دانشگاه آزاد اسلامی

تحلیل آماری(قسمت اول)

کارشناسی ارشد مدیریت و حسابداری

دکتر احمد زنده دل

پائیز 1392

کلیات:

کلیات آمار

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 بپوشانند. 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 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.

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 و همه آنها در یک ویژگی مشترک هستند را جامعه آماری می‌گوییم.

واحد آماری (فرد آماری)

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

صفت مشخصه (صفت ثابت)

آن ویژگی که در بین همه 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 گوئیم که چون مقدار این صفت از فردی به فرد دیگر تغییر می کند لذا آن را صفت متغیر یا به اختصار متغیر می نامیم. 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.

مقیاس‌سازی

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 یا سطوح اندازه گیری 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 (سطح اندازه گیری اسمی) (Nominal 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. عبارت ۱+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 (سطح اندازه گیری ترتیبی یا رتبه ای) (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. اما این اعداد فاصله و نسبت را رعایت نمی کنند.

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 داده ها فاصله را رعایت نمی کنند و این به آن معنی است که اختلاف میزان مهارت دو کارگر که کد 1 و 2 گرفته اند دقیقا برابر با اختلاف میزان مهارت دو کارگر که کدهای 2 و 3 گرفته اند نیست. همچنین این اعداد نسبت را رعایت نمی کنند و این به آن معنی است که میزان مهارت کارگری که کد 2 گرفته است دو برابر میزان مهارت کارگری که کد 1 گرفته است نیست. بنابر این چون این داده ها فاصله و نسبت را رعایت نمی کنند لذا بر روی آنها چهار عمل اصلی تعریف نمی شود.

مقیاس فاصله‌ای (سطح اندازه گیری فاصله ای) (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. بنابر این داده های حاصل از اندازه گیری دما بر حسب درجه سانتی گراد از نوع فاصله ای است و بر این اساس بر روی این داده ها عمل جمع و تفریق تعریف می شود اما این داده ها نسبت را رعایت نمی کنند و لذا ضرب و تقسیم روی آنها تعریف نمی شود. بنابراین اگر دمای جسم الف 20 درجه سانتی گراد و دمای جسم ب 10 درجه سانتی گراد باشد دراینصورت چون $20-10=10$ لذا می توان نتیجه گرفت که دمای جسم الف به اندازه 10 درجه بیشتر از دمای جسم ب است. اما اگر چه $20/10=2$ ولی چون این داده ها نسبت را رعایت نمی کنند لذا نمی توان نتیجه گرفت که دمای جسم الف دو برابر دمای جسم ب است.

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 $\frac{۵۰-۳۲}{۵۹-۵۰}=\frac{۱۰-۰}{۱۵-۱۰}$.

مقیاس نسبتی (سطح اندازه گیری نسبتی یا 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 علاوه بر اینکه می توان ادعا کرد که طول جسم A به اندازه ۵۰ سانتیمتر بیشتر از طول جسم B است، 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 1

نوع مقیاس اندازه‌گیری تنها به صفت آماری بستگی نداشته بلکه به چگونگی اندازه‌گیری آن صفت نیز بستگی دارد. و بخصوص صفاتی که ذاتا کمی هستند را می توان طوری اندازه گیری کرد که سطح اندازه گیری نسبتی باشد یا فاصله ای یا ترتیبی و یا حتی اسمی باشد. 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. یا مثلا در اندازه گیری دما بر حسب درجه سانتی گراد همانطریکه بحث شد سطح اندازه گیری از نوع فاصله ای بود. اما اگر دما را بر حسب درجه کلوین اندازه گیری کنیم چون صفر درجه کلوین مطلق بوده و به معنای عدم وجود دما است لذا سطح اندازه گیری نسبتی است. و اگر دما را به صورت سرد، معتدل، گرم، داغ اندازه گیری کنیم، سطح اندازه گیری ترتیبی است.

توجه 2

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 3: 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 (متغیر مقوله ای یا عامل)

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. این متغیرها را متغیر مقوله ای یا عامل (فاکتور) نیز می نامند.

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.

سرشماری

در یک مطالعه آماری هرگاه تمام افراد جامعه آماری مورد بررسی و مطالعه قرار گیرند، آن را سرشماری گویند.

نمونه گیری

در یک مطالعه آماری هرگاه قسمتی از جامعه مورد بررسی و مطالعه قرار گیرد، آن را نمونه گیری گویند که البته در آمار همواره نمونه باید به صورت تصادفی انتخاب شود.

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 مورد نیاز از بین 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، و 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) : که چون معمولا این جداول بر اساس فراوانی ها تنظیم می شود لذا آنها را جدول فراوانی یا جدول توزیع فراوانی نیز می نامند. 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): نمودارهای آماری در واقع بیان هندسی جداول آماری بوده و 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 داده 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 اسمی هستند ناچارا از مد به عنوان شاخص تمرکز استفاده می شود. مثلا شاخص تمرکز مناسب برای اندازه قد، 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 برای اندازه گیری میزان عدم تقارن یا کجی یک توزیع است. کشیدگی شاخصی برای اندازه گیری میزان پراکندگی یک توزیع نسبت به توزیع نرمال است. 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:

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

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:

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

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

همانطوریکه در مقدمه و در بخش کلیات بیان شد موضوع آمار استنباطی که آن را آمار تحلیلی نیز می نامند و قسمت عمده علم آمار امروزی را تشکیل می دهد مطالعۀ 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 مشاهدۀ جامعۀ آماری نبوده و بلکه تنها می توان به یک نمونه از آن جامعه دسترسی داشت. لذا هر گونه استنباط و استنتاج دربارۀ 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 رابطۀ بین متغیرهای انسانی وطبیعی است.

بنابر این 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. 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 به صورت 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. 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:

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 $\left(s^2, \bar{x}\right)$ آماره هستند.

برآوردگر(برآورد یاب- تخمین زن):

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 برآوردگرهایی برای میانگین و واریانس جامعه هستند.

توجه:

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

نحوهٔ برآورد پارامترها:

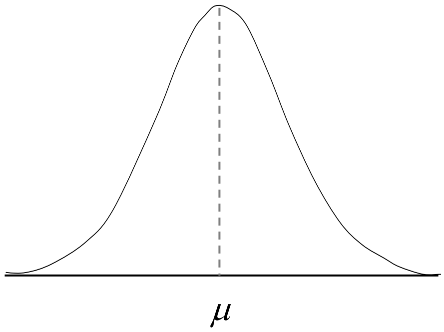
بطور کلی پارامترها را به دو روش نقطه ای و فاصله ای برآورد می کنند.

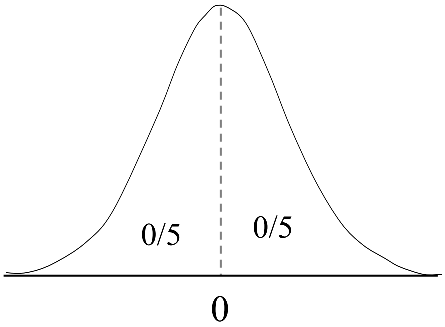
در روش نقطه ای پارامتر جامعه را توسط آمارهٔ همنامش در نمونه برآورد می کنند. مثلا میانگین نمونه یک برآورد نقطه ای برای میانگین جامعه است. اما در روش فاصله ای پارامتر جامعه توسط یک فاصله برآورد می شود. فاصله ای که با ضریب اطمینان خوبی پارامتر جامعه را در بر می گیرد. مثلا در نمونه گیری از یک جامعه اگر میانگین نمونه 12 به دست آید این عدد یک برآورد نقطه ای برای میانگین جامعه است و لذا ادعا می شود که برآورد می شود که میانگین جامعه نیز 12 باشد. اما این برآورد یا تخمین فاقد ضریب اطمینان است. به عبارت دیگر چون مقدار میانگین نمونه وابسته به افراد نمونه است و با تغییر نمونه این میانگین نیز عوض خواهد شد لذا به این عدد نمی توان اطمینان داشت. روش دیگر در برآورد میانگین آن است که میانگین را با یک فاصله برآورد کنیم. فاصله ای که دارای ضریب اطمینان باشد. مثلا به جای اینکه بیان کنیم " برآورد می شود که میانگین جامعه 12 باشد" بیان کنیم که " با 95 درصد اطمینان برآورد می شود که میانگین جامعه در فاصله 10 تا 15 باشد". منظور از 95 درصد اطمینان آن است که اگر چه با تغییر نمونه میانگین آن عوض می شود اما اگر 100 نمونه از این جامعه بگیریم مطمئنیم که میانگین 95 تا از آنها در این فاصله قرار می گیرد. چنین فاصله ای را یک فاصله اطمینان 95 درصد و یا یک برآورد فاصله ای با ضریب اطمینان 95 درصد برای میانگین گویند.

توزیع نرمال :

یکی از توزیع های پیوسته و مهم در آمار توزیع نرمال است . این توزیع دارای دو پارامتر با نام μ و σ^2 است که μ میانگین این توزیع و σ^2 واریانس آن می باشد . نمودار این توزیع یک منحنی زنگوله شکلی است که نسبت به میانگین خود (خط $x=\mu$) مقارن است .

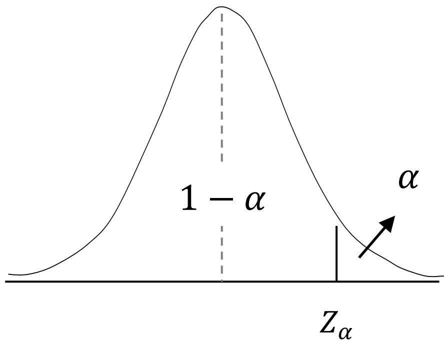
در توزیع نرمال اگر میانگین صفر و واریانس یک باشد ، آن را نرمال استاندارد گفته و متغیر تصادفی مربوط به آن را معمولاً با Z نشان می دهند . از آنجائیکه منحنی نرمال نمودار یک تابع چگالی احتمال است ، لذا کل مساحت زیر منحنی مساوی یک است و چون منحنی نرمال نسبت به میانگین خود متقارن است . لذا این خط مساحت زیر منحنی را به دو قسمت مساوی تقسیم می کند . بنابراین در منحنی نرمال استاندارد مساحت زیر منحنی سمت راست و سمت چپ عدد صفر هر دو مساوی $0/5$ است .



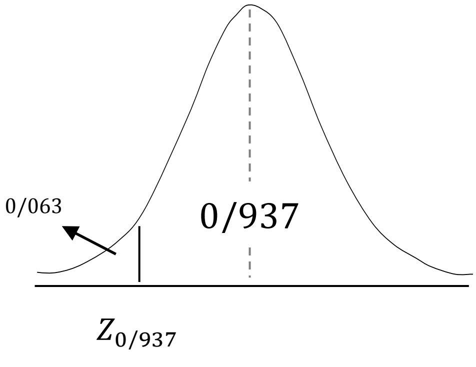


$\mathbf{Z_\alpha}$: اگر α عددی بین صفر و یک باشد $(0 < \alpha < 1)$ آنگاه Z_α عددی است در دامنه توزیع نرمال استاندارد بطوریکه مساحت زیر منحنی سمت راست این عدد مساوی α می شود و در نتیجه به مساحت زیر منحنی سمت چپ این عدد $1 - \alpha$ خواهد شد .

و این عدد از جدول توزیع نرمال بدست می آید .



مثال 1) با استفاده از جدول نرمال مقادیر زیر را بدست آورید .



$$Z_{0.937} = -1.53$$

1- از آنجائیکه جدول نرمالی که در اختیار می باشد برای هر عدد مانند Z مساحت زیر منحنی ، سمت چپ این عدد را داده است لذا در جدول نرمال بجای یافتن 0/937 بایستی به دنبال 0/063=1-0/937 پیدا کنیم ، Z مربوط به 0/063 برابر با 1/53- می باشد لذا $Z_{0.937} = -1.53$.

2- با توجه به اینکه 0.937 بزرگتر از 0.5 می باشد لذا Z مربوط به این عدد یک عدد منفی است. .

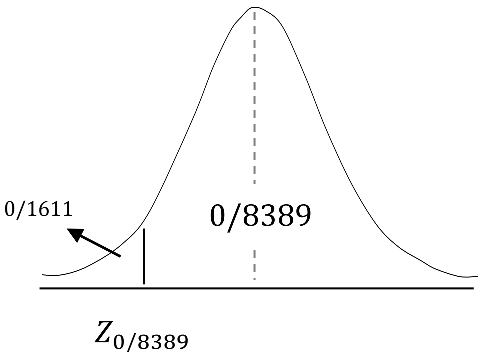
مثال 2) $Z_{0/5}$: با توجه به تعریف Z_α ، واضح است که $Z_{0.5} = 0$ (صفر) است .

توجه : اگر α بیشتر از 0.5 باشد Z_α عددی منفی است و اگر α کوچکتر از 0.5 باشد ، Z_α عددی مثبت خواهد بود .

مثال) $Z_{0/8389}$:

$Z_{0/8389}$ = -0/99

1 -0/8389 = 0/1611



مثال) $Z_{0/025}$ = 1/96

تمرین : مقادیر زیر را بدست آورید :

$$Z_{0.9}\ ,\quad Z_{0.05}\ ,\quad Z_{0.88}\ ,\quad Z_{0.1}$$

توجه: توصیه می شود مقادیر زیر را از بر کنید.

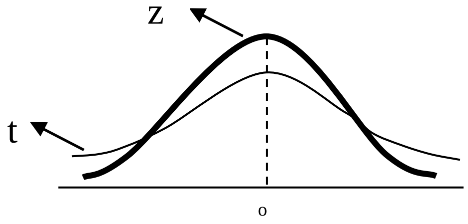
$$Z_{0.005} = 2.58\ ,\quad Z_{0.01} = 2.33\ ,\quad Z_{0.05} = 1.64\ ,\quad Z_{0.025} = 1.96$$

توزیع استیودنت (توزیع t):

یکی دیگر از توزیعهای پیوسته و مهم در آمار توزیع استیودنت یا توزیع t است .این توزیع دارای پارامتری به نام درجه آزادی است که درجهٔ آزادی، یک مقدار صحیح مثبت مانند k است. میانگین این توزیع همواره صفر و واریانس آن $\frac{k}{k-2}$ است. دامنه این توزیع نیز مانند توزیع نرمال R بوده و نمودار آن نیز شبیه به نمودار نرمال یک منحنی زنگوله ای شکل متقارن است با این تفاوت که در

این توزیع پراکندگی از توزیع نرمال بیشتر است و لذا کشیدگی این توزیع از توزیع نرمال کمتر است .در واقع منحنی t در قسمتهای میانی،کوتاه تر از نرمال بوده اما دنباله های آن چاق تر از نرمال است.

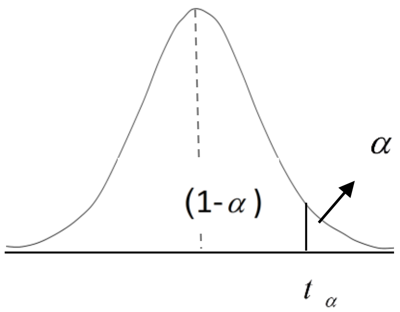
هر چقدر درجه آزادی این توزیع افزایش یابد، پراکندگی آن کمتر شده و منحنی آن به منحنی نرمال نزدیکتر می شود بطوریکه برای درجه آزادی بیشتر از 30 این دو توزیع خیلی به هم نزدیک می شوند.



$t_{\alpha,k}$:

مانند توزیع نرمال در توزیع استیودنت نیز $t_{\alpha,k}$ عددی است در دامنهٔ توزیع استیودنت (t) با درجهٔ آزادی k به طوریکه مساحت زیر منحنی سمت راست این عدد مساوی α می شود یعنی:

$$P(T > t_\alpha) = \alpha$$



و این عدد برای مقادیر مختلف K,α از جدول توزیع t بدست می آید.

مثال: مقادیر زیر را از جدول استیودنت بدست آورید؟

$$t_{0.05,4} = 2.132, \quad t_{0.05,14} = 1.761, \quad t_{0.01,18} = 2.552, \quad t_{0.025,8} = 2.306$$

توجه:

از آنجائیکه با افزایش درجه آزادی توزیع استیودنت به توزیع نرمال میل می کند لذا برای درجات آزادی بزرگ و به خصوص درجات آزادی بیشتر از 30 مقادیر t_α و Z_α خیلی به هم نزدیک بوده و تا دو رقم اعشار با هم برابر هستند. بنابر این $t_{0.05}$ با درجات آزادی بیشتر از 30 برابر با 1.64 است و نیز $t_{0.025}$ با درجات آزادی بیشتر از 30 برابر 1.96 است.

آزمون فرضیه :

در آزمون فرضیه 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 این ادعا را آزمون کنیم. در واقع فرضیه آماری حدسی زیرکانه در خصوص پارامترهای جامعه است، بنابر یک تعریف دیگر" هر حکمی دربارۀ جامعه را یک فرض آماری می نامیم که باید توسط اطلاعات یک نمونه ازآن جامعه آن حکم را رد کرده یا رد نکرد".

آزمون فرض آماری:

قاعده یا قانونی است که توسط آن و با استفاده از داده های نمونه راجع به فرض آماری تصمیم گیری می شود.

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

فرضیه ای که بیانگر عدم وجود اختلاف یا ارتباط بین پارامترها است را فرض صفر گفته و آن را با H_0 نشان می دهیم، نقیض فرض صفر را فرض خلاف یا جانشین گفته و آنرا با H_1 یا H_A نشان می دهیم.

توجه:

در تبدیل فرضیه های پژوهشی به فرضیه های آماری، اگر آن فرضیه حالت تساوی را شامل باشد در فرض صفر قرار گرفته و نقیض آن در فرض خلاف قرار می گیرد و در غیر این صورت خلاف آن عمل میشود.

مثال: فرضیه های پژوهشی زیر را به فرضیه های آماری تبدیل کنید.

1. میانگین معدل لیسانس دانشجویان حسابداری (μ) کمتر از 16 است.

این فرضیه به صورت $\mu<16$ است که چون حالت تساوی را شامل نیست لذا فرض صفر عبارت است از $\mu \geq 16$ و فرض مقابل به صورت $\mu<16$ می باشد. یعنی فرضیه های آماری به صورت زیر است.

$$\mu < 16 \rightarrow \begin{cases} H_0 : \mu \geq 16 \\ H_1 : \mu \langle 16 \end{cases}$$

2. میانگین معدل لیسانس دانشجویان حسابداری (μ) حداکثر 16 است.

این فرضیه به صورت $\mu \leq 16$ است که چون حالت تساوی را شامل است لذا در فرض صفر قرار گرفته و نقیض آن در فرض مقابل قرار می گیرد یعنی فرضیه های آماری به صورت زیر است.

$$\mu \leq 16 \rightarrow \begin{cases} H_0 : \mu \leq 16 \\ H_1 : \mu \rangle 16 \end{cases}$$

3. میانگین معدل دانشجویان دختر (μ_1) بیشتر از میانگین معدل دانشجویان پسر (μ_2)است.

فرضیه های آماری به صورت زیر است.

$$\mu_1 \rangle \mu_2 \rightarrow \begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 \rangle \mu_2 \end{cases}$$

4. نسبت مبتلایان به آنفولانزا در بین زنان (P_1) متفاوت ازنسبت مبتلایان به آنفولانزا در بین مردان (P_2) است.

فرضیه های آماری به صورت زیر است.

$$P_1 \neq P_2 \rightarrow \begin{cases} H_0 : P_1 = P_2 \\ H_1 : P_1 \neq P_2 \end{cases}$$

خطاها:

در هر آزمون فرض همواره فرض صفر آزمون می شود و این فرضیه گزاره ای است که ممکن است درست یا نادرست باشد از طرفی مطابق با داده های نمونه و روشی که در آزمون فرضیه بکار می بریم ممکن است ما فرضیه را رد کرده و یا آن را رد نکنیم، بنابراین 4 حالت زیر اتفاق می افتد.

1. فرض صفر درست است و آنرا می پذیریم. (AH_0 / H_0)
2. فرض صفر درست است و آنرا رد می کنیم. (RH_0 / H_0)
3. فرض صفر نادرست است و آنرا قبول می کنیم. $(AH_0 / H_0') = (AH_0 / H_1)$
4. فرض صفر نادرست است و آنرا رد می کنیم. $(RH_0 / H_0') = (RH_0 / H_1)$

واضح است که در حالات 2 و 3 دچار خطا شده ایم ، حالت 2 را خطای نوع اول و حالت 3 را خطای نوع دوم گوئیم. به عبارت دیگر:

خطای نوع اول: خطای ناشی از رد کردن فرض صفر وقتی این فرض درست است.

خطای نوع دوم: خطای ناشی از پذیرفتن فرض صفر وقتی این فرض غلط (نادرست) است.

سطح خطای آزمون (سطح معنی داری آزمون):

احتمال رخ دادن خطای نوع اول را سطح معنی داری یا سطح خطای آزمون گفته و آنرا با α (آلفا) نشان می دهیم و احتمال رخ دادن خطای نوع دوم را با β (بتا) نشان می دهیم.

توجه: *بین α و β رابطه ای معکوس برقرار است ، اما حاصل جمع آنها لزوما مساوی یک نیست یعنی در حالت کلی خطای نوع اول و خطای نوع دوم دو پیشامد مکمل هم نیستند.*

__توجه__ : در آمار (α) را ثابت در نظر گرفته و برای آن معمولا مقادیری همچون 0/05 یا 0/01 را در نظر می گیرند که با ثابت بودن (α) رابطه ای معکوس بین حجم نمونه و بتا (β) برقرار است.

توان آزمون (π):

احتمال رد کردن فرض صفر وقتی این فرض غلط است را توان آزمون گفته و آنرا با π نشان می دهیم. بین توان آزمون و β رابطه ای معکوس برقرار است و حاصل جمع آنها مساوی یک می باشد.

$$\pi+\beta=1\rightarrow\pi=1-\beta$$

__توجه__ : با ثابت بودن (α) بین حجم نمونه و (β) رابطه معکوس برقرار است یعنی افزایش حجم نمونه باعث کاهش (β) می شود و از طرفی چون $(1-\beta=\pi)$ لذا بین حجم نمونه و توان آزمون رابطه مستقیم برقرار است.

سطح اطمینان:

$1-\alpha$ را سطح اطمینان آزمون می نامیم ، بنابراین اگر مثلا آزمونی در سطح خطای $\alpha=0.05$ باشد ، سطح اطمینان آن 0.95 می باشد.

آماره آزمون :

ملاک یا معیاری است که از روی داده های نمونه محاسبه می شود و توسط آن در مورد رد کردن یا رد نکردن فرض صفر تصمیم گیری می شود ، آماره آزمون یک آماره و در نتیجه یک متغیر تصادفی است که در اکثر موارد توزیع احتمال آن نرمال (z) ، استیودنت (t) ، کی دو (χ^2) ، فیشر (F) است.

ناحیۀ بحرانی یا ناحیه رد :

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

:P-value

p-value که آن را P- مقدار یا مقدار احتمال نیز می نامند، عددی است بین صفر و یک که هرچقدر مقدار آن بزرگتر باشد به معنای درست بودن فرض صفر است و لذا فرضیه پذیرفته می شود و هر چه مقدار آن کوچکتر باشد به معنای نادرست بودن فرض صفر بوده و لذا باید آن فرضیه رد شود. در عمل P-value را با سطح آزمون یعنی α مقایسه کرده اگر P-value کمتر از α شد فرض صفر را رد کرده و در غیر این صورت فرض صفر 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.

توجه: در آزمون فرضیه های آماری باید به این نکته توجه داشت که فرضیه های آماری قابل اثبات به روش ریاضی نیست. یعنی بر خلاف فرضیه ها و قضایای ریاضی که اثبات پذیر بوده و پس از اثبات همواره می توان از آنها به عنوان یک حکم قطعی استفاده کرده و قابل نقض نیستند، فرضیه های آماری را نمی توان اثبات کرد و به همین دلیل این فرضیه ها هیچگاه پذیرفته نمی شوند بلکه رد شده یا رد نمی شوند. اگر یک فرضیه آماری رد نشود نمی توان نتیجه گرفت که لزوما این فرضیه درست است بلکه تنها می توان نتیجه گرفت که داده ها برای رد کردن آن کافی نبود. یا به عبارت دیگر با توجه به این داده ها نمی توان این فرضیه را رد کرد اما این به آن معنی نیست که این فرضیه لزوما درست است. البته در مسائل عملی و کاربردی از واژه پذیرفته شدن استفاده می شود و اگر فرضیه ای رد نشود خود به خود به معنای پذیرفته شدن آن بوده و ادعا می شود که فرضیه پذیرفته شد. ولی باید به این نکته توجه داشت که آن فرضیه تنها رد نشده است و رد نشدن آن لزوما به معنای درست بودن آن نیست.

آزمون تساوی میانگین (μ) با یک مقدار ثابت (μ_0) :

فرض کنید μ و σ^2 میانگین و واریانس یک جامعه نرمال و μ_0 یک مقدار ثابت باشد می خواهیم هر کدام از فرضیه های زیر را آزمون کنیم:

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

یکطرفه جهت دار   یکطرفه جهت دار   دوطرفه بدون جهت

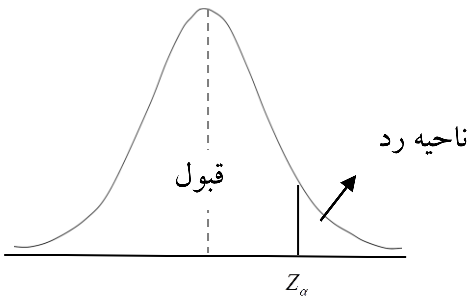
برای آزمون هر کدام از این فرضیه ها ابتدا یک نمونه تصادفی n تایی از این جامعه می گیریم. فرض کنید $\overline{X}, S^2$ میانگین و واریانس نمونه باشد. برای آزمون هر کدام ازاین فرضیه ها موارد زیر را در نظر می گیریم:

الف) واریانس جامعه (σ^2) معلوم: در این حالت آماره آزمون به صورت زیر می باشد.

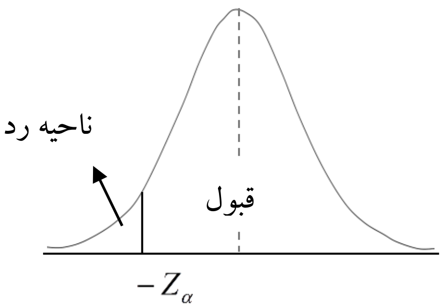
$$Z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$$

که تحت فرض صفر توزیع این آماره نرمال است و لذا نواحی بحرانی یا همان نواحی رد با استفاده از توزیع نرمال و به صورت زیر تعیین می شود.

1. **برای آزمون** $\begin{vmatrix} H_0 : \mu \leq \mu_0 \\ H_1 : \mu \rangle \mu_0 \end{vmatrix}$**:** در صورتی فرض صفر را رد می کنیم که Z بیشتر شود از Z_α، $(Z \rangle Z_\alpha)$، یعنی ناحیه بحرانی به صورت زیر است.

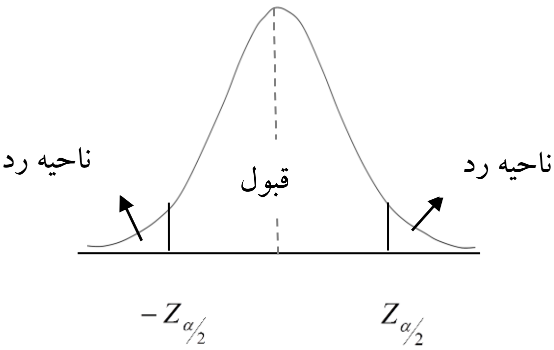


2. **برای آزمون** $\begin{vmatrix} H_0 : \mu \geq \mu_0 \\ H_1 : \mu \langle \mu_0 \end{vmatrix}$**:** در صورتی فرض را رد می کنیم که Z یعنی آماره آزمون کمتر شود از منفی Z_α یعنی $(Z \langle -Z_\alpha)$



3. **برای آزمون فرضیه دوطرفه** $\begin{vmatrix} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{vmatrix}$**:** درصورتی فرض صفر را رد می کنیم که $Z > Z_{\alpha/2}$ یا $Z < -Z_{\alpha/2}$ شود

یا به عبارت دیگر فرض صفر در صورتی رد می شود که $|Z| > Z\alpha/_2$، یعنی ناحیه رد بصورت زیر است.



__توجه:__ همان طوریکه مشاهده می شود ناحیه رد هم جهت با فرض مقابل است. اگر فرض مقابل بصورت بزرگتری باشد ناحیه رد در دُم سمت راست توزیع قرار دارد و اگر فرض مقابل بصورت کوچکتری باشد، ناحیه رد در دُم سمت چپ توزیع قرار دارد و اگر فرض مقابل بصورت نامساوی باشد، سطح آزمون (α) نصف میشود، نیمی در دُم سمت چپ و نیمی در دم سمت راست قرار می گیرد. این قاعده در مورد کلیه آزمونهای آماری برقرار است.

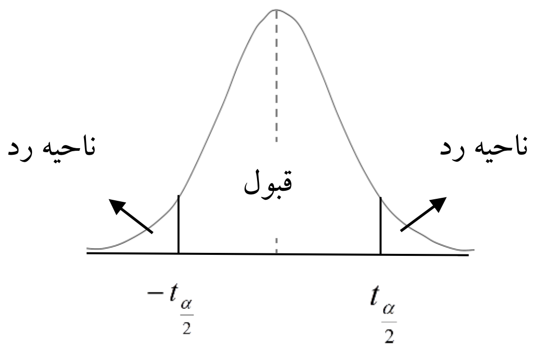
ب. واریانس جامعه (σ^2) مجهول :

در این حالت آماره آزمون بصورت زیر است:

$$T=\frac{\bar{x}-\mu_0}{S_{\bar{x}}}=\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}=\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$$

که توزیع این آماره ، استیودنت با درجه آزادی n-1 است لذا نواحی بحرانی با استفاده از توزیع استیودنت با درجه آزادی n-1 و دقیقاً مانند آزمون قبل بدست می آید.

مثلا برای آزمون دو طرفه $\begin{vmatrix} H_0: \mu=\mu_0 \\ H_1: \mu\neq\mu_0 \end{vmatrix}$ ، در صورتی فرض صفر رد می شود که $T\langle -t_{\frac{\alpha}{2}}$ یا $T\rangle t_{\frac{\alpha}{2}}$



البته باید توجه داشت که اگر حجم نمونه به قدر کافی بزرگ باشد (n> 30) ، چون توزیع استیودنت به سمت توزیع نرمال میل می کند، لذا این آزمون را نیز می توان توسط توزیع نرمال انجام داد.

مثال 11-4 کتاب عادل آذر صفحه111 :

فرضیه ای به این صورت بیان شده است: (میانگین نمره مسئولیت پذیری مدیران در کشور دست کم 50 است) ، شخصی برای بررسی فرضیه فوق یک نمونه 64 تایی از بین مدیران کشور بطور تصادفی انتخاب کرده که میانگین و انحراف معیار آنها به ترتیب 45 و 16 است ، درسطح خطای 5 درصد صحت فرضیه فوق را بررسی کنید.

حل:

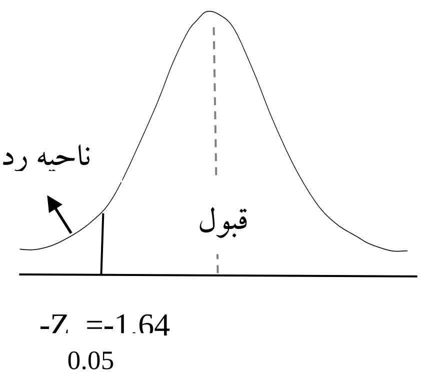
فرض کنید μ و σ^2 به ترتیب بیانگر میانگین و واریانس نمره مسئولیت پذیری مدیران کشور باشد، این ادعا بصورت $\mu \geq 50$ است.لذا فرضیه های آماری به شکل $\Leftarrow$ $\begin{vmatrix} H_0 : \mu \geq 50 \\ H_1 : \mu < 50 \end{vmatrix}$ می باشد و چون واریانس جامعه مجهول است لذا آماره آزمون بصورت زیر است: $T = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{S}$

چون حجم نمونه از 30 بیشتر است در نتیجه به جای توزیع استیودنت می توان از توزیع نرمال استفاده کرد.

> ***توجه:*** *در حالتی که واریانس جامعه مجهول است اگر حجم نمونه از 30 بیشتر باشد به جای توزیع استیودنت می توان از توزیع نرمال استفاده کرد.*

$$n = 64, \quad \bar{x} = 45, \quad S = 16, \quad \alpha = \%5 \quad Z_{0.05} = 1.64$$

$$T = \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} = \frac{\sqrt{64}(45 - 50)}{16} = -2.5$$



چون $T=-2.5\langle -1.64$ ، لذا فرض صفر در سطح خطای 5 درصد رد می شود.به عبارت دیگر چون فرض صفر رد شد می توان نتیجه گرفت که 5 واحد اختلاف مشاهده شده بین آنچه که ادعا شده است $(\mu_0 \geq 50)$ و آن چه که در نمونه $(\bar{x}=45)$ مشاهده گردیده است، در سطح خطای 5 درصد معنی دار است.

مثال : ادعا شده است که میانگین معدل لیسانس دانشجویان ارشد حسابداری کمتر از 16 است در یک نمونه تصادفی 10 تایی از این دانشجویان معدل لیسانس آنها به صورت زیر بدست آمده است.

18.00 – 17.73 – 17.80 – 15.00 – 18.13 – 15.98 – 17.60 – 15.93 – 15.00 – 16.18

حل: فرض کنید μ بیانگر میانگین معدل لیسانس کل دانشجویان ارشد حسابداری باشد. این ادعا به صورت $\mu < 16$ است لذا فرضیه های آماری به صورت زیر است.

$H_0:\ \mu \geq 16$

$H_1:\ \mu < 16$

چون واریانس جامعه مجهول است آماره آزمون به صورت زیر است .

$$T = \frac{\sqrt{N}\ (\bar{X}-\mu_0)}{S}$$

برای محاسبه این آماره لازم است ابتدا میانگین و واریانس نمونه را به دست آوریم. بدین منظور داریم:

$n=10 \quad , \quad \sum x = 167.35 \quad , \quad \bar{x} = \frac{167.35}{10} = 16.735$

$\sum(x-\bar{x})^2 = (16.18-16.74)^2 + (15.00-16.74)^2 + \ldots\ldots + (18.00-16.74)^2 = 13.96$

$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{13.96}{9} = 1.55 \quad , \quad s = \sqrt{1.55} = 1.24$

حال با معلوم بودن میانگین و واریانس نمونه آماره آزمون به صورت زیر به دست می آید.

$$T = \frac{\sqrt{10}\ (16.74-16)}{1.24} = 1.88$$

که این آماره دارای توزیع استیودنت با درجه آزادی n-1=9 است.سطح خطا $\propto = .05$ است لذا مقدار بحرانی برابر است با $t_{\propto,n-1} = t_{.05,9} = 1.83$.

چون $T=1.89 > t_{.05,9} = 1.83$ لذا فرض صفر در سطح خطای 0.05 رد نمی شود (پذیرفته می شود) و در نتیجه فرضیه مقابل رد می شود. اما در این مثال فرضیه پژوهشی در فرض مقابل قرار داشت لذا در سطح خطای 0.05 فرضیه پژوهشی یا همان ادعای مورد نظر رد می شود.

آزمون تساوی نسبت (P) با یک مقدار ثابت (P_0):

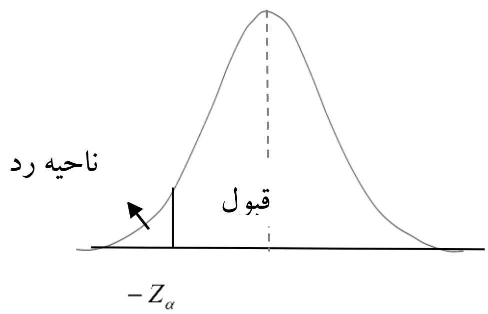
فرض کنید p نسبت یک صفت در یک جامعه (نسبت موفقیت) و p_0 یک مقدار ثابت باشد می خواهیم هر کدام از فرضیه های زیر را آزمون کنیم:

$$1)\ \begin{vmatrix} H_0 : P \le P_0 \\ H_1 : P \rangle P_0 \end{vmatrix} \qquad 2)\begin{vmatrix} H_0 : P \ge P_0 \\ H_1 : P \langle P_0 \end{vmatrix} \qquad 3)\ \begin{vmatrix} H_0 : P = P_0 \\ H_1 : P \ne P_0 \end{vmatrix}$$

برای آزمون هرکدام ازاین فرضیه ها آماره آزمون بصورت زیر است:

$$Z = \frac{\sqrt{n}\left(\overline{P} - P_0\right)}{\sqrt{P_0\left(1 - P_0\right)}}$$

که در آن $\overline{P}$ نسبت همان صفت در یک نمونه تصادفی n تائی است. توزیع این آماره ،نرمال است. لذا نواحی بحرانی با استفاده از توزیع نرمال و دقیقاً مانند آزمون های قبلی بدست می آید. مثلاً برای آزمون $\begin{vmatrix} H_0 : P \ge P_0 \\ H_1 : P \langle P_0 \end{vmatrix}$ در صورتی فرض صفر را رد می کنیم که $Z \langle -Z_\alpha$ باشد.



مثال 8-11 کتاب عادل آذر صفحه 125:

فرضیه ای به این صورت بیان شده است: (60٪ مدیران کشور از شیوه S_1 برخوردارند). تحلیل گری برای بررسی فرضیهٔ فوق یک نمونهٔ 200 تایی از بین مدیران کشور انتخاب کرده است که نیمی از آنها از شیوهٔ S_1 برخوردارند ، در سطح خطای 0/05 صحت این فرضیه را بررسی کنید؟

حل:

فرض کنید p بیانگر نسبتی از مدیرانی کشور باشد که از شیوه S_1 برخوردارند این فرضیه بصورت p=0.6 بیان شده است

، لذا فرضیه های آماری به صورت زیر است:

$$\begin{vmatrix} H_0 : P = 0.6 \\ H_1 : P \neq 0.6 \end{vmatrix}$$

$n = 200$ ، $x = 100$ ، $\overline{P} = \frac{x}{n} = 0.5$ ، $\alpha = 0.05$

$$Z = \frac{\sqrt{n}\left(\overline{P} - P_0\right)}{\sqrt{P_0\left(1 - P_0\right)}} \qquad Z = \frac{\sqrt{200}\left(0.5 - 0.6\right)}{\sqrt{0.6\left(1 - 0.6\right)}} = -2.89$$

$$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow Z_{\alpha/2} = Z_{0.025} = 1.96$$

در صورتی فرض صفر رد می شود که $Z > Z_{0.025}$ یا $Z < -Z_{0.025}$.

چون $Z = -2/89 < -1/96$ بنابراین فرض صفر رد می شود.

مثال : ادعا شده است که دست کم 35٪ از دانشجویان ارشد حسابداری واحد نیشابور ساکن شهر مشهد هستند. در یک نمونه تصادفی 20 تایی از این دانشجویان 6 نفر از آنها ساکن شهر مشهد بودند . این ادعا را در سطح خطای 5٪ آزمون کنید.

حل: فرض کنید p نسبتی از دانشجویان ارشد حسابداری واحد نیشابور باشد که ساکن مشهد هستند . این ادعا به صورت $P \geq 0.35$ است لذا فرضیه های آماری به صورت زیر است.

$H_0 : P \geq 0.35$

$H_1 : P < 0.35$

برای محاسبه آماره آزمون داریم:

$n = 20 \quad , \quad x = 6 \quad , \quad \bar{P} = \frac{x}{n} = \frac{6}{20} = 0.3$

آماره آزمون $Z = \frac{\sqrt{20}\ (0.3 - 0.35)}{\sqrt{0.35(1 - 0.35)}} = -0.47$

$\propto = 0.5 \quad , \quad Z_{\propto} = Z_{0.05} = 1.64$

چون $Z = -0.47 > -1.64$ لذا فرض صفر و در نتیجه ادعای محقق در سطح خطای 5٪ پذیرفته می شود.

آزمون مقایسه دو میانگین:

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 و نیز فرض کنید $S_1^2, \bar{X}_1$ میانگین و واریانس یک نمونه تصادفی n_1 تایی از جامعه اول و $S_2^2, \bar{X}_2$ میانگین و واریانس یک نمونه تصادفی n_2 تایی از جامعه دوم باشد و نیز فرض کنید این دو نمونه مستقل از یکدیگرباشند ، می خواهیم هر کدام از فرضیه زیر را آزمون کنیم.

$$1)\ \begin{cases} H_0 : \mu_1 \le \mu_2 \\ H_1 : \mu_1 \rangle \mu_2 \end{cases} \qquad 2)\ \begin{cases} H_0 : \mu_1 \ge \mu_2 \\ H_1 : \mu_1 \langle \mu_2 \end{cases} \qquad 3)\ \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \ne \mu_2 \end{cases}$$

برای آزمون هر کدام از این فرضیه ها موارد زیر را در نظر می گیریم.

الف)واریانس دو جامعه (σ_2^2, σ_1^2)**معلوم باشد:** در این حالت آزمون آماره آزمون به صورت زیر است.

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

که توزیع این آماره نرمال می باشد. لذا نواحی بحرانی با استفاده از توزیع نرمال و دقیقا مانند آزمونهای قبلی بدست می آید.

ب) واریانس دوجامعه (σ_2^2, σ_1^2)**مجهول باشد:** در این حالت بنابراین که حجم دو نمونه بزرگ یا کوچک باشد موارد زیر را در نظر می گیریم.

ب.I) حجم نمونه ها بزرگ $(n_1 + n_2 - 2 \ge 30)$ **:** در این حالت آماره آزمون به صورت زیر است.

$$T = Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

که توزیع آن نرمال است. لذا نواحی بحرانی با استفاده از توزیع نرمال دقیقا مانند آزمون های قبلی بدست می آید.

ب.II)حجم نمونه ها کوچک $(n_1 + n_2 - 2 \langle 30)$**:** در این حالت بنابر اینکه واریانس های دو جامعه یکسان بوده یا یکسان نباشد موارد زیر را در نظر می گیریم.

ب.I-II) واریانس دو جامعه مساوی $(\boldsymbol{\sigma_1^2 = \sigma_2^2})$**:**در این حالت فرمول آماره آزمون بصورت زیر است.

$$T=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_P^2}{n_1}+\frac{S_P^2}{n_2}}}=\frac{\bar{x}_1-\bar{x}_2}{S_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

که توزیع این آماره،استیودنت با درجه آزادی n_1+n_2-2 است. لذا نواحی بحرانی با استفاده از توزیع استیودنت و دقیقاً مانند آزمون های قبلی بدست می آید و S_p^2 واریانس ادغام شده دو نمونه است که فرمول آن به صورت زیر می باشد.

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \qquad \text{یا} \qquad S_p^2=\frac{1}{n_1+n_2-2}\left[\sum_i(x_{i1}-\bar{x}_1)^2+\sum_i(x_{i2}-\bar{x}_2)^2\right]$$

ب. II-II: واریانس های دو جامعه مساوی نیستند $(\sigma_1^2\neq\sigma_2^2)$: در این حالت فرمول آماره آزمون به صورت زیر است.

$$T=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

که توزیع این آماره استیودنت (t) با درجه آزادی k است و K از فرمول زیر بدست می آید:

$$K=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

بنابر این روش آزمون مقایسه دو میانگین در نمونه های مستقل را می توان بصورت الگوریتم شاخه ای زیر نشان داد.

آزمون مقایسه دو میانگین

σ_1^2,σ_2^2 معلوم: $Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0،1)$

σ_1^2,σ_2^2 مجهول:

$n_1+n_2-2\geq 30$: $Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\sim N(0،1)$

$n_1+n_2-2<30$:

$\sigma_1^2=\sigma_2^2$: $T=\frac{\bar{X}_1-\bar{X}_2}{S_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{(n_1+n_2-2)}$

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_k \qquad \sigma_1^2 \neq \sigma_2^2$$

تمرین 1 کتاب عادل آذر صفحه 122 :

در مطالعه حقوق کارکنان یک شرکت بزرگ ، نمونه های تصادفی مرکب از 150 کارمند متخصص به طور مستقل از دو بخش مهندسی و حسابداری انتخاب شده اند . اطلاعات مربوط به حقوق این کارکنان به شرح زیر است.

بخش حسابداری $\begin{vmatrix} n_1 = 150 \\ \bar{x}_1 = 37250 \\ s_1 = 5541 \end{vmatrix}$ بخش مهندسی $\begin{vmatrix} n_2 = 150 \\ \bar{x}_2 = 39212 \\ s_2 = 5356 \end{vmatrix}$

در سطح خطای دو درصد فرضیه زیر راآزمون کنید : "بین میانگین حقوق کارمندان بخش مهندسی و حسابداری تفاوتی دیده نمی شود."

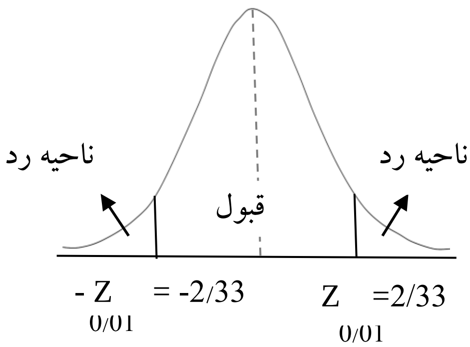
حل :

فرض کنید μ_1و σ_1^2 ، میانگین و واریانس حقوق کارکنان بخش حسابداری(جامعه کارکنان بخش حسابداری) و μ_2و σ_2^2 میانگین و واریانس حقوق کارکنان بخش مهندسی باشند. این ادعا بصورت $\mu_1 = \mu_2$ می باشد ، لذا فرضیه های آماری بصورت زیراست.

$$\begin{vmatrix} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{vmatrix}$$

چون واریانس های دو جامعه مجهول اند و حجم دو نمونه ها به قدر کافی بزرگ است لذا آماره آزمون به صورت زیر است:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{37250 - 39212}{\sqrt{\frac{(5541)^2}{150} + \frac{(5356)^2}{150}}} = -3.12$$



$$\alpha = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01 \Rightarrow Z_{\frac{\alpha}{2}} = Z_{0.01} = 2.33$$

چون $Z = -3/12 < -2/33$ شد لذا فرض صفر در سطح خطای 2 درصد رد می شود.به عبارت دیگر 1962واحد اختلاف مشاهده شده در میانگین های این دو نمونه معنی دار است.

آزمون مقایسه دو نسبت:

فرض کنید P_2, P_1 نسبت های یک صفت در دو جامعه و $\overline{P}_1 = \frac{X_1}{n_1}$ و $\overline{P}_2 = \frac{X_2}{n_2}$ نسبت های همان صفت در دو نمونه تصادفی n_1 تایی و n_2 تایی از این دو جامعه باشند.می خواهیم هر کدام از فرضیه های زیر را آزمون کنیم.

$$1)\begin{cases} H_0 : p_1 \le p_2 \\ H_1 : p_1 > p_2 \end{cases} \qquad 2)\begin{cases} H_0 : p_1 \ge p_2 \\ H_1 : p_1 < p_2 \end{cases} \qquad 3)\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \ne p_2 \end{cases}$$

برای آزمون هر کدام ازاین فرضیه ها، آماره آزمون به صورت زیر است.

$$Z = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\overline{P}\left(1-\overline{P}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

که $\overline{P}$ نسبت صفت مورد نظر در نمونه ادغام شده بوده و از فرمول زیر بدست می آید.

$$\overline{P} = \frac{n_1\overline{P_1} + n_2\overline{P_2}}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$$

و توزیع این آماره نیز نرمال است لذا نواحی بحرانی با استفاده از توزیع نرمال و دقیقا مانند آزمونهای قبلی بدست می آید.

تمرین 1 کتاب عادل آذر صفحه 128 :

ادعا شده است که " سطح آمادگی کارمندان سازمان الف از کارمندان سازمان ب بالاتر است ." برای بررسی این ادعا ، از بین کارمندان سازمان الف یک نمونه 100 تایی انتخاب شده است که 60 نفر از آنها از سطح آمادگی بالا برخوردارند.در حالی که فقط نیمی از یک نمونه تصادفی 200 تایی از کارمندان سازمان ب از سطح آمادگی بالا برخوردارند . در سطح خطای 5 درصد صحت ادعای فوق را آزمون کنید .

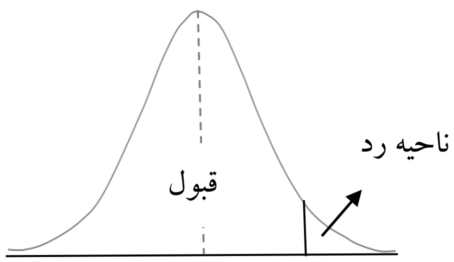
حل:

فرض کنید P_2, P_1 بیانگر نسبت افرادی باشند که به ترتیب در سازمانهای الف و ب از سطح آمادگی بالایی برخوردارند، طبق ادعای مسئله که طرح شده $P_1 > P_2$ بوده و لذا فرضیه های آماری بصورت زیر است:

$$\begin{cases} H_0 : p_1 \leq p_2 \\ H_1 : p_1 \rangle p_2 \end{cases}$$

(الف) $\begin{cases} n_1 = 100 \\ x_1 = 60 \\ \overline{P}_1 = \frac{60}{100} = \frac{6}{10} = 0.6 \end{cases}$

(ب) $\begin{cases} n_2 = 200 \\ x_2 = 100 \\ \overline{P}_2 = \frac{x_2}{n_2} = \frac{100}{200} = \frac{1}{2} = 0.5 \end{cases}$



$Z_{0.05} = 1.64$

$\alpha = 0.05$

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{60+100}{100+200} = \frac{160}{300} = \frac{40}{75} = 0.53$$

$$Z = \frac{0.6 - 0.5}{\sqrt{0.53(1-0.53)\left(\frac{1}{100} + \frac{1}{200}\right)}} = 1.67$$

$$Z_\alpha = Z_{0.05} = 1.64$$

چون $Z=1/67>1/64$ لذا فرض صفر رد می شود، یعنی این ادعای مورد نظر در سطح خطای $0/05$ پذیرفته می شود.

آزمون مقایسه دو میانگین در نمونه های همبسته (جفت شده، طرح پیش آزمون- پس آزمون):

در آزمون های مقایسه دو میانگین فرض بر این بود که دو نمونه مستقل از یکدیگرند اما اگر نمونه ها مستقل نبوده و وابسته باشند، از آزمونهای قبلی نمی توان استفاده کرد. منظور از دو نمونه غیر مستقل که آنها را نمونه های هم بسته یا وابسته می نامند ، نمونه هایی هستند که در آن اعضای هر 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.

مثلا فرض کنیم بخواهیم تاثیر یک روش آموزشی را در بالا بردن معلومات مالی مدیران مالی شرکت های سرمایه گذاری مورد بررسی قرار دهیم. دراین صورت ابتدا یک نمونه تصادفی از مدیران شرکت های سرمایه گذاری انتخاب کرده و

معلومات مالی آنها را اندازه گیری می کنیم. سپس دوره آموزشی را برگزار کرده و مجددا معلومات مالی این مدیران را اندازه گیری می کنیم. در این صورت دو نمونه داریم(یکی اندازه معلومات مالی قبل از برگزاری دوره و دیگری بعد از برگزاری دوره) که این دو نمونه وابسته هستند.

البته در بسیاری از موارد عملاً امکان اینکه بتوان هر دو اندازه گیری را روی یک نمونه انجام داد نیست، مثلاً فرض کنید بخواهیم تاثیر دو روش مختلف آموزشی الف وب را در بالا بردن میزان معلومات مالی مدیران شرکت های سرمایه گذاری با یکدیگر مقایسه کنیم.در این صورت واضح است که عملاً امکان این که بتوان هر دو روش الف و ب را روی یک نمونه به کار برد و نتایج را با یکدیگر مقایسه کرد نیست. چون روش آموزشی اول بر روی نتیجه روش دوم تاثیر می گذارد. در چنین مواردی از نمونه های جفت شده یا زوج شده یا همتراز شده استفاده می کنیم. به این ترتیب که دو نفر مدیر که از لحاظ ویژگی های اثر گذار بر معلومات مالی شبیه به یکدیگر هستند (مثلا هر دو دارای تحصیلات یکسان و جنسیت یکسان و...) را انتخاب کنیم این دو نفر را یک جفت یا یک زوج می نامیم.حال یکی از آنها را در گروه الف ودیگری را در گروه ب قرار می دهیم. این عمل زوج یابی را آن قدر تکرار می کنیم تا نمونۀ نهایی تکمیل گردد. چنین نمونه هایی را نمونه های جفت شده یا زوج شده می نامیم ،که نمونه های جفت شده نیز نمونه هایی وابسته هستند.

فرض کنید $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ یک نمونه تصادفی n تایی همبسته از دو صفت x و y در یک جامعه باشند،مثلا x نمره یک فرد در قبل از برگزاری دوره و y نمره همان فرد در بعد از برگزاری دوره باشد. فرض کنید μ_x,μ_y،میانگین های این دو متغیر باشند، می خواهیم هر کدام از فرضیه های زیر را آزمون کنیم.

$$\begin{vmatrix} H_0:\mu_x=\mu_y \\ H_1:\mu_x\neq\mu_y \end{vmatrix} \qquad \begin{vmatrix} H_0:\mu_x\geq\mu_y \\ H_1:\mu_x<\mu_y \end{vmatrix} \qquad \begin{vmatrix} H_0:\mu_x\leq\mu_y \\ H_1:\mu_x>\mu_y \end{vmatrix}$$

ابتدا اختلاف هر زوج داده را به صورت $d=x-y$ حساب می کنیم. اگر μ_D بیانگر میانگین اختلاف ها باشد در این صورت داریم: $\mu_D=\mu_X-\mu_Y$ و بنابراین فرضیه های فوق بصورت زیر تغییر می کنند.

$$\begin{vmatrix} H_0:\mu_D=0 \\ H_1:\mu_D\neq 0 \end{vmatrix} \qquad \begin{vmatrix} H_0:\mu_D\geq 0 \\ H_1:\mu_D<0 \end{vmatrix} \qquad \begin{vmatrix} H_0:\mu_D\leq 0 \\ H_1:\mu_D>0 \end{vmatrix}$$

آماره این آزمون عبارت است از:

$$T = \frac{\sqrt{n}\,\bar{d}}{s_d}$$

که در آن $\bar{d}$ و S_d میانگین و انحراف معیار اختلاف ها بوده که به صورت زیر هستند.

$$\bar{d} = \frac{1}{n}\sum d_i \qquad S_d^2 = \frac{1}{n-1}\sum\left(d_i - \bar{d}\right)^2$$

و توزیع این آماره، استیودنت (t) با درجه آزادی $n\text{-}1$ است، لذا نواحی بحرانی با استفاده از توزیع استیودنت با درجه آزادی $n\text{-}1$ ودقیقا مانند آزمون های قبلی بدست می آید.

تمرین 4 کتاب عادل آذر صفحه 123:

ادعا شده است که یک برنامه ایمنی صنعتی در کاهش تضییع ساعات کار ناشی از نقص در ماشین های کارخانه موثر است . داده های زیر مربوط به ضایع شدن ساعت های کار هفتگی به واسطه نقص در 6 دستگاه است که یکی قبل و دیگری بعد از اجرای برنامه ایمنی جمع آوری شده است. این ادعا را در سطح خطای 0/05 آزمون کنید.

حل :

فرض کنید μ_y, μ_x به ترتیب بیانگر میانگین ضایعات در قبل و بعد از برگذاری برنامه ایمنی باشد، فرضیه پژوهشی به صورت $\mu_x > \mu_y$ می باشد. لذا فرضیه های آماری بصورت زیر می باشد:

$$\begin{vmatrix} H_0 : \mu_x \le \mu_y \\ H_1 : \mu_x \rangle \mu_y \end{vmatrix}$$

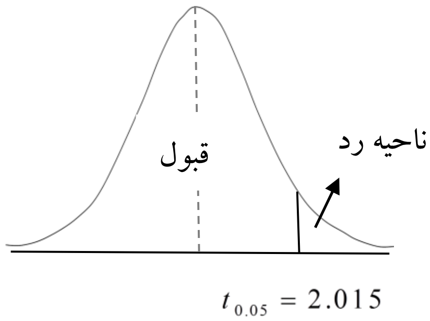
حال فرض کنید $D=x\text{-}y$ ، اختلاف این دو صفت باشد در این صورت فرضیه های فوق بصورت زیر تبدیل می شود.

$$\begin{vmatrix} H_0 : \mu_D \le 0 \\ H_1 : \mu_D \rangle 0 \end{vmatrix}$$

حال برای انجام این آزمون ابتدا اختلاف هر زوج داده را محاسبه کرده ،میانگین و واریانس اختلاف ها را حساب کرده و سپس آماره آزمون را محاسبه می کنیم :

دستگاه	1	2	3	4	5	6	
X (قبل)	12	29	16	37	28	15	
Y (بعد)	10	28	17	35	25	16	

d=x-y	2	1	-1	2	3	-1	$\sum d_i = 6$
$(d_i - \bar{d})^2$	1	0	4	1	4	4	$\sum (d_i - \bar{d})^2 = 14$



$n = 6 \qquad \bar{d} = \frac{1}{n}\sum d_i \qquad \bar{d} = \frac{1}{6} \times 6 = 1$

$S_d^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2 = \frac{1}{5}(14) = \frac{14}{5} = 2.8 \qquad S_d = 1.67$

$df = n - 1 = 5 \qquad \alpha = 0.05, \quad t_{0.05,5} = 2.015$

$T = \frac{\sqrt{n}\bar{d}}{S_d} = \frac{\sqrt{6}(1)}{1.67} = 1.47$

چون $\mu_x > \mu_y$ ناحیه رد دم سمت راست است. چون $T = 1.47 \langle 2.015$ لذا فرض صفر در سطح خطای 0/05 پذیرفته می شود و بنابراین ادعای مورد نظر در سطح خطای 0/05 رد می شود . به عبارت دیگر اختلاف مشاهده شده در ضایعات این ماشین ها در قبل و بعد از اجرای برنامه معنی دار نیست . یا به عبارت دیگر این برنامه اثر معنی داری در کاهش تضییع ساعات کار ماشینها ندارد.

مثال : به منظور تعیین اثر مقطع تحصیلی بر میزان علاقه دانشجویان رشته حسابداری به رشته تحصیلی شان ، گروهی از دانشجویان کارشناسی به تصادف انتخاب شده و میزان علاقه آنها به رشته تحصیلی شان اندازه گیری شده است. سپس در مقطع کارشناسی ارشد مجدداً میزان علاقه به رشته همان دانشجویان اندازه گیری شده است . میزان علاقه به رشته در مقیاس 1 تا 10 اندازه گیری شده است به طوری که عدد 1 بیانگر کمترین علاقه و عدد 10 بیشترین علاقه است . نتیجه در جدول زیر آمده است.

دانشجو	1	2	3	4	5	6	7	8	9	10	11
علاقه به رشته (کارشناسی)	7	5	5	10	6	6	8	5	8	5	2
علاقه به رشته (کارشناسی ارشد)	7	7	8	10	8	7	8	8	9	8	8

این فرضیه را که تحصیل در مقطع بالاتر باعث افزایش میزان علاقه دانشجویان به رشته تحصیلی شان می شود را در سطح خطای 5٪ آزمون کنید. فرض می کنیم میزان علاقه به رشته به صورت نرمال توزیع شده باشد.

حل : فرض کنیم y , x به ترتیب بیانگر میزان علاقه به رشته دانشجویان حسابداری در مقطع کارشناسی و کارشناسی ارشد و μ_x و μ_y میانگین این دو متغیر باشد. فرضیه پژوهشی به صورت $\mu_x < \mu_y$ است، بنابراین فرضیه آماری به صورت زیر است:

$$\begin{cases} H_0: \mu_x \geq \mu_y \\ H_1: \mu_x < \mu_y \end{cases}$$

فرض کنید D = x – y بیانگر اختلاف علاقه به رشته در مقطع کارشناسی و کارشناسی ارشد و μ_D میانگین این اختلاف ها باشد فرضیه فوق را می توان به صورت زیر نوشت.

$$\begin{cases} H_0 : \mu_D \geq 0 \\ H_1: \mu_D < 0 \end{cases}$$

برای آزمون این فرضیه لازم است ابتدا اختلاف هر زوج داده و میانگین و واریانس این اختلاف ها را بدست آوریم. بدین منظور داریم :

x_i	7	5	5	10	6	6	8	5	8	5	2	
y_i	7	7	8	10	8	7	8	8	9	8	8	
$d_i = x_i - y_i$	0	-2	-3	0	-2	-1	0	-3	-1	-3	-6	$\sum d_i$ = -21
$(d_i - \bar{d})^2$												$\sum (d_i - \bar{d})^2 = 32.91$

n = 11 , $\sum d_i$ = -21 , $\bar{d} = \frac{-21}{11}$ = -1.9

$s^2{}_d = \frac{1}{n-1}\sum(d_i - \bar{d})^2 = \frac{32.91}{10} = 3.29$, $s_d = \sqrt{3.29} = 1.81$

$T = \frac{\sqrt{n}\bar{d}}{s_d} = \frac{\sqrt{11}(-1.9)}{1.81} = -3.48$

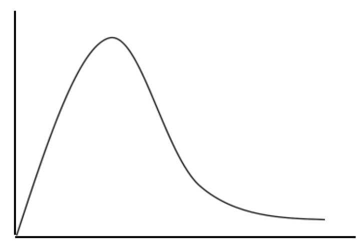
df = n-1 = 10 , $\propto = 0.05$, $t_{\propto,n-1} = t_{0.05,10} = 1.81$

چون $T = -3.48 < -t_{0.05} = -1.81$ لذا فرض صفر در سطح خطای 5٪ رد شده و فرض مقابل و در نتیجه فرضیه پژوهشی در سطح خطای 5٪ پذیرفته میشود.

توزیع کی دو (کای اسکور ، کای مربع، خی دو):

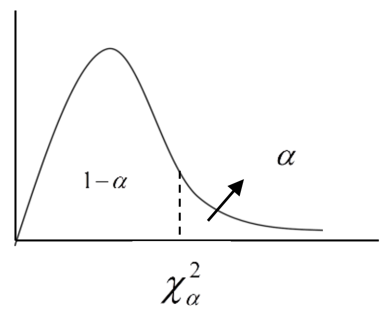
یکی دیگر از توزیع های پیوسته در آمار توزیع کی دو است. این توزیع دارای پارامتری با نام درجه آزادی است، که درجه آزادی یک مقدار صحیح مثبت مانند k است. در یک توزیع کی دو با درجه آزادی k میانگین برابر است با k و واریانس برابر است با $2k$. دامنه این توزیع بر خلاف دامنه توزیع های نرمال و استیودنت که R بود، فقط مقادیر مثبت است و نمودار آن یک منحنی نامتقارن چوله به راست است کی دو با درجه آزادی k را با نماد χ_k^2 (***chi-square***) نشان می دهند.

که هر چه قدر درجه آزادی این توزیع افزایش یابد چولگی آن کاهش می یابدو نمودار آن متقارن تر می شود. متغیر



0

مانند توزیع های نرمال و استیودنت اگر α(آلفا) عددی بین صفر و یک باشد $0\langle\alpha\rangle 1$ ؛ $\chi^2_{\alpha,k}$ عددی است در دامنهٔ توزیع کی دو با درجه آزادی k به طوری که مساحت زیر منحنی سمت راست این عدد مساوی α می شود.



و این عدد برای درجات آزادی مختلف ازر جدول کی دو در انتهای کتاب بدست می آید. مثلا با استفاده از این جدول برخی مقادیر کی دو به صورت زیر است .

$$\chi^2_{0.95\,,14} = 6.571$$

$$\chi^2_{0.05\,,10} = 18.307$$
$$\chi^2_{0.975\,,20} = 9.5908$$

آزمون تساوی واریانس(σ^2) با یک مقدار ثابت(σ_0^2):

فرض کنید σ^2 واریانس یک جامعه نرمال و σ_0^2 یک مقدار ثابت باشد،می خواهیم هرکدام از فرضیه های زیررا آزمون کنیم:

$$\begin{vmatrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{vmatrix} \qquad \begin{vmatrix} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{vmatrix} \qquad \begin{vmatrix} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{vmatrix}$$

برای آزمون هر کدام از این فرضیه ها آماره آزمون عبارت است از:

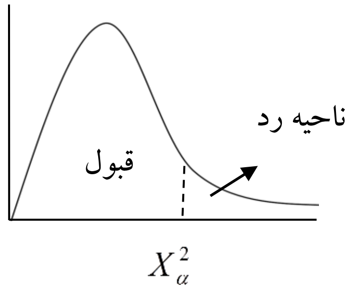
$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

که در آن S^2 واریانس یک نمونه تصادفی n تایی از آن جامعه است و توزیع این آماره کی دو با درجه آزادی $n-1$ است

لذا نواحی بحرانی با استفاده از توزیع کی دو با درجه آزادی n- 1 و به صورت زیر به دست می آید.

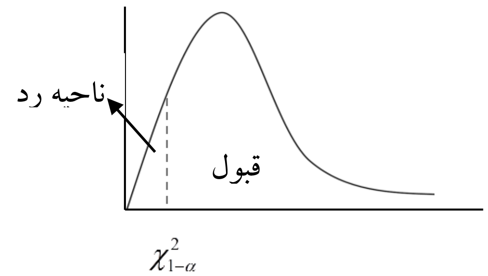
1. **برای آزمون** $\begin{vmatrix} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{vmatrix}$ در صورتی فرض صفر را رد می کنیم که آماره آزمون بیشتر شود از کی دو آلفا یعنی

یعنی $\chi_0^2 > \chi_\alpha^2$



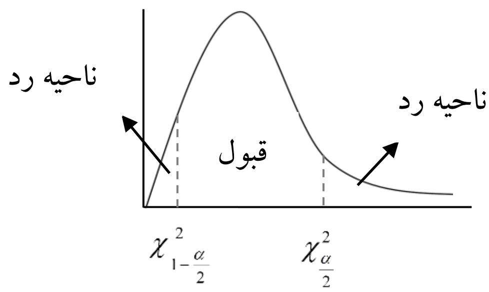
2. **برای آزمون** $\begin{vmatrix} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{vmatrix}$ در صورتی فرض صفر رد می کنیم که آماره آزمون کمتر شود از کی دوی یک

منهای آلفا یعنی $\chi_0^2 < \chi_{1-\alpha}^2$



3. **برای آزمون** $\begin{vmatrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{vmatrix}$ در صورتی فرض صفر رارد می کنیم که آماره آزمون از $\chi_{\alpha/2}^2$ بیشتر و یا از

$X_{1-\alpha/2}^2$ کمتر باشد یعنی $\chi_0^2 < \chi_{1-\frac{\alpha}{2}}^2$ یا $\chi_0^2 > \chi_{\frac{\alpha}{2}}^2$



توجه: *آزمون تساوی واریانس را برای انحراف معیار نیز می توان انجام داد که در چنین صورتی آماره آزمون و نواحی بحرانی هیچ تغییری نمی کند.*

مثال 11-10 کتاب عادل آذر صفحه 131 :

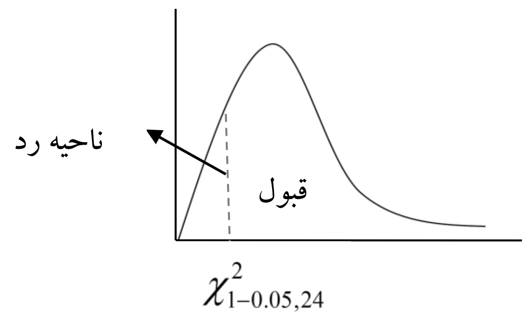
مدیرعامل بازار بورس تهران ادعا کرده که ریسک (انحراف معیار) بازده سهام شرکتهای عرضه کنندهٔ سهام دربازار بورس کمتر از 5 تومان است. بدین منظور یکی از کارگزاران ، 25 شرکت را به طور تصادفی از بین شرکتهای عرضه کنندهٔ سهام در بازار بورس انتخاب کرده که میانگین بازده آنها 14 و انحراف معیارشان 4 تومان است. اگر بازده سهام شرکتها از توزیع نرمال برخوردار باشد، ادعّا را در سطح خطای 0/05 بررسی کنید.

حل:

فرض کنید σبیانگر انحراف معیار بازده سهام شرکت های فوق الذکر باشد این ادعا به صورت $\sigma < 5$ است . لذا فرضیه های آماری به صورت مقابل است:

$$\begin{vmatrix} H_0 : \sigma \geq 5 \\ H_1 : \sigma \langle 5 \end{vmatrix} \Longrightarrow \begin{vmatrix} H_0 : \sigma^2 \geq 25 \\ H_1 : \sigma^2 \langle 25 \end{vmatrix}$$

$n = 25 \qquad \overline{X} = 14 \qquad S = 4 \qquad \alpha = 0.05$



$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{24(16)}{25} = 15.36$$
$$df = n-1 = 25-1 = 24$$
$$\chi^2_{0.95,24} = 13.848$$

چون آماره آزمون $\chi_0^2 = 15.36 > \chi_{0.95}^2 = 13.84$ شده است ، لذا فرض صفر در سطح خطای 0/05 پذیرفته می شود و در نتیجه ادعای مورد نظر در سطح خطای 0/05 رد می شود.

مثال: ادعا شده است که میزان پراکندگی در زمان صرف شده برای مطالعه (برحسب ساعت) در بین دانشجویان ارشد حداکثر 3 ساعت است. داده های زیر زمان مطالعه یک نمونه از دانشجویان ارشد حسابداری برحسب ساعت در هفته گذشته است. با فرض اینکه این متغییر دارای توریع نرمال باشد این ادعا را در سطح خطای 5٪ آزمون نمائید.

5 - 3 - 10 - 2.5 - 5 - 1.5- 1 - 2 - 7.5 - 0 - 3 - 7 - 8 - 6 - 2

حل: فرض کنید σ بیانگر پراکندگی در زمان صرف شده برای مطالعه دانشجویان ارشد حسابداری باشد. این ادعا به صورت $\sigma \leq 3$ می باشد ولذا فرضیه آماری به صورت زیر است .

$$\begin{cases} H_0: \sigma \leq 3 \\ H_1: \sigma > 3 \end{cases} \longrightarrow \begin{cases} H_0: \sigma^2 \leq 9 \\ H_1: \sigma^2 > 9 \end{cases}$$

داریم :

$$n = 15 \quad , \sum x_i = 58.5 \quad , \bar{x} = \frac{58.5}{15} = 3.90$$

$$\sum(x_i - \bar{x})^2 = (5-3.9)^2 + (2-3.9)^2 + \ldots\ldots + (2-3.9)^2 = 92.66$$

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{92.66}{14} = 6.61 \quad , \quad s = \sqrt{6.61} = 2.57$$

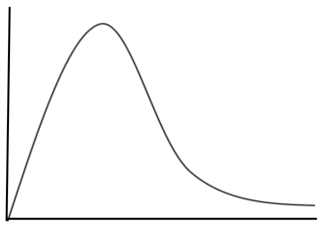
آماره آزمون $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{14 \times 6.61}{9} = 10.28$

$$\propto = 0.05 \quad , \quad {\chi^2}_{0.05,14} = 23.6$$

چون آماره آزمون $\chi^2 = 10.28 < {\chi^2}_{0.05,14} = 23.6$ لذا فرض صفر در سطح خطای 5٪ پذیرفته و براساس این فرضیه پژوهشی در سطح خطای 5٪ پذیرفته می شود.

توزیع فیشر:

فیشر نیز یکی دیگر از توزیع های پیوسته و مهم در آمار است که این توزیع دارای دو پارامتر به نام های درجات آزادی است. که درجات آزادی عدد صحیح مثبت بوده و دارای ترتیب می باشد و معمولا آنها را درجه آزادی اول و دوم یا درجه آزادی صورت و مخرج می نامند، دامنۀ این توزیع ، مقادیر مثبت بوده ونمودار آن شبیه به نمودار منحنی کی دو، یک منحنی نامتقارن چوله به راست است به شکل زیر.



نکات قابل توجه:

1) *اگر* $x_1 \sim \chi^2_{(n_1)}$ *و* $x_2 \sim \chi^2_{(n_2)}$ *و* x_1 *و* x_2 *مستقل از هم باشند آنگاه:*

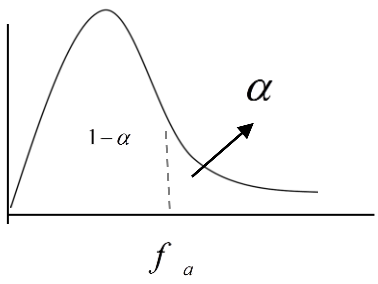
$$\frac{\frac{x_1}{n_1}}{\frac{x_2}{n_2}} = \frac{x_1}{x_2}.\frac{n_2}{n_1} \sim f_{n_1,n_2}$$

یعنی $\frac{x_1}{x_2}.\frac{n_2}{n_1}$ دارای توزیع فیشر با درجات آزادی به ترتیب n_2, n_1 خواهد بود .

2) عکس توزیع فیشر نیز دارای توزیع فیشر خواهد بود که درجات آزادی آن عوض می شود.

$$f_{n_1,n_2} = \frac{1}{f_{n_2,n_1}}$$

3)در توزیع فیشر نیز f_α عددی است که مساحت زیر منحنی سمت راست این عدد مساوی آلفا α می شود و این عدد برای درجات آزادی مختلف از جدول فیشر که در انتهای کتابهای آماری است به دست می آید.



مثلا مطابق این جدول برخی از مقادیر فیشر به شرح زیر است.

$$f_{0.05,3,10} = 3.71 \qquad\qquad f_{0.025,15,20} = 2.57$$

4) از آنجائیکه عکس توزیع فیشر خود دارای توزیع فیشر است لذا $f_{1-\alpha,n_1,n_2} = \frac{1}{f_{\alpha,n_2,n_1}}$ مثلا

$$f_{0.95,10,3} = \frac{1}{f_{0.05,3,10}} = \frac{1}{3.71} = 0.269$$

$$f_{0.975,20,15} = \frac{1}{f_{0.025,15,20}} = \frac{1}{2.57} = 0.389$$

آزمون تساوی دو واریانس:

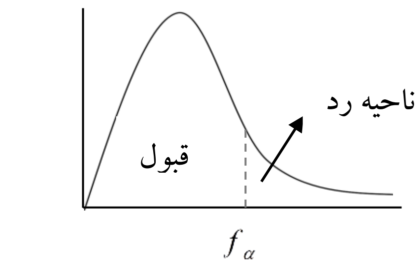
فرض کنید σ_1^2,σ_2^2 واریانس های دو جامعهٔ نرمال باشند و نیز فرض کنید S_1^2,S_2^2 واریانس های دو نمونه n_1 تایی و n_2 تایی از این دو جامعه باشند. می خواهیم هرکدام از فرضیه های زیر را آزمون کنیم.

$$\begin{cases} H_0:\sigma_1^2=\sigma_2^2 \\ H_1:\sigma_1^2\neq\sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0:\dfrac{\sigma_1^2}{\sigma_2^2}=1 \\ H_1:\dfrac{\sigma_1^2}{\sigma_2^2}\neq 1 \end{cases} \qquad \begin{cases} H_0:\sigma_1^2\le\sigma_2^2 \\ H_1:\sigma_1^2>\sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0:\dfrac{\sigma_1^2}{\sigma_2^2}\le 1 \\ H_1:\dfrac{\sigma_1^2}{\sigma_2^2}>1 \end{cases}$$

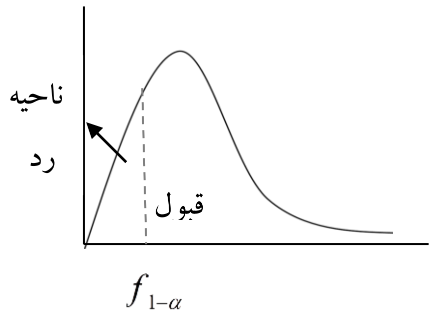
$$\begin{cases} H_0:\sigma_1^2\ge\sigma_2^2 \\ H_1:\sigma_1^2<\sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0:\dfrac{\sigma_1^2}{\sigma_2^2}\ge 1 \\ H_1:\dfrac{\sigma_1^2}{\sigma_2^2}<1 \end{cases}$$

برای آزمون هرکدام از این فرضیه ها ، آماره آزمون $F_o=\frac{S_1^2}{S_2^2}$ است که توزیع این آماره ، فیشر با درجات آزادی به ترتیب n_1-1 و n_2-1 می باشد. لذا نواحی بحرانی با استفاده از توزیع فیشر با درجات آزادی فوق و بصورت زیر تعیین می شود.

1) برای آزمون فرضیه $\begin{cases} H_0:\sigma_1^2\le\sigma_2^2 \\ H_1:\sigma_1^2>\sigma_2^2 \end{cases}$ در صورتی فرض صفر را رد می کنیم که $F_0>F_\alpha$ باشد.



2) برای آزمونی فرضیهٔ $\begin{cases} H_0:\sigma_1^2\ge\sigma_2^2 \\ H_1:\sigma_1^2<\sigma_2^2 \end{cases}$ درصورتی فرض صفر را رد می کنیم که $F_0<f_{1-\alpha}$ باشد.



3) برای آزمون فرضیهٔ $\begin{cases} H_0:\sigma_1^2=\sigma_2^2 \\ H_1:\sigma_1^2\neq\sigma_2^2 \end{cases}$ در صورتی فرض صفر را رد می کنیم که

$$F_0<f_{1-\alpha/2} \quad \text{یا} \quad F_0>f_{\alpha/2}$$

مثال 11-11 کتاب عادل آذر صفحه 134 :

در گزارش مدیریت ، جمله ای به این صورت توسط مامور کنترل کیفیت آمده است :"پراکندگی وزن محصولات تولید شده به وسیله ماشین الف بیشتر از ماشین ب است ." مدیر کارخانه به منظور بررسی گزارش مامور کنترل کیفیت از ماشین الف و ب نمونه هایی انتخاب کرده که حاصل اطلاعات آن به شرح زیر است.توزیع وزن محصولاتی که به وسیله ماشین الف و ب تولید می شوند نرمال است . صحت جمله را در سطح خطای 5 درصد آزمون کنید .

ماشین الف $\left| \begin{array}{l} n_A = 16 \\ \bar{x}_A = 15 \\ s_A = 4.5 \end{array} \right.$ ماشین ب $\left| \begin{array}{l} n_B = 8 \\ \bar{x}_B = 15.5 \\ s_B = 2.25 \end{array} \right.$

حل :

فرض کنید σ_B, σ_A به ترتیب بیانگر پراکندگی در وزن محصولات تولید شده ،توسط ماشین های الف و ب باشد، این ادعا به صورت $\sigma_A \rangle \sigma_B$ است. لذا فرضیه مورد آزمون بصورت زیر می باشد.

$\left| \begin{array}{l} H_0 : \sigma_A \leq \sigma_B \\ H_1 : \sigma_A \rangle \sigma_B \end{array} \right.$ که می توان آنها را بر حسب واریانس به صورت زیر نوشت. $\left| \begin{array}{l} H_0 : \sigma_A^2 \leq \sigma_B^2 \\ H_1 : \sigma_A^2 \rangle \sigma_B^2 \end{array} \right.$

$$\alpha = 0.05, \qquad f_{0.05,15,7} = 3.511 \ , \qquad F_o = \frac{S_1^2}{S_2^2} = \frac{(4.5)^2}{(2.25)^2} = 4$$

چون آماره آزمون $F_0 = 4 \rangle 3.511$ لذا فرض صفر در سطح خطای 0/05 رد می شود و در نتیجه ادعای مورد نظر پذیرفته می شود.

مثال: داده های زیر زمان صرف شده برای مطالعه (بر حسب ساعت) در هفته گذشته در یک نمونه تصادفی از دانشجویان ارشد حسابداری دختر و پسر می باشد.

دختر ها: $5 - 3 - 10 - 2.5 - 5 - 1 - 1.5 - 2 - 2.5 - 0 - 3 - 7 - 8 - 6 - 2$

پسرها: $5 - 4 - 4 - 2 - 3 - 0$

با فرض اینکه زمان صرف شده برای مطالعه دارای توزیع نرمال باشد.

الف) فرضیه یکسان بودن پراکندگی مطالعه در بین دانشجویان دختر و پسر را در سطح خطای 5٪ آزمون کنید.

ب) با فرض اینکه پراکندگی زمان مطالعه در بین دانشجویان دختر و پسر یکسان باشد این فرض که متوسط زمان مطالعه در بین دانشجویان دختر و پسر یکسان است را در سطح خطای 5٪ آزمون کنید.

حل : فرض کنید μ_1 و ${\sigma_1}^2$ میانگین و واریانس زمان صرف شده برای مطالعه در بین دانشجویان دخترو μ_2 و ${\sigma_2}^2$ همان پارامترها در بین دانشجویان پسر باشد. برای انجام هر کدام از این آزمونها لازم است ابتدا میانگین و واریانس دو نمونه را حساب کنیم. بدین منظور داریم:

دختر ها: $n_1 = 15$ و $\bar{x}_1 = 3.90$ و $s_1^2 = 6.61$ و $s_1 = 2.57$

پسرها: $n_2 = 6$ و $\bar{x}_2 = 3$ و $s_2^2 = 3.2$ و $s_2 = 1.79$

الف)

$$\sigma_1 = \sigma_2 \longrightarrow \left| \begin{array}{l} H_0: \sigma_1 = \sigma_2 \\ H_1: \sigma_1 \neq \sigma_2 \end{array} \right. \longrightarrow \left| \begin{array}{l} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{array} \right.$$

آماره آزمون $F_0 = \frac{s_1^2}{s_2^2} = \frac{6.61}{3.2} = 2.06$

$\alpha = 0.05$, $\alpha/2 = 0.025$, df: $n_1 - 1$, $n_2 - 1 = 14, 5$, $f_{0.025\ ,14\ ,5} = 6.43$

$f_{0.975\ ,14\ ,5} = \frac{1}{f_{0.025\ ,5,14}} = \frac{1}{3.66} = 0.27$

چون $0.27 < f_o = 2.06 < 6.43$ لذا فرض صفر و در نتیجه فرضیه پژوهشی در سطح خطای 0.05 پذیرفته می شود.

ب)

$$\mu_1 = \mu_2 \longrightarrow \left| \begin{array}{l} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{array} \right.$$

آماره آزمون $T = \frac{\overline{x_1} - \overline{x_2}}{sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \quad = \frac{1}{n_1+n_2-2}\left[\sum(x-\bar{x})^2+\sum(y-\bar{y})^2\right]$

$= \frac{1}{19}\ [92.66+16] = \frac{108.66}{19} = 5.72 \qquad , \qquad s_p = \sqrt{5.72} = 2.39$

$T = \frac{3.9-3}{2.39\sqrt{\frac{1}{15}+\frac{1}{6}}} = \frac{0.9}{1.15} = 0.78 \qquad , \qquad df = n_1+n_2-2 = 19$

$\alpha = 0.05 \quad , \quad \alpha/2 = 0.025 \quad , \quad t_{0.25\,,19} = 2.09$

چون $-2.09 < T=0.78 < 2.09$ لذا فرض صفر و در نتیجه فرضیه پژوهشی در سطح خطای 5٪ پذیرفته می شود.

روش تعیین حجم نمونه :

به طور کلی هیچ فرمول خاصی برای تعیین حجم نمونه ای که همواره در هر تحقیقی بتوان از آن فرمول استفاده کرد وجود ندارد . در تحقیقات میدانی که محقق باید در صحنه عملی اقدام به جمع آوری اطلاعات کند توصیه می شود که حجم نمونه حداقل 100 باشد، هم چنین در آزمون فرضیه ها به خصوص در فرضیه های تطبیقی که قرار است گروههای مختلفی با یکدیگر مقایسه شوند توصیه می شود که حجم نمونه در هر کدام از این گروهها دست کم 30 باشد .

اما اگر در یک تحقیق هدف صرفاً برآورد میانگین یک متغیر باشد و فرض کنیم واریانس آن متغیر در جامعه σ^2 باشد و بخواهیم میانگین را به نحوی برآورد کنیم که با ضریب اطمینان $100(1-\alpha)$٪ دقت برآورد یا حداکثر خطای حدی e باشد (دقت برآورد یا حداکثر خطای حدی، حداکثر فاصله بین مقدار برآورد شده یک پارامتر با مقدار واقعی آن است) در این صورت حداقل حجم نمونه لازم از فرمول زیر به دست می آید .

$$n \geq \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2}$$

در استفاده از این فرمول لازم است واریانس جامعه معلوم باشد که واریانس نیز پارامتری است که عموماً مجهول است . بنابراین یا باید از مطالعات از قبل انجام شده استفاده کرد و برآوردی برای واریانس قرار داد و یا باید در یک نمونه اولیه واریانس را برآورد کرده و از برآورد آن در این فرمول استفاده کرد. در مواردی نیز می توان از خاصیت توزیع نرمال استفاده کرده و مطابق با آن برآوردی برای واریانس به دست آورد. به این ترتیب که اگر توزیع صفت تحت مطالعه نرمال باشد و

بتوان دو مقدار مانند a و b را به نحوی تعیین کرد که احتمال اینکه مقداری کمتر از a و بیشتر از b وجود داشته باشد خیلی کوچک باشد، یعنی در واقع اکثریت حدود 95 درصد از افراد جامعه مقادیری بین دو عدد a و b را اختیار کرده و وجود مقداری کمتر از a و بیشتر از b نه غیرممکن بلکه غیر عادی باشد، در اینصورت فاصله بین این دو عدد تقسیم بر 4 یک برآورد از انحراف معیار جامعه است یعنی در این حالت می توان انحراف معیار جامعه را به صورت زیر برآورد کرد.

$$\hat{\sigma} = \frac{b-a}{4}$$

حال اگر در برآورد میانگین، حجم جامعه معلوم و مقدار متناهی N باشد در این صورت حداقل حجم نمونه لازم در برآورد میانگین از فرمول زیر به دست می آید .

$$n \geq \frac{NZ_{\frac{\alpha}{2}}^2 \sigma^2}{e^2(N-1)+Z_{\frac{\alpha}{2}}^2 \sigma^2}$$

مثال (تمرین 1 صفحه 78- عادل آذر)

مدیر کنترل کیفیت کارخانه ای در صدد برآورد میانگین وزن محصولات تولید شده کارخانه است. او در این زمینه یک نمونه مقدماتی انتخاب کرده که میانگین آن 50 گرم و انحراف معیارش 12 گرم است. اگر دقت برآورد 14 گرم در نظر باشد، حجم نمونه مناسب را با ضریب اطمینان 99 درصد برآورد کنید.

حل: داریم

$$\bar{x} = 50 \;,\; S = 12 \;,\;\; e = 14 \;,\;\; 1-\alpha = 0.99 \;,\;\; \alpha = 0.01 \;,\; z_{\frac{\alpha}{2}} = z_{.005} = 2.58$$

$$n = \frac{z_{\frac{\alpha}{2}}^2 \; S^2}{e^2} = \frac{(2.58)^2\;(12)^2}{(14)^2} = 4.89$$

بنابر این حداقل حجم نمونه 5 می باشد.

تعیین حجم نمونه در برآورد نسبت:

اگر هدف از نمونه گیری برآورد یک نسبت مانند p باشد و بخواهیم با ضریب اطمینان (1-α)100٪ حداکثر خطای حدی یا دقت برآورد e باشد در این صورت حداقل حجم نمونه لازم از فرمول زیر بدست می آید .

$$n \ge \frac{Z_{\frac{\alpha}{2}}^{2}\, p(1-p)}{e^{2}}$$

اما در استفاده از این فرمول لازم است p معلوم باشد که p همان پارامتر مجهولی است که می خواهیم توسط نمونه گیری آن را برآورد کنیم لذا یا از مطالعات از قبل انجام شده استفاده کرده و برآوردی برای آن قرار می دهیم یا آن را در یک نمونه اولیه برآورد کرده و از برآورد آن استفاده می کنیم ویا از اطلاعات تجربی افراد خبره استفاده می کنیم و یا در صورت فقدان هر اطلاعی راجع به p آن را مساوی 0/5 قرار می دهیم که البته در چنین صورتی مقداری که برای حجم نمونه بدست می آید حداکثر مقدار ممکن است.

حال اگر در برآورد p حجم جامعه معلوم و مقدار متناهی N باشد در این صورت حداقل حجم نمونه لازم از رابطه زیر بدست می آید .

$$n \ge \frac{N Z_{\frac{\alpha}{2}}^{2}\, p(1-p)}{e^{2}(N-1) + Z_{\frac{\alpha}{2}}^{2}\, p(1-p)}$$

مثال (تمرین 4 صفحه 78- عادل آذر):

مدیر شرکت الف در نظر دارد شرکت ب را که سماور برقی تولید می کند خریداری کند. شرکت ب ادعا می کند که 30 درصد کل بازار سماور برقی را در اختیار دارد. مدیر شرکت بزرگ الف می خواهد بوسیله نمونه گیری صحت این ادعا را بررسی کند. اگر مدیر شرکت الف بخواهد سهم شرکت ب را تا 0/025 اختلاف با 98 درصد اطمینان برآورد کند، حجم نمونه انتخابی چقدر باید باشد.

حل: داریم

$$e = 0.025 \ , \quad 1-\alpha = .98 \ , \quad \alpha = .02 \ , \ z_{\frac{\alpha}{2}} = z_{.01} = 2.33$$

برای برآورد p در فرمول حجم نمونه یا باید یک نمونه اولیه گرفت و در آن p را برآورد کرد و یا می توان از ادعای شرکت ب در این خصوص استفاده کرده و آن را به عنوان یک برآورد اولیه برای p در نظر گرفت. بنابر این با در نظر گرفتن این ادعا به عنوان برآوردی اولیه برای p داریم $\bar{p} = 0.3$ و لذا

$$n = \frac{z_{\frac{\alpha}{2}}^2\ \bar{p}\ (1-\bar{p})}{e^2} = \frac{(2.33)^2\ 0.3\ (1-0.3)}{(0.025)^2} = 1824.11$$

بنابراین حداقل حجم نمونه لازم 1825 است.

مثال : حسابرسی می خواهد توسط نمونه گیری از حسابهای یک شرکت بزرگ درصد تحریف شده حسابهای آن شرکت را برآورد کند. فرض کنید این حسابرس بخواهد نسبت حسابهای تحریف شده این شرکت را طوری برآورد کند که با ضریب اطمینان 95٪ دقت برآورد یا حداکثر خطای حدی 0.03 باشد. هر کدام از حالات زیر حداقل حجم نمونه لازم را بدست آورید.

الف: فرض کنید در یک نمونه مقدماتی این نسبت 0.2 برآورد شده باشد.

ب: فرض کنید حسابداری که آشنا به حسابهای این نوع شرکتها است معتقد باشد که 80٪ از حسابهای این شرکت تحریف شده باشد و بخواهیم از اطلاعات تجربی این حسابدار استفاده کنیم.

ج: فرض کنید هیچکونه اطلاعاتی در خصوص این نسبت نداشته باشیم.

حل:

فرض کنید P نسبت حسابهای تحریف شده توسط این شرکت باشد. داریم:

$1-\alpha = 0.95$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$, $e = 0.03$

الف: در این حالت داریم $\bar{P} = 0.2$ بنابراین:

$$n \geq \frac{z_{\frac{\alpha}{2}}^2\ (\bar{p})\ (1-\bar{p})}{e^2} = \frac{1.96^2 \times 0.2 \times (1-0.2)}{0.03^2} = 682.95 \sim 683$$

بنابراین حداقل حجم نمونه 683 می باشد.

ب: در این حالت داریم $\bar{P} = 0.8$ بنابراین:

$$n \geq \frac{(1.96)^2 \times (0.8) \times (1-0.8)}{(0.03)^2} = 683.95$$

و لذا حداقل حجم نمونه در این حالت 683 نیز می باشد.

ج: در این حالت چون هیچگونه اطلاعائی در خصوص P نداریم آنرا مساوی 0.5 قرار می دهیم بنابراین:

$$n \geq \frac{(1.96)^2 \times (0.5) \times (1-0.5)}{(0.03)^2} = 1067.11$$

بنابراین حداقل حجم نمونه لازم 1068 می باشد.

تحلیل واریانس (آنالیز واریانس) یک عاملی(ANOVA):

در تحلیل واریانس می خواهیم اثر یک عامل را بر روی یک متغیر وابسته مورد بررسی قرار دهیم که عامل یک متغیر کیفی طبقه ای (رسته ای) است و متغییر وابسته یک متغییر کمی و در مقیاس فاصله ای یا نسبتی است که فرض می شود توزیع آن نرمال است. در حالت ساده تحلیل واریانس را به صورت یک عاملی در نظر می گیریم. در چنین حالتی می خواهیم میانگین متغیر وابسته را در سطوح مختلف متغیر طبقه ای که معمولا آنها را تیمار می نامیم با یکدیگر مقایسه کنیم. بنابراین اگر عامل مورد نظر دارای K سطح (تیمار)باشد در واقع می خواهیم میانگین متغیر وابسته را در این K جمعیت با یکدیگر مقایسه کنیم. بر این اساس تحلیل واریانس یک عاملی در واقع شکل تعمیم یافته آزمون مقایسه دومیانگین یعنی *T-Test* است که در این حالت میانگین های بیش از دو جمعیت با یکدیگر مقایسه می شوند.

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 $0.14 = (1-0.5)^3 - 1$ 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 $6 = \binom{4}{2}$ آزمون انجام دهیم خطای نوع اول کل این آزمونها $0.26 = (1-0.5)^6 - 1$ 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 $\mu_1, \mu_2, ..., \mu_K$ میانگین های یک متغیر پیوسته در K جمعیت نرمال باشند فرضیه مورد آزمون به صورت زیر است:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_K \\ H_1 : \exists i, j : \mu_i \neq \mu_k \end{cases}$$

به این منظور ابتدا از هر کدام از این جوامع یک نمونه تصادفی می گیریم.

فرض کنید :

$x_{11}, x_{12}, ..., x_{1n}$ یک نمونه تصادفی n_1 تایی از جامعه اول

$x_{21}, x_{22}, ..., x_{2n}$ یک نمونه تصادفی n_2 تایی از جامعه دوم

و نیز $x_{k1}, x_{k2}, ..., x_{Kn}$ یک نمونه تصادفی n_K تایی از جامعهٔ k ام باشد .

فرض می کنیم این نمونه ها همه مستقل از هم باشند و صفت تحت مطالعه پیوسته و در سطح اندازه گیری فاصله ای یا نسبتی

و توزیع آن نرمال و نیز واریانس K جمعیت با یکدیگر برابر باشد.

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

مجموع مربعات انحرافات از میانگین کل را بصورت زیر تعریف می کنیم:

مجموع مربعات انحرا فات از میانگین کل$=\sum_{i=1}^{K}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_{..}\right)^2$

که در آن $\bar{x}_{..}$ میانگین کل است که به صورت زیر تعریف می شود:

میانگین کل نمونه ها $\bar{x}_{..} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n}x_{ij}$

تعداد کل نمونه ها $n = \sum_{i-1}^{K} n_i$

حال مجموع مربعات کل را بصورت زیر تجزیه می کنیم:

مجموع مربعات خطا (باقیمانده) + مجموع مربعات تیمار = مجموع مربعات کل

$$SST = SS(Tr) + SSE$$

که در آن:

$$SS(Tr) = \sum_{i-1}^{K} n_i\left(\bar{x}_{i.} - \bar{x}_{..}\right)^2$$

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i.}\right)^2$$

میانگین مشاهدات در تیمار iام ⇘ $\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n} x_{ij}$

در عمل برای محاسبه مجموع مربعات به صورت زیر عمل می شود.

حجم کل نمونه ها $n = \sum_{i-1}^{K} n_i$

مجموع مشاهدات در تیمارiام $T_{i.} = \sum_{j=1}^{n_i} x_{ij}$

مجموع کل مشاهدات $T_{..}=\sum_{i=1}^{K}\sum_{j=1}^{n_i}X_{ij}=\sum_{i=1}^{k}T_{i.}$

مجموع مربعات کل $SST=\sum_i\sum_j x_{ij}^{2}-\frac{T_{..}^{2}}{N}$

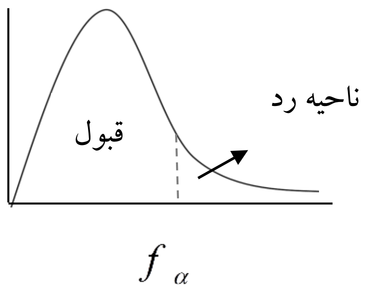
مجموع مربعات تیمار $SS(Tr)=\sum_{i=1}^{K}\frac{T_i^2}{n_i}-\frac{T_{..}^2}{N}$

$$SSE=SST-SS(Tr)$$

پس از به دست آوردن مجموع مربعات، جدول زیر را که آن را جدول تحلیل واریانس یا جدول ANOVA می نامیم به صورت زیر تنظیم می کنیم .

منبع تغییرات(پراکندگی)	مجموع مربعات(SS)	درجه آزادی(df)	میانگین مربعات(MS)	آماره آزمون F_0
بین تیمارها(اثر عامل)	$SS(Tr)$	k-1	$MS(Tr)=\frac{SS(Tr)}{k-1}$	
داخل تیمارها (خطا)	SSE	(N-1)-(k-1)=N-k	$MSE=\frac{SSE}{N-k}$	$F_0=\frac{MS(Tr)}{MSE}$
کل	SST	N-1	....	

آماره آزمون دارای توزیع فیشر با درجات آزادی به ترتیب k-1 و N-k است(درجه آزادی صورت و مخرج). لذا اگر F_0 از F_α بیشتر شود ($F_0 \rangle F_\alpha$) فرض صفر رد می شود.



توجه 1:

در تحلیل واریانس یک عاملی، اگر حجم نمونه ها با هم برابر باشد طرح را متعادل و در غیر اینصورت آن را نامتعادل گویند.

توجه 2:

همانطوریکه قبلا نیز بیان شده است در تحلیل واریانس یک عاملی در واقع اثر یک عامل بر روی یک متغیر وابسته مورد بررسی قرار می گیرد که عامل یک متغیر کیفی رسته ای است. مثلا فرض کنیم بخواهیم تاثیر نوع فعالیت در بازده سهام

شرکت های بورسی را مورد بررسی قرار دهیم. در این صورت متغیر وابسته بازده سهام است که این متغیر کمی و فرض می کنیم توزیع آن نرمال باشد و متغیر رسته ای یا همان عامل نوع فعالیت شرکت است که این عامل داری سیزده تیمار می باشد و تیمارها شامل تولیدی،خدماتی،بازرگانی،سرمایه گذاری و . . . است. یا مثلا فرض کنید بخواهیم تاثیر سطح تحصیلات بر حقوق دریافتی کارکنان یک شرکت بزرگ را مورد بررسی قرار دهیم. در اینصورت متغیر وابسته حقوق دریافتی کارکنان است که یک متغیر کمی است و فرض می شود توزیع آن نرمال باشد و عامل نیز سطح تحصیلات کارکنان است که مثلا دارای 4 تیمار دیپلم، فوق دیپلم، لیسانس و فوق لیسانس است.

بنابر این یک مدل تحلیل واریانس یک عاملی به صورت زیر است.

$$x_{ij} = \mu + \alpha_i + e_{ij} \qquad i = 1,2,\dots,k \qquad j = 1,2,\dots,n_i$$

که در آن:

x_{ij} : اندازه متغیر وابسته نزد فرد j ام در تیمار i ام است و فرض می شود که x_{ij} ها همگی مستقل از یکدیگر و هر کدام دارای توزیع نرمال با میانگین μ_i و واریانس ثابت و مشترک σ^2 هستند. یعنی $x_{ij} \sim N(\mu_i, \sigma^2)$.

μ: میانگین کل

α_i: اثرتیمار iاُم که فرض می شود $\sum_{i=1}^{K} \alpha_i = 0$

e_{ij} : مقدار باقیمانده یا خطا که فرض می شود e_{ij} ها متغیرهای تصادفی مستقل از یکدیگر و دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 هستند یعنی $e_{ij} \sim N(0, \sigma^2)$.

در چنین صورتی و با در نظر گرفتن مدل فوق فرضیه های آماری به صورت زیرتبدیل می شود:

$$\begin{cases} H_0 : \mu_1 = \mu_2, \dots, = \mu_K \\ H_1 : \text{دست کم 2 میانگین نابرابرند} \end{cases} \Rightarrow \begin{cases} H_0 : \alpha_1 = \alpha_2 = \dots, = \alpha_k = 0 \\ H_1 : \alpha_j \neq 0 \quad \text{دست کم یکی از تیمارها موثر است} \end{cases}$$

مثال کتاب عادل آذر صفحه 147 :

فرض کنید بخواهیم تعداد ضایعات سه ماشین را با هم مقایسه کنیم از هر کدام از این ماشین ها و در پنج روز ضایعات آنها اندازه گیری شده است و نتیجه به صورت زیر بوده است :

ماشین اول	86	79	81	70	84
ماشین دوم	89	82	88	76	90
ماشین سوم	82	68	73	71	81

در سطح خطای 5 درصد مشخص کنید که آیا متوسط ضایعات این 3 ماشین تفاوت معناداری با یکدیگر دارند یا خیر ؟

حل :

دراین مثال عامل یا فاکتور، نوع ماشین است که دارای سه سطح یا سه تیمار ماشین 1 ، ماشین 2 و ماشین 3 است ومتغیر پاسخ یا وابسته میزان ضایعات این ماشینهاست. فرض کنید: μ_3, μ_2, μ_1 به ترتیب بیانگر میانگین ضایعات ماشینهای 1 ، 2 ، 3 باشد. فرضیهٔ مورد آزمون بصورت زیر است:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 :$ میانگین ضایعات دست کم2ماشین نابرابراست

در این مثال مدل تحلیل واریانس عبارتست از: $x_{ij} = \mu + \alpha_i + e_{ij}$ $\quad i = 1,2,3 \quad j = 1,...,5$ که در آن :

x_{ij} : میزان ضایعات در روز j اُم درماشین i اُم .

μ : میانگین کل ضایعات.

α_i=اثر ماشین i اُم بر ضایعات

e_{ij} = میزان باقی مانده یا خطا

با در نظر گرفتن این مدل فرضیه های آماری بصورت زیر تبدیل می شود:

$H_0 :$ عامل ماشین اثری برضایعات ندا، د

$H_1 :$ عامل ماشین برضایعات اثردارد دا، د.

که آنرا بر حسب مدل تحلیل واریانس می توان به صورت زیر نوشت.

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 : \exists i : \alpha_i \neq 0 \end{cases}$$

پس از تشریح مسئله و تعیین مدل آنالیز واریانس حال می خواهیم آزمون را انجام دهیم که برای انجام آزمون باید جدول تحلیل واریانس را تشکیل دهیم و برای تشکیل جدول تحلیل واریانس ، نیز لازم است که ابتدا مجموع مربعات را حساب کنیم، بدین منظور محاسبات زیر را انجام دهیم.

ماشینها (تیمارها)	نمونه	Ti.
ماشین 1	86, 79, 81, 70, 84	$T_{1.} = 400$
ماشین 2	89, 83, 88, 76, 90	$T_{2.} = 425$
ماشین 3	82, 68, 73, 71, 81	$T_{3.} = 375$
		$T_{..} = 1200$

$n_1 = n_2 = n_3 \rightarrow N = \sum n_i = 15$ ، $k = 3$

$$\sum_i^k \sum_j^{ni} x_{ij}^2 = 86^2 + 79^2 + \ldots + 81^2 = 96698$$

$$\frac{T..^2}{N} = \frac{(1200)^2}{15} = 96000$$

$$\sum \frac{T_{i.}^2}{n_i} = \frac{400^2}{5} + \frac{425^2}{5} + \frac{375^2}{5} = 96250$$

$$SST = \sum \sum x_{ij}^2 - \frac{T..^2}{N} = 96698 - 96000 = 698$$

$$SS(Tr) = \sum \frac{T_i^2}{n_i} - \frac{T..^2}{N} = 96250 - 96000 = 250$$

$$SSE = SST - SS(Tr) = 698 - 250 = 448$$

حال جدول *ANOVA* را تشکیل می دهیم.

منبع تغییرات(پراکندگی)	SS	df	MS	Fo
بین تیمارها (اثرماشین)	250	3 -1=2	$\frac{250}{2} = 125$	
داخل تیمارها (خطا)	448	14 -2=12	$\frac{448}{12} = 37.33$	$Fo = \frac{125}{37.33} = 3.35$
کل	698	15 -1=14		

$f_{0.05,2,12} = 3.88$

چون $f_{0.05,2,12} = 3.88 \rangle Fo = 3.35 \leftarrow$ لذا در سطح خطای 0.05 فرض صفر مبنی بر یکسان بودن میانگین ضایعات هر سه ماشین، یا بی تاثیر بودن اثر عامل ماشین برضایعات پذیرفته می شود.

مثال: برای مقایسه معدل لیسانس دانشجویان ارشد حسابداری که فارغ التحصیل دانشگاه آزاد (تیمار1) دانشگاه پیام نور(تیمار 2) و سایر دانشگاهها (تیمار3) هستند یک نمونه تصادفی ار دانشجویان فارغ التحصیل هر کدام از این دانشگاه ها گرفته شده و معدل لیسانس آنها در جدول زیر آمده است.

دانشگاه	معدل لیسانس
آزاد	17.60 – 15.00 – 17.80 – 17.78 – 16.24
پیام نور	15.56 – 17.00 – 16.54 – 16.00
سایر	– 17.00 – 16.70 – 17.70 – 17.10 – 17.50 17.29

فرض یکسان بودن معدل لیسانس فارغ التحصیلان دانشگاههای مختلف یا بی تاثیر بودن نوع دانشگاه بر معدل لیسانس را در سطح خطای 5٪ آزمون کنید.

حل:

فرض کنید μ_1 میانگین معدل لیسانس فارغ التحصیلان دانشگاه آزاد، μ_2 میانگین معدل لیسانس فارغ التحصیلان دانشگاه پیام نور، و μ_3 میانگین معدل لیسانس فارغ التحصیلان سایر دانشگاهها باشد . فرضیه مورد آزمون به صورت زیر است.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1 : دست کم 2 میانگین نابرابرند

برای آزمون این فرضیه لازم است جدول $ANOVA$ را تشکیل داد و برای تشکیل این جدول نیز باید ابتدا مجموع مجذورات را بدست بیاوریم. بدین منظور داریم:

دانشگاه	نمونه	n_i	T_{00}	$(T_{00})^2$
آزاد	*16.24,……., 17.60*	$n = 5$	$T_{10} = 84.42$	*7126.73*
پیام نور	*16.00 ,………, 15.56*	$n = 4$	$T_{20} = 65.10$	*4238.01*
سایر	*17.50 , ………, 17.29*	$n = 6$	$T_{30} = 103.29$	*10669*
		$N = 15$		

$N=\sum n = 15$, $\sum T_{ij} = 252.81$, $\frac{T_{00}^2}{N} = \frac{(252.81)^2}{15} = 4260.85$

$\sum_i \sum_j X_{IJ}^2 = (16.24)^2 + \ldots\ldots\ldots + (17.29)^2 = 4270.94$

$\sum \frac{T_i^2}{N} = \frac{7126.73}{5} + \frac{4238.01}{4} + \frac{10669}{6} = 4263.01$

$SST = \sum_i \sum_j X_{ij}^2 - \sum \frac{T_i^2}{N} = 4270.94 - 4260.85 = 10.09$

$SS(Tr) = \sum \frac{T_i^2}{N} - \frac{T_{00}^2}{N} = 4263.01 - 4260.85 = 2.16$

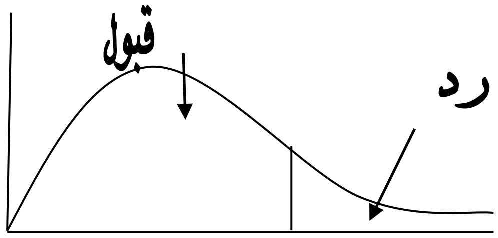
$SSE = SST - SS(Tr) = 10.09 - 2.16 = 7.93$

حال می توان جدول *ANOVA* را به صورت زیر شکل داد.

منبع تغییرات(پراکندگی)	*SS*	*df*	*MS*	
بین گروهها(اثر دانشگاه)	*2.16*	*3-1=2*	$\frac{2.16}{2} = 1.08$	$F_0 = \frac{1.08}{.066} = 1.63$
داخل گروهها(خطا)	*7.93*	*14-2=12*	$\frac{7.93}{12} = 0.64$	
جمع (کل)	*10.9*	*15-1=14*		

آماره آزمون دارای توزیع فیشر با درجات آزادی به ترتیب 2 و 12 است .

$\alpha = 0.05$, $f_{0.05,2,12} = 3.88$



لذا چون آماره آزمون $F_0 = 1.63 < f_{0.05} = 3.88$ لذا فرض صفر در سطح خطای 5٪ پذیرفته میشود. یعنی در سطح خطای 5٪ تفاوت معنی داری بین میانگین معدل لیسانس فارغ التحصیلان دانشگاههای مختلف نیست. یا به عبارت دیگر در سطح خطای 5٪ دانشگاه محل تحصیل اثر معنی داری بر معدل لیسانس ندارد.

رگرسیون و همبستگی :

فرض کنید $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ یک نمونه تصادفی n تایی از مقادیر دو متغیر x و y در یک جامعه باشند. یعنی یک نمونه داریم که روی هر فرد نمونه دو ویژگی اندازه گیری شده است و زوج مرتب (x_i , y_i) ، بیانگر مقادیر این دو صفت روی فرد i اُم نمونه است. در این صورت همان طوری که از قبل می دانیم، میانگین و واریانس های نمونه بصورت زیر تعریف می شود.

$$\bar{x} = \frac{1}{n}\sum xi \qquad \bar{y} = \frac{1}{n}\sum yi$$

$$S_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}(\bar{x})^2 \quad \text{و} \quad S_x = \sqrt{S_x^2}$$

$$S_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\sum {y_i}^2 - \frac{n}{n-1}(\bar{y})^2 \quad \text{و} \quad S_y = \sqrt{S_y^2}$$

کواریانس نمونه که آن را با $S_{x,y}$ نشان می دهیم به صورت زیر تعریف می شود.

$$s_{x,y} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$S_{x,y} = \left(\frac{1}{n-1}\sum x_i y_i\right) - \frac{n}{n-1}\bar{x}\bar{y}$$

توجه کنید که در فرمول کواریانس اگر هر دو متغیر یکی باشد در این صورت فرمول کوواریانس به واریانس تبدیل می شود به عبارت دیگر کواریانس یک متغیر با خودش در واقع همان واریانس آن متغیر است به عبارت دیگر.

$$S_{x,x} = S_x^2 \quad , \quad S_{y,y} = S_y^2$$

کواریانس معیاری است برای سنجش رابطه خطی بین دو متغیر اما چون این شاخص به واحد اندازه گیری داده ها بستگی دارد، لذا به جای آن از شاخص دیگری با نام ، ضریب همبستگی استفاده می شود.

ضریب همبستگی خطی پیرسون :

که آن را به اختصار ضریب همبستگی نیز نامیده و با $r_{x,y}$ یا به اختصار با r نشان می دهیم. بصورت زیر تعریف می شود.

$$r = \frac{S_{x,y}}{S_x . S_y}$$

که با جایگذاری فرمول های کوواریانس و انحراف معیار ، فرمول های زیر برای ضریب هم بستگی بدست می آید.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

از نمادهای زیر برای نشان دادن مجموع مربعات استفاده می شود.

مجموع مربعات x $\qquad SS_x = \sum(x_i - \bar{x})^2$

مجموع مربعات y $\qquad SS_y = \sum(y_i - \bar{y})^2$

مجموع حاصلضرب $\qquad SP_{x.y} = \sum(x_i - \bar{x})(y_i - \bar{y})$

با داشتن چنین نمادهایی واریانس ،کوواریانس و ضریب همبستگی را می توان به صورت زیر نوشت.

$$s_x^2 = \frac{ss_x}{n-1} \quad ، \quad s_y^2 = \frac{ss_y}{n-1} \quad ، \quad s_{x,y} = \frac{sp_{x,y}}{n-1} \quad ، \quad r = \frac{SP_{xy}}{\sqrt{SS_x . SS_y}}$$

توجه:

1. ترتیب قرار گرفتن x,y در محاسبه کواریانس و ضریب همبستگی بی تاثیر است یعنی:

$$S_{x,y} = S_{y,x} \quad , \quad r_{x,y} = r_{y,x}$$

2. ضریب همبستگی به واحد اندازه گیری داده ها بستگی ندارد.

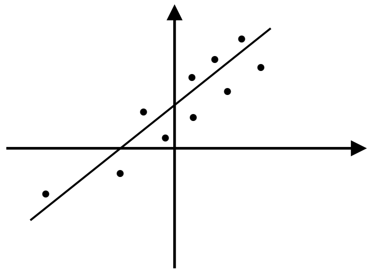
3. مقیاس یا سطح اندازه گیری ضریب همبستگی ، فاصله ای است (مقیاس فاصله ای، نسبت را رعایت نمی کند) و بنابراین این شاخص نسبت را رعایت نمی کند بر این اساس اگر مثلا ضریب همبستگی بین قد و وزن دانشجویان در کلاس الف مساوی 0/3 و در کلاس ب برابر 0/9 باشد ، در این صورت نمی توان نتیجه گرفت که شدت همبستگی بین قد و وزن دانشجویان در کلاس ب، سه برابر کلاس الف است بلکه تنها می توان نتیجه گرفت که شدت این همبستگی در کلاس ب به اندازه 0/6 بیشتر از کلاس الف است.

4. وجود همبستگی بین دو متغیر لزوماً به معنای وجود رابطه علت و معلولی بین آن دو متغیر نیست در مواردی ممکن است بین دو متغیر به تنهایی هیچ رابطه ای برقرار نباشد اما ضریب همبستگی آنها عدد بالا یی را نشان دهد که این امر می تواند به جهت وجود متغیر سوم دیگری باشد. یعنی در واقع رابطه بین این دو متغیر ناشی از وجود متغیر دیگر است که اگر آن متغیر را کنترل کنیم بین این دو متغیر رابطه ای برقرار نیست.

5. اگر دو متغیر مستقل از یکدیگر باشند یعنی هیچ رابطه ای بین آنها برقرار نباشد، کوواریانس وضریب همبستگی آنها صفر است.عکس این قضیه برقرار نیست. یعنی اگر کوواریانس یا ضریب همبستگی بین دو متغیر مساوی صفر شود نمی توان نتیجه گرفت که آن دو متغیر مستقل از یکدیگرند بلکه تنها می توان نتیجه گرفت که بین آن دو متغیر رابطه خطی برقرارنیست چیزی که آن را اصطلاحا نا همبسته گویند. البته اگر هر دو متغیر دارای توزیع نرمال باشند این قضیه دو طرفه می شود.به عبارت دیگر اگر xو yدارای توزیع نرمال باشند آنگاه x و y مستقل از یکدیگراند ، اگر و فقط اگر $r = 0$ باشد.

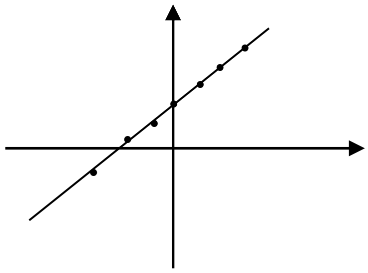
6. ضریب همبستگی حداکثر 1+ و حداقل 1- است یعنی: $-1 \le r \le 1$ است. هر چه قدر r به 1+ نزدیکتر باشد بیانگر وجود رابطه خطی ومستقیم بین آن دو متغیر است، یعنی افزایش یکی از متغیرها باعث افزایش متغیر دیگر می شود و همین طور درباره کاهش. هر چه قدر r به 1- نزدیکتر باشد بیانگر وجود رابطه خطی و معکوس بین آن دو متغیر است. یعنی افزایش یکی از متغیرها باعث کاهش متغیر دیگر می شود و بالعکس. و هرچه قدر r به صفر نزدیکتر باشد بیانگر عدم وجود رابطه خطی بین دو متغیراست.که در مورد صفر شدن ضریب همبستگی در بند 5 بطور مفصل بحث شد.

در نمودارهای زیر که آنها را نمودار پراکنش نیز می نامیم مقادیر مختلف r نشان داده شده است:



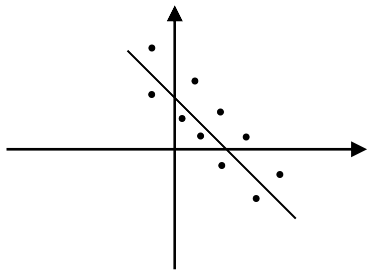
$0 \langle r \langle 1 \quad (r \simeq 1)$

رابطه خطی ناقص و مستقیم



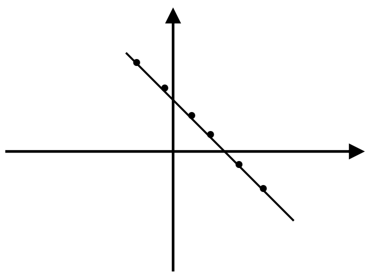
r = +1

رابطه خطی کامل و مستقیم



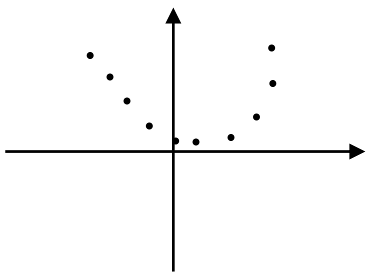
$-1 \langle r \langle 0 \qquad r \simeq -1$

رابطه خطی ناقص ومعکوس



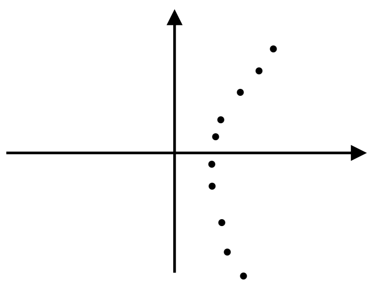
r = -1

رابطه خطی کامل و معکوس



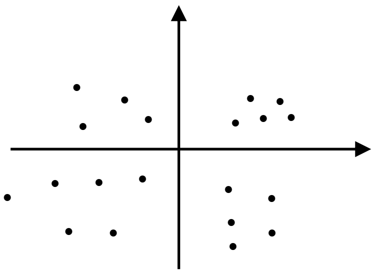
$r \simeq 0$

عدم وجود رابطه خطی (ناهمبسته)



$r \simeq 0$

عدم وجود رابطه خطی (ناهمبسته)

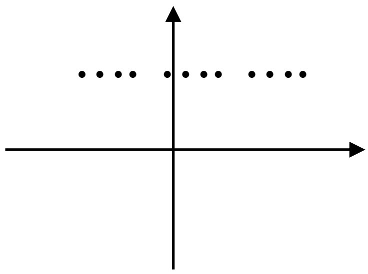


$r \simeq 0$

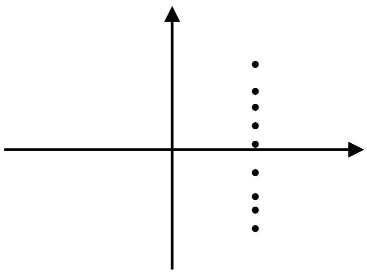
عدم وجود هر نوع رابطه ای (استقلال)

توجه:

یک متغیر و یک مقدار ثابت همواره ناهمبسته هستند یعنی $S_{x,a} = 0$ و $r_{x,a}=0$ بنابراین اگر نمودار پراکنش خطی موازی یکی از محورها شد در این حالت نیز ضریب همبستگی صفر است.



$r=0$



$r=0$

مثال : در داده های زیر ضریب همبستگی را بدست آورید:

1.

x	2	-1	3	5
y	4	-2	6	10

$$\Rightarrow y = 2x + 1 \Rightarrow r = 1$$

توجه کنید که در این مثال و با کمی دقت متوجه می شوید که بین این دو متغیر یک رابطه خطی کامل با شیب مثبت برقرار است. بنا براین بدون محاسبه ضریب همبستگی آنها مساوی یک است.

2.

x	-1	2	3	-5
y	2	-4	-6	10

$$\Rightarrow y = 2 - x \Rightarrow r = -1$$

توجه کنید که در این مثال و با کمی دقت متوجه می شوید که بین این دو متغیر یک رابطه خطی کامل با شیب منفی برقرار است. بنا براین بدون محاسبه ضریب همبستگی آنها مساوی منفی یک است.

آزمون معنی داری ضریب همبستگی:

پس ازمحاسبه ضریب همبستگی در نمونه لازم است معنی داری آن را آزمون کنیم و منظور از معنی داری ضریب همبستگی یعنی آیا همبستگی مشاهده شده در نمونه را می توان به کل جامعه تعمیم داد؟ چون همان طور که می دانیم ضریب همبستگی محاسبه شده مربوط به نمونه است. بنابراین فرض می کنیم ρ همبستگی بین دو متغیر x و y در یک جامعه و r برآورد آن در یک نمونه تصادفی n تایی باشد.آزمون معنی داری ضریب هم بستگی به صورت زیر است:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases} \Rightarrow \begin{cases} H_0 : & \text{بین دو متغیر x و y رابطه ای وجود ندارد} \\ H_1 : & \text{بین دو متغیر x و y رابطه وجود دارد} \end{cases}$$

H_0 : y, x ناهمبسته اند

H_1 : y , x همبسته اند

H_0 : ضریب همبستگی محاسبه شده درنمونه معنی دار نیست

H_1 : ضریب همبستگی محاسبه شده درنمونه معنی دار است

آماره این آزمون به صورت روبرو است. $T = r\sqrt{\frac{n-2}{1-r^2}}$

که توزیع این آماره تحت فرض صفر، استیودنت با درجه آزادی n-2 است بنابراین اگر $|T| > t_{\frac{\alpha}{2},n-2}$ باشد فرض صفر رد می شود.

مثال: در داده های زیر ضریب همبستگی را بدست آورید و معنی دار بودن آن را آزمون کنید.

x	5	7	9
y	20	15	13

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	x^2	y^2	x.y
5	20	–2	4	4	16	–8	25	400	100
7	15	0	–1	0	1	0	49	225	105
9	13	2	–3	4	9	–6	81	169	117
21	480			8	26	–14	155	794	322

$\bar{x} = \frac{21}{3} = 7 \quad , \quad \bar{y} = \frac{48}{3} = 16$

$SS_x = \sum(x_i - \bar{x})^2 = 8 \quad , \quad SS_y = \sum(y_i - \bar{y})^2 = 26 \quad , \quad SP_{x.y} = \sum(x_i - \bar{x})(y_i - \bar{y}) = -14$

$$r = \frac{SP_{xy}}{\sqrt{SS_x . SS_y}} = \frac{-14}{\sqrt{8 \times 26}} = -0.97$$

ضریب همبستگی نمونه که نزدیک به منفی یک است . حال معنی دار بودن آن را آزمون می کنیم.

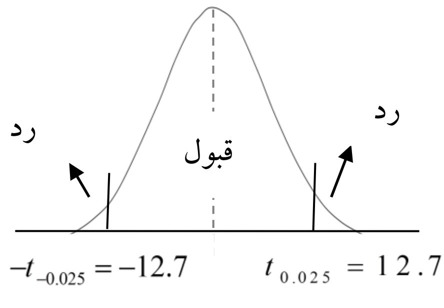
$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$

H_0 : ضریب همبستگی محاسبه شده درنمونه معنی دار نسیت

H_1 : ضریب همبستگی محاسبه شده درنمونه معنی دار است

$T = r\sqrt{\frac{n-2}{1-r^2}} \Rightarrow \qquad T = -0.970\sqrt{\frac{3-2}{1-(-0.970)}} = -0.970\sqrt{\frac{1}{0.0591}} = -4.04$

dF=n-2=3-2=1 $\qquad$ *α=0.05* $\qquad t_{\frac{\alpha}{2}, n-2} = t_{.025,1} = 12.706$



چون $-12.7 \le T = -4.04 \le 12.7$ شد فرض صفر پذیرفته می شود. یعنی ضریب هم بستگی محاسبه شده در نمونه معنی دار نیست.

تمرین: در مثال قبل ضریب همبستگی را مجددا محاسبه کنید و این بار آن رافرمول زیر بدست آورید.

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}} = \frac{-14}{\sqrt{8 \times 26}} = \frac{-14}{14.42} = -0.97$$

مثال: داده های زیر نمرات درس پژوهش عملیاتی تعدادی از دانشجویان ارشد حسابداری در مقطع کارشناسی(x) و مقطع کارشناسی ارشد (y) می باشد.

x	18	18	18	20	16	17	15	20	15	18
y	12	16.5	15	19	19	15	15	18	16	19.5

با فرض اینکه این داده ها را بتوان یک نمونه تصادفی از کل دانشجویان ارشد حسابداری این دانشگاه دانست و نیز با فرض اینکه توزیع نمرات درس پژوهش عملیاتی در هر دو مقطع نرمال باشد تعیین کنید آیا در سطح خطای 5٪ ارتباط معنی داری بین نمرات این درس در این دو مقطع وجود دارد یا خیر؟

حل:

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
18	12	0.5	-4.5	0.25	20.25	-2.25
18	16.5	0.5	0	0.25	0	0
18	15	0.5	-1.5	0.25	2.25	-0.75
20	19	2.5	2.5	6.25	6.25	6.25
16	19	- 1.5	2.5	2.25	6.25	-3.75
17	15	-0.5	-1.5	0.25	2.25	0.75
15	15	-2.5	-1.5	6.25	2.5	3.75
20	18	2.5	1.5	6.25	2.25	3.75
15	16	-2.5	-0.5	6.25	0.25	1.25
18	19.5	0.5	3	0.25	9	1.5
175	165			28.5	51.25	10.5

$n = 10$, $\bar{x} = \frac{175}{10} = 17.5$, $\bar{y} = \frac{165}{10} = 16.5$

$ss_x = \sum(x_i - \bar{x})^2 = 28.5$, $ss_y = \sum(y_i - y)^2 = 51.25$,

$sp_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 10.5$

$s_x^2 = \frac{ss_x}{n-1} = \frac{28.5}{9} = 3.16$, $s_x = 1.78$,

$s_y^2 = \frac{ss_y}{n-1} = \frac{51.25}{9} = 5.67$, $s_y = 2.38$

$s_{xy} = \frac{sp_{xy}}{n-1} = \frac{10.5}{9} = 1.17$

$r = \frac{sp_{xy}}{\sqrt{ss_x ss_y.}} = \frac{1.17}{\sqrt{3.16*5.67}} = 0.23$

حال معنی دار بودن این ضریب را آزمون می کنیم .

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

H_0 : ضریب همبستگی محاسبه شده درنمونه معنی دار نسیت

H_1 : ضریب همبستگی محاسبه شده درنمونه معنی دار است

آماره آزمون $T = r\sqrt{\frac{n-2}{1-r^2}} = 0.23\sqrt{\frac{8}{1-(0.23)^2}} = 0.67$

$\propto = 0.05$, $\frac{\propto}{2} = 0.025$, $df = n-2 = 8$

$t_{0.025,8} = 2.306$

چون $-2.3 < T = 0.67 < 2.3$ لذا فرض صفر در سطح خطای 5٪ پذیرفته می شود. یعنی ضریب همبستگی محاسبه شده در نمونه در سطح خطای 5٪ معنی دار نبوده و بر این اساس بین نمرات این درس در این دو مقطع رابطه معنی داری برقرار نیست.

رگرسیون:

در تحلیل رگرسیون می خواهیم رابطه بین یک یا بیش از یک متغیر مستقل که معمولا آنها را متغیر پیشگو می نامند را با یک متغیر وابسته که آن را متغیر پاسخ می نامند مورد بررسی قرار دهیم. متغیر های مستقل یا همان متغیرهای پیشگو، متغیرهای

کمی یا کیفی هستند اما متغیر پاسخ یک متغیر کمی است که فرض می شود دارای توزیع نرمال است. البته حالتی که در آن متغیر پاسخ یک متغیر کیفی است نیز وجود دارد که این حالت قسمت خاصی از رگرسیون به نام رگرسیون لجستیک یا رگرسیون لجیت است.

در حالت ساده تعداد یک متغیر مستقل (که آن را متغیر توضیحی یا توصیفی یا متغیر پیشگو نیز می نامند)و یک متغیر وابسته یا همان متغیر پاسخ در نظر می گیریم . متغیر پیشگورا با x و متغیر پاسخ را با y نشان می دهیم و فرض می کنیم بین این دو متغیر یک رابطه خطی به شکل زیر برقرار است.

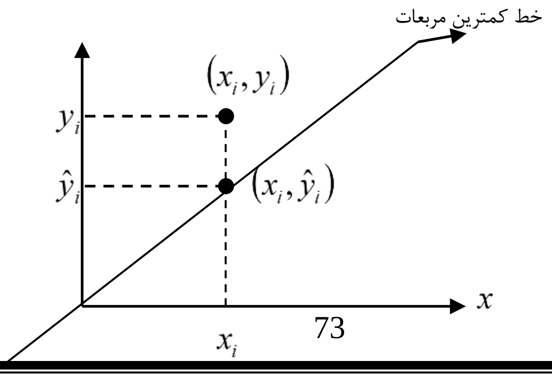
$$Y = \alpha + \beta X + \varepsilon$$

که در آن α و β پارامترهای ثابت و مجهول و ε یک متغیر تصادفی دارای توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 است یعنی $\varepsilon \sim N(0,\sigma^2)$

در رگرسیون هدف برآورد پارامترهای α و β از روی داده های یک نمونه تصادفی از این دو متغیر است.

فرض کنید $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ یک نمونه تصادفی n تایی از مقادیر این دو صفت باشند. برای تعیین این که آیا بین این دو متغیر رابطه خطی برقرار است یا خیر ابتدا لازم است نمودار پراکنش این داده ها را رسم کنیم ، فرض کنید از روی نمودار پراکنش رابطه بین این دو متغیر خطی تشخیص داده شود. بنابراین باید ضریب زاویه و عرض از مبدا این خط(همان α و β) را از روی داده های نمونه برآورد کنیم.

یکی از عمده ترین روش های برآورد β,α روشی موسوم به روش کمترین مربعات(OLS) می باشد که مطابق این روش β,α به نحوی برآورد می شوند که خطی که از بین نقاط نمونه عبور می کند به طور متوسط کمترین فاصله را با همه نقاط نمونه داشته باشد که این خط همان خط رگرسیون است که آن را خط کمترین مربعات نیز می گوییم. برای تشریح روش کمترین مربعات ، نمودار پراکنش زیر را در نظر می گیریم.



در این نمودار نقطه (x_i , y_i) نقطه i اُم نمونه و نقطه ($x_i , \hat{y}_i$) نقطه متناظر با آن روی خط رگرسیون است. یعنی $\hat{y}_i$ برآورد y_i توسط خط رگرسیون است.اختلاف این دو مقدار را با e_i نشان می دهیم یعنی $e_i = y_i - \hat{y}_i$ است و آن را باقیمانده یا خطا می نامیم.چون خط رگرسیون باید از بین تمام نقاط عبور کند لذا مجموع باقی مانده ها صفر است یعنی $\sum e_i = 0$.

در روش کمترین مربعات، ضریب زاویه و عرض از مبدا خط رگرسیون را به نحوی برآورد می کنیم که مجموع مربعات یا مجموع مجذورات باقی مانده ها یعنی $\sum e_i^2$ مینیمم شود اما

$$\sum e_i^{\,2} = \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum\left(y_i - (\alpha + \beta x)\right)^2$$

اما رابطه فوق تابعی است از β,α، بنابراین با استفاده از روش های ریاضی β,α را به نحوی بدست می آوریم که این تابع مینیمم شود.

$$f(\alpha,\beta) = \sum e_i^2 = \sum\left(y_i - (\ \alpha + \beta x_i)\ \right)^2 = \sum(y_i - \alpha - \beta x_i\ )^2$$

$$\begin{cases} \dfrac{\partial f}{\partial \alpha} = -2\sum\left(y_i - \alpha - \beta x_i\right) = 0 \\ \dfrac{\partial f}{\partial \beta} = -2\sum x_i\left(y_i - \alpha - \beta x_i\right) = 0 \end{cases}$$

که با حل دستگاه معادله فوق که آنها را معادلات نرمال نیز می نامند ، فرمول های برآورد β,α که آنها را با $\hat{\alpha},\hat{\beta}$ یا a و b نشان می دهیم به صورت زیر به دست می آید.

برآورد شیب خط رگرسیون $$\hat{\beta} = \frac{s_{x,y}}{s_x^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum xi^2 - n\bar{x}^2} = \frac{SP_{x,y}}{SS_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$r = \frac{S_{x,y}}{S_x S_y} \quad \Rightarrow \quad \hat{\beta} = r\frac{s_y}{s_x}$$

برآورد عرض از مبدا خط رگرسیون $$\hat{\alpha} = a = \bar{y} - b\bar{x}$$

پس از برآورد β,α معادله خط رگرسیون به شکل زیر برآورد می شود: $\hat{y} = \hat{\alpha} + \hat{\beta}x$

که با داشتن چنین معادله ای و معلوم بودن یک مقدار برای متغیر مستقل می توان یک مقدار برای متغیر وابسته پیش بینی کرد.

توجه: معادله خط رگرسیون همواره از نقطه $(\bar{x}\,,\bar{y})$ عبور می کند.

ضریب تعیین (ضریب تشخیص) :

ضریب تعیین معیاری است که برای سنجش دقت رابطه برازش شده بین متغیرها استفاده می شود و آن را با R^2 یا R- Square نشان می دهیم. برای تعریف ضریب تعیین تعاریف زیر را در نظر می گیریم.

انحرافات کل $= (y-\bar{y})$

انحرافات بیان شده توسط رگرسیون $= (\hat{y}-\bar{y})$

انحرافات بیان نشده $= (y-\hat{y})$

مجموع مربعات انحرافات کل $= \sum_{i=1}^{n_i}(y_i-\bar{y})^2$

مجموع مربعات انحرافات بیان شده $= \sum_{i=1}^{n_i}(\hat{y}_i-\bar{y})^2$

مجموع مربعات انحرافات بیان نشده $= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

ثابت می شود که :

مجموع مربعات انحرافات بیان نشده + مجموع مربعات انحرافات بیان شده = مجموع مربعات انحرافات کل

یا

پراکندگی بیان نشده + پراکندگی بیان شده = پراکندگی کل

ضریب تعیین به صورت زیر تعریف می شود :

$$R^2 = \frac{\text{پراکندگی بیان شده توسط رگرسیون}}{\text{پراکندگی کل}} \quad \text{یا} \quad R^2 = 1-\frac{\text{پراکندگی بیان نشده}}{\text{پراکندگی کل}}$$

بنابراین در این جا ضریب تعیین در واقع نسبت پراکندگی تبیین شده توسط معادله رگرسیون به پراکندگی کل است یعنی ضریب تعیین مشخص میکند که چند درصد از تغییرات یا پراکندگی در متغیر وابسته (پاسخ) توسط متغیر مستقل (پیشگو) تبیین می شود .به عنوان مثال اگر در یک معادله رگرسیونی ضریب تعیین 0/6 باشد این بدان معنی است که 60٪ از تغییرات (پراکندگی) در متغیر وابسته توسط معادله رگرسیون به متغیر مستقل مربوط می شود و 40٪ دیگر به عوامل دیگری مرتبط است که در این مدل در نظر گرفته نشده است .

توجه *: در رگرسیون خطی ساده ضریب تعیین برابر است با مجذور ضریب همبستگی یعنی :* $R^2 = r_{x,y}^2$ *ولی در سایر رگرسیون ها (چندگانه و غیرخطی) ضریب تعیین مجذور ضریب همبستگی چند گانه است.*

ضریب تعیین حداقل صفر و حداکثر یک است یعنی $0 \le R^2 \le 1$ *هر چه قدر* R^2 *به یک نزدیک تر باشد بیانگر معتبر بودن معادله رگرسیون است و هرچه قدر ضریب تعیین به صفر نزدیک تر باشد بیانگر عدم اعتبار معادله رگرسیون است و مثلا اگر در یک معادله رگرسیون ضریب تعیین برابر 0/7 باشد این به آن معنی است که 70 درصد از واریانس یا پراکندگی در متغیر وابسته توسط این مدل رگرسیونی به متغیر مستقل بستگی دارد یا به عبارت دیگر این مدل 70 درصد از واریانس یا پراکندگی در متغیر وابسته را تبیین می کند.*

مثال : اگر معادله رگرسیون به شکل $\hat{y} = 3 - 2x$ و ضریب تعیین این مدل $R^2 = 0.81$ باشد ضریب هم بستگی x و y را به دست آورید.

$$R^2 = 0.81 \Longrightarrow r_{x,y} = \pm 0.9 \xRightarrow{\text{شيب خط منفي}} r_{x,y} = -0.9$$

چون شیب خط منفی است رابطه بین این دو متغیر معکوس است و لذا $r_{x,y} = -0.9$ قابل قبول است .

نتیجه این که در رگرسیون خطی ساده شیب خط و ضریب هم بستگی همواره هم علامت هستند .

توجه:

در یک رگرسیون خطی ساده $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ *،* $\hat{\beta}_1$ *یا همان شیب خط رگرسیون میزان تغییر در میانگین متغیر پاسخ (y) است هرگاه متغیر پیشگو (x) یک واحد تغییر کند مثلا در معادله* $\hat{y} = 2 + 3x$ *می توان نتیجه گرفت که هر یک واحد افزایش درمتغیر x، باعث افزایش 3 واحد در y می شود.*

اگر دامنه داده ها شامل x=0 باشد آنگاه $\hat{\beta}_0$ *یا همان عرض از مبدا میانگین متغیر پاسخ به ازای x=0 است در غیر این صورت* $\hat{\beta}_0$ *هیچ گونه تعبیر عملی ندارد.*

آزمون معنی داری رگرسیون :

در یک رگرسیون خطی ساده پس از برآورد ضرایب خط رگرسیون لازم است معنی دار بودن مدل رگرسیون و نیز معنی دار بودن ضرایب مدل را آزمون کنیم که فرضیات مورد آزمون به صورت زیر است :

$$\begin{cases} H_0: & \text{مدل رگرسیون معني دارنیست} \\ H_1: & \text{مدل رگرسیون معني دار است} \end{cases} \Longrightarrow \begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$\begin{cases} H_0: & \text{عرض از مبداء معني دارنیست} \\ H_1: & \text{عرض از مبداء معني دار است} \end{cases} \qquad \begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

که برای انجام این آزمون ها باید آماره آزمون را به دست آورده و مطابق آزمون های قبلی عمل کنیم که آماره این آزمون ها نیز آماره t است که در اینجا از ذکر فرمول های آن خودداری می شود.

مثال: در داده های مربوط به نمرات درس پژوهش عملیاتی تعدادی از دانشجویان ارشد حسابداری، فرض کنید نمره درس پژوهش در مقطع ارشد تابعی از نمره این درس در مقطع کارشناسی باشد. در این صورت معادله رگرسیونی این دو متغیر را بدست آورده، ضریب تعیین آن را محاسبه کرده و آن را تفسیر کنید.

حل:

$b_1 = \frac{sp_{xy}}{ss_x} = \frac{8.625}{28.5} = 0.3$, $b_0 = \bar{y} + b\bar{x} = 16.58 - 0.3(17.5) = 11.33$

$\hat{y} = 11.33 + 0.3\,x$ معادله خط رگرسیون

$R^2 = r^2 = (0.23)^2 = 0.05$ ضریب تعیین

حال با داشتن چنین مدلی و با معلوم بودن نمره پژوهش در مقطع کارشناسی برای یک دانشجو می توان نمره پژوهش مقطع ارشد او را پیش بینی کرد. مثلاً اگر دانشجویی در مقطع کارشناسی در درس پژوهش نمره 10 گرفته باشد در اینصورت پیش بینی می کنیم نمره آن در مقطع ارشد 14.33 شود. ($y = 11.33 + 0.3 * 10= 14.33$)

که البته دقت این پیش بینی بسیار پایین است چون ضریب تعیین این مدل عدد خیلی کوچکی به دست آمده است. از آنجایی که ضریب تعیین این مدل 5٪ شده است لذا تنها 5٪ از تغییرات در نمره مقطع ارشد به نمره مقطع کارشناسی بستگی دارد که مقدار بسیار پایین است .