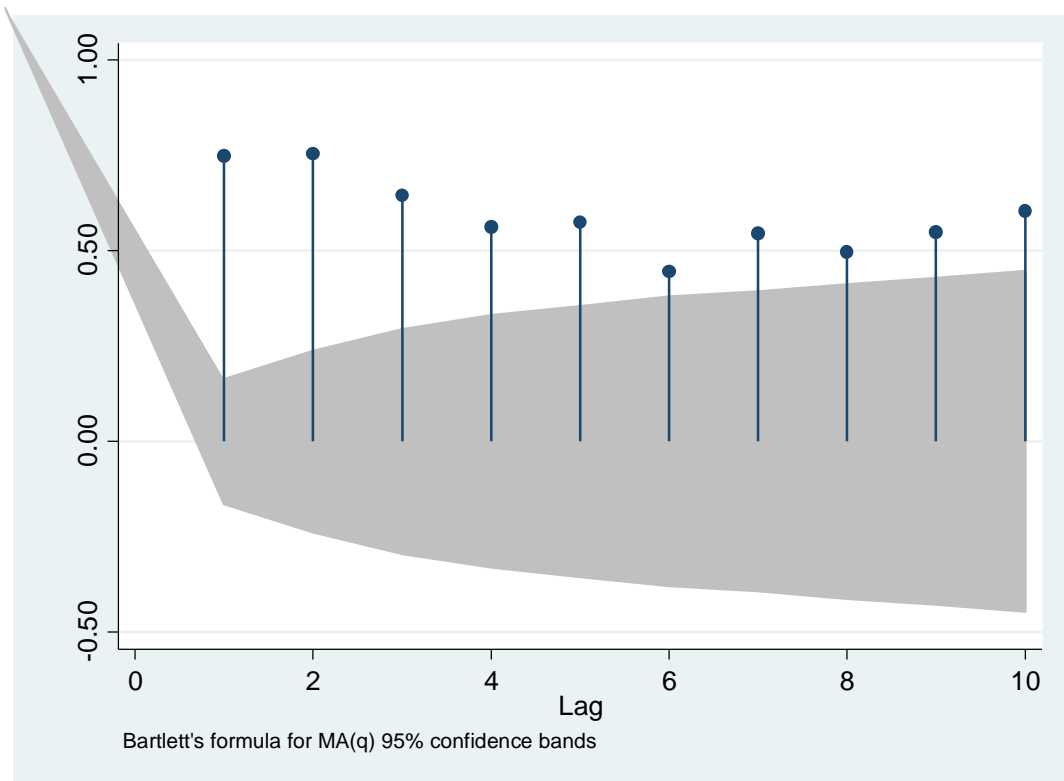
مانند ضریب همبستگی، خودهمبستگی از -1 تا +1 تغییر می کند که +1 یعنی همبستگی مثبت ۱۰۰٪ و -1 یعنی همبستگی معکوس ۱۰۰٪ و صفر به معنای عدم همبستگی است. آزمون Q اختلاف معنی دار مقدار خودهمبستگی مشاهده شده از صفر را نشان می دهد.  $Q = ۸۲,۴۲۷$

اگر مانند بالا ضرایب خودهمبستگی و خودهمبستگی پارشیال برای زمان های تاخیر یا لگ های ۱ تا ۵ را محاسبه کنید. خواهید دید که همه لگ ها معنی دار است.

. corrgram skindisease, lags(5)

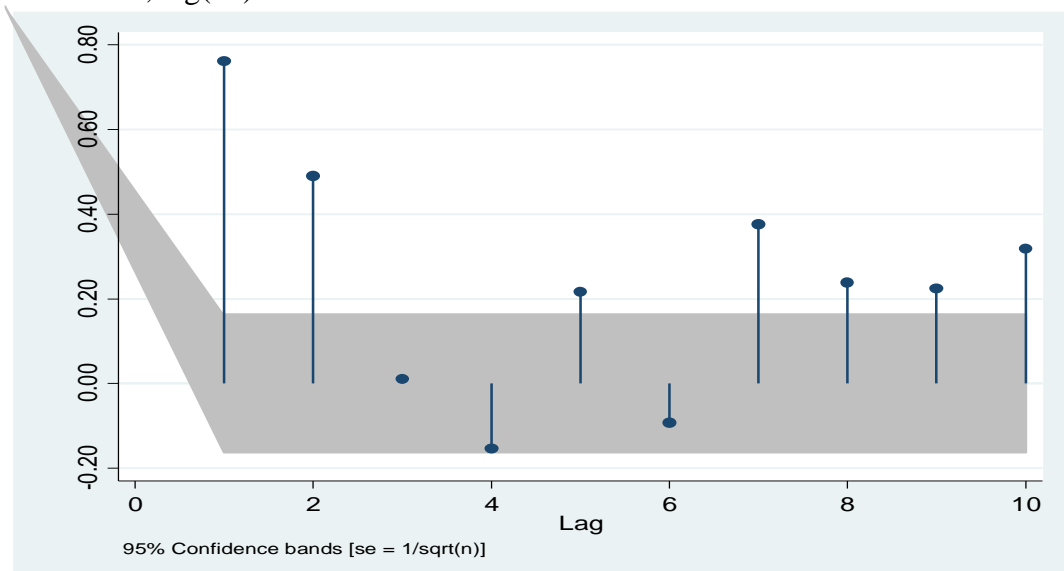
LAG	AC	PAC	Q	Prob>Q	-1	0	1	-1	0	1
					[Autocorrelation]			[Partial Autocor]		
1	0.7488	0.7611	82.427	0.0000		—			—	
2	0.7537	0.4901	166.52	0.0000		—			—	
3	0.6449	0.0102	228.53	0.0000		—			—	
4	0.5627	-0.1543	276.08	0.0000		—			—	
5	0.5747	0.2165	326.03	0.0000		—			—	

دستورات دیگری هم برای نمایش الگوی خودهمبستگی و خودهمبستگی پارشیال در داده ها وجود دارند.  
ac skindisease , lag(10)



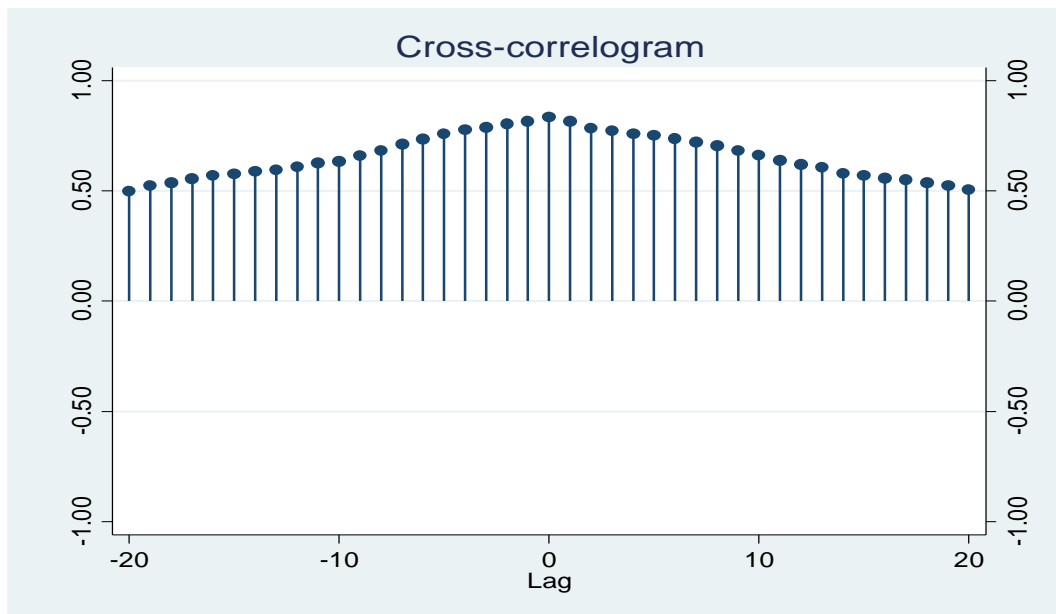
در نمودار بالا و همچنین نمودار زیر، مقادیری از لگ ها که از نمودار خاکستری بیرون آمده باشند معنی دار است. و نشان از ترند قابل توجه دارد. لذا این ترند را باید در داده ها حذف و کاهش داد تا به مدل مناسب دست یافت. قسمت خاکستری رنگ حدود اطمینان برای ضریب صفر است.

pac skindisease , lag(10)



برای پی بردن به اینکه کدام مقدار lag بیشترین خودهمبستگی را بین مشاهدات ایجاد می کند، دستور xcorr توصیه می شود.

xcorr time skindisease



برای درک بهتر مفهوم autocorrelation و Partial autocorrelation با اجرای دستورات زیر مشخص می شود که همبستگی لگ های ۱ و ۲ چگونه است. متغیری بنام لگ اول ایجاد می شود که با متغیر تعداد بیماری پوستی (یک سال قبل از خود) برابر است. سپس ارتباط لگ ساخته شده با متغیر بیماری پوستی در سال مورد نظر بررسی می شود.

```
. gen lag1= skindisease[_n-1]
(1 missing value generated)
```

```
. correlate skindisease lag1
(obs=143)
```

	skindi~e	lag1
skindisease	1.0000	
lag1	0.7613	1.0000

```
. gen lag2= skindisease[_n-2]
(2 missing values generated)
```

```
. correlate skindisease lag2
(obs=142)
```

	skindi~e	lag2
skindisease	1.0000	
lag2	0.7873	1.0000

```
. pcorr skindisease lag1 lag2
(obs=142)
```

Partial and semipartial correlations of skindisease with

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
lag1	0.3969	0.2447	0.1575	0.0599	0.0000
lag2	0.4984	0.3254	0.2484	0.1059	0.0000



نتایج نشان می دهد که همبستگی داده های بیماری پوستی در هر سال با سال قبل از خود وابسته می باشد. این وابستگی در لگ های ساخته شده معنی دار است. متغیرهای ایجاد شده را با دستور list فهرست نمایید. هر مشاهده با یک مشاهده قبل از خودش ارتباط دارد.

```
list
```

	skindi~e	time	lag1	lag2
1.	1343	1	.	.
2.	1236	2	1343	.
3.	1401	3	1236	1343
4.	1396	4	1401	1236
5.	1457	5	1396	1401
6.	1388	6	1457	1396
7.	1389	7	1388	1457
8.	1369	8	1389	1388
9.	1318	9	1369	1389

more—

مجدد با استفاده از دستور corrgram، و lag برابر با ۲ و ۵ یافته ها را اگر با نتایج بدست آمده از دستورات بالا مقایسه شود، نتایج یکسان شده و مقادیر خود همبستگی تغییر ننموده است.

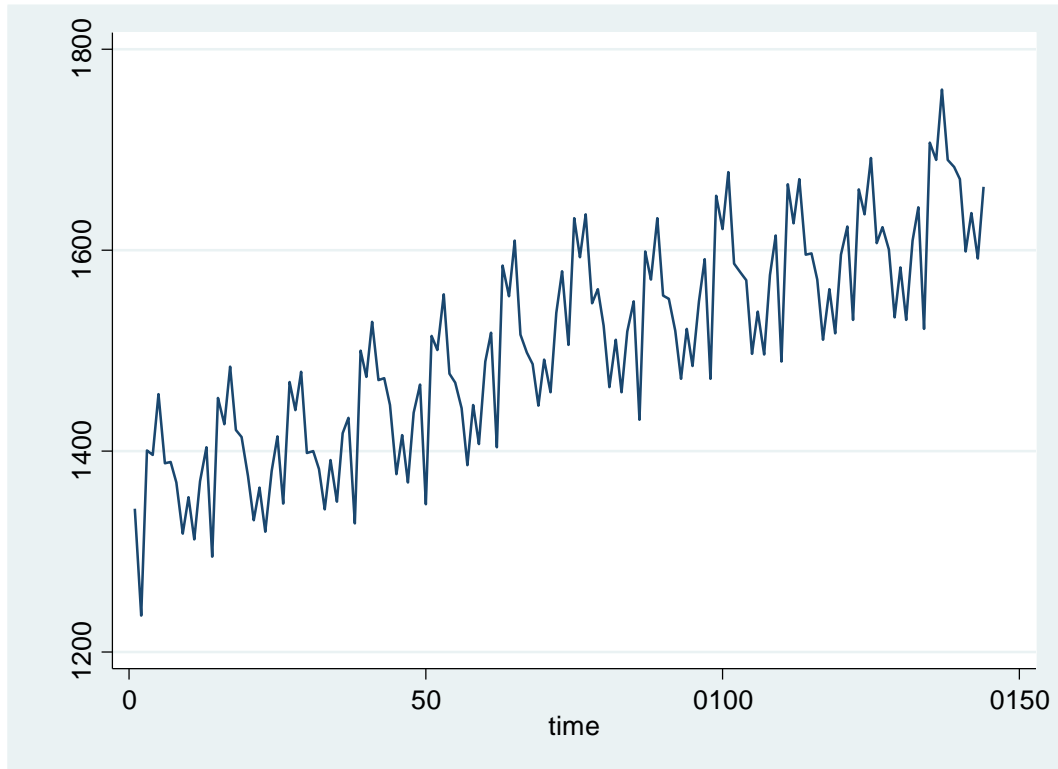
```
. corrgram skindisease, lags(2)
```

LAG	AC	PAC	Q	Prob>Q	-1	0	1	-1	0	1
					[Autocorrelation]			[Partial Autocor]		
1	0.7488	0.7611	82.427	0.0000						
2	0.7537	0.4901	166.52	0.0000						

### مدل (ARIMA (Autoregressive Integrated Moving Average

قبل از برازش مدل، مجدداً نمودار سری زمانی برای داده های بیماری های پوستی در آفریقا ترسیم می شود:

tsline skindisease



نموداری که مشاهده می شود نشان می دهد که یک تغییرات سالیانه یا دقیق تر فصلی، در نمودار مشاهده می شود. برای یافتن مدل مناسب دستورات زیر اجرا می شوند.

ac skindisease  
pac skindisease

یافتن مدل برای یافتن مدل سری زمانی هنگامی معنی دار است که حداقل همبستگی بین مشاهده های سری نهایی (ایستا شده) « مرتبه یک » بوجود آید. همبستگی در lagهای ۱۲، ۲۴، ۳۶ و... در داده های ماهانه نشان از وجود روند فصلی در داده هاست. وجود موج سینوسی در داده ها نیز فصلی بودن را نشان می دهد. برای پیشگویی و برازش مدل، باید فصلی بودن داده ها را کنترل کنیم. با دستور boxcox می توان ایستایی داده ها را بررسی کرد. حل مشکل نایستایی در میانگین با difference گرفتن است. معمولاً با یکبار difference (تفاضلی کردن مرتبه اول) مشکل حل می شود. نایستایی در میانگین ممکن است ناشی از روند خطی و فصلی به طور توأم باشد. در بررسی نایستایی در میانگین مقادیر پارامترهای مدل ARIMA مشخص می شود. مدل های ARIMA برای توصیف رفتار بسیاری از سری های زمانی مفید می باشند. در بررسی و پیش بینی بیماریهای عفونی و در مدیریت سلامت و بهداشت کاربرد فراوانی دارد. مدل ARIMA دارای سه پارامتر است.

- PAC-AR(p) : تعداد lag های معنی دار پیوسته (روی داده های تفاضلی شده در صورت اصلاح نایستایی در میانگین).
- I(difference) : تعداد difference ها برای ایستا کردن میانگین. بعضاً d هم نامیده می شود.
- AC-MA(q) : تعداد lag های معنی دار پیوسته (روی داده های تفاضلی شده در صورت اصلاح نایستایی در میانگین).

**ARIMA(p,d,q) model:**

p=degree of autoregressive model

d = degree of integration (d=number of differences needed to get stationarity).

q = degree of moving average term.

با توجه به نتایج این دو نمودار، مدل احتمالی می تواند arima(10,0,0) باشد. به عبارتی هر ده سال روند داده ها تغییر نموده است. برای اجرای آن دستور زیر را اجرا کنید:

arima skindisease , arima(10,0,0)

ARIMA regression

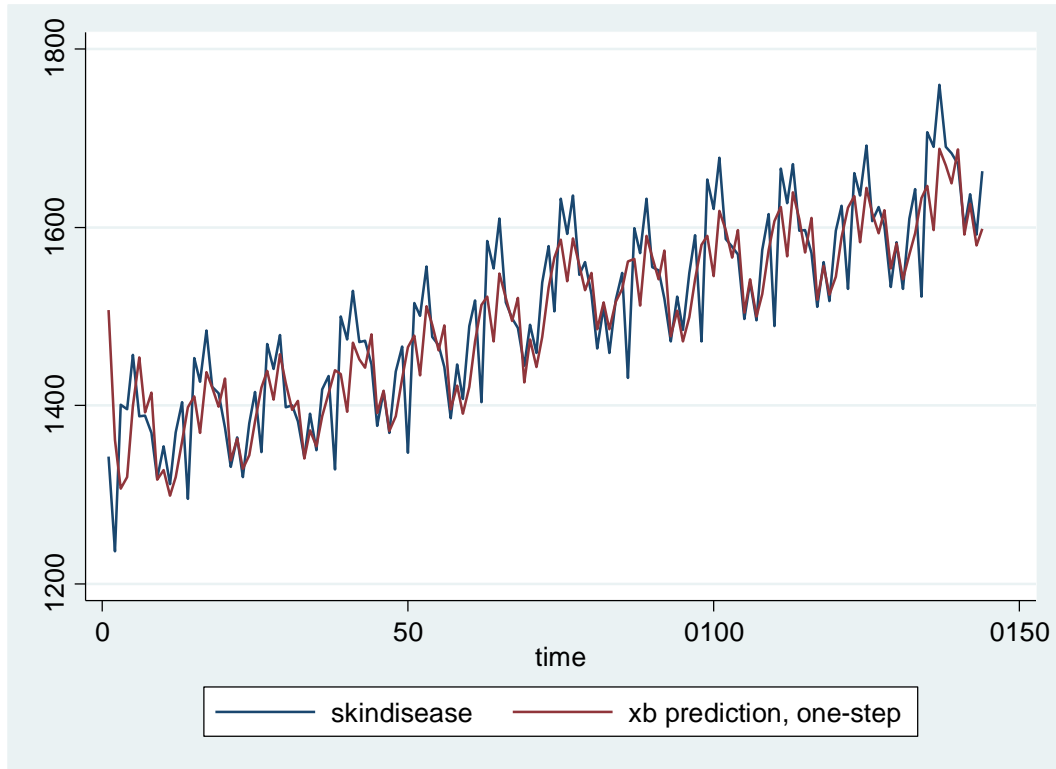
```
Sample: 1 - 0144          Number of obs   =      144
                          Wald chi2(10)         =     2085.25
Log likelihood = -762.996  Prob > chi2       =      0.0000
```

skindisease	OPG					[95% Conf. Interval]	
	Coef.	Std. Err.	z	P> z			
skindisease _cons	1507.195	143.1638	10.53	0.000	1226.599	1787.791	
<b>ARMA</b>							
ar							
L1.	.2829466	.1100967	2.57	0.010	.067161	.4987322	
L2.	.4046142	.1203392	3.36	0.001	.1687538	.6404747	
L3.	.0464321	.1158058	0.40	0.688	-.180543	.2734073	
L4.	-.037035	.1825573	-0.20	0.839	-.3948407	.3207708	
L5.	.0492812	.1890701	0.26	0.794	-.3212893	.4198518	
L6.	-.3304356	.1321658	-2.50	0.012	-.5894759	-.0713953	
L7.	.1400618	.2524049	0.55	0.579	-.3546427	.6347663	
L8.	.0308011	.2122555	0.15	0.885	-.3852121	.4468142	
L9.	.1151612	.1880452	0.61	0.540	-.2534005	.483723	
L10.	.2857494	.1330334	2.15	0.032	.0250087	.5464901	
/sigma	47.48191	4.310746	11.01	0.000	39.033	55.93082	

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

مقدار برازش (goodness of fit) این مدل با مقدار log likelihood نمایش داده شده است. مدل معنی دار است. اگر مدل های دیگری را مناسب می دانید نیز آنها را تعریف نمایید. و مدل را ذخیره نمایید و سپس با مدل های دیگر مقایسه نمایید و مدل مناسب را انتخاب نمایید. اگر نتیجه گرفتید که مدل مناسب است دستور (predict fit) را اجرا نمایید.

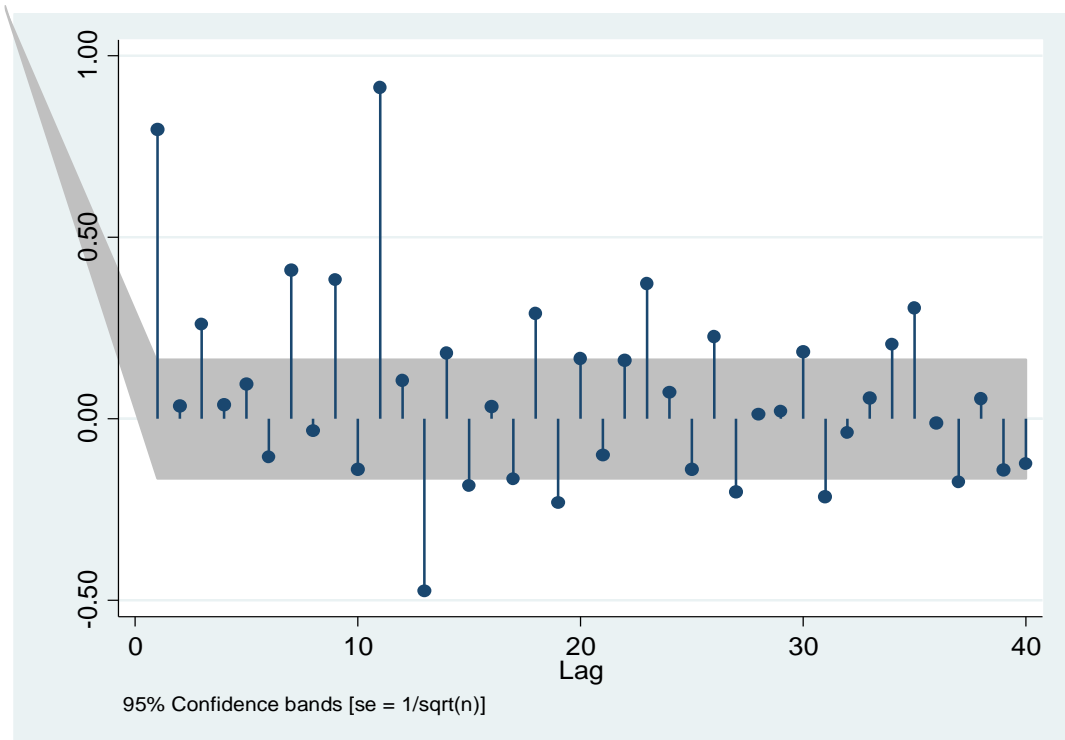
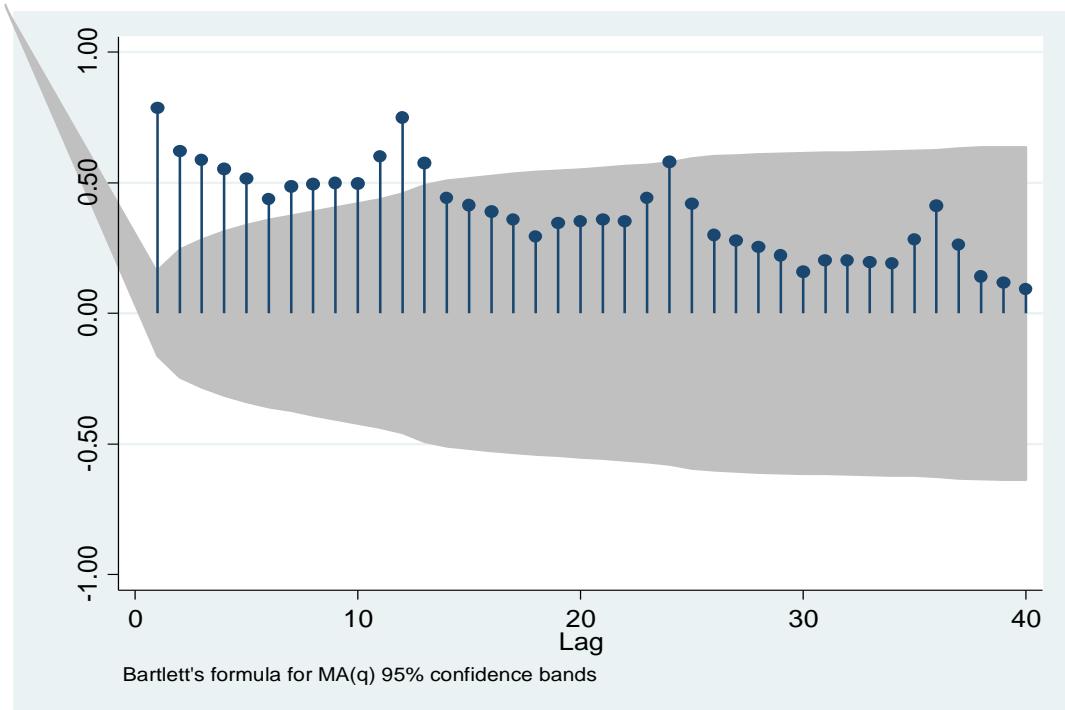
این دستور ستونی بنام fit را ایجاد می کند که همان مقادیر پیش بینی شده است. برای رسم نمودار سری زمانی که هم سری واقعی و هم سری پیش بینی شده را نشان دهد از دستور `tsline skindisease fit` استفاده نمایید.

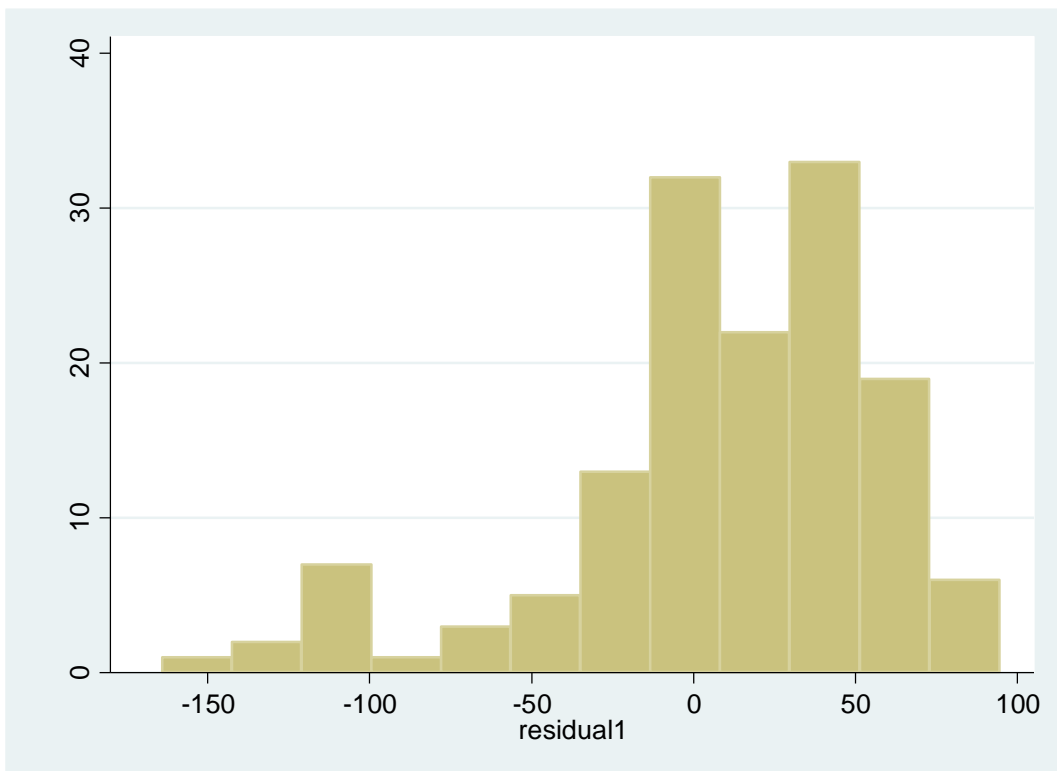
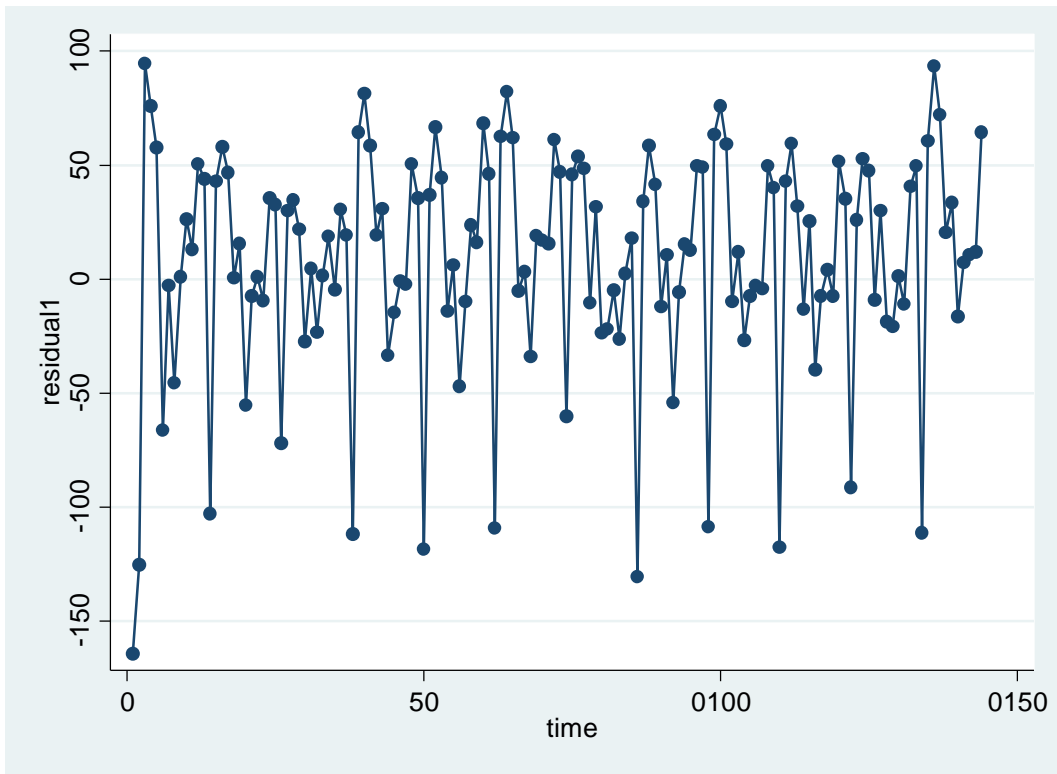


همانگونه که نمودار بالا نشان می دهد مقادیر مشاهده شده و مقادیر برازش شده به نظر نزدیک به هم هستند و مدل تقریباً مناسب است! دستورات زیر را برای ایجاد و بررسی و ارزیابی باقی مانده ها اجرا کنید:

```
gen residual= skindisease-fit
ac residual
pac residual
scatter residual time
scatter residual time, connect(1)
histogram residual, frequency
```

نتایج دستورات بالا فرض استقلال، نرمالیتی داده ها و همبستگی آنها را چک میکند. نمودارها در زیر آمده است.





## روش های اپیدمیولوژی و پژوهش با Stata

اگر به داده ها برگردید مقادیر مشاهده شده و مقادیر برازش شده و مقادیر باقی مانده مشخص است. اما آنچه با توجه به این داده ها و مدل بالا مشخص است این است که مدل تقریبا مناسب است! لذا مجدد باید در خواندن و درک لگ ها و ترند آنها بررسی مجدد نمود و مدل مناسب تری را انتخاب نمود. این داده ها صرفا برای تمرین با نرم افزار بوده است. اما علت ممکن است ناپایداری در داده ها باشد که در اولین گام باید داده ها را ایستا نمود.

	skindisease	time	lag1	lag2	_est_rm	_est_fm	fit	residual
1	1343	1	.	.	0	0	1507.195	-164.1951
2	1236	2	1343	.	1	0	1361.173	-125.1735
3	1401	3	1236	1343	1	0	1306.499	94.50073
4	1396	4	1401	1236	1	0	1319.976	76.02356
5	1457	5	1396	1401	1	0	1399.119	57.88147
6	1388	6	1457	1396	1	0	1454.335	-66.33533
7	1389	7	1388	1457	1	0	1391.955	-2.955322
8	1369	8	1389	1388	1	0	1414.628	-45.62805
9	1318	9	1369	1389	1	0	1317.005	.9952393
10	1354	10	1318	1369	1	0	1327.66	26.34045
11	1312	11	1354	1318	1	0	1298.976	13.02405
12	1370	12	1312	1354	1	0	1319.697	50.30347
13	1404	13	1370	1312	1	0	1360.145	43.85474
14	1295	14	1404	1370	1	1	1397.656	-102.656
15	1453	15	1295	1404	1	1	1410.161	42.83936
16	1427	16	1453	1295	1	1	1368.868	58.1322
17	1484	17	1427	1453	1	1	1437.31	46.68958
18	1421	18	1484	1427	1	1	1420.44	.5601807

برای بررسی ایستایی و ناپایداری داده ها از دستور boxcox skindisease استفاده می شود. نتایج زیر حاصل می شود.

Fitting full model

```
Iteration 0: log likelihood = -874.30894
Iteration 1: log likelihood = -874.30795
Iteration 2: log likelihood = -874.30795
```

```
Log likelihood = -874.30795
Number of obs = 144
LR chi2(0) = 0.00
Prob > chi2 = .
```

skindisease	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	1.050334	1.132811	0.93	0.354	-1.169934 3.270602

Estimates of scale-variant parameters

	Coef.
Notrans	
_cons	2069.225
/sigma	151.5282

Test H0:	Restricted log likelihood	LR statistic chi2	P-value Prob > chi2
theta = -1	-875.96541	3.31	0.069
theta = 0	-874.74045	0.87	0.352
theta = 1	-874.30894	0.00	0.965

در خروجی اگر تنها برابر با یک معنی دار شد یعنی نایستایی در داده ها وجود دارد. خروجی بالا نشان می دهد داده ها نایستا هستند. اگر معنی دار نشد یعنی ایستایی وجود دارد. لذا ممکن است دلیل مدل نامناسب پیش بینی شده نایستا بودن داده ها باشد. برای ایستا کردن داده اگر در دیتا ترند وجود داشته باشد مانند مثال بالا، بهترین کار اختلاف گرفتن است. و با اختلاف گرفتن ترند دیتا از بین می رود و داده ها ایستا می شوند (استراتژی جیکنز باکس). از آزمون های دیکی - فولر می توان برای آزمون پایایی داده ها استفاده نمود. این آزمون فرض اینکه میانگین داده ها صفر است یا خیر را آزمون می کند. از دستور زیر یا از منوی نرم افزار قابل انجام است.

```
. dfuller skindisease, lags(0)
```

Dickey-Fuller test for unit root		Number of obs = 143		
Test Statistic	Interpolated Dickey-Fuller			
	1% Critical Value	5% Critical Value	10% Critical Value	
Z(t)	-4.377	-3.496	-2.887	-2.577

MacKinnon approximate p-value for Z(t) = 0.0003

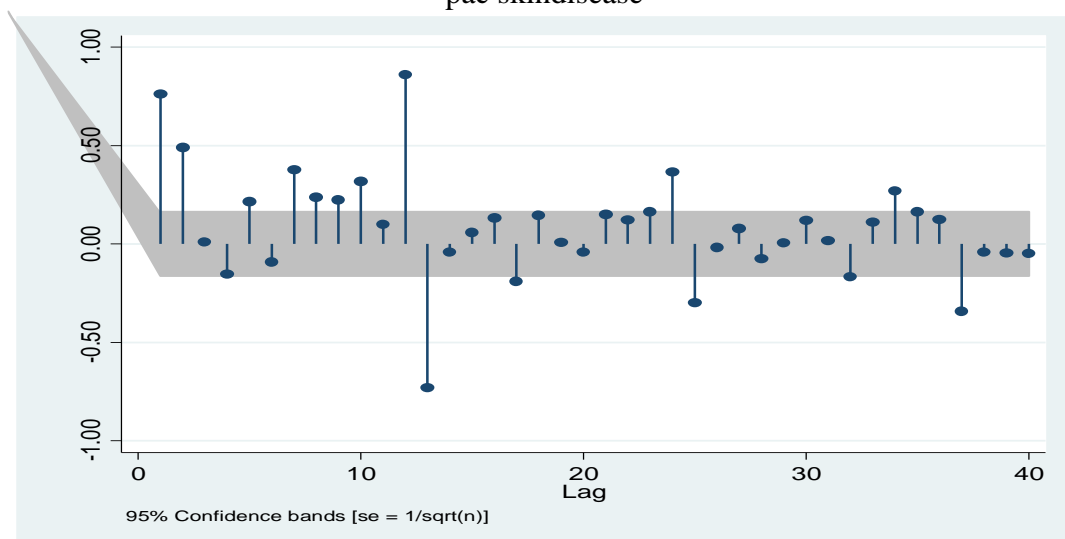
بنابراین در اینجا نتایج نشان می دهد که داده های مورد استفاده نایستا هستند.

برای ایستا کردن داده ها با استفاده از روش تفاضل گیری از دستورات زیر استفاده می شود.

```
tsline d2. skindisease
tsline d3. skindisease
tsline d3. skindisease
arima skindisease time
```

برای ایستایی در میانگین از دستورات `pac d.skindisease` و `ac d.skindisease` استفاده می شود.

`pac skindisease`



اینک نمودار بالا با نمودار قبل از ایستا کردن داده تفاوت هایی را در لگ های معنی دار دارد.



چنانچه مجدد با دستور `boxcox skindisease` مقدار تنا محاسبه شود روشن است که تنای یک معنی دار نیست.

Test H0:	Restricted log likelihood	LR statistic chi2	P-value Prob > chi2
theta = -1	-875.96541	3.31	0.069
theta = 0	-874.74045	0.87	0.352
theta = 1	-874.30894	0.00	0.965

برای ارزیابی مدل از معیار AIC استفاده می شود.

- با استفاده از معیار AIC مناسب ترین مدل را از بین مدل‌های قابل قبول برازش شده انتخاب می شود (مدلی که AIC کمتری داشته باشد مناسب تر است). مثال در جدول زیر آمده است.

$$AIC = -2 \ln(MLE) + 2(p+q)$$

**Table 1 Comparisons of tested models**

Model	Ljung-Box Q statistic	P value	AIC
ARIMA(0,2,1)	5.78	0.3869	117.7200
ARIMA(1,2,1)	7.97	0.0488	119.6435
ARIMA(0,3,1)	5.95	0.3113	116.7075
ARIMA(1,3,1)	8.46	0.0453	118.6647
ARIMA(2,3,1)	9.33	0.0260	120.1425

### خلاصه و نتیجه گیری:

در این فصل به معرفی سری های زمانی و مفاهیم مرتبط پرداخته شد. دستورات تحلیل و استفاده از روش تحلیل سری های زمانی با نرم افزار بیان گردید. مدل باکس جیکنز نیز معرفی گردید. با این نوع تحلیل با توجه به اطلاعات گذشته، می توان به پیش بینی وقایع در آینده پرداخت. این نوع تحلیل یکی از تحلیل های پیشرفته محسوب شده و کاربردی زیادی در مدیریت سلامت و پیش بینی وقایع مرتبط با سلامت و بیماریها در آینده دارد. البته با توجه به آمار و روند داده ها در گذشته، می توان پیش بینی را انجام داد. در اقتصاد و حسابداری این تحلیل نیز جایگاه ویژه ای دارد.

**تمرین ها :**

- ۱- سری های زمانی چه کاربردی در بهداشت دارند؟
- ۲- سری های زمانی چه پیش فرض هایی دارند؟
- ۳- گام های انجام یک مطالعه سری زمانی چیست؟
- ۴- ضریب خود همبستگی چه مفهومی دارد؟
- ۵- کاربرد دستورات نرم افزار در تحلیل یک سری زمانی و تفسیر نمودارهای مربوطه را مجدد بازخوانی نمایید. تفاوت دستورات و مدل ها و بهترین مدل برای انتخاب چیست؟
- ۶- پارامترهای مدل ARIMA را توضیح دهید.
- ۷- فایل مربوط به سری های زمانی را باز نمایید. دستورات را گام به گام برای تحلیل مذکور انجام دهید.
- ۸- فایل مربوط به گزارش بیماری های پوستی را یکبار به صورت گزارش ماهیانه و یکبار به صورت گزارش فصلی به نرم افزار معرفی نمایید. تحلیل و پیش بینی بیماریهای پوستی را انجام دهید.
- ۹- تفاوت های بین نتایج تحلیل ماهیانه و تحلیل سالیانه را بیان نمایید.
- ۱۰- تفاوت های بین نتایج تحلیل فصلی و تحلیل سالیانه را بیان نمایید.
- ۱۱- در بین سه تحلیل انجام شده کدام مدل برای داده ها مناسب تر است؟

## فصل هشتم: حجم نمونه و توان در مطالعات اپیدمیولوژیک .....

روش های نمونه گیری  
رویکردهای تعیین حجم نمونه  
محاسبه حجم نمونه  
محاسبه توان در انواع مطالعات  
کار با نرم افزار

### • اهداف:

#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند روش های نمونه گیری برای انواع مطالعات اپیدمیولوژی را بیان و استفاده نماید.
- کاربر بتواند برای انواع مطالعات اپیدمیولوژیک حجم نمونه را با نرم افزار تعیین کند.
- کاربر بتواند توان یا قدرت مطالعه را با توجه به حجم نمونه بکار رفته محاسبه نماید.



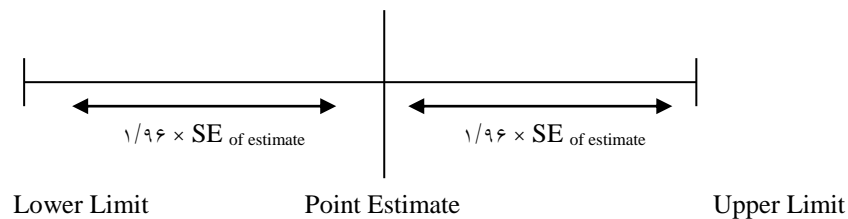
انواع روش های نمونه گیری :

- ۱- نمونه گیری احتمالی  
 Probability Sampling  
 تصادفی ساده (Simple Random S.)  
 طبقه بندی (Stratified S.)  
 دو مرحله ای (Two stage S)  
 خوشه ای (Cluster)  
 منظم (Systematic)  
 چند مرحله ای (Multi Stage S)
- ۲- نمونه گیری غیر احتمالی  
 NonProbability Sampling  
 هدفدار (Purposive S)  
 آسان (Convenience S)  
 سهمیه ای (Quota)  
 شکار - شکار مجدد  
 (Capture-Recapture)  
 گلوله برفی (Snow Ball S)  
 نمونه گیری مورد شایع (Modal Instance S)  
 نمونه گیری فرد ماهر (Expert S)  
 نمونه گیری قضاوتی (Judgmental S)

رویکرد محاسبه حجم نمونه: دقت (Precision) در برابر قدرت (Power)

دو رویکرد برای محاسبه حجم نمونه وجود دارد:

۱) رویکرد برای برآورد یک نسبت، میانگین، تفاوت، خطر نسبی و... همراه با درجاتی از دقت. به عنوان مثال، برای برآورد نسبت نوزادان واکسینه شده با شیوع ۱۰٪ این نکته را باید مد نظر قرار داد: **محدوده اطمینان مورد نیاز چه میزان باید باشد؟** پایه و اساس این مبحث محدوده اطمینان مورد نیاز برای بدست آوردن اثر و رابطه مورد مطالعه می باشد.



فاصله مقدار برآورد شده تا حد بالای محدوده اطمینان، ۱/۹۶ (به طور تقریبی ۲) برابر انحراف استاندارد (SE) می باشد. همین مساله در مورد حد پایین محدوده اطمینان نیز صادق می باشد. این بدان معنی است که به احتمال ۹۵٪ میزان واقعی (پارامتر جمعیت) در این محدوده برآورد شده قرار گرفته است. به عبارت دیگر  $1/96 \times SE \text{ of estimate}$  = دقت (Precision)

همانطور که می دانیم ۱/۹۶، Z-score مربوط به سطح معنی داری ۹۵٪ می باشد (خطای  $\alpha = 5\%$ ).

به عنوان مثال، فرمول برای محاسبه حجم نمونه در برآورد یک نسبت به صورت زیر می باشد:

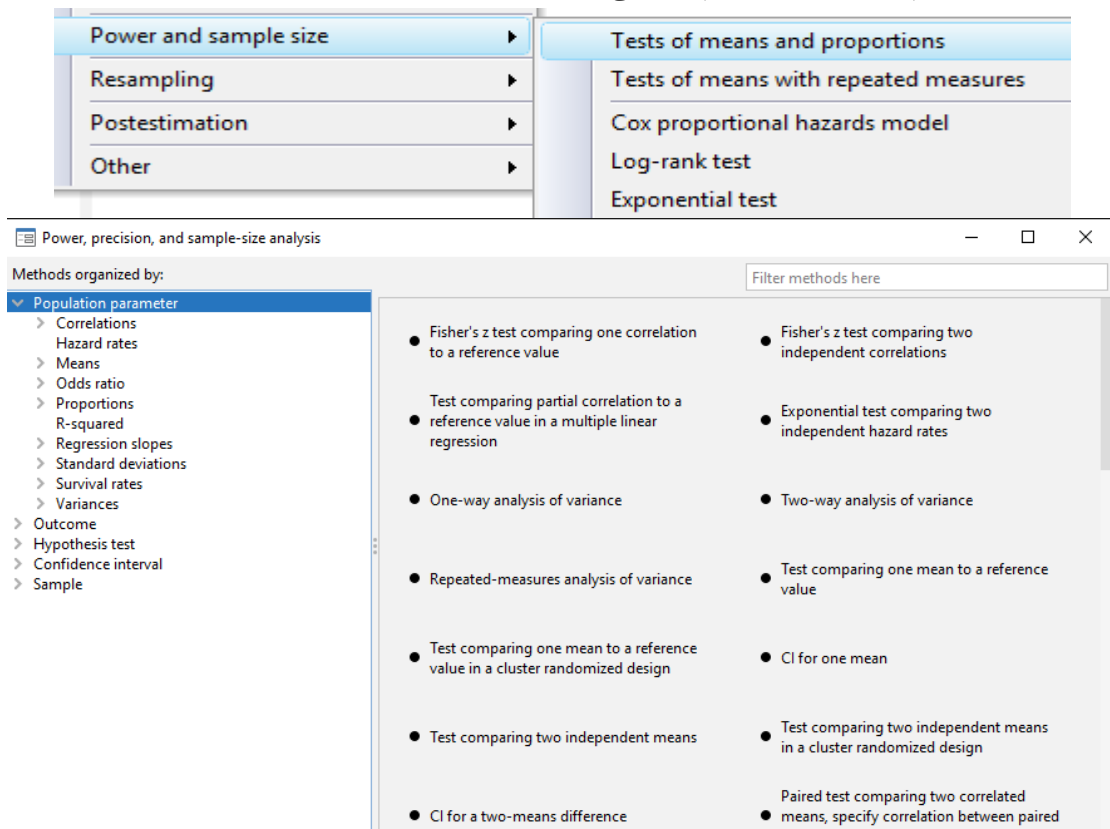
محدوده اطمینان ۹۵٪ برای شیوع (p) :  $1.96 \times \sqrt{p(1-p)/n}$  = دقت

دقت شود که برای استفاده از فرمول فوق مقدار p باید مشخص باشد که این مقدار بر اساس مطالعات قبلی ارزیابی می شود.

۲) رویکرد آزمون یک فرضیه. به عنوان مثال، برای مقایسه مدت زمان شیردهی بین گروه مداخله و شاهد، این نکته را باید مد نظر قرار داد: چقدر احتمال دارد که تفسیر نادرست از یک آزمون آماری معنی دار صورت گیرد؟

محاسبه توان، حجم نمونه و کار با نرم افزار:

برای محاسبه حجم نمونه و توان در نرم افزار می توان از مسیر زیر استفاده نمود.



بسته به نوع مطالعه، اندازه اثر، آزمون مورد استفاده، مقدار خطای نوع اول، توان مطالعه، گروه های مورد مطالعه، استقلال یا وابستگی اندازه اثر و گروه های مطالعه، نوع پیامد و طراحی مطالعه می توان حجم نمونه را تعیین نمود.

یا با دستور db sampsi کادر محاوره ای زیر را باز نمود.

The screenshot shows the 'sampszi' dialog box with the 'Input' section active. The 'Two-sample comparison of means' radio button is selected. The 'Mean one' and 'Mean two' fields are empty. The 'Std. deviation one' and 'Std. deviation two' fields both contain the value '1'. Other options like 'One-sample comparison of mean to hypothesized value' and 'Two-sample comparison of proportions' are unselected.

**در کادر بالا چهار گزینه برای تعیین حجم نمونه وجود دارد:**

مقایسه حجم نمونه برای مقایسه دو میانگین یا دو نسبت، مقایسه حجم نمونه برای یک میانگین یا نسبت با یک عدد مشخص. حجم نمونه در محاسبات بالا برای یک گروه تعیین می گردد وقتی قرار است دو گروه مقایسه شود تعداد نمونه در هر گروه برابر با حجم نمونه برآورد شده از طریق گزینه های بالا می باشد. در گزینه های بالا محقق باید میانگین یا نسبت و انحراف معیار میانگین و عدد مشخص را برای نرم افزار تعریف نماید. سپس با انتخاب گزینه option در کادر محاوره ای بالا می توان نوع خروجی مورد نظر (محاسبه حجم نمونه یا محاسبه توان)، مقدار خطای نوع اول، توان و یک دامنه یا دو دامنه بودن آزمون برای نرم افزار تعریف تا خروجی مورد حاصل گردد. در زیر این گزینه نمایش داده شده است.

The screenshot shows the 'Options' section of the 'sampszi' dialog box. The 'Compute sample size' radio button is selected. The 'Significance level (alpha)' field is empty, and the 'Power of the test' field contains '.90'. Under 'Sample-based calculations', 'Sample one size' and 'Sample two size' are both set to '100', and 'Ratio of sample sizes' is set to '1'. The 'Type of test' dropdown menu is open, showing 'Two-sided test' (selected) and 'One-sided test' as options.

**برآورد حجم نمونه برای مقایسه میانگین یک گروه با مقدار مشخص:** برای محاسبه حجم نمونه در طرحهایی که هدف آنها برآورد یک میانگین می باشد، از رابطه زیر استفاده میشود:

$$n = Z^2 \times \delta^2 / d^2$$

n: تعداد نمونه‌های مورد نیاز

Z: بیان کننده ضریب سطح اطمینان طرح در تعمیم نتایج به جامعه می باشد و بر اساس احتمال خطای نوع اول ( $\alpha$ ) مشخص می شود. مثلاً در صورتیکه  $\alpha$  برابر ۰/۰۵ باشد،  $Z_{1-\alpha/2}$  برابر با ۱/۹۶ می شود.

$\delta$ : انحراف معیار متغیر مورد نیاز در جامعه هدف

d: دقت (Precision) مدنظر است که نهایتاً جهت گزارش نتایج طرح و تعمیم آنها به جامعه، به صورت محدوده اطمینان گزارش خواهد شد. مقدار d معمولاً بر اساس مطالعات قبلی مشخص میگردد و هرگاه محقق اطلاع در مورد میزان فاصله اطمینان مطلوب نداشته باشد، می توان d را عددی بین ۰/۱۵ تا ۰/۳۲ $\delta$  قرار دهد. از آنجا که هر چه مقدار خطای قابل قبول مطالعه کمتر و پراکندگی متغیر مورد پژوهش در نمونه ها بیشتر باشد، حجم نمونه بیشتری مورد نیاز است، لذا Z و  $\delta$  با حجم نمونه رابطه مستقیم دارند. از طرفی هر چه فاصله اطمینان کمتری مورد نظر باشد، دقت مطالعه بیشتر خواهد بود (یعنی d کمتر) و نیاز به حجم نمونه بیشتری می باشد. لذا d با حجم نمونه رابطه عکس دارد. رابطه فوق صرفاً جهت محاسبه حجم نمونه در مطالعات مقطعی توصیفی کاربرد دارد.

به عنوان مثال برای محاسبه حجم نمونه در یک مطالعه برای مقایسه میانگین فشارخون آن جامعه با عدد ۱۴ ابتدا در نرم افزار میانگین و انحراف معیار که از مطالعات قبلی یا مطالعه پایلوت بدست آمد در قسمت کادر محاوره ای وارد می شود و حجم نمونه به شرح زیر بدست می آید.

The screenshot shows the 'sampsi' dialog box in Stata. The 'Options' tab is selected. Under the 'Input' section, the radio button for 'One-sample comparison of mean to hypothesized value' is checked. The 'Hypothesized' field contains the value 13, and the 'Postulated' field contains 14. The 'Std. deviation one' field contains 3. The other radio buttons for 'Two-sample comparison of means' and 'Two-sample comparison of proportions' are unselected. At the bottom, there are 'OK', 'Cancel', and 'Submit' buttons.

همانگونه که نتایج زیر نشان می دهد به حجم نمونه ۹۵ نفری در این پژوهش نیاز می باشد.  
`. sampsi 13 14, sd(3) onesample`

Estimated sample size for one-sample comparison of mean  
to hypothesized value

Test Ho: m = 13, where m is the mean in the population

Assumptions:

alpha = 0.0500 (two-sided)  
 power = 0.9000  
 alternative m = 14  
 sd = 3

Estimated required sample size:

n = 95

چنانچه در حالت دیگر محقق با اعداد بالا و حجم نمونه فوق، بخواهد توان مطالعه خویش را آزمون کند نتیجه زیر حاصل می شود. عبارتی توان ۹۰ درصدی بدست می آید.

Assumptions:

alpha = 0.0500 (two-sided)  
 alternative m = 14  
 sd = 3  
 sample size n = 95

Estimated power:

power = 0.9013

### حجم نمونه برای برآورد یک نسبت :

در مطالعاتی که هدف آنها برآورد نسبت (p) یک صفت مشخص در جامعه است، برای محاسبه حجم

$$n = z^2 \times P(1-P) / d^2$$

نمونه از رابطه زیر استفاده می شود:

p پیش فرضی از نسبت فراوانی وجود صفت مورد نظر در جامعه می باشد که بر اساس بررسی متون و مطالعات قبلی مشخص می شود.

سایر پیش فرضها مانند رابطه قبل می باشند. تنها، هرگاه محقق پیش فرضی در مورد میزان d مطلوب نداشته باشد، می تواند d را برابر با عددی بین ۰/۱P تا ۰/۲P انتخاب کند.

۱. هرگاه پژوهشگر بخواهد فراوانی چند متغیر با نسبتهای پیش فرض متفاوت را در یک مطالعه

تعیین کند، حجم نمونه بر مبنای نسبت پیش فرضی که به عدد ۰/۵۰٪ نزدیکتر است محاسبه می -

گردد. مثلاً در مطالعه ای جهت تعیین فراوانی انواع گروههای خونی، حجم نمونه بر اساس گروه

خونی ای که بیشترین فراوانی را دارد تعیین می شود.



۲. چنانچه محقق هیچ برآوردی از مقدار P نداشته باشد برای بدست آوردن بیشترین حجم نمونه کافی است مقدار آنرا برابر ۵۰٪ فرض کند.

۳. رابطه فوق صرفاً برای محاسبه حجم نمونه در مطالعات مقطعی توصیفی کاربرد دارد. برای محاسبه حجم نمونه در نرم افزار با داشتن نسبت 0.4 در مطالعه پایلوت و مقایسه آن با 0.2 حجم نمونه مورد نیاز ۵۵ نفر به شرح زیر خواهد بود.

Estimated sample size for one-sample comparison of proportion to hypothesized value

Test Ho:  $p = 0.4000$ , where  $p$  is the proportion in the population

Assumptions:

alpha = 0.0500 (two-sided)  
 power = 0.9000  
 alternative p = 0.2000

Estimated required sample size:

n = 55

حجم نمونه برای مقایسه دو میانگین

برای محاسبه حجم نمونه در مطالعاتی که هدف از انجام آنها مقایسه دو میانگین می باشد از رابطه زیر استفاده می شود:

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \delta^2}{d^2}$$

$\alpha$ : احتمال خطای نوع اول  $\beta$ : احتمال خطای نوع دوم  
 $\delta$ : انحراف معیار متغیر مورد نظر (معمولاً  $S_1^2$  و  $S_2^2$  را برابر فرض کرده و بنابراین به جای  $(S_1^2 + S_2^2)$  مقدار  $2S^2$  گذاشته می شود.

$d$ : حداقل اختلاف دارای ارزش کلینیکی هنگام مقایسه دو میانگین  
 میزان  $d$  را به چند روش می توان بدست آورد: انجام بررسی متون و استفاده از مطالعات قبلی  
 بر طبق مطالعه ای در گذشته در یک شهر، میانگین و انحراف معیار برای فشارخون سیستولیک در بیماران دیابتی نوع دو تحت درمان با متفورمین به ترتیب ۱۵۰ و ۴۵ بدست آمد. این ارقام برای کسانی که دیابتی نبودند اما به فشارخون بالا مبتلا بودند و دارو مصرف می کردند ۱۸۰ و ۲۵ بدست آمد. برای محاسبه حجم نمونه مقایسه دو میانگین در جامعه و انجام یک مطالعه تحلیلی مورد شاهد حجم نمونه مورد نیاز بر اساس محاسبات زیر ۳۱ نفر در هر گروه خواهد بود.

**حجم نمونه برای مقایسه دو نسبت و حجم نمونه در مطالعات همگروهی و مورد شاهد:**

هرگاه هدف از انجام مطالعه‌ای مقایسه دو نسبت باشد، برای محاسبه حجم نمونه از رابطه زیر استفاده می‌شود:

$$n = \frac{[Z_{1-\alpha/2} \sqrt{2\bar{P}(1-\bar{P})} + Z_{1-\beta} \sqrt{P_0(1-P_0) + P_1(1-P_1)}]^2}{(P_1 - P_0)^2}$$

$\bar{P}$ : میانگین حسابی دو نسبت

$P_0$ : در مطالعات با جهت رو به عقب، بیانگر احتمال مواجهه یافتن با عامل خطر در گروه کنترل و در مطالعات با جهت رو به جلو، بیانگر احتمال ایجاد بیماری در گروه مواجهه نیافته از جامعه هدف است.  
 $P_1$ : در مطالعات با جهت رو به عقب، بیانگر احتمال مواجهه یافتن با عامل خطر در گروه مورد و در مطالعات با جهت رو به جلو، بیانگر احتمال ایجاد بیماری در گروه مواجهه یافته از جامعه هدف است.

نکته: در جایی که  $P_1$  و  $P_0$  تقریباً برابر باشند، می‌توان با جایگذاری  $\bar{P}$  به جای  $P_0$  و  $P_1$ ، محاسبات بالا را ساده تر کرد:

$$n = \frac{2 \bar{P} (1-\bar{P})(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(p_1 - p_0)^2}$$

- چنانچه محقق برآوردی از احتمال مواجهه یافتن (یا ایجاد بیماری) در گروه مورد را نداشته باشد، می‌توان  $P_1$  را با استفاده از  $P_0$  و Odds Ratio یا در مطالعات مورد شاهدی یا RR در مطالعات همگروهی و کارآزمایی بالینی بدست آورد:

$$P_1 = \frac{P_0 \times OR}{1 + P_0 (OR - 1)} \quad P_1 = \frac{P_0 \times RR}{1 + P_0 (RR - 1)}$$

رابطه فوق برای محاسبه حجم نمونه در مطالعات مقطعی-تحلیلی، مورد-شاهدی، همگروهی و کارآزمایی بالینی مورد استفاده قرار می‌گیرد. مانند رابطه قبل، تعداد افراد مورد مطالعه در هر دو گروه مورد بررسی برابر حجم نمونه محاسبه شده در نظر گرفته می‌شود.

برای محاسبه حجم نمونه با نرم افزار کادر محاوره ای زیر باز و به ترتیب زیر اقدام می شود.

The screenshot shows the 'sampsi' dialog box in Stata. The 'Main' tab is selected. Under the 'Input' section, the radio button for 'Two-sample comparison of proportions (values in [0,1])' is checked. The input fields are filled with the following values: Mean one: 150, Mean two: 180, Std. deviation one: 45, Std. deviation two: 25, Proportion one: 0.5, Proportion two: 0.2, Hypothesized: 0.4, and Postulated: 0.2. At the bottom, there are buttons for 'OK', 'Cancel', and 'Submit'.

بر طبق مطالعه ای در یک شهر شیوع کم خونی در مادران باردار 0.5 و مادران غیرباردار 0.2 بدست آمد. حجم نمونه مورد نیاز برای اجرای یک مطالعه کوهورت برای تفاوت این دو نسبت در سطح اطمینان ۹۹ درصد و توان ۸۵ درصد با وارد نمودن در کادر محاوره ای بالا برای هر گروه ۴۲ نفر خواهد بود.

**Assumptions:**

```
alpha = 0.1000 (two-sided)
power = 0.8500
p1 = 0.5000
p2 = 0.2000
n2/n1 = 1.00
```

**Estimated required sample sizes:**

```
n1 = 42
n2 = 42
```

حجم نمونه برای اندازه گیری های تکراری برای میانگین و همچنین تحلیل کاکس و یکسان بودن خطر در طی زمان های متعدد نیز با نرم افزار قابل محاسبه است.

Power Analysis

کار با نرم افزار و محاسبه توان مطالعه

از منوی **Statistics > Power, precision, and sample size** می توان محاسبات را انجام داد. وقتی که  $m_0$  میانگین در فرضیه صفر (فرض شده) یا مقدار میانگین تحت فرضیه صفر است و  $m_a$  میانگین جایگزین (هدف) یا مقدار میانگین تحت فرضیه جایگزین است.  $m_0$  و  $m_a$  هر یک ممکن است به عنوان یک عدد یا به عنوان لیستی از مقادیر داخل پرانتز مشخص شوند.

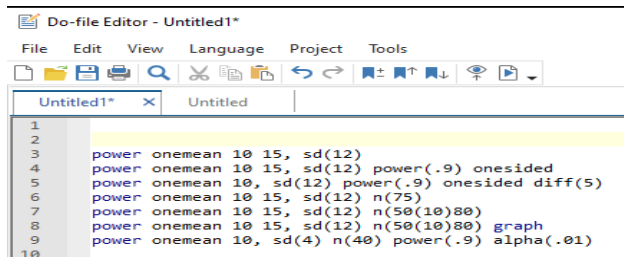
توان مطالعه را می توان با دستور `power onemean m0, n(numlist) power(numlist) [options]` محاسبه نمود. در قسمت `option` دستور، می توان مقادیر خطای نوع اول، توان، خطای نوع دوم، حجم نمونه مورد نیاز، مقدار اختلاف میانگین در فرضیه صفر و جایگزین، انحراف معیار، حجم نمونه در جامعه محدود، یک طرفه بودن فرضیه را نوشت. دستورات مورد نیاز در این گزینه در جدول زیر آمده است.

options	Description
Main	
* <u>a</u> lpha( <i>numlist</i> )	significance level; default is <code>alpha(0.05)</code>
* <u>p</u> ower( <i>numlist</i> )	power; default is <code>power(0.8)</code>
* <u>b</u> eta( <i>numlist</i> )	probability of type II error; default is <code>beta(0.2)</code>
* <u>n</u> ( <i>numlist</i> )	sample size; required to compute power or effect size
<u>n</u> fractional	allow fractional sample size
* <u>d</u> iff( <i>numlist</i> )	difference between the alternative mean and the null mean, $m_a - m_0$ ; specify instead of the alternative mean $m_a$
* <u>s</u> d( <i>numlist</i> )	standard deviation; default is <code>sd(1)</code>
<u>k</u> nownsd	request computation assuming a known standard deviation; default is to assume an unknown standard deviation
* <u>f</u> pc( <i>numlist</i> )	finite population correction (FPC) as a sampling rate or as a population size
<u>d</u> irection( <u>u</u> pper  <u>l</u> ower)	direction of the effect for effect-size determination; default is <code>direction(upper)</code> , which means that the postulated value of the parameter is larger than the hypothesized value
<u>o</u> nesided	one-sided test; default is two sided
<u>p</u> arallel	treat number lists in starred options or in command arguments as parallel when multiple values per option or argument are specified (do not enumerate all possible combinations of values)

دستورات بالا را به صورت زیر برای محاسبه با نرم افزار استفاده می کنند.

```
Power onemean 1 2
power onemean 1, diff(1)
power onemean 1 2, power(0.9)
power onemean 1 2, sd(2) knownsd
power onemean 1 2, sd(2) knownsd fpc(200)
power onemean 1 2, alpha(0.01) onesided
power onemean 1 (1.5(0.5)3), power(0.8 0.9) graph
```

اندازه نمونه برای آزمون دو طرفه  $H_0: \mu = 10$  در مقابل  $H_a: \mu \neq 10$  با میانگین جایگزین  $m_a = 15$ ، و انحراف استاندارد ۱۲ با استفاده از توان های پیش فرض  $0.8$  و  $0.9$  و سطح معنی داری  $\alpha = 0.05$  را با دستورات ذخیره شده در `do-file` زیر اجرا نمایید.



وقتی حجم نمونه ۷۵ نفر باشد توان مطالعه چقدر است؟

```
. power onemean 10 15, sd(12)
```

Performing iteration ...

Estimated sample size for a one-sample mean test

t test

H0: m = m0 versus Ha: m != m0

Study parameters:

```
alpha = 0.0500
power = 0.8000
delta = 0.4167
m0 = 10.0000
ma = 15.0000
sd = 12.0000
```

Estimated sample size:

```
N = 48
```

```
. power onemean 10 15, sd(12) n(75)
```

Estimated power for a one-sample mean test

t test

H0: m = m0 versus Ha: m != m0

Study parameters:

```
alpha = 0.0500
N = 75
delta = 0.4167
m0 = 10.0000
ma = 15.0000
sd = 12.0000
```

Estimated power:

```
power = 0.9453
```

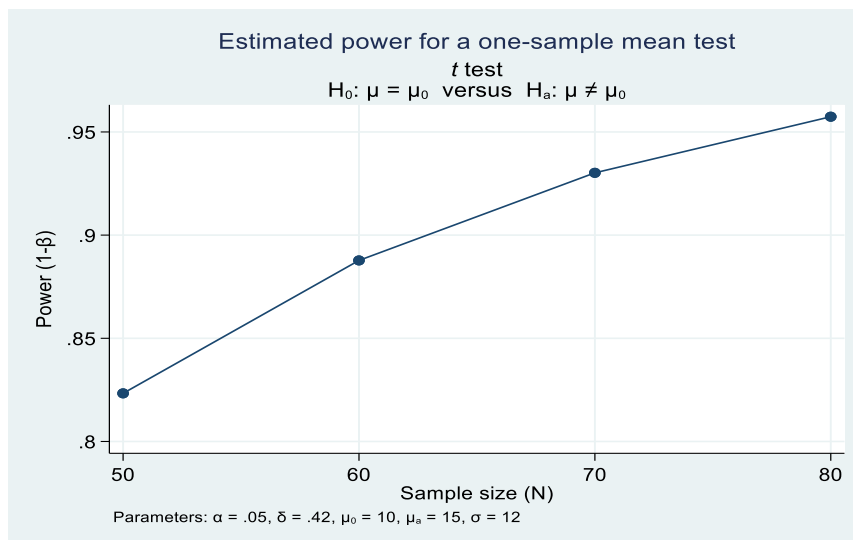
```
. power onemean 10 15, sd(12) n(50(10)80)
```

Estimated power for a one-sample mean test

t test

H0: m = m0 versus Ha: m != m0

alpha	power	N	delta	m0	ma	sd
.05	.8233	50	.4167	10	15	12
.05	.8877	60	.4167	10	15	12
.05	.9302	70	.4167	10	15	12
.05	.9574	80	.4167	10	15	12



## ..... فصل نهم: متاآنالیز

مقدمه ای بر متا آنالیز  
دستورات متاآنالیز برای متغیرهای کمی و کیفی  
کار با نرم افزار Stata  
تمرین ها

### • اهداف:

**در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:**

- کاربر بتواند هدف از انجام متاآنالیز را بیان نماید.
- مراحل انجام متاآنالیز را تشریح نماید.
- متاآنالیز را برای مطالعات مختلف انجام و تفسیر نماید.
- انواع نمودارهای رایج در متاآنالیز را ترسیم و تفسیر نماید.

## مقدمه ای بر متا آنالیز Meta Analysis

قبل از انجام متاآنالیز، انجام مطالعه مرور نظام مند ضروری است. در این مرور از روش های نظام مند و شفاف جستجو، برای مرور متونی که در رابطه با سؤال مورد نظر می باشند، با استراتژی مناسب جستجو (PICOTS<sup>۲۲</sup>) انجام می شود. پس از اتمام جستجو با حساسیت و ویژگی مورد قبول، منابع شناسایی شده انتخاب و مورد ارزیابی نقادانه قرار می گیرند.

مرور های نظام مند، پژوهش های انجام شده در رابطه با یک موضوع خاص را بصورت مختصر و مفید در اختیار می گذارند. از ویژگی های مرور نظام مند این است که از آنجا که کلیه مراحل و روش های آن از قبل طراحی شده و شفاف است، اگر افراد دیگری با همان مراحل و روش ها اقدام به مرور همان مجموعه از داده ها نمایند، به نتایج یکسانی می رسند و به عبارتی دیگر مرور های نظام مند قابلیت تولید مجدد را داشته و از پایایی مناسبی برخوردارند. مرور منظم می تواند منجر به (نه لزوماً) متا آنالیز گردد. متاآنالیز مشتمل بر روش های آماری است که به کمک این روش ها سعی می شود تا اطلاعات بدست آمده در مرور ساختاریافته ترکیب و نتیجه گیری مبتنی بر مستندات ارایه شود.

افزایش توان آماری (Statistical Power) و افزایش دقت (توان آماری عبارت است از احتمال کشف یک رابطه آماری معنی دار، در صورت وجود آن رابطه. توان آماری رابطه مستقیمی با حجم نمونه دارد و با افزایش حجم نمونه دامنه اطمینان دقیق تر و باریک تر می شود.)، پاسخ به سؤالاتی که مطالعات اولیه نمی توانند به آنها پاسخ دهند. و ارزیابی دقیق تر تفاوت های موجود بین یافته های مطالعات، از کارکردها و نقش های اصلی متاآنالیز می باشند. سایت و کتاب راهنمای کاکرین (Cochrane Handbook) برای انجام یک تحقیق مروری به شیوه ساختاریافته و متاآنالیز شش مرحله را پیشنهاد داده است:

۱) تعریف دقیق موضوع و سؤال پژوهش (۲) انتخاب اولیه مقالات در قالب یک بررسی ساختاریافته اولیه (۳) بررسی اعتبار علمی مقالات و منابع (۴) استخراج اطلاعات (۵) تحلیل آماری و ارائه آن (۶) تفسیر و بحث نتایج. در این کتاب به مرحله ۵ یعنی تحلیل آماری و انجام متاآنالیز با ذکر یک مثال و استفاده از دستورات نرم افزار پرداخته می شود.

برای کسب اطلاعات بیشتر درباره مرور سیستماتیک و متاآنالیز به کتاب موجود در باب این موضوع مخصوصاً کتاب انگلیسی متاآنالیز در Stata و کتاب فارسی آقای دکتر حقدوست و همکاران مراجعه شود. به طور کلی دو نوع متاآنالیز در مطالعات انجام می شود. توصیفی و تحلیلی. در مطالعات توصیفی، هدف صرفاً برآورد یک پارامتر ساده است. رایج ترین آماره های مورد سنجش در مطالعات توصیفی را به دو دسته میانگین و فراوانی و درصد تقسیم می کنند.

در متاآنالیز در مطالعات تحلیلی هدف با متاآنالیز مطالعات توصیفی متفاوت است و در این دسته مطالعات سعی می شود که رابطه بین متغیرها و اندازه اثر مورد بررسی و تحلیل قرار گیرد.



شاخص ها در متاآنالیز:

شاخص های آماری مورد نیاز بر حسب نوع داده ها و نوع مدل که در متاآنالیز استفاده می شود در جدول زیر آمده است.

نوع داده	شاخص آماری	روش آماری	نوع
		Woolf	ثابت
		Mantel-Haenzel	ثابت
		Peto	ثابت
	نسبت شانس (Odds Ratio)	Der simonian-Laird	تصادفی
		Meta-Regression	تصادفی
دو حالته		Mantel-Haenzel	ثابت
		Inverse Variance	ثابت
	خطر نسبی (Risk Ratio)	Der Simonian-laird	تصادفی
		Meta-Regression	تصادفی
عددی		Mantel-Haenzel	ثابت
		Inverse Variance	ثابت
	تفاضل میانگین ها و یا تفاضل میانگین های استاندارد	Der Simonian-laird	تصادفی
		Meta-Regression	تصادفی

به هر گونه عدم یکنواختی در نتایج مطالعات وارد شده در متاآنالیز ناهمگونی (Heterogeneity) گفته می شود. که ممکن است منشاء بالینی یا متدولوژیک داشته باشد. مهمترین دلایل ناهمگونی تفاوت در نمونه های مورد مطالعه، تفاوت در نوع درمان ها و تفاوت در متغیرهای وابسته است. برای بررسی این ناهمگونی از آزمون  $\tau^2$  (tau<sup>2</sup>) استفاده می شود. واریانس کل مشاهده شده در مطالعات متاآنالیز دارای دو منشاء است. یکی واریانس درون مطالعات (Within Studies) و دیگری واریانس بین مطالعات (Between Studies). واریانس درون مطالعات حاصل تفاوت بین داده های هر مطالعه است و واریانس بین مطالعات حاصل اختلاف بین نتایج مطالعات مختلف است. آزمون  $\tau^2$  واریانس بین مطالعات را بررسی می کند. شاخص  $I^2$  مقدار ناهمگونی را به صورت کمی با درصد مشخص می کند. دامنه تغییرات شاخص  $I^2$  معمولاً بین صفر تا ۱۰۰٪ می باشد و مقادیر منفی آن را صفر در نظر می گیرند. وقتی مقدار عددی این شاخص صفر باشد، مبین همگونی نتایج است و هر چه مقدار آن بزرگتر شود، مبین میزان ناهمگونی نتایج مطالعات می باشد.

Random Effect and Fixed Effect

در مدل ثابت، فرض بر آن است که پارامتر مورد بررسی در تمام مطالعات اولیه مقدار حقیقی ثابتی داشته و تفاوت بین مقادیر محاسبه شده در مطالعات مختلف، تنها به دلیل شانس و تصادف حاصل از نمونه گیریهای مکرر (Sampling Variation) بوده است. در مدل تصادفی فرض بر این است که پارامتر اندازه-

گیری شده خود توزیع نرمالی داشته و هر مطالعه مقداری از آن را اندازه گرفته است. لذا در مدل تصادفی، تفاوت‌های دیده شده در بین نتایج مطالعات اولیه دو ریشه دارند؛ یکی به دلیل تصادف حاصل از نمونه-گیری‌های مکرر و دیگری به دلیل تغییرات تصادفی مقدار واقعی پارامتر در مطالعات مختلف.

### در گزارش نتایج متاآنالیز چند نمودار مهم است:

۱- نمودار انباشت (Forest Plot): در این نمودار، اطلاعات تک تک مطالعات وارد شده به متاآنالیز و برآیند نهایی آنها نمایش داده می شود. برآورد هر مطالعه با یک مربع سیاه رنگ و حدود اطمینان ۹۵٪ برآورد با یک خط عرضی نمایش داده می شود. اندازه مربع سیاه رنگ، وزن هر مطالعه است. در پایین نمودار انباشت، یک لوزی ترسیم می گردد که در واقع مقدار برآیند وزن داده شده مربع‌ها است. قطر افقی لوزی نیز نشان دهنده دامنه اطمینان ۹۵٪ برآیند شاخص مورد مطالعه است. معمولاً دو خط عمود برهم رسم می شود؛ یکی خط عمودی منقطع که منطبق بر محور عمودی لوزی است و نشان می دهد که در کل مقدار شاخص مورد مطالعه در متاآنالیز چقدر است و به عبارت دیگر برآیند کلی متاآنالیز را نشان می دهد. خط عمود ممتد دیگر خطی است که مبین فرضیه صفر یا فقدان اثر (حالت خنثی) می باشد که مقدار عددی آن برای خطر نسبی (RR) و نسبت شانس (OR) یک و برای اختلاف خطر (RD)، شیوع، بروز و میانگین‌ها صفر است.

۲- نمودار کیفی (Funnel Plot): یک نمودار پراکنش ساده است که مقدار شاخص برآورد شده از هر مطالعه را نسبت به حجم نمونه یا دقت (precision) آن مطالعه ترسیم می کند. معمولاً شاخص در محور افقی (X) و حجم نمونه یا دقت مطالعه در محور عمودی (y) نمایش داده می شود.

### متاآنالیز با Stata:

در نسخه ۱۷ نرم افزار استاتا، از منوی statistics میتوان گزینه Meta Analysis را انتخاب و جعبه محاوره ای معرفی متغیرها را بر حسب نوع متاآنالیز تکمیل و اجرا نمود. این جعبه در زیر قابل مشاهده است.

ابتدا بسته به اینکه چه نوع متاآنالیزی قرار است انجام شود و مقایسه اندازه اثر برای دو گروه با پیامد پیوسته یا کیفی دوحالته است یا اینکه متاآنالیز اندازه اثر به صورت عمومی مد نظر است یکی از سه حالت زیر باید انتخاب و تعریف گردد.

Declare meta-analysis data

- Compute and declare effect sizes for two-group comparison of continuous outcomes
- Compute and declare effect sizes for two-group comparison of binary outcomes
- Declare generic, precomputed effect sizes (in the metric closest to normality)

متاآنالیز متغیرهای پیوسته

فایل meta2 را باز نمائید. و متغیرها را به شکل زیر ثبت نمائید.

در این فایل به متاآنالیز انجام شده در سال ۱۹۹۵ توسط وینترا (weintraub) و همکاران در خصوص اثر آنتی هیستامین در سرماخوردگی و شدت عطسه و آبریزش بینی پرداخته می شود. در این مطالعه بعد از مصرف یک روزه آنتی هیستامین در گروه مداخله و مقایسه با گروه پلاسبو، میانگین استاندارد شده شدت عطسه و آبریزش بینی و انحراف معیار آن در هر دو گروه، بوسیله مقیاس سنجش کمی اندازه گیری شد. اطلاعات مطالعات بدست آمده مشتمل بر حجم نمونه، میانگین و انحراف معیار در هر یک از گروه های مداخله و کنترل در جدول زیر نشان داده شده است. فایل meta2 این اطلاعات را نشان می دهد.

چنانچه گزینه submit را کلیک نمائید در واقع دستور زیر اجرا می شود:

meta esize n1 mean1 sd1 n2 mean2 sd2

خروجی زیر را مشاهده می کنید.

```
. meta esize n1 mean1 sd1 n2 mean2 sd2

Meta-analysis setting information

Study information
  No. of studies: 9
  Study label: Generic
  Study size: _meta_studysize
  Summary data: n1 mean1 sd1 n2 mean2 sd2

Effect size
  Type: hedgesg
  Label: Hedges's g
  Variable: _meta_es
  Bias correction: Approximate

Precision
  Std. err.: _meta_se
  Std. err. adj.: None
  CI: [_meta_cil, _meta_ciu]
  CI level: 95%

Model and method
  Model: Random effects
  Method: REML
```

می توان نوع مدل تصادفی، ثابت یا ترکیب هر دو را از قسمت مدل انتخاب نمائید.

Main	if/in	Model	Options
Meta-analysis model			
<input type="radio"/> Random effects <input checked="" type="radio"/> Common effect <input type="radio"/> Fixed effects		Method: <input type="text" value="Inverse-variance"/>	

سپس از قسمت summary نوع مدل ثابت را انتخاب و دستور اجرا می گردد. می توان متانالیز بر حسب زیرگروه ها نیز انجام داد که در قسمت مربوطه میتوان متغیر گروه بندی شده را معرفی نمود. نتایج به صورت زیر است.

```
. meta summarize, fixed(invvariance)

Effect-size label: Hedges's g
Effect size: _meta_es
Std. err.: _meta_se
Study label: id

Meta-analysis summary
Fixed-effects model
Method: Inverse-variance

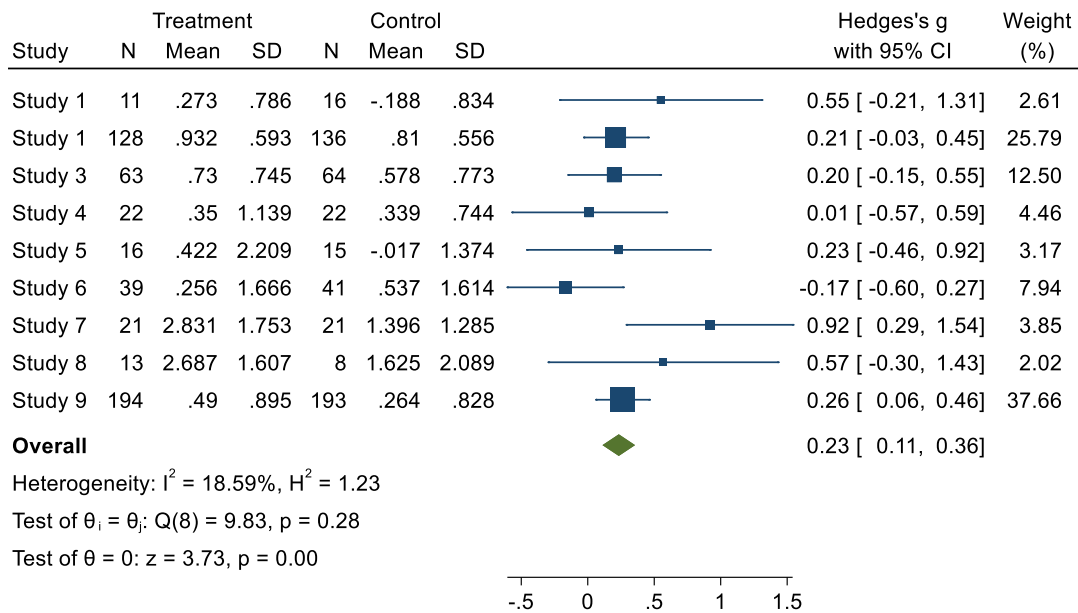
Number of studies = 9
Heterogeneity:
  I2 (%) = 18.59
  H2 = 1.23
```

Study	Hedges's g	[95% conf. interval]	% weight
Study 1	0.548	-0.210 1.307	2.61
Study 1	0.212	-0.030 0.453	25.79
Study 3	0.199	-0.148 0.546	12.50
Study 4	0.011	-0.569 0.592	4.46
Study 5	0.231	-0.458 0.919	3.17
Study 6	-0.170	-0.605 0.265	7.94
Study 7	0.916	0.291 1.541	3.85
Study 8	0.567	-0.296 1.429	2.02
Study 9	0.262	0.062 0.461	37.66
theta	0.233	0.111 0.356	

```
Test of theta = 0: z = 3.73 Prob > |z| = 0.0002
Test of homogeneity: Q = chi2(8) = 9.83 Prob > Q = 0.2774
```

اندازه اثر ۹ مطالعه ترکیب گردیده و برابند کلی نتایج نشان می دهد مصرف آنتی هیستامین در ۲۴ ساعت اول درمان می تواند شدت علائم سرماخوردگی را به میزان قابل توجهی کاهش دهد. این کاهش از نظر آماری معنی دار بوده است. اختلاف بین مطالعات یا یکنواختی آنها، نیز آزمون گردیده و ناشی از تغییرات تصادفی و بکار بردن مدل ثابت است.

در قسمت بعد با انتخاب گزینه Forest Plot نمودار انباشت را به صورت زیر می توان ترسیم نمود.



Fixed-effects inverse-variance model

با انتخاب گزینه های بعدی می توان ارزیابی هموژنیته و ترسیم نمودار تورش نشر را اجرا نمود.

### متا آنالیز متغیرهای کیفی دو حالته

مثال مورد کاربرد درباره متاآنالیز مطالعات مختلف برای تعیین اثر توصیه های پزشکان برای ترک و توقف سیگار کشیدن در افراد است. در سال ۱۹۹۷ توسط سیلاچی (silagy) و همکاران مطالعه مرور نظام مندی با جمع آوری کارآزمایی های بالینی انجام شده درباره موضوع مورد اشاره انجام گردید. ۱۶ مطالعه در این خصوص جمع آوری گردید. جدول زیر این مطالعات را نشان می دهد. این مطالعات در فایل meta1 نیز پیوست می باشند. r1 و r2 هر کدام، جمع گروه های مداخله و شاهد مطالعه است.  $\alpha$  نیز هر کدام وقوع یا عدم وقوع واقعه در گروه های مداخله و شاهد را نشان می دهند. از پانل معرفی متغیرها به نرم افزار گزینه دوم به صورت زیر انتخاب می شود.

Note: Multivariate meta-analysis does not require any setup. Proceed to the Multivariate pane.

Declare meta-analysis data

- Compute and declare effect sizes for two-group comparison of continuous outcomes
- Compute and declare effect sizes for two-group comparison of binary outcomes
- Declare generic, precomputed effect sizes (in the metric closest to normality)

مانند گام های متاآنالیز در مثال قبلی، مراحل اجرا می گردد.

خروجی در زیر قابل مشاهده است.

```
. meta esize a r1 c r2

Meta-analysis setting information

Study information
  No. of studies: 16
  Study label: Generic
  Study size: _meta_studysize
  Summary data: a r1 c r2

Effect size
  Type: lnoratio
  Label: Log odds-ratio
  Variable: _meta_es
  Zero-cells adj.: None; no zero cells

Precision
  Std. err.: _meta_se
  CI: [_meta_cil, _meta_ciu]
  CI level: 95%

Model and method
  Model: Random effects
  Method: REML

. meta summarize, fixed(mhaenszel)

Effect-size label: Log odds-ratio
Effect size: _meta_es
Std. err.: _meta_se

Meta-analysis summary
Fixed-effects model
Method: Mantel-Haenszel

Number of studies = 16
Heterogeneity:
  I2 (%) = 24.25
  H2 = 1.32
```

Study	Log odds-ratio	[95% conf. interval]		% weight
Study 1	0.019	-2.766	2.804	0.43
Study 2	0.108	-1.237	1.453	1.79
Study 3	1.099	0.074	2.124	2.14
Study 4	0.020	-1.136	1.177	2.52
Study 5	-0.047	-1.204	1.110	2.58
Study 6	1.267	0.336	2.197	3.06
Study 7	0.704	-0.191	1.599	3.29
Study 8	1.518	0.743	2.293	3.34
Study 9	0.637	-0.141	1.415	4.24
Study 10	0.001	-0.873	0.875	4.44
Study 11	0.541	-0.180	1.262	5.27
Study 12	0.847	0.273	1.421	7.42
Study 13	0.518	-0.063	1.099	7.98
Study 14	0.459	0.014	0.904	13.85
Study 15	0.062	-0.396	0.520	15.71
Study 16	0.353	-0.008	0.715	21.94
theta	0.523	0.358	0.688	

```
Test of theta = 0: z = 6.21 Prob > |z| = 0.0000
Test of homogeneity: Q = chi2(15) = 19.80 Prob > Q = 0.1796
```

نسبت های خطر برای مطالعات بالا و نسبت خطر یک کاسه شده (pooled RR) با روش مانتل – هنزل و روش ثابت محاسبه و در قالب جدول بالا گزارش شده است. نتیجه نشان داده است که مداخله برای ترک سیگار موثر و معنی دار بوده است. هتروژینیتی مطالعات ارزیابی و به صورت زیر گزارش شده است. مقدار پی آن معنی دار نشده است. که این عدم معنی داری نشان می دهد که اختلاف بین مطالعات بوسیله تغییرات تصادفی قابل توجیه می باشد. به عبارت دیگر می توان از مدل ثابت استفاده نمود.

## متاآنالیز با برنامه نویسی و نوشتن دستورات

برای کار با نرم افزار ابتدا محقق باید سه دستور زیر را یک به یک اجرا نموده تا فایل های مربوط به دستورات متاآنالیز را از اینترنت دانلود و در یک دایرکتوری به نام c:\ado\plus ذخیره شوند. برای انجام متاآنالیز نیاز به نصب این دستورات است.

```
net from http://www.Stata-press.com/data/mais
net install mais
spinst_mais
```

وقتی دستورات بالا اجرا می شوند نتایج زیر نمایش داده می شود.

```
. net from http://www.Stata-press.com/data/mais

http://www.Stata-press.com/data/mais/
Meta-Analysis in Stata: An Updated Collection from the Stata Journal
```

Edited by Jonathan A. C. Sterne

```
PACKAGES you could -net describe-:
  mais          Files associated with mais
```

```
.
. net install mais
checking mais consistency and verifying not already installed...
installing into C:\Users\Tehran\ado\plus\...
installation complete.

.
. spinst_mais
Installing ...
  package sbe24_3 from http://www.stata-journal.com/software/sj9-2
  package sbe22_1 from http://www.stata-journal.com/software/sj9-1
  package sbe28_1 from http://www.stata.com/stb/stb56
  package sbe23_1 from http://www.stata-journal.com/software/sj8-4
  package st0061 from http://www.stata-journal.com/software/sj4-2
  package gr0033_1 from http://www.stata-journal.com/software/sj9-2
  package sbe19_6 from http://www.stata-journal.com/software/sj9-2
  package sbe39_2 from http://www.stata.com/stb/stb61
  package st0163 from http://www.stata-journal.com/software/sj9-2
  package st0096_2 from http://www.stata-journal.com/software/sj9-2
  package st0157 from http://www.stata-journal.com/software/sj9-1
  package st0156 from http://www.stata-journal.com/software/sj9-1
Installation complete.
```

خروجی بالا نشان می دهد که بسته دستورات متاآنالیز قبلا در نرم افزار سیستم مورد استفاده وجود داشته و هم اینک این دستورات به روز رسانی و به صورت کامل نصب شده اند. دستورات اصلی در نرم افزار Stata برای انجام متاآنالیز metan و metacum و metap می باشند. دستور metan تحلیل اصلی را برای مطالعات با پیامدهای دو حالت و پیوسته را انجام می دهد.

**برای کار با نرم افزار و متاآنالیز، اولین گام** آشنایی و معرفی ساختار داده ها است. در متاآنالیز مطالعات مختلف (k مطالعه) وارد می شوند. چنانچه پیامد مطالعات متغیر دو حالت باشد ساختار داده ها به شکل جدول الف و چنانچه پیامد مطالعه با متغیر کمی پیوسته باشد ساختار داده ها به شکل جدول ب به شرح زیر نمایش داده می شوند.

**جدول الف) ساختار داده های دو حالت (Binary data)**

	Event	no event
Intervention	ai	bi
Control	ci	di

**جدول ب) ساختار داده های عددی یا کمی پیوسته (Continuous data)**

	حجم نمونه گروه ها	میانگین متغیر پاسخ	انحراف معیار
Intervention	n1i	m1i	sd1i
Control	n2i	m2i	sd2i

**کاربرد دستورات و انجام متآنالیز با نرم افزار**

برای کار با نرم افزار و انجام متآنالیز ۵ مثال به شرح زیر نحوه انجام متآنالیز توصیفی و تحلیلی در مطالعات مورد - شاهد، کوهورت و کارآزمایی های بالینی با داده های مختلف را نشان می دهد. در طی این مثال ها دستورات مختلف انجام متآنالیز بیان می شود.

**مثال اول:**

کار با داده های دو حالت و انجام متآنالیز هدف محقق است. مثال مورد کاربرد درباره متآنالیز مطالعات مختلف برای تعیین اثر توصیه های پزشکان برای ترک و توقف سیگار کشیدن در افراد است. در سال ۱۹۹۷ توسط سیلاجی (silagy) و همکاران مطالعه مرور نظام مندی با جمع آوری کارآزمایی های بالینی انجام شده درباره موضوع مورد اشاره انجام گردید. ۱۶ مطالعه در این خصوص جمع آوری گردید. جدول زیر این مطالعات را نشان می دهد. این مطالعات در فایل meta1 نیز پیوست می باشند. r1 و r2 هر کدام، جمع گروه های مداخله و شاهد مطالعه است. Coa نیز هر کدام وقوع یا عدم وقوع واقعه در گروه های مداخله و شاهد را نشان می دهند.

	name	year	a	r1	c	r2
1	Slama	1990	1	104	1	106
2	Porter	1972	5	101	4	90
3	Demers	1990	15	292	5	292
4	Stewart	1982	11	504	4	187
5	Page	1986	8	114	5	68
6	Slama	1995	42	2199	5	929
7	Haug	1994	20	154	7	109
8	Russell	1979	34	1031	8	1107
9	Wilson	1982	21	106	11	105
10	McDowell	1985	12	85	11	78
11	Janz	1987	28	144	12	106
12	Wilson	1990	43	577	17	532
13	Vetter	1990	34	237	20	234
14	Higashi	1995	53	468	35	489
15	Russell	1983	43	761	35	659
16	Jamrozik	1984	77	512	58	549



## روش های اپیدمیولوژی و پژوهش با Stata

با توجه به اینکه جدول بالا مربوط به مطالعات با پیامد دو حالته است، برای کار با این داده ها ابتدا باید ساختار آنها مطابق جدول الف تغییر داده شود.  $r1$  و  $r2$  جمع هر کدام از گروه های مداخله و کنترل است. لذا باید متغیرهایی ساخته شود و مانند فرمت جدول الف مقادیر  $b$  و  $d$  را نشان دهد. برای این کار باید دو متغیر  $d$  و  $b$  ساخته شوند. به این منظور دستورات زیر را باید بکار برد:

```
gen b=r1-a
gen d=r2-c
label var name "study" و label var year "year of publication"
```

	name	year	a	r1	c	r2	b	d
1	Slama	1990	1	104	1	106	103	105
2	Porter	1972	5	101	4	90	96	86
3	Demers	1990	15	292	5	292	277	287
4	Stewart	1982	11	504	4	187	493	183
5	Page	1986	8	114	5	68	106	63
6	Slama	1995	42	2199	5	929	2157	924
7	Haug	1994	20	154	7	109	134	102
8	Russell	1979	34	1031	8	1107	997	1099
9	Wilson	1982	21	106	11	105	85	94
10	McDowell	1985	12	85	11	78	73	67
11	Janz	1987	28	144	12	106	116	94
12	Wilson	1990	43	577	17	532	534	515
13	Vetter	1990	34	237	20	234	203	214
14	Higashi	1995	53	468	35	489	415	454
15	Russell	1983	43	761	35	659	718	624
16	Jamrozik	1984	77	512	58	549	435	491

با تعریف متغیرها و ساخت اطلاعات بالا، اکنون می توان متآنالیز را با دستور زیر انجام داد.

```
. metan a b c d , rr lcols(name year)xlabel(0.1, 0.2 , 0.5 , 1, 2
> , 5, 10)
```

Study	RR	[95% Conf. Interval]		% Weight
-----				
> -----				
Slama	1.019	0.065	16.081	0.40
Porter	1.114	0.309	4.021	1.71
Demers	3.000	1.105	8.147	2.02
Stewart	1.020	0.329	3.165	2.36
Page	0.954	0.325	2.800	2.53
Slama	3.549	1.409	8.941	2.84
Haug	2.022	0.886	4.615	3.32
Russell	4.563	2.122	9.811	3.12
Wilson	1.891	0.960	3.724	4.47
McDowell	1.001	0.469	2.137	4.64
Janz	1.718	0.917	3.219	5.59
Wilson	2.332	1.347	4.038	7.16
Vetter	1.678	0.996	2.829	8.14
Higashi	1.582	1.053	2.379	13.85
Russell	1.064	0.689	1.642	15.18
Jamrozik	1.424	1.035	1.958	22.65
-----				
> -----				
M-H pooled RR	1.676	1.440	1.951	100.00

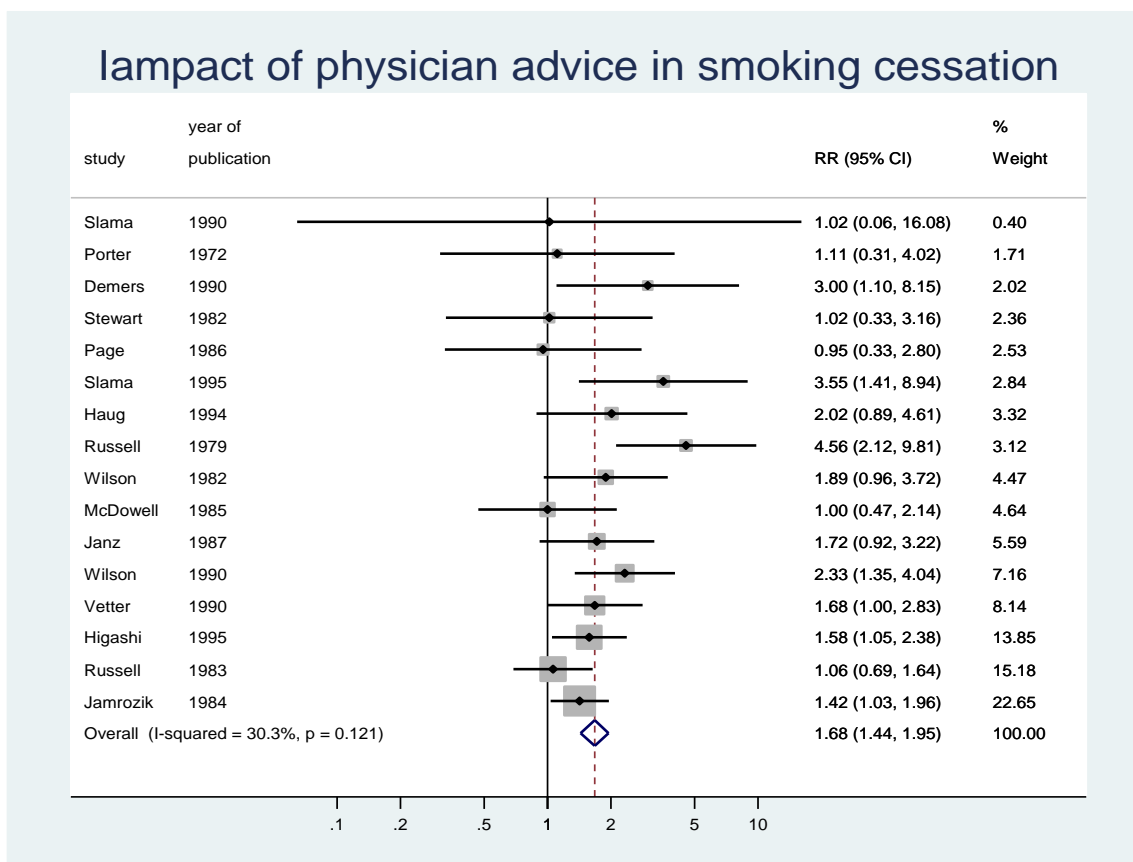
نسبت های خطر برای مطالعات بالا و نسبت خطر یک کاسه شده (pooled RR) با روش مانتل - هنزل و روش ثابت محاسبه و در قالب جدول بالا گزارش شده است. نتیجه نشان داده است که مداخله برای ترک سیگار موثر و معنی دار بوده است. هتروژینیتی مطالعات ارزیابی و به صورت زیر گزارش شده است. مقدار پی آن معنی دار نشده است. که این عدم معنی داری نشان می دهد که اختلاف بین مطالعات بوسیله تغییرات تصادفی قابل توجیه می باشد. به عبارت دیگر می توان از مدل ثابت استفاده نمود.

Heterogeneity chi-squared = 21.51 (d.f. = 15) p = 0.121

I-squared (variation in RR attributable to heterogeneity) = 30.3%

Test of RR=1 : z= 6.66 p = 0.000

شکل زیر نمودار انباشت برای مثال تاثیر توصیه های پزشکان برای ترک سیگار را نشان می دهد.



### مثال دوم:

در مثال دوم به متآنالیز انجام شده در سال ۱۹۹۵ توسط وینتراوب (weintraub) و همکاران در خصوص اثر آنتی هیستامین در سرماخوردگی و شدت عطسه و آبریزش بینی پرداخته می شود. در این مطالعه بعد از مصرف یک روزه آنتی هیستامین در گروه مداخله و مقایسه با گروه پلاسبو، میانگین استاندارد شده شدت عطسه و آبریزش بینی و انحراف معیار آن در هر دو گروه، بوسیله مقیاس سنجش کمی اندازه گیری شد. اطلاعات مطالعات بدست آمده مشتمل بر حجم نمونه، میانگین و انحراف معیار در هر یک از گروه های مداخله و کنترل در جدول زیر نشان داده شده است.

فایل meta2 این اطلاعات را نشان می دهد.

	id	n1	mean1	sd1	n2	mean2	sd2	n
1.	Study 1	11	.273	.786	16	-.188	.834	27
2.	Study 1	128	.932	.593	136	.81	.556	264
3.	Study 3	63	.73	.745	64	.578	.773	127
4.	Study 4	22	.35	1.139	22	.339	.744	44
5.	Study 5	16	.422	2.209	15	-.017	1.374	31
6.	Study 6	39	.256	1.666	41	.537	1.614	80
7.	Study 7	21	2.831	1.753	21	1.396	1.285	42
8.	Study 8	13	2.687	1.607	8	1.625	2.089	21
9.	Study 9	194	.49	.895	193	.264	.828	387

ساختار داده ها به صورت کمی پیوسته است. با دستور زیر می توان متاآنالیز این مطالعات را به صورت زیر انجام داد.

```
. metan n1 mean1 sd1 n2 mean2 sd2, xlabel (-1.5, -1 , -0.5, 0, 0.5 ,1, 1.5
> )title("effect of antihistamin on cold severity") xtitle("standardizatio
> n mean difference")textsize(125)
```

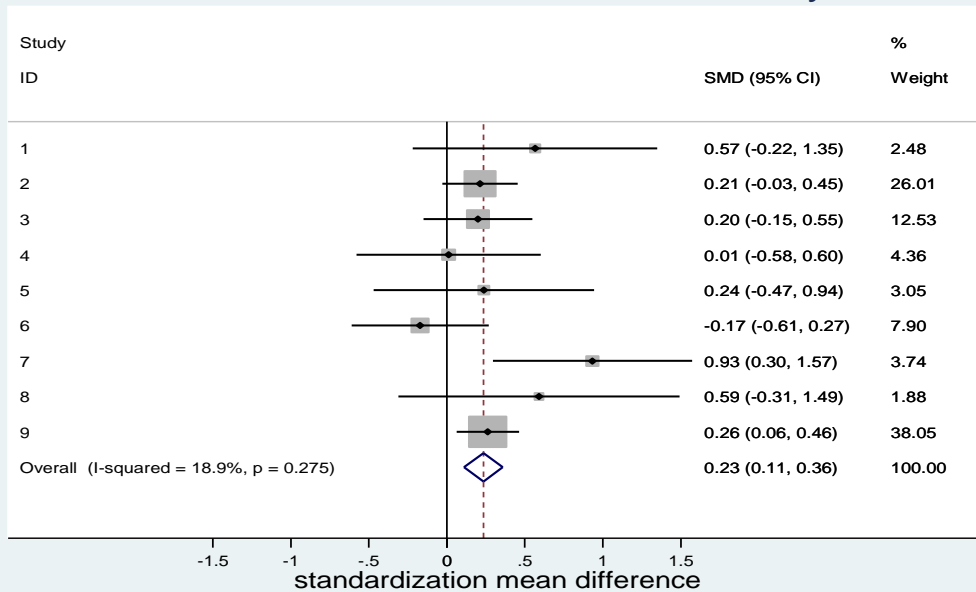
Study	SMD	[95% Conf. Interval]	% Weight
1	0.566	-0.218 1.349	2.48
2	0.212	-0.030 0.455	26.01
3	0.200	-0.149 0.549	12.53
4	0.011	-0.580 0.602	4.36
5	0.237	-0.470 0.944	3.05
6	-0.171	-0.611 0.268	7.90
7	0.934	0.295 1.572	3.74
8	0.590	-0.310 1.491	1.88
9	0.262	0.062 0.462	38.05
I-V pooled SMD	0.234	0.111 0.358	100.00

Heterogeneity chi-squared = 9.86 (d.f. = 8) p = 0.275  
I-squared (variation in SMD attributable to heterogeneity) = 18.9%

Test of SMD=0 : z= 3.72 p = 0.000

نتایج نشان می دهد مصرف آنتی هیستامین در ۲۴ ساعت اول درمان می تواند شدت علائم سرماخوردگی را به میزان قابل توجهی کاهش دهد. این کاهش از نظر آماری معنی دار بوده است. اختلاف بین مطالعات یا یکنواختی آنها، نیز آزمون گردیده و ناشی از تغییرات تصادفی و بکار بردن مدل ثابت است. نمودار انباشت این تحلیل در زیر آمده است.

effect of antihistamin on cold severity



مثال سوم:

در اینجا برای کار با نرم افزار از مطالعاتی که با هدف اثربخشی واکسن ب ت ژ برای بیماری سل در سال ۱۹۹۴ توسط colditz و همکاران انجام شده، استفاده می شود. این متآنالیز اطلاعات ۱۱ کارآزمایی انجام شده در این خصوص را نشان می دهند. فایل meta3 مربوط به این اطلاعات است.

	trial	startyr	tcases	tnonca~s	ccases	cnonca~s
1.	2	1933	6	300	29	274
2.	1	1935	4	119	11	128
3.	8	1941	17	1699	65	1600
4.	10	1947	5	2493	3	2338
5.	9	1949	186	50448	141	27197
6.	11	1950	27	16886	29	17825
7.	4	1950	33	5036	47	5761
8.	3	1950	62	13536	248	12619
9.	7	1965	29	7470	45	7232
10.	5	1965	8	2537	10	619
11.	6	1968	505	87886	499	87892

تعداد افراد بیمار و غیربیمار در هر یک از گروه های واکسن گرفته و واکسن نگرفته در جدول بالا آمده است. برای انجام متآنالیز با روش ثابت ( روش واریانس معکوس یا عکس واریانس به عنوان وزن مطالعات) به صورت زیر عمل می شود.

## روش های اپیدمیولوژی و پژوهش با Stata

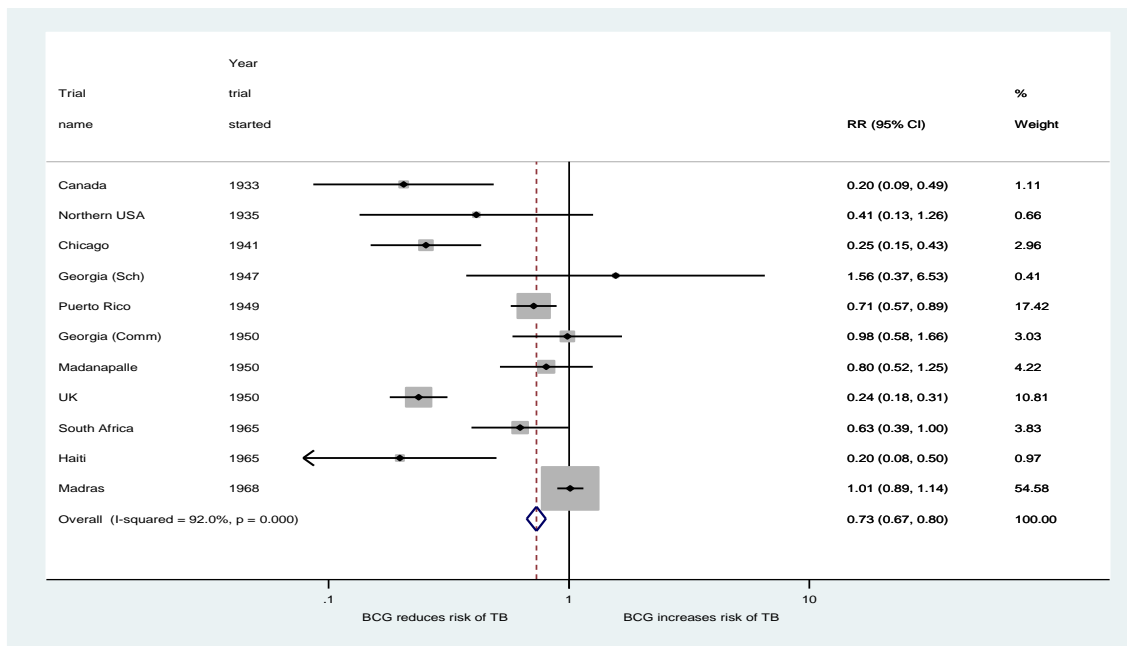
```
. metan tcases tnoncases ccases cnoncases , rr fixedi lcols (trialna
> m startyr) xlabel(0.1, 10) favours (BCG reduces risk of TB# BCG incr
> eases risk of TB)
```

Study	RR	[95% Conf. Interval]		% Weight
Canada	0.205	0.086	0.486	1.11
Northern USA	0.411	0.134	1.257	0.66
Chicago	0.254	0.149	0.431	2.96
Georgia (Sch)	1.562	0.374	6.528	0.41
Puerto Rico	0.712	0.573	0.886	17.42
Georgia (Comm)	0.983	0.582	1.659	3.03
Madanapalle	0.804	0.516	1.254	4.22
UK	0.237	0.179	0.312	10.81
South Africa	0.625	0.393	0.996	3.83
Haiti	0.198	0.078	0.499	0.97
Madras	1.012	0.895	1.145	54.58
<b>I-V pooled RR</b>	<b>0.730</b>	<b>0.667</b>	<b>0.800</b>	<b>100.00</b>

Heterogeneity chi-squared = 125.63 (d.f. = 10) p = 0.000  
 I-squared (variation in RR attributable to heterogeneity) = 92.0%

Test of RR=1 : z= 6.75 p = 0.000

در جدول بالا برای هر یک از مطالعات شاخص اثر و حدود اطمینان و وزن نشان داده شده است. شاخص اثر کلی نیز محاسبه و در انتهای جدول آمده است. آزمون یکنواخت بودن یا همگن بودن مطالعات نیز انجام گردید. ناهمگنی بین مطالعات نیز معنی دار شده است. و این ناهمگنی در مطالعات علاوه بر تغییرات تصادفی بین مطالعات، بیشتر به تغییرات اثر واقعی درمان نیز منتسب می باشد. به عبارت دیگر از مدل تصادفی می توان استفاده نمود. زیرا مطالعات هتروژن نیستند. شاخص کلی نیز برای اثر بخشی واکسن (فرض  $RR=1$ ) آزمون گردیده و نتیجه نشان می دهد که اثر بخشی واکسن معنا دار بوده است. نمودار انباشت (Forest Plot) متاآنالیز انجام شده برای اثر بخشی واکسن سل در زیر آمده است.



این نمودار با استفاده از روش عکس واریانس به عنوان وزن مطالعات در تحلیل، و استفاده از روش ثابت اثر واکسن ب ت ژ را در پیشگیری از بروز موارد بیماری سل نشان می دهد.

### مثال چهارم، کار با نرم افزار و معرفی دستورات مهم:

محققی قصد دارد تاثیر استفاده از ویتامین دی را در کیفیت زندگی بیماران دیابتی نوع ۲ پیشرفته را بمدت ۵ سال پیگیری و ارزیابی نماید. با جستجوی سیستماتیک در پایگاه های مهم از جمله ISI, Pubmed, Scopus, ScienceDirect, Magiran, Iranmedex, Google Scholar کلیه مطالعات کارآزمایی بالینی انجام شده در این خصوص و مستندات منتشر شده را در فاصله ۱۹۸۵ میلادی تا اول سپتامبر ۲۰۱۳ جمع آوری نموده است. در کل ۳۵ مقاله حاصل گردید. پس از ارزیابی و اعتبارسنجی کمی و کیفی مقالات، ۱۴ مطالعه انتخاب و اطلاعات مورد نظر بر اساس فرم جمع آوری اطلاعات استخراج گردید. اطلاعات مزبور و مطالعات پیدا شده در این زمینه در جدول زیر که به همین گونه در نرم افزار اطلاعات وارد می شوند، نمایش داده شده است. در صورت ناخوانا بودن جدول زیر، فایل meta4 را باز نمایید. پنجره ویراستار را باز کرده و اطلاعات را مشاهده نمایید. جدول زیر فرمت گزارش نتایج چهارده کارآزمایی های انجام شده برای بررسی تاثیر ویتامین دی در کیفیت زندگی بیماران دیابتی نوع دو که بعد از پی گیری طولانی مدت زنده یا فوت نموده اند را نشان می دهد.

id	trial	yr	a	b	c	d	ss	or	cillor	ciulor	logor	selogor
1	trial	1988	11	138	12	160	321	1.062802	.4556953	2.477814	.0609088	.4332884
2	Mat	1988	11	138	12	160	321	1.062802	.4556953	2.477814	.0609088	.4332884
3	vahid	1999	18	190	30	220	458	.6947368	.376443	1.285011	-.3642221	.3141616
4	sabor	1999	10	78	10	90	188	1.153846	.4574876	2.907617	.1431008	.4732141
5	ali	2000	35	350	29	300	714	1.034483	.6185421	1.729766	.0339015	.2631438
6	madi	2000	67	150	58	189	464	1.455517	.9644071	2.191691	.3753613	.2100582
7	Shaid	2004	9	57	15	57	138	.6	.2439481	1.480331	-.5108256	.4613735
8	fortr	2005	54	156	58	144	412	.8594164	.5576156	1.325788	-.1515017	.2216181
9	Piamw	2007	60	200	25	120	405	1.44	.8577399	2.412143	.3646431	.2645751
10	Sezklo	2009	12	145	10	138	305	1.142069	.4790756	2.720373	.1328415	.444383
11	bin	2010	80	380	140	1500	2100	2.255639	1.67362	3.025007	.8134333	.1514632
12	Kamal	2011	23	52	19	50	144	1.163968	.5669487	2.387457	.1518345	.3678867
13	Niz	2012	160	1500	200	1800	3660	.96	.7718843	1.194259	-.040822	.1116791
14	Habib	2013	89	800	184	800	1873	.4836957	.3698589	.6353796	-.7262994	.1384584

در جدول بالا ستون ها به صورت زیر تعریف شده اند.

id: Study ID  
 yr: Year of publication  
 a: Intervention group  
 (dead)

or: Odds Ratio (OR)  
 cillor: lower limit of 95% CI of OR  
 ciulor: Upper limit of 95% CI of OR  
 logor: Natural logarithm of OR

b: Intervention group  
(alive)  
c: Control group (dead)  
d: Control group (alive)  
ss: Sample size

اکنون که بر اساس جدول بالا، داده ها برای متاآنالیز آماده گردید، گام به گام متاآنالیز با دستورات زیر انجام می شود.

### اولین دستور: meta

دستور meta logor selogor اولین دستور برای انجام متاآنالیز و محاسبه نسبت شانس تجمعی با در نظر گرفتن متغیر selogor به عنوان وزن مطالعات برای متغیر logor بکار می رود. نتیجه این دستور برای مطالعه بالا در زیر آمده است. نسبت شانس برای ویتامین دی در دو مدل تصادفی و ثابت معنی دار نگردید.

#### Meta-analysis

Method	Pooled Est	95% CI Lower	95% CI Upper	Asymptotic z_value	p_value	No. of studies
Fixed	0.008	-0.107	0.123	0.142	0.887	14
Random	0.032	-0.257	0.320	0.215	0.830	

Test for heterogeneity: Q= 65.009 on 13 degrees of freedom (p= 0.000)

Moment-based estimate of between studies variance = 0.211

با ازمون انجام شده مشخص است که فرض همگون بودن مطالعات برقرار نیست و لذا مدل تصادفی انتخاب می شود. البته با این فرض که سایر حربه ها برای بررسی هتروزینیتی مطالعات کارساز نباشند. مثلا تبدیل متغیر به مقیاس لگاریتمی.

چنانچه محقق بخواهد تحلیل را فقط بر مطالعاتی که از سال ۲۰۰۰ منتشر شده انجام بدهد باید در دستور مذکور از شرط استفاده نماید. دستور و خروجی در زیر نمایش داده می شود.

```
. meta logor selogor if yr>=2000
```

#### Meta-analysis

Method	Pooled Est	95% CI Lower	95% CI Upper	Asymptotic z_value	p_value	No. of studies
Fixed	0.018	-0.102	0.139	0.296	0.767	10
Random	0.055	-0.296	0.407	0.310	0.757	

Test for heterogeneity: Q= 63.467 on 9 degrees of freedom (p= 0.000)

Moment-based estimate of between studies variance = 0.250

چنانچه بخواهید علاوه بر اجرای دستور متاآنالیز، نمودار انباشت نیز با مدل تصادفی (random) نمایش داده شود به طریق زیر عمل می شود.

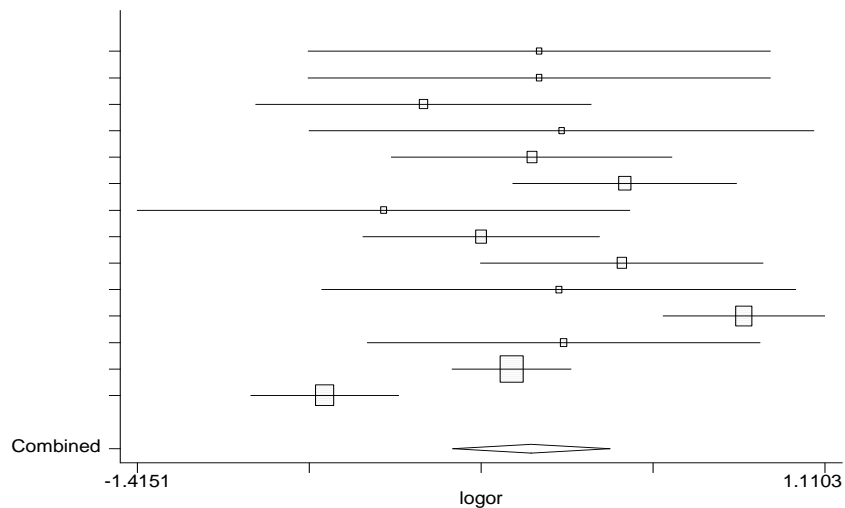
. meta logor selogor, graph(r)

Meta-analysis

Method	Pooled Est	95% CI		Asymptotic		No. of studies
		Lower	Upper	z_value	p_value	
Fixed	0.008	-0.107	0.123	0.142	0.887	14
Random	0.032	-0.257	0.320	0.215	0.830	

Test for heterogeneity: Q= 65.009 on 13 degrees of freedom (p= 0.000)  
 Moment-based estimate of between studies variance = 0.211

نمودار انباشت دستور بکار رفته در زیر آمده است.



در معرفی دستورات به نرم افزار، شاخص مورد مطالعه با علامت  $\theta$  و خطای استاندارد آن با علامت  $se\theta$  مشخص شده است که باید هنگام اجرای دستورات به جای این علامت، اسم یکی از شاخص های مورد مطالعه نظیر خطر نسبی (RR)، نسبت شانس (OR)، تفاضل خطر (RD) و یا شیوع و بروز جایگزین شوند. در برخی از دستورات علامت «|» آورده شده است؛ مفهوم این علامت آن است که فقط یکی از گزینه های داخل پرانتز را که با این علامت جدا شده اند می توانید انتخاب کنید و استفاده همزمان آنها مجاز نیست. دستور meta در زیر به صورت کامل آمده است. با انتخاب سایر گزینه نمودار دلخواه و خروجی مورد نظر حاصل می شود.

meta  $\theta$   $se\theta$ , eform graph(f|r) cline xline(#) id(varname) print level(#) stlevel(#)  
 ltrunc(#) rtrunc(#) xlab(##,##) b2title(name) fmult(#) ebay



در دستور بالا گزینه های زیر قابل تعریف است.

option	عملکرد
eform	مقیاس شاخص مورد مطالعه را از لگاریتمی به عددی تبدیل می کند.
graph(f r)	نمایش نمودار انباشت برای مدل ثابت (f) یا مدل تصادفی (r)
cline	رسم خط عمود منقطع برای نشان دادن برآیند شاخص مورد مطالعه
xline(#)	رسم خط عمود ممتد برای نشان دادن فرضیه صفر
id(varname)	نمایش اطلاعات متغیر (varname) در محور افقی Y نمودار انباشت
print	نمایش اطلاعات تک تک مطالعات در خروجی نرم افزار
level(#)	تعیین سطح اطمینان برای برآیند کلی شاخص مورد مطالعه
stlevel(#)	تعیین سطح اطمینان شاخص برای تک تک مطالعات
ltrunc(#)	تعیین حد پایین محور افقی X
rtrunc(#)	تعیین حد بالای محور افقی X
xlab(##,##,##)	تعیین مقیاس دلخواه برای محور X
b2title(name)	افزودن عنوان دلخواه به محور X
fmult(#)	بزرگ نمودن فونت در نمودار انباشت به میزان دلخواه
ebay	برآورد شاخص و حدود اطمینان مربوطه برای تک تک مطالعات به روش بیزین

در مثال مذکور با انجام این دستور نتیجه زیر بدست می آید.

```
. meta logor selogor, eform graph(r) cline xline(1) id(trial) b2title(0d
> ds Ratio)
```

Meta-analysis (exponential form)

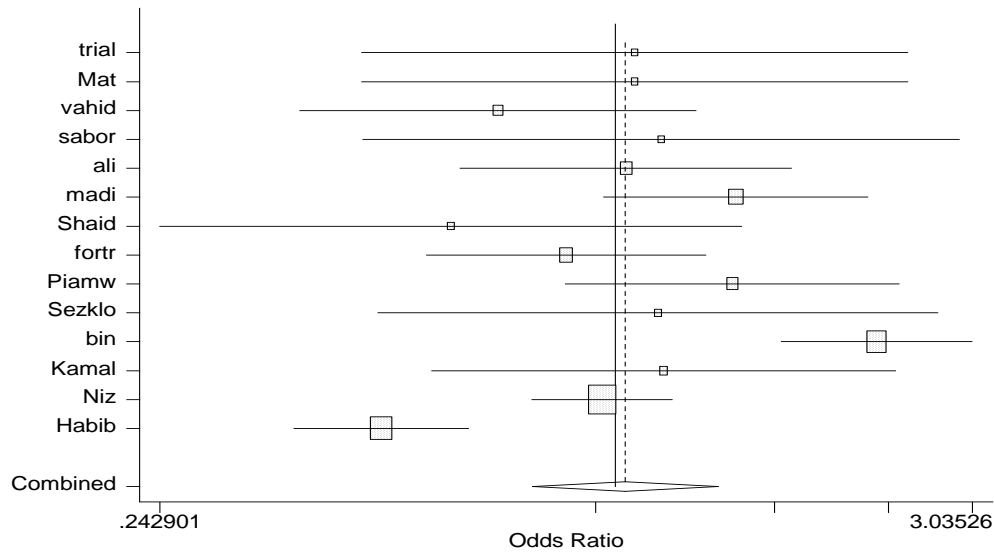
Method	Pooled Est	95% CI		Asymptotic		No. of studies
		Lower	Upper	z_value	p_value	
Fixed	1.008	0.899	1.131	0.142	0.887	14
Random	1.032	0.774	1.377	0.215	0.830	

Test for heterogeneity: Q= 65.009 on 13 degrees of freedom (p= 0.000)

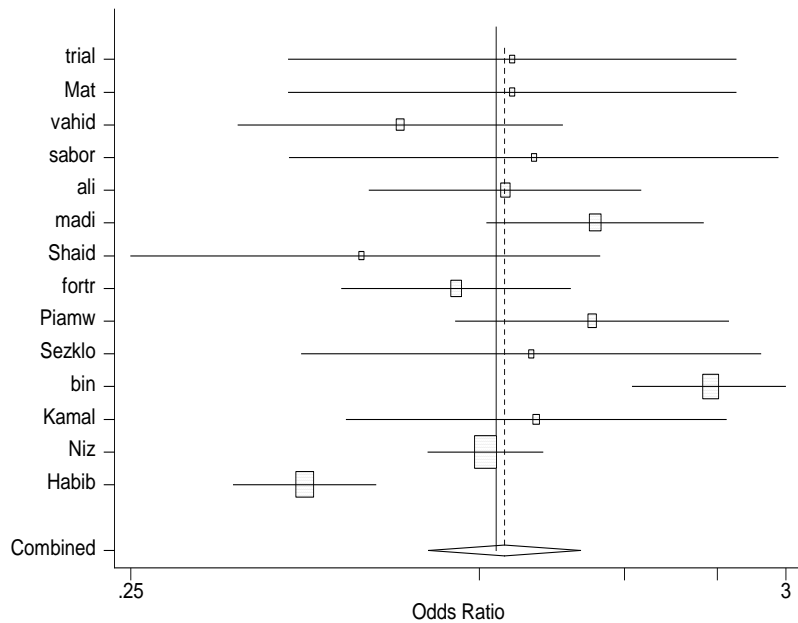
Moment-based estimate of between studies variance = 0.211

نتیجه ارزیابی همگونی مطالعات نشان داده است که مطالعات همگن و یکنواخت نیستند. P=0/000  
 مقادیر نسبت شانس کلی در هر یک از مدل های ثابت و تصادفی در جدول آمده است.  
 این مقادیر معنی دار نیستند. نتایج نشان می دهد که تجویز ویتامین دی بر کیفیت زندگی بیماران دیابتی  
 نوع ۲ تاثیری معنا دار آماری نداشته است.

نمودار انباشت دستور بالا در زیر آمده است.



برای زیباتر شدن نمودار بالا و نمایش به صورت زیر می توان مقادیر را گرد نمود به نحوی که در حد پایین محور افقی به عدد 0.25 و حد بالای آن به عدد 3 گرد شوند. دستور زیر را اجرا نمایید.  
`meta logor selogor, eform graph(r) cline xline(1) id(trial) b2title(Odds Ratio) ltr(0.25) rtr(3)`



در اجرای دستور `meta logor` تنها برای شاخص‌هایی که به صورت نسبت بیان می‌شوند نظیر: Odds Ratio, Risk Ratio, Rate Ratio و Hazard Ratio باید لگاریتم گرفته شود. برای بررسی تفاضل

میانگین‌ها نیازی به چنین کاری نیست و به راحتی می‌توان تفاضل میانگین‌ها و خطای معیار آنها را در دستور وارد نمود.

### دومین دستور: metan

دستوری که هم برای داده‌های گسسته (Binary) و هم برای داده‌های پیوسته (Continuous) قابل استفاده می‌باشد دستور metan است. شاخص مورد مطالعه در داده‌های گسسته ممکن است نسبت خطر (Risk Ratio) یا نسبت شانس (Odds Ratio) و یا تفاضل خطر (Risk Difference) باشد. در داده‌های پیوسته شاخص مورد مطالعه ممکن است تفاضل میانگین‌ها (Difference in Mean) و یا تفاضل استاندارد شده میانگین‌ها (Standardized Difference in Mean) باشد. شکل کامل این دستور به صورت زیر است.

```
metan a b c d, rr|or|rd fixed|random label(namevar=namevariable,
yearvar=yearvariable) xlabel(##,##) ilevel(#) olevel(#) texts(#) nooverall notable
nograph nowt nostats
```

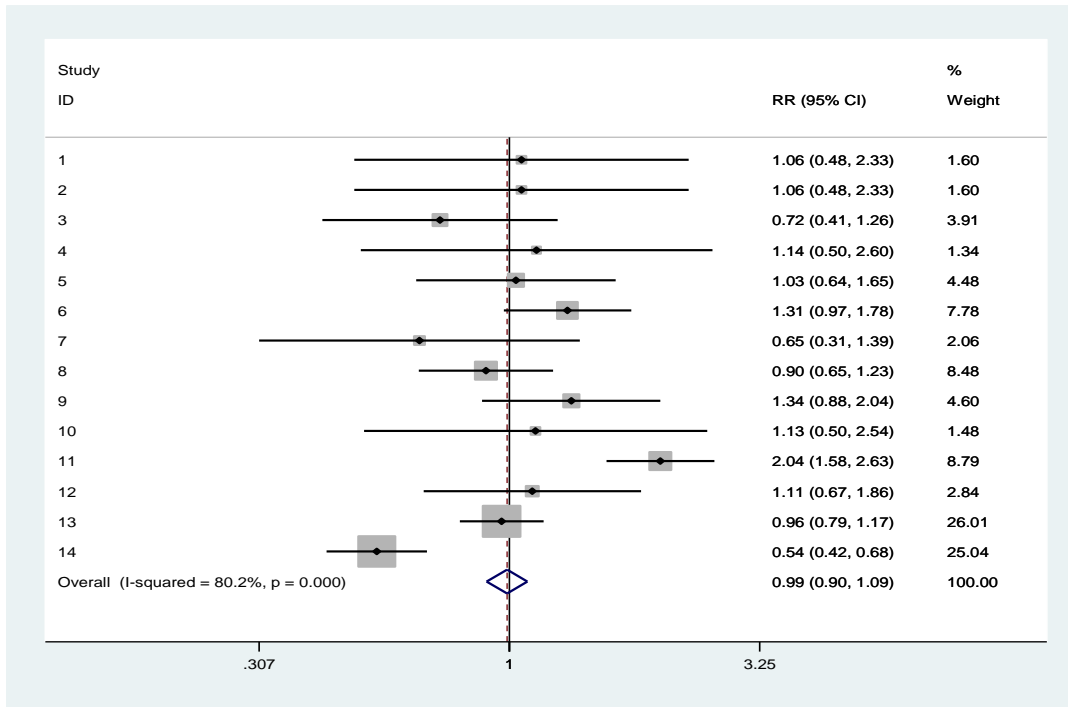
گزینه‌های دستور metan همانند دستور meta است. گزینه‌های جدید در این دستور عبارتند از:

option	عملکرد
rr or rd	محاسبه شاخص مورد مطالعه بر حسب RR, OR, و یا RD
fixed random	محاسبه شاخص مورد مطالعه با مدل ثابت یا تصادفی
label	نمایش اسمی اول نویسندگان مقالات و سال انتشار آن در نمودار انباشت
ilevel(#)	تعیین سطح اطمینان شاخص مورد مطالعه برای تک تک مطالعات
olevel(#)	تعیین سطح اطمینان برآیند کلی شاخص مورد مطالعه
nooverall	عدم نمایش برآیند کلی در نمودار انباشت
notable	عدم نمایش جدول نتایج تک تک مطالعات
nograph	عدم نمایش نمودار انباشت
nowt	عدم نمایش وزن اختصاص داده شده به هر یک از مطالعات در نمودار انباشت
nostats	عدم نمایش نتایج تک تک مطالعات در نمودار انباشت

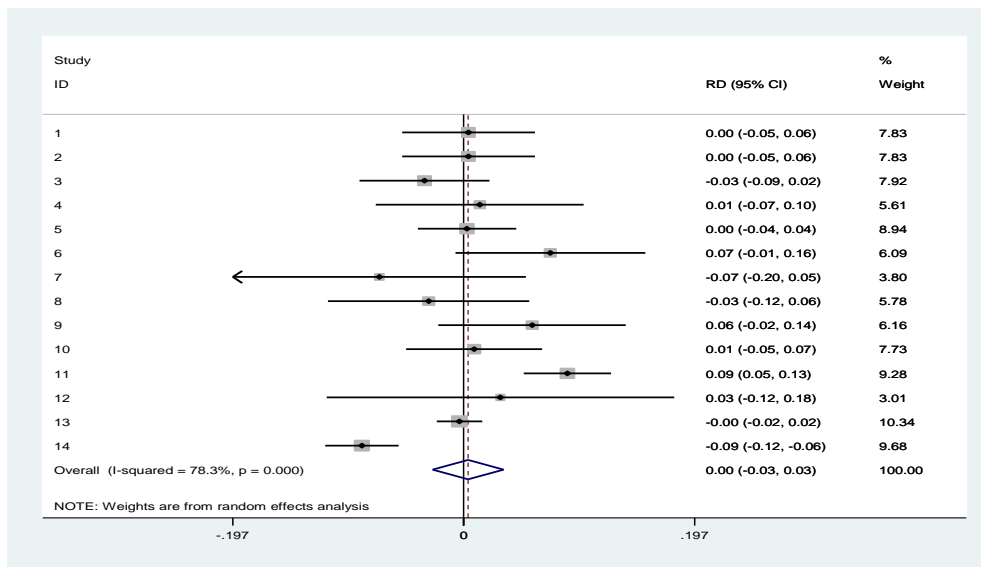
وقتی دستور metan اجرا شود برآورد خطر نسبی در مدل ثابت محاسبه می‌شود. `metan a b c d`

Heterogeneity chi-squared = 65.69 (d.f. = 13) p = 0.000  
I-squared (variation in RR attributable to heterogeneity) = 80.2%

Test of RR=1 : z= 0.23 p = 0.820



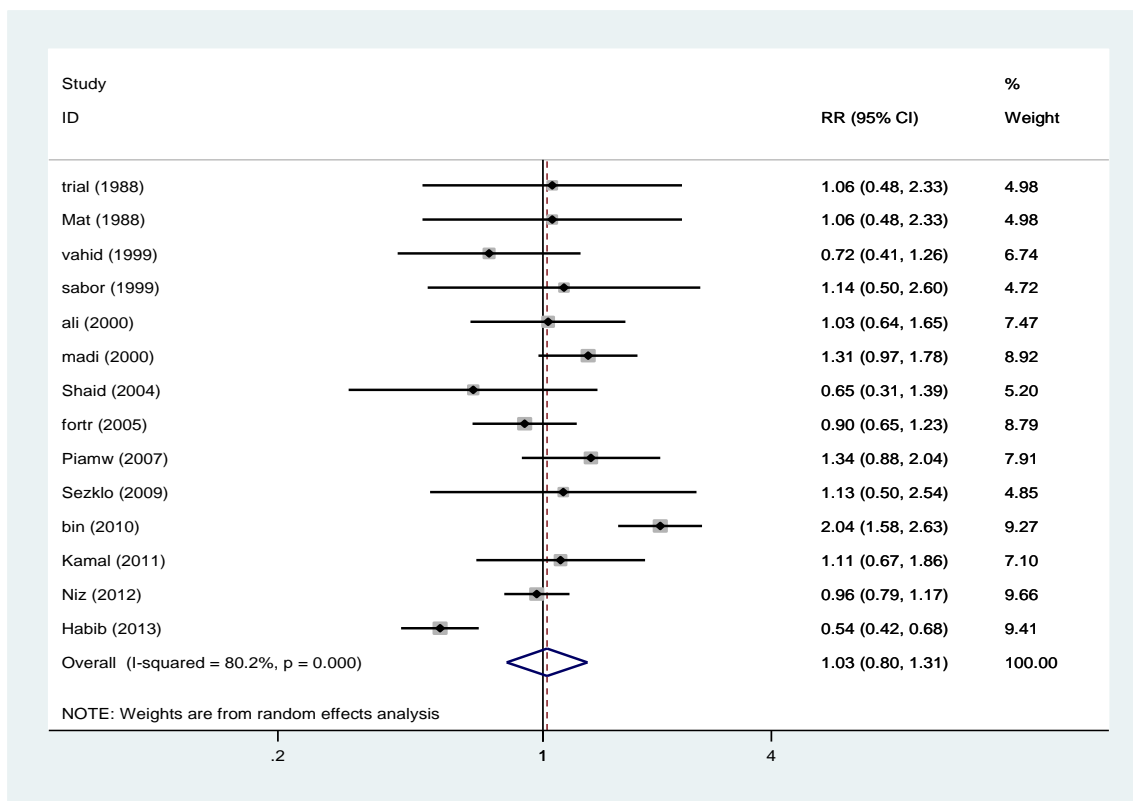
یا می توان با دستور `metan a b c d, rr fixed` نسبت محاسبه نمود. بجای خطر نسبی می توان نسبت شانس را نیز در مدل تصادفی به روش «Mantel and Haenszel» با دستور `metan a b c d, or random` محاسبه نمود. یا به روش عکس واریانس (inverse variance) نسبت شانس را با دستور `metan a b c d, or randomi` محاسبه نمود. اختلاف خطر نیز در مدل تصادفی با دستور `metan a b c d, rd random` محاسبه می شود. نتایج نشان می دهد که خطر نسبی معنی دار نیست. نمودار انباشت اختلاف خطر با دستور بالا در زیر آمده است.



برای نشان دادن اسامی نویسندگان اول مقالات و سال انتشار آنها در نمودار انباشت و تغییر ارقام محور X از دستور `metan a b c d, rr random label(namevar=trial, yearvar=yr) xlabel(0.2, 1, 4)` استفاده نمایید.

Heterogeneity chi-squared = 65.69 (d.f. = 13) p = 0.000  
 I-squared (variation in RR attributable to heterogeneity) = 80.2%  
 Estimate of between-study variance Tau-squared = 0.1511

Test of RR=1 : z= 0.20 p = 0.841



با افزودن گزینه `counts` به دستور بالا می توان از نرم افزار خواست تا مقادیر گروه های مداخله و کنترل را به تفکیک وقوع یا عدم وقوع پیامد گزارش نماید.

نحوه نگارش دستور `metan` برای متغیرهای عددی (مثلاً تفاضل میانگین ها) نیز `metan n1 m1 sd1 n2 m2 sd2` می باشد. در این دستور `n`، `m` و `sd` به ترتیب حجم نمونه، میانگین و انحراف معیار در گروه های مداخله و کنترل می باشد.

تمام گزینه هایی که برای دستور `metan` در بالا گفته شد برای این دستور هم قابل اجرا می باشند.

سومین دستور: metacum

این دستور نیز گام به گام اطلاعات مطالعات جدیدتر را با اطلاعات مطالعات قبلی ترکیب می نماید و نتیجه را بیان می کند. در واقع متآنالیز تجمعی را انجام می دهد. به همین دلیل با اضافه شدن مطالعات جدید و افزایش حجم نمونه اولاً به تدریج دامنه اطمینان برآورد پارامتر مورد مطالعه باریکتر می شود و همچنین پایایی آن نیز بالا می رود.

به عبارت دیگر برآیند کلی شاخص از نتایج مطالعات جدیدتر کمتر متأثر می شود. برای انجام این تحلیل باید ابتدا با استفاده از دستور sort ترتیب مطالعات را بر اساس فاکتور مورد نظر که در اکثر اوقات سال انتشار است مرتب نمود.

sort yr

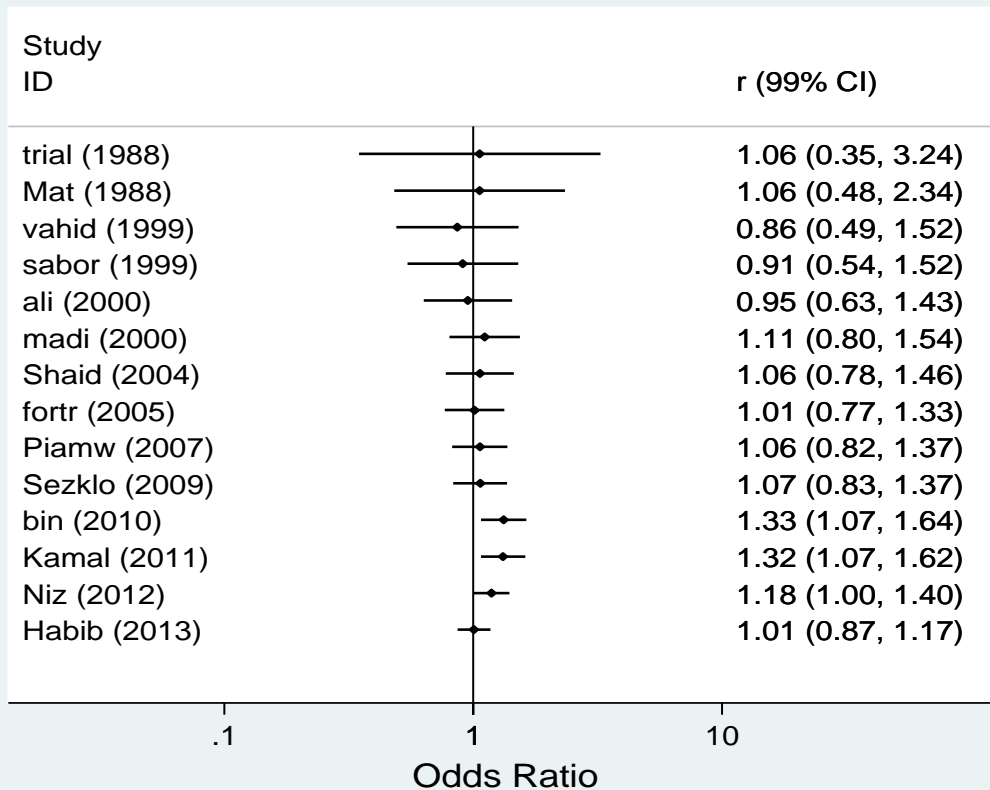
سپس دستور metacum را اجرا کرد.

شکل کلی این دستور به صورت زیر است:

metacum  $\theta$  se $\theta$ , effect(f|r) eform graph label(namevar=trial, yearvar=yr) ilevel(#) b2title(Odds Ratio) xlab(##,##)

اکثر گزینه های این دستور در دستورات قبلی معرفی گردید. تنها گزینه ناآشنا و جدید عبارت effect است که به نرم افزار می گوید که برآورد شاخص مورد مطالعه را با مدل ثابت (f) یا تصادفی (r) گزارش نماید. گزینه ilevel هم سطح اطمینان را برای برآورد هر یک از مطالعه تعیین می کند.

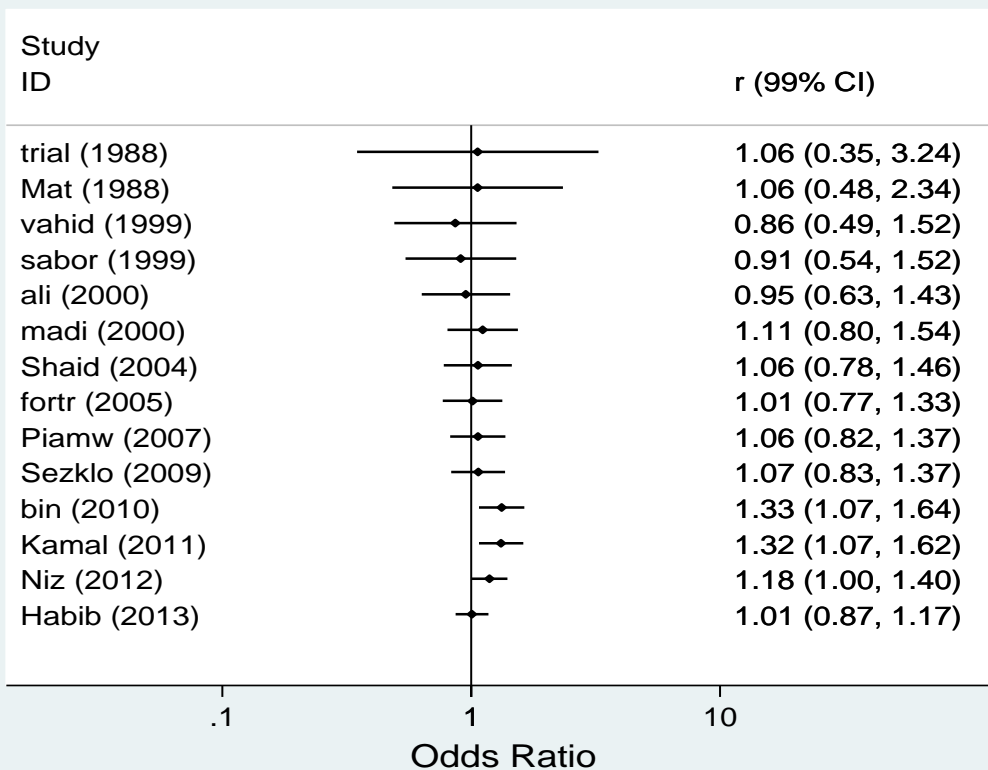
**چنانچه دستور بالا برای فایل داده ها اجرا شود نتایج زیر بدست می آید.**



با دستور زیر می توان اطلاعات وارد شده به مطالعه را به صورت تک تک بر حسب مطالعه دید.

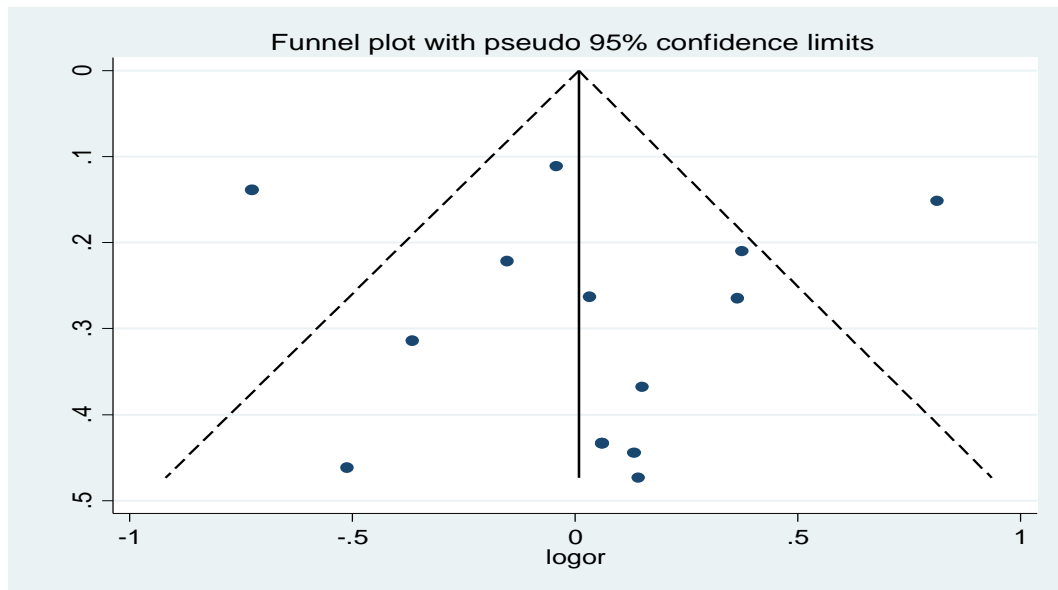
```
. metacum logor selogor, effect(r) eform graph label(namevar=trial, year
> var=yr) ilevel(99) b2title(Odds Ratio) xlab(.1,1,10)
```

Study	ES	[99% Conf. Interval]	
trial (1988)	1.063	0.348	3.245
Mat (1988)	1.063	0.483	2.340
vahid (1999)	0.864	0.491	1.520
sabor (1999)	0.909	0.545	1.518
ali (2000)	0.953	0.633	1.434
madi (2000)	1.112	0.802	1.540
Shaid (2004)	1.065	0.777	1.458
fortr (2005)	1.013	0.769	1.334
Piamw (2007)	1.064	0.824	1.374
Sezklo (2009)	1.068	0.832	1.370
bin (2010)	1.326	1.075	1.636
Kamal (2011)	1.318	1.074	1.618
Niz (2012)	1.185	1.002	1.400
Habib (2013)	1.008	0.867	1.173

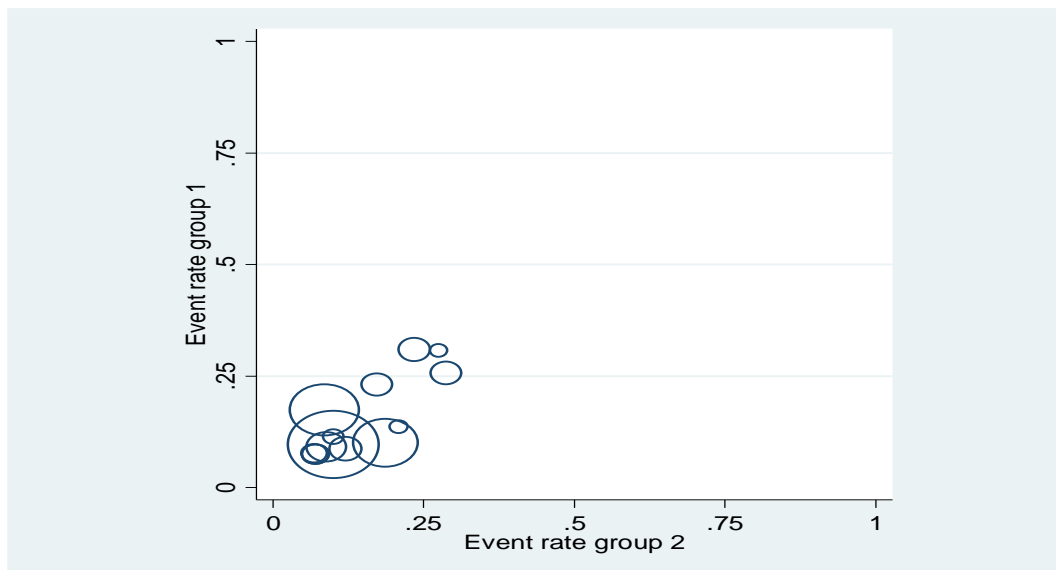


با دستورات meta و metan آزمون های آماری  $\chi^2$  و  $\tau^2$  که جهت بررسی ناهمگونی استفاده می شوند گزارش می شوند و نیازی به اجرای دستور دیگری نیست. اما برای بررسی شاخص  $I^2$  باید دستور heterogi (#Q #df, level(#)) را نوشت.

دستور funnel  $\theta$  se $\theta$ , sample overall(#) xlabel(##,##) نمودار کیفی را رسم می کند. نمودار کیفی به صورت دقیق تر و کامل تر با استفاده از دستور metafunnel رسم می شود. دستور metafunnel logor در متاآنالیز برای بررسی سوگیری در انتشار بیشترین استفاده را دارد. نمودار کیفی مطالعات مذکور در زیر آمده است.



برای رسم نمودار L'Abbe نیز دستور labbe a b c d که برای بررسی تورش نشر بکار می رود. به صورت زیر ترسیم می شود.





## روش های اپیدمیولوژی و پژوهش با Stata

دستور `metabias θ seθ, graph(begglegger) gweight` برای بررسی احتمال وجود سوگیری در انتشار نتایج بکار می رود و علاوه بر رسم نمودارهای مربوطه، نتایج آزمون های آماری Egger و Begg را هم به طور همزمان گزارش می کند.

```
. metabias a b c d, or egger graph xlabel(trial)
```

Note: data input format `tcases tnoncases ccases cnoncases` assumed.  
 Note: Peters or Harbord tests generally recommended for binary data

Egger's test for small-study effects:  
 Regress standard normal deviate of intervention  
 effect estimate against its standard error

```
Number of studies = 14                                Root MSE =
> 2.322
```

Std_Eff	Coef.	Std. Err.	t	P> t	[95% Conf. Int	erval]
slope	-.0536113	.2982862	-0.18	0.860	-.7035211	.5
bias	.3169343	1.357783	0.23	0.819	-2.64142	3.

```
Test of H0: no small-study effects                    P = 0.819
```

.

**مثال پنجم:**

در این مثال به متآنالیز کارآزمایی های بالینی در ارتباط با اثر درمانی استرپتوکیناز در درمان سکنه قلبی است که در سال ۱۹۵۹ گزارش گردید پرداخته می شود.  
 ۲۱ کارآزمایی بالینی در فاصله ۱۹۵۹ تا سال ۱۹۶۸ انجام گردید. دو کارآزمایی چندمرکزی با بیش از ۱۷۰۰۰ بیمار و ۱۸۰۰ مورد مرگ نیز در این فاصله انجام گردید.  
 نتایج نشان داد که درمان شانس مرگ را کاهش می دهد. اطلاعات این مطالعات در قالب فایل meta5 در جدول زیر نشان داده شده است.

	trialnam	year	pop1	deaths1	pop0	deaths0
1.	Fletcher	1959	12	1	11	4
2.	Dewar	1963	21	4	21	7
3.	1st European	1969	83	20	84	15
4.	Heikinheimo	1971	219	22	207	17
5.	Italian	1971	164	19	157	18
6.	2nd European	1971	373	69	357	94
7.	2nd Frankfurt	1973	102	13	104	29
8.	1st Australian	1973	264	26	253	32
9.	NHLBI SMIT	1974	53	7	54	3
10.	Valere	1975	49	11	42	9
11.	Frank	1975	55	6	53	6
12.	UK Collab	1976	302	48	293	52
13.	Klein	1976	14	4	9	1
14.	Austrian	1977	352	37	376	65
15.	Lasierra	1977	13	1	11	3
16.	N German	1977	249	63	234	51
17.	Witchitz	1977	32	5	26	5
18.	2nd Australian	1977	112	25	118	31
19.	3rd European	1977	156	25	159	50
20.	ISAM	1986	859	54	882	63
21.	GISSI-1	1986	5860	628	5852	758
22.	ISIS-2	1988	8592	791	8595	1029

قبل از انجام متآنالیز ، برای هر مطالعه log odds ratio و همچنین واریانس باید محاسبه گردد. این متغیرها با دستورات زیر ساخته می شوند.

```
gen logor=log( ( deaths1/( pop1- deaths1))/ (( deaths0/( pop0- deaths0))))
gen selogor= sqrt(1/ deaths1+1/( pop1- deaths1)+1/ deaths0+1/( pop0- deaths0))
```

نتایج تحلیل متانالیز به شکل زیر آمده است.

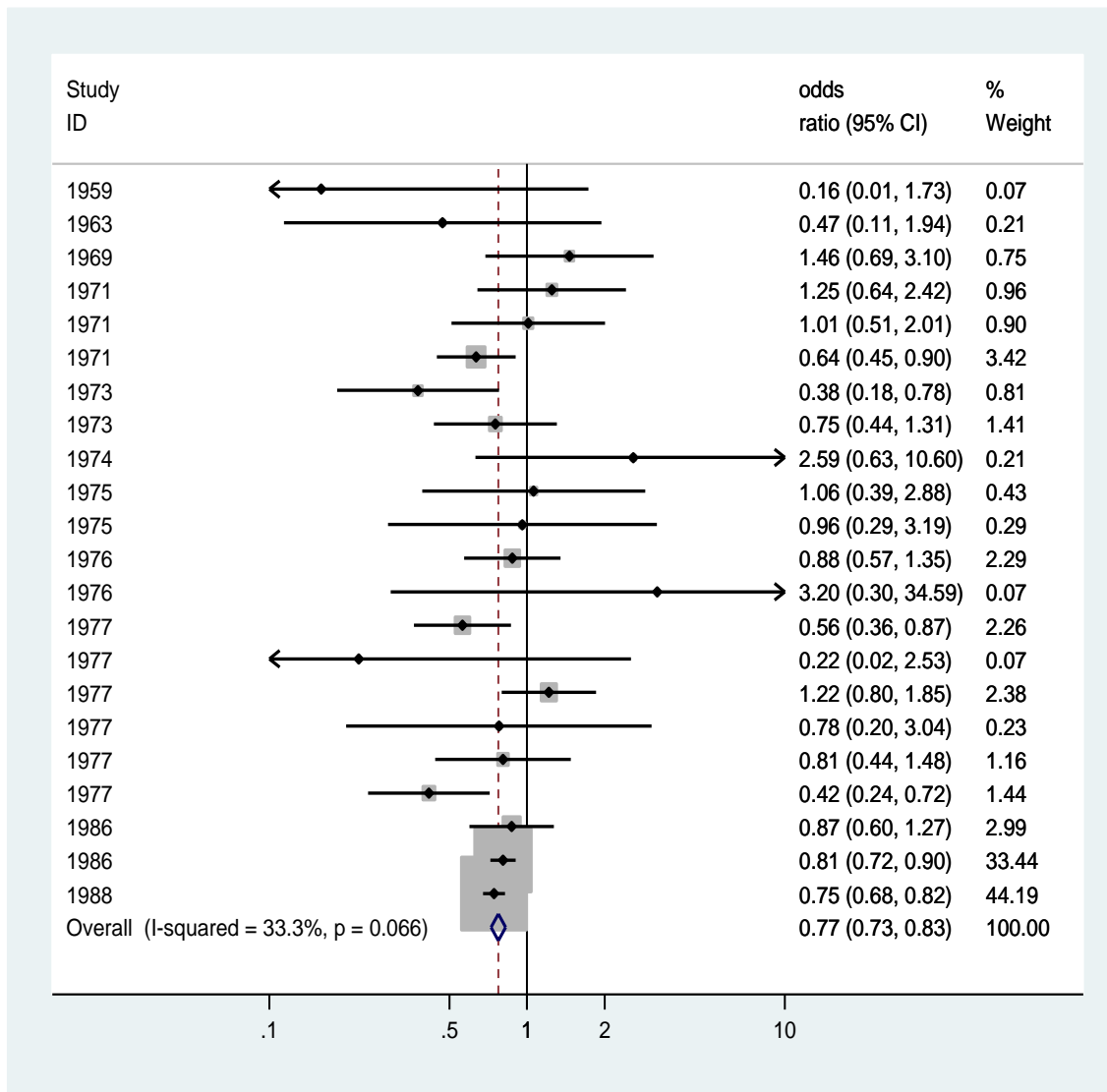
```
. metan logor selogor, eform fixed label (namevar=trialnam,yearvar=y
> ear) xlabel(0.1 , 0.5, 1, 2 , 10)force effect("odds ratio")
```

Study	ES	[95% Conf. Interval]	% Weight
1959	0.159	0.015 1.732	0.07
1963	0.471	0.114 1.942	0.21
1969	1.460	0.689 3.096	0.75
1971	1.248	0.643 2.423	0.96
1971	1.012	0.510 2.008	0.90
1971	0.635	0.447 0.903	3.42
1973	0.378	0.183 0.778	0.81
1973	0.754	0.436 1.306	1.41
1974	2.587	0.632 10.596	0.21
1975	1.061	0.392 2.876	0.43
1975	0.959	0.289 3.185	0.29
1976	0.876	0.570 1.346	2.29
1976	3.200	0.296 34.588	0.07
1977	0.562	0.365 0.867	2.26
1977	0.222	0.019 2.533	0.07
1977	1.215	0.797 1.853	2.38
1977	0.778	0.199 3.044	0.23
1977	0.806	0.440 1.477	1.16
1977	0.416	0.242 0.716	1.44
1986	0.872	0.599 1.270	2.99
1986	0.807	0.721 0.903	33.44
1988	0.746	0.676 0.822	44.19
I-V pooled ES	0.774	0.725 0.826	100.00

```
Heterogeneity chi-squared = 31.50 (d.f. = 21) p = 0.066
I-squared (variation in ES attributable to heterogeneity) = 33.3%

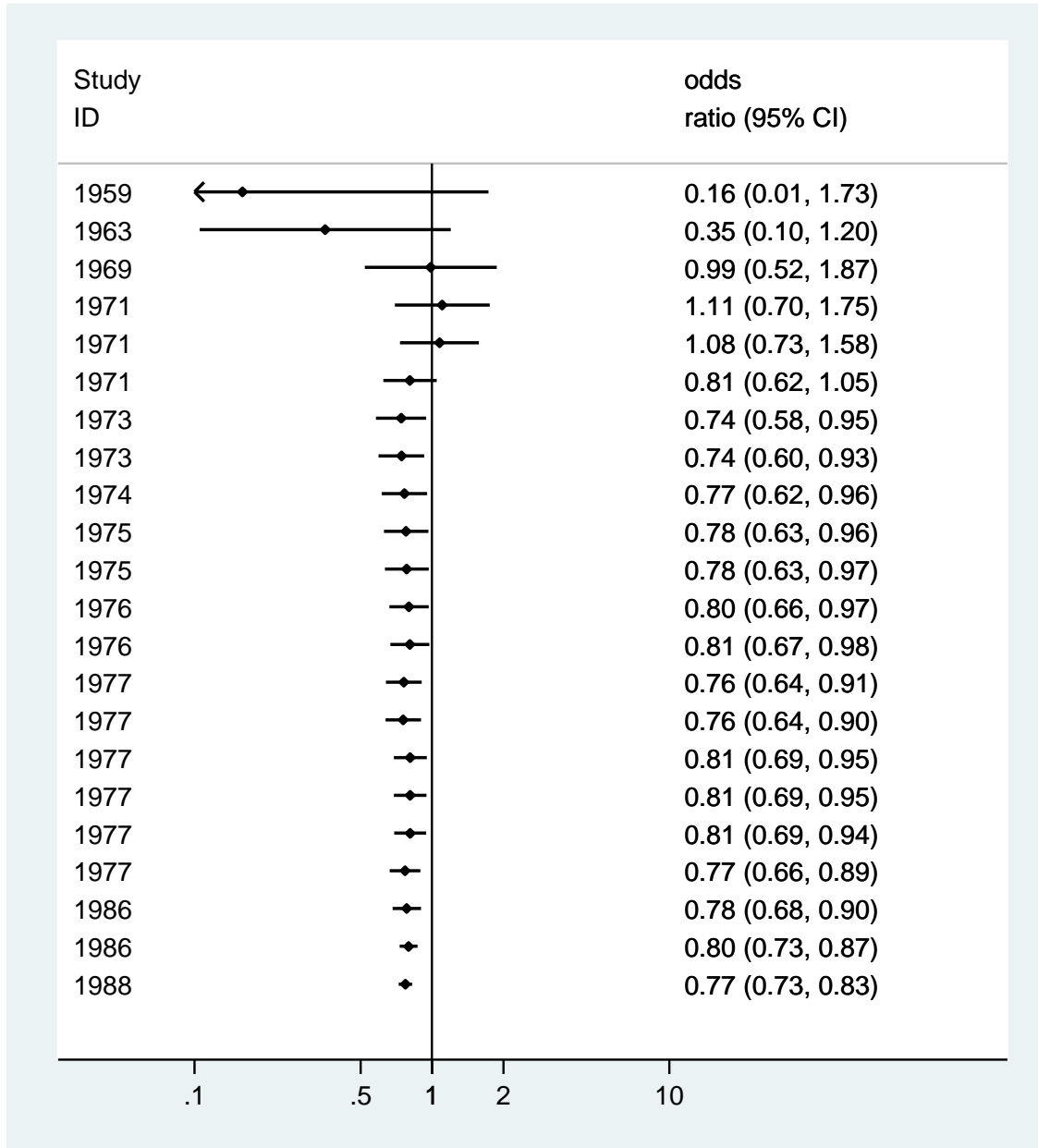
Test of ES=1 : z= 7.71 p = 0.000
```

نتایج مذکور نشان می دهد که مطالعات همخوان یا یکنواخت هستند. و ناهمگنی فقط ناشی از تغییرات تصادفی در بین مطالعات مختلف است. آزمون برای ارزیابی معنی داری شاخص اثر انجام گردید و نتیجه نشان داد که مقدار اثر معنی دار بوده است. ( $p=0.000$ ) یا به عبارتی اثر درمانی استرپتوکیناز بر پیشگیری از بروز مرگ در بیماران سکته قلبی موثر بوده است.



نمودار بالا نشان می دهد که مطالعات مربوط به سال ۱۹۸۶ و ۱۹۸۸ وزن بیشتری نسبت به سایر مطالعات داشته اند. متاآنالیز تجمعی برای اطلاعات مذکور انجام و از این دید نیز تحلیل را می توان دید.

نتیجه متاآنالیز تجمعی در زیر آمده است. از اواخر سال ۱۹۷۷ به روشنی نتایج نشان می دهند که استرپتوکیناز در پیشگیری از مرگ بیماران سکته قلبی موثر بود.



برآورد نقطه ای نسبت شانس برای سال ۱۹۷۷ در نمودار بالا برابر با 0.771 می باشد. بعد از این نتیجه، کارآزمایی های بزرگ در سال ۱۹۸۸ انجام و نسبت شانس برابر با 0.774 بدست آمد.

متاآنالیز P- Values :

فیشر از سال ۱۹۳۲ برای ترکیب p-value ها پیشنهاداتی را برای انجام متاآنالیز ارائه نمود. روشهای ناپارامتریک برای این ترکیب ها وجود دارد. ادینگتون (Edgington) نیز روشی را در سال ۱۹۷۲ برای ترکیب مطالعات کوچک و با حجم نمونه کمتر پیشنهاد نمود که شبیه روش فیشر است. و روش سومی نیز بر مبنای منحنی نرمال و متوسط مقادیر پی وجود دارد. این روش برای تعداد مطالعات زیاد پیشنهاد شده است. برای انجام این نوع متاآنالیز از اطلاعات ۷ مطالعه کارآزمایی بالینی کنترل شده با گروه پلاسبو استفاده می شود. این مطالعات برای بررسی اثر آسپیرین در پیشگیری از مرگ بعد از سکته قلبی انجام گردیدند. فایل اطلاعات مذکور به نام meta6 در پیوست آمده است. این اطلاعات در جدول زیر آمده است. ابتدا تابع normprob برای p value یکطرفه ساخته می شود. این تابع با دستور gen pvar=normprob(- logrr/ logse) ساخته می شود.

. list studyid logrr logse pvar, noobs

studyid	logrr	logse	pvar
MCR-1	0.3289	0.1972	.0476728
CDP	0.3853	0.2029	.0287845
MRC-2	0.2192	0.1432	.0629185
GASP	0.2229	0.2545	.1905599
PARIS	0.2261	0.1876	.1140584
AMIS	-0.1249	0.0981	.8985248
ISIS-2	0.1112	0.0388	.0020786

. metap pvar

Meta-analysis of p-values

Method	chi2	p_value	studies
Fisher	38.938235	.00037283	7

. metap pvar,e(a)

Meta-analysis of p-values

Method	chi2	p_value	studies
Fisher	38.938235	.00037283	7

نتایج بالا با هم همخوان بوده و نشان از معنی داری ترکیب مقادیر p در هفت مطالعه بالا که ۴ مطالعه مقادیر معنی داری نداشت دارد.

## خلاصه و نتیجه گیری:

در این فصل مفاهیم و نحوه انجام متآنالیز بیان گردید. برای انواع مطالعات همگروهی، مورد شاهد، مقطعی و کارآزمایی های بالینی با گزارش پیامدهای با داده های کمی پیوسته و دو حالت در قالب ۵ مثال متآنالیز انجام گردید. در طی یک مثال نیز دستورات رایج و مهم متآنالیز با نرم افزار Stata تشریح گردید. نمودارهای مهم برای گزارش نتایج متآنالیز و گزارش تورش نشر نیز با ذکر دستورات و مثال کاربردی بیان گردیدند.

## تمرین ها :

- ۱- برای انجام مرور سیستماتیک چه اقداماتی لازم است؟
- ۲- متآنالیز با مرور سیستماتیک چه تفاوتی دارد؟
- ۳- کاربرد متآنالیز چیست؟
- ۴- انواع متآنالیز و مراحل انجام آن چیست؟
- ۵- شاخص های مورد نیاز برای انجام متآنالیز توصیفی چیست؟
- ۶- شاخص های مورد نیاز برای انجام متآنالیز تحلیلی چیست؟
- ۷- برای کار با نرم افزار و انجام متآنالیز به چه دستوراتی نیاز است؟
- ۸- انواع نمودارهای رایج در متآنالیز و کاربرد آنها را توضیح دهید.
- ۹- برای ارزیابی یکنواخت بودن مطالعات چه باید انجام داد؟
- ۱۰- مدل های تصادفی و ثابت را توضیح دهید.
- ۱۱- یک بانک اطلاعاتی پیرامون یک سوال پژوهشی تهیه و متآنالیز را با نرم افزار انجام دهید.
- ۱۲- فایل meta7 را باز نمایید. این فایل مربوط به درمان با منیزیم برای سکتة قلبی مربوط به ۱۵ کارآزمایی بالینی است. با این درمان موارد مرگ سکتة قلبی در بیماران کاهش می یابد. نمودار کیفی برای این متآنالیز ترسیم نمایید. متآنالیز را قبل از آن نیز انجام دهید. نمودار کیفی نمایش داده شده متقارن است یا نامتقارن؟

## فصل دهم: روایی و پایایی .....

مقدمه ای بر روان سنجی ابزارها

روایی

پایایی

مقدمه ای بر تحلیل عاملی

دستورات نرم افزار

کار با نرم افزار

تحلیل کلاس پنهان

تمرین ها

### • اهداف:

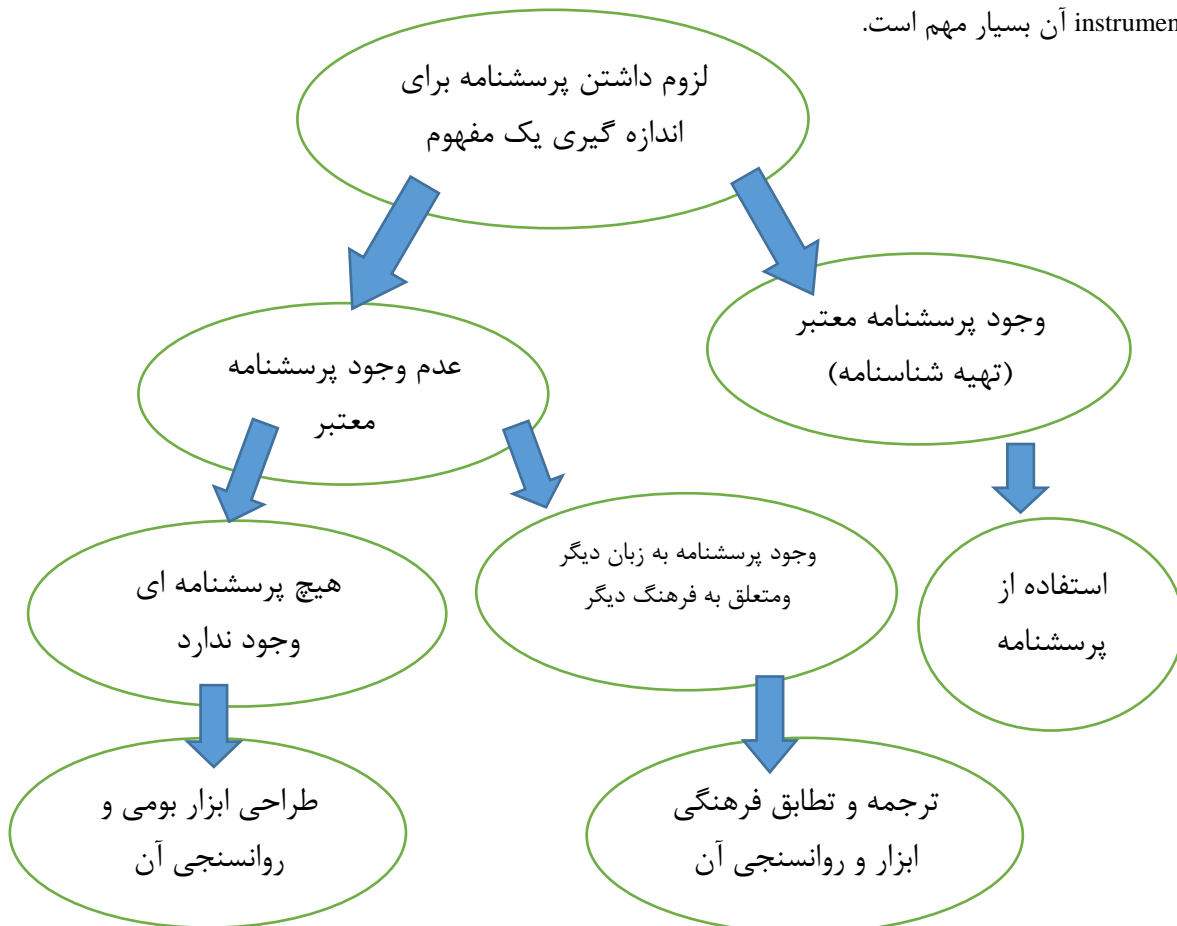
#### در پایان مطالعه این فصل انتظار می رود فراگیر با مباحث زیر آشنا گردد:

- کاربر بتواند مفاهیم روایی و پایایی و هدف از انجام تحلیل عاملی را بیان نماید.
- مراحل انجام تحلیل عاملی را تشریح نماید.
- برای پرسشنامه های مختلف، تحلیل عاملی را انجام و تفسیر نماید.
- شاخص های اندازه گیری روایی و پایایی را شناخته و آنها را بتواند با نرم افزار انجام دهد.
- کاربر بتواند مدل تحلیل کلاس پنهان را برای اندازه گیری متغیرهای پنهان بکار ببرد.

مقدمه ای بر روان سنجی ابزارهای جمع آوری داده ها:



برای جمع آوری اطلاعات در پژوهش های علوم پزشکی و اپیدمیولوژی روش های متعددی از جمله استفاده از پرونده ها و مستندات موجود، مدارک پزشکی، فرم ها، مشاهده، مصاحبه و پرسشنامه وجود دارد. هر کدام از این روشها معایب و مزایایی دارند. آنچه مهم است صحت، پایایی و اعتبار این روش ها و ابزارها است. یکی از رایجترین این ابزارها پرسشنامه است. پرسشنامه ها معمولا رایجترین ابزار و مخصوصا در مطالعات تعیین آگاهی، نگرش، رفتار و خطر درک شده در حوزه آموزش بهداشت و ارتقاء سلامت جایگاه و کاربرد ویژه ای دارند. برای طراحی یک پرسشنامه با تعیین هدف پرسشنامه، ترسیم نمودار درختی، وزن دهی و امتیاز گذاری نمودار، تعیین اعتبار نمودار درختی، تعیین بانک سوالات، تعیین اعتبار ظاهری، تعیین اعتبار محتوایی، ارزیابی پایایی سوالات و با انتخاب نهایی سوالات، پرسشنامه تهیه می شود. بنابراین برای جمع آوری داده ها باید از ابزاری استفاده نمود که معتبر و متناسب با بستر فرهنگی جامعه باشد. چنانچه چنین ابزار مناسبی وجود نداشت میبایست ابزاری معتبر را که در بستر فرهنگی جامعه تدوین نشده، ترجمه نمود. اگر چنین ابزاری نیز وجود نداشته باشد باید با رعایت شرایط و ضوابط مخصوص، ابزاری معتبر و پایا طراحی نمود. در صورت عدم رعایت این ضوابط نتایج پژوهش مخدوش می شود و مورد سوال قرار می گیرد. در طراحی یک ابزار، توجه به کیفیت روان سنجی (Psychometric properties of health instruments) آن بسیار مهم است.



روایی درجه ای از صحت نتایج مطالعه است و نشان میدهد که مطالعه تا چه اندازه ای چیزی را که قصد اندازه گیری آن را داشته است به درستی سنجیده است؟  
برای دستیابی به اعتبار قابل قبول در مطالعات پرسشنامه ای، لازم است که سوالات به گونه ای طراحی شوند که علاوه بر تناسب و انطباق با موضوع مورد سنجش، اولاً برداشت همه پاسخگویان از سوالات با یکدیگر یکسان بوده و ثانیاً برداشت پاسخگویان با برداشت طراح پرسشنامه مطابقت داشته باشد.

### انواع روایی

- ۱- روایی صوری یا ظاهری..... face validity
- ۲- روایی محتوا..... content validity
- ۳- روایی ضابطه یا ملاکی یا معیار: ..... Criterion validity
- الف) روایی همزمان،
- ب) روایی پیش بینی
- ج) روایی همگرایی
- د) روایی تمایز
- ۴- روایی سازه یا ساختار..... construct validity

### ۱- روایی صوری یا ظاهری (face validity)

روایی صوری: روایی صوری قضاوت عینی در مورد ساختار ابزار است و از نگاه گروه هدف به این سوالات پاسخ می دهد که:

آیا ابزار طراحی شده به صورت ظاهری با هدف مطالعه مرتبط است؟  
آیا افرادی که قرار است به ابزار پاسخ دهند، با عبارات و جمله بندی ابزار موافق هستند؟  
آیا اجزاء و کلیت ابزار قابل پذیرش پاسخ دهندگان می باشد؟  
جذاب بودن و توالی منطقی سوالات، چاپ مناسب و مقدمه گویا و مختصر، تناسب شکل سوالات با شخصیت پاسخ دهندگان نکات مهمی است که می تواند اعتبار ظاهری پرسشنامه را بالا ببرد.

### روش های تعیین روایی صوری

#### نحوه اندازه گیری روایی صوری با روش کیفی

در روش کیفی و در آغاز فرآیند تعیین روایی صوری، با کمک متخصصین امر در رشته های مرتبط با زمینه تحقیقاتی و پرسشنامه، از متخصصین آمار و اپیدمیولوژی و بر حسب موضوع از متخصصین روانشناسی یا علوم دیگر استفاده میشود.

۱- پرسشنامه فارسی به دو فرد هیأت علمی متبحر در زبان انگلیسی و واژه های تخصصی داده میشود تا نسخه نهایی فارسی را از نظر مطلوب بودن عبارات به لحاظ وضوح (استفاده از واژه های ساده و قابل فهم)، کاربرد زبان مشترک (پرهیز از به کارگیری واژه های فنی و تخصصی) مورد بررسی قرار دهند. در صورت لزوم میتوان تغییراتی را در سنجه مورد استفاده در جهت ساده تر کردن و قابلیت درک بیشتر انجام داد.

۲- همچنین، برای دریافت نظرات گروه هدف، مصاحبه ای با نمونه ای از آنها جهت یافتن دشواری در درک عبارات و کلمات، تناسب و ارتباط مطلوب آیتم ها، احتمال وجود ابهام و برداشت های نارسا از عبارات و یا وجود نارسایی در معانی کلمات انجام میشود

۳- در صورت وجود مشکلات بایستی نظرات آنان به صورت تغییراتی جزئی در پرسشنامه اعمال شود.

### نحوه اندازه گیری روایی صوری با روش کمی

روایی صوری سنجه ها به صورت کمی را با استفاده از روش تأثیر آیتم اندازه گیری می کنند. بدین منظور برای هر یک از آیتم های سنجه، طیف لیکرتی ۵ قسمتی در نظر گرفته میشود: کاملاً مهم است (امتیاز ۵)، مهم است (امتیاز ۴)، به طور متوسطی مهم است (امتیاز ۳)، اندکی مهم است (امتیاز ۲) و اصلاً مهم نیست (امتیاز ۱). سپس پرسشنامه جهت تعیین روایی صوری در اختیار گروه هدف قرار گرفته و پس از تکمیل پرسشنامه ها توسط گروه هدف، با استفاده از فرمول روش تأثیر آیتم، روایی صوری محاسبه میگردد:

$$Importance\ Score = \frac{\sum(f \times Importance)}{N}$$

f در فرمول تأثیر آیتم، فراوانی تکرار امتیاز توسط گروه هدف را نشان میدهد.

### مثال از اندازه گیری روایی ظاهری با روش کیفی (میرزاخانی و همکاران در سال ۲۰۱۶)

ترجمه، روایی ظاهری و محتوایی نسخه فارسی آزمون ارزیابی بازی وانمودین خودانگیخته ی کودکان ۴-۷ ساله چگونگی اندازه گیری روایی ظاهری با روش کیفی در این مقاله

برای بررسی روایی ظاهری پنج تن از کاردرمانگران در حیطه ی کودک و بازی در ارتباط با واضح و قابل فهم بودن پرسشنامه نظر دادند. نظر سه نفر از اساتید حوزه ی کودکان و بازی نیز در ارتباط با شکل کلی و قابل فهم بودن پرسشنامه نیز استخراج گردید و سپس با حضور اساتید راهنما و پژوهشگر نظرات این افراد مورد بررسی قرار گرفت و پرسشنامه نهایی فارسی تنظیم شد.

### مثالی از اندازه گیری روایی ظاهری با روش کمی (عبداللهی پور و همکاران در سال ۲۰۱۶)

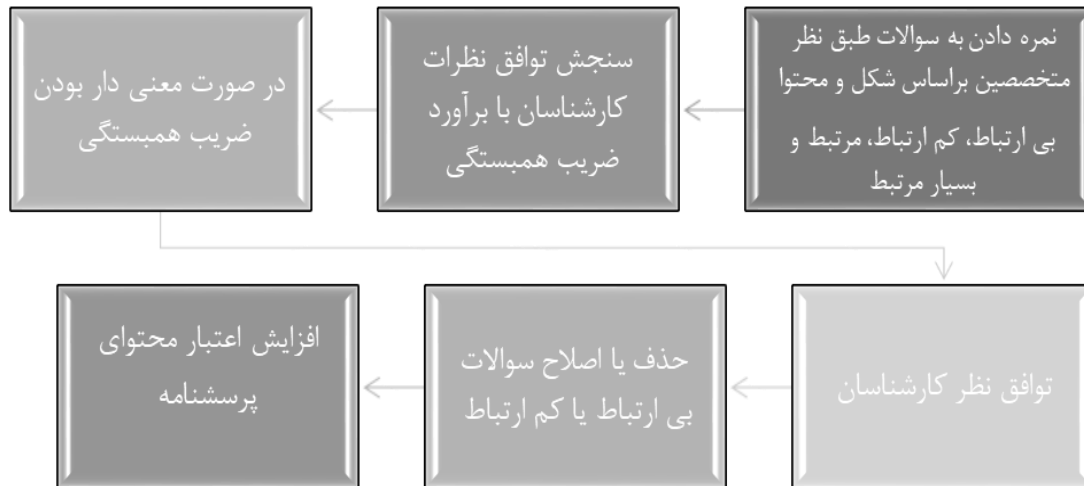
تهیه نسخه فارسی «پرسشنامه سنجش رفتاری کارکرد اجرایی-نسخه پیش دبستانی» و بررسی روایی ظاهری و محتوایی آن.

چگونگی اندازه گیری روایی ظاهری با روش کمی در این مقاله:

انتخاب ۲۰ نمونه (از مادران) و دادن فرم روایی صوری پرسشنامه ترجمه شده به آنان که در این فرم برای هر یک از ۶۳ سوال پرسشنامه سه قسمت مجزا طراحی شده بود که دشواری درک مفاهیم، ابهام و برداشت های ناروا و تناسب و ارتباط موارد را به صورت مجزا در یک طیف لیکرتی ۵ قسمتی مورد بررسی قرار میداد. توافق بالای ۸۰ درصد و نمره اثر بالای ۱,۵ بدین معناست که مورد از روایی ظاهری مناسبی برخوردار است و برای گروه نمونه به راحتی ساده قابل فهم، رسا و واضح است.

میزان کافی و متناسب تحت پوشش قرار دهند. داشتن یک اعتبار محتوای بالا به این معنی است که محقق به تمام جوانب موضوع مورد بررسی احاطه کامل داشته باشد. انواع روش های تعیین روایی محتوا:

### نحوه اندازه گیری روایی محتوا با روش کیفی



### نحوه اندازه گیری روایی محتوا با روش کمی

1-Content Validity Ratio (CVR) نسبت روایی محتوا

- جهت اطمینان از اینکه مهم ترین و صحیح ترین محتوا (ضرورت آیتم) انتخاب شده است

2-Content Validity Index (CVI) شاخص روایی محتوا

- برای اطمینان از این که آیتم های ابزار به بهترین نحو جهت اندازه گیری محتوا طراحی شده اند

روایی محتوا به صورت کمی نیز بر اساس نظرات متخصصین و با محاسبه دو شاخص اندازه گیری می شود:

### الف: نسبت روایی محتوی (CVR) Content Validity Ratio

جهت تعیین نسبت روایی محتوی به صورت کمی از تعدادی از متخصصین و صاحب نظران استفاده میشود تا در مورد هر کدام از آیتم های ابزار یا سنجه مورد استفاده به سه طیف آیتم شامل ضروری است، مفید اما ضروری نیست و ضرورتی ندارد، پاسخ دهند. بهتر آن است که از تعداد ارزیابان بیشتری (بالاتر از ۱۰ نفر) استفاده شود. CVR بر اساس نظر داوران به قرار ذیل محاسبه میشود:

$$CVR = \frac{(ne - \frac{N}{2})}{\frac{N}{2}}$$

ne = تعداد ارزیابانی که گویه مورد نظر را ضروری می دانند  
N = تعداد کل ارزیابان یا داوران

حداقل مقادیر CVR در تست های یک طرفه مورد نظر Lawshe جهت مناسب بودن روایی محتوی در جدول زیر آمده است:

حداقل مقدار نسبت روایی محتوی	تعداد ارزیابان
------------------------------	----------------

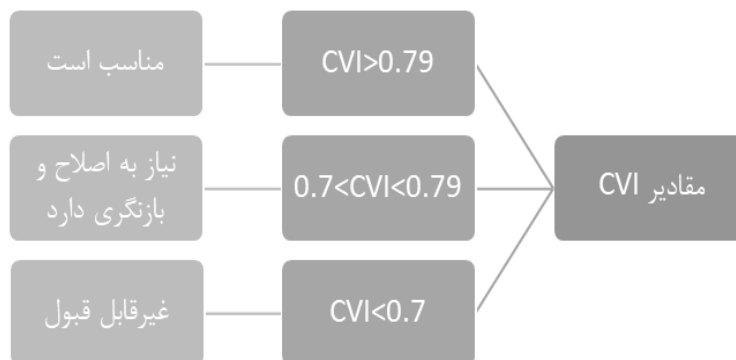
5	0/99
6	0/99
7	0/99
8	0/75
9	0/78
10	0/62
15	0/49
20	0/42
25	0/37
30	0/33
35	0/31
40	0/29

**ب: شاخص روایی محتوی (CVI) Content Validity Index**

برای شاخص روایی محتوا، ارزیابان می بایست به هر آیتم ابزار مورد استفاده، در خصوص سه معیار مربوط یا اختصاصی بودن، سادگی و روان بودن و وضوح یا شفاف بودن، بر اساس طیف لیکرتی ۴ قسمتی اظهار نظر می نمایند. مثلاً جهت معیار مربوط بودن گزینه های، مربوط نیست=۱، نسبتاً مربوط است =۲، و مربوط است =۳ و کاملاً مربوط است =۴، به کار گرفته می شود و سپس با استفاده از فرمول CVI شاخص روایی محتوا محاسبه میگردد.

$$CVI = \frac{\text{نسبت تعداد ارزیابی که به آیتم نمره ۳ و ۴ داده اند}}{\text{تعداد کل ارزیابان}}$$

مقادیر مناسب برای CVI



مثالی از اندازه گیری روایی محتوا با روش کمی (بغوری و همکاران در سال ۲۰۱۶) بررسی روایی ظاهری و محتوایی نسخه فارسی مقیاس خستگی پارکینسون (PFS-16) در افراد مبتلابه پارکینسون

**چگونگی اندازه گیری روایی محتوا با روش کمی**

الف) نسبت روایی محتوایی CVR :

در این روش از متخصصین خواسته شد بر پایه طبقه بندی زیر، هر یک از سوالات پرسشنامه را ارزیابی نمایند :

۱. ضروری است
۲. ضرورتی ندارد، اما میتوان استفاده کرد
۳. غیر مرتبط و بی اهمیت است.

**ب) شاخص روایی محتوایی CVI :**

مقدار قابل قبول برای هر سؤال براساس تعداد متخصصینی که روایی محتوا را تعیین می کنند، متفاوت می باشد که با فرمول زیر محاسبه می شود

$$CVI = \frac{\text{نسبت تعداد ارزیابی که به آیتم نمره ۳ و ۴ داده اند}}{\text{تعداد کل ارزیابان}}$$

نتایج روایی محتوا

نتیجه گیری: نسبت روایی محتوایی (CVR) بر اساس جدول لاوشه برای ۱۵ نفر متخصص حداقل ۰/۴۹ قابل قبول است که در این مطالعه برای ۱۶ سوال که حاصل پاسخ ۱۵ متخصص کاردرمانی حاصل شد در محدوده (۰/۷-۱) قرار گرفت. همچنین شاخص روایی محتوایی نیز بر اساس روش لاوشه در محدوده قابل قبول (۰/۹۴) قرار گرفت.

جدول شماره ۱: شاخص روایی محتوایی (CVI) به تفکیک سوالات (n= ۱۵)

سوال	مربوط بودن	واضح بودن	ساده بودن
سوال ۱	۱	۱	۱
سوال ۲	۱	۱	۱
سوال ۳	۱	۱	۱
سوال ۴	۰/۹۳	۰/۴۶	۰/۲۲
سوال ۵	۱	۰/۸۶	۰/۹۳
سوال ۶	۱	۱	۱
سوال ۷	۱	۱	۱
سوال ۸	۱	۰/۸۰	۰/۸۰
سوال ۹	۱	۱	۱
سوال ۱۰	۱	۰/۹۳	۰/۸۶
سوال ۱۱	۱	۱	۱
سوال ۱۲	۱	۰/۸۶	۰/۸۶
سوال ۱۳	۱	۱	۱
سوال ۱۴	۱	۱	۱
سوال ۱۵	۱	۱	۱
سوال ۱۶	۰/۹۳	۱	۱

**۳- روایی ضابطه (Criterion-related Validity)**

درجه ای از توافق بین یک تست با یک یا تعداد بیشتری تست معتبر دیگر در همان زمینه (که بهتر است دقیق و ایده آل باشد) است که معمولاً با استفاده از ضریب همبستگی بین آنها سنجیده و ارزیابی می شود.

روایی ضابطه به ۴ دسته تقسیم می شود:

۱- روایی همزمان concurrent validity.....

- ۲- روایی پیش بینی کننده..... predictive validity  
 ۳- روایی همگرایی..... convergence validity  
 ۴- روایی تمایز..... Discriminant Validity

### الف - روایی همزمان (Concurrent Validity)

روایی همزمان: در این جا توان یک پرسشنامه در افتراق بین گروه هایی که از نظر تئوری، باید توان افتراق آن ها را داشته باشد، ارزیابی می شود. برای مثال، پرسشنامه ای که برای شناخت افراد مانیک دپرسیو تهیه شده است باید توان افتراق این افراد را از افراد اسکیزوفرنیک پارانوئید داشته باشد. هرچه پرسشنامه توان افتراق گروه های مشابه تری را داشته باشد، اعتبار همزمان آن بالاتر است.

### ب-روایی پیش بینی کننده (Predictive Validity)

ارزیابی توان پیش بینی پرسشنامه ؛ مثلا پیش بینی نمره پایان ترم از میان ترم . یکی از ارزشمندترین اعتبارات پرسشنامه است. مثال: پرسشنامه شخصیت شناسی. اعتبار سنجی پرسشنامه درمورد اینکه " آیا پرسشنامه شخصیت شناسی قادر به پیش بینی رنگ مورد علاقه افراد هست؟"  
 پرسشنامه ای در اختیار افراد قرار داده شده که به بررسی ساختار شخصیتی آن ها می پرداخت، درضمن در مورد رنگ مورد علاقه افراد هم سوالاتی مطرح شده بود. بعد از پرشدن پرسشنامه، افراد به اتاقی راهنمایی شدند که وسایل گوناگون در رنگ های مختلف در آن چیده شده بود و از افراد خواسته شد که به عنوان هدیه و قدردانی از شرکت در آن تحقیق چند وسیله را انتخاب نمایند. البته هدف قسمت آخر یعنی انتخاب هدیه کاملا مخفی بود و شرکت کنندگان تصور می کردند که بعد از اتمام پرسشنامه، تحقیق بر روی آن ها پایان پذیرفته است، در حالیکه رفتارشان در انتخاب رنگ وسایل، به وسیله دوربین های مخفی کنترل می شد. در انتها نتایج پرسشنامه، یعنی ساختار شخصیت شرکت کنندگان ، با نتایج قسمت دوم یعنی انتخاب رنگ ها مقایسه شد، و نتیجه آن بود که پرسشنامه مذکور اعتبار کافی برای پیش بینی رنگ مورد علاقه افراد ندارد.

ج- روایی همگرایی (Convergent Validity): عبارتست از همبستگی نسبتاً قوی میان سؤال مربوط به یک حیطة با همان حیطة. که بنا بر توصیه سازمان اروپایی تحقیقات و درمان سرطان ضریب همبستگی پیرسون باید بزرگتر از ۰/۴ باشد. اصل در جهت گروه بندی متغیرها نیز در یک حیطة خاص بر این منوال است که چنانچه ضریب همبستگی جدول ماتریس همبستگی:

کمتر از ۰/۳ ± باشد حداقل بوده	۰/۴ ± بیانگر مهم بودن متغیر	۰/۵ ± بیانگر اهمیت اساسی متغیر
-------------------------------	-----------------------------	--------------------------------

### د- روایی تمایز، تشخیصی، واگرا یا تفکیکی: (Discriminant Validity)

بر همبستگی ضعیف بین سؤالات مربوط به یک حیطة با دیگر حیطة ها دلالت دارد. ضریب همبستگی یک سؤال با حیطة های دیگر باید کمتر از ضریب همبستگی آن سؤال با حیطة های مربوط به خودش باشد.

شاخص رایج برای تعیین میزان روایی تشخیصی حساسیت (sensitivity) و ویژگی (specificity) آزمون است. منظور از حساسیت، این است که آزمون چند درصد گروه بیمار را بیمار تشخیص می دهد و منظور از ویژگی آزمون، این است که چند درصد از افراد گروه سالم نمره پایین تر از نقطه برش (cut off point) کسب کرده اند.

یک ابزار موفق علاوه بر آنکه باید با ابزارهای دیگر که سازه مشابه را اندازه گیری میکند همگرا باشد باید با ابزارها یا مقیاس های مفاهیم دیگری که یک نظریه را بامفهوم کانونی پیوند میزند نیز همبستگی بالایی داشته باشد. علاوه بر آن هرچند اعتبار همگرا لازم است اما کافی نیست و اعتبار تفکیکی در حد اعتبار همگرا لازم است زیرا به بررسی حساسیت ابزارها در سنجش پدیده ای غیرمربوط می پردازد و میزان ابهام یا گسترده بودن مفهوم را نشان میدهند.

#### ۴-روایی ساختار (Construct Validity)

#### اعتبار سازه ای نشانگر توافق بین میزان هایی است که مطالعه عملاً قادر به اندازه گیری آن هاست، با آنچه

ظواهر مورد ارزیابی قرار گرفته است. بدین معنی که پرسشنامه معتبر از نظر اعتبار سازه ای، باید قادر باشد همان چیزی را بسنجد که برای آن طراحی شده است. آیتم های یک ابزار میتواند بر اساس تحلیل عاملی تعیین شود. هدف اصلی تحلیل عاملی تلخیص تعداد زیادی از متغیرها در تعداد محدودی از عامل ها می باشد، بطوریکه کمترین میزان از دست رفتن اطلاعات را داشته باشیم. تحلیل عاملی به دو شکل کلی انجام میگردد که شامل تحلیل عاملی اکتشافی و تأییدی است.

#### تحلیل عاملی اکتشافی Exploratory factor analysis

#### در تحلیل اکتشافی پژوهشگر به دنبال بررسی داده های تجربی به منظور کشف و شناسایی شاخص ها و نیز روابط بین آنهاست. اگر هیچ حدسی از ساختار روابط میان گویه ها نباشد از تحلیل عاملی اکتشافی استفاده می شود. به این صورت آیتم هایی از ابزار که در هر فاکتور بیشترین همبستگی را با یکدیگر نشان میدهند میتوانند آیتم های مورد استفاده در تبیین هر فاکتور یا مؤلفه ابزار قرار گیرند.

تحلیل عاملی اکتشافی معمولاً با روش مؤلفه های اصلی (Principle Component Analysis (PCA انجام می گیرد.

تحلیل عاملی اکتشافی معمولاً با روش مؤلفه های اصلی (Principle Component Analysis (PCA انجام می گیرد.

#### تحلیل عاملی تأییدی Confirmatory factor analysis

#### تحلیل عاملی تأییدی هنگامی کاربرد دارد که ابزار یا پرسشنامه مورد استفاده از روایی سازه مناسب برخوردار باشد به منظور تأیید کارهای قبلی به کار می رود. بنابراین امکان استخراج فاکتورهای جدید در این نوع تحلیل عاملی وجود ندارد. در تحلیل عاملی تأییدی پژوهشگر به دنبال تهیه مدلی است که فرض می شود داده های تجربی را بر پایه چند پارامتر نسبتاً اندک، توصیف تبیین یا توجیه می کند. روشهای تأییدی (آزمون فرضیه) تعیین می کنند که داده ها با یک ساختار عاملی معین (که در فرضیه آمده) هماهنگ هستند یا نه.

تحلیل عاملی تأییدی هنگامی کاربرد دارد که ابزار یا پرسشنامه مورد استفاده از روایی سازه مناسب برخوردار باشد به منظور تأیید کارهای قبلی به کار می رود. بنابراین امکان استخراج فاکتورهای جدید در این نوع تحلیل عاملی وجود ندارد. در تحلیل عاملی تأییدی پژوهشگر به دنبال تهیه مدلی است که فرض می شود داده های تجربی را بر پایه چند پارامتر نسبتاً اندک، توصیف تبیین یا توجیه می کند. روشهای تأییدی (آزمون فرضیه) تعیین می کنند که داده ها با یک ساختار عاملی معین (که در فرضیه آمده) هماهنگ هستند یا نه.

#### حجم نمونه تحلیل عاملی

در خصوص حجم نمونه لازم جهت انجام تحلیل عاملی تناقضات متفاوتی وجود دارد . Tabachnick . حداقل ۳۰۰ نمونه را برای انجام تحلیل عاملی ضروری میدانند. اما Hair و همکاران بیان میکنند که حجم



## روش های اپیدمیولوژی و پژوهش با Stata

نمونه باید بیش از ۱۰۰ باشد. برخی منابع دیگر از جمله Lee و Comrey طبقه بندی انجام داده اند که ۱۰۰ را ضعیف، ۲۰۰ را نسبتاً خوب، ۳۰۰ را خوب، ۵۰۰ را خیلی خوب و ۱۰۰۰ و بیشتر را عالی معرفی کرده اند.

### بررسی مسیر سنجش همبستگی

پرسشنامه ای با ۱۲ سوال که دارای مقیاس لیکرتی ۴ گزینه ای : موافقم (کد۴)، بی نظر(کد۳)، مخالفم (کد۲)، کاملاً مخالفم (کد۱) مربوط به سلامت عمومی (GHQ) است را باز کنید.

Data Editor (Edit) - [Untitled]  
File Edit View Data Tools

ID[1]	GHQ1	GHQ2	GHQ3	GHQ4	GHQ5	GHQ6	GHQ7	GHQ8	GHQ9	GHQ10	GHQ11	GHQ12
1	2	2	2	2	2	2	1	1	1	1	1	1
2	3	2	3	3	2	3	1	3	1	2	2	2
3	2	2	2	2	2	2	2	1	2	1	1	1
4	2	2	2	2	2	2	2	1	2	2	2	1
5	2	2	1	3	3	2	2	1	2	2	2	1
6	3	2	2	4	2	2	2	3	3	1	3	3
7	3	2	2	2	2	2	2	2	3	2	2	3
8	2	2	2	2	2	3	1	2	3	2	2	2
9	2	2	2	2	2	2	2	1	1	2	1	1
10	2	2	1	2	2	2	1	1	1	1	1	1
11	2	2	2	3	3	2	1	1	1	1	1	1
12	2	2	1	1	1	2	3	2	2	1	1	1
13	2	1	2	2	2	1	1	1	1	1	1	1
14	2	2	2	2	2	2	1	1	1	1	1	1
15	3	2	2	3	3	3	3	2	3	3	2	2
16	2	2	2	1	2	2	2	2	1	1	1	1
17	2	2	2	2	2	2	2	2	1	2	1	1
18	1	1	2	1	3	2	1	1	1	1	1	1
19	2	2	2	2	2	2	3	2	2	3	2	2
20	2	2	3	2	1	2	3	2	2	2	2	2
21	2	2	2	2	2	3	1	1	3	2	1	1
22	2	2	2	2	2	2	3	2	1	1	1	1
23	2	2	2	2	3	3	3	3	3	3	3	3
24	2	2	2	2	3	1	3	2	2	1	2	2
25	2	2	2	2	2	2	3	1	2	2	1	1
26	2	2	2	2	2	2	2	2	2	1	1	1
27	2	2	2	2	2	2	1	2	2	2	1	1

پرسشنامه در مورد سنجش سلامت عمومی است. محقق قصد دارد از بین این ۱۲ سوال، ۳ سوال را انتخاب کند.

مرحله اول: ساختن ستون جمع نمرات هر سوال با دستور زیر در استاتا است:

```
gen TotalGHQ = ( GHQ1 + GHQ2 + GHQ3 + GHQ4 + GHQ5 + GHQ6 + GHQ7 + GHQ8 + GHQ9 + GHQ10 + GHQ11 + GHQ12 )
```

مرحله دوم: به دست آوردن همبستگی بین ستون total با تک تک سوالات. با دستور زیر در استاتا این کار انجام می شود.

```
pwcorr TotalGHQ GHQ1 GHQ1 GHQ2 GHQ3 GHQ4 GHQ5 GHQ6 GHQ7 GHQ8 GHQ9 GHQ10
GHQ11 GHQ12, sig
. pwcorr TotalGHQ GHQ1 GHQ1 GHQ2 GHQ3 GHQ4 GHQ5 GHQ6 GHQ7 GHQ8 GHQ9 GHQ10 GHQ11 GHQ12, sig
```

	TotalGHQ	GHQ1	GHQ1	GHQ2	GHQ3	GHQ4	GHQ5
TotalGHQ	1.0000						
GHQ1	0.3818 0.0000	1.0000					
GHQ1	0.3818 0.0000	1.0000 0.0000	1.0000				
GHQ2	0.3886 0.0000	0.2171 0.0000	0.2171 0.0000	1.0000			
GHQ3	0.4688 0.0000	0.2185 0.0000	0.2185 0.0000	0.2712 0.0000	1.0000		
GHQ4	0.3797 0.0000	0.1946 0.0000	0.1946 0.0000	0.2615 0.0000	0.2186 0.0000	1.0000	

از طریق ضریب همبستگی مشخص می شود که همه سوالات همبستگی معنی داری با جمع سوالات دارند در نتیجه سوالی حذف نمی شود و همه سوالات می توانند در پرسشنامه باشند. چنانچه سوالی همبستگی غیرمعنی داری داشت باید حذف شود. فرض کنید ۳ سوال همبستگی معنی داری نداشتند. در مرحله روایی ۳ سوال حذف می شوند، حال در مرحله پایایی باید بررسی شود که ۳ سوال برتر که مدنظر محقق بودند کدام سوالات هستند.

## Reliability

## پایایی

پایایی آزمون یعنی توافق و همخوانی بین نتایج حاصله از دفعات متعدد انجام یک آزمون، نتایج به دست آمده حاصل از چند بار اندازه گیری تا چه حد: مشابه (similar)، قابل اعتماد (dependable)، دقیق (accurate) و پیش بینی پذیر (predictable) هستند.

### Reliability $\neq$ stability

کلمه Reliability دقیقاً به معنای پایایی و ثبات نیست بلکه مفاهیم زیر را در بر میگیرد و کلمه پایایی بیشتر رایج شده است: ثبات یا هماهنگی، حساسیت، اعتمادپذیری، همگونی، بازپدیدآوری، دقت، تکرارپذیری.

برای رسیدن به پایایی چند شرط لازم است:

نتایج تحقیق باید از فرد انجام دهنده ی تحقیق مستقل باشد. بهترین شرایط در هر تحقیق این است که نتایج تحقیق به فرد انجام دهنده و مهارت و نظر او بستگی نداشته باشد.

نتایج تحقیق باید از زمان تحقیق مستقل باشد.

نتایج تحقیق باید از مکان و محل انجام تحقیق مستقل باشد

نتایج تحقیق باید مستقل از ابزارهای خاص استفاده شده در تحقیق باشد.

وجود پایایی شرط لازم برای روایی است. بدون پایایی، یک ابزار اعتبار نیز نخواهد داشت.

### انواع خطا و رابطه با پایایی و روایی

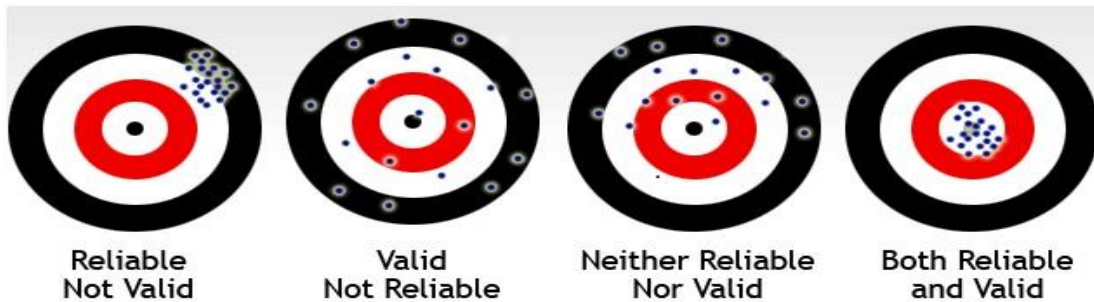
پایایی با خطای تصادفی در ارتباط است و روایی با خطای منظم و بنابراین هرچقدر که تعداد حجم نمونه بالاتر برود میتواند بر کاهش خطای تصادفی تأثیر گذارد و بنابراین پایایی ابزار افزایش می یابد و بنابراین بر دقت اندازه گیری مؤثر است. پایایی سنجشی از وجود خطای اندازه گیری تصادفی random measure error را منعکس می کند که این خطا به ۳ دلیل می تواند ایجاد شود:

**الف)** فاکتورهای مربوط به آزمون شونده، مثل تمرکز سوالات، خستگی، ملالت، بی دقتی در پاسخگویی، حواس پرتی، بدشانسی در انتخاب تصادفی (حدس. luck in guessing)

**ب)** فاکتورهای مربوط به آزمون: مثل سوالات اختصاصی انتخاب شده، سوالات مبهم یا مجهول و یا مسیرهای نامشخص poor directions

**ج)** فاکتورهای مرتبط با امتیاز دهی مثل دستورالعمل امتیاز دهی غیر یکنواخت، بی دقتی و خطاهای شمارشی یا محاسباتی (A.Wollack)

مثال معروف سیبل تیراندازی را به خاطر بیاورید. رابطه Validity and Reliability در زیر اشاره شد.



رابطه بین دقت و صحت را نیز در شکل زیر مشاهده و با شکل بالا تطبیق و مقایسه نمایید.



### ضریب پایایی

$$X = \text{true} + \text{error}$$

مقدار اندازه گیری شده  $X$  = مقدار صحیح و واقعی = true

error و true هیچ همبستگی باهم ندارند

وقتی دو متغیر بایکدیگر همبسته نباشند واریانس مجموع آن ها برابر است با جمع واریانس

$$= V_t + V_e V_0 \quad \text{هریک از آن ها}$$

با تقسیم دو طرف معادله به  $V_0$  خواهیم داشت :

$$V_0/V_e = 1 - V_0/V_t$$

$$\text{ضریب پایایی} = \frac{V_t}{V_0}$$

ضریب پایایی نشان دهنده میزان دقت و پایایی ابزار اندازه گیری است. این ضریب عددی بین ۰ تا ۱ می باشد. ضریب پایایی ۰ نشان دهنده پایا نبودن ابزار و ضریب پایایی ۱ نشان دهنده پایایی کامل ابزار است. و عدد منفی دلیل این است که کد بندی سوالات جابجا بوده و به شکل معکوس انجام شده است. همه سنجش های توانایی های بشری، مقداری خطا را شامل می شوند و ۱۰۰ درصد پایا و معتبر نیستند.

### انواع روش های سنجش پایایی

تکنیک های مورد استفاده برای سنجش پایایی یک ابزار می تواند به دلیل عملکرد و طبیعت ابزار (پرسشنامه) و چگونگی استفاده از آن در طرح تحقیقاتی متفاوت باشد. معمول ترین آن ها عبارت اند از :

#### ۱- پایایی باز آزمون **test- retest reliability** :

این روش تعیین پایایی با اندازه گیری مکرر (دو نوبت) به یک گروه واحد، تحت شرایط مشابه داده می شود و نمرات حاصل با هم مقایسه می شوند. در نهایت ضریب همبستگی بین نمرات حاصل از اجرای دو بار آزمون برآورد شده تا چگونگی شباهت امتیازات مشخص گردد و به عنوان ضریب پایایی به کار می رود. این روش فرض می کند که متغیرها یا مفاهیم مورد اندازه گیری و همچنین خصوصیات آزمون شوندگان در طول دوره تغییر نخواهند کرد. و این فرض هنگامی که موضوع تحت بررسی یک جسم بی جان است راحتتر محقق می شود. قابل قبول ترین آزمون جهت تعیین ثبات آزمون (آزمون پایایی)، شاخص همبستگی درون خوشه ای (Interclass Correlation Coefficient) است . چنانچه این شاخص بالاتر از ۰/۸ باشد میزان ثبات مطلوب است.

مشکلات روش: **test- retest**

آشنایی آزمون شوندگان با محتوی سوالات در اثر تجربه اول که بر روی پاسخگویی در نوبت دوم تأثیر می گذارد. به عبارتی آزمون ممکن است به دلیل اثر عمل - یادگیری **learning-practice effect** ناشی از آزمون اول نامعتبر شود. تعیین فاصله زمانی مناسب برای اجرای دو نوبت آزمون یعنی فاصله زمانی کوتاه می تواند سبب برآورد پایایی بسیار بالا شود و از طرفی فاصله زمانی طولانی به دلیل پختگی **maturation effect** تغییر افراد تحت بررسی در طول دوره ، می تواند پایایی را نامعتبر سازد. عدم علاقه آزمون شوندگان به اجرای نوبت دوم به دلیل تکراری بودن و اتلاف وقت و... می تواند سبب از دست رفتن بعضی از افراد شده و یا باعث تورش عدم پاسخگویی **non response bias** شود .

#### ۲- پایایی بین مشاهده کنندگان **inter rater-observer reliability**

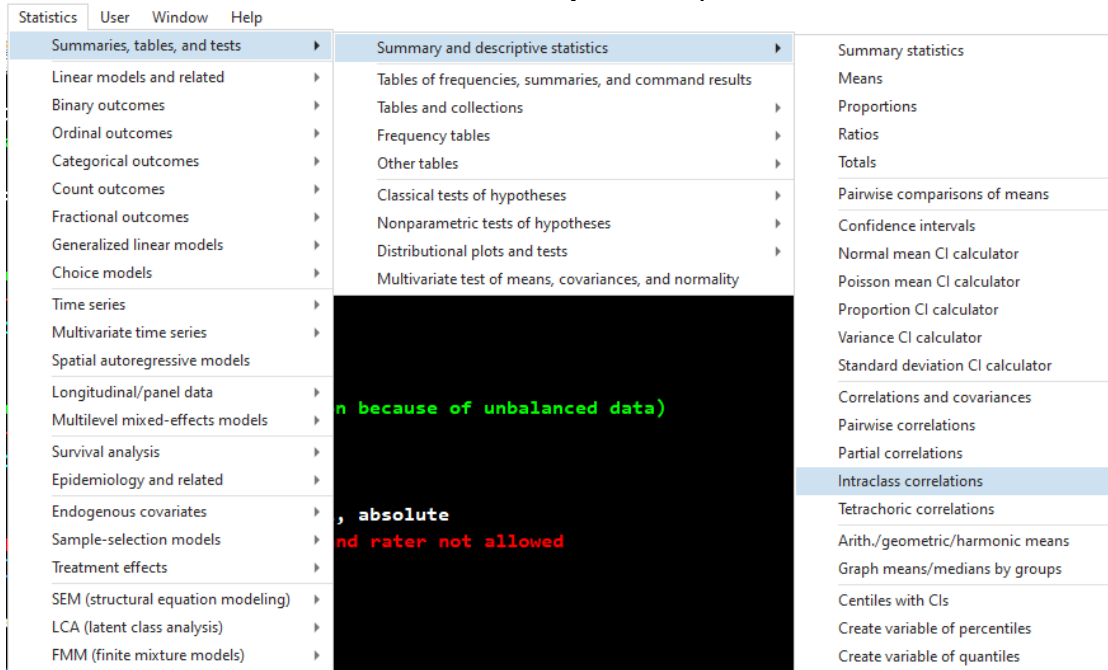
این ضریب، ثبات و همبستگی نمرات بین مشاهده کنندگان یا داوران یا مصححان مختلف یک موضوع را می سنجد و ضریب توافق هم نامیده می شود. زمانیکه یک ابزار یا پرسشنامه بر روی افراد مشابهی توسط دو یا چند مصاحبه گر یا پرسشگر به کار رود و بخواهیم به یک توافق کلی در مورد ابزار دست یابیم.

مثال: خواندن ۱۰۰ کلیشه رادیولوژی عکس قفسه سینه توسط دو رادیولوژیست، برای این منظور یک بار تعداد ۱۰۰ کلیشه را در اختیار یکی از رادیولوژیست ها قرار می دهیم و از وی می خواهیم که آن ها را بررسی کند و در نهایت گزارش کند که آیا ۱۰۰ کلیشه ضایعه ای دارند یا نه. بدین ترتیب وی ۱۰۰ نتیجه به ما گزارش خواهد داد که هر کدام یکی از این دو حالت خواهند بود: ضایعه دار ، بدون ضایعه. سپس کلیشه ها را به رادیولوژیست دوم می دهیم و از او می خواهیم ۱۰۰ نتیجه به ما گزارش دهد. حال این دو

گزارش باهم متناظرند ولی وابسته نیستند و مستقل اند. در این جا پایایی به این معنی است که چقدر دو نفر مشاهده گر در تفسیر کلیشه یکسان هستند و به نتایج مشابه می رسند. برای محاسبه میزان توافق می توان از آماره کاپا، ضرایب فی و کرامر استفاده کرد.

برای محاسبه ICC از مسیر زیر اقدام نمائید. یا با دستور db icc جعبه گفتگو را فراخوانی کنید.

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Intraclass correlations



جعبه گفتگو متغیرها را معرفی و شاخص ICC را محاسبه یا با دستورات زیر محاسبات را انجام دهید. فایل ICC را باز نمائید و دستورات زیر را اجرا کنید. برای محاسبه ICC برای مدل اثرات تصادفی یک طرفه

```
. icc rating target
. icc rating target

Intraclass correlations
One-way random-effects model
Absolute agreement

Random effects: target      Number of targets =      6
                          Number of raters =       4

+-----+-----+-----+
| rating |      ICC | [95% conf. interval] |
+-----+-----+-----+
| Individual | .1657418 | -.1329323 .7225601 |
| Average   | .4427971 | -.8844422 .9124154 |
+-----+-----+-----+

F test that
ICC=0.00: F(5.0, 18.0) = 1.79          Prob > F = 0.165

Note: ICCs estimate correlations between individual measurements
and between average measurements made on the same target.
```

برای محاسبه ICC برای مدل اثرات تصادفی یک یک طرفه و آزمون برابری با  $\alpha = 0.5$  (ICCs equal 0.5)

```
. icc rating target, testvalue(.5)
. icc rating target, testvalue(.5)

Intraclass correlations
One-way random-effects model
Absolute agreement

Random effects: target      Number of targets =      6
                          Number of raters   =      4
```

rating	ICC	[95% conf. interval]	
Individual	.1657418	-.1329323	.7225601
Average	.4427971	-.8844422	.9124154

```
F test that
ICC(1)=0.50: F(5.0, 18.0) = 0.36      Prob > F = 0.870
ICC(k)=0.50: F(5.0, 18.0) = 0.90      Prob > F = 0.504
```

برای محاسبه ICC برای مدل اثرات تصادفی دو طرفه

```
. icc rating target judge

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: target      Number of targets =      6
Random effects: judge      Number of raters   =      4
```

rating	ICC	[95% conf. interval]	
Individual	.2897638	.0187865	.7610844
Average	.6200505	.0711368	.927232

```
F test that
ICC=0.00: F(5.0, 15.0) = 11.03      Prob > F = 0.000
```

برای محاسبه ICC برای مدل اثرات تصادفی دو طرفه و پایداری توافق

```
. icc rating target judge, consistency
. icc rating target judge, consistency

Intraclass correlations
Two-way random-effects model
Consistency of agreement

Random effects: target      Number of targets =      6
Random effects: judge      Number of raters   =      4
```

rating	ICC	[95% conf. interval]	
Individual	.7148407	.3424648	.9458583
Average	.9093155	.6756747	.9858917

```
F test that
ICC=0.00: F(5.0, 15.0) = 11.03      Prob > F = 0.000
```

برای محاسبه ICC برای مدل اثرات مختلط (mixed-effects model) دو طرفه

```
. icc rating target judge, mixed
```

## روش های اپیدمیولوژی و پژوهش با Stata

```
. icc rating target judge, mixed

Intraclass correlations
Two-way mixed-effects model
Consistency of agreement

Random effects: target      Number of targets =      6
Fixed effects: judge       Number of raters   =      4
```

rating	ICC	[95% conf. interval]	
Individual	.7148407	.3424648	.9458583
Average	.9093155	.6756747	.9858917

```
F test that
ICC=0.00: F(5.0, 15.0) = 11.03      Prob > F = 0.000
```

برای محاسبه ICC برای مدل اثرات مختلط (mixed-effects model) دو طرفه

```
. icc rating target judge, mixed absolute

Intraclass correlations
Two-way mixed-effects model
Absolute agreement

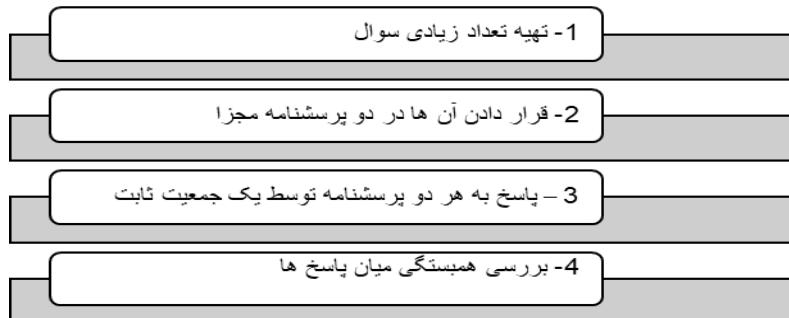
Random effects: target      Number of targets =      6
Fixed effects: judge       Number of raters   =      4
```

rating	ICC	[95% conf. interval]	
Individual	.2897638	.0187865	.7610844
Average	.6200505	.0711368	.927232

```
F test that
ICC=0.00: F(5.0, 15.0) = 11.03      Prob > F = 0.000
```

### ۳- پایایی تست های معادل یا موازی parallel

طراحی دو پرسشنامه کاملا هماهنگ و هم ارز برای بررسی یک موضوع



### ۴- روش همسانی درونی Internal consistency

تاکنون در مورد روش هایی صحبت شد که بر تکرار یک تست یا انجام بیش از یک سنجش تاکید داشتند ولی معمولا مشکل اصلی این است که محقق در بیشتر مواقع امکان تکرار تست یا انجام تست های معادل را ندارد. این روش بیش از سایر روش ها مورد توجه است زیرا محقق تنها یک ابزار را در گروه واحدی از آزمون شوندگان مورد بررسی قرار می دهد. برای محاسبه همسانی درونی روش های آماری متفاوتی وجود دارد:

الف - ضریب آلفای کرونباخ ← معمول ترین روش

ب - ضریب کودر ریچاردسون

ج - روش دونیمه کردن

## Cronbach alpha

## الف - ضریب آلفای کرونباخ

Lee J Cronbach در سال ۱۹۵۱ در دانشگاه Stanford روش آماری ضریب آلفا را برای حل مشکل تعیین پایایی آزمونهای چند سوالی ابداع کرد. مثال: پرسشنامه اندازه گیری سلامت عمومی (GHQ) را باز نمائید. منوی محاسبه مقدار آلفای کرونباخ با دستور db alpha باز نمائید. متغیرها را معرفی کنید. با دستورهایی زیر قابل محاسبه است.

.alpha GHQ1 - GHQ12

.alpha GHQ1 GHQ2 GHQ3 GHQ4 GHQ5 GHQ6 GHQ7 GHQ8 GHQ9 GHQ10 GHQ11 GHQ12, asis item

. alpha GHQ1 - GHQ12

Test scale = mean(unstandardized items)

Average interitem covariance: .1037504  
Number of items in the scale: 12  
Scale reliability coefficient: 0.7967

در جدول همبستگی تمامی مقادیر معنی دار بوده است. همبستگی منفی نیز به چشم نمی خورد. کمترین مقدار همبستگی برابر ۰/۳۸ و بیشترین برابر ۰/۷۱ است. میانگین تمام اعداد همبستگی برابر ۰/۷۹ است. به طور معمول مقادیر همبستگی مورد جمع بالای ۰/۳ مورد قبول است و کمتر از ۰/۳ را از مجموع سوالات باید حذف کرد.

. alpha GHQ1 GHQ2 GHQ3 GHQ4 GHQ5 GHQ6 GHQ7 GHQ8 GHQ9 GHQ10 GHQ11 GHQ12, asis item

Test scale = mean(unstandardized items)

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	alpha
GHQ1	7610	+	0.3821	0.2880	.1146323	0.7942
GHQ2	7610	+	0.3875	0.2732	.1132762	0.7959
GHQ3	7610	+	0.4690	0.3536	.1091223	0.7898
GHQ4	7610	+	0.3793	0.2476	.1129865	0.7993
GHQ5	7610	+	0.4427	0.3352	.1110548	0.7910
GHQ6	7608	+	0.4700	0.3626	.1096845	0.7889
GHQ7	7610	+	0.5900	0.4469	.0988277	0.7828
GHQ8	7606	+	0.7155	0.6100	.091924	0.7630
GHQ9	7609	+	0.6819	0.5623	.0932038	0.7687
GHQ10	7609	+	0.7070	0.6080	.0939487	0.7640
GHQ11	7609	+	0.6994	0.6035	.0952743	0.7651
GHQ12	7607	+	0.6257	0.5270	.1010707	0.7742
Test scale					.1037504	0.7967

## فرمول آلفای کرونباخ

حال شاخص همسانی درونی آلفای کرونباخ را به شکل فرمول زیر می توان تعریف کرد:

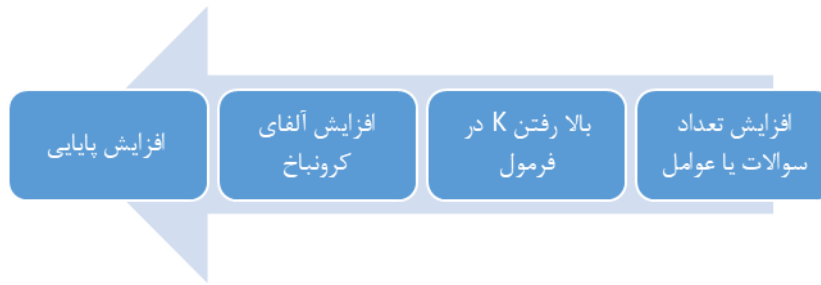
$$\alpha = \frac{K}{K-1} \times \left( 1 - \frac{K}{K+K(K-1)r^2} \right)$$

k = تعداد سوالات که برای سنجش یک موضوع به کار می رود ← در این مثال ۱۲ سوال

r<sup>۲</sup> = میانگین ضرایب همبستگی ← در این مثال برابر ۰/۷۹

ارتباط بین آلفای کرونباخ و تعداد سوالات پرسشنامه





$$\alpha = \frac{K}{K-1} \times \left( 1 - \frac{K}{K+K(K-1)r^2} \right)$$

اگر پرسشنامه چند حیطة داشته باشد باید برای هر حیطة آلفای مجزا حساب نمود و بعد میانگین وزن داده شده ی آلفاهای حیطة های مختلف به عنوان آلفای کل پرسشنامه بیان می شود. مبنای وزن دادن، تعداد سوالات هر حیطة می باشد و در این میان هر حیطة حداقل باید ۷ سوال داشته باشد.

### فرمول ارتباط بین طول پرسشنامه و ضریب پایایی

$$R = \frac{nr}{1+(n-1)r}$$

R = پایایی پرسشنامه جدید    r = ضریب پایایی پرسشنامه    n = حجم پرسشنامه یا تعداد سوالات  
 مثال : اگر ضریب پایایی پرسشنامه ای با ۲۰ سوال برابر ۰/۵ باشد، با افزایش چه تعداد سوال پایایی آن به ۰/۸۳ می رسد؟ با افزایش تعداد سوالات به ۱۰۰ سوال

نکته : با افزایش طول پرسشنامه و زیاد کردن تعداد سوالات کیفیت پاسخ گویی پرسش شونده، به دلیل خستگی و بی حوصلگی کاهش می یابد. بهترین حالت این است که در نهایت نزدیک به ۵ تا ۱۰ سوال برای هر موضوع طراحی شود. Gorge و Mallery در سال ۲۰۰۳ قاعده عملی زیر در مورد آلفای کرونباخ را ارائه کردند :

• >0.9	= Exellent
• 0.8-0.9	= good
• 0.7-0.8	= acceptable
• 0.6-0.7	= questionable
• 0.5-0.6	= poor
• <0.5	= unacceptable

### ب) ضریب کودرریچاردسون Kuder - Richardson reliability coefficient

دقیقا معادل ضریب آلفای کرونباخ است. در مواقعی که سوالات به صورت متغیر اسمی دوحالته تعریف شده از این ضریب استفاده می شود. یعنی سوالات آزمون باید به صورت پاسخ صحیح ۱ و پاسخ غلط ۰ ، تصحیح شده باشند. دارای ۲ فرمول KR-20 و KR-21 است.

$$KR-20 = \frac{n}{n-1} \left( 1 - \frac{\sum pq}{\sigma^2} \right)$$

n = تعداد سوالات

$\sigma^2$  = واریانس کل سوالات آزمون

$P$  = نسبت افرادی که به سوال پاسخ صحیح داده اند  
 $q$  = نسبت افرادی که به سوال پاسخ غلط داده اند.

$$KR-21 = \frac{n}{n-1} \left( 1 - \frac{npq}{\sigma^2} \right)$$

در صورتی از این فرمول استفاده می کنیم که درجه دشواری سوالات آزمون شبیه به هم باشند.

$P$  = میانگین دشواری سوالات (ضریب دشواری سوالات)

ضریب دشواری = (تعداد سوالات صحیح) / (کل سوالات)

هرچقدر ضریب دشواری بالاتر باشد سوالات آسان تر است.  $q = 1 - p$

$\sigma^2$  = واریانس کل سوالات آزمون

آزمون یا ابزار هنگامی مورد پذیرش است که ضریب پایایی کادر ریچاردسن حداقل برابر با ۰/۶۴ باشد

### ج ( روش دو نیمه کردن Split half reliability)

در این روش سوالات بر اساس قرارداد محقق به دو نیمه تقسیم می شود و همبستگی هر دو نیمه بایکدیگر سنجیده می شود. منظور از دو نیمه کردن، دونیمه کردن تمام سوالات پرسشنامه نیست. منظور دونیمه کردن سوالاتی است که یک موضوع خاص را می سنجد. با به دست آوردن ضریب همبستگی بین دو نیمه پرسشنامه، پایایی بین دو نیمه به دست می آید. برای به دست آوردن پایایی کل پرسشنامه یا موضوع از فرمول اسپیرمن براون Spearman Brown استفاده می کنیم.

$$r_w = \frac{n r_p}{1 + (n-1)r_p}$$

$n$  = تعداد بخش های برابر ( که در دو نیمه کردن همواره برابر ۲ است)

$r_p$  = همبستگی بین دونیمه ها

$r_w$  = پایایی برآورد شده کل

### تحلیل عامل Factor Analysis

یکی از روش های مهم برای تلخیص، کاهش و سازماندهی داده ها روش تحلیل عاملی است. در این روش امکان درک و فهم نتایج تحقیق برای محقق بیشتر فراهم می شود. در این روش از میان مجموعه متغیرهای مورد بررسی، مجموعه هایی را در قالب عامل (Factor) تشخیص و جدا می کند. دو نوع تحلیل عاملی اکتشافی و تاییدی قابل انجام است.

تحلیل عاملی بر اساس آزمون همبستگی کار می کند. بنابراین بین متغیرها باید روابط خطی برقرار باشد. متغیرهای مورد استفاده باید کمی فاصله ای باشند. سوالات با مقیاس لیکرت بهترین متغیرها برای تحلیل عاملی هستند. توزیع متغیرها در صورت نرمال بودن جواب دقیق تری را به همراه دارد. اما توزیع نرمال الزامی نیست. در صورتی که متغیری با دیگر متغیرها همبستگی آماری معنی داری نداشته باشد و یا میزان همبستگی آن بسیار پایین باشد در هیچ یک از عوامل قرار نخواهد گرفت.

لذا در ابتدای تحلیل پژوهشگر باید به بررسی همبستگی بین متغیرها بپردازد. متغیرهایی که شرایط همبستگی را ندارند در ماتریس همبستگی شناسایی و وارد تحلیل عاملی نشوند. معمولاً همبستگی زیر ۰/۷ را وارد تحلیل عاملی نمی کنند. این مقادیر قراردادی و بستگی به نظر محقق دارد. در یکی از متون

همبستگی بالاتر از ۰/۳ و در جای دیگری همبستگی بالاتر از ۰/۵ به عنوان شرط ورود به تحلیل ذکر گردید.

حجم نمونه فاکتور مهم دیگری است که در تحلیل عاملی نیازمند توجه است. نسبت تعداد متغیرهای مورد بررسی که وارد تحلیل عاملی می شوند به تعداد نمونه های مورد مطالعه حداقل ۱۰ به ۱ باشد. به عنوان مثال اگر محقق قصد دارد ۲۰ متغیر را وارد تحلیل عاملی کند حداقل باید ۲۰۰ نمونه داشته باشد. دیدگاه دیگری حداقل حجم نمونه ۱۰۰ را کافی می داند. دیدگاه دیگری حجم نمونه را بین ۱۰۰ تا ۲۰۰ نمونه ذکر می نماید. بیشتر بودن حجم نمونه باعث دقیق شدن نتایج می شود. کفایت حجم نمونه با Kaiser-Meyer-Olkin یا KMO که برگرفته از حرف اول سه ابداع کننده این آزمون است ارزیابی می شود. مقدار بین صفر تا یک دارد. اگر این مقدار از ۰/۸ بیشتر باشد تحلیل عاملی مطلوب است. ماتریس همبستگی مناسب بین متغیرها با آزمون Bartlett انجام می شود. این دو آزمون در تصمیم گیری نرم افزار برای انجام تحلیل عاملی مهم و به عنوان پیش شرط وجود دارند. نرم افزار در صورت برقراری این پیش فرض ها تحلیل عاملی را انجام می دهد.

در تحلیل عاملی استخراج عامل ها (Factor extraction) و سپس چرخش عامل ها (Rotation) انجام می شود. بر اساس متغیرهای موجود، عامل ها تشخیص داده می شوند. ممکن است یک متغیر در بیش از یک عامل قرار گیرد. که با انجام چرخش در انواع مختلف، سعی می شود متغیرهایی که در بین بیش از یک عامل مشترک هستند به یک عامل اختصاص یابند.

### کار با نرم افزار:

فایل validityReliability را باز نمایید. این مجموعه داده ها را توصیف نمایید. در این فایل ۷ متغیر وجود دارد. پرسشنامه مذکور برای ۵۶۵ نفر تکمیل گردید. برچسب متغیرهای این فایل در زیر آمده است. جدول زیر نتایج ماتریس همبستگی بین متغیرها را نشان می دهد. مقدار معنی داری نیز با علامت ستاره مشخص شده است.

```
. pwcorr bg2cost1- bg2cost7, star (0.5)
```

	bg2cost1	bg2cost2	bg2cost3	bg2cost4	bg2cost5	bg2cost6	bg2cost7
bg2cost1	1.0000						
bg2cost2	0.5953*	1.0000					
bg2cost3	0.4550*	0.5286*	1.0000				
bg2cost4	0.4015*	0.4910*	0.6573*	1.0000			
bg2cost5	0.4424*	0.5042*	0.4480*	0.4100*	1.0000		
bg2cost6	-0.1259*	-0.0744*	-0.0300*	-0.0151	-0.0474*	1.0000	
bg2cost7	0.4550*	0.5286*	1.0000*	0.6573*	0.4480*	-0.0300*	1.0000

با دستور alpha نیز می توان مقدار ضریب پایایی را محاسبه نمود. خروجی در زیر آمده است.

```
. alpha bg2cost1- bg2cost7
```

```
Test scale = mean(unstandardized items)
Reversed item: bg2cost6
```

```
Average interitem covariance: .3762426
Number of items in the scale: 7
Scale reliability coefficient: 0.8273
```

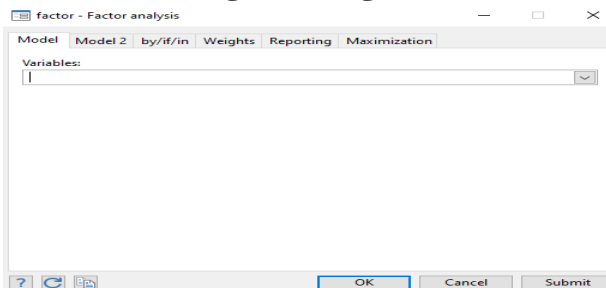
همچنین می توان با دستور alpha bg2cost1 bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6, item نیز مقدار alpha را برای هر یک از متغیرها محاسبه و به صورت زیر نشان داد.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. alpha bg2cost1 bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6, item
Test scale = mean(unstandardized items)
```

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	alpha
bg2cost1	565	+	0.7159	0.5801	.3006044	0.7074
bg2cost2	568	+	0.7697	0.6406	.2750108	0.6886
bg2cost3	568	+	0.7754	0.6211	.2578507	0.6895
bg2cost4	567	+	0.7370	0.5848	.2841436	0.7049
bg2cost5	564	+	0.6913	0.5376	.3074065	0.7199
bg2cost6	561	-	0.3109	0.0730	.4406431	0.8291
Test scale					.3110422	0.7634

تعداد مشاهدات و ضریب همبستگی و علامت آن در آزمون اول و مجدد و متوسط ضرایب همبستگی و مقدار آلفا در جدول بالا آمده است. جمع ستون آلفا نیز همانند دستور قبلی است. برای انجام تحلیل عاملی فایل factor analysis را باز نمایید. این فایل مربوط به ۱۳۶۵ دانش آموز است. متغیرهای زیادی در این فایل وجود دارد. برای تحلیل عاملی از متغیرهای ۱۳ تا ۲۴ استفاده می شود. در تحلیل عاملی ساخت ۳ فاکتور مد نظر محقق است که برای کار با نرم افزار با دستور db factor کادر محاوره ای یا جعبه گفتگو را باز و متغیرها معرفی و تعریف می شوند.



```
. factor item13 item14 item15 item16 item17 item18 item19 item20 item21 item22 item23 item24, ipf factor (3)
(obs=1,365)
```

```
Factor analysis/correlation          Number of obs =      1,365
Method: iterated principal factors   Retained factors =    3
Rotation: (unrotated)                Number of params =   33
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	5.85150	5.04464	0.8336	0.8336
Factor2	0.80687	0.44540	0.1149	0.9485
Factor3	0.36146	0.23001	0.0515	1.0000
Factor4	0.13146	0.07619	0.0187	1.0187
Factor5	0.05527	0.02362	0.0079	1.0266
Factor6	0.03164	0.02946	0.0045	1.0311
Factor7	0.00218	0.00658	0.0003	1.0314
Factor8	-0.00440	0.01466	-0.0006	1.0308
Factor9	-0.01906	0.02688	-0.0027	1.0281
Factor10	-0.04594	0.01440	-0.0065	1.0215
Factor11	-0.06035	0.03050	-0.0086	1.0129
Factor12	-0.09084	.	-0.0129	1.0000

```
LR test: independent vs. saturated: chi2(66) = 8683.10 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
item13	0.7134	-0.3987	0.0923	0.3236
item14	0.7032	-0.3391	0.0978	0.3810
item15	0.7212	-0.2450	0.1057	0.4086
item16	0.6478	-0.1890	0.1114	0.5322
item17	0.7831	-0.0734	0.0667	0.3770
item18	0.7395	0.3448	0.1129	0.3216
item19	0.6165	0.4159	0.1551	0.4228
item20	0.5501	0.2392	0.0932	0.6315
item21	0.7317	0.1168	0.0007	0.4509
item22	0.6128	0.2609	-0.0228	0.5559
item23	0.8194	-0.0262	-0.3454	0.2086
item24	0.6952	0.0183	-0.3873	0.3665

در جدول خروجی بالا ستون های جدول به شرح زیر توصیف می شوند.

Eigenvalue یا مقدار ویژه ، مقادیر واریانس هر عامل است که آن را توصیف می کنند. معمولاً عواملی که مقدار Eigenvalue بالاتر از ۱ دارند به عنوان عامل مناسب شناخته می شوند. فاکتور یک بالاترین مقدار واریانس را دارد. فاکتور دوم مرتبه بعدی واریانس را دارد. بعضی از Eigenvalue ها منفی هستند. برای اینکه در رتبه بندی کامل از چرخش استفاده نشد. ۱۲ متغیر وجود دارد که مقادیر Eigenvalue آنها خیلی کم است. Difference مقدار اختلاف بین Eigenvalue مورد نظر و Eigenvalue بعدی است.

Proportion سهم واریانسی را که هر فاکتور تبیین می کند نشان می دهد. Cumulative نسبت تجمعی سهم واریانس های فاکتور مورد نظر و فاکتورهای قبلی است. مقدار درصد واریانس از تقسیم عدد Eigenvalue هر عامل بر مجموع کل مقادیر اعداد Eigenvalue بدست می آید. اگر این درصد کمتر از ۵٪ باشد عامل مناسب نیست. این درصد نشان می دهد که واریانس های متغیرهای مورد بررسی چگونه توصیف و تبیین می شوند.

در جدول دوم خروجی نیز:

Factor Loadings برای وزن دادن به متغیرها برای هر فاکتور با توجه به همبستگی بین متغیر و فاکتور انجام می شود. برای این منظور از روش orthogonal استفاده می شود. در حقیقت ضریب همبستگی هر یک از متغیرها با هر یک از عوامل است. در صورتی که جمع جبری توان دوم مقادیر همبستگی هر متغیر در عوامل مختلف را جمع نمایید به مقداری خواهید رسید که اشتراک یا Communality آن متغیر در تمام عوامل نامیده می شود. اگر مقادیر ضرایب مذکور را به صورت عمودی در جدول بالا جمع شوند مجموع مقادیر واریانس هایی به وجود می آیند که هر عامل توصیف می کند. این مقدار همان Eigenvalue است. Uniqueness درصد واریانس مشترک هر متغیر است که با فاکتور مرتبط نیست. به عبارتی این مقدار برابر است با یک منهای اشتراکات (1 - communality).

. rotate, varimax horst

Factor analysis/correlation  
Method: iterated principal factors  
Rotation: orthogonal varimax (Kaiser on)

Number of obs = 1365  
Retained factors = 3  
Number of params = 33

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.94943	0.29428	0.4202	0.4202
Factor2	2.65516	1.23992	0.3782	0.7984
Factor3	1.41524	.	0.2016	1.0000

LR test: independent vs. saturated: chi2(66) = 8683.10 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
item13	0.7714	0.1740	0.2260	0.3236
item14	0.7256	0.2130	0.2171	0.3810
item15	0.6756	0.2950	0.2187	0.4086
item16	0.5908	0.2926	0.1820	0.5322
item17	0.5867	0.4461	0.2825	0.3770
item18	0.2865	0.7386	0.2255	0.3216
item19	0.1702	0.7281	0.1343	0.4228
item20	0.2278	0.5396	0.1594	0.6315
item21	0.4020	0.5333	0.3210	0.4509
item22	0.2178	0.5584	0.2913	0.5559
item23	0.4488	0.3769	0.6692	0.2086
item24	0.3235	0.3205	0.6528	0.3665

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.6584	0.6121	0.4381
Factor2	-0.6840	0.7294	0.0088
Factor3	0.3141	0.3055	-0.8989

در خروجی بالا که با دستور چرخش انجام شده است **Uniqueness** همانند مقدار e. (Eigenvalue) است. در این جدول نیز سه فاکتور درست گردید. در دستور زیر نیز چرخش با واریماکس و در گزینه blanks به نرم افزار گفته شده است که در نمایش فاکتورها انهایی که بیشتر از 0.3 هستند نشان داده شوند.

```
. rotate, varimax horst blanks(.3)
```

```
Factor analysis/correlation          Number of obs =      1365
Method: iterated principal factors   Retained factors =      3
Rotation: orthogonal varimax (Kaiser on) Number of params =     33
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.94943	0.29428	0.4202	0.4202
Factor2	2.65516	1.23992	0.3782	0.7984
Factor3	1.41524	.	0.2016	1.0000

LR test: independent vs. saturated:  $\chi^2(66) = 8683.10$  Prob> $\chi^2 = 0.0000$

Rotated factor loadings (pattern matrix) and unique variances

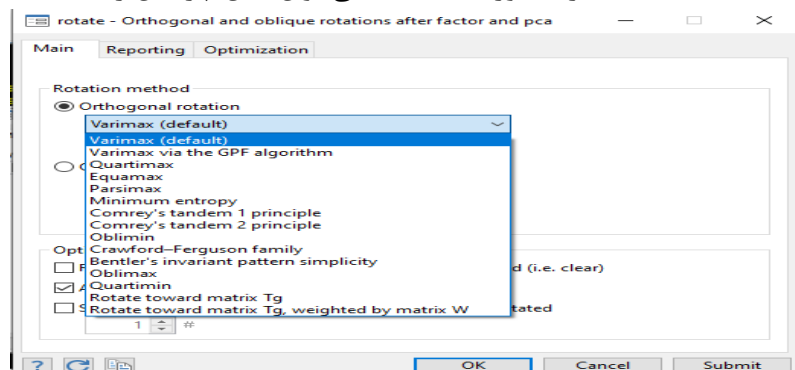
Variable	Factor1	Factor2	Factor3	Uniqueness
item13	0.7714			0.3236
item14	0.7256			0.3810
item15	0.6756			0.4086
item16	0.5908			0.5322
item17	0.5867	0.4461		0.3770
item18		0.7386		0.3216
item19		0.7281		0.4228
item20		0.5396		0.6315
item21	0.4020	0.5333	0.3210	0.4509
item22		0.5584		0.5559
item23	0.4488	0.3769	0.6692	0.2086
item24	0.3235	0.3205	0.6528	0.3665

(blanks represent  $\text{abs}(\text{loading}) < .3$ )

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.6584	0.6121	0.4381
Factor2	-0.6840	0.7294	0.0088
Factor3	0.3141	0.3055	-0.8989

در دستور بالا از واریماکس استفاده شده ولی در دستور زیر از پریماکس (promax) برای چرخش متغیرها و ساختن فاکتورها استفاده شده است. از دستور db rotate می توان مدل چرخش را انتخاب نمود.



## روش های اپیدمیولوژی و پژوهش با Stata

```
. rotate, promax horst blanks(.3)
```

```
Factor analysis/correlation          Number of obs   =   1365
Method: iterated principal factors   Retained factors =     3
Rotation: oblique promax (Kaiser on) Number of params =   33
```

Factor	Variance	Proportion	Rotated factors are correlated
Factor1	4.86265	0.6927	
Factor2	4.52052	0.6440	
Factor3	4.30842	0.6138	

```
LR test: independent vs. saturated:  chi2(66) = 8683.10 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
item13	0.8518			0.3236
item14	0.7855			0.3810
item15	0.6969			0.4086
item16	0.6044			0.5322
item17	0.5087			0.3770
item18		0.7626		0.3216
item19		0.8200		0.4228
item20		0.5541		0.6315
item21		0.4298		0.4509
item22		0.5265		0.5559
item23			0.7187	0.2086
item24			0.7502	0.3665

(blanks represent abs(loading)<.3)

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.8977	0.8593	0.8479
Factor2	-0.4157	0.4864	0.0071
Factor3	0.1462	0.1581	-0.5301

در خروجی بالا می توان هر کدام از متغیرهای ۱۳ تا ۲۴ را در یک فاکتور ملاحظه نمود. این فاکتورها هر کدام یک بعد از پرسشنامه را اندازه گیری و نشان می دهند. به عنوان مثال در ارزیابی پرسشنامه های کیفیت زندگی یا بررسی ابعاد سلامت از اینگونه تحلیل ها استفاده می شود. برای نمایش فاکتورها از روش *promax oblique rotation* استفاده شده است. این روش می گوید هر یک از متغیرها وزن هر یک از فاکتورها هستند. اما باید توجه داشت که این همبستگی میان متغیرها و فاکتورها نیست.

*promax rotation* به فاکتورها اجازه می دهد برای همبستگی بهتر، ساده و تقریبی با مدل اشباع شده. *Uniqueness* نیز مشابه مقدار *e* در بالا می باشد زیرا سه فاکتور را نمایش داده است. برای ارزیابی بعد از تحلیل و برآوردها از دستور *estat* استفاده می شود. نتیجه همبستگی میان فاکتورها در چرخش را نشان می دهد. خروجی این دستور در زیر آمده است. هر همبستگی میان فاکتورها کمتر باشد مدل تحلیل مناسب تر است.

## روش های اپیدمیولوژی و پژوهش با Stata

. estat common

Correlation matrix of the promax(3) rotated common factors

Factors	Factor1	Factor2	Factor3
Factor1	1		
Factor2	.5923	1	
Factor3	.6807	.6482	1

برای انجام تحلیل عاملی مجدد فایل قبلی یعنی validityReliability را باز نمایید و تحلیل عامل را به صورت زیر انجام و اعتبار پرسشنامه را بررسی نمایید. خروجی این تحلیل و دستورات در زیر آمده است.

. factor bg2cost1-bg2cost6, ml

(obs=568)

number of factors adjusted to 3

Iteration 0: log likelihood = -28.605724

Iteration 1: log likelihood = -.66919454

Iteration 2: log likelihood = -.00634396

Iteration 3: log likelihood = -6.465e-06

Iteration 4: log likelihood = -3.342e-12

Factor analysis/correlation

Method: maximum likelihood

Rotation: (unrotated)

Log likelihood = -3.34e-12

Number of obs = 568

Retained factors = 3

Number of params = 15

Schwarz's BIC = 95.1318

(Akaike's) AIC = 30

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	1.07659	0.31904	0.5229	0.5229
Factor2	0.75755	0.53268	0.3679	0.8908
Factor3	0.22487	.	0.1092	1.0000

LR test: independent vs. saturated: chi2(15) = 269.07 Prob>chi2 = 0.0000  
(the model with 3 factors is saturated)

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
bg2cost1	0.2315	0.3840	-0.1662	0.7713
bg2cost2	-0.4172	0.3811	-0.2541	0.6161
bg2cost3	-0.4431	0.4662	0.1354	0.5680
bg2cost4	-0.3133	0.2014	0.2578	0.7948
bg2cost5	0.5564	0.3600	0.1749	0.5302
bg2cost6	0.4948	0.2780	-0.1315	0.6606

در خروجی مذکور نیز ۳ فاکتور معرفی گردید. فاکتور سوم مقدار کمی از واریانس متغیرها را توصیف نموده است. بنابراین با دستور زیر به نرم افزار گفته می شود که با پیش فرض تهیه ۲ فاکتور تحلیل انجام شود. خروجی و نتیجه به شرح زیر است.



## روش های اپیدمیولوژی و پژوهش با Stata

```
. factor bg2cost1-bg2cost6, factor(2) ml
(obs=568)
Iteration 0: log likelihood = -28.702162
Iteration 1: log likelihood = -7.0065234
Iteration 2: log likelihood = -6.8513798
Iteration 3: log likelihood = -6.8429502
Iteration 4: log likelihood = -6.8424747
Iteration 5: log likelihood = -6.8424491
Iteration 6: log likelihood = -6.8424477
```

```
Factor analysis/correlation
Method: maximum likelihood
Rotation: (unrotated)
Log likelihood = -6.842448
Number of obs = 568
Retained factors = 2
Number of params = 11
Schwarz's BIC = 83.4482
(Akaike's) AIC = 35.6849
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	1.02766	0.28115	0.5792	0.5792
Factor2	0.74651	.	0.4208	1.0000

```
LR test: independent vs. saturated: chi2(15) = 269.07 Prob>chi2 = 0.0000
LR test: 2 factors vs. saturated: chi2(4) = 13.58 Prob>chi2 = 0.0087
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
bg2cost1	-0.1371	0.4235	0.8018
bg2cost2	0.4140	0.1994	0.7888
bg2cost3	0.6199	0.3692	0.4794
bg2cost4	0.3577	0.0909	0.8638
bg2cost5	-0.3752	0.4355	0.6695
bg2cost6	-0.4295	0.4395	0.6224

از دستور چرخش نیز استفاده می شود. این دستور در نرم افزار انجام و خروجی زیر حاصل می شود.

```
. rotate, varimax
```

```
Factor analysis/correlation
Method: maximum likelihood
Rotation: orthogonal varimax (Kaiser off)
Log likelihood = -6.842448
Number of obs = 568
Retained factors = 2
Number of params = 11
Schwarz's BIC = 83.4482
(Akaike's) AIC = 35.6849
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	0.89639	0.01863	0.5052	0.5052
Factor2	0.87777	.	0.4948	1.0000

```
LR test: independent vs. saturated: chi2(15) = 269.07 Prob>chi2 = 0.0000
LR test: 2 factors vs. saturated: chi2(4) = 13.58 Prob>chi2 = 0.0087
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
bg2cost1	0.4276	0.1239	0.8018
bg2cost2	-0.0670	0.4547	0.7888
bg2cost3	-0.0417	0.7203	0.4794
bg2cost4	-0.1253	0.3471	0.8638
bg2cost5	0.5710	-0.0666	0.6695
bg2cost6	0.6047	-0.1093	0.6224

Factor rotation matrix

	Factor1	Factor2
Factor1	-0.5606	
Factor2	0.8281	0.5606

## خلاصه و نتیجه گیری:

روان سنجی ابزارهای اندازه گیری سلامت مرحله ای بسیار مهم و حیاتی برای معتبرکردن ابزارها و جلوگیری از مخدوش شدن می باشد. ابزارهایی که ترجمه می شوند یا از ابتدا تدوین می شوند نیاز به اندازه گیری روایی و پایایی دارند. پایایی درجه ای از ثبات نتایج طی زمان و قابلیت تکرار آنها میباشد که سنجش پایایی علاوه بر روش معمول آلفای کرونباخ با روشهای دیگر نیز قابل اندازه گیری است. هنگامی که یک ابزار پایاست به این معنی است که تحت تاثیر شرایط زمانی و مکانی و یا سایر شرایط نبوده و نتایج در موقعیت های مختلف یکسان است. در حالی که روایی درجه ای از صحت نتایج می باشد و بیان کننده ی تطابق بین هدف مطالعه و آنچه مطالعه اندازه گیری می کند می باشد که ابعاد گوناگونی دارد. لذا توجه محققین و پژوهشگران و دست اندرکاران نظام سلامت به ارائه نتایج و آمار دقیق و صحیح میتواند در نتیجه استفاده از سنجش های پایا و قابل اعتماد به دست آید. سنجش هایی که علاوه بر کم کردن خطاهای تصادفی از خطای منظم نیز بتواند دوری کند. برای درک مباحث عمیق تهیه پرسشنامه، انواع آن (رودرو، تلفنی، پستی، رایانه ای، پست الکترونیک، اینترنت و مبتنی بر صفحه وب) و انواع مقیاس های پرسشنامه (لیکرت، گاتمن، ترستون) به کتاب های مرتبط مراجعه شود. اما به طور خلاصه در این قسمت به نحوه اعتبارسنجی، پایایی و تحلیل عاملی پرداخته می شود. تصمیم به اینکه قرار است یک ابزار را بومی کرده یا اینکه یک ابزار ساخته شود مراحل اعتبار سنجی متفاوت است.

منظور از اعتبار یا روایی پرسشنامه یا یک ابزار دیگر، این است که آنچه پرسشنامه یا هر ابزاری، بررسی و اندازه گیری می کند همانی باشد که قرار است سنجیده شود و برای آن تهیه شده است. اعتبار یک پرسشنامه به پایایی آن وابسته است و اگر پرسشنامه یا ابزاری پایا نباشد انجام بررسی برای اعتبار آن جایگاهی ندارد. لذا برای اعتبار قابل قبول در مطالعات پرسشنامه ای لازم است سوالات به گونه ای طراحی شوند تا علاوه بر تناسب و انطباق با موضوع مورد اندازه گیری، برداشت همه پاسخگویان از سوالات با یکدیگر یکسان بوده و منطبق با برداشت طراح پرسشنامه نیز باشند.

### اعتبار انواع مختلفی دارد.

- اعتبار ظاهری (Face Validity): این اعتبار یک قضاوت شخصی و ذهنی است. ضعیف ترین نوع اعتبار است. بستگی به شکل ظاهری پرسشنامه، توالی سوالات و جذابیت آنها دارد.
- اعتبار سازه ای (Construct Validity): نشانگر توافق بین میزان هایی است که مطالعه عملا قادر به اندازه گیری آنهاست، با آنچه ظاهرا مورد ارزیابی قرار گرفته است. یعنی پرسشنامه معتبر از نظر اعتبار سازه ای باید قادر باشد همان چیزی را اندازه گیری کند که برای آن طراحی شده است. اعتبار محتوی (Content)، اعتبار همگرایی (Convergence) و همخوانی با سایر روشهای اندازه گیری و بررسی موضوع، اعتبار تمایز (Discriminant) و اعتبار همزمان (Concurrent) و توان یک پرسشنامه در افتراق

بین افراد از جمله روش هایی هستند که اعتبار سازه ای را بررسی و اندازه گیری می کنند.

- اعتبار پیش گویی کننده (Predictive) : توانایی پیشگویی برای آنچه قرار است پیش بینی شود را ارزیابی می کند. مقایسه با یک ملاک استاندارد معمولاً انجام می شود.
- سازگاری درونی (Internal Consistency) : یعنی نتایج تمامی سوالاتی که برای اندازه گیری یک متغیر طراحی و اجرا می شوند موید یکدیگر باشند. روش های مختلفی برای ارزیابی سازگاری درونی وجود دارد. روش دو نیمه کردن سوالات (Half Split) و روش استفاده از ضریب آلفای کرونباخ (Chronbach s Alpha) دو روش مهم برای ارزیابی و اندازه گیری سازگاری درونی است.

پایایی به معنی عدم وجود خطای تصادفی در اندازه گیری ها است. به عبارت دیگر پایایی یا اعتماد یعنی توافق و همخوانی بین نتایج حاصله از دفعات متعدد انجام یک تحقیق. بدین منظور باید نتایج پژوهش از فرد انجام دهنده تحقیق (مجری)، از زمان انجام تحقیق، از مکان و محل انجام تحقیق و ابزارهای مورد استفاده مستقل باشند. به عنوان مثال ترازویی که وزن یک کودک را در یک درمانگاه در یک روستا اندازه گیری می کند به مهارت پرسنل آن درمانگاه و زمان و مکان وابسته نباشد. وزنی که این ترازو اندازه گیری می کند و نشان می دهد همان اندازه گیری و نشان دادن باشد که در یک شهر یا روستای دیگر استفاده می شود. بنابراین پایایی شرط لازم برای روایی است. در اعتبار مجموعه خطاهای تصادفی و منظم مد نظر است. در پایایی تنها خطای تصادفی مد نظر است. و لذا وقتی پایایی بیان می شود فرض بر این است که ابزار خطای منظم یا سیستماتیک و یا تورش ندارد. روش های مختلفی برای محاسبه پایایی بسته به مقیاس متغیرها وجود دارد. روش اجرای مجدد یک آزمون واحد (Test - Retest) و محاسبه ضریب همبستگی آنها یکی از روش هاست. روش دیگر محاسبه ضریب همبستگی داخل گروهی (Intra Class Correlation Coefficient- ICC) است. این روش در دفعات اندازه گیری بالای دوبار استفاده می شود. از روشهای همسانی درونی (Intral Consistency) نیز استفاده می شود. این روشها ساده تر از روش های قبلی است. ضریب آلفای کرونباخ و ضریب کودر ریچاردسون (وقتی متغیر اسمی دو حالتی وجود داشته باشد) دو روش مهم برای اندازه گیری همسانی درونی است.

با دستور **db factor** و معرفی متغیرها در آن، می توان نوع مدل انالیز را انتخاب و تحلیل عامل را انجام داد. چنانچه قصد محقق انجام تحلیل عاملی (principle components analysis) باشد با اضافه نمودن **pca** به ادامه دستور **db** می توان تحلیل را انجام داد.

به عنوان مثال برای داده های فایل بالا این کار انجام و خروجی زیر بدست می آید.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. factor bg2cost1 bg2cost2 bg2cost4 bg2cost6, pcf
(obs=568)
```

```
Factor analysis/correlation          Number of obs   =    568
Method: principal-component factors   Retained factors =     2
Rotation: (unrotated)                Number of params =     6
```

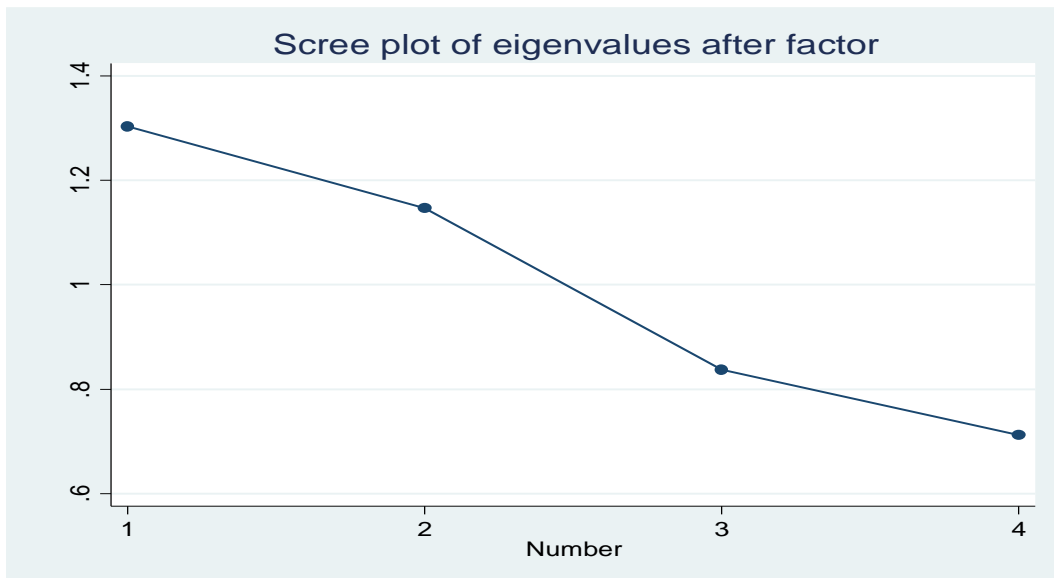
Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	1.30342	0.15675	0.3259	0.3259
Factor2	1.14668	0.30903	0.2867	0.6125
Factor3	0.83765	0.12539	0.2094	0.8219
Factor4	0.71225	.	0.1781	1.0000

```
LR test: independent vs. saturated:  chi2(6) = 64.86 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
bg2cost1	0.6157	0.5518	0.3164
bg2cost2	-0.2292	0.7663	0.3602
bg2cost4	-0.5232	0.4905	0.4856
bg2cost6	0.7733	0.1196	0.3877

با دستور `loadingplot` می توان نمودار فاکتورها را مشاهده نمود. سپس با دستور `scree` می توان نمودار `eigenvalues` را به صورت زیر نمایش داد. در این نمودار مشخص است که دو فاکتور بیشترین واریانس را تبیین می کنند.



با دستور زیر چرخش برای فاکتور دوم انجام داده می شود.

## روش های اپیدمیولوژی و پژوهش با Stata

```
. rotate, factor(2) blank(.3)
```

```
Factor analysis/correlation          Number of obs   =    568
Method: principal-component factors  Retained factors =     2
Rotation: orthogonal varimax (Kaiser off) Number of params =     6
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	1.26933	0.08856	0.3173	0.3173
Factor2	1.18077	.	0.2952	0.6125

```
LR test: independent vs. saturated: chi2(6) = 64.86 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
bg2cost1	0.8020		0.3164
bg2cost2		0.7848	0.3602
bg2cost4		0.6779	0.4856
bg2cost6	0.7398		0.3877

(blanks represent abs(loading)<.3)

Factor rotation matrix

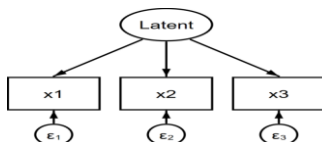
	Factor1	Factor2
Factor1	0.8846	-0.4664
Factor2	0.4664	0.8846

در پایان نیز در صورتیکه محقق بخواهد امتیاز فاکتورهای پیش بینی کننده را محاسبه و نمایش دهد از دستور **predict** استفاده می نماید. در دستور بالا سه متغیر مبتنی بر بارگذاری فاکتورها ساخته شده است. نرم افزار از روش رگرسیونی برای پیش بینی این متغیرها استفاده می کند. نتیجه روش **pca** دقیق تر است. با استفاده از این فرایند تحلیل، عوامل انتخاب شده از هم تفکیک و متغیرها به یک عامل اختصاص می یابند. بنابراین با تحلیل عاملی ارتباطات و روابط احتمالی بین تعداد زیادی متغیر شناسایی و کشف می شود. کاربرد این نوع تحلیل در طراحی ابزارهای سنجش، طراحی نظریه ها، تلخیص داده ها و طراحی و آنالیز پرسشنامه است.

### تمرین

- ۱- روش های اندازه گیری روایی با نرم افزار STATA را برای پرسشنامه سلامت عمومی (GHQ) با وارد کردن داده های فرضی برای ۳۰ نفر را انجام و شاخص ها را گزارش نمایید.
- ۲- روش های اندازه گیری پایایی با نرم افزار STATA را برای پرسشنامه سلامت عمومی (GHQ) با وارد کردن داده های فرضی برای ۳۰ نفر را انجام و شاخص ها را گزارش نمایید.
- ۳- فایل داده های ساخته شده را با تحلیل عاملی آنالیز و نتایج را گزارش نمایید.

از دیدگاه علم اپیدمیولوژی، در برخی موارد گروه های ناهمگنی در جمعیت وجود دارد و افراد این گروه ها رفتار متفاوتی دارند. اما متغیری نداریم که گروه ها را مشخص کند. این گروه ها ممکن است مصرف کنندگان با ترجیحات خرید متفاوت، نوجوانان با الگوهای رفتاری متفاوت یا طبقه بندی وضعیت سلامتی باشند. LCA به ما امکان می دهد این گروه های مشاهده نشده را شناسایی و درک کنیم. به ما این امکان را می دهد که بدانیم چه کسانی در یک گروه هستند و تفاوت ویژگی های آن گروه با سایر گروه ها چگونه است. در مدل های کلاس پنهان، از یک متغیر پنهان استفاده می کنیم که برای نمایش گروه ها به صورت طبقه بندی شده است و از گروه ها به عنوان کلاس ها نام می بریم. مدل های کلاس نهفته شامل دو قسمت است. یکی با احتمالاتی که مربوط به کدام طبقه است متناسب است و دیگری رابطه بین کلاسها و متغیرهای مشاهده شده را توصیف می کند. دیاگرام زیر را مشاهده کنید.



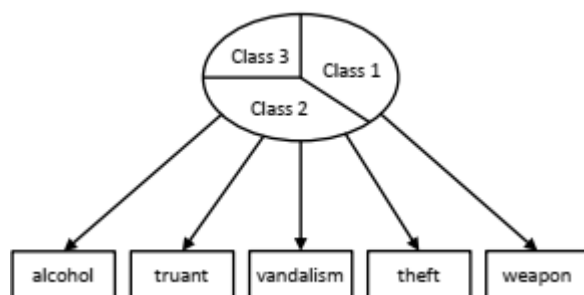
متغیرهای پنهان، متغیرهای تصادفی هستند که به صورت مستقیم قابل مشاهده یا قابل اندازه گیری نبوده و آنها را میتوان به صورت غیرمستقیم بر مبنای دیگر کمیت های مشاهده شده اندازه گیری کرد. مدل هایی که امکان بررسی رابطه ی بین متغیرهای پنهان و متغیرهای مشاهده شده را فراهم می نمایند، مدل های متغیر پنهان نامیده میشوند. بعضی از مفاهیم مرتبط با سلامتی و بیماری و اختلالات موجود در جامعه مانند ناتوانی، استرس، افسردگی، اضطراب، اسکیزوفرنی و سایر پدیده های ذهنی، به طور واضح و مستقیم قابل مشاهده نیستند و برای تفسیر و استنباط این مفاهیم از روش مدل کلاس پنهان و رگرسیون کلاس پنهان استفاده میشود.

حجم نمونه ی مورد نیاز در مدل های تحلیل رگرسیون کلاس پنهان، به تعداد کلاس هایی که از داده های مطالعه استخراج می گردد و تعداد متغیرهای پیش بینی کننده در مدل بستگی دارد. غالباً حجم نمونه بیش از ۵۰۰ نفر در مطالعات از توان کافی برای برآزش مدل برخوردار است.

در هر کلاس پنهان افرادی که بسیار شبیه به یکدیگر هستند و بطور منحصر به فردی از سایر کلاسها در سراسر مجموعه ای از متغیرها متفاوتند، بررسی شوند. به عنوان مثال برای ارزیابی شدت اختلالات خفیف روانی به عنوان یک متغیر پنهان، می توان از آنالیز کلاس پنهان استفاده کرد. تحلیل کلاس پنهان روش آماری پیشرفته ای است که ارتباط بین متغیرهای مشاهده شده گسسته و یک متغیر پنهان گسسته را مدل بندی می کند.

افراد بر طبق الگوی ریاضی پاسخ به سوالات (متغیر مشاهده شده گسسته)، در یک کلاس (متغیر پنهان) واقع میشوند، به گونهای که افراد درون یک کلاس از لحاظ پاسخگویی به سوالات مشابه، با افراد واقع در کلاس دیگر حداکثر تفاوت را دارند. مدل رگرسیون کلاس پنهان، تعمیم یافته تحلیل کلاس پنهان است که امکان ارزیابی تأثیر متغیرهای مختلف را بر متغیرهای پنهان فراهم

می کند. مدل رگرسیون کلاس پنهان شبیه مدل رگرسیون لوجستیک است، با این تفاوت که در رگرسیون لوجستیک متغیر پاسخ یک متغیر گسسته و دو حالت است که به طور مستقیم قابل مشاهده و اندازه گیری است، در حالی که در رگرسیون کلاس پنهان متغیر پاسخ یک متغیر گسسته است که به طور مستقیم قابل مشاهده یا قابل اندازه گیری نیست، بلکه براساس نشانگرهای مرتبط که به طور مستقیم قابل مشاهده یا اندازه گیری هستند، مورد ارزیابی قرار می گیرد. در یک مثال ما مجموعه ای از متغیرهای مشاهده شده داریم که نشان می دهد نوجوانان مشروبات الکلی (الکل) مصرف کرده اند، بیش از ۱۰ غیبت غیر موجه از مدرسه (فرار از خانه)، استفاده از سلاح در نبرد (سلاح)، خرابکاری (بزهکاری) و دارای اشیای سرقت شده به ارزش بیش از ۲۵ دلار (سرقت) بوده اند. ما از این موارد برای متناسب کردن یک مدل کلاس نهفته با ۳ کلاس رفتاری مشاهده نشده استفاده خواهیم کرد. دیاگرام این مثال را در زیر ببینید.



### کار با نرم افزار و تحلیل کلاس پنهان

فایل داده های `gsem_lca1` را باز نمائید. متغیرها پاسخ های دانشجویانی از دانشگاه های هاروارد و رادکلیف را گزارش و نشان می دهند که از آنها خواسته شد به چهار موقعیت چگونه واکنش نشان دهند. پاسخ دهندگان یا یک پاسخ خاص (بر اساس تعهدات نسبت به یک دوست) یا یک پاسخ جهانی (بر اساس تعهد به جامعه) را انتخاب کردند. هر متغیر با مقدار ۰ نشان دهنده یک پاسخ خاص و مقدار ۱ نشان دهنده یک پاسخ جهانی است.

سوال این است که چگونه گروه های افراد در واکنش به موقعیت هایی که نیاز به تصمیم گیری بین کمک به دوست (انتخاب خاص) و انجام آنچه برای جامعه مناسب است (انتخاب جهانی)، متفاوت و پاسخ می دهند؟ از افراد سوال شد که در چنین شرایطی چگونه واکنش نشان می دهند و پاسخ آنها در متغیرهای صحنه تصادف، هنگام بازی، بیمه و سهام ثبت شد. این متغیرها کدگذاری می شوند. به طوری که ۱ پاسخ جهان شمول و ۰ یک پاسخ خاص است. برای تناسب با مدل کلاس نهفته، باید تعداد کلاس ها را در متغیر پنهان مشخص کنیم. در شکل اصلی مدل کلاس پنهان نشان داده شده در اینجا، ما یک متغیر پنهان طبقه ای با دو کلاس پارامترهای موجود در مدل، یعنی رهگیری در مدل های رگرسیون لجستیک برای چهار متغیر مشاهده شده داریم. دستور زیر را اجرا کنید:

```
. gsem (accident play insurance stock <- ), logit lclass(C 2)
```





## روش های اپیدمیولوژی و پژوهش با Stata

```
. . estat lcmean
```

Latent class marginal means

Number of obs = 216

	Delta-method			
	Margin	std. err.	[95% conf. interval]	
1				
accident	.7135879	.0403588	.6285126	.7858194
play	.3296193	.0496984	.2403573	.4331299
insurance	.3540164	.0485528	.2655049	.4538042
stock	.1323726	.0383331	.0734875	.2268872
2				
accident	.9931933	.0253243	.0863544	.9999956
play	.9397644	.0659957	.6135685	.9935191
insurance	.9265309	.0656538	.6557086	.9881667
stock	.769132	.0952072	.5380601	.9050206

بخش اول جدول خروجی بالا، احتمالات برای کلاس یا طبقه ۱ را گزارش می کند. در این کلاس، احتمال پاسخگویی جهانی به اولین سوال که مربوط به شهادت در برابر دوستی که تصادف کرده بود ۰/۷۱۳ و احتمال دادن یک پاسخ جهانی در آخرین سوال درباره هشدار به یک دوست در مورد قیمت سهام در حال سقوط برابر ۰/۱۳۲ بود. بخش دوم جدول احتمالات مربوط به کلاس ۲ را گزارش می کند. ما در می یابیم که احتمال پاسخگویی جهانی به هر سوال در کلاس ۲ بیشتر از کلاس ۱ است. به نظر می رسد که کلاس ۲ نمایانگر گروهی متمایل به جهان شمول بیشتر است.

ما می توانیم با استفاده از دستور estat lcprob احتمال حضور در هر کلاس را برآورد کنیم

```
. estat lcprob
```

Latent class marginal probabilities

Number of obs = 216

C	Delta-method			
	Margin	std. err.	[95% conf. interval]	
1	.7207539	.0580926	.5944743	.8196407
2	.2792461	.0580926	.1803593	.4055257

این نشان می دهد که انتظار می رود ۷۲٪ از افراد در طبقه کمتر جهان شمول و ۲۸٪ در طبقه جهانی تر باشند. ما می توانیم از پیش بینی های احتمال عضویت در کلاس، برای ارزیابی احتمال حضور هر فرد در هر کلاس استفاده کنیم.

دستورات زیر را اجرا کنید.

```
. predict classpost*, classposteriorpr
. list in 1, abbrev(10)
```

```
. predict classpost*, classposteriorpr
```

```
. . list in 1, abbrev(10)
```

	accident	play	insurance	stock	classpost1	classpost2
1.	0	0	0	0	.999975	.000025

برای اولین فرد در مجموعه داده، که با پاسخ خاص به هر چهار سوال پاسخ داد، احتمال حضور در کلاس ۱ بیشتر از کلاس ۲ است. ما می توانیم کلاس مورد انتظار برای هر فرد را بر اساس اینکه احتمال بعدی بیشتر از ۰.۵ باشد را تعیین کنیم.

دستورات زیر را اجرا کنید:

```
. generate expclass = 1 + (classpost2>0.5)
. tabulate expclass
```

```
. . generate expclass = 1 + (classpost2>0.5)
```

```
. . tabulate expclass
```

expclass	Freq.	Percent	Cum.
1	145	67.13	67.13
2	71	32.87	100.00
Total	216	100.00	

در مجموعه داده های ما، ۱۴۵ نفر در کلاس ۱ و ۷۱ نفر در کلاس ۲ قرار دارند. برای مدل های استاندارد کلاس پنهان با متغیرهای مشاهده شده که همه طبقه بندی شده اند، یکی از راه های ارزیابی برازش مدل، مقایسه مدلی است که به تازگی اجرا شده با مدل اشباع شده است. می توانیم از estat استفاده کنیم. دستور lcgof برای انجام یک نسبت احتمال به اینکه آیا مدل ما با مدل اشباع شده مطابقت دارد یا خیر بکار می رود.

```
. estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(6)	2.720	model vs. saturated
p > chi2	0.843	
Information criteria		
AIC	1026.935	Akaike's information criterion
BIC	1057.313	Bayesian information criterion

با دستور estimates store twoclass مدل دو کلاسه ذخیره می شود. با دستور quietly gsem (accident play insurance stock <- ), logit lclass(C 1) مدل با کلاس اجرا می شود. با دستور estimates store oneclass مدل ۱ کلاسی ذخیره می شود. اکنون دو مدل را می توانید با دستور estat lcgof مقایسه کنید.

```
. . estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(11)	81.084	model vs. saturated
p > chi2	0.000	
Information criteria		
AIC	1095.300	Akaike's information criterion
BIC	1108.801	Bayesian information criterion

```
. estimates store twoclass
```

```
. quietly gsem (accident play insurance stock <- ), logit lclass(C 1)
```

```
. . estimates store oneclass
```

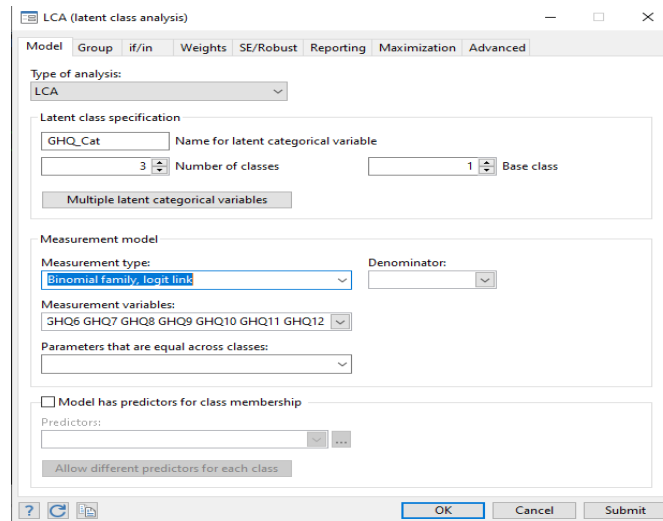
```
. . estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(11)	81.084	model vs. saturated
p > chi2	0.000	
Information criteria		
AIC	1095.300	Akaike's information criterion
BIC	1108.801	Bayesian information criterion

### مثال سوم

در این مثال، شیوع اختلالات روانی که بر مبنای پاسخ به پرسش های پرسشنامه استاندارد اندازه گیری سلامت عمومی (GHQ-12) ارزیابی گردید به عنوان متغیر پاسخ پنهان مورد استفاده قرار گرفت. تعداد کلاسها با استفاده از معیار آماری (Bayesian Criterion Information (BIC) و قابلیت تفسیرپذیری تعیین می گردد. در پرسشنامه مذکور، نقطه برش بالینی، حساسیت و ویژگی حاصل از این پرسشنامه به عوامل فرهنگی و اجتماعی بستگی دارد و در نتیجه از جامعه ای به جامعه دیگر متفاوت خواهد بود. از طرفی در بررسی شدت اختلالات خفیف روانی در جوامع مختلف به دلیل یکتا نبودن شیوه نمره گذاری و نقطه برش متفاوت که تفاسیر متفاوتی در بردارد، استفاده از روشهای مرسوم با محدودیت و نقص هایی همراه است. مدل رگرسیون کلاس پنهان بدون نیاز به نقطه برش و نحوه امتیازدهی پرسشنامه، یکی از روش آماری پیشرفته است که توانایی از بین بردن معایب یاد شده را دارد. نحوه نمره دهی سؤالات پرسشنامه سلامت عمومی ۱۲ سوالی براساس طیف لیکرت (۳-۲-۱-۰) است. کسب نمره بالاتر از این پرسشنامه نشانگر فقدان اختلالات خفیف روانی در فرد است. بنابراین در این فایل تعداد کلاس را ابتدا ۳ کلاس فرض می کنیم. افراد دارای اختلال شدید، متوسط و خفیف یا بدون اختلال. دستور زیر را اجرا کنید. از منوی `statistics` گزینه LCA را انتخاب و متغیرها و نوع مدل را انتخاب و اجرا کنید.

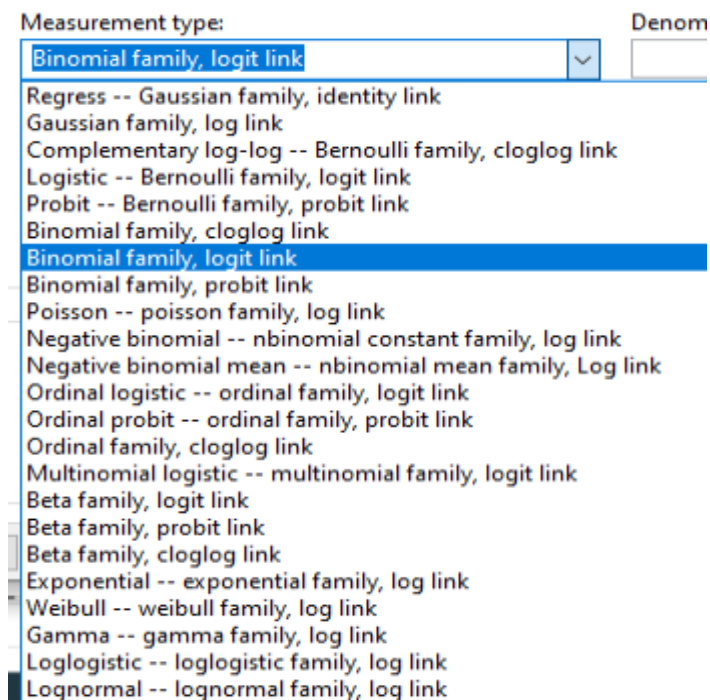
## روش های اپیدمیولوژی و پژوهش با Stata



دستور زیر را اجرا کنید.

```
gsem (GHQ1 GHQ2 GHQ3 GHQ4 GHQ5 GHQ6 GHQ7 GHQ8 GHQ9 GHQ10 GHQ11 GHQ12<-),  
logit lclass(C 3)
```

برای برآزش نوع مدل مناسب می توانید با انتخاب یکی از مدل های زیر در جعبه محاوره ای، نوع مدل مناسب را برآزش دهید.

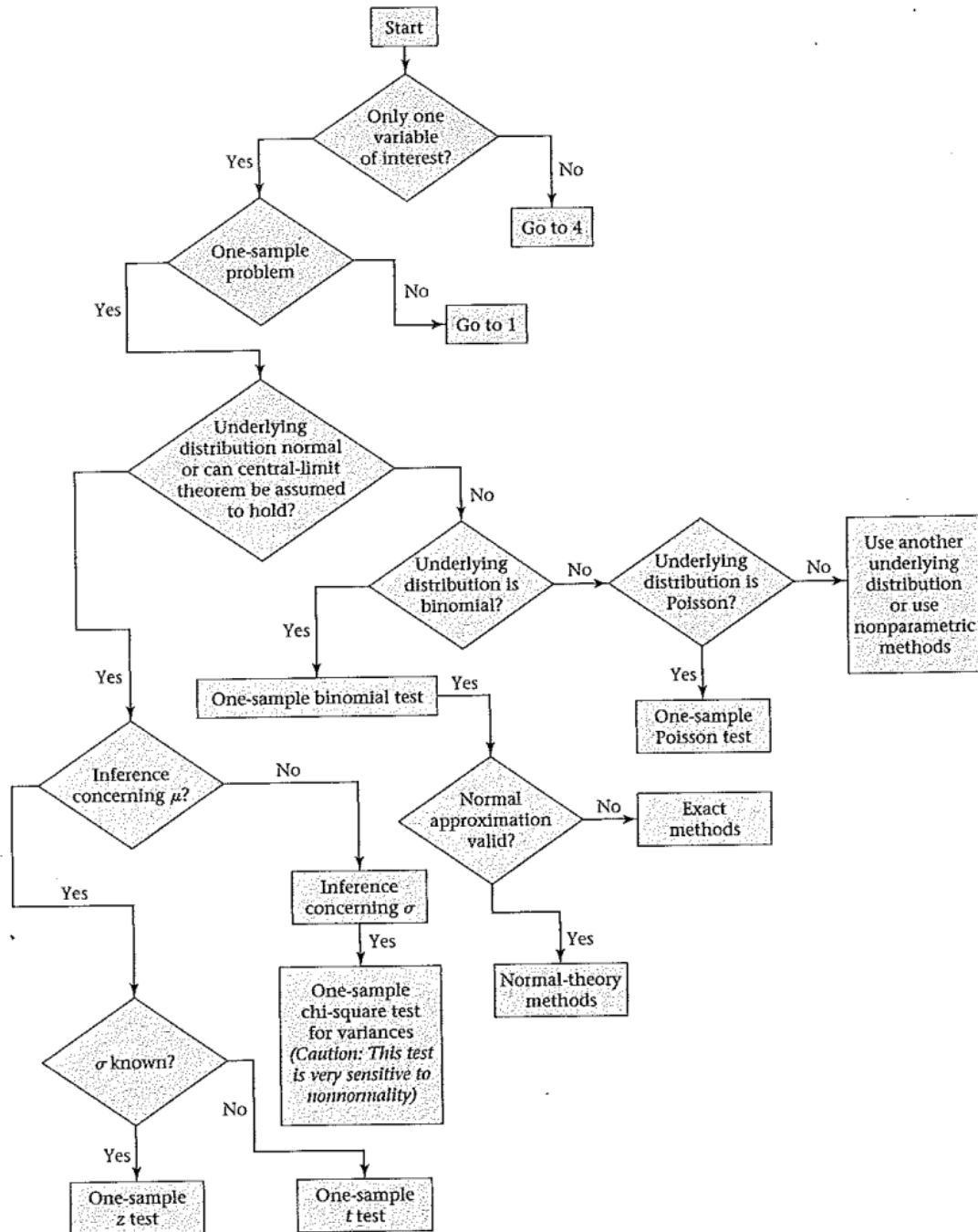


## پیوست ها :

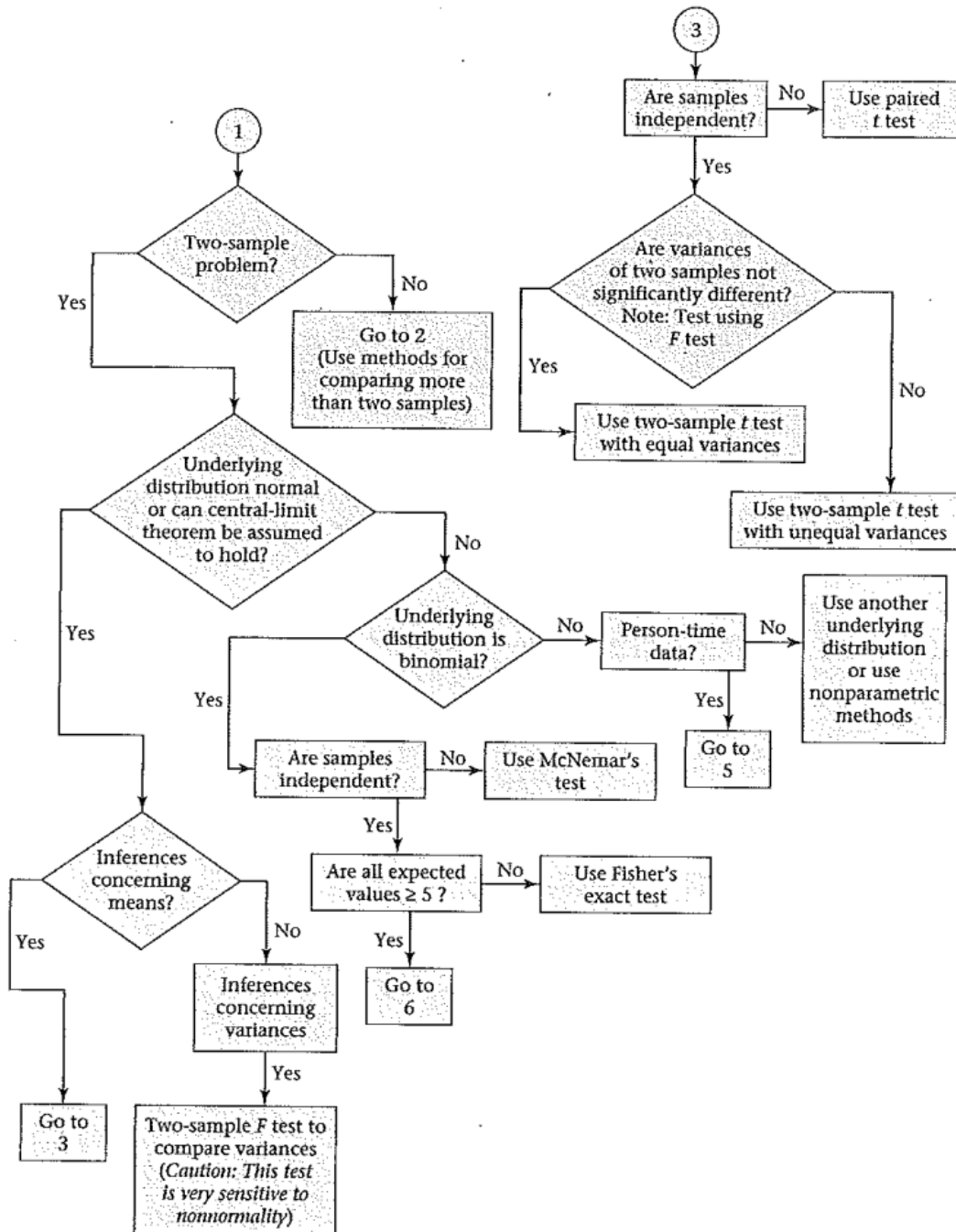
## پیوست ۱ - جدول داده های (.dta) مورد استفاده برای کار با نرم افزار در فصل های مختلف کتاب

ردیف	نام مجموعه داده	توضیحات
۱	Partone	این داده ها در فصل اول کاربرد دارند.
۲	Parttwo	این داده ها در فصل دوم کاربرد دارند.
۳	Partthree	این داده ها در فصل سوم کاربرد دارند.
۴	Partfour	این داده ها در فصل چهارم کاربرد دارند.
۵	Partfive	این داده ها در فصل پنجم کاربرد دارند.
۶	Part6regression	این داده ها در فصل ششم کاربرد دارند.
۷	Lifetable	این داده ها برای کار با جدول عمر کاربرد دارند.
۸	Manova2	این داده ها برای آنالیز واریانس چندمتغیره کاربرد دارند.
۹	Manovapartfive	این داده ها برای آنالیز واریانس چندمتغیره کاربرد دارند.
۱۰	Coxregression	این داده ها برای رگرسیون کاکس کاربرد دارند.
۱۱	Coxregression2	این داده ها برای رگرسیون کاکس کاربرد دارند.
۱۲	Logistic	این داده ها برای رگرسیون لجستیک کاربرد دارند.
۱۳	Matcheddata	این داده های برای تحلیل همسان شده کاربرد دارند.
۱۴	Partexam	این داده ها برای ارزیابی و تمرین کاربرد دارند.
۱۵	Poissonregression	این داده ها برای رگرسیون پواسون کاربرد دارند.
۱۶	Poissonregressionrate	این داده ها برای رگرسیون پواسون کاربرد دارند.
۱۷	Polytomuslogistic	این داده ها برای رگرسیون چندوجهی کاربرد دارند.
۱۸	Repeatedmeasure	این داده ها برای تحلیل داده های تکراری کاربرد دارند.
۱۹	Tb	این داده ها برای تمرین GEE کاربرد دارند.
۲۰	Timeseries	این داده ها برای سری زمانی کاربرد دارند.
۲۱	Timeseriesskindisease	این داده ها برای سری زمانی کاربرد دارند.
۲۲	Casecontrol	این داده ها برای تحلیل مورد شاهد کاربرد دارند.
۲۳	Cohort1	این داده ها برای تحلیل کوهورت کاربرد دارند.
۲۴	GEE	داده ها برای تحلیل داده های بهم وابسته کاربرد دارند.
۲۵	Epidemiology	داده ها برای تحلیل مطالعات اپیدمیولوژیک کاربرد دارند.
۲۶	Meta1-7	داده ها در قالب ۷ فایل برای متاآنالیز کاربرد دارند.
۲۷	factoranalysis	داده ها برای تحلیل عاملی بکار می روند.
۲۸	Validity reliability	داده ها برای ارزیابی روایی و پایایی کاربرد دارند.
۲۹	comparisonROC	داده ها برای مقایسه منحنی ROC کاربرد دارند.
۳۰	Inequality	داده ها برای محاسبه شاخصهای نابرابری کاربرد دارند.
۳۱	Multilevel 1 , 2	داده ها برای تحلیل چند سطحی بکار می رود.
۳۲	فایل اکسل و سایر فایل ها	فراخوانی در نرم افزار
۳۳	فایل do-file	ذخیره دستورات و اجرا

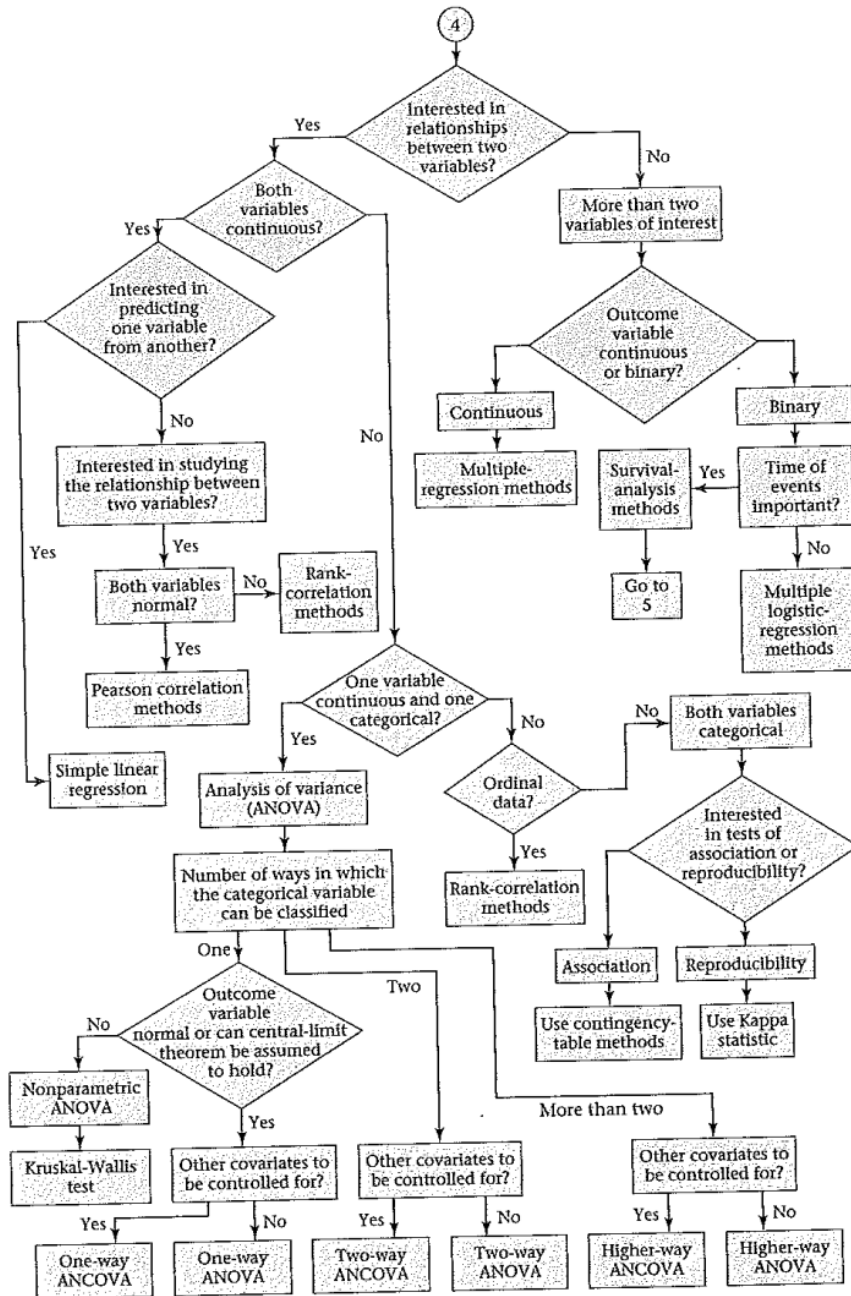
## پیوست ۲- فلوجارت های استنباط آماری (۱-۶):



پیوست ۲- ادامه فلوجارت های استنباط آماری :

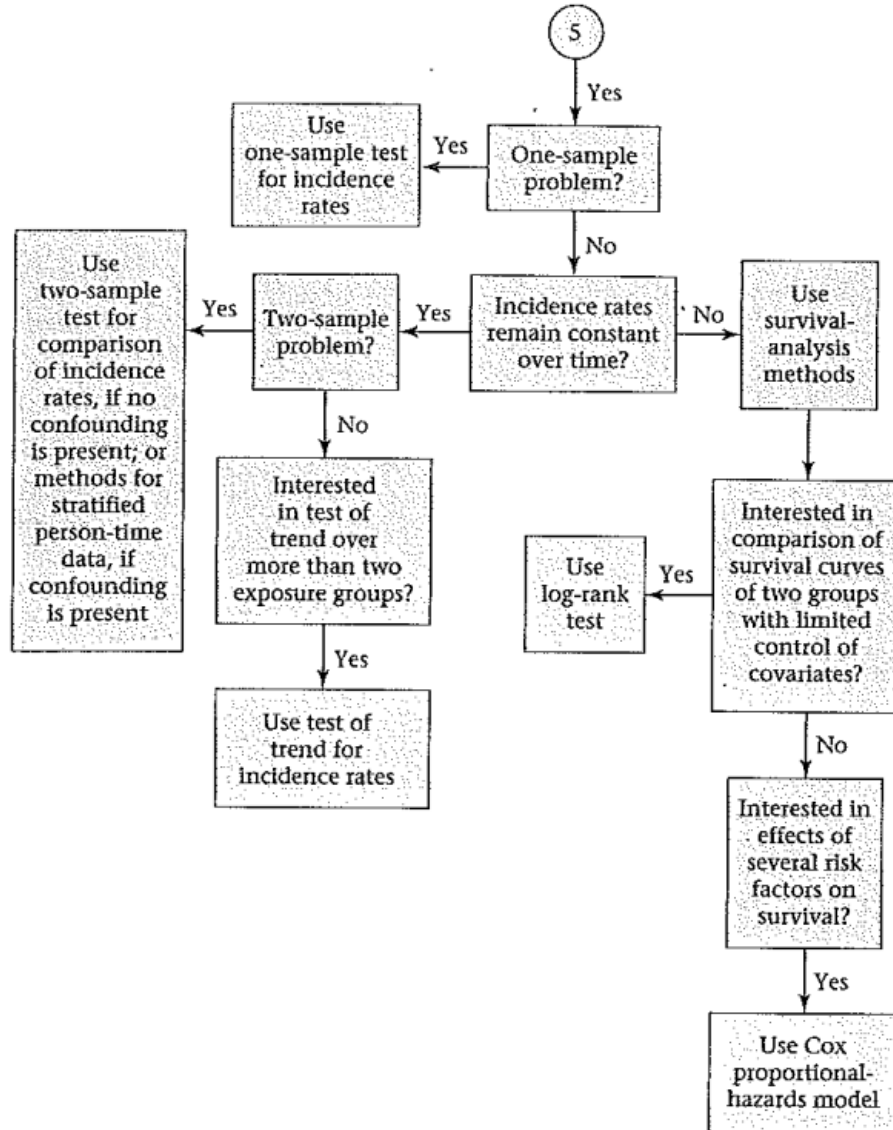


پیوست ۲- ادامه فلوجارت های استنباط آماری :

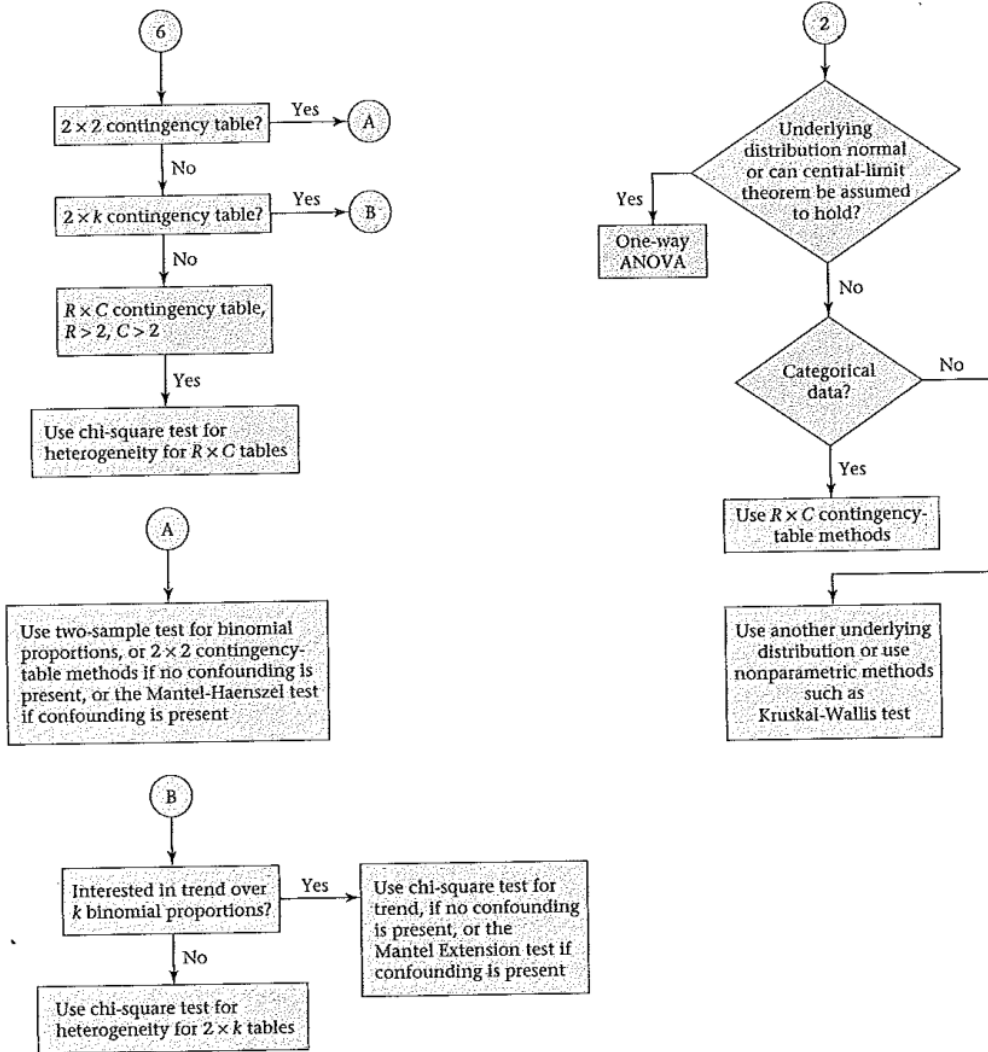




## پیوست ۲- ادامه فلوجارت های استنباط آماری :



## پیوست ۲- ادامه فلوجارت های استنباط آماری :



### پیوست ۳: فرمول های مهم در مطالعات اپیدمیولوژیک

۱- شیوع نقطه ای

$$\text{Point prevalence} = \frac{\text{No. of existing cases in a defined population at one point in time}}{\text{No. of people in the defined population at the same point in time}}$$

۲- خطر

$$\text{Risk} = \frac{\text{No. of new cases of disease arising in a defined population over a given period of time}}{\text{No. of disease free people in that population at the beginning of that time period}}$$

۳- شانس بیماری

$$\text{Odds of disease} = \frac{\text{No. of new cases of disease arising in a defined population over a given period of time}}{\text{No. of people in that population who remain disease-free during that period}}$$

۴- میزان بروز

$$\text{Incidence rate} = \frac{\text{No. of new cases of disease arising in a defined population over a given time period}}{\text{Total person-time at risk during that period}}$$

۵- رابطه خطر و میزان بروز

$$\text{Risk} = 1 - \exp(-\text{incidence rate} \times \text{duration of the period at risk})$$

۶- رابطه خطر و دوره بیماری

$$\text{Risk} = \text{incidence rate} \times \text{duration of the period at risk.}$$

۷- میزان بروز خام در یکصد هزار شخص - سال

$$\text{Crude incidence rate per 100 000 pyrs} = \frac{\text{No. of new cases arising in a defined population in a specific period of time}}{\text{Total person - years at risk in that population during that period of time}} \times 100\ 000$$

۸- میزان بروز اختصاصی - سنی در یکصد هزار نفر شخص - سال

$$\text{Age-specific incidence rate per 100 000 pyrs} = \frac{\text{No. of new cases arising in a certain age-group in a defined population and in a specific period of time}}{\text{Person - years at risk in that age group in the same population and during that period of time}} \times 100\ 000$$

۹- درصد بروز نسبی یا تناسب ( سهم ) بروز

$$\text{Proportional incidence (\%)} = \frac{\text{No. of cases of the disease of interest in a specified time period}}{\text{Total number of cases of all conditions in the same time period}} \times 100$$

۱۰- نسبت بروز نسبی

$$\text{PIR (\%)} = \frac{\text{Proportion of cases from a specific cause in population A}}{\text{Proportion of cases from the same cause in population B}} \times 100$$

۱۱- نسبت خطر

$$\text{Risk ratio} = \frac{\text{Risk in the exposed group}}{\text{Risk in the unexposed group}}$$

۱۲- نسبت میزان

$$\text{Rate ratio} = \frac{\text{Incidence rate in the exposed group}}{\text{Incidence rate in the unexposed group}}$$

۱۳- نسبت شانس یا نسبت بخت بیماری

$$\text{Odds ratio (of disease)} = \frac{\text{Odds of disease in the exposed group}}{\text{Odds of disease in the unexposed group}}$$

۱۴- فرمول های ۱۱، ۱۲ و ۱۳ اندازه های خطر نسبی (Relative Risk) نامیده می شوند.

۱۵- اختلاف خطر

$$\text{Risk difference} = \text{risk in the exposed} - \text{risk in the unexposed}$$

۱۶- اختلاف میزان

$$\text{Rate difference} = \text{rate in the exposed} - \text{rate in the unexposed}$$

۱۷- کسر افزوده یا خطر اضافه

$$\text{Excess fraction (\%)} = 100 \times (\text{excess risk/risk (or rate) in the exposed})$$

۱۸- کاهش خطر

$$\text{Risk reduction} = \text{risk (or rate) in the unexposed} - \text{risk (or rate) in the exposed}$$

۱۹- رابطه بروز، شیوع و طول دوره بیماری

$$\frac{\text{Point Prevalence}}{1 - \text{Point Prevalence}} = \text{Incidence} \times \text{duration}$$

۲۰- کسر پیشگیری شده

$$\text{Prevented fraction (\%)} = 100 \times (\text{risk reduction/risk (or rate) in the unexposed})$$

۲۱- فرمول ۱۸ و ۱۹ اندازه های مطلق اثر ، وقتی خطر مواجهه برای ایجاد بیماری نسبت به مواجهه نیافتگان کم باشد در نظر گرفته می شوند.

۲۲- خطای استاندارد نسبت (p) در جامعه و نمونه از یکی از سه فرمول زیر بدست می آید. n حجم نمونه است. P می تواند شیوع یا خطر باشد.

$$SE(p) = \sqrt{\frac{p^2(1-p)}{a}} \quad SE(p) = \sqrt{\frac{p(1-p)}{n}} \quad SE(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

where a is the numerator of the proportion  $p = a/n$

۲۳- حدود اطمینان ۹۹ و ۹۰٪ برای نسبت. در ۹۵% CI بجای 1.64 باید 1.96 گذاشته شود.

$$p \pm 2.58 \times \sqrt{\frac{p^2(1-p)}{a}} \quad p \pm 1.64 \times \sqrt{\frac{p^2(1-p)}{a}}$$

۲۴- حدود اطمینان ۹۵٪ برای میانگین نمونه

$$\text{Sample mean} \pm 1.96 \times \text{SE of the mean}$$

$$\text{SE (mean)} = \text{SD}/\sqrt{n}$$

۲۵- فرمول کلی محاسبه حدود اطمینان ۹۵٪ برای برآوردهای حاصل از نمونه

$$\text{Sample estimate} \pm 1.96 \times \text{SE(sample estimate)}$$

۲۶- حدود اطمینان برای اختلاف میزان

$$\text{Rate difference} \pm 1.96 \times \text{SE(rate difference)}$$

-۲۷

تفسیر P-Value بر اساس نظر کاکس و اسنیل در سال ۱۹۸۱

If  $P > 0.1$ , the results are reasonably consistent with the null hypothesis.

If  $P \approx 0.05$ , there is moderate evidence against the null hypothesis.

If  $P \leq 0.01$ , there is strong evidence against the null hypothesis.

چنانچه مقدار P مساوی یا کمتر از 0.01 باشد مستند قوی برای رد فرض صفر است. مقادیر P کمتر از 0.05 به طور معمول، معنی دار در نظر گرفته می شوند.

هر چند این مقادیر بسته به سطح اطمینان و نظر محقق قراردادی هستند.

-۲۸

فرمول محاسبه میزان (rate) و حدود اطمینان ۹۵٪ برای آن، a تعداد موارد بیماری

یا cases می باشد. Y شخص زمان در معرض خطر است.

$$r \pm 1.96 \times SE(r) \quad SE(r) = r/\sqrt{a} \quad r = a/y$$

-۲۹ فرمول محاسبه نسبت خطر یا نسبت شیوع (R) و حدود اطمینان آنها با استفاده از نمادهای

جدول زیر:

$$\ln R = \ln (p_1/p_0) \quad R = p_1/p_0$$

		مواجهه		جمع
		دارد	ندارد	
پیامد	دارد	a	b	n1
	ندارد	c	d	n0
جمع		m1	m0	N

$$(\ln R) \pm 1.96 SE(\ln R)$$

$$SE(\ln R) = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{m_1} - \frac{1}{m_0}}$$

-۳۰ خطای معیار اختلاف خطر یا اختلاف نسبت ها یا اختلاف شیوع و حدود اطمینان ۹۵٪

$$SE(p_1 - p_0) = \sqrt{\frac{p_1^2(1-p_1)}{a} + \frac{p_0^2(1-p_0)}{b}}$$

$$95\% \text{ CI } (p_1 - p_0) = (p_1 - p_0) \pm 1.96 SE(p_1 - p_0)$$

۳۱- حدود اطمینان برای نسبت میزان با استفاده از نمادهای جدول زیر :

	مواجهه		جمع
	دارد	ندارد	
موارد	a	b	n1
شخص - زمان در معرض خطر	y1	y0	y
میزان در ۱۰۰۰ شخص - زمان	r1	r0	r

$$\ln RR = \ln (r_1/r_0) \quad SE (\ln RR) = \sqrt{(1/a + 1/b)}$$

$$95\% \text{ CI } (\ln RR) = \ln RR \pm 1.96 SE (\ln RR)$$

سپس با آنتی لگاریتم گرفتن از مقادیر بدست آمده یا Exponential آنها حدود اطمینان بدست می آید.

۳۲- خطای معیار برای اختلاف دو میزان و حدود اطمینان ۹۵٪ برای آن

$$SE (r_1 - r_0) = \sqrt{\frac{r_1^2}{a} + \frac{r_0^2}{b}}$$

$$95\% \text{ CI } (r_1 - r_0) = (r_1 - r_0) \pm 1.96 SE(r_1 - r_0)$$

۳۳- محاسبه نسبت شانس (بخت) و حدود اطمینان آن بر اساس جدول زیر

		مواجهه		جمع
		دارد	ندارد	
پیامد	موردها	a	b	n1
	شاهدها	c	d	n0
جمع		m1	m0	N

$$OR = \frac{\text{odds of exposure among the cases}}{\text{odds of exposure among the controls}}$$

$$SE (\ln OR) = \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

$$95\% \text{ CI } (\ln OR) = \ln OR \pm SE (\ln OR)$$

سپس با آنتی لگاریتم گرفتن از مقادیر بدست آمده یا Exponential آنها حدود اطمینان بدست می آید.

۳۴- فرمول دیگر برای محاسبه میزان در زیر آمده است.

$$\text{Rate} = \frac{\text{No. events (x)}}{\frac{\text{average population (n)}}{\text{time (t)}}} = \frac{x}{n \times t} = \text{Density}$$

۳۵- فرمول محاسبه میزان سالیانه

$$\text{Yearly Rate} = \frac{\frac{x}{n}}{t} = \frac{x}{n \times t} = \frac{x}{\text{person years}}$$

۳۶- مقایسه دو نسبت : مقایسه شیوع یا خطر

به این منظور با توجه به نمادهای جدول زیر و محاسبه مقادیر مشاهده شده بیماری (O) که با تعداد موارد بیماری در میان مواجهه یافتگان (a) برابر است، مقادیر منتظره (E) برای خانه a و محاسبه واریانس (V) آن، فرضیه صفر که وجود اختلاف بین دو نسبت است با آزمون کاس اسکور به شرح زیر ارزیابی می شود.

		مواجهه		جمع
		دارد	ندارد	
پیامد	دارد	a	b	n1
	ندارد	c	d	n0
جمع		m1	m0	N

$$\chi^2 = \frac{(O - E)^2}{V} \quad V = \frac{n_1 n_0 m_1 m_0}{N^2 (N - 1)} \quad O = a \quad E = m_1 n_1 / N$$

۳۷- مقایسه دو میزان : به این منظور با توجه به نمادهای جدول زیر اختلاف بین دو میزان ارزیابی می شود. در اینجا نیز مانند جدول بالا مقادیر مورد انتظار و واریانس خانه را محاسبه و با آزمون کاس اسکور ارزیابی می شود. این آزمون در اپیدمیولوژی، مانند - هنزل نام دارد.

		مواجهه		جمع
		دارد	ندارد	
موارد		a	b	n1
شخص - زمان در معرض خطر		y1	y0	y
میزان در ۱۰۰۰ شخص - زمان		r1	r0	r

$$\chi^2 = \frac{(O - E)^2}{V} \quad E = n y_1 / y \quad \text{and} \quad V = n y_1 y_0 / y^2$$



۳۸- فرمول محاسبه کسر پیشگیری شده در مطالعات مداخله ای و کارآزمایی های بالینی

$$\text{Prevented fraction (\%)} = 100 \times [\text{rate (or risk) difference} / \text{rate (or risk) in the unexposed}]$$

این فرمول نسبتی از کسانی است که از بیمار شدن بوسیله مداخله انجام شده، پیشگیری شده اند. در مطالعات ارزیابی واکسن این نسبت کارایی واکسن (Vaccine Efficacy) نام دارد.

۳۹- فرمول شیوع ظاهری:

$$\text{شیوع ظاهری} = \text{شیوع واقعی (p)} * \text{حساسیت} + (1-p) * (1-\text{Spesifity})$$

۴۰- فرمول محاسبه شیوع واقعی بیماری وقتی جمع حساسیت و ویژگی یک نباشد.

$$\text{ویژگی (۱-)} + \text{حساسیت} / (\text{ویژگی} - ۱) + \text{شیوع ظاهری} = \text{شیوع واقعی}$$

۴۱- فرمول خطر نسبی افزوده

$$\text{میزان بروز در جمعیت عمومی} - \text{میزان بروز در گروه مورد مقایسه} /$$

$$\text{میزان بروز در جمعیت عمومی} - \text{میزان بروز در جمعیت} - \text{مورد مطالعه} = \text{خطر افزوده نسبی}$$

۴۲- فرمول جز قابل پیشگیری در جمعیت

$$\text{میزان بروز در کل جمعیت} / \text{میزان بروز در افراد مواجهه یافته} - \text{میزان بروز در کل جمعیت} = \text{جز قابل پیشگیری}$$

۴۳- جز پیشگیری شده جمعیت

$$\text{میزان بروز بیماری در افراد مواجهه نیافته} / \text{میزان بروز در افراد مواجهه نیافته} - \text{میزان بروز در کل جمعیت} =$$

۴۴- ارزش اخباری مثبت

$$(\text{ویژگی} - ۱) * (1-p) + \text{حساسیت} * p = \text{ارزش اخباری مثبت}$$

۴۵- ارزش اخباری منفی

$$\text{ارزش اخباری منفی} = sp(1-p) / p(1-sen) + sp(1-p)$$

۴۶- آماره کاپا

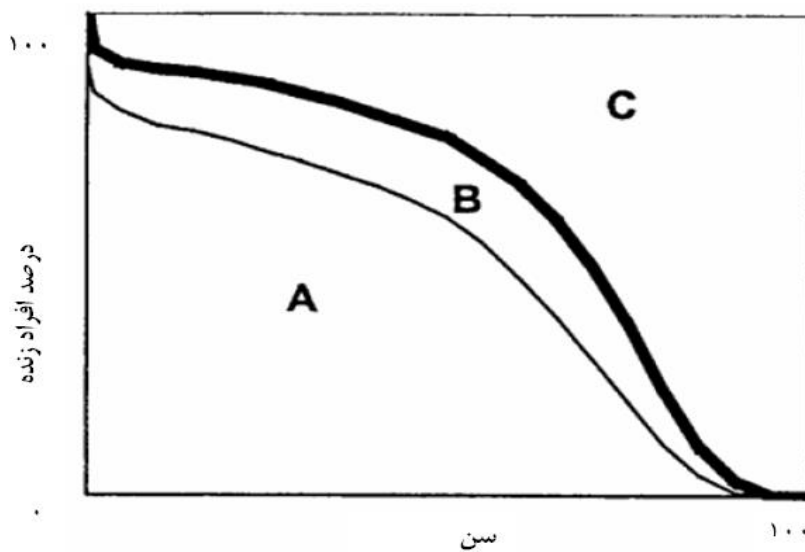
درصد توافق بر مبنای شانس - ۱۰۰٪ / توافق بر مبنای شانس٪ - توافق مشاهده شده = kappa

۴۷- مقدار بازده مفید یک مداخله، روش یا رژیم درمانی در شرایط مطلوب را کارایی (Efficacy) می نامند. یا عبارتی سنجشی از دامنه بهره وری در شرایط مطلوب است.

۴۸- اندازه گیری درجه انطباق نتایج حاصل از کاربرد یک مداخله در یک جمعیت با نتایج مورد انتظار کارسازی یا اثر بخشی (Effectiveness) نام دارد. یا عبارتی سنجشی از دامنه تاثیر یک خدمت است.

۴۹- تاثیرات یا نتایج نهایی که در رابطه با منابع صرف شده (پول، زمان و...) برای ارائه یک خدمت یا مداخله بدست می آید کارآمدی (Efficiency) نام دارد. یا عبارتی نسبت هزینه به منفعت را بدست می آورند.

۵۰- شاخص سال های عمر از دست رفته یا DALYs از دو جزء تشکیل شده است:  
 سال های از دست رفته به واسطه مرگ زود هنگام : YLL : Years of Life Lost (بخش c در شکل زیر که منحنی بقا را در یک جمعیت فرضی نشان می دهد. و سال های از دست رفته به واسطه زندگی توام با ناتوانی : YLD : Years Lived with Disability (تابعی از بخش B در شکل زیر)



۵۱- رابطه تبدیل OR به RR : اگر نوع مطالعه اجازه دهد، به راحتی و با استفاده از فرمول های زیر می توان این دو شاخص را به یکدیگر تبدیل نمود.

$$RR = \frac{OR}{1 - (BR \cdot (1 - OR))} \quad \text{and} \quad OR = \frac{RR(1 - BR)}{1 - (BR \cdot RR)}$$

*BR = Baseline Risk or Risk in group without risk factor*

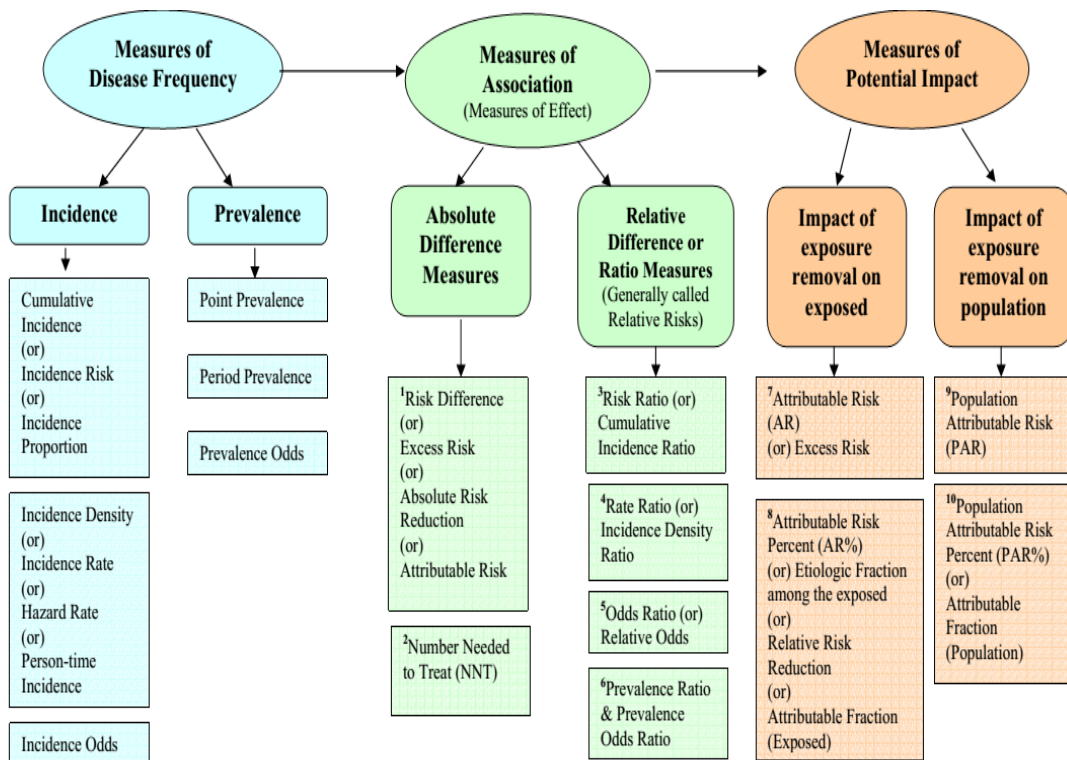
۵۲- محاسبه تفاوت استاندارد شده میانگین ها با استفاده از نسبت شانس وقتی که یک متغیر عددی را به صورت دو حالت در آورده و نسبت شانس را گزارش نمایند.

$$SMD = \frac{\sqrt{3}}{\pi} \log OR$$

$$Se(SMD) = \frac{\sqrt{3}}{\pi} Se(\log OR)$$

ادامه پیوست شماره ۳:

سنجش ها و اندازه های اساسی در مطالعات اپیدمیولوژیک



**ادامه پیوست شماره ۳:** برای محاسبه اندازه های اپیدمیولوژیک از جدول دو در دو و فرمول های زیر استفاده می شود.

		Outcome (Disease)		
		Yes	No	
Exposure	Yes	a	b	a + b
	No	c	d	c + d
		a + c	b + d	

Other notation used:

- $I_o$  = Incidence of outcome among the unexposed (baseline risk)
- $I_e$  = Incidence of outcome among the exposed
- $I_t$  = Incidence of outcome in the total population (exposed and unexposed)
- $P_{exp}$  = Prevalence of exposure in the population
- $P_o$  = Prevalence of outcome among the unexposed
- $P_e$  = Prevalence of outcome among the exposed
- RR = Relative Risk (could refer to a Risk Ratio or a Rate Ratio)
- PR = Prevalence Ratio
- OR = Odds Ratio
- AR = Attributable Risk
- PAR = Population Attributable Risk
- ARR = Absolute Risk Reduction
- RRR = Relative Risk Reduction
- NNT = Number Needed to Treat
- CIR = Cumulative Incidence Ratio
- IDR = Incidence Density Ratio
- PF = Prevented Fraction

$$^1 \text{Risk Difference (ARR, AR)} = \frac{a}{a+b} - \frac{c}{c+d} = I_e - I_o$$

$$^2 \text{Number Needed to Treat (NNT)} = 1 / \text{ARR}$$

$$^3 \text{Risk Ratio (RR, CIR)} = \frac{a/(a+b)}{c/(c+d)} = I_e / I_o$$

$$^4 \text{Rate Ratio (RR, IDR)} = \text{see end of this handout}$$

$$^5 \text{Odds Ratio (OR)} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

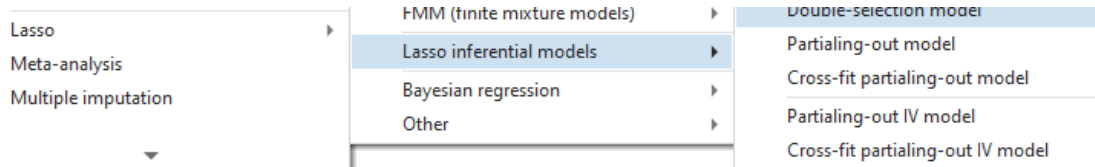
<sup>6</sup> Prevalence Ratio (PR)	=	$P_c / P_o$
<sup>7</sup> Attributable Risk (AR)	=	Same formula as Risk Difference
<sup>8</sup> Attributable Risk Percent (AR%)*	=	$\frac{I_c - I_o}{I_c} * 100 = \frac{AR}{I_c} * 100$
	=	$\frac{a/(a+b) - c/(c+d)}{a/(a+b)}$
Alternative formula for AR%	=	$\frac{(RR - 1)}{RR} * 100$
AR% in a case-control study	=	$\frac{(OR - 1)}{OR} * 100$
<sup>9</sup> Population Attributable Risk (PAR)	=	$I_t - I_o$
Alternative formula for PAR	=	$AR * P_{exp}$
<sup>10</sup> Population Attributable Risk Percent (PAR%)	=	$\frac{I_t - I_o}{I_t} * 100$
Alternative formula for PAR%	=	$\frac{P_{exp}(RR-1)}{P_{exp}(RR-1) + 1} * 100$
<sup>4</sup> Rate Ratio (RR, IDR)	=	$\frac{a/N1}{b/N2}$

This formula for Rate Ratio is based on the following 2 x 2 table format:

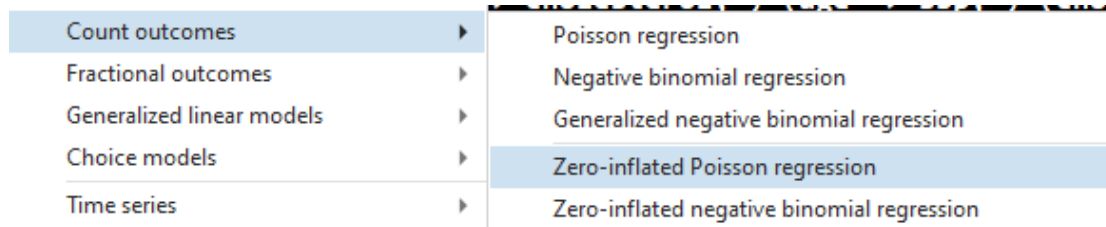
	Cases (Outcome)	Person-time
Exposed	a	N1
Unexposed	b	N2

## پیوست تکمیلی شماره ۳ و معرفی روش های جدیدتر تحلیل در نرم افزار stata

۱) **لاسو (lasso)** روشی برای انتخاب متغیرها برای اجرای مدل های رگرسیونی است. در نسخه ۱۷ نرم افزار استاتا برای رگرسیون های خطی، پواسون و لجستیک می توان از این روش استفاده نمود.

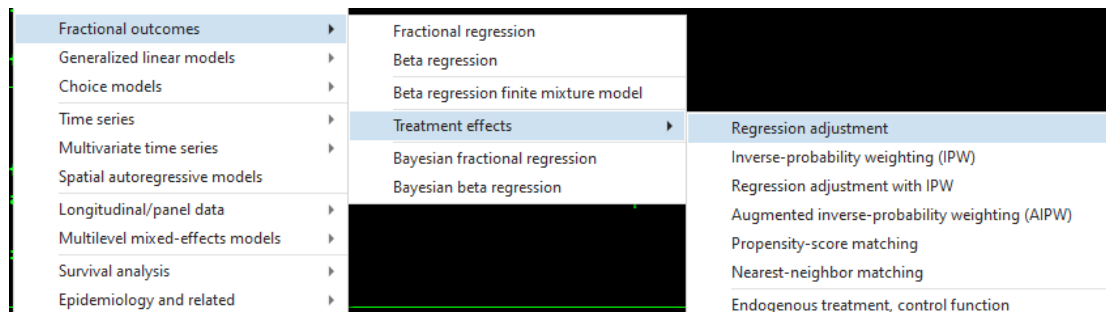


۲) وقتی پیامد مدل رگرسیونی ما شمارشی باشد و تعداد صفرها در متغیر پاسخ بیش از حد و زیاد باشد از روش **Zero-inflated** در رگرسیون پواسون برای برآورد پارامترهای مدل استفاده می شود. مسیرهای اجرا را در منوی **statistics** در زیر آمده است. در واقع همان رگرسیون پواسن با قدرت بیشتر است.

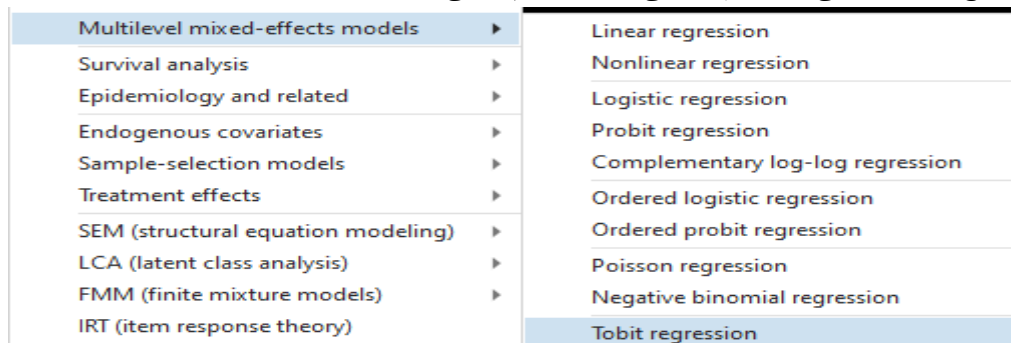


### ۳) رگرسیون بتا **Beta regression**

مسیر اجرا را در نرم افزار ببینید. وقتی متغیر پاسخ کسری یا بین صفر و یک باشد از این روش استفاده می شود. مثال درصد هماتوکریت مثلا ۴۵٪ باشد. بریا تعیین عوامل موثر بر آن می توان از رگرسیون بتا استفاده نمود.

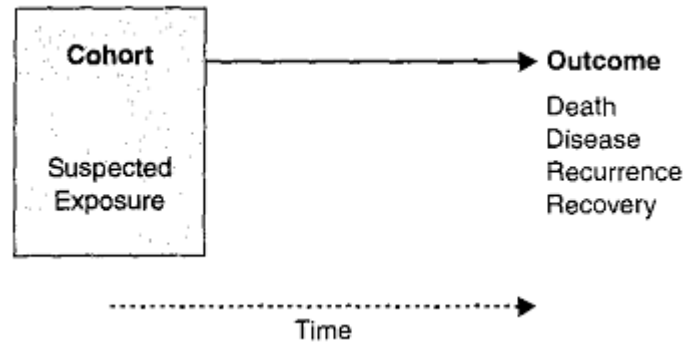


۴) **رگرسیون توبیت**، وقتی متغیر پاسخ متغیری عددی و کرانه دار باشد. مثلا در اقتصاد سلامت سوال این است در صورت ابتلا به بیماری یا برای سلامتی چقدر حاضر به پرداخت پول برای هزینه درمان بیماری X هستید؟ برخی پاسخ ها هیچ و صفر و برخی حداکثر پول و بی نهایت و تمام دارایی است. دستور **db tobit**

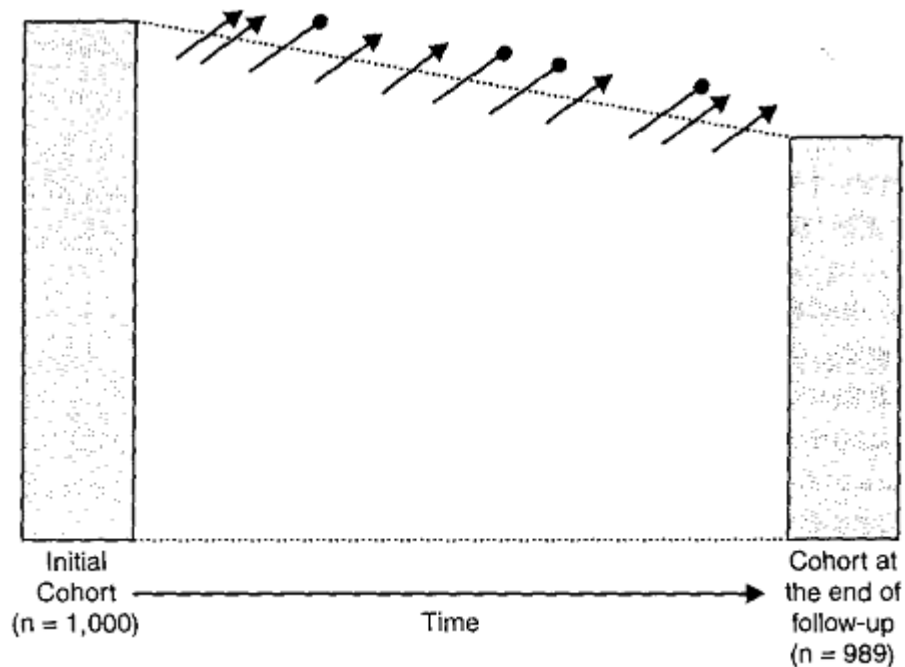


## پیوست شماره ۴: شکل های مربوط به طراحی انواع پژوهش ها و مطالعات اپیدمیولوژیک

الف- مطالعه همگروهی (کوهورت) با نمایش عناصر اصلی مطالعه: مواجهه، زمان پیگیری و پیامد

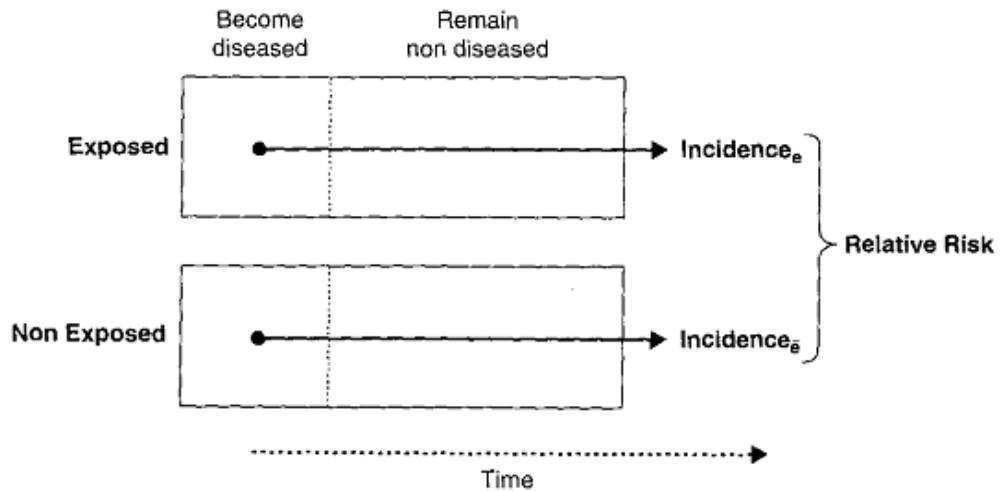


ب- مطالعه کوهورت فرضی با حجم نمونه ۱۰۰۰ نفر در ابتدای مطالعه و از دست رفتن ۱۱ نفر در طی مطالعه (۴ نفر ابتلا به بیماری و ۷ نفر گمشده) و رسیدن ۹۸۹ نفر به انتهای مطالعه

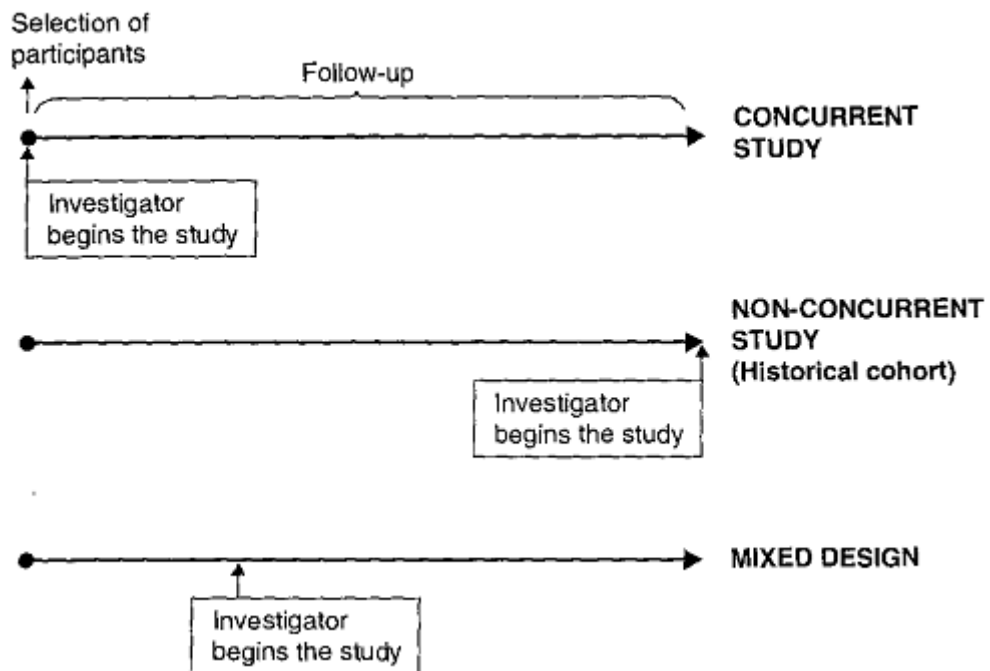




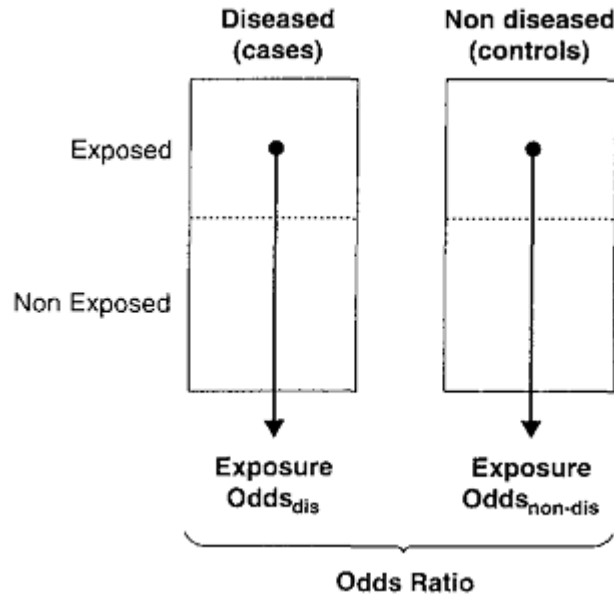
پ- رویکرد تحلیلی در مطالعات کوهورت : محاسبه بروز در گروه های مواجهه یافته و نیافته و تعیین خطر نسبی



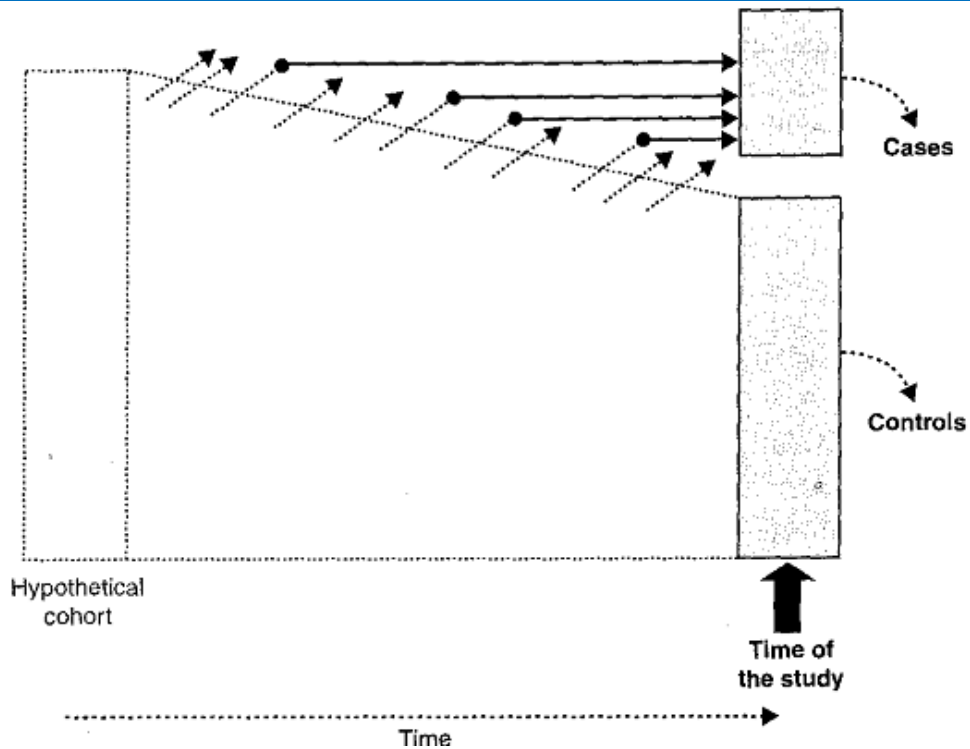
ت- انواع مطالعات کوهورت : از نظر زمان شروع مطالعه و انتخاب افراد و پیگیری آنها که به سه طراحی همزمان، تاریخی یا غیرهمزمان، طراحی ترکیبی یا مخلوط در زیر تقسیم شده اند.



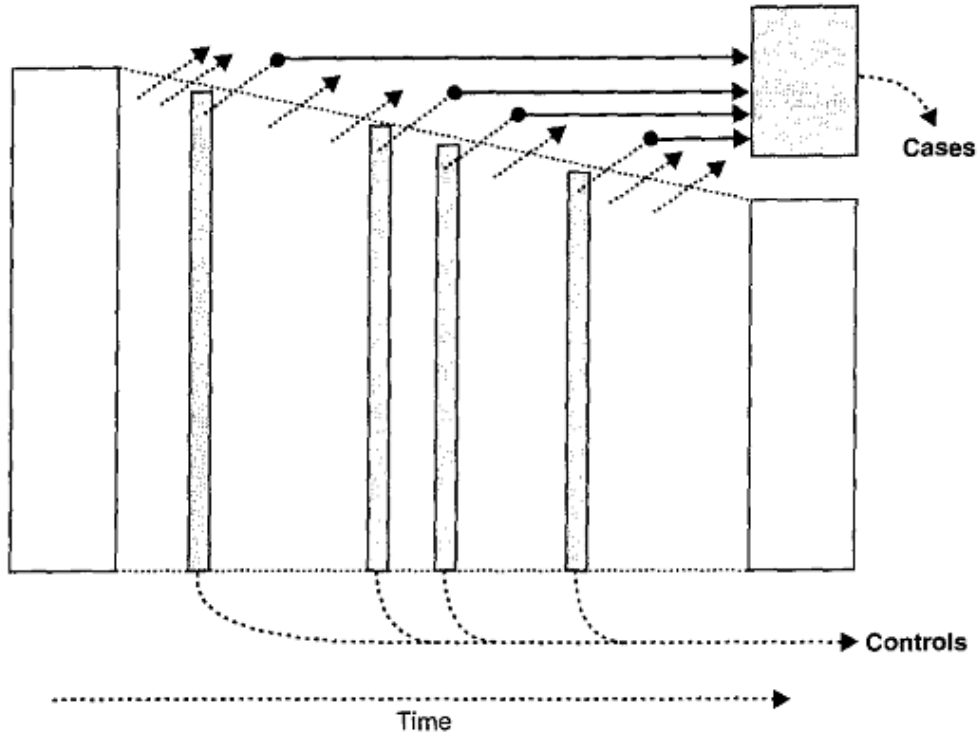
ث- مطالعه مورد - شاهد و رویکرد تحلیل آن و محاسبه نسبت شانس در بیماران و افراد شاهد با توجه به سابقه مواجهه در آنها



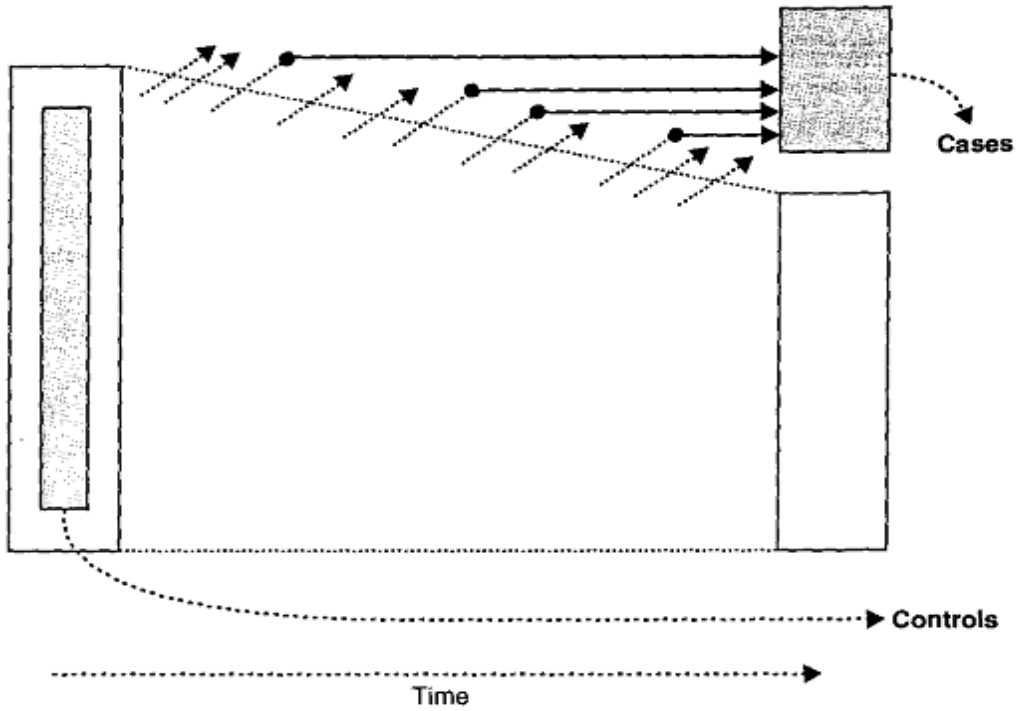
د- مطالعه مورد - شاهد مبتنی بر موردها که در آن موارد و شاهدها از یک کوهورت فرضی انتخاب شده اند.



ر- طراحی مطالعه مورد - شاهد لانه ای : در طی مدت زمان مطالعه، همزمان با بروز یک مورد، کنترل نیز انتخاب می شود.

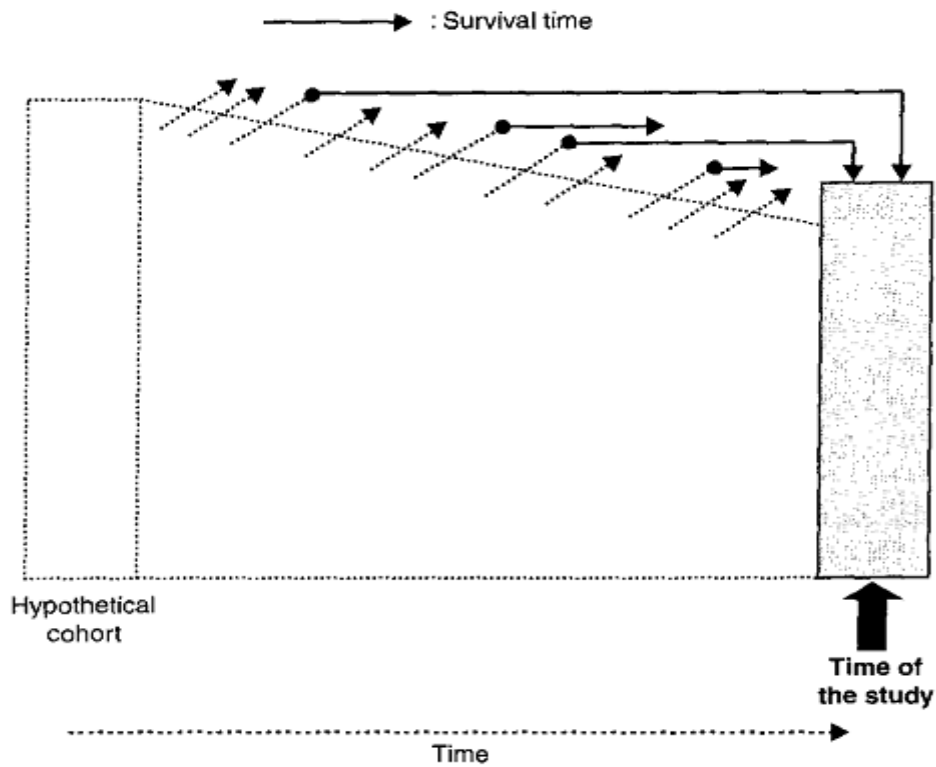


ذ- طراحی مورد - همگروهی : کنترل ها مبتنی بر گروه کوهورت پایه انتخاب می شوند.



### ح- طراحی مطالعه مقطعی

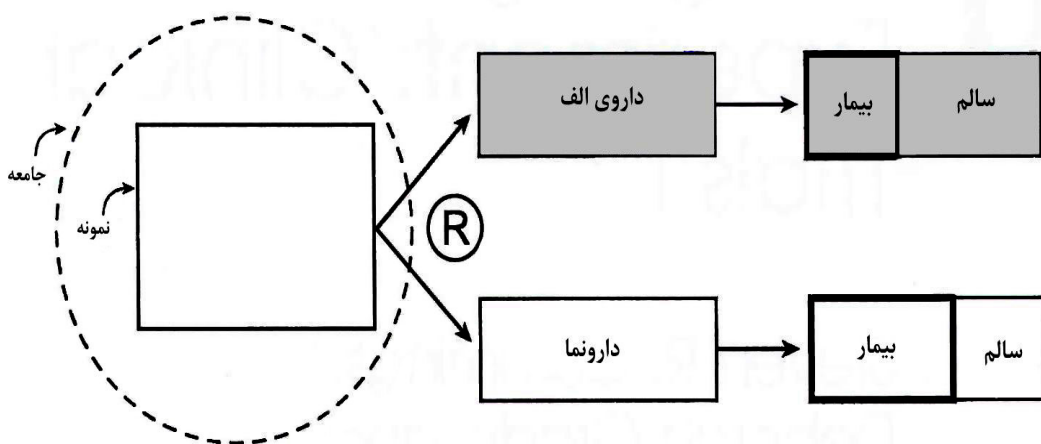
در یک زمان یا دوره زمانی نمونه ای از جامعه انتخاب و وضعیت آنها درخصوص بیمار یا غیربیمار و وضعیت مواجهه آنها بررسی می شود. تورش بقا مهمترین تورش این مطالعات است. بررسی همزمان مواجهه و پیامد نیز مشخصه اصلی این مطالعه است. تقدم زمانی مواجهه و پیامد نامعلوم است.



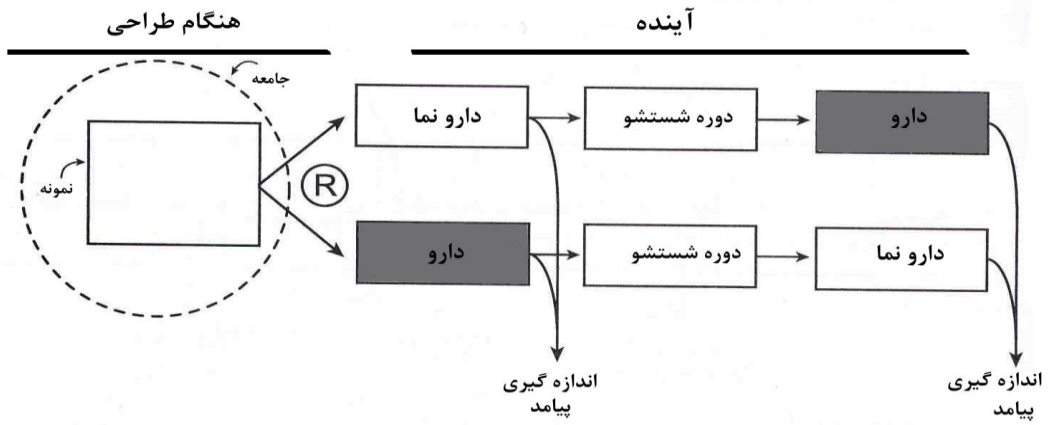
### چ- طراحی مطالعه مداخله ای - کار آزمایی بالینی

هنگام طراحی

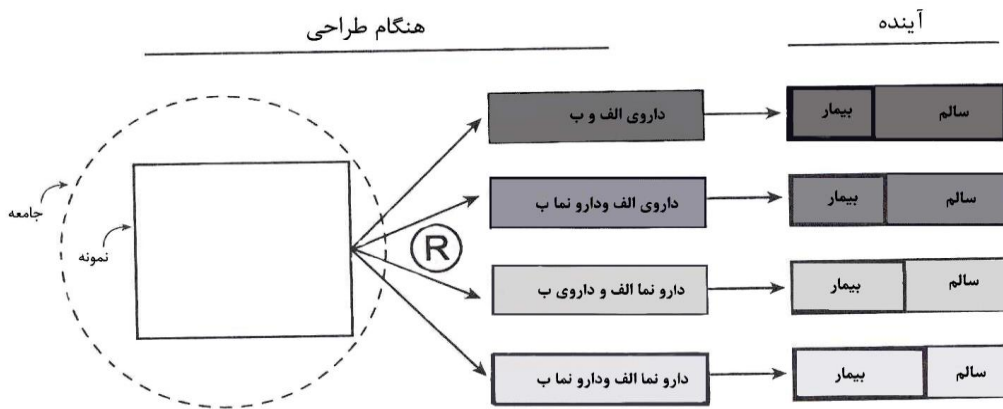
آینده



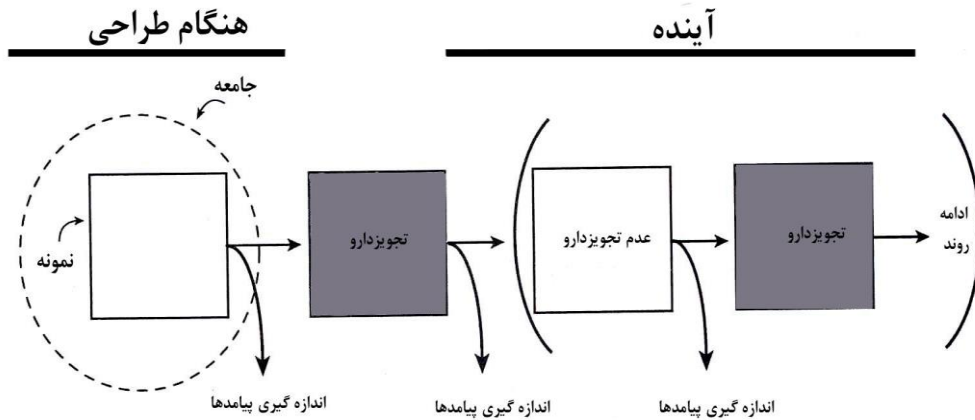
ع - طراحی کار آزمایی متقاطع



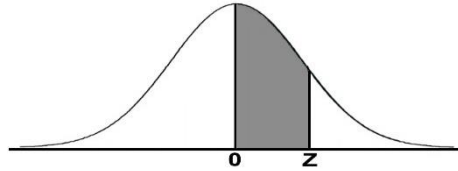
غ - طراحی کار آزمایی بالینی به صورت فاکتوریل



ی - طراحی کار آزمایی بالینی متوالی



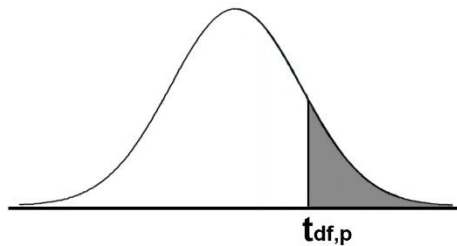
## پیوست ۵ - جدول توزیع Z



Area between 0 and z

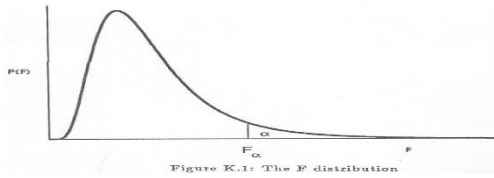
	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

## پیوست ۶ - جدول توزیع t



t table with right tail probabilities

df/p	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
1	0.32492	1	3.077684	6.313752	12.7062	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.5407	5.84091	12.924
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.94318	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.306	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.36343	1.795885	2.20099	2.71808	3.10581	4.437
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.681	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.34503	1.76131	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.75305	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.015
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.25658	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.31946	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.68485	1.317836	1.710882	2.0639	2.49216	2.79694	3.7454
25	0.25606	0.68443	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75	3.646
inf	0.253347	0.67449	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905



## پیوست ۷- جدول توزیع F

**Critical values of F for the 0.05 significance level:**

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
31	4.16	3.31	2.91	2.68	2.52	2.41	2.32	2.26	2.20	2.15
32	4.15	3.30	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.29	2.89	2.66	2.50	2.39	2.30	2.24	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11



**پیوست شماره ۸: جدول عمر خلاصه شده**

X	N	nPx	nDx	nMx	a	nqx	np <sub>x</sub>	l <sub>x</sub>	nd <sub>x</sub>	nL <sub>x</sub>	T <sub>x</sub>	e <sub>x</sub>
0	1	369300	2022	0.005	0.1	0.005	0.995	100000	545	99510	7471075	75.08
1	4	1462700	476	0.000	0.4	0.001	0.999	99455	129	397510	7371566	18.54
5	5	1755600	306	0.000	0.5	0.001	0.999	99326	87	496413	6974056	14.05
10	5	1974200	394	0.000	0.5	0.001	0.999	99239	99	495949	6477643	13.06
15	5	1997800	1315	0.001	0.5	0.003	0.997	99140	326	494887	5981694	12.09
20	5	2041900	2377	0.001	0.5	0.006	0.994	98815	573	492639	5486807	11.14
25	5	2155700	2824	0.001	0.5	0.007	0.993	98241	641	489602	4994168	10.20
30	5	2193000	3659	0.002	0.5	0.008	0.992	97600	811	485971	4504566	9.27
35	5	2150700	4764	0.002	0.5	0.011	0.989	96789	1066	481279	4018595	8.35
40	5	2114300	6906	0.003	0.5	0.016	0.984	95723	1551	474737	3537316	7.45
45	5	2167100	9944	0.005	0.5	0.023	0.977	94172	2136	465520	3062578	6.58
50	5	1608000	10873	0.007	0.5	0.033	0.967	92036	3060	452530	2597058	5.74
55	5	1341000	13304	0.010	0.5	0.048	0.952	88976	4307	434113	2144528	4.94
60	5	1362300	20756	0.015	0.5	0.073	0.927	84669	6213	407813	1710414	4.19
65	5	1266000	28937	0.023	0.5	0.108	0.892	78456	8482	371075	1302602	3.51
70	5	1056300	36075	0.034	0.5	0.157	0.843	69974	11009	322349	931527	2.89
75	5	651300	33265	0.051	0.5	0.226	0.774	58965	13353	261443	609178	2.33
80	5	409100	37544	0.092	0.5	0.373	0.627	45612	17024	185501	347735	1.87
+85	5	346500	61059	0.176	0.5	1.000	0.000	28588	28588	162234	162234	1.00

## ایندکس ها (واژه نامه فارسی - انگلیسی):

انواع برنامه نرم افزار استتا ۹	Stata (MP/SE/IC/Small)
پنجره مرور ۱۳	review window
پنجره متغیرها ۱۳	Variable Window
پنجره دستور یا فرمان ۱۳	Command Window
پنجره نتایج ۱۳	Result Window
ویراستار داده ها ۱۵	Data Editor
جستجو گر یا گشت زن داده ها ۱۵	Data Browser
مدیریت داده ها ۲۲	Data Management
داده های گمشده ۲۵	Missing Data
آمار ۴۴	Statistic
تحقیق ۴۴	Research
آمار زیستی ۴۴	Biostatistic
جامعه آماری ۴۵	Statistical Population
نمونه ۴۵	Sample
متغیر ۴۵	Variable
متغیر کمی ۴۵	Quantitative Variable
مقیاس های اندازه گیری ۴۵	Scale of Measurement
مقیاس اسمی ۴۶	Nominal Scale
مقیاس رتبه ای ۴۶	Ordinal Scale
مقیاس فاصله ای ۴۶	Interval Scale
مقیاس نسبتی ۴۶	Ratio Scale
آمار توصیفی ۵۰	Descriptive Statistic
شاخص های تمایل مرکزی ۵۵	Measure of Central Tendency
میانگین ۵۵	Mean
انحراف معیار ۵۶	standard deviation
خطای معیار ۵۷	standard error
واریانس یا پراش ۵۸ و ۵۵	variance
نمودار ۶۳	Graph
نمودار ستونی ۶۳	Bar Chart
نمودار دایره ای ۶۳	Pie Chart
هیستوگرام ۶۹	Histogram

Kaplan- Meier Graph	۷۸ و ۲۴۲
Life Table	۲۴۳ و ۸۰
Hypothesis Testing	۸۰ فرضیه
Null Hypothes	۸۶ فرضیه صفر
Alternative Hypothes	۸۷ فرضیه جایگزین
Normal Distribution	۸۵ توزیع استاندارد
Confidence Interval	۳۶۱ و ۸۶ حدود یا فاصله اطمینان
ANOVA	۱۲۴ تحلیل واریانس یکطرفه
MANOVA	۱۳۷ تحلیل واریانس چند متغیره
ANCOVA	۱۳۳ تحلیل کوواریانس
Homoscedasticity	۱۴۷ یکسان بودن واریانس
Sphericity	۱۴۷ کرویت و تقارن
Wide and Long Data	۱۵۰ داده های پهن و دراز
Mixed Model	۱۵۲ مدل مخلوط و پیچیده
Modeling	۱۷۴ مدل سازی
Correlation	۱۶۴ همبستگی
Linear Regression	۱۷۷ رگرسیون خطی
Logistic Regression	۱۹۰ رگرسیون لجستیک
Model Making	۲۰۱ ساخت مدل
Maximum Likelihood Method	۲۰۳ روش حداکثر درست نمایی
ROC	۲۰۶ منحنی راک
Conditional Logistic Regression	۲۱۳ رگرسیون لجستیک شرطی
Ordinal Logistic Regression	۲۲۰ رگرسیون لجستیک رتبه ای
Poisson Regression	۲۳۰ رگرسیون پواسون
Cox Regression	۲۳۷-۲۶۳ رگرسیون کاکس
Cusation	۳۰۹ علیت
Epidemiologic Study	۳۰۹ مطالعات اپیدمیولوژیک
Cross-Sectional Study	۳۱۶ مطالعه مقطعی
Cohort Study	۳۱۸ مطالعه همگروهی
Case Control Study	۳۲۵ مطالعه مورد شاهد
Relative Risk	۳۲۰ خطر نسبی
Odds Ratio	۳۳۲ نسبت شانس
Randomization	۳۴۰ تصادفی سازی

Clinical Trial	۳۴۱	کارآزمایی بالینی
Blinding	۳۴۵	کورسازی
Selection Bias	۳۶۹	خطای (تورش یا بایاس) انتخاب
Information Bias	۳۷۲	خطای اطلاعات
Confounding	۳۷۴	مخدوش کننده
Intracation	۳۶۵	اثر متقابل یا اینترکشن
Diagnosis Test	۳۷۷	آزمون های تشخیصی
Sensitivity/ Specificity	۳۷۸	حساسیت/ویژگی
Inequality	۳۸۵	نابرابری
Equality	۳۸۵	برابری
Equity	۳۸۵	عدالت
Concentration Curve	۳۳۹۱	منحنی تمرکز
Time Series	۳۹۶	سری های زمانی
ARIMA Model	۴۰۹	مدل سری زمانی (آریمادل)
Sampling Methodes	۴۲۰	روشهای نمونه گیری
MetaAnalysis	۴۲۹	متاآنالیز
Random/ Fix Model	۴۳۱	مدل تصادفی/ ثابت
Validity	۴۵۹	روایی
Reliability	۴۶۱	پایایی
Factor Analysis	۴۶۱	تحلیل عامل

## References:

1. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Second ed. USA: John Wiley & Sons; 2008.
2. Szklo M, Nieto FJ. Epidemiology: Beyond the Basics. USA: John Wiley & Sons; 2004.
3. Woodward M. Epidemiology: Study Design and Data Analysis. Second, editor. USA: Chapman & Hall/CRC; 2004.
4. Newman SC. Biostatistical Methods in Epidemiology. USA: John Wiley & Sons; 2003.
5. Stanley L, editor Logistic Regression Text Book. Erasmus Summer Program; Erasmus - Holand: John Wiley & Sons. 2012.
6. Landon school of hygiene & tropical medicine. Advance Statistical methods in epidemiology (ASME): Unit study 2001. Landon school of hygiene & tropical medicine ; 2001.
7. Jewell NP. Statistics for Epidemiology. USA: John Wiley & Sons; 2003.
8. Ahmadi A, Manual of Stata. first, editor. esfahan: Asef; 2006. [Persian]
9. Moghimi B, Safaei A. Epidemiology.blogfa.com: Stata Slide, IRAN: www.epidemiology.blogfa.com; 2008 [cited 2013].
10. Hamilton LC. Statistics with Stata Version 12 USA: John Wiley & Sons; 2012.
11. Longest Kyle C. Using Stata for Quantitative Analysis. USA: Chapman & Hall/CRC; 2011.
12. Rabe-Hesketh S, Everitt B. A Handbook of Statistical Analyses Using Stata. Fourth ed. USA: Chapman & Hall/CRC; 2007.
13. Malekafzali H, Majdzadeh R, Tavakoli S, Fotohi A. Research Methodology in medical Science. Tehran: Ministry of Health; 2003.[Persian]
14. Daneil W. Biostatistic: a foundation for analysis in the health science. Translated by Ayatollahi SMT. USA John Wiley & Sons 1990. [Persian]
15. Ahmadi A. The incidence of depression and related causes among Bakhteyari tribal population, Iran – 2006. Journal of Gorgan University of Medical Sciences. 2008;10(2):1-7. [Persian]
16. Mitchel Michael N. Visual Guide to Stata Graphics. Third ed. USA: Stata Press; 2012.
17. Mohammad K, Malekafzali H, Nahaptiyan V. Statistical methods and health indexes Tehran: Salman; 1993. [Persian]
18. Ahmadi A, Hasanzadeh J. to Determine the relative factor on hypertension in kohrang. Iranian Journal of Epidemiology. 2008;4(2):19-25. [Persian]
19. Ahmadi A, Hasanzadeh J. Metabolic Control and Care Assessment in Patients with Type 2 Diabetes in Chaharmahal and Bakhtiyari Province 2008. IJEM. 2009;11(1):33-9.
20. Ahmadi A, Hasanzadeh J. quality of life in diabetic patient in Chaharmahal and Bakhtiyari Province. Journal of north khorasan university. 2011;3(1):7-13. [Persian]

21. Ahmadi A, Hasanzadeh J . The survey of family history of diabetes in patients with type 2 diabetes in Chaharmahal va Bakhteyari province, Iran. Journal of Shahrekord University of Medical Sciences. 2009;11(2):1-7. [Persian]
22. Ahmadi A, Mobasheri M, Maboudi A. Knowledge generation, Knowledge transfer, Knowledge promotion and the application of Science evidence in SKUMS International Journal of Current Research. 2013;5(8):2275-8.
23. Ahmadi A, Mousaviyan R. Survey About 1107 Case Of Brucellosis. 1st International Congress of Zoonotic Disease In The Universities Of Iran; Karaj-Iran: Islamic Azad University of Karaj 2005. p. 14-5. [Persian]
24. Kahn HA, Sempos CT. Statistical methods in epidemiology. USA: John Wiley & Sons; 1989.
- 25 . Bernard R. fundamental of biostatistics. 5th ed. USA: Brooks cole, Harvard University; 2000.
26. Breslow N E, Day NE. Statistical Methods in Cancer Research. Volume I - The analysis of case-control studies. IARC Scientific Publications 2004.
27. Ahmadi A, Khadivi R. An epidemiological survey on clustering appendicitis in Chaharmahal and Bakhtiari Province. ISMJ. 2008;11(1):40-6.
28. Beckett S. Introduction to Time Series using Stata. USA: John Wiley & Sons; 2013.
29. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. USA: John Wiley & Sons; 2005.
30. Kleinbaum DG, Klein M. Logistic Regression: A Self-Learning Text. USA: John Wiley & Sons; 2010.
31. Reisi R, Yousefi H, Shojaei H, Ahmadi A. Improving the capabilities of the physicians working in health care centers of Chaharmahal va Bakhtiari province for doing applied researches. Journal of Shahrekord University of Medical Sciences. 2010;11(4):42-8. [Persian]
32. KUMS (Kerman University Medical of Sciences). ICEP:The breif Epidemiology and biostatistics intensive program in the Kerman University Medical of Sciences (Unpublished). 2006. [Persian]
33. Ahmadi A, Shojaei M. Estimation of life expectancy changes in Chaharmahal va Bakhtiyary Province, Iran, during 2005-2009. J Shahrekord Univ Med Sci. 2011;13(4):74-80. [Persian]
34. Haghdoost A.K. Systematic review & Meta analysis. Kerman: Pezeshkan bedon Marz; 2011. [Persian]
35. Soori H, Sanago M. Common errors in biomedical research and their control. Tehran: Shahid Beheshti University of Medical Sciences; 2008.
36. Oakes JM, Kaufman JS. Methods in Social Epidemiology: John Wiley & Sons; 2006.
37. Sterne J. Meta-Analysis: An Updated Collection from the Stata Journal. USA: Stata; 2009.
38. Cryer JD. Time Series Analysis Translated by: Niromand HA. Mashhad: Ferdowsi University of Mashhad; 1992. [Persian]
39. Dos Santos I. Cancer Epidemiology : Principles and Methods. France: International Agency for Research on Cancer: World Health Organization. IARC Publication 1999.

40. Lashkardoost H, Zeighami B, Mahmoudi M, Hassanzadeh J, Hamed A, et al. Assessment of Quanti FERON-TB Gold (In-Tube) Test in Tuberculosis Diagnosis. *Iranian Journal of Epidemiology*. 2010;6(1):26-32. [Persian]
41. Ansarimoghadam A, Pourolajal J, Haghdooost A, Sadeghirad B, Najafi F. Systematic review & Meta analysis. Kerman; 2012. [Persian]
- 42- Rezaei S, Tahmasebi R. *Statistical Simulation*. Tehran; 2011. [Persian]
- 43- Orsini Nicola, Bellocco Rino, Bottai Matteo, Greenland Sander. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *The Stata Journal*, 2008: 29–48
- 44- Rafieian-Kopaei M, Behradmanesh S, Ahmadi A, Nasri H. Association of serum lipids with level of blood pressure in type 2 diabetic patients. in Press
- 45- Nasri H, Ahmadi A, Baradaran A, Nasri P, Hajian S, Pour-Arian A, et al. A biochemical study on ameliorative effect of green tea against contrast media induced acute kidney injury. *J Ren Inj Prev* 2014; 3(1):19-21.
- 46- Ahmadi A, Montazeri A. Determinants of health inequalities in urban, rural and tribal population: a qualitative study. *Iranian Epidemiology Meeting*. Yasoj. 2013
- 47- Ahmadi A, Montazeri A, Shakiba M. Inequalities regression based measures for determining socioeconomic inequality: a methodologic approach in social epidemiology. *Iranian Epidemiology Meeting*. Yasoj. 2013
- 48- Ahmadi A, Eshrati B. Root Cause Analysis(RCA) for Determinat of Maternal Mortality in Iran. in Press
- 49- Ahmadi A, Hashemi S, Norozi M. Survival Analysis in Colorectal Cancer in Iran: Cohort Study. in Press
- 50- Knol M J, Vandenbroucke J P, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. 2008, 168(9):1073-81
- 51- Kiani FZ, Ahmadi A, Babadi AS, Rouhi H. Profile and preliminary results of Iranian sub cohort chronic obstructive pulmonary disease (COPD) in Shahrekord PERSIAN cohort in southwest Iran. *BMC Pulm Med*. 2021;21(1):105.
- 52- Kiani FZ, Ahmadi A. Prevalence of different comorbidities in chronic obstructive pulmonary disease among Shahrekord PERSIAN cohort study in southwest Iran. *Sci Rep*. 2021;11(1):1548.
- 53- Khamseh ME, Sepanlou SG, Hashemi-Madani N, Joukar F, Mehrparvar AH, Faramarzi E, et al. Nationwide Prevalence of Diabetes and Prediabetes and Associated Risk Factors Among Iranian Adults: Analysis of Data from PERSIAN Cohort Study. *Diabetes Ther*. 2021.
- 54- Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *Lancet*. 2021;398(10304):957-80.
- 55- Soofi M, Pasdar Y, Karami Matin B, Hamzeh B, Rezaei S, Kazemi Karyani A, et al. Socioeconomic-related inequalities in oral hygiene behaviors: a cross-sectional analysis of the PERSIAN cohort study. *BMC Oral Health*. 2020;20(1):63.
- 56- Samaniyan Bavarsad P, Kheiri S, Ahmadi A. Estimation of the 10-Year Risk of Cardiovascular Diseases: Using the SCORE, WHO/ISH, and Framingham

- Models in the Shahrekord Cohort Study in Southwestern Iran. *J Tehran Heart Cent.* 2020;15(3):105-12.
- 57- Najafi F, Rezaei S, Hajizadeh M, Soofi M, Salimi Y, Kazemi Karyani A, et al. Decomposing socioeconomic inequality in dental caries in Iran: cross-sectional results from the PERSIAN cohort study. *Arch Public Health.* 2020;78:75.
- 58- Ahmadi A, Kheiril S, Mohammadian-Hafshejani A, Raeisi H, Mosavi M, Sharifi H. The accuracy of "Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: A modelling analysis based on overseas cases and air travel data". *Int J Infect Dis.* 2020;95:319-20.
- 59- Ahmadi A, Fadaei Y, Shirani M, Rahmani F. Modeling and forecasting trend of COVID-19 epidemic in Iran until May 13, 2020. *Med J Islam Repub Iran.* 2020;34:27.
- 60- NCD risk Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: a pooled analysis of 2181 population-based studies with 65 million participants. *Lancet.* 2020;396(10261):1511-24.
- 61- NCD risk. Repositioning of the global epicentre of non-optimal cholesterol. *Nature.* 2020;582(7810):73-7.
- 62- Kazemi Karyani A, Karmi Matin B, Soltani S, Rezaei S, Soofi M, Salimi Y, et al. Socioeconomic gradient in physical activity: findings from the PERSIAN cohort study. *BMC Public Health.* 2019;19(1):1312.
- 63- Ahmadi A, Sodejani SA, Malekzadeh R, Poustchi H, Solati K. Study of correlation between chronic stressor, biochemical markers and hematologic indices in Shahrekord Cohort Study: A population-based cross-sectional study. *Diabetes Metab Syndr.* 2019;13(3):2170-4.
- 64- Khaledifar A, Hashemzadeh M, Solati K, Poustchi H, Bollati V, Ahmadi A, et al. The protocol of a population-based prospective cohort study in southwest of Iran to analyze common non-communicable diseases: Shahrekord cohort study. *BMC Public Health.* 2018;18(1):660.
- 65- Ahmadi A, Salehi F, Ahmadimirghaed Z. Cancer trends in a province of Southwest of Iran, 2003-2016. *J Res Med Sci.* 2018;23:80.
- 66- Ahmadi A, Salehi F. Evaluation of observed and the expected incidence of common cancers: An experience from Southwestern of Iran, 2010-2014. *J Res Med Sci.* 2018;23:4.
- 67- Ahmadi A, Ramazani R, Rezaghali T, Yavari P. Incidence pattern and spatial analysis of breast cancer in Iranian women: Geographical Information System applications. *East Mediterr Health J.* 2018;24(4):360-7.
- 68- Mohammadbeigi A, Mohammadsalehi N, Moshiri E, Anbari Z, Ahmadi A, Ansari H. The prevalence of phantom vibration/ringing syndromes and their related factors in Iranian' students of medical sciences. *Asian J Psychiatr.* 2017;27:76-80.
- 69- Mahmoodnia L, Sadoughi M, Ahmadi A, Kafeshani M. Relationship between serum irisin, glycemic indices, and renal function in type 2 diabetic patients. *J Renal Inj Prev.* 2017;6(2):88-92.



- 70- Eghtesad S, Mohammadi Z, Shayanrad A, Faramarzi E, Joukar F, Hamzeh B, et al. The PERSIAN Cohort: Providing the Evidence Needed for Healthcare Reform. *Arch Iran Med.* 2017;20(11):691-5.
- 71- Nasri H, Hasanpour Z, Nematbakhsh M, Ahmadi A, Rafieian-Kopaei M. The effect of the various doses of atorvastatin on renal tubular cells; an experimental study. *J Nephropathol.* 2016;5(3):111-5.
- 72- Mohammadbeigi A, Valizadeh F, Mirshojaee SR, Ahmadli R, Mokhtari M, Ghaderi E, et al. Self-rated Health and Internet Addiction in Iranian Medical Sciences Students; Prevalence, Risk Factors and Complications. *Int J Biomed Sci.* 2016;12(2):65-70.
- 73- Mohammadbeigi A, Absari R, Valizadeh F, Saadati M, Sharifimoghadam S, Ahmadi A, et al. Sleep Quality in Medical Students; the Impact of Over-Use of Mobile Cell-Phone and Social Networks. *J Res Health Sci.* 2016;16(1):46-50.
- 74- Ahmadi A, Khaledifar A, Etemad K. Modeling of in hospital mortality determinants in myocardial infarction patients, with and without stroke: A national study in Iran. *J Res Med Sci.* 2016;21:74.
- 75- Ahmadi A, Etemad K, Khaledifar A. Risk factors for heart failure in a cohort of patients with newly diagnosed myocardial infarction: a matched, case-control study in Iran. *Epidemiol Health.* 2016;38:e2016019.
- 76- Sadeghi M, Ahmadi A, Baradaran A, Masoudipoor N, Frouzandeh S. Modeling of the relationship between the environmental air pollution, clinical risk factors, and hospital mortality due to myocardial infarction in Isfahan, Iran. *J Res Med Sci.* 2015;20(8):757-62.
- 77- Nasri H, Hajian S, Ahmadi A, Baradaran A, Kohi G, Nasri P, et al. Ameliorative effect of green tea against contrast-induced renal tubular cell injury. *Iran J Kidney Dis.* 2015;9(6):421-6.
- 78- Nasri H, Ahmadi A, Rafieian-Kopaei M, Bashardoust B, Nasri P, Mubarak M. Association of glomerular C4d deposition with various demographic data in IgA nephropathy patients; a preliminary study. *J Nephropathol.* 2015;4(1):19-23.
- 79- Hasanpour Z, Nasri H, Rafieian-Kopaei M, Ahmadi A, Baradaran A, Nasri P, et al. Paradoxical effects of atorvastatin on renal tubular cells: an experimental investigation. *Iran J Kidney Dis.* 2015;9(3):215-20.
- 80- Amani S, Kheiri S, Ahmadi A. Honey versus diphenhydramine for post-tonsillectomy pain relief in pediatric cases: a randomized clinical trial. *J Clin Diagn Res.* 2015;9(3):Sc01-4.
- 81- Ahmadi A, Soori H, Sajjadi H, Nasri H, Mehrabi Y, Etemad K. Current status of the clinical epidemiology of myocardial infarction in men and women: a national cross-sectional study in Iran. *Int J Prev Med.* 2015;6:14.
- 82- Ahmadi A, Soori H, Sajjadi H. Modeling of in hospital mortality determinants in myocardial infarction patients, with and without type 2 diabetes, undergoing pharmaco-invasive strategy: the first national report using two approaches in Iran. *Diabetes Res Clin Pract.* 2015;108(2):216-22.
- 83- Ahmadi A, Soori H, Mehrabi Y, Etemad K, Sajjadi H, Sadeghi M. Predictive Factors of Hospital Mortality Due to Myocardial Infarction: A Multilevel Analysis of Iran's National Data. *Int J Prev Med.* 2015;6:112.

- 84- Ahmadi A, Soori H, Mehrabi Y, Etemad K, Khaledifar A. Epidemiological pattern of myocardial infarction and modelling risk factors relevant to in-hospital mortality: the first results from the Iranian Myocardial Infarction Registry. *Kardiol Pol.* 2015;73(6):451-7.
- 85- Ahmadi A, Soori H, Mehrabi Y, Etemad K. Spatial analysis of myocardial infarction in Iran: National report from the Iranian myocardial infarction registry. *J Res Med Sci.* 2015;20(5):434-9.
- 86- Ahmadi A, Soori H. Important statistical points to improve and promote the methodology of the articles on medical sciences, particularly nephrology and kidney; a review article. *J Renal Inj Prev.* 2015;4(1):4-8.
- 87- Ahmadi A, Noroozi M, Pourhoseingholi MA, Hashemi-Nazari SS. Effect of metabolic syndrome and its components on survival in colorectal cancer: a prospective study. *J Renal Inj Prev.* 2015;4(1):15-9.
- 88- Ahmadi A, Mosavi-Jarrahi A, Pourhoseingholi MA. Mortality determinants in colorectal cancer patients at different grades: a prospective, cohort study in Iran. *Asian Pac J Cancer Prev.* 2015;16(3):1069-72.
- 89- Nasri H, Tavakoli M, Ahmadi A, Baradaran A, Nematbakhsh M, Rafieian-Kopaei M. Ameliorative effect of melatonin against contrast media induced renal tubular cell injury. *Pak J Med Sci.* 2014;30(2):261-5.
- 90- Nasri H, Behradmanesh S, Ahmadi A, Rafieian-Kopaei M. Impact of oral vitamin D (cholecalciferol) replacement therapy on blood pressure in type 2 diabetes patients; a randomized, double-blind, placebo controlled clinical trial. *J Nephrothol.* 2014;3(1):29-33.
- 91- Nasri H, Behradmanesh S, Ahmadi A, Baradaran A, Nasri P, Rafieian-Kopaei M. Association of serum lipids with level of blood pressure in type 2 diabetic patients. *J Renal Inj Prev.* 2014;3(2):43-6.
- 92- Nasri H, Ahmadi A, Baradaran A, Momeni A, Nasri P, Mardani S, et al. Clinicopathological correlations in lupus nephritis; a single center experience. *J Nephrothol.* 2014;3(3):115-20.
- 93- Mardani S, Nasri H, Hajian S, Ahmadi A, Kazemi R, Rafieian-Kopaei M. Impact of *Momordica charantia* extract on kidney function and structure in mice. *J Nephrothol.* 2014;3(1):35-40.
- 94- Hafizi M, Mardani S, Borhani A, Ahmadi A, Nasri P, Nasri H. Association of helicobacter pylori infection with serum magnesium in kidney transplant patients. *J Renal Inj Prev.* 2014;3(4):101-5.
- 95- Ahmadi A, Mobasheri M, Hashemi-Nazari SS, Baradaran A, Choobini ZM. Prevalence of hypertension and type 2 diabetes mellitus in patients with colorectal cancer and their median survival time: A cohort study. *J Res Med Sci.* 2014;19(9):850-4.
- 96- Ahmadi A, Khaledifar A, Sajjadi H, Soori H. Relationship between risk factors and in-hospital mortality due to myocardial infarction by educational level: a national prospective study in Iran. *Int J Equity Health.* 2014;13:116.
- 97- Ahmadi A, Hashemi Nazari SS, Mobasheri M. Does ethnicity affect survival following colorectal cancer? A prospective, cohort study using Iranian cancer registry. *Med J Islam Repub Iran.* 2014;28:83.

- علی اکبر سیف. ۱۳۷۱. اندازه گیری و ارزشیابی پیشرفت تحصیلی. تهران، تهران، انتشارات آگاه.
- احمدی علی. گزارش مطالعه کوهورت شهرکرد ۱۳۹۴-۱۳۹۹
- بغوری، دلارام، رضایی، میرزاخانی، نوید، ایرانی، اشکان باغبان، ا. ۲۰۱۶. بررسی روایی ظاهری و محتوایی نسخه فارسی مقیاس خستگی پارکینسون (PFS-16) در افراد مبتلابه پارکینسون. فصلنامه علمی-پژوهشی طب توانبخشی، ۵، ۸۵-۹۲.
- عبدالهی پور، فروزان، زارعی، ع. ز. اکبرفهمی، ملاح، اسماعیلی، ک. سمانه ۲۰۱۶. تهیه نسخه فارسی «پرسشنامه سنجش رفتاری کارکرد اجرایی-نسخه پیش دبستانی» و بررسی روایی ظاهری و محتوایی آن. فصلنامه علمی پژوهشی توانبخشی، ۱۷، ۱۲-۱۹.
- حقدوست، اسیوند سرور اصول طراحی پرسشنامه در مطالعات علوم پزشکی، انتشارات نور دانش.
- روانسنجی، ا. ص. ف. ل. یک آزمون خوب دارای چه ویژگی هایی است. زیبا، تقی زاده، عباس، عبادی، علی، منتظری، زهرا، شاهواری، محمود، طاووسی. راضیه، باقرزاده. روانسنجی ابزارهای سنجش سلامت (۱): ترجمه، طراحی ابزار، بررسی روایی محتوا و صوری.
- محمدباقر، کجباف، ایلناز، سجادیان، ابوالقاسم، نوری. بررسی ساختار عاملی، روایی و اعتبار مقیاس گرایشهای اجتماعی مطلوب در بین دانشجویان.
- محمدبیگی، محمدصالحی گل، ع. ۲۰۱۵. روایی و پایایی ابزارها و روش های مختلف اندازه گیری آنها در پژوهش های کاربردی در سلامت. مجله علمی دانشگاه علوم پزشکی رفسنجان، ۱۳، ۱۱۵۳-۱۱۷۰.
- میرزاخانی، نوید، دبیری گل چین، مینو، رضائی، طباطبائی، زینعلی، رزی، یاسائی & راما ۲۰۱۶. ترجمه، روایی ظاهری و محتوایی نسخه فارسی آزمون ارزیابی بازی وانمودین خودانگیخته ی کودکان ۷-۴ ساله. فصلنامه علمی-پژوهشی طب توانبخشی، ۵، ۲۱-۳۰.

.....یادداشت



North Khorasan University  
of Medical Sciences  
Deputy of Research



Shahrekord University  
of Medical Sciences  
Deputy of Research

## Principles and Statistical Methods in Medical Research & Epidemiology through Stata

Ali Ahmadi, Hossein Lashkardoost

With guidance and introduction:

Ali Akbar Haghdoost

این کتاب مشتمل بر پنج بخش: اصول و مبانی کار با نرم افزار Stata، مبانی آمار توصیفی و کار با Stata، آمار تحلیلی در Stata، مدل سازی در اپیدمیولوژی با Stata و روشهای آماری در اپیدمیولوژی می باشد. در این کتاب علاوه بر تسلط کامل بر نرم افزار، موضوعات زیر را با مثال کاربردی می توان انجام داد:

- Descriptive Statistics, Univariate & Multivariate Analysis
- ANOVA, ANCOVA, MANOVA, MANCOVA, Repeated Measures, Chi-square
- Non Parametric Tests, Correlation, Graphics, Maximum Likelihood, Modeling
- Regression Analysis: Linear, Logistic, Conditional, Ordinal, Multinomial, Poisson
- Survival Analysis: Life Table, Kaplan Meier, Cox Regression
- Correlated Data Analysis: Robust Standard Error, GEE, Random Effect, Multi-Level Analysis
- Time Series, Diagnostic Tests Analysis, ROC Curve, Sensitivity Analysis
- Measures of Effect & Association in Epidemiologic Study: Cross Sectional, Case Control, Cohort, RCT, Error, Bias, Interaction, Measuring Inequality in Health
- Sample Size & Power Analysis, Factor Analysis
- Meta Analysis: meta, metan, metacum, metap, Random & Fixed effect



www.ElyasBook.com



9786005727210

جلد کتاب سال ۱۳۹۲ مولفین

# EPIDEMIOLOGY & RESEARCH with Stata



**By: Ahmadi A , lashkardost H**

- Descriptive Statistics, Univariate & Multivariate Analysis
- ANOVA, ANCOVA, MANOVA, MANCOVA, Repeated Measures, Chi-square
- Non Parametric Tests, Correlation, Graphics, Maximum Likelihood, Modeling
- Regression Analysis: Linear, Logistic, Conditional, Ordinal, Multinomial, Poisson, Cox,..
- Survival Analysis: Life Table, Kaplan Meier, Cox
- Correlated Data Analysis: Robust Standard Error, GEE, Random Effect, Multi- Level Analysis, Path Analysis, DDD model,
- Time Series, Diagnostic Tests Analysis, ROC
- Measure of Effect & Association in Epidemiologic Study: Cross Sectional, Case Control, Cohort, RCT, Error, Bias, Interaction, Measuring Inequality in Health
- Sample Size & Power Analysis, Factor Analysis
- Meta Analysis: meta