فصل نهم : نمونه گیری و توزیعهای نمونه گیری

❖ چند تعریف

- **جامعه آماری (Population)**: تعدادی از عناصر مطلوب مورد نظر که دارای یک صفت مشخصه باشند.
  - ➢ **صفت مشخصه:** صفتی که بین همه عناصر جامعه اماری مشترک و متمایز کننده جامعه آماری از سایر جوامع باشد.
- **سر شماری (Census)**: بررسی تمام اعضای جامعه آماری.
- **پارامتر (Parameter)**: شاخص به دست آمده از جامعه آماری با استفاده از سرشماری.
- **نمونه (Sample)**: هر بخشی از جامعه آماری.
- **نمونه گیری (Sampling)**: بررسی اعضای نمونه.
- **آماره (Statistic)**: شاخص به دست آمده از یک نمونه nتایی از جامعه.

یک نمونه ۶ تایی از جامعه آماری

- ➢ از آنجایی که اعضای یک نمونه با اعضای یک نمونه دیگر متفاوت است .در نتیجه مقدار ه اماره از یک نمونه به نمونه دیگر تغییر می کند.

❖ دلایل نمونه گیری :

۱) **هزینه:** اغلب نمونه می تواند اطلاعات قابل اعتماد و مفیدی با هزینه ای کمتر از سرشماری فراهم کند .

۲) **به روز بودن:** نمونه اغلب اطلاعات به هنگام تری از سرشماری به دست می دهد. زیرا داده های کمتری جمع آوری و تجزیه و تحلیل می شوند .(این ویژگی بخصوص زمانی که اطلاعات برای تصمیم گیری سریع مورد نیاز باشد اهمیت بیشتری می یابد).

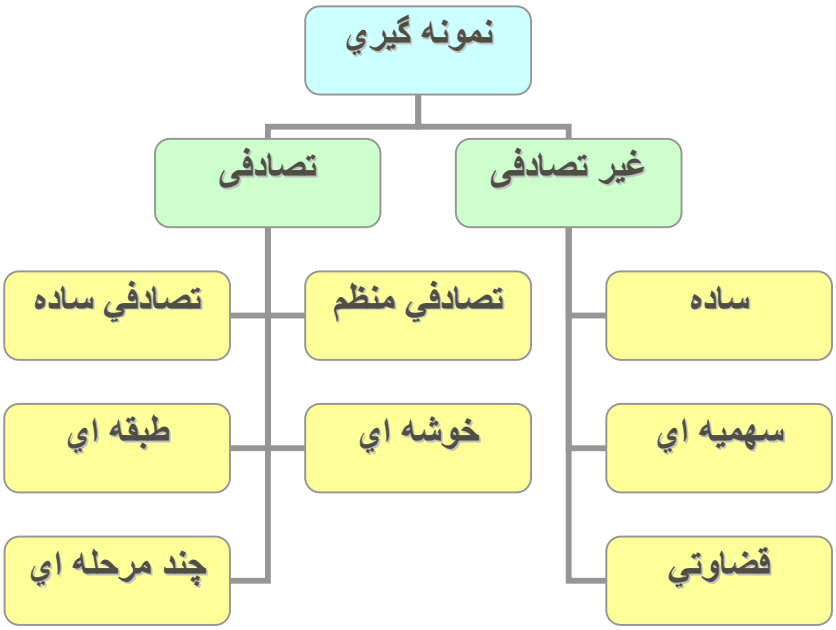
۳) **درستی:** چون برای انجام یک نمونه گیری به دلیل حجم کار کمتر از سرشما ری، امکان آموزش افراد برای تهیه پرسشنامه ها و انجام مصاحبه ها وجود دارد ، لذا صحت عمل در نمونه گیری به اندازه و یا حتی بیشتر از سرشماری است.

۴) **زمان:** سرشماری کل جامعه آماری به زمان طولانی تری نسبت به نمونه گیری نیاز دارد .

۵) **آزمون تخریب کننده:** وقتی آزمونی موجب خراب شده یک کالا می شود، باید نمونه گیری را بکار برد.

❖ **روشهای نمونه گیری:**

➢ به منظور انجام یک بررسی درست در مور د پارامترهای جامعه ، نمونه انتخاب شده باید نماینده واقعی جامعه باشد در غیر این صورت پیش بینی صحیح و دقیق درباره پارامترهای جامعه امکان نخواهد داشت.



- نمونه گیری تصادفی : نمونه گیری بر پایه شانس و احتمال.
- نمونه گیری غیر تصادفی : نمونه گیری بر پایه شانس و احتمال نبوده و محقق نظریات خود را دخالت می دهد و به افراد خاصی در جامعه چشم می دوزد.

➢ روش های نمونه گیر ی تصادفی، علمی محسوب شده و نتایج نمونه حاصل از آن با اطمینان مشخص به کل جامعه آماری قابل تعمیم است.

❖ **انواع روش های نمونه گیری تصادفی**

۱) نمونه گیری تصادفی ساده
۲) نمونه گیری منظم
۳) نمونه گیری گروهی
۴) نمونه گیری خوشه ای
۵) نمونه گیری مرحله ای

۱) **نمونه گیری تصادفی ساده:** هریک از عناصر جامعه مورد نظر برای انتخاب شدن شانس مساوی دارند .دراین روش ، افراد یا اشیاء مورد نیاز از فهرست جامعه آماری که به همین منظور شماره گذاری و تهیه شده به صورت تصادفی انتخاب می شوند این انتخاب یا به صورت قرعه کشی و یا با استفاده از جدول اعداد تصادفی انجام می گیرد .

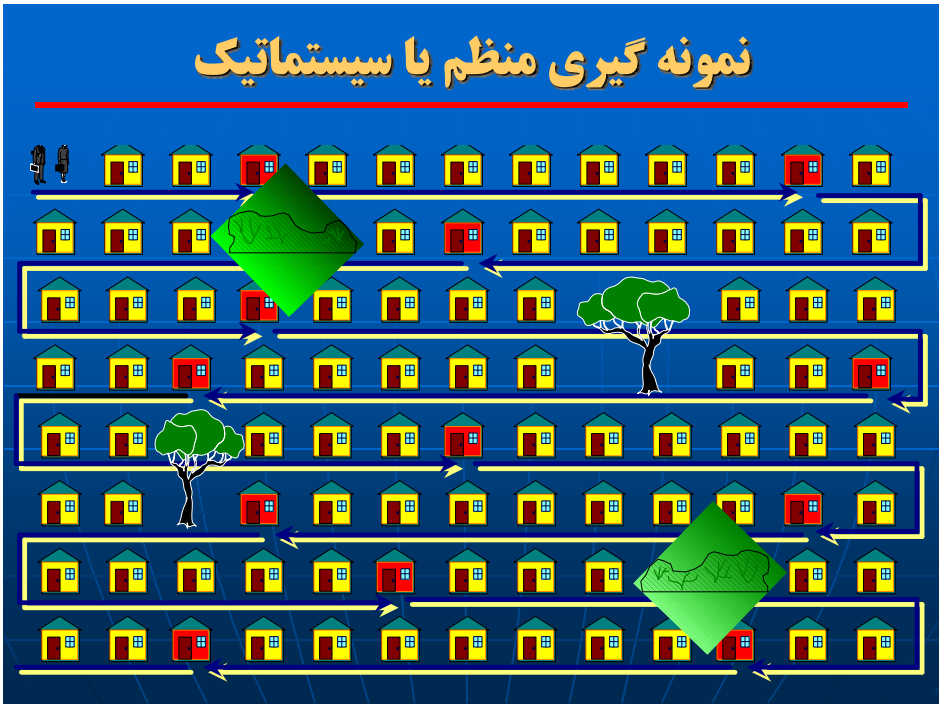
➢ یکی از مشکلات نمونه گیری تصادفی ساده تهیه و تدوین فهرست افراد جامعه آماری است.

۲) **نمونه گیری منظم**: این روش نمونه گیری زمانی که بین اعضای جامعه آماری نوعی وابستگی وجود دارد کاربرد دارد. در این روش عناصر نمونه از فهرست افراد یا اعضای جامعه آماری که به همین منظور آماده شده است، انتخاب می شوند.این روش برای آن دسته از جوامع آماری که از پیش تعیین شده و مرتبی دارند (همانند کد کارمندی ، دانشجویی و پلاک منازل) کاربرد فراوان دارد.

برای اجرای این نمونه گیری فرض کنیدN تعداد اعضای جامعه, n تعداد اعضای نمونه باشد.

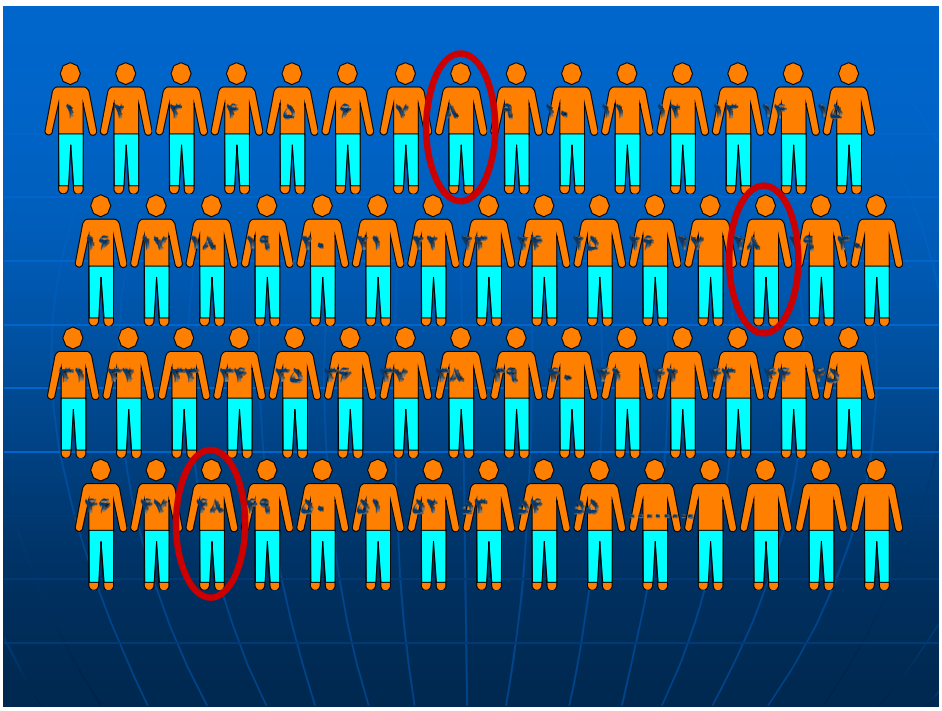
یک عدد تصادفی از ۱ تا k انتخاب کرده و k تا، K تا به نمونه انتخاب شده اضافه می کنیم.

مثال ۱: فرض کنید شما نظرات اعضای یک محله رادر مورد یک برنامه رادیویی جویا شوید، از آنجا که اعضای یک خانواده تقریبا نظرات مشابهی دارند بنابراین اعضای جامعه آماری به نوعی وابستگی دارند و نمونه گیری منظم بهتر است.



مثال۲: می خواهیم شیوع پوسیدگی دندان (برمبنای DMF) را در بین ۱۲۰۰ نفر دانش آموزان یک مدرسه تعیین کنیم.

- فهرست دانش آموزان مدرسه را تهیه کنیم
- دانش آموزان را از ۱ تا ۱۲۰۰ شماره گذاری کنیم
- اگر حجم نمونه ۶۰ نفر باشد، بایستی با توجه به ۲۰ = ۶۰ / ۱۲۰۰، از هر ۲۰ نفر یکی انتخاب کنیم.
- یک عدد تصادفی بین ۱ تا ۲۰ انتخاب می کنیم (مثلاً ۸)
- سپس ۲۰ تا ۲۰ تا به عدد فوق (۸) اضافه می کنیم به این ترتیب نمونه اول فرد شماره ۸، بعدی ۲۸، بعدی ۴۸ و .... خواهند بود.



۳) نمونه گیری طبقه بندی شده (stratified)

نمونه گیری گروهی یا طبقه ای زمانی به کار می رود که جامعه دارای ساختار همگن و متجانسی نیست. این نوع نمونه گیری را می توان به صورت زیر انجام داد.

- جامعه را به تعدادی طبقه افراز می کنیم.
  - بین طبقه ها نباید هم پوشانی وجود داشته باشد.
  - هیچ فردی از جامعه نبایستی بیرون از طبقه بندی قرار بگیرد.
- از هر طبقه تعدادی نمونه انتخاب می کنیم.

یافته های حاصل از طبقات را روی هم می ریزیم.

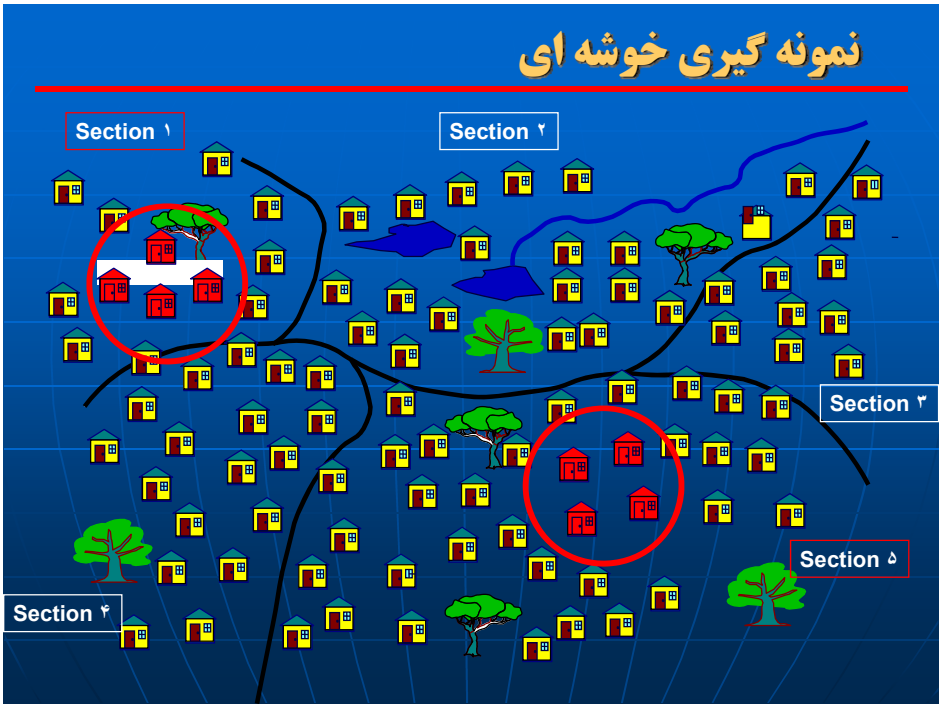
هر چقدر داخل طبقات افراد بیشتر به هم شبیه باشند ولی بین طبقه ها تفاوت زیاد وجود داشته باشد، نمونه گیری طبقه ای بهتر است.

۴) نمونه گیری خوشه ای (cluster)

هر گاه جامعه مورد بررسی خیلی وسیع و گسترده با شد. و تهیه فهرست تمامی اعضای جامعه امکان پذیر نباشد انتخاب نمونه از نظر اجرایی مشکل به نظر می رسد. در این حالت می توان از روش نمونه گیری خوشه ای استفاده کرد برای این منظور

- جامعه را به تعدادی گروه (خوشه) از واحدها تقسیم می کنیم
- تعدادی از خوشه ها را به طور تصادفی انتخاب می کنیم
- در هر خوشه همه واحدها (یا تعدادی از آنها) را انتخاب می کنیم

هر چقدر داخل خوشه ها افراد کمتر به هم شبیه باشند نمونه گیری خوشه ای موفقتر است.



۵) نمونه برداری خوشه‌ای چند مرحله‌ای

✓ بواسطه گستردگی بیش از حد جامعه محقق ناگزیر می‌گردد نمونه را طی دو یا چند مرحله انتخاب کند .

مراحل:

- ✓ جامعه موردنظر را دقیقا تعریف کنید .
- ✓ واحدها یا خوشه‌های نمونه برداری را تعریف کنید .
- ✓ تعدادی از خوشه‌ها یا واحدها را به صورت تصادفی انتخاب کنید .
- ✓ از میان خوشه‌های انتخاب شده تعداد افراد مورد نظر را به روش تصادفی انتخاب کنید.

❖ روش نمونه گیری تصادفی ساده باجایگذاری و بدون جایگذاری

اگر در یک جامعه آماری با N عضو، یک نمونه تصادفی به حجم n انتخاب شود، در صورتی که نمونه انتخابی به جامعه باز گردانده شود و دوباره نمونه بعدی انتخاب شود و این رویه تا انتخاب nامین نمونه ادامه داشته باشد، نمونه گیری را با جایگذاری و در غیر این صورت نمونه گیری را بدون جایگذاری گویند. در نمونه گیری با جایگذاری احتمال انتخاب هر یک از اعضا برابر 1/N و در صورتی که نمونه گیری بدون جایگذاری باشد شانس انتخاب اولین عضو نمونه 1/N، شانس انتخاب دومین نمونه $1/N-1$ ، ....و شانس انتخاب nامین عضو نمونه برابر $1/N-n+1$ است. تفاوت این دو روش نمونه گیری در این است که در نمونه گیری با جایگذاری امکان انتخاب یک فرد برای چندین مرتبه وجود دارد، اما در نمونه گیری بدون جایگذاری امکان تکرار یک عضو از جامعه هرگز وجود ندارد. در نظر اول روش نمونه گیری بدون جایگذاری معقول تر است ولی به دلیل راحتی کار و محاسبات آسان تر از روش نمونه گیری باجایگذاری بیشتر استفاده می شود. لازم به ذکر است در حالتی که تعداد اعضای جامعه زیاد باشد حتی اگر نمونه گیری بدون جایگذاری انجام گیرد درعمل، شبیه نمونه گیری با جایگذاری است.نمونه گیری از جامعه های بزرگ را "نمونه گیری از جامعه نامحدود" گویند که بر اساس این قاعده تعریف می شود:

$n/N<=0.05$

چنانچه قاعده فوق برقرار نباشد. حجم نمونه نسبت به جامعه آنقدر بزرگ است که می توانیم آن را "نمونه گیری از جامعه محدود" بنامیم.

❖ توزیع نمونه گیری آماره

- ➢ **برای استنباط هر پارامتر جامعه باید از یک آماره (که از نمونه حاصل می شود.) استفاده کرد، بنابراین متناظر با هر پارامتری یک آماره وجود دارد که خود یک متغیر تصادفی است.**

تابع احتمال یک آماره، تابعی است که بر اساس نمونه های تصادفی nتایی که به طور مکرر از جامعه اماری انتخاب شده است به دست می آید. این تابع را "توزیع نمونه گیری آماره" گویند.
توزیع نمونه گیری یک آماره به جامعه ای بستگی دارد که نمونه از آن حاصل شده و ممکن است به شور ریاضی و یا تقریب تجربی استنباط شود. زمانی که توزیع نمونه گیری، به طور تجربی تقریب زده می شود، چنانچه اندازه نمونه nتایی بزرگ باشد، آماره محاسبه شده از نمونه های مکرر که به صورت بافت نگار فراوانی نسبی ترسیم می شود، ممکن است تقریب بسیار خوبی برای توریع نمونه گیری واقعی باشد.

- هر گونه استنباط از آماره به توزیع آن بستگی دارد. توزیع آماره نشان دهنده رفتار آن در نمونه گیری های مکرر است.

در نمونه گیری تصادفی ساده، شانس انتخاب هر یک از عناصر جامعه آماری در نمونه n تایی یکسان است. برای انتخاب یک نمونه nتایی از یک جامعه N عضوی، نمونه ممکن وجود دارد. با این توصیف شانس انتخاب هر یک از نمونه های nتایی عبارت است از .

همانطور که می دانیم دو تا از مهمترین پارامترهایی که در بررسی های آماری مطلوب است میانگین جامعه و واریانس جامعه است که به صورت زیر تعریف می شوند.

همانطور که می دانیم به دلایل ذکر شده در قسمت های قبل امکان بررسی کل جامعه به منظور محاسبه میانگین و واریانس آن وجود ندارد. بر همین اساس سعی می شود با استفاده از نمونه برآوردی برای آنها محاسبه شود. با توجه به اینکه میانگین جامعه یک شاخص مرکزی است، میانگین نمونه یا میانه نمونه می توانند آماره هایی باشند که از نمونه برای برآورد میانگین جامعه معرفی می شوند. حال سوالی که مطرح می شود این است که کدامیک بهتر هستند و کمتر خطا دارند. برای پاسخ به این سوال برخی از ویژگی های توزیع میانگین نمونه و میانه نمونه را بررسی می کنیم.
در مورد میانگین نمونه داریم:

در مورد میانه نمونه نیز داریم:

و اثبات می شود همواره

عبارت فوق به این معنی است که یک آماره بد میانه ممکن است در فاصله دورتری از میانگین جامعه قرار گیرد تا یک آماره بد میانگین نمونه.

در نتیجه در صورتی که تعریف کنیم:

اگر دو برآوردگر نااریب برای برآورد پارامتر جامعه داشته باشیم برآوردگری بهتر است که واریانس کوچکتری داشته باشد.

بنابراین کارایی نسبی دو برآوردگر نااریب را می توان به صورت زیر تعریف کرد.

در مورد پارامتر واریانس جامعه ، یک برآوردگر مناسب می تواند واریانس نمونه باشد که به صورت زیر تعریف می شود.

در مورد واریانس نمونه در صورتی که جامعه محدود باشد. داریم:

همانطور که ملاحظه می شود این برآوردگر نااریب نیست. البته بدیهی است هرگاه اندازه جامعه آماری بزرگ یا نامحدود شود، مقدار به سمت یک میل می کند . در نتیجه خواهد شد.

با این مقدمه می توان گفت در بعضی از موارد یک برآوردگر ممکن است مقدار کمی اریب باشد ولی واریانس خیلی کوچکتری نسبت به برآوردگر دیگری داشته باشد، که نااریب است. در این حالت برآوردگر اول بهتر خواهد بود بر همین اساس معیار دیگری تحت عنوان حداقل میانگین مجذور خطا به صورت زیر تعریف می شود.

که به طور همزمان هم اریبی و هم واریانس برآوردگر را در نظر می گیرد. بنابراین برای دو برآوردگر و (که ممکن است اریب یا نااریب باشند.) داریم:

❖ قضیه حد مرکزی

به نام خدا

فصل دهم: تخمین آماری

هم استنباط ها و پیش بینی ها بر پایه اطلاعات خاصی است که آنها را مشاهده (Observation) یا داده (Data) می نامیم.

هدف دانشمندان آماری آن است که روشهای ریاضی ای ارائه دهند که استنباط درباره جامعه آماری بر پایه اطلاعات حاصل از نمونه هر چه مطلوبتر و بهتر انجام گیرد.

هدف از تحلیل و توصیف در آمار، جامعه آماری است که به دو طریق توصیف می شود یا با استفاده از سرشماری کلیه عناصر جامعه و محاسبه پارامتر که در این صورت فنون آمار توصیفی به کار خواهد رفت و یا با استفاده از نمونه که شامل محاسبه تخمین زننده برای برآورد پارامتر است.

اگر تحقیق از نوع سوالی و صرفا حاوی پرسش درباره پارامتر باشد، برای پاسخ به سوالات از تخمین آماری استفاده می شود و اگر حاوی فرضیه ها بوده و از مرحله سوال گذر کرده باشد، آزمون فرضیه ها و فنون آماری آن به کار می رود.

انواع تخمین

۱- تخمین نقطه ای : در این رویکرد، فقط یک مقدار برای پارامتر مورد نظر جامعه برآورد و ارائه می شود. (برای مثال آماره $\bar{X}$ برای برآورد پارامتر μ).

۲- تخمین فاصله ای: در این رویکرد، پارامترهای مورد نظر جامعه محدوده ای شامل حد پائین (L) و حد با لا (H) تعیین می گردد.

در این فصل محاسبه تخمین فاصله ای برای پارامترهای زیر مدنظر است:

۱) میانگین یک جامعه (μ)

۲) تفاضل میانگین دو جامعه $(\mu_1 - \mu_2)$

۳) نسبت موفقیت در یک جامعه (P)

۴) تفاضل نسبت موفقیت دو جامعه $(P_1 - P_2)$

۵) واریانس یک جامعه (σ^2)

۶) نسبت واریانس دو جامعه $(\frac{\sigma_1^2}{\sigma_2^2})$

❖ تخمین میانگین جامعه (μ)

- بهترین تخمین نقطه ای برای میانگین جامعه آماری (μ)، میانگین نمونه ($\bar{X}$) است.
- در یک توزیع پیوسته، احتمال اینکه $\bar{X}$ با میانگین جامعه مساوی باشد، تقریبا صفر است.

تخمین فاصله ای پارامتر میانگین جامعه (μ)، قاعده ای است که به ما می گوید چگونه دو مقدار را بر پایه داده های نمونه محاسبه کنیم تا $\bar{X}$ در وسط آن قرار گیرد. وقتی یک تخمین فاصله ای برای پارامتر جامعه آماری به کار رود یک جفت عدد از تخمین زننده به دست می آید که آن را تخمین فاصله ای یا فاصله اطمینان برای پارامتر گویند. عدد بزرگی که حد بالای فاصله را می سازد ، حد بالای اطمینان (UCL) و عدد کوچکی که حد پایین فاصله را می سازد ، حد پایین اطمینان (LCL) گفته می شود.
یک تخمین فاصله ای برای میانگین به طور کلی به صورت زیر بیان می شود.

$\bar{X} \pm \varepsilon$

εمقدار ثابتی است که می توان به کمک آنUCL و LCL را تعریف کرد و آن را "دقت برآورد" نامید.

سطح اطمینان $-$۱$=\alpha \leftarrow$ سطح خطا

- سطح اطمینان نسبتی از فاصله های اطمینان است که توسط نمونه هایn تایی(هم حجم) ایجاد شده و در برگیرنده پارامترجامعه باشد .
- یک فاصله اطمینان خوب، فاصله ای است که با کوچک ترین عرض برآورد، در برگیرنده پارامتر باشد.
- تخمین فاصله ای μو یا به عبارتی مقدار ε تحت تاثیر **سطح اطمینان** و **توزیع** $\bar{X}$ است.

مقدار ε به شرایط زیر بستگی دارد: ۱- نوع توزیع جامعه آماری نرمال یا غیر نرمال
۲- کیفیت انحراف معیار جامعه معلوم یا غیر معلوم
۳- اندازه نمونه کوچک یا بزرگ

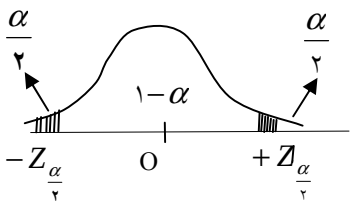
❖ توزیع جامعه آماری نرمال با انحراف معیار معلوم: n هر اندازه باشد

در این حالت می دانیم

$$X \sim N(\mu,\sigma^2) \Rightarrow \bar{X} \sim N(\mu,\frac{\sigma^2}{n}) \Rightarrow Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

و

$$P(-Z_{\frac{\alpha}{2}} \le Z \le Z_{\frac{\alpha}{2}}) = 1-\alpha$$



در نتیجه با ترکیب دو فرمول خواهیم داشت

$$P(-Z_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z_{\frac{\alpha}{2}}) = 1-\alpha \Rightarrow P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

در رابطه فوق $\varepsilon = Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$است. به عبارت دیگر، با احتمال $100(1-\alpha)$ درصد میانگین جامعه در فاصله برآورد شده قرار می گیرد و با 100α در صد خطا ، میانگین جامعه خارج دامنه فوق قرار خواهد گرفت.

مثال ۱: نمونه ای مرکب از ۹ بسته ماکارونی به تصادف از خط تولید انتخاب می شود اگر میانگین وزن ۹ بسته ۲۶۸ گرم باشد. یک فاصله اطمینان ۹۰ درصد برای میانگین وزن بسته ها چقدر است؟ تجربه گذشته حاکی از آن است که توزیع بسته های ماکارونی دارای توزیع نرمال با انحراف معیار ۲۵/۹ هستند.

حل: جامعه نرمال و انحراف معیار جامعه معلوم در نتیجه بر طبق فرمول فوق داریم.

$$268 - Z_{0.05}\frac{25.9}{\sqrt{9}} \le \mu \le 268 + Z_{0.05}\frac{25.9}{\sqrt{9}}$$

$$268 - 1.65\frac{25.9}{\sqrt{9}} \le \mu \le 268 + 1.65\frac{25.9}{\sqrt{9}}$$

$$253.755 \le \mu \le 282.245$$

❖ توزیع جامعه آماری نرمال، انحراف معیار نامعلوم و n کوچک (n<30):

هرگاه انحراف معیار جامعه نامعلوم باشد در تخمین فاصله ای μ ناچار $S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$ جایگزین $\sigma_{\bar{x}}$ خواهد شد. $S_{\bar{x}}$ برآورد نقطه ای انحراف معیار توزیع $\bar{X}$ است . به عبارت دیگر، علاوه بر برآورد نقطه ای μ یعنی $\bar{X}$ باید از آماره ای دیگر به جای $\sigma_{\bar{x}}$یعنی $S_{\bar{x}}$ استفاده کرد.

این عمل موجب خواهدشد که رابطه $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ جایگزین رابطه $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ شود.

بدیهی است رابطه اول که دو آماره دارد، دارای دقتی کمتر از رابطه بعدی است، یعنی ریسک آن بیشتر است. از سوی د یگر، چون از یک جامعه نرمال نمونه گیری شده است می توان متقارن بودن را برای توزیع $\bar{X}$ تصور کرد. پس در این حات توزیع $\bar{X}$ یک توزیع متقارن است که دارای پراگندگی(ریسک) بیشتری به واسطه نامعلوم بودن σ_x است. ثابت می شود آماره $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$دارای توزیع-t استیودنت است که t ، توزیعی قرینه است واز توزیع نرمال دقت کمتری پراگندگی بیشتری دارد.

$$t=\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$$

ضریب چگونگی توزیعt، صفر، ولی ضریب کشیدگی آن منفی است.

گامت، واضع توزیع t، کشف کرد که توزیع t ، شدیدا تحت تاثیر حجم نمونه است. آماره $\bar{X}$ نااریب بوده، مخرج n-1 که در فرمول S_x^2 مشاهده می شود، در جه آزادی (d.f) نامیده می شود که با S_x^2 ارتباط دارد. مبنای اصطلاح درجه آزادی به تعداد انحرافات مستقلی که در توزیع آماره S_x^2 برای تخمین σ_x^2 به کار می روند، اشاره دارد.

در نتیجه در این حالت فاصله اطمینان در سطح $100(1-\alpha)$ اطمینان برای μ به صورت زیر خواهد شد.

$$P(\bar{X}-t_{\frac{\alpha}{2},df}.\frac{S}{\sqrt{n}}\le\mu\le\bar{X}+t_{\frac{\alpha}{2},df}.\frac{S}{\sqrt{n}})=1-\alpha$$

مثال ۲: نمونه ای مرکب از ۹ بسته ماکارونی به تصادف از خط تولید انتخاب می شود اگر میانگین وزن ۹ بسته ۲۶۸ و انحراف معیار آنها ۲۵/۹ گرم باشد. یک فاصله اطمینان ۹۰ درصد برای میانگین وزن بسته ها چقدر است؟ تجربه گذشته حاکی از آن است که توزیع بسته های ماکارونی دارای توزیع نرمال هستند.

حل: جامعه نرمال، انحراف معیار جامعه نامعلوم و حجم نمونه کمتر ۳۰ است. در نتیجه بر طبق فرمول فوق داریم.

$$268 - t_{0.05,8}\frac{25.9}{\sqrt{9}} \le \mu \le 268 + t_{0.05,8}\frac{25.9}{\sqrt{9}}$$
$$268 - 1.86\frac{25.9}{\sqrt{9}} \le \mu \le 268 + 1.86\frac{25.9}{\sqrt{9}}$$
$$251.94 \le \mu \le 284.06$$

همانطور که ملاحظه می شود در این حالت فاصله اطمینان بزرگتری نسبت به حالت قبل(مثال۱) حاصل شده است.

❖ **توزیع جامعه آماری نرمال، انحراف معیار نامعلوم و n بزرگ** ($n \ge 30$):

براساس قضیه حد مرکزی همچنان که حجم نمونه بزرگ می شود، توزیع t استیودنت همچون دیگر توزیعها به سمت توزیع نرمال میل می کند، به طوری که در $n \ge 30$ می توان به جای توزیع t استیودنت از توزیع نرمال برای تخمین μ استفاده کرد.

$$Z \simeq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = t$$

در نتیجه در این حالت فاصله اطمینان در سطح $100(1-\alpha)$ اطمینان برای μ به صورت زیر خواهد شد.

$$P(\bar{X} - Z_{\frac{\alpha}{2}}.\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}}.\frac{S}{\sqrt{n}}) = 1-\alpha$$

مثال ۳: نمونه ای مرکب از ۹۰ بسته ماکارونی به تصادف از خط تولید انتخاب می شود اگر میانگین وزن ۹۰ بسته ۲۶۸ و انحراف معیار آنها ۲۵/۹ گرم باشد. یک فاصله اطمینان ۹۰ درصد برای میانگین وزن بسته ها چقدر است؟ تجربه گذشته حاکی از آن است که توزیع بسته های ماکارونی دارای توزیع نرمال هستند.

حل: جامعه نرمال، انحراف معیار جامعه نامعلوم و حجم نمونه بزرگتر از ۳۰ است. در نتیجه بر طبق فرمول فوق داریم.

$$268 - Z_{0.05}\frac{25.9}{\sqrt{90}} \le \mu \le 268 + Z_{0.05}\frac{25.9}{\sqrt{90}}$$
$$268 - 1.65\frac{25.9}{\sqrt{90}} \le \mu \le 268 + 1.65\frac{25.9}{\sqrt{90}}$$
$$262.9 \le \mu \le 273.08$$

در این حالت افزایش حجم نمونه سبب شده است که فاصله اطمینان به طور قابل ملاحظه ای کوچک شود.

❖ **جامعه غیر نرمال و n کوچک (n<30):**

چنانچه حجم نمونه کوچک **(n<30)** است و جامعه به طور نرمال توزیع نشده است برای تنظیم فاصله اطمینان نمی توان از توزیع نرمال و t استیودنت استفاده کرد، در این حالت از قضیه "چی بی شف" استفاده می شود. براساس این قضیه احتمال قرار گر فتن میانگین نمونه در بین k انحراف استاندارد، σ_x برابراست با :

$$P(|\bar{X}-\mu|\leq k\,\sigma_{\bar{x}})\geq 1-\frac{1}{k^2}$$

در این عبارت، σ_x معلوم تلقی شده است. اگر σ_x معلوم نباشد از S_x استفاده می شود. برای ساختن فاصله اطمینان ابتدا $1-\frac{1}{k^2}$ برابر درجه مطلوب اطمینان $(1-\alpha)$، قرار می گیرد و k به دست می آید.

برای σ_x معلوم داریم:

$$P(\bar{X}-\sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{X}+\sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}})=1-\alpha$$

برای σ_x نامعلوم داریم:

$$P(\bar{X}-\sqrt{\frac{1}{\alpha}}\frac{S}{\sqrt{n}}\leq\mu\leq\bar{X}+\sqrt{\frac{1}{\alpha}}\frac{S}{\sqrt{n}})=1-\alpha$$

❖ **جامعه غیر نرمال و n بزرگ** ($n\geq 30$):

در صورتی که توزیع جامعه غیر نرمال و حجم نمونه بزرگ تر از ۳۰ باشد می توان بر حسب مورد از رابطه های قبل استفاده کرد.

برای σ_x معلوم داریم:

$$P(\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})=1-\alpha$$

برای σ_x نامعلوم داریم:

$$P(\bar{X}-Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\leq\mu\leq\bar{X}+Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}})=1-\alpha$$

❖ تخمین فاصله ای تفاضل میانگین دوجامعه :

می دانیم $\bar{X}_1 - \bar{X}_2$ یک آماره نااریب برای $\mu_1 - \mu_2$ باکمترین واریانس خواهد بود.

➢ اگر $\bar{X}_1$ و $\bar{X}_2$ دو متغیر تصادفی مستقل از هم باشند، میانگین و واریانس $\bar{X}_1 - \bar{X}_2$ به صورت زیر تعریف می شوند.

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

➢ ترکیب خطی دو متغیر نصادفی مستقل نرمال است، اگر توزیع هر یک از آنها نرمال باشد.

در نتیجه تخمین فاصله ای برای $\mu_1 - \mu_2$ را به طور کلی می توان به صورت زیر بیان کرد.

$$\bar{X}_1 - \bar{X}_2 \pm \varepsilon$$

➢ مقدار ε به شرایط زیر بستگی دارد :

۱- نوع توزیع آماری دو جامعه مورد نمونه گیری (نرمال یا غیر نرمال)

۲- کیفیت انحراف معیار دوجامعه مورد نمونه گیری (معلوم یا نامعلوم ، مساوی یا نامساوی)

۳- مقدار درجه آزادی $df = n_1 + n_2 - 2$. (بزرگ یا کوچک)

❖ توزیع دو جامعه آماری نرمال و σ_1 و σ_2 معلوم:

می دانیم در این حالت داریم:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma^2_{\bar{x}_1 - \bar{x}_2}}$$

پس فاصله اطمینان در این حالت را می توان به صورت زیر بیان کرد.

$$P((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1 - \alpha$$

➢ به طور کلی برآورد فاصله ای $\mu_1 - \mu_2$ براساس عرض برآورد تفسیر می شود، به طوری که:

الف) اگر هر دو دامنه مثبت باشد، در سطح اطمینان مورد نظر μ_1 بزرگ تر از μ_2 است

ب) اگر هر دو دامنه منفی باشد در سطح اطمینان مورد نظر μ_1 کوچک تر از μ_2 است

ج) در غیر اینصورت (موارد الف و ب) بین μ_1 و μ_2 اختلاف معناداری دیده نمی شود، چون

نخمین به عمل آمده از نقطه صفر می گذرد.

- هر چه مطلق LCL و UCL در حالتهای الف و ب بیشتر باشد از شدت اختلاف μ_1 و μ_2 ناشی می شود.

❖ **توزیع دو جامعه آمار ی نرمال، $df = n_1 + n_2 - 2 < 30$ و انحرا ف معیار ها نامعلوم و مساوی ($\sigma_1^2 = \sigma_2^2$):**

در این حالت

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$$

که

$$S_{\bar{x}_1 - \bar{x}_2} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ , \qquad S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

پس فاصله اطمینان برای $\mu_1 - \mu_2$ در این حالت را می توان به صورت زیر بیان کرد.

$$P((\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},(n_1+n_2-2)} . S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},(n_1+n_2-2)} . S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) = 1 - \alpha$$

❖ **توزیع دو جامعه آمار ی نرمال، $df = n_1 + n_2 - 2 < 30$ و انحرا ف معیار ها نامعلوم و نامساوی ($\sigma_1^2 \ne \sigma_2^2$):**

در این حالت

$$t' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$$

که

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

پس فاصله اطمینان برای $\mu_1 - \mu_2$ در این حالت را می توان به صورت زیر بیان کرد.

$$P((\bar{X}_1 - \bar{X}_2) - t'_{\frac{\alpha}{2},df'} . \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + t'_{\frac{\alpha}{2},df'} . \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}) = 1 - \alpha$$

$$df' = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2 - 1}}$$

توزیع t' همان توزیع t است برای اینکه بین فرمول این قسمت و قسمت قبل تفاوت قائل شویم تفکیک از t' استفاده می‌کنیم.

❖ **توزیع دو جامعه آمار ی نرمال، $df = n_1 + n_2 - 2 \geq 30$ و انحرا ف معیار ها نامعلوم:**

لازم به ذکر است در این حالت تساوی یا عدم تساوی انجراف معیارها اهمیت ندارد.

هم t و هم t' با افزایش حجم نمونه ها بر اساس قضیه حد مرکزی از تفریب Z برخوردار خواهد بود، بنابراین چنانچه $df = n_1 + n_2 - 2$ دست کم ۳۰ باشد، فاصله اطمینان برای $\mu_1 - \mu_2$ در این حالت را می توان به صورت زیر بیان کرد.

$$P((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}) = 1 - \alpha$$

❖ تخمین نسبت موفقیت در جامعه (p):

نسبت موفقیت یکی از مهمترین پارامترهای توصیف مشاهدات با مقیاس اسمی یا رتبه ای است.

$$p = \frac{X}{N}$$

توزیع نمونه گیری $\bar{p}$ در نمونه های بزرگ از تقریب نرمال برخوردار است و متغیر استاندارد آن به صورت زیر می باشد:

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}$$

فاصله اطمینان برای p را می توان به صورت زیر بیان کرد.

$$P(\bar{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \leq p \leq \bar{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}) = 1 - \alpha$$

مثال۴ : می‌خواهیم نسبت افراد بیکار را با اطمینان ۹۹ درصد برآورد کنیم. می دانیم از یک نمونه ۴۰۰ نفری ۳۰ نفر بیکار هستند.

$n = ٤٠٠$

$x = ٣٠$

$\bar{\rho} = \frac{x}{n} = \frac{٣٠}{٤٠٠} = ٠/٠٧٥$

$\alpha = ٠/٠١ \rightarrow \frac{\alpha}{٢} = ٠/٠٠٥ \rightarrow Z_{\frac{\alpha}{٢}} = Z_{٠/٠٠٥} = ٢/٥٨$ (جدول)

$\Rightarrow ٠/٠٧٥ \pm ٢/٥٨ \times \sqrt{\frac{٠/٠٧٥(١ - ٠/٠٨٥)}{٤٠٠}} = ٠/٠٧٥ \pm ٠/٠٣٤ \Rightarrow (٠/٠٤١$ و $٠/١١)$

یعنی با اطمینان ۹۹ درصد نسبت افراد بیکار بین ۴/۱ ٪ و ۱۱٪ قرار دارد.

❖ تخمین فاصله ای تفاضل نسبت موفقیت در دو جامعه $(\rho_1 - \rho_2)$:

چنانچه داده های جمع آوری شده است از دوجامعه آماری میانگین پذیر باشند (داده هایی که از نوع کمی باشند یا دارای مقیاس نسبی یا فاصله ای باشند داده های میانگین پذیر هستند.) از تخمین فاصله ای $\mu_1 - \mu_2$ برای مقایسه آنها استفاده می شود ولی چنانچه داده ها از نوع کیفی باشند به ناچار، باید از مقایسه نسبت موفقیت دو جامعه آماری استفاده کرد. در این حالت

$$\bar{p}_1 = \frac{X_1}{n_1}, \qquad \bar{p}_2 = \frac{X_2}{n_2}$$

آماره $\bar{p}_1 - \bar{p}_2$ برآورد کننده نااریب از $p_1 - p_2$ با کمترین واریانس خواهد بود. ویژگی های آماره $\bar{p}_1 - \bar{p}_2$ عبارتند از:

$$\mu_{\bar{p}_1 - \bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

$$\sigma^2_{\bar{p}_1 - \bar{p}_2} = \sigma^2_{\bar{p}_1} - \sigma^2_{\bar{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

در این حالت آماره $\bar{p}_1 - \bar{p}_2$ دارای توزیع نرمال است:

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

برای برآورد فاصله ای $p_1 - p_2$ ناچار باید یک آماره نااریب برای $\sigma^2_{\bar{p}_1 - \bar{p}_2}$ تعریف کرد:

$$S_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

فاصله اطمینان برای $p_1 - p_2$ را می توان به صورت زیر بیان کرد.

$$P((\bar{p}_1 - \bar{p}_2) - Z_{\frac{\alpha}{2}} S_{\bar{p}_1 - \bar{p}_2} \le p_1 - p_2 \le (\bar{p}_1 - \bar{p}_2) + Z_{\frac{\alpha}{2}} S_{\bar{p}_1 - \bar{p}_2}) = 1 - \alpha$$

❖ تعیین اندازه نمونه برای تخمین میانگین جامعه (μ):

داده هایی که دارای مقیاس نسبی و فاصله ای هستند از نوع داده های میانگین پذیرند .در این نوع داده ها برای تعیین اندازه نمونه از تخمین فاصله ای میانگین استفاده می شود:

$$\bar{X} \pm \varepsilon \Rightarrow \varepsilon = Z_{\frac{\alpha}{2}} \sigma_{\bar{x}} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

چون تمام عرض حدود اطمینان دو برابر این مقدار است اگر ضریب انحراف استاندارد مقداری ثابت در نظر گرفته شود، تنها راهی که برای کم کردن عرض حدود می ماند، این است که انحراف استاندارد را کاهش دهیم. از این رو، چون انحراف معیار $\bar{X}$ برابر $\frac{\sigma}{\sqrt{n}}$ است و σ عدد ثابتی است، تنها را ه دستیابی به انحراف معیار کوچک تر این است که نمونه بزرگ تری انتخاب شود.

بزرگی اندازه نمونه به مقدار σ ، سطح اطمینان و عرض مطلوب حدود اطمینان بستگی دارد.

هنگامی که نمونه گیری با جایگذاری از یک جامعه محدود و یا نمونه گیری بدون جایگذاری از یک جامعه نامحدود انجام گیرد :

$$n = \frac{Z_{\frac{\alpha}{2}}^{2}\sigma^2}{\varepsilon^2}$$

هرگاه نمونه برداری بدون جایگذاری جایگذاری از یک جامعه محدود انجام شود، اصلاح جمعیت محدود لازم می آید .در نتیجه مقدار ε یا عامل تصحیح $\sqrt{\frac{N-n}{N-1}}$ باید تعریف شود، یعنی:

$$\varepsilon = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

که وقتی برای n حل شود:

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

- روشهای برآورد σ^2:

۱- می توان نمونه مقدماتی را از جامعه آماری انتخاب کرده با استفاه از آن انحراف معیار جامعه را محاسبه کرد و به عنوان برآورد σ^2 به کار برد. مشاهداتی که در نمونه مقدماتی به کار می روند، می توان به صورت بخشی از نموه هایی شمرد به طوری که:

$$n = n_1 - n_2$$

که در آن n، اندازه نمونه محاسبه شده، n_1، اندازه نمونه مقدماتی و n_2، تعداد مشاهدات لازم که برای تکمیل نمونه لازم است.

۲- ممکن است برآوردهای σ^2 از مطالعه مشابه در دسترس باشد.

۳- اگر شواهدی وجود داشته باشدکه جامعه مورد نمونه گیری به طور تقریبی از توزیع نرمال برخوردار است، می توان این حقیقت را به کار برد که عرض دامنه، تقریبا

معادل ۴ یا ۶ انحراف معیار است و $\sigma \simeq \frac{R}{4}$ یا $\sigma \simeq \frac{R}{6}$ خواهد بود. R دامنه تغییرات است که محاسبه آن مستلزم در دسترس بودن کوچکترین و بزرگ ترین مقدار متغیر در جامعه آماری است.

نامعلوم بودن انحراف معیار جامعه موجب می شود که اگر $n \leq 30$ باشد، توزیع $\bar{X}$ یک توزیع t استیودنت شود.

پس اگر $\bar{X}$ دارای توزیع t باشد باید n براساس رابطه زیر :

$$n = \left(\frac{t_{\frac{\alpha}{2},df}\,\sigma_0}{\varepsilon} \right)^2$$

تعیین شود که درآن σ_0 برآوردی از انحراف معیار واقعی جامعه آماری است.

از آنجا که مجهول n تحت تاثیر درجه آزادی (df=n-1) است پس فرمول محاسبه n در یک دور قرار می گیرد که با استفاده از رابطه توزیع Z و t باید به این دور پایان داد.

نظر به اینکه رابطه $t_{\frac{\alpha}{2},df} \geq Z_{\frac{\alpha}{2}}$ برقرار است، یک تخمین مقدماتی برای n عبارت است از :

$$n \geq \left(\frac{Z_{\frac{\alpha}{2}}\sigma_0}{\varepsilon} \right)^2$$

به ازای مقادیر بزرگ n (مثلا $n \geq 50$)، $t_{\frac{\alpha}{2},df} \simeq Z_{\frac{\alpha}{2}}$ است.

در تمام روابط فوق فرض نرمال بودن جامعه آماری یک فرض اساسی است.

در صورتی که نرمال بودن توزیع $\bar{X}$ و یا برخورداری آن از توزیع t استیودنت برما معلوم نباشد، می توان به کمک قضیه چی بی شف به راه حلی محافظه کارانه دست یافت، داریم :

$$P\left(|\bar{X} - \mu| \leq d\right) = 1 - \frac{\sigma_{\bar{x}}^2}{\varepsilon^2} = 1 - \frac{\sigma_x^2}{n\varepsilon^2}$$

$(1 - \frac{\sigma_x^2}{n\varepsilon^2})$ را با سطح اطمینان $1-\alpha$ مساوی قرار می دهیم و حجم نمونه به این صورت به دست می آید:

$$1 - \frac{\sigma_x^2}{n\varepsilon^2} = 1 - \alpha \Rightarrow \sigma_x^2 = \alpha n \varepsilon^2 \Rightarrow n = \frac{\sigma_x^2}{\alpha \varepsilon^2}$$

با فرض اینکه نمونه گیری از جامعه بطور تصادفی صورت می گیرد و توزیع $\bar{p}$ از توزیع نرمال برخوردار است .از فرمولهای فوق در صورتی استفاده می شود که نمونه گیری با جایگذاری از جامعه محدود صورت گیرد و یا جامعه مورد نظر به حد کافی بزرگ باشد که بتوان استفاده از اصلاح جامعه محدود را غیر لازم شمرد.

در صورتی که $\frac{n}{N} \leq 0.05$ نسبت به بزرگ باشد اصلاح جامعه محدود را می توان نادیده گرفت.

$$\varepsilon = Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-P)}{n}} \Rightarrow n = \frac{Z_{\frac{\alpha}{2}} p(1-p)}{\varepsilon^2}$$

اگر اصلاح جامعه محدود نایدده گرفته نشود.

$$n = \frac{N Z^2_{\frac{\alpha}{2}} p(1-p)}{\varepsilon^2 (N-1) + Z^2_{\frac{\alpha}{2}} p(1-p)}$$

❖ تخمین فاصله ای واریانس یک جامعه آماری (σ_x^2)

S_x^2 واریانس نمونه و برآورد نقطه ای واریانس جامعه است.

موفقیت در تعیین حدود اطمینان برای واریانس جامعه به میزان توانایی ما در دسترسی به یک توزیع نمونه گیری مناسب بستگی خواهد داشت. معمولاً حدود اطمینان برای σ^2 بر مبنای توزیع نمونه‌گیری زیر قرار دارد که آماره مزبور از توزیعی موسوم به «کای – مربع» با درجه آزادی (n – 1) برخوردار است .

$$\chi^2 = \frac{(n-1)S_x^2}{\sigma_x^2}$$

کای مربع(Chi-Square) مقدار منفی ندارد و نامتقارن است (چوله به راست) .در صورتی که توزیع جامعه نرمال باشد، فاصله اطمینان برای σ_x^2 را می توان به صورت زیر بیان کرد.

$$P\left(\chi^2_{1-\frac{\alpha}{2},df} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},df}\right) = 1-\alpha \Rightarrow P\left(\frac{(n-1)S_x^2}{\chi^2_{\frac{\alpha}{2},df}} \leq \sigma_x^2 \leq \frac{(n-1)S_x^2}{\chi^2_{1-\frac{\alpha}{2},df}}\right) = 1-\alpha$$

مثال ۵ : می‌خواهیم ریسک سود سالانه شرکت‌های صنعت غذایی را برآورد کنیم. بر اساس نمونه‌ای از ۱۵ شرکت، واریانس ۱۲۲۵ محاسبه شده است. توزیع سود سالانه از توزیع نرمال تبعیت می‌کند. با اطمینان ۹۵ درصد، ریسک سود سالانه را برآورد کنید.

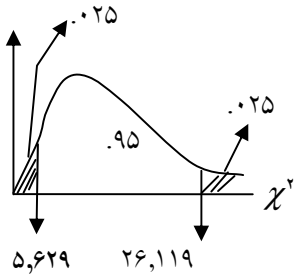
$n=۱۵\ ,\ S^۲=۱۲۲۵$

$$\begin{cases} \frac{\alpha}{۲}=.۰۲۵ \\ df=n-۱=۱۴ \\ ۱-\frac{\alpha}{۲}=۰۹۷۵ \end{cases} \xrightarrow{\text{جدول}} \begin{cases} \chi^۲_{(.۰۲۵\ ,\ ۱۴)}=۲۶.۱۱۹ \\ \chi^۲_{(.۹۷۵\ ,\ ۱۴)}=۵.۶۲۹ \end{cases}$$

$$\Rightarrow \frac{(۱۴)(۱۲۲۵)}{۲۶.۱۱۹} \le \sigma^۲ \le \frac{(۱۴)(۱۲۲۵)}{۵.۶۲۹} \Rightarrow ۶۵۶.۶ \ \le \sigma^۲ \le ۳۰۴۶.۷$$

و انحراف معیار جامعه با اطمینان ۹۵٪ :

$$\Rightarrow ۲۵.۶ \le \sigma \le ۵۵.۲$$



❖ تخمین فاصله ای نسبت واریانس دو جامعه آماری ($\frac{\sigma_1^2}{\sigma_2^2}$):

اگر دو واریانس با هم مساوی باشند، نسبت آنها مساوی یک خواهد بود.

واریانسهای دو جامعه معمولا معلوم نیستند، در نتیجه هرگونه مقایسه ای بر مبنای واریانسهای نمونه قرار می گیرد. برای مقایسه واریانس‌ها بر اساس نمونه‌گیری، از توزیعی به نام «F» استفاده می‌کنیم. این توزیع نیز چوله به راست بوده و به دو درجه آزادی بستگی دارد: یکی به مقدار (n_1-1) که در محاسبه S_1^2 و دیگری به مقدار (n_2-1) که در محاسبه S_2^2 به کار می رود معمولا این دو به ترتیب درجه آزادی صورت و به درجه آزادی مخرج معروفند. در صورت نرمال بودن توزیع دو جامعه، حدود اطمینان برای نسبت واریانس دو جامعه با اطمینان $(۱-\alpha)$ به صورت زیر می توان محاسبه کرد.

آماره Fبه صورت زیر تعریف می شود.

$$F=\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}=\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$

بر اساس توزیع F می دانیم.

$$P\left(F_{(1-\frac{\alpha}{2}),df_1,df_2} \le F \le F_{(\frac{\alpha}{2}),df_1,df_2}\right)=1-\alpha$$

با فرض اینکه $df_1 = n_1 - 1$ و $df_2 = n_2 - 1$ باشند.داریم:

$$P\left(\frac{\frac{S_1^2}{S_2^2}}{F_{(\frac{\alpha}{2}),df_1,df_2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\frac{S_1^2}{S_2^2}}{F_{(1-\frac{\alpha}{2}),df_1,df_2}}\right)=1-\alpha$$

$$F_{(1-\frac{\alpha}{2}),df_1,df_2} = \frac{1}{F_{(\frac{\alpha}{2}),df_2,df_1}}$$

$$P\left(\frac{\frac{S_1^2}{S_2^2}}{F_{(\frac{\alpha}{2}),df_1,df_2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{(\frac{\alpha}{2}),df_2,df_1}\right)=1-\alpha$$

مثال ۶: هدف تحقیق مقایسه پراکندگی نمره های دانشجویان دو دانشکده **"الف و ب"** است. از دانشکده الف یک نمونه ۱۶ تایی انتخاب شده که میانگین آن ۱۴ و واریانس ان ۱۶ است در حالی که میانگین و واریانس یک نمونه تصادفی ۱۰ تایی از دانشجویان ب به ترتیب ۱۵ و ۱۲ بوده است. در سطح اطمینان ۹۰ درصد، پراکندگی نمره ها را در دو دانشکده مقایسه کنید.(توزیع نمرات دو دانشکده را نرمال فرض کنید).

$n_۱ = ۱۶$ و $S_۱^۲ = ۱۶$

$n_۲ = ۱۰$ و $S_۲^۲ = ۱۲$

$$\begin{cases} df_۱ = n_۱ - ۱ = ۱۵ \\ df_۲ = n_۲ - ۱ = ۹ \\ \alpha = .۱۰ \rightarrow \frac{\alpha}{۲} = .۰۵ \end{cases} \Rightarrow \begin{cases} F_{(.۰۵,۱۵,۹)} = ۳.۰۱ \\ F_{(.۰۵,۹,۱۵)} = ۲.۵۹ \end{cases}$$

بنابراین با اطمینان ۹۵ درصد :

$$\Rightarrow \frac{\frac{۱۶}{۱۲}}{۳.۰۱} \le \frac{\sigma_۱^۲}{\sigma_۲^۲} \le \frac{۱۶}{۱۲} \times ۲.۵۹$$

$$\Rightarrow ۰.۴۴ \le \frac{\sigma_۱^۲}{\sigma_۲^۲} \le ۳.۴۵$$

با توجه به اینکه بازه عدد یک را شامل می‌شود نمی‌توان گفت که پراکندگی (واریانس) نمرات دو دانشکده الف و ب اختلاف معنی‌داری با هم دارند.

- تحلیل فاصله اطمینان نسبت واریانس :

الف) اگر هر دو دامنه UCL و LCL ، بزرگتر از یک باشد در سطح اطمینان مورد نظر می توان گفت σ_1^2 بزرگ تر از σ_2^2 است.

ب) اگر UCL و LCL هردو کوچکتر از یک باشد در سطح اطمینان مورد نظر می توان گفت σ_1^2 کوچک تراز σ_2^2 است.

ج) در حالتی غیر از موارد الف و ب یعنی در حالتی که فاصله به دست آمده در برگزیده یک است نمی توان ادعا کرد که تفاوت معناداری بین واریانس دو جامعه وجود دارد .

- توزیع t استیودنت یک توزیع قرینه است که بر حسب α و $df = n-1$ تعریف می شود، این توزیع کوتاهتر از توزیع نرمال است. توزیع کای – مربع یک توزیع نامتقارن است که برای تخمین فاصله ای واریانس جامعه به کار می رود.توزیع F نیز یک توزیع نامتقارن است که چوله به راست است. این توزیع بر حسب α درجه آزادی صورت df_1 و درجه آزادی مخرج df_2 تعریف و برای تخمین فاصله ای نسبت واریانس دو جامعه آماری از آن استفاده می شود.

فصل یازدهم: آزمون فرض آماری

- فرضیه حدسی زیرکانه در خصوص پارامتر جامعه است.
- فنون آماری مناسب برای بررسی صحت یا سقم فرضیه ها، فنون "آزمون فرض آماری" هستند.
- آزمون زمانی استفاده می شود که علاوه بر سوال در تحقیق خود فرضیه نیز داشته باشیم که بطور کلی هدف از آزمون آماری آن است که با توجه به اطلاعات بدست آمده از داده های نمونه، حدس خود را در مورد جامعه به طور قوی رد یا قبول کنیم.
- در واقع هر حکمی درباره جامعه را یک فرض آماری می نامند که قابل قبول بودن آن باید برمبنای اطلاعات حاصل از نمونه گیری از جامعه بررسی شود.
- فرضیه های آماری بر دو نوع فرضیه صفر(H_0) و فرضیه مقابل(H_1) می باشد

H_0 : Null Hypothesis
H_1 :Alternative Hypothesis

بر اساس برهان خلف H_0 بایستی نقیض ادعا(فرضیه پژوهش) باشد و H_1 همان ادعای ما می باشد. اما چنانچه علامت مساوی در H_1 قرار گرفت بایستی جای ادعا و نقیض ادعا عوض شود یعنی ملاک اصلی علامت مساوی(=) می باشد که بایستی حتما در H_0 قرار گیرد.
بنابراین قاعده این است که همواره باید H0 در بر گیرنده علامت های (≤ یا = یا ≥) و H1 در برگیرنده علامت (< یا = یا >) باشد. در زمینه نظریه اخیر میتوان پذیرفت که H_0گاهی بیان کننده ادعا و گاهی بیان کننده نقیض ادعا است.

مثال۱. فرضیه پژوهشی زیر را در نظر گرفته و فرضیه های صفر و مقابل آنرا صورت بندی نمایید.
"نسبت مدیران مشارکت جو در سازمان بیش از ۷۰ درصد می باشد"
ادعا: $p > 0.7$ *چون علامت مساوی ندارد لذا در فرضیه مقابل قرار می گیرد.پس داریم:*

$$\begin{cases} H_0 : p \le 0.7 \\ H_1 : p > 0.7 \end{cases}$$

مثال ۲. فرضیه پژوهشی زیر را در نظر گرفته و فرضیه های صفر و مقابل آنرا صورت بندی نمایید.
"میانگین معدل دانشجویان کلاس دست کم ۱۳ می باشد"
ادعا: $\mu \ge 13$ *چون علامت مساوی دارد لذا در فرضیه صفر قرار می گیرد.پس داریم:*

$$\begin{cases} H_0 : \mu \ge 13 \\ H_1 : \mu < 13 \end{cases}$$

❖ سطح معنی داری (Significant Level)

روش کار این است که فرض H_0 را بنفع H_1 رد کنیم بشرط اینکه از یک آزمون آماری، مقداری بدست آوریم که احتمال وقوع آن مقدار با توجه به H_0 برابر یا کمتر از یک احتمال بسیار کوچک باشد که با α نشان داده می شود . این احتمال وقوع کوچک را سطح معنی داری گویند. مقادیر مرسوم برای α، ۰/۰۱ و ۰/۰۵ است.

از آنجا که مقدار α در تعیین اینکه H_0 باید رد شود یا نه دخالت مستقیم دارد، الزاما رعایت عینیت در تحقیق ایجاب می کند که α را پیش از شروع به جمع آوری داده ها مشخص کنیم.

❖ خطاهای آماری

هنگام اتخاذ تصمیم درباره H_0 ممکن است دو نوع خطا پیش می آید:

- خطای نوع اول(α): احتمال رد کردن H_0 در حالی که H_0 درست است.
- خطای نوع دوم(β): احتمال پذیرفتن H_0 در حالی که H_0غلط است.

	گزینه های صحیح	
نتیجه گیری از نمونه	**H_0درست است**	**H_0غلط است**
H_0پذیرفته می شود	**تصمیم درست است**	β
H_0رد می شود	**α**	**تصمیم درست است**

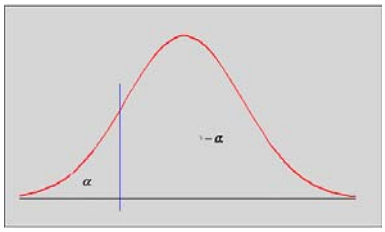
- بین α و β یک رابطه معکوس وجود دارد. با بالا رفتن α مقدار β کاهش می یابد و بالعکس. آنچه مسلم است مجموع α وβ الزاما یک نیست.
- توان یا قدرت آزمون عبارت است از احتمال رد کردن H_0 وقتی که در حقیقت H_0 نادرست باشد، یعنی :

  توان آزمون = ۱− احتمال وقوع خطای نوع دوم = ۱− β
- آنچه باعث کاهش خطای نوع اول و دوم و همچنین موجب افزایش توان آزمون می شود، افزایش حجم نمونه است.
- اگر θ پارامتر جامعه باشد و مقدار عددی آن θ_0 باشد آنگاه انواع فرضیه های آماری به صورت زیر است.

- آزمون فرض یک دنباله چپ

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

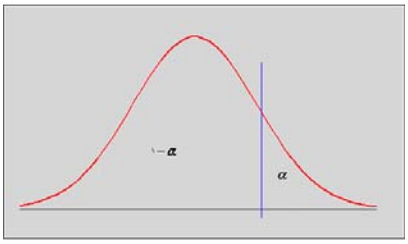
در این حالت ناحیه رد و پذیرش آزمون به صورت زیر خواهد شد.



- آزمون فرض یک دنباله راست

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

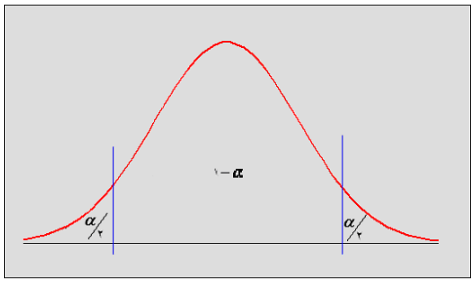
در این حالت ناحیه رد و پذیرش آزمون به صورت زیر خواهد شد.



- آزمون فرض دو دنباله

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

در این حالت ناحیه رد و پذیرش آزمون به صورت زیر خواهد شد.



❖ **مراحل انجام یک آزمون فرض آماری**

برای انجام یک آزمون فرض می بایستی مراحل زیر را به صورت گام به گام طی نمود.

۱) تعیین فرض های صفر (H_0) و فرض مقابل (H_1).

۲) تعیین یک سطح معنی دار α که معمولا یکی از اعداد ۰/۰۱، ۰/۰۵ و ۰/۱ در نظر گرفته می شود.

۳) تعیین **آماره آزمون** که عموما بر اساس وضعیت جامعه آماری مورد بررسی و اطلاعات موجود به دست می آید.

۴) تعیین ناحیه بحرانی آزمون که از روی آماره آزمون، فرض مقابل آزمون و سطح معنی دار α به دست می آید.

۵) محاسبه مقدار آماره آزمون که از روی نمونه های تصادفی $X_1, X_2, ..., X_n$ به دست می آید.

۶) نتیجه گیری: اگر مقدار محاسبه شده آماره آزمون درون ناحیه بحرانی باشد فرض H_0 را رد و در غیر اینصورت فرض H_0 را می پذیریم.

توجه کنید که عموما در حل مسائل آزمون فرض، با در نظر گرفتن انواع آزمون فرض ها محاسبه مراحل ۳و ۴ بسیار ساده تر می شود.

❑ **آزمونهای آماری مورد بررسی در این فصل**

❖ آزمون فرض آماری میانگین یک جامعه

❖ آزمون آماری مقایسه میانگین دو جامعه

❖ آزمون مقایسه زوج ها

❖ آزمون فرض آماری نسبت موفقیت در جامعه

❖ آزمون فرض مقایسه نسبت موفقیت دردو جامعه آماری

❖ آزمون فرض آماری برای واریانس جامعه

❖ آزمون فرض آماری برای مقایسه واریانس دو جامعه

❖ آزمون فرض آماری میانگین یک جامعه

اگر فرضیه ای خصوص میانگین یک جامعه آماری طراحی شود، با استفاده از آزمون فرض آماری میتوان صحت و سقم فرضیه را در سطح معنی داری α تعیین کرد. شرایطی که بررسی انجام شده، نوع آماره آزمون و فرضیه مطرح شده، ناحیه بحرانی راتعیین می کند. در جدول زیر فرضیه های آماری ممکن، شرایطی که ممکن است بررسی انجام شود، آماره آزمون مناسب و ناحیه بحرانی (ناحیه رد فرض صفر) ارائه شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	جامعه نرمال و واریانس جامعه معلوم	$Z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	$Z_0 > Z_{\frac{\alpha}{2}}$ or $Z_0 < -Z_{\frac{\alpha}{2}}$	
$\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$			$Z_0 > Z_\alpha$	
$\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$			$Z_0 < -Z_\alpha$	
$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	جامعه نرمال، واریانس جامعه نامعلوم و $n < 30$	$t_0 = \dfrac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$t_0 > t_{n-1,\frac{\alpha}{2}}$ or $t_0 < -t_{n-1,\frac{\alpha}{2}}$	
$\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$			$t_0 > t_{(n-1,\alpha)}$	
$\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$			$t_0 < -t_{(n-1,\alpha)}$	
$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	جامعه نرمال، واریانس جامعه نامعلوم و $n \geq 30$	$Z_0 = \dfrac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$Z_0 > Z_{\frac{\alpha}{2}}$ or $Z_0 < -Z_{\frac{\alpha}{2}}$	
$\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$			$Z_0 > Z_\alpha$	
$\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$			$Z_0 < -Z_\alpha$	

مثال ۳. از جامعه ی نرمالی نمونه ای تصادفی به اندازه ی ۲۵ انتخاب کرده ایم که بر اساس این نمونه ی تصادفی میانگین نمونه ای برابر ۷۶ بدست آمده است. اگر واریانس جامعه برابر ۱۰ باشد فرض $H_0: \mu \le 75$ را در برابر فرض $H_1: \mu > 75$ در سطح معنی دار ۰/۱ آزمون کنید.

$$\begin{cases} H_0: \mu \le 75 \\ H_1: \mu > 75 \end{cases}$$

$$Z_0 = \frac{76-75}{\frac{10}{\sqrt{25}}} = 0.5,\ Z_\alpha = Z_{0.1} = 1.29$$

در نتیجه $Z_0 < Z_{0.1}$ پس H_0رد نمی شود .

مثال ۴. ادعا شده است که میانگین وزن مردان در جامعه ای برابر با 70 کیلوگرم است. یک نمونه تصادفی به حجم ۲۰ نفر از مردان جامعه انتخاب و مشاهده می شود که میانگین و انحراف معیار وزن این افراد به ترتیب برابر با ۶۴ و ۵ کیلوگرم است. با فرض نرمال بودن وزن مردان در سطح معنی داری ۰/۰۵ این ادعا را بررسی کنید.

$$H_0: \mu = 70$$
$$H_1: \mu \ne 70$$
$$t_0 = \frac{64-70}{5/\sqrt{20}} = -5.4,\ t_{(19,0.025)} = 2.093,\ -t_{(19,0.025)} = -2.093$$

در نتیجه $t_0 < -t_{(19,0.025)}$ بنابراین در سطح معنی داری ۰/۰۵ فرض صفر رد می شود یعنی میانگین وزن مردان درجامعه براب ۷۰ نیست.

❖ آزمون آماری میانگین دو جامعه

بخش اعظم فرضیه های پژوهشی در مدیریت و علوم رفتاری به منظور مقایسه دو جامعه آماری انجام می گیرد. این نوع فرضیه ها را "فرضیه های تطبیقی" گویند. برای آزمون این نوع فرضیه ها (چنانچه میانگین پذیر باشند) و تعیین صحت و سقم آنها را می توان از آزمون فرض آماری برای میانگین دو جامعه استفاده کرد.

به طور کلی فرضیه های آماری میانگین دو جامعه را می توان به صورت زیر بیان کرد.

$$(1)\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \ne \mu_2 \end{cases},\quad (2)\begin{cases} H_0: \mu_1 \le \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases},\quad (3)\begin{cases} H_0: \mu_1 \ge \mu_2 \\ H_1: \mu_1 < \mu_2 \end{cases}$$

در صورتی که در فرضیه های فوق μ_2 به شمت چپ منتقل شود فرضیه های معادلی به صورت زیر حاصل می شود.

$$(1)\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \ne 0 \end{cases},\quad (2)\begin{cases} H_0: \mu_1 - \mu_2 \le 0 \\ H_1: \mu_1 - \mu_2 > 0 \end{cases},\quad (3)\begin{cases} H_0: \mu_1 - \mu_2 \ge 0 \\ H_1: \mu_1 - \mu_2 < 0 \end{cases}$$

در جدول زیر خلاصه ای از فرضیه های آماری مقایسه دو جامعه، شرایط مختلف بررسی، آماره آزمون مناسب و ناحیه رد بیان شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	توضیحات
$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2\neq 0 \end{cases}$	دوجامعه نرمال و واریانس دو جامعه معلوم	$Z_0=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$Z_0>Z_{\frac{\alpha}{2}}$ $or\ Z_0<-Z_{\frac{\alpha}{2}}$	که $\bar{x}_1$، σ_1^2 و n_1 به ترتیب میانگین نمونه، واریانس جامعه و تعداد نمونه مربوط به جامعه اول و $\bar{x}_2$، σ_2^2 و n_2 به ترتیب میانگین نمونه، واریانس جامعه و تعداد نمونه مربوط به جامعه دوم است.
$\begin{cases} H_0: \mu_1-\mu_2\leq 0 \\ H_1: \mu_1-\mu_2> 0 \end{cases}$			$Z_0>Z_{\alpha}$	
$\begin{cases} H_0: \mu_1-\mu_2\geq 0 \\ H_1: \mu_1-\mu_2< 0 \end{cases}$			$Z_0<-Z_{\alpha}$	
$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2\neq 0 \end{cases}$	دو جامعه نرمال، واریانس دو جامعه نامعلوم و برابر و $n_1+n_2-2<30$	$t_0=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$	$t_0>t_{(n_1+n_2-2,\frac{\alpha}{2})}$ $or\ t_0<-t_{(n_1+n_2-2,\frac{\alpha}{2})}$	که $\bar{x}_1$، S_1^2 و n_1 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه اول و $\bar{x}_2$، S_2^2 و n_2 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه دوم است. و $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
$\begin{cases} H_0: \mu_1-\mu_2\leq 0 \\ H_1: \mu_1-\mu_2> 0 \end{cases}$			$t_0>t_{(n_1+n_2-2,\alpha)}$	
$\begin{cases} H_0: \mu_1-\mu_2\geq 0 \\ H_1: \mu_1-\mu_2< 0 \end{cases}$			$t_0<-t_{(n_1+n_2-2,\alpha)}$	
$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2\neq 0 \end{cases}$	دو جامعه نرمال، واریانس دو جامعه نامعلوم و نابرابر و $n_1+n_2-2<30$	$t_0=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$	$t_0>t_{(d.f',\frac{\alpha}{2})}$ $or\ t_0<-t_{(d.f',\frac{\alpha}{2})}$	که $\bar{x}_1$، S_1^2 و n_1 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه اول و $\bar{x}_2$، S_2^2 و n_2 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه دوم است. و $d.f'=\dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(\frac{S_1^2}{n_1})^2}{n_1-1}+\dfrac{(\frac{S_2^2}{n_2})^2}{n_2-1}}$
$\begin{cases} H_0: \mu_1-\mu_2\leq 0 \\ H_1: \mu_1-\mu_2> 0 \end{cases}$			$t_0>t_{(d.f',\alpha)}$	
$\begin{cases} H_0: \mu_1-\mu_2\geq 0 \\ H_1: \mu_1-\mu_2< 0 \end{cases}$			$t_0<-t_{(d.f',\alpha)}$	
$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2\neq 0 \end{cases}$	دو جامعه نرمال، واریانس دو جامعه نامعلوم و $n\geq 30$	$Z_0=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$	$Z_0>Z_{\frac{\alpha}{2}}$ $or\ Z_0<-Z_{\frac{\alpha}{2}}$	که $\bar{x}_1$، S_1^2 و n_1 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه اول و $\bar{x}_2$، S_2^2 و n_2 به ترتیب میانگین نمونه، واریانس نمونه و تعداد نمونه مربوط به جامعه دوم است.
$\begin{cases} H_0: \mu_1-\mu_2\leq 0 \\ H_1: \mu_1-\mu_2> 0 \end{cases}$			$Z_0>Z_{\alpha}$	
$\begin{cases} H_0: \mu_1-\mu_2\geq 0 \\ H_1: \mu_1-\mu_2< 0 \end{cases}$			$Z_0<-Z_{\alpha}$	

- در صورتی که هدف مقایسه اختلاف میانگین های دو جامعه با مقدار مشخصی باشد فرضیه های آماری به ضورت زیر بیان می شود.

$$(1)\begin{cases} H_0 : \mu_1 - \mu_2 = d \\ H_1 : \mu_1 - \mu_2 \neq d \end{cases}, \quad (2)\begin{cases} H_0 : \mu_1 - \mu_2 \leq d \\ H_1 : \mu_1 - \mu_2 > d \end{cases}, \quad (3)\begin{cases} H_0 : \mu_1 - \mu_2 \geq d \\ H_1 : \mu_1 - \mu_2 < d \end{cases}$$

- در صورتی که سمت چپ تساوی ها در فرضیه های آماری برابر صفر باشد. مقدار $(\mu_1 - \mu_2)$ در آماره آزمون همواره صفراست و در صورتی که سمت چپ تساوی ها در فرضیه های آماری برابر d باشد، مقدار $(\mu_1 - \mu_2)$ در آماره آزمون برابر d خواهد شد.

مثال ۵. فرض کنید محققی بخواهد بداند که آیا میانگین هموگلوبین خون مردان و زنان یکی است. ضمنا از تجارب گذشته اطلاع کسب می کند که دو جامعه نرمال و واریانس هر دو جامعه برابر ۴ است. بدین منظور نمونه هایی به حجم n_1، برابر با ۱۲ از جامعه مردان و n_2 برابر با ۸ از جامعه زنان انتخاب می کند و مشاهده می کند که به ترتیب $\bar{x}_1 = 13.5$ و $\bar{x}_2 = 11.5$ ، در سطح معنی داری ۰/۰۵ فرض برابر میانگین هموگلوبین خون زنان و مردان را آزمون کنید.

حل: لازم به ذکر است که آزمون های (۱) و (۲) معادل هستند.

$$(1)\begin{cases} H_0 : \mu_1 \neq \mu_2 \\ H_1 : \mu_1 = \mu_2 \end{cases} \Rightarrow (2)\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \Rightarrow (3)\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

$$Z_0 = \frac{(13.5\text{-}11.5) - 0}{\sqrt{4(\frac{1}{12} + \frac{1}{8})}} = 2.19, \quad Z_{0.025} = 1.96, -Z_{0.025} = -1.96$$

در نتیجه $|Z_0| > Z_{0.025}$ بنابراین در سطح معنی داری ۰/۰۵ فرض صفر رد می شود یعنی میانگین هموگلوبین مردان و زنان برابر نیست.

مثال ۶. از دو جامعه نرمال نمونه های تصادفی و مستقل از یکدیگر با اندازه های ۳۲ اختیار کرده ایم که بر اساس آن ها میانگین های نمونه ای به ترتیب برابر ۰/۱۳۶ و ۰/۰۸۳ بدست آمده است . ادعا شده است که تفاوت میانگین های واقعی بیشتر از ۰/۰۵ است . آزمون فرض مربوط به این ادعا را در سطح خطای ۵٪ درصد انجام دهید. با فرض این که انحراف معیارهای جامعه اصلی به ترتیب برابر ۰/۰۰۴ و ۰/۰۰۵ می باشد .

$$n_1 = n_2 = 32 \qquad \bar{X}_1 = 0.136 \qquad \bar{X}_2 = 0.083$$

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0.05 \\ H_1 : \mu_1 - \mu_2 > 0.05 \end{cases}$$

$$Z_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad Z_0 = \frac{(0.136 - 0.083) - (0.05)}{\sqrt{\frac{(0.004)^2}{32} + \frac{(0.005)^2}{32}}} = 2.65 \qquad Z_{0.05} = 1.64$$

در نتیجه چون $Z_\alpha < Z_0$ پس H_0 رد می شود.

مثال ۷. فرض کنید محققی بخواهد بداند که آیا میانگین فشار خون جامعه مردان و زنان یکی است. بدین منظور نمونه هایی به حجم n_1 برابر با ۱۵ از جامعه مردان و n_2 برابر با ۱۰ از جامعه زنان انتخاب می کند و مشاهده می کند که به ترتیب $\bar{x}_1 = 125$ و $\bar{x}_2 = 140$ میلیمتر جیوه و برآورد واریانس در دو نمونه به ترتیب $S_1^2 = 225$ و $S_2^2 = 400$ است. با فرض برابر بودن واریانس دو جامعه در سطح معنی داری ۰/۰۵، ادعا را بررسی کنید.

$$(1)\begin{cases} H_0 : \mu_1 \neq \mu_2 \\ H_1 : \mu_1 = \mu_2 \end{cases} \Rightarrow (2)\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \Rightarrow (3)\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

$$t_0 = \frac{125 - 140}{\sqrt{\frac{225(15-1) + 400(10-1)}{15+10-2}\left(\frac{1}{15} + \frac{1}{10}\right)}} = -2.14$$

$$t_{(23,0.025)} = 2.069 \ , \ -t_{(23,0.025)} = -2.069$$

در نتیجه در سطح خطای ۰/۰۵ با توجه به اینکه $t_0 < -t_{(23,0.025)}$ شده است، فرض صفر رد می شود. یعنی میانگین فشار خون جامعه مردان و زنان متفاوت است.

❖ آزمون مقایسه زوج‌ها

در آزمون مقایسه میانگین دو جامعه، فرض بر مستقل بودن دو جامعه و نمونه‌های آن دو است. اگر جامعه‌ای در اختیار داشته باشیم و از آن جامعه نمونه‌ای انتخاب و در دو وضعیت مختلف (مثلاً قبل و بعد : Pre – Post) آنها را بررسی کنیم، در این صورت این دو وضعیت مستقل نبوده و نمی‌توان از آزمون مقایسه دو جامعه استفاده کرد. در این حالت، از آزمون مقایسه زوج‌ها استفاده می‌کنیم.

در آزمون مقایسه زوج‌ها (Paires)، برای هر عضو نمونه، دو مشاهده وجود دارد : مشاهده یا رفتار اول را با x و مشاهده دوم را با y نشان می‌دهیم. تفاضل این دو را با d نشان می‌دهیم : $d_i = y_i - x_i$ ، اگر توزیع دو متغیر x و y نرمال باشد، آن گاه توزیع d نیز نرمال خواهد بود. در صورتی که تعریف کنیم، $\bar{d} = \frac{\sum d_i}{n}$ ، $S_d^۲ = \frac{\sum (d_i - \bar{d})^۲}{n-۱}$ و μ_d میانگین d در جامعه آماری.

دربیانی دیگر می توان گفت μ_d، میانگین جامعه آماری اختلاف مشاهده دوم (بعد) از مشاهده اول (قبل) است.

- اگر $\mu_d = 0$، یعنی اختلاف بین اختلاف مشاهده دوم (بعد) از مشاهده اول (قبل) برابرصفر است، یعنی دو مشاهده دوم و اول تفاوتی با هم ندارند.
- اگر $\mu_d > 0$، یعنی اختلاف بین اختلاف مشاهده دوم (بعد) از مشاهده اول (قبل) بزرگترصفر است، یعنی به طور متوسط مشاهده دوم بزرگتر از مشاهده اول است.
- اگر $\mu_d < 0$، یعنی اختلاف بین اختلاف مشاهده دوم (بعد) از مشاهده اول (قبل) کوچکتر از صفر است، یعنی به طور متوسط مشاهده دوم کوچکتر از مشاهده اول است.

در جدول زیر فرضیه های آماری ممکن در مورد μ_d، شرایط استفاده از این آزمون، آماره آزمون و ناحیه رد بیان شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases}$	زمانی که در مساله هدف مقایسه میانگین **یک جامعه** در دو حالت مختلف (به طور مثال قبل و بعد یا مطلوب و نامطلوب) باشد.	$t_0 = \dfrac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$	$t_0 > t_{(n-1,\frac{\alpha}{2})}$ $or\ t_0 < -t_{(n-1,\frac{\alpha}{2})}$	
$\begin{cases} H_0 : \mu_d \leq 0 \\ H_1 : \mu_d > 0 \end{cases}$			$t_0 > t_{\alpha,n-1}$	
$\begin{cases} H_0 : \mu_d \geq 0 \\ H_1 : \mu_d < 0 \end{cases}$			$t_0 < -t_{\alpha,n-1}$	

مثال۸. در مقایسه ای بین نمرات امتحانهای پاییز و بهار دانشجویان ، ۴ دانشجو انتخاب و نمره ی امتحانات بهار و پاییز آنها ثبت شده با فرض نرمال بودن نمره های پاییز و بهار دانشجویان، آیا می توان گفت به طور معنی داری نمرات بهار دانشجویان بهتر از نمرات پاییزآنهاست؟(در سطح معنی داری ۰/۰۵)

دانشجو	A	B	C	D
نمره بهار (بعد)	۶۴	۶۶	۸۹	۷۷
نمره پاییز(قبل)	۵۴	۵۴	۷۰	۶۲
d_i	۱۰	۱۲	۱۹	۱۵

حل: به عنوان یک توافق همواره بعد از قبل و مطلوب از نامطلوب کم می شود.

$$\begin{cases} H_0 : \mu_d \le 0 \\ H_1 : \mu_d > 0 \end{cases}$$

$$t_0 = \frac{14}{\frac{4.79}{2}} = 5.83 \ , \ t_{0.05,3} = 2.35 \Rightarrow t_0 > t_{0.05,3}$$

در سطح معنی داری ۰/۰۵ با توجه به اینکه $t_0 > t_{0.05,3}$ ، فرض صفر رد می شود یعنی نمره بهار دانشجویان بهتر از نمره پاییز آنهاست.

❖ آزمون آماری نسبت جمعیت

بعضی از فرضیه های تحقیق به صورت نسبت یا درصد بیان می شوند. مانند نسبت بیکاران، نسبت محصلین، درصد کالاهای معیوب و ... فرضیه های مربوط به تحقیقات با مقیاس کیفی با استفاده از آزمون نسبت بررسی می شوند. در صورتی که نسبت مورد نظر در جامعه آماری را با p و نسبت مورد نظر در نمونه موردبررسی را با $\bar{p}$ نشان دهیم. در صورتی که n اندازه نمونه و x تعداد افرادی که در نمونه دارای ویژگی مورد نظر هستند باشد، $\bar{p} = \frac{x}{n}$.

در جدول زیر فرضیه های آماری، شرایط استفاده از این آزمون، آماره آزمون و ناحیه بحرانی دیده می شود.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : p = p_0 \\ H_1 : p \ne p_0 \end{cases}$	زمانی که در مساله هدف بررسی فرضی در نسبتی در جامعه است.	$Z_0 = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$Z_0 > Z_{\frac{\alpha}{2}}$ $or\ Z_0 < -Z_{\frac{\alpha}{2}}$	
$\begin{cases} H_0 : p \le p_0 \\ H_1 : p > p_0 \end{cases}$			$Z_0 > Z_\alpha$	
$\begin{cases} H_0 : p \ge p_0 \\ H_1 : p < p_0 \end{cases}$			$Z_0 < -Z_\alpha$	

مثال ۸. محموله ای شامل ۵۰ رایانه است. اگر ۸ رایانه در این محموله معیوب باشد، آیا در سطح خطای ۵ درصد می توان گفت نسبت معیوب ها کمتر از ۲۰ درصد است ؟

حل :

$$\begin{cases} H_0 : p \geq 0.2 \\ H_1 : p < 0.2 \end{cases}, \qquad \bar{p} = \frac{8}{50}$$

$$Z_0 = \frac{\frac{8}{50} - 0.2}{\sqrt{\frac{0.2 \times (1-0.2)}{50}}} = -0.71 \qquad Z_{0.05} = 1.64 \Rightarrow -Z_{0.05} = -1.64$$

در نتیجه ملاحظه می شود که $Z_0 > -Z_{0.05}$ شده است پس H_0 رد نمی شود (یعنی پذیرفته می شود).

مثال۹. از۵۰۰ نوزاد که در یک بیمارستان به دنیا آمده اند ۲۷۰ نفر آنها پسر است. آیا می توان نسبت به دنیا آمدن نوزاد پسر را با دختر یکسان دانست؟ $\alpha = 0.05$

$$(1)\begin{cases} H_0 : p \neq 0.5 \\ H_1 : p = 0.5 \end{cases} \Rightarrow (2)\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

$$\hat{p} = \frac{270}{500} = 0.54$$

$$Z_0 = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 * 0.5}{500}}} = 0.022, \qquad Z_{0.025} = 1.96, \quad -Z_{0.025} = -1.96$$

در نتیجه در سطح معنی داری ۰/۰۵ با توجه به اینکه $Z_0 > -Z_{0.025}$ و $Z_0 < Z_{0.025}$ شده است و در هیچ یک از شرط های رد فرض صفر قرار ندارد پس فرض صفر رد نمی شود به عبارت دیگر نسبت به دنیا آمدن نوزاد پسر را با دختر یکسان است.

❖ آزمون آماری نسبت دوجمعیت

اگر فرضیه ای پژوهشی در خصوص مقایسه دو جامعه آماری با استفاده از داده های کیفی وجود داشته باشد، می توان به مقایسه آن دو جامعه با استفاده از تفاضل نسبت $p_1 - p_2$ که p_1، نسبت موفقیت در جامعه اول و p_2، نسبت موفقیت در جامعه دوم است. ودرصورتی که n_1 اندازه نمونه از جامعه اول و x_1، تعداد افرادی که در نمونه انتخابی از جامعه دوم دارای ویژگی مورد نظر

هستند و n_2 اندازه نمونه از جامعه دوم و x_2، تعداد افرادی که در نمونه انتخابی از جامعه دوم دارای ویژگی مورد نظر هستند، در نظر بگیریم، تعریف می کنیم $\bar{p}_1 = \frac{x_1}{n_1}$ و $\bar{p}_2 = \frac{x_2}{n_2}$.

مانند حالت مقایسه میانگین های دو جامعه، در این حالت نیز فرضیه های آماری میانگین ممکن را می توان به صورت زیر بیان کرد.

$$(1)\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}, \quad (2)\begin{cases} H_0 : p_1 \leq p_2 \\ H_1 : p_1 > p_2 \end{cases}, \quad (3)\begin{cases} H_0 : p_1 \geq p_2 \\ H_1 : p_1 < p_2 \end{cases}$$

در صورتی که در فرضیه های فوق p_2 به شمت چپ منتقل شود فرضیه های معادلی به صورت زیر حاصل می شود.

$$(1)\begin{cases} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \end{cases}, \quad (2)\begin{cases} H_0 : p_1 - p_2 \leq 0 \\ H_1 : p_1 - p_2 > 0 \end{cases}, \quad (3)\begin{cases} H_0 : p_1 - p_2 \geq 0 \\ H_1 : p_1 - p_2 < 0 \end{cases}$$

در جدول زیر فرضیه های آماری، شرایط استفاده از این آزمون، آماره آزمون و ناحیه بحرانی دیده می شود.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \end{cases}$	زمانی که در مساله هدف مقایسه دو نسبت در دو جامعه است.	$Z_0 = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}$	$Z_0 > Z_{\frac{\alpha}{2}}$ $or\ Z_0 < -Z_{\frac{\alpha}{2}}$	
$\begin{cases} H_0 : p_1 - p_2 \leq 0 \\ H_1 : p_1 - p_2 > 0 \end{cases}$			$Z_0 > Z_\alpha$	
$\begin{cases} H_0 : p_1 - p_2 \geq 0 \\ H_1 : p_1 - p_2 < 0 \end{cases}$			$Z_0 < -Z_\alpha$	

مثال ۱۰. کارخانه سازنده مگس کش افشان می خواهد دو فرمول جدید ۱و ۲ را با هم مقایسه کند. دو اتاق با اندازه های برابر در نظر می گیرد و در هر کدام ۱۰۰۰ مگس را رها می کند. در یکی از این اطاق ها از افشان شماره ۱و در اطاق دیکر از افشان شماره ۲ استفاده می کند. در اطاق های ۱و ۲ به ترتیب ۸۲۵و ۷۶۰ مگس از پا درآمدند. آیا می توان ادعا کرد فرمول ۱ بهتر از فرمول ۲ است؟ (در سطح خطای ۰/۰۵)

$$\begin{cases} H_0 : p_1 - p_2 \le 0 \\ H_1 : p_1 - p_2 > 0 \end{cases}$$

$$\bar{p}_1 = \frac{825}{1000}\ ,\ \bar{p}_2 = \frac{760}{1000}$$

$$Z_0 = \frac{(0.825 - 0.760) - 0}{\sqrt{\frac{0.825(1-0.825)}{1000} + \frac{0.760(1-0.760)}{1000}}} = 3.59$$

$$Z_\alpha = Z_{0.05} = 1.65 \ \Rightarrow Z_0 = 3.59 > Z_{0.05} = 1.65$$

در نتیجه در سطح معنی داری ۰/۰۵ با توجه به اینکه $Z_0 > Z_{0.05}$ شده است فرض صفر رد می شود به عبارت دیگر نسبت مگس هایی که توسط فرمول ۱ از پا درآمده اند بیشتر از نسبت مگس هایی که توسط فرمول ۲ از پا درآمده اند در نتیجه فرمول ۱ بهتر از فرمول ۲ است.

❖ آزمون فرض آماری برای واریانس یک جامعه

هر گاه فرضیه ای دربازه پراکندگی جامعه وجود داشته باشد، ضحت وسقم آن را می توان با استفاده از آزمون فرض بررسی کرد.

همان طور که در مورد برآورد فاصله ای گفته شد استنباط در مورد واریانس یک جامعه (σ^2) با استفاده از توزیع کای ۲ با درجه n-1 آزادی خواهد بود که داریم :

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

آزمون فرض در مورد σ^2 نیز بر اساس همین توزیع کای ۲ می باشد هرگاه آن را تحت فرض تنظیم کنیم. بنا براین داریم:

در جدول زیر فرضیه های آماری، شرایط استفاده از این آزمون، آماره آزمون و ناحیه بحرانی دیده می شود.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$	زمانی که در مساله هدف بررسی واریانس یک جامعه است.	$\chi_0^2 = \frac{(n-1)S}{\sigma_0^2}$	$\chi_0^2 > \chi^2_{\frac{\alpha}{2},n-1}$ or $\chi_0^2 < \chi^2_{1-\frac{\alpha}{2},n-1}$	
$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$			$\chi_0^2 > \chi^2_{\alpha,n-1}$	
$\begin{cases} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$			$\chi_0^2 < \chi^2_{1-\alpha,n-1}$	

مثال ۱۱. فرض می کنیم از جامعه نرمالی نمونه تصادفی به اندازه n=18 انتخاب کرده باشیم . بر اساس این نمونه تصادفی انحراف معیار نمونه ای برابر ۰/۶۸ محاسبه شده است. فرض $\sigma^2 > 0.36$ را در برابر فرض $\sigma^2 \leq 0.36$ در سطح معنی دار ۰/۰۵ محاسبه کنید.

حل :

$$\begin{cases} H_0 : \sigma^2 \leq 0.36 \\ H_1 : \sigma^2 > 0.36 \end{cases}$$

$$\chi_0^2 = \frac{(18-1)(0.68)^2}{0.36} = 21.83,\ \chi^2_{\alpha,n-1} = \chi^2_{0.05,17} = 27.58$$

در نتیجه با توجه به اینکه $\chi_0^2 < \chi_\alpha^2$، بنابراین فرض H_0 رد نمی شود.

❖ آزمون فرض آماری برای مقایسه واریانس دو جامعه

زمانی که بخواهیم واریانس دو جامعه آماری را مقایسه کنیم از این آزمون استفاده می کنیم، در این حالت آمار آزمون داراب توزیع F با $n_1 - 1$ و $n_2 - 1$ درجه آزادی است.

در جدول زیر فرضیه های آماری، شرایط استفاده از این آزمون، آماره آزمون و ناحیه بحرانی دیده می شود.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$	زمانی که در مساله هدف مقایسه واریانس دو جامعه است.	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\frac{\alpha}{2},n_1-1,n_2-1}$ or $F_0 < F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$	
$\begin{cases} H_0 : \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{cases}$			$F_0 > F_{\alpha,n_1-1,n_2-1}$	
$\begin{cases} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases}$			$F_0 < F_{1-\alpha,n_1-1,n_2-1}$	

مثال۱۲. فرض می کنیم از دو جامعه نرمال نمونه های تصادفی به اندازه های ۱۳ و ۱۶ انتخاب کرده باشیم که بر اساس آن ها واریانس های نمونه ای به ترتیب برابر ۱۹/۲ و ۳/۵ محاسبه شده اند . در سطح معنی دار ٪۵ فرض برابری واریانس ها را در برابر فرض مخالف آن که واریانس ها مساوی نباشند آزمون کنید .

حل :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{19.2}{3.5} = 5.49$$

$$F_{0.025,(12,15)} = 2.96, \quad F_{0.975,(12,15)} = \frac{1}{F_{0.025,(15,12)}} = \frac{1}{3.28} = 0.30$$

در نتیجه چون $F_0 > F_{0/025,(12,15)}$ حاصل شده است، H_0 رد می شود.

فصل دوازدهم: تحلیل واریانس

❖ مقدمه

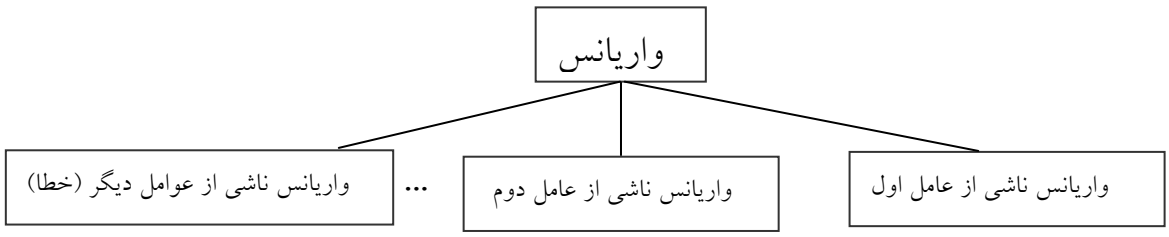
در فصل قبل درباره تفاوت های مشاهده شده بین دو میانگین نمونه ای بحث شد. در این فصل به بررسی و تحلیل تفاوت بین بیش از دو میانگین نمونه ای می پردازیم. به طور مثال ممکن است بخواهیم بدانیم آیا بین میزان محصول ناشی از چهار نوع بذر گندم تفاوت چشم گیری وجود دارد یا خیر. نکته ای که در چنین بررسی هایی وکود دارد این است که زمانی که هدف مقایسه میانگین یک متغیر (میزان محصول) در چند سطح یک متغیر مستقل یا جالت های مختلف یک متغیر است (چهار نوع بذر گندم)، وجود تفاوت بین میانگین ها معادل موثر بودن، متغیر مستقل مورد بررسی است از این رو روشی که ما برای بررسی موثر بودن متغیر استفاده می کنیم "تحلیل واریانس" نامیده می شود.

پایه اصلی تحلیل واریانس بر این اصل استوار است که تفاوت مقدار هر مشاهده از میانگین کل بوسیله عوامل متعددی ایجاد می شود. در مورد مثال چهار نوع بذر اگر ما ۱۲ قطعه زمین داشته باشیم.

بذر 1	بذر 2	بذر 3	بذر 4
x11	x21	x31	x41
x12	x22	x32	x42
x13	x23	x33	x43

مطابق شکل فوق x_{11} ، نشان دهنده میزان محصول اولین زمینی است که بذر نوع اول در آن کاشته شده است، x_{12}، میزان محصول دومین زمینی است که بذر نوع اول در آن کاشته شده است و x_{21} ، میزان محصول اولین زمینی است که بذر نوع ۲ در آن کاشته شده است و ...حال اگر $\bar{x}$ متوسط محصول زمین های مورد بررسی باشد. x_{11} با $\bar{x}$ متفاوت است و دلیل این تفاوت عوامل متعددی مانند نوع بذر، نوع زمین، میزان نور و ... است. این موضوع در مورد تمام قطعات زمین برقرار است، بنابراین در مورد تمام قطعات زمین، قسمتی از این تفاوت به خاطر نوع بذر و بقیه مربوط به عوامل دیگر است. حال اگر تفاوت ناشی از نوع بذر نسبت به تفاوت ناشی از عوامل دیگر بیشتر باشد یعنی عامل نوع بذر مهم است و همانطور که در بالا اشاره شد این بدین معنی است که محصولی که از سه بذر تولید می شود یکی نیست و این به این معنی است که فرض $(\mu_1=\mu_2=\mu_3)$ برقرار نیست. حال در بیان تئوری اگر فرض کنیم داده های x_{ij} برای $i=1,2,...,k$ و $j=1,2,...,n$ بوده، میانگین کل آنها $\bar{x}_{..}$ باشد، در این صورت کل تغییر نسبت به میانگین به صورت $\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{ij}-\bar{x}_{..})^2$ در می آید که مجموع کل توان دوم انحرافات نامیده می

شود. روش تحلیل واریانس این مقدار را به قسمت هایی تجزیه می کند که در شکل زیر سه منبع تغییر قابل شناسایی به علاوه مولفه خطا مشخص شده است.



در حالت کلی داریم:

SST=SS(Tr1)+SS(Tr2)+…+SSE

واریانس ناشی از خطا + ... +واریانس ناشی از عامل دوم +واریانس ناشی از عامل اول= واریانس کل

نکته کلیدی در تحلیل واریانس این است که هر چقدر عاملی تاثیر بیشتری داشته باشد، واریانسی که ایجاد می کند بیشتر خواهد بود.

در هر تحلیل واریانسی چنین چیزی بررسی می شود: واریانسی که عامل ایجاد می کند نسبت به واریانس خطا در جه وضعی قرار دارد، اگر واریانس بزرگ تر باشد عامل مورد نظر موثر تلقی می شود.

$$\begin{cases} H_0 : \sigma^2_{Treatment} \le \sigma^2_{Error} \\ H_1 : \sigma^2_{Treatment} > \sigma^2_{Error} \end{cases}$$

به طور معادل می توان گفت:

H0: عامل مورد نظر موثر نیست.

H1: عامل مورد نظر موثر است.

همانطور که در فصل قبل بیان شد آزمون مناسب برای مقایسه دو واریانس، آزمون فیشر (F) است و به فرضیه آماری بیان شده در بالا آزمون یک دنباله راست است.

چنانچه تحلیل واریانس بر اساس مشاهداتی صورت گیرد که بر مبنای معیار واحدی برای مثال نوع بذر ،طبقه بندی شده اند، آنرا تحلیل واریانس یک عامله می گویند، ولی اگر براساس مشاهداتی صورت گیرد که بر مبنای ۲ معیار نوع بذر ونوع کود طبقه بندی شده اند آنرا تحلیل واریانس دو عامله گویند.

❖ تحلیل واریانس یک عامل

فرض کنید عامل (تیمار) مورد بررسی k سطح داشته باشد و هدف مقایسه میانگین متغیر پاسخ در این k سطح باشد. (در مثال بذر هدف مقایسه میانگین های میزان محصول چهار نوع بذر است.) داده های حاصل از بررسی را می توان به صورت ماتریس زیر بیان کرد.

مشاهدات	(تیمار)

سطوح عامل	۱	۲	...	n
۱	$x_{۱۱}$	$x_{۱۲}$	...	$x_{۱n}$
۲	$x_{۲۱}$	$x_{۲۲}$	...	$x_{۲n}$
⋮	⋮	⋮	...	⋮
k	$x_{k۱}$	$x_{k۲}$	...	x_{kn}

در جدول فوق x_{ij}، نشان دهنده j امین مشاهده، تحت i امین تیمار (سطح) عامل مورد بررسی است. به طور کلی در هر سطح یا تیمار عامل مورد بررسی n مشاهده داریم. در مورد هر مشاهده می توان نوشت:

$$x_{ij} = \mu + \alpha_i + e_{ij}\ , \qquad \begin{cases} i = 1,2,...,k \\ j = 1,2,...,n \end{cases}$$

عبارت فوق بدین معنی است که هر مشاهده برابر میانگین کل به اضافه تغییری که به واسطه تاثیر تیمار iام ایجاد شهد که به آن اثر iامین تیمار گفته می شود. به اضافه تغییری که به واسطه عوامل دیگر (خطا) ایجاد شده است.

حال اگر $\mu_i = \mu + \alpha_i$، داریم:

$$x_{ij} = \mu_i + e_{ij}\ , \qquad \begin{cases} i = 1,2,...,k \\ j = 1,2,...,n \end{cases}$$

مراحل انجام تحلیل واریانس را در چهار مرحله خلاصه می کنیم.

مرحله اول: مشخص نمودن فرضیه آماری

فرض صفری که در تحلیل واریانس بررسی می شود عبارت است از

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

یا به طور معادل

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k$$

یعنی تمام میانگین های جامعه به هم برابرند (و یا همه اثرهای تیماری صفرند.). فرض مقابل در این حالت عبارت است از:

H0=دست کم دو تا از میانگین ها برابر نیستند.

یا

H0=دست کم یکی از اثرهای تیماری مخالف صفر است.

- لازم به ذکر است که در تحلیل واریانس فرض می کنیم متغیرهای تصادفی متناظر (یعنی x_{ij}) که همه مستقل هستند دارای توزیع نرمال با میانگین های μ_i و واریانس مشترک σ^2 هستند.

مرحله دوم:تعیین آماره آزمون

حال به منظور آزمون چنین فرضیه ای تغییر پذیری کل را به دو جز انحرافات حاصل از عامل مورد بررسی و انحرافات باقی مانده تفکیک می کنیم:

مجموع توان دوم باقی مانده + مجموع توان دوم حاصل از تیمار = مجموع کل توان دوم

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(xij-\bar{x}_{..})^2 = n\times\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(xij-\bar{x}_{i.})^2$$

که در آن $\bar{x}_{i.}$ میانگین مشاهدات جامعه iام و $\bar{x}_{..}$ میانگین همه nk مشاهده است. فرمول فوق را به صورت $SST = SS(Tr)+SSE$ نشان می دهیم.

بنابراین داریم:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{..})^2$$

$$SS(Tr) = n\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x}_{..})^2$$

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

در جه آزادی مجموع توان دوم تیمارهای (SS(Tr))، k-1 و درجه آزادی مجموع توان دوم خطا k(n-1) است. در این صورت ، میانگین توان دوم تیمارها و میانگین توان دوم خطا که آنها را به ترتیب با $MS(Tr)$ و MSE نشان می دهیم به این صورت خواهند بود .

$$MS(Tr) = \frac{SS(Tr)}{k-1}$$

$$MSE = \frac{SSE}{k(n-1)}$$

اگر میانگین توان دوم تیمارها نسبت به میانگین توان دوم خطا کم باشد نتیجه می شود که میانگین های جامعه متفاوت هستند در این صورت فرض H_oرد می شود. حال اگر تعریف کنیم:

$$F_0 = \frac{MS(Tr)}{MSE} = \frac{SS(\mathrm{Tr})/(\mathrm{k-1})}{SSE/k(n-1)}$$

F_0 (آماره آزمون) دارای توزیع F ، با درجات آزادی k-1 برای صورت وk(n-1) برای مخرج است.

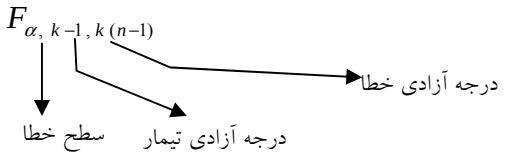
معمول است که تجزیه مجموع توان های دوم و درجات آزادی به همراه میانگین های توان دوم را به شکل جدولی به نام «جدول تحلیل واریانس»یا به صورت ساده تر جدول (ت . و)ارائه می کنند.

منبع پراکندگی	SS	df	MS	F

تیمار	SS(Tr)	$k-۱$	MS (Tr)	$F_0=\frac{MS\ (Tr)}{MSE}$
خطا	SSE	$k\,(n-۱)$	MSE	
جمع	SST	$kn-۱$	-	

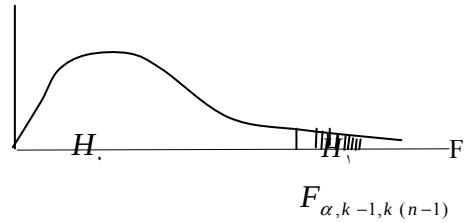
مرحله سوم: تعیین مقدار بحرانی

مقدار بحرانی در این آزمون برابر است با



مرحله چهارم: تصمیم گیری

حال اگر $F_0 \geq F_{\alpha,k-1,k\,(n-1)}$، فرض صفر رد می شود.



برای ساده تر کردن محاسبه مجموعتوان هـای دوم فـوق، معمـولا از فرمـول هـای محاسـباتی زیـر استفاده می شود.

$$SST=\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}^{2}-\frac{1}{kn}.T_{..}^{2}$$

$$SS\ (Tr)=\frac{1}{n}\sum_{i=1}^{k}T_{i.}^{2}-\frac{1}{kn}.T_{..}^{2}$$

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

که در آن $T_{i.}$ مجموع مشاهدات جامعه iام و $T_{..}$ مجموع همه nk مشاهده است.

مثال۱ : می‌خواهیم این فرضیه را آزمون کنیم « میانگین محصول حاصل از چهار نوع بذر یکسان است». برای بررسی این فرضیه، هر نوع بذر را در ۳ قطعه زمین هم مساحت کاشته ایم و میزان محصول حاصل به صورت جدول زیر است. با فرض نرمال بودن توزیع محصول حاصل از بذرها، در سطح خطای ۵ درصد فرض آماری را ؟ آزمون مناسب را برگزار کنید.

نوع بذر	نمونه		
	۱	۲	۳

۱	۸۶	۷۹	۸۱
۲	۸۹	۸۲	۸۸
۳	۸۲	۶۸	۷۳
۴	۸۱	۷۲	۸۰

حل:

مرحله اول: تعریف فرضیه آماری

فرضیه آماری معادل فرضیه فوق عبارت است:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1: \text{دست کم میانگین حاصل از دو نوع بذر یکسان نیست.} \end{cases}$$

مرحله دوم:محاسبه آماره آزمون

برای محاسبه آماره آزمون ابتدا SS ها را محاسبه، سپس جدول ANOVA را تشکیل می‌دهیم :

$$k=4 \quad , \quad n=3 \qquad , \quad \sum\sum x^2_{ij} = 86^2 + ... + 80^2 = 77409$$

جمع مقادیر بذر ۱، ۲، ۳ و ۴ به ترتیب :

$$T_{1.} = 246 \quad , T_{2.} = 259 \ , T_{3.} = 223 \ , T_{4.} = 233$$

جمع مقادیر (۱۲ عدد) :

$$T_{..} = 961$$

$$SST = \Sigma\Sigma x^2_{ij} - \frac{1}{kn}.T_{00}{}^2 = 77409 - \frac{1}{12}(961^2) = 448.917$$

$$SS(Tr) = \frac{1}{n}\sum T_{i.}^2 - \frac{1}{kn}.T_{00}^2 = \frac{1}{3}(246^2 + 259^2 + 223^2 + 233^2) - \frac{1}{12}(961^2) = 244.917$$

$$SSE = SST - SS(Tr) = 448.917 - 244.917 = 204.00$$

یعنی از مجموع ۴۴۸/۹۱۷ واحد توان دوم، ۲۴۴/۹۱۷ واحد را عامل مورد نظر و بقیه (۲۰۴/۰۰) واحد را عوامل دیگر ایجاد کرده است. اکنون جدول را تهیه می‌کنیم

جدول ANOVA

منبع پراکندگی	SS	Df	MS	F
بذر	۲۴۴/۹۱۷	$4-1=3$	۸۱/۶۴	۳/۲۰۲

خطا	۲۰۴/۰۰	$4(3-1)=8$	۲۵/۵۰	-
جمع	۴۴۸/۹۱۷	۱۱	-	-

بنابراین مقدار آماره آزمون (F) عدد ۳/۲۰۲ محاسبه شده است.

مرحله ۳) مقدار بحرانی :داریم $\alpha = 0.05$، $n = 3$ و $k = 4$، پس:

$$F_{0/05\,,\,3\,,\,8} = 4/07$$

مرحله ۴) تصمیم‌گیری : با توجه به اینکه مقدار آماره آزمون ۳/۲۰۲ کمتر از ۴/۰۷ شده است در ناحیه H_0 قرار می‌گیرد دلایل و شواهد کافی بر رد H_0 وجود ندارد. به عبارت دیگر نوع بذر تاثیری در میزان محصول ندارد. (میانگین چهار نوع بذر می‌تواند برابر باشد).

❖ تحلیل واریانس دو عامله

گاهی هدف بررسی دو تاثیر دو عامل است و یا گاهی اوقات عامل غیر مرتبط تاثیر زیادی داشته و این امر سبب بزرگ شده واریانس خط می شود .(چون عوامل غیر مرتبط در داخل خطا قرار می گیرند).

در نتیجه هر چقدر عامل مورد بررسی موثر باشد چون واریانس کوچکتری نسبت به خطا دارد همواره در تحلیل واریانس فرض موثر بودن عامل مورد بررسی رد می شود برای احتراز از چنین وضعی می توان عامل غیر مربوط را ثابت گرفت ولی با این کار به ندرت به اطلاعاتی که لازم داریم می رسیم. امکان دیگر آن است که عامل غیر مربوط را به عنوان یک عامل مهم در نظر گرفته، آزمایش را طوری طراحی می کنیم. که بتوانیم اثر آن را اندازه بگیریم در این صورت تحلیل واریانس دو عامله به کمک ما می آید که در آن تغییرات کل داده ها به سخ جزء افراز می شود: تیمار، عامل غیر مربوط و خطای آزمایش یا تصادف.

همانطور که بیان شد تحلیل واریانس دوعامله به بررسی اثر دو عامل در ایجاد تغییرات می پردازد. مثلا بررسی عامل نوع بذر و نوع خاک در میزان محصول.

مفهوم اساسی که در تحلیل واریانس دو عامله باید معرفی شود تاثیر متقابل است. در آزمایش دو عامله زمانی که دو عامل مستقل نیستند اثر خاص سطوح مختلف تیمار در یک عامل براساس سطوح عامل دیگر تغییر می کند گفته می شود دو عامل تاثیر متقابل دارند. در مثال نوع بذر زمانی که یک نوع بذر در نوع خاک خاصی بهتر رشد می کند ولی در نوع دیگر نه پس تاثیر متقابل داشته و تاثیر متقابل آنها بر متغیر وابسته (میزان محصول) تاثیر گذار است. لازم به ذکر است در بسیاری از آزمایش ها در عامل مستقل بوده و تاثیر متقابل وجود ندارد. در این بخش ما صرفا آزمایش های دو عامله که دو عامل از هم مستقل هستند بررسی می کنیم.

- **تحلیل واریانس دو عامله بدون تاثیر متقابل**

در تحلیل واریانس دو عامله می توانیم دو متغیر (عامل) را تیمارها و بلوک ها- یـا عامـل A و عامـل B – بنامیم و تیمارها را در سطر و بلوک ها را در ستون نشان دهیم.

- فرض می شود، x_{ij} به ازای $i=1,2,...,k$ و $j=1,2,...,n$، مقادیر n متغیر تصادفی مستقا نرمال با میانگین های μ_{ij} و واریانس مشترک σ^2 با شند. ماتریس زیر را در نظر بگیرید.

(تیمار)	سطوح عامل B			
سطوح عامل A	۱	۲	…	n
۱	$x_{۱۱}$	$x_{۱۲}$	…	$x_{۱n}$
۲	$x_{۲۱}$	$x_{۲۲}$	…	$x_{۲n}$
⋮	⋮	⋮	…	⋮
k	$x_{k۱}$	$x_{k۲}$	…	x_{kn}

که x_{ij} نشان دهنده مقدار متغیر (در مثال بذر میزان محصول) در i امین سطح عامل A و j امین سطح بلوک یا عامل B است. هر مشاهده را می توان به این صورت نوشت:

$$x_{ij}=\mu+\propto_i+\beta_j+e_{ij}$$

که μ میانگین کل ، $\propto_i$ اثرهای تیمار A و β_j اثرهای بلوکی (اثر تیماری B) می باشد. دو فرض صفری که آزمون می کنیم عبارت اند از صفر بودن اثرهای تیماری و اثرهای بلوکی

$$H_o:\propto_1=\propto_2=...=\propto_k=0$$
$$H_o^{'}:\beta_1=\beta_2=....=\beta_n=0$$

فرض مقابل H_o آن است که همه اژهای تیماری برابر صفر نیستند و فرض مقابل $H_o^{'}$ آن است که همه اژهای بلوکی صفر نیستند یعنی :

$H_۱=\propto_i\neq ۰, i$ دست کم به ازای یک مقدار

$H_1^{'}=\beta_j\neq 0, j$ دست کم به ازای یک مقدار

می توان به راحتی ثابت کرد که :

مجموع توان دوم + مجموع توان دوم انحرافات + مجموع توان دوم انحرافات = مجموع توان انحرافات

انحرافت باقی مانده حاصل بلوک حاصل تیمار

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{..})^2=n.\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x}_{..})^2+k\sum_{j=1}^{n}(\bar{x}_{.j}-\bar{x}_{..})^2+\sum_{i=1}^{k}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{i.}-\bar{x}_{j.}+\bar{x}_{..})^2$$

که $\bar{x}_{i.}$، میانگین مشاهدات برای تیمار iام، $\bar{x}_{.j}$، میانگین مشاهدات برای بلـوک jام و $\bar{x}_{..}$، میـانگین همه nk مشاهده است.

فرمول به این صورت می شود که

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

حاصل جدول(ت . و) که در قبل اشاره کردیم به این گونه می شود

منبع تغییرات	جمع توانهای دوم	درجه آزادی	میانگین توانهای دوم	F
تیمارها	$SS(Tr)$	$k-۱$	$MS(Tr)=\frac{SS(Tr)}{k-۱}$	$F_{Tr}=\frac{MS(Tr)}{MSE}$
بلوک ها	SSB	$n-۱$	$MSB\frac{SSB}{n-۱}$	$F_B=\frac{MSB}{MSE}$
خطا	SSE	$(n-۱)(k-۱)$	$MSE=\frac{SSE}{(n-۱)(k-۱)}$	
جمع	SST	$nk-۱$		

در این حالت دو F محاسبه می شود (F_{Tr} و F_B). فرض H_0 در صورتی رد می شود که $F_{Tr} \geq F_{\alpha,k-1,(n-1)(k-1)}$ باشد و فرض H_0' را در صورتی رد می کنیم که $F_B \geq F_{\alpha,n-1,(n-1)(k-1)}$ باشد.

طرز محاسبه SSها به این صورت است :

$$SST = \Sigma\Sigma x_{ij}^{۲} - \frac{۱}{kn}.T_{..}^{۲}$$

$$SS(Tr) = \frac{۱}{n}\Sigma T_{i.}^{۲} - \frac{۱}{kn}.T_{..}^{۲}$$

$$SSB = \frac{۱}{k}\Sigma T_{.j}^{۲} - \frac{۱}{kn}.T_{..}^{۲}$$

$$SSE = SST - [SS(Tr) + SSB]$$

مثال۲ : در مثال قبل فرض کنید با وجود اینکه هدف تحقیق بررسی تاثیر نوع بذر بر میزان محصول است، می دانیم نوع خاک نیز بر میزان محصول تاثیر دارد و در منطقه مورد بررسی سه نوع خاک نیز وجود دارد. حال می خواهیم می‌خواهیم این فرضیه ها را آزمون کنیم که « میانگین محصول حاصل از چهار نوع بذر یکسان است» و « میانگین محصول حاصل از سه نوع خاک یکسان است» . برای بررسی این فرضیه ها، هر نوع بذر را در ۳ قطعه زمین که هر یک دارای یک نوع خاک هستند کاشته ایم و میزان محصول حاصل به صورت جدول زیر است. با فرض نرمال بودن توزیع محصول حاصل از بذرها و نوع خاک ها، در سطح خطای ۵ درصد فرض آماری را ؟ آزمون مناسب را برگزار کنید.

نوع بذر	نوع خاک		
	۱	۲	۳

۱	۸۶	۷۹	۸۱
۲	۸۹	۸۲	۸۸
۳	۸۲	۶۸	۷۳
۴	۸۱	۷۲	۸۰

چهار مرحله آزمون فرض‌ها را انجام می‌دهیم :

مرحله اول: تعریف فرض‌ها :

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \text{دست کم میانگین دو نوع بذر برابر نیست} \end{cases}$$

$$\begin{cases} H'_0 : \mu_{k1} = \mu_{k2} = \mu_{k3} \\ H'_1 : \text{دست کم میانگین دو نوع خاک برابر نیست} \end{cases}$$

۲. آماره آزمون

محاسبات اولیه

نوع بذرها (تیمارها)	نوع خاک (بلوکها) نوع ۱	نوع ۲	نوع ۳	جمع، $T_{i.}$	میانگین
نوع اول	۸۶	۷۹	۸۱	۲۴۶	۸۲
نوع دوم	۸۹	۸۲	۸۸	۲۵۹	۸۶/۳۳
نوع سوم	۸۲	۶۸	۷۳	۲۲۳	۷۴/۳۳
نوع چهارم	۷۲	۷۲	۸۰	۲۳۳	۷۷/۶۷
جمع، $T_{.j}$	۳۳۸	۳۰۱	۳۲۲	$T_{00} = 961$	
میانگین					$x_{00} = 80.08$

$$k = 4 , \qquad n = 3$$

$$\sum_{i=1}^{4} \sum_{j=1}^{3} x_{ij}^2 = 86^2 + 89^2 + ... + 80^2 = 77409$$

$$SST = 77409 - \frac{1}{4 \times 3}(961)^2 = 448.917$$

$$SS\ (Tr) = \frac{1}{3}(246^2 + 259^2 + 223^2 + 233^2) - \frac{1}{4 \times 3}(961)^2 = 244.92$$

$$SSB = \frac{1}{4}(338^2 + 301^2 + 322^2) - \frac{1}{4\times 3}(961)^2 = 172.17$$

$$SSE = 448.917 - 244.92 - 172.17 = 31.828$$

جدول ANOVA

منبع تغییرات	مجموع توانهای دوم	درجه آزادی	میانگین توان دوم	F
تیمار	۲۴۴/۹۲	۳	$\frac{244.92}{2}=81.64$	$\frac{81.64}{5.30}=15.40$
بلوکها	۱۷۲/۱۷	۲	$\frac{172.17}{2}=86.085$	$\frac{86.085}{5.30}=16.24$
خطا	۳۱/۸۳	۶	$\frac{31.828}{6}=5.30$	
جمع	۴۴۸/۹۱۷	۱۱		

مرحله سوم: مقادیر بحرانی که داریم $k=4$ و $n=3$ α=۰/۰۵؛ پس :

برای تیمارها : $F_{\alpha,k-1,(n-1)(k-1)} = F_{0/05,3,6} = 4/76$

برای بلوکها : $F_{\alpha,n-1,(n-1)(k-1)} = F_{0/05,2,6} = 5.14$

۴- تصمیم‌گیریها : چون $F_{Tr} = 15/40$ از $F_{0/05,3,6} = 4.76$ بیشتر است و $F_B = 16.24$ از $F_{0/05,2,6} = 5.14$ بیشتر است؛ پس نتیجه می‌گیریم که هر دو فرض صفر باید رد شوند. به عبارت دیگر، اختلافهای بین میانگینهای حاصل برای چهار نوع بذر و سه نوع خاک معنی‌دار است.

آزمون‌های پس از تجربه

در تحلیل واریانس زمانی که فرض صفر(H_0) رد می شود و فریضه مخالف (یعنی H_1) پذیرفته می شود (عامل مورد بررسی معنی دار تشخیص داده می شود) ،یعنی دست کم میانگین دو گروه (جامعه) با هم اختلاف دارند و سوالی که مطرح می شود این است که کدام دو گروه یا جامعه با هم متفاوت هستند برای پاسخ به این سوال در مقایسه های بتعی یا پس از تجزیه استفاده می شود .

تعداد این آزمون‌ها زیادمی باشد. سه آزمون معروف عبارتند از :

- LSD
- HSD توکی (Tukey)
- شِفِه (Scheffe)

برای انجام این منظور روش های متعددی وجود دارد که اینجا ما صرفا روش HSD بررسی می کنیم .

❖ **روش HSD (تفاوت معنی‌دار راستین)**

HSD = Honestly Significant Difference

روش HSD یا تفاوت معنی دار راستین که توکی ارائه کرده مستلزم محاسبه کوچک ترین تفاوت معنی دار بین دو میانگین است .در این روش قدر مطلق تفاضل میانگین های تمام زوج های ممکن $|x_i - x_j|$ محاسبه می شود. در صورتی که برای هر زوج $|x_i - x_j|$ از مقدار HSD بیشتر باشد تفاوت بین میانگین های آن زوج معنی دار خواهد بود یا به عبارتی می توان گفت $\mu_i \neq \mu_j$ است که مقدار HSD زمانی که حجم نمونه های گروه ها با هم برابر باشد. عبارت است از

$$HSD = q_{(\alpha,k,df)} \times \sqrt{\frac{MSE}{n}}$$

و اگر حجم نمونه از گروه‌های مختلف برابر نباشد، از این فرمول استفاده می‌کنیم.

$$HSD = q_{(\alpha,k,df)} \times \sqrt{\frac{MSE}{k}(\frac{۱}{n_۱} + \frac{۱}{n_۲} + ... + \frac{۱}{n_k}}$$

که در فرمول‌های فوق :

q : از روی جدول به دست می آید. (جدول ۹ انتهای کتاب).

α : سطح معنی‌دار بودن

K : تعداد گروه‌ها (جامعه‌های) مورد مقایسه

Df : درجه آزادی خطا

MSE : میانگین توان دوم خطاها (واریانس خطاها) است.

مثال :

با مراجعه به مثال قبل (بررسی عامل‌های بذر و نوع خاک)، فرضیه $H_۰$ یعنی تساوی میانگین ۴ نوع بذر رد شد، نمونه از هر نوع بذر ۳ تایی بود و میانگین‌های نمونه‌ها به صورت زیر بود :

$$\bar{X}_1 = 82,\ \bar{X}_2 = 86.33 \quad , \quad \bar{X}_3 = 74.33 \qquad \bar{X}_4 = 77.67$$

$$n_1 = n_2 = n_3 = n_4 = 3 \quad , \quad MSE = 5.30$$

حال می خواهیم ببینیم اختلاف بین کدام میانگین‌ها معنی‌دار است (سطح خطا ۵ درصد) :

در اینجا k=۴ و $df = 6$ و α=۰/۰۵ بنابراین مقدار جدول q :

$$q_{(.05,4,12)} = 4.896$$

با توجه به این که نمونه‌ها برابر است، از فرمول اول استفاده می‌کنیم. مقدار HSD به این صورت محاسبه می‌شود :

$$HSD = q_{(.05,4,6)} \times \sqrt{\frac{MSE}{n}} = 4.896 \times \sqrt{\frac{5.30}{3}} = 6.51$$

بنابراین اختلاف بین هر دو میانگینی که بیشتر از (5.51) باشد معنی‌دار است. در جدول زیر اختلاف‌ها محاسبه شده و آنهایی که معنی‌دار هستند با (*) مشخص شده است.

		بذر ۱	بذر ۲	بذر ۳	بذر ۴
	میانگین نمونه	۸۲	۸۶/۳۳	۷۴/۳۳	۷۷/۶۷
بذر ۱	۸۲	–	۴/۳۳*	۷/۶۷*	۴/۳۳
بذر ۲	۸۶/۳۳	–	–	۱۲*	۸/۶۶*
بذر ۳	۷۴/۳۳	–	–	–	۳/۳۴
بذر ۴	۷۷/۶۷	–	–	–	–

بنابراین اختلاف بین میانگین محصول بذر اول و بذر سوم، اختلاف بین میانگین محصول بذر دوم و بذر سوم و دست آخر اختلاف بین میانگین محصول بذر دوم و بذر چهارم معنی دار است یعنی :

$$\begin{cases} \mu_1 \neq \mu_3 \\ \mu_2 \neq \mu_3 \\ \mu_2 \neq \mu_4 \end{cases}$$

به نام خدا

فصل ۱۳: رگرسیون و همبستگی

مدیران هر روز تصمیماتی شخصی و حرفه‌ای می‌گیرند که مبتنی بر پیش‌بینی وضع آینده است. در بسیاری از موارد پیش‌بینی آینده بر مبنای گذشته و حال است. در واقع آنان سعی می‌کنند بین دو یا چند متغیر به نحوی ارتباطی منطقی برقرار نمایند تا بتوانند از آن در پیش‌بینی آینده استفاده کنند. مثلاً ممکن است مدیری به رابطهٔ بین میزان پولی که صرف تحقیق و توسعه شرکت می‌شود و سود خالص شرکت علاقه‌مند باشد، و یا به رابطهٔ بین میزان پاداش و تولید، میزان رضایت شغلی و بهره‌وری کارکنان، میزان تبلیغات و فروش، میزان انتظارات از کارکنان و عملکرد و بازدهی آنها، میزان حقوق و کارآیی، میزان درآمد سرانه و فروش محصولات شرکت و غیره علاقه نشان دهد. نحوهٔ ایجاد ارتباط خطی بین دو متغیر و تحلیل آن موضوع این فصل است. به عبارت دیگر، اگر دو متغیر x,y صفات متغیر مورد بررسی باشند همبستگی بین آنها برازش خط رگرسیون، پیش بینی، و آزمون فرض ها روی پارامترهای خط رگرسیون در این فصل مورد بحث قرار می گیرند.

رابطه بین دو متغیر

انواع رابطه بین دو متغیر

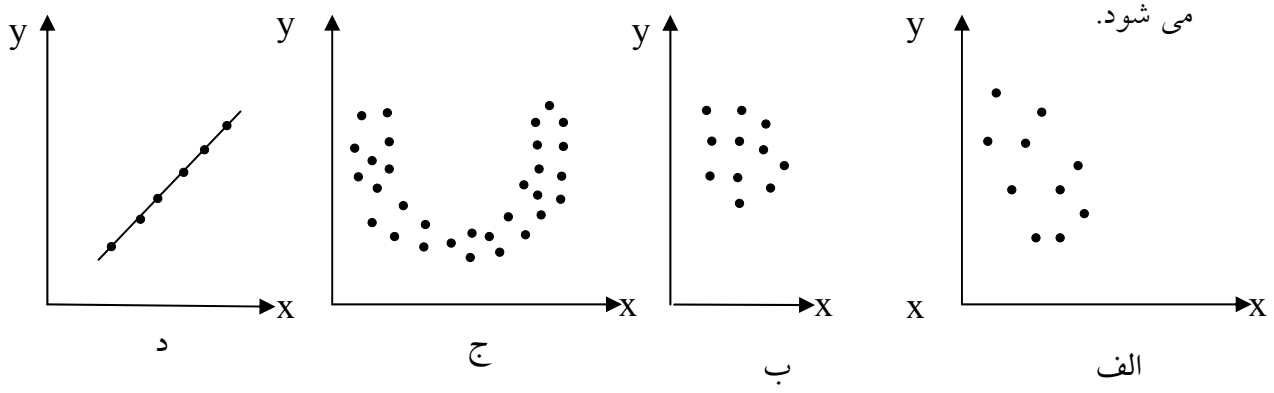
دقیق : مثل رابطه بین شعاع و محیط دایره

غیر دقیق : مثل رابطه بین میزان تبلیغات و فروش شرکت

- رابطه دقیق در ریاضیات و غیردقیق در آمار بحث می شود.

گاهی مشاهده می‌شود که تغییرات یک متغیر بطور مستقیم یا معکوس در تغییرات متغیر دیگر مؤثر است. در این صورت گفته می شود بین دو متغیر رابطه علت و معلولی وجود دارد و به منظور تحلیل رابطه بین این دو متغیر از رگرسیون استفاده می شود. به متغیر علت، متغیر مستقل (X)، و به معلول متغیر وابسته (y) گفته می‌شود.

بین دو متغیر x و y روابط مختلفی ممکن است وجود داشته باشد که در نمودارهای پراکنش زیر نمونه هایی دیده می شود.



نمودار پراکنش (الف) نشان می‌دهد که رابطه بین x و y رابطه‌ای تقریباً خطی و معکوس وجود دارد. نمودار پراکنش (ب) نشان می‌دهد که رابطه‌ای بین x و y وجود ندارد. نمودار (ج) نشان می‌دهد که رابطه‌ای غیرخطی از نوع سهمی بین x و y وجود دارد. و نمودار پراکنش (د) نشان می‌دهد که رابطه‌ای دقیق بین x و y وجود دارد و شیب خط نیز مثبت است.

برآورد رابطه بین دو متغیر، امکان پذیر نخواهد بود مگر آنکه ابتدا فرض کنیم رابطه بین دو متغیر دارای فرم خاصی است. یکی از معمول‌ترین این فرمها،تابع خطی ساده است. یک چنین توابعی در اقتصاد از اهمیت بسیاری برخوردارند، زیرا کار کردن با آنها نسبتاً ساده است و اغلب می‌توانند بعنوان تقریبی از توابع غیرخطی بکار روند. فرم ریاضی یک تابع خطی ساده بصورت زیر است :

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

که در آن مقادیر α و β ثابت هستند. ضریب α که عرض از مبدأ نامیده می‌شود، مقدار Y به ازاء X مساوی صفر را نشان می‌دهد. ضریب β که نمایانگر شیب خط است، میزان تغییرات Y را به ازای یک واحد تغییر در X مشخص می‌کند.

می خواهیم معادله خط را پیدا کنیم. بدیهی است وقتی این معادله مشخص می شود که α و β معلوم باشند، از آنجا که α و β اطلاعات مربوط به جامعه هستند، تنها زمانی بطور دقیق قابل محاسبه هستند که کل جامعه در اختیار باشد، که در اکثر موارد چنین امکانی وجود ندارد در نتیجه، در این جا آن ها را به کمک نمونه ای تصادفی برآورد خواهیم کرد. بنابراین برآورد های حاصل را به صورت زیر نشان می دهیم.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = a + bx \qquad (1)$$

که هدف محاسبه a و b به عناون براوردهای α و β است. روشی که در این مورد انتخاب می کنیم، روش حداقل توان های دوم می باشد.

روش حداقل توان های دوم:

معمولاً مقادیر X را با $x_1, x_2, \ldots, x_n$ و مقادیر Y را با $y_1, y_2, \ldots, y_n$ نشان می دهیم.

پس از انتخاب نمونه می توانیم داده ها را به زیر در نظر گرفت کنیم:

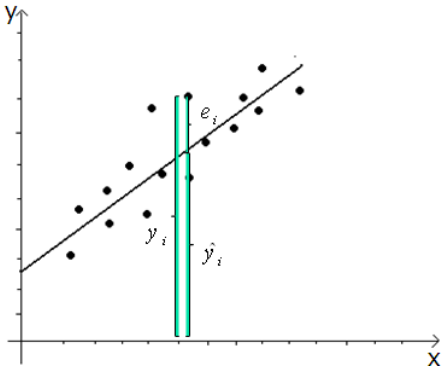
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

همان‌طور که در شکل دیده می شود، تمام نقاط در نمودار پراکنش کاملا بر خط منطبق نیستند و با خط فاصله دارند (این مفهمو رابطه غیردقیق است.) بنابراین در واقعیت داریم:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

که ε_iها خطاهای تصادفی می باشند.

- با فرض اینکه هر $\varepsilon \sim N(0, \sigma^2)$ شکل بعدی رادر نظر می گیریم.



با وتوجه به اینکه ما فقط نمونه را داریم خطای موجود در نمونه را با e_i نمایش می دهیم. مطابق شکل داریم:

$$e_i = y_i - \hat{y}_i$$

مسلما خطی به عنوان بهترین خط شناخته می شود که تا حد ممکن به نقاط نزدیک باشد یا به عبارتی e_i های کوچکی داشته باشد. ازآنجا که مجموع e_iها معمولاً صفرمی باشد (مقادیر مثبت و منفی همدیگر را خنثی می کنند.) بنابراین ازمجموع توان دوم آن $\sum e_i^۲$ استفاده می کنیم. و سعی می کنیم a و b ای را پیدا کنیم که معادله خط حاصل از آنها دارای کمترین مقدار $\sum e_i^۲$ باشد. روش حداقل مربعات عبارت است ازمینیمم کردن رابطه زیر:

$$A(a,b) = \sum e_i^۲ = \sum (y_i - \hat{y}_i)^۲$$

با حداقل کردن عبارت فوق، مقادیر a و b به صورت زیر خواهد شد:

$$a = \bar{y} - b\bar{x}, \qquad b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

به این ترتیب معادله خط برآورد شده عبارت است از:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = a + bx$$

مثال۱) با توجه به داده های زیر معادله خط رگرسیون را محاسبه کنید:

x:	۹	۱۱	۲۱	۱۴	۶	۱۳	۲۷	۱۵
y:	۱۱۵	۶۶	۹۰	۱۳۹	۹۸	۴۹	۹۳	۱۵۹

حل:

$$\sum x_i = 116 \qquad \sum y_i = 809 \qquad \sum x_i y_i = 1339$$
$$\sum x_i^2 = 1998 \quad \sum y_i^2 = 90937$$
$$\bar{x} = \frac{\sum x_i}{n} = \frac{162}{15} = 10.8 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1840.5}{15} = 122.7$$
$$\sum x_i^2 - n\bar{x}^2 = 316 \qquad \sum x_i y_i - n\overline{xy} = 1665.5$$
$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2} = \frac{1665.5}{316} = 5.27$$
$$a = \bar{y} - b\bar{x} = 122.7 - 5.27 \times 10.8 = 24.71$$

در نتیجه معادله خط رگرسیون برآورد شده عبارت است از:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = a + bx = ۲۴/۷۱ + ۵/۲۷x$$

مثال۲) فرض می کنیم بخواهیم رابطه خطی یا معادله رگرسیونی بین دو متغیر Y و Xرا بر اساس نمونه تصادفی بـه اندازه ۱۵ بدست آوریم. بر اساس این نتایج حاصله از نمونه مسئله را حل کنید:

$$\sum x_i = ۱۶۲ \quad \sum y_i = ۱۸۴۰/۵ \quad \sum x_i y_i = ۱۹۹۴۵/۷$$
$$\sum x_i^۲ = ۱۸۲۰/۲ \quad \sum y_i^۲ = ۲۲۵۹۲۷/۸۵$$

حل:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{۱۶۲}{۱۵} = ۱۰/۸ \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{۱۸۴۰/۵}{۱۵} = ۱۲۲/۷$$
$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^۲ - n\bar{x}^۲} = \frac{۱۹۹۴۵/۷ - ۱۵ \times (۱۰/۸ \times ۱۲۲/۷)}{۱۸۲۰/۲ - ۱۵ \times (۱۰/۸)^۲} = \frac{۶۸/۳}{۷۰/۶} = ۰/۹۶$$
$$a = \bar{y} - b\bar{x} = ۱۲۲/۷ - ۰/۹۶ \times ۱۰/۸ = ۱۱۲/۲۵۶$$

در نتیجه معادله خط رگرسیون برآورد شده عبارت است از:

$$\hat{y} = a + bx = ۱۱۲/۲۵۶ + ۰/۹۶x$$

تذکر:به کمک معادله رگرسیون می توانیم به ازای مقدار خاصی از x مانندx^* مقدار y را پیش بینی کنیم:

$$y^* = a + b\ x^*$$

مثلاً در مثال قبل به ازای $x^*=2$ مقدار پیش بینی برابر است با

$$y^* = a + b\ x^* = ۱۱۲/۲۵۶ + ۰/۹۶ \times ۲ = ۱۱۴/۱۷$$

تذکر: گاهی اوقات هدف مقایسه چند خط رگرسیونی است که از نمونه های مختلف به دست آمده اند، معیاری که برای مقایسه این خط ها پیشنهاد می شود، مجموع مانده ها که آن را با نماد S_e نشان می دهیم، عبارت

است از:

$$S_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

خطی که S_e کمتری داشته باشد، بهتر است.

به عنوان نمونه می توانیم مجموع مانده ها در مثال قبل را محاسبه کنیم:

$$S_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$
$$= \sqrt{\frac{225927/85 - 112/256\times 1840/5 - 0/96\times 19945/7}{15-2}} = 3/65$$

مثال۴) برای تعیین رابطه خطی بین دو متغیر تصادفی X و Y نمونه تصادفی به اندازه ۲۰ از جامعه استخراج کرده ایم که بر اساس آن نتایج زیر حاصل شده است.

$$\sum x_i = ۳۵ \quad \sum y_i = ۴۸ \quad \sum x_i y_i = ۹۶۰$$
$$\sum x_i^۲ = ۶۸۰ \quad \sum y_i^۲ = ۱۳۴۸$$

معادله خط رگرسیون و مجموع مربعات مانده ها را بدست آورید.

حل:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{۳۵}{۲۰} = ۱/۷۵ \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{۴۸}{۲۰} = ۲/۴$$

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^۲ - n\bar{x}^۲} = \frac{۸۷۶}{۶۱۸/۷۵} = ۱/۴۱ \quad a = \bar{y} - b\bar{x} = -۰/۰۷$$

در نتیجه معادله خط رگرسیون برآورد شده عبارت است از:

$$\hat{y} = a + bx = -۰/۰۷ + ۱/۴۱x$$

مثال۵) برای تعیین رابطه هزینه حمل یک کالا که آن را با y نشان می دهیم, با فاصله فروشگاه که آن را با x نشان می دهیم, نمونه ای تصادفی شامل ۸ فروشگاه را انتخاب و براساس آن نتایج زیر را بدست آورده ایم:

x:	۸	۱۰	۲۱	۱۵	۶	۱۲	۲۱	۱۴
y:	۱۱۳	۶۸	۹۱	۱۳۹	۹۵	۴۹	۵۶	۱۶۹

معادله خط رگرسیون را مشخص کنید.

حل:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{۱۰۷}{۸} = ۱۳/۳۷ \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{۷۸۰}{۸} = ۹۷/۵$$

$$b=\frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^{۲} - n\overline{x}^{۲}}=\frac{۱۴۶۶/۴}{۱۱۸/۰۴}=۱۲/۴۲ \quad a=\overline{y}-b\overline{x}=-۶۸/۵$$

$$S_e=\sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}=۲/۶۶$$

در نتیجه معادله خط رگرسیون برآورد شده عبارت است از:

$$\hat{y}=a+bx=-۶۸/۵+۱۲/۴۲x$$

خواص برآورد کننده ها

همانطور که بیان شد α و β باید از طریق جامعه محاسبه شوند. به عبارتی پارامتر هستند، بنابراین برای α و β برآورد نقطه ای، برآورد فاصله ای و آزمون فرض را می توان بررسی کرد.

برآورد نقطه ای α و β همانطور که در بالا بحث شد برابرندبا:

$$\hat{\alpha}=a=\overline{y}-b\overline{x}, \qquad \hat{\beta}=b=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}x_i y_i - n\overline{xy}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2}$$

اگر تعریف کنیم:

$$S_a=S_e\sqrt{\frac{۱}{n}+\frac{\overline{X}^{۲}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}}, \qquad S_b=\frac{S_e}{\sqrt{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}}$$

داریم:

$$T=\frac{a-\alpha}{S_a}\sim t_{df=(n-۲)}, \quad T=\frac{b-\beta}{S_b}\sim t_{df=(n-۲)}$$

بنابراین یک فاصله اطمینان $(۱-\alpha)$٪۱۰۰ برای α عبارت است از:

$$\alpha : a \pm t_{\alpha/۲}.S_a$$

به همین ترتیب یک فاصله اطمینان $(۱-\alpha)$٪۱۰۰برای β عبارت است از:

$$\beta : b \pm t_{\alpha/۲}.S_b$$

همچنین مانند درس های گذشته می توانید برای α,β آزمون فرض انجام دهید.

در مورد پارامتر α، فرضیه های آماری، آماره آزمون و ناحیه بحرانی در جدول زیر ارائه شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \alpha = \alpha_0 \\ H_1 : \alpha \neq \alpha_0 \end{cases}$	زمانی که در مساله هدف انجام آزمونی در مورد عرض از مبدا خط رگرسیونی باشد..	$t_{\cdot} = \frac{a - \alpha_{\cdot}}{S_a}$	$t_0 > t_{(n-2,\frac{\alpha}{2})}$ $or\ t_0 < -t_{(n-2,\frac{\alpha}{2})}$	
$\begin{cases} H_0 : \alpha \leq \alpha_0 \\ H_1 : \alpha > \alpha_0 \end{cases}$			$t_0 > t_{\alpha,n-2}$	
$\begin{cases} H_0 : \alpha \geq \alpha_0 \\ H_1 : \alpha < \alpha_0 \end{cases}$			$t_0 < -t_{\alpha,n-2}$	

در مورد پارامتر β ، فرضیه های آماری، آماره آزمون و ناحیه بحرانی در جدول زیر ارائه شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$	زمانی که در مساله هدف انجام آزمونی در مورد شیب خط رگرسیونی باشد..	$t_{\cdot} = \frac{b - \beta_{\cdot}}{S_b}$	$t_0 > t_{(n-2,\frac{\alpha}{2})}$ $or\ t_0 < -t_{(n-2,\frac{\alpha}{2})}$	
$\begin{cases} H_0 : \beta \leq \beta_0 \\ H_1 : \beta > \beta_0 \end{cases}$			$t_0 > t_{\alpha,n-2}$	
$\begin{cases} H_0 : \beta \geq \beta_0 \\ H_1 : \beta < \beta_0 \end{cases}$			$t_0 < -t_{\alpha,n-2}$	

مثال۶) برای جدول مقادیر زیر پارامترها را برآورد کنید. معادله خط رگرسیون را بنویسید.

آزمون فرض زیر را در سطح $\alpha = ۰/۰۵$ انجام دهید.

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

x: ۵۲ ۷۵ ۳۴ ۴۷ ۵۷ ۲۸ ۳۹ ۲۱ ۴۳ ۶۴

y: ۷۵ ۹۸ ۵۶ ۸۹ ۹۲ ۷۳ ۶۵ ۵۲ ۷۸ ۸۲

حل:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{460}{10} = 46, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{760}{10} = 76$$

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{1894}{2474} = 0/76, \quad a = \bar{y} - b\bar{x} = 40/8$$

در نتیجه معادله خط رگرسیون برآورد شده عبارت است از:

$\hat{y} = a + bx = 40.8 + 0.76x$

$S_e = 9.95$

$t_0 = \dfrac{0.76 - 0}{\dfrac{9.95}{49.74}} = 3.79 \ , \qquad t_{0/025, df=8} = 2.306$

بنابراین با توجه به اینکه $t_0 = 3/79 > 2/306 = t_{0025, df=8}$ ، فرض صفر رد می شود. یعنی مقدار β مخالف صفر است.

- زمانی که در معادله خط رگرسیونی ساده، فرض $\beta=0$ در مقابل $\beta\neq0$ آزمون می شود، به عبارتی وجود رابطه خطی بین x و y بررسی می شود. در صورتی که فرض صفر تایید شود یعنی $\beta=0$، به این معنی است که بین دو متغیر x و y رابطه خطی وجود ندارد و در صورتی که فرض مقابل تایید شود یعنی $\beta\neq0$، به این معنی است که بین دو متغیر x و y رابطه خطی وجود دارد.

مثال ۷) جدول زیر تعداد ساعات مطالعه را در برابر نمره دریافتی برای درس زبان برای ۱۰ نفر می باشد. با توجه به این داده ها معادله خط رگرسیون را که در حقیقت رگرسیون نمرات امتحانی روی تعداد ساعات مطالعه را تقریب می کند, بیابید.

X	*Watch*	۴	۹	۱۰	۱۴	۴	۷	۱۲	۲۲	۱	۱۷
Y	*number*	۳۱	۵۸	۶۵	۷۳	۳۷	۴۴	۶۰	۹۱	۲۱	۸۴

سپس آزمون $\beta \leq 3$ را در برابر فرض $\beta > 3$ در سطح ۰/۰۱ انجام دهید. همچنین یک فاصله ی اطمینان ۹۵٪ برای β بسازید.

$\sum x_i = ۱۰۰ \quad \sum y_i = ۵۶۴ \quad \sum x_i^۲ = ۱۳۷۶ \quad \sum x_i y_i = ۶۹۴۵$

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^۲ - n\bar{x}^۲} = \frac{۶۹۴۵ - ۱۰\times۱۰\times۵۶/۴}{۱۳۷۶ - ۱۰\times(۱۰)^۲} = ۳/۴۷$$

$$a = \bar{y} - b\bar{x} = ۵۶/۴ - ۳/۴۷\times۱۰ = ۲۱/۷$$

$$\Rightarrow \qquad \hat{y} = a + bx = ۲۱/۷ + ۳/۴۷x$$

حال می خواهیم آزمون زیر را در سطح ۰/۰۱ بیازماییم:

$$\begin{cases} H_۰ : \beta \leq ۳ \\ H_۱ : \beta > ۳ \end{cases}$$

برای این کار ابتدا به یک آماره نیاز داریم تا آزمون را بر اساس آن بنا کنیم. آماره آزمون عبارت است از:

$$t_. = \frac{b - \beta}{S_b} \sim t_{df=(n-۲)}$$

برای محاسبه آماره آزمون به یک سری محاسبات نیاز داریم:

$$S_e = \sqrt{\frac{36562 - 21/7 \times 564 - 3/47 \times 6945}{8}} = 5/29$$

$$t_0 = \frac{3/47 - 3}{\frac{5/29}{\sqrt{1376 - 10 \times 10^2}}} = \frac{0/47}{\frac{5/29}{\sqrt{376}}} = 1/723 \ , \ \ t_{0/01,8} = 2/89$$

بنابراین با توجه به اینکه $t_0 = 1/723 < 2/89 = t_{0/01,8}$ نتیجه می گیریم که نمی توانیم ادعا کنیم که بطور متوسط یک ساعت مطالعه بیشتر نمره امتحانی را سه نمره افزایش خواهد داد.

حال می خواهیم یک فاصله اطمینان ۹۵٪ برای β بسازیم, برای این کار داریم:

$$\beta : \hat{\beta} \pm t_{\alpha/۲, df=(n-۲)} . S_b = ۳/۴۷ \pm ۲/۳۰۶ \times ۰/۲۷$$

$$\Rightarrow \quad \beta : (۳/۴۷ - ۰/۶۳ \ , \ ۳/۴۷ + ۰/۶۳) = (۲/۸۴ \ , \ ۴/۱)$$

❖ رگرسیون و تحلیل واریانس

همانطور که در قسمت قبل بیان شد یکی از راه های آزمون وجود رابطه خطی بین دو متغیر x و y ، انجام آزمون فرض $\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$ است، یک راه دیگر استفاده از تحلیل واریانس به منظور آزمون وجود رابطه خطی بین x و y است. همانطور که در فصل قبل بیان شد، در تحلیل واریانس، واریانس کل به چند جزء تفکیک شده و در صورتی که واریانسی که بواسطه متغیر مستقل (عامل) ایجاد شده نسبت به واریانس خطا بیشتر باشد. وجود رابطه خطی بین x و y پذیرفته می شود. بنابراین داریم:

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

که در رگرسیون

$$SST = \sum (y - \bar{y})^2 = \sum y^2 - \frac{1}{n} (\sum y)^2$$

$$SS(Tr) = \sum (\hat{y} - \bar{y})^2 = a \sum y + b \sum xy - \frac{1}{n} (\sum y)^2$$

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

در این حالت نیز جدول تحلیل واریانس یا ANOVA به صورت زیر خواهد بود:

منبع پراکندگی	مجموع توان های دوم	درجه آزادی	میانگین توان های دوم	F
تیمار(x)	SS(Tr)	۱	$MS(Tr)=\frac{SS(Tr)}{1}$	$F_0=\frac{MS\ (Tr)}{MSE}$
خطا	SSE	n-2	$MSE=\frac{SSE}{n-2}$	
جمع	SST	n-1	-	

چنانچه مقدار آماره آزمون (F_0)، بزرگتر از مقدار بحرانی ($F_{\alpha,1,n-2}$) باشد فرض صفر در سطح معنی داری α رد می شود و می توان فرض وجود رابطه خطی بین x و y را پذیرفت.

مثال۸) در مثال ۶ در صورتی که بخواهیم مدل (خط رگرسیونی) را با استفاده از تحلیل واریانس آزمون کنیم داریم:

$$\sum y^2=59816,\quad \sum y=760,\quad \sum xy=36854,\quad a=40.8,\quad b=0.76$$

$$SST=\sum y^2-\frac{1}{n}(\sum y)^2=59816-(\frac{1}{10})(760)^2=2056$$

$$SS(Tr)=a\sum y+b\sum xy-\frac{1}{n}(\sum y)^2=40.8\times 760+0.76\times 36854-(\frac{1}{10})(760)^2=1257.04$$

$$SSE=SST-SS(Tr)=2056-1257.04=798.96$$

منبع پراکندگی	مجموع توان های دوم	درجه آزادی	میانگین توان های دوم	F
تیمار(x)	۱۲۵۷/۰٤	۱	۱۲۵۷/۰۴	۱۲/۵۹
خطا	۷۹۸/۹٦	۸	۹۹/۸۷	
جمع	۲۰۵٦	۱۰	-	

با توجه به اینکه مقدار آماره آزمون ($F_0=12.59$)، بزرگتر از مقدار بحرانی ($F_{0.05,1,8}=5.32$) حاصل شده در سطح معنی داری $\alpha=0.05$ فرض صفر رد می شود و می توان فرض وجود رابطه خطی بین x و y را پذیرفت.

- **ضریب همبستگی**

ضریب همبستگی شاخصی است ریاضی که جهت و مقدار رابطه ی خطی بین دو متغیر را توصیف می کند. به عبارتی همبستگی برای تعیین میزان ارتباط دو متغیر استفاده می شود. در صورتی که با افزایش مقدار یکی

دیگری نیز افزایش پیدا کند گویند بین دو متغیر رابطه مستقیم یا مثبت وجود دارد و در صورتی که با افزایش یکی دیگری کاهش یابد گویند بین دو متغیر رابطه معکوس یا منفی وجود دارد. در همبستگی دو معیار بحث می شود: ۱) ضریب تعیین و ۲) ضریب همبستگی

- **ضریب تعیین**

ضریب تعیین معیاری برای بررسی خوبی (نیکویی برازش) معادله رگرسیون است. هر چقدر مقدار آن بیشتر باشد، خطاها کمتر و مدل رگرسیون قابل اعتمادتر است.

ضریب تعیین با علامت r^2 نشان داده شده و به صورت زیر تعریف می شود:

$$r^2 = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$$

همواره $0 \le r^2 \le 1$ است. و مقدار آن نشان دهنده درصدی از تغییرات y است که توسط متغیر x قابل توضیح است.

در صورتی که $r^2 = 0$ باشد. نشان می دهد که خط رگرسیون هرگز نتوانسته تغییرات y را به تغییرات x نسبت دهد.

در صورتی که $r^2 = 1$ باشد. نشان می دهد که خط رگرسیون دقیقا توانسته تغییرات y را به تغییرات x نسبت دهد.

برای محاسبات ساده تر از فرمول زیر برای محاسبه r^2 استفاده می شود:

$$r^2 = \frac{a\sum y + b\sum xy - n\bar{y}}{\sum y^2 - n\bar{y}^2}$$

مثال۹) در مثال ۶ مقدار ضریب تعیین برابر است با:

$$\sum y^2 = 59816, \quad \sum y = 760, \quad \sum xy = 36854, \quad \bar{y} = 76, \quad a = 40.8, \quad b = 0.76$$

$$r^2 = \frac{40.784 \times 760 + 0.766 \times 36854 - 10 \times 76^2}{59816 - 10 \times 76^2} = 0.71$$

مقدار r^2 حاصل به این معنی است که ۷۱٪ تغییرات متغیر y توسط x قابل توجیه است. (در فرمول فوق به منظور دقت در محاسبات مقادیر a و b بدون گرد کردن گذاشته شده است.)

- **ضریب همبستگی پیرسون**

ضریب همبستگی، شدت رابطه و همچنین نوع رابطه (مستقیم یا معکوس) را نشان می دهد. برای محاسبه ضریب همبستگی پیرسون دو روش وجود دارد.

۱) با استفاده از خط رگرسیونی

۲) به صورت کلی

۱) در این حالت اگر از ضریب تعیین، ریشه دوم بگیریم مقدار ضریب همبستگی به دست می آید. ($r = \sqrt{r^2}$) بنابراین همواره $-1 \le r \le 1$ است. علامت ضریب همبستگی (r)، همان علامت شیب خط رگرسیونی (b) است. یعنی اگر شیب خط رگرسیون مثبت باشد، ضریب همبستگی نیز مثبت است و اگر

شیب خط رگرسیونی منفی باشد، ضریب همبستگی نیز منفی است. همچنین اگر شیب خط رگرسیون صفر باشد (b=0)، ضریب همبستگی نیز صفر می شود. (r=0).

مثال ۱۰) در مورد مثال ۶ داریم:

$$r=\sqrt{r^2}=\sqrt{0.71}=0.84$$

۲) به صورت کلی: به طور کلی می توان از فرمول زیر برای محاسبه ضریب همبستگی استفاده کرد:

$$r=\frac{\sum xy-n\bar{x}\bar{y}}{\sqrt{\sum x^2-n\bar{x}}\sqrt{\sum y^2-n\bar{y}^2}}$$

مثال۱۱) در مورد مثال ۶ داریم:

$$r=\frac{36854-10\times46\times76}{\sqrt{23634-10\times46^2}\sqrt{59816-10\times76^2}}=\frac{1894}{2255.3368}=0.84$$

- **آزمون معنی دار بودن r**

همانوطر که ملاحظه شد، ضریب همبستگی محاسبه شده در بالا از روی نمونه به دست آمد، بنابراین آماره است. ضریب همبستگی جامعه را با ρ نشان می دهند که در واقع پارامتر است. و r در واقع برآورد نقطه ای برای پارامتر ρ است. انجام آزمون در مورد ضریب همبستگی جامعه ρ در دو حالت مطرح می شود.

۱) آزمون فرض ρ در مقایسه با صفر

۲) آزمون فرض ρ در مقایسه با مقادیر غیرصفر

۱) در این حالت، در مورد پارامتر ρ، فرضیه های آماری، آماره آزمون و ناحیه بحرانی در جدول زیر ارائه شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases}H_0:\rho=0\\H_1:\rho\neq0\end{cases}$	زمانی که در مساله هدف انجام آزمونی در مورد ضریب همبستگی جامعه در مقایسه با صفر باشد.	$t_0=\frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$	$t_0>t_{(n-2,\frac{\alpha}{2})}$ or $t_0<-t_{(n-2,\frac{\alpha}{2})}$	
$\begin{cases}H_0:\rho\leq0\\H_1:\rho>0\end{cases}$			$t_0>t_{\alpha,n-2}$	
$\begin{cases}H_0:\rho\geq0\\H_1:\rho<0\end{cases}$			$t_0<-t_{\alpha,n-2}$	

مثال ۱۲) در مورد مثال ۶ اگر بخواهیم فرض $\begin{cases}H_0:\rho=0\\H_1:\rho\neq0\end{cases}$ را آزمون کنیم داریم:

$$t_0=\frac{0.84}{\sqrt{\frac{1-0.84^2}{8}}}=4.37,\quad t_{0.025,8}=2.306\ ,\quad -t_{0.025,8}=-2.306$$

با توجه به اینکه $t_0=4.37>2.306=t_{\frac{\alpha}{2}},n-2$ ، فرض صفر رد می شود، یعنی $\rho\neq0$ است.

۳) در این حالت، در مورد پارامتر ρ، فرضیه های آماری، آماره آزمون و ناحیه بحرانی در جدول زیر ارائه شده است.

فرضیه آماری	شرایط	آماره آزمون	رد فرض صفر اگر	ناحیه رد و پذیرش فرض صفر
$\begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho \neq \rho_0 \end{cases}$	زمانی که در مساله هدف انجام آزمونی در مورد ضریب همبستگی جامعه در مقایسه با صفر باشد.	$Z_0 = \dfrac{\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}}{\frac{1}{\sqrt{n-3}}}$	$Z_0 > Z_{\frac{\alpha}{2}}$ $or \quad Z_0 < -Z_{\frac{\alpha}{2}}$	
$\begin{cases} H_0 : \rho \leq \rho_0 \\ H_1 : \rho > \rho_0 \end{cases}$			$Z_0 > Z_\alpha$	
$\begin{cases} H_0 : \rho \geq \rho_0 \\ H_1 : \rho < \rho_0 \end{cases}$			$Z_0 < -Z_\alpha$	

مثال ۱۳)) در مورد مثال ۶ اگر بخواهیم

فرض $\begin{cases} H_0 : \rho = 0.5 \\ H_1 : \rho \neq 0.5 \end{cases}$ را آزمون کنیم داریم:

$$Z_0 = \frac{\frac{1}{2}\ln\frac{1+0.84}{1-0.84} - \frac{1}{2}\ln\frac{1+0.5}{1-0.5}}{\frac{1}{\sqrt{7}}} = 1.777, \quad Z_{\frac{\alpha}{2}} = 1.96\ , \quad -Z_{\frac{\alpha}{2}} = -1.96$$

با توجه به اینکه $Z_0 = 1.77 < 1.96 = Z_{\frac{\alpha}{2}}$ و $Z_0 = 1.77 > -1.96 = -Z_{\frac{\alpha}{2}}$ در نتیجه در ناحیه پذیرش فرض صفر قرار دارد پس فرض صفر رد نمی شود. یعنی $\rho = 0.5$ است.

به نام خدا

فصل چهاردهم: رگرسیون چندگانه و غیرخطی

❖ مقدمه

گاهی اوقات دو یا چند متغیر تاثیر عمده ای روی متغیر وابسته ای دارند. مثلا شما ممکن است معتقد با شید که تعداد فروش، هم به میزان تبلیغات و هم به تعداد فروشندگان بستگی دارد، در فصل قبل، تنها تاثیر خطی یک متغیر مستقل (x) را روی متغیر وابسته (y) بررسی کردیم در این فصل می خواهیم، روشی را ارائه دهیم که بتوانیم تاثیر همزمان و خطی دو یا چند متغیر را روی متغیر وابسته ای اندازه بگیریم.

در رگرسیون خطی ساده یا یک متغیره، معادله خط زیر بررسی می شد:

$$y = \alpha + \beta x + \varepsilon$$

که α و β پارامتر بودند که باید برآورد می شدند که ما با حداقل کردن عبارت:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

برآوردهای a و b را برای پارامترهای α و β معرفی کردیم.

در رگرسیون خطی دو متغیره، مدل مورد بررسی به صورت زیر است:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

که α ، β_1 و β_2 پارامترهایی هستند، که باید برآورد شوند در این حالت نیز با حداقل کردن عبارت:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

این کار را انجام می دهند، که حاصل آن دستگاه معادلات زیر است که با حل آن مقدار a ، b_1 و b_2 به عنوان برآوردهای α ، β_1 و β_2 حاصل می شود:

$$na + b_1 \Sigma x_1 + b_2 \Sigma x_2 = \Sigma y$$

$$a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$

$$a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y$$

در حالت کلی مدل خط رگرسیون، زمانی که k متغیر مستقل در مدل باشد به صورت زیر بیان می شود.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

α ، β_1، β_2، . . . و β_k پارامترهایی هستند که باید برآورد شوند در این حالت نیز با حداقل کردن عبارت:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

این کار را انجام می شود، حاصل آن یک دستگاه معادلات است که حل آن بدون کمک رایانه کار بسیار مشکل و وقت گیری است.

❖ رایانه و رگرسیون (اختیاری)

همانطور که در بخش قبل ملاحظه شد، در رگرسیون دو متغیره، حتی اگر تعداد داده ها زیاد نباشد، حجم زیاد و کار پرزحمتی است. حال اگر تعداد داده ها زیاد یا تعداد متغیرهای مستقل بیشتر باشد حجم محاسبات به طور قابل ملاحظه ای افزایش می یابد.

براین اساس بسته های رایانه ای بسیاری مانند SPSS، Minitab، R و SPLUS وجود دارند که به راحتی می توان بسیاری از مسائل آماری را با آنها تحلیل کرد.

در زیر نمونه ای از خروجی نرم افزار SPSS، در ارتباط با تحلیل رگرسیونی ارائه شده که در این بخش نحوه تحلیل خروجی ها به اختصار شرح داده می شود.

به منظور راحتی مراحل تحلیل را با استفاده ار یک مثال شرح می دهیم.

مثال(۱): یک توزیع کننده نوشابه، سیستم تحویل نوشابه را تحلیل می کند. مشخصا توزیع کننده علاقه مند به پیشگویی مدت زمان لازم برای ارائه سرویس به خرده فروش هاست. مهندس عهده دار این مطالعه دو تا از مهمترین عوامل موثر در مدت زمان تحویل را تعداد جعبه های نوشابه برای تحویل و ماکسیمم طول مسیر می داند. این مهندس نمونه ای از داده های مدت زمان تحویل که آنها را در جدول زیر نشان داده ایم، جمع آوری کرده است. معادله خط رگرسیونی و نتایج حاصل از بررسی این خط را بیان کنید.

Time	Longitude	Nbox
24.00	30.00	10.00
27.00	25.00	15.00
29.00	40.00	10.00
31.00	18.00	20.00
25.00	22.00	25.00
33.00	31.00	18.00
26.00	26.00	12.00
28.00	34.00	14.00
31.00	29.00	16.00
39.00	37.00	22.00
33.00	20.00	24.00
30.00	25.00	17.00
25.00	27.00	13.00
42.00	23.00	30.00
40.00	33.00	24.00

حل: در زیر جداول خروجی حاصل از تحلیل رگرسیونی در نرم افزار SPSS دیده می شود.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.858[a]	.737	.693	3.14079

a. Predictors: (Constant), Nbox, longitude

متغیرهای مستقل موجود در مدل که در این مثال
شامل عرض ارضه . تعداد جعبه ها و تلفن متغیر است

خطای معیار متن ، S_e ، که در این مثال برابر 3.1407 شده است

ضریب تعیین تعدیل شده ، r^2_{adj} ، که در این مثال برابر 0.693 شده است

ضریب تعیین متن ، r^2 است که در این مثال برابر 0.737 شده است

ضریب همبستگی متن ، r ، است که در این مثال برابر 0.858 است

لازم به ذکر است که ضریب تعیین تعدیل شده از رابطه زیر به دست می آید:

$$\tilde{r}^2 = 1 - \frac{\frac{\sum (y-\hat{y})^2}{n-2}}{\frac{\sum (y-\bar{y})^2}{n-1}}$$

و S_e یا خطای معیار مدل در این حالت از رابطه زیر به دست می آید:

$$S_e = \sqrt{\frac{\sum (y-\hat{y})^2}{n-k-1}}$$

که n تعداد مشاهدات و k تعداد متغرهای مستقل است.

در زیر جدول ANOVA حاصل از خروجی دیده می شود.

جدول زیر فرضیه زیر را آزمون می کند.

$$\begin{cases} H_0 : \alpha = \beta_1 = \beta_2 = ... = \beta_k \\ H_1 : \boxed{\text{حداقل یکی از تساوی ها برقرار نباشد}} \end{cases}$$

در مورد این مثال فرضیه فوق به صورت زیر بیان می شود.

$$\begin{cases} H_0 : \alpha = \beta_1 = \beta_2 \\ H_1 : \boxed{\text{حداقل یکی از تساوی ها برقرار نباشد}} \end{cases}$$

با توجه به نتایج آزمون، آماره آزمون $F_0 = 16.795$ حاصل شده است که باید با مقدار بحرانی $F_{0.05,2,12} = 3.89$ مقایسه شود. با توجه به اینکه آماره آزمون بزرگتر از مقدار بحرانی شده است فرض صفر رد می شود.

یک راه دیگر برای آزمون فرض آماری فوق، استفاده از p- مقدار حاصل از خروجی است که در سطح معنی دار ۰/۰۵، با توجه به اینکه p-مقدار کمتر از ۰/۰۵ شده است فرض صفر رد می شود.

درجه آزادی که k نشان‌دهنده تعداد متغیرهای مستقل و n تعداد نمونه است

مجموع مربعات ناگکها

مقدار آماره F که از تقسیم ستون اول ستون قبل به ستون دوم آن به دست می آید.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	331.359	k= 2	165.679	16.795	.000[a]
	Residual	118.375	n-k-1=12	9.865		
	Total	449.733	n-1= 14			

a. Predictors: (Constant), Nbox, longitude

b. Dependent Variable: time

رگرسیون یا متغیر مستقل

باقیمانده یا خطا

کل

میانگین مربعات، که از تقسیم ستون مجموع مربعات بر درجه آزادی حاصل می شود

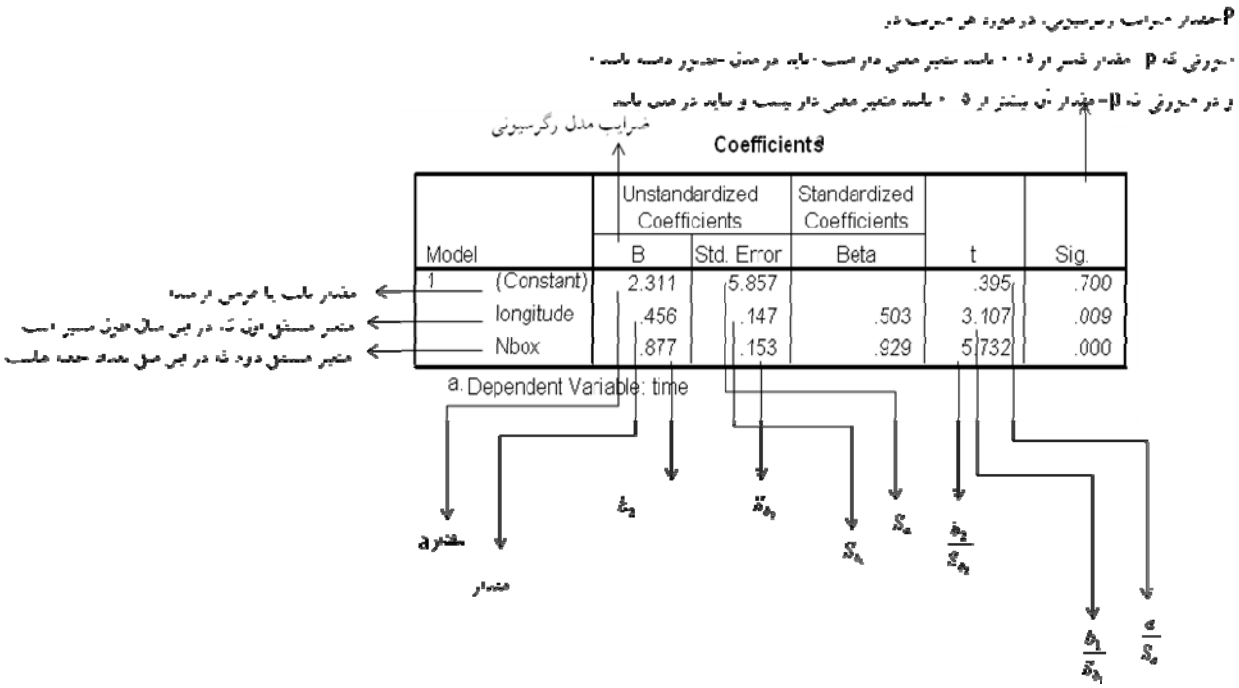
متغیرهای پیش بینی یا به عبارتی متغیرهای مستقل سامل عرض از مبدا تعداد جمله ها و تعاون متغیر

متغیر وابسته که در این مثال زمان است

P مقدار معنی دار مشروطی که کمتر از ۰۰۵ باشد معنی دار است به عبارت دیگر متغیر یا متغیرهای مستقل بر متغیر وابسته اثر دارند و در غیر این صورت یعنی

P مقدار بیشتر از ۰۰۵ باشد معنی دار نیست یعنی متغیر یا متغیرهای مستقل بر متغیر وابسته اثر ندارند

جدول زیر در واقع قسمت اصلی تحلیل رگرسیونی است. با استفاده از آن می توان معادله خط رگرسیونی، را نوشت و در مورد هر یک از پارامترها استنباط انجام داد.



با توجه به نتایج حاصل از جدول فوق معادله خط رگرسیونی عبارت است از:

$$y = 2.311 + 0.147x_1 + 0.153x_2$$

در مورد آزمون فرض آماری

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

آماره آزمون برابر $t_0 = \frac{a}{S_a} = 0.395$ شد که باید با مقدار بحرانی $t_{\frac{\alpha}{2},n-k-1} = t_{0.025,12} = \pm 2.179$ مقایسه شود. با توجه به اینکه در ناحیه پذیرش H0 قرار دارد، فرض صفر رد نمی شود.

یک راه دیگر برای آزمون فرض آماری فوق، استفاده از p- مقدار حاصل از خروجی است که در سطح معنی دار ۰/۰۵، با توجه به اینکه p-مقدار بیشتر از ۰/۰۵ شده است فرض صفر رد نمی شود.

در مورد آزمون فرض آماری

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

آماره آزمون برابر $t_1 = \frac{b_1}{S_{b_1}} = 3.107$ شد که باید با مقدار بحرانی $t_{\frac{\alpha}{2},n-k-1} = t_{0.025,12} = \pm 2.179$ مقایسه شود. با توجه به اینکه در ناحیه پذیرش H1 قرار دارد، فرض صفر رد می شود.

مانند قبل یک راه دیگر برای آزمون فرض آماری فوق، استفاده از p- مقدار حاصل از خروجی است که در سطح معنی دار ۰/۰۵، با توجه به اینکه p-مقدار کمتر از ۰/۰۵ شده است فرض صفر رد می شود.

در مورد آزمون فرض آماری

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

نیز روند مانند فوق است.

❖ رگرسیون ناخطی

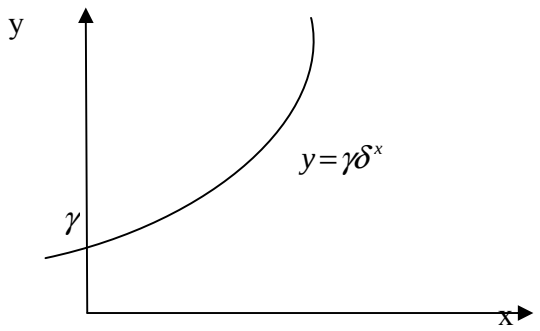
در بخش های گذشته در مورد رگرسیون خطی یک متغیر مستقل در مقابل یک متغیر وابسته یا چند متغیر مستقل در مقابل یک متغیر وابسته بحث شد. ولی همانطور که واضح است. ارتباط بین دو متغیر لزوما خطی نیست و حالت های متفاوتی می تواند داشته باشد. و این حالت ها بسیار متنوع و گسترده هستند و بحث در مورد تمام آنها امکان پذیر نیست. بنابراین در این بخش ما در مورد ۲ شکل غیرخطی که نسبتا ساده بوده و از طریق رابطه خطی قابل برآورد هستند بحث می کنیم یکی از این حالت ها نمایی و دیگری سهمی است.

- **رگرسیون نمایی**

معادله تابع نمایی به صورت زیر است:

معادله تابع نمایی $\rightarrow$ $y=\gamma\delta^x$

که در آن (γ) گاما و (δ) دلتا دو پارامتر تابع هستند.



شکل تابع نمایی

اگر c و d به ترتیب برآوردکننده های γ و δ باشند، در این صورت معادله نمایی به صورت زیر خواهد شد.

$$\hat{y}=cd^x$$

هدف یافتن c و d می‌باشد پس از طرفین معادله لگاریتم می‌گیریم:

$$\log\hat{y}=\log cd^x=\log c+x\ \log d \ \rightarrow \log\hat{y}=\log c+(\log d)x$$

اگر $a=\log c$ و $b=\log d$ باشد داریم.

$$\log\hat{y}=a+bx$$

رابطه فوق خطی است و می‌توان با استفاده از نقاط $(x_i\ ,\ \log y_i)$ مقادیر a و b را محاسبه کرد.

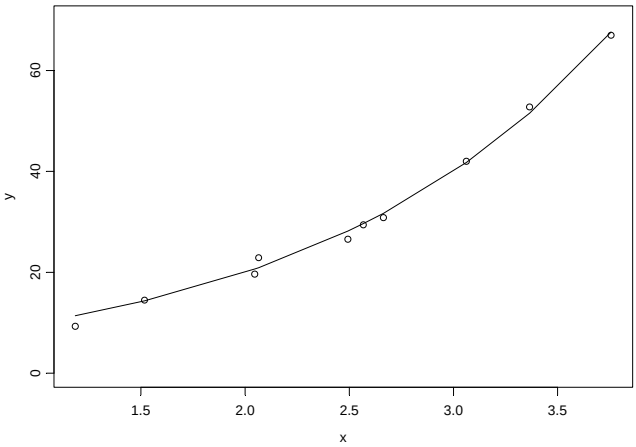
$$b=\frac{\Sigma x\log y-n\,\bar{x}\,(\overline{\log y})}{\Sigma x^2-n\bar{x}^2}\qquad,\qquad a=(\overline{\log y})-b\bar{x}$$

مثال(۲): داده‌های زیر را در نظر بگیرید :

y	30.82	22.87	14.47	52.75	19.62	29.38	9.27	26.54	41.99	66.98
x	2.66	2.07	1.52	3.37	2.05	2.57	1.19	2.49	3.06	3.76

بعد از رسم نمودار پراکنش مشخص شده است که تابع فوق تقریباً نمایی است. معادله رگرسیون نمایی را برای داده‌های فوق برآورد کنید.

حل :



x	y	$\log y$	x^2	$x(\log y)$
2.66	30.82	3.43	7.10	9.13
2.07	22.87	3.13	4.26	6.64
1.52	14.47	2.67	2.30	4.05
3.37	52.57	3.97	11.33	13.35
2.05	19.62	2.98	4.19	6.09
2.57	29.38	3.38	6.60	8.68
1.19	9.27	2.23	1.40	2.64
2.49	26.54	3.28	6.22	8.17
3.06	41.99	3.74	9.38	11.45
3.76	66.98	4.20	14.12	15.80
24.7	–	33.0		85.82

$$\bar{x} = \frac{24.7}{10} = 2.47 \quad , \overline{\log y} = \frac{\Sigma \log y}{n} = \frac{33.0}{10} = 3.30$$

$$b = \frac{\sum x \log y - n(\bar{x})(\overline{\log y})}{\sum x^2 - n(\bar{x})^2} = \frac{85.82 - (10)(2.47)(3.30)}{66.9 - (10)(2.47)^2} \rightarrow b = 0.73$$

$$a = \overline{Logy} - b(\bar{x}) = 3.30 - 0.73(2.47) = 1.50$$

$$b = \log d \quad \rightarrow d = 10^b \ \rightarrow d = 10^{(0.73)} = 5.37$$

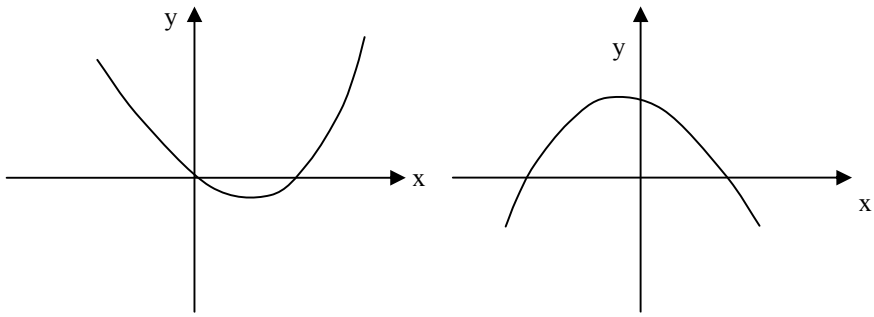
$$\rightarrow \ \hat{y} = 31.62(5.37)^x$$

$$a = \log c \ \rightarrow c = 10^a \ \rightarrow c = 10^{1.5} \ = 31.62$$

- **رگرسیون سهمی**

معادله سهمی را می توان به صورت زیر نوشت که در آن x ، متغیر پاسخ و y متغیر وابسته است:

معادله سهمی → $y=\alpha+\beta_1 x+b_2 x^2 \qquad ,\beta \neq 0$



$\beta_۲ >۰$ (الف) $\qquad \beta_۲ <۰$

شکل کلی نمودار سهمی

برآورد معادله تابع سهمی → $\hat{y}=a+b_۱ x+b_۲ x^۲$

که در آن $b_۲,b_۱,a$ به ترتیب برآورد کننده $\beta_۲,\beta_۱,\alpha$ هستند.

اگر در معادله سهمی مقدار $x_۱$, $x_۲$ را به صورت $x_۱=x$ و $x_۲=x^۲$ تعریف کنیم. در این صورت داریم :

$\hat{y}=a+b_۱ x_۱+b_۲ x_۲$

که مشابه با شکل کلی معادله خطی دو متغیره است و از طریق حل دستگاه زیر $b_۲,b_۱,a$ را بدست می‌آوریم.

$$\begin{cases} na+b_1\Sigma x_1 +b_2\Sigma x_2=y \\ a\Sigma x_1+b_1\Sigma x_1^2+b_1\Sigma x_1 x_2=\Sigma x_1 y \\ a\Sigma x_2+b_1\Sigma x_1 x_2+b_2\Sigma x_2^2=\Sigma x_2 y \end{cases}$$

مثال (۳): این داده‌ها را در نظر بگیرید :

x	۴	۴	۱۰	۳	۶	۵	۵	۳	۷	۷	۸	۱۰	۸
y	۱۸	۱۲	۱۰	۶	۲۴	۲۰	۱۸	۱۰	۲۰	۲۲	۲۰	۸	۱۷

پس از رسم نمودار پراکنش مشخص شد که تابع آنها تقریباً سهمی است معادله رگرسیون آن را برآورد کنید. با فرض $x_1 = x$, $x_2 = x^2$ داریم:

$$\hat{y} = -24/368 + 14/132x_1 - 1/083x_2$$

بنابراین معادله رگرسیون سهمی برآوردی به این ترتیب است :

$$\hat{y} = -۲۴/۳۶۸ + ۱۴/۱۳۲x - ۱/۰۸۳x^۲$$

که در آن بجای x_1, x_2 به ترتیب مقادیر آنها یعنی x و x^2 گذاشته شده است.

❖ رگرسیون و متغیرهای مستقل کیفی

می‌توان متغیرهای مستقل کیفی را هم در مدل رگرسیون وارد کرد. متغیرهایی مثل شیوه مدیریت (رابطه‌مداری و وظیفه‌مداری)، وضعیت تأهل (مجرد و متأهل) که از نوع متغیرهای کیفی هستند.

- به هر یک از حالات ممکن یک متغیر کیفی، یک سطح می‌گویند. مثلاً متغیر شیوه‌ی مدیریت (رابطه‌مداری و وظیفه‌مداری) دو سطح دارد و یا نوع گزارش حسابرسی (قبول، رد، شرط و عدم اظهار نظر) چهار سطح دارد.

نکته : متغیرهای مستقل کیفی با استفاده از متغیرهای مجازی وارد مدل می‌شوند.

۱- تعداد سطوح متغیر کیفی = تعداد متغیرهای مجازی برای یک متغیر کیفی

متغیر مجازی متغیری است دو ارزشی (۰، ۱) که مقدار صفر برای یک حالت و مقدار یک برای حالت دیگر قابل تعریف است. مثلاً اگر متغیر کیفی دارای سه سطح باشد (مثلاً مکان A و B و C) تعداد متغیرهای مجازی لازم، طبق تعریف برابر ۲ است (۲=۱-۳) که در جدول زیر آمده است :

مکان	x_1	x_2
A	۰	۰
B	۱	۰
C	۰	۱

مثال (۴): میزان فروش ۱۰ شعبه یک فروشگاه زنجیره ای در ماه گذشته همراه با مساحت شعبه، میزان تبلیغات شعبه و مکان شعبه به صورت زیر است:

مکان	میزان تبلیغات (x_2)	مساحت (x_1)	فروش	شماره شعبه

۱	۴۰۰	۲۰۰۰	۲۰	A
۲	۲۸۵	۱۵۰۰	۱۰	B
۳	۴۵۰	۲۹۰۰	۳۵	C
۴	۳۸۰	۱۵۰۰	۲۸	A
۵	۲۶۷	۲۰۰۰	۱۰	B
۶	۵۲۰	۳۸۰۰	۳۰	A
۷	۳۸۰	۳۰۰۰	۳۲	B
۸	۹۲۰	۴۶۰۰	۴۵	C
۹	۸۳۰	۳۹۰۰	۴۱	C
۱۰	۶۲۰	۴۰۵۰	۳۴	C

اگر فروش را y، مساحت را x_1، میزان تبلیغات را x_2 تعریف کنیم و برای متغیر کیفی مکان (که دارای ۳ سطح است و نیاز به دو متغیر مجازی دارد.) دو متغیر مجازی x_3 و x_4 را به گونه زیر تعریف می کنیم.

$$x_3 = \begin{cases} 0 & \text{اگر در مکان B نباشد.} \\ 1 & \text{اگر در مکان B باشد.} \end{cases}$$

$$x_4 = \begin{cases} 0 & \text{اگر در مکان C نباشد.} \\ 1 & \text{اگر در مکان C باشد.} \end{cases}$$

روشن است که اگر x_3 و x_4 هر دو صفر باشند، مکان A است.

بنابراین داده های مورد نظر به صورت زیر خواهد شد:

شماره شعبه	فروش	مساحت (x_1)	میزان تبلیغات (x_2)	(x_3)	(x_4)
۱	۴۰۰	۲۰۰۰	۲۰	۰	۰
۲	۲۸۵	۱۵۰۰	۱۰	۱	۰
۳	۴۵۰	۲۹۰۰	۳۵	۰	۱
۴	۳۸۰	۱۵۰۰	۲۸	۰	۰
۵	۲۶۷	۲۰۰۰	۱۰	۱	۰
۶	۵۲۰	۳۸۰۰	۳۰	۰	۰
۷	۳۸۰	۳۰۰۰	۳۲	۱	۰
۸	۹۲۰	۴۶۰۰	۴۵	۰	۱
۹	۸۳۰	۳۹۰۰	۴۱	۰	۱
۱۰	۶۲۰	۴۰۵۰	۳۴	۰	۱

با استفاده از نرم افزار SPSS، معادله رگرسیون برآوردی به صورت زیر خواهد شد.

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 = 95.2 + 0.09x_1 + 4.33x_2 - 60.44x_3 + 83.96x_4$$

❖ رگرسیون و تحلیل تمایزات

یکی از مسائلی که در بعضی از پژوهش ها پیش می آید طبقه بندی کردن سوژه ها (افراد، سازمان ها و غیره) بر حسب نمره هایشان است. برای این منظور می توان از تحلیل تمایزات (که گاهی به آن تحلیل ممیزی یا تحلیل تشخیص می گویند.) استفاده کرد. در تحلیل تمایزات می توانیم افراد، سازمان ها، کالاها و غیره را بر مبنای دو یا چند متغیر مستقل (پیش بین) به دو یا چند گروه منتسب کنیم.

در شیوه تحلیل تمایزات، تابع ممیز، معادله ای رگرسیونی با یک متغیر وابسته (y) است که این متغیر عضویت گروهی را نشان می دهد. متغیر وابسته y تنها دو مقدار صفر و یک را می گیرد. بنابراین تفاوت تحلیل تمایزات با قسمت قبل (متغیرهای کیفی) این است که در قسمت قبل، متغیر وابسته y متغیری پیوسته بود ولی در اینجا (تحلیل تمایزات)، متغیر وابسته y ، متغیری کیفی دو ارزشی است.

با داشتن مجموعه ای از سوژه ها (افراد، شرکت ها و غیره) که عضویت گروهی آنها از قبل مشخص است، تابع ممیز را برآورد می کنیم و بر اساس سوژه های جدید را در یکی از طبقات قرار می دهیم.
مثال: فرض کنید ۱۰ نفر از دانشجویان کلاس آمار را برگزیده و میزان علاقه و تلاش و وضعیت قبولی شان را مشخص کرده ایم که نتایج آن در جدول زیر آمده است.(ق=قبولی و ر=رد)

میزان علاقه	میزان تلاش	وضعیت قبولی
3	8	ق
4	7	ق
5	5	ق
4	3	ق
2	3	ق
2	4	ر
1	3	ر
2	3	ر
2	2	ر
5	2	ر

مشخص است که وضعیت قبولی فرد (y) تابعی از میزان تلاش (x_1) و میزان علاقه (x_2) است. در این حالت متغیر وابسته y متغیر کیفی دو سطحی (قبولی و ردی) است. اگر به قبولی نمره ۱۰ و به ردی نمره صفر را اختصاص دهیم و معادله رگرسیون را برآورد کنیم، تابع رگرسیون برآوردی، که به آن تابع ممیز می گوییم، به صورت زیر خواهد شد:

$$\hat{y} = -0.417 + 0.139x_1 + 0.120x_2$$

حال اگر دانشجوی دیگری مثلا با میزان تلاش ۴ و علاقه ۵ داشته باشیم، آیا آن را موفق پیش بینی خواهیم کرد یا خیر، یعنی آن را در طبقه قبولی ها پیش بینی می کنیم یا ردیها؟ از معادله ممیز بدین منظور استفاده می کنیم:

$$\hat{y} = -0.417 + 0.139(4) + 0.120(5) = 0.739$$

عدد ۰/۷۳۹ بیشتر از ۰/۵ است، بنابراین پیش بینی ما آن است که در طبقه قبولی ها قرار گیرد.

به نام خدا

فصل پانزدهم: کاربردهای آزمون کای- مربع در مدیریت

مقدمه: در فصول مربوط به تخمین و آزمون فرض آماری در تعیین فاصله اطمینان برای واریانس و نیز آزمون فرضیه درباره آن از توزیع کای-مربع به اختصار یاد کردیم . توزیع کای مربع یک توزیع آماری است که در تحلیل های آماری کاربردهای فراوان دارد. میانگین و واریانس توزیع کای- مربع با k درجه آزادی، به ترتیب عبارت است از: k و 2k.

توزیع کای- مربع علاوه بر تخمین و آزمون آماری واریانس، کاربردهای بسیاری دارد. از توزیع کای-مربع در آزمون فرضیه هایی که داده های مورد تجزیه و تحلیل به صورت فراوانی ارائه شده اند، می توان استفاده کرد. در این فصل درباره شیو های آزمون فرضیه های مزبور تحت عناوین آزمون استقلال و آزمون نیکویی برازش به تفصیل بحث خواهد شد.

آزمون استقلال:

متداولترین استفاده از توزیع کای – مربع آزمون فرضیه، وجود استقلال بین دو متغیر کیفی (طبقه ای) است.
داده های جمع آوری شده برای متغیرهای کیفی در یک جدول که شامل r سطر و c ستون است خلاصه می شوند. به طور کلی چنین جدولی را جدول توافقی گویند.

متغیر دوم / متغیر اول	۱ ۲ ۳ ... C	مجموع
۱	$n_{۱۱}$ $n_{۱۲}$ $n_{۱۳}$ ... $n_{۱c}$	$n_{۱.}$
۲	$n_{۲۱}$ $n_{۲۲}$ $n_{۲۳}$ ... $n_{۲c}$	$n_{۲.}$
۳	$n_{۳۱}$ $n_{۳۲}$ $n_{۳۳}$ ... $n_{۳c}$	$n_{۳.}$
⋮	⋮ ⋮ ⋮ ⋮ ⋮	⋮
r	$n_{r۱}$ $n_{r۲}$ $n_{r۳}$ ... n_{rc}	$n_{r.}$
مجموع	$n_{.۱}$ $n_{.۲}$ $n_{.۳}$ ... $n_{.c}$	n

❖ جدول توافقی

مقادیر n_{ij} فراوانی‌های مشاهده شده در هر سلول است. که فصل مشترک هر سلول i و ستون j می‌باشد. $n_{i.}$، جمع مقادیر سطر iام و $n_{.j}$، جمع مقادیر ستون jام است.

برای انجام آزمون استقلال باید مراحل زیر را دنبال کرد :

۱- تعریف فرضیه آماری

$$\begin{cases} H_0 : & \text{دو متغیر سطر و ستون مستقل هستند} \\ H_1 : & \text{دو متغیر سطر و ستون مستقل نیستند} \end{cases}$$

۲-محاسبه آماره آزمون

آماره آزمون در این حالت به صورت زیر تعریف می شود .

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(Fo_i - Fe_i)^2}{Fe_i}$$

که در فرمول فوق Fo_i، همان Fo_{ij}، فراوانی های مشاهده شده متناظر با سلول ij و Fe_i، همان Fe_{ij}، مقادیر مورد انتظار سلول ij است که به صورت زیر تعریف می شود.

$$Fe_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

که $n_{i.}$، جمع سطری و $n_{j.}$، جمع ستونی و در انتها n تعداد کل مشاهدات است.

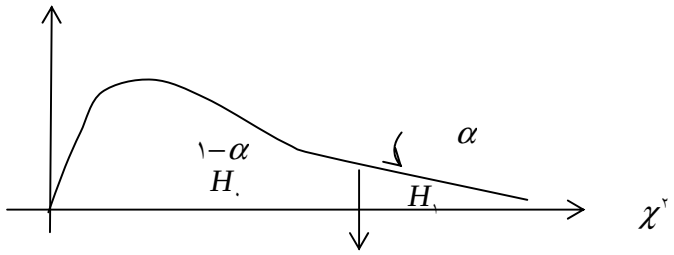
اگر مجموع تمایز مقادیر مشاهده شده و مقادیر مورد انتظار، «کوچک» باشد. فرضیه صفر قابل قبول است و دو متغیر مستقلند ولی اگر مجموع تمایز مقادیر مشاهده شده و مقادیر مورد انتظار «بزرگ» باشد. فرضیه صفر رد می‌شود و دو متغیر مستقل از هم نیستند.

۳- مقدار بحرانی

مقدار بحرانی در این حالت برابر است با $\chi^2_{\alpha,(r-1)(c-1)}$ که α سطح خطا، r تعداد سطر جدول توافقی و c تعداد ستون جدول توافقی است.

۴- تصمیم‌گیری

در صورتی که مقدار آماره آزمون بیشتر از مقدار بحرانی باشد. ($\chi^2_0 > \chi^2_{\alpha,(r-1)(c-1)}$) فرض صفر رد می شود و فرض مقابل پذیرفته می شود و در غیر اینصورت فرض صفر تایید و فرض مقابل رد می شود. ناحیه پذیرش H1 و H0 در شکل زیر دیده می شود.



مقدار بحرانی (χ^2_α)

مثال۱: در تحقیقی از ۱۰۰۰ مورد مرگ مردان بین سنین ۴۵ تا ۶۴ ساله، علل مرگ همراه عادت سیگار کشیدن آنها ثبت شده است. پژوهشگر به دنبال پاسخ این سوال است که آیا براساس داده های حاصل می توان نتیجه گرفت که رابطه ای بین علت مرگ و سیگار کشیدن افراد وجود دارد یا نه؟ از داده های جدول زیر برای آزمون این ادعا که علت مرگ مستقل از سیگار کشیدن است استفاده می کنیم.

حل:

۱- فرضیه آماری

$$\begin{cases} H_0: & \text{اعتیاد به سیگار و علت مرگ از هم مستقل هستند} \\ H_1: & \text{اعتیاد به سیگار و علت مرگ از هم مستقل نیستند} \end{cases}$$

۲- آماره آزمون

	علت مرگ			
	سرطان	بیماریهای قلبی	سایر امراض	جمع سطری
سیگاری	۱۳۵ (۱۲۳/۵)	۳۱۰ (۳۰۲/۲۵)	۲۰۵ (۲۲۴/۲۵)	۶۵۰
غیرسیگاری	۵۵ (۶۶/۵)	۱۵۵ (۱۶۲/۷۵)	۱۴۰ (۱۲۰/۷۵)	۳۵۰
جمع ستونی	۱۹۰	۴۶۵	۳۴۵	۱۰۰۰

مقادیر داخل پرانتز مقادیر مورد انتظار هستند.

$$Fe_1 = \frac{n_{1.} \times n_{.1}}{n} = \frac{650 \times 190}{1000} = 123.5$$

محاسبه آماره آزمون

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(Fo_i - Fe_i)^2}{Fe_i} = \left[\frac{(135-123.5)^2}{123.5} + \frac{(310-302.25)^2}{302.25} + \ldots + \frac{(140-120.75)^2}{120.75} \right] = 8.349$$

۳- مقدار بحرانی

$$\chi^2_{\alpha,[(r-1)(c-1)]} = \chi^2_{0.05,1\times 2} = \chi^2_{(0.05,2)} = 5.991$$

۴- تصمیم گیری

همانطور که بیان شد در صورتی که $\chi_0^2 > \chi^2_{\alpha,(r-1)(c-1)}$ باشد فرض صفر رد می شود در این مثال نیز با توجه به اینکه $\chi_0^2 = 8.349 > 5.991 = \chi^2_{\alpha,(r-1)(c-1)}$ فرض صفر رد می شود پس فرض صفر استقلال بین دو متغیر را رد می کنیم سپس به نظر می رسد که مصرف سیگار و علت مرگ وابسته اند.

❖ آزمون نیکویی برازش

در این آزمون به تطابق توزیع نمونه با توزیع نظری با استفاده از ملاک (کای دو) می پردازیم.
هدف ما آزمون معنی دار بودن اختلاف بین فراوانی های مشاهده شده و فراوانی هایی است که از نظر تئوری انتظار داریم. چون آزمون می کنیم که تا چند اندازه توزیع فراوانی مشاهده شده (فراوانی تجربی) بر توزیع فراوانی مورد انتظار (فراوانی نظری) منطبق می شود (برازنده است) این روش اغلب به عنوان آزمون نیکویی برازش معروف است.
چهار مرحله انجام این آزمون را می توان به صورت زیر بیان کرد.

۱- فرضیه آماری

فرضیه آماری به صورت زیر است:

H_0: توزیع نمونه با توزیع موردنظر تطابق دارد (مثلاً نرمال است)

H_1:توزیع نمونه با توزیع موردنظر تطابق ندارد.

۲- آماره آزمون

آماره آزمون نیکویی برازش به صورت زیر است:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(Fo_i - Fe_i)^2}{Fe_i}$$

که در آن Fo_i و Fe_i به ترتیب فروانی های مشاهده شده (فراوانی تجربی وفراوانیهای مورد انتظار (فراوانی نظری) می باشند.

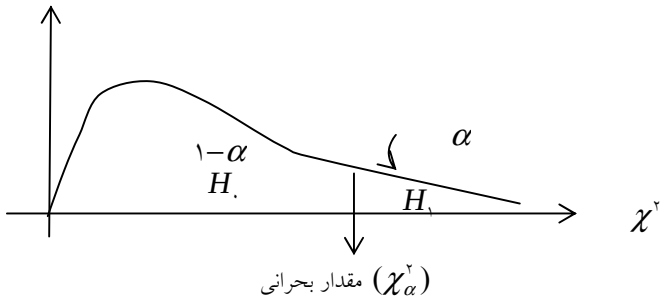
اگر فراوانی های مشاهده شده نزدیک به مقادیر مورد انتظار مربوطه باشند، مقدار χ_0^2 کوچک شده و آن خود نشانه خوبی انطباق (یا برازش) است. اگر مقادیر مشاهده شده به طور قابل ملاحظه ای از مقادیر مورد انتظار دور باشند مقدار χ_0^2 شده و انطباق (یا برازش) ضعیف خواهد شد. خوبی برازش (مقدار χ_0^2 کوچک) منجر به رد نکردن H0 شده و حال آنکه ضعیف بودن برازش (مقدار χ_0^2 بزرگ) منتهی به رد H0 خواهد شد.

۳- مقدار بحرانی

مقدار بحرانی در این حالت برابر است با $\chi^2_{\alpha,k-p-1}$ که α سطح خطا، k تعداد رسته های مختلف یا تعداد گروههای مختلف و p تعداد پارامترهای جامعه است. که برمبنای داده های نمونه ای برآورد می شود.

۴- تصمیم‌گیری

در صورتی که مقدار آماره آزمون بیشتر از مقدار بحرانی باشد. ($\chi^2_0 > \chi^2_{\alpha,k-p-1}$) فرض صفر رد می شود و فرض مقابل پذیرفته می شود و در غیر اینصورت فرض صفر تایید و فرض مقابل رد می شود. ناحیه پذیرش H1 و H0 در شکل زیر دیده می شود.



مثال۲: تاسی را ۱۲۰ بار پرتاب می کنیم. نتایج زیر مشاهده شده است.

روس تاس	1	2	3	4	5	6
تعداد دفعات	20	22	17	18	19	24

آیا می توان ادعا کرد تاس سالم است. و در بیان آماری آیا می توان گفت نتایج حاصل از پرتاب تاس دارای توزیع یکنواخت است؟

حل:

۱- فرضیه آماری

H_0: نتایج حاصل از پرتاب تاس دارای توزیع یکنواخت است.

H_1: نتایج حاصل از پرتاب تاس دارای توزیع یکنواخت نیست.

۲- آماره آزمون

از نظر تئوری اگر تاس سالم باشد. (دارای توزیع یکنواخت است.) انتظار خواهیم داشت که هر کدام از اعداد ۱ تا ۶ را ۲۰ بار نشان دهد. ($120\times\frac{1}{6}=20$).

روی تاس	1	2	3	4	5	6
فراوانی تجربی	20	22	17	18	19	24
فراوانی مورد انتظار	20	20	20	20	20	20

$$\chi_0^2=\sum\frac{(Fo_i-Fe_i)^2}{Fe_i}=\frac{(20-20)^2}{20}+\frac{(22-20)^2}{20}+\frac{(17-20)^2}{20}+\frac{(18-20)^2}{20}+\frac{(19-20)^2}{20}+\frac{(24-20)^2}{20}$$
$$=1.7$$

۳- مقدار بحرانی

نیز برابر است با $\chi^2_{\alpha,k-p-1}$ که طبق مسئله α برابر ۰/۰۵ و k که برابر با تعداد رسته هاست در این مثال برابر ۶ است در مورد p با توجه به اینکه در توزیع یکنواخت ما هیچ پارامتری را برآورد نکردیم پس برابر صفر است. پس داریم:

$$\chi^2_{0.05,6-0-1}=\chi^2_{0.05,5}=11.07$$

۴- تصمیم گیری

با توجه به اینکه ۱/۷ <۱۱/۰۷ بنابراین $\chi_0^2<\chi^2_{0.05,5}$، یعنی مقدار عددی آماره آزمون در ناحیه بحرانی قرار نمی گیرد. لذا فرض صفر رد نمی شود. در نتیجه می توان گفت مشاهدات دارای توزیع یکنواخت است و به عبارتی تاس سالم است.

مثال ۲: نمونه‌ای شامل ۱۰۰ کارت منگنه شده را مورد بررسی قرار داده‌ایم. توزیع تعداد اشتباهات به شرح جدول زیر است.

$x =$ تعداد اشتباهات	۰	۱	۲	۳	۴	۵
$Fo_i =$ تعداد کارتهای منگنه شده	۲۵	۳۰	۲۵	۱۲	۶	۲

فرضیه اینکه توزیع فراوانی تعداد اشتباهات از قانون پواسون تبعیت می‌کند را در سطح معنی‌دار بودن α=۰/۱ آزمون نمائید.

۱- فرضیه آماری

$$\begin{cases} H_۰ & \text{توزیع اشتباهات از قانون پواسون تبعیت می‌کند} \\ H_۱ & \text{توزیع اشتباهات از قانون پواسون تبعیت نمی‌کند} \end{cases}$$

۲- آماره آزمون

محاسبه مقادیر مورد انتظار:

$$P(X=x)=\frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

$$\gamma=\lambda=\frac{\Sigma x_i F_i}{n}=\frac{۰\times۲۵+۱\times۳۰+۲\times۲۵+۳\times۱۲+۴\times۶+۵\times۲}{۱۰۰}=۱/۵$$

و حال با استفاده از فرمول احتمال مربوط به توزیع جدول زیر را ایجاد می‌کنیم :

X	Fo	$P=\frac{[e^{-1/5}\,(1/5)^{x}]}{x!}$	Fe=۱۰۰×P
۰	۲۵	۰/۲۲۳۱	۲۲/۳
۱	۳۰	۰/۳۳۴۷	۳۳/۵
۲	۲۵	۰/۲۵۱۰	۲۵/۱
۳	۱۲	۰/۱۲۵۵	۱۲/۶
۴	۶	۰/۰۴۷۱	۴/۷
$x\geq$۵	۲	۰/۰۱۷۶	۱/۸
Σ	۱۰۰	۱	۱۰۰

- در صورتی که در بعضی از رسته ها، فراوانی های مورد انتظار کوچک باشد، (کمتر از ۵ باشد.) تقریب بوسیله χ^2 چندان معتبر نیست. در مواجهه با چنین مواقعی رسته های مجاور را ترکیب می کنند تا به حداقل مقدار پیشنهادی برسد.

چون فراوانی سطری دو طبقه آخر از ۵ کمتر است بنابراین با هم ادغام می‌شوند و جدول جدید یک طبقه کمتر خواهد شد.

x	Fo	Fe	$Fo-Fe$	$(Fo-Fe)^۲$	$(Fo-Fe)^۲ / Fe$
۰	۲۵	۲۲/۳	۲/۷	۷/۲۹	۰/۳۳
۱	۳۰	۳۳/۵	−۳/۵	۱۲/۲۵	۰/۳۷
۲	۲۵	۲۵/۱	−۰/۱	۰/۰۱	۰
۳	۱۲	۱۲/۶	−۰/۶	۰/۳۶	۰/۰۳
$x \geq ۴$	۸	۶/۵	۱/۵	۲/۲۵	۰/۳۵
	۱۰۰	۱۰۰	۰		

بنابراین مقدار آماره آزمون عبارت است از:

$$\chi_0^2 = 0.33 + 0.37 + 0 + 0.03 + 0.35 = 1.08$$

۳- مقدار بحرانی

مقدار بحرانی نیز برابر است با $\chi^2_{\alpha,k-p-1}$ که طبق مسئله α برابر ۰/۱ و k که برابر با تعداد رسته هاست در این مثال برابر ۵ است در مورد p با توجه به اینکه در توزیع پواسون ما مجبور به برآورد λ با استفاده از مشاهدات شدیم پس برابر یک است. پس داریم:

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.1,5-1-1} = \chi^2_{0.1,3} = 6.25$$

۴- تصمیم گیری

با توجه به اینکه ۶/۲۵> ۱/۰۸ بنابراین $\chi_0^2 < \chi_{0.1,3}^2$، یعنی مقدار عددی آماره آزمون در ناحیه بحرانی قرار نمی گیرد. لذا فرض صفر رد نمی شود. در نتیجه می توان گفت توزیع اشتباهات کارت‌های منگنه شده از توزیع احتمال پواسون تبعیت می‌کنند.

مثال ۳: اطلاعات جدول زیر که مربوط بـه فشـارخون سیسـتولیک نمونـه ای ازمـردان ۳۵ سـال بـه بـالای روستایی است رادرنظربگیرید. تطابق توزیع صفت فشـارخون رادرایـن جامعـه بـاتوزیع نظـری نرمـال آزمون کنید.

گروه	۱	۲	۳	۴	جمع
فشار خون سیستولیک	۶۰-۹۰	۹۰-۱۲۰	۱۲۰-۱۵۰	۱۵۰-۱۸۰	
فراوانی مشاهده شده	۰	۱۵	۵۷	۲۸	۱۰۰

حل:

۱- فرضیه آماری

$$\begin{cases} H_0: & \text{توزیع داده ها نرمال است.} \\ H_1: & \text{توزیع داده ها نرمال نیست.} \end{cases}$$

۲- آماره آزمون

محاسبه مقادیر مورد انتظار

$$X \sim N(\mu,\sigma)$$

$$p = p(X < x) = p(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = p(Z < z)$$

گروه	فشارخون سیستولیک	فراوانی مشاهده شده	x_i	$Fo_i x_i$	$(x_i-\bar{x})^2$	$Fo_i(x_i-\bar{x})^2$
۱	۶۰–۹۰	۰	۷۵	۰	۴۰۸۳/۳۲	۰
۲	۹۰–۱۲۰	۱۵	۱۰۵	۱۵۷۵	۱۱۴۹/۲۱	۱۷۲۳۸/۱۵
۳	۱۲۰–۱۵۰	۵۷	۱۳۵	۷۶۹۵	۱۵/۲۱	۸۶۶/۹۷
۴	۱۵۰–۱۸۰	۲۸	۱۶۵	۴۶۲۰	۶۸۱/۲۱	۱۹۰۷۳/۸۸
جمع		۱۰۰		۱۳۸۹۰		۳۷۱۷۹

$$\bar{x}=\frac{13890}{100}=138.9, \qquad S^2=\frac{\sum Fo_i(x_i-\bar{x})^2}{n-1}=\frac{37179}{99}=375.54, \quad S=19.38$$

$$\Rightarrow X \approx N(138.9, 19.38)$$

$$p_1=p(X<90)=p(Z<\frac{90-138.9}{19.38})=p(Z<-2.52)=0.0059$$

$$p_2=p(X<120)=p(Z<\frac{120-138.9}{19.38})=p(Z<-0.98)=0.1635$$

$$p_3=p(X<150)=p(Z<\frac{150-138.9}{19.38})=p(Z<0.58)=0.7190$$

$$p_4=p(X<180)=p(Z>\frac{180-138.9}{19.38})=p(Z<2.12)=0.9834$$

$$p_0=p(X<60)=p(Z<\frac{60-138.9}{19.38})=p(Z<-4.07)=0.00$$

$$p_5=p(X<\infty)=p(Z<\frac{\infty-138.9}{19.38})=p(Z<\infty)=1$$

گروه	فشارخون سیستولیک	حد بالا	Z	$\Phi(Z)$	فراوانی تجمعی $100\times\Phi(Z)$	Fe_i
۰	کمتر از ۶۰	۶۰	–۴/۰۷	۰	۰	۰
۱	۶۰–۹۰	۹۰	–۲/۵۲	۰/۰۰۵۹	۰/۵۹	۰/۵۹
۲	۹۰–۱۲۰	۱۲۰	–۰/۹۸	۰/۱۶۳۵	۱۶/۳۵	۱۵/۷۶
۳	۱۲۰–۱۵۰	۱۵۰	۰/۵۸	۰/۷۱۹۰	۷۱/۹۰	۵۵/۵۵
۴	۱۵۰–۱۸۰	۱۸۰	۲/۱۲	۰/۹۸۳۴	۹۸/۳۴	۲۶/۴۴
۵	بیشتر از ۱۸۰	∞	∞	۱	۱۰۰	۱/۶۶

گروه	حد بالا	Fo_i	Fe_i	$\frac{(Fo_i - Fe_i)^2}{Fe_i}$
۱	۱۲۰	۱۵	۱۶/۷۶	۰/۱۸۵
۲	۱۵۰	۵۷	۵۵/۵۵	۰/۰۳۸
۳	∞	۲۸	۲۸/۱۰	۰/۰۰۰۳۵

$$\chi_0^2 = \sum_{i=1}^{4} \frac{(Fo_i - Fe_i)^2}{Fe_i} = 0.185 + 0.038 + 0.00035 = 0.223$$

۳- مقدار بحرانی

مقدار بحرانی نیز برابر است با $\chi^2_{\alpha,k-p-1}$ که طبق مسئله α برابر ۰/۰۵ و k که برابر با تعداد رسته هاست در این مثال برابر ۳ است در مورد p با توجه به اینکه در توزیع نرمال ما مجبور به برآورد μ و σ با استفاده از مشاهدات شدیم پس برابر ۲ است. پس داریم:

$$\chi^2_{\alpha,k-p-1} = \chi^2_{0.05,3-2-1} = \chi^2_{0.05,1} = 3.84$$

۴- تصمیم گیری

با توجه به اینکه ۳/۸۴> ۰/۲۲۳ بنابراین $\chi_0^2 < \chi^2_{0.05,1}$، یعنی مقدار عددی آماره آزمون در ناحیه بحرانی قرار نمی گیرد. لذا فرض صفر رد نمی شود. در نتیجه می توان گفت تطابق توزیع صفت فشارخون رادراین جامعه باتوزیع نظری نرمال قابل قبول است.