« 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 مي باشد.
هر ستون یک field است که به عنوان متغیر یا variable مطرح مي باشد و هر سطر یک record است که به عنوان یک مورد یا cace مطرح مي باشد.
به پنجره باز شده Datasheet مي گویند. و کار برنامه SPSS ویرایش اطلاعات این صفحه است.هر Datasheet دوقسمت دارد:
۱- Data view که مقادیر را درآن وارد مي کنیم
۲- Variable view که متغیر هارا تعریف مي کنیم.

در SPSS متغیر ها خصوصیاتي دارند که مهمترین خاصیت آنها مقیاس آنهاست. و هر بار که متغیر تعریف مي شود باید مقیاس آن را تعریف کنیم.

انواع مقیاس ها:

۱. مقیاس اسمي (صوري) : Nominal
مانند شهري که درآن زندگي مي کنیم که میتوانیم به آن کد داده و تبدیل به مقیاس اسمي شود در این مقیاس کوچک و بزرگ بودن اعداد تفاوتي ندارد و هر طور مي توان کد گذاري کرد و مناسب گروه بندي است ومیانگین و جمع در آن معني ندارد.

۲. مقیاس رتبه اي (ترتیبي) : Ordinal
کد گذاري بر اساس درجه اهمیت مي باشد و اولویت یا ترتیب را نشان مي دهد اگر کد ها جابجا شوندفرق مي کند. مانند مهارت خوب ، بد و عالي.فاصله قابل اندازه گیری بین گروهها وجود ندارد.

۳. مقیاس نسبي : Scale
هر عددي که از طریق اندازه گیري بدست بیاید مقیاس نسبي دارد مثل قد افراد .

شاخص تمرکز :

نشانه تمرکز داده هاست و معروفترین آن میانگین (**Mean**) است که فقط براي مقیاس نسبي بکار مي رود.بسیار به داده ها حساس است و اگر مقادیر بزرگ باشد به طرف آن مي رود.ولي اگر داده ها خیلي پراکنده باشند میانگین گویا نخواهد بود.
شاخص دیگر میانه **(Median)** است که رتبه میانی داده ها ست و از روي رتبه بدست مي آیداگر تعداد داده ها زوج باشد رتبه $\frac{n+1}{2}$ میانه است و اگر فرد باشد وسطي میانه است.میانه به داده بزرگ یا کوچک حساس نمي باشد.
شاخص دیگر مد **(Mode)** است که براي داده هاي اسمي یا Nominal هم بکار مي رود میزان فراواني داده ها را نشان مي دهد و بیشترین مشاهدات را دارد.

شاخص پراکندگی :

نشانه پراکندگي داده هاست و براي داده هاي نسبي يا Scale بكار ميرود و معروفترين آن واريانس است . كه f_i فراواني هر داده است ميانگين $\text{Mean} = \frac{\sum x_i}{n} = \frac{\sum x_i f_i}{\sum f_i}$

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

و جذر واريانس انحراف معيار است كه واحد آن با واحد داده ها يكي است اگر واحد داده ها كيلوگرم باشد انحراف معيار هم كيلوگرم است. واريانس و انحراف معيار شاخص هاي خوبي هستند چون داده ها درآن موثرند.

شاخص ديگر پراكندگي دامنه تغييرات يا Range مي باشد. R=max-min

شاخص ديگر چارک است كه اگر داده ها خيلي پراكنده باشنداز آن استفاده مي كنيم و ميانبر چارکي :

$$R = \overline{Q_3} - \overline{Q_1}$$

چارک سوم يعني ۷۵% داده ها از آن كمتر و ۲۵% داده ها ازآن بيشتر هستند و چارک دوم همان ميانه است.

اگر ميانه و ميانگين داده ها به هم نزديک باشند آن جامعه نرمال مي باشد.

مثال۱ :

متغير هاي نام كشور و درآمد ناخالص ملي را تعريف كرده و اطلاعات زير را در صفحه وارد كنيد سپس شاخص هاي مختلف آماري، توزيع و پراكندگي را محاسبه نماييد.

نام كشور	درآمد ناخالص ملي
ايران	۱۵
تركيه	۲۰
افغانستان	۵
عراق	۱۰
پاكستان	۱۳

متغير هارا در variable view تعريف مي كنيم:

متغير اول نام كشورها به نام country

در تايپ اسم متغير نبايد از فاصله استفاده شود ولي خط تيره اشكالي ندارد و فارسي تايپ نشود.

در Type يا نوع متغير STRING انتخاب مي كنيم.

در label ميتوان توضيحات كافي در مورد متغير داد. مثلاً، name of country

در قسمت value چون متغير اسمي تعريف شده چيزي نمي زنيم.و مخصوص كد بندي متغير هاي گروهي است كه عددي باشند.

در قسمت missing يعني اعدادي معرفي شود كه در محاسبات در نظر گرفته نمي شودمثلاً در پرسشنامه ها برخي گزينه ها ناخواناست يا پر نشده است ما عدد ۱۰۰۰ را معرفي ميكنيم كه هر وقت درج شد به معني آن باشد و حساب نشود.

پهناي ستون را ۱۰ تعريف مي كنيم.

Align را انتخاب ميكنيم راست ، چپ يا وسط

Measure يا مقياس متغير را Nominal و اسمي انتخاب مي كنيم.

متغیر دوم میزان درآمد ناخالص ملي يا GNP
در Type يا نوع متغير Numeric انتخاب مي كنيم.كه اختصاص به مقادير عددي دارد.
در label ميتوان توضيحات كافي در مورد متغير داد. مثلاً، Gross National Product
در قسمت value چيزي نمي زنيم.
پهناي ستون را ۱۰ تعريف مي كنيم.
Align را انتخاب ميكنيم راست ، چپ يا وسط
Measure يا مقياس متغير را Scale و نسبي انتخاب مي كنيم.
Decimal را اگر ۲ بزنيم تا ۲رقم مميز مي 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 :

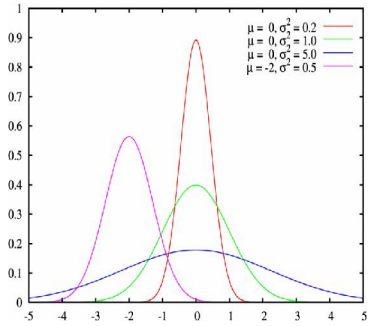
Analyze – Report –Case Summaries

پنجره اي باز مي شود كه متغير مورد نظر جهت تنظيم شاخص هاي آماري را توسط فلش از پنجره موجود به پنجره سمت راست Variable انتقال مي دهيم و سپس با زدن دگمه statistic شاخص هاي آماري مورد نظر را مانند ميانه ، ميانگين، واريانس ، چولگي (skewness) ، كشيدگي (kutosis) و . . . انتخاب مي كنيم.
خروجي spss در پنجره جداگانه به نام output1 ظاهر مي شود.بهتر است اين پنجره را نبنديد و هرچه اظافه كنيد در همين پنجره اضافه خواهد شد.

روش دوم محاسبه شاخص ها :

Analyze – Descriptive statistics - frequencies

مي توانيم جدول فراواني را نيز رسم كنيم.
Percentile Values شامل Quartiles چاركها ، Cutpoints دهک ها و Percentile صدكها هستند.
Central Tendency شاخص هاي تمركز ، Dispersion معيار پراكندگي ، Distribution معيار توزيع مي باشند.
اگر داده ها پراكنده و متفاوت باشند مي توانيم چولگي و كشيدگي را حساب كنيم.
اگر داده ها تكراري باشند و جدول فراواني را بخواهيم بايد گزينه Display frequency table را تيک بزنيم.



نمودار نرمال ، كشيدگي

در حالت پخي kutosis<0 و در حالت كشيده kutosis>0 مي باشد.
در حالت چولگي به چپ skewness<0 و ميانگين<ميانه<مد است.
در حالت چولگي به راست skewness>0 و مد<ميانه<ميانگين است.

مثال۲ :

متغير هاي نوع ماشين ، زمان و مكان را تعريف كرده و اطلاعات زير را در صفحه وارد كنيد سپس شاخص هاي مختلف آماري، توزيع و پراكندگي را محاسبه نماييد.نتيجه ميزان تردد در كدام زمان و توسط چه نوع ماشيني را مشخص نماييد

نوع ماشين	مكان	زمان	ميزان سوخت
شخصي	داخل طرح	۱۲	۶.۵
شخصي	خارج طرح	۱۳	۱۵
شخصي	داخل طرح	۱۴	۱۴
تاكسي	خارج طرح	۱۳	۱۴
تاكسي	داخل طرح	۱۴	۲۰
اتوبوس	خارج طرح	۱۵	۲۶
وانت	داخل طرح	۱۶	۳۰
اتوبوس	خارج طرح	۱۶	۳۰
تاكسي	داخل طرح	۱۷	۵۰
شخصي	خارج طرح	۱۵	۷
وانت	داخل طرح	۱۳	۸

انتخاب متغير 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

براي متغير هايي كه كد بندي شده اند شاخص هاي آماري مهم نيست و جدول فراواني اهميت دارد.
اما در اينجا اگر missing تعريف مي كرديم در جدول فراواني بين دو ستون percent و valid percent تفاوت ايجاد مي شد.
ستون Cumulative percent در جدول فراواني نشان دهنده درصد تجمعي است كه اگر داده ها Nominal باشد اين ستون معني ندارد.
جدول فراواني سوخت چيزي نشان نميدهد و تقريباً هر مورد يک بار تكرار شده است اما ما مي خواهيم در فاصله هاي مختلف بسنجيم لذا متغير جديدي به نام Gfuel را معرفي مي كنيم:
Transform-Record-Into Different Variable
نام و برچسب متغيرجديد را درج مي كنيم و change را زده، سپس دگمه old and new value را مي زنيم. در قسمت :
Old value : اگر مي خواهيم به هر مقدار متغير قديم متغير جديد نسبت دهيم و تناظر يک به يک برقرارا كنيم گزينه value را تيک مي زنيم و اگر مي خواهيم به هر مقدار متغير قديم يک رنج جديد تعريف كنيم از گزينه Range استفاده مي كنيم.كه در اينجا:

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

و add را مي زنيم. و سپس در قسمت تعريف متغير Gfuel در ستون ششم در value تعريف مي كنيم منظور از اعداد ۱،۲،۳ و ۴ چيست؟

	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 فراواني را براي متغير Gfuel و شاخص هاي آماري را براي متغيرfuel محاسبه مي كنيم تا معني داشته باشد.

سدوال۱ : برای فر دگزارش گیرنده صفحه اصلی ورود اطلاعات مهم است یا نتایج و خروجی؟

سئوال۲: اگر جدول فراوانی را داشته باشیم چطور می توانیم داده هارا در صفحه اصلی درج کنیم؟

سئوال۳: طریقه رده بندی صحیح برای داده ها چیست؟از نظر طول فاصله و از نظر تعداد گروهها؟

« جلسه دوم – ۸۶/۱۱/ ۲۵ »

جدول فراوانی :

متغیر ها (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 ۴ = فوق لیسانس ۵ = دکترا	.....	.....	۵۰	۰.۲

فراواني نسبي 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 بالابرای 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

مثال۳ :
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 متغیر ها :

	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 فروانی 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.

مثال۴ :

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

انتخاب متغیر ها :

	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 :

برای رده بندی متغیر سن از این فرمول استفاده می کنیم: $\frac{Max-Min}{5}=\frac{50-20}{5}=6$

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	$\frac{27+33}{2}=30$	

و طبق فرمول میانگین، x_i نماینده رده i ام است. $\bar{X}=\frac{\sum f_i x_i}{\sum f_i}$

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.

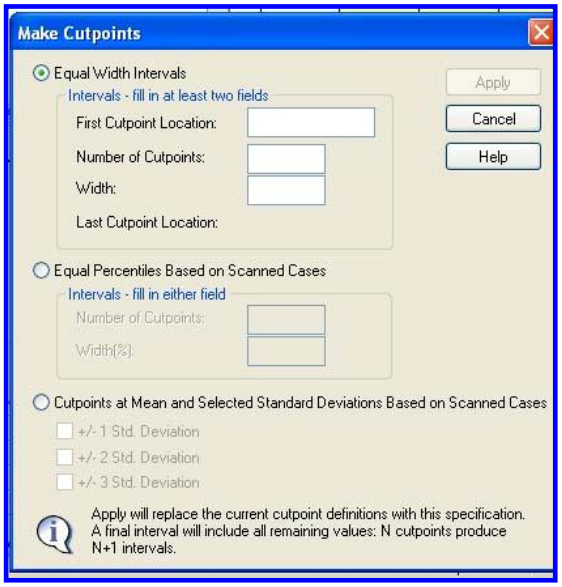
grouping variable for age

		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 جدیدvisual Binning اسم و برچسب متغیر جدید را وارد می کنیم gweight حال با زدن دگمه Make cutpoints نقاط و محل شکست را می سازیم.

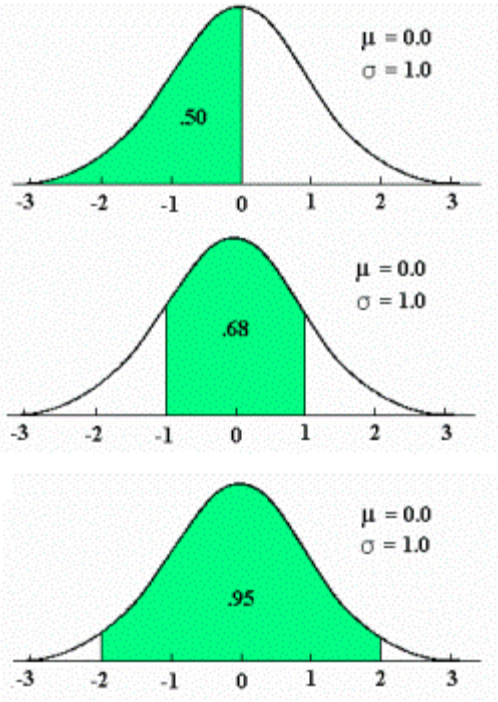


سه راه برای تعیین نقاط شکست وجود دارد:
۱- Equal Width Intervals
نقطه اول را min ما باشد و در اینجا ۴۰ است میدهیم که بهتر است برای پیوستگی عدد ۳۹.۵ را وارد کنیم بعدی تعداد رده ها ست که ۳ وارد می کنیم و سومی طول رده است که خودش می نویسد اگر طول رده را بدهید تعداد رده را خودش می دهدو آخرین عددرده را در 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 میدهیم.
۳- Cutpoints at Mean and Selected Standard Deviation Based on Scanned Cases

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

نکته: به هرحال از هر کدام این سه گزینه که استفاده شود تعداد n نقطه شکست تعداد n+1 رده را خواهد داشت.
با زدن دگمه Make Label در همان پنجره رده ها هم ظاهر می شوند.



ارتفاع این منحنی با مقادیر میانگین μ (و انحراف معیار) σ ارتباط دارد.
μ میزان تراکم داده ها حول یک مقدار و σ (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 × 1000 یا تابعی خاص را انتخاب کنیم .

نکته : فرق فرمول نویسی در SPSS با Excel این است که در SPSSبا 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 مثالاً تغییر واحد وزن به گرم فقط برای خانمها انجام شود باید از دگمه if پایین پنجره Compute استفاده کنیم مثلاً : شرط بگزاریم که Gender =1 یا Gender =1& Age>30

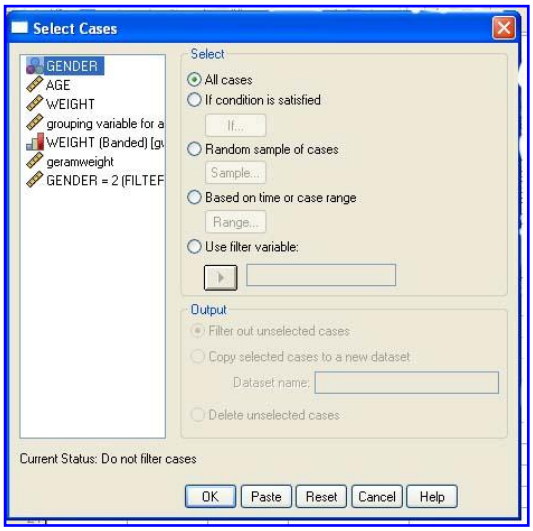
راه دیگر اینکه می توانیم یک متغیر به نام select تعریف کنیم برای داده های خاصی ۱ بگذرایم و برای داده های دیگر ۰ بعد فرمول را برای یکها حساب کنیم یا برای صفرها.

روش ثبت انواع دستورات اجرا شده در SPSS :

با استفاده از دگمهpaste 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.



<u>رو ش اول :</u>

Data – Select Cases

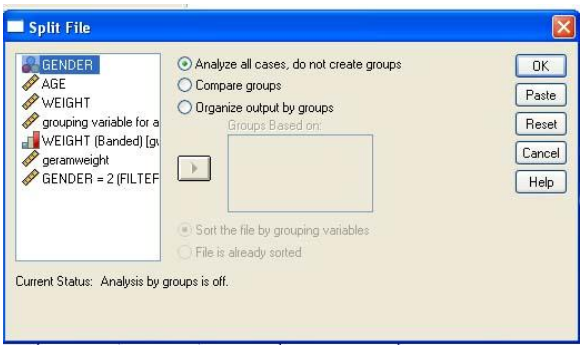
پنجره روبرو باز می شود که اگر All Cases فعال باشد به معنای عدم فیلتر است و با فعال کردن if condition is satisfied شرط مورد نظر را درج می کنیم .مثلاً شرط می گذاریم فقط مردها را محاسبه کند که اگر حال با انجام فیلتر اگر میانگین بگیریم فقط برای مردها محاسبه میکند و در Data view هم رکورد های زنهارا خط می 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 بازدن ok فیلتر انجام می شود و در پایین صفحه نیز filter on نمایش داده خواهد شد.

<u>رو ش دوم :</u>

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 دوقسمت مردها و زنها نشان خواهد داد.

مثال5 : 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 ۴۹۰۰۰-۲۵۰۰۰ 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 ۲۸-۱۹ ، ۵۲.۳٪ زیر ۲۵۰۰۰ دلار درآمد دارند در حالیکه در رده سنی ۵۸-۴۸ ، نزدیک ۵۰٪ درآمد بالای ۷۵۰۰۰ دلار دارند..

« جلسه سوم – ۲ /۸۶/۱۲ »

اقسام نمونه :

۱. نمونه گیری تصادفی ساده : شامل n عنصر از یک جامعه است که :
ویژگیها: شانس مساوی برای انتخاب هر عنصردارد.
- شانس مساوی برای انتخاب هر نمونه n تایی دارد. (آزمایش تصادفی)
- موارد استفاده: جامعه شماره گذاری شده یا قابل 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 ودر 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 ۵و۱۰و۱۵ و ... را انتخاب می کنیم و در نمومه 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 آماري 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 هايي كه اصطلاحاً خوشه ناميده مي شود و مبناي تشكيل خوشه ها در اغلب اوقات، تقسيم بندي هاي جغرافيايي ازقبيل استان، شهرستان، دهستان، شهر و بلوك يا آبادي است، تقسيم بندي مي كنند .خوشه ها بايد كل جامعه را پوشش دهند و فاقد هم پوشاني باشند .به بيان ديگر هر كدام از عناصر جامعه بايد به يكي و فقط يكي از اين خوشه ها تعلق داشته باشد.
بسته ي نرم افزاري SPSS طرح نمونه گيري پيچيده ، نمونه گيري خوشه اي يك يا چند مرحله اي، نمونه گيري طبقه بندي،را با Complex Samples فراهم مي كند.

سئوال : 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 متغیر های اسمی بکار می رود جدول توافقی زیر را در صفحه نرم افزار وارد کنید :

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

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

	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=emploee	-	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
	emploee	20	10	30
Total		25	25	50

با زدن دگمه 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

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

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

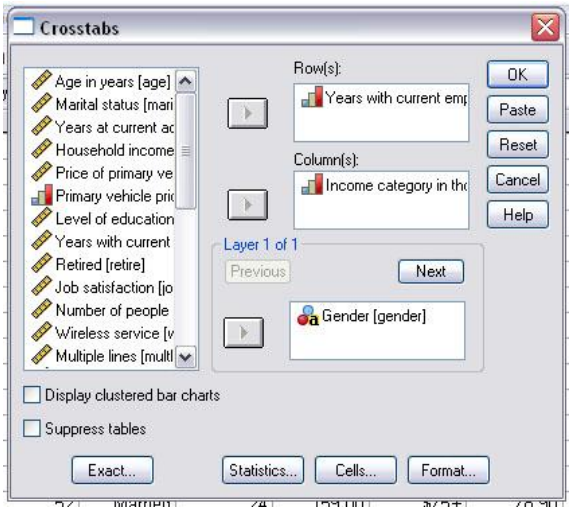
سئوال : تحقیق کنید بین جنسیت و نوع اسباب بازی ارتباطی وجود دارد یا نه؟

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 inccatرا بدست آورید.و متغیر جنسیتgender را به عنوان متغیر سوم وارد کنید.
اگر در پنجره crosstabs دگمه 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

Gender			Income category in thousands				Total
			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) ، 000. است یعنی استقلالی 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		

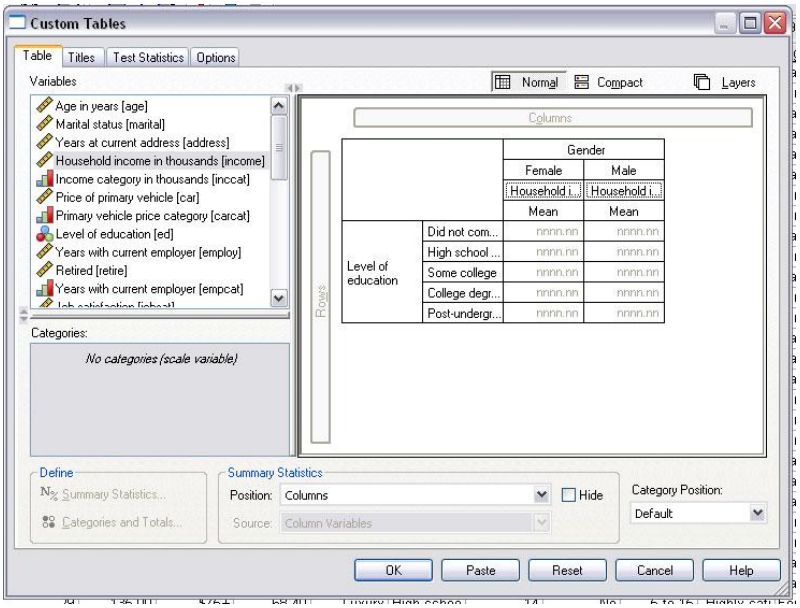
a 0 cells (.0%) have expected count less than 5. The minimum expected count is 318.50.

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 دگمه summary statistics در همان پنجره محاسبه نمود.

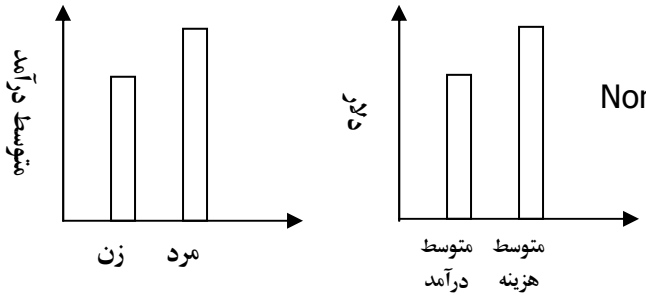
تمرین : در فایل Demo جدولی رسم کنید که ارتباط بین جنسیت gender و انواع ماشین یا carcat Primary vehicle price category را نشان دهد ولی به جای تعداد میزان متوسط درآمد آنها در جدول درج کند.

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 خرند ولی آقایان با بالاترین درآمد ۱۱۱۶ هزار دلارماشین لوکس می خرند
میانگین درآمد خانم هایی که ماشین لوکس می خرند کمتر از آقایان است.



انواع نمودار:

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

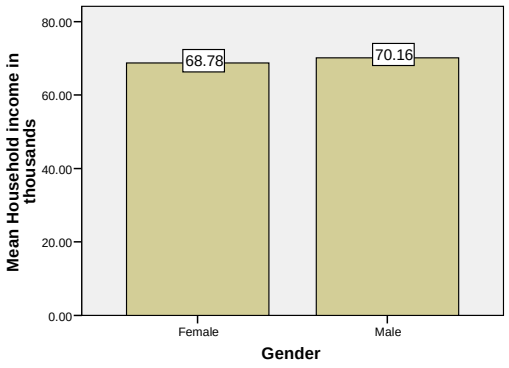
تمرین : این نمودارها را درست کنید:

ساخت نمودار نوع اول:

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

یا

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



ساخت جدول نوع دوم:

Data – split files - gender

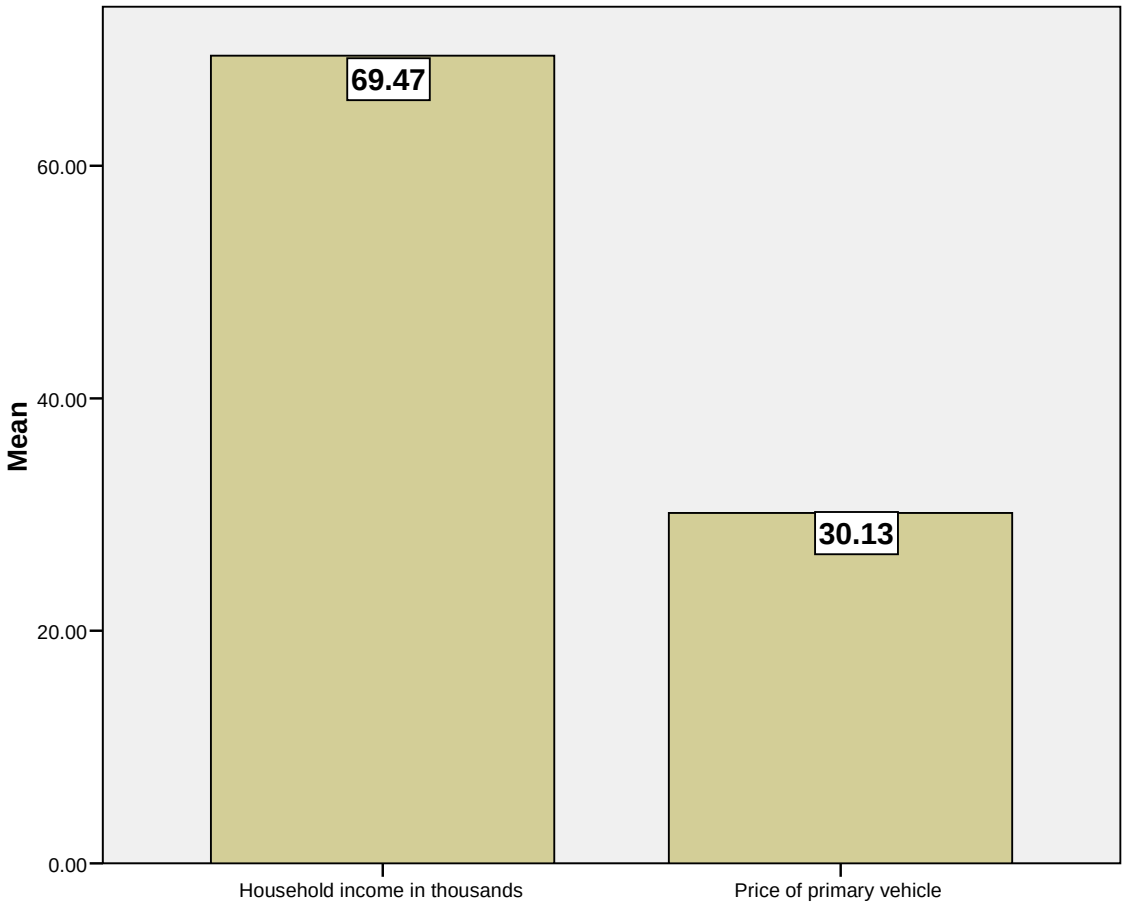
Analyze – Descriptive statistics – Descriptive - Household income in thousands , Price of primary vehicle

		Household income in thousands	Price of primary vehicle
		Mean	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 ودر قسمت سطر جدول قرارداده 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 متغیرهارا در پنجره ببرید و ok رابزنید.

۲. اندازه نمونه را چگونه محاسبه کنیم؟

الف – برای متغیر کمی:

* در جامعه نامحدود:

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

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

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

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

$$\sqrt{n}d=z.\sigma$$

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

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

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

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

برای $\alpha=0.05$ ، $z=1.64$

ب- متغیر کیفی :

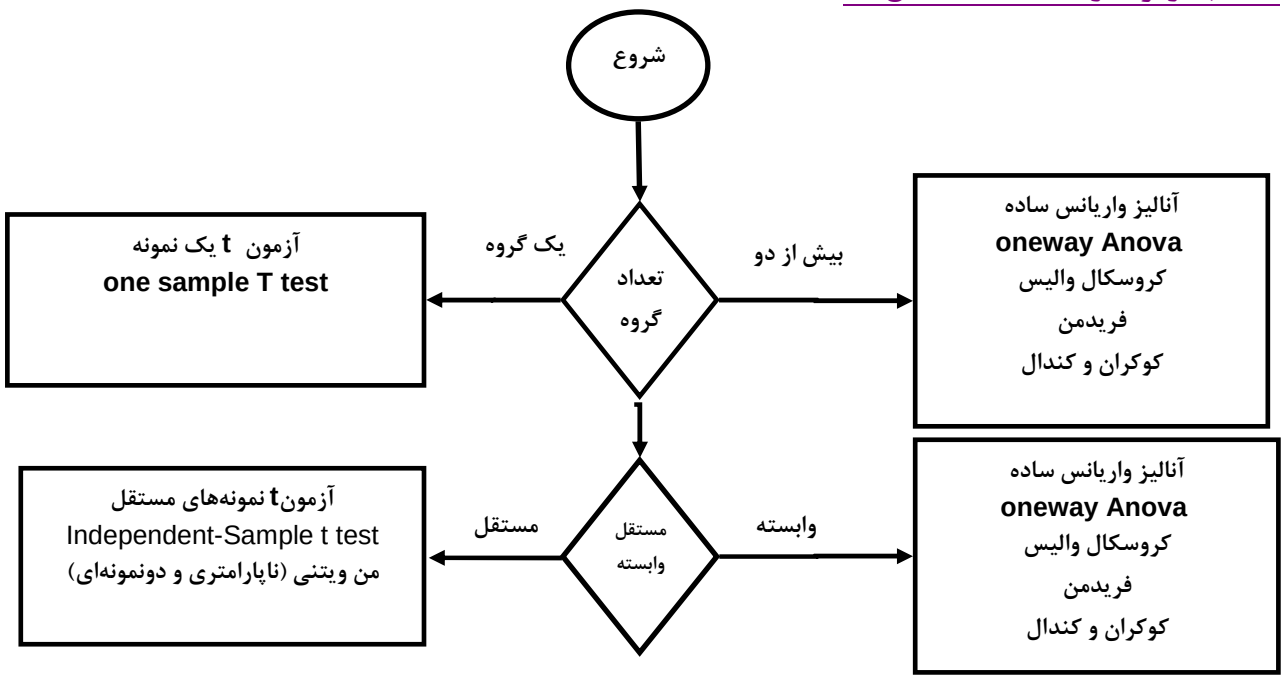
به جای σ ، pq را قرار می دهیم که p احتمال داشتن صفت مورد نظر و q=1-p احتمال نداشتن آن صفت است.

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

آزمون فرض :

آزمون فرض براي مقايسه شاخص جمعيت با نمونه و انتخاب مقدار شاخص نمونه به عنوان شاخص جمعيت به كار مي‌رود.
دقت اين انتخاب با توجه به اندازه جمعيت و نمونه و ميزان خطاي است كه مي‌توانيم متحمل شويم.
هرچه اندازه نمونه نسبت به جمعيت بيشتر باشد دقت آزمون بيشتر خواهد بود. هرچه جمعيت به توزيع نرمال نزديكتر باشد آزمون‌ها پرتوان‌تر خواهند بود.

انتخاب آزمون برای مقایسه میانگین ها:



آزمون های مربوط به میانگین:

زمانی که جامعه را نداریم و یک نمونه انتخاب کرده ایم و می خواهیم خصوصیت نمونه را به جامعه نسبت بدهیم لذا آزمون انجام می دهیم. و با استفاده از قابلیت آزمون فرضیات میتوان اختلاف بین میانگین ها 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 می شویم.
هرچه توزیع به طرف نرمال پیش برود آزمون ها پرتوان تر خواهند بود و خطای نوع دوم کمتر است.خطای نوع دوم یعنی احتمال رد H1 به شرط اینکه درست باشد.

فرض کنیم $X_1, X_2, ..., X_n$ یک نمونه n تایی از توزیع نرمال با میانگین مجهول μ و واریانس معلوم σ^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 قراردارد.

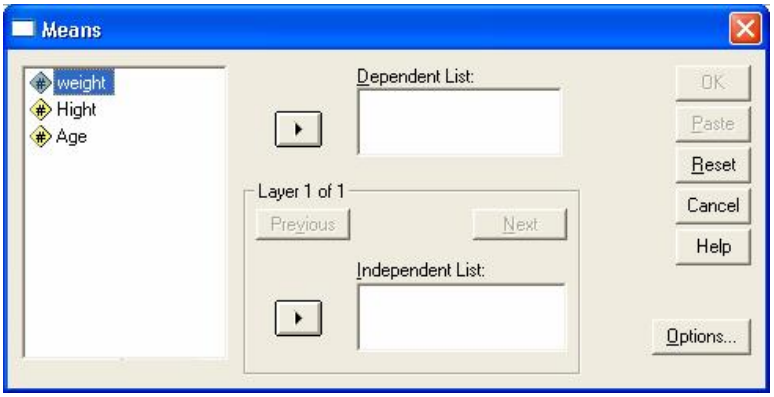
$$\begin{cases} H0: \mu = 0 \\ H1: \mu > 0 \end{cases}$$

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 H0 است به شرطی که صحیح نباشد.)
یکی از روشهای آزمون فرض آماری استفاده از $p-value$ یا مقدار احتمال است اگر $p-value \le \alpha$ 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 مقايسه ميانگين به عنوان شاخص تمركز جمعيت مي‌توان مقايسه مناسبي بين دو گروه از جمعيت (نمونه) بدست آورد.
از آنجايي كه ميانگين يا ديگر معيارهاي تمركز مثل ميانه و مد به عنوان شاخص و نماينده جمعيت شناخته شده هستند، براي مقايسه دو ويژگي يا دو جمعيت بهتر است معيارهاي تمركز آنها را مقايسه كرد.

Analyze – Compare Means – Means



در قسمت Dependent list :
متغير يا ويژگي اندازه گيري شده را وارد كنيد. اين متغير بايد داراي مقادير ***عددی*** باشد.
و در قسمت Independent list :
متغير دسته بندي يا ***گروه‌بندی*** را معرفي كنيد. با اين كار ميانگين ويژگي براساس متغير دسته‌بندي براي هر گروه بدست آمده و مقايسه مي‌شوند.
مثلاً مقايسه درآمد براساس وضعيت تاهل

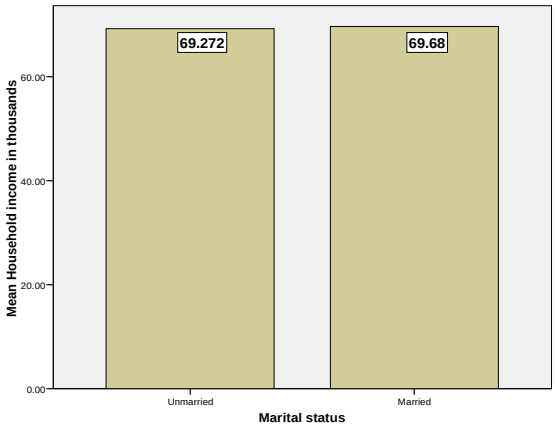
آماره ميانگين : $T = \frac{x - \overline{x}}{\frac{s}{\sqrt{n}}}$ می باشد.

مقادير مربوط به ويژگيهاي يک *نمونه* (ميانگين، انحراف معيار و غيره) را آماره (Statistics) و ويژگيهاي معادل آنها در *جامعه* اصلی را پارامتر (Parameters) می گویند.
نمودار ميانگين متغير income را نيز بر اساس وضعيت تاهل می کشيم :

Report

Household income in thousands

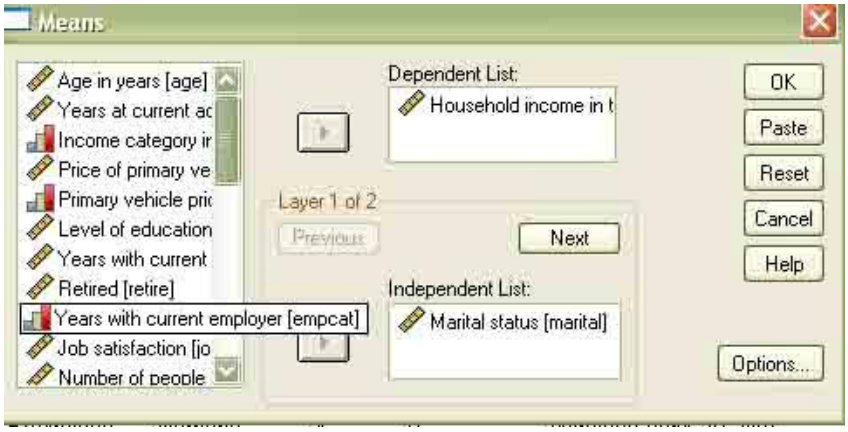
Marital status	Mean	N	Std. Deviation
Unmarried	69.2723	3224	78.32925
Married	69.6804	3176	79.12361
Total	69.4748	6400	78.71856

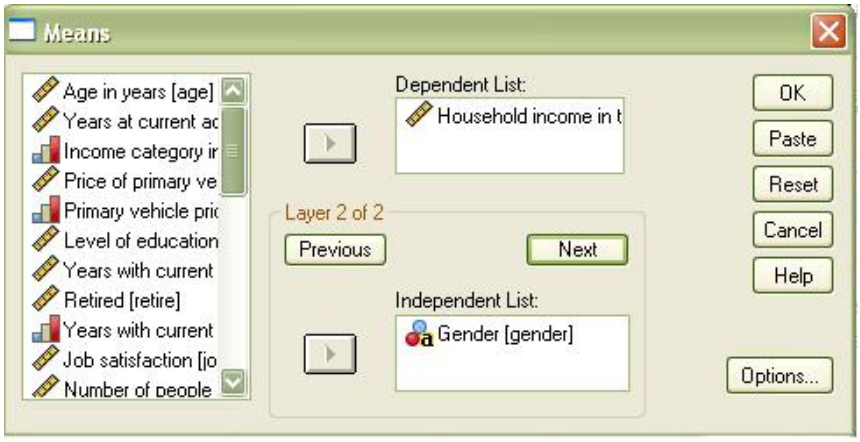


لایه بندی متغیرها:

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

Analyze – Compare Means – Means- dependent list .....income
Independent list.....Martial
Next
Independent list.....Genderl





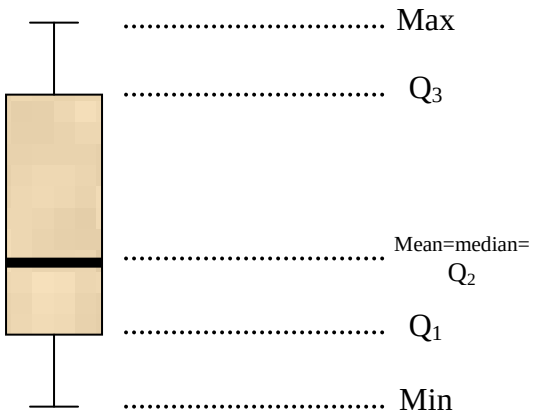
Report

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



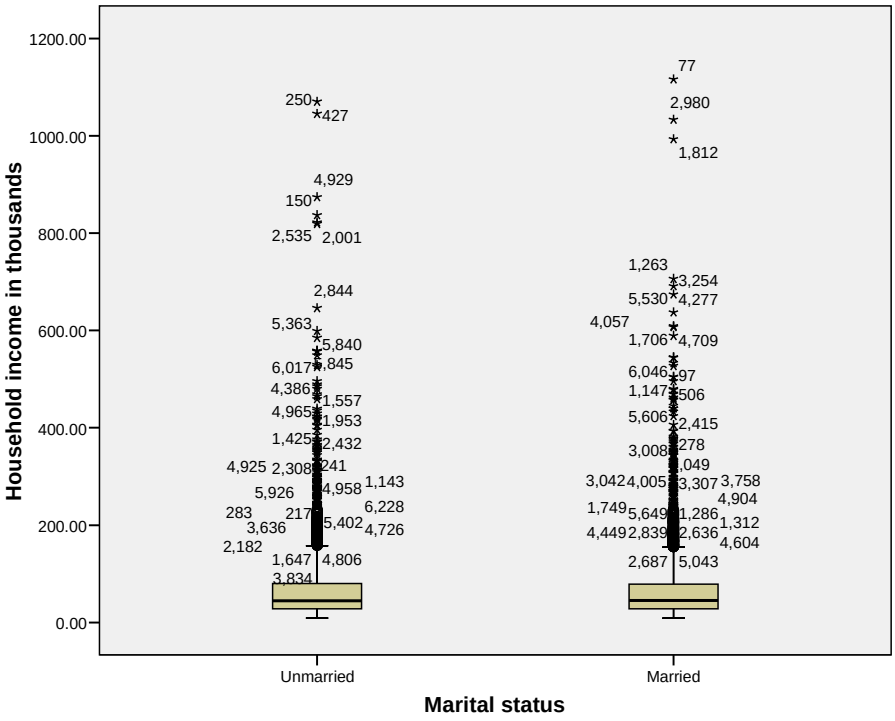
اگر میانگین وسط box باشد متقارن است و اگر بالاتر یا پایین تراز وسط باشد نشانه چولگی است.
☆نشان دهنده یک داده پرت است که شماره ۱ آن نشان میدهد که مربوط به case شماره ۱ است.
به داده پرت outliar می گویند.
به طور کلی داده ای که این شرط را داشته باشد پرت حساب می شود:

$$Q = Q_3 - Q_1$$

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

به داده های خیلی پرت نیز Extreme variable می گویندکه 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 باشدیعنی 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 ……..Martial

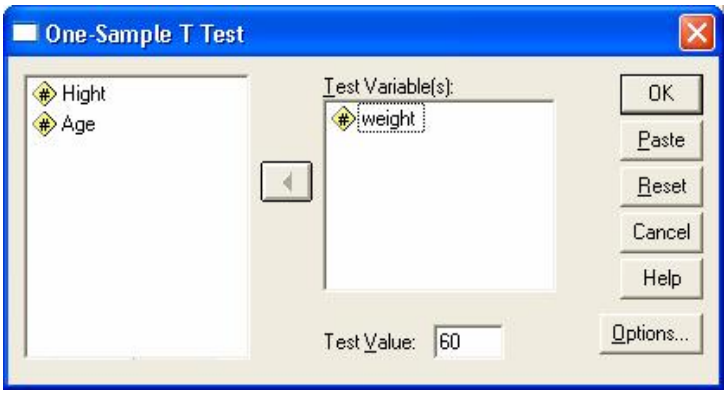


این نمودار نشان میدهد که نقاط پرت بسیاری وجود دارد

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 اجراي اين دستور امكان مقايسه و آزمون t براي سنجش آزمون به كار مي‌رود.
- در اين آزمون فرض بر نرمال بودن داده‌ها است.
- متغير معرفي شده بايد مقادير ***عددي*** داشته باشد.
- اگر داده‌ها مقداري چولگي هم داشته باشند اين آزمون دربرابر اشتباه مقاومت خواهد كرد.

آزمون به شكل زير در نظر گرفته شده است. (μ مقدارميانگين نمونه است و 80 مقدار حدسي است كه از جمعيت براي ميانگين مي‌زنيم.)

Analyze – compare Means – One_ sample T Test / Test variable .....Income
Test value 80

در قسمت Test variables متغير يا متغيرهايي كه ميانگينشان بايد با مقدار test value (در اينجا مثلا 80) مقايسه شود قرار مي‌گيرد.
با انتخاب دكمه options نيز امكان تعريف درصد فاصله اطمينان (مثلا يك فاصله اطمينان 95%) وجود دارد.
دقت داشته باشيم فاصله اطمينان نشان دهنده درصد فواصلي از نمونه‌هاي مختلف است كه شامل ميانگين مي‌باشند. براي مثال يك فاصله اطمينان ۹۵% نشان مي‌دهد، در ۱۰۰ نمونه ديگر از جمعيت، ۹۵ نمونه، ميانگين نمونه‌ها در اين فاصله خواهند بود.

نكته : وقتی می گویین 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؟

$$\begin{cases} H0: \mu = 80 \\ H1: \mu \neq 80 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H0: \mu - 80 = 0 \\ H1: \mu - 80 \neq 0 \end{cases}$$

```
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					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean 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.

$$T = \frac{x - \bar{x}}{\frac{s}{\sqrt{n}}} = \frac{69.47.48 - 80}{0.98398} = \frac{-10.52516}{0.98398} = -10.696 < 0$$

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

۱.چون مقدار $p - value$ از α 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 tبرابر یا بزرگتر از ۲ باشد، معنی دار است.

این آزمون را برای عدد ۷۰ نیز تکرار می کنیم:

$$\begin{cases} H0: \mu = 70 \\ H1: \mu \neq 70 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H0: \mu - 70 = 0 \\ H1: \mu - 70 \neq 0 \end{cases}$$

```
T-TEST
  /TESTVAL = 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 :

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

۱.چون مقدار $p-value$ از α 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 tبرابر یا بزرگتر از ۲ باشد، معنی دار است.

نکته : در پنجره One_ sample T Test در قسمت Option میتوان مقدار خطای نوع اول یا α را تغییر داد. اگر ضریب اطمینان را تا ۹۹% بالابریم یعنی α را ۱% بگذاریم بیشتر فرضیه ها رد می شوند.

مثال :

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

Analyze- report – frequency

آزمونی انجام دهید با فرض اینکه میانگین وزن ۷۵ است.

$$\begin{cases} H0: \mu = 75 \\ H1: \mu \neq 75 \end{cases} \quad SPSS \Rightarrow \quad \begin{cases} H0: \mu - 75 = 0 \\ H1: \mu - 75 \neq 0 \end{cases}$$

```
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 :

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

۱.چون مقدار $p-value$ از α 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 tبرابر یا بزرگتر از ۲ 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 :

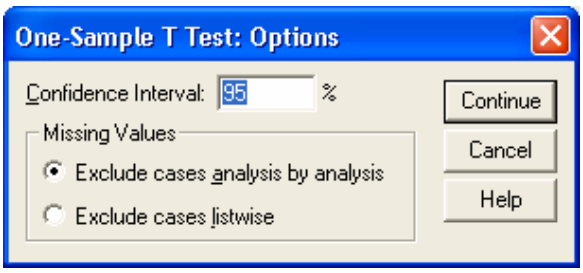
Sig.(2-tailed) = $p-value = 0.024$ <0.05 = α

۱.چون مقدار $p-value$ (متغیرweight-after) از α 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 tبرابر یا بزرگتر از ۲ باشد، معنی دار است.



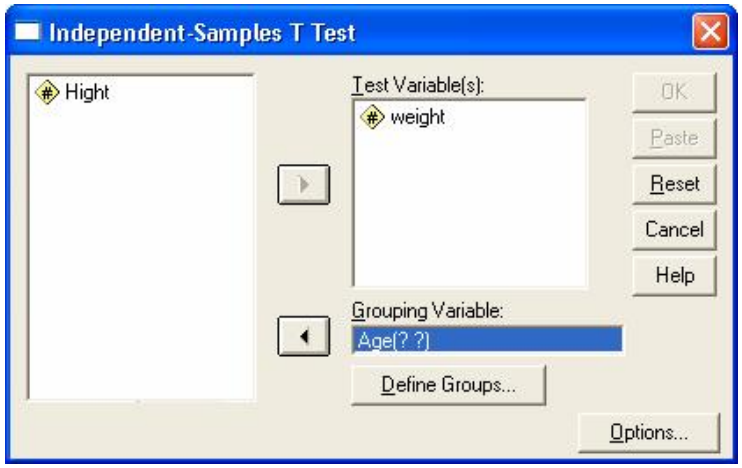
نکته: اگر در گزینه option exclude cases analysis by را تیک بزنیم و missing داشته باشیم در analysis مقایسه دو متغیر هر کدام را جداگانه حساب می کند و آزمون براي متغيرهاي به کار رفته را بدون در نظر گرفته مقادیر گمشده براي متغيرهاي ديگر به کار مي‌رود.
ولی اگر گزینه دوم یعنیexclude cases listwise را تیک بزنیم کل مورد یا case ی کهmissing دارد را حذف می کند. مورد با مقدار گمشده براي همه متغيرهاي به کار رفته در آزمون در نظر گرفته نخواهد شد. این missing چه در اثرتایپ 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) :

- براي آزمون برابر بودن ميانگين يك ويژگي جمعيت در بين دو گروه از اين روش استفاده مي‌كنيم.
- فرض براين است كه ويژگي جمعيت به صورت نرمال توزيع شده است و متغير يك مقدار كمي (**عددی**) است.
- همچنين نمونه به صورت مستقل و تصادفي انتخاب شده است.
- تقارن و عدم وجود نقاط پرت در اين آزمون از اهميت برخوردار است.
- اين آزمون نسبت به چولگي حساس نيست.
- **برابر بودن واریانس‌ها** با برابر نبودن آن‌ها در دو گروه، خروجي آزمون را تغيير مي‌دهد.

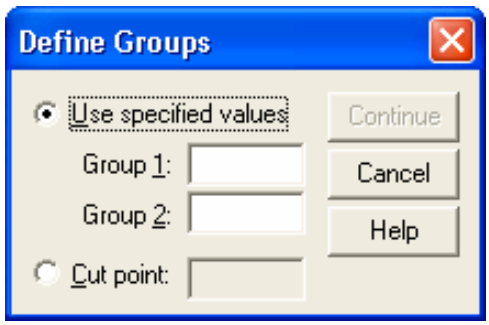
براي دسترسي به اين دستور مسير زير را طي كنيد.

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



در قسمت test variable :
متغير يا متغيرهايي كه بايد آزمون شوند را وارد كنيد.
در قسمت Grouping Variable :
متغير **گروه‌بندی** را مشخص كنيد.
اين آزمون فقط مي‌تواند مقايسه بين دو گروه را بدست آورد.
با انتخاب دكمه define group براي متغير گروه‌بندي مقدار مورد نظر براي تعريف دو گروه را وارد كنيد. حتی اگر متغير گروه بندی بيشتر از دو گروه داشته باشد اين آزمون فقط در دو گروه محاسبه خواهد شد.
همچنين با تنظيم مقدار cut point نيز امكان تفكيك گروه‌ها به صورت كمتر و بيشتر و مساوي مقدار cut point نيز هست.

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



- مقدار Group1 براي متغير گروه‌بندي، دسته اول را مشخص مي‌كند.
- مقدار Group2 براي متغير گروه‌بندي، دسته دوم رامشخص مي‌كند.
- اگر متغير رده بندی، عددی بود ميتوان با تعيين مقدار Cut Point دو گروه را به صورت زير تعريف كرد:
- گروه ۱ مقادير بيشتر و مساوي از cut point
- گروه ۲ مقادير كمتر از cut point

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

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

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						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error 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 $p-value$ (متغیرweight) که مربوط به واریانس است از α 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 tبرابر 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.

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

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						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error 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 $p-value$ (متغیرweight) که مربوط به واریانس است از α 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 tبرابر یا بزرگتر از ۲ باشد، معنی دار است.

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

m_1 میانگین متغیر درآمد income در گروه تحصیلات تکمیل نشده

m_2 میانگین متغیر درآمد income در گروه تحصیلات عالیه

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

```
.
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						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error 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 α 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 α 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 عالیه برابر نیست.

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

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

- اگر ميانگين دو متغير در نمونه بايد مقايسه شوند از اين آزمون استفاده مي‌كنيم.
- در اين آزمون مشخصه‌هاي اندازه گيري شده به صورت زوج مرتب (x,y) در نظر گرفته مي‌شوند كه در آن x ويژگي اول و y ويژگي دوم در نمونه است.
- براي مثال مي‌خواهيم اثر يك دارو را در كاهش فشار خون مشخص كنيم. به افراد نمونه در ابتدا دارونما داده فشار خون را اندازه‌گيري مي‌كنيم. سپس از خوردادن دارو واقعي نيز فشار خون اندازه‌گيري خواهد شد. آزمون مي‌خواهد ميانگين فشار خون را قبل و بعد از داروي واقعي مقايسه كند.
- فرض بر اين است مقادير ***عددی*** هستند.
- داده‌ها هر دو متغير توزيع نرمال دارند. (يا تفاوت آنها داراي توزيع نرمال است.)
- نمونه‌ها به صورت مستقل و تصادفي انتخاب شده‌اند.
- تعداد نمونه‌ها براي هر دو متغير بايد ***برابر*** باشد. در صورت وجود مقدار گمشده در يك متغير به طور كامل آن مورد ناديده گرفته خواهد شد.
- تفاوت در مقدار واريانس دو متغير اشكالي ندارد.

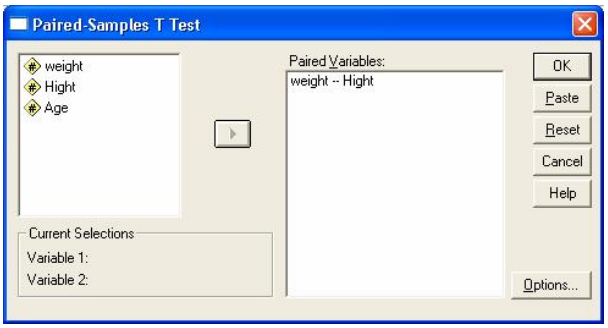
براي دسترسي به اين آزمون مراحل زير را طي كنيد.

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

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

m_1 ميانگين متغيراول

m_2 ميانگين متغير دوم



هر دو متغير را در قسمت سمت چپ انتخاب كرده و به قسمت pair Variable انتقال دهيد.

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

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

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

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 α ، 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 tبرابر یا بزرگتر از ۲ باشد، معنی دار است.

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

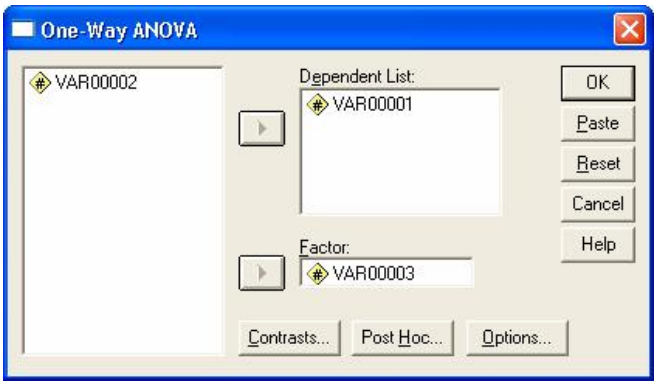
با استفاده از جدول آنالیز واریانس یك طرفه امکان مقایسه مقدار يك شاخص در بین چندین گروه وجود دارد.
با توجه به آماره‌هاي جدول آنالیز واریانس اگر نسبت مقدار میانگین مربعات خطاي درون گروهي (مجموع مربعات اختلاف مقادیر هر گروه از میانگین گروه) با مقدار میانگین مربعات خطاي بین گروهي (مجموع مربعات اختلاف میانگین هر گروه از میانگین کل) داراي اختلاف زیادي باشد نشان دهنده نقش عامل در تغییر میانگین گروه را نشان مي‌دهد.

فرض‌هاي اولیه :

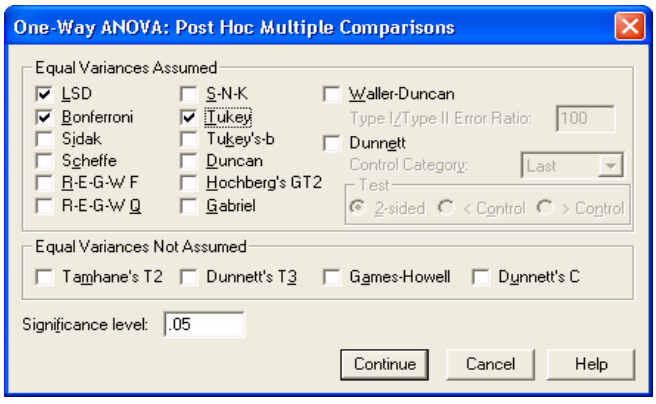
- متغیر فاکتور باید مقادیر صحیح داشته باشد.
- متغیر وابسته باید داراي مقادیر پیوسته باشد.
- نمونه در هر گروه باید داراي توزیع نرمال تصادفي باشند.
- آزمون آنالیز واریانس نسبت به شرط نرمال بودن مقاوم است ولي باید توزیع داراي تقارن باشد.
- گروه‌ها باید از جمعیتي با ***واریانس برابر*** انتخاب شده باشند. (این شرط را با استفاده از آماره Levens مي‌توان اندازه گیري کرد)

با استفاده از این جدول امکان اندازه گیري تغییر در مقدار میانگین در بین گروه‌ها وجود دارد.
براي دسترسي به این دستور مراحل زیر را طي کنید.

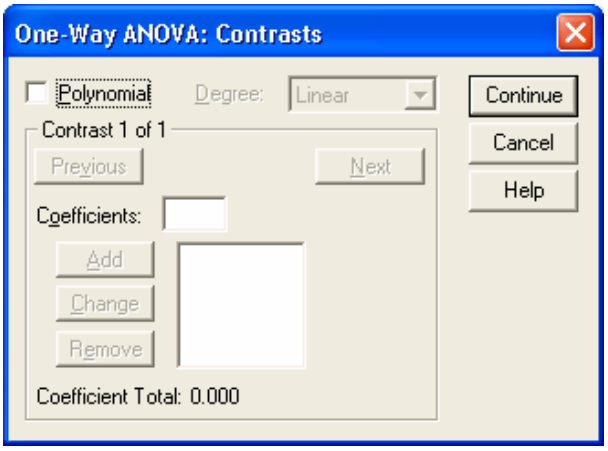
Analyze - Compare Means - One-Way ANOVA...



متغیر وابسته را در محل Dependent List وارد کنید.
متغیر عامل (متغیري که مقدار متغیر وابسته را نسبت به گروه‌هاي مختلف آن مي‌خواهید آزمون کنید.) را در قسمت Factor وارد کنید.
مي‌توان ***چندین*** متغیر وابسته را در لیست وارد کنید ولي براي متغیر فاکتور فقط از ***يك*** متغیر مي‌توان استفاده کرد.
با این کار میانگین هر متغیر در سطوح متغیر فاکتور مقایسه مي‌شوند.



اگر مي‌خواهید آزمون مربوط به دوتایي‌ها را انجام دهید گزینه Post Hoc را انتخاب کنید.
نوع آزمون را مي‌توانید در صورت برابر بودن واریانس‌ها یکي از گزینه‌هاي کادر اولي مانند LSD- tukey-Bunfferoni انتخاب کنید.
سپس دکمه Continueرا بزنید.
اگر مي‌خواهید ترکیبي خطي از میانگین‌هاي گروه‌ها را با هم مقایسه کنید از دکمه Contrast استفاده کنید.
توجه داشته باشید که بهتر است مجموع ضرایب ترکیب خطي 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 يك رابطه خطي بين ميانگين‌ها احتياج داريد گزينه Polynomial را انتخاب کنید.
دکمه Continue را بزنید.

آزمون مقایسه میانگین چند متغیر :

فرض آزمون به صورت زیر خواهد بود.
H0: m1=m2=m3
H1: m1#m2#m3

در صورتي كه فرض صفر رد شود مي‌توان با استفاده از آزمون‌هاي دوتايي مشخص كرد كه علت عدم تساوي براي كدام زوج بوده‌است.
اين آزمون را به عنوان Post Hoc نيز مي‌شناسند.
همچنين امكان آزمون تركيبي از ميانگين‌ها (Contrast) وجود دارد.

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

Analyze - Compare Means - One-Way Anova

$$\begin{cases} H0: m_1 = m_2 = m_3 = m_4 = m_5 \\ H1: m_1 \neq m_2 \neq m_3 \neq m_4 \neq m_5 \end{cases}$$

m_1 میانگین متغیر درآمد در گروه اول (did not completed high school)

m_2 میانگین متغیر درآمد در گروه دوم(high sckool 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 α ، 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 خطاي درون گروهي (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 $Sig. = 0.000 = p-value < \alpha$

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 Tamhane's T2 ، Dunnet's T3 ، Games Howell ، Dunnet's C استفاده می شوند.

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

نکته: درنتایج 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 کردهو 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	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
	m_1	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	Did not complete high school-m1	6.34094	2.33139	.064	-.1905	12.8724
		Some college-m3	-3.92743	2.74970	.811	-11.6318	3.7769
	m_2	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	Did not complete high school-m1	10.26837(*)	2.74450	.002	2.5781	17.9587
		High school degree-m2	3.92743	2.74970	.811	-3.7769	11.6318
	m_3	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	Did not complete high school-m1	18.78400(*)	3.01158	.000	10.3447	27.2233
		High school degree-m2	12.44306(*)	3.01632	.000	3.9909	20.8952
	m_4	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	Did not complete high school-m1	27.30373(*)	5.26659	.000	12.4846	42.1229
		High school degree-m2	20.96279(*)	5.26930	.001	6.1362	35.7894
	m_5	Some college-m3	17.03536(*)	5.46465	.019	1.6691	32.4016
		College degree-m4	8.51973	5.60355	.749	-7.2309	24.2704

نتیجه هر چهار آزمون :
۱.میانگین درآمد در گروه هاي پشت سر هم برابر است مثل m1 و m2
ولي در گروههاي با فاصله برابر نيست مثل m1 و m3

$m_1 \neq m_3$ $m_1 = m_2$
$m_1 \neq m_4$ $m_2 = m_3$
$m_1 \neq m_5$ $m_4 = m_3$
$m_2 \neq m_4$ $m_4 = m_5$
$m_2 \neq m_5$
$m_5 \neq m_3$

	(I) Level of education	(J) Level of education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Upper Bound	Lower Bound
Dunnett T3	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
	m_1	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
	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
	m_2	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
	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
	m_3	College degree-m4	-8.51563	3.34591	.104	-17.8905	.8593
		Post-undergraduate degree-m5	-17.03536(*)	5.46465	.019	-32.3985	-1.6723
	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
	m_4	Some college-m3	8.51563	3.34591	.104	-.8593	17.8905
		Post-undergraduate degree-m5	-8.51973	5.60355	.747	-24.2675	7.2280
	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
	m_5	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
Dunnett C	Did not complete high school	High school degree-m2	-6.34094	2.33139		-12.7076	.0257
		Some college-m3	-10.26837(*)	2.74450		-17.7647	-2.7720
	m_1	College degree-m4	-18.78400(*)	3.01158		-27.0099	-10.5581
		Post-undergraduate degree-m5	-27.30373(*)	5.26659		-41.7380	-12.8695
	High school degree	Did not complete high school-m1	6.34094	2.33139		-.0257	12.7076
		Some college-m3	-3.92743	2.74970		-11.4370	3.5821
	m_2	College degree-m4	-12.44306(*)	3.01632		-20.6810	-4.2051
		Post-undergraduate degree-m5	-20.96279(*)	5.26930		-35.4039	-6.5217
	Some college	Did not complete high school-m1	10.26837(*)	2.74450		2.7720	17.7647
		High school degree-m2	3.92743	2.74970		-3.5821	11.4370
		College degree-m4	-8.51563	3.34591		-17.6548	.6236
	m_3	Post-undergraduate degree-m5	-17.03536(*)	5.46465		-32.0089	-2.0618
	College degree	Did not complete high school-m1	18.78400(*)	3.01158		10.5581	27.0099
		High school degree-m2	12.44306(*)	3.01632		4.2051	20.6810
	m_4	Some college-m3	8.51563	3.34591		-.6236	17.6548
		Post-undergraduate degree-m5	-8.51973	5.60355		-23.8715	6.8320
	Post-undergraduate degree	Did not complete high school-m1	27.30373(*)	5.26659		12.8695	41.7380
		High school degree-m2	20.96279(*)	5.26930		6.5217	35.4039
	m_5	Some college-m3	17.03536(*)	5.46465		2.0618	32.0089
		College degree-m4	8.51973	5.60355		-6.8320	23.8715

* 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	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
	m_1	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
	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
	m_2	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
	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
	m_3	College degree-m4	-8.51563	3.34591	.081	-17.6488	.6175
		Post-undergraduate degree-m5	-17.03536(*)	5.46465	.016	-31.9958	-2.0749
	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
	m_4	Some college-m3	8.51563	3.34591	.081	-.6175	17.6488
		Post-undergraduate degree-m5	-8.51973	5.60355	.550	-23.8554	6.8160
	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
	m_5	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.
بامحاسبه $x-\bar{x}$ 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 :

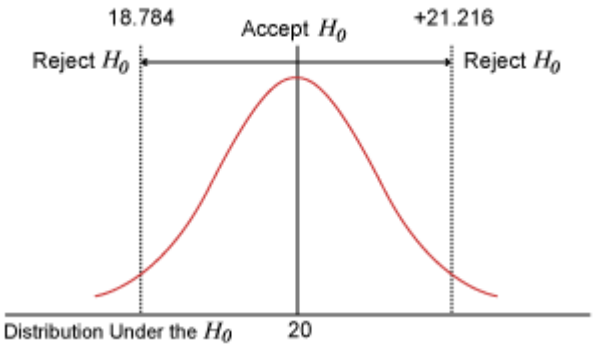
دلائل استفاده از آزمون ناپارامتری:
- توزیع داده ها نرمال نباشد.
- اختلاف پراکندگی زیاد باشد.
- اندازه نمونه کم باشد.

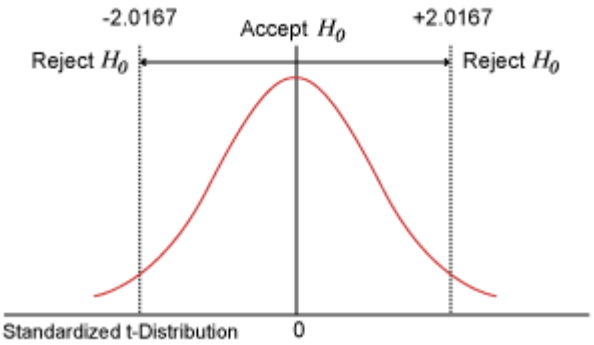
چرا از ابتدا بدون دانستن شرایط لازم از آزمون ناپارامتری استفاده نمی کنیم؟
زیرا توان این آزمون در شرایطی که توزیع داده ها نرمال باشد کم است.
فرض صفر توزیع جمعیت را مشخص می کند اما فرض مقابل، فرض ادعایی است و برای رد فرض صفر ادعا شده است. و هدف آزمون این است که با ارائه یک نمونه فرض صفر را رد کنیم.
هیچ وقت نمی گوییم 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 بااین آزمون کاربیهوده ای انجام شده است.برای جلوگیری از اینکاربهتر است ازاول فرضیه را درست انتخاب کنیم نمونه معمولی بگیریم وشاخص را استخراج کنیم بعدحجم نمونه و بقیه را انتخاب کنیم.(α هم خطایی است که محقق میتواندتحمل کند).

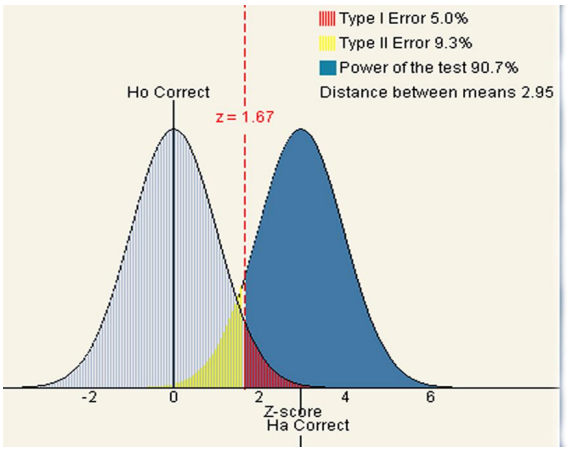
توان آزمون :
اگر واقعاً فرض صفر غلط باشد و توسط آزمون نمونه هم غلط شود بدون خطا بوده ایم.(تصمیم درست) وگرنه در غیر اینصورت دچار خطا شده ایم :
خطای نوع اول:رد فرض صفر توسط نمونه به شرطی که فرض صفر در جمعیت صحیح بوده است.(α)
خطای نوع دوم:قبول فرض صفر توسط نمونه به شرطی که فرض صفر در جمعیت غلط بوده است.(β)
سطح اطمینان ($1-\alpha$)
توان آزمون ($1-\beta$)
سطح معنی داری و خطای نوع اول (α)
خطای نوع دوم (β)

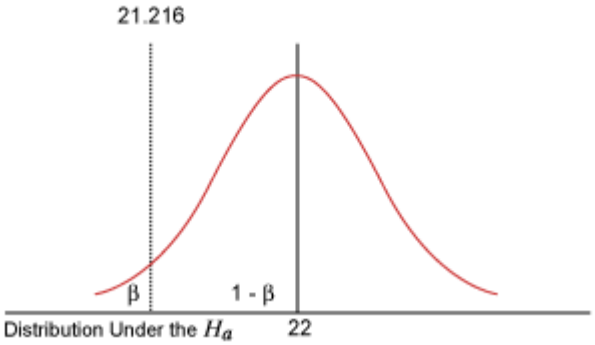
	H_0 رد	H_0 عدم رد
H_0 صحیح	α	$1-\alpha$
H_0 غلط	$1-\beta$	β

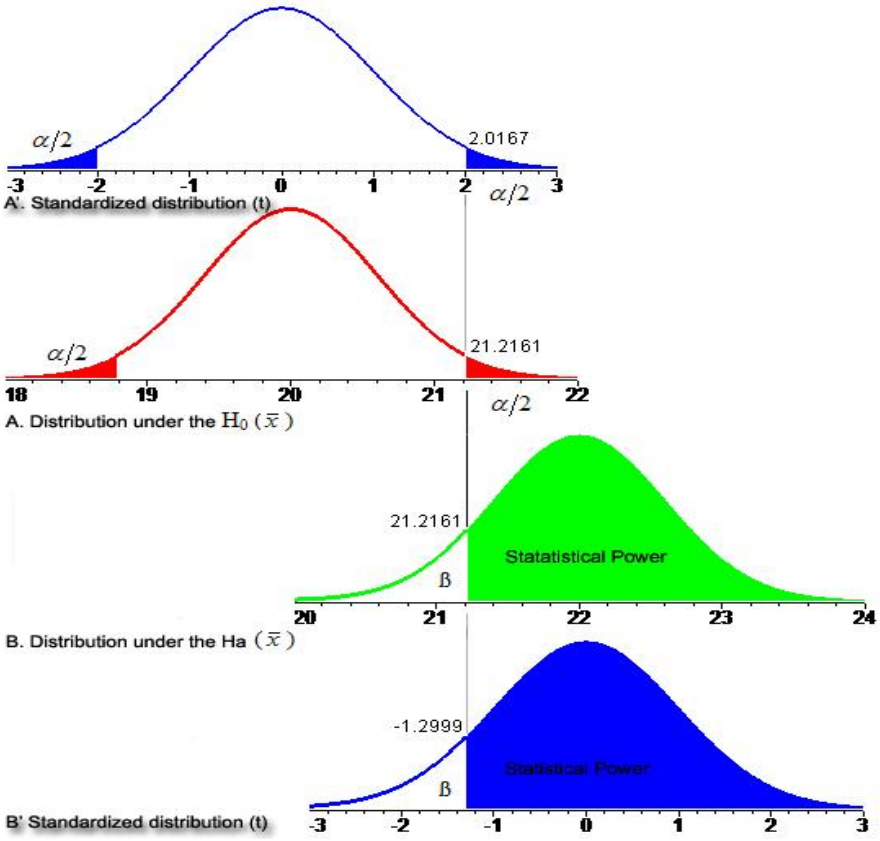
سطح اطمینان ۹۵% یعنی اگر ۱۰۰ نمونه بگیریم میانگین ۹۵ تای آن ($\bar{x}$) در فاصله اطمینان می افتد.
از آزمونهایی باید استفاده کنیم که خطای نوع اول کمتری داشته باشنددر پزشکی(۰.۰۱) و در صنعتی(۰.۰۵)
پرتوان ترین آزمون آزمونی است که $1-\beta$ آن بیشترین باشد. یعنی رد فرض صفر به شرط غلط بودن آن.
سئوال : در باره رابطه $\alpha+\beta\prec 1$ تحقیق کنید.


18.784
Accept H_0
+21.216
Reject H_0
Reject H_0
Distribution Under the H_0
20


-2.0167
Accept H_0
+2.0167
Reject H_0
Reject H_0
Standardized t-Distribution
0


Type I Error 5.0%
Type II Error 9.3%
Power of the test 90.7%
Distance between means 2.95
Ho Correct
z = 1.67
-2
0
2
4
6
Z-score
Ha Correct


21.216
β
1 - β
Distribution Under the H_a
22


$\alpha/2$
2.0167
-3
-2
-1
0
1
2
$\alpha/2$
3
A'. Standardized distribution (t)
$\alpha/2$
21.2161
18
19
20
21
$\alpha/2$
22
A. Distribution under the H_0 ($\bar{x}$)
21.2161
Statatistical Power
ß
20
21
22
23
24
B. Distribution under the Ha ($\bar{x}$)
-1.2999
Statistical Power
ß
-3
-2
-1
0
1
2
3
B' Standardized distribution (t)

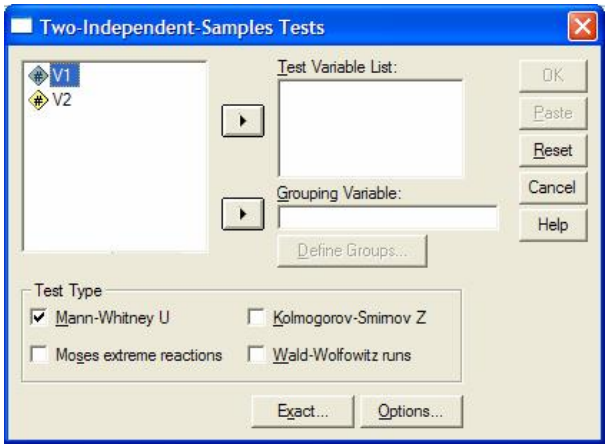
برای انجام آزمون های ناپارامتری در SPSS از دستور زیر استفاده می کنیم :

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 وو 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 یاتوزیع نرمال نباشداز 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.درصورتی 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.

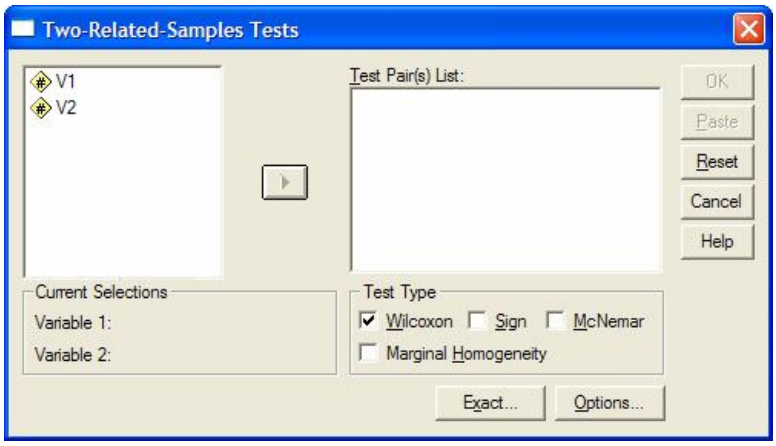
برای آزمون مقایسه میانگین بین چند گروه از دستور زیر استفاده می کنیم:

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 درنظر 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.

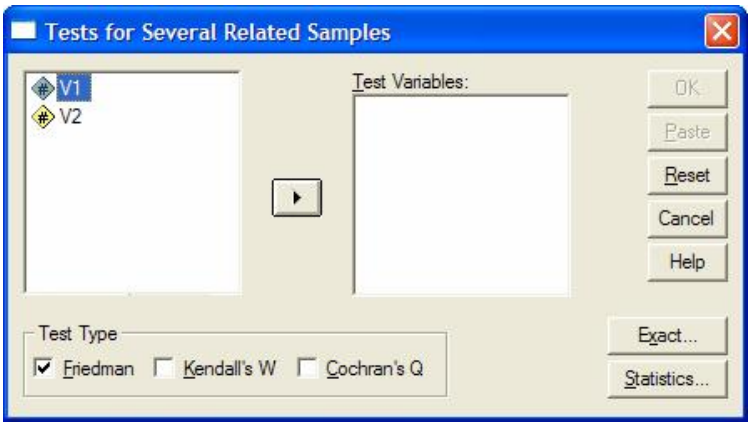
۳. McNemar اگرداده‌ها به صورت درست و غلط هستند (۰و۱) بهتر است ازاین آزمون استفاده کنید. معمولا از این آزمون براي مقايسه قبل و بعد از يك رفتار به كار گرفته مي‌شود.

اگر داده‌هاي چندتايي باشند (تعميمي از روش McNemar) از اين روش استفاده كنيد. داده‌ها، گروه بندي هستند.

آزمون مقايسه ميانگين نمونه‌هاي چندتايي :

Analyze – Nonparametric Tests – K Related samples

انواع آزمون :



۱. آزمون فريدمن مانند روشهاي پارامتري مقايسه ميانگين (توزيع) چند متغير را آزمون مي‌كند. با استفاده از رتبه گذاري براي هر گروه و مقايسه رتبه‌ها آزمون انجام مي‌شود.

۲. آزمون كندال با استفاده از توافق (همبستگي رتبه‌اي) اين آزمون را انجام مي‌دهد.

۳. آزمون كاكران نيز مانند آزمون فريدمن عمل كرده ولي با داده‌هاي دوتايي (۰و۱) محاسبات را انجام مي‌دهد. براي مقايسه درست و غلط مناسب است.

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

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

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؟
٤. 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 دوراه می توان استفاده کرد:
Analyze – Descriptive statistic- descriptive
Analyze – Descriptive statistic- Explore
فرق این دوراه این است که با استفاده از Explore میتوان متغیر زمان را در سه قسمت objects جداگانه مشاهده نمود.

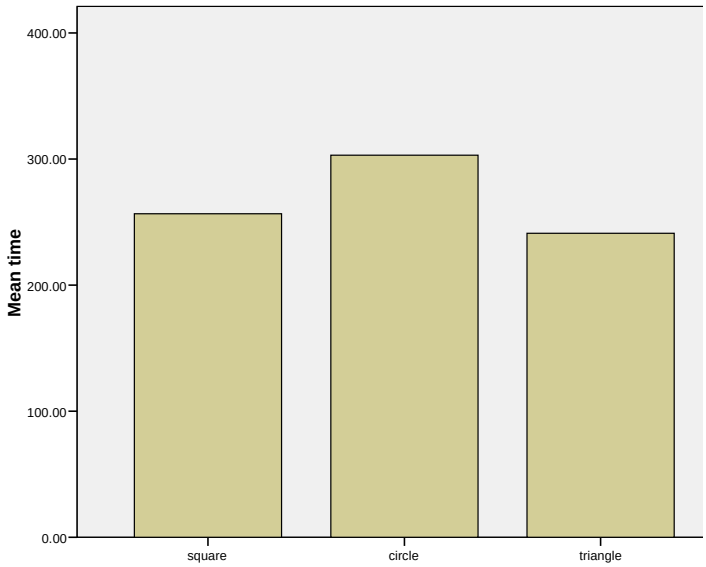
Descriptive Statistics

	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

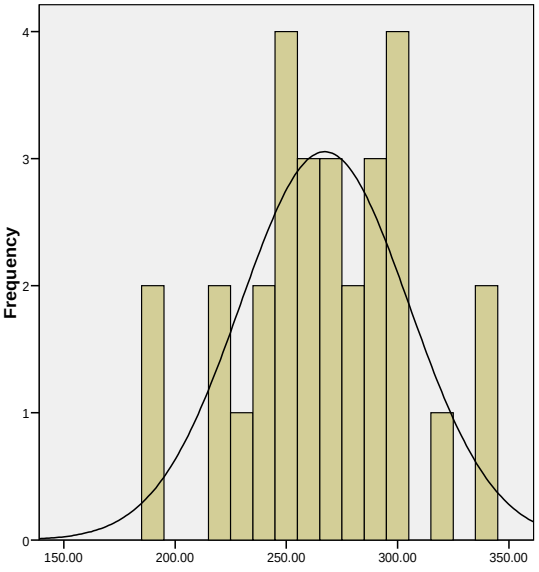
Category Axis……objects

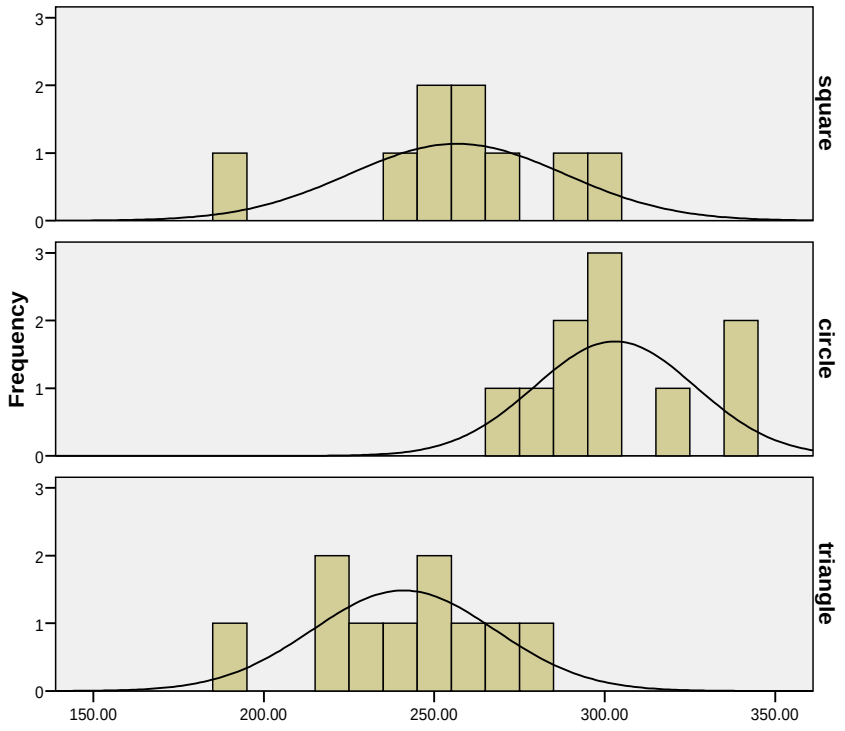


۳. وجود توزیع نرمال:

Graph – Legacy Dialogs - Histogram-display normal curve- variable……time

Panel……..objects





الف-حالت اول(بدون گروه بندی) : Analyze – Nonparametric Tests – 1-sample k-s...

One-Sample Kolmogorov-Smirnov Test

		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 α ، فرض H_0 (نرمال بودن توزیع متغیر زمان) رد نمی شود.

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

ب- حالت دوم (با گروه بندی) : Data- split files – variable ....objects

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 α ، 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 α ، فرض H_0 (نرمال بودن توزیع متغیر زمان) رد نمی شود. (در گروه دایره)

Sig. = $p-value = 0.556 > 0.05 = \alpha$

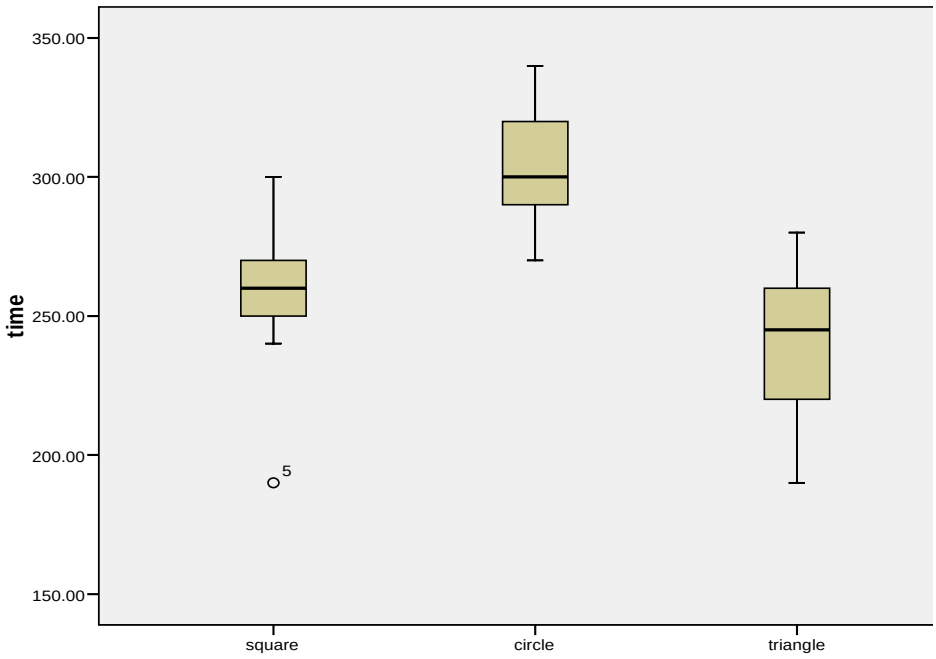
۳. با توجه به بزرگتر بودن $p-value$ از α ، فرض H_0 (نرمال بودن توزیع متغیر زمان) رد نمی شود. (در گروه مثلث)

Sig. = $p-value = 0.995 > 0.05 = \alpha$

٤. 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 :

حالت اول- ناپارامتری :

$$\begin{cases} H\ 0 : m_1 = m_2 = m_3 \\ H\ 1 : m_1 \neq m_2 \neq m_3 \end{cases}$$

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 α ، فرض H_0 (نرمال بودن توزیع متغیر زمان) رد می شود.

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

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 α ، فرض H_0 (نرمال بودن توزیع متغیر زمان) رد می شود.

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

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

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 Sig. = $p-value = 0.936 > \alpha$

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** دوبه 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.

.

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)

$m_1 = m_3$

$m_1 \neq m_2$

$m_2 \neq m_3$

« جلسه ششم – ۸۷/۱/۲۹ »

آزمون Chi-square (χ^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 (متغیر های گسسته)

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

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

$H\ 0$: نمونه پارامتر جمعیت را تایید می کند
$H\ 1$: نمونه پارامتر جمعیت را تایید نمی کند

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

- آزمون استقلال – نيكويي برازش
- فرضها
  - نمونه تصادفي
  - شکل یا توزیع داده ها اهمیت ندارد.
  - حداقل فراواني در هر گروه حداقل ۱
  - ۲۰% گروه ها فراواني بيش از ۵ داشته باشند.

این آزمون با استفاده از آماره خي ۲ مقادير مشاهده شده را با مقادير مورد انتظار مقايسه مي‌كند.
روش دسترسي به آزمون :
- Descriptive Statistics - Crosstabs... ... Analyze

بعدازتعیین متغیرها دکمه statistic را بزنید و گزینه chi square را انتخاب کنید.

یا Analyze – nonparametric – chi-square

مثال: فرض کنید توزیع یکنواخت است و احتمال مشاهده هر یک از داده ها با دیگری برابر است. ۵ 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؟

نوع اسباب بازی	تفنگ	بازی کامپیوتر	عروسک	ماشین	سه چرخه
مقدار خریداری شده	٤٥	٢٠	١٥	١٥	٥

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.
۴. Analyze – nonparametric – chi-square – Test variable .....Typeoftoy
۵. 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
bycicle	5	20.0	-15.0
Total	100		

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

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:

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

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

در پنجره آزمون Chi-Square در قسمت Expected Range ، گزینه Used specified range ، برای lower=1 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 و۴۳.۳ را وارد می کنیم.

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 دانشگاههاي مركز استان و شهرستان به تفكيك دانشكده آمار ثبت نام آورده شده است. آيا نوع دانشكده در انتخاب دانشجو تاثير دارد يانه؟ (نسبت قبولی توزیع یکنواخت دارد یا نه یعنی تعداد قبولی به یک نسبت بین دانشگاهها تقسیم می شوند؟) آيا نسبت قبولی بین دانشکده مهندسی و هنر ۸۰ به ۲۰ است؟ (يعني مهندسي ۸۰% و هنر ۲۰% توسط دانشجويان انتخاب مي شود.)

نوع دانشكده	مهندسي	هنر	اقتصاد	ساير
مركز استان	١٦	١٤	١٣	١٣
شهرستان	١٤	٦	١٠	٨

انتخاب متغیر ها :

	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.

facaulty

	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

	facaulty
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 $50*0.8=40$
۴۰ و۱۰را وارد می کنیم. یا نسبت آنها را به درصد یعنی خود ۰.۸ ۰.۲ را وارد کنیم. $50*0.2=10$

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 دوجمله 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 پيش آمد سنجيده مي‌شود. از آنجايي كه رخداد يا عدم رخداد پيش آمد توزيع دوجمله اي (Binomial) از اين توزيع براي آزمون استفاده مي‌شود.
روش دسترسي به اين آزمون

Analyze - Nonparametric Tests - Binomial...

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

$$\begin{cases} H0: p = \frac{1}{2} \\ H1: p \neq \frac{1}{2} \end{cases}$$

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	.50	.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 و0.5 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 ۱و۲ 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	.7	.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 بندي داده ها امکان تصادفي بودن داده‌ها آزمون مي‌شود.
تصادفی بودن داده ها زمانی حاصل می شود که دو داده متوالی یکی بالای میانگین و یکی پایین تر از آن باشد.
روش دسترسي به این آزمون
Analyze – Nonparametric Tests – Runs

با استفاده از اکسل (تابع Rand() =)۱۰۰ عدد تصادفي ایجاد کنید و با امکانات spss آزمون کنید که این اعداد تصادفي هستند.
نحوه ورود اطلاعات از اکسل را مرور کنید. کافي است در هنگام باز کردن سند اطلاعاتي (data) در قسمت file type 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
	شیفت۱	شیفت۲	شیفت۳	شیفت۴

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 گوییم 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 همبستگي به 1یا 1- نزدیک باشد رابطه خطي بین دو متغیر وجود دارد.

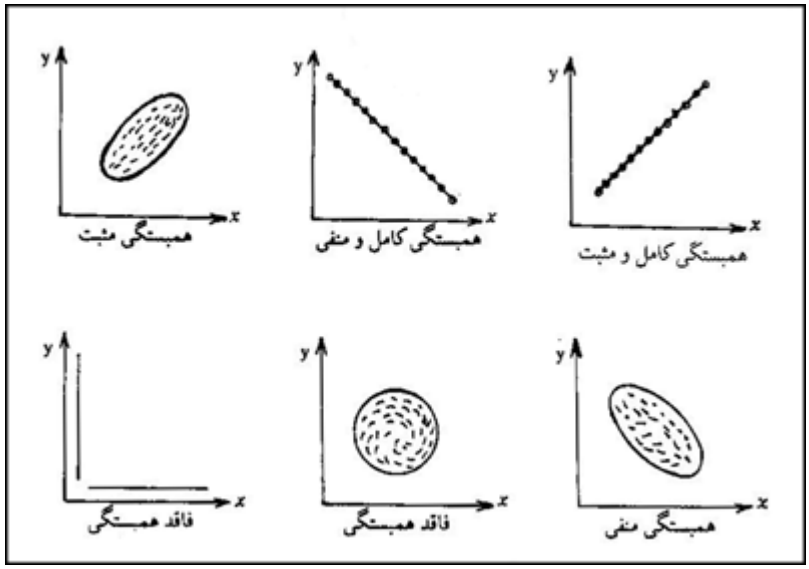
ضریب همبستگی :

- این ضریب نشان دهنده وجود ارتباط بین دو متغیر است.
- هرچه ارتباط دو متغیر شدید باشد مقدار ضریب به ۱+ و یا ۱- نزدیکتر است.
- با كاهش ارتباط بين دو متغير مقدار ضريب به صفر نزديك مي‌شود.
- اگر دو متغیر مستقل باشند مقدار ضریب همبستگي برابر صفر خواهد بود.
- اگر دو متغیر نرمال باشند اگر ضریب همبستگي برابر صفر باشد مي‌توان استقلال را نتیجه گرفت در غیر این صورت خیر.

محاسبه ضریب همبستگی :

- براي محاسبه ضريب همبستگي پيرسون از مقدار Covariance استفاده مي‌شود.
- مقدارCovariance براي اندازه‌گيري ميزان ارتباط مناسب است ولي از آنجايي كه با واحد و مقياس مشخصه اندازه‌گيري شده بدست مي‌آيد بايد براي مقايسه ميزان همبستگی از یک ضریب بدون واحد (درصد) استفاده کرد.
- Correlation یا ضریب همبستگی این خاصیت را دارد.
- معمولا براي مشاهده ارتباط دو متغير از نمودار نقطه‌اي استفاده مي‌شود.
- 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 رویدادنشان با هم ارتباط دارد.

نمودار خط همبستگی:



محاسبه ضریب همبستگی پیرسون:

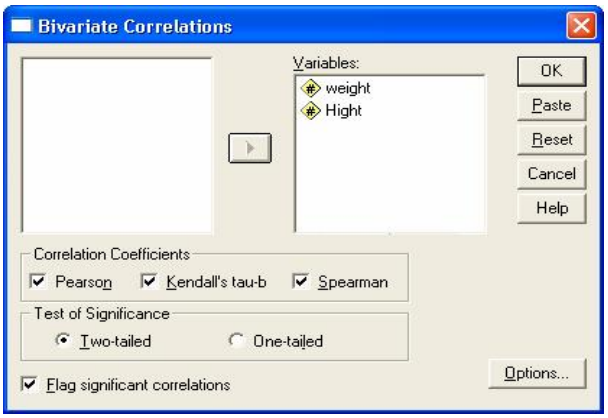
$$Co\,var\,iance = \frac{E(xy) - E(x).E(y)}{n-1}$$

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

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

براي دسترسي به اين دستور مسير زير را طي كنيد. ... Analyze –Corrolate – Bivariate

براي اجراي اين دستور حداقل به دو متغير عددي احتياج داريد.



انواع ضريب همبستگي :

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

- Covariance
- همبستگي Correlation
- ضريب همبستگي پارامتري :
- همبستگي خطي (ضريب همبستگي پيرسون – Pearson)

فرض بر این است که هر دو متغیرعددي داراي توزيع نرمال دو متغیره هستند.

ضريب همبستگي ناپارامتري :
- همبستگي رتبه‌اي (ضريب همبستگي اسپيرمن- Spearman)
- همبستگي ترتيبي (ضريب همبستگي كندال– Kendall's Tau_b)

آزمون ضريب همبستگي:

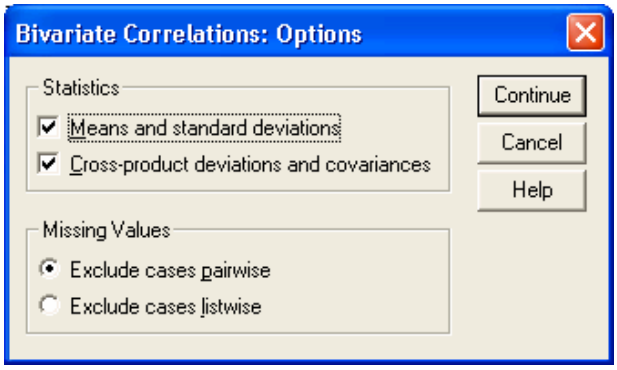
$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

-با انتخاب گزينه two tail يك آزمون دو طرفه براي مقدار ضريب همبستگي انجام مي‌شود.

-اگر علاوه بر مقدار ضريب همبستگي جهت آن نيز در نظر است بهتر است از آزمون يكطرفه (One Tail) استفاده شود.

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

-با انتخاب گزينه ...flagامكان تعيين معنادار بودن مقدار ضريب همبستگي در خروجي داده مي‌شود.



-با انتخاب option امكان نمايش آمار توصيفي (ميانگين و انحراف استاندارد) و نمايش مقدار كواريانس و مضرب تفاضلات وجود دارد. ضريب تفاضلات همان صورت محاسبه كواريانس است.

- با انتخاب گزينه Exclude Cases pairwise محاسبه براي جفت ها بدون در نظر گرفتن داده گمشده در ديگر متغيرهاي هر مورد محاسبه مي شود.

- با انتخاب گزينه Exclude cases listwise با وجود يك متغير كه مقدار گم شده در بررسي دارد كل مورد ناديده گرفته خواهد شد.

- امكان انتخاب بيش از دو متغير نيز در اين بررسي وجود دارد. به اين ترتيب يك ماتريس دو بعدي براي بررسي و محاسبه مقدار ضريب همبستگي جفت جفت متغيرها ايجاد خواهد شد.

براي رسم نمودار همبستگی از دستور زير استفاده می كنيم:
Graph – Legacy dialogs – scatter plot– simple يا 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☑

facaulty * place Crosstabulation
Count

		place		Total
		center	outer	center
facaulty	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:
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)

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

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 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 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_i فراوانی مورد انتظار i ام

مرکز استان :

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

شهرستان :

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

Pearson Chi – Square=0.62+0.91=1.524

نرم افزار SPSS می تواند مقادیر مورد انتظار یعنی E_I ها را حساب کند.

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

place * facaulty Crosstabulation

			facaulty				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 بررسي كنيد.

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

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.

نتيجه : $p-value=0$ پس نمونه فرض صفر را رد می کند و استقلال دو متغیر رد می شود میزان همبستگي در جدول چون دو متغير اسمي هستند اسپيرمن مناسب است كه عدد 0.105 را نشان مي دهد.

ضریب همبستگی جزیی:

با استفاده از اين ضريب امكان سنجش ميزان همبستگي خطي بين دو متغير در صورت كنترل يك متغير سوم (يا حذف تاثير آن) وجود دارد.

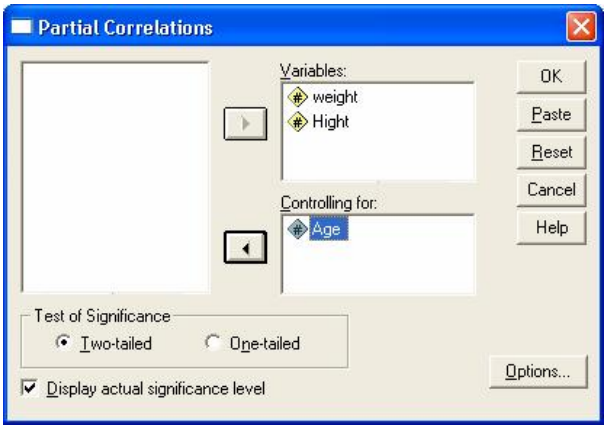
ممكن است دو متغير به واسطه متغير ديگري نيز رابطه داشته باشند. با محاسبه ضريب همبستگي جزئي ميزان رابطه خطي بين دو متغير با حذف اثر متغير سوم محاسبه خواهد شد.
روش محاسبه به صورت زير است:

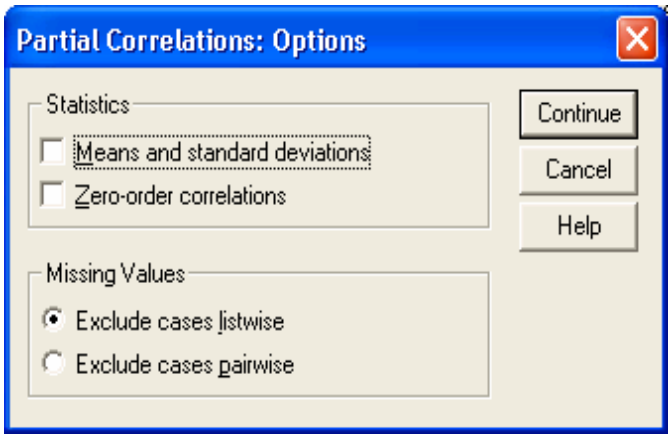
$$r_{ABC} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{(1-r^2_{AC})(1-r^2_{BC})}}$$

براي انجام اين دستور احتياج به حداقل ۳ متغير عددي است.
براي دسترسي به اين دستور مسير زير را طي كنيد.

Analyze - Correlate - Partial...

با انتخاب گزينه Option امكان نمايش آماره هاي توصيفي و نمايش جدول ضريب همبستگي ساده (Zero-order...) نيز وجود دارد.





مثـال- در فایـل demo سـه متغیـر سـن ، درآمـد و هزینـه خریـد ماشـین (age,income,car)را در نظـر بگیریدابتدا نمودار scatterplot را به طور ماتریسي براي اين سه متغير بكشيد.

Grafh - scatterplot - matrix

ضريب همبستگي بين سه متغير را حساب مي كنيم:

Analyze - correlate- bivariate …

بـا اسـتفاده از ضـريب همبـستگي جزيـي متغيـر income را ثابـت گرفتـه و مجـددا رابطـه بـين دو متغيـر age,car را بدست مي آوريم.

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).

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

كنترل 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

ضريب همبستگی age و car..... 0.193

نتيجه : در حالت اول ضريب همبـستگي بـين دو متغيـر age,car ، 0.376 بـود ولـي پـس از كنتـرل متغيـر income ، به 0.193 تغيير پيدا كرد كه اين عدد ارتباط خالص را نشان مي دهد.

رگرسیون Regression (معادله خط برگشت) :

براي سنجش ارتباط دو متغير سه راه وجود دارد:

۱. رسم نمودار (scatterplot)
۲.اندازه گيري ضريب همبستگي (Corrolation)
۳.محاسبه رگرسيون (خط برگشت)

رگرسيون پركاربردترين روش آماري است كه براي سنجش و ارائه مدل ارتباط دو متغير بكار مي رود.
رگرسيون خطي، ضرايب معادله خطي كه نزديكترين هماهنگي با داده‌هاي مشاهده شده را دارا است برآورد مي‌كند.
با استفاده از معادله خط برگشت امكان پيشگويي مقادير بعدي نيز وجود دارد.
در معادله رگرسيون يك متغير مستقل (Independent) و يك متغير وابسته (Dependent) داريم.
در معادله رگرسيون چند متغيره تعداد متغيرهاي مستقل بيش از يكي است.

مثال:
- برآورد ميزان فروش يك كالاي لوكس براساس درآمد خانوار يك كشور
- برآورد ميزان فروش فروشندگان يك فروشگاه براساس، سن، تجربه، ميزان تحصيلات.
- برآورد ميزان محصول براساس ميزان بارندگي، نوع كود، نوع بذر و استفاده از روشهاي مخصوص رفع آفات
- برآورد مساحت روشن شدن يك فضا توسط تير چراغ برق براساس ارتفاع تير

مدل خطي ساده :

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

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

- كه در آن x متغير مستقل و تحت كنترل ماست.
- a مقدار عرض از مبدا يا constant ناميده مي‌شود .
- b مقدار شيب خط يا ضريب متغير مستقل ناميده مي‌شود.
- e خطاي مشاهده و مدل رگرسيون است.
- Y|x متغير تصادفي وابسته به x است و داراي ميانگين a+bx و واريانس σ^2 است.
- متغير تصادفي ε خطاي تصادفي است و داراي ميانگين صفر و واريانس σ^2 است.
- پارامترهاي اين مدل عبارتند از a, b, σ^2
- Y مقدار واقعي و $\hat{Y}$ مقدار برآورد شده است.

مفروضات :
- براي هر مقدار متغير مستقل، توزيع متغير وابسته يك توزيع نرمال است. يا توزيع باقي مانده‌ها بايد نرمال و مستقل از مشاهدات ديگر باشد.
- براي همه مقادير متغير مستقل بايد واريانس متغير وابسته ثابت باشد.
- همه مشاهدات بايد تصادفي باشند.
- ارتباط دو متغير وابسته و مستقل بايد خطي 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 مي‌شوند.

$$\hat{b} = r\frac{s_y}{s_x} \qquad \hat{a} = \overline{y} - \hat{b}\overline{x}$$

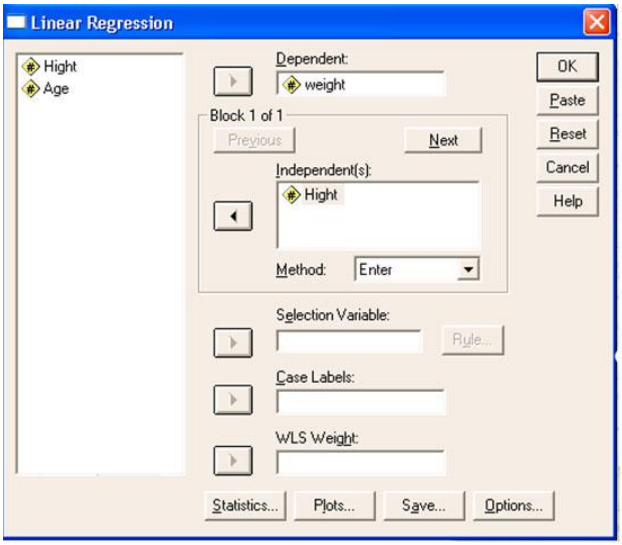
برآورد مقدار خطا (Residual) به صورت زیر محاسبه مي‌شوند:

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

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

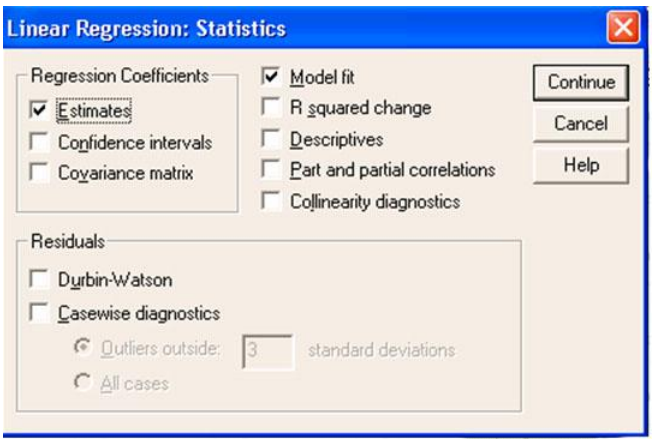
براي دسترسي به اين دستور مسير زير را طي كنيد:

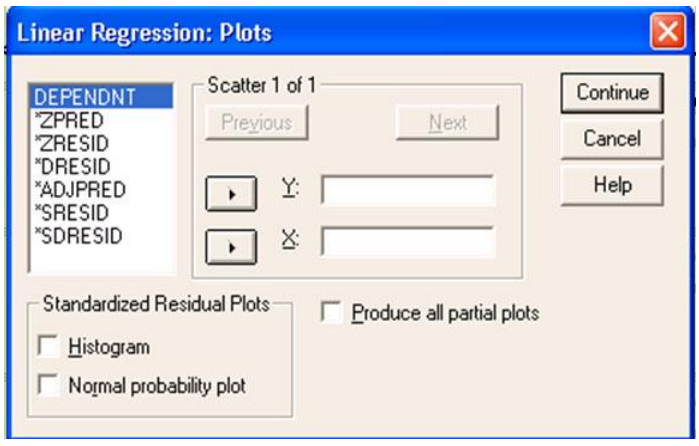
Analyze - Regression - Linear...



در صورتي كه به عنوان متغير مستقل چندين متغير را وارد كنيد spss محاسبات مربوط به رگرسيون چند متغيره را محاسبه خواهد كرد.

- در قسمت dependent متغير وابسته را وارد كنيد.
- در قسمت independent متغير يا متغيرهاي مستقل را وارد كنيد.
- در قسمت selection variable متغير و مقداري از آن كه براي بررسي مورد نظر شما است وارد كنيد. (اين متغير نبايد در رگرسيون به عنوان متغير مستقل به كار رفته باشد.)
- در قسمت Case lableمتغيري كه به عنوان اسامي موردها به كار رفته است مشخص كنيد.
- در قسمت WLS Wieght متغير وزن دهي به موردها را وارد كنيد. (اين وزن براي يكسان سازي مقدار واريانس به كار خواهد رفت.)
- با انتخاب دكمه Statistic امكان نمايش مقادير آماره‌هاي ديگري در خروجي وجود دارد.
- با انتخاب دكمه Plotامكان نمايش نمودارهاي بررسي معادله رگرسيون وجود دارد.
- با دكمه Save امكان ذخيره سازي مقادير برآورده شده از متغير وابسته و باقي مانده (استاندارد-واقعي) وجود دارد





H_0 : مدل مناسب نیست وتاثیري درتغییرات Y نسبت به X ندارد.

H_1 : مدل رگرسیون مناسب است.

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

نتیجه :

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

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

- جدول آنالیز واریانس مقدار خطای رگرسیون را حساب می کند.
- برای هر یک از ضرائب 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 martial را جدا می کنیم بعد مراحل محاسبه خط رگرسیون را تکرار می کنیم.

Coefficients[a]

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؟

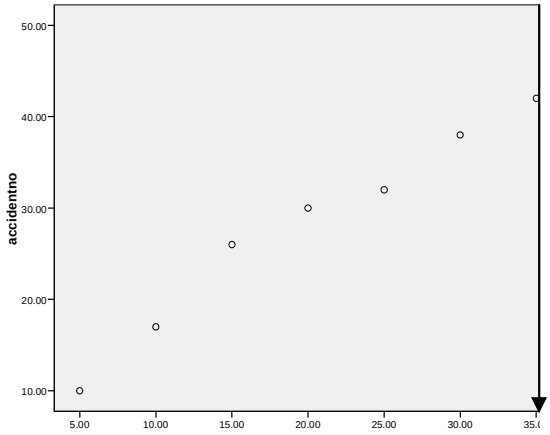
$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_k X_k + \varepsilon$$

بزرگی یا کوچکی ضرایب نشانه 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.آیامشاهدات تصادفی هستند؟

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

جواب۱.



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

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

با $p-value = 0.000$ فرض صفر یعنی استقلال دو متغیر توسط نمونه رد می شود

$\rho = 0.984$ که نشانه همبستگی بالاست.

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	B	Std. Error
1	(Constant)	7.286	1.863		3.912	.011
	alckol	1.029	.083	.984	12.348	.000

a Dependent Variable: accidentno

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

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

- در جدول آنالیز واریانس برای هر یک از ضرائب 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 ☑

جواب۵.

با استفاده از آزمون Runs :

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

$p - value = 0.704 > 0.05$

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

Standard Residual $\sim N(0,1)$
Unstandard Residual $\sim N(0,\sigma^2)$