تحلیل آماری با SPSS

(قسمت اول)

کارشناسی ارشد روانشناسی و علوم تربیتی

دکتر احمد زنده دل

پائیز 1391

کلیات و تعاریف

آمار علمی است که در هر دو جنبه نظری و کاربردی از اهمیت زیادی برخوردار بوده و به خصوص طی دهه‌های اخیر توسعه زیادی پیدا کرده است. آمار را می‌توان یکی از علوم واسطه یا یکی از رشته‌های میان رشته دانست به طوری که یکی از ابزارهای عمده شناخت در بسیاری از علوم دیگر است. علم آمار نیز مانند علوم دیگر در نتیجه نیازهای بشر به وجود آمده است و تاریخی طولانی داشته و از دوران‌های گذشته تا کنون رشد و توسعه آن ادامه یافته است. پیدایش این علم را می‌توان مقارن با بدو تشکیل دولت‌های اولیه دانست. دولت‌های اولیه نیاز به آگاهی از جمعیت تحت سلطه، میزان دارائی و ثروت قلمرو حکومت خود و نیز نیروهای نظامی تحت امر خود داشته‌اند و بنابراین به اموری مانند سرشماری، اندازه‌گیری و تحلیل توصیفی اطلاعات اقدام می‌کرده‌اند. در آن زمان منظور از آمار، ارقام و اطلاعات مورد نیاز دولت‌ها جهت گرفتن مالیات، سربازی و سایر امور مربوط به کشورداری و سیاست بوده است اما همین اندازه‌گیری‌ها و شمارش‌های ابتدایی پایه و اساس آمار امروزی را بنیان نهاده است. در قرن گذشته، علم آمار همراه با سایر علوم سیر صعودی را پیموده و در مواردی پیشتاز بسیاری از علوم دیگر بوده است. اغلب علوم از قبیل فیزیک، زیست‌شناسی، اقتصاد، مدیریت، علوم بهداشتی و نیز علوم اجتماعی با استفاده از آمار توانسته‌اند سرعت پیشرفت و توسعه خود را چند برابر کنند. چون روش‌ها و فنونی که برای تحقیقات علمی ضروری هستند از علم آمار بدست می‌آید. امروزه علم آمار در اکثر شاخه‌های علوم مانند، کشاورزی، اقتصاد و بازرگانی، روانشناسی، پزشکی و... کاربردهای گوناگونی داشته و کمتر رشته‌ای را می‌توان یافت که بی‌نیاز از بکارگیری روش‌ها و فنون آماری باشد.

واژه آمار معادل لاتین statistics است که این کلمه نیز از status به معنی دولت گرفته شده است.

از آمار یک تعریف یکتا و مشخصی ارائه نشده است و از آنجایی که در اکثر شاخه‌های علوم از آمار استفاده شده است، لذا در هر شاخه‌ای از علم بنابر استفاده‌ای که از آمار شده است برای آن تعریف یا تعبیری ارائه کرده‌اند که تعداد این تعاریف از ده‌ها

مورد متجاوز بوده و هیچ کدام از آن‌ها نمی‌توانند جامعیت این علم وسیع را بپوشاند. برخی از تعاریف عمومی‌تری که برای آمار ارائه شده است به شرح زیر است.

- *آمار علمی است که خواص جامعه را مورد مطالعه قرار می‌دهد.*

- *آمار علمی است که مشخصات جوامع را به صورت کمّی ولی با در نظر گرفتن اوضاع کیفی آن‌ها مورد بررسی قرار می‌دهد.*

- *آمار علمی است که جنبه‌های کمّی نمودهای اجتماعی را در ارتباط با کیفیت آن‌ها، با هم مطالعه می‌کند.*

- *آمار عبارت است از مجموعه‌ای از روش‌ها برای طرح آزمایش‌ها، بدست آوردن داده‌ها و سپس تجزیه و تحلیل، تعبیر و استخراج نتایج آن‌ها.*

در یک جمع‌بندی کلی آمار را می‌توان به دو معنی زیر به کار برد:

الف) به معنی اعداد و ارقام واقعی یا تقریبی در خصوص مسأله یا فعالیتی مانند اموری از قبیل زاد و مرگ، میزان محصولات کشاورزی، میزان فروش کالایی خاص و....

ب) به معنی تکنیک‌ها و روش‌هایی جهت جمع‌آوری، تنظیم، تلخیص و تجزیه و تحلیل اطلاعات عددی دربارهٔ موضوعات مختلف مانند بررسی قیمت کالایی خاص در گذشته، حال و آینده و رابطه آن با برخی ویژگی‌های دیگر، مطالعه اثر یک دارو بر روی گروهی بیمار، مطالعه اثر یک کود بر میزان محصولات کشاورزی و....

در یک تقسیم‌بندی کلی، آمار را می‌توان به دو بخش آمار توصیفی و آمار استنباطی (یا آمار تحلیلی) تقسیم کرد.

آمار توصیفی

آمار توصیفی دارای پیشینه بسیار دراز بوده و موضوع آن کمّی کردن ویژگی یا ویژگی‌های یک جامعه و سپس استخراج اطلاعات مورد نیاز از بین این کمیت‌ها است. آمار توصیفی جزء کوچکی از علم آمار امروزی را تشکیل داده اما عمر آن بسیار طولانی بوده و هنوز نیز دارای کاربردهای زیادی است. موضوع آمار توصیفی مطالعه کمّی جوامع است. بدین منظور ابتدا ویژگی

یا ویژگی‌های یک جامعهٔ آماری کمّی شده و سپس اطلاعات مورد نیاز از بین این کمیت‌ها استخراج می‌شود. کمیت‌های مورد بحث را داده‌ها یا داده‌های خام گویند چون این داده‌ها محتوی اطلاعاتی درباره جامعهٔ مورد نظر هستند که این اطلاعات به صورت یک پتانسیل در درون داده‌ها نهفته بوده و لذا به راحتی قابل استفاده نیست. برای استفاده از اطلاعات این داده‌ها باید بتوان آنها را به نحو مناسبی پالایش کرد. مراحلی که در آمار توصیفی برای پالایش داده‌ها و استخراج اطلاعات آنها، بر روی آنها انجام می‌گیرد به شرح زیر است:

الف- (جداول آماری) داده‌ها را در جدول یا جداولی تنظیم می‌کنند. چون این جداول اکثراً برپایه فراوانی‌ها تنظیم می‌شود لذا آنها را جدول فراوانی یا جدول توزیع فراوانی نیز می‌نامند. این جداول اطلاعات نهفته در داده‌ها را به صورت خلاصه شده‌ای در اختیار می‌گذارند.

ب- (نمودارهای آماری) از روی جداول بند الف نمودارهایی رسم می‌کنند که آنها را نمودارهای آماری می‌نامند. اکثر نمودارهای آماری چون از روی جداول آماری رسم می‌شوند لذا همان اطلاعات جداول را در اختیار می‌گذارند که چون این اطلاعات را به صورت تصویری در اختیار می‌گذارند لذا استفاده از آنها راحت‌تر است. بنابراین نمودارهای آماری را می‌توان بیان هندسی جداول آماری دانست.

ج- (تلخیص داده‌ها) داده‌ها را در یک یا چند عدد خلاصه می‌کنند به طوری که این اعداد اطلاعاتی دربارهٔ کل داده‌ها را شامل است. این بخش را تلخیص داده‌ها می‌نامند. اعدادی که اطلاعاتی دربارهٔ تمرکز داده‌ها را در اختیار می‌گذارند شاخص‌های تمرکز یا معیارهای مرکزی گفته، اعدادی که اطلاعاتی دربارهٔ پراکندگی داده‌ها در اختیار می‌گذارند را شاخص‌های پراکندگی و اعدادی که اطلاعاتی دربارهٔ چگونگی توزیع داده‌ها را در اختیار می‌گذارند شاخص‌های توزیع می‌نامند.

بنابراین موضوع آمار توصیفی خود به سه قسمت اصلی جداول آماری، نمودارهای آماری و تلخیص داده‌ها تقسیم می‌شود.

آمار استنباطی

آمار استنباطی که آن را آمار تحلیلی نیز می‌نامند، برخلاف آمار توصیفی عمری کوتاه داشته و موضوع اصلی علم آمار امروزی را تشکیل می‌دهد. آمار استنباطی بر پایه تئوری احتمال بنا شده و موضوع آن چگونگی تعمیم نتایج حاصل از یک نمونه به جامعه است به طوری که خطای حاصل از این تعمیم می‌نیمم شود. به عبارت دیگر در آمار استنباطی اطلاعات از نمونه به دست آمده و به جامعه تعمیم داده می‌شود لذا همواره میزانی خطا در این تعمیم دادن وجود دارد. موضوع آمار استنباطی کم کردن (به حداقل رساندن) این خطا و افزایش اطمینان در تعمیم نتایج حاصل از نمونه به جامعه است. همچنین تعیین مدل‌ها و الگوهای ریاضی برای فعالیت‌های مختلف طبیعی و انسانی قسمت دیگری از موضوع آمار استنباطی است. آمار استنباطی خود به دو شاخهٔ آمار پارامتری و آمار ناپارامتری تقسیم می‌شود. هرگاه توزیع جامعه نرمال باشد، مطالعه آن در حوزه آمار پارامتری است و در غیر این صورت مطالعه آن در حوزه آمار ناپارامتری که آن را آمار آزاد توزیع نیز می‌نامند، می‌باشد.

جامعه آماری (جمعیت)

مجموعه‌ای از افراد یا اشیایی که قرار است یک یا چند خصوصیت آن‌ها مورد مطالعه قرار گیرد را جامعه آماری می‌گوییم. به عبارت دیگر جامعهٔ آماری مجموعه‌ای از افراد یا اشیایی است که حداقل در یک خاصیت با هم مشترک باشند.

واحد آماری (فرد آماری)

هر عضوی از جامعهٔ آماری را یک واحد آماری یا یک فرد آماری می‌گوییم.

صفت مشخصه (صفت ثابت)

آن خصوصیتی که در بین همهٔ افراد یک جامعه آماری مشترک است و در واقع وجه تمایز افراد این جامعه با سایر جوامع است را صفت مشخصه یا صفت ثابت می‌گوییم.

صفت آماری (صفت متغیر)

آن خصوصیتی که از هر فرد آماری مورد نظر است و قرار است آن ویژگی مورد مطالعه قرار گیرد را صفت آماری یا صفت متغیر می‌نامیم. در واقع صفت آماری یا صفت متغیر، از هر فردی به فرد دیگر تغییر می‌کند. صفات متغیر بر دو نوع‌اند، صفت کمّی و صفت کیفی.

صفت کمّی

صفاتی که به طور مستقیم قابل سنجش و اندازه‌گیری باشند را صفت کمّی گوییم. مانند سن افراد، اندازه قد، درآمد، تعداد فرزندان خانوار و مانند این‌ها.

صفت کیفی

صفتی که به طور مستقیم قابل سنجش و اندازه‌گیری نباشند را صفت کیفی گوییم. مانند جنسیت، نوع شغل، رنگ چشم، مزه و مانند این‌ها.

مثال 1:

فرض کنید در یک شهرستان بخواهیم متوسط درآمد ماهانه خانوارهای آن شهر را به دست آوریم. یعنی می‌خواهیم مطالعه‌ای بر روی درآمد ماهانه خانوارهای آن شهر انجام دهیم. در این صورت مجموعه تمام خانوارهای آن شهرستان جامعهٔ آماری را تشکیل می‌دهد. هر خانواری در آن شهرستان یک واحد آماری یا یک فرد آماری می‌باشد. خانوار بودن و ساکن آن شهرستان بودن صفت مشخصه را تشکیل می‌دهد. چون می‌خواهیم درآمد ماهانه را مورد مطالعه قرار دهیم لذا درآمد ماهانه هر خانوار صفت آماری یا صفت متغیر را تشکیل می‌دهد که این صفت یک صفت کمّی است. اما در همین مثال اگر بخواهیم نظر خانوارهای این شهرستان را راجع به یک رویداد اجتماعی (مثلاً چگونگی شرکت آن‌ها در انتخابات شورای شهر) جویا شویم در این صورت صفت متغیر، نظر خانواده در مورد موضوع مورد بحث بوده که این صفت کیفی است و معمولاً آن را به صورت گزینه‌های جدا از هم مانند، شرکت نمی‌کنم، شاید شرکت کنم، حتماً شرکت خواهم کرد، اندازه‌گیری می‌کنند.

مقیاس سازی

از آنجایی که تحلیل‌های آماری بر روی اعداد و ارقام انجام می‌گیرد لذا هر صفت آماری را باید بتوان به نحو مناسبی به عدد تبدیل کرد. مثلاً در یک تحقیق اجتماعی بر روی گروهی از دانشجویان فرض کنید بخواهیم، وزن، قد، معدل، رنگ چشم، شهرستان محل سکونت و نظر آن‌ها را راجع به یک مسأله اجتماعی مورد مطالعه قرار دهیم. در این صورت وزن، قد و معدل هر دانشجو چون صفات کمّی هستند، به راحتی به عدد تبدیل می‌شوند. معدل دانشجو به صورت یک نمره ثبت شده موجود است. برای اندازه‌گیری قد و وزن هر دانشجو نیز وسیله اندازه‌گیری مناسب و نیز واحد اندازه‌گیری خاصی وجود دارد که توسط آن‌ها می‌توان قد و وزن را نیز کمّی کرد. اما وسیله اندازه‌گیری خاص و نیز واحد اندازه‌گیری مناسبی برای اندازه‌گیری و به عدد تبدیل کردن رنگ چشم، شهرستان محل سکونت و نظر آن‌ها راجع به مسأله اجتماعی وجود ندارد. در تحلیل آماری باید بتوان این گونه صفات را نیز به نحو مناسبی مورد اندازه‌گیری قرار داد. یا مثلاً فرض کنید بخواهیم وزن و تُردی سیب‌های یک درخت را مورد مطالعه قرار دهیم. در این صورت برای هر سیب وزن آن را می‌توان توسط یکی از واحدهای اندازه‌گیری وزن اندازه گرفت و آن را به عدد تبدیل کرد اما برای اندازه‌گیری تُردی سیب برخلاف وزن آن، وسیله اندازه‌گیری مناسب و واحد اندازه‌گیری خاصی وجود ندارد.

منظور از مقیاس‌سازی، اندازه‌گیری صفت مورد مطالعه و نسبت دادن اعداد بر طبق قواعد معین به اشیاء یا افراد می‌باشد. به عبارت دیگر، اطلاق یک عدد به یک فرد را طبق قاعدای مشخص، مقیاس‌سازی گفته و قاعده را یک مقیاس گوییم. مقیاس‌ها به طور کلی به چهار دسته تقسیم می‌شوند، مقیاس اسمی (Nominal Scale)، مقیاس ترتیبی یا رتبه‌ای (Ordinal Scale)، مقیاس فاصله‌ای (Interval Scale) و مقیاس نسبتی یا نسبی (Ratio Scale).

مقیاس اسمی (Ordinal Scale)

در اندازه‌گیری صفات توسط مقیاس اسمی، به هر فرد آماری یک عدد نسبت داده می‌شود که این اعداد صرفاً جهت شناسایی و جدا کردن آن‌ها می‌باشد. مثلاً به هر مرد عدد 1 و به هر زن عدد 2 را نسبت می‌دهیم. این اعداد جامعه را به دو طبقه تقسیم می‌کند. از این مقیاس عموماً در اندازه‌گیری صفات کیفی که فاقد شدت و ضعف بوده و شامل اسامی، نشان‌ها یا طبقات و گروه‌ها هستند استفاده می‌شود. مثلاً صفاتی مانند رنگ چشم، جنسیت، ملیت، گروه خونی و مانند این‌ها توسط مقیاس

اسمی اندازه‌گیری می‌شوند. اعداد حاصل از مقیاس اسمی صرفاً بیانگر یک اسم بوده و بر روی آن‌ها چهار عمل اصلی را نمی‌توان انجام داد. همچنین از این اعداد جهت مقایسه آن‌ها با یکدیگر نیز نمی‌توان استفاده کرد.

مثلاً در یک مطالعه بر روی شهرستان محل سکونت گروهی از دانشجویان، محل سکونت آن‌ها را توسط اعداد 1، 2،...، n کُد گذاری می‌کنیم. مثلاً فرد مشهدی را با کد 1، فرد اصفهانی را با کد 2، فرد تهرانی را با کد 3 و الی آخر تعیین می‌کنیم. در این صورت داده‌ها شامل مجموعه‌ای از اعداد 1، 2،...، n می‌باشد. اما این اعداد صرفاً بیانگر اسم یک شهرستان بوده و چهار عمل اصلی را نمی‌توان روی آن‌ها انجام داد. عبارت۱+۲ در این داده‌ها کاملاً بی‌معنی است. همچنین از این اعداد برای مقایسه با یکدیگر نیز نمی‌توان استفاده کرد. در این جا فردی که اندازه صفت او 2 می‌باشد به معنی بیشتر بودن مقدار صفتش از فردی که اندازه صفت او 1 می‌باشد نیست.

مقیاس ترتیبی (رتبه‌ای) (Ordinal Scale)

در اندازه‌گیری صفات توسط مقیاس ترتیبی، به هر فرد آماری یک عدد نسبت داده می‌شود که این اعداد علاوه بر شناسایی افراد، قابل مقایسه با یکدیگر نیز هستند. یعنی از این اعداد برای مقایسه عناصر از نظر کوچکتر یا بزرگتر یا برابر بودن نیز می‌توان استفاده کرد. بنابراین این اعداد نیز افراد جامعه را طبقه‌بندی می‌کند به نحوی که این طبقات دارای ترتیب هستند. اما بر روی این اعداد نیز چهار عمل اصلی را نمی‌توان انجام داد.

از این مقیاس اکثراً در اندازه‌گیری صفات کیفی که دارای شدت و ضعف هستند استفاده می‌شود. صفاتی که توسط مقیاس ترتیبی اندازه‌گیری می‌شوند نیز مانند مقیاس اسمی شامل اسامی، نشان‌ها و طبقات می‌باشند، با این تفاوت که این صفات دارای شدت و ضعف هستند یعنی طبقات یا گروه‌ها رتبه‌بندی شده‌اند.

مثلاً میزان تحصیلات در افراد یک جامعه که به طبقات « عالی، متوسطه، سیکل، ششم ابتدایی، بی‌سواد» تقسیم‌بندی شده است، کیفیت یک محصول که به طبقات « عالی، خوب، متوسط، ضعیف» تقسیم‌بندی شده است، توسط مقیاس ترتیبی اندازه‌گیری می‌شوند. مثلاً فرض کنید کارگران یک کارخانه را بر حسب میزان مهارتشان به چهار گروه ضعیف، متوسط، خوب و عالی طبقه‌بندی کرده و هر یک از این گروه‌ها را به ترتیب با کدهای 1، 2، 3 و 4 اندازه‌گیری کنیم. یعنی اگر کارگری دارای مهارت عالی بود اندازه او 4 می‌باشد. در این صورت این اعداد قابل مقایسه با یکدیگر هستند چون کارگری که اندازه او 2 می‌باشد ماهرتر از کارگری است

که اندازه او 1 است، اما این به آن معنی نیست که مهارت او دو برابر مهارت کارگر دومی است. ضمناً در این داده‌ها ۱+۲ معنی نمی‌دهد.

مقیاس فاصله‌ای (Distance Scale)

در اندازه‌گیری صفات توسط مقیاس فاصله‌ای، به هر فرد آماری یک عدد نسبت داده می‌شود که این اعداد علاوه بر شناسایی و مقایسه آن‌ها از نظر ترتیب (بیشتر یا کمتر بودن)، قابل مقایسه بر حسب طول فاصله‌ها نیز هستند (چقدر بیشتر یا چقدر کمتر). مقیاس فاصله‌ای از مفهوم واحد فاصله استفاده می‌کند و بدین جهت می‌تواند فاصله بین دو عدد را نیز اندازه‌گیری کند. در مقیاس ترتیبی اگر اندازه فرد A، 2 و اندازه فرد B، 1 بود در این صورت فقط می‌توانستیم بگوییم که اندازه صفت فرد A بیشتر از اندازه صفت فرد B است. اما در همین مثال، اگر اعداد از مقیاس فاصله‌ای باشند در این صورت می‌توانیم بگوییم اندازه صفت فرد A به اندازه یک واحد بیشتر از فرد B است. در داده‌های فاصله‌ای علاوه بر انجام مقایسه (مقایسه به صورت ترتیبی و فاصله‌ای)، عمل جمع و تفریق نیز می‌توان انجام داد. از این مقیاس اکثراً در اندازه‌گیری صفات کمی استفاده می‌شود و اندازه‌گیری نیز به نحوی است که در آن عدد صفر جنبه مطلق نداشته بلکه قراردادی است. به عبارت دیگر در مقیاس فاصله‌ای، نقطه‌ای را به عنوان مبداء قرارداد کرده و آن را با صفر نشان می‌دهند و سپس توسط یک واحد اندازه‌گیری خاص و به فواصل مساوی در طرفین مبداء، صفت آماری را اندازه‌گیری می‌کنند.

بهترین مثالی که در این خصوص می‌توان زد اندازه‌گیری میزان دما توسط درجه سانتی‌گراد است. در اندازه‌گیری دما توسط سانتی‌گراد، قرارداد شده است که دمایی که در آن آب شروع به یخ زدن می‌کند را به عنوان مبداء در نظر گرفته و با صفر نشان می‌دهند و دمایی که در آن آب شروع به جوشیدن می‌کند را با ۱۰۰ نشان می‌دهند. سپس این فاصله را به ۱۰۰ قسمت مساوی تقسیم کرده و هر قسمت را یک درجه سانتی‌گراد گویند. لذا عدد صفر جنبه مطلق نداشته بلکه قراردادی است. به عبارت دیگر، اگر دمای جسمی ۰ درجه سانتی‌گراد باشد، به این معنی نیست که این جسم هیچ‌گونه دمایی ندارد.

به طور خلاصه، از مقیاس فاصله‌ای در اندازه‌گیری صفات کمی استفاده می‌شود و نحوه اندازه‌گیری به گونه‌ای است که:

الف) در این داده‌ها صفر جنبهٔ مطلق نداشته بلکه قراردادی است.

ب) این داده‌ها قابل مقایسه و جمع جبری با یکدیگر بوده اما نسبت در این داده‌ها بی‌معنی است. مثلاً اگر اندازه یک فرد 4 و فرد دیگر 2 باشد در این صورت ۲+۴ و ۲−۴ معنی‌دار است اما عبارت $\frac{۴}{۲} = ۲$ بی‌معنی است. یعنی نمی‌توان گفت فردی که دارای اندازه صفت 4 است میزان صفتش دو برابر فردی است که اندازه صفت او 2 است.

ج) در این داده‌ها نسبت فواصل با تغییر واحد اندازه‌گیری تغییر نمی‌کند.

مثال 2

فرض کنید درجه حرارت چهار جسم خاص بر حسب سانتی‌گراد عبارت باشد از ۰، ۱۰، ۱۵ و ۳۰. درجه حرارت همین چهار جسم بر حسب فارانهایت و طبق رابطهٔ $y = \frac{۹}{۵}x + ۳۲$ به ترتیب برابر است با ۳۲، ۵۰، ۵۹ و ۸۶. حال موارد زیر را در نظر می‌گیریم:

الف) جسمی که درجه حرارت آن بر حسب سانتی‌گراد ۰ می‌باشد، به این معنی نیست که فاقد حرارت است چون درجه حرارت همین جسم بر حسب فارنهایت 32 می‌باشد. پس صفر در این جا صرفاً قراردادی است.

ب) اگر دو درجه حرارت ۳۰ و ۱۰ درجه سانتی‌گراد را با یکدیگر جمع یا تفریق کنیم درجه حرارت به ترتیب ۴۰ و یا ۲۰ سانتی‌گراد به دست می‌آید که این اعداد معنی‌دار هستند.

ج) نسبت درجه حرارت جسمی که بر حسب سانتی‌گراد حرارتش ۳۰ می‌باشد به جسمی که دارای ۱۰ درجه سانتی‌گراد است $\frac{۳۰}{۱۰} = ۳$ است، اما این به آن معنی نیست که حرارت این جسم 3 برابر جسم دیگر است. چون نسبت درجه حرارت همین دو جسم بر حسب فارانهایت $\frac{۸۶}{۵۰} = ۱/۷۲$ می‌باشد. اگر دمای جسم دوم، واقعاً سه برابر دمای جسم اول بود، این دما را با هر واحدی که اندازه‌گیری کنیم، نسبت 3 باید ثابت باشد در صورتی که دیدیم چنین نیست. پس اعداد 3 و 1/72 در این مورد بی‌معنی هستند یعنی بر روی این داده‌ها نسبت تعریف نمی‌شود.

د) با کمی دقت مشاهده می‌شود که نسبت فواصل هر دو عدد در هر دو واحد اندازه‌گیری ثابت است مثلاً $\frac{۱۰-۰}{۱۵-۱۰} = \frac{۵۰-۳۲}{۵۹-۵۰}$.

توجه 1

توجه کنید که اگر درجه حرارت را بر حسب واحد کلوین که دارای صفر مطلق است اندازه‌گیری کنیم آنگاه مقیاس اندازه‌گیری فاصله‌ای نبوده بلکه مقیاس نسبتی است که در قسمت بعد بحث می‌شود.

مقیاس نسبتی (مقیاس نسبی) (Ratio Scale)

در اندازه‌گیری صفات توسط مقیاس نسبتی که آن را مقیاس نسبی نیز می‌نامند، به هر فرد آماری یک عدد نسبت داده می‌شود که این اعداد علاوه بر شناسایی و مقایسه آن‌ها از نظر ترتیب (بیشتر یا کمتر یا برابر بودن) و نیز قابل مقایسه بر حسب طول فاصله‌ها (چقدر بیشتر یا چقدر کمتر)، قابل مقایسه از نظر نسبت (چند برابر بزرگتر یا کوچکتر) نیز هستند.

در مقیاس ترتیبی اگر اندازه فرد A، 2 و اندازه فرد B، 1 بود، در این صورت فقط می‌توانستیم بگوییم که اندازه صفت فرد A بیشتر از اندازه صفت فرد B است. اما اگر اعداد از مقیاس فاصله‌ای باشند در این صورت می‌توانیم بگوییم اندازه صفت فرد A به اندازه یک واحد بیشتر از فرد B است. حال اگر اعداد از مقیاس نسبتی باشند علاوه بر این که می‌توان گفت فرد A به اندازه یک واحد بیشتر از فرد B است، می‌توان گفت که اندازه صفت A، 2 برابر اندازه صفت فرد B است، یعنی روی این داده‌ها نسبت نیز تعریف می‌شود.

در داده‌های نسبتی علاوه بر انجام مقایسه (مقایسه به صورت ترتیبی و فاصله‌ای) و نیز عمل جمع و تفریق، اعمال ضرب و تقسیم نیز تعریف می‌شود.

این مقیاس که عالی‌ترین نوع مقیاس می‌باشد در اندازه‌گیری صفات کمّی استفاده می‌شود و اندازه‌گیری نیز به نحوی است که در آن عدد صفر جنبه مطلق دارد. صفاتی مانند وزن، طول، سطح، حجم و مانند این‌ها. در این مقیاس عدد صفر به معنی نقطه شروع بوده و جنبه مطلق دارد. مثلاً اگر طول جسم A، ۱۰۰ سانتی‌متر و طول جسم B، ۵۰ سانتی‌متر باشد آن گاه طول جسم حاصل از ردیف کردن A و B، ۱۵۰ سانتی‌متر می‌باشد. همچنین می‌توان ادعا کرد که طول جسم A، 2 برابر طول جسم

B است. در صورتی که اگر درجه حرارت جسم A، ۱۰۰ درجه سانتی‌گراد و درجه حرارت جسم B، ۵۰ درجه سانتی‌گراد باشد نمی‌توان ادعا کرد که درجه حرارت جسم A، دو برابر درجه حرارت جسم B است.

توجه 2

نوع مقیاس اندازه‌گیری تنها به صفت آماری بستگی نداشته بلکه به چگونگی اندازه‌گیری نیز بستگی دارد. مثلاً فرض کنید بخواهیم اندازه قد دانشجویان یک کلاس را اندازه‌گیری کنیم. اگر دانشجویان یک کلاس را از لحاظ اندازه قد به گروه‌های، کوتاه قد، قد متوسط و بلند قد تقسیم کنیم، از مقیاس ترتیبی استفاده کرده‌ایم. اگر اندازه قد دانشجویان را بر حسب اندازه قد یکی از دانشجویان اندازه‌گیری کنیم، مثلاً افرادی که هم قد او هستند، اندازه قد آن‌ها صفر، افرادی که قدشان بلندتر از اوست بر حسب میزان بلندی بر حسب سانتی‌متر عدد بدهیم، مثلاً 2 سانتی‌متر، 1 سانتی‌متر و... و افرادی که قدشان کوتاه‌تر از اوست بر حسب میزان کوتاهی بر حسب سانتی‌متر عدد بدهیم مثلاً 1- سانتی‌متر، 2- سانتی‌متر و غیره. در این صورت از مقیاس فاصله‌ای استفاده کرده‌ایم. اما اگر اندازه قد افراد را به صورت متریک و با یکی از واحدهای اندازه‌گیری مانند سانتی‌متر اندازه‌گیری کنیم (چیزی که مرسوم است) در این صورت از مقیاس نسبتی استفاده کرده‌ایم. به عنوان مثال دیگر می توان از اندازه گیری دما بر حسب درجه سانتی گراد و درجه کلوین نام برد که در اندازه گیری دما بر حسب سانتی گراد همانطوریکه بحث شد مقیاس اندازه گیری فاصله ای اما در اندازه گیری توسط کلوین مقیاس نسبتی است. و نیز اگر دما را به صورت سرد، معتدل، گرم اندازه گیری کنیم در اینصورت مقیاس ترتیبی است.

توجه 3

یک رابطه ترتیبی بین مقیاسهای اندازه گیری برقرار است. بدین ترتیب که مقیاس اسمی پست ترین مقیاس، مقیاس ترتیبی همان مقیاس اسمی است با یک خاصیت اضافه تر، مقیاس فاصله ای همان مقیاس ترتیبی است با یک خاصیت اضافه تر و خلاصه مقیاس نسبتی که عالی ترین مقیاس اندازه گیری است همان مقیاس فاصله ای است با یک خاصیت اضافه تر.

توجه 4: مقیاس جمع پذیر یا مقیاس لیکرت (Summative Scale)

نوع دیگری از مقیاس اندازه گیری که با نام مقیاس جمع پذیر و اکثرا با نام مقیاس لیکرت از آن یاد می شود سطحی از اندازه گیری است که در آن یک صفت به صورت ترتیبی اندازه گیری می شود با این تفاوت که فاصله بین گزینه های مختلف یکسان در نظر گرفته می شود. به عبارت دیگر در سطح اندازه گیری ترتیبی یک صفت به نحوی به عدد تبدیل می شود که این اعداد علاوه بر شناسائی افراد دارای یک رابطه ترتیبی نیز هستند اما فاصله بین این مقادیر لزوما یکسان نیست. مثلا اگر کیفیت یک محصول را به صورت " ضعیف، متوسط، خوب، عالی" اندازه گیری کنیم از سطح اندازه گیری ترتیبی استفاده کرده ایم. در این حالت اختلاف کیفیت دو محصول که به صورت " ضعیف" و " متوسط" اندازه گیری شده است لزوما برابر با اختلاف کیفیت دو محصول که به صورت " متوسط" و " خوب" اندازه گیری شده است نیست. اما اگر این اختلافها یکسان باشد آنگاه می توان این سطح از اندازه گیری را جمع پذیر یا لیکرت دانست. بر این اساس مقیاس لیکرت در واقع بین دو مقیاس ترتیبی و فاصله ای قرار می گیرد. از این مقیاس اکثرا در اندازه گیری صفات کیفی که مقادیر آن دارای شدت و ضعف است و یا صفاتی که ذاتا کمی بوده اما قادر به اندازه گیری کمی آنها نیستیم استفاده می شود و بخصوص از این مقیاس در طراحی پرسشنامه ها و گزینه های سئوالات یک پرسشنامه استفاده می شود. مثلا فرض کنید بخواهید توسط یک سئوال در پرسشنامه میزان رضایت پاسخگو از کیفیت یک محصول را اندازه گیری کنید. و فرض کنید سئوال را به صورت " شما تا چه حد از کیفیت این محصول رضایت دارید" قرار داده باشید. در اینصورت اگر پاسخ این سئوال را به صورت گزینه های "خیلی کم"، "کم"، "متوسط"، "زیاد"، "خیلی زیاد" قرار داده باشید از طیف اندازه گیری لیکرت استفاده کرده اید. در طیف لیکرت اکثرا پاسخهای سئوالات پرسشنامه ها را در پنج و یا هفت گزینه قرار می دهند. مثلا گزینه هائی مانند "کاملا موافقم"، "موافقم"، "نظری ندارم"، "مخالفم"، "کاملا مخالفم" قرار می دهند. در طیف لیکرت به گزینه ها به ترتیب کدهای 1 الی 5 اختصاص می دهند. تفاوت اصلی بین طیف لیکرت با طیف ترتیبی در آن است که می توان متغیرهای لیکرت را با یکدیگر جمع کرد و به همین دلیل آن را جمع پذیر نیز می نامند. در طراحی پرسشنامه ها معمولا برای اندازه گیری برخی صفات که دارای جنبه های مختلف است تعدادی سئوال در پرسشنامه قرار داده و پاسخها را در طیف لیکرت قرار می دهند. سپس از جمع بستن پاسخهای داده شده به این سئوالات یک اندازه کمی برای آن صفت به دست می آید. حاصل جمع چند متغیر لیکرت را می توان یک متغیر در سطح اندازه گیری فاصله ای دانست و به همین دلیل از روشهای آماری خیلی بیشتری می توان استفاده کرد. مثلا فرض کنید توسط سئوالات یک

پرسشنامه بخواهیم میزان رضایت شغلی را اندازه گیری کنیم. فرض کنید رضایت شغلی دارای ابعاد رضایت از حقوق، رضایت از همکار، رضایت از سرپرست و . . . باشد. در اینصورت برای هرکدام از این ابعاد یک یا بیش از یک سئوال با پاسخهائی در طیف لیکرت در پرسشنامه قرار می دهند. حال می توان از جمع بستن پاسخهای داده شده به این سئوالات یک اندازه کمی از میزان رضایت شغلی پاسخگو به دست آورد.

داده‌ها

مجموعه اعداد به دست آمده از اندازه‌گیری یک صفت آماری با یک مقیاس مناسب را داده‌ها یا data گوییم. داده‌ها بر دو نوع‌اند، داده‌های گسسته و داده‌های پیوسته.

داده‌های گسسته

داده‌های حاصل از اندازه‌گیری با مقیاس‌های اسمی و ترتیبی و نیز داده‌های شمارشی را داده‌های گسسته گوییم. به بیان دیگر، داده‌های گسسته، داده‌هایی هستند که در آن‌ها بین هر دو دادهٔ متوالی دادهٔ دیگری از صفت مورد نظر را نتوان قرار داد. مثلاً اگر تعداد فرزندان خانوارهای یک محله را ثبت کرده باشیم، این داده‌ها گسسته هستند چون بین دو دادهٔ 2 و 3 که دو دادهٔ متوالی هستند، هیچ دادهٔ دیگری از تعداد فرزندان را نمی‌توان قرار داد. به عبارت دیگر هیچ خانواری را نمی‌توان یافت که تعداد فرزندان آن عددی بین 2 و 3 (مثلاً 2/5) باشد.

داده‌های پیوسته

داده‌هایی که از راه اندازه‌گیری با مقیاس‌های فاصله‌ای و نسبتی به دست می‌آیند را داده‌های پیوسته گوییم. به بیان دیگر داده‌های پیوسته، داده‌هایی هستند که در آن‌ها بین هر دو دادهٔ متوالی، دادهٔ دیگری از صفت مورد نظر را بتوان قرار داد. مثلاً فرض کنید اندازه قد دانشجویان یک دانشگاه را ثبت کنیم. در این صورت این داده‌ها از نوع پیوسته‌اند. چون فرض کنید اندازه قد دو نفر از دانشجویان 165 و 166 سانتی‌متر باشد. در این صورت می‌توان فرد دیگری را یافت که اندازه قد او بین 165 و 166 سانتی‌متر مثلاً 165/5 سانتی‌متر باشد. کافی است دقت وسیله اندازه‌گیری را افزایش داد. در این صورت هر قدر این دو

داده به هم نزدیک‌تر شوند، باز هم می‌توان فردی را یافت که اندازه قدش عددی بین آن دو عدد باشد. داده‌های حاصل از وزن و زمان نیز وقتی به صورت متریک اندازه‌گیری شوند از نوع پیوسته هستند.

متغیر

از آن جایی که صفت آماری از فردی به فرد دیگر تغییر می‌کند لذا آن را متغیر نیز می‌نامند. متغیرها را به طور کلی به دو دسته متغیر گروهی و متغیر عددی تقسیم می‌کنند.

متغیر گروهی

متغیری که با مقیاس اسمی یا ترتیبی سنجیده شده و بر اساس آن جمعیت گروه‌بندی می‌شود را متغیر گروهی می‌نامند، مانند گروه خونی، ملیت، نژاد و مانند این‌ها.

متغیر عددی

متغیری که از راه شمارش یا اندازه‌گیری با مقیاس‌های فاصله‌ای و نسبتی به دست می‌آید را متغیر عددی می‌نامند مانند تعداد فرزندان، وزن و مانند این‌ها.

کلیات آمار توصیفی

در یک تقسیم بندی کلی روشهای آماری به دو دسته توصیفی و استنباطی تقسیم می شوند. موضوع آمار توصیفی، کمّی کردن ویژگی یا ویژگیهای یک جامعه و سپس استخراج اطلاعات موردنیاز از روی آن کمیتها می باشد. کمیتهایی که آنها را داده ها یا داده های خام می نامند چون این داده ها محتوی اطلاعاتی درباره ویژگی مورد نظر هستند اما این اطلاعات بصورت یک پتانسیل در درون داده ها نهفته است. برای استفاده از این اطلاعات باید بتوان داده ها را به نحو مناسبی پالایش کرد. مراحلی که در آمار توصیفی برای پالایش داده ها و استخراج اطلاعات مورد نیاز بر روی آنها به کار می برند به شرح زیر است.

الف: داده ها را در جدول یا جداولی تنظیم میکنند(جداول آماری). این جداول اطلاعات نهفته در داده ها را به صورت دسته بندی شده در اختیار می گذارند و معمولا دارای ستونهای فراوانی، فراوانی نسبی، درصد، فراوانی تجمعی و ستونهای دیگری که بنا به نیاز به این جداول اضافه می شود.

ب: از روی جداول آماری نمودارهائی رسم می کنند (نمودارهای آماری). این نمودارها همان اطلاعات جداول را منتها به صورت تصویری در اختیار می گذارند و لذا استفاده از آنها راحتتر است. نمودارهای آماری خیلی زیاد و متنوع بوده و بنا بر نوع داده ها و اطلاعات مورد نیاز این نمودارها متفاوت است.

ج: داده ها را در یک یا چند عدد خلاصه می کنند(تلخیص داده ها). بطوریکه این اعداد یک اطلاعات کلی در مورد همه داده ها را شامل باشد. اعدادی که اطلاعاتی در مورد تمرکز داده ها را شامل است شاخصهای تمرکز، اعدادی که اطلاعاتی در مورد پراکندگی داده ها را شامل است شاخصهای پراکندگی و اعدادی که اطلاعاتی در مورد چگونگی توزیع داده ها را شامل است شاخصهای توزیع می نامند.

شاخصهای تمرکز:

شاخصهای تمرکز اعدادی هستند که اطلاعاتی در باره تمرکز داده ها در اختیار می گذارند و عمده ترین آنها شامل میانگین (حسابی، وزنی، هندسی، هارمونیک و . . .)، میانه و نما (مد) می باشد. میانگین عمده ترین و پرکاربردترین این شاخصها بوده و دارای انواع مختلف است. میانگین در واقع مرکز ثقل داد ها است یعنی نقطه ای است که وزن داده ها در سمت چپ و راست آن با هم برابر است. یا به عبارت دیگر مجموع انحرافات داده ها در سمت چپ و راست میانگین با هم برابر است. میانه نقطه وسط داده ها است. یعنی داده ای که تعداد داده های سمت چپ و راست آن با هم برابر است. مد یا نما داده ای است که بیشترین فراوانی را دارد. از بین این شاخصها میانگین دارای ارزش بیشتری بوده و به همین جهت پر کاربرد تر از میانه و نما است. اما با این حال هنگامی که داده ها از یک صفت کیفی و در سطح اندازه گیری ترتیبی هستند از میانه به عنوان شاخص تمرکز استفاده می شود. مد یا نما ارزش خیلی کمتری نسبت به میانگین و میانه داشته و در نمونه های کوچک فاقد اعتبار است. اما با این حال هنگامی که داده ها از یک صفت کیفی و در سطح اندازه گیری ترتیبی هستند ناچارا از مد به عنوان شاخص تمرکز استفاده می شود. مثلا شاخص تمرکز مناسب برای قدف وزن، نمرات و ویژگی هائی از این قبیل میانگین است. شاخص تمرکز مناسب برای دیدگاه دانشجویان نسبت به کیفیت برگزاری کلاس درس آمار که به صورت رتبه ای خیلی کم، کم، متوسط، خوب، خیلی خوب اندازه گیری شده است، میانه و شاخص تمرکز مناسب برای گروه خونی دانشجویان مد است.

شاخصهای پراکندگی:

شاخصهای پراکندگی اعدادی هستند که اطلاعاتی در باره پراکندگی داده ها در اختیار می گذارند و عمده ترین آنها واریانس و انحراف معیار است. واریانس در واقع میانگین مجذور تفاضل داد ها از میانگینشان است و شاخص خوبی برای اندازه گیری پراکندگی است منتها ایراد واریانس در آن است که واحد اندازه گیری آن با واحد اندازه گیری داده ها یکسان نبوده بلکه مجذور یا مربع واحد اندازه گیری داده ها است. مثلا اگر داده ها همه بر حسب متر باشد آنگاه میانگین آنها نیز بر حسب متر خواهد بود حال آنکه واریانس آنها بر حسب متر مربع خواهد بود. به همین دلیل و به جهت یکسان سازی واحد اندازه گیری این شاخص با

واحد اندازه گیری داده ها از واریانس جذر گرفته و جذر آن را انحراف معیار یا انحراف استاندارد نامیدند که واحد اندازه گیری انحراف معیار با واحد اندازه گیری داده ها یکسان است.

شاخصهای توزیع:

شاخصهای توزیع اعدادی هستند که اطلاعاتی در باره توزیع داده ها در اختیار می گذارند و این شاخصها شامل چولگی و کشیدگی است. چولگی شاخصی است برای اندازه گیری میزان عدم تقارن یا کجی یک توزیع نسبت به توزیع نرمال. کشیدگی نیز شاخصی است برای اندازه گیری میزان پراکندگی نسبت به توزیع نرمال. راجع به چولگی و کشیدگی بعدا و در بخش SPSS به طور مفصل بحث خواهد شد.

به جهت اهمیت، نحوه محاسبه میانگین و واریانس با ذکر مثال در زیر خواهد آمد.

محاسبه میانگین و واریانس(جامعه) :

فرض کنید x_1، x_2، ...، x_N، داده‌های یک جامعهٔ N نفره باشد. میانگین و واریانس این جامعه که آنها را به ترتیب با μ و σ^2 نشان می‌دهیم به صورت زیر به دست می‌آید.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

حال اگر این داده ها یک نمونه از یک جامعه باشد فرمول محاسبه واریانس آن کمی متفاوت است. به عبارت دیگر فرض کنید x_1، x_2، ...، x_n یک نمونه تصادفی n تائی از یک جامعه باشد. در اینصورت میانگین و واریانس نمونه که آنها را به ترتیب با $\bar{x}$ و S^2 نشان می دهیم به صورت زیر تعریف می شود.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

مثال 3:

فرض کنید 2 ، 5 ، 7 ، 0 ، 6 مقادیر یک جامعه محدود پنج نفره باشد. میانگین و واریانس این جامعه را به دست آورید. واریانس را از هر دو فرمول آن به دست آورید.

حل: داریم:

$$N = 5\,, \quad \sum x = 20\,, \quad \mu = \frac{20}{5} = 4$$

واریانس (فرمول اول):

$$\sum (x-\mu)^2 = (2-4)^2 + (5-4)^2 + \ldots + (6-4)^2$$

$$= 4 + 1 + \ldots + 4 = 34$$

$$\sigma^2 = \frac{34}{5} = 6.8$$

$$\sigma = \sqrt{6.8} = 2.61$$

واریانس (فرمول دوم):

$$\sum x^2 = 2^2 + 5^2 + \ldots + 6^2 = 114$$

$$\sigma^2 = \frac{114}{5} - 4^2 = 6.8$$

مثال 4:

فرض کنید 2 ، 5 ، 7 ، 0 ، 6 مقادیر یک نمونه تصادفی از یک جامعه باشد. میانگین و واریانس این نمونه را به دست آورید.

واریانس نمونه را از هر دو فرمول آن به دست آورید.

حل: داریم:

$$n = 5\,, \quad \sum x = 20\,, \quad \bar{x} = \frac{20}{5} = 4$$

واریانس نمونه(فرمول اول):

$$\sum (x - \bar{x})^2 = (2-4)^2 + (5-4)^2 + \ldots + (6-4)^2$$

$$= 4 + 1 + \ldots + 4 = 34$$

$$S^2 = \frac{34}{5-1} = 8.5$$

$$\mathrm{S} = \sqrt{8.5} = 2.92$$

واریانس نمونه (فرمول دوم):

$$\sum x^2 = 2^2 + 5^2 + \ldots + 6^2 = 114$$

$$S^2 = \frac{114}{5-1} - \frac{5}{5-1}(4)^2 = 8.5$$

کلیات آمار استنباطی

موضوع آمار استنباطی که آنرا آمار تحلیلی نیز می نامند و قسمت عمده علم آمار امروزی را تشکیل می دهد مطالعهٔ یک جامعه از روی داده های یک نمونه تصادفی از آن جامعه است یعنی در آمار استنباطی فرض براین است که به دلایل مختلف امکان مشاهدهٔ جامعهٔ آماری نبود و بلکه تنها می توان یک نمونه تصادفی از آن جامعه را در دسترس داشت. لذا هر گونه استنباط و استنتاج دربارهٔ جامعه باید از طریق داده های آن نمونه انجام گیرد. واضح است که تعمیم نتایج حاصل از یک نمونه به کل جامعه توام با میزانی خطاست و لذا قسمت عمده ای از موضوع آمار استنباطی استفاده از روشهای ریاضی در به حداقل رساندن این خطا است و البته قسمت دیگری از موضوع آمار استنباطی که بخصوص طی سالهای اخیر بسیار مورد توجه واقع شده است بحث مدل سازی است یعنی استفاده از مدلها و الگوهای ریاضی در تعیین رابطهٔ بین متغیرهای انسانی و (طبیعی) است.

یعنی در آمار استنباطی فرض بر این است که به دلائل مختلف ،که این دلائل به خصوص در بحث آمار اقتصادی اجتماعی بحث هزینه و زمان است، جامعه سرشماری نمی شود بلکه از جامعه یک نمونه انتخاب می شود. منظور از سرشماری نیز مورد مطالعه قرار دادن کل افراد جامعه است. یعنی در یک مطالعه آماری هر گاه همه افراد جامعه مورد مطالعه قرار گیرند آن را سرشماری گویند. بنابراین در بحث آمار استنباطی اولین گام چگونگی نمونه گیری از یک جامعه است.واضح است که از یک جامعه به روشهای مختلف می توان نمونه گرفت اما در آمار این نمونه باید طوری باشد که بتواند به خوبی معرف جامعه اش باشد که چنین نمونه ای را نمونهٔ ناریب می نامند. نمونه تصادفی نمونه ای **ناریب** است بنابراین روش نمونه گیری در آمار باید بطور تصادفی باشد و منظور از نمونه تصادفی ، نمونه ای است که در آن هر فرد جامعه شانسی برای انتخاب در نمونه را داشته و این شانس برای همه افراد جامعه یکسان باشد. برای نمونه گیری تصادفی از یک جامعه بنابر چگونگی توزیع آن جامعه و موضوع تحت مطالعه ، روشهای مختلفی ارائه شده است که عمده ترین روشهای کلاسیک نمونه گیری شامل موارد زیر است:

1. نمونه گیری تصادفی ساده
2. نمونه گیری تصادفی منظم (سیستماتیک)
3. نمونه گیری تصادفی طبقه ای
4. نمونه گیری تصادفی خوشه ای

نمونه گیری تصادفی ساده:

نمونه گیری تصادفی ساده پایه و اساس همه روشهای نمونه گیری آماری است. در این روش نمونه به نحوی انتخاب می شود که همه افراد جامعه شانسی برای انتخاب در نمونه را داشته و این شانس برای همه افراد یکسان باشد. نمونه گیری تصادفی ساده را می توان به دو روش با جایگذاری و بدون جایگذاری انجام داد. در روش با جایگذاری هر فرد جامعه می تواند بیش از یکبار در نمونه انتخاب شود حال آنکه در روش بدون جایگذاری هر فرد جامعه فقط یکبار شانسی برای انتخاب در نمونه را دارد.

نمونه گیری تصادفی منظم (سیستماتیک):

از این روش هنگامی استفاده می شود که لیستی از افراد جامعه در دسترس باشد. در این روش ابتدا گامی از قبل تعیین میشود و گام یک عدد صحیح مثبت مانند k است که معمولا آن را از رابطه $k=N/n$ به دست می آورند که در آن N حجم جامعه و n حجم نمونه است. سپس از بین اعداد صحیح 1 تا k یک عدد به تصادف انتخاب می شود. فرد متناظر با عدد انتخاب شده اولین عضو نمونه است و سایر اعضای نمونه به تعداد گام مورد نظر به نمونه اضافه می شوند. بنابر این در این روش اولین عضو نمونه بطور تصادفی انتخاب شده و سایر اعضای نمونه وابسته به نفر اول به نمونه وارد می شوند. مثلا فرض کنید بخواهیم از یک جامعه $N=1000$ نفره یک نمونه $n=100$ تائی به روش سیستماتیک انتخاب کنیم و نیز فرض کنید لیستی از افراد جامعه در دسترس باشد. در این صورت ابتدا گام حرکت را به صورت $k=N/n=1000/100=10$ تعیین می کنیم. سپس از بین اعداد صحیح 1 تا 10 یک عدد به تصادف انتخاب می کنیم. فرض کنید عدد انتخاب شده 4 باشد در این صورت فرد ردیف چهارم در لیست افراد جامعه اولین عضو نمونه است. سایر اعضای نمونه افراد ردیفهای $4+10=14$ ، $14+10=24$ ، $24+10=34$ ، ... ، $984+10=994$ هستند.

نمونه گیری طبقه ای:

از این روش هنگامی استفاده می شود که افراد جامعه به گروههای مجزا از یکدیگر افراز شده باشند و این گروه بندی به نحوی باشد که افرادی که داخل هر گروه قرار می گیرند متجانس اما گروههای مختلف نا متناجس باشند. به عبارت دیگر واریانس یا

پراکندگی در داخل هر گروه کم اما بین گروهها زیاد باشد. چنین گروههائی را طبقه گویند. در روش طبقه ای نمونه به نحوی انتخاب می شود که از هر طبقه ای به نسبت حجم آن طبقه تعدادی در نمونه داشته باشیم. مثلا فرض کنید بخواهیم توسط نمونه گیری از یک شهر بزرگ درآمد ماهانه خانوارهای آن شهر را برآورد کنیم و نیز فرض کنید خانوارهای این شهر را بتوان از لحاظ درآمد به گروههای کم درآمد، درآمد متوسط، درآمد بالا و درآمد خیلی بالا تقسیم کرد. در این صورت هر کدام از این گروههای درآمدی یک طبقه را تشکیل می دهند چون خانوارهائی که در این گروهها قرار میگیرند از لحاظ درآمدی شبیه به یکدیگر بوده اما خانوارهای گروههای مختلف از لحاظ درآمدی متفاوت از یکدیگر هستند. به عبارت دیگر واریانس یا پراکندگی درآمد در بین خانوارهای هر کدام از این گروهها کم اما بین گروهها زیاد است. بنابر این این گروهها را طبقه گفته و برای نمونه گیری از روش نمونه گیری طبقه ای استفاده می کنیم. به این ترتیب که از همه این گروههای درآمدی یا همان طبقات نمونه میگیریم اما از هر طبقه ای به نسبت حجم آن طبقه. مثلا اگر طبقه افراد با درآمد پایین 20 درصد از افراد جامعه را تشکیل می دهند در این صورت 20 درصد از حجم نمونه را به این طبقه اختصاص می دهیم و به همین ترتیب الی آخر.

نمونه گیری خوشه ای:

از این روش نیز هنگامی استفاده می کنیم که افراد جامعه به گروههای مجزا از یکدیگر افراز شده باشند با این تفاوت که در این حالت گروه بندی به نحوی است که افرادی که داخل هر گروه قرار می گیرند نامتجانس اما گروههای مختلف متجانس هستند. به عبارت دیگر واریانس یا پراکندگی در داخل هر گروه زیاد اما بین گروهها کم است. یعنی در واقع هر کدام از این گروهها را می توان نمونه کوچک شده ای از کل جامعه در نظر گرفت که همه خواص جامعه را شامل است. چنین گروههائی را خوشه می نامیم. در نمونه گیری خوشه ای تعدادی از خوشه ها به تصادف انتخاب شده و آنها را سرشماری می کنند. و یا در صورت بزرگ بودن خوشه ها پس از انتخاب خوشه های نمونه از آنها یک نمونه تصادفی انتخاب می شود(خوشه ای دو مرحله ای) و یا ممکن است پس از انتخاب خوشه های نمونه از هر کدام از این خوشه ها نمونه به روش طبقه ای انتخاب کرد و یا حتی مجددا خوشه های نمونه را خوشه بندی کرده و از آنها به روش خوشه ای نمونه گرفت(خوشه ای چند مرحله ای). در هر صورت در روش خوشه ای از همه خوشه های جامعه نمونه گرفته نمی شود بلکه تنها از تعدادی از آنها نمونه گرفته می شود. مثلا فرض کنید بخواهیم در یک تحقیق دیدگاه دانشجویان یک دانشگاه بزرگ را در باره یک رویداد اجتماعی جویا شویم. فرض کنید این

دانشگاه دارای چند دانشکده بوده که هر دانشکده ای دارای یک ساختمان مجزائی است. از آنجائی که از نظر موضوع تحت مطالعه تفاوتی بین دانشجویان دانشکده های مختلف با یکدیگر نیست لذا هر دانشکده ای را می توان به عنوان یک خوشه در نظر گرفت و بنابر این برای نمونه گیری از این دانشگاه نیازی به اینکه از همه این دانشکده ها نمونه گرفت نیست بلکه کافی است یک یا چند دانشکده را به عنوان خوشه های نمونه انتخاب کرد. از آنجائیکه دانشجویان هر دانشکده زیاد بوده و امکان سرشماری همه آنها نیست (امکان اینکه همه دانشجویان را مورد پرسش و بررسی قرار داد) لذا می توان از هر کدام از دانشکده های نمونه یک نمونه تصادفی گرفت. فرض کنید ساختمان هر دانشکده ای چند طبقه بوده و کلاسهای دانشجویان در طبقات مختلف این ساختمان باشد. از آنجائیکه از لحاظ موضوع تحت مطالعه تفاوتی بین دانشجویان در طبقات مختلف نیست لذا هر کدام از این طبقات را می توان به عنوان یک خوشه در نظر گرفت و بر این اساس نیازی به اینکه از دانشجویان همه طبقات نمونه گرفت نیست بلکه کافی است یکی از این طبقات را به عنوان خوشه نمونه در نظر گرفته و دانشجویان آن طبقه را مورد سرشماری قرار داد. یا حتی اگر دانشجویان هر طبقه زیاد باشند می توان از آنها یک نمونه تصادفی انتخاب کرده و تنها آنها را مورد بررسی قرار داد.

توجه 4

در مسائل عملی و کاربردی ممکن است از یکی از این روشها و یا از ترکیبی از آنها استفاده شود.

پارامتر:

هر ویژگی عددی یک جامعه را یک پارامتر گویند به عبارت دیگر پارامتر عددی است که خاصیتی از جامعه را مشخص می کند. مثلا میانگین و واریانس جامعه (μ, σ^2) پارامتر هستند ، پارامترها مقادیر ثابتی هستند که در عمل همیشه مجهولند.

آماره: (استاتیستیک statistic):

آماره عددی است که خاصیتی از نمونه را بیان و مشخص می کند. به عبارت دیگر هر تابعی از نمونه که به هیچ پارامتر مجهولی بستگی نداشته باشد را آماره گویند. مثلا میانگین و واریانس نمونه $(\bar{x}, s^2)$ آماره هستند.

برآوردگر(برآورد یاب):

آماره ای است که از آن برای برآورد یا تخمین پارامتری از جامعه استفاده می شود. به عبارت دیگر هرگاه از یک آماره برای برآورد یا تخمین پارامتری از جامعه استفاده شود، آن آماره را برآورد گر یا تخمین زن آن پارامتر گویند مثلا میانگین و واریانس نمونه برآوردگرهائی برای میانگین و واریانس جامعه هستند.

توجه 5:

پارامترها مقادیر ثابت و مجهولند اما برآورد گرها و بطور کلی آماره ها ثابت نبوده بلکه متغیر تصادفی هستند.

نحوهٔ برآورد پارامترها:

بطور کلی پارامترها را به دو روش نقطه ای و فاصله ای برآورد می کنند.

در روش نقطه ای پارامتر جامعه را توسط یکی از آماره های نمونه برآورد می کنند. مثلا میانگین نمونه یک برآورد نقطه ای برای میانگین جامعه است. اما در روش فاصله ای پارامتر جامعه توسط یک فاصله برآورد می شود. فاصله ای که با ضریب اطمینان خوبی پارامتر جامعه را در بر بگیرد.

نرم افزار SPSS

SPSS مخفف Statistical Package for Social Science یکی از قدیمی ترین برنامه های آماری است که حتی قبل از پیدایش کامپیوتر های شخصی نیز این نرم افزار وجود داشته است و توسط این برنامه آماری تقریباً همه محاسبات و تحلیل ها و آزمون های آماری بخصوص رشته های غیر آماری را به راحتی می توان انجام داد.

پایه اولیه این نرم افزار در سال 1968 توسط سه دانشجوی دکتری در رشته های علوم اجتماعی، تحقیق در عملیات و برنامه نویسی در دانشگاه استنفورد گذاشته شد. پس از آن این برنامه کامپیوتری که با ایده تبدیل داده های خام به اطلاعات ضروری برای انجام تصمیم گیری ها بنا شده بود در سایر دانشگاههای امریکا نیز مورد استفاده قرار گرفت. در سال 1975 دو نفر از سه نفر جوان فوق شرکت SPSS را بنا نهادند و این شرکت به موفقیتهای مالی زیادی از طریق فروش این برنامه کامپیوتری دست یافت. از استفاده کنندگان اولیه SPSS که وابسته به ارگانهای دولتی بودند می توان سازمان ناسا را نام برد. اولین نسخه های این نرم افزار برای کامپیوترهای بزرگ (Main frame) که اطلاعات را از طریق کارت پانچ دریافت می کردند طراحی شده بود. پس از پیدایش کامپیوترهای شخصی در سال 1982 اولین نرم افزار آماری تحت سیستم عامل DOS نرم افزار SPSS بود. با ظهور Windows به عنوان یک سیستم عامل، نسخه تحت ویندوز این نرم افزار نیز ارائه شد. طی این سالها با توسعه امکانات سخت افزاری و نرم افزاری کامپیوترها این نرم افزار نیز توسعه یافت. در سال 2009 شرکت SPSS به کمپانی بزرگ IBM واگذار شد. نسخه های جدید این نرم افزار (از نسخه 18 به بعد) با نام PASW به بازار عرضه شده است و نام شرکت SPSS از ژانویه 2010 به شرکت SPSS An IBM تغییر پیدا کرد.

در ابتدای ورود به محیط SPSS دو پنجره با نام های Data View و Variable View باز می شود که پنجره Data ماتریسی است شامل تعداد زیادی سطر و ستون برای ورود داده ها و پنجره Variable برای معرفی متغیرها و ویرایش فایل داده است. نحوه ورود داده ها در پنجره Data طوری است که اطلاعات مربوط به هر فرد نمونه در یک ردیف و اطلاعات مربوط به هر متغیر در یک ستون از این جدول قرار می گیرد. مثلاً اگر در تحقیقی بر روی 20 نفر دانشجو برخی ویژگی های آنان مانند جنسیت، قد، وزن و ... اندازه گیری شده باشند و تعداد این ویژگی ها 12 صفت باشد در اینصورت فایل داده ماتریسی است شامل 20 سطر و 12 ستون.

برای آشنایی با نحوه ورود داده ها در پنجره Data مثال زیر را در نظر بگیرید:

مثال:

فرض کنید در یک تحقیق بر روی گروهی از دانشجویان این کلاس برخی ویژگی های آنان توسط ابزار پرسشنامه اندازه گیری شده باشد، در اینصورت پس از جمع آوری پرسشنامه ها ابتدا لازم است داده های این پرسشنامه ها را استخراج کرده و در جدولی به شکل

ردیف	جنسیت	گروه خونی	قد	معدل لیسانس	علاقه به رشته	علاقه آمار
۱	۲	۴	۱۸۰	۱۸/۸۲	۴	۳
۲	۲	۳	۱۸۶	۱۶/۸۲	۳	۲
۳	۱	۲	۱۸۵	۱۵/۵۰	۴	۴
۴	۱	۲	۱۶۸	۱۶/۷۰	۴	۳
۵	۱	۱	۱۶۸	۱۵/۸۶	۲	۴
۶	۱	۴	۱۶۸	۱۶/۴۰	۴	۳
۷	۲	۴	۱۶۷	۱۶/۵۰	۴	۴
۸	۲	۳	۱۶۵	۱۶/۸۵	۵	۴
۹	۱	۲	۱۶۵	۱۶/۵۰	۴	۴
۱۰	۱	۳	۱۶۰	۱۶/۶۸	۴	۳

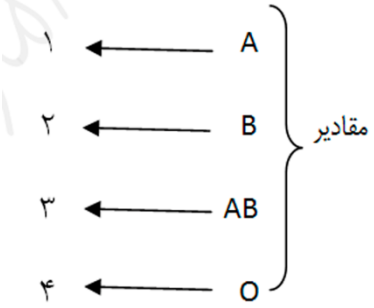
جدول زیر تنظیم کنیم:

در این تحقیق فرض شده است که تعداد دانشجویان یا همان نمونه 10 نفر بوده و بر روی هر فرد نمونه تعداد شش ویژگی اندازه گیری شده است. به عبارت دیگر پرسشنامه ای که در اختیار هر فرد نمونه قرار گرفته است دارای شش سئوال جنسیت، گروه خونی و . . . بوده است. پس از جمع آوری پرسشنامه ها و قبل از اقدام به ورود داده ها به محیط SPSS توصیه می شود ابتدا آنها را در جدولی به شکل جدول فوق وارد کنید. بدین منظور لازم است متغیرها را نامگذاری کرده و در مورد متغیرهای کیفی مانند جنسیت، گروه خونی، پاسخهای داده شده به سئوالات پرسشنامه و مانند اینها مقادیر این متغیرها را نیز کد گذاری کنید. در مورد متغیرهای کمی فقط کافی است متغیرها را نامگذاری کنید چون مقادیر این متغیرها کمیت بوده و نیازی به کد گذاری ندارند. در این مثال برای نامگذاری متغیرها و نیز کد گذاری مقادیر متغیرهای کیفی به صورت زیر عمل شده است.

متغیر جنسیت: نام: X1 Lable: Jensiat

مقادیر
- ۱ ← زن
- ۲ ← مرد

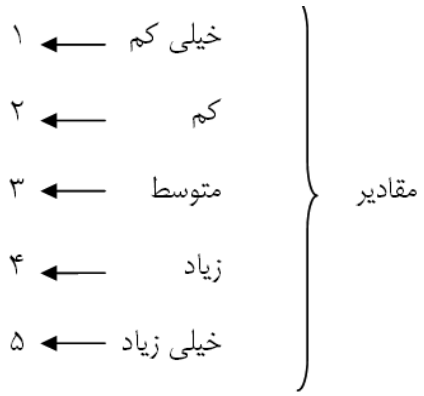
متغیر گروه خونی: نام: X_2 Lable: Goroh



متغیر قد: نام: X_3 Lable: Ghad

متغیر معدل لیسانس: نام: X_4 Lable: Moaddel

متغیرمیزان علاقه به رشته: نام: X_5 Lable: Resht



متغیرمیزان علاقه به رشته آمار: نام: X_6 مقادیر همان مقادیرX_5 Lable: Amar

توجه : ویژگی های کمی را حتماً به صورت کمی اندازه گیری کنید. به عبارت دیگر در طراحی پرسشنامه برای اندازه گیری ویژگی های مختلف، تا جائیکه بتوان آنها را به صورت کمی اندازه گیری کرد، به هیچ عنوان آنها را به صورت کیفی و طبقه ای اندازه گیری نکنید. چون در چنین صورتی قسمتی از اطلاعات موجود در داده ها را دور ریخته اید. مثلا برای اندازه گیری وزن آن را به صورت

طبقه ای مانند " کمتر از 50"، "50–60"، "60–70" و . . . اندازه گیری نکنید بلکه عدد وزن را ثبت کنید. بعدا هرگونه طبقه بندی را به راحتی می توان توسط SPSS انجام داد.

پس از تنظیم داده ها در جدول فوق اطلاعات این جدول را به همین صورتی که هست در پنجره Data وارد می کنیم. یعنی در این پنجره داده های مربوط به جنسیت را در یک ستون، داده های مربوط به گروه خونی در ستون دیگر و الی آخر. در چنین صورتی همانطوریکه مشاهده می شود اگر مثلا حجم نمونه 10 و تعداد متغیرها 20 باشد آنگاه فایل داده جدولی شامل 10 سطر و 20 ستون خواهد بود.

پس از ورود داده ها در پنجره Data لازم است این فایل را به نحو مناسبی ویرایش کنیم که برای ویرایش فایل به پنجره Variable View می رویم. در این پنجره:

پنجره Variable View

در این پنجره:

ستون اول با عنوان **Name** برای نامگذاری متغیر ها است، که معمولاً در آمار از حرف X_1 و X_2 و ... برای نامگذاری متغیر ها استفاده می شود.

ستون دوم بعنوان **Type** برای تعیین نوع متغیر است که SPSS به صورت پیش فرض متغیرها را عددی 8 کاراکتری با دو نقطه اعشار در نظر می گیرد اما با کلیک کردن بر روی این بخش می توان نوع متغیر را تغییر داد. با کلیک کردن بر روی این بخش پنجره ای باز می شود که در این پنجره می توان تعداد کاراکترهای متغیرهای عددی و نیز تعداد نقاط اعشار آنها را تغییر داد. مثلا اگر داده ها بدون نقطه اعشار هستند می توان Decimal places ان را صفر کرد تا داده ها در فایل داده بدون نقطه اعشار نشان داده شوند. همچنین برای داده های بزرگ می توان از جدا کننده Dot یا Comma استفاده کرد. همچنین می توان از نمادهای علمی (Scientific notation) استفاده کرد. از نمادهای علمی نیز برای نشان دادن اعداد بزرگ یا اعداد کوچک بخصوص در علوم ریاضی استفاده می شود. مطابق با نمادهای علمی مثلا عدد 1230 به صورت 1.23E+3 یعنی 1.23×10^3 نشان داده می شود. همچنین عدد 0.0345 به صورت 3.45E-2 یعنی 3.45×10^{-2} نشان داده می شود..

دو ستون بعدی با عناوین Width و Decimals تعداد کارکترهای متغیر عددی و نیز تعداد نقاط اعشاری آنها است. یعنی همان ویژگی هایی که در پنجره Type نیز دیده می شود.

ستون بعدی با عنوان Lable نیز برای نامگذاری متغیرها است. Lable در واقع نام دیگری برای متغیر است، منتها نامی که در خروجی نشان داده می شود و توصیه می شود برای Lable گذاری متغیرها از اسامی با مفهوم استفاده شود. Name نامی است که در فایل داده نشان داده می شود و Lable نامی است که در خروجی نشان داده می شود. مثلا اگر متغیر جنسیت را در فایل داده با X_1 نامگذاری کردید برای Lable آن توصیه می شود از " جنسیت" استفاده شود که هم می توان آن را به فارسی و یا انگلیسی وارد کرد. استفاده از فونتهای انگلیسی دارای مشکلات کمتری است و توصیه می شود از فونتهای انگلیسی استفاده شود.

ستون بعدی با عنوان **Values** برای Lable گذاری مقادیر متغیر های کیفی است. متغیرهای کیفی که مقادیر آن را کد گذاری کرده و کدها را در فایل داده وارد کرده اید توصیه می شود در این بخش کدها را Lable گذاری کنید. چون در اینصورت در خروجی به جای نشان دادن آن کدها Lable ها را نشان خواهد داد.

ستون بعدی با عنوان Missing برای تعیین کد داده های گم شده است. در SPSS اگر خانه ای از جدول خالی گذاشته شود، SPSS آنرا به طور پیش فرض Missing فرض می کند. اما در این بخش می توان برای داده های گم شده کد خاصی را در نظر گرفت. در حالت مبتدی توصیه می شود اگر مقداری از یک متغیر نامعلوم بود خانه مربوط به آن داده را خالی بگذارید

ستون بعدی با عنوان Columns ، تعداد ستون های در نظر گرفته شده برای هر متغیر است که به صورت پیش فرض SPSS برای هر متغیر 8 ستون در نظر می گیرد. اما می توان آنرا تغییر داد.

ستون بعدی با عنوان Align ، مکان قرار گرفتن داده ها در داخل متغیرها است، که به صورت پیش فرض داده ها در سمت راست هر متغیر قرار می گیرد. اما می توان مکان آنرا تغییر داد.

ستون بعدی با عنوان Measure ، برای تعیین مقیاس یا سطح اندازه گیری داده ها است.SPSS متغیرها را در 3 سطح اسمی، ترتیبی، و کمی (فاصله ای و نسبتی) شناسایی می کند.

تمرین :

داده های مربوط به ویژگی های دانشجویان این کلاس را در پنجره Data وارد کرده و این فاصله را به نحو مناسبی ویرایش نمایید.

معرفی فایل Test 2

این فایل محتوی داده های مربوط به تحقیقی بر روی یک نمونه تصادفی از دانش آموزان مقطع راهنمایی شهرستان مشهد در دو منطقه شمال و جنوب (مرفه و محروم) این شهرستان است. متغیرهای این فایل شامل:

منطقه X_1، جنسیتX_2 ، پایه تحصیلی X_3 ، و نمرات دروس ریاضی، علوم ، حرفه و فن زبان فارسی است، به ترتیب X_4 .... X_8

توجه: بطور کلی در تهیه گزارش های آماری مثلاً فصل چهار پایان نامه لازم است ابتدا و قبل از انجام آزمون فرضیه ها توصیفی از نمونه بعمل آید، که توصیف نمونه در مورد متغیرهای کیفی در قالب تنظیم داده ها در جداول توزیع فراوانی و رسم نمودارهای میله ای و دایره ای است و در مورد متغیر های کمی در قالب محاسبه شاخص های آماری و رسم هیستوگرام فراوانی است.

برای تنظیم داد ه های متغیرهای کیفی در جداول توزیع فراوانی و رسم نمودار های میله ای و دایره ای از مسیر زیر استفاده می کنیم:

Analyze ⟶ Descriptive Statistics ⟶ Frequency

مثلاً فرض کنید بخواهیم جدول توزیع فراوانی نمونه برحسب پایه تحصیلی را بدست آوریم در اینصورت پس از رفتن به مسیر فوق، متغیر مربوط به پایه تحصیلی (X_3) را به قسمت Variable می بریم و بعد آن را OK می کنیم. پس از اجرای این فرمان نتیجه در قالب دو جدول و در خروجی خواهد آمد، که در اولین جدول گزارشی از داده های گم شده و داده های معتبر خواهد آمد و در دومین جدول توزیع فراوانی نمونه بر حسب پایه تحصیلی خواهد آمد. این جدول در زیر آمده است.

GRADE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FIRST	425	31.5	31.5	31.5
	SECOND	484	35.9	35.9	67.3
	THIRD	441	32.7	32.7	100.0
	Total	1350	100.0	100.0	

در این جدول ستون اول فراوانی، ستون های دوم و سوم به ترتیب درصد و درصد معتبر و آخرین ستون نیز درصد تجمعی شده است.

تفاوت درصد و درصد معتبر در آن است که در محاسبه درصد Missing ها به حساب می آید، اما در محاسبه درصد معتبر Missing ها به حساب نمی آیند.

برای رسم نمودارهای میله ای و دایره ای در پنجره Frequencies بر روی گزینه Charts کلیک کرده و در پنجره باز شده نمودار مورد نظر را انتخاب می کنیم. از نمودارهای میله ای (***Bar Chart***) و دایره ای (***Pie Chart***) برای متغیرهای کیفی و از نمودار هیستوگرام برای متغیرهای کمی استفاده می شود. معمولا از نمودار میله ای برای نشان دادن فراوانی ها و از نمودار دایره ای برای نشان دادن فراوانی های نسبی یا درصدها استفاده می شود.

توجه: پس از رسم نمودار و در پنجره خروجی با دو بار کلیک کردن بر روی هر نمودار به پنجره دیگری با نام Chart Editor می رویم که در این پنجره می توان نمودارمورد نظر را ویرایش نمود.

تمرین:

جدول توزیع فراوانی متغیرهای جنسیت، منطقه و پایه تحصیلی را بدست آورده و در هر حالتی نمودارهای میله ای و دایره ای را نیز رسم کنید. نمودار میله ای را بر اساس فراوانی ها و نمودار دایره ای را بر اساس درصدها رسم کنید و این نمودارها را کمی ویرایش کنید.

برای محاسبه شاخص های آماری متغیرهای کمی از مسیر زیر استفاده می کنیم.

Analyze ⟶ Descriptive Statistics ⟶ Descriptive

مثلاً فرض کنید بخواهیم شاخص های آماری نمرات درس ریاضی دانش آموزان را بدست آوریم در این صورت پس از رفتن به مسیر فوق متغیر مربوط به نمرات درس ریاضی(X_4) را به قسمت Variable می بریم. با کلیک کردن بر روی گزینه Option می توان شاخصهای آماری که در خروجی نشان داه خواهد شد را تغییر داد.

پس از اجرای این فرمان نتیجه در قالب یک جدول و در خروجی خواهد آمدکه این جدول و به صورت عمودی در زیر آمده است. لازم به ذکر است که در حالت معمولی این جدول و نیز سایر جداول به صورت افقی در خروجی نشان داه خواهد شد که در اینجا به جهت محدودیت مکانی آن را به صورت عمودی نشان داه ایم.

Descriptive Statistics

		MATH.	Valid N (listwise)
N	Statistic	1311	1311
Range	Statistic	19.00	
Minimum	Statistic	1.00	
Maximum	Statistic	20.00	
Mean	Statistic	12.9605	
	Std. Error	.13454	
Std. Deviation	Statistic	4.87145	
Variance	Statistic	23.731	
Skewness	Statistic	-.182	
	Std. Error	.068	
Kurtosis	Statistic	-1.002	
	Std. Error	.135	

در این جدول سطر اول به عنوان N حجم نمونه است ، نمونه ای که محاسبه بر روی آن انجام شده است یعنی حجم نمونه بدون در نظر گرفتن داه های گمشده.. سطر دوم با عنوان Range ، دامنه تغییرات نمونه است که دامنه تغییرات فاصله بین کمترین و بیشترین مقدار نمونه بوده و یکی از شاخصهای پراکندگی است.. دو سطر بعدی جدول نیز مقادیر کمینه و بیشینه نمونه است. دو سطر بعدی این جدول با عنوان Mean به ترتیب میانگین نمونه و خطای استاندارد میانگین نمونه است.

خطای استاندارد یک آماره در واقع انحراف معیار آن آماره است. یعنی شاخصی است که مشخص می کند مقدار یک آماره به طور متوسط چقدر تا مقدار واقعی آن پارامتر فاصله دارد. مثلاً در این مثال میانگین نمونه 12/9605 و خطای استاندارد آن حدود

0/13454 می باشد. این بدان معنا است که میانگین واقعی جامعه که یک پارامتر مجهول است برآورد می شود که حدود 12/96 باشد و نیز برآورد می شود که این میانگین به طور متوسط حدود 0/134 کمتر یا بیشتر از 12/96 باشد. ستون بعدی این جدول با عنوان Std .Deviation انحراف معیار نمونه(S) و ستون بعدی نیز واریانس نمونه است.

توجه: همانطوریکه می دانیم واریانس شاخص پراکندگی است اما واحد اندازه گیری آن با واحد اندازه گیری داده ها یکسان نبوده بلکه مجذور یا مربع واحد اندازه گیری داده ها است. مثلا اگر داده ها اندازه قد دانشجویان یک کلاس بر حسب متر باشد آنگاه واحد اندازه گیری میانگین داده ها نیز متر است حال آنکه واحد اندازه گیری واریانس آنها متر مربع خواهد بود. به همین منظور از واریانس جذر گرفته و جذر آن را انحراف معیار یا انحراف استاندارد نامیده اند که واحد اندازه گیری انحراف معیار با واحد اندازه گیری داده ها یکسان و از انحراف معیار به عنوان شاخص پراکندگی استفاده شده است.

توجه: اگر میانگین و انحراف معیار اندازه قد دانشجویان یک کلاس بر حسب سانتی متر به ترتیب 170 و 12 باشد این به آن معنی است که اندازه قد هر دانشجو در این کلاس به طور متوسط 170 سانتی متر است و نیز به طور متوسط قد هر دانشجو در این کلاس حدود 12 سانتی متر بیشتر یا کمتر از 170 است.

دو سطر بعدی این جدول با عنوان skewness به ترتیب چولگی و خطای استاندارد آن است.

چولگی شاخصی است برای اندازه گیری ***میزان تقارن*** یا ***عدم تقارن*** یک توزیع که اگر توزیع متقارن باشد مانند توزیع نرمال چولگی آن صفر است. اما اگر توزیع نا متقارن بوده و دنباله آن به سمت راست کشیده شده باشد(چولگی به راست) ضریب چولگی آن عددی مثبت است و اگر توزیعی نامتقارن بوده و دنباله آن به سمت چپ کشیده شده باشد(چولگی به چپ) ضریب چولگی آن عدد منفی خواهد بود مانند اشکال زیر.

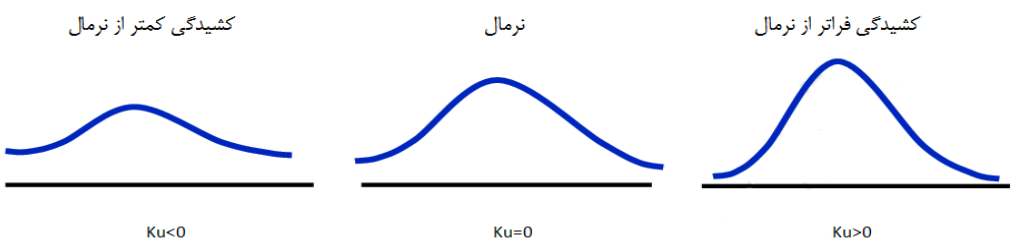
توجه کنید که در یک توزیع چوله به راست وزن یا فراوانی داده های کوچک بیشتر از داده های بزرگ است و در توزیع چوله به چپ خلاف آن است. مثلا اگر چولگی نمرات درس آمار در یک کلاس مثبت باشد (چوله به راست) این به آن معنی است که در این کلاس فراوانی نمرات کوچک بیشتر از فراوانی نمرات بزرگ است یا به عبارت دیگر در این کلاس اکثر دانشجویان نمراتی پایین تر از میانگین گرفته اند. در مورد چولگی منفی خلاف آن است.

تقسیم بندی زیر در مورد ضریب چولگی برقرار است:

۱) اگر چولگی بین	۰/۱≤SK≤ ۰/۱ـ	توزیع تقریبا متقارن است.
۲) اگر چولگی بین	۰/۵≤SK≤ ۰/۱	توزیع کمی چولگی به راست است.
۳) اگر چولگی	SK >۰/۵	توزیع چولگی به راست است.
۴) اگر چولگی بین	۰/۱ ـ ≤SK≤ ـ ۰/۵	توزیع کمی چولگی به سمت چپ است.
۵) اگر چولگی	SK< ـ۰/۵	توزیع چولگی به چپ است.

در این مثال همانطور یکه مشاهده می شود، ضریب چولگی 0/182 ـ می باشد. که بیانگر کمی چوله به چپ بودن توزیع نمرات درس ریاضی دانش آموزان است یعنی در این نمونه نمرات بزرگ کمی بیشتر از نمرات کوچک است.

دو سطر آخر این جدول با عنوان Kurtosis ، کشیدگی و خطای استاندارد آن است. کشیدگی شاخصی است که میزان برجستگی یا ارتفاع یک توزیع نسبت به توزیع نرمال را نشان می دهد،که در مورد توزیع نرمال کشیدگی **صفر** است، و اگر کشیدگی یک توزیع بیشتر از نرمال باشد یعنی پراکندگی آن کمتر از نرمال باشد ضریب کشیدگی عددی **مثبت** است و بالعکس اگر کشیدگی یک توزیع کمتر از نرمال باشد یا پراکندگی آن بیشتر از نرمال باشد ضریب کشیدگی عددی **منفی** خواهد بود.



تقسیم بندی زیر در مورد ضریب کشیدگی برقرار است:

1- اگر $-.1 \leq KU \leq +.1$ توزیع تقریبا نرمال است.

2- اگر $+.1 < \mathrm{KU} \leq +.5$ کشیدگی کمی بیشتر از نرمال یا پراکندگی کمی کمتر از نرمال است.

3- اگر $KU > +.5$ کشیدگی بیشتر از نرمال و یا پراکندگی کمتر از نرمال است.

4- اگر $-.5 \leq KU < -.1$ کشیدگی کمی کمتر از نرمال و یا پراکندگی کمی بیشتر از نرمال است.

5- اگر $KU < -.5$ کشیدگی کمتر از نرمال و یا پراکندگی بیشتر از نرمال است.

در این مثال همانطوری که مشاهده می شود ضریب کشیدگی 1 ـ است که این بدان معنا است که توزیع نمره ریاضی دانش آموزان دارای کشیدگی خیلی کمتری نسبت به توزیع نرمال است، یا به عبارت دیگر پراکندگی در توزیع نمره ریاضی دانش آموزان نسبت به توزیع نرمال خیلی بیشتر است.

تمرین:

شاخص های آماری نمرات دروس مختلف را توسط فرمان Descriptive محاسبه کرده و نتایج حاصل را به نحو مناسبی تفسیر کنید.

محاسبه شاخص های آماری یک متغیر کمی به تفکیک گروه های مختلف یک متغیر کیفی:

برای محاسبه شاخص های آماری یک متغیر کمی به تفکیک گروه های مختلف یک متغیر کیفی از مسیر زیر استفاده می کنیم.

Analyze ⟶ Descriptive ⟶ Explore

مثلاً فرض کنید بخواهیم شاخص های آماری نمرات درس ریاضی دانش آموزان را به تفکیک جنسیت محاسبه کنیم در اینصورت پس از رفتن به مسیر فوق متغیر مربوط به نمرات درس ریاضی دانش آموزان X_4 را به قسمت Dependent List و متغیر جنسیت را به قسمت Factor List می بریم (توجه داشته باشید که بطورکلی متغیرهای کیفی که افراد جامعه را به گروه های مجزا از هم تقسیم می کند مانند جنسیت یا گروه خونی را فاکتور یا عامل نیز می نامند).

در قسمت پایین این پنجره و در بخش Display می توان مشخص کرد که این فرمان تنها شاخص های آماری را محاسبه کند یا نمودارهای مربوطه را رسم کند، و یا هر دو کار را انجام دهد. پیش فرض آن Both است. پس از اجرای این فرمان نتیجه در قالب یک جدول و در خروجی خواهد آمد که این جدول در زیر آمده است.

Descriptives

	SEX			Statistic	Std. Error
MATH.	GIRL	Mean		14.8820	.18632
		95% Confidence Interval for Mean	Lower Bound	14.5161	
			Upper Bound	15.2478	
		5% Trimmed Mean		15.1295	
		Median		16.5000	
		Variance		22.426	
		Std. Deviation		4.73562	
		Minimum		2.00	
		Maximum		20.00	
		Range		18.00	
		Interquartile Range		8.50	
		Skewness		-.565	.096
		Kurtosis		-1.025	.192
	BOY	Mean		11.0940	.16431
		95% Confidence Interval for Mean	Lower Bound	10.7714	
			Upper Bound	11.4166	
		5% Trimmed Mean		11.1186	
		Median		11.2500	
		Variance		17.953	
		Std. Deviation		4.23714	
		Minimum		1.00	
		Maximum		20.00	
		Range		19.00	
		Interquartile Range		6.00	

	Skewness	-.116	.095
	Kurtosis	-.660	.189

این جدول دارای 2 بخش اصلی است که در هر بخشی شاخص های آماری نمرات درس ریاضی به تفکیک برای دخترها و پسرها آمده است. در هر بخش از این جدول اولین ردیف میانگین نمونه و خطای استاندارد آن است. در این مثال میانگین نمونه و خطای استاندارد آن برای نمرات درس ریاضی دخترها به ترتیب 14/88 و 0/186 و در مورد پسرها به ترتیب 11/09 و 0/164 است.

دو ردیف بعدی این جدول با عناوین Upper Bound و Lower Bound به ترتیب کرانهای پایین و بالای یک فاصله اطمینان 95٪ برای میانگین جامعه است، که مثلاً در مورد دخترها این دو کران به ترتیب14/5161 و 15/2478 است. این بدان معنی است که با 95٪ اطمینان میانگین نمره ریاضی دانش آموزان دختر (جامعه دانش آموزان دختر) حداقل 14/5161 و حداکثر 15/2478 می باشد. یعنی در فاصله (15/2478 و 14/5478) قرار دارد.

ردیف بعدی این جدول با عنوان 5% Trimmed mean میانگین پیراسته 5 درصدی است. یعنی 5٪ از داده های خیلی بزرگ و 5٪ از داده های خیلی کوچک کنار گذاشته شده و میانگین داده های 90٪ وسط محاسبه شده است که این عمل به جهت حذف اثر داده های دور افتاده بر میانگین انجام می گیرد. ردیف بعدی با عنوان Median میانه نمونه است. میانه نقطه وسط داه ها است یعنی داده ای که نیمی از داده ها کوچکتر یا مساوی آن است و یکی از شاخصهای تمرکز است. تفاوت میانگین با میانه در آن است که میانگین مرکز ثقل داده ها است یعنی میانگین داده ای است که وزن داده های سمت چپ و راست آن با یکدیگر برابر است یا به عبارت دیگر میانگین داده ای است که مجموع انحرافات داده ها در سمت چپ و راست آن (وزن داده ها در سمت چپ و راست آن) با یکدیگر برابر است حال آنکه میانه داده ای است که تعداد داده های سمت چپ و راست آن با یکدیگر برابر است. استفاده از میانگین بسیار متداول تر از میانه است چون از لحاظ تئوری ارزش میانگین از میانه بیشتر است اما با این حال مواردی وجود دارد که استفاده از میانه به جای میانگین ترجیح داده می شود. مثلا هنگامی که داده ها دارای داده های پرت یا دور افتاده (داده هائی که نسبت به سایر داده ها خیلی بزرگ یا خیلی کوچک است) است توصیه می شود به جای میانگین از میانه استفاده شود. همچنین هنگامی که داده ها از یک متغیر ترتیبی هستند بهتر است از میانه استفاده شود.

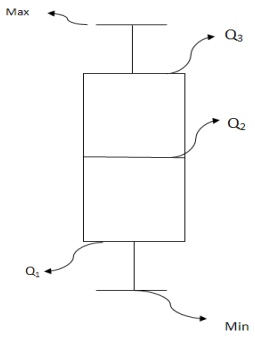
ردیف های بعدی به ترتیب واریانس، انحراف معیار، کمترین، بیشترین و دامنه تغییرات(Range) نمونه است که قبلا راجع به آنها صحبت شد. ردیف بعدی با عنوان Interquartile Range دامنه چارک ها است. یعنی فاصله بین چارک اول تا چارک سوم که یکی از شاخصهای پراکندگی بوده و پراکندگی در نیمه میانی داده ها را اندازه گیری می کند.

دو ردیف بعدی این جدول نیز چولگی و کشیدگی و خطای استاندارد آنها است که راجع به آنها قبلا صحبت شد.

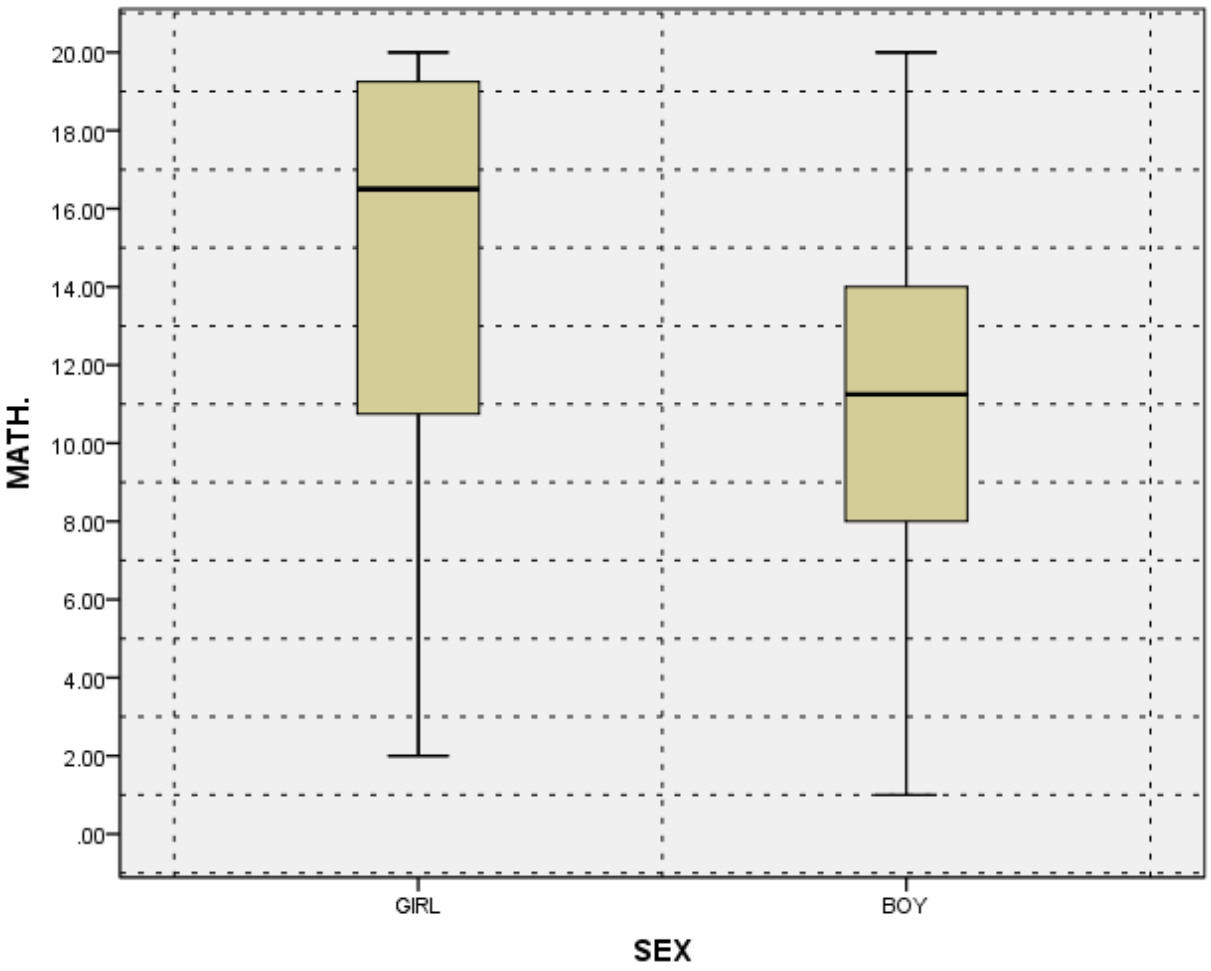
این فرمان نمودارهای جعبه ای (Box Plot) و نیز نمودار ساقه و برگ (Stem and Leaf Plot) را نیز رسم می کند که این نمودارها جزو نمودارهای اکتشافی است. نمودار ساقه و برگ در واقع شکل دیگری از هیستوگرام فراوانی است که فراوانی ها نه به شکل ستونهای پیوسته (مانند آنچه که در هیستوگرام رسم می شود) بلکه توسط اعداد نشان داده خواهد شد. کاربرد نمودار جعبه ای بیشتر از نمودار ساقه و برگ است و به همین دلیل راجع به آن کمی صحبت می کنیم.

نمودار جعبه ای Box Plot

این نمودار از روی مقادیر کمترین، بیشترین و چارک های اول، دوم و سوم رسم می شود و توسط این نمودار علاوه بر اینکه توزیع یک متغیر دریک گروه مورد بررسی قرار می گیرد گروه های مختلف را نیز می توان توسط آن با یکدیگر مقایسه کرد. نمونه ای از یک نمودار جعبه ای در زیر رسم شده است. در این نمودار خط پایین نمودار بیانگر کمترین مقدار نمونه، قاعده پائین نمودار بیانگر چارک اول، خط میانی بیانگر چارک دوم یا همان میانه، قاعده بالا بیانگر چارک سوم و خط بالای نمودار نیز بیانگر بیشترین مقدار نمونه است. توجه کنید که چارک اول داده ای است که یک چهارم یا 25 درصد از داده ها کوچکتر یا مساوی آن است، چارک دوم یا همان میانه داده ای است که دو چهارم یا پنجاه درصد از داده ها حداکثر مساوی آن است و چارک سوم داده ای است که سه چهارم یا 75 درصد از داده ها حداکثر مساوی آن است.



نمودار جعبه ای نمرات درس ریاضی دخترها و پسرها در زیر آمده است.



توجه کنید که مطابق این نمودار نیز می توان چوله به چپ بودن توزیع نمره ریاضی دخترها و تقریبا متقارن بودن توزیع نمره ریاضی پسرها (چیزی که از مقدار شاخص چولگی آنها نیز مشخص بود) را مشاهده کرد. همچنین مطابق آنچه که از این نمودار پیداست، توزیع نمره ریاضی دخترها از وضعیت خیلی بهتری نسبت به پسرها برخوردار است.

تمرین: توسط فرمان Explore شاخص های آماری نمرات دروس مختلف را به تفکیک جنسیت، منطقه، پایه تحصیلی محاسبه کرده و آنها را تفسیر کنید. همچنین نمودار جعبه ای مربوط را رسم کرده و آنرا کمی ویرایش کنید.

تشکیل جدول توزیع فراوانی بر حسب دو متغیر کیفی (جدول متقاطع):

در مواردی ممکن است بخواهیم توزیع فراوانی نمونه بر حسب دو متغیر کیفی را به صورت یک جدول متقاطع (توافقی) بدست آوریم بدین منظور از مسیر زیر استفاده می کنیم:

Analyze ⟶ Descriptive Statistics ⟶ Cross Tabs

مثلاً فرض کنید بخواهیم توزیع فراوانی نمونه بر حسب جنسیت و پایه تحصیلی را به شکل جدولی مانند جدول زیر بدست آوریم:

پایه تحصیلی / جنسیت	اول	دوم	سوم
دختر			
پسر			
جمع			

بدین منظور پس از رفتن به مسیر فوق یکی از این دو متغیر را به قسمت Row (سطر) و دیگری را به قسمت Column (ستون) می بریم و پس از اجرای این فرمان نتیجه در قالب یک جدول متقاطع و در خروجی خواهد آمد که این جدول به شکل انگلیسی آن در زیر نیز آمده است.

SEX * GRADE Crosstabulation

Count

		GRADE			
		FIRST	SECOND	THIRD	Total
SEX	GIRL	210	239	226	675
	BOY	215	245	215	675
Total		425	484	441	1350

توجه کنید که در یک جدول متقاطع دو بعدی سه نوع درصد می توان محاسبه کرد:

1) درصدی که بر پایه جمع های سطری محاسبه می شود بنام درصد سطری.

2) درصدی که بر پایه جمع های ستونی محاسبه می شود بنام درصد ستونی.

3) درصدی که بر پایه جمع کل نمونه محاسبه می شود بنام درصد کل.

برای محاسبه چنین درصدهایی در پنجره Cross Tabs بر روی گزینه Cell کلیک کرده و در قسمت Percentages نوع درصد مورد نظر را انتخاب می کنیم.

تمرین: جدول توزیع فراوانی نمونه بر حسب متغیر های (جنسیت/منطقه)، (جنسیت/پایه تحصیلی) و (پایه تحصیلی و منطقه) را به صورت یک جدول متقاطع بدست آورده و در هر حالتی درصدهای سطری، ستونی و کل را نیز محاسبه کنید.

«آمار استنباطی»

همانطور که در مقدمه نیز بیان شد موضوع آمار استنباطی ***مطالعه یک جامعه از روی داده های یک نمونه از آن جامعه*** است. به عبارت دیگر در آمار استنباطی فرض بر این است که به دلایل مختلف امکان مشاهده جامعه آماری نبوده بلکه تنها می توان به یک نمونه از آن جامعه دسترسی داشت . لذا هر گونه استنباط و استنتاج درباره جامعه باید از طریق داده های یک نمونه از آن جامعه انجام گیرد. بنابر این **اولین گام** در موضوع آمار استنباطی چگونگی و روش نمونه گیری از یک جامعه است. واضح است که به روش های مختلف می توان از یک جامعه آماری نمونه گرفت اما در آمار، نمونه باید **تصادفی** باشد و منظور از یک نمونه تصادفی نمونه ای است که در آن همه افراد جامعه شانسی برای انتخاب در نمونه را داشته باشند و این شانس برای همه افراد جامعه یکسان باشد که چنین نمونه ای را نمونه **نا اُریب** نیز می نامند. یعنی نمونه ای که می تواند به خوبی معرف جامعه اش باشد و همه خواص جامعه را شامل است. برای اخذ یک نمونه تصادفی از یک جامعه روش های مختلفی وجود دارد که این روش ها

بنابر نوع موضوع تحت مطالعه و چگونگی توزیع جامعه متفاوت از یکدیگرند. اما عمده ترین روش های کلاسیک نمونه گیری که اکثراً از این روش ها و یا ترکیبی از آنها استفاده می شود به شرح زیر است:

1) **نمونه گیری تصادفی ساده**
2) **نمونه گیری تصادفی منظم (سیستماتیک)**
3) **نمونه گیری طبقه ای**
4) **نمونه گیری خوشه ای(شکل کوچک شده از جامعه)**

نمونه گیری تصادفی ساده

نمونه گیری تصادفی ساده پایه و اساس همه روشهای نمونه گیری آماری است. در این روش نمونه به نحوی انتخاب می شود که همه افراد جامعه شانسی برای انتخاب در نمونه را داشته و این شانس برای همه افراد یکسان باشد. نمونه گیری تصادفی ساده را می توان به دو روش با جایگذاری و بدون جایگذاری انجام داد. در روش با جایگذاری هر فرد جامعه می تواند بیش از یکبار در نمونه انتخاب شود حال آنکه در روش بدون جایگذاری هر فرد جامعه فقط یکبار شانسی برای انتخاب در نمونه را دارد. برای انتخاب یک نمونه تصادفی ساده از یک جامعه می توان از روشهای مختلفی استفاده کرد. مثلا می توان به هر فرد جامعه یک شماره نسبت داده و سپس با استفاده از جدول اعداد تصادفی و به تعداد نمونه مورد نظر افرادی از آن جامعه را انتخاب کرد. البته امروزه با وجود برنامه های کامپیوتری می توان به جای استفاده از جدول اعداد تصادفی از این برنامه ها در تولید اعداد تصادفی کمک گرفت.

نمونه گیری تصادفی منظم (سیستماتیک)

از این روش هنگامی استفاده می شود که لیستی از افراد جامعه در دسترس باشد. در این روش ابتدا گامی از قبل تعیین می شود و گام یک عدد صحیح مثبت مانند k است که معمولا آن را از رابطه $k=N/n$ به دست می آورند که در آن N حجم جامعه

و n حجم نمونه است. سپس از بین اعداد صحیح 1 تا k یک عدد به تصادف انتخاب می شود. فرد متناظر با عدد انتخاب شده اولین عضو نمونه است و سایر اعضای نمونه به تعداد گام مورد نظر به نمونه اضافه می شوند. بنابر این در این روش اولین عضو نمونه بصورت تصادفی انتخاب شده و سایر اعضای نمونه وابسته به نفر اول به نمونه وارد می شوند. مثلا فرض کنید بخواهیم از یک جامعه N = 1000 نفره یک نمونه n = 100 تائی به روش سیستماتیک انتخاب کنیم و نیز فرض کنید لیستی از افراد جامعه در دسترس باشد. در این صورت ابتدا گام حرکت را به صورت K=N/n=1000/100=10 تعیین می کنیم. سپس از بین اعداد صحیح 1 تا 10 یک عدد به تصادف انتخاب می کنیم. فرض کنید عدد انتخاب شده 4 باشد. در این صورت فرد ردیف چهارم در لیست افراد جامعه اولین عضو نمونه است. سایر اعضای نمونه افراد ردیفهای 14=10+ 4 ، 24 =10+ 14 ، 34=10+24 ، ... ، 994=10+984 هستند.

نمونه گیری طبقه ای

از این روش هنگامی استفاده می شود که افراد جامعه به گروههای مجزا از یکدیگر افراز شده باشند و این گروه بندی به نحوی باشد که افرادی که داخل هر گروه قرار می گیرند متجانس اما گروههای مختلف نا متجانس باشند. به عبارت دیگر واریانس یا پراکندگی در داخل هر گروه کم اما بین گروهها زیاد باشد. چنین گروههائی را طبقه گویند. در روش طبقه ای نمونه به نحوی انتخاب می شود که از هر طبقه ای به نسبت حجم آن طبقه تعدادی در نمونه داشته باشیم. مثلا فرض کنید بخواهیم توسط نمونه گیری از یک شهر بزرگ درآمد ماهانه خانوارهای آن شهر را برآورد کنیم و نیز فرض کنید خانوارهای این شهر را بتوان از لحاظ درآمد به گروههای کم درآمد، درآمد متوسط ، درآمد بالا و درآمد خیلی بالا تقسیم کرد. در این صورت هر کدام از این گروههای درآمدی یک طبقه را تشکیل می دهند چون خانوارهائی که در این گروهها قرار میگیرند از لحاظ درآمدی شبیه به یکدیگر بوده اما خانوارهای گروههای مختلف از لحاظ درآمدی متفاوت از یکدیگر هستند. به عبارت دیگر واریانس یا پراکندگی درآمد در بین خانوارهای هر کدام از این گروهها کم اما بین گروهها زیاد است. بنابر این این گروهها را طبقه گفته و برای نمونه گیری از روش نمونه گیری طبقه ای استفاده می کنیم. به این ترتیب که از همه این گروههای درآمدی یا همان طبقات نمونه می گیریم ، اما از هر طبقه ای به نسبت حجم آن طبقه نمونه می گیریم. مثلا اگر طبقه افراد با درآمد پایین 20 درصد از افراد جامعه را تشکیل دهند در این صورت 20 درصد از حجم نمونه را به این طبقه اختصاص می دهیم و به همین ترتیب الی آخر.

نمونه گیری خوشه ای

از این روش نیز هنگامی استفاده می کنیم که افراد جامعه به گروه های مجزا از یکدیگر افراز شده باشند با این تفاوت که در این حالت گروه بندی به نحوی است که افرادی که داخل هر گروه قرار می گیرند نامتجانس اما گروه های مختلف متجانس هستند. به عبارت دیگر واریانس یا پراکندگی در داخل هر گروه زیاد اما بین گروهها کم است. یعنی در واقع هر کدام از این گروه ها را می توان نمونه کوچک شده ای از کل جامعه در نظر گرفت که همه خواص جامعه را شامل است. چنین گروه هایی را خوشه می نامیم. در نمونه گیری خوشه ای تعدادی از خوشه ها به تصادف انتخاب شده و آنها را سرشماری می کنند. و یا در صورت بزرگ بودن خوشه ها پس از انتخاب خوشه های نمونه از آنها یک نمونه تصادفی انتخاب می شود (خوشه ای دو مرحله ای) و یا ممکن است پس از انتخاب خوشه های نمونه از هر کدام از این خوشه ها نمونه به روش طبقه ای انتخاب کرد و یا حتی مجددا خوشه های نمونه را خوشه بندی کرده و از آنها به روش خوشه ای نمونه گرفت (خوشه ای چند مرحله ای) . در هر صورت در روش خوشه ای از همه خوشه های جامعه نمونه گرفته نمی شود بلکه تنها از تعدادی از آنها نمونه گرفته می شود.

در مسائل عملی و کاربردی ممکن است از یکی از این روشها و یا از ترکیبی از آنها استفاده شود.

مثلا فرض کنید در یک تحقیق هدف تعیین میزان تاثیر برگزاری دوره های آموزشی برای کارشناسان کنترل کیفیت صنایع مختلف باشد. به عبارت دیگر می خواهیم مشخص کنیم که آیا دوره های آموزشی برگزار شده توسط سازمان صنایع برای کارشناسان کنترل کیفیت کارخانجات تولیدی در افزایش معلومات و بهبود کارکرد آنها موثر بوده است یا خیر؟ فرض کنید واحدهای تولیدی در پنج صنعت مختلف صنایع غذائی، صنایع نساجی، قطعات خودرو، سیمان و صنایع مبلمان مشغول فعالیت باشند. در این تحقیق جامعه آماری مجموعه کلیه کارشناسان کنترل کیفیت واحدهای تولیدی این پنج صنعت است. اگر بپذیریم که چگونگی کنترل کیفیت (توزیع متغیر تحت مطالعه) در صنایع مختلف متفاوت از یکدیگر است آنگاه هر کدام از این صنایع تشکیل یک طبقه می دهند و در اولین گام برای نمونه گیری باید از روش طبقه ای استفاده کرد. بدین ترتیب که از واحدهای تولیدی هر کدام از این صنایع باید نمونه گرفت و حجم نمونه اختصاص داده شده به هر صنعت باید متناسب با حجم واحدهای تولیدی آن صنعت باشد. در هر کدام از این صنایع تعدادی واحد تولیدی مشغول به فعالیت هستند. اگر بپذیریم که در هر صنعتی واحدهای تولیدی مختلف از لحاظ وضعیت کنترل کیفیت شبیه به هم هستند (توزیع متغیر تحت مطالعه در این واحدها یکسان است) در اینصورت هر کدام از واحدهای تولیدی در هر صنعتی تشکیل یک خوشه می دهد و بنابراین برای نمونه گیری از این

واحدها نیازی به اخذ نمونه از همه آنها نیست بلکه کافی است تعدادی از این واحدها را به عنوان خوشه های نمونه انتخاب کرده و آن خوشه ها را سرشماری کرد. یعنی کارشناسان کنترل کیفیت آن واحدها را مورد بررسی و مطالعه قرار داد. در برخی موارد حجم خوشه های انتخابی در نمونه زیاد است که در چنین صورتی پس از انتخاب خوشه های نمونه از هر خوشه ای یک نمونه تصادفی گرفته می شود. بر این اساس در این تحقیق با توجه به توزیع متغیر تحت مطالعه برای نمونه گیری از این جامعه آماری در ابتدا جامعه طبقه بندی شده و نمونه به روش طبقه ای انتخاب شده است. سپس در هر طبقه ای واحدها تشکیل خوشه داده اند و بنابر این در هر طبقه ای برای نمونه گیری از روش خوشه ای استفاده شده است. اگر تعداد افراد در هر خوشه (کارشناسان کنترل کیفیت در هر واحد انتخابی در نمونه) کم باشد خوشه ها سرشماری می شوند(همه کارشناسان کنترل کیفیت در واحدهای تولیدی انتخاب شده در نمونه مورد بررسی و مطالعه قرار می گیرند) اما اگر حجم خوشه ها زیاد باشد پس از انتخاب خوشه های نمونه از هر کدام از آنها یک نمونه تصادفی گرفته می شود.

پارامتر (Parameter**)**

هر ویژگی عددی **یک جامعه** را یک « *پارامتر* » گوییم. پارامترها مقادیر ثابتی هستند که در عمل همیشه مجهولند. مثلاً میانگین و واریانس یک جامعه (μ و σ^2) پارامترند (برای جامعه ثابت می باشند).

آماره (Statistic)

هر ویژگی عددی **یک نمونه** از یک جامعه را « آماره » می گویند. مثلاً میانگین و واریانس نمونه ($\bar{x}, S^2$) آماره هستند. آماره ها برخلاف پارامترها ثابت نبوده بلکه متغیر تصادفی هستند.

روش های برآورد پارامترها

بطورکلی پارامترها را به دو روش *نقطه ای* و *فاصله ای* برآورد می کنند. در روش نقطه ای، پارامتر جامعه توسط آماره ای در نمونه برآورد میشود. مثلاً میانگین جامعه توسط **میانگین نمونه** برآورد می شود. اما در روش فاصله ای که آنرا **فاصله اطمینان** نیز می نامند، پارامتر جامعه توسط یک فاصله برآورد می شود. فاصله ای که با ضریب اطمینان خوبی پارامتر جامعه را در بر می گیرد.

آزمون فرض

یکی دیگر از روشهای برآورد و تعیین پارامترها استفاده از آزمون فرض است. در آزمون فرض از قبل یک ادعا یا حدس یا گمانی بر روی پارامتر یا پارامترهای جامعه وجود دارد که می خواهیم توسط داده های یک نمونه تصادفی از آن جامعه آن ادعا را مورد بررسی قرار دهیم.

فرض صفر یا H_0 و فرض خلاف (جانشین H_1 یا H_A):

فرضیه ای که قرار است مورد آزمون قرار گیرد و معمولاً بیانگر **عدم وجود اختلاف** یا ارتباط بین پارامترها است را **فرض صفر** گفته و آن را با H_0 نشان می دهند. نقیض فرض صفر را ***فرض خلاف*** یا ***فرض جانشین*** گفته و آن را با H_1 یا H_A نشان می دهیم.

خطا ها:

در هر آزمون فرضیه همواره *فرض صفر آزمون* می شود و این فرضیه گزاره ای است که ارزش آن درست یا نادرست است. از طرفی مطابق با داده های نمونه و روشی که در آزمون فرضیه به کار میبریم ما فرضیه را رد کرده یا آن را رد نمی کنیم. لذا چهار حالت ممکن است بوجود آید به شرح زیر.

1) فرض صفر (H_0) درست است و آن را می پذیریم.
2) فرض صفر (H_0) درست است و آن را رد می کنیم.
3) فرض صفر (H_0) نادرست است و آن را می پذیریم.
4) فرض صفر (H_0) نادرست است و آن را رد می کنیم.

واضح است که از این چهار حالت ممکن فوق در حالات 2 و 3 دچار خطا شده ایم. حالت 2 را ***خطای نوع اول*** و حالت 3 را ***خطای نوع دوم*** گویند. به عبارت دیگر:

خطای نوع اول :

خطای ناشی از رد کردن فرض صفر، وقتی که این فرض درست است.

خطای نوع دوم:

خطای ناشی از پذیرفتن فرض صفر وقتی این فرض غلط است.

سطح خطای آزمون (سطح معنی داری آزمون):

α: احتمال رخ دادن خطای نوع اول را سطح خطا یا سطح معنی داری آزمون گفته و آن را با α نمایش می دهیم.

β : احتمال رخ دادن خطای نوع دوم را با β نشان می دهند.

توجه : بین α و β رابطه معکوس برقرار بوده اما حاصل جمع آن ها لزوماً مساوی یک نیست. یعنی این دو پیش آمد لزوماً پیشامدهای مکمل یکدیگر نیستند. در آمار α را از قبل ثابت در نظر گرفته و برای آن معمولاً مقادیر 0/01 یا 0/05 را در نظر می گیرند که با ثابت بودن α رابطه ای معکوس بین حجم نمونه و β برقرار است و لذا یکی از راه های کاهش β افزایش حجم نمونه است.

توان آزمون:

احتمال رد کردن فرض صفر وقتی این فرض غلط است را « توان آزمون » گفته و معمولاً آن را با π (عدد پی) نشان می دهند.

توجه: بین توان آزمون و β رابطه ای معکوس برقرار بوده و حاصل جمع آنها مساوی یک است.

$$\pi + \beta = 1 \longrightarrow \text{توان } \pi = 1 - \beta$$

از آنجایی که بین حجم نمونه و β رابطه ای معکوس برقرار است، لذا بین حجم نمونه و توان آزمون رابطه ای مستقیم برقرار است. یعنی با افزایش حجم نمونه β کاهش یافته و توان آزمون افزایش می یابد.

آماره آزمون:

ملاک یا معیار ی است که از روی داده های نمونه محاسبه شده است و توسط آن در مورد رد کردن یا رد نکردن فرض صفر تصمیم گیری می شود.

ناحیه رد (ناحیه بحرانی):

ناحیه ای است در دامنه توزیع آماره آزمون بطوریکه هرگاه مقدار آماره در آن ناحیه قرار گیرد فرض صفر رد می شود.

توجه: روش سنتی در انجام آزمون فرضیه ها استفاده از آماره آزمون و تعیین ناحیه بحرانی است. بدین ترتیب که ابتدا آماره آزمون توسط فرمول مربوطه محاسبه شده و سپس مقدار آن با مقدار بحرانی جدول توزیع احتمال مربوط به آن مقایسه شده و بنابراینکه مقدار آماره در ناحیه بحرانی قرار گیرد یا خیر در مورد آن فرضیه تصمیم گیری می شود. اما در استفاده از نرم افزار ها در انجام آزمون های آماری اگرچه این نرم افزار ها آماره آزمون را محاسبه کرده و در خروجی ارائه میدهند اما نیازی به این آماره و تعیین ناحیه بحرانی نیست. چون این نرم افزارها شاخصی با نام P.Value را در خروجی ارایه می دهند که تصمیم گیری در مورد فرضیه توسط این شاخص انجام می گیرد.

:P-Value

P-Value که آن را P–مقدار یا مقدار احتمال نیز می نامند، عددی است بین صفر و یک که هر چقدر این عدد بزرگتر باشد به معنای آن است که فرض صفر درست بوده و نباید آن را رد کرد و هر چقدر این عدد کوچک باشد به معنای نادرست بودن فرض صفر بوده و لذا باید آن فرضیه را رد کرد. در عمل P.Value را با سطح خطای آزمون (α) مقایسه می کنند.

اگر P.Value کمتر از α شد ($P.Value<\alpha$) فرض صفر را رد کرده و در غیر این صورت آن را رد نمی کنیم.

توجه: اگر فرضیه ای در سطح خطای 5٪ پذیرفته شود (P.Value>0/05) آنگاه P-value ی آن از 0/05 بزرگتر بوده و لذا از 0/01 نیز بزرگتر خواهد بود و در نتیجه این فرضیه در سطح خطای 0/01 نیز پذیرفته خواهد شد. اگر فرضیه ای در سطح خطای 1٪ رد شود(P.Value<0/01) آنگاه P-value ی آن از 0/01 کوچکتر است و لذا از 0/05 نیز کوچکتر خواهد بود و در نتیجه این فرضیه در سطح خطای 0/05 نیز رد خواهد شد. اما اگر فرضیه ای در سطح خطای 5٪ رد شود در سطح خطای 1٪ نتیجه ای از آن بدست نمی آید، و به طور مشابه اگر فرضیه ای در سطح خطای 1٪ پذیرفته شود آنگاه در سطح خطای 5٪ نتیجه ای از آن بدست نمی آید.

آزمون تساوی میانگین با یک عدد ثابت:

فرض کنید μ میانگین یک جامعه نرمال و μ_0 یک مقدار ثابت باشد، می خواهیم فرضیه $H_0:\mu=\mu_0$ را آزمون کنیم. بدین منظور از مسیر زیر استفاده می کنیم:

Analyze ⟶ Compare Mean ⟶ One Sample T test

مثلاً فرض کنید بخواهیم میانگین نمرات درس ریاضی دانش آموزان را با عدد 12 مقایسه کنیم یعنی در واقع می خواهیم فرضیه $H_0: \mu= 12$ که در آن μ بیانگر میانگین نمرات درس ریاضی دانش آموزان است را آزمون کنیم.

بدین منظور پس از رفتن به مسیر فوق متغیر مربوط به نمرات درس ریاضی دانش آموزان (X_4) را به قسمت Test Variable برده و عدد ثابت و یا همان 12 را در قسمت Test Value وارد می کنیم و فرمان را اجرا می کنیم. پس از اجرای این فرمان نتیجه در قالب دو جدول و در خروجی خواهد آمد که این جداول در زیر آمده است.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
MATH.	1311	12.9605	4.87145	.13454

One-Sample Test

	Test Value = 12					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
MATH.	7.139	1310	.000	.96053	.6966	1.2245

در اولین جدول گزارشی از برخی شاخص های آماری این متغیر یعنی نمرات درس ریاضی دانش آموزان آمده است. سپس در دومین جدول نتیجه انجام این آزمون آمده است. در این جدول ستونهای اول و دوم با عناوین t و df به ترتیب مقدار آماره آزمون (آماره t) و درجه آزادی آن است. سومین ستون این جدول با نام Sig (2-Tailed) همان P.Value ی این آزمون است که در

این مثال همانطور که مشاهده می شود P.Value خیلی کوچک بوده و دست کم تا سه رقم اعشار صفر است. بنابراین فرض صفر مبنی بر یکسان بودن این میانگین با عدد 12 در سطح خطای 0/01 (و در نتیجه در سطح خطای 0/05) رد می شود.

دو ستون آخر این جدول با عناوین Lower و Upper به ترتیب کران های پایین و بالای یک فاصله اطمینان 95٪ برای اختلاف میانگین با عدد μ_0 یعنی $\mu - \mu_0$ است. در این فاصله:

1) اگر هر دو کران فاصله مثبت باشد این به آن معنی است که با 95٪ اطمینان $\mu - \mu_0 > 0$ (مثبت) است. یعنی $\mu > \mu_0$ می باشد. بنابر این فرضیه $H_0: \mu = \mu_0$ رد شده و به جای آن فرضیه $H_1: \mu > \mu_0$ پذیرفته می شود.
2) اگر هر دو کران فاصله منفی باشد این به آن معنی است که با 95٪ اطمینان $\mu - \mu_0 < 0$ (منفی) است، یعنی $\mu < \mu_0$ است. بنابر این فرضیه $H_0: \mu = \mu_0$ رد شده و به جای آن فرضیه $H_1: \mu < \mu_0$ پذیرفته می شود.
3) اگر کران پایین فاصله منفی و کران بالای آن مثبت باشد، یعنی صفر در این فاصله قرار گیرد، این به آن معنی است که با 95٪ اطمینان $\mu - \mu_0 = 0$ است. یعنی $\mu = \mu_0$ است. بنابر این فرضیه $H_0: \mu = \mu_0$ پذیرفته می شود.

$$\mu - \mu_0 \in (a, b) \Rightarrow \mu - \mu_0 > 0 \rightarrow \mu > \mu_0$$
$$\Rightarrow \mu - \mu_0 < 0 \rightarrow \mu < \mu_0$$
$$\Rightarrow \mu - \mu_0 \rightarrow \mu = \mu_0$$

(a, b) =کرانه بالا و پایین

در این مثال همانطور که مشاهده می شود هر دو کران این فاصله مثبت است و بنابر این با 95٪ اطمینان فرضیه $H_0: \mu = 12$ رد شده و به جای آن فرضیه $H_1: \mu > 12$ پذیرفته می شود. یعنی میانگین نمره ریاضی دانش آموزان بطور معنی داری از 12 بیشتر است. از طرفی چون کران پایین این فاصله حدود 0/7 ~ 0/69 و کران بالای آن حدود 1/2 ~ 1/22 است . لذا می توان نتیجه گرفت که با 95٪ اطمینان اختلاف این میانگین با عدد 12 حداقل 0/7 و حداکثر 1/2 است. یا به عبارت دیگر با 95٪ اطمینان میانگین نمره ریاضی دانش آموزان حداقل 0/7 و حداکثر 1/2 از 12 بیشتر است یعنی این میانگین با 95 درصد اطمینان بین دو عدد 12/7 تا 13/2 قرار دارد.

انتخاب گروهی خاص در نمونه :

در برخی موارد ممکن است بخواهیم یک آزمون یا تحلیل بخصوصی را تنها بر روی گروهی خاص در نمونه انجام دهیم. بدین منظور از فرمان Select Case در مسیر Data ⟶ Select Case همچنین در برخی موارد ممکن است بخواهیم آزمون یا تحلیل بخصوصی را بر روی گروه های مختلف بطور مجزا انجام دهیم در چنین صورتی از Split کردن فایل در مسیر Data ⟶ Split استفاده می کنیم.

تمرین

میانگین نمرات دروس مختلف را با عدد μ_0 مقایسه کنید. μ_0 را به نحوی انتخاب کنید که یکبار فرضیه رد شده و به جای آن فرضیه $H_1: \mu > \mu_0$ پذیرفته شود. یکبار فرضیه رد شده و به جای آن فرضیه $H_1 : \mu<\mu_0$ پذیرفته شود و بار دیگر نیز فرضیه $H_0: \mu = \mu_0$ پذیرفته شود. این تمرین را یکبار بر روی کل نمونه انجام دهید. یکبار به تفکیک جنسیت، بار دیگر به تفکیک پایه تحصیلی. یکبار به تفکیک جنسیت و منطقه و یکبار نیز به تفکیک جنسیت و منطقه و پایه تحصیلی انجام دهید. در انجام این تمرین یکبار از فرمان Select Case استفاده کرده و یکبار از Split کردن فایل استفاده کنید.

آزمون تساوی دو میانگین در نمونه های مستقل:

فرض کنید μ_1 و σ_1^2 میانگین و واریانس یک جامعه نرمال و μ_2 و σ_2^2 میانگین و واریانس یک جامعه نرمال دیگر باشند. می خواهیم فرضیه یکسان بودن میانگین های این دو جمعیت یعنی $H_0 : \mu_1 = \mu_2$ را آزمون کنیم. قبل از آزمون این فرضیه لازم است فرض یکسان بودن واریانس های این دو جمعیت یعنی فرضیه $H_0 :\sigma_1^2=\sigma_2^2$ آزمون شود. چون بنابر این که واریانس های این دو جمعیت یکسان بوده یا یکسان نباشد، روش انجام آزمون تساوی دو میانگین کمی متفاوت است، که البته SPSS بطور خودکار قبل از آزمون تساوی دو میانگین فرضیه یکسان بودن واریانس ها را آزمون می کند. برای انجام آزمون تساوی دو میانگین در نمونه های مستقل از مسیر زیر استفاده می کنیم.

Analyze ⟶ Compare means ⟶ Indepent – Samples T Test

مثلاً فرض کنید بخواهید نمره ریاضی دانش آموزان دختر و پسر را با یکدیگر مقایسه کنیم، یعنی در واقع می خواهیم فرضیه $H_0 : \mu_1 = \mu_2$ که در آن μ_1 و μ_2 به ترتیب بیانگر میانگین نمره ریاضی دانش آموزان دختر و پسر است را آزمون کنیم. به

این منظور پس از رفتن به مسیر فوق متغیر مربوط به نمره ریاضی دانش آموزان (X_4) را به قسمت Test Variable برده و متغیر جنسیت را به قسمت Grouping Variable برده و کد مربوط به جنسیت دختر و پسر را برای آن تعریف می کنیم.

پس از اجرای این فرمان نتیجه در قالب دو جدول و در خروجی خواهد آمد که این جداول در زیر آمده است. متذکر می شویم که در حالت معمولی جدول دوم در خروجی به صورت سطری (افقی) نشان داده خواهد شد که به جهت کمبود جا در اینجا جای سطرها و ستونهای آن عوض شده و به صورت ستونی (عمودی) نشان داده شده است.

Group Statistics

	SEX	N	Mean	Std. Deviation	Std. Error Mean
MATH.	GIRL	646	14.8820	4.73562	.18632
	BOY	665	11.0940	4.23714	.16431

Independent Samples Test

			MATH.	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		37.154	
	Sig.		.000	
t-test for Equality of Means	T		15.273	15.248
	Df		1309	1283.981
	Sig. (2-tailed)		.000	.000
	Mean Difference		3.78798	3.78798
	Std. Error Difference		.24802	.24842
	95% Confidence Interval of the Difference	Lower	3.30142	3.30063
		Upper	4.27454	4.27534

در اولین جدول گزارش از برخی از شاخص های آماری نمرات درس ریاضی به تفکیک جنسیت آمده است. در دومین جدول نتیجه آزمون تساوی دو واریانس و دو میانگین آمده است. دو سطر اول این جدول با عنوان Levene's Test for Equality of variances نتیجه آزمون تساوی دو واریانس یا همان آزمون فرضیه $H_0: \sigma_1^2=\sigma_2^2$ است که دومین سطر آن با عنوان Sig. همان P-value ی این آزمون است و در این مثال طوری که مشاهده می شود P.value ی این آزمون خیلی کوچک بوده و دسته کم

تا سه رقم اعشاری است. بنابر این این فرض یکسان بودن واریانس های دو جمعیت در سطح خطای 0/01 (1٪) و در نتیجه سطح خطای 5٪ رد می شود. بر این اساس واریانس یا پراکندگی در توزیع نرمال نمرات درس ریاضی دانش آموزان دختر و پسر یکسان نبوده و بطور معنی داری متفاوت از یکدیگر است. ادامه این جدول نتیجه آزمون تساوی دو میانگین یا همان آزمون فرضیه $H_0 : \mu_1 = \mu_2$ است که در دو ستون آمده است.

ستون اول با فرض یکسان بودن واریانس ها و ستون دوم با فرض عدم تساوی واریانس ها است. به عبارت دیگر اگر فرض یکسان بودن واریانس ها پذیرفته شود در ادامه این جدول از اطلاعات ستون اول آن استفاده می کنیم و در غیر این صورت از اطلاعات ردیف دوم آن استفاده می کنیم. در این مثال با توجه به اینکه فرض یکسان بودن واریانس ها رد شد، لذا در ادامه این جدول از اطلاعات ستون دوم باید استفاده کنیم. در ادامه این جدول ردیفهای اول و دوم با عناوین t و df به ترتیب آماره آزمون (آماره t) و درجه آزادی آن است که اولین آن با فرض یکسان بودن واریانسها و دومی با فرض عدم تساوی واریانسها است. سومین ردیف در این قسمت از جدول با عنوان Sig.(2-Taied) همان P-valueی آزمون تساوی دو میانگین است که در این مثال همانطور که مشاهده می شود مقدار آن در هر دو حالت یکسان بودن یا یکسان نبودن واریانسها خیلی کوچک بوده و تا سه رقم اعشار صفر است. بنابر این، فرض یکسان بودن میانگین های این دو جمعیت در سطح خطای 1٪ (0/01) و در نتیجه در سطح خطای 5 درصد رد می شود. به عبارت دیگر میانگین نمره ریاضی دخترها و پسرها یکسان نبوده بلکه به طور کاملا معنی داری متفاوت از یکدیگر است.

در انتهای این جدول و آخرین دو سطر آن با عناوین Lower و Upper به ترتیب کران های و پایین و بالای یک فاصله اطمینان 95٪ برای اختلاف این دو میانگین یعنی μ_1-μ_2 آمده است در این فاصله :

1) اگر هر دو کران فاصله مثبت باشند این به آن معنی است که با 95٪ اطمینان μ_1-μ_2 >0 است یعنی $\mu_1 > \mu_2$ است و بنابر این فرضیه $H_0 : \mu_1 = \mu_2$ رد شده و به جای آن فرضیه $H_1: \mu_1 > \mu_2$ پذیرفته می شود.
2) اگر هر دو کران فاصله منفی باشد این به آن معنا است که با 95٪ اطمینان $\mu_1 - \mu_2 < 0$ است. یعنی $\mu_1 < \mu_2$ است و بنابرا ین فرضیه $H_0 : \mu_1 = \mu_2$ رد شده و به جای آن فرضیه $H_1: \mu_1 < \mu_2$ پذیرفته می شود.
3) اگر کران پایین فاصله منفی و کران بالای آن مثبت باشد یعنی صفر در این فاصله قرار گیرد این به آن معنی است که با 95٪ اطمینان $\mu_1 = \mu_2 = 0$ است. یعنی $\mu_1 = \mu_2$ و بنابر این فرضیه $H_0 : \mu_1 = \mu_2$ پذیرفته می شود.

$$\mu_1 - \mu_2 \in (a^+, b^+) \longrightarrow \mu_1 - \mu_2 > 0 \longrightarrow \mu_1 > \mu_2$$

$$\mu_1 - \mu_2 \in (a^-, b^-) \longrightarrow \mu_1 - \mu_2 < 0 \longrightarrow \mu_1 < \mu_2$$

$$\mu_1 - \mu_2 \in (a^+, b^-) \longrightarrow \mu_1 - \mu_2 = 0 \longrightarrow \mu_1 = \mu_2$$

در این مثال همانطور که مشاهده می شود هر دو کران فاصله مثبت است و بنابر این فرض صفر مبنی بر یکسان بودن میانگین های این دو جمعیت رد شده و به جای آن فرضیه $H_1 : \mu_1 > \mu_2$ پذیرفته می شود. یعنی میانگین نمره ریاضی دانش آموزان دختر به طور معنی داری بیشتر از میانگین نمره ریاضی دانش آموزان پسر است. از طرفی چون کران پایین این فاصله 3/3 و کران بالای آن تقریباً 4/3 است. لذا با 95٪ اطمینان می توان نتیجه گرفت که اختلاف میانگین نمره ریاضی دخترها و پسرها حداقل 3/3 و حداکثر4/3 است یا به عبارت دیگر با 95٪ اطمینان میانگین نمره ریاضی دخترها حداقل 3/3 و حداکثر 4/3 از میانگین نمره ریاضی پسرها بیشتر است.

تمرین

میانگین نمرات دروس مختلف دانش آموزان را در گروههای (دختر–پسر)، (شمال–جنوب)، (دختر شمال– دختر جنوب)، (پسر شمال–پسر جنوب)، (دختر شمال–پسر شمال)، (دختر جنوب–پسر جنوب) با یکدیگر مقایسه کرده و در هر حالتی نتیجه به دست آمده را به طور کامل تشریح و تفسیر کنید.

آزمون تساوی دو میانگین در نمونه های همبسته (جفت شده، زوج شده، طرح پیش آزمون–پس آزمون) :

در آزمون قبل، میانگین های دو جمعیت از طریق دو نمونهٔ تصادفی مستقل از هم از آن دو جمعیت با یکدیگر مقایسه شدند. اما اگر بخواهیم میانگین های دو جمعیت را از طریق دو نمونه وابسته از آن دو جمعیت مورد بررسی و آزمون قرار دهیم از آزمون قبل نمی توان استفاده کرد. بطور خلاصه وقتی که نمونه ها همبسته است از آزمون قبل نمی توان استفاده نمود.

منظور از دو نمونه همبسته نمونه هایی است که در آن اعضای دو نمونه از لحاظ فیزیکی یکی هستند و روی هر فرد نمونه دو ویژگی اندازه گیری شده است. . به خصوص از این آزمون برای بررسی تاثیر یک متغیر مستقل یا آزمایشی بر روی یک متغیر وابسته استفاده می شود. همچنین از این آزمون در طرحهای پیش آزمون، پس آزمون استفاده می شود.

مثلاً فرض کنید بخواهیم تاثیر یک برنامه تمرینی خاص را در بهبود انعطاف پذیری دانش آموزان مورد بررسی قرار دهیم. در این صورت یک نمونهٔ تصادفی از دانش آموزان را انتخاب کرده و میزان انعطاف پذیری آنان را اندازه گیری می کنیم، سپس برنامه تمرینی را اجرا کرده و پس از آن مجددا انعطاف پذیری این دانش آموزان را اندازه گیری می کنیم. در چنین صورتی دو نمونه داریم که اعضای این دو نمونه از لحاظ فیزیکی یکی هستند. یکی از این نمونه ها میزان انعطاف پذیری قبل از برنامه تمرینی و دیگری میزان انعطاف پذیری پس از برنامه تمرینی است. این دو نمونه را نمونه های **همبسته** یا **وابسته** می گویند.

البته در بسیاری از موارد عملاً امکان اینکه بتوان هر دو اندازه گیری را روی یک نمونه بکار برد نیست. در چنین مواردی از نمونه های جفت شده یا زوج شده استفاده می شود. مثلاً فرض کنید بخواهیم تاثیر دو برنامه تمرینی مختلف بر روی انعطاف پذیری دانش آموزان را مورد بررسی قرار دهیم. در اینصورت از لحاظ تئوری باید یک نمونهٔ تصادفی انتخاب کرده و هر دو برنامهٔ تمرینی را بر روی این نمونه پیاده کرده و سپس نتایج را با یکدیگر مقایسه کرد. اما واضح است که عملاً امکان اینکه بتوان هر دو نوع اندازه گیری را روی یک نمونه بکار برد نیست. چون برنامهٔ تمرینی اول بر روی نتیجهٔ برنامه دوم اثر می گذارد. در چنین مواردی از نمونه های جفت شده یا زوج شده استفاده می کنیم. بدین ترتیب که دو نفر دانش آموز که از نظر ویژگی های اثر گذار در انعطاف پذیری شبیه یکدیگر هستند را انتخاب کنید (جنسیت یکسان، سن یکسان و . . .) و این دو نفر را یک جفت یا یک زوج می نامیم. سپس یکی از آنها را در برنامه تمرینی « الف » و دیگری را در برنامهٔ تمرینی« ب » قرار می دهیم این عمل زوج یابی را آنقدر تکرار می کنیم تا نمونه نهایی تکمیل گردد. چنین نمونه هایی را نمونه های جفت شده یا زوج شده می نامیم. که نمونه های جفت شده نیز نمونه های همبسته هستند.

فرض کنید μ_1 و μ_2 میانگین های دو جمعیت باشند و می خواهیم این میانگین ها را از طریق دو نمونهٔ همبسته از آن دو جمعیت با یکدیگر مقایسه کنیم، یعنی می خواهیم فرضیه $H_0 : \mu_1 = \mu_2$ را از طریق دو نمونه همبسته مورد آزمون قرار دهیم، بدین منظور از مسیر زیر استفاده می کنیم:

Analyze ⟶ Compare Means ⟶ Parled- Samples Test

مثلاً فرض کنیم بخواهیم میانگین نمره ریاضی و علوم دانش آموزان را با یکدیگر مقایسه کنیم، یعنی می خواهیم فرضیه $H_0 : \mu_1 = \mu_2$ که در آن μ_1 و μ_2 به ترتیب بیانگر میانگین نمره ریاضی و علوم دانش آموزان است را آزمون کنیم. بدین منظور پس از رفتن به مسیر فوق متغیر مربوط به نمرات دروس ریاضی و علوم را به عنوان یک جفت به قسمت Parled Variable برده و فرمان را اجرا می کنیم. پس از اجرای این فرمان نتیجه در قالب سه جدول و در خروجی خواهد آمد. اولین جدول گزارشی از برخی شاخص های آماری نمرات دروس ریاضی و علوم دانش آموزان است. جدول دوم ضریب همبستگی بین نمرات دروس ریاضی و علوم که راجع به ضریب همبستگی بعداً و بطور مفصل بحث خواهد شد. سومین جدول نتیجه انجام آزمون فرضیه فوق است. جداول اول و سوم در زیر آمده است. متذکر می شویم که در حالت معمولی جدول دوم در خروجی به صورت سطری (افقی) نشان داده خواهد شد که به جهت کمبود جا در اینجا جای سطرها و ستونهای آن عوض شده و به صورت ستونی (عمودی) نشان داده شده است.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MATH.	12.9818	1307	4.86115	.13446
	SCEIN.	13.5145	1307	4.23704	.11720

Paired Samples Test

			Pair 1
			MATH. - SCEIN.
Paired Differences	Mean		-.53271
	Std. Deviation		3.10102
	Std. Error Mean		.08578
	95% Confidence Interval of the Difference	Lower	-.70098
		Upper	-.36443
T			-6.210
Df			1306
Sig. (2-tailed)			.000

در این جدول سطرهای اول، دوم و سوم به ترتیب میانگین، انحراف معیار و خطای استاندارد میانگین اختلاف هر زوج داده است. در قسمت دوم جدول سطرهای اول و دوم جدول با عناوین t و df به ترتیب آماره t و درجه آزادی آن است. آخرین سطر این جدول با عنوان Sig. (2-tailed) همان P.Value ی این آزمون است که در این مثال همانطور که مشاهده می شود P.Value خیلی کوچک بوده و دست کم تا سه رقم اعشاری صفر است. بنابر این فرض صفر مبنی بر یکسان بودن میانگین های این دو جمعیت در سطح خطای 0/01 (و در نتیجه در سطح خطای 0/05) رد می شود.

دو آخر در قسمت اول این جدول با عناوین Lower و Upper به ترتیب کران های پایین و بالای یک فاصله اطمینان 95٪ برای اختلاف این دو میانگین یعنی $\mu_1 - \mu_2$ است، که در این مثال کران های فاصله منفی است. بنابراین فرضیه

$H_0 : \mu_1 = \mu_2$ رد شده و به جای آن فرضیه $H_1 : \mu_1 < \mu_2$ پذیرفته می شود. یعنی میانگین ریاضی دانش آموزان بطور معنی داری کمتر از میانگین نمره علوم آن ها است، و با 95٪ اطمینان اختلاف این دو میانگین حداقل 0/26 و حداکثر 0/7 می باشد.

تمرین

میانگین نمرات دروس مختلف را دو به دو با یکدیگر مقایسه کنید. این عمل را یکبار روی کل نمونه، یکبار به تفکیک جنسیت، یکبار به تفکیک منطقه، یکبار به تفکیک جنسیت و منطقه و یکبار به تفکیک جنسیت و منطقه و پایه تحصیلی انجام دهید.

تحلیل واریانس یا آنالیز واریانس یک طرفه

همانطور که از قبل می دانیم برای مقایسه میانگین های دو جمعیت از طریق دو نمونه تصادفی مستقل از هم از آن دو جمعیت از آزمون T دو نمونه ای استفاده می کنیم. اما اگر تعداد جمعیت ها از دو تا بیشتر شود یک روش آن است که این جمعیت ها را توسط همان آزمون t و به صورت دو به دو با یکدیگر مقایسه کرد، که در چنین صورتی برای مقایسه k جمعیت به صورت دو به دو با یکدیگر باید تعداد $\binom{k}{2} = \frac{k!}{2!(k-2)!}$ آزمون t انجام داد. مثلاً اگر بخواهیم سه جمعیت را به صورت دو به دو با یکدیگر

مقایسه کنیم در این صورت باید تعداد $\binom{3}{2} = 3$ بار از آزمون t استفاده کنیم و اگر این جمعیت ها چهار تا شود، برای مقایسه دو به دوی آنها باید تعداد $\binom{4}{2} = 6$ آزمون t انجام داد. چنین عملی باعث افزایش خطای نوع اول می شود.

فرض کنید بخواهیم میانگینهای k جمعیت را به صورت دو به با یکدیگر مقایسه کنیم. در اینصورت تعداد $\binom{k}{2}$ آزمون باید انجام دهیم. اگر خطای نوع اول در هر آزمون را α قرار دهیم، خطای نوع اول کل این آزمونها $1 - (1 - \alpha)^n$ خواهد شد که در آن n تعداد کل آزمونهای انجام شده یعنی همان $\binom{k}{2}$ است. این عدد کمی کمتر از k برابر α بوده و بنابر این خطای نوع اول کل این آزمونها به شدت افزایش می یابد. مثلا فرض کنید بخواهیم میانگینهای سه جمعیت را به صورت دو به با یکدیگر مقایسه کنیم و نیز فرض کنید خطای نوع اول برای هر آزمون را 0.05 تعیین کنیم. چون باید تعداد سه آزمون انجام دهیم خطای نوع اول کل این آزمونها $1 - (1 - 0.5)^3 = 0.14$ خواهد شد یعنی کمی کمتر از سه برابر خطای نوع اول در هر آزمون. به طور مشابه اگر بخواهیم میانگینهای چهار جمعیت را به صورت دو به دو با یکدیگر مقایسه کنیم و نیز فرض کنید خطای نوع اول برای هر آزمون را 0.05 تعیین کنیم. چون باید تعداد $\binom{4}{2} = 6$ آزمون انجام دهیم خطای نوع اول کل این آزمونها $1 - (1 - 0.5)^6 = 0.26$ خواهد شد یعنی کمی کمتر از 6 برابر خطای نوع اول در هر آزمون. به همین دلیل و به جهت ثابت نگاه داشتن میزان خطا روش دیگری با نام تحلیل واریانس یکطرفه ابداع شد که مطابق این روش میانگین های بیش از دو جمعیت نه به صورت دو به دو بلکه به صورت توام با یکدیگر مقایسه می شوند و لذا خطای نوع اول افزایش نیافته بلکه در همان سطح α ثابت می ماند. بنابراین تحلیل واریانس یکطرفه در واقع شکل تعمیم یافته آزمون مقایسه دو میانگین است که در آن میانگینهای بیش از دو جمعیت به صورت توام با یکدیگر مقایسه می شوند.

فرض کنید $\mu_1, \mu_2, \ldots, \mu_k$ میانگین های K جمعیت نرمال باشد. فرضیه مورد آزمون به صورت زیر است.

$H_0: \mu_1 = \mu_2 = \ldots = \mu_K$ و فرض مقابل به صورت زیر است

H_1: میانگین های دست کم 2 جمعیت نابرابرند

برای انجام این آزمون از مسیر زیر استفاده می کنیم.

Analyze ⟶ Compare Means ⟶ One-way ANOVA

مثلاً فرض کنید بخواهیم میانگین نمره ریاضی دانش آموزان پایه های اول، دوم، و سوم را با یکدیگر مقایسه کنیم. یعنی در واقع می خواهیم فرضیه $H_0 : \mu_1 = \mu_2 = \mu_3$ که در آن μ_1، μ_2 و μ_3 به ترتیب بیانگر میانگین نمرهٔ ریاضی دانش آموزان پایه های اول، دوم، و سوم است را آزمون کنیم. بدین منظور پس از رفتن به مسیر فوق متغیر مربوط به نمره ریاضی دانش آموزان (X_4) را به قسمت Dependent List برده و متغیر مربوط به پایه تحصیلی دانش آموزان X_3 را به قسمت factor می بریم. پس از اجرای این فرمان نتیجه در قالب جدولی با نام ANOVA در خروجی داده خواهد شد که این جدول در زیر آمده است.

ANOVA

MATH.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	98.465	2	49.233	2.078	.126
Within Groups	30989.179	1308	23.692		
Total	31087.645	1310			

در این جدول دومین ستون با عنوان Sum of Square مجموع مربعات یا همان SSها است. ردیف اول مجموع مربعات بین گروهها یا همان اثر عامل (SSTr)، ردیف دوم مجموع مربعات داخل گروهها یا همان اثر خطا (SSE) و آخرین ردیف نیز مجموع مربعات کل (SST) است. سومین ستون با عنوان df درجه آزادی، ستون بعدی میانگین مربعات یا همان MS ها، ستون بعدی با عنوان F آماره آزمون یا همان کسر F و آخرین ستون با عنوان Sig نیز همان P.Value ی این آزمون است. در این مثال همانطور که مشاهده می شود P.Value این آزمون 0/126 بوده که چون از 0/05 بزرگتر است این فرضیه در سطح خطای 0/05 (و در نتیجه در سطح خطای 0/01) پذیرفته می شود. یعنی تفاوت معنی داری بین میانگین نمره ریاضی دانش آموزان پایه های اول، دوم، و سوم وجود ندارد. برای مشاهده این میانگینها می توان در همان مسیر و در پنجره One-Way ANOVA روی گزینه Options کلیک کرده و در پنجره باز شده گزینه Descriptive را علامت زد که در چنین صورتی علاوه بر انجام این آزمون شاخصهای میانگین و انحراف معیار این متغیر در هر گروه نیز در خروجی داده خواهد شد.

آزمون های تعقیبی (مقایسه های پس از تجزیه)

در آنالیز واریانس یکطرفه اگر فرض صفر پذیرفته شود(مانند مثال قبل) به معنای آن است که تفاوت معنی داری بین میانگین ها برقرار نیست و بنابر این آزمون خاتمه یافته است. اما اگر فرض صفر رد شود به معنای آن است که میانگین در دست کم دو

جمعیت نا برابر است و بنابر این آزمون خاتمه نیافته است، چون می خواهیم ببینیم میانگین های کدام جمعیت ها متفاوت از یکدیگر است. به همین منظور آزمون هایی با نام **آزمون های تعقیبی** وجود دارد، که تعداد این آزمون ها خیلی زیاد بوده و برخی از آنها در شرایط خاص نسبت به برخی دیگر ارجحیت دارند. اما متداول ترین این آزمون ها، آزمون های **توکی** و **شفه** است. آزمون توکی دارای توان بیشتری نسبت به آزمون شفه است. اما استفاده از این آزمون مستلزم برقرار بودن فرض نرمال بودن توزیع داده ها و نیز یکسان بودن حجم نمونه در گروهها است. یعنی در استفاده از این آزمون توزیع داده ها حتماً باید نرمال باشد و نیز حجم نمونه در گروه های مختلف باید یکسان باشد. اما آزمون شفه نسبت به فرض نرمال بودن توزیع داده ها کمتر حساس بوده و استفاده از آن نیز مستلزم یکسان بودن حجم نمونه در گروه های مختلف نیست و لذا کاربرد این آزمون از آزمون توکی بیشتر است. مسیر انجام آزمون های تعقیبی همان مسیر ANOVA است. در پنجره ANOVA بر روی گزینهٔ Post Hoc کلیک کرده و در پنجرهٔ ظاهر شده آزمون تعقیبی مورد نظر را انتخاب می کنیم.

مثال

فرض کنید بخواهیم میانگین نمرات درس حرفه و فن دانش آموزان پایه های اول، دوم، و سوم را با یکدیگر مقایسه کنیم. یعنی در واقع می خواهیم فرضیه $H_0 : \mu_1 = \mu_2 = \mu_3$ که در آن μ_1، μ_2، و μ_3 به ترتیب بیانگر میانگین نمرات حرفه و فن دانش آموزان پایه های اول، دوم، و سوم است را آزمون کنیم. نتیجه انجام این آزمون توسط ANOVA در زیر آمده است.

ANOVA

TECH.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	120.159	2	60.079	4.209	.015
Within Groups	18611.629	1304	14.273		
Total	18731.788	1306			

در این جدول همانطوری که مشاهده می شود P.Value این آزمون 0/015 بوده و بر این اساس فرض صفر در سطح خطای 5٪ رد می شود. یعنی میانگین نمرات حرفه و فن دست کم در دو پایه تحصیلی متفاوت است. حال برای تعیین اینکه میانگین کدامیک از پایه ها متفاوت از یکدیگر است، از آزمون تعقیبی شفه استفاده می کنیم. بدین منظور در پنجره ANOVA بر روی گزینه Post Hoc کلیک کرده و در پنجره ظاهر شده آزمون تعقیبی شفه (Scheffe) را انتخاب می کنیم. پس از اجرای این فرمان

ابتدا ANOVA انجام شده و جدول آن در خروجی خواهد آمد. سپس آزمون تعقیبی شفه انجام می گیرد و نتیجه آن در قالب دو جدول در خروجی خواهد آمد که دومین جدول آن در زیر آمده است.

TECH.

Scheffe[a,b]

GRADE	N	Subset for alpha = 0.05	
		1	2
FIRST	416	14.7530	
SECOND	473	15.2072	15.2072
THIRD	418		15.5066
Sig.		.209	.506

مطابق این جدول پایه های اول، دوم و سوم از لحاظ درس حرفه و فن در دو گروه 1 و 2 افراز شده اند. در گروه 1 پایه های اول و دوم قرار گرفته و در گروه 2 پایه های دوم و سوم قرار گرفته اند. این به آن معنی است که میانگین نمره درس حرفه و فن در پایه های اول و دوم تفاوت معنی داری با یکدیگر نداشته و آنها را می توان در یک گروه قرار داد. پایه های دوم و سوم نیز تفاوت معنی داری با یکدیگر نداشته و آنها را نیز می توان در یک گروه دیگر و متفاوت از گروه قبل قرار داد. یعنی پایه دوم هم می تواند در گروه 1 قرار گیرد و هم می تواند در گروه 2 قرار گیرد و بر این اساس پایه دوم نه با پایه اول و نه با پایه سوم تفاوت معنی داری ندارد بلکه این تفاوت در مورد پایه های اول و سوم است. توجه کنید که نتیجه این آزمون به هر شکلی می تواند باشد. ممکن است هر کدام از پایه ها به تنهائی در یک گروه قرار گیرند که در چنین صورتی میانگین در همه پایه ها متفاوت از یکدیگر خواهد بود. ممکن است دو پایه در یک گروه و پایه ای دیگر در یک گروه مجزا قرار گیرد و ممکن است پایه ای در برخی گروهها مشترک باشد مانند آنچه که در این مثال اتفاق افتاد. هر کدام از این حالتها تفسیر خاص خودش را دارد.

تمرین

میانگین نمرات دروس مختلف را در پایه های اول، دوم و سوم توسط ANOVA با یکدیگر مقایسه کنید و در صورت معنی دار بودن تفاوت بین آنها توسط آزمون تعقیبی شفه مشخص کنید که کدامیک از پایه ها متفاوت از یکدیگرند. این عمل را یکبار روی کل نمونه، یکبار به تفکیک جنسیت، یکبار به تفکیک منطقه و یکبار نیز به تفکیک جنسیت و منطقه انجام دهید.

نحوه Recode کردن متغیرها

در برخی موارد ممکن است بخواهیم مقادیر یک متغیر را به مقادیر جدیدی تغییر دهیم. بخصوص ممکن است بخواهیم یک متغیر کمی که در سطح اندازه گیری لااقل فاصله ای است را توسط کدگذاری به یک متغیر ترتیبی تبدیل کنیم. مثلا فرض کنید متغیر وزن را به صورت کمی اندازه گیری کرده ایم اما می خواهیم وزن را به صورت طبقه ای کمتر از 50 ، 50-60 ، 60-70 و 70 به بالا تبدیل کنیم. در صورتی که بخواهیم متغیر جدید روی متغیر قبلی تشکیل شده و متغیر قبلی از بین برود از فرمان Recode کردن در مسیر زیر استفاده می کنیم.

Recode into same variable → Transform

اما اگر بخواهیم متغیر قبلی حفظ شده و متغیر جدیدی تشکیل شود از فرمان زیر استفاده می کنیم.

Recode into different variable → Transform

مثلا فرض کنید در فایل test 2 بخواهیم نمرات درس ریاضی دانش آموزان را به صورت زیر طبقه بندی کنیم.

$$Y=\begin{cases}1 & 0\le X<5\\ 2 & 5\le X<10\\ 3 & 10\le X<12\\ 4 & 12\le X<14\\ 5 & 14\le X<17\\ 6 & 17\le X\le 20\end{cases}$$

که در آن x نمرات درس ریاضی دانش آموزان در فاصله 0 تا 20 و y نمره تبدیل یافته درس ریاضی دانش آموزان است.

بدین منظور پس از رفتن به مسیر فوق متغیر مربوط به نمره درس ریاضی دانش آموزان (x4) را به قسمت Input Variable برده و در بخش output variable یک نام برای متغیر جدید وارد می کنیم. (مثلا y4) سپس با کلیک کردن بر روی old and new values پنجره ای باز می شود که در سمت چپ این پنجره مقادیر قدیم و یا فاصله های متغیر قدیم را وارد کرده و در بخش new variable مقادیر جدید و یا کد مربوط به آن فاصله ها را وارد می کنیم.

در قسمت old values و در قسمت range فاصله بین هر گروه را وارد کرده و سپس در قسمت new values کد هر گروه را وارد کرده و add را می زنیم.

تمرین:

نمرات دروس مختلف دانش آموزان را مطابق با تقسیم بندی فوق طبقه بندی کرده و کد مربوط به مقادیر جدید را در متغیرهای به ترتیب Y1 ، Y2 ، Y3، Y4 قرار دهید.

تمرین:

با استفاده از فرمان ANOVA میانگین نمره درس علوم دانش آموزانی که در درس ریاضی خیلی ضعیف، ضعیف، متوسط پائین، متوسط بالا، خوب و خیلی خوب است (مقادیر مختلف متغیر Y1) را با یکدیگر مقایسه کرده و در صورت معنی دار بودن اختلاف بین این میانگینها توسط آزمون شفه مشخص کنید میانگین کدامیک از گروهها با یکدیگر متفاوت است. چه نتیجه ای به دست می آید؟ این مقایسه را برای نمرات سایر دروس نیز انجام دهید.

تشکیل یک متغیر جدید توسط ترکیبی از متغیرهای قدیم

در مواردی ممکن است بخواهیم توسط ترکیبی از متغیرهای موجود در فایل یک متغیر جدیدی را ایجاد کنیم. مثلا ممکن است بخواهیم تعدادی از متغیرهای موجود در فایل را با یکدیگر جمع کرده و حاصل جمع آنها را در یک متغیر جدید قرار دهیم. یا مثلا فرض کنید بخواهیم میانگین مقادیر تعدادی از متغیرها را محاسبه کرده و این میانگینها را در یک متغیر جدید قرار دهیم. در این صورت از فرمان Compute variable در مسیر زیر استفاده می کنیم.

Transform → Compute variable

مثلا فرض کنید در در فایل test 2 بخواهیم معدل نمرات دروس مختلف هر دانش آموز را محاسبه کرده و آنها را در متغیری با نام Ave قرار دهیم در این صورت پس از رفتن به مسیر فوق در قسمت Target variable اسم متغیر جدید (Ave) را وارد

کرده و در قسمت numeric expresion عبارت مورد نظر با همان فرمول محاسبه میانگین نمرات دروس مختلف [5/(X4+X5+X6+X7+X8)] را وارد می کنیم. در چنین صورتی میانگین نمرات دروس مختلف برای هر دانش آموز محاسبه شده و این میانگینها در متغیر جدید Ave قرار می گیرد. البته در انجام این محاسبه اگر مقدار یکی از متغیرها برای یک نفر از افراد نمونه Missing باشد برای آن شخص محاسبه ای انجام نمی گیرد. در این فایل نیز چون دانش آموزان کلاس اول فاقد نمره زبان هستند لذا برای آنها محاسبه ای انجام نگرفته است. اما واضح است که مایلیم برای این دانش آموزان نیز معدل دروسشان محاسبه شود منتها بدون در نظر گرفتن نمره زبان. یعنی در واقع برای کلاس اولی ها می خواهیم معدل دروس توسط فرمول Ave=(X4+X5+X6+X8)/4 محاسبه شود. بدین منظور مجددا این فرمان را اجرا می کنیم منتها از فرمول فوق استفاده کرده و توسط گزینه If در پائین این پنجره مشخص می کنیم که این فرمان فقط برای کلاس اولی ها محاسبه شود. نتیجه اجرای این فرمان را نیز در همان متغیر Ave قرار می دهیم.

ضرایب همبستگی:

در این بحث میخواهیم رابطه بین متغیرها را توسط شاخصی با نام ضریب همبستگی مورد بررسی قرار دهیم. برای محاسبه میزان رابطه و همبستگی بین متغیرها شاخص های مختلفی وجود دارد که عمده ترین آنها، شاخص های **ضریب همبستگی پیرسون**، **ضریب همبستگی تاو کندال** و **ضریب همبستگی رتبه ای اسپیرمن** است.

ضریب همبستگی پیرسون:

این ضریب شاخصی است برای اندازه گیری میزان رابطه بین دو متغیر کمی که در سطح اندازه گیری لااقل فاصله ای هستند. به عبارت دیگر هرگاه متغیرهایی را که می خواهیم میزان رابطه بین آنها را اندازه گیری کنیم، هر دو کمی بوده و سطح اندازه گیری آنها لااقل فاصله ای باشد (فاصله ای یا نسبتی) و نیز توزیع آنها نرمال باشد در این صورت از **ضریب همبستگی پیرسون** استفاده می کنیم. البته اگر حجم نمونه به قدر کافی بزرگ باشد و هر دو متغیر در کمی بوده و در سطح اندازه گیری لااقل فاصله ای باشند، حتی در صورت عدم برقراری شرط نرمال بودن باز هم می توان از این شاخص استفاده کرد. مقدار این ضریب حداقل منهای 1 و حداکثر 1 است. هر چه قدر مقدار این ضریب به 1 نزدیک تر باشد، بیانگر وجود رابطه خطی و مستقیم بین دو متغیر است. یعنی افزایش یکی از متغیرها باعث افزایش متغیر دیگر می شود و همینطور در مورد کاهش. هر چقدر مقدار این ضریب به 1ـ نزدیک تر باشد بیانگر وجود رابطه خطی و معکوس بین دو متغیراست، یعنی افزایش یکی از متغیرها باعث کاهش متغیر دیگر می شود و یا بالعکس. و هر چقدر مقدار این ضریب به صفر نزدیک تر باشد، بیانگر عدم وجود رابطه خطی بین دو متغیر است. پس از محاسبه ضریب همبستگی در نمونه لازم است معنی دار بودن آن را آزمون کنیم که آزمون معنی دار بودن ضریب همبستگی دارای فرضیه ای به صورت زیر است.

H_0 : ضریب همبستگی محاسبه شده در نمونه معنی دار نیست(بین این دو متغیر رابطه برقرار نیست)

H_1 : ضریب همبستگی محاسبه شده در نمونه معنی دار است(بین این دو متغیر رابطه برقرار است)

مانند سایر آزمونها اگر P.valueی این آزمون از 0/05 کمتر شود فرض صفر رد شده که بیانگر معنی دار بودن رابطه بین آن دو متغیر است. SPSS بطور خودکار پس از محاسبه ضریب همبستگی نمونه، معنی دار بودن آن را آزمون کرده و P.value ی آن را در خروجی ارائه می دهد.

ضریب همبستگی رتبه ای اسپیرمن و ضریب همبستگی تاو کندال:

این دو ضریب نیز شاخص هایی برای اندازه گیری رابطه بین دو متغیر است. با این تفاوت که از این شاخص ها هنگامی استفاده می شود که دست کم یکی از متغیرها در سطح اندازه گیری ترتیبی باشد. به عبارت دیگر برای اندازه گیری میزان رابطه بین دو متغیر وقتی *دست کم یکی از متغیرها* ***ترتیبی*** است از « **ضرایب همبستگی رتبه ای اسپیرمن و یا تاو کندال** » استفاده می کنیم، که البته استفاده از اسپیرمن متداولتر است. این دو ضریب نیز مانند پیرسون حداقل 1 ـ و حداکثر 1 است . بنابراین هر چقدر مقدار این ضرایب به 1 نزدیکتر باشد بیانگر وجود رابطه ای مستقیم بین دو متغیر است و هر چقدر به 1 ـ نزدیک تر باشد، بیانگر وجود رابطه ای معکوس بین دو متغیر است. و هر چقدر مقدار این ضرایب به صفر نزدیک تر باشد، بیانگر عدم وجود رابطه بین دو متغیر است. در مورد این دو ضریب نیز مانند پیرسون لازم است پس از محاسبه آنها در نمونه معنی دار بودن آنها آزمون شود و آزمون معنی دار بودن آنها مانند آزمون معنی دار بودن پیرسون دارای فرضیه ای به صورت زیر است.

H_0 : ضریب همبستگی محاسبه شده در نمونه معنی دار نیست(بین این دو متغیر رابطه برقرار نیست)

H_1 : ضریب همبستگی محاسبه شده در نمونه معنی دار است(بین این دو متغیر رابطه برقرار است)

SPSS بطور خودکار پس از محاسبه ضریب همبستگی نمونه، معنی دار بودن آن را آزمون کرده و P.value ی آن را در خروجی ارائه می دهد.

مثلاً اگر بخواهیم تعیین کنیم که آیا بین اندازه *قد* دونده ها و *مسافت طی شده* توسط آنها (که هر دو متغیر کمی است) رابطه معنی داری وجود دارد یا خیر؟ از **ضریب همبستگی پیرسون** استفاده می کنیم. و اگر بخواهیم رابطه بین *مقام کسب شده* در مسابقه شنا (که به صورت ترتیبی اول، دوم و سوم است) با *طول قد* شناگر را اندازه گیری کنیم از **ضریب همبستگی اسپیرمن** استفاده می کنیم. اگر بخواهیم تعیین کنیم که آیا بین نمرات درس ریاضی و نمرات درس علوم دانش آموزان رابطه ای برقرار است یا خیر؟ چون نمرات این دو درس هر دو متغیرهای کمی و در سطح اندازه گیری فاصله ای هستند لذا از ضریب همبستگی پیرسون استفاده می کنیم. اما اگر بخواهیم رابطه بین میزان علاقه دانشجو به رشته تحصیلی اش و نمره درس آمار او را تعیین کنیم و میزان علاقه به رشته نیز به صورت "خیلی کم"، "کم"، " متوسط"، "زیاد" و "خیلی زیاد" اندازه گیری شده باشد، چون یکی از این متغیرها به صورت ترتیبی و دیگری به صورت فاصله ای اندازه گیری شده اند لذا از ضریب همبستگی اسپیرمن استفاده می کنیم.

همچنین اگر بخواهیم رابطه بین میزان علاقه دانشجو به رشته تحصیلی اش و میزان علاقه او به درس آمار را اندازه گیری کنیم و میزان علاقه دانشجو به رشته تحصیلی اش و نیز میزان علاقه به درس آمار به صورت ترتیبی "خیلی کم"، "کم"، "متوسط" و . . . اندازه گیری شده باشد، چون هر دو متغیر در سطح اندازه گیری شده اند از ضریب همبستگی اسپیرمن استفاده می کنیم.

مسیر محاسبه ضرایب همبستگی در SPSS به صورت زیر است.

Analyze ⟶ Correlate ⟶ Bivariate

مثلاً فرض کنید بخواهیم ضریب همبستگی بین نمرات دروس مختلف را محاسبه کنیم. در این صورت پس از رفتن به مسیر فوق متغیرهای مربوط به نمرات دروس را به قسمت Variables می بریم. در قسمت پایین و در بخش Correlation Coefficientes مشخص می کنیم که کدامیک از این ضرایب را می خواهیم محاسبه کنیم. در این مثال چون نمرات دروس متغیرهای کمی و در سطح اندازه گیری فاصله ای است از ضریب همبستگی پیرسون که پیش فرض SPSS نیز پیرسون است استفاده می کنیم. نتیجه به صورت یک ماتریس متقارن با نام ماتریس همبستگی در خروجی خواهد آمد که این ماتریس در زیر آمده است.

Correlations

		MATH.	SCEIN.	TECH.	FOREI.	PERSIAN
MATH.	Pearson Correlation	1	.776**	.715**	.790**	.678**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1311	1307	1293	880	1301
SCEIN.	Pearson Correlation	.776**	1	.756**	.778**	.648**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1307	1322	1301	883	1312
TECH.	Pearson Correlation	.715**	.756**	1	.709**	.613**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1293	1301	1307	871	1297
FOREI.	Pearson Correlation	.790**	.778**	.709**	1	.729**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	880	883	871	890	886
PERSIAN	Pearson Correlation	.678**	.648**	.613**	.729**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	1301	1312	1297	886	1326

Correlations

		MATH.	SCEIN.	TECH.	FOREI.	PERSIAN
MATH.	Pearson Correlation	1	.776**	.715**	.790**	.678**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1311	1307	1293	880	1301
SCEIN.	Pearson Correlation	.776**	1	.756**	.778**	.648**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1307	1322	1301	883	1312
TECH.	Pearson Correlation	.715**	.756**	1	.709**	.613**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1293	1301	1307	871	1297
FOREI.	Pearson Correlation	.790**	.778**	.709**	1	.729**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	880	883	871	890	886
PERSIAN	Pearson Correlation	.678**	.648**	.613**	.729**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	1301	1312	1297	886	1326

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

در جدول فوق که در واقع ماتریس همبستگی نمرات دروس مختلف است، در هرکدام از خانه های این جدول ضریب همبستگی پیرسون بین نمرات دو درس، P-valueی آزمون معنی دار بودن این ضریب و نیز حجم نمونه آمده است. همانطوریکه مشاهده می شود عناصر قطر اصلی این ماتریس همگی یک بوده و P-value برای آن محاسبه نشده است. چون عناصر قطر اصلی در واقع ضریب همبستگی یک متغیر با خودش است. همچنین این ماتریس متقارن است یعنی مثلا خانه سطر دوم، ستون سوم آن با خانه سطر سوم، ستون دوم آن یکی است. چون ضریب همبستگی نسبت به دو متغیرش متقارن است. به عبارت دیگر ضریب همبستگی x و y برابر است با ضریب همبستگی y و x .

مطابق این جدول مشاهده می شود که مثلا ضریب همبستگی بین نمرات دروس ریاضی و علوم دانش آموزان 0/776 بوده و p-valueی آزمون معنی دار بودن این ضریب تا سه رقم اعشار صفر است. بنابر این این ضریب در سطح خطای 0/01 (و در نتیجه در سطح خطای 0/05) معنی دار است. چون مقدار این ضریب مثبت است می توان نتیجه گرفت که رابطه ای مستقیم و معنی دار بین نمرات دروس ریاضی و علوم دانش آموزان برقرار است. با توجه به اعداد این جدول این نتیجه در مورد سایر دروس نیز برقرار است (همانطوریکه انتظار داشتیم). همچنین مطابق اطلاعات سطر اول این جدول می توان نتیجه گرفت که بیشترین

همبستگی نمرات درس ریاضی با نمرات درس زبان و پس از آن با نمرات درس علوم است. پس از آن نیز نمرات دروس حرفه و فن و فارسی قرار می گیرد. به همین ترتیب در مورد سایر دروس نیز می توان نتایجی مشابه به دست آورد.

ضریب همبستگی جزئی (Partial Correlation)

از این شاخص برای محاسبه همبستگی بین دو متغیر پس از حذف اثر یک یا بیش از یک متغیر استفاده می شود.

گاهی اوقات به نظر می رسد بین دو متغیر رابطه معنی داری وجود دارد، حال آنکه وجود رابطه بین آن دو متغیر متاثر از وجود یک یا بیش از یک متغیر دیگر است که اگر اثر آن متغیرها حذف شود ممکن است بین آن دو متغیر رابطه معنی داری وجود نداشته باشد. در چنین مواردی و برای محاسبه همبستگی بین دو متغیر پس از حذف اثر یک یا بیش از یک متغیر از ضریب همبستگی جزئی استفاده می شود که مسیر آن در SPSS به صورت زیر است.

Analyze → Correlate → Partial . . .

مثال (مثال صفحه 84 از کتاب تحلیل رگرسیون خطی)

داده های زیر "تعداد گلهای زده شده"، " میزان فروش چتر" و " میزان بارندگی" در یکی از میادین فوتبال را در دوازده مسابقه برگزار شده در آن استادیوم و در یک فاصله زمانی مشخص نشان می دهد.

تعداد گلهای زده شده	میزان فروش چتر	میزان بارندگی
0	96	319
1	80	280
0	69	288
4	65	100
1	65	130

150	58	2
120	51	3
95	63	6
90	40	6
90	20	3
60	26	5
60	3	5

ضریب همبستگی بین "تعداد گلهای زده شده" و "میزان فروش چتر" و نیز ضریب همبستگی بین این دو متغیر پس از حذف اثر " میزان بارندگی" را محاسبه کرده و نتیجه گیری کنید.

حل: ابتدا توسط فرمان Bivariate ضریب همبستگی بین این دو متغیر را به دست می آوریم. نتیجه در جدول زیر آمده است.

Correlations

		Goal	Umberela
Goal	Pearson Correlation	1	-.636*
	Sig. (2-tailed)		.026
	N	12	12
Umberela	Pearson Correlation	-.636*	1
	Sig. (2-tailed)	.026	
	N	12	12

*. Correlation is significant at the 0.05 level (2-tailed).

همانطوریکه مشاهده می شود ضریب همبستگی بین این دو متغیر 0.636- و p-valueی معنی دار بودن این ضریب 0.026 است که بیانگر وجود رابطه ای معکوس و معنی دار (در سطح خطای 0.05) بین این دو متغیر است. یعنی مطابق با این ضریب نتیجه می شود که اگر در یک روز چتر بیشتری فروخته شود در آن روز گل کمتری زده خواهد شد و بر این اساس اگر بخواهیم

در یک روز خاص بازی خیلی پر هیجان بوده و گلهای زیادی بین دو تیم رد و بدل شود باید در آن روز فروش چتر در اطراف استادیوم را ممنوع کنیم!!!!

اما واضح است که این نتیجه نامعقول بوده و اساسا فروش چتر ارتباطی با گلهای زده شده ندارد. فروش چتر می تواند به جهت بارندگی در آن روز باشد و بر این اساس لازم است ضریب همبستگی بین این دو متغیر پس از حذف اثر بارندگی محاسبه شود. نتیجه در زیر آمده است.

Correlations

Control Variables			Goal	Umberela
Rain	Goal	Correlation	1.000	.075
		Significance (2-tailed)	.	.826
		Df	0	9
	Umberela	Correlation	.075	1.000
		Significance (2-tailed)	.826	.
		Df	9	0

همانطوریکه مشاهده می شود پس از حذف اثر بارندگی ضریب همبستگی بین این دو متغیر به 0.075 کاهش یافته و p-valueی آزمون معنی دار بودن آن 0.829 است که بیانگر معنی دار نبودن این ضریب است. بنابر این وجود همبستگی بین این دو متغیر به جهت وجود متغیر "میزان بارندگی" بود که اگر اثر بارندگی حذف شود رابطه معنی داری بین تعداد گلهای زده شده در مسابقه و میزان فروش چتر در آن روز وجود ندارد.

ضریب همبستگی فاصله ای (Distance Correlation)

این ضریب شاخصی است برای محاسبه تشابه یا عدم تشابه (فاصله) بین هر جفت مشاهده یا هر جفت متغیر. یعنی این شاخص هم برای جفت مشاهدات و هم برای جفت متغیرها به کار می رود. از این شاخص به طور مستقیم استفاده نشده بلکه از آن برای استفاده در سایر تحلیلها مانند تحلیل عاملی، آنالیز خوشه ای و تحلیل ممیزی استفاده می شود.

رگرسیون :

در تحلیل رگرسیون می خواهیم رابطه بین یک متغیر وابسته و چند متغیر مستقل را توسط یک فرمول ریاضی نشان دهیم. در ساده ترین حالت تعداد یک متغیر وابسته و یک متغیر مستقل در نظر گرفته و رابطه ی بین آنها را یک رابطه خطی به شکل $y = \beta_0 + \beta_1 x$ در نظر می گیریم. چنین مدلی را یک مدل خطی ساده می نامند. در این مدل x متغیر مستقل یا متغیر پیشگو و y متغیر وابسته یا متغیر پاسخ است و β_0 و β_1 ضرایب خط رگرسیون که پارامتر های ثابتی هستند که آنها را از روی یک نمونه تصادفی بر آورد می کنیم. برای برآورد این پارامترها از روش کمترین مربعات استفاده می شود. β_1 ضریب زاویه یا شیب و β_0 عرض از مبدا خط رگرسیون است. پس از بر آورد این ضرائب ، آنها را با $\hat{\beta}_0$ یا b_0 ونیز $\hat{\beta}_1$ یا b_1 نشان می دهیم و برآورد معادله خط رگرسیون را به صورت $y = b_0 + b_1 x$ و یا $\hat{y} = b_0 + b_1 x$ نشان می دهیم. با داشتن چنین معادله ای و با معلوم بودن یک مقدار برای متغیر مستقل می توان یک مقدار برای متغیر وابسته پیش بینی کرد که اگر مدل معنی دار باشد در اینصورت هر چقدر ضریب تعیین مدل بالاتر باشد مقدار پیش بینی شده توسط این مدل به مقدار واقعی نزدیکتر خواهد بود. در حالت ایده آل و با شرط معنی دار بودن مدل، اگر ضریب تعیین مساوی یک باشد در اینصورت پیش بینی که توسط این مدل انجام می گیرد به طور کامل (100 درصد) با مقدار واقعی یکسان خواهد بود.

در یک مدل رگرسیون خطی ساده شیب خط بیانگر میزان تغییرات در متغیر وابسته به ازای یک واحد تغییر در متغیر مستقل است. اگر متغیر مستقل دارای صفر معنی دار باشد یعنی بر روی عدد صفر تعریف شود در اینصورت عرض از مبدا خط بیانگر مقدار متغیر وابسته به ازای x=0 است، اما اگر متغیر مستقل بر روی صفر تعریف نشود در اینصورت عرض از مبدا هیچگونه تعبیر عملی خاصی ندارد.

پس از بر آورد معادله خط رگرسیون لازم است معنی دار بودن این معادله و نیز معنی دار بودن هر کدام از این ضرایب آزمون شود که آزمون معنی دار بودن مدل رگرسیون دارای فرضیه ای به صورت زیر است.

$$\begin{cases} H_0: & \text{مدل رگرسیونی برازش شده معنی دار نیست} \\ H_1: & \text{مدل رگرسیونی برازش شده معنی دار هست} \end{cases}$$

که این فرضیه توسط یک تحلیل واریانس (ANOVA) آزمون می شود هم چنین آزمون معنی دار بودن ضرایب خط رگرسیون دارای فرضیه هایی به صورت زیر است.

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0: & \text{وجود } \beta_1 \text{ یا همان شیب خط رگرسیون در معادله معنی دار نسیت} \\ H_1: & \text{وجود } \beta_1 \text{ یا همان شیب خط رگرسیون در معادله معنی دار هست} \end{cases}$$

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0: & \text{وجود } \beta_0 \text{ یا همان عرض از مبدا در مدل معنی دار نیست} \\ H_1: & \text{وجود } \beta_0 \text{ یا همان عرض از مبدا در مدل معنی دار هست} \end{cases}$$

که این آزمون ها توسط آماره t صورت می گیرد.

ضریب تعیین (ضریب تشخیص)

نسبت پراکندگی بیان شده توسط مدل رگرسیون به پراکندگی کل را ضریب تعیین یا ضریب تشخیص گفته و آنرا با R^2 (R-square) نشان می دهیم. ضریب تعیین مشخص می کند که چه نسبتی از تغییرات یا پراکندگی در متغیر وابسته مطابق با مدل رگرسیونی به دست آمده به متغیر مستقل مربوط است یا به عبارت دیگر متغیر مستقل چه نسبتی از تغییرات متغیر وابسته را تبیین می کند. مثلا اگر در یک مدل رگرسیونی ضریب تعیین $R^2 = 0.7$ بدست آمده باشد این به آن معنی است که ٪70 از تغییرات یا پراکندگی در متغیر وابسته توسط این مدل رگرسیونی به متغیر مستقل مربوط می شود و ٪30 باقی مانده به سایر متغیرها بستگی دارد که در این مدل نادیده گرفته شده است. مقدار این ضریب حداقل صفر و حداکثر 1 است. هر چقدر مقدار این ضریب به صفر نزدیکتر باشد بیانگر نا معتبر بودن مدل رگرسیون و هر چقدر مقدار آن به یک نزدیکتر باشد بیانگر معتبر بودن مدل است. در رگرسیون خطی ساده ضریب تعیین برابر است با مجذور ضریب همبستگی.

مسیر انجام تحلیل رگرسیون ساده (با یک متغیر مستقل) به صورت زیر است:

$$Analyze \rightarrow Regression \rightarrow Curve\ Estimation$$

مثلاً در فایل $Test_2$ فرض کنید نمرات درس علوم تابعی از نمرات درس ریاضی باشد یعنی نمرات درس علوم متغیر وابسته و نمرات درس ریاضی متغیر مستقل باشد. برای برازش یک مدل رگرسیونی به این دو متغیر به مسیر فوق رفته و در پنجره ظاهر شده متغیر وابسته یا همان نمرات درس علوم را به قسمت $Dependent$ و متغیر مستقل یا نمرات درس ریاضی را به قسمت

$independent$ می بریم. در قسمت سمت راست این پنجره می توان تعیین کرد که این مدل با وجود عرض از مبدا یا بدون وجود آن باشد و اکیداً توصیه می شود که همواره هر مدل رگرسیونی را با وجود عرض از مبدا بدست آوریم مگر آنکه دلایل کافی برای عدم وجود عرض از مبدا وجود داشته باشد. در قسمت پایین این پنجره و در بخش Models می توان تعیین کرد که مدل برازش شده یک مدل خطی(Linear) یا مدل غیر خطی باشد و در قسمت انتهایی این پنجره می توان مشخص کرد که جدول ANOVA مربوط به آزمون معنی دار بودن مدل رگرسیون نیز در خروجی داده شود. پس از اجرای این فرمان نتیجه در قالب سه جدول اصلی و در خروجی خواهد آمد که این جداول در زیرآمده است.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.776	.602	.602	2.673

The independent variable is MATH..

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	14119.122	1	14119.122	1975.528	.000
Residual	9326.852	1305	7.147		
Total	23445.974	1306			

The independent variable is MATH..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
MATH.	.676	.015	.776	44.447	.000
(Constant)	4.734	.211		22.441	.000

در اولین جدول با عنوان $mode\ summary$ ستون اول با عنوان R قدر مطلق ضریب همبستگی بین این دو متغیر است که در این مثال مقدار این ضریب 0.776 شده است. دومین ستون با عنوان $R\ Square$ ضریب تعیین این مدل است که در این مثال ضریب تعیین 0.602 شده است و می توان نتیجه گرفت که حدود 60 ٪ از تغییرات یا پراکندگی در نمرات درس علوم دانش آموزان مطابق با این مدل رگرسیونی به نمرات درس ریاضی آنان بستگی دارد. ستون بعدی با نام $Adjusted\ R-Square$

ضریب تعیین تعدیل شده است که کاربرد آن در رگرسیونیهای چند گانه بوده و بعدا راجع به آن صحبت خواهد شد. آخرین ستون این جدول خطای استاندارد برآورد است. خطای استاندارد برآورد شاخصی است که توسط آن می توان مشخص کرد که مقدار برآورد شده و یا پیش بینی شده متغیر وابسته توسط مدل رگرسیون به طور متوسط چقدر تا مقدار واقعی فاصله دارد.

جدول بعدی با نام جدول $ANOVA$ نتیجه آزمون معنی دار بودن مدل رگرسیونی است که در این مثال P_value این آزمون خیلی کوچک بوده و دست کم تا سه رقم اعشار صفر است. بنابر این فرض صفر مبنی بر معنی دار نبودن مدل رد می شود یعنی مدل رگرسیونی به دست آمده معنی دار است.

سومین جدول با نام $Coefficient$ برآورد ضرایب خط رگرسیون و نتیجه آزمون معنی دار بودن آنهاست. در این جدول ستون اول با عنوان B برآورد ضرایب رگرسیون است که در این مثال $\hat{\beta}_0 = \mathrm{b}_0 = 4.734$ و $\hat{\beta}_1 = b_1 = 0.676$ است.

ستون دوم خطای استاندارد این ضرایب است. ستون بعدی با عنوان Beta ضرایب استاندارد شده است که کاربرد آن در رگرسیونیهای چندگانه بوده و بعدا راجع به آنها صحبت خواهد شد. ستون بعدی با عنوان t آماره آزمون معنی دار بودن این ضرایب است. آخرین ستون این جدول با عنوان Sig ، P_value ی آزمون معنی دار بودن این ضرایب است که در این مثال هر دو p_value خیلی کوچک بوده و دست کم تا سه رقم اعشار صفر است. بنابراین فرض صفر مبنی بر معنی دار نبودن این ضرائب در سطح خطای 0.01 (و در نتیجه در سطح خطای 0.05) رد می شود. بر این اساس وجود هر کدام از این ضرایب در مدل معنی دار است.

خلاصه نتیجه ای که از این تحلیل بدست می آید آن است که: فرض کنید y بیانگر نمره درس علوم و x بیانگر نمره درس ریاضی دانش آموزان باشد. و نیز فرض کنید نمرات درس علوم تابعی از نمرات درس ریاضی باشد یعنی نمرات درس علوم متغیر وابسته و نمرات درس ریاضی متغیر مستقل باشد. در این صورت یک مدل خطی رگرسیونی

برازش شده بین این دو متغیر به صورت زیر است.

$$y = b_0 + b_1 x = 4.734 + 0.676\, x$$

که این مدل معنی دار است و نیز وجود هر کدام از این ضرایب در مدل معنی دار است و ضریب تعیین این مدل $R^2 = 0.6$ است. یعنی 60 ٪ از تغییرات یا پراکندگی در نمرات علوم دانش آموزان مطابق با این مدل به نمره ریاضی آنها بستگی داشته و

40 درصد به سایر عوامل بستگی داشته که در این مدل نادیده گرفته شده اند. یا به عبارت دیگر حدود 60 درصد از پراکندگی در نمرات درس علوم دانش آموزان را می توان مطابق با این مدل به نمرات درس ریاضی آنها مربوط دانست. حال با داشتن چنین مدلی و با معلوم بودن نمره ریاضی یک دانش آموز می توان نمره درس علوم او را پیش بینی کرد. مثلا فرض کنید دانش آموزی در درس ریاضی نمره 10 بگیرد در اینصورت داریم.

$$x = 10 \quad \rightarrow \quad \hat{y} = 4.734 + 0.676\,(10) \simeq 11.5$$

بنابراین اگر دانش آموزی در درس ریاضی نمره 10 بگیرد پیش بینی می شود که نمره علوم او حدود 11.5 شود که البته این پیش بینی توام با میزانی خطاست که هر چقدر ضریب تعیین مدل بالاتر باشد این خطا کمتر است. همچنین چون خطای استاندارد برآورد در اولین جدول 2.67 می باشد لذا می توان نتیجه گرفت که اگر نمره ریاضی یک دانش آموز 10 باشد پیش بینی می شود که نمره علوم او حدود 11.5 باشد و نیز به طور متوسط نمره واقعی علوم این دانش آموز حدود 2.67 بیشتر یا کمتر از 11.5 است.

چون شیب خط رگرسیون حدوداً 0.7 به دست آمده است این به آن معنی است که هر یک واحد افزایش در نمره درس ریاضی باعث ایجاد حدود 0.7 افزایش در نمره علوم خواهد شد.

عرض از مبدا خط حدود 4.7 است لذا نمره علوم دانش آموزانی که از لحاظ ریاضی خیلی ضعیف بوده و نمره آنها در حد صفر است حدود 4.7 برآورد می شود.

در مسیر انجام رگرسیون ساده و در پنجره *Curve Estimation* با کلیک کردن بر روی گزینه *Save* در گوشه سمت راست این پنجره می توان مقدار پیش بینی شده متغیر وابسته به ازای هرکدام از مشاهدات و نیز باقیمانده مقدار پیش بینی شده با مقدار واقعی را محاسبه کرده و به عنوان متغیرهای جدید در فایل داده ذخیره کرد. از این متغیرها در انجام برخی تحلیلها و به خصوص تحلیلهای مربوط به آزمون پیش فرضهای رگرسیون استفاده می شود.

مدلهای غیر خطی

در مدلهای غیر خطی رابطه بین متغیر مستقل با متغیر وابسته به صورت یک معادله غیر خطی در نظر گرفته می شود. برای برازش مدل غیر خطی به داده ها از همان مسیر قبل استفاده می کنیم.

$$Analyze \rightarrow Regression \rightarrow Curve\ Estimation$$

و در قسمت $models$ نوع مدل مورد نظر را انتخاب می کنیم. برای آگاهی از فرمول های این مدل ها می توان از $help$ کمک گرفت. مثلاً فرض کنید مدل های خطی($Linear$)، لگاریتمی($Logarithmic$)، درجه دو($Quadratic$) و درجه سه($Cubic$) را به متغیرهای نمرات درس ریاضی (به عنوان متغیر مستقل یا x) و نمرات درس علوم (به عنوان متغیر وابسته یا y) برازش کنیم. نتیجه انجام این تحلیل در زیر آمده است.

Linear

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.776	.602	.602	2.673

The independent variable is MATH..

ANOVA

	Sum of Squares	Df	Mean Square	F	Sig.
Regression	14119.122	1	14119.122	1975.528	.000
Residual	9326.852	1305	7.147		
Total	23445.974	1306			

The independent variable is MATH..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
MATH.	.676	.015	.776	44.447	.000
(Constant)	4.734	.211		22.441	.000

این سه جدول نتیجه برازش مدل خطی به این دو متغیر است که قبلا راجع به آنها به طور مفصل بحث شد. در ادامه سه جدول زیر نتیجه برازش مدل لگاریتمی به این سه متغیر آمده است.

Logarithmic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.719	.517	.516	2.946

The independent variable is MATH..

ANOVA

	Sum of Squares	Df	Mean Square	F	Sig.
Regression	12116.987	1	12116.987	1395.771	.000
Residual	11328.986	1305	8.681		
Total	23445.974	1306			

The independent variable is MATH..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
ln(MATH.)	6.571	.176	.719	37.360	.000
(Constant)	-2.734	.442		-6.179	.000

در این سه جدول نیز مانند جداول مربوط به مدل خطی، در اولین جدول ضریب همبستگی، ضریب تعیین و خطای استاندارد برآورد آمده است. جدول دوم (جدول $ANOVA$) نتیجه آزمون معنی دار بودن مدل و سومین جدول شامل برآورد ضرائب، خطای استاندارد آنها، آماره آزمون معنی دار بودن این ضرائب و نیز $P\text{-}value$ی این آزمون است. مطابق این تحلیل مدل لگاریتمی برازش شده به این دو متغیر به صورت زیر است.

$$y = -2.734 + 6.571\ Lnx$$

$p - value$ آزمون معنی دار بودن این مدل در جدول $ANOVA$ تا سه رقم اعشار صفر است که بیانگر معنی دار بودن مدل در هر کدام از سطوح خطای 0.01 و 0.05 است. همچنین $P\text{-}value$ی آزمون معنی دار بودن هر کدام از ضرائب در سومین

جدول نیز تا سه رقم اعشار صفر است که بیانگر معنی دار بودن وجود هر کدام از این ضرائب در مدل است. ضریب تعیین این مدل $R^2 = 0.517$ ودقت برآورد 2.95 است.

در ادامه نتیجه برازش مدل درجه دو به این سه متغیر آمده است.

Quadratic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.785	.616	.615	2.628

The independent variable is MATH..

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	14439.317	2	7219.658	1045.275	.000
Residual	9006.657	1304	6.907		
Total	23445.974	1306			

The independent variable is MATH..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
MATH.	.141	.080	.162	1.759	.079
MATH. ** 2	.021	.003	.625	6.809	.000
(Constant)	7.582	.467		16.239	.000

در این سه جدول نیز مانند جداول مربوط به مدلهای خطی و لگاریتمی، در اولین جدول ضریب همبستگی، ضریب تعیین و خطای استاندارد برآورد آمده است. جدول دوم (جدول *ANOVA*) نتیجه آزمون معنی دار بودن مدل و سومین جدول شامل برآورد ضرائب، خطای استاندارد آنها، آماره آزمون معنی دار بودن این ضرائب و نیز *P-value*ی این آزمون است. مطابق این تحلیل مدل درجه دو برازش شده به این دو متغیر به صورت زیر است.

$$y = 7.582 + 0.141\,x + 0.021\,x^2$$

$p-value$ آزمون معنی دار بودن این مدل در جدول $ANOVA$ تا سه رقم اعشار صفر است که بیانگر معنی دار بودن مدل در هر کدام از سطوح خطای 0.01 و 0.05 است. همچنین $P\text{-}value$ی آزمون معنی دار بودن ضریب ثابت و ضریب جمله درجه دو در سومین جدول تا سه رقم اعشار صفر است که بیانگر معنی دار بودن وجود هر کدام از این ضرائب در مدل است. اما $P\text{-}value$ی ضریب جمله درجه اول 0.079 است که اگر سطح معنی داری را 0.05 در نظر بگیریم وجود این جمله در مدل معنی دار نیست اما می توان با افزایش سطح معنی داری به 0.08 وجود این جمله در مدل را نیز معنی دار کرد.ضریب تعیین این مدل $R^2 = 0.616$ و دقت برآورد آن 2.63 است.

در ادامه نتیجه برازش مدل درجه دو به این سه متغیر آمده است.

Cubic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.785	.616	.615	2.629

The independent variable is MATH..

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	14439.754	3	4813.251	696.371	.000
Residual	9006.220	1303	6.912		
Total	23445.974	1306			

The independent variable is MATH..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
MATH.	.207	.274	.237	.754	.451
MATH. ** 2	.015	.025	.441	.598	.550
MATH. ** 3	.000	.001	.111	.252	.801
(Constant)	7.387	.905		8.164	.000

در این سه جدول نیز مانند جداول مربوط به مدلهای قبلی، در اولین جدول ضریب همبستگی، ضریب تعیین و خطای استاندارد برآورد آمده است. جدول دوم (جدول $ANOVA$) نتیجه آزمون معنی دار بودن مدل و سومین جدول شامل برآورد ضرائب، خطای استاندارد آنها، آماره آزمون معنی دار بودن این ضرائب و نیز $P\text{-}value$ی این آزمون است. مطابق این تحلیل مدل درجه سه برازش شده به این دو متغیر به صورت زیر است.

$$y = 7.387 + 0.207\,x + 0.015\,x^2 + 0.000\,x^3$$

در این مدل ضریب جمله درجه سه خیلی کوچک بوده و تا سه رقم اعشار صفر است. $p - value$ آزمون معنی دار بودن این مدل در جدول $ANOVA$ تا سه رقم اعشار صفر است که بیانگر معنی دار بودن مدل در هر کدام از سطوح خطای 0.01 و 0.05 است. اما مطابق جدول سوم و با توجه به $P\text{-}value$ی آزمون معنی دار بودن ضرائب در مدل می توان نتیجه گرفت که تنها وجود ضریب ثابت در مدل معنی دار بوده و وجود سایر جملات در مدل معنی دار نیست. بنا بر این این مدل درجه سه مدل مناسبی برای برازش به این دو متغیر نیست. ضریب تعیین این مدل $R^2 = 0.616$ و دقت برآورد آن 2.63 است.

رگرسیون چندگانه

در بحث رگرسیون چندگانه می خواهیم رابطهٔ بین یک متغیر وابسته و چند متغیر مستقل را توسط یک مدل ریاضی نشان دهیم. در ساده ترین حالت این مدل را به صورت خطی در نظر می گیریم. فرض کنید y متغیر وابسته یا متغیر پاسخ که یک متغیر کمی پیوسته و در سطح اندازه گیری لااقل فاصله ای (فاصله ای یا نسبتی) است و نیز فرض کنید $x_k, \dots, x_2, x_1$ متغیرهای مستقل یا متغیرهایی پیشگو باشند که این متغیر ها می توانند متغیرهای پیوسته یا متغیرهای گسسته باشند. فرض می کنیم متغیر وابسته یک متغیر تصادفی و دارای توزیع نرمال است اما متغیرهای مستقل می توانند متغیر تصادفی و یا غیر تصادفی باشند. یک مدل خطی چندگانه مدلی به شکل زیر است.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

که در آن β_i ها ضرایب مدل هستند که این ضرائب پارامتر های ثابتی هستند و هدف از انجام تحلیل رگرسیون برآورد این پارامترها از روی داده های یک نمونه تصادفی است. پس از برآورد این پارامترها، مقدار برآورد شده آنها را با $\hat{\beta}_k, \dots, \hat{\beta}_1, \hat{\beta}_0$ و یا با $b_k, \dots, b_1, b_0$ نشان می دهیم. همچنین مدل برآورد شده را به صورت زیر نشان می دهیم.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

پس از برآورد مدل لازم است معنی دار بودن این مدل و نیز معنی دار بودن وجود هر کدام از این ضرائب در مدل آزمون شود که آزمون معنی دار بودن مدل رگرسیون دارای فرضیه ای به صورت است.

$$\begin{cases} H_0: & \text{مدل رگرسیون معنی دار نیست} \\ H_1: & \text{مدل رگرسیون معنی دار است} \end{cases}$$

و نیز آزمون معنی دار بودن هر کدام از این ضرایب در مدل دارای فرضیه ای به صورت زیر است.

$$\begin{cases} H_0: & \text{وجود } \beta_i \text{ در مدل معنی دار نیست} \\ H_1: & \text{وجود } \beta_i \text{ در مدل معنی دار است} \end{cases}$$

مسیر انجام رگرسیون چندگانه در $spss$ به صورت زیر است.

$$Analyze \rightarrow regression \rightarrow Linear$$

مثلاً فرض کنید نمرات درس ریاضی تابعی از نمرات دروس علوم، حرفه و فن، زبان و فارسی باشد. یعنی نمرات درس ریاضی متغیر پاسخ(y) و نمرات دروس علوم، حرفه و فن، زبان و فارسی متغیرهای پیشگو (x_4،x_3،x_2،x_1) باشد. برای برازش یک مدل چندگانه به این متغیرها پس از رفتن به مسیر فوق متغیر مربوط به نمرات درس ریاضی را به قسمت $Dependent$ و متغیرهای مربوط به نمرات سایر دروس را به قسمت $Independent$ می بریم. پس از اجرای این فرمان نتیجه در قالب چند جدول و در خروجی خواهدآمد که این جداول در زیر آمده است.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841[a]	.708	.706	2.65422

a. Predictors: (Constant), PERSIAN, TECH., SCEIN., FOREI.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14595.714	4	3648.929	517.956	.000[a]
	Residual	6030.400	856	7.045		
	Total	20626.114	860			

a. Predictors: (Constant), PERSIAN, TECH., SCEIN., FOREI.

b. Dependent Variable: MATH.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.347	.485		-2.774	.006
	SCEIN.	.364	.036	.333	10.098	.000
	TECH.	.103	.038	.079	2.740	.006
	FOREI.	.329	.032	.350	10.127	.000
	PERSIAN	.240	.038	.174	6.363	.000

a. Dependent Variable: MATH.

در اولین جدول با عنوان $Mode\ Summary$:

ستون اول با عنوان R ضریب همبستگی چندگانه است. ضریب همبستگی چندگانه شاخصی است که توسط آن میزان همبستگی بین متغیر وابسته و متغیرهای مستقل به طور توام اندازه گیری می شود. در این مثال مقدار این ضریب 0.841 است که بیانگر همبستگی بالا بین متغیر وابسته و متغیرهای مستقل به طور توام می باشد.

ستون بعدی با عنوان R_Square ضریب تعیین مدل است که در این مثال مقدار آن حدود 0.71 است و همانطوریکه از قبل می دانیم این به آن معنی است که 71٪ از تغییرات یا پراکندگی در نمرات درس ریاضی (متغیر پاسخ) توسط نمرات این چهار درس (متغیر های پیشگو) به طور توام تبیین می شود. ضریب تعیین برابر با مجذور ضریب همبستگی چندگانه است.

ستون بعدی با عنوان $Adjusted\ R\ Square$ ضریب تعیین تعدیل شده است. ضریب تعیین تعدیل شده شاخص است صرفاً برای مقایسه ضریب تعیین دو مدل مختلف که دارای تعداد متغیرهای مستقل متفاوتی هستند. به عبارت دیگر برای مقایسه

ضریب تعیین در دو مدل مختلف که این دو مدل دارای متغیر های مستقل متفاوتی هستند از ضریب تعیین تعدیل شده استفاده می شود و کاربرد دیگری ندارد.

آخرین ستون این جدول نیز خطای استاندارد بر آورد است. در این مثال مقدار این شاخص حدود 2.65 شده است. این به آن معنی است که مقدار پیش بینی شده متغیر پاسخ توسط این مدل به طور متوسط حدود 2.65 واحد با مقدار واقعی فاصله دارد.

دومین جدول با عنوان جدول $ANOVA$ جدول تحلیل واریانس مربوط به آزمون معنی دار بودن مدل رگرسیون است. در این مثال P_value این آزمون خیلی کوچک بوده و دست کم تا سه رقم اعشار صفر است. بنابراین فرض صفر مبنی بر معنی دار نبودن این مدل در سطح خطای 1٪ (و در نتیجه 5٪) رد می شود. یعنی مدل معنی دار است.

سومین جدول با عنوان $coefficient$ برآورد ضرایب مدل و نیز نتیجه آزمون معنی دار بودن این ضرایب است. در این جدول ستون اول به عنوان B همان برآورد ضرایب مدل است که در این مثال داریم $\hat{\beta}_0 = -1.347$ ، $\hat{\beta}_1 = 0.364$ ، $\hat{\beta}_2 = 0.103$ ، $\hat{\beta}_3 = 0.329$ ، $\hat{\beta}_4 = 0.240$ بنابراین برآورد مدل رگرسیون به شکل زیر است.

$$\hat{y} = -1.35 + 0.36\,x_1 + 0.10\,x_2 + 0.33\,x_3 + 0.24\,x_4$$

در یک مدل رگرسیون چندگانه ضریب یک متغیر مشخص می کند که با فرض ثابت بودن سایر متغیرها هر یک واحد تغییر در آن متغیر مستقل باعث ایجاد چقدر تغییر در متغیر وابسته خواهد شد. مثلاً در این مثال ضریب X_1 حدوداً 0.36 است این به آن معنی است که با فرض ثابت بودن سایر متغیرها هر یک واحد افزایش در نمره درس علوم (x_1) باعث ایجاد 0.36 افزایش در نمره درس ریاضی خواهد شد. اگر علامت یک ضریب منفی باشد به معنی کاهش متغیر وابسته به ازای افزایش آن متغیر مستقل خواهد بود. مثلا اگر ضریب x_1 عدد منفی 0.36- بود نتیجه می شد که با فرض ثابت بودن سایر متغیرها هر یک واحد افزایش در نمره درس علوم (x_1) باعث ایجاد 0.36 کاهش در نمره درس ریاضی خواهد شد.

دومین ستون با عنوان $Std\ error$ خطای استاندارد این برآورد هاست.

سومین ستون با عنوان $Beta$ برآورد پارامترهای استاندارد شده است. در محاسبه این برآوردها ابتدا داده ها استاندارد شده و سپس برآورد پارامترها از روی داده های استاندارد شده محاسبه می شود. از این برآوردها برای مقایسه اثر متغیرهای پیشگو بر متغیر پاسخ با یکدیگر استفاده می شود. به عبارت دیگر برای مقایسه اثر متغیر های مستقل بر متغیر وابسته از ضرایب این

متغیرها نمی توان استفاده کرد چون ممکن است متغیرهای مستقل دارای واحدهای اندازه گیری متفاوتی باشند که این واحدها در محاسبه ضرایب متغیر مؤثر است. مثلاً در یک مدل رگرسیون ممکن است ضریب یکی از متغیرهای مستقل از سایر ضرایب بزرگتر باشد حال آنکه این متغیر نسبت به سایر متغیرهای مستقل اثر کمتری بر متغیر وابسته داشته باشد. به همین دلیل و به جهت حذف اثر واحد اندازه گیری داده ها در بر آورد ضرایب ابتدا داده ها را استاندارد کرده و سپس ضرایب بر آورد می شود که این ضرایب استاندارد شده قابل مقایسه با یکدیگر هستند. یعنی اگر ضریب استاندارد شده یک متغیر از سایر ضرایب بزرگتر باشد می توان نتیجه گرفت که آن متغیر اثر بیشری بر متغیر وابسته دارد. در محاسبه ضرایب استاندارد شده همواره مقدار ثابت وجود نخواهد داشت یعنی برآورد مقدار ثابت استاندارد شده همواره صفر خواهد بود. در این مثال با توجه به ضرایب استاندارد شده می توان نتیجه گرفت که نمره درس زبان بیشترین تأثیر را بر نمره ریاضی دارد و پس از آن به ترتیب نمرات دروس علوم، فارسی و حرفه و فن قرار دارند. حال آنکه اگر از برآورد ضرائب غیر استاندارد شده استفاده شود نتیجه می شود که نمره درس علوم بیشترین تاثیر را بر نمره درس ریاضی دارد که این نتیجه اشتباه است.

ستون بعدی با عنوان t آماره آزمون معنی دار بودن این ضرایب در مدل است و آخرین ستون این جدول با عنوان sig ، P_value ی آزمون معنی داری هر کدام از این ضرایب در مدل است که در این مثال همانطور که مشاهده می شود کلیه این P_value ها از 0.01 کوچکتر بوده و بر این اساس فرض صفر مبنی بر معنی دار نبودن این ضرائب برای کلیه این ضرائب در سطح خطای 0.01 (و در نتیجه در سطح خطای 0.05) رد می شود. بر این اساس وجود همهٔ این ضرایب در مدل معنی دار است.

به طور خلاصه نتیجه ای که از انجام این تحلیل رگرسیونی به دست آمده آن است که رابطه بین نمرات درس ریاضی (بعنوان متغیر وابسته) و نمرات سایر دروس (به عنوان متغیرهای مستقل) را می توان به شکل مدل ریاضی زیر نشان داد .

$$\hat{y} = -1.35 + 0.36\,x_1 + 0.10\,x_2 + 0.33\,x_3 + 0.24\,x_4$$

که این مدل معنی دار است و نیز وجود هر کدام از ضرایب در مدل معنی دار است. ضریب تعیین مدل $R^2 = 0.71$ است یعنی 71 درصد از تغییرات یا پراکندگی در نمرات درس ریاضی توسط مدل فوق به نمرات این چهار درس بستگی دارد. بنابراین با معلوم بودن نمرات این چهار درس برای یک دانش آموز و قرار دادن آنها در مدل فوق می توان نمره درس ریاضی آن دانش آموز را پیش بینی کرد که خطای استاندارد این پیش بینی 2.65 است. مثلا فرض کنید دانش آموزی در دروس علوم، حرفه وفن،

زبان و فارسی به ترتیب نمرات 16، 15، 8.75 و 13 گرفته باشد. در اینصورت با قرار دادن این مقادیر به جای متغیرهای x_4, x_3, x_2, x_1 در معادله فوق مقدار $\hat{y}$ برابر خواهد شد با 11.93 . یعنی پیش بینی می شود که نمره ریاضی این دانش آموز 11.93 شود و نیز نمره واقعی ریاضی این دانش آموز به طور متوسط 2.65 بیشتر یا کمتر از 11.93 خواهد بود.

توجه 1:

در یک مدل چندگانه اگر وجود برخی از متغیرهای مستقل در مدل معنی دار نبود یعنی باید آن متغیرها را از مدل خارج کرده و مجدداً مدل دیگری بر متغیرهای باقی مانده برازش کرده و معنی دار بودن متغیرها مجددا آزمون شود. اما در یک مدل رگرسیونی صرفنظر از اینکه وجود ثابت (β_0) در مدل معنی دار بوده یا معنی دار نباشد همواره باید ثابت معادله یا همان β_0 در مدل وجود داشته باشد. مگر آنکه دلیل نظری قوی مبنی بر عدم وجود آن در مدل وجود داشته باشد.

توجه 2:

پس از برازش مدل رگرسیون و با قرار دادن مقادیر متغیرهای مستقل در این مدل مقادیری برای متغیر وابسته بدست می آید که آنها را مقادیر پیش بینی شده یا $Prediet\ value$ گفته و با $\hat{y}_i$ نشان می دهیم. اختلاف مقادیر پیش بینی شده با مقادیر واقعی متغیر وابسته را باقیمانده یا مانده یا $Residual$ گوییم و آنها را با e_i نشان می دهیم یعنی $e_i = y_i - \hat{y}_i$ که اگر مدل، مدل مناسبی باشد انتظار داریم e_i *ها کوچک باشد.*

توجه 3 (تعیین سهم هر متغیر در پراکندگی کل):

همانطوریکه می دانیم ضریب تعیین بیانگر میزان از تغییرات یا پراکندگی در متغیر وابسته است که توسط متغیرهای مستقل تبیین می شود و در یک رگرسیون چندگانه ضریب تعیین در واقع مجذور ضریب همبستگی چندگانه است. با محاسبه برخی ضرائب همبستگی و مجذور کردن آنها می توان سهم هر کدام از متغیرهای پیشگو را در پراکندگی کل تعیین کرد. بدین منظور کافی است در مسیر رگرسیون و در پنجره Linear Regression بر روی تب Statistics کلیک کرده و در پنجره ظاهر شده گزینه Part and partial correlations را تیک بزنیم. در خروجی و در جدول coefficients سه ستون با نام correlations داده خواهد شد. ستون اول با نام Zero-order ستون دوم با نام Partial و ستون سوم با نام Part

خواهد بود. ستون Zero-order ضرائب همبستگی مرتبه صفر است. این ضریب در واقع ضریب همبستگی ساده بین هرکدام از متغیرهای پیشگو با متغیر پاسخ است که در تحلیل رگرسیون چندگانه کاربردی ندارد. ستون دوم با نام Partial ضریب همبستگی جزئی بین هر متغیر پیشگو با متغیر پاسخ است که سایر متغیرهای پیشگو به عنوان متغیر کنترل در نظر گرفته شده اند. راجع به ضریب همبستگی جزئی و کاربرد آن قبلا بطور مفصل بحث شده است. با مجذور کردن این ضرائب می توان سهم هر متغیر در پراکندگی کل را پس از حذف اثر سایر پیشگوها تعیین کرد. مثلا فرض کنید در یک رگرسیون چندگانه با چهار متغیر پیشگوی X1,X2,X3,X4 و متغیر پاسخ Y ضریب همبستگی جزئی X1 عدد 0.32 به دست آمده باشد. مجذور این عدد حدودا 0.10 خواهد بود و این به آن معنی است که 10 درصد از پراکندگی باقیمانده در متغیر پاسخ را که سایر متغیرهای پیشگو نتوانسته اند تبیین کنند توسط متغیر پیشگوی X1 تبیین می شود. هرچقدر این ضریب برای متغیری بزرگتر باشد بیانگر تاثیر بیشتر آن متغیر در پراکندگی کل است. ستون سوم با نام Part ضریب همبستگی مولفه هرکدام از متغیرهای پیشگو است. ضریب همبستگی مولفه که آن را ضریب همبستگی نیمه جزئی (Semi-Partial Correlation) نیز می نامند بیانگر همبستگی منحصر به فرد یک متغیر پیشگو با متغیر پاسخ است که از مجذور کردن آن سهم آن متغیر در پراکندگی کل متغبر پاسخ به دست می آید.

پیش فرض های رگرسیون (فرضیه های اولیه)

در انجام یک تحلیل رگرسیونی بعضی فرضیه های اولیه وجود دارد که این فرضیه ها باید برقرار باشد و در صورت عدم برقراری این فرضیه ها اعتبار مدل رگرسیون زیر سؤال می رود این پیش فرض ها به صورت زیر است.

1. متغیر وابسته یا همان متغیر پاسخ یک متغیر تصادفی پیوسته در سطح اندازه گیری لااقل فاصله ای و توزیع آن نرمال است.
2. متغیرهای مستقل یا همان متغیرهای پیشگو متغیرهایی تصادفی یا غیر تصادفی و نا هم بسته اند.
3. باقیمانده های مدل (e_i ها) متغیرهای تصادفی دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 هستند. یعنی $e_i \sim N(0\,,\sigma^2)$ و نیز این باقیمانده ها دو به دو نا همبسته اند یعنی $Cov(e_i, e_j) = 0$.

بررسی پیش فرض های رگرسیون

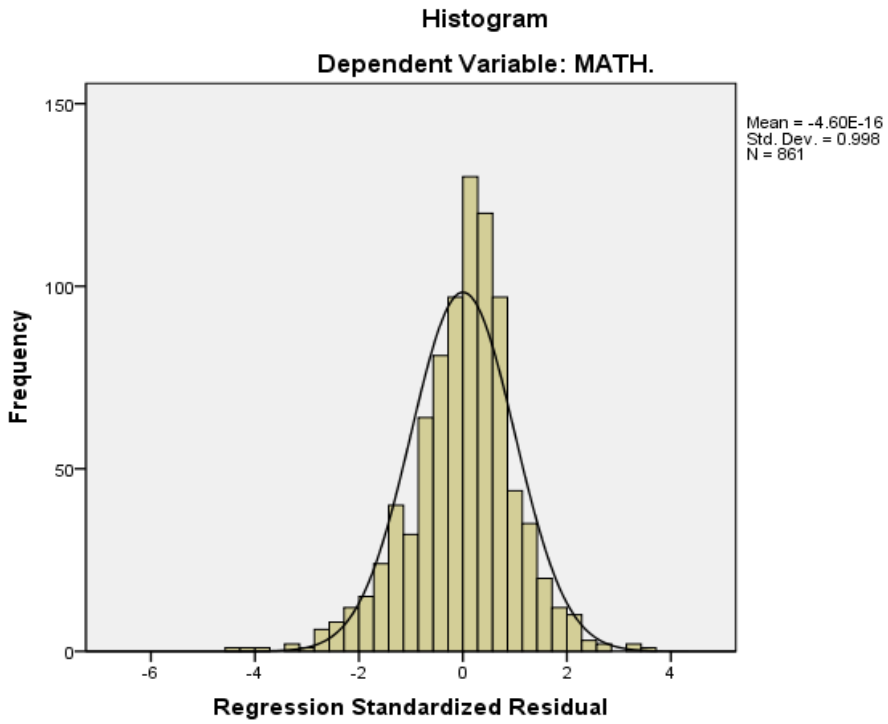
1 . بررسی فرض نرمال بودن: برای بررسی فرض نرمال بودن توزیع متغیر پاسخ کافی است نرمال بودن توزیع باقیمانده ها مورد بررسی قرار گیرد چون اگر توزیع باقیمانده ها نرمال باشد توزیع متغیر پاسخ نیز نرمال خواهد بود. برای بررسی فرض نرمال بودن توزیع باقیمانده ها هم می توان هیستوگرام و نمودار احتمال نرمال ($p-p\ plot$) باقیمانده های استاندارد شده را رسم کرد که در چنین صورتی فرض نرمال بودن این متغیر به صورت چشمی مورد بررسی قرار می گیرد. علاوه بر رسم نمودار می توان از آزمونهای آماری نیز کمک گرفت.

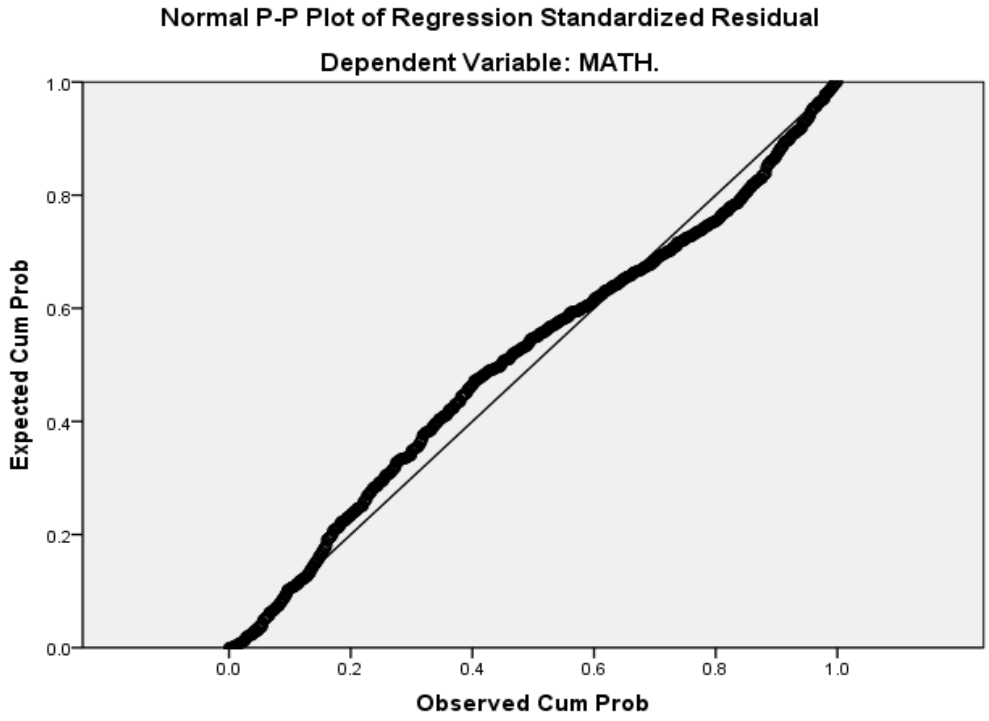
بررسی نموداری فرض نرمال بودن توزیع باقیمانده ها:

برای بررسی فرض نرمال بودن توزیع باقیمانده ها هم می توان از هیستوگرام و هم می توان از نمودار احتمال نرمال ($p-p\ plot$) باقیمانده های استاندارد شده استفاده کرد.

نمودار احتمال نرمال ($p-p\ plot$) نموداری است که در آن مقادیر تابع توزیع تجربی نمونه در مقابل تابع توزیع نرمال در صفحه رسم می شود اگر توزیع متغیر مورد نظر نرمال باشد این نقاط بر روی نیمساز ربع اول قرار می گیرند. هر چقدر این نقاط از نیمساز ربع اول انحراف بیشتری داشته باشد بیانگر عدم نرمال بودن توزیع آن متغیر است.

برای رسم هستیوگرام و نمودار $p-p\ plot$ باقیمانده های استاندارد شده در پنجره رگرسیون خطی $Linear\ regression$ بر روی گزینه $plot$ کلیک و در پنجره ظاهر شده و در بخش $Standardized\ Residual\ Plots$ می توان تعیین کرد که این دو نمودار در خروجی رسم شود این دو نمودار در مدل مثال قبل رسم شده و در زیر آمده است.


Histogram
Dependent Variable: MATH.
Mean = -4.60E-16
Std. Dev. = 0.998
N = 861
Frequency
150
100
50
0
-6
-4
-2
0
2
4
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: MATH.
Expected Cum Prob
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Observed Cum Prob

اولین نمودار هیستوگرام و منحنی فراوانی باقیمانده های استاندارد شده است که چون حجم داده ها خیلی زیاد است به نظر می رسد این نمودار شبیه به منحنی نرمال است. البته کشیدگی آن بیشتر از نرمال به نظر می رسد و نیز به جهت حجم زیاد داده ها نمی تواند به خوبی چولگی داده ها را نشان دهد.در نمودار $p-p\ plot$ و با توجه به اینکه حجم داده ها خیلی زیاد است می توان تشخیص داد که انحراف این نقاط از نیمساز ربع اول کمی زیاد است و بنابراین فرض نرمال بودن توزیع باقیمانده ها به طور چشمی قابل قبول نیست.

بررسی فرض نرمال بودن توزیع باقیمانده ها توسط آزمونهای آماری:

علاوه بر بررسی نموداری توزیع باقیمانده ها می توان فرض نرمال بودن آنها را توسط برخی آزمونهای آماری مانند آزمون ناپارامتری با نام **آزمون نیکویی برازش کولموگروف–اسمیرنف** و نیز **آزمون شاپیرو ویلک** انجام داد. آزمون شاپیرو–ویلک قویتر از کولموگروف– اسمیرنف است. در هر کدام از این آزمونها فرضیه مورد بررسی به صورت زیر است.

$$\begin{cases} H_0 : \text{توزیع متغیر مورد نظر نرمال است} \\ \mathrm{H}_1 : \text{توزیع متغیر مورد نظر نرمال نیست} \end{cases}$$

برای انجام آزمون نرمال بودن توزیع یک متغیر توسط آزمون نیکویی برازش کولموگروف– اسمیرنف و نیز آزمون شاپیرو–ویلک از مسیر زیر استفاده می کنیم:

$$Analyze \rightarrow Descriptive\ statistics \rightarrow Explore$$

در این مثال برای بررسی نرمال بودن توزیع باقیمانده ها و باقیمانده های استاندارد شده لازم است ابتدا آنها را به عنوان متغیرهای جدید در فایل داده ذخیره کرده و سپس نرمال بودن آنها را مورد آزمون قرار می دهیم. لازم به ذکر است که برای آزمون نرمال بودن توزیع باقیمانده ها می توان از باقیمانده های استاندارد شده استفاده کرد و در هر صورت نتیجه یکسان است یعنی اگر توزیع باقیمانده ها نرمال باشد، توزیع باقیمانده های استاندارد شده نیز نرمال خواهد بود و بالعکس. برای ذخیره کردن باقیمانده ها و باقیمانده های استاندارد شده در فایل داده در پنجره رگرسیون خطی بر روی گزینه $save$ کلیک کرده و در پنجره ظاهر شده و در قسمت Residuals تعیین می کنیم که باقیمانده ها و باقیمانده های استاندارد شده ذخیره شود. پس از ذخیره کردن باقیمانده ها و باقیمانده های استاندارد شده برای بررسی نرمال بودن آنها به مسیر فوق الذکر رفته و در پنجره ظاهر شده

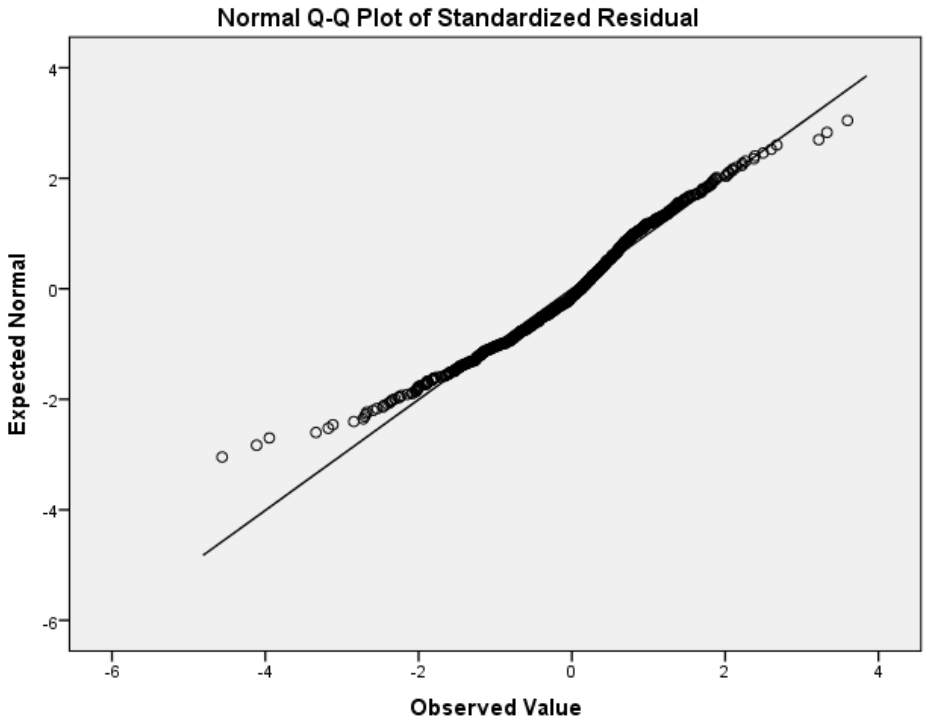
متغیر یا متغیرهائی که می خواهیم نرمال بودن توزیع آنها را آزمون کنیم (در اینجا متغیر باقیمانده های استاندارد شده) را به قسمت Dependent List می بریم. سپس بر روی تب Plots کلیک کرده و در پنجره ظاهر شده گزینه Normality plots with tests را تیک می زنیم. پس از اجرای این فرمان شاخصهای آماری این متغیر، نمدار ساقه و برگ، نمودار جعبه ای، نموداری با نام نمودار Q-Q Plot و نیز نتیجه آزمون نرمالیتی در خروجی داده خواهد شد. نتیجه آزمون نرمالیتی در جدولی به صورت زیر در خروجی خواهد آمد.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.069	861	.000	.977	861	.000

a. Lilliefors Significance Correction

در این جدول سه ستون اول آن نتیجه آزمون کولموگروف–اسمیرنف و سه ستون بعدی آن نتیجه آزمون شاپیرو–ویلک است. در هر کدام از این دو آزمون ستون آخر با عنوان .Sig همان p-value های این دو آزمون است که در این مثال p-value برای هر دو آزمون تا سه رقم اعشار صفر بوده و بنابراین فرض نرمال بودن توزیع باقیمانده های استاندارد شده توسط هر دو آزمون رد می شود. این نتیجه از نمودار Q-Q plot که در خروجی اجرای این فرمان آمده است نیز به دست می آید. نمودار Q-Q plot شبیه به نمودار P-P plot است با این تفاوت که در این نمودار صدکهای به دست آمده در توزیع تجربی نمونه در مقابل صدکهای توزیع نرمال رسم می شود. اگر توزیع متغیر نرمال باشد باید نقاط نمونه بر روی نیمساز ربع اول و سوم قرار گیرد. هر چقدر انحراف نقاط از نیمساز بیشتر باشد بیانگر انحراف توزیع آن متغیر از توزیع نرمال است. نمودار Q-Q plot در این مثال به صورت زیر است.



در این نمودار و با توجه به حجم زیاد داده ها می توان مشاهده کرد که انحراف نقاط از نیمساز زیاد بوده و بر این اساس فرض نرمال بودن توسط این نمودار نیز رد می شود.

تبدیلات توانی (تبدیلات باکس-کاکس)

اما اگر توزیع باقیمانده ها نرمال نباشد باید با انجام یک تبدیل مناسب بر روی متغیر وابسته توزیع آنرا نرمال کرد. بدین منظور چولگی باقیمانده ها را محاسبه می کنیم. اگر باقیمانده های چوله به راست باشند (چولگی مثبت) اکثراً انجام تبدیل لگاریتمی باعث نرمال شدن آنها می شود یعنی در چنین حالتی به جای اینکه از متغیر وابسته y در مدل استفاده کنیم از لگاریتم آن یعنی Lny در مدل استفاده می کنیم که در چنین صورتی مدل رگرسیون به شکل زیر تبدیل می شود.

$$Lny = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

و اگر باقیمانده ها چوله به چپ باشد (چولگی منفی) اکثراً تبدیل y^2 باعث نرمال شدن توزیع باقیمانده ها می شود.یعنی در چنین حالتی به جای استفاده از متغیر وابسته y از توان دوم آن یعنی y^2 در مدل استفاده می کنیم که در چنین صورتی مدل رگرسیون به صورت زیر تبدیل می شود.

$$y^2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

به طور کلی اگر توزیع متغیری مانند y نرمال نبوده و چوله به چپ باشد انجام تبدیل y^λ که $\lambda > 0$ توزیع آنرا نرمال می کند و اگر چولگی به راست باشد تبدیل y^λ که $\lambda \leq 0$ توزیع آنرا نرمال می کند. و توجه کنید $\lambda=0$ معادل تبدیل لگاریتمی است یعنی *Lny* .

گاهی اوقات نیز عدم نرمال بودن توزیع باقیمانده ها به جهت وجود داده های دور افتاده است. برای شناسایی داده های دور افتاده کافی است مشاهداتی که باقیمانده استاندارد شده آنها خارج از فاصله (3+ ، 3-) و به خصوص خارج از فاصله ی (4+ ، 4-) است را شناسایی کرده و آنها را به عنوان داده دور افتاده از تحلیل خارج کنیم. البته همیشه حذف دور افتاده ها جایز نیست و گاهی اوقات چنین داده هایی نباید حذف شود و در چنین مواردی باید از مدلهای غیر خطی استفاده کرد.

مثال: در مثال قبل همانطوری که مشاهده شد فرض نرمال بودن توزیع باقیمانده ها رد شد حال با محاسبه چولگی باقیمانده های استاندارد شده و انجام یک تبدیل مناسب بر روی متغیر وابسته نرمال بودن توزیع باقیمانده ها را مجدداً بررسی می کنیم. با محاسبه چولگی باقیمانده های استاندارد شده مشاهده می شود میزان چولگی 0.47- بوده و بنابراین باقیمانده ها چوله به چپ است و لذا در این مدل به جای y از y^λ که $\lambda > 0$ است استفاده می کنیم. با قرار دادن مقادیر مختلف مثبت برای λ و به روش آزمون و خطا بهترین مقدار برای λ بطوریکه مدل و کلیه متغیرها در مدل معنی دار بوده و به علاوه توزیع باقیمانده ها نیز نرمال باشد، مقدار 3.2 برای λ به دست آمد. بنابراین برای $\lambda = 3.2$ این مدل را برازش کرده و فرض نرمال بودن باقیمانده های آنرا آزمون می کنیم. بدین منظور ابتدا توسط فرمان *Compute variable* در مسیر *Transform → Computr variable* متغیر جدیدی ایجاد می کنیم که این متغیر توان 3.2 متغیر مربوط به نمرات درس ریاضی دانش آموزان (*x4*) است. نام این متغیر را *Y3.2* می گذاریم. سپس مجددا رگرسیون چندگانه قبل را انجام می دهیم با این تفاوت که اینبار متغیر وابسته متغیر *Y3.2* ایجاد شده است. نتیجه انجام این رگرسیون در زیر آمده است.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.852[a]	.726	.725	2610.82424

a. Predictors: (Constant), PERSIAN, TECH., SCEIN., FOREI.

b. Dependent Variable: Y3.2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15498119980.887	4	3874529995.222	568.413	.000[b]
	Residual	5834841129.498	856	6816403.189		
	Total	21332961110.385	860			

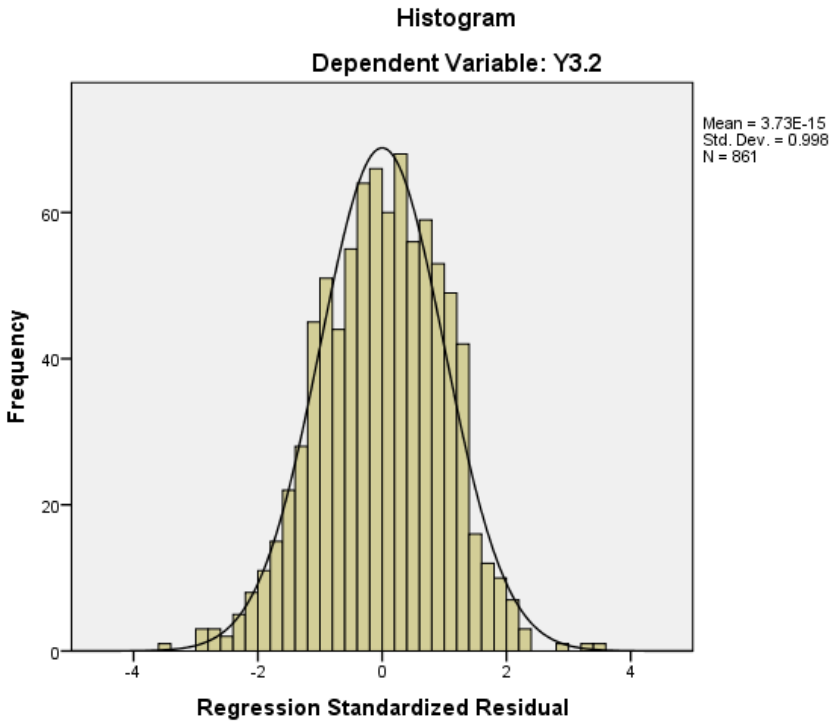
a. Dependent Variable: Y3.2

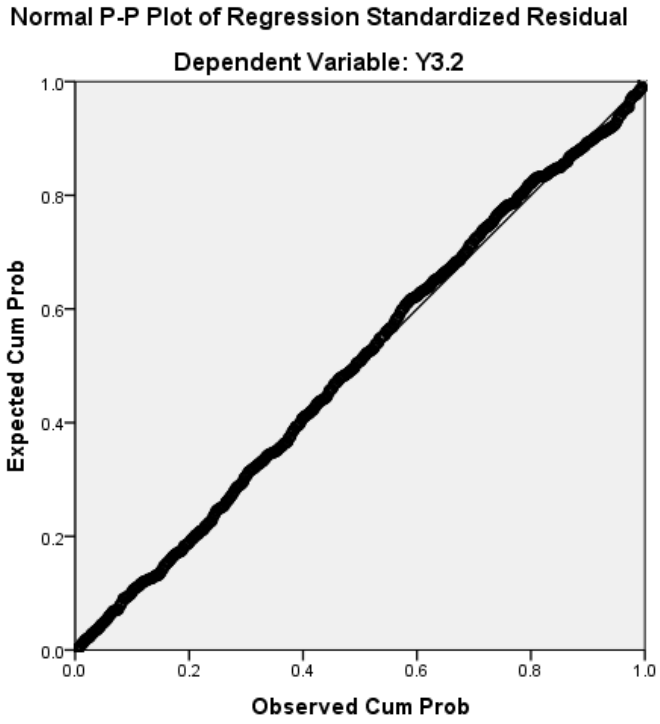
b. Predictors: (Constant), PERSIAN, TECH., SCEIN., FOREI.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9617.035	477.545		-20.138	.000
	SCEIN.	327.960	35.422	.295	9.259	.000
	TECH.	98.190	37.094	.074	2.647	.008
	FOREI.	343.815	31.931	.360	10.767	.000
	PERSIAN	312.922	37.132	.222	8.427	.000

a. Dependent Variable: Y3.2


Histogram
Dependent Variable: Y3.2
Mean = 3.73E-15
Std. Dev. = 0.998
N = 861
Frequency
0
20
40
60
-4
-2
0
2
4
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Y3.2
Expected Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0
Observed Cum Prob

مطابق با تحلیل انجام شده مدل رگرسیونی زیر به دست آمده است.

$$\hat{y}^{3.2} = -96.17 + 327.96\,X_5 + 98.19\,X_6 + 343.82\,X_7 + 312.92\,X_8$$

که در آن $\hat{y}$ بیانگر نمره درس ریاضی و X_5 ، X_6 ، X_7 و X_8 به ترتیب بیانگر نمرات دروس علوم، حرفه وفن، زبان و فارسی است. با توجه به *P-value* ی به دست آمده برای آزمونهای معنی دار بودن مدل و نیز معنی دار بودن ضرائب در مدل نتیجه می شود که این این مدل معنی دار بوده و نیز وجود همه ضرائب در مدل نیز معنی دار است. ضریب تعیین این مدل 0.73 است. که نسبت به مدل خطی قبل 2 درصد افزایش داشته است. هیستوگرام و نمودار *p-p plot* باقیمانده های استاندارد شده به طور چشمی فرض نرمال بودن باقیمانده ها را تائید می کنند. علاوه بر این نمودارها، فرض نرمال بودن توزیع باقیمانده های استاندارد شده توسط آزمون نیکوئی برازش کولموگروف- اسمیرنف و آزمون شاپیرو-ویلک نیز آزمون شده است که نتیجه در زیر آمده است.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.029	861	.075	.996	861	.031

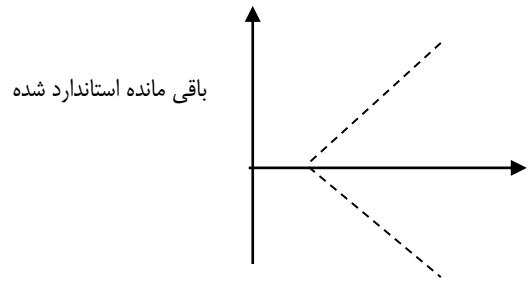
a. Lilliefors Significance Correction

همانطوریکه مشاهده می شود *p-value* ی مربوط به آزمون کولموگروف-اسمیرنف از 0.05 و در مورد آزمون شاپیرو-ویلک از 0.01 بزرگتر است بر این اساس فرض نرمال بودن توزیع باقیمانده های استاندارد شده توسط یکی از آزمونها در سطح خطای 0.05 و توسط آزمون دیگر در سطح خطای 0.01 پذیرفته می شود.

بررسی فرض ثابت بودن واریانس:

همانطوری که قبلاً بیان شد یکی دیگر از پیش فرض های انجام تحلیل رگرسیون فرض ثابت بودن واریانس است. برای بررسی این فرضیه کافی است نمودار پراکنش مانده های استاندارد شده در مقابل متغیرهای پیشگو و نیز نمودار پراکنش مانده های

استاندارد شده در مقابل مقادیر پیش بینی شده($\hat{y}$ ها $Predicted\ value$) را رسم کنید. اگر این نمودار ها هیچ طرح خاصی را نشان ندهد و پراکندگی نقاط تصادفی بود دال بر ثابت بودن واریانس است. اما اگر طرح خاصی مشاهده شود مثلاً با تغییر مقادیر متغیر های پیشگو باقیمانده ها افزایش یا کاهش یابد یعنی این نمودار قیفی شکل شود در اینصورت فرض ثابت بودن واریانس مورد تردید قرار می گیرد. مثلا نموداری به شکل زیر:



در چنین حالتی می توان:

1. نقاط دور افتاده را تشخیص داده و آنها را خذف کرد.
2. از تبدیلات تثبیت واریانس استفاده کرد که شکل ساده این تبدیلات به صورت y^{λ} است که در آن $\lambda = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$ می باشد. در اینجا $\lambda=0$ متناظر با تبدیل لگاریتمی یعنی Lny است.

بررسی فرض نا همبسته بودن باقیمانده ها:

همانطور که در قبل بیان شد یکی دیگر از پیش فرض های انجام تحلیل رگرسیون ناهمبسته بودن باقی مانده ها است. یعنی کوواریانس دو به دوی آنها باید مساوی صفر شود. البته این فرضیه تنها در مورد داده هایی است که به طور تدریجی و در طی یک دوره زمانی جمع آوری شده باشد مانند داده های بورس (یعنی سری زمانی باشد) و در مورد داده هایی که در تحقیق های میدانی و به صورت مقطعی ($Sectional$) جمع آوری می شود نیازی به بررسی این فرضیه نیست. در مورد داده های سری زمانی اگر این فرضیه برقرار نباشد گوییم داده ها دارای خود همبستگی ($Auto\ Correlation$) است. برای بررسی این فرضیه از آماره دوربین واتسون استفاده می شود. برای محاسبه این آماره در مسیر رگرسیون خطی بر روی گزینه $Statistics$ کلیک کرده و در پنجره ظاهر شده و در بخش $Residuals$ گزینه $Durbin\text{-}Watson$ را انتخاب می کنیم. مقدار این آماره در فاصله 0

تا 4 قرار دارد که اگر مقدار آن در بازه*1.5* تا *2.5* باشد فرض عدم همبستگی بین باقیمانده ها پذیرفته می شود. اما اگر باقیمانده ها دارای خود هم بستگی باشند در این صورت :

1. از تابع تأخیر زمانی (Lag) متغیر وابسته به جای آن متغیر در مدل استفاده می شود.
2. از تابع تأخیر زمانی (Lag) متغیرهای پیشگو، به جای آن متغیر ها در مدل استفاده می شود.

مثال: در مثال قبل آماره دوربین واتسون را محاسبه کرده و در صورت وجود خود همبستگی توسط تابع تاخیر زمانی خود همبستگی را از بین ببرید.

بررسی فرض استقلال یا عدم وجود هم خطی بین متغیرهای پیشگو:

یکی دیگر از پیش فرض های تحلیل رگرسیون آن است که متغیرهای پیشگو مستقل از هم باشند. گو اینکه در اکثر موارد این فرضیه بر قرار نبوده و این متغیرها دارای ارتباطی با یکدیگر هستند که اگر این رابطه به صورت وجود یک رابطه خطی بین آنها باشد رگرسیون را شدیداً تحت تاثیر قرار می دهد. در چنین موردی گویند بین متغیرهای پیشگو همخطی وجود دارد. در حالتی که بین متغیرهای پیشگو هم خطی وجود داشته باشد آنگاه ممکن است ضرایب رگرسیون معنی دار نبوده حال آنکه ضریب تعیین عددی بزرگ باشد و اگر این همخطی کامل باشد یعنی دست کم یکی از متغیرهای پیشگو یک ترکیب خطی از سایر متغیرها باشد آنگاه در چنین حالتی محاسبه ضرایب به روش کمترین مربعات امکان پذیر نخواهد بود. بنابراین بین متغیرهای پیشگو نباید همخطی وجود داشته باشد. برای تشخیص هم خطی بین متغیرهای پیشگو از آماره های $Tolerance$ (تحمل) و VIF (عامل تورم واریانس) استفاده می شود.

$Tolerance$ بین صفر و یک در نوسان است هر چقدر T یک متغیر به صفر نزدیکتر باشد بیانگر آن است که آن متغیر را می توان توسط یک ترکیب خطی از سایر متغیرها به دست آورد و لذا بین آن متغیر با سایر متغیرها همخطی وجود دارد. و هر چقدر T به یک نزدیکتر باشد به معنی نا همبسته بودن آن متغیر با سایر متغیرهاست. نقطه برش را معمولاً 0.5 در نظر می گیرند یعنی اگر آماره T یک متغیر از 0.5 بیشتر باشد ($T>0.5$) به معنای عدم وجود هم خطی بین آن متغیر با سایر متغیرهاست و در غیر این صورت بین آن متغیر و سایر متغیرها همخطی وجود دارد. VIF عکس T است بنابراین اگر VIF کمتر از 2 باشد ($VIF<2$) به معنای عدم وجود همخطی است و در غیر اینصورت به معنی وجود همخطی بین آن متغیر با سایر متغیرها است.

برخی مولفین به جای 0.5 عدد 0.1 را پیشنهاد داده اند. بنابراین مطابق با این دیدگاه اگر آماره تلرانس برای یک متغیر از 0.1 بیشتر باشد به معنای عدم وجود هم خطی بین آن متغیر با سایر متغیرهاست و در غیر این صورت بین آن متغیر و سایر متغیرها همخطی وجود دارد و به طور مشابه در این حالت اگر VIF کمتر از 10 باشد ($VIF<10$) به معنای عدم وجود همخطی است و در غیر اینصورت به معنی وجود همخطی بین آن متغیر با سایر متغیرها است.

برای محاسبه این دو آماره در مسیر رگرسیون خطی بر روی گزینه *Statistics* کلیک کرده و در پنجره ظاهر شده گزینه *Collinearity diagnostics* را انتخاب می کنیم. در اینصورت برای هر کدام از متغیرهای پیشگو این دو ضریب محاسبه شده و در خروجی داده خواهد شد.

اگر بین یک متغیر با سایر متغیرها هم خطی وجود داشته باشد در این صورت :

1. آن متغیر را از مدل خارج می کنیم.
2. حجم نمونه را افزایش می دهیم (البته اگر همخطی کامل باشد این عمل بی تاثیر است).
3. از روش گام به گام ($Step\ wise$) در ورود متغیرها به مدل استفاده می کنیم.
4. از رگرسیون ستیغی ($Ridge$) استفاده می کنیم.
5. با استفاده از تحلیل مولفه اصلی (PC) متغیرهای مستقل را با یکدیگر ترکیب می کنیم.

کشف مشاهدات دور افتاده :

در حالت چند متغیره و بخصوص در رگرسیون چند متغیره یکی از روشهای کشف مشاهدات دور افتاده استفاده از شاخصهای فاصله است که یکی از این شاخصها شاخص فاصله ماهالا نوبیس موسوم به شاخص M است. در این حالت برای هر مشاهده ای این شاخص محاسبه شده و در یک متغیر جدید در فایل قرار می گیرد. سپس توسط فرمان Explore و با تیک زدن Outlier در بخش Statistics می توان مشاهداتی که نسبت به سایرین فاصله زیادی دارند را تعیین کرد. توسط این فرمان تعداد پنج مشاهده که دارای کمترین فاصله و نیز تعداد پنج مشاهده که دارای بیشترین فاصله هستند مشخص خواهد شد. در

تحلیل رگرسیون داده های پرت داده هائی هستند که دارای بیشترین فاصله با سایر داده ها هستند. حال برای تعیین اینکه آیا این پنج مقدار دور افتاده هستند یا خیر کافی است هرکدام از این فاصله ها را با مقدار بحرانی توزیع کی دو در سطح خطای 0.05 یا 0.01 و با درجه آزادی k که در آن k تعداد متغیرهای مستقل است مقایسه کرد. فاصله هائی که از این مقدار بزرگتر باشند را می توان به عنوان داده های دور افتاده در نظر گرفت. در جدول زیر مقدار بحرانی توزیع کی دو در سطوح خطای 0.05 و 0.01 و برای درجات آزادی 1 الی 10 آمده است.

مقادیر بحرانی توزیع کی دو در سطوح 0.05 و 0.01 برای درجات آزادی 1 الی 10

درجه آزادی	0.05	0.01
1	3.84	6.64
2	5.99	9.21
3	7.82	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.47
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

علاوه بر روش فوق در تعیین داده های دورافتاده لازم است پس از انجام تحلیل رگرسیون و بخصوص هنگامی که برخی از فرضیات پایه ای رگرسیون نقض شده است مشاهدات دور افتاده رگرسیونی را نیز تشخیص داده و در صورت لزوم آن مشاهدات را از تحلیل رگرسیون خارج کرد. توصیه می شود ابتدا و قبل از انجام تحلیل رگرسیون توسط فاصله ماهالانوبیس مشاهدات دور افتاده را تشخیص داده و مدل رگرسیون را بدون دورافتاده ها انجام داد. پس از انجام تحلیل رگرسیون و بخصوص در صورت نقض برخی از پیش فرضهای رگرسیون مشاهدات دور افتاده را از روی باقیمانده ها و باقیمانده های استاندارد شده مشخص کرده و مدل رگرسیون را پس از حذف این دور افتاده ها برازش کرد.

برای تشخیص مشاهدات دور افتاده در تحلیل رگرسیون باید مشاهداتی که باقیمانده استاندارد شده آنها از 3+ بزرگتر بوده و یا از 3- کوچکتر است (در برخی موارد ممکن است به جای عدد 3 از عدد 4 استفاده کرد) را مشخص کرد. برای اینکار می توان باقیمانده های استاندارد شده برای کلیه مشاهدات را محاسبه کرده و در یک متغیر جدید در فایل داده قرار داد و سپس بطور چشمی مشخص کرد که کدامیک از مشاهدات دور افتاده هستند. بدین منظور در مسیر انجام رگرسیون خطی و در پنجره Linear Regression بر روی تب Save کلیک کرده و در پنجره ظاهر شده و در بخش Residuals گزینه Standardized را تیک می زنیم. با اجرای این فرمان علاوه بر انجام تحلیل رگرسیون یک متغیر جدید در فایل داده ایجاد خواهد شد که این متغیر حاوی باقیمانده های استاندارد شده برای کلیه مشاهدات است.

اما اگر هدف صرفا مشخص کردن مشاهدات دورافتاده باشد، نیازی به محاسبه باقیمانده های استاندارد شده برای کلیه مشاهدات نیست. بلکه کافی است تنها مشاهدات دورافتاده را تعیین کرد که بدین منظور در مسیر انجام رگرسیون خطی و در پنجره Linear Regression بر روی تب Statistics کلیک کرد ه و در پنجره ظاهر شده و در بخش Residual گزینه Casewise diagnostics را تیک می زنیم. در خروجی جدولی با همین عنوان داده خواهد شد که در آن جدول شماره ردیف مشاهداتی که باقیمانده استاندارد شده آنها خارج از دامنه تعیین شده است نشان داده خواهد شد.

نمودار باقیمانده های استاندارد شده در مقابل مقادیر پیش بینی شده استاندارد شده

در این نمودار مقادیر باقیمانده های استاندارد شده در محور عمود و مقادیر پیش بینی شده استاندارد شده در محور افق قرار می گیرد. مشاهدات نیز به صورت نقاطی در صفحه رسم می شوند. در حالت ایده آل کلیه این نقاط روی خط صفر قرار می گیرد. این

حالتی است که ضریب تعیین مساوی یک است. انتظار داریم این نقاط حول خط صفر و به صورت تصادفی توزیع گردد بطوریکه تقریبا تشکیل یک مستطیل بدهد. در چنین حالتی می توان نرمال بودن توزیع باقیمانده ها را پذیرفت. علاوه بر این توسط این نمودار می توان مشاهدات دورافتاده را نیز تشخیص داد. نقاطی که از خط صفر خیلی دور هستند چه به طرف بالا و چه به طرف پائین همان مشاهدات دورافتاده هستند. انحراف این نقاط از شکل مستطیلی چه به طرف بالا و چه به طرف پائین بیانگر انحراف توزیع باقیمانده ها از نرمال است. همچنین قیفی شکل شدن آن بیانگر نقض فرض همگنی واریانس است.

برای رسم این نمودار در مسیر انجام رگرسیون خطی و در پنجره Linear Regression بر روی تب Plots کلیک کرده و در پنجره ظاهر شده متغیر ZRESID* را به قسمت Y و متغیر ZPRED* را به قسمت X برده و فرمان را اجرا می کنیم.

روش های ورود های متغیرهای مستقل (پیشگو)

به طور کلی در یک رگرسیون چندگانه برای ورود متغیرهای پیشگو در مدل روش های مختلفی وجود دارد که عمده ترین و متداول ترین آنها روش $Enter$ است. در این روش همه متغیرهای پیشگو به طور یکجا به مدل وارد می شوند و مدل مورد بررسی و آزمون قرار می گیرد. اگر ضرائب مربوط به بعضی از متغیرهای پیشگو در مدل معنی دار نبود، آن متغیرها از مدل خارج شده و سپس مجددا مدل دیگری به متغیرها برازش می شود و این عمل همینطور ادامه میابد تا مدل نهائی استخراج گردد. توصیه می شود تا جای ممکن از روش $Enter$ استفاده کنیم. یعنی تا هنگامیکه الگوی خاصی در ورود متغیرها مورد نظر نباشد(که در اکثر موارد چنین است) از روش$Enter$ استفاده کنید. برخی روش های دیگر در ورود متغیرهای پیشگو به صورت زیر است.

روش پیشرو (Forward)

در این روش ابتدا متغیر پیشگوئی که دارای بیشترین همبستگی با متغیر پاسخ است تعیین شده و در صورت معنی دار بودن این همبستگی به مدل وارد می شود. این متغیر در مدل باقی می ماند. سپس در بین متغیرهای پیشگوی باقیمانده متغیری که دارای بیشترین همبستگی جزئی با متغیر پاسخ است در صورت معنی دار بودن این همبستگی به مدل وارد شده و در مدل باقی می

ماند. این عمل آنقدر تکرار می شود تا دیگر متغیری که دارای ضریب همبستگی جزئی معنی دار با متغیر پاسخ است باقی نماند. در محاسبه ضریب همبستگی جزئی متغیر یا متغیرهائی که قبلا به مدل وارد شده اند به عنوان متغیرهای کنترل یا متغیرهای همپراش در نظر گرفته می شوند. در این روش برای معنی دار بودن ضرائب همبستگی معمولا از سطح خطای 0.05 استفاده می شود.

روش پسرو (Backward)

در این روش ابتدا کلیه متغیرهای پیشگو به مدل وارد می شوند. سپس متغیرهائی که وجودشان در مدل معنی دار نیست مورد ارزیابی قرار می گیرند. در بین چنین متغیرهائی متغیری که نبودش در مدل باعث کمترین کاهش در ضریب تعیین می شود از مدل خارج می گردد. سپس یک مدل جدید بدون متغیر حذف شده برازش شده و این فرایند حذفی تکرار می گردد. این فرآیند حذف و بازسازی معادله تا زمانی ادامه می یابد که فقط پیشگوهای معنی دار در مدل باقی بمانند. در این روش برای آزمون معنی دار بودن وجود متغیرها در مدل معمولا از سطح خطای 0.1 استفاده می شود.

روش گام به گام (Stepwise)

این روش ترکیبی از روشهای پیشرو و پسرو است. در این روش نیز مانند روش پیشرو ابتدا متغیری که دارای بیشترین همبستگی معنی دار با متغیر پاسخ است به مدل وارد می شود. سپس در بین متغیرهای پیشگوی باقیمانده متغیری که دارای بیشترین همبستگی جزئی معنی دار با متغیر پاسخ است تعیین شده و به مدل وارد می شود. در این مرحله مانند روش پسرو مدل مورد ارزیابی قرار گرفته و در صورت معنی دار نبودن متغیری در مدل این متغیر از مدل خارج می شود. این فرآیند اضافه و حذف همچنان تکرار می شود تا مدل نهائی استخراج گردد.

توجه:

نتیجه ای که از روشهای پیشرو و پسرو به دست می آید لزوما یکسان نیست. بخصوص به جهت اینکه سطح معنی داری برای باقی ماندن یا حذف متغیرها در مدل یکسان نیست. در واقع معیار ورود متغیرها به مدل در روش پیشرو سختتر از معیار خروج متغیرها از مدل در روش پسرو است. به عبارت دیگر ورود به معادله سختتر از خروج از آن است.

توجه:

از روشهای مرحله ای در ورود متغیرها به مدل مانند روشهای پیشرو، پسرو و گام به گام تنها هنگامی استفاده شود که محقق مایل باشد با کمترین تعداد از متغیرهای پیشگو بزرگترین ضریب تعیین را به دست آورد. اما در صورتی که متغیرهای پیشگو بر اساس مبانی تئوریک و یا یافته های تجربی انتخاب شده باشند و محقق مایل باشد که توان پیش بینی کننده این مجموعه از متغیرهای مستقل را ارزیابی کند لازم است که از روش معمولی و استاندارد یا همان روش Enter استفاده شود. این روش به محقق امکان می دهد تا فرضیات مربوط به مدل را به صورت کلی آزمون کند.

متغیرهای پیشگوی کیفی (رگرسیون با متغیرهای نشانگر)

همانطور که قبلاً بیان شد در یک مدل رگرسیونی متغیرهای پیشگو یا همان متغیرهای مستقل می توانند از نوع متغیرهای کیفی رسته ای نیز باشند. یعنی از متغیرهای کیفی نیز می توان در یک مدل رگرسیونی استفاده کرد. مثلاً متغیرهایی مانند جنسیت، گروه خونی ، وضعیت تاهل، رشته تحصیلی و مانند اینها.

البته برای استفاده از متغیرهای کیفی رسته ای در یک مدل رگرسیونی لازم است ابتدا آنها را به متغیرها نشانگر $Dummy\ variable$ تبدیل کرد. یک متغیر نشانگر متغیری است که تنها دو مقدار صفر و یک را اختیار می کند که این مقادیر بیانگر آن است که مشاهده به یکی از دو رسته ممکن تعلق دارد یا خیر. اگر مشاهده به رسته مورد نظر تعلق داشته باشد این متغیر 1 و در غیر این صورت صفر خواهد بود به صورت زیر:

$$z=\begin{cases}1 & \text{اگر مشاهده به رسته مورد نظر تعلق داشته باشد}\\ 0 & \text{در غیر این صورت}\end{cases}$$

متغیر نشانگر یک متغیر اسمی و دو مقداری است یعنی مقادیر آن هیچگونه ترتیبی را منعکس نمی کند، بلکه تنها عضویت رسته یا طبقه را مشخص می کند. اگر متغیر کیفی که می خواهیم از آن در مدل رگرسیون استفاده کنیم یک متغیر دو مقداری باشد(مانند جنسیت) نیاز به تعریف یک متغیر نشانگر داریم اما اگر متغیر کیفی دارای k مقدار باشد در این صورت تعداد $k\text{-}1$ متغیر رسته ای باید تعریف کرد. مثلاً فرض کنید در یک مدل رگرسیونی بخواهیم جنسیت را نیز به عنوان یک متغیر در مدل وارد کنیم. در این صورت باید یک متغیر نشانگر به صورت زیر تعریف کرد:

$$z=\begin{cases}1 & \text{اگر فرد مرد باشد}\\ 0 & \text{در غیر این صورت}\end{cases}$$

حال فرض کنید بخواهیم رشته تحصیلی را به عنوان یک متغیر در مدل وارد کنیم و نیز فرض کنید رشته تحصیلی دارای سه مقدار مدیریت، حسابداری و اقتصاد باشد در این صورت باید تعداد دو متغیر نشانگر به صورت زیر تعریف کنیم:

$$z_1=\begin{cases}1 & \text{اگر رشته مدیریت باشد}\\ 0 & \text{در غیر این صورت}\end{cases} \quad \text{و} \quad z_2=\begin{cases}1 & \text{اگر رشته حسابداری باشد}\\ 0 & \text{در غیر این صورت}\end{cases}$$

توجه کنید که در این حالت نیاز به تعریف متغیر سوم نداریم. چون وقتی $z_1 = 1$ و $z_2 = 0$ است بیانگر مدیریت بودن رشته آن فرد است. وقتی $z_1 = 0$ و $z_2 = 1$ است بیانگرحسابداری بودن رشته فرد و در غیر اینصورت یعنی وقتی هر دو متغیر مساوی صفر است بیانگر اقتصاد بودن رشته آن فرد است.

مثال (مدل رگرسیون با یک متغیر رسته ای دو مقداری)

فرض کنید بخواهیم یک مدل رگرسیونی به متغیر پاسخ برازش کنیم که در این مدل تنها یک متغیر رسته ای دو مقداری داریم. چنین مدلی معادل آزمون Tی مستقل است یعنی در واقع این مدل همان کار آزمون Tی دو نمونه ای مستقل را انجام می دهد.

مثلاً فرض کنید متغیر پاسخ نمرات درس ریاضی و متغیر پیشگو جنسیت دانش آموز باشد. بنابراین باید ابتدا یک متغیر متغیر نشانگر به صورت زیر تعریف کرد.

$$Z = \begin{cases} 1 & \text{اگر دانش آموز دختر باشد} \\ 0 & \text{در غیر این صورت (پسر)} \end{cases}$$

برای تشکیل چنین متغیری از فرمان *Recode* کردن در مسیر زیر استفاده می کنیم.

Transform → Recode into different variables..

حال یک مدل رگرسیونی بین نمرات درس ریاضی به عنوان متغیر پاسخ و این متغیر نشانگر به عنوان متغیر پیشگو برازش می کنیم. نتیجه این تحلیل رگرسیونی به صورت زیر است .

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.389	.151	.151	4.490

The independent variable is Z.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	4701.831	1	4701.831	233.258	.000
Residual	26385.813	1309	20.157		
Total	31087.645	1310			

The independent variable is Z.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Z	3.788	.248	.389	15.273	.000
(Constant)	11.094	.174		63.721	.000

p-value ی آزمون معنی دار بودن این مدل رگرسیون و نیز *p-value* های آزمون معنی دار بودن ضرایب تا سه رقم اعشار صفر است که این امر بیانگر معنی دار بودن مدل رگرسیون و نیز معنی دار بودن وجود ضرایب در مدل است. مدل رگرسیونی به دست آمده به صورت زیر است.

$$y = 11.094 + 3.788\,Z$$

برای دانش آموزان دختر ($z=1$) مدل فوق به صورت زیر تبدیل می شود.

$$y = 11.094 + 3.788 = 14.882$$

یعنی به طور متوسط نمره ریاضی دخترها حدود 14.88 برآورد می شود. برای دانش آموزان پسر($z=0$) داریم.

$$y = 11.094$$

یعنی به طور متوسط نمره ریاضی پسرها حدود 11.09 برآورد می شود.یا به عبارت دیگر میانگین نمره ریاضی دخترها 14.88 و پسر ها 11.09 است. یعنی به طور متوسط حدود 3.79 واحد اختلاف بین این دو میانگین وجود دارد که چون ضریب z در این مدل معنی دار است لذا اختلاف بین این دو میانگین معنی دار است. این همان نتیجه ای است که قبلاً و در آزمون مقایسه دو میانگین توسط *T-test* به دست آورده بودیم.

مدل رگرسیون با یک متغیر کیفی رسته ای بیش از دو مقداری :

در این حالت می خواهیم یک مدل رگرسیون برازش کنیم که در آن تنها یک متغیر پیشگو داریم که این متغیر یک متغیر کیفی رسته ای بیش از دو مقداری است. در واقع چنین مدلی معادل تحلیل واریانس یک طرفه یا همان *ANOVA* است.

مثلاً فرض کنید متغیر پاسخ (Y) نمرات درس ریاضی و متغیر پیشگو پایه تحصیلی دانش آموز باشد. چون این متغیر دارای سه مقدار اول، دوم و سوم است لذا باید تعداد دو متغیر نشانگر به صورت زیر تعریف کرد.

$$z_1 = \begin{cases} 1 & \text{اگر دانش آموز کلاس اولی باشد} \\ 0 & \text{در غیر این صورت} \end{cases} \quad \text{و} \quad z_2 = \begin{cases} 1 & \text{اگر دانش آموز کلاس دومی باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

پس از برازش مدل رگرسیونی نتیجه به صورت است.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.056[a]	.003	.002	4.86745

a. Predictors: (Constant), z2, z1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	98.465	2	49.233	2.078	.126[a]
	Residual	30989.179	1308	23.692		
	Total	31087.645	1310			

a. Predictors: (Constant), z2, z1

b. Dependent Variable: MATH.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.299	.238		55.993	.000
	z1	-.310	.336	-.030	-.920	.358
	z2	-.663	.326	-.065	-2.033	.042

a. Dependent Variable: MATH.

مدل رگرسیونی به دست آمده به صورت زیر است.

$$y = 13.299 - 0.310z_1 - 0.663z_2$$

مطابق این مدل برای کلاس اولی ها ($z_1 = 1\ ,\ z_2 = 0$) داریم:

$$y = 13.299 - 0.310 = 12.889$$

برای کلاس دومی ها($z_1 = 0\ ,\ z_2 = 1$) داریم.

$$y = 13.299 - 0.663 = 12.636$$

و برای کلاس سومی ها ($\{z_1 = z_2 = 0$) داریم.

$$y = 13.299$$

یعنی میانگین نمره ریاضی کلاس اولی ها 12.89، کلاس دومی ها 12.64 و کلاس سومی ها 13.30 است.

اما از آنجایی که *P-value*ی آزمون معنی دار بودن این مدل 0.126 است لذا این مدل معنی دار نیست. یعنی پایه تحصیلی اثر معنی داری بر نمره ریاضی ندارد. یا به عبارت دیگر میانگین نمره ریاضی در بین این گروهها تفاوت معنی داری با یکدیگر ندارد. *p-value*ی ضریب z_1 مساوی 0.358 است که بیانگر معنی دار نبودن این ضریب است. یعنی اختلاف بین میانگین نمره ریاضی کلاس اولی ها با کلاس سومی ها که حدود 0.35 است معنی دار نیست. *P-value*ی ضریب z_2 مساوی 0.04 است که به صورت ضعیفی معنی دار است. یعنی متوسط اختلاف بین میانگین نمره ریاضی کلاس دومی ها با کلاس سومی ها که حدود 0.663 است به صورت ضعیفی معنی دار است.

مدل رگرسیون با دو متغیر کیفی رسته ای :

در این حالت یک مدل رگرسیون با دو متغیر کیفی رسته ای در نظر می گیریم. توجه کنید که در آزمون مقایسه میانگین های بیش از دو جمعیت با یکدیگر یا همان تحلیل واریانس یک طرفه در واقع اثر یک عامل بر روی یک متغیر وابسته مورد بررسی قرار می گیرد. مثلاً در مقایسه میانگین نمرات درس آمار دانشجویان رشته های مدیریت، حسابداری و اقتصاد توسط *ANOVA* در واقع داریم اثر عامل رشته تحصیلی بر نمره آمار را مورد بررسی قرار می دهیم که در *ANOVA* یک طرفه تعداد یک عامل داریم. حال اگر تعداد عوامل از یکی بیشتر شود در واقع یک آنالیز واریانس چند عاملی داریم. حال آنالیز واریانس چند عاملی را

می توان توسط یک مدل رگرسیونی با بیش از یک متغیر رسته ای انجام داد. به طور خلاصله یک مدل رگرسیونی با بیش از یک متغیر رسته ای در واقع همان کار آنالیز واریانس چند عاملی را انجام می دهد. مثلاً فرض کنید بخواهیم به طور هم زمان اثر جنسیت و پایه تحصیلی را بر نمرات درس ریاضی دانش آموزان توسط یک مدل رگرسیونی مورد بررسی قرار دهیم. در این صورت ابتدا باید متغیرهای نشانگر زیر را تعریف کنیم.

$$Z_1 = \begin{cases} 1 & \text{اگر زن باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{اگر کلاس اولی باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{اگر کلاس دومی باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

نتیجه برازش این مدل رگرسیونی به صورت زیر است.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.394[a]	.155	.153	4.48350

a. Predictors: (Constant), z3, z1, z2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4814.664	3	1604.888	79.838	.000[a]
	Residual	26272.981	1307	20.102		
	Total	31087.645	1310			

a. Predictors: (Constant), z3, z1, z2

b. Dependent Variable: MATH.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.447	.250		45.796	.000
	z1	3.794	.248	.390	15.317	.000
	z2	-.314	.310	-.030	-1.013	.311
	z3	-.708	.300	-.070	-2.358	.019

a. Dependent Variable: MATH.

مدل رگرسیونی به دست آمده به صورت زیر است.

$$y = 11.447 + 3.794\, z_1 - 0.0314\, z_2 - 0.708\, z_3$$

مطابق مدل فوق برای دخترهای کلاس اولی ($z_3 = 0 \, , z_2 = 1 \, , z_1 = 1$) این مدل به صورت زیر تبدیل می شود.

$$y = 11.447 + 3.794 - 0.314 = 14.93$$

برای پسر های کلاس اولی ($z_3 = 0 \, , z_2 = 1 \, , z_1 = 0$) این مدل به صورت زیر تبدیل می شود.

$$y = 11.447 - 0.314 = 11.13$$

برای دخترهای کلاس دومی ($z_3 = 1 \, , z_2 = 0 \, , z_1 = 1$) این مدل به صورت زیر تبدیل می شود.. $y = 11.447 +$

$$3.794 - 0.708 = 14.53$$

برای پسرهای کلاس دومی ($z_3 = 1 \, , z_2 = 0 \, , z_1 = 0$) این مدل به صورت زیر تبدیل می

$y = 11.447 - 0.708 = 10.74$ شود.

دخترهای کلاس سومی ($z_3 = 0 \, , z_2 = 0 \, , z_1 = 1$) این مدل به صورت زیر تبدیل می شود.

$$y = 11.447 + 3.794 = 15.24$$

برای پسر های کلاس سومی ($z_3 = 0 \, , z_2 = 0 \, , z_1 = 0$) این مدل به صورت زیر تبدیل می شود.

$$y = 11.447$$

*p-value*ی آزمون های معنی دار بودن این مدل تا سه رقم اعشار صفر است یعنی این مدل معنی دار است. بنابر این جنسیت و پایه تحصیلی به طور توام اثر معنی داری بر نمره ریاضی دارد. *P-value*ی آزمون معنی دار بودن ضریب z_1 تا سه رقم اعشار

صفر است یعنی وجود این ضریب در مدل معنی دار است و بنابراین جنسیت اثر کاملاً معنی داری بر نمره ریاضی دارد. *p-value* ی ضریب z_2 مساوی 0.311 است که این ضریب در مدل معنی دار نیست. یعنی نمره درس ریاضی کلاس اولی ها تفاوت معنی داری با کلاس سومی ها ندارد. *p- value* ی ضریب z_3 مساوی 0.019 است. یعنی نمره درس ریاضی کلاس دومی ها در سطح خطای 5 درصد تفاوت معنی داری با نمره ریاضی کلاس سومی ها دارد.

مدل رگرسیون با متغیرهای پیشگوی کمی و کیفی رسته ای

در این حالت می خواهیم مدلی را در نظر بگیریم که در آن متغیرهای پیشگو از هر دو نوع کمی و کیفی هستند. بدین منظور از داده های مثال صفحه 145 کتاب تحلیل رگرسیون با مثال (ترجمه دکتر نیرومند) استفاده می کنیم.

در این مثال تحقیقی بر روی گروهی از متخصصین رایانه در یک شرکت بزرگ انجام شده است و هدف از اجرای این تحقیق بررسی میزان تاثیر تجربه کاری (x_1) میزان تحصیلات (x_2) و سمت مدیریت (x_3) بر میزان حقوق دریافتی(y) این متخصصین می باشد. حقوق دریافتی یا همان متغیر پاسخ یک متغیر کمی ، تجربه کاری یک متغیر کمی، میزان تحصیلات یک متغیر کیفی رسته ای سه مقداری است که مقادیر آن شامل دیپلم(کد 1)، لیسانس(کد 2)، و بالا تر از لیسانس(کد 3) می باشد و متغیر مدیریت نیز یک متغیر کیفی رسته ای دو مقداری است که مقادیر آن شامل داشتن پست مدیریت(کد 1) و نداشتن آن(کد 0) است.داده ها در جدول زیر آمده است.

x_3	x_2	x_1	y	x_3	x_2	x_1	Y	x_3	x_2	x_1	y
1	2	10	27780	0	2	4	12884	1	1	1	13876
1	2	11	25410	0	2	5	13245	0	3	1	11608
0	1	11	14861	0	3	5	13677	1	3	1	18701
0	2	12	16862	1	1	5	15965	0	2	1	11283
1	3	12	24170	0	1	6	12336	0	3	1	11767

0	1	13	15990	1	3	6	21352	1	2	2	20872
1	2	13	26330	0	2	6	13839	0	2	2	11772
0	2	14	17949	1	2	6	22884	0	1	2	10535
1	3	15	25685	1	1	7	16978	0	3	2	12195
1	2	16	27837	0	2	8	14803	0	2	3	12313
0	2	16	18838	1	1	8	17404	1	1	3	14975
0	1	16	17483	1	3	8	22184	1	2	3	1371
0	2	17	19207	0	1	8	13548	1	3	3	19800
0	1	20	19364	0	1	10	14467	0	1	4	11417
				0	2	10	15942	1	3	4	20263
				1	3	10	23174	0	3	4	13231

برای انجام تحلیل رگرسیون ابتدا متغیرهای نشانگر زیر را تعریف می کنیم.

$$z_1 = \begin{cases} 1 & \text{تحصیلات دیپلم} \\ 0 & \text{در غیر اینصورت} \end{cases} \quad \text{و} \quad z_2 = \begin{cases} 1 & \text{تحصیلات لیسانس} \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{اگر پست مدیریت دارد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

نتیجه برازش این مدل رگرسیونی به صورت زیر است.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850[a]	.723	.696	2963.229

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850[a]	.723	.696	2963.229

a. Predictors: (Constant), z3, Tajrebeh, z1, z2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.397E8	4	2.349E8	26.753	.000[a]
	Residual	3.600E8	41	8780727.826		
	Total	1.300E9	45			

a. Predictors: (Constant), z3, Tajrebeh, z1, z2

b. Dependent Variable: Hoghogh

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11055.474	1105.236		10.003	.000
	Tajrebeh	647.134	88.020	.623	7.352	.000
	z1	-3502.396	1187.535	-.303	-2.949	.005
	z2	-1214.895	1118.047	-.113	-1.087	.284
	z3	5936.522	905.373	.554	6.557	.000

a. Dependent Variable: Hoghogh

نتیجه برازش این مدل به صورت زیر است.

$$y = 11055.47 + 647.13x_1 - 3502.4z_1 - 1214.89z_2 + 5936.52z_3$$

p-value این مدل تا سه رقم اعشار صفر است و لذا این مدل معنی دار است.

ضریب x_1، 647.13 است این به آن معنی است که با فرض ثابت بودن سایر عوامل، هر یک سال افزایش در تجربه کاری باعث حدود 647 واحد افزایش در حقوق خواهد شد که این میزان معنی دار است. یعنی تجربه کاری اثر معنی داری بر حقوق دریافتی دارد.

داشتن مدرک دیپلم نسبت به کارکنانی که دارای مدرک لیسانس هستند باعث کاهش حدود 3502 واحد در حقوق خواهد شد که این تفاوت معنی دار است. یعنی اختلاف بین متوسط حقوق دیپلمه ها با لیسانس به بالاها حدود 3502 واحد بوده که این اختلاف معنی دار است. داشتن مدرک لیسانس نسبت به کارکنانی که مدرک بالاتر از لیسانس دارند باعث کاهش حدود 1215 واحد در حقوق خواهد شد که این تفاوت معنی دار نیست. به عبارت دیگر اختلاف حقوق کارکنان دارای مدرک لیسانس با کارکنانی که مدرک بالاتر از لیسانس دارند حدود 1215 واحد بوده که این اختلاف معنی دار نیست.

داشتن پست مدیریت نسبت به نداشتن آن باعث افزایش 5936 واحد در حقوق خواهد شد که این تفاوت کاملاً معنی دار است. یعنی اختلاف حقوق مدیران با سایرین حدود 5936 واحد بوده که این تفاوت معنی دار است.

از مقایسه ضرائب استاندارد شده با یکدیگر نتیجه می شود که داشتن تجربه کاری بیشترین تاثیر را بر حقوق دریافتی دارد و این تاثیر مثبت است. پس از آن داشتن سمت مدیریت بیشرین تاثیر را داشته که اثر آن نیز مثبت است. پس از آن بیشترین تاثیر بر حقوق را دارا بودن مدرک دیپلم دارد که اثر آن منفی است و در انتها نیز داشتن مدرک لیسانس بر حقوق موثر است که این اثر نیز منفی است اما معنی دار نیست.

مدل رگرسیونی فوق برای دیپلمه های فاقد پست مدیریت ($z_3 = 0 \ , \ z_2 = 0 \ , \ z_1 = 1$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1 - 3502.4 = 7553.07 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 7553.07 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

مدل رگرسیونی فوق برای دیپلمه های دارای پست مدیریت ($z_3 = 1 \ , \ z_2 = 0 \ , \ z_1 = 1$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1 - 3502.4 + 5936.5 = 13489.6 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 13489.6 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

مدل رگرسیونی فوق برای لیسانسیه های فاقد پست مدیریت ($z_3 = 0 \ , \ z_2 = 1 \ , \ z_1 = 0$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1 - 1214.89 = 9840.6 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 9840.6 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

مدل رگرسیونی فوق برای لیسانسیه های دارای پست مدیریت ($z_1 = 0$ ، $z_2 = 1$ ، $z_3 = 1$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1 - 1214.89 + 5936.5 = 15777 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 15777 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

مدل رگرسیونی فوق برای افراد دارای مدرک لیسانس به بالای فاقد پست مدیریت ($z_1 = 0$ ، $z_2 = 0$ ، $z_3 = 0$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 11055.5 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

مدل رگرسیونی فوق برای افراد دارای مدرک لیسانس به بالای دارای پست مدیریت ($z_1 = 0$ ، $z_2 = 0$ ، $z_3 = 1$) به صورت زیر تبدیل می شود.

$$y = 11055.47 + 647.13x_1 + 5936.5 = 16992 + 647.13x_1$$

این به آن معنی است که متوسط حقوق چنین افرادی 16992 واحد بوده و هر یک سال افزایش در تجربه کاری باعث افزایش حدود 647 واحد در حقوق آنها خواهد شد.

رگرسیون لجستیک

در رگرسیون لوجستیک متغیر پاسخ یک متغیر کیفی چند مقداری است. برخلاف رگرسیون قبل که در آن متغیر پاسخ یک متغیر کمی پیوسته و دارای توزیع نرمال بود. در رگرسیون لوجستیک متغیر پاسخ یک متغیر کیفی دو مقداری و یا چند مقداری است که اگر متغیر پاسخ دو مقداری باشد آن را رگرسیون لجستیک دودویی گفته و در غیر این صورت آن را رگرسیون لجستیک چند مقداری می گوئیم. در رگرسیون معمولی هدف از برازش مدل رگرسیون پیش بینی مقدار متغیر وابسته با معلوم بودن مقادیر متغیرهای مستقل است. اما در رگرسیون لجستیک هدف پیش بینی احتمال عضویت یک نمونه در یکی از دو گروه ویا چند گروه مورد نظر است.

مثلاً از رگرسیون لجستیک دودویی می توان در پیش بینی ورشکسته شدن یا ورشکسته نشدن یک شرکت با معلوم بودن نسبت های مالی آن شرکت استفاده کرد.

همچنین در پیش بینی احتمال موفقیت یا عدم موفقیت یک داوطلب در یک آزمون با معلوم بودن ویژگی هایی مانند جنسیت، معدل و . . . از رگرسیون لجستیک می توان استفاده کرد.

همچنین در پیش بینی احتمال ابتلا به سرطان یا عدم ابتلا به آن با معلوم بودن برخی ویژگی ها مانند سن، جنسیت، سابقه فامیلی، استعمال دخانیات و . . . از رگرسیون لجستیک می توان استفاده کرد.

روش انجام تحلیل رگرسیون لجستیک با رگرسیون معمولی متفاوت است. در رگرسیون معمولی برای برآورد پارامترها از روش حداقل کردن مجذور خطاها (روشOLS) استفاده می شود. حال آنکه در رگرسیون لجستیک برای برآورد پارمترها از روش حداکثر درستنمایی (ML) استفاده می شود. دررگرسیون معمولی برای بررسی معنی دار بودن مدل رگرسیون از آزمونF یا همان ANOVA وبرای بررسی معنی دار بودن وجود هر کدام از ضرائب در مدل از آزمونT(t_value) استفاده می شود. اما در رگرسیون لجستیک برای بررسی معنی دار بودن مدل رگرسیون از آزمون کی دو (کای اسکور) و برای بررسی معنی دار بودن وجود هر کدام از ضرائب از آزمون والد (wald) استفاده می شود.

در رگرسیون معمولی فرض بر این بود که متغیر پاسخ دارای توزیع نرمال با واریانس ثابت است اما در این رگرسیون فرض نرمال بودن متغیر پاسخ و نیز فرض ثابت بودن واریانس مطرح نیست. اما فرض عدم وجود همخطی در بین متغیرهای پیشگو باید برقرار باشد.

در رگرسیون لجستیک برای معتبر بودن تفسیر نتایج حجم نمونه باید بزرگتر از حجم نمونه در رگرسیون خطی باشد. مطابق قواعد تجربی توصیه شده است که در این رگرسیون حجم نمونه حداقل باید 30 برابر پارامترهای برآورد شده باشد. مثلا اگر در این رگرسیون تعداد سه متغیر پیشگو وجود داشته باشد چون تعداد چهار پارامتر برآورد می شود (ضرائب هرکدام از متغیرهای پیشگو ونیز ضریب ثابت در معادله) لذا حجم نمونه باید حداقل 120 باشد.

فرض کنید متغیر پاسخ ،y، یک متغیردو مقداری باشد که مقادیر آن را با صفر(بیانگر شکست) و یک (بیانگرموفقیت) نشان می دهیم. فرض کنیم احتمال موفقیت p و احتمال شکست q=1-p باشد یعنی:

P(y=0)=1-p ؛ P(y=1)=p نسبت p به 1-p را بخت موفقیت می نامند.

بخت و لگاریتم بخت (odd)

به طورکلی احتمال رخ دادن یک پیشامد به احتمال عدم رخ دادن آن را بخت آن پیشامد می نامند. مثلاً اگر احتمال رخ دادن یک پیشامد 0.2 باشد، بخت رخ دادن آن پیشامد $\frac{0.2}{1-0.2} = 0.25$ خواهد بود.

بخت یک پیشامد در فاصله صفر تا بی نهایت تغییر می کند و متقارن نیست. مثلاً اگر بخت رخ دادن یک پیشامد $\frac{0.2}{1-0.2} = 0.25$ باشد، آنگاه بخت عدم رخ دادن آن $\frac{1-0.2}{0.2} = 4$ خواهد بود. همین امر تفسیر بخت را مشکل می سازد. به همین دلیل به جای بخت یک پیشامد از لگاریتم بخت آن استفاده می کنند. پایه لگاریتم را معمولاً عدد2 یا 10 یا e قرار می دهند. لگاریت بخت یک پیشامد در فاصله $-\infty$ تا $+\infty$ تغییر کرده ونسبت به صفر متقارن است. مثلاً اگر احتمال رخ دادن یک پیشامد 0.2 باشد لگاریتم بخت آن 0.602- و لگاریتم بخت عدم رخ دادن آن پیشامد 0.602+ خواهد بود.

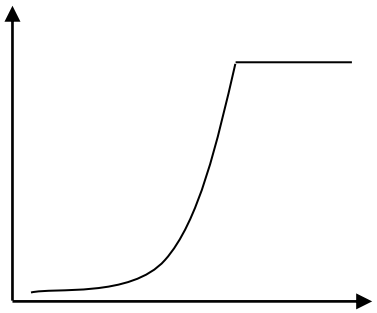
اگر لگاریتم بخت یک پیشامد مثبت باشد به معنی آن است که احتمال وقوع آن پیشامد از احتمال عدم وقوع آن پیشامد بیشتر است. حال هر چقدر این عدد بزرگتر باشد شانس رخ دادن آن پیشامد بیشتر است و به طور مشابه اگر لگاریتم بخت یک پیشامد

منفی باشد به آن معنی است که احتمال وقوع آن پیشامد از احتمال عدم وقوع آن کمتر است و هرچقدر این عدد کوجکتر باشد شانس رخ دادن آن پیشامد کمتر خواهد بود.

فرض کنید متغیر پاسخ که آن را با Y نشان می دهیم مقادیر یک و صفر را با احتمال های به ترتیب p و 1-p اختیار کرده و تنها یک متغیر پیشگو داشته باشیم که آن را با X نشان می دهیم. در این رگرسیون به دنبال مدل سازی احتمال موفقیت هستیم. مدلی که برای احتمال موفقیت در نظر می گیریم به صورت زیر است.

$$\boldsymbol{p = P(y = 1) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}}$$

که در آن β_0 و β_1 پارامترهای ثابتی هستند. این تابع را تابع لجستیک می نامیم. نمودار این تابع به شکل S و به صورت زیر است.



همین طور که ازاین شکل پیداست احتمالp ابتدا با افزایش x به کندی افزایش یافته سپس این افزایش شتاب می گیرد و سرانجام پایدار می شود ولی بیشتر از یک نمی شود.

البته مدل های دیگری برای تعیین رابطه بین p و x در نظر می گیرند. مثلاً یکی دیگر از این مدل ها استفاده از توزیع تجمعی نرمال است که چنین الگویی را الگوی پرابیت (probit) می نامند. استفاده ازالگوی لجستیک ساده تر و بهتراز الگوی پرابیت است.

تابع لجستیک نسبت به پارامترهای β_0 و β_1 غیر خطی است. اما همانطوری که گفتیم به جای استفاده از احتمال وقوع یک پیشامد از لگاریتم بخت آن پیشامد استفاده می شود. بدین ترتیب داریم.

$$\ln(\frac{p}{1-p}) = \boldsymbol{\beta_0 + \beta_1 X} \quad ; \quad \frac{p}{1-p} = e^{e^{\beta_0+\beta_1 X}} \quad ; \quad p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

تبدیل فوق را تبدیل لوجیت نیز می نامند. با انجام چنین تبدیلی و استفاده از لگاریتم بخت رابطه فوق نسبت بـه پارامترهـای β_0 و β_1 خطی می شود. بنابراین یک مدل رگرسیون لوجستیک با یک متغیر پیشگو، مدلی به صورت زیر است.

$$Z = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X}$$

که در آن $Z = \ln(\frac{p}{1-p})$ لگاریتم بخت موفقیت بوده ومقدارآن در فاصله $-\infty$ تا $+\infty$ تغییر می کند.

به طور مشابه در یک رگرسیون چندگانه لجستیک متغیر پاسخ یک متغیر دو مقداری است که مقادیر آن یک(بیانگر موفقیـت) و صفر (بیانگر شکست) با احتمال های به ترتیب p و 1-p هستند و نیز تعداد k متغیر پیشگو داریم که این متغیرها را بـا $x_1, x_2, \ldots, x_k$ نشان می دهیم. مدلی که احتمال موفقیت را به متغیرهای پیشگو مرتبط می سازد توسط یک تابع لجستیک و به شکل زیر نشان می دهیم.

$$\boldsymbol{p = P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}}$$

که درآن $\beta_0, \beta_1, \ldots, \beta_k$ پارامترهای ثابت هستند. با انجام تبدیل لوجیت(استفاده از لگاریتم بخت) تابع فـوق بـه صـورت زیـر تبدیل می شود.

$$Z = Ln(\frac{p}{1-p}) = \boldsymbol{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}$$

که تابع فوق یک تابع خطی از پارامترهای $\beta_0, \beta_1, \ldots, \beta_k$ است.

دررگرسیون لجستیک برای برآورد پارامترها ازروش حداکثردرست نمایی(Maximum Likelihood) برای برآورد پارامترها استفاده می شود.

انجام رگرسیون لجستیک توسط SPSS

درspss برای انجام رگرسیون لجستیک ازمسیرزیراستفاده می کنیم.

Analyze ⟶ **Regression** ⟶ **Binary Logistic**

مثال:(کتاب تحلیل رگرسیون ترجمه دکتر نیرومند)

داده های زیر عملکرد نسبت های مالی33 کارخانه که بعد از 2 سال ورشکست شده اند و33 کارخانـه کـه در طـول همـان دوره فعال باقی مانده اند را نشان می دهد.

X3	X2	X1	Y	X3	X2	X1	Y
1.7	-89.5	-62.8	0	1.3	16.4	43	1
1.1	-3.5	3.3	0	1.9	16	47	1
2.5	-103.2	-120.8	0	2.7	4	-3.3	1
1.1	-28.8	-18.1	0	1.9	20.8	35	1
0.9	-50.6	-3.8	0	0.9	12.6	46.7	1
1.7	-56.2	-61.2	0	2.4	12.5	20.8	1
1	-17.4	-20.3	0	1.5	23.6	33	1
0.5	-25.8	-194.5	0	2.1	10.4	26.1	1
1	-4.3	20.8	0	1.6	13.8	68.6	1
1.5	-22.9	-106.1	0	3.5	33.4	37.3	1
1.2	-35.7	-39.4	0	5.5	23.1	59	1
1.3	-17.7	-164.1	0	1.9	23.8	49.6	1
0.8	-65.8	-308.9	0	1.8	7	12.5	1
2	-22.6	7.2	0	1.5	34.1	37.3	1
1.5	-34.2	-118.3	0	0.9	4.2	35.3	1
6.7	-280	-185.9	0	2.6	25.1	49.5	1
3.4	-19.3	-34.6	0	4	13.5	18.1	1
1.3	6.3	-27.9	0	1.9	15.7	31.4	1
1.6	6.8	-48.2	0	1	-14.4	21.5	1

0.3	-17.2	-19.2	0	1.5	5.8	8.5	1
0.8	-36.7	-19.2	0	1.8	5.8	40.6	1
0.9	-6.5	-18.1	0	1.8	26.4	34.6	1
1.7	-20.8	-98	0	2.3	26.7	19.9	1
1.3	-14.2	-129	0	1.3	12.6	17.4	1
2.1	-15.8	-0.4	0	1.7	14.6	54.7	1
2.8	-36.3	-8.7	0	1.1	20.6	53.5	1
2.1	-12.8	-59	0	2	26.4	35.9	1
0.9	-17.6	-13.1	0	1.9	30.5	39.4	1
1.2	1.6	-38	0	1.9	7.1	53.1	1
0.8	0.7	-57.9	0	1.2	13.8	39.8	1
0.9	-9.1	-8.8	0	2	7	59.4	1
0.1	-4	-64.7	0	1	20.4	16.3	1
0.9	4.8	-11.4	0	1.6	-7.8	21.7	1

در این داده ها $X_3=\frac{\text{فروش}}{\text{کل سرمایه}}$ ؛ $\mathbf{X_2}=\frac{\text{عایدی قبل از سود و مالیات}}{\text{کل دارایی}}$ ؛ $\mathbf{X_1}=\frac{\text{عایدی باقیمانده}}{\text{کل دارایی}}$ می باشد. می خواهیم توسط رگرسیون لجستیک مدلی بدست آوریم که توسط آن ورشکستی شرکت ها را به 3 نسبت مالی فوق نسبت داده و بر این اساس با معلوم بودن نسبت های فوق بتوان احتمال ورشکسته شدن یک شرکت را پیش بینی کرد. بدین منظور متغیر پاسخ را به صورت زیر تعریف می کنیم.

$$y=\begin{cases}1 & \text{فعال بودن شرکت پس از دو سال}\\ 0 & \text{ورشکسته شدن شرکت پس از دو سال}\end{cases}$$

بدین منظور پس از رفتن به مسیر رگرسیون لجستیک متغیر پاسخ را به قسمت dependent و متغیرهای پیشگو را به قسمت covariates می بریم. پس از اجرای این فرمان نتیجه در قالب چند جدول ودر خروجی خواهد آمد که عمده ترین این جداول در زیر آمده است.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	85.582	3	.000
	Block	85.582	3	.000
	Model	85.582	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5.914[a]	.727	.969

a. Estimation terminated at iteration number 13 because parameter estimates changed by less than .001.

Classification Table[a]

		Observed	Predicted		
			Y		Percentage Correct
			.0	1.0	
Step 1	Y	.0	32	1	97.0
		1.0	1	32	97.0
	Overall Percentage				97.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	.346	.323	1.151	1	.283	1.414
	X2	.186	.109	2.943	1	.086	1.205
	X3	5.171	5.386	.922	1	.337	176.162
	Constant	-10.497	11.595	.820	1	.365	.000

a. Variable(s) entered on step 1: X1, X2, X3.

اولین جدول با عنوان Omnibus Tests of Model Coefficients نتیجه آزمون معنی دار بودن مدل لجستیک است. هرسه آماره دراین جدول یکسان است. این آماره ها تنها زمانی متفاوت از یکدیگر خواهند بود که در ورود متغیرهای پیشگو از روش های پیشرو، پسرو وگام به گام استفاده شود. درغیراین صورت هرسه آماره یکسان است. همان طورکه مشاهده می شود در این جدول p-value خیلی کوچک بوده وتا سه رقم اعشار صفر است. بنابراین فرض صفر مبنی بر معنی دار نبودن مدل رگرسیون رد می شود، یعنی مدل معنی دار است.

جدول بعدی با عنوان Model Summary شامل تقریب هایی از ضریب تعیین این مدل است. در رگرسیون لجستیک امکان محاسبه مقدار دقیق ضریب تعیین نیست و ضریب تعیین های ارائه شده دراین جدول(ستون های سوم وچهارم) تقریب هایی از ضریب تعیین هستند که در این مثال همان طورکه مشاهده می شود یکی از تقریب ها برای ضریب تعیین حدود 0.73 و تقریب دیگر حدود0.97 است بنابراین می توان نتیجه گرفت که ضریب تعیین واقعی این مدل بین 0.73 تا 0.97 قرار دارد.

جدول بعدی با عنوان **Classification Table** در واقع دقت مدل لجستیک برازش شده در پیش بینی ورشکسته شدن و یا سالم بودن شرکت ها از روی نسبت های مالی آنها را نشان می دهد. مطابق با این جدول مشاهده می شود که در بین تمام شرکت های ورشکسته (که تعداد آنها 33 شرکت بود) این مدل32 تا از آنها را ورشکسته و1 شرکت را سالم پیش بینی کرد . یعنی دقت پیش بینی این مدل در مورد شرکتهای ورشکسته0.97 است. همچنین در میان تمام شرکتهای سالم که تعداد آنها 33 شرکت بود، این مدل یکی از آن را ورشکسته و32 تای دیگر را فعال پیش بینی کرد. یعنی دقت پیش بینی این مدل برای شرکتهای سالم نیز 0.97 است. و به طورکلی دقت پیش بینی این مدل برای ورشکسته شدن و یا فعال بودن شرکت 0.97 است.

آخرین جدول ضرائب مدل لجستیک ونیز نتیجه آزمون معنی دار بودن آنها را نشان می دهد. در این جدول ستون دوم با عنوانβ همان ضرائب مدل لوجستیک است و ستون ششم با عنوان sig همان p-value ی آزمون معنی دار بودن این ضرائب است. با توجه به اینکه همه این ضرائب از 0.05 بزرگتر است لذا فرض صفر مبنی بر معنی دار نبودن ضریب برای کلیه این ضرائب

پذیرفته شده و بر این اساس می توان نتیجه گرفت که وجود هیچ کدام از این متغیرها در مدل معنی دار نیست. به طور خلاصه، مدل رگرسیون لجستیک برازش شده به صورت زیر است.

$$z = -10.497 + 0.35\,x_1 + 0.19\,x_2 + 5.17\,x_3$$

که درآن کل مدل معنی دار است، اما وجود هرکدام از ضرائب به تنهایی در مدل معنی دار نیست و نیز $Z=\ln(\frac{p}{1-p})$ لگاریتم بخت فعال بودن شرکت پس از دو سال است.

حال با داشتن چنین مدلی و با معلوم بودن سه نسبت مالی برای یک شرکت می توان لگاریتم بخت فعال ماندن آن شرکت و به تبع آن شانس ورشکسته نشدن شرکت را پیش بینی کرد. مثلا فرض کنید اندازه این سه نسبت برای شرکتی برابر باشد با $x_1 = -5.2\ ,\ x_2 = 15\ ,\ x_3 = 1.5$ با قرار دادن این مقادیر در معادله فوق داریم.

$$z = -10.497 + 0.35(-5.2) + 0.19(15) + 5.17(1.5) = -1.712$$

بنابراین لگاریتم بخت فعال بودن چنین شرکتی منفی است و لذا بخت این شرکت برای فعال بودن کمتر از بخت آن برای ورشکسته شدن است. از طرفی چون $z = Ln(\frac{p}{1-p})$ لذا با انجام کمی عملیات ریاضی می توان نشان داد که $p = \frac{e^z}{1+e^z}$. حال با قرار دادن مقدار لگاریتم بخت در این رابطه می توان احتمال ورشکسته نشدن (فعال ماندن) این شرکت پس از دو سال را به دست آورد. در مورد این شرکت داریم.

$$p = \frac{e^{-1.172}}{1 + e^{-1.172}} = 0.15$$

لذا شانس فعال بودن این شرکت پس از دو سال 0.15 خواهد بود.

آمار پارامتری و آمار ناپارامتری

به طور کلی آزمون ها و روش های آماری به دو دسته پارامتری و ناپارامتری تقسیم می شوند. آمار پارامتری؛ مستلزم نرمال بودن توزیع متغیرهاست حال آنکه آمار ناپارامتری هیچگونه فرض خاصی بر روی توزیع متغیرها در نظر نمی گیرد. به عبارت دیگر هرگاه توزیع متغیرهای تحت مطالعه نرمال باشد که در نتیجه این متغیرها کمی پیوسته و در سطح اندازه گیری لااقل فاصله ای هستند مطالعه ی آن ها در حوزه ی آمار پارامتری قرار می گیرد و در غیر این صورت و به خصوص هنگامی که متغیرهای تحت مطالعه کیفی و در سطح اندازه گیری اسمی یا ترتیبی هستند مطالعه ی آن ها در حوزه ی آمار ناپارامتری قرار می گیرد. آمار پارامتری نسبت به آمار ناپارامتری توسعه ی خیلی بیشتری یافته است و توان آزمون های پارامتری نسبت به آزمون های ناپارامتری بیشتر است. به همین دلیل تا جایی که بتوان از روش های پارامتری استفاده کرد نباید سراغ روش های ناپارامتری رفت. بنابراین؛ هرگاه متغیرهای تحت مطالعه کمی پیوسته هستند حتی اگر توزیع آن ها نرمال نباشد با شرط اینکه حجم نمونه به قدر کافی بزرگ باشد باز هم از روش های پارامتری استفاده می کنیم. چون روش های ناپارامتری هیچ توزیع خاصی را بر روی داده ها در نظر نمی گیرد لذا از این روش ها همواره و بر روی هر نوع متغیری می توان استفاده کرد. مثلاً اگر متغیرها کمی و توزیع آن ها نرمال باشد نیز می توان از روش های ناپارامتری استفاده کرد. اما چون توان این روش ها از توان روش های پارامتری کمتر است لذا از روش های ناپارامتری تنها هنگامی استفاده می کنیم که نتوانیم روش های پارامتری را به کار ببریم. به خصوص از روش های ناپارامتری بر روی متغیرهای کیفی و سطوح اندازه گیری اسمی و ترتیبی استفاده می کنیم.

آزمون هایی که تا به حال دیده ایم از قبیل آزمون t یک نمونه ای، آزمون t دو نمونه ای مستقل، آزمون t دو نمونه ای همبسته، آنالیز واریانس، ضریب همبستگی پیرسون و آنالیز رگرسیون همگی از نوع روش های پارامتری هستند که برای برخی از این آزمون ها معادل ناپارامتری نیز وجود دارد. اما ضرایب همبستگی اسپیرمن و تاوکندال و نیز آزمون نیکویی برازش کولموگروف–اسمیرنف از نوع روش های ناپارامتری هستند.

آزمون استقلال (آزمون کی دو)

این آزمون یکی از آزمونهای ناپارامتری است. در این آزمون می خواهیم وجود یا عدم وجود ارتباط بین دو متغیرمقوله ایی (متغیرهای اسمی یا ترتیبی) را مورد بررسی و آزمون قرار می دهیم .

مثلا فرض کنید بخواهیم تعیین کنیم آیا بین رشته تحصیلی(مدیریت – حسابداری –روانشناسی) ومیزان علاقه دانشجو به رشته(خیلی کم – کم –متوسط – زیاد – خیلی زیاد) . در این مثال دو متغیر مقوله ایی داریم که یکی از آنها رشته تحصیلی, یک متغیر اسمی با سه سطح و دیگری میزان علاقه دانشجو به رشته تحصیلی اش که یک متغیر ترتیبی با پنج سطح است. برای تعیین وجود یا عدم وجودرابطه بین این دو متغیر پس از اخذ یک نمونه تصادفی از این جامعه ابتدا نمونه را در یک جدول متقاطع و به صورت زیر تنظیم می کنیم.

رشته / میزان علاقه به رشته	مدیریت	حسابداری	روانشناسی	جمع
خیلی کم	n_{11}	n_{12}	n_{13}	$n_{1.}$
کم	n_{21}	...	...	$n_{2.}$
متوسط	n_{31}	...	...	$n_{3.}$
زیاد	n_{41}	...	...	$n_{4.}$
خیلی زیاد	n_{51}	...	...	$n_{5.}$
جمع	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

که برای تنظیم چنین جدولی همان طور ی که از قبل می دانیم از فرمان crosstab در مسیرزیر

Analyze → descriptive statistics → crosstab

استفاده می کنیم.در این آزمون فرضیه مورد آزمون به صورت زیر است:

H_0 : بین دو متغیر ارتباطی وجود ندارد(دو متغیر مستقل از یکدیگر هستند)

H_1 : $\sim H_0$

مثلا در این مثال فرضیه مورد آزمون به صورت زیر است.

H_0 : نوع رشته ومیزان علاقه به آن ارتباطی با یکدیگر ندارند یا رشته تحصیلی و میزان علاقه به آن , مستقل از یکدیگر هستند

H_1 : H_0

برای آزمون فرضیه فوق از آزمون کای اسکور استفاده می کنیم که در spss در پنجره crosstab بر روی statistics کلیک کرده ودر پنجره ظاهر شده گزینه chi square را تیک می زنیم . نتیجه این آزمون در قالب یک جدول و در خروجی خواهد آمد که مانند سایر آزمونهای آماری اگر p-value کوچک باشد فرض H_0 رد می شود و در غیر این صورت فرض صفر رد نمی شود.

توجه

برای تعیین ارتباط یا عدم ارتباط بین دو متغیر یا همان رابطه سنجی اگرهر دو متغیر کمی بوده وسطح اندازه گیری آنها لااقل فاصله ای باشد و توزیع آنها نرمال باشد (که البته اگر حجم نمونه به قدر کافی بزرگ باشد می توان شرط نرمال بودن توزیع آنها را در نظر نگرفت) در چنین صورتی از ضریب همبستگی پیرسون استفاده می شود که اگر یکی از این متغیرها , متغیر مستقل و دیگری متغیر وابسته باشد از تحلیل رگرسیون نیز برای تعیین رابطه سنجی بین این دو متغیر می توان استفاده کرد. اما اگردست کم یکی از این دو متغیر ترتیبی باشد و یا هردو متغیر کمی وفاصله ای , اما توزیع آنها نرمال نباشد وهردو نمونه نیز کوچک باشد

در این صورت از یکی از ضرایب همبستگی اسپیرمن و تاو کندال استفاده می شود. اما اگر هر دو متغیر اسمی یا یکی اسمی ودیگری ترتیبی باشد از آزمون استقلال (کای اسکور) برای تعیین رابطه بین آنها استفاده می کنیم.

آزمون نیکوئی برازش کولموگروف- اسمیرنف

این آزمون یکی از آزمونهای ناپارامتری است. توسط این آزمون می توان فرض نرمال بودن توزیع متغیرها را آزمون کرد . فرضیه مورد آزمون به صورت زیر است :

H_0: توزیع متغیر مورد نظر نرمال است

H_1: $\sim$ H_0

که برای انجام این آزمون از مسیر زیر استفاده می کنیم.

Analyze ⟶ Nonparametric Tests ⟶ Legacy Dialog ⟶ 1 Sample k-s

توجه:

راجع به آزمون دو نمونه ای T وابسته (مقایسه میانگین های دو جمعیت از طریق دو نمونه همبسته) آزمون دو نمونه ای T مستقل (مقایسه میانگین های دو جمعیت از طریق دو نمونه مستقل) وآنالیز واریانس یک طرفه (ANOVA) قبلا و به طور مفصل بحث شده است حال برای این آزمونها معادل ناپارامتری نیز وجود دارد. یعنی هر گاه متغیری که می خواهیم آن را در بین دو جمعیت یا بیش از دو جمعیت با یکدیگر مقایسه کنیم کمی و نرمال نباشد در این صورت از معادل ناپارامتری آزمونهای فوق می توان استفاده کرد که کلیه این آزمونها در مسیر زیر قرار دارند.

Analyze ⟶ Nonparametric Tests ⟶ Legacy Dialog

مقیاس ها یا سطوح اندازه گیری **(ا.ت.ف.ن)**

اتفن	مقیاس اسمی	مقیاس ترتیبی(رتبه ای)	مقیاس فاصله ای	مقیاس نسبی(نسبتی)
برای شناسایی افراد	+	+	+	+
اندازه گیری نوع صفت	کیفی	کیفی	کمی	کمی
شدت و ضعف صفت کیفی	ندارد	دارد	-	-
رعایت ترتیب	نمی شود	می شود	می شود	می شود
رعایت فاصله	نمی شود	نمی شود	می شود	می شود
رعایت نسبت	نمی شود	نمی شود	نمی شود	می شود
صفر مطلق	ندارد	ندارد	صفر دارد(ولی مطلق نیست)	دارد
4 عمل اصلی	تعریف نمی شود	تعریف نمی شود	فقط جمع و تفریق	دارد(جمع تفریق ضرب و تقسیم)
نوع داده ها	گسسته	گسسته	پیوسته	پیوسته
نوع متغیر	گروهی	گروهی	عددی	عددی
مثال	ملیت، جنسیت، رنگ، گروه خونی،نژاد، داده های شمارشی	رتبه اول و دوم و سوم ورزشکاران، مهارت دانش آموزان(خوب متوسط عالی)، کوتاه قد، بلند قد، و متوسط	نمرات دروس،اندازه دما، اندازه قد در مقایسه با قد یک فرد خاص	ـ مقایسه به صورت ترتیبی، فاصله ای و نسبتی ـ عالی ترین سطح اندازه گیری صفات کمی ـ در اندازه گیری صفر جنبه مطلق و شروع دارد. مانند وزن، طول، حجم، و سطح ـ اندازه گیری با متر ـ اندازه گیری بر حسب کلوین

ویژگی های شاخص های پراکندگی و تمرکز

میانگین	میانه	مد (نما)	واریانس	انحراف معیار یا استاندارد
ـ مرکز ثقل ـ مجموع وزن داده های سمت چپ و راست برابرند ـ عمده ترین و متداولترین شاخص تمرکز ـ ارزش بیشتری نسبت به میانه و مد دارد ـ برای محاسبه: داده ها حداقل فاصله ای(نسبی و فاصله ای) ـ میانگین خطی است اگر یک مقدار ثابت به هر داده های اضافه و کم و یا ضرب و تقسیم شود به میانگین هم همان مقدار ثابت اعمال می شود ـ میانگین داده های استاندارد شده صفر است	ـ نقطه وسط داده ها ـ تعداد داده های طرف و چپ برابرند ـ ارزش کمتری نسبت به میانگین دارد. ولی در انحرافات بزرگ ترجیحاً به جای میانگین مورد استفاده قرار می گیرد ـ برای محاسبه: داده ها حداقل ترتیبی(نسبتی، فاصله و ترتیبی) ـ صدک پنجاهم یعنی میانه	ـ بیشترین فراوانی را دارد ـارزش کمتری نسبت به میانگین و میانه دارد ـ ناچاراً به عنوان شاخص تمرکز استفاده می شود مخصوصاً زمانی که داده ها از نوع صفت کیفی و مقیاس اسمی باشد مثلاً شاخص تمرکز مناسب برای گروه خونی دانشجویان یک کلاس ـ برای محاسبه: هر نوع داده ای(اسمی، ترتیبی، فاصله ای و نسبتی)	ـ شاخص خوب برای پراکندگی ـ مربع یا مجذور واحد اندازه گیری داده ها مثلاً اگر داده ها متر بود واریانس متر مربع می شود ـ جذر واریانس= انحراف معیار ـ واریانس یک عدد ثابت صفر است. مثلاً واریانس 2 عدد صفر است. ـ واریانس همیشه مثبت است. فقط در یک صورت صفر می شود که همه داده ها با هم برابر باشند ـ اگر هر داده ای را در یک مقدار ثابت ضرب و تقسیم کنیم، واریانس در مجذور آن عدد ثابت ضرب و تقسیم می شود ـ واریانس داده های استاندارد شده 1 است	جذر واریانس است ـ با واحد اندازه گیری داده ها یکسان است ـ متداول ترین شاخص پراکندگی انحراف معیار داده های استاندارد شده 1 است ـ اگر هر داده ای را در یک مقدار ثابت ضرب یا تقسیم کنیم انحراف معیار در قدر مطلق آن عدد ثابت ضرب و تقسیم می شود