

۷ کارنیل، بزرگترین شبکه موفقیت ایرانیان می باشد، که افرادی زیادی توانسته اند با آن به موفقیت برسند، فاطمه رتبه ۱۱ کنکور کارشناسی، محمد حسین رتبه ۶۸ کنکور کارشناسی، سپیده رتبه ۳ کنکور ارشد، مریم و همسرش راه اندازی تولیدی مانتو، امیر راه اندازی فروشگاه اینترنتی، کیوان پیوستن به تیم تراکتور سازی تبریز، میلاد پیوستن به تیم صبا، مهسا تحصیل در ایتالیا، و.... این موارد گوشه از افرادی بودند که با کارنیل به موفقیت رسیده اند، شما هم می توانید موفقیت خود را با کارنیل شروع کنید.

برای پیوستن به تیم کارنیلی های موفق روی لینک زیر کلیک کنید.

[www.karnil.com](http://www.karnil.com)

همچنین برای ورود به کانال تلگرام کارنیل روی لینک زیر کلیک کنید.

<https://telegram.me/karnil>

## « جلسه اول – ۸۶/۱۱/۱۸ »

تعاریف اولیه :
**SPSS : Statistical Package For Social Science**

این نرم افزار شامل نوار عنوان ، فهرست منو ، نوار ابزار و نوار نشان دهنده مقادیر می باشد. هر ستون یک field است که به عنوان متغیر یا variable مطرح می باشد و هر سطر یک record است که به عنوان یک مورد یا case مطرح می باشد. به پنجره باز شده Datasheet می گویند. و کار برنامه SPSS ویرایش اطلاعات این صفحه است. هر Datasheet دو قسمت دارد:

۱- Data view که مقادیر را در آن وارد می کنیم

۲- Variable view که متغیرها را تعریف می کنیم.

در SPSS متغیرها خصوصیات دارند که مهمترین خاصیت آنها مقیاس آنهاست. و هر بار که متغیر تعریف می شود باید مقیاس آن را تعریف کنیم.

انواع مقیاس ها:
**۱. مقیاس اسمی (صوری) : Nominal**

مانند شهری که در آن زندگی می کنیم که میتوانیم به آن کد داده و تبدیل به مقیاس اسمی شود در این مقیاس کوچک و بزرگ بودن اعداد تفاوتی ندارد و هر طور می توان کد گذاری کرد و مناسب گروه بندی است و میانگین و جمع در آن معنی ندارد.

**۲. مقیاس رتبه ای ( ترتیبی ) : Ordinal**

کد گذاری بر اساس درجه اهمیت می باشد و اولویت یا ترتیب را نشان می دهد اگر کد ها جایجا شوند فرق می کند. مانند مهارت خوب ، بد و عالی. فاصله قابل اندازه گیری بین گروهها وجود ندارد.

**۳. مقیاس نسبی : Scale**

هر عددی که از طریق اندازه گیری بدست بیاید مقیاس نسبی دارد مثل قد افراد .

شاخص تمرکز :

نشانه تمرکز داده هاست و معروفترین آن میانگین (**Mean**) است که فقط برای مقیاس نسبی بکار می رود. بسیار به داده ها حساس است و اگر مقادیر بزرگ باشد به طرف آن می رود. ولی اگر داده ها خیلی پراکنده باشند میانگین گویا نخواهد بود.

شاخص دیگر میانه (**Median**) است که رتبه میانی داده ها ست و از روی رتبه بدست می آید اگر تعداد داده ها زوج باشد رتبه  $\frac{n+1}{2}$  میانه است و اگر فرد باشد وسطی میانه است. میانه به داده بزرگ

یا کوچک حساس نمی باشد.

شاخص دیگر مد (**Mode**) است که برای داده های اسمی یا Nominal هم بکار می رود میزان فراوانی داده ها را نشان می دهد و بیشترین مشاهدات را دارد.

### شاخص پراکندگی :

نشانه پراکندگی داده هاست و برای داده های نسبی یا Scale بکار میرود و معروفترین آن واریانس است. که  $f_i$  فراوانی هر داده است

$$\text{Mean} = \frac{\sum x_i}{n} = \frac{\sum x_i f_i}{\sum f_i} \quad \text{میانگین}$$

$$\text{variance} = \frac{\sum (x_i - \bar{x})^2}{n} \quad \sum (x_i - \bar{x}) = 0 \quad \text{واریانس}$$

و جذر واریانس انحراف معیار است که واحد آن با واحد داده ها یکی است اگر واحد داده ها کیلوگرم باشد انحراف معیار هم کیلوگرم است. واریانس و انحراف معیار شاخص های خوبی هستند چون داده ها در آن موثرند.

شاخص دیگر پراکندگی دامنه تغییرات یا Range می باشد.  $R = \text{max} - \text{min}$  شاخص دیگر چارک است که اگر داده ها خیلی پراکنده باشند آن استفاده می کنیم و میانبر چارکی :

$$R = \bar{Q}_3 - \bar{Q}_1$$

چارک سوم یعنی ۷۵% داده ها از آن کمتر و ۲۵% داده ها از آن بیشتر هستند و چارک دوم همان میانه است.

اگر میانه و میانگین داده ها به هم نزدیک باشند آن جامعه نرمال می باشد.

### مثال ۱ :

متغیر های نام کشور و درآمد ناخالص ملی را تعریف کرده و اطلاعات زیر را در صفحه وارد کنید سپس شاخص های مختلف آماری، توزیع و پراکندگی را محاسبه نمایید.

نام کشور	درآمد ناخالص ملی
ایران	۱۵
ترکیه	۲۰
افغانستان	۵
عراق	۱۰
پاکستان	۱۳

متغیر هارا در variable view تعریف می کنیم:

متغیر اول نام کشورها به نام country

در تایپ اسم متغیر نباید از فاصله استفاده شود ولی خط تیره اشکالی ندارد و فارسی تایپ نشود.

در Type یا نوع متغیر STRING انتخاب می کنیم.

در label میتوان توضیحات کافی در مورد متغیر داد. مثلاً، name of country،

در قسمت value چون متغیر اسمی تعریف شده چیزی نمی زنیم. و مخصوص کد بندی متغیر های گروهی است که عددی باشند.

در قسمت missing یعنی اعدادی معرفی شود که در محاسبات در نظر گرفته نمی شود مثلاً در پرسشنامه ها برخی گزینه ها ناخواناست یا پر نشده است ما عدد ۱۰۰۰ را معرفی میکنیم که هر وقت درج شد به معنی آن باشد و حساب نشود.

پهنای ستون را ۱۰ تعریف می کنیم.

Align را انتخاب میکنیم راست ، چپ یا وسط

Measure یا مقیاس متغیر را Nominal و اسمی انتخاب می کنیم.

متغیر دوم میزان درآمد ناخالص ملی یا GNP در Type یا نوع متغیر Numeric انتخاب می کنیم. که اختصاص به مقادیر عددی دارد. در label میتوان توضیحات کافی در مورد متغیر داد. مثلاً، Gross National Product، در قسمت value چیزی نمی زنیم. پهنای ستون را ۱۰ تعریف می کنیم. Align را انتخاب می کنیم راست ، چپ یا وسط Measure یا مقیاس متغیر را Scale و نسبی انتخاب می کنیم. Decimal را اگر ۲ بزنیم تا ۲ رقم ممیز می خورد.

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1	country	String	8	-	country	-	-	10	left	Nominal
2	GNP	Numeric	8	2	GNP	-	-	10	left	Scale

روش اول محاسبه شاخص ها :

### Analyze – Report –Case Summaries

پنجره ای باز می شود که متغیر مورد نظر جهت تنظیم شاخص های آماری را توسط فلش از پنجره موجود به پنجره سمت راست Variable انتقال می دهیم و سپس با زدن دکمه statistic شاخص های آماری مورد نظر را مانند میانه ، میانگین ، واریانس ، چولگی (skewness) ، کشیدگی (kurtosis) و . . انتخاب می کنیم.

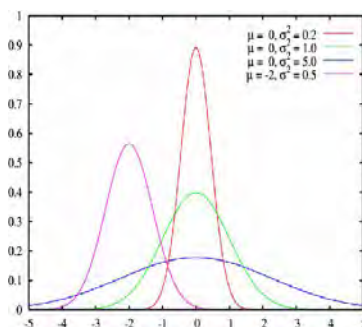
خروجی spss در پنجره جداگانه به نام output1 ظاهر می شود. بهتر است این پنجره را ببندید و هرچه اضافه کنید در همین پنجره اضافه خواهد شد.

روش دوم محاسبه شاخص ها :

### Analyze – Descriptive statistics - frequencies

می توانیم جدول فراوانی را نیز رسم کنیم. Percentile Values شامل Quartiles چارکها ، Cutpoints دهک ها و Percentile صدکها هستند. Central Tendency شاخص های تمرکز ، Dispersion معیار پراکندگی ، Distribution معیار توزیع می باشند.

اگر داده ها پراکنده و متفاوت باشند می توانیم چولگی و کشیدگی را حساب کنیم. اگر داده ها تکراری باشند و جدول فراوانی را بخواهیم باید گزینه Display frequency table را تیک بزنیم.



### نمواد نرمال ، کشیدگی

در حالت پخ  $kurtosis < 0$  و در حالت کشیده  $kurtosis > 0$  می باشد. در حالت چولگی به چپ  $skewness < 0$  و میانگین > میانه > مد است. در حالت چولگی به راست  $skewness > 0$  و مد > میانه > میانگین است.

مثال ۲ :

متغیر های نوع ماشین ، زمان و مکان را تعریف کرده و اطلاعات زیر را در صفحه وارد کنید سپس شاخص های مختلف آماری، توزیع و پراکندگی را محاسبه نمایید. نتیجه میزان تردد در کدام زمان و توسط چه نوع ماشینی را مشخص نمایید

نوع ماشین	مکان	زمان	میزان سوخت
شخصی	داخل طرح	۱۲	۶.۵
شخصی	خارج طرح	۱۳	۱۵
شخصی	داخل طرح	۱۴	۱۴
تاکسی	خارج طرح	۱۳	۱۴
تاکسی	داخل طرح	۱۴	۲۰
اتوبوس	خارج طرح	۱۵	۲۶
وانت	داخل طرح	۱۶	۳۰
اتوبوس	خارج طرح	۱۶	۳۰
تاکسی	داخل طرح	۱۷	۵۰
شخصی	خارج طرح	۱۵	۷
وانت	داخل طرح	۱۳	۸

انتخاب متغیر ها :

Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1 car	Numeric	8	0	car	1=ca 2=taxi 3=bus 4=vanet	-	10	left	Nominal
2 place	Numeric	8	0	place	0=in center 1=out center	-	10	left	Nominal
3 time	Numeric	8	2	time		-	10	left	Scale
4 fuel	Numeric	8	2	fuel		-	10	left	Scale

برای متغیر هایی که کد بندی شده اند شاخص های آماری مهم نیست و جدول فراوانی اهمیت دارد. اما در اینجا اگر missing تعریف می کردیم در جدول فراوانی بین دو ستون percent و valid percent تفاوت ایجاد می شد.

ستون Cumulative percent در جدول فراوانی نشان دهنده درصد تجمعی است که اگر داده ها Nominal باشد این ستون معنی ندارد.

جدول فراوانی سوخت چیزی نشان نمیدهد و تقریباً هر مورد یک بار تکرار شده است اما ما می خواهیم در فاصله های مختلف بسنجیم لذا متغیر جدیدی به نام Gfuel را معرفی می کنیم:

Transform-Record-Into Different Variable

نام و برجسب متغیر جدید را درج می کنیم و change را زده، سپس دگمه old and new value را می زنیم. در قسمت :

Old value : اگر می خواهیم به هر مقدار متغیر قدیم متغیر جدید نسبت دهیم و تناظر یک به یک برقرار کنیم گزینه value را تیک می زنیم و اگر می خواهیم به هر مقدار متغیر قدیم یک رنج جدید تعریف کنیم از گزینه Range استفاده می کنیم. که در اینجا:

- Range<10 1
- 10<Range<20 2
- 20<Range<30 3
- Range>30 4

و add را می زنیم. و سپس در قسمت تعریف متغیر Gfuel در ستون ششم در value تعریف می کنیم منظور از اعداد ۱،۲،۳ و ۴ چیست؟

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1	Gfuel	Numeric	8	0	Gfuel	1=low 2=middle 3=high 4=very high	-	10	left	Nominal

حال جدول فراوانی را برای متغیر Gfuel و شاخص های آماری را برای متغیر fuel محاسبه می کنیم تا معنی داشته باشد.

- سؤال ۱ : برای فر دگزارش گیرنده صفحه اصلی ورود اطلاعات مهم است یا نتایج و خروجی؟
- سؤال ۲: اگر جدول فراوانی را داشته باشیم چطور می توانیم داده هارا در صفحه اصلی درج کنیم؟
- سؤال ۳: طریقه رده بندی صحیح برای داده ها چیست؟ از نظر طول فاصله و از نظر تعداد گروهها؟

« جلسه دوم – ۲۵ / ۱۱ / ۸۶ »

جدول فراوانی :

متغیرها ( X )	فراوانی $f_i$	احتمال یا فراوانی نسبی $\frac{f_i}{\sum f_i}$	فراوانی تجمعی	فراوانی نسبی تجمعی
۱ = دیپلم				
۲ = فوق دیپلم				
۳ = لیسانس	.....	.....	۵۰	۰.۲
۴ = فوق لیسانس				
۵ = دکترا				

فراوانی نسبی نشان دهنده تعداد حالات مطلوب به کل حالات است که همان احتمال است. فراوانی تجمعی مجموع حالات قبل از خود می باشد و فراوانی نسبی تجمعی درصد آن لذا اگر طبق جدول بالا برای لیسانس مشخص می شود که ۵۰ نفر یا ۲۰% دارای تحصیلات زیر لیسانس هستند . ستون Invalid Percent مربوط به داده هایی است که missing وارد شده است

مثال ۲ :

لیست تحصیلات کارکنان یک اداره به شرح زیر است که این جدول فراوانی می باشد حال جدول داده اصلی را ایجاد کنید.

کد مدرک تحصیلی	تعداد
۱ = دیپلم	۱۵
۲ = فوق دیپلم	۱۷
۳ = لیسانس	۱۰
۴ = فوق لیسانس	۵
۵ = دکترا	۳

در اینجا ما یک متغیر داریم و آن مدرک تحصیلی است اما مجبوریم متغیری به نام فراوانی را تعریف کنیم تا اطلاعات را در جدول وارد کنیم:

انتخاب متغیرها :

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1	madrak	Numeric	8	-	-	-	-	10	left	Scale
2	frequency	Numeric	8	-	-	-	-	10	left	Scale

اگر الان برای متغیر madrak جدول فراوانی تشکیل دهیم همه را یک می زند چون هر کدام یکبار تکرار شده است در صورتی که ما ۱۵ نفر دیپلم داریم پس باید متغیر وزن دار تعریف کنیم : وزن را به متغیر فراوانی نسبت می دهیم. ولی جدول فراوانی را بر اساس متغیر madrak می گیریم و نتیجه درست نشان داده می شود.

Data – Weight Case

هر زمان که این صفحه را باز کنیم با نشانه Weight On در سمت راست پایین صفحه می فهمیم که داده ها وزن دارند.

مثال ۴ :

سه متغیر جنسیت، سن و وزن را داریم برای متغیر سن و وزن میانگین و برای متغیر جنسیت مد را حساب کنید.

انتخاب متغیرها :

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1	Gender	Numeric	8	-	-	1=male 2=female	-	10	left	Nominal
2	Age	Numeric	8	-	-	-	-	10	left	Scale
3	Weight	Numeric	8	-	-	-	-	10	left	Scale

پس از مشخص شدن جدول فراوانی مشخص می شود که هر داده یکبار تکرار شده است ولی جدول فراوانی خوب است که حداکثر ۲۰٪ داده های آن کمتر از ۵ باشد . برای رفع این مشکل باید داده ها را رده بندی کنیم تا فراوانی بیشتر شود.

روش اول رده بندی :

$$\frac{Max - Min}{5} = \frac{50 - 20}{5} = 6$$

برای رده بندی متغیر سن از این فرمول استفاده می کنیم:

تعداد رده را ۵ انتخاب می کنیم و طول بازه ۶ به دست می آید. برای ایجاد رده بندی از گزینه Transform – recode- into different variable

متغیر جدید را به نام Age group تعریف می کنیم :

۲۰-۲۶    ۲۷-۳۳    ۳۴-۴۰    ۴۱-۴۷    ۴۸-۵۴

\* نکته : برای محاسبه جدول فراوانی از متغیر جدید Age group که رده بندی شده است استفاده می کنیم ولی برای محاسبه بقیه شاخص های آماری مثل میانگین از متغیر اولی یعنی Age استفاده خواهیم کرد زیرا میانگین متغیر رده بندی شده معنی ندارد مگر برای هر رده نماینده تعیین کنیم.

رده بندی	نماینده رده	فراوانی
۲۰ - ۲۶	$\frac{20 + 26}{2} = 23$	
۲۷ - ۳۳	$\frac{27 + 33}{2} = 30$	

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

و طبق فرمول میانگین،  $x_i$  نماینده رده  $i$  ام است.

با تعریف متغیر جدید Age group در قسمت Variable view آن را در قسمت value مشخص می کنیم:

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1	Gender	Numeric	8	-	-	1=male 2=female	-	10	left	Nominal
2	Age	Numeric	8	-	-	-	-	10	left	Scale
3	Weight	Numeric	8	-	-	-	-	10	left	Scale
4	Agegroup	Numeric	8	-	grouping variable for age	1=20-26 2=27-33 3=34-40 4=41-47 5=48-54	-	10	left	Scale



برای متغیر جدید جدول فراوانی رسم می کنیم که از جدول قبلی بهتر خواهد شد. جدول فراوانی به رده های ما بستگی دارد میتوان رده ها را کمتر و طول رده را بزرگتر کرد تا جدول فراوانی بهتری بدست آورد.

grouping variable for age

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 20-26	1	10.0	10.0	10.0
27-33	5	50.0	50.0	60.0
34-40	2	20.0	20.0	80.0
41-47	1	10.0	10.0	90.0
48-54	1	10.0	10.0	100.0
Total	10	100.0	100.0	

### روش دوم رده بندی :

Transform - visual Bander یا visual Binning

دقت داشته باشید که رده بندی برای متغیر Nominal انجام نمی شود. در پنجره باز شده متغیر weight را انتخاب می کنیم و continue میزنیم در پنجره جدید visual Binning اسم و برجسب متغیر جدید را وارد می کنیم gweight حال با زدن دکمه Make cutpoints نقاط و محل شکست را می سازیم.



سه راه برای تعیین نقاط شکست وجود دارد:

#### 1- Equal Width Intervals

نقطه اول را min ما باشد و در اینجا ۴۰ است می دهیم که بهتر است برای پیوستگی عدد ۳۹.۵ را وارد کنیم بعدی تعداد رده ها ست که ۳ وارد می کنیم و سومی طول رده است که خودش می نویسد اگر طول رده را بدهید تعداد رده را خودش می دهد و آخرین عدد رده را در قسمت Last cutpoint location را خودش نشان میدهد.

#### 2- Equal Percentiles Based on Scanned Cases

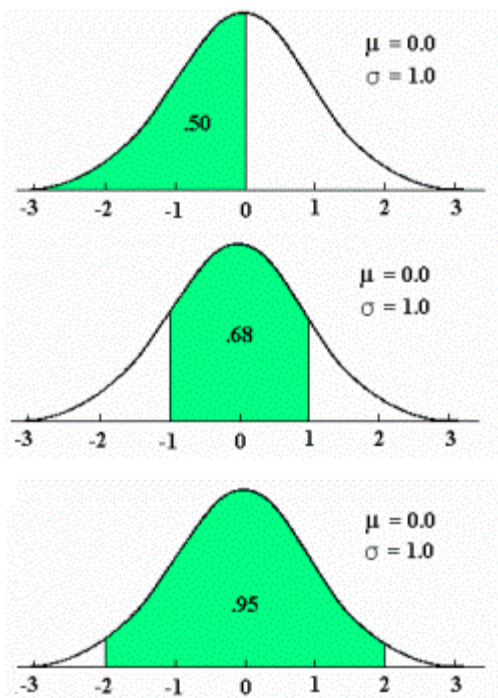
تعداد رده و درصدی که هر رده از فاصله max-min میگیرد را می دهیم.

#### 3- Cutpoints at Mean and Selected Standard Deviation Based on Scanned Cases

بر اساس +/- یک، دو یا سه انحراف معیار بازه ها را تعیین میکند.

نکته: به هر حال از هر کدام این سه گزینه که استفاده شود تعداد n نقطه شکست تعداد n+1 رده را خواهد داشت.

با زدن دکمه Make Label در همان پنجره رده ها هم ظاهر می شوند.



ارتفاع این منحنی با مقادیر میانگین  $\mu$  (و انحراف معیار)  $\sigma$  ارتباط دارد.  
 $\mu$  میزان تراکم داده ها حول یک مقدار و  $\sigma$  (انحراف معیار) میزان پراکندگی داده ها از میانگین است.

به عنوان ی مثال در یک امتحان درسی نمرات دانش آموزان اغلب اطراف میانگین بیشتر می باشد و هر چه به سمت نمرات بالا یا پایین پیش برویم تعداد افرادی که این نمرات را گرفته اند کمتر می شود. این رفتار را بسهولت می توان با یک توزیع نرمال مدل کرد.

**روش ایجاد متغیر جدید و انجام محاسبات مشروط بر روی متغیر جدید:**

با استفاده از روش Transform – recode- into different variable متغیر جدیدی را تعریف می کردیم ولی حال می خواهیم با روش محاسبه متغیر جدیدی را بدست بیاوریم:

**Transform – Compute**

در قسمت Target Variabele نام متغیر جدید را درج می کنیم و فرمول را در قسمت Numeric Expression می نویسیم مثلاً اگر می خواهیم متغیر Weight را در ستون دیگری بر حسب گرم حساب کنیم می توانی این فرمول را بنویسیم:  $Weight \times 1000$  یا تابعی خاص را انتخاب کنیم.

نکته: فرق فرمول نویسی در SPSS با Excel این است که در SPSS با تغییر داده فرمول اجرا نمی شود و نتیجه به روز نمی گردد و می بایست مجدداً فرمول اجرا شود ولی در Excel هر گاه داده را تغییر دهیم نتیجه نیز تغییر می کند.

اگر بخواهیم این فرمول تحت شرایط خاصی و فقط برای داده های خاصی انجام شود مثلاً تغییر واحد وزن به گرم فقط برای خانمها انجام شود باید از دکمه if پایین پنجره Compute استفاده کنیم مثلاً: شرط بگذاریم که  $Gender = 1$  یا  $Gender = 1 \& Age > 30$

راه دیگر اینکه می توانیم یک متغیر به نام select تعریف کنیم برای داده های خاصی ۱ بگذاریم و برای داده های دیگر ۰ بعد فرمول را برای یکها حساب کنیم یا صفرها.

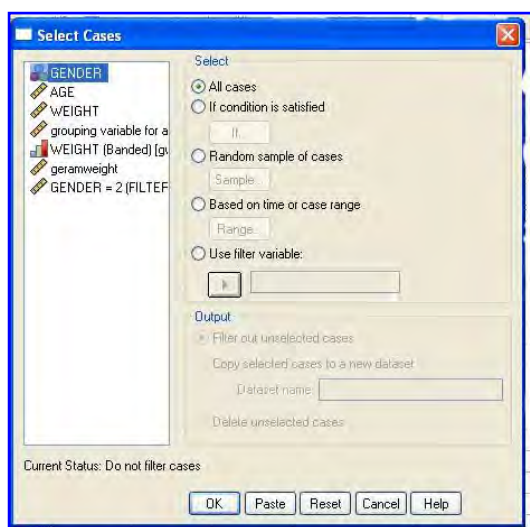
**روش ثبت انواع دستورات اجرا شده در SPSS :**

با استفاده از دکمه paste در هر پنجره ای از دستورات ، پنجره syntax باز شده و فرمان ما درج می شود حتی با بازکردن مجدد این پنجره می توان با دستور Run فرامین را مجدداً اجرا کرد. بهتر است موقع باز کردن صفحه spss صفحه syntax هم باز باشد تا دستورات در آن درج شود.

**تفکیک داده ها و بدست آوردن خروجی های مجزا :**

فرق این راه با راه قبلی اینست که با استفاده از پنجره Compute متغیر جدید در صفحه Data view ایجاد میشود ولی با این راه در خروجی و output تغییرات حاصل می شود. و این یک نوع فیلتر کردن است.

**روش اول :**



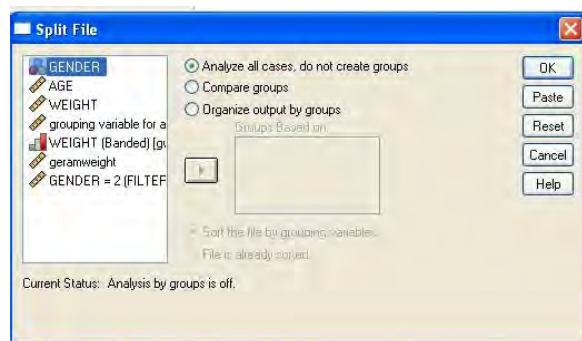
**Data – Select Cases**

پنجره روبرو باز می شود که اگر All Cases فعال باشد به معنای عدم فیلتر است و با فعال کردن if condition is satisfied شرط مورد نظر را درج می کنیم. مثلاً شرط می گذاریم فقط مردها را محاسبه کند که اگر حال با انجام فیلتر اگر میانگین بگیریم فقط برای مردها محاسبه میکند و در Data view هم رکورد های زنهارا خط می کشد و ستونی ایجاد می کند که در آن انتخاب شدن برای محاسبات را نشان می دهد.

با انتخاب Random of cases یک نمونه تصادفی انتخاب می کند. با انتخاب متغیر مورد فیلتر در use filter available و باززدن ok فیلتر انجام می شود و در پایین صفحه نیز filter on نمایش داده خواهد شد.

**روش دوم :**

اول فیلتر های قبلی را با روش Data – Select Cases او زدن گزینه All Cases بر می داریم بعد:  
**Data – Split Files**



گزینه اول عدم فیلتر است. گزینه دوم جداول را در حالت مقایسه ای در خروجی نشان می دهد .  
گزینه سوم کاملاً جداگانه نشان داده خواهد شد. که متغیر gender را به دو قسمت مردها و زنها نشان خواهد داد.

مثال 5 : فایل Demo را از مسیر زیر باز کنید:  
 My computer – win(programming) - program files - spssEVAL – Tutorial – sample files- Demo  
 اطلاعات لازم را در مورد تعداد case ها ، تعداد متغیر ها به دست آورید. سپس میانگین میزان درآمد خانم ها و آقایان را جداگانه حساب کنید و جدول فراوانی درآمد را بر اساس نوع ماشین بدست بیاورید. جدول فراوانی بر اساس درآمد رده بندی شده بدست بیاورید.

اخذ اطلاعات از یک فایل :

روش اول : File – display data files information  
 روش دوم : Analyze-Report – Case Summaries

الف- برای میانگین میزان درآمد مردها و زنها به شکل جداگانه :

Data – Split Files – gender- organize output by groups- ok  
 سپس جدول فراوانی برای متغیر درآمد یا inccat حساب می کنیم. و لی برای محاسبه میانگین فراوانی را بر حسب متغیر income در قسمت statistics گزینه Mean را تیک می زنیم و سپس ok . میانگین برای متغیر رده بندی شده محاسبه نمی شود.

**Gender = Female  
Statistics(a)**

Household income in thousands		
N	Valid	3179
	Missing	0
Mean		68.7798

a Gender = Female

**Gender = Male  
Statistics(a)**

Household income in thousands		
N	Valid	3221
	Missing	0
Mean		70.1608

a Gender = Male

**Income category in thousands(a)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under \$25	563	17.7	17.7	17.7
	\$25 - \$49	1208	38.0	38.0	55.7
	\$50 - \$74	572	18.0	18.0	73.7
	\$75+	836	26.3	26.3	100.0
	Total	3179	100.0	100.0	

a Gender = Female

**Income category in thousands(a)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under \$25	611	19.0	19.0	19.0
	\$25 - \$49	1180	36.6	36.6	55.6
	\$50 - \$74	548	17.0	17.0	72.6
	\$75+	882	27.4	27.4	100.0
	Total	3221	100.0	100.0	

a Gender = Male

ب- جدول فراوانی بر حسب نوع ماشین:  
 اولاً باید انواع ماشین را فیلتر کنیم پس متغیر carcat که انواع ماشین را نشان میدهد توسط این دستور  
 Data – Split Files –carcat - organize output by groups- ok  
 فیلتر می کنیم. بعد دستور فراوانی متغیر درآمد رده بندی شده یا inccat را حساب می کنیم.

Income category in thousands(a) - a Primary vehicle price category = Economy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under \$25	1174	63.8	63.8	63.8
	\$25 - \$49	667	36.2	36.2	100.0
	Total	1841	100.0	100.0	

Income category in thousands(a) - a Primary vehicle price category = Standard

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	\$25 - \$49	1721	75.6	75.6	75.6
	\$50 - \$74	554	24.4	24.4	100.0
	Total	2275	100.0	100.0	

Income category in thousands(a) - a Primary vehicle price category = Luxury

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	\$50 - \$74	566	24.8	24.8	24.8
	\$75+	1718	75.2	75.2	100.0
	Total	2284	100.0	100.0	

محاسبه فراوانی درآمد رده بندی شده یا متغیر inccat براساس نوع ماشین یا متغیر carcat در حالت split یا فیلتر شده:

#### Income category in thousands

Primary vehicle price category			Frequency	Percent	Valid Percent	Cumulative Percent
Economy	Valid	Under \$25	1174	63.8	63.8	63.8
		\$25 - \$49	667	36.2	36.2	100.0
		Total	1841	100.0	100.0	
Standard	Valid	\$25 - \$49	1721	75.6	75.6	75.6
		\$50 - \$74	554	24.4	24.4	100.0
		Total	2275	100.0	100.0	
Luxury	Valid	\$50 - \$74	566	24.8	24.8	24.8
		\$75+	1718	75.2	75.2	100.0
		Total	2284	100.0	100.0	

نتایج:

۱- ۶۳.۸٪ که نوع ماشین مدل Economy دارند، درآمد زیر ۲۵۰۰۰ دلار دارند.

۲- ۷۵.۲٪ که از نوع ماشین Luxury استفاده می کنند درآمد بالای ۷۵۰۰۰ دلار دارند

محاسبه میانگین درآمد رده بندی شده یا متغیر inccat براساس نوع تاهل یا متغیر martial در حالت split یا فیلتر شده:

Income category in thousands

Marital status			Frequency	Percent	Valid Percent	Cumulative Percent
Unmarried	Valid	Under \$25	578	17.9	17.9	17.9
		\$25 - \$49	1228	38.1	38.1	56.0
		\$50 - \$74	552	17.1	17.1	73.1
		\$75+	866	26.9	26.9	100.0
		Total	3224	100.0	100.0	
Married	Valid	Under \$25	596	18.8	18.8	18.8
		\$25 - \$49	1160	36.5	36.5	55.3
		\$50 - \$74	568	17.9	17.9	73.2
		\$75+	852	26.8	26.8	100.0
		Total	3176	100.0	100.0	

نتایج :

- ۱ - ۳۸.۱٪ که مجرد هستند درآمد بین ۲۵۰۰۰-۴۹۰۰۰ دلار دارند.
- ۲ - ۲۶.۸٪ که متاهل هستند درآمد بین بالای ۷۵۰۰۰ دلار دارند.

رده بندی متغیر سن و ایجاد متغیر جدید agegroup :

Transform-Visual binning – cutpoint 17.5

محاسبه میانگین درآمد رده بندی شده یا متغیر inccat براساس رده سنی یا متغیر agegroup در حالت split یا فیلتر شده:

Income category in thousands

Age in years (Binned)			Frequency	Percent	Valid Percent	Cumulative Percent
19 - 28	Valid	Under \$25	410	52.3	52.3	52.3
		\$25 - \$49	319	40.7	40.7	93.0
		\$50 - \$74	41	5.2	5.2	98.2
		\$75+	14	1.8	1.8	100.0
		Total	784	100.0	100.0	
29 - 38	Valid	Under \$25	331	19.1	19.1	19.1
		\$25 - \$49	929	53.5	53.5	72.6
		\$50 - \$74	293	16.9	16.9	89.5
		\$75+	182	10.5	10.5	100.0
		Total	1735	100.0	100.0	
39 - 48	Valid	Under \$25	124	7.0	7.0	7.0
		\$25 - \$49	641	36.0	36.0	43.0
		\$50 - \$74	453	25.4	25.4	68.4
		\$75+	563	31.6	31.6	100.0
		Total	1781	100.0	100.0	
49 - 58	Valid	Under \$25	73	5.6	5.6	5.6
		\$25 - \$49	341	26.0	26.0	31.6
		\$50 - \$74	247	18.8	18.8	50.4
		\$75+	650	49.6	49.6	100.0
		Total	1311	100.0	100.0	
59 - 68	Valid	Under \$25	169	26.0	26.0	26.0
		\$25 - \$49	135	20.8	20.8	46.8
		\$50 - \$74	75	11.5	11.5	58.3
		\$75+	271	41.7	41.7	100.0
		Total	650	100.0	100.0	
69+	Valid	Under \$25	67	48.2	48.2	48.2
		\$25 - \$49	23	16.5	16.5	64.7
		\$50 - \$74	11	7.9	7.9	72.7
		\$75+	38	27.3	27.3	100.0
		Total	139	100.0	100.0	

نتایج :

۱- در رده سنی ۱۹-۲۸ ، ۵۲.۳٪ زیر ۲۵۰۰۰ دلار درآمد دارند در حالیکه در رده سنی ۴۸-۵۸ ، نزدیک ۵۰٪ درآمد بالای ۷۵۰۰۰ دلار دارند..

## « جلسه سوم - ۲ / ۱۲ / ۸۶ »

## اقسام نمونه :

۱. نمونه گیری تصادفی ساده : شامل  $n$  عنصر از یک جامعه است که :
  - ویژگیها: شانس مساوی برای انتخاب هر عنصر دارد.
  - شانس مساوی برای انتخاب هر نمونه  $n$  تایی دارد. ( آزمایش تصادفی)
  - موارد استفاده: جامعه شماره گذاری شده یا فابل شماره گذاری با هزینه کم ( با استفاده از جدول اعداد تصادفی)
  - دارای قابلیت دسترسی به سرعت و هزینه کم

۲. نمونه گیری تصادفی سیستماتیک : نمونه ای است که هر  $I$  مین عنصر جامعه را شامل می شود یعنی شامل هر عنصر از جامعه که معلوم و غیر صفر است. برای جوامع ترتیب پذیر خیلی مناسب است و در این نمونه گیری هر نمونه با دیگری کاملاً تفاوت دارد. مثلاً میخواهیم بدانیم چه تعدادی از واگن های قطار مسافر بری رجا بعد از مسافتی خراب می شوند. در نمونه گیری اول واگن های ۱۵۰ و ۱۵۱ و ... را انتخاب می کنیم و در نمونه گیری دوم واگن های دیگری با تفاوت عدد مثلاً ۵.

در این روش جامعه را از ۱ تا  $N$  شماره گذاری کرده بعد  $k = \frac{N}{n}$  فاصله نمونه گیری مشخص شده سپس یک عدد به تصادف از ۱ تا  $k$  انتخاب شده که اولین واحد نمونه ماست و با افزودن  $k$  به اولین واحد ، سایر واحدها هم بدست می آید.

۳. نمونه گیری تصادفی طبقه بندی شده: تقسیم یک جامعه آماری به طبقه های متمایز و انتخاب یک نمونه تصادفی از هر طبقه  
مزایا:

- اطمینان از وجود عناصر کافی در هر طبقه
  - بدست آوردن تخمین های بهتر از پارامترهای جامعه
- مثلاً اگر متوسط قد انسانهای یک جامعه را بگیریم یک نتیجه ای بدست می آید ولی اگر انسانها را به دو طبقه زن و مرد تقسیم کنیم و متوسط قد را در نمونه گیری طبقه بندی شده محاسبه کنیم نتیجه بهتری بدست می آید.
- اصل کلی: در صورتی که اختلاف عناصر بین طبقات بیشتر از اختلاف عناصر در داخل هر طبقه باشد این روش نسبت به روش نمونه گیری تصادفی ساده نتایج دقیقتری را می دهد.
- قضیه حد مرکزی : در هر نمونه گیری تصادفی اگر اندازه نمونه بزرگ شود توزیع آن به سمت نرمال میل خواهد کرد.

۴. نمونه گیری تصادفی خوشه ای: انتخاب خوشه هایی از عناصر جامعه است که شرط همگن بودن در آن باید رعایت شود.

جامعه را به قسمت هایی که اصطلاحاً خوشه نامیده می شود و مبنای تشکیل خوشه ها در اغلب اوقات، تقسیم بندی های جغرافیایی از قبیل استان، شهرستان، دهستان، شهر و بلوک یا آبادی است، تقسیم بندی می کنند. خوشه ها باید کل جامعه را پوشش دهند و فاقد هم پوشانی باشند. به بیان دیگر هر کدام از عناصر جامعه باید به یکی و فقط یکی از این خوشه ها تعلق داشته باشد.

بسته ی نرم افزاری SPSS طرح نمونه گیری پیچیده ، نمونه گیری خوشه ای یك یا چند مرحله ای، نمونه گیری طبقه بندی، را با Complex Samples فراهم می کند.

سؤال : روش های یافتن اندازه نمونه را توضیح دهید و اینکه به چه عواملی بستگی دارد؟



**جداول توافقی Crosstabs :**

جداول توافقی برای نشان داده ارتباط بین متغیرهای اسمی بکار می رود جدول توافقی زیر را در صفحه نرم افزار وارد کنید :

	زن	مرد	
بیکار	۱۵	۵	۲۰
شاغل	۱۰	۲۰	۳۰
	۲۵	۲۵	

دو متغیر جنسیت و شغل را انتخاب می کنیم:  
 که متغیر تعداد را نیز برای وزن دادن تعریف می کنیم  
 بعد در صفحه ورود داده ها هر حالت را فقط یکبار انتخاب می کنیم :  
 Male – jobless  
 Male – employee  
 Female –jobless  
 Female – employee

	Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
۱	Gender	Numeric	8	0	-	1=male 2=female	-	10	left	Nominal
۲	job	Numeric		0	-	1=jobless 2=employee	-	10	left	Nominal
3	frequency	Numeric	8	0			-	10	left	Scale

سپس با استفاده از منوی Data – Weightcases – frequency به داده ها وزن می دهیم. جهت تشکیل جدول توافقی از منو :

Analyze –Descriptive Statistics- Crosstabs

در قسمت سطر و ستون پنجره باز شده باید دو متغیر Nominal یا اسمی بگذاریم که جدول زیر در output ظاهر می شود:

job \* gender Crosstabulation

Count

		gender		Total
		male	female	male
job	jobless	5	15	20
	employee	20	10	30
Total		25	25	50

با زدن دکمه statistics در پنجره crosstabs میتوان از آزمونی به نام خی دو که فرضیه استقلال دو متغیر اسمی را آزمون میکند، استفاده نمود .

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.333(b)	1	.004		
Continuity Correction(a)	6.750	1	.009		
Likelihood Ratio	8.630	1	.003		
Fisher's Exact Test				.009	.004
Linear-by-Linear Association	8.167	1	.004		
N of Valid Cases	50				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.00.

ما دنبال ستون (Asymp. Sig. (2-sided) هستیم که سطح معنی داری را برای استقلال دو متغیر نشان می دهد در اینجا بسیار کم است که به معنی وابستگی دو متغیر می باشد. آزمون خی دو که آزمون استقلال دو متغیر اسمی می باشد توسط این فرمول حساب می شود:

$O_i$  = مشاهده شده

$E_i$  = مورد انتظار

$$\sum \frac{(O_i - E_i)^2}{E_i} > \chi^2 \rightarrow \text{وابسته اند}$$

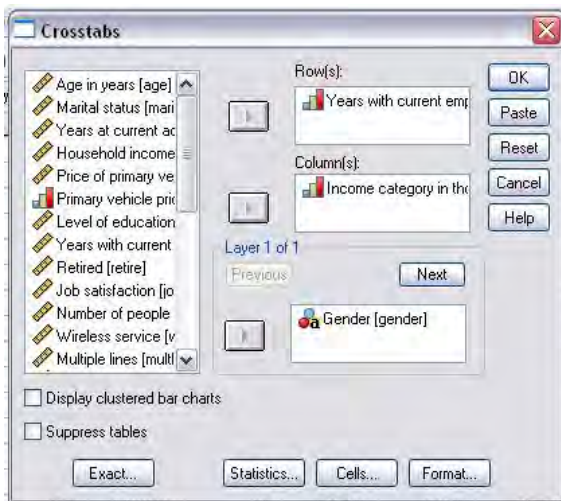
$$\sum \frac{(O_i - E_i)^2}{E_i} < \chi^2 \rightarrow \text{مستقل اند}$$

**سؤال :** تحقیق کنید بین جنسیت و نوع اسباب بازی ارتباطی وجود دارد یا نه؟

تغییر سطر و ستون در خروجی:

در جدول خروجی اگر بخواهیم جای سطر و ستون را عوض کنیم کافی است دو بار روی آن کلیک کرده و با استفاده از نوار ابزار formatting toolbar و پنجره pivoting tray جای سطر و ستون را عوض کرد. همین جدول را میتوان با روش های قبلی و استفاده از split files نیز کشید.

ایجاد یک لایه متغیر Layer:



میتوان از متغیر سومی در جداول توافقی استفاده نمود یعنی یک لایه متغیر را اضافه کنیم تا یک جدول سه راهه بسازیم و نشان می دهد که وقتی تاثیرات متغیر سوم را کنترل می کنیم چگونه رابطه بین متغیرهای سطر و ستون تغییر می یابد.

فایل Demo را باز کنید و میزان سوابق شغلی empcat و درآمد incat را بدست آورید. و متغیر جنسیت gender را به عنوان متغیر سوم وارد کنید.

اگر در پنجره crosstabs دکمه cell را زده و درصد سطر را فعال کنیم اعداد را با درصد نیز نشان می دهد.

Years with current employer \* Income category in thousands \* Gender Crosstabulation

Count			Income category in thousands				Total
Gender			Under \$25	\$25 - \$49	\$50 - \$74	\$75+	Under \$25
Female	Years with current employer	Less than 5	369	557	130	61	1117
		5 to 15	148	553	277	211	1189
		More than 15	46	98	165	564	873
	Total		563	1208	572	836	3179
Male	Years with current employer	Less than 5	409	516	107	67	1099
		5 to 15	131	562	274	208	1175
		More than 15	71	102	167	607	947
	Total		611	1180	548	882	3221

نتیجه : در گروه زنها و مردها بالاترین تعداد مربوط به سابقه شغلی بیشتر از ۱۵ سال و در آمد بالای ۷۵۰۰۰ دلار است. در گروه زنها و مردها پایین ترین تعداد مربوط به سابقه شغلی کمتر از ۵ سال و در آمد بالای ۷۵۰۰۰ دلار است.

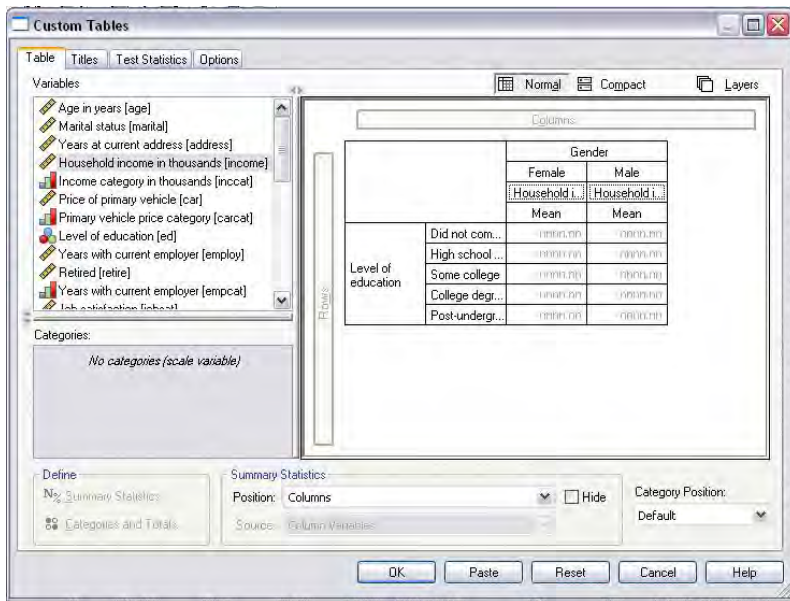
با آزمون خی دو متوجه می شویم که عدد ستون (Asymp. Sig. (2-sided) ، 0.000 است یعنی استقلال بین دو متغیر سوابق شغلی و درآمد وجود ندارد و ایندو وابسته اند.  
**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2506.010(a)	6	.000
Likelihood Ratio	2527.590	6	.000
Linear-by-Linear Association	2012.691	1	.000
N of Valid Cases	6400		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 318.50.

تمرین : در فایل Demo جدولی رسم کنید که ارتباط بین جنسیت و میزان تحصیلات را نشان دهد ولی

به جای تعداد میزان متوسط درآمد آنها در جدول درج کند.  
 در اینجا نمی توان از جدول توافقی استفاده نمود بلکه باید جدول سفارشی ایجاد کرد:



Analyze – Tables – Custom Table  
 در ستون gender جنسیت و در سطر سطح تحصیلات یا edu و متغیر درآمد income که حتماً باید scale باشد در قسمت count می گذاریم .

نکته بسیار مهم :متغیر سطر و ستون حتماً باید متغیر رده بندی شده باشد اگر نبود با کلیک راست آنرا به nominal تبدیل می کنیم. و متغیر جهت محاسبه حتماً باید scale باشد.

		Gender	
		Female	Male
		Household income in thousands	Household income in thousands
		Mean	Mean
Level of education	Did not complete high school	59.20	60.49
	High school degree	65.12	67.27
	Some college	71.63	68.60
	College degree	76.27	81.02
	Post-undergraduate degree	84.81	89.62

می توان به جای میانگین شاخص های آماری دیگری را با استفاده از summary statistics در همان پنجره محاسبه نمود.

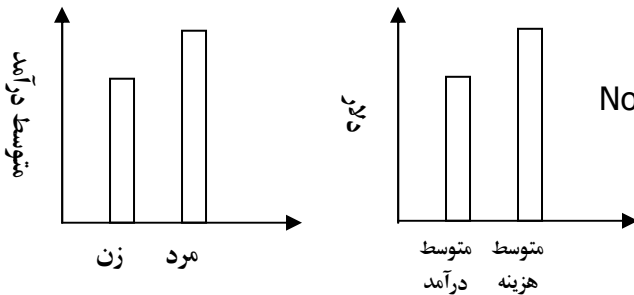
تمرین : در فایل Demo جدولی رسم کنید که ارتباط بین جنسیت gender و انواع ماشین یا carcat Primary vehicle price category را نشان دهد ولی به جای تعداد میزان متوسط درآمد آنها در جدول درج کند.

Analyze – Tables – Custom Table

در ستون gender جنسیت و در سطر انواع ماشین یا carcat و متغیر درآمد income که حتماً باید scale باشد در قسمت count می گذاریم . و با زدن دگمه summary statistics شاخص های میانگین، ماکزیمم، مینیمم و مد را هم انتخاب می کنیم.

		Gender							
		Female				Male			
		Household income in thousands				Household income in thousands			
		Mean	Maximum	Minimum	Mode	Mean	Maximum	Minimum	Mode
Primary vehicle price category	Economy	22.17	31.00	9.00	25.00	21.62	31.00	9.00	23.00
	Standard	42.45	61.00	29.00	34.00	42.67	61.00	29.00	31.00
	Luxury	132.60	1070.00	59.00	65.00	136.65	1116.00	58.00	62.00

نتیجه : خانمها با بالاترین درآمد ۱۰۷۰ هزار دلار بهترین ماشین لوکس را می خردند ولی آقایان با بالاترین درآمد ۱۱۱۶ هزار دلار ماشین لوکس می خردند میانگین درآمد خانم هایی که ماشین لوکس می خردند کمتر از آقایان است.

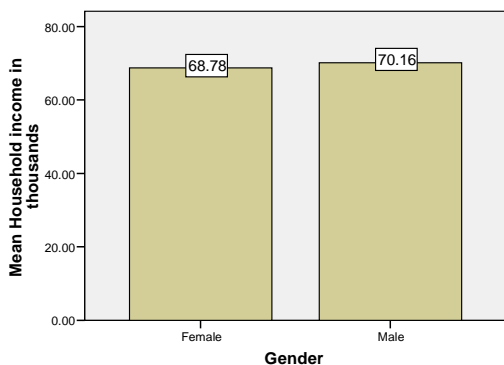


انواع نمودار:

۱. مقایسه یک متغیر scale در بین گروههای مختلف Nominal
۲. مقایسه دو متغیر scale یا بیشتر با هم
۳. مقایسه یک متغیر در بین تمام case ها

تمرین : این نمودارها را درست کنید:

ساخت نمودار نوع اول:



Graph - Chart builder

۱. متغیر Gender در محور x ها
۲. متغیر income در محور y ها
۳. در دگمه Groups/point id گزینه Clustering variable on X را از فعال بودن خارج کنید.
۴. یا

Graph – Legacy dialogs – bar – variable ..... mean Household income in thousands / category Axis ..... gender

ساخت جدول نوع دوم:

Data – split files - gender

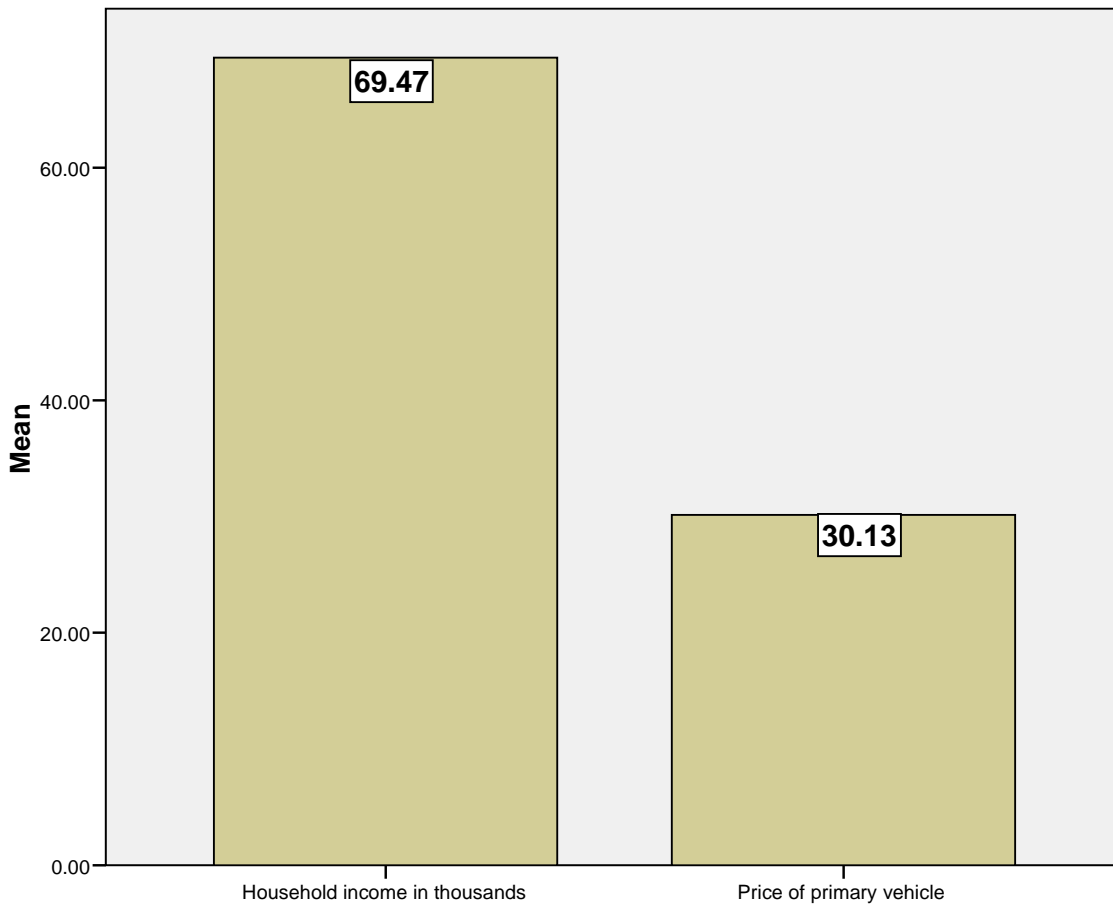
Analyze – Descriptive statistics – Descriptive - Household income in thousands , Price of primary vehicle

		Household income in thousands Mean	Price of primary vehicle Mean
Gender	Female	68.78	29.95
	Male	70.16	30.31

ساخت نمودار دوم:

Graph – Legacy dialogs – bar – simple – summaries of seprate variables - variable ..... mean

Household income in thousands / Price of primary vehicle- ok



« جلسه چهارم – ۸۶/۱۲/۱۶ »

پاسخ به سئوالات جلسه قبل :

۱. چگونه میتوان در فایل Demo متغیرهایی که پاسخ یکسان دارند و چند گزینه ای هستند را در یک جدول جمع آوری کرد؟

Analyze – Tables – Custom Tables

روش اول : سپس کلیه متغیرهای با پاسخ یکسان که فقط جواب بله یا خیر دارند را ابتدا با کلیک راست به نوع category تبدیل کرده و در قسمت سطر جدول قراردادده به حالت stacking سپس از سمت راست پایین پنجره قسمت category position گزینه row labels in columns را انتخاب می کنیم.

	No	Yes
	Count	Count
Wireless service	3853	2547
Multiple lines	3709	2691
Voice mail	3645	2755
Paging service	4819	1581
Internet	4509	1636

روش دوم :

Analyze – Tables – Tables of frequency

همه متغیرها را در پنجره ببرید و ok رابزنید.

۲. اندازه نمونه را چگونه محاسبه کنیم؟

**الف – برای متغیر کمی:**

\* در جامعه نامحدود:

$$z = \frac{x - \bar{x}}{\sigma_{\bar{x}}}$$

$$\sqrt{nd} = z \cdot \sigma$$

$$nd^2 = z^2 \sigma^2$$

$$x - \bar{x} = d$$

$$n = \frac{z^2 \sigma^2}{d^2}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$z = \frac{d}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{nd}}{\sigma}$$

\* در جامعه محدود : ( که N تعداد عناصر جامعه مشخص است )

$$n = \frac{Nz^2\sigma^2}{(N-1)d^2 + z^2\sigma^2}$$

برای  $z = 1.64$  ،  $\alpha = 0.05$

**ب- متغیر کیفی :**

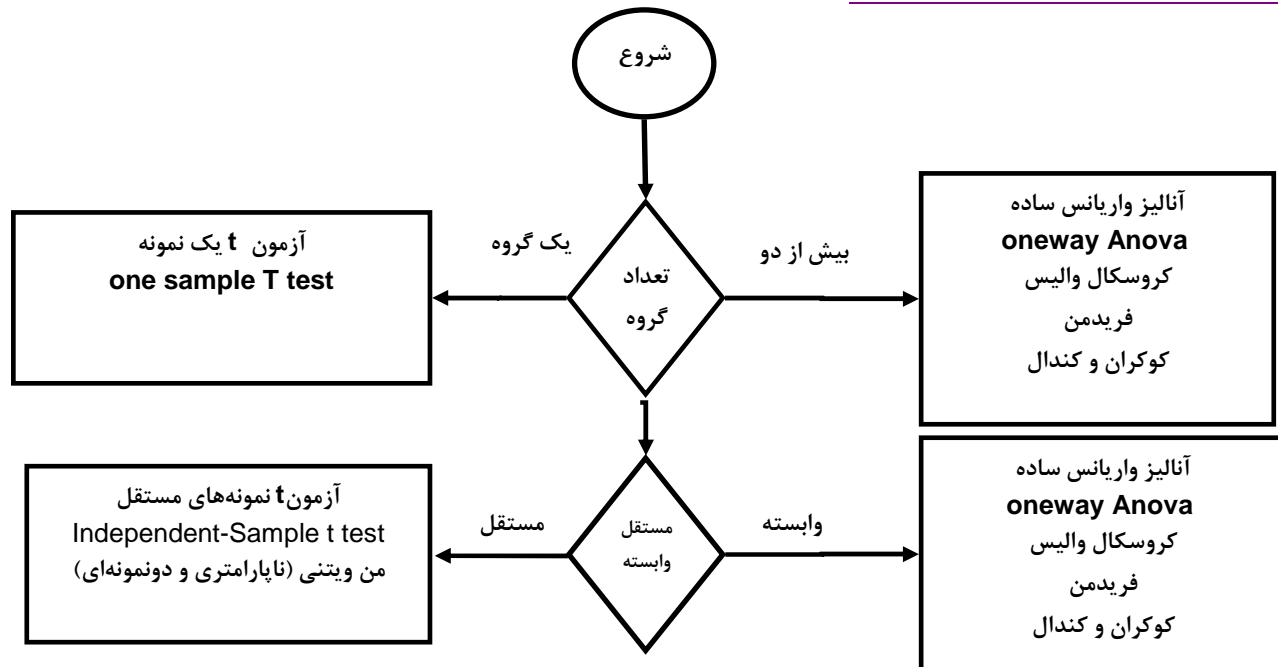
به جای  $\sigma$  ،  $pq$  را قرار می دهیم که  $p$  احتمال داشتن صفت مورد نظر و  $q=1-p$  احتمال نداشتن آن صفت است.

$$n = \frac{Nz^2 pq}{(N-1)d^2 + z^2 pq}$$

آزمون فرض :

آزمون فرض برای مقایسه شاخص جمعیت با نمونه و انتخاب مقدار شاخص نمونه به عنوان شاخص جمعیت به کار می‌رود. دقت این انتخاب با توجه به اندازه جمعیت و نمونه و میزان خطای است که می‌توانیم متحمل شویم. هرچه اندازه نمونه نسبت به جمعیت بیشتر باشد دقت آزمون بیشتر خواهد بود. هرچه جمعیت به توزیع نرمال نزدیکتر باشد آزمون‌ها پرتوان‌تر خواهند بود.

انتخاب آزمون برای مقایسه میانگین‌ها:



آزمون‌های مربوط به میانگین:

زمانی که جامعه را نداریم و یک نمونه انتخاب کرده ایم و می‌خواهیم خصوصیت نمونه را به جامعه نسبت بدهیم لذا آزمون انجام می‌دهیم. و با استفاده از قابلیت آزمون فرضیات میتوان اختلاف بین میانگین‌ها را سنجید. در انجام آزمون فرض ، اگر فرض آماری درست باشد و ما آن را رد کنیم مرتکب خطای نوع اول و اگر فرض آماری را که پذیرفتیم نادرست باشد مرتکب خطای نوع دوم می‌شویم. هرچه توزیع به طرف نرمال پیش برود آزمون‌ها پرتوان‌تر خواهند بود و خطای نوع دوم کمتر است. خطای نوع دوم یعنی احتمال رد  $H_1$  به شرط اینکه درست باشد.

فرض کنیم  $X_1, X_2, \dots, X_n$  یک نمونه  $n$  تایی از توزیع نرمال با میانگین مجهول  $\mu$  و واریانس معلوم  $\sigma^2$  باشند هدف آزمون فرض زیر در مورد میانگین جامعه است. فرض  $H_0$  یک فرض ساده در مقابل فرض  $H_1$  که یک فرض مرکب یک طرفه است قرارداد.

$$\begin{cases} H_0: \mu = 0 \\ H_1: \mu > 0 \end{cases}$$

در توزیع نرمال توان بیشتر است و نتیجه آزمون از یک نمونه به نمونه دیگر تفاوت نمی کند. و به طور کلی در جوامع بزرگ با فرض نرمال بودن از آزمون های پارامتری استفاده می شود. که به ترتیب قرارگرفتن اطلاعات تکیه دارند.

در جمعیت های با توزیع غیر نرمال و کوچک از آزمون های ناپارامتری استفاده می کنند. آزمونهای ناپارامتری در باره توزیعهای جامعه و واریانس آن فرض خاصی نمی کنند. رویکرد دیگر (که به همان میزان موجب افت توان نمی شود) خارج کردن مقادیر پرت و به کار بردن آزمون t با مجموعه داده های جدید است.

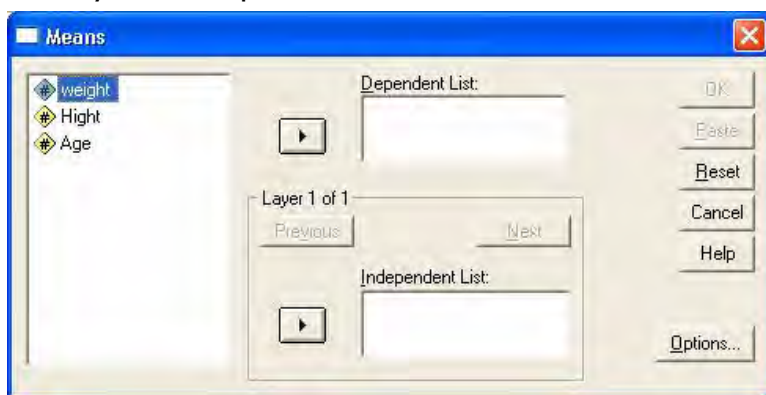
(توان یا power یک آزمون آماری احتمال رد کردن  $H_0$  است به شرطی که صحیح نباشد.) یکی از روشهای آزمون فرض آماری استفاده از  $p$ -value یا مقدار احتمال است اگر  $p$ -value  $\leq \alpha$  مقدار فرض  $H_0$  رد می شود.

### آزمون پارامتری مقایسه میانگین یک متغیر عددی برای دو یا چند ویژگی در یک متغیر رده بندی شده (Means) :

با مقایسه میانگین به عنوان شاخص تمرکز جمعیت می توان مقایسه مناسبی بین دو گروه از جمعیت (نمونه) بدست آورد.

از آنجایی که میانگین یا دیگر معیارهای تمرکز مثل میانه و مد به عنوان شاخص و نماینده جمعیت شناخته شده هستند، برای مقایسه دو ویژگی یا دو جمعیت بهتر است معیارهای تمرکز آنها را مقایسه کرد.

#### Analyze – Compare Means – Means



در قسمت Dependent list : متغیر یا ویژگی اندازه گیری شده را وارد کنید. این متغیر باید دارای مقادیر **عددی** باشد.

و در قسمت Independent list : متغیر دسته بندی یا **گروه بندی** را معرفی کنید. با این کار میانگین ویژگی براساس متغیر دسته بندی برای هر گروه بدست آمده و مقایسه می شوند. مثلاً مقایسه درآمد براساس وضعیت تاهل

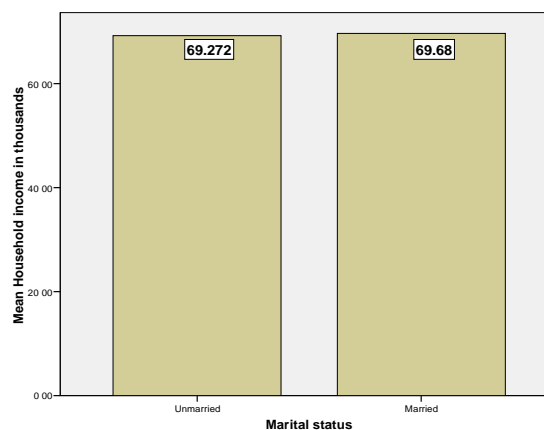
$$\text{آماره میانگین} : T = \frac{x - \bar{x}}{\frac{s}{\sqrt{n}}} \text{ می باشد.}$$

مقادیر مربوط به ویژگیهای یک نمونه (میانگین، انحراف معیار و غیره) را آماره (Statistics) و ویژگیهای معادل آنها در جامعه اصلی را پارامتر (Parameters) می گویند. نمودار میانگین متغیر income را نیز بر اساس وضعیت تاهل می کشیم :

#### Report

Household income in thousands

Marital status	Mean	N	Std. Deviation
Unmarried	69.2723	3224	78.32925
Married	69.6804	3176	79.12361
Total	69.4748	6400	78.71856

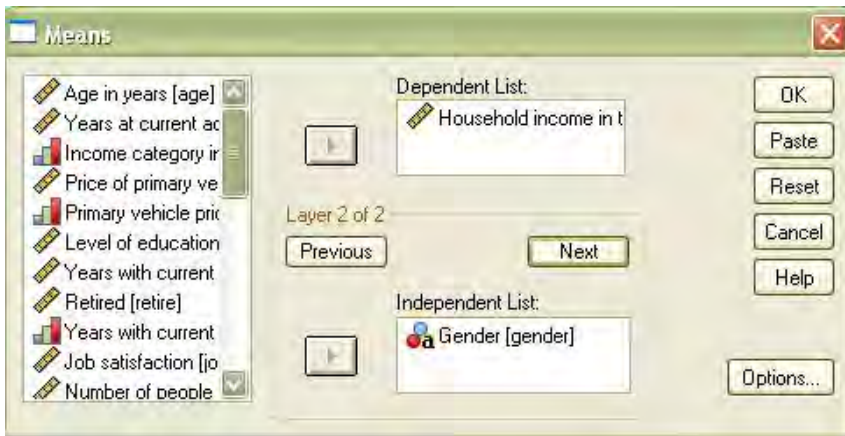




لایه بندی متغیرها:

با استفاده از دستورالعمل Mean موقعیت را به همراه جنسیت میتوان لایه بندی کرد.

Analyze – Compare Means – Means- dependent list .....income  
Independent list.....Marital  
Next  
Independent list.....Gender



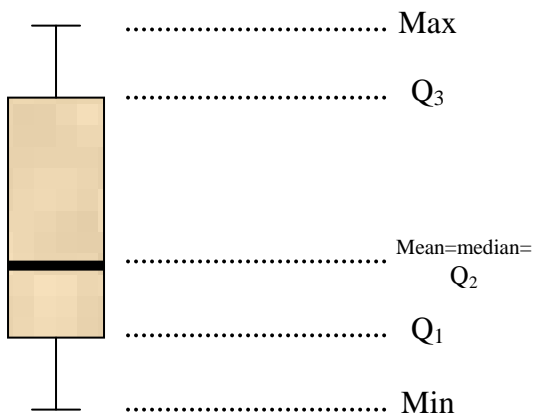
Report

Household income in thousands

Marital status	Gender	Mean	N	Std. Deviation
Unmarried	Female	72.0633	1533	83.02021
	Male	66.7422	1691	73.75302
	Total	69.2723	3224	78.32925
Married	Female	65.7179	1645	68.15853
	Male	73.9392	1530	89.27551
	Total	69.6797	3175	79.13606
Total	Female	68.7788	3178	75.74700
	Male	70.1608	3221	81.56216
	Total	69.4744	6399	78.72471

برای مقایسه میانگین ها بهترین نمودار Boxplot است. در این نوع نمودار برای هر متغیر یک box می کشد که هر چه ارتفاع box بیشتر باشد پراکندگی داده ها بیشتر است.

☆ 1



اگر میانگین وسط box باشد متقارن است و اگر بالاتر یا پایین تر از وسط باشد نشانه چولگی است.  
 ☆ نشان دهنده یک داده پرت است که شماره ۱ آن نشان میدهد که مربوط به case شماره ۱ است.  
 به داده پرت outlier می گویند.  
 به طور کلی داده ای که این شرط را داشته باشد پرت حساب می شود:

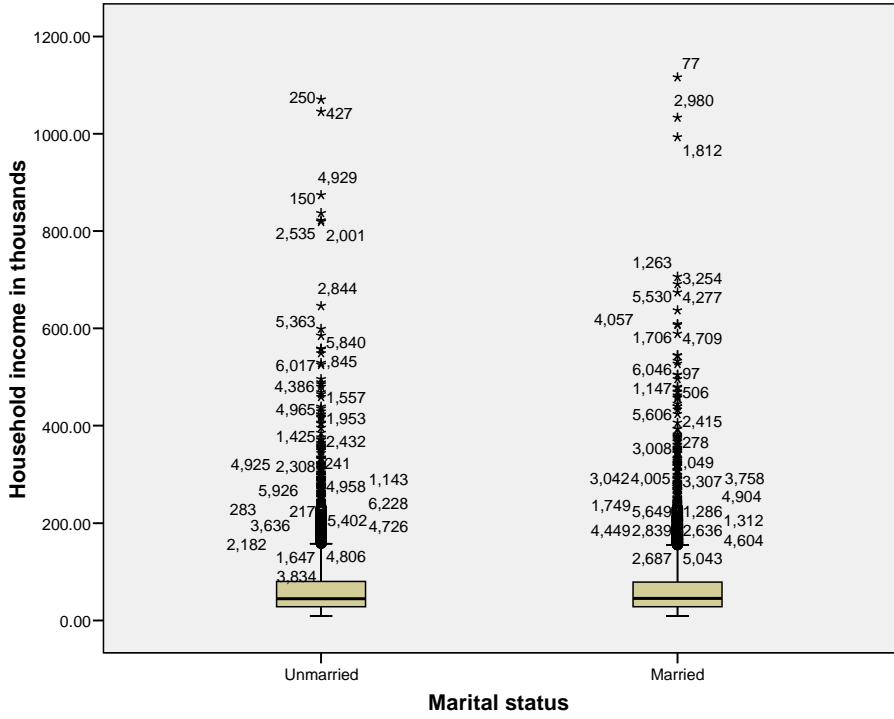
$$Q = Q_3 - Q_1$$

$$\frac{1}{2}Q + Q_3 < Outliar < Q_1 - \frac{1}{2}Q$$

به داده های خیلی پرت نیز Extreme variable می گویند که بیشتر از سه برابر Q فاصله دارند. با وجود داده های پرت ممکن است میانگین به سمت آنها متمایل شود و نتیجه آزمون را درست بدست نیاوریم پس بهتر است داده های پرت را حذف کنیم.

در مقایسه boxplot های بین میانگین دومتغیر اگر همپوشانی وجود داشته باشد یعنی پراکندگی اولی شامل پراکندگی دومی است لذا اختلاف زیادی بین میانگین ها نمی باشد و می شود آزمون را انجام داد.

Graph –legacy Dialogs - Boxplot – simple – variable.....income  
 Category Axis .....Marital

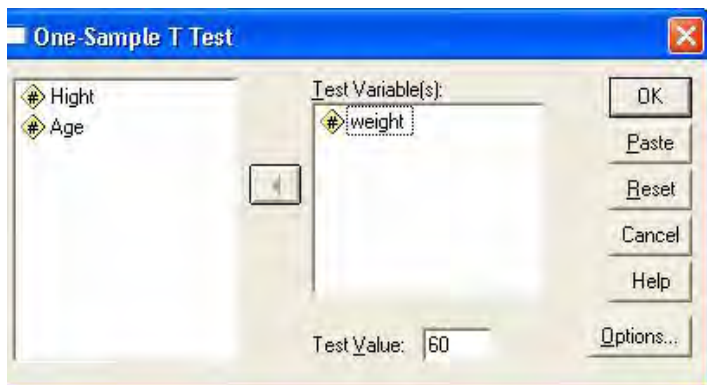


این نمودار نشان میدهد که نقاط پرت بسیاری وجود دارد

برای حذف داده های پرت از منوی: Data-select cases-if income<200 را انتخاب می کنیم.

تا بیشتر از ۲۰۰ را که خیلی پرت هستند در محاسبات و نمودار نیاورد.

آزمون میانگین نمونه در مقابل یک مقدار شناخته شده ( One\_ sample T Test ) :



- ✦ با اجرای این دستور امکان مقایسه و آزمون t برای سنجش آزمون به کار می‌رود.
- ✦ در این آزمون فرض بر نرمال بودن داده‌ها است.
- ✦ متغیر معرفی شده باید مقادیر **عددی** داشته باشد.
- ✦ اگر داده‌ها مقادیری چولگی هم داشته باشند این آزمون در برابر اشتباه مقاومت خواهد کرد.
- ✦ آزمون به شکل زیر در نظر گرفته شده است. ( $\mu$  مقدار میانگین نمونه است و 80 مقدار حدسی است که از جمعیت برای میانگین می‌زنیم.)

Analyze – compare Means – One\_ sample T Test / Test variable .....Income  
Test value .....80

در قسمت Test variables متغیر یا متغیرهایی که میانگینشان باید با مقدار test value (در اینجا مثلا 80) مقایسه شود قرار می‌گیرد.  
با انتخاب دکمه options نیز امکان تعریف درصد فاصله اطمینان (مثلا يك فاصله اطمینان 95% ) وجود دارد.  
دقت داشته باشیم فاصله اطمینان نشان دهنده درصد فواصلی از نمونه‌های مختلف است که شامل میانگین می‌باشند. برای مثال يك فاصله اطمینان 95% نشان می‌دهد، در 100 نمونه دیگر از جمعیت، 95 نمونه، میانگین نمونه‌ها در این فاصله خواهند بود.

*نکته : وقتی می‌گوییم اختلاف یا تساوی معنی دار است یعنی ناشی از نمونه گرفته شده نیست بلکه اگر با نمونه دیگری هم آزمون شود نتیجه همین خواهد بود. اگر معنی دار نباشد یعنی اختلاف یا تساوی ناشی از جمعیت است. روش تشخیص ، آزمون است.*

می خواهیم ببینیم میانگین حدس زده شده با مقدار ۸۰ در مورد متغیر درآمد معنی دار است یا نه؟

$$\begin{cases} H_0: \mu = 80 \\ H_1: \mu \neq 80 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H_0: \mu - 80 = 0 \\ H_1: \mu - 80 \neq 0 \end{cases}$$

```
T-TEST
/TESTVAL = 80
/MISSING = ANALYSIS
/VARIABLES = income
/CRITERIA = CI(.95) .
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Household income in thousands	6400	69.4748	78.71856	.98398

One-Sample Test

	Test Value = 80					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Household income in thousands	-10.696	6399	.000	-10.52516	-12.4541	-8.5962

نتیجه :  
درجه آزادی n-1 یعنی 6400-1=6399 است.

$$T = \frac{x - \bar{x}}{\frac{s}{\sqrt{n}}} = \frac{69.4748 - 80}{\frac{0.98398}{\sqrt{6400}}} = \frac{-10.52516}{0.98398} = -10.696 < 0$$

$$\text{Sig. (2-tailed)} = p - \text{value} = 0.000 < 0.05 = \alpha$$

۱. چون مقدار  $p - \text{value}$  از  $\alpha$  کوچکتر شده است پس دلیلی برای تایید  $H_0$  وجود ندارد. (رد فرض  $H_0$ ).
۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر نمی باشد که این خود عامل رد کننده فرض  $H_0$  است و نشان دهنده تفاوت آشکار است. چون این بازه منفی است یعنی  $\mu - 80$  منفی شده و یعنی میانگین از ۸۰ کمتر است.
۳. اختلاف معنی دار است.
۴. قدر مطلق  $t$  برابر یا بزرگتر از ۲ باشد، معنی دار است.

این آزمون را برای عدد ۷۰ نیز تکرار می کنیم:

$$\begin{cases} H_0: \mu = 70 \\ H_1: \mu \neq 70 \end{cases} \Rightarrow \begin{cases} H_0: \mu - 70 = 0 \\ H_1: \mu - 70 \neq 0 \end{cases}$$

T-TEST

```

/TTESTVAL = 70
/MISSING = ANALYSIS
/VARIABLES = income
/CRITERIA = CI(.95) .
    
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Household income in thousands	6400	69.4748	78.71856	.98398

One-Sample Test

	Test Value = 70					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Household income in thousands	-.534	6399	.594	-.52516	-2.4541	1.4038

نتیجه :

$\text{Sig. (2-tailed)} = p\text{-value} = 0.594 > 0.05 = \alpha$

۱. چون مقدار  $p\text{-value}$  از  $\alpha$  بزرگتر شده است پس دلیلی برای رد  $H_0$  وجود ندارد.
۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر می باشد که این خود عامل تایید کننده فرض  $H_0$  توسط نمونه است.
۳. اختلاف معنی دار نیست.
۴. قدر مطلق  $t$  برابر یا بزرگتر از ۲ باشد، معنی دار است.

نکته : در پنجره *One\_sample T Test* در قسمت *Option* میتوان مقدار خطای نوع اول یا  $\alpha$  را تغییر داد. اگر ضریب اطمینان را تا ۹۹% بالا ببریم یعنی  $\alpha$  را ۱% بگذاریم بیشتر فرضیه ها رد می شوند.

مثال :

فایل spss sample data.sav را باز کنید و میانگین وزن یا متغیر weight را بدست آورید.

Analyze- report – frequency

آزمونی انجام دهید با فرض اینکه میانگین وزن ۷۵ است.

$$\begin{cases} H_0: \mu = 75 \\ H_1: \mu \neq 75 \end{cases} \Rightarrow \begin{cases} H_0: \mu - 75 = 0 \\ H_1: \mu - 75 \neq 0 \end{cases}$$

T-TEST

```

/TESTVAL = 75
/MISSING = ANALYSIS
/VARIABLES = weight wieght_after
/CRITERIA = CI(.95) .

```

Statistics

weight		
N	Valid	11
	Missing	0
Mean		68.3182

One-Sample Test

Test Value = 75						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
weight	-1.220	10	.250	-6.68182	-18.8805	5.5169

نتیجه :

$$\text{Sig. (2-tailed)} = p\text{-value} = 0.250 > 0.05 = \alpha$$

۱. چون مقدار  $p\text{-value}$  از  $\alpha$  بزرگتر شده است پس دلیلی برای رد  $H_0$  وجود ندارد.
۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر می باشد که این خود عامل تایید کننده فرض  $H_0$  توسط نمونه است.
۳. اختلاف معنی دار نیست.
۴. قدر مطلق  $t$  برابر یا بزرگتر از  $t$  باشد، معنی دار است.

میانگین وزن بعد از رژیم یا متغیر weight -after را بدست آورید.

Analyze – compare Means – One\_ sample T Test / Test variable .....weight , weight - after  
Test value .....75

آزمونی انجام دهید با فرض اینکه میانگین وزن بعد از رژیم ۷۵ است.

```
T-TEST
/TESTVAL = 75
/MISSING = ANALYSIS
/VARIABLES = wieght_after weight
/CRITERIA = CI(.95) .
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
weight	11	68.3182	18.15802	5.47485
wieght_after	11	63.5455	14.29240	4.30932

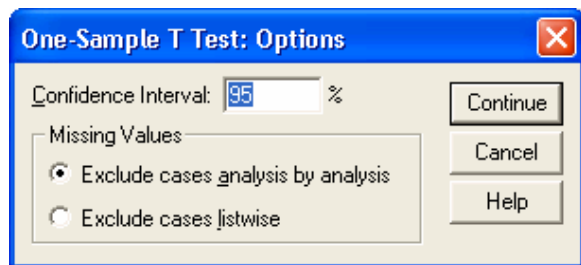
One-Sample Test

	Test Value = 75					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
weight	-1.167	9	.273	-7.05000	-20.7111	6.6111
wieght_after	-2.658	10	.024	-11.45455	-21.0563	-1.8528

نتیجه :

$Sig.(2-tailed) = p - value = 0.024 < 0.05 = \alpha$

۱. چون مقدار  $p - value$  (متغیر weight-after) از  $\alpha$  کوچکتر شده است پس دلیلی برای تایید  $H_0$  وجود ندارد. (رد فرض  $H_0$ )
۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر نمی باشد که این خود عامل رد فرض  $H_0$  است و نشان دهنده تفاوت آشکار است.
۳. اختلاف معنی دار است. پس میانگین وزن بعد از رژیم نمیتواند ۷۵ کیلوگرم باشد.
۴. قدر مطلق برابر یا بزرگتر از ۲ باشد، معنی دار است.



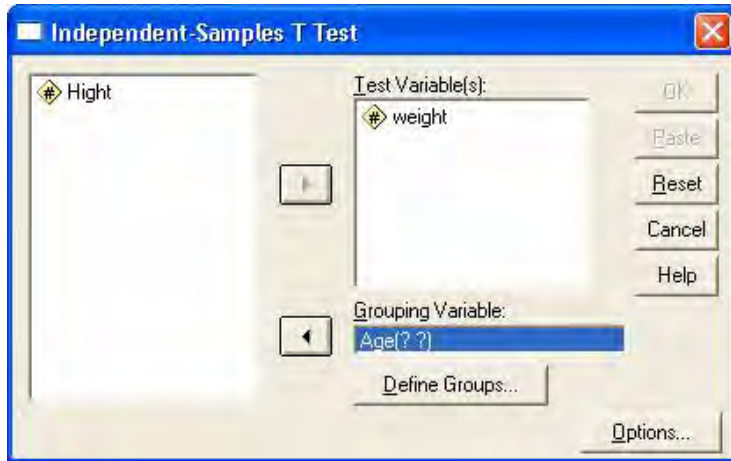
نکته: اگر در گزینه *exclude cases analysis by analysis* را تیک بزیم و *missing* داشته باشیم در مقایسه دو متغیر هر کدام را جداگانه حساب می کند و آزمون برای متغیرهای به کار رفته را بدون در نظر گرفته مقادیر گمشده برای متغیرهای دیگر به کار می رود. ولی اگر گزینه *exclude cases listwise* دوم یعنی *missing* را تیک بزیم کل مورد یا *case* ی که *missing* دارد را حذف می کند. مورد با مقدار گمشده برای همه متغیرهای به کار رفته در آزمون در نظر گرفته نخواهد شد. این *missing* چه در اثر تایپ نکردن داده ای باشد چه در *missing value* تعریف شده باشد فرقی نمی کند.

آزمون میانگین یک متغیر بین دو گروه یا رده ( Independent-Samples T Test ) :

- برای آزمون برابر بودن میانگین يك ويژگي جمعيت در بين دو گروه از اين روش استفاده مي كنيم.
- فرض براین است که ویژگی جمعیت به صورت نرمال توزیع شده است و متغیر يك مقدار کمی (عددی) است.
- همچنین نمونه به صورت مستقل و تصادفی انتخاب شده است.
- تقارن و عدم وجود نقاط پرت در این آزمون از اهمیت برخوردار است.
- این آزمون نسبت به چولگی حساس نیست.
- برابر بودن **واریانسها** با برابر نبودن آنها در دو گروه، خروجی آزمون را تغییر می دهد.

برای دسترسی به این دستور مسیر زیر را طی کنید.

Analyze - Compare Means - Independent-Samples T Test/Test variables.....weight  
Grouping variables.....gender



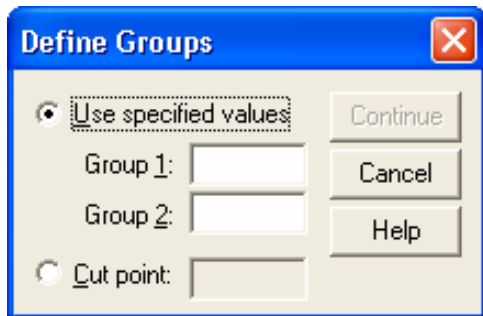
در قسمت test variable :  
متغیر یا متغیرهایی که باید آزمون شوند را وارد کنید.

در قسمت Grouping Variable :  
متغیر **گروه بندی** را مشخص کنید.  
این آزمون فقط می تواند مقایسه بین دو گروه را بدست آورد.

با انتخاب دکمه define group برای متغیر گروه بندی مقدار مورد نظر برای تعریف دو گروه را وارد کنید. حتی اگر متغیر گروه بندی بیشتر از دو گروه داشته باشد این آزمون فقط در دو گروه محاسبه خواهد شد.

همچنین با تنظیم مقدار cut point نیز امکان تفکیک گروه ها به صورت کمتر و بیشتر و مساوی مقدار cut point نیز هست.

$$\begin{cases} H0: m_1 = m_2 \\ H1: m_1 \neq m_2 \end{cases} \quad SPSS \Rightarrow \begin{cases} H0: m_1 - m_2 = 0 \\ H1: m_1 - m_2 \neq 0 \end{cases}$$



■ مقدار Group1 برای متغیر گروه بندی، دسته اول را مشخص می کند.

■ مقدار Group2 برای متغیر گروه بندی، دسته دوم را مشخص می کند.

■ اگر متغیر رده بندی، عددی بود میتوان با تعیین مقدار Cut Point دو گروه را به صورت زیر تعریف کرد:

- گروه ۱ مقادیر بیشتر و مساوی از cut point
- گروه ۲ مقادیر کمتر از cut point



میانگین متغیر وزن را در دو گروه زنان و مردان بررسی کنید:

$$\begin{cases} H_0: m_1 = m_2 \\ H_1: m_1 \neq m_2 \end{cases} \quad SPSS \Rightarrow \begin{cases} H_0: m_1 - m_2 = 0 \\ H_1: m_1 - m_2 \neq 0 \end{cases}$$

$m_1$  میانگین متغیر وزن در گروه مردان  
 $m_2$  میانگین متغیر وزن در گروه زنان

T-TEST

```
GROUPS = gender(1 2)
/MISSING = ANALYSIS
/VARIABLES = weight
/CRITERIA = CI(.95) .
```

**Group Statistics**

gender		N	Mean	Std. Deviation	Std. Error Mean
weight	male	4	75.7500	19.12023	9.56012
	female	6	62.7500	18.86730	7.70254

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
weight	Equal variances assumed	.009	.928	1.062	8	.319	13.00000	12.24027	-15.22611	41.22611
	Equal variances not assumed			1.059	6.512	.327	13.00000	12.27701	-16.47694	42.47694

نتیجه :

$Sig. = p - value = 0.928 > 0.05 = \alpha$

۱. چون مقدار  $p - value$  (متغیر weight) که مربوط به واریانس است از  $\alpha$  بزرگتر شده است پس برابری واریانس ها رد نمی شود. لذا به  $p - value$  مربوط به میانگین در سطر اول نگاه می کنیم. در غیر اینصورت سطر دوم را بررسی خواهیم کرد.

$Sig(2-tailed). = p - value = 0.319 > 0.05 = \alpha$

۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر می باشد که این خود عامل تایید فرض  $H_0$  توسط نمونه است .

۳. اختلاف معنی دار نیست. پس میانگین وزن در دو گروه مردان و زنان برابر است.

۴. قدر مطلق برابر یا بزرگتر از ۲ باشد، معنی دار است.

میانگین متغیر وزن را در دو گروه سنی بالای ۳۰ سال و پایین ۳۰ سال بررسی کنید:  
 برای گروه بندی متغیر وزن به دو گروه بالای ۳۰ سال و پایین ۳۰ سال میتوان از راه recode یا compute استفاده نمود. ولی ساده ترین راه این است که در پنجره Independent-Samples T Test و دگمه option از گزینه cutpoint استفاده نمود و عدد ۳۰ را وارد کرد.

$$\begin{cases} H_0: m_1 = m_2 \\ H_1: m_1 \neq m_2 \end{cases} \quad SPSS \Rightarrow \begin{cases} H_0: m_1 - m_2 = 0 \\ H_1: m_1 - m_2 \neq 0 \end{cases}$$

$m_1$  میانگین متغیر وزن در گروه سنی پایین تر از ۳۰ سال  
 $m_2$  میانگین متغیر وزن در گروه سنی بالاتر از ۳۰ سال

T-TEST

```
GROUPS = age(30)
/MISSING = ANALYSIS
/VARIABLES = weight
/CRITERIA = CI(.95) .
```

Group Statistics

age		N	Mean	Std. Deviation	Std. Error Mean
weight	>= 30.00	7	74.5714	18.36534	6.94145
	< 30.00	3	52.5000	10.85127	6.26498

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
weight	Equal variances assumed	1.334	.281	1.903	8	.093	22.07143	11.59642	-4.66997	48.81283
	Equal variances not assumed			2.360	6.606	.052	22.07143	9.35060	-.30937	44.45223

نتیجه :

۱. چون مقدار  $p - value$  (متغیر weight) که مربوط به واریانس است از  $\alpha$  بزرگتر شده است پس برابری واریانس ها رد نمی شود. لذا به  $p - value$  مربوط به میانگین در سطر اول نگاه می کنیم. در غیر اینصورت سطر دوم را بررسی خواهیم کرد.

$Sig(2-tailed). = p - value = 0.093 > 0.05 = \alpha$

۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر می باشد که این خود عامل تایید فرض  $H_0$  توسط نمونه است.

۳. اختلاف معنی دار نیست. پس میانگین وزن در دو گروه سنی بالاتر از ۳۰ سال و پایین تر از ۳۰ سال برابر است.

۴. قدر مطلق برابر یا بزرگتر از ۲ باشد، معنی دار است.

میانگین متغیر درآمد را در فایل Demo در دو گروه کسانی که تحصیلات عالی دارند و کسانی که تحصیلاتشان را تکمیل نکرده اند بررسی کنید:

$$\begin{cases} H_0: m_1 = m_2 \\ H_1: m_1 \neq m_2 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H_0: m_1 - m_2 = 0 \\ H_1: m_1 - m_2 \neq 0 \end{cases}$$

$m_1$  میانگین متغیر درآمد income در گروه تحصیلات تکمیل نشده  
 $m_2$  میانگین متغیر درآمد income در گروه تحصیلات عالی

T-TEST

```
GROUPS = ed(1 5)
/MISSING = ANALYSIS
/VARIABLES = income
/CRITERIA = CI(.95) .
```

**Group Statistics**

Level of education		N	Mean	Std. Deviation	Std. Error Mean
Household income in thousands	Did not complete high school	1390	59.8662	61.30036	1.64420
	Post-undergraduate degree	359	87.1699	94.79998	5.00335

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Household income in thousands	Equal variances assumed	35.128	.000	-6.636	1747	.000	-27.30373	4.11419	-35.37298	-19.23448
	Equal variances not assumed			-5.184	438.180	.000	-27.30373	5.26659	-37.65464	-16.95282

نتیجه :

$Sig. = p - value = 0.000 < 0.05 = \alpha$

۱. چون مقدار  $p - value$  (متغیر income) که مربوط به واریانس است از  $\alpha$  کوچکتر شده است پس برابری واریانس ها رد می شود. لذا به  $p - value$  مربوط به میانگین در سطر دوم نگاه می کنیم. کوچکتر از  $\alpha$  است پس فرض  $H_0$  رد می شود.  
 $Sig(2-tailed). = p - value = 0.000 < 0.05 = \alpha$

۲. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر نمی باشد که این خود عامل رد فرض  $H_0$  است و نشان دهنده تفاوت آشکار است.

۳. اختلاف معنی دار است. پس میانگین درآمد در دو گروه تحصیلات تکمیل نشده و عالی برابر نیست.

۴. قدر مطلق برابر یا بزرگتر از ۲ باشد، معنی دار است.

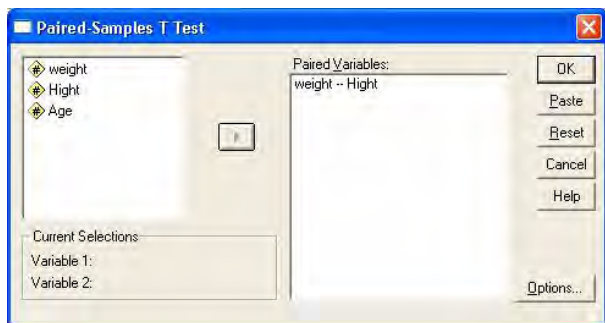
### آزمون مقایسه میانگین دو متغیر جفت ( Paired-Samples T Test ) :

- ✦ اگر میانگین دو متغیر در نمونه باید مقایسه شوند از این آزمون استفاده می‌کنیم.
  - ✦ در این آزمون مشخصه‌های اندازه‌گیری شده به صورت زوج مرتب (X,Y) در نظر گرفته می‌شوند که در آن X ویژگی اول و Y ویژگی دوم در نمونه است.
  - ✦ برای مثال می‌خواهیم اثر یک دارو را در کاهش فشار خون مشخص کنیم. به افراد نمونه در ابتدا دارونما داده فشار خون را اندازه‌گیری می‌کنیم. سپس از خوردن دارو واقعی نیز فشار خون اندازه‌گیری خواهد شد. آزمون می‌خواهد میانگین فشار خون را قبل و بعد از داروی واقعی مقایسه کند.
  - ✦ فرض بر این است مقادیر **عددی** هستند.
  - ✦ داده‌ها هر دو متغیر توزیع نرمال دارند. (یا تفاوت آنها دارای توزیع نرمال است).
  - ✦ نمونه‌ها به صورت مستقل و تصادفی انتخاب شده‌اند.
  - ✦ تعداد نمونه‌ها برای هر دو متغیر باید **برابر** باشد. در صورت وجود مقدار گمشده در یک متغیر به طور کامل آن مورد نادیده گرفته خواهد شد.
  - ✦ تفاوت در مقدار واریانس دو متغیر اشکالی ندارد.
- برای دسترسی به این آزمون مراحل زیر را طی کنید.

Analyze - Compare Means - Paired-Samples T Test...

$$\begin{cases} H_0: m_1 = m_2 \\ H_1: m_1 \neq m_2 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H_0: m_1 - m_2 = 0 \\ H_1: m_1 - m_2 \neq 0 \end{cases}$$

$m_1$  میانگین متغیر اول  
 $m_2$  میانگین متغیر دوم



هر دو متغیر را در قسمت سمت چپ انتخاب کرده و به قسمت pair Variable انتقال دهید.

دو متغیر weight و weight-after را با هم مقایسه کنید.

Analyze - Compare Means - Paired-Samples T Test...

$$\begin{cases} H0: m_1 = m_2 \\ H1: m_1 \neq m_2 \end{cases} \Rightarrow \begin{cases} H0: m_1 - m_2 = 0 \\ H1: m_1 - m_2 \neq 0 \end{cases}$$

$m_1$  میانگین متغیروزن weight  
 $m_2$  میانگین متغیر وزن بعد از رژیم weight-after

```
T-TEST
  PAIRS = weight WITH wieght_after (PAIRED)
  /CRITERIA = CI (.95)
  /MISSING = ANALYSIS.
```

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 weight & wieght_after	10	.941	.000

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 weight - wieght_after	5.65000	7.37884	2.33339	.37150	10.92850	2.421	9	.039

نتیجه :

۱. ضریب همبستگی یا correlation عدد 0.941 را نشان میدهد که هر چه این عدد به عدد 1 نزدیکتر باشد یعنی دو متغیر اثر بیشتری روی هم دارند. یعنی در تمام موارد کاهش وزن پس از رژیم وجود داشته است و کاملاً با هم هماهنگ بودند.
۲. با توجه به کوچکتر بودن  $p-value$  از  $\alpha$  ، فرض  $H_0$  رد می شود.  
 $Sig(2-tailed). = p-value = 0.039 < 0.05 = \alpha$
۳. دو عدد نشان داده شده در ستون ( 95% Confidence Interval of the Difference ) شامل صفر نمی باشد که این خود عامل رد فرض  $H_0$  است و نشان دهنده تفاوت آشکار است.
۴. اختلاف معنی دار است. پس میانگین جفت متغیر وزن و وزن بعد از رژیم برابر نیست.
۵. قدر مطلق t برابر یا بزرگتر از ۲ باشد، معنی دار است.

### آزمون مقایسه میانگین یک متغیر در چندین گروه ( One-Way Anova ) :

با استفاده از جدول آنالیز واریانس یک طرفه امکان مقایسه مقدار یک شاخص در بین چندین گروه وجود دارد.

با توجه به آماره‌های جدول آنالیز واریانس اگر نسبت مقدار میانگین مربعات خطای درون گروهی (مجموع مربعات اختلاف مقادیر هر گروه از میانگین گروه) با مقدار میانگین مربعات خطای بین گروهی (مجموع مربعات اختلاف میانگین هر گروه از میانگین کل) دارای اختلاف زیادی باشد نشان دهنده نقش عامل در تغییر میانگین گروه را نشان می‌دهد.

#### فرض‌های اولیه :

- متغیر فاکتور باید مقادیر صحیح داشته باشد.
- متغیر وابسته باید دارای مقادیر پیوسته باشد.
- نمونه در هر گروه باید دارای توزیع نرمال تصادفی باشند.
- آزمون آنالیز واریانس نسبت به شرط نرمال بودن مقاوم است ولی باید توزیع دارای تقارن باشد.
- گروه‌ها باید از جمعیتی با **واریانس برابر** انتخاب شده باشند. (این شرط را با استفاده از آماره Levens می‌توان اندازه گیری کرد)

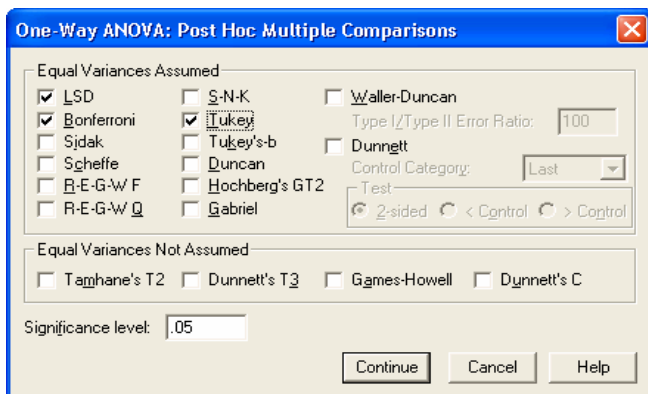
با استفاده از این جدول امکان اندازه گیری تغییر در مقدار میانگین در بین گروه‌ها وجود دارد. برای دسترسی به این دستور مراحل زیر را طی کنید.

#### Analyze - Compare Means - One-Way ANOVA...



متغیر وابسته را در محل Dependent List وارد کنید.

متغیر عامل (متغیری که مقدار متغیر وابسته را نسبت به گروه‌های مختلف آن می‌خواهید آزمون کنید) را در قسمت Factor وارد کنید. می‌توان چندین متغیر وابسته را در لیست وارد کنید ولی برای متغیر فاکتور فقط از یک متغیر می‌توان استفاده کرد. با این کار میانگین هر متغیر در سطوح متغیر فاکتور مقایسه می‌شوند.

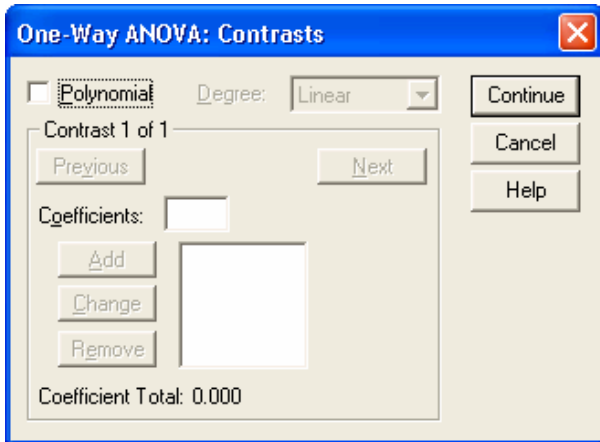


اگر می‌خواهید آزمون مربوط به دوتایی‌ها را انجام دهید گزینه Post Hoc را انتخاب کنید.

نوع آزمون را می‌توانید در صورت برابر بودن واریانس‌ها یکی از گزینه‌های کادر اولی مانند LSD-tukey-Bunfferoni- انتخاب کنید. سپس دکمه Continue را بزنید.

اگر می‌خواهید ترکیبی خطی از میانگین‌های گروه‌ها را با هم مقایسه کنید از دکمه Contrast استفاده کنید.

توجه داشته باشید که بهتر است مجموع ضرایب ترکیب خطی برابر صفر باشد.



به ترتیب ضرایب هر میانگین را در قسمت Coefficient وارد کرده و دکمه Add را بزنید.  
 دقت کنید که مجموع ضرایب در قسمت Coefficient Total محاسبه شده است و باید برابر صفر باشد.  
 اگر یک رابطه خطی بین میانگین‌ها احتیاج دارید گزینه Polynomial را انتخاب کنید.  
 دکمه Continue را بزنید.

آزمون مقایسه میانگین چند متغیر :

فرض آزمون به صورت زیر خواهد بود.  
 $H_0: m_1=m_2=m_3$   
 $H_1: m_1 \neq m_2 \neq m_3$

در صورتی که فرض صفر رد شود می‌توان با استفاده از آزمون‌های دوتایی مشخص کرد که علت عدم تساوی برای کدام زوج بوده است.  
 این آزمون را به عنوان Post Hoc نیز می‌شناسند.  
 همچنین امکان آزمون ترکیبی از میانگین‌ها (Contrast) وجود دارد.

در فایل Demo مقایسه کنید که میانگین درآمد بر سطوح مختلف تحصیلی چه اثری دارد؟

Analyze - Compare Means - One-Way Anova

$$\begin{cases} H_0 : m_1 = m_2 = m_3 = m_4 = m_5 \\ H_1 : m_1 \neq m_2 \neq m_3 \neq m_4 \neq m_5 \end{cases}$$

- $m_1$  میانگین متغیر درآمد در گروه اول ( did not completed high school )
- $m_2$  میانگین متغیر درآمد در گروه دوم ( high school degree )
- $m_3$  میانگین متغیر درآمد در گروه سوم ( some college )
- $m_4$  میانگین متغیر درآمد در گروه چهارم ( college degree )
- $m_5$  میانگین متغیر درآمد در گروه پنجم ( post-undergraduate degree )

```
GET
  FILE='C:\Program Files\SPSS\Tutorial\sample_files\demo.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
ONEWAY
  income BY ed
  /STATISTICS HOMOGENEITY
  /MISSING ANALYSIS .
```

**Test of Homogeneity of Variances**

Household income in thousands

Levene Statistic	df1	df2	Sig.
14.766	4	6395	.000

**ANOVA**

Household income in thousands

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	376079.699	4	94019.925	15.309	.000
Within Groups	39276042.251	6395	6141.680		
Total	39652121.950	6399			

نتیجه :

۱. با توجه به کوچکتر بودن  $p$ -value از  $\alpha$  ، فرض  $H_0$  رد می شود.

Sig. =  $p$ -value = 0.000 < 0.05 =  $\alpha$

۲. اختلاف معنی دار است و ناشی از گروه بندی تحصیلات است. پس میانگین متغیر درآمد در بین گروههای تحصیلات برابر نیست.

۳. تفاوت زیاد مربعات خطای بین گروهی (مجموع مربعات اختلاف میانگین هر گروه از میانگین کل) - between Groups با مربعات خطای درون گروهی (مجموع مربعات اختلاف مقادیر هر گروه از میانگین گروه) - within Groups نشان دهنده نقش عامل ( factor ) در تغییر میانگین گروه است.



۴. نتایج بدست آمده از آزمون مقایسه واریانسها در جدول **Test of Homogeneity of Variances** نشان میدهد که واریانس های متغیر در آمد در ۵ گروه متغیر تحصیلات اختلاف معنی داری دارند و  $\text{Sig} < \alpha$  ،  $p\text{-value} = 0.000$

۵. نتایج بدست آمده از آنالیز واریانس در جدول **ANOVA** نشان میدهد که بین میانگین های متغیر درآمد در ۵ گروه متغیر تحصیلات اختلاف معنی داری وجود دارد. پس باید به دنبال اختلافها باشیم و با استفاده از گزینه **Post HOC** دوبه دو با هم مقایسه کنیم.

در پنجره **Post HOC** دو حالت وجود دارد :

قسمت بالا مربوط به آزمونهای مورد استفاده در حالتی که واریانس جوامع تفاوتی نداشته باشند:  
(Equal Variances Assumed)

قسمت پایین مربوط به آزمونهای مورد استفاده در حالتی که واریانس جوامع متفاوت باشند:  
(Equal Variances Not Assumed)

در قسمت بالا (برابری واریانسها) در حالتی که واریانس گروهها با هم اختلاف معنی داری ندارند به طور نمونه سه آزمون متداول (Dunnett، Duncan، Tukey) را برای تجزیه و تحلیل آماری انتخاب می شود. در مورد آزمون Dunnett این نکته را باید مورد توجه قرار داد که از میان یکی از گروهها (۵ گروه تحصیلی) یکی را به عنوان گروه کنترل (شاهد) در نظر می گیریم تا سایر گروهها را با آن بسنجند . این گروه می توان گروه اول (First) یا گروه آخر (Last) باشد. انتخاب هر کدام از این گروهها به عنوان گروه اول یا آخر در نتایج تغییری ایجاد نمی کند. برای این کار با فعال کردن آزمون Dunnett گزینه Control Category فعال شده و در مربع روبروی آن گروه کنترل را انتخاب می کنیم. در مرحله بعد با کلیک بر روی Continue به پنجره قبلی می رویم و با کلیک بر روی Ok می توانیم خروجیهای مورد نظر را مشاهده کنیم.

در قسمت پایین (نابرابری واریانسها) آزمونهای (Games Howell ، Dunnett's T3 ، Tamhane's T2 ، Dunnett's C استفاده می شوند.

در این مثال چون فرض همگنی واریانسها پذیرفته نشد، به بیان دیگر فرض  $H_0$  که برابری واریانسها را مطرح می کند رد شده است از آزمونهای پایینی استفاده می کنیم.

نکته: در نتایج **post hoc** ، آماره F تا موقعی که اندازه های نمونه مساوی یا تقریباً مساوی هستند در برابر واریانس های نامساوی نیز مقاومت می کند ولی با اندازه نمونه نامساوی فاقد توان است و نتایج ناصحیح می دهد.

اگر واریانس ها برابر نبودند و در مورد آماره F ، sig نزدیک ۰.۰۵ باشد ما دودل هستیم که به این نتایج اعتماد بکنیم یا نه؟ لذا در دگمه option گزینه های Welch, Brown-Forsythe را نیز انتخاب کرده و بررسی می کنیم. در صورت تایید باید مقادیر پرت را حذف کرده و دوباره ANOVA را اجرا کنیم.

## Post Hoc Tests

### Multiple Comparisons

Dependent Variable: Household income in thousands

(I) Level of education	(J) Level of education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
					Upper Bound	Lower Bound	
Tamhon's $m_1$	Did not complete high school	High school degree-m2	-6.34094	2.33139	.064	-12.8724	.1905
		Some college-m3	-10.26837(*)	2.74450	.002	-17.9587	-2.5781
		College degree-m4	-18.78400(*)	3.01158	.000	-27.2233	-10.3447
		Post-undergraduate degree-m5	-27.30373(*)	5.26659	.000	-42.1229	-12.4846
High school degree $m_2$	Did not complete high school-m1	High school degree-m2	6.34094	2.33139	.064	-.1905	12.8724
		Some college-m3	-3.92743	2.74970	.811	-11.6318	3.7769
		College degree-m4	-12.44306(*)	3.01632	.000	-20.8952	-3.9909
		Post-undergraduate degree-m5	-20.96279(*)	5.26930	.001	-35.7894	-6.1362
Some college $m_3$	Did not complete high school-m1	High school degree-m2	10.26837(*)	2.74450	.002	2.5781	17.9587
		High school degree-m2	3.92743	2.74970	.811	-3.7769	11.6318
		College degree-m4	-8.51563	3.34591	.105	-17.8907	.8594
		Post-undergraduate degree-m5	-17.03536(*)	5.46465	.019	-32.4016	-1.6691
College degree $m_4$	Did not complete high school-m1	High school degree-m2	18.78400(*)	3.01158	.000	10.3447	27.2233
		High school degree-m2	12.44306(*)	3.01632	.000	3.9909	20.8952
		Some college-m3	8.51563	3.34591	.105	-.8594	17.8907
		Post-undergraduate degree-m5	-8.51973	5.60355	.749	-24.2704	7.2309
Post-undergraduate degree $m_5$	Did not complete high school-m1	High school degree-m2	27.30373(*)	5.26659	.000	12.4846	42.1229
		High school degree-m2	20.96279(*)	5.26930	.001	6.1362	35.7894
		Some college-m3	17.03536(*)	5.46465	.019	1.6691	32.4016
		College degree-m4	8.51973	5.60355	.749	-7.2309	24.2704

$m_1 \neq m_3$        $m_1 = m_2$

$m_1 \neq m_4$        $m_2 = m_3$

$m_1 \neq m_5$        $m_4 = m_3$

$m_2 \neq m_4$        $m_4 = m_5$

$m_2 \neq m_5$

$m_5 \neq m_3$

نتیجه هر چهار آزمون :  
 ۱. میانگین درآمد در گروه های پشت سر هم برابر است مثل  $m_1$  و  $m_2$   
 ولی در گروههای با فاصله برابر نیست مثل  $m_1$  و  $m_3$

(I) Level of education	(J) Level of education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
					Upper Bound	Lower Bound	
Dunnett T3  $m_1$	Did not complete high school	High school degree-m2	-6.34094	2.33139	.064	-12.8721	.1902
		Some college-m3	-10.26837(*)	2.74450	.002	-17.9585	-2.5782
		College degree-m4	-18.78400(*)	3.01158	.000	-27.2230	-10.3450
		Post-undergraduate degree-m5	-27.30373(*)	5.26659	.000	-42.1194	-12.4881
$m_2$	High school degree	Did not complete high school-m1	6.34094	2.33139	.064	-.1902	12.8721
		Some college-m3	-3.92743	2.74970	.810	-11.6317	3.7768
		College degree-m4	-12.44306(*)	3.01632	.000	-20.8950	-3.9911
		Post-undergraduate degree-m5	-20.96279(*)	5.26930	.001	-35.7859	-6.1397
$m_3$	Some college	Did not complete high school-m1	10.26837(*)	2.74450	.002	2.5782	17.9585
		High school degree-m2	3.92743	2.74970	.810	-3.7768	11.6317
		College degree-m4	-8.51563	3.34591	.104	-17.8905	.8593
		Post-undergraduate degree-m5	-17.03536(*)	5.46465	.019	-32.3985	-1.6723
$m_4$	College degree	Did not complete high school-m1	18.78400(*)	3.01158	.000	10.3450	27.2230
		High school degree-m2	12.44306(*)	3.01632	.000	3.9911	20.8950
		Some college-m3	8.51563	3.34591	.104	-.8593	17.8905
		Post-undergraduate degree-m5	-8.51973	5.60355	.747	-24.2675	7.2280
$m_5$	Post-undergraduate degree	Did not complete high school-m1	27.30373(*)	5.26659	.000	12.4881	42.1194
		High school degree-m2	20.96279(*)	5.26930	.001	6.1397	35.7859
		Some college-m3	17.03536(*)	5.46465	.019	1.6723	32.3985
		College degree-m4	8.51973	5.60355	.747	-7.2280	24.2675

(I) Level of education	(J) Level of education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
					Upper Bound	Upper Bound	
Dunnnett C  $m_1$	Did not complete high school	High school degree-m2	-6.34094	2.33139		-12.7076	.025
		Some college-m3	-10.26837(*)	2.74450		-17.7647	-2.772
		College degree-m4	-18.78400(*)	3.01158		-27.0099	-10.558
		Post-undergraduate degree-m5	-27.30373(*)	5.26659		-41.7380	-12.869
$m_2$	High school degree	Did not complete high school-m1	6.34094	2.33139		-.0257	12.707
		Some college-m3	-3.92743	2.74970		-11.4370	3.582
		College degree-m4	-12.44306(*)	3.01632		-20.6810	-4.205
		Post-undergraduate degree-m5	-20.96279(*)	5.26930		-35.4039	-6.521
$m_3$	Some college	Did not complete high school-m1	10.26837(*)	2.74450		2.7720	17.764
		High school degree-m2	3.92743	2.74970		-3.5821	11.437
		College degree-m4	-8.51563	3.34591		-17.6548	.623
		Post-undergraduate degree-m5	-17.03536(*)	5.46465		-32.0089	-2.061
$m_4$	College degree	Did not complete high school-m1	18.78400(*)	3.01158		10.5581	27.009
		High school degree-m2	12.44306(*)	3.01632		4.2051	20.681
		Some college-m3	8.51563	3.34591		-.6236	17.654
		Post-undergraduate degree-m5	-8.51973	5.60355		-23.8715	6.832
$m_5$	Post-undergraduate degree	Did not complete high school-m1	27.30373(*)	5.26659		12.8695	41.738
		High school degree-m2	20.96279(*)	5.26930		6.5217	35.403
		Some college-m3	17.03536(*)	5.46465		2.0618	32.008
		College degree-m4	8.51973	5.60355		-6.8320	23.871

\* The mean difference is significant at the .05 level.

(I) Level of education	(J) Level of education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
					Upper Bound	Upper Bound	
Games-Howell $m_1$	Did not complete high school	High school degree-m2	-6.34094	2.33139	.051	-12.7040	.0221
		Some college-m3	-10.26837(*)	2.74450	.002	-17.7602	-2.7766
		College degree-m4	-18.78400(*)	3.01158	.000	-27.0053	-10.5627
		Post-undergraduate degree-m5	-27.30373(*)	5.26659	.000	-41.7298	-12.8776
$m_2$	High school degree	Did not complete high school-m1	6.34094	2.33139	.051	-.0221	12.7040
		Some college-m3	-3.92743	2.74970	.609	-11.4330	3.5782
		College degree-m4	-12.44306(*)	3.01632	.000	-20.6770	-4.2091
		Post-undergraduate degree-m5	-20.96279(*)	5.26930	.001	-35.3961	-6.5295
$m_3$	Some college	Did not complete high school-m1	10.26837(*)	2.74450	.002	2.7766	17.7602
		High school degree-m2	3.92743	2.74970	.609	-3.5782	11.4330
		College degree-m4	-8.51563	3.34591	.081	-17.6488	.6175
		Post-undergraduate degree-m5	-17.03536(*)	5.46465	.016	-31.9958	-2.0749
$m_4$	College degree	Did not complete high school-m1	18.78400(*)	3.01158	.000	10.5627	27.0053
		High school degree-m2	12.44306(*)	3.01632	.000	4.2091	20.6770
		Some college-m3	8.51563	3.34591	.081	-.6175	17.6488
		Post-undergraduate degree-m5	-8.51973	5.60355	.550	-23.8554	6.8160
$m_5$	Post-undergraduate degree	Did not complete high school-m1	27.30373(*)	5.26659	.000	12.8776	41.7298
		High school degree-m2	20.96279(*)	5.26930	.001	6.5295	35.3961
		Some college-m3	17.03536(*)	5.46465	.016	2.0749	31.9958
		College degree-m4	8.51973	5.60355	.550	-6.8160	23.8554

« جلسه پنجم - ۸۷/۱/۲۲ »

نکته: اگر جهت مقایسه، دو متغیر هم واحد نباشند میتوان از متغیر استاندارد استفاده کرد. با محاسبه  $x - \bar{x}$  عرض از مبدا منحنی صفر می شود و با تقسیم بر  $s$  واحد صورت کسر و مخرج آن یکی شده و در نهایت  $Z$  بی واحد خواهد بود.

$$z = \frac{x - \bar{x}}{s}$$

برای محاسبه  $Z$  در نرم افزار SPSS از دستور COMPUTE استفاده می کنیم.

آزمون های ناپارامتری :

دلایل استفاده از آزمون ناپارامتری:

- ✦ توزیع داده ها نرمال نباشد.
- ✦ اختلاف پراکندگی زیاد باشد.
- ✦ اندازه نمونه کم باشد.

چرا از ابتدا بدون دانستن شرایط لازم از آزمون ناپارامتری استفاده نمی کنیم؟ زیرا توان این آزمون در شرایطی که توزیع داده ها نرمال باشد کم است.

فرض صفر توزیع جمعیت را مشخص می کند اما فرض مقابل، فرض ادعایی است و برای رد فرض صفر ادعا شده است. و هدف آزمون این است که با ارائه یک نمونه فرض صفر را رد کنیم. هیچ وقت نمی گوئیم فرض مقابل صحیح است چون فقط با یک نمونه نشان داده ایم که فرضیه صفر رد شد ولی این دلیل اثبات فرضیه مقابل نیست.

$p$ -value کمترین مقدار برای رد فرض صفر است. و از نمونه بدست میآید. هدف اصلی ما برای آزمون کوچک شدن  $p$ -value و رد فرضیه  $H_0$  است وگرنه باین آزمون کاربیهوده ای انجام شده است. برای جلوگیری از اینکار بهتر است از اول فرضیه را درست انتخاب کنیم نمونه معمولی بگیریم و شاخص را استخراج کنیم بعد حجم نمونه و بقیه را انتخاب کنیم. ( $\alpha$  هم خطایی است که محقق میتواند تحمل کند).

توان آزمون :

اگر واقعاً فرض صفر غلط باشد و توسط آزمون نمونه هم غلط شود بدون خطا بوده ایم. (تصمیم درست) وگرنه در غیر اینصورت دچار خطا شده ایم :

خطای نوع اول: رد فرض صفر توسط نمونه به شرطی که فرض صفر در جمعیت صحیح بوده است. ( $\alpha$ )  
خطای نوع دوم: قبول فرض صفر توسط نمونه به شرطی که فرض صفر در جمعیت غلط بوده است. ( $\beta$ )

سطح اطمینان ( $1 - \alpha$ )

توان آزمون ( $1 - \beta$ )

سطح معنی داری و خطای نوع اول ( $\alpha$ )

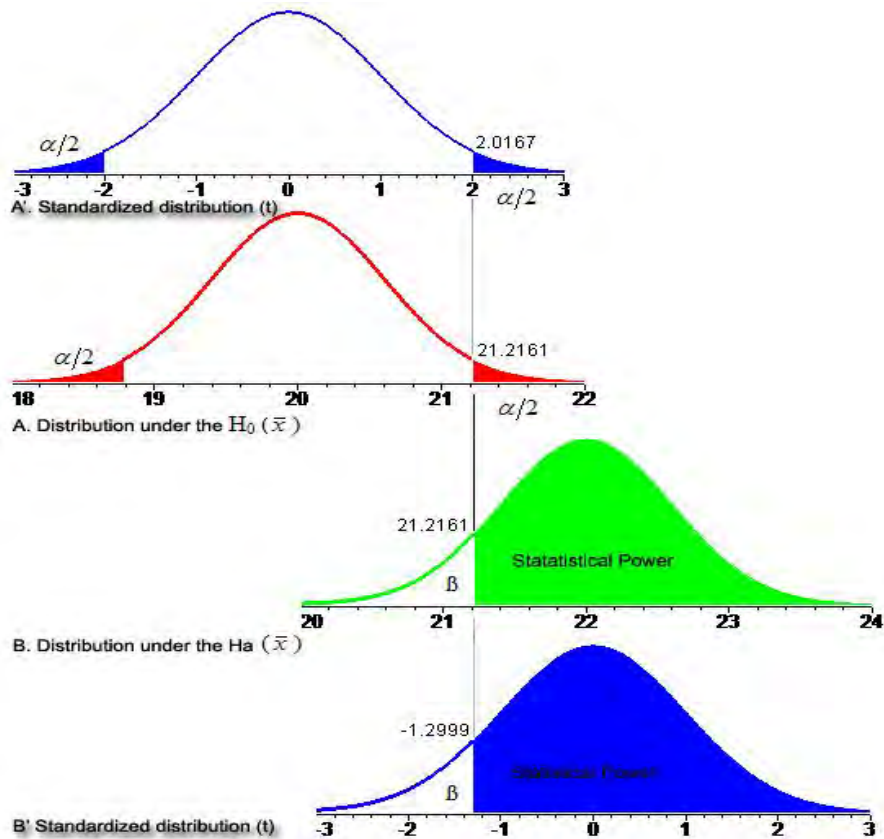
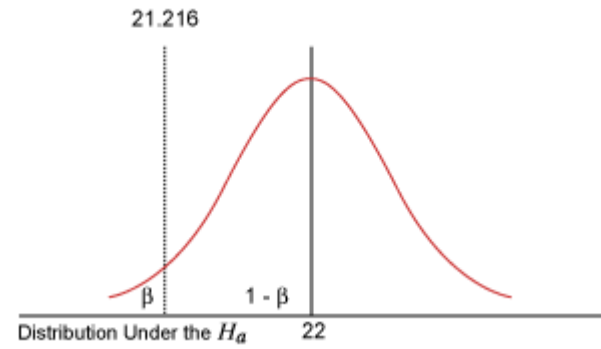
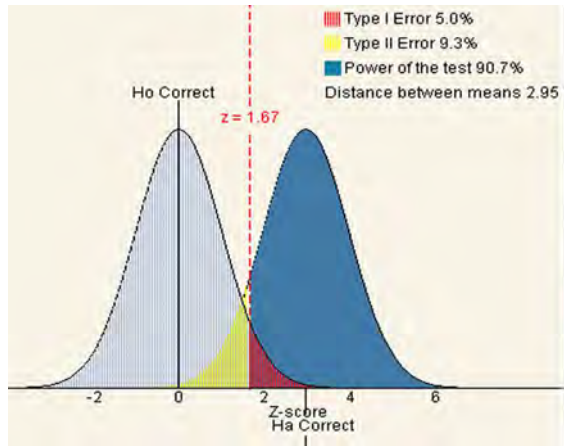
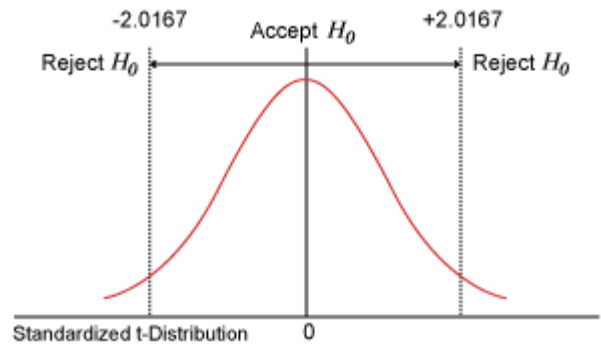
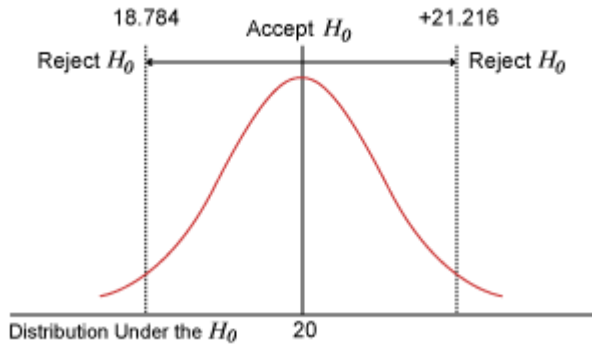
خطای نوع دوم ( $\beta$ )

	عدم رد $H_0$	رد $H_0$
صحیح $H_0$	$1 - \alpha$	$\alpha$
غلط $H_0$	$\beta$	$1 - \beta$

سطح اطمینان ۹۵% یعنی اگر ۱۰۰ نمونه بگیریم میانگین ۹۵ تا آن ( $\bar{x}$ ) در فاصله اطمینان می افتد. از آزمونهایی باید استفاده کنیم که خطای نوع اول کمتری داشته باشند در پزشکی (۰.۰۱) و در صنعتی (۰.۰۵)

پرتوان ترین آزمون آزمونی است که  $1 - \beta$  آن بیشترین باشد. یعنی رد فرض صفر به شرط غلط بودن آن.

سؤال : در باره رابطه  $\alpha + \beta < 1$  تحقیق کنید.



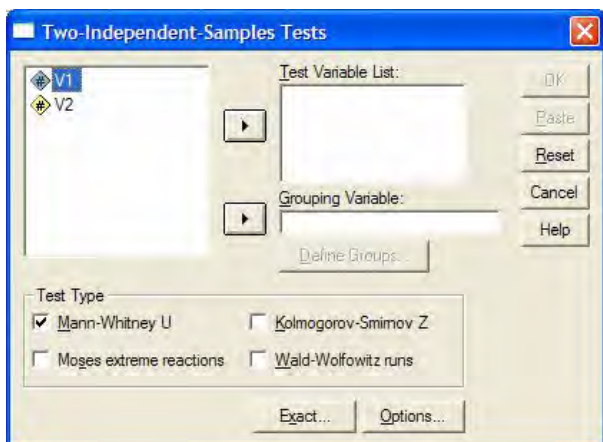
برای انجام آزمون های ناپارامتری در SPSS از دستور زیر استفاده می کنیم :

Analyze – Nonparametric Tests – 1-sample k-s...

این آزمون با استفاده از روش کولموگروف – اسمیرنوف ( Kolmogorov- Smirnov Z ) نشان میدهد که توزیع متغیر انتخاب شده چه نوعی است؟ نرمال، یکنواخت، پواسون یا نمایی  
 آزمون ناپارامتری از رتبه بندی داده ها و میانه استفاده می کند، میانه را بدست می آورد و اختلاف رتبه میانه را با داده ها محاسبه می کند و اگر خیلی اختلاف داشته باشد در نتیجه آزمون در Most Extreme می نویسد. ( اولین مقدار اختلاف با عنوان Absolute مقدار مطلق بزرگترین دو مقدار اختلاف چاپ شده در زیر آن می باشد لازم دارد.)

برای آزمون مقایسه میانگین بین دو نمونه وقتی تعداد نمونه کم یا توزیع نرمال نباشد از دستور زیر استفاده می کنیم:

Analyze – Nonparametric Tests – 2 Independent samples



**انواع آزمون:**

۱- کولموگروف – اسمیرنوف ( Kolmogorov- Smirnov Z ) علاوه بر شاخص مرکزی شکل توزیع را نیز برای مقایسه در نظر می گیرد. در صورتی که توزیع هر دو نمونه شبیه نباشد این آماره مناسبتر است.

۲- من ویتنی ( Man-Whitney U ) برای مقایسه شاخص های مرکزی مناسب است و با استفاده از جمع رتبه ها آماره را حساب می کند.

۳- موزس ( Moses extreme reactions ) بر مقدار انتهایی و رتبه آنها تاکید دارد و مقایسه بین گروه شاهد و آزمایش را روی مقدار انتهایی انجام می دهد.

۴- والد ( Wald wolfowitz runs ) از رتبه های مخلوط استفاده می کند و مانند کولموگروف شکل توزیع را مقایسه می کند.

برای آزمون مقایسه میانگین بین چند گروه از دستور زیر استفاده می کنیم:

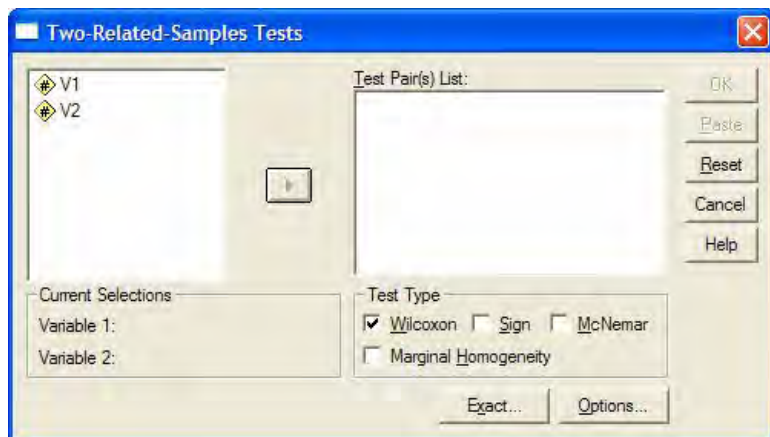
Analyze – Nonparametric Tests – K Independent samples

مشابه این آزمون ها در حالت نرمال هم بود ولی در آنجا نمونه ها مستقل بودند ولی اگر نمونه ها مرتبط باشند از دستورات زیر استفاده می کنیم. مثلاً آزمایش تاثیر دارو برای دو نفر نمونه های مستقل است اما برای یک نفر قبل و بعد از دارو مرتبط است و از هم تاثیر می پذیرند.



### آزمون مقایسه میانگین جفت ها :

Analyze – Nonparametric Tests – 2 Related samples



#### انواع آزمون :

۱. Wilcoxon وقتی است که داده ها پیوسته باشند مثل زمان و در این آزمون اختلاف رتبه ها و جهت اختلاف رتبه ها در نظر گرفته می شود. ( استفاده از این آزمون ارجح است).

۲. Sign مثل بالایی است ولی فقط جهت اختلاف رتبه ها در نظر گرفته می شود.

۳. McNemar اگر داده ها به صورت درست و غلط هستند (۱ و ۰) بهتر است از این آزمون استفاده کنید. معمولا از این آزمون

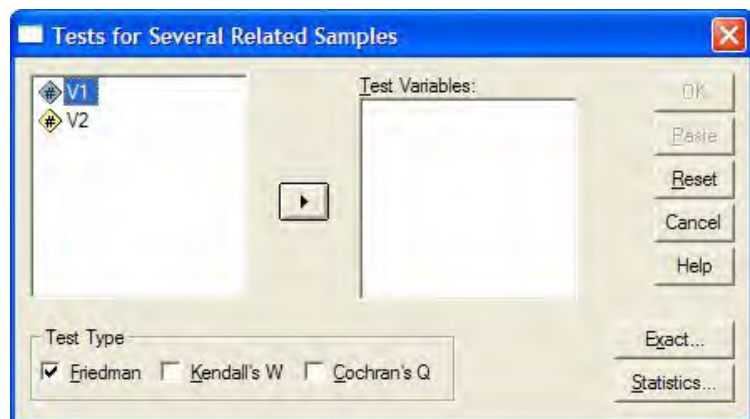
برای مقایسه قبل و بعد از یک رفتار به کار گرفته می شود.

اگر داده های چندتایی باشند (تعمیمی از روش McNemar) از این روش استفاده کنید. داده ها، گروه بندی هستند.

### آزمون مقایسه میانگین نمونه های چندتایی :

Analyze – Nonparametric Tests – K Related samples

#### انواع آزمون :



۱. آزمون فریدمن مانند روشهای پارامتری مقایسه میانگین (توزیع) چند متغیر را آزمون می کند. با استفاده از رتبه گذاری برای هر گروه و مقایسه رتبه ها آزمون انجام می شود.

۲. آزمون کندال با استفاده از توافق (همبستگی رتبه ای) این آزمون را انجام می دهد.

۳. آزمون کاکران نیز مانند آزمون فریدمن عمل کرده ولی با داده های دوتایی (۱ و ۰) محاسبات را انجام می دهد. برای مقایسه درست و غلط مناسب است.

مثال : زمانی را که ۱۰ بچه صرف میکنند تا قطعه مناسب را در جای خود بگذارند بر حسب ثابته ثبت کرده ایم.

مثلث	دایره	مربع
۲۲۰	۳۰۰	۲۶۰
۲۵۰	۲۹۰	۳۰۰
۲۶۰	۲۸۰	۲۹۰
۲۳۰	۳۴۰	۱۹۰
۱۹۰	۳۰۰	۲۵۰
۲۲۰	۲۷۰	۲۴۰
۲۵۰	۳۲۰	۲۷۰
۲۸۰	۲۹۰	۲۶۰
۲۷۰	۳۴۰	۲۵۰
۲۴۰	۳۰۰	۲۵۰

۱. آمارهای توصیفی برای زمان تشخیص اشیاء را محاسبه نمایید. ( معیارهای تمرکز و پراکندگی )
۲. یک نمودار مقایسه ای برای این سه زمان رسم کنید.
۳. آیا زمان تشخیص برای هر سه شکل دارای توزیع نرمال است؟
۴. با استفاده از نمودار Boxplot نشان دهید بین میانگین زمان تشخیص در این سه گروه اختلاف معنی داری وجود دارد.
۵. با استفاده از آزمون های آماری فرض بالا را آزمون کنید. ( پارامتری و ناپارامتری).

جواب:

برای ثبت داده ها و تعیین متغیرها دو راه وجود دارد:  
حالت اول : دو متغیر در نظر بگیریم اولی متغیر شکل شامل ( مربع، دایره، مثلث) و دومی زمان که در اینصورت ۳۰ مورد داریم.  
حالت دوم : سه متغیر در نظر بگیریم مربع ، دایره، مثلث که در این حالت ۱۰ مورد داریم.

۱. برای آمارهای توصیفی از دوره می توان استفاده کرد:

Analyze – Descriptive statistic- descriptive

Analyze – Descriptive statistic- Explore

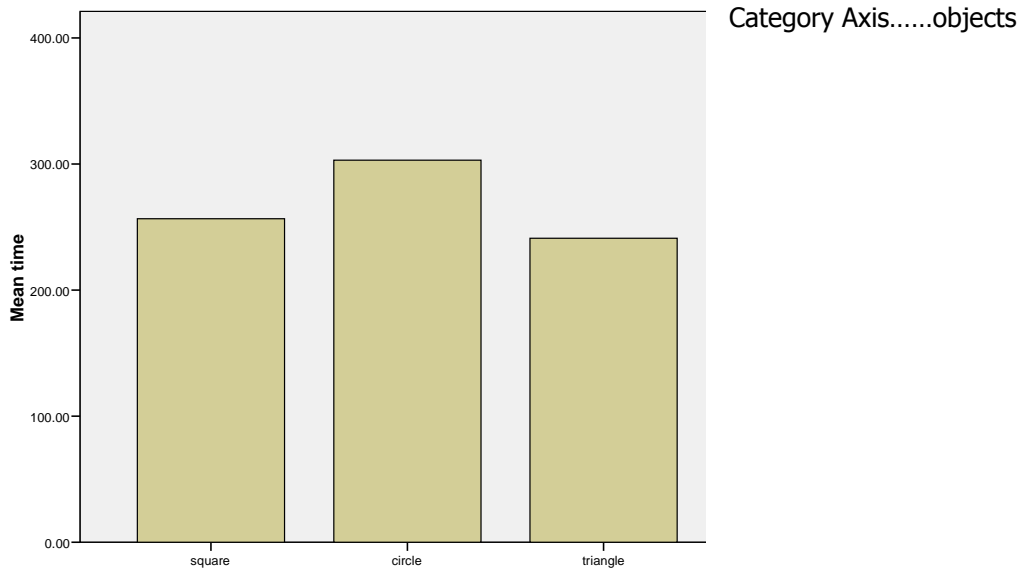
فرق این دوره این است که با استفاده از Explore میتوان متغیر زمان را در سه قسمت objects جداگانه مشاهده نمود.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
square	10	110.00	190.00	300.00	256.0000	29.88868	893.333
triangle	10	90.00	190.00	280.00	241.0000	26.85351	721.111
circle	10	70.00	270.00	340.00	303.0000	23.59378	556.667
Valid N (listwise)	10						

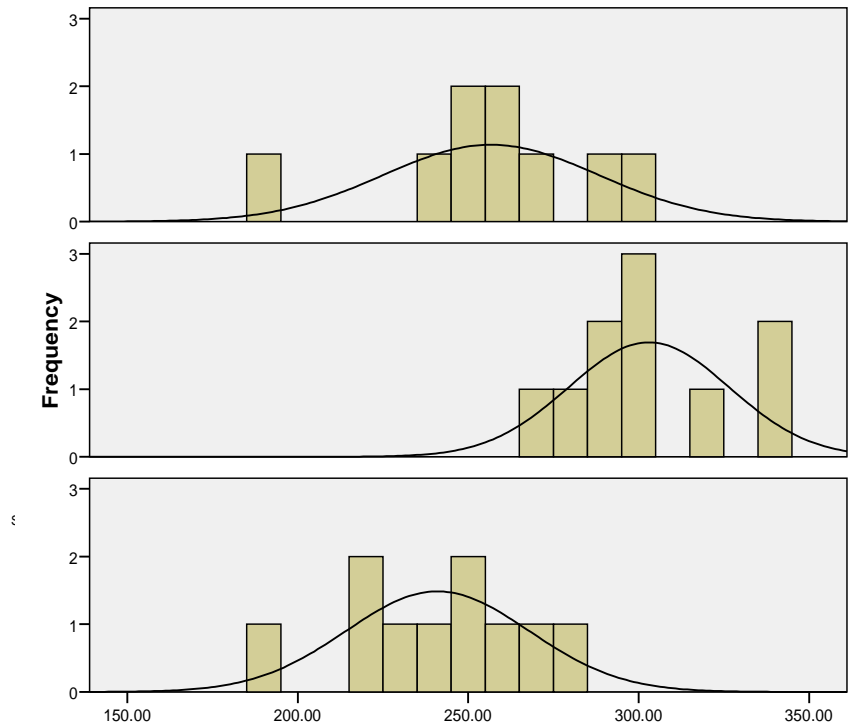
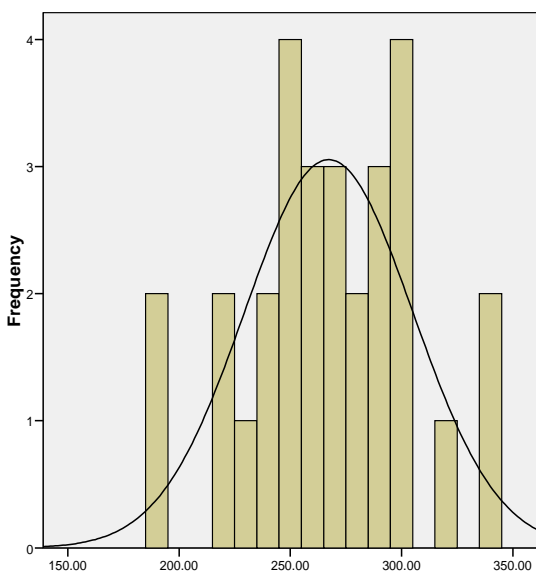
۲. نمودار مقایسه ای:

Graph – Legacy Dialogs- Bar – simple - Define –other statistic- variable .....time



۳. وجود توزیع نرمال:

Graph – Legacy Dialogs - Histogram-display normal curve- variable.....time  
Panel.....objects



Analyze – Nonparametric Tests – 1-sample k-s...

الف-حالت اول(بدون گروه بندی) :

One-Sample Kolmogorov-Smirnov Test

objects		time
N		29
Normal Parameters(a,b)	Mean	267.2414
	Std. Deviation	37.88107
Most Extreme Differences	Absolute	.090
	Positive	.090
	Negative	-.083
Kolmogorov-Smirnov Z		.485
Asymp. Sig. (2-tailed)		.973

a Test distribution is Normal.b Calculated from data.

نتیجه :

۱. با توجه به بزرگتر بودن  $p-value$  از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد نمی شود.

$$Sig. = p - value = 0.973 > 0.05 = \alpha$$

Data- split files – variable ....objects

ب- حالت دوم ( با گروه بندی ) :

Analyze – Nonparametric Tests – 1-sample k-s...

One-Sample Kolmogorov-Smirnov Test

objects		time	
square	N	10	
	Normal Parameters(a,b)	Mean	256.6667
		Std. Deviation	31.62278
	Most Extreme Differences	Absolute	.194
		Positive	.125
		Negative	-.194
	Kolmogorov-Smirnov Z		.655
	Asymp. Sig. (2-tailed)		.784
circle	N	10	
	Normal Parameters(a,b)	Mean	303.0000
		Std. Deviation	23.59378
	Most Extreme Differences	Absolute	.251
		Positive	.251
		Negative	-.142
	Kolmogorov-Smirnov Z		.792
	Asymp. Sig. (2-tailed)		.556
Triangle	N	10	
	Normal Parameters(a,b)	Mean	241.0000
		Std. Deviation	26.85351
	Most Extreme Differences	Absolute	.131
		Positive	.083
		Negative	-.131
	Kolmogorov-Smirnov Z		.415
	Asymp. Sig. (2-tailed)		.995

a Test distribution is Normal.

b Calculated from data.

نتیجه :

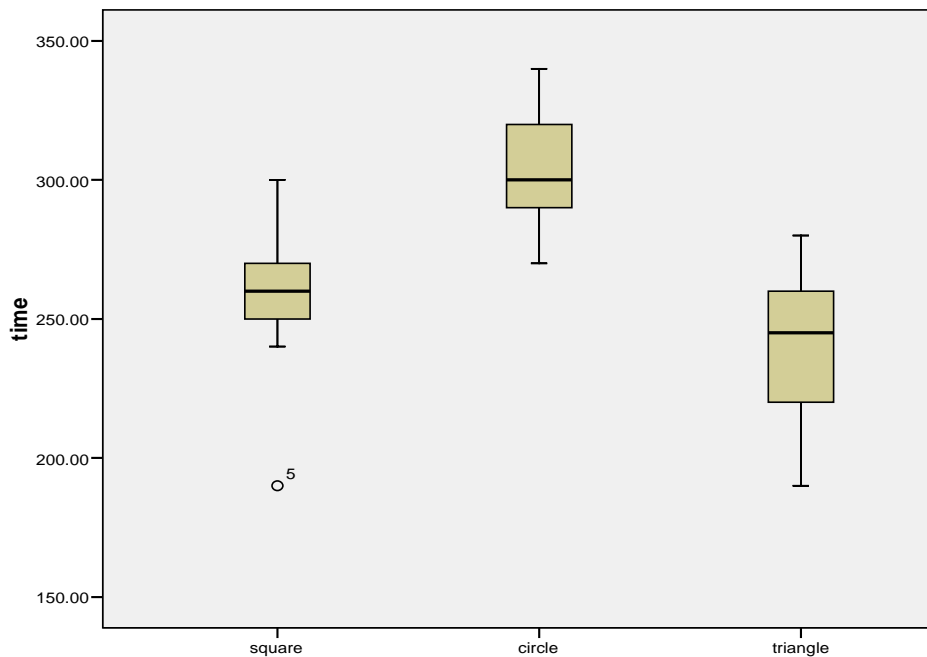
۱. با توجه به بزرگتر بودن  $p-value$  از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد نمی شود. ( در گروه مربع )  
Sig. =  $p-value = 0.784 > 0.05 = \alpha$

۲. با توجه به بزرگتر بودن  $p-value$  از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد نمی شود. ( در گروه دایره )  
Sig. =  $p-value = 0.556 > 0.05 = \alpha$

۳. با توجه به بزرگتر بودن  $p-value$  از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد نمی شود. ( در گروه مثلث )  
Sig. =  $p-value = 0.995 > 0.05 = \alpha$

۴. اختلاف میانگین زمان در سه گروه با نمودار Boxplot :  
نکته : در حالت split files نباشد.

Graph – Legacy Dialogs Boxplots - variable.....time  
category.....objects



نتیجه: میانگین متغیر زمان در سه گروه اشکال متفاوت نشان میدهد ولی قضاوت با نمودار کافی نیست. باید آزمون کرد.

۵. آزمون :

$$\begin{cases} H_0 : m_1 = m_2 = m_3 \\ H_1 : m_1 \neq m_2 \neq m_3 \end{cases}$$

**حالت اول- ناپارامتری :**

$m_1$  میانگین متغیر زمان در گروه مربع (Square)  
 $m_2$  میانگین متغیر درآمد در گروه دایره (circle)  
 $m_3$  میانگین متغیر درآمد در گروه مثلث (Triangle)

 Analyze – Nonparametric Tests – K Independent samples- test variable..... time  
 Grouping variable.....object(1,3)

**Kruskal-Wallis Test**

Ranks			
	objects	N	Mean Rank
time	square	10	13.20
	circle	10	24.40
	triangle	10	8.90
	Total	30	

Test Statistics(a,b)	
	time
Chi-Square	16.683
df	2
Asymp. Sig.	.000

a Kruskal Wallis Test  
 b Grouping Variable: objects

نتیجه :

 ۱. با توجه به کوچکتر بودن  $p$ -value از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد می شود.

$$\text{Sig.} = p\text{-value} = 0.000 < 0.05 = \alpha$$

**حالت دوم- پارامتری :**

 Analyze - Compare Means - One-Way Anova  
**ANOVA**

time					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20720.000	2	10360.000	14.337	.000
Within Groups	19510.000	27	722.593		
Total	40230.000	29			

**Test of Homogeneity of Variances**

time			
Levene Statistic	df1	df2	Sig.
.067	2	27	.936

نتیجه :

 ۱. با توجه به کوچکتر بودن  $p$ -value از  $\alpha$  ، فرض  $H_0$  ( نرمال بودن توزیع متغیر زمان ) رد می شود.

$$\text{Sig.} = p\text{-value} = 0.000 < 0.05 = \alpha$$

۲. اختلاف معنی دار است و ناشی از گروه بندی اشیاست.

 ۳. نتایج بدست آمده از آزمون لون مقایسه واریانسها در جدول **Test of Homogeneity of Variances** نشان میدهد که واریانس های متغیر زمان در ۳ گروه متغیراشیا اختلاف معنی داری ندارند و  $\text{Sig.} > \alpha = 0.936 = p\text{-value}$ 

 ۴. نتایج بدست آمده از آنالیز واریانس در جدول **ANOVA** نشان میدهد که باید به دنبال اختلافها میانگینها با فرض برابری واریانسها باشیم و با استفاده از گزینه **Post HOC** دویه دو با هم مقایسه کنیم.

برای یافتن علت اختلاف از دستور postHoc استفاده می کنیم:

Multiple Comparisons

Dependent Variable: time

Tukey HSD

(I) objects	(J) objects	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
square	circle	-46.00000*	12.02159	.002	-75.8065	-16.1935
	triangle	16.00000	12.02159	.391	-13.8065	45.8065
circle	square	46.00000*	12.02159	.002	16.1935	75.8065
	triangle	62.00000*	12.02159	.000	32.1935	91.8065
triangle	square	-16.00000	12.02159	.391	-45.8065	13.8065
	circle	-62.00000*	12.02159	.000	-91.8065	-32.1935

\*. The mean difference is significant at the .05 level.

$$m_1 = m_3$$

$$m_1 \neq m_2$$

$$m_2 \neq m_3$$

نتیجه آزمون مقایسه دو به دو:

۱. میانگین زمان در گروه های مربع و دایره برابر نیست.

۲. میانگین زمان در گروه های مثلث و دایره برابر نیست.

۳. میانگین زمان فقط در گروه های مربع و مثلث برابر است.

۴. علت اختلاف گروه دایره می باشد.

$m_1$  میانگین متغیر زمان در گروه مربع (Square)

$m_2$  میانگین متغیر درآمد در گروه دایره (circle)

$m_3$  میانگین متغیر درآمد در گروه مثلث (Triangle)

## « جلسه ششم – ۸۷/۱/۲۹ »

 آزمون Chi-square ( $\chi^2$ ) :

- این آزمون برای دو منظور به کار میرود:
۱. تطابق توزیع (نمودار از توزیع خاصی پیروی می کند یا نه؟)
  ۲. استقلال دو متغیر تصادفی (متغیرهای گسسته)

$O_i$  مقادیر مشاهده شده

$E_i$  مقادیر مورد انتظار

(متوسط اختلاف مشاهده شده از سوی نمونه مقادیر مورد انتظار)

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

- $$\left\{ \begin{array}{l} H_0 : \text{نمونه پارامتر جمعیت را تایید می کند} \\ H_1 : \text{نمونه پارامتر جمعیت را تایید نمی کند} \end{array} \right.$$

اگر نمونه همان توزیع جمعیت را داشت پس اختلافات کم می شود و کم بودن مقدار  $\chi^2$  نمونه جمعیت را تایید می کند. البته مقدار آن با جدول آماری مقایسه می شود.

- آزمون استقلال – نیکویی برازش
- فرضها
  - نمونه تصادفی
  - شکل یا توزیع داده ها اهمیت ندارد.
  - حداقل فراوانی در هر گروه حداقل ۱
  - ۲۰٪ گروه ها فراوانی بیش از ۵ داشته باشند.

این آزمون با استفاده از آماره  $\chi^2$  مقادیر مشاهده شده را با مقادیر مورد انتظار مقایسه می کند.  
روش دسترس به آزمون :

Analyze - Descriptive Statistics - Crosstabs... .. Analyze

بعد از تعیین متغیرها دکمه statistic را بزنید و گزینه chi square را انتخاب کنید.

Analyze – nonparametric – chi-square

یا



مثال: فرض کنید توزیع یکنواخت است و احتمال مشاهده هر یک از داده ها با دیگری برابر است. ه گروه اسباب بازی داریم :  
 سئوال اول - می خواهیم بدانیم تمایل بچه ها به خریدن اینها یکسان است یا نه؟  
 سئوال دوم - می خواهیم مقایسه بین سه گروه اول انجام شود.  
 سئوال سوم- آیا نسبت خرید بین تفنگ و بازی کامپیوتر یک سوم به دو سوم است؟

سه چرخه	ماشین	عروسک	بازی کامپیوتر	تفنگ	نوع اسباب بازی
۵	۱۵	۱۵	۲۰	۴۵	مقدار خریداری شده

جواب اول :

۱. دو نوع متغیر به نام Typeoftoy و num تعریف می کنیم.
۲. متغیر اول را در ۵ نوع اسباب بازی کد می دهیم.
۳. به متغیر num وزن می دهیم.

Analyze – nonparametric – chi-square – Test variable .....Typeoftoy

۴. مقدار Chi-Square(a) عدد ۴۵ شده است.

typeoftoy

	Observed N	Expected N	Residual
gun	45	20.0	25.0
camputer	20	20.0	.0
doll	15	20.0	-5.0
car	15	20.0	-5.0
bycycle	5	20.0	-15.0
Total	100		

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} = \frac{25^2 + 0^2 + 5^2 + 5^2 + 15^2}{20} = \frac{900}{20} = 45$$

۶. مقدار Asymp. Sig. = 0.000 و کوچکتر از 0.05 است

لذا نمونه فرض صفر رد می کند و توزیع نمونه با پارامتر جمعیت یکی نیست.

Test Statistics

	typeoftoy
Chi-Square(a)	45.000
df	4
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 20.0.

جواب دوم:

در پنجره آزمون Chi-Square در قسمت Expected Range ، گزینه Used specified range ، برای lower=1 و برای upper=3 را انتخاب کرده تا مقایسه بین گروه ۱ تا ۳ صورت گیرد.  
 اگر گروههای مورد نظر ما مرتب نبود بایستی اول مرتبشان کنیم.  
 میتوان داده هارا با دستور select cases اول انتخاب کرد بعد آزمون را انجام داد.  
 با انجام این آزمون همچنان فرض صفر توسط نمونه رد شد.

**Frequencies**

	typeoftoy			
	Category	Observed N	Expected N	Residual
1	gun	45	26.7	18.3
2	camputer	20	26.7	-6.7
3	doll	15	26.7	-11.7
Total		80		

**Test Statistics**

	typeoftoy
Chi-Square(a)	19.375
df	2
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 26.7.

جواب سوم:

نسبت یک سوم به دو سوم را اینطور حساب می کنیم که مجموع خرید تفنگ و بازی کامپیوتر ۶۵ است این عدد را به نسبت یک سوم به دو سوم تقسیم می کنیم:  
 در پنجره آزمون Chi-Square در قسمت Expected Range ،  
 گزینه Used specified range ، برای 1=lower و برای  
 upper=2 را انتخاب کرده تا مقایسه بین گروه ۱ ، ۲ صورت گیرد.  
 در پنجره آزمون Chi-Square در قسمت Expected Value اعداد  
 ۲۱.۷ و ۴۳.۳ را وارد می کنیم.

$$65 * \frac{1}{3} = 21.7$$

$$65 * \frac{2}{3} = 43.3$$

**Frequencies**

	typeoftoy			
	Category	Observed N	Expected N	Residual
1	gun	45	21.7	23.3
2	camputer	20	43.3	-23.3
Total		65		

مقدار Asymp. Sig. = 0.655 و بزرگتر از 0.05 است  
 لذا نمونه فرض صفر رد نمی کند .

**Test Statistics**

	typeoftoy
Chi-Square(a)	.200
df	1
Asymp. Sig.	.655

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 21.7.

مثال: در دانشگاه‌های مرکز استان و شهرستان به تفکیک دانشکده آمار ثبت نام آورده شده است. آیا نوع دانشکده در انتخاب دانشجو تاثیر دارد یا نه؟ (نسبت قبولی توزیع یکنواخت دارد یا نه یعنی تعداد قبولی به یک نسبت بین دانشگاه‌ها تقسیم می شوند؟) آیا نسبت قبولی بین دانشکده مهندسی و هنر ۸۰ به ۲۰ است؟ (یعنی مهندسی ۸۰٪ و هنر ۲۰٪ توسط دانشجویان انتخاب می شود.)

نوع دانشکده	مهندسی	هنر	اقتصاد	سایر
مرکز استان	۱۶	۱۴	۱۳	۱۳
شهرستان	۱۴	۶	۱۰	۸

انتخاب متغیرها :

Name	Type	Width	Decimal	Label	Value	Missing	column	Align	Measure
1 faculty	Numeric	8	-	-	1=eng 2=art 3=eco 4=etc	-	10	left	Scale
place	Numeric	8	-	-	1=center 2=outer	-	10	left	Scale
2 frequency	Numeric	8	-	-	-	-	10	left	Scale

به متغیر frequency وزن می دهیم. برای آزمون توزیع یکنواخت از آزمون chi-square استفاده می کنیم.

Analyze – nonparametric – chi-square – Test variable ..... faculty

همانطور که مشاهده می شود جودش تعداد قبولی هارا جمع می زند و توزیع آن را آزمون می کند. نتیجه این است که نمونه فرض صفر را رد نمی کند و می توان گفت نمونه پارامترجمعیت را تایید می کند.

faculty

	Observed N	Expected N	Residual
engineering	30	23.5	6.5
art	20	23.5	-3.5
ecoomic	23	23.5	-.5
etc	21	23.5	-2.5
Total	94		

Test Statistics

	faculty
Chi-Square(a)	2.596
df	3
Asymp. Sig.	.458

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 23.5.

برای آزمون نسبت ۸۰ به ۲۰ :  
 در پنجره آزمون Chi-Square :  
 در قسمت Expected Range ، (1,2)  
 در قسمت Expected Value اعداد

$$50 * 0.8 = 40$$

$$50 * 0.2 = 10$$

۴۰ و ۱۰ را وارد می کنیم. یا نسبت آنها را به درصد یعنی خود ۰.۸ و ۰.۲ را وارد کنیم.  
**Frequencies**

facaulty				
	Category	Observed N	Expected N	Residual
1	engineering	30	40.0	-10.0
2	art	20	10.0	10.0
Total		50		

**Test Statistics**

facaulty	
Chi-Square(a)	12.500
df	1
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10.0.

نتیجه اینکه این فرض توسط نمونه رد شد.

**آزمون Binomial یا دو جمله ای :**

تعریف متغیر تصادفی دو جمله ای:

آزمایش تصادفی آزمایشی است که اگر در شرایط یکسان بارها تکرار شود نتایج متفاوتی داشته باشد  
 P حدس زده شده از روی جمعیت توسط نمونه تایید می شود یا نه؟

در این آزمون نسبت احتمال رخداد يك پیش آمد سنجیده می شود. از آنجایی که رخداد یا عدم رخداد  
 پیش آمد توزیع دو جمله ای (Binomial) از این توزیع برای آزمون استفاده می شود.  
 روش دسترسبی به این آزمون

Analyze - Nonparametric Tests - Binomial...

مثال: آیا نسبت قبولی دانشجویان در مرکز استان و شهرستانها ۵۰ به ۵۰ است؟

$$\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p \neq \frac{1}{2} \end{cases}$$

Analyze - Nonparametric Tests - Binomial... variable.....place  
 Test propertion.....0.5

**Binomial Test**

	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
place	Group 1	center	56	.60	.079(a)
	Group 2	outer	38	.40	
Total		94	1.00		

a Based on Z Approximation.

نتیجه نمونه فرض صفر را رد نمی کند.

راه دیگر استفاده از آزمون Chi-Square :

در قسمت Expected Range ، (1,2)

در قسمت Expected Value اعداد 0.5 و 0.5 را وارد می کنیم.

**نکته : آزمون Chi-Square عمومیت بیشتری دارد ولی دو جمله ای محدودتر است.**

مثال: آیا نسبت قبولی دانشجویان گروههای مهندسی و هنر با گروههای اقتصاد و سایر به نسبت ۷۰ به ۳۰ هست یا نه؟

- $H_0 : p =$  دانشجو مهندسی و هنر را انتخاب می کند
- $H_1 : q =$  دانشجو اقتصاد و سایر را انتخاب می کند

Analyze - Nonparametric Tests - Binomial... variable.....place  
 Test propertion.....0.7  
 Cutpoint .....2

وقتی cutpoint=2 است یعنی گروه ۱ و ۲ را یک گروه و بعد از ۲ را یک گروه دیگر حساب کند.

**Binomial Test**

	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
facaulty	Group 1	<= 2	50	.5	.000(a,b)
	Group 2	> 2	44	.5	
Total		94	1.0		

a Alternative hypothesis states that the proportion of cases in the first group < .7.

b Based on Z Approximation.

نتیجه اینکه نمونه فرض صفر را رد می کند.

**آزمون Runs :**

با استفاده از این آزمون و رتبه بندی داده ها امکان تصادفی بودن داده ها آزمون می شود. تصادفی بودن داده ها زمانی حاصل می شود که دو داده متوالی یکی بالای میانگین و یکی پایین تر از آن باشد.

روش دسترسبی به این آزمون

Analyze – Nonparametric Tests – Runs

با استفاده از اکسل ( تابع Rand() = ۱۰۰ عدد تصادفی ایجاد کنید و با امکانات spss آزمون کنید که این اعداد تصادفی هستند.

نحوه ورود اطلاعات از اکسل را مرور کنید. کافی است در هنگام باز کردن سند اطلاعاتی (data) در قسمت file گزینه excel را انتخاب کنید.

مثال: در چهار شیفت کاری متوسط قطر لوله های ساخته شده را اندازه گرفته اند می خواهیم ببینیم این اعداد نسبت به میانگینشان تصادفی هستند یا دستگاه نیاز به کالیبره شدن دارد. قطر لوله ها باید ۴ باشد.

متوسط قطرها	3.8	4.2	3.3	4.5
	شیفت ۱	شیفت ۲	شیفت ۳	شیفت ۴

### آزمون همبستگی :

رابطه همبستگی به بررسی ارتباط بین دو یا چند متغیر می پردازد و ضریب آن را محاسبه می کند. همبستگی بین متغیرها ممکن است مثبت یا منفی باشد. اگر تغییرات یک متغیر با تغییرات متغیر دیگری همراه باشد و افزایش یکی با افزایش دیگری یا بالعکس کاهش یکی با کاهش دیگری همراه بشود می گوئیم همبستگی بین آنها مثبت است مثل افزایش درآمد و تقاضا برای خرید. که از 0 تا +1 نوسان دارد.

اگر تغییر و افزایش یک متغیر با کاهش متغیر دیگری همراه شود گفته می شود که همبستگی بین آنها منفی است و از 0 تا -1 تغییر می کند. اگر بین دو متغیر رابطه ای وجود نداشته باشد ضریب همبستگی بین آنها صفر است. اگر ضریب همبستگی به 1 یا -1 نزدیک باشد رابطه خطی بین دو متغیر وجود دارد.

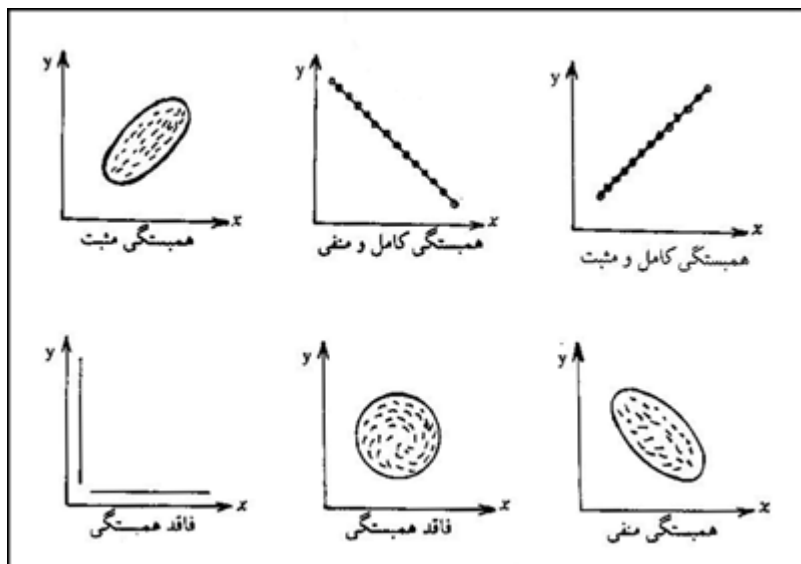
### ضریب همبستگی :

- این ضریب نشان دهنده وجود ارتباط بین دو متغیر است.
- هرچه ارتباط دو متغیر شدید باشد مقدار ضریب به +1 و یا -1 نزدیکتر است.
- با کاهش ارتباط بین دو متغیر مقدار ضریب به صفر نزدیک می شود.
- اگر دو متغیر مستقل باشند مقدار ضریب همبستگی برابر صفر خواهد بود.
- اگر دو متغیر نرمال باشند اگر ضریب همبستگی برابر صفر باشد می توان استقلال را نتیجه گرفت در غیر این صورت خیر.

### محاسبه ضریب همبستگی :

- برای محاسبه ضریب همبستگی پیرسون از مقدار Covariance استفاده می شود.
- مقدار Covariance برای اندازه گیری میزان ارتباط مناسب است ولی از آنجایی که با واحد و مقیاس مشخصه اندازه گیری شده بدست می آید باید برای مقایسه میزان همبستگی از یک ضریب بدون واحد (درصد) استفاده کرد.
- Correlation یا ضریب همبستگی این خاصیت را دارد.
- معمولا برای مشاهده ارتباط دو متغیر از نمودار نقطه ای استفاده می شود.
- استقلال ( Independent ) دو متغیر تصادفی یعنی :  $p(y/x) = p(y)$
- لازم نیست دو متغیر کاملاً همبسته علت و معلول هم باشند مثل قد و وزن بلکه رویدادشان با هم ارتباط دارد.

نمودار خط همبستگی:



محاسبه ضریب همبستگی پیرسون:

$$Co\ variance = \frac{E(xy) - E(x).E(y)}{n - 1}$$

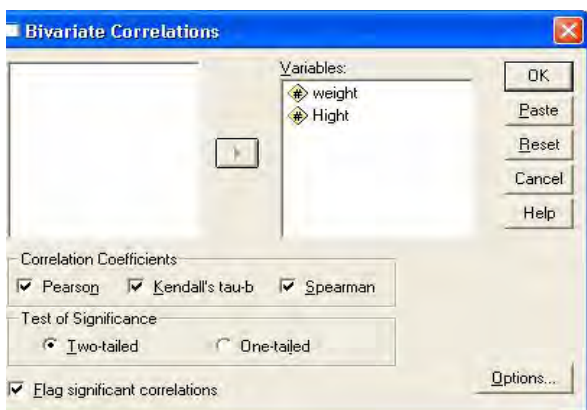
$$Corrolation = \rho = \frac{\frac{E(xy) - E(x).E(y)}{n - 1}}{\sqrt{\frac{Var(x)}{n - 1} \frac{Var(y)}{n - 1}}} = \frac{E(xy) - E(x).E(y)}{\sqrt{Var(x)Var(y)}}$$

اگر همبستگی وجود داشته باشد آنگاه  $\rho \neq 0$  و  $E(xy) \neq E(x)E(y)$

برای دسترسی به این دستور مسیر زیر را طی کنید. Analyze - Corrolate - Bivariate ...  
برای اجرای این دستور حداقل به دو متغیر عددی احتیاج دارید.

انواع ضریب همبستگی :

برای هر یک از مقیاسهای عددی ، رتبه ای و اسمی باید از روش محاسبه خودشان استفاده کرد.



- Covariance
- همبستگی Corrolation
- ضریب همبستگی پارامتری :
- همبستگی خطی (ضریب همبستگی پیرسون - Pearson)
- فرض بر این است که هر دو متغیر عددی دارای توزیع نرمال دو متغیره هستند.

ضریب همبستگی ناپارامتری :

- همبستگی رتبه‌ای (ضریب همبستگی اسپیرمن - Spearman)
- همبستگی ترتیبی (ضریب همبستگی کندال - Kendall's Tau\_b)

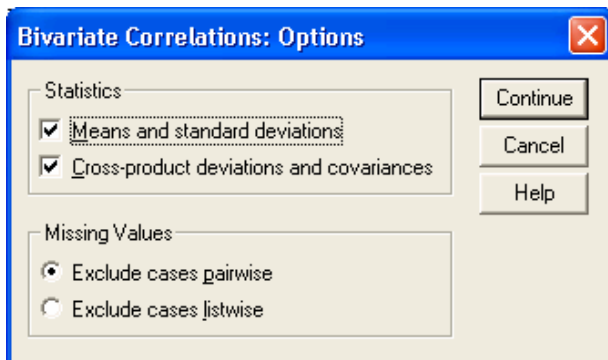
### آزمون ضریب همبستگی:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$
 - با انتخاب گزینه two tail يك آزمون دو طرفه برای مقدار ضریب همبستگی انجام می‌شود.

اگر علاوه بر مقدار ضریب همبستگی جهت آن نیز در نظر است بهتر است از آزمون یکطرفه ( One Tail) استفاده شود.

آزمون یکطرفه	آزمون یکطرفه
$H_0 : \rho = 0$	$H_0 : \rho = 0$
$H_1 : \rho > 0$	$H_1 : \rho < 0$

- با انتخاب گزینه flag... امکان تعیین معنادار بودن مقدار ضریب همبستگی در خروجی داده می‌شود.



- با انتخاب option امکان نمایش آمار توصیفی (میانگین و انحراف استاندارد) و نمایش مقدار کواریانس و مضرب تفاضلات وجود دارد. ضریب تفاضلات همان صورت محاسبه کواریانس است.

- با انتخاب گزینه Exclude Cases pairwise محاسبه برای جفت‌ها بدون در نظر گرفتن داده گمشده در دیگر متغیرهای هر مورد محاسبه می‌شود.

- با انتخاب گزینه Exclude cases listwise با وجود يك متغیر که مقدار گم شده در بررسی دارد کل مورد نادیده گرفته خواهد شد.

- امکان انتخاب بیش از دو متغیر نیز در این بررسی وجود دارد. به این ترتیب يك ماتریس دو بعدی برای بررسی و محاسبه مقدار ضریب همبستگی جفت جفت متغیرها ایجاد خواهد شد.

برای رسم نمودار همبستگی از دستور زیر استفاده می‌کنیم:

Graph – Legacy dialogs – scatter plot– simple یا Matrix



« جلسه هفتم – ۸۷/۲/۵ »

در جلسه قبل گفته شد که آزمون Chi-square بهم برای برازش توزیع و هم برای آزمون استقلال دو متغیر تصادفی بکار می رود. یعنی برای زوج مرتب (x,y) اگر  $\rho \neq 0$  بود نشانه وجود رابطه بین دو متغیر می باشد.

در مثال جلسه قبل مقایسه نسبت قبولی دانشجویان در دانشکده های مختلف از مرکز استان و شهرستانها می خواهیم بدانیم که محل دانشکده و رشته های مختلف که هر دو متغیر اسمی هستند مستقلند یا نه؟

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Analyze –Descriptive Statistics- Crosstabs- statistic.....chi-square

faculty \* place Crosstabulation

Count		place		Total
		center	outer	
faculty	engineering	16	14	30
	art	14	6	20
	ecoomic	13	10	23
	etc	13	8	21
Total		56	38	94

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.524(a)	3	.677
Likelihood Ratio	1.551	3	.671
Linear-by-Linear Association	.153	1	.696
N of Valid Cases	94		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.09.

وقتی مقدار Pearson Chi-Square=1.524 کوچک است یعنی اختلافات همدیگر را خنثی کرده اند و اگر بزرگ باشد نشان دهنده رابطه بین دو متغیر است.

Asymp. Sig. (2-sided)=0.677 که بزرگتر از 0.05 بوده و نشان می دهد که نمونه فرض صفر را رد نمی کند یعنی وجود استقلال بین دو متغیر.

می خواهیم بدانیم که عدد 1.524 از کجا به دست آمده است؟ فرمول محاسبه به این شرح است:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$O_i$  مقادیر مشاهده شده ( ستون center و ستون outer )  
 $E_i$  مقادیر مورد انتظار  
 ( متوسط اختلاف مشاهده شده از سوی نمونه مقادیر مورد انتظار )

برای ستون شهرستان ( outer )

$$E_1 = \frac{30 * 38}{94} = 12.13$$

$$E_2 = \frac{20 * 38}{94} = 8.09$$

$$E_3 = \frac{23 * 38}{94} = 9.30$$

$$E_4 = \frac{21 * 38}{94} = 8.49$$

برای ستون مرکز استان (center)

$$E_1 = \frac{30 * 56}{94} = 17.87$$

$$E_2 = \frac{20 * 56}{94} = 11.91$$

$$E_3 = \frac{23 * 56}{94} = 13.7$$

$$E_4 = \frac{21 * 56}{94} = 12.51$$

برای محاسبه فراوانی مورد انتظار از مجموع فراوانی های سطر و ستوت از فرمول زیر استفاده می شود :

$$E_i = \frac{A * B}{N}$$

A جمع فراوانی مشاهده شده در سطر  
B جمع فراوانی های مشاهده شده در ستون  
N جمع فراوانی مشاهده شده  
E<sub>i</sub> فراوانی مورد انتظار i ام

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(16-17.87)^2}{17.87} + \frac{(14-11.91)^2}{11.91} + \frac{(13-13.7)^2}{13.7} + \frac{(13-12.51)^2}{12.51} = 0.62 \quad \text{مرکز استان :}$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(14-12.13)^2}{12.13} + \frac{(6-8.09)^2}{8.09} + \frac{(10-9.30)^2}{9.30} + \frac{(8-8.49)^2}{8.49} = 0.91 \quad \text{شهرستان :}$$

Pearson Chi – Square=0.62+0.91=1.524

نرم افزار SPSS می تواند مقادیر مورد انتظار یعنی E<sub>i</sub> ها را حساب کند.

Analyze –Descriptive Statistics- Crosstabs- statistic.....chi-square   
Cell .....observed  , expected

place \* faculty Crosstabulation

			faculty				Total
			engineering	art	ecoomic	etc	engineering
place	center	Count	16	14	13	13	56
		Expected Count	17.9	11.9	13.7	12.5	56.0
	outer	Count	14	6	10	8	38
		Expected Count	12.1	8.1	9.3	8.5	38.0
Total		Count	30	20	23	21	94
		Expected Count	30.0	20.0	23.0	21.0	94.0

مثال: در فایل demo استقلال دو متغیر میزان تحصیلات و نوع ماشین (ed,carcat) را بررسی کنید.

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Analyze –Descriptive Statistics- Crosstabs- statistic.....chi-square   
correlation

Level of education \* Primary vehicle price category Crosstabulation

Count		Primary vehicle price category			Total
		Economy	Standard	Luxury	Economy
Level of education	Did not complete high school	483	481	426	1390
	High school degree	587	689	660	1936
	Some college	390	485	485	1360
	College degree	312	508	535	1355
	Post-undergraduate degree	69	112	178	359
Total		1841	2275	2284	6400

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	85.689(a)	8	.000
Likelihood Ratio	85.520	8	.000
Linear-by-Linear Association	75.149	1	.000
N of Valid Cases	6400		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 103.27.

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Interval by Interval	Pearson's R	.108	.012	8.720	.000(c)
Ordinal by Ordinal	Spearman Correlation	.105	.012	8.453	.000(c)
N of Valid Cases		6400			

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

c Based on normal approximation.

نتیجه :  $p - value = 0$  پس نمونه فرض صفر را رد می کند و استقلال دو متغیر رد می شود میزان همبستگی در جدول چون دو متغیر اسمی هستند اسپیرمن مناسب است که عدد 0.105 را نشان می دهد.

### ضریب همبستگی جزئی:

با استفاده از این ضریب امکان سنجش میزان همبستگی خطی بین دو متغیر در صورت کنترل یک متغیر سوم (یا حذف تاثیر آن) وجود دارد.

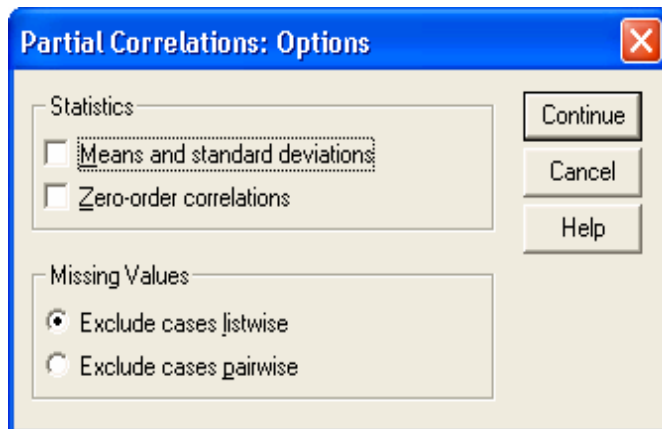
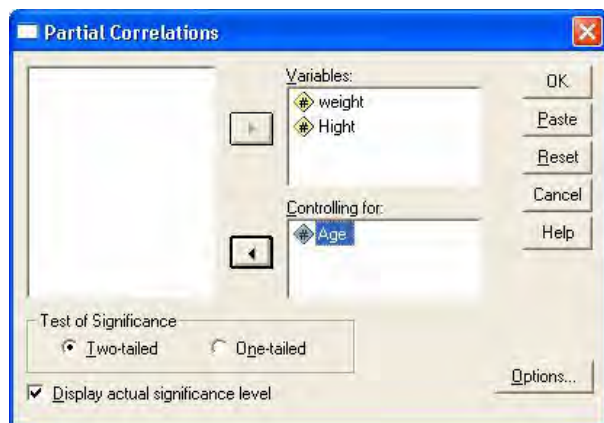
ممکن است دو متغیر به واسطه متغیر دیگری نیز رابطه داشته باشند. با محاسبه ضریب همبستگی جزئی میزان رابطه خطی بین دو متغیر با حذف اثر متغیر سوم محاسبه خواهد شد. روش محاسبه به صورت زیر است:

$$r_{ABC} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}}$$

برای انجام این دستور احتیاج به حداقل ۳ متغیر عددی است. برای دسترسی به این دستور مسیر زیر را طی کنید.

Analyze - Correlate - Partial...

با انتخاب گزینه Option امکان نمایش آماره های توصیفی و نمایش جدول ضریب همبستگی ساده (Zero-order...) نیز وجود دارد.



مثال- در فایل demo سه متغیر سن ، درآمد و هزینه خرید ماشین ( age,income,car) را در نظر بگیرید ابتدا نمودار scatterplot را به طور ماتریسی برای این سه متغیر بکشید.

Grafh - scatterplot - matrix

ضریب همبستگی بین سه متغیر را حساب می کنیم:

Analyze - correlate- bivariate ...

با استفاده از ضریب همبستگی جزئی متغیر income را ثابت گرفته و مجددا رابطه بین دو متغیر age,car را بدست می آوریم.

Correlations

		Age in years	Household income in thousands	Price of primary vehicle
Age in years	Pearson Correlation	1	.335(**)	.376(**)
	Sig. (2-tailed)		.000	.000
	N	6400	6400	6400
Household income in thousands	Pearson Correlation	.335(**)	1	.792(**)
	Sig. (2-tailed)	.000		.000
	N	6400	6400	6400
Price of primary vehicle	Pearson Correlation	.376(**)	.792(**)	1
	Sig. (2-tailed)	.000	.000	
	N	6400	6400	6400

\*\* Correlation is significant at the 0.01 level (2-tailed).

ضریب همبستگی age و income ..... 0.335  
ضریب همبستگی age و car ..... 0.376  
ضریب همبستگی car و income ..... 0.792

کنترل income:

Correlations

Control Variables			Age in years	Price of primary vehicle
Household income in thousands	Age in years	Correlation	1.000	.193
		Significance (2-tailed)	.	.000
		df	0	6397
Price of primary vehicle		Correlation	.193	1.000
		Significance (2-tailed)	.000	.
		df	6397	0

ضریب همبستگی age و car ..... 0.193

نتیجه : در حالت اول ضریب همبستگی بین دو متغیر age,car ، 0.376 بود ولی پس از کنترل متغیر income ، به 0.193 تغییر پیدا کرد که این عدد ارتباط خالص را نشان می دهد.

**رگرسیون Regression ( معادله خط برگشت ) :**

برای سنجش ارتباط دو متغیر سه راه وجود دارد:

- ۱. رسم نمودار ( scatterplot )
- ۲. اندازه گیری ضریب همبستگی ( Corrolation )
- ۳. محاسبه رگرسیون ( خط برگشت )

رگرسیون پرکاربردترین روش آماری است که برای سنجش و ارائه مدل ارتباط دو متغیر بکار می رود. رگرسیون خطی، ضرایب معادله خطی که نزدیکترین هماهنگی با داده‌های مشاهده شده را دارا است برآورد می‌کند.

با استفاده از معادله خط برگشت امکان پیشگویی مقادیر بعدی نیز وجود دارد. در معادله رگرسیون یک متغیر مستقل (Independent) و یک متغیر وابسته (Dependent) داریم. در معادله رگرسیون چند متغیره تعداد متغیرهای مستقل بیش از یکی است.

مثال:

- برآورد میزان فروش یک کالای لوکس براساس درآمد خانوار یک کشور
- برآورد میزان فروش فروشندگان یک فروشگاه براساس، سن، تجربه، میزان تحصیلات.
- برآورد میزان محصول براساس میزان بارندگی، نوع کود، نوع بذر و استفاده از روشهای مخصوص رفع آفات
- برآورد مساحت روشن شدن یک فضا توسط تیر چراغ برق براساس ارتفاع تیر

**مدل خطی ساده :**

$$Y / x = a + bX + \varepsilon$$

$$\varepsilon^2 = (Y - \hat{Y})^2$$

- که در آن x متغیر مستقل و تحت کنترل ماست.
- a مقدار عرض از مبدا یا constant نامیده می‌شود .
- b مقدار شیب خط یا ضریب متغیر مستقل نامیده می‌شود.
- e خطای مشاهده و مدل رگرسیون است.
- Y|x متغیر تصادفی وابسته به x است و دارای میانگین a+bx و واریانس  $\sigma^2$  است.
- متغیر تصادفی  $\varepsilon$  خطای تصادفی است و دارای میانگین صفر و واریانس  $\sigma^2$  است.
- پارامترهای این مدل عبارتند از  $a, b, \sigma^2$
- Y مقدار واقعی و  $\hat{Y}$  مقدار برآورد شده است.

**مفروضات :**

- برای هر مقدار متغیر مستقل، توزیع متغیر وابسته یک توزیع نرمال است. یا توزیع باقی مانده‌ها باید نرمال و مستقل از مشاهدات دیگر باشد.
- برای همه مقادیر متغیر مستقل باید واریانس متغیر وابسته ثابت باشد.
- همه مشاهدات باید تصادفی باشند.
- ارتباط دو متغیر وابسته و مستقل باید خطی باشد.

مثلاً دو متغیر فشارخون و مصرف نمک را در نظر بگیرید که فشار خون متغیر مستقل و مصرف نمک متغیر وابسته است.

هرچه اندازه متوسط مربعات خطا کمتر باشد معادله رگرسیون بهتر مشاهدات را مدل کرده است. برآورد ضرایب رگرسیون به صورت زیر محاسبه می‌شوند.

$$\hat{b} = r \frac{S_y}{S_x} \quad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

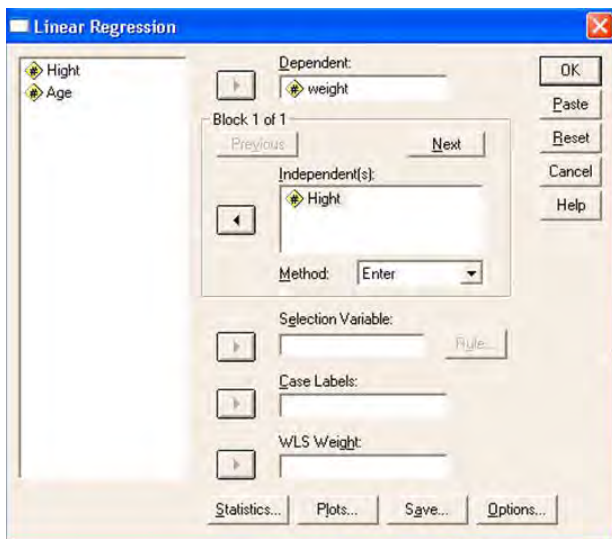
برآورد مقدار خطا (Residual) به صورت زیر محاسبه می‌شوند:

$$\hat{e} = y_i - \hat{y} = y_i - \hat{a} - \hat{b}x_i$$

هدف کمینه کردن مجموع مربعات خطا است. (که برآوردها نیز به همین روش بدست آمده‌اند)

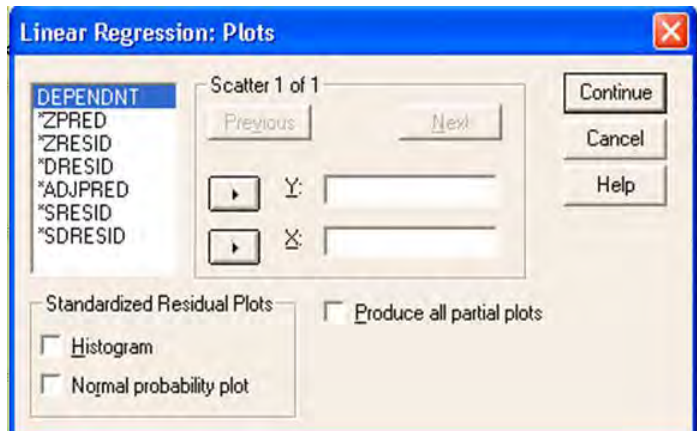
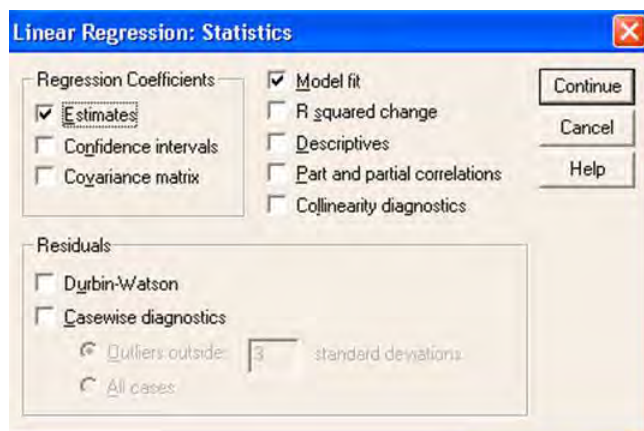
برای دسترسی به این دستور مسیر زیر را طی کنید:

Analyze - Regression - Linear...



در صورتی که به عنوان متغیر مستقل چندین متغیر را وارد کنید spss محاسبات مربوط به رگرسیون چند متغیره را محاسبه خواهد کرد.

- در قسمت dependent متغیر وابسته را وارد کنید.
- در قسمت independent متغیر یا متغیرهای مستقل را وارد کنید.
- در قسمت selection variable متغیر و مقداری از آن که برای بررسی مورد نظر شما است وارد کنید. (این متغیر نباید در رگرسیون به عنوان متغیر مستقل به کار رفته باشد).
- در قسمت Case label متغیری که به عنوان اسامی موردها به کار رفته است مشخص کنید.
- در قسمت WLS Weight متغیر وزن دهی به موردها را وارد کنید. (این وزن برای یکسان سازی مقدار واریانس به کار خواهد رفت).
- با انتخاب دکمه Statistic امکان نمایش مقادیر آماره‌های دیگری در خروجی وجود دارد.
- با انتخاب دکمه Plot امکان نمایش نمودارهای بررسی معادله رگرسیون وجود دارد.
- با دکمه Save امکان ذخیره سازی مقادیر برآورده شده از متغیر وابسته و باقی مانده (استاندارد-واقعی) وجود دارد



$$\begin{cases} H_0 : \text{مدل مناسب نیست و تاثیر در تغییرات Y نسبت به X ندارد.} \\ H_1 : \text{مدل رگرسیون مناسب است.} \end{cases}$$

مثال- معادله خطی رگرسیون را برای متغیر های income و car بدست آورید. مدل را یکبار دیگر فقط برای مجرد ها حساب کنید

Analyze - Regression - Linear..... Dependent.....car  
Independent.....income

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.792(a)	.627	.627	13.38386

a Predictors: (Constant), Household income in thousands

**ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1930513.413	1	1930513.413	10777.296	.000(a)
	Residual	1146059.754	6398	179.128		
	Total	3076573.167	6399			

a Predictors: (Constant), Household income in thousands

b Dependent Variable: Price of primary vehicle

**Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	14.799	.223		66.319	.000
	Household income in thousands	.221	.002	.792	103.814	.000

a Dependent Variable: Price of primary vehicle

$$(car)Y = 14.799 + 0.221 * X (income)$$

نتیجه :

- جدول آنالیز واریانس مقدار خطای رگرسیون را حساب می کند.  
- برای هر یک از ضرایب این آزمون را انجام داده است  
- چون  $p - value = 0.000$  پس نمونه فرض صفر را رد می کند.

- برای مجرد ها اول با split file متغیر marital را جدا می کنیم بعد مراحل محاسبه خط رگرسیون را تکرار می کنیم.

**Coefficients<sup>a</sup>**

Marital status	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
Unmarried	1	(Constant)	14.755	.318		46.468	.000
		Household income in thousands	.223	.003	.791	73.323	.000
Married	1	(Constant)	14.840	.314		47.327	.000
		Household income in thousands	.219	.003	.794	73.504	.000

a. Dependent Variable: Price of primary vehicle



فرض کنید می خواهیم با توجه به مدل بدست آمده برای مقدار درآمد  $x=73$  مقدار هزینه ماشین را حساب کنیم در حالی که این داده جزء داده های ما نمی باشد.  
 ۱. عدد ۷۳ را در صفحه Dataview آخرین ردیف در ستون income وارد می کنیم.  
 ۲.

Analyze - Regression - Linear..... Dependent.....car  
 Independent.....income  
 Save .....predicted value.....standardized

۳. مشاهده مقدار پیش بینی شده در ستون جدید به نام ZPR در آخرین ردیف :  $y=0.3208$

۴. برای مشاهده باقی مانده ها :

Save .....Residuals .....standardized

ستونی از باقیمانده ها در صفحه Dataview نشان داده می شود.

### محاسبه ضرایب خط رگرسیون بدون مقدار ثابت:

Analyze - Regression - Linear.....option.....Include constant equation

Coefficients(a,b)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	Household income in thousands	.314	.002	.885	151.668	.000

a Dependent Variable: Price of primary vehicle  
 b Linear Regression through the Origin

سئوال: در مدل خطی چند متغیره کدام متغیر اهمیت بیشتری دارد یعنی تغییر کوچک آن باعث تغییرات بزرگ معادله می شود؟

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + \varepsilon$$

بزرگی یا کوچکی ضرایب نشانده اهمیت آنها نیست چون اگر واحد متغیر هاراعوض کنیم ضریب نیز عوض می شود.

متغیر اگر بدون واحد باشد ضریب آن در جدول Coefficients, ستون Standardized Coefficients تحت عنوان Beta درج می شود. این ضریب اهمیت تغییرات است که در مدل نمی آید. مثلاً در مثال بالا که فقط یک متغیر مستقل وجود دارد یک Beta هم در جدول درج شده است که اگر چند متغیر باشد به Beta های مختلف تقسیم می شد.

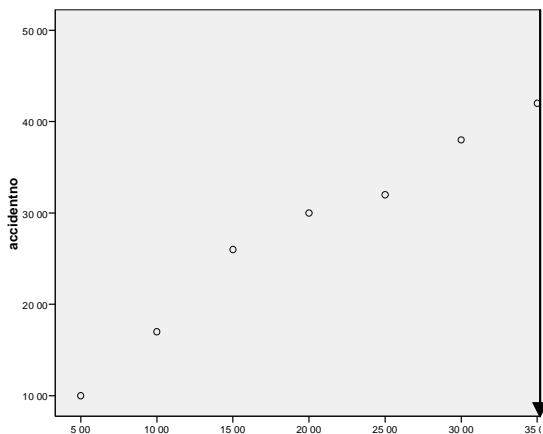
مثال: میزان الکل خون رانندگانی که تصادف کرده اند را اندازه گرفته اند. میخواهیم بدانیم:

۱. آیا ارتباطی بین میزان الکل و تعداد تصادفات هست یا نه؟ با نمودار نشان دهید.
۲. ضریب همبستگی خطی بین دو متغیر را حساب کنید و فرضیات را تایید کنید
۳. با استفاده از spss رابطه خطی بین دو متغیر را برآورد کنید و برای سطوح مختلف الکل خون تعداد تصادفات را برآورد کنید.
۴. آیا باقیمانده ها دارای توزیع نرمالند؟
۵. آیا مشاهدات تصادفی هستند؟

الکل خون	۵	۱۰	۱۵	۲۰	۲۵	۳۰	۳۵
تعداد تصادفات	۱۰	۱۷	۲۶	۳۰	۳۲	۳۸	۴۲

جواب ۱.

نتیجه: به نظر می رسد رابطه خطی وجود دارد.



Graph – Legacy dialogs – scatter plot– simple

X Axis.....alckol  
Y Axis.....accidentno

X متغیر مستقل الکل و Y متغیر وابسته تعداد تصادفات است.

جواب ۲.

Correlations

		alckol	accidentno
alckol	Pearson Correlation	1	.984(**)
	Sig. (2-tailed)		.000
	N	7	7
accidentno	Pearson Correlation	.984(**)	1
	Sig. (2-tailed)	.000	
	N	7	7

با  $p - value = 0.000$  فرض صفر یعنی استقلال دو متغیر توسط نمونه رد می شود.  $\rho = 0.984$  که نشانه همبستگی بالاست.

\*\* Correlation is significant at the 0.01 level (2-tailed).

جواب ۳.

Analyze - Regression - Linear..... Dependent.....accidentno  
Independent.....alckol

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	740.571	1	740.571	152.471	.000(a)
	Residual	24.286	5	4.857		
	Total	764.857	6			

a Predictors: (Constant), alckol  
b Dependent Variable: accidentno

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.286	1.863		3.912	.011
	alckol	1.029	.083	.984	12.348	.000

a Dependent Variable: accidentno

$$(accidentno)Y = 7.286 + 1.029 * X (alckol)$$

نتیجه :

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases} \quad \begin{cases} H_0 : b = 0 \\ H_1 : b \neq 0 \end{cases}$$

- در جدول آنالیز واریانس برای هر یک از ضرایب این آزمون را انجام داده است - چون  $p - value = 0.000$  پس نمونه فرض صفر را رد می کند. یعنی ضرایب صفر نیستند.

جواب ۴. برای ایجاد ستون باقیمانده ها:

Analyze - Regression - Linear..... Dependent.....accidentnor  
Independent.....alckol  
Save ..... Residuals .....standardized

Runs Test 2

	Standardized Residual
Test Value(a)	.0000000
Cases < Test Value	5
Cases >= Test Value	2
Total Cases	7
Number of Runs	3
Z	-.380
Asymp. Sig. (2-tailed)	.704

a Mean

جواب ۵.  
با استفاده از آزمون Runs :

$$p - value = 0.704 > 0.05$$

پس فرض صفر توسط نمونه رد نمی شود و باقیمانده ها تصادفی اند.

$$\begin{aligned} \text{Standard Residual} &\sim N(0,1) \\ \text{Unstandard Residual} &\sim N(0, \sigma^2) \end{aligned}$$



در کانال تلگرام کارنیل هر روز انگیزه خود را شارژ کنید 😊

<https://telegram.me/karnil>

