

فرادرس

فراتر از یک کلاس درس
www.faradars.org

مجموعه فرادرس های تحلیل داده ها با نرم افزار آماری Minitab

مدرس:

علی رجبی خانقاهی

مدرس فرادرس

کارشناسی ارشد مهندسی معدن

FaraDars.org

1- توزیع دو جمله ای

در این توزیع متغیر تصادفی X پیشامد تعداد موفقیت ها در n آزمایش می باشد به طوریکه در هر آزمایش با احتمال P می بریم و با احتمال $1-P$ می بازیم. تابع توزیع X به صورت زیر می باشد:

$$P(x) = \binom{n}{x} P^x (1-P)^{n-x} \quad x = 0, 1, 2, \dots, n$$

میانگین واقعی و واریانس برای متغیر تصادفی X با توزیع دوجمله ای به صورت زیر می باشد:

$$\mu = nP$$

$$\sigma^2 = nP(1-P)$$

در نظر داشته باشید که هرگاه تعداد نمونه زیاد و احتمال موفقیت در هر نمونه پایین باشد ($np < 5$)، برای محاسبه احتمال یک متغیر تصادفی با توزیع دوجمله ای از تقریب های پواسون و نرمال استفاده می کنند.

2- توزیع پواسون

در این توزیع متغیر تصادفی X پیشامد تعداد رویدادها در یک بازه زمانی یا مکانی می باشد به طوریکه λ متوسط تعداد رویدادها در همان بازه ی زمانی یا مکانی می باشد. تابع توزیع پواسون به صورت زیر است:

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

در صورتیکه X متغیر تصادفی با توزیع پواسون باشد، میانگین واقعی و واریانس این متغیر تصادفی به قرار زیر است:

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

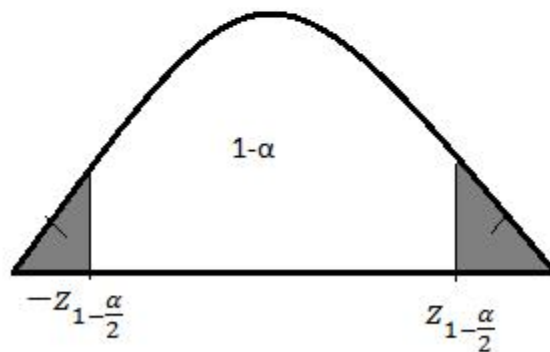
توزیع پواسون تنها تابع توزیعی می باشد که میانگین و واریانس آن با هم برابر است.

1- توزیع نرمال

توزیع نرمال حالت خاصی از توزیع گاما می باشد به طوریکه اگر X پیشامد یک متغیر تصادفی پیوسته با میانگین μ و انحراف معیار σ باشد، آن گاه تابع توزیع X به صورت زیر می باشد:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty \leq x \leq +\infty$$

از آنجاییکه برای محاسبه ی هر احتمالی باید از تابع توزیع فوق انتگرال گرفت، و هر انتگرالی از تابع فوق غیر قابل حل می باشد، پس از تغییر متغیر $z = \frac{x-\mu}{\sigma}$ برای محاسبه ی هر احتمالی استفاده می شود که به Z مقدار استاندارد شده گویند و در صورتیکه X دارای توزیع نرمال با میانگین μ و انحراف معیار σ باشد، آن گاه Z دارای توزیع نرمال با میانگین صفر و انحراف معیار یک می باشد. نمودار تابع چگالی احتمال نرمال به صورت زیر می باشد:



نمودار تابع چگالی احتمال نرمال

قسمت ج نیز احتمال $p(\bar{x} \leq 2.1)$ را خواسته است از آن جاییکه ضخامت نورد توزیع نرمال دارد پس میانگین نمونه این 9 نمونه نیز توزیع نرمال با میانگین 2 و انحراف معیار $\frac{0.1}{\sqrt{9}}$ دارد پس می توان برای حل قسمت ج احتمال زیر را حل کرد:

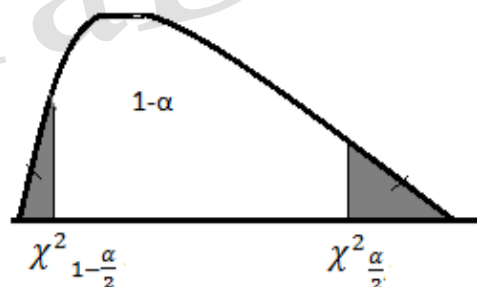
$$p\left(z \leq \frac{2.1 - 2}{0.03}\right) = p(z \leq 3)$$

2- توزیع مربع کای

در صورتیکه $X_1, X_2, \dots, X_n$ هر کدام توزیع نرمال با میانگین صفر و واریانس یک داشته باشند و از یک جامعه باشند، آن گاه مجموع مربعات آن ها توزیع کای دو با درجه آزادی n دارند. از آن جاییکه یک متغیر تصادفی کای دو با n درجه آزادی از n متغیر نرمال تشکیل شده است پس توزیع متغیر تصادفی مربع کای اختیاری نبوده و تابع چگالی آن حالت خاصی از توزیع گاما می باشد. اگر متغیر تصادفی X دارای توزیع کای دو با n درجه آزادی باشد پس تابع چگالی X به صورت زیر تعریف می شود:

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} \quad x \geq 0$$

که در رابطه فوق Γ معرف تابع گاما می باشد. نمودار زیر، نمودار تابع چگالی احتمال توزیع مربع کای با n درجه آزادی می باشد:



نمودار تابع چگالی احتمال توزیع مربع کای با n درجه آزادی

در صورتیکه X یک متغیر تصادفی مربع کای با n درجه ی آزادی باشد، آنگاه میانگین واقعی این متغیر تصادفی برابر n و واریانس آن به ازای n های بزرگ تر مساوی n برابر $2n$ می باشد.

یکی از کاربردهای اصلی توزیع مربع کای، این است که برای توزیع واریانس نمونه (S^2) در یک جامعه نرمال به کار می رود به طوریکه اگر یک نمونه ی n تایی از یک جامعه نرمال با انحراف معیار σ تهیه کنیم، آن گاه

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

برای درک بهتر رابطه فوق، فرض کنید نمونه های 5 تایی از یک جامعه نرمال با انحراف معیار 0.005 برداشته شده باشد. احتمال اینکه واریانس نمونه حداکثر 0.008^2 باشد به صورت زیر محاسبه می شود:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{4(0.008)^2}{0.005^2} = 10.24$$

$$p(\chi^2_{(4)} < 10.24) = 0.96$$

رابطه فوق بدین معنی می باشد که یک نمونه ی 5 تایی از یک جامعه نرمال تهیه می کنیم و احتمال اینکه حداکثر واریانس نمونه برابر 0.008^2 باشد، 0.96 می باشد.

در قسمت الف باید احتمال $p(s^2 \leq 1.1)$ را محاسبه کنیم از طرفی می دانیم که

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

پس

$$\frac{(8)s^2}{1} \sim \chi^2_{(8)}$$

در نتیجه می توان گفت برای محاسبه احتمال $p(s^2 \leq 1.1)$ ، $p(\chi^2_{(8)} \leq 8 * 1.1)$ را محاسبه کنیم.

تحلیل: مقدار X که به دست آمده مربوط به رابطه $\frac{(8)s^2}{1}$ می باشد پس s^2 با $15.50/8$ برابر است که مقدار آن 1.93 به دست خواهد آمد. یعنی اگر واریانس یک نمونه ی 9 تایی حداکثر برابر 1.93 باشد با احتمال 0.95 قابل کشف خواهد بود.

توزیع T

یکی از کاربردهای توزیع t برای بیان توزیع میانگین نمونه از یک جامعه نرمال وقتی که انحراف معیار مجهول است به کار می رود به صورتیکه :

$$\frac{(\bar{x} - \mu)\sqrt{n}}{s} \sim t(n - 1)$$

برای درک بهتر رابطه فوق، فرض کنید یک نمونه ی 4 تایی از یک جامعه نرمال با میانگین 5 انتخاب کردید و مقدار انحراف معیار نمونه به ترتیب برابر 0.5 به دست آمده است، حال احتمال اینکه میانگین نمونه حداکثر برابر 5.5 شود به صورت زیر محاسبه می شود

$$t = \frac{(5.5 - 5)\sqrt{4}}{0.5} = 2$$

$$p(t_3 \leq 2) = 0.93$$

احتمال فوق بدین معنی می باشد که اگر یک نمونه ی 4 تایی از یک جامعه نرمال تهیه کنیم که انحراف معیار آن مجهول است، احتمال اینکه میانگین نمونه این 4 نمونه حداکثر برابر 5.5 شود برابر با 0.93 می باشد.

یکی دیگر از کاربردهای توزیع t برای بیان توزیع اختلاف میانگین نمونه‌ی دو جامعه‌ی نرمال مستقل از هم وقتی که واریانس هر دو جامعه نامعلوم ولی مساوی می‌باشد به کار می‌رود به طوریکه:

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2)$$

در رابطه فوق n_x برابر اندازه نمونه‌ای می‌باشد که از جامعه x و n_y اندازه نمونه‌ای می‌باشد که از جامعه y تهیه کردیم و S_p برابر انحراف معیار مشترک دو جامعه x و y می‌باشد که از رابطه زیر به دست می‌آید:

$$S_p = \sqrt{\frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}}$$

در قسمت الف باید احتمال زیر را محاسبه کنیم:

$$p(\bar{x} \leq 1.5)$$

از آنجاییکه انحراف معیار فرآیند نامعلوم می‌باشد پس باید از رابطه زیر برای محاسبه احتمال فوق استفاده کنیم:

$$\frac{(\bar{x} - \mu)\sqrt{n}}{s} \sim t(n - 1)$$

در نتیجه

$$p\left(\frac{(\bar{x} - \mu)\sqrt{n}}{s} \leq \frac{(1.5 - \mu)\sqrt{n}}{s}\right) = p\left(t_9 \leq \frac{(1.5 - 1)\sqrt{10}}{0.5}\right) = p(t_9 \leq 3.16)$$

تحلیل: عدد X به دست آمده برای رابطه $\frac{(\bar{X}-1)\sqrt{10}}{0.5}$ می باشد پس خود $\bar{X}$ از رابطه $1 + \left(1.83 * \frac{0.5}{\sqrt{10}}\right)$ به دست می آید که برابر 1.2 خواهد بود، یعنی اگر میانگین نمونه عمر لامپ ها حداکثر 1.2 سال باشد با احتمال 0.95 قابل کشف خواهد بود.

در قسمت ج باید احتمال $P(\bar{x} - \bar{y} \leq \frac{1}{6})$ را محاسبه کنیم از طرفی میدانیم که

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2)$$

و از آنجاییکه

$$S_p = \sqrt{\frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}} = \sqrt{\frac{(9)0.5^2 + (15)(8/12)^2}{24}} = 0.6$$

پس:

$$P\left(\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \leq \frac{(\frac{1}{6}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}\right) = P\left(t_{24} \leq \frac{(\frac{1}{6}) - 0}{0.6 \sqrt{\frac{1}{10} + \frac{1}{16}}}\right) = P(t_{24} \leq 0.68)$$

FISHER DIS.

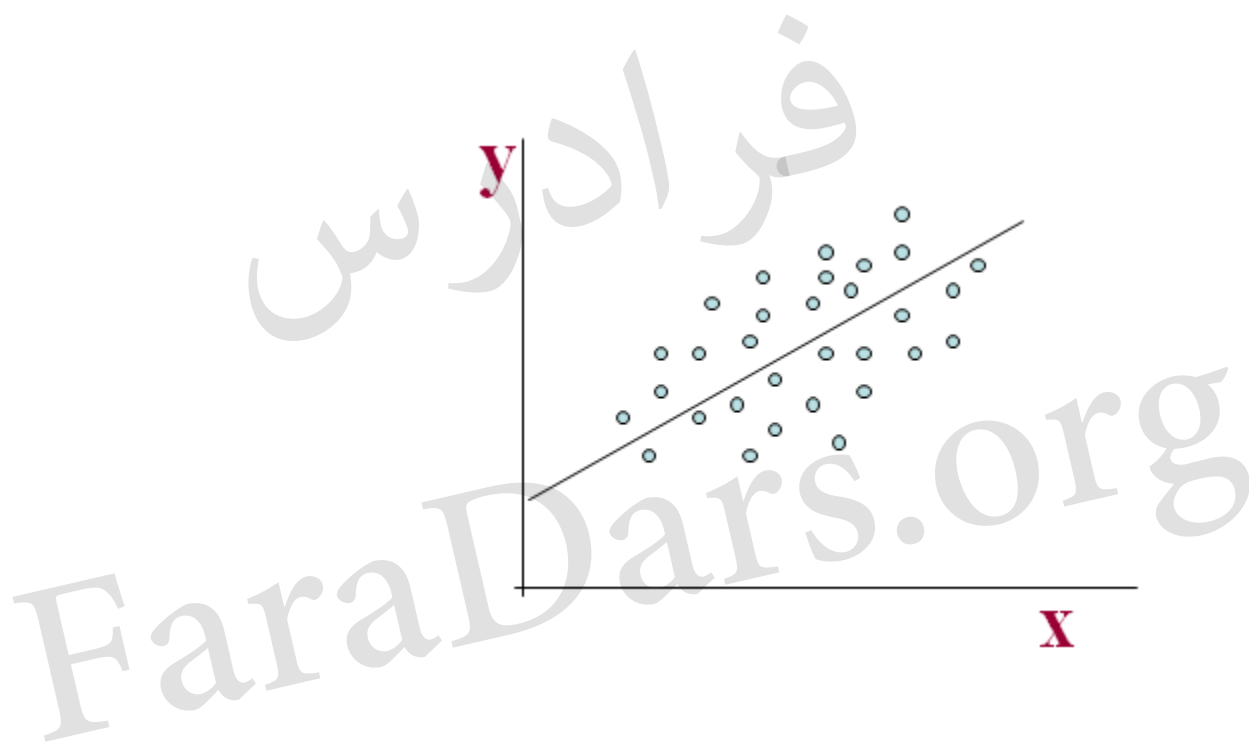
$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1)$$

دارای توزیع F با $n_1 - 1$ و $n_2 - 1$ درجه ی آزادی می باشد.

$$\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1)$$

پس می توان گفت :

$$p\left(\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \leq 1.5 * \frac{2}{1}\right) = p(F(15, 8) \leq 3)$$

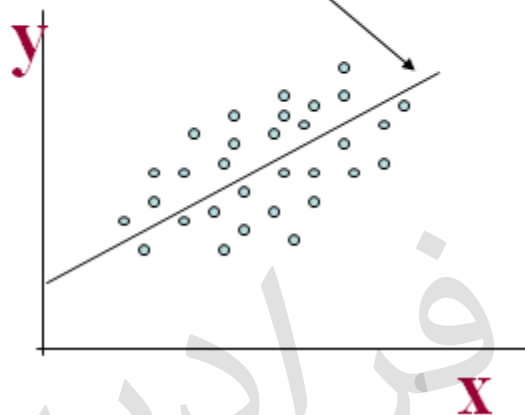


این فایل بر مبنای نکات مطرح شده در فرادرس «تحلیل داده ها با نرم افزار آماری Minitab» تهیه شده است.

<http://www.faradars.org/fvst94031>

$$y = A + Bx + \varepsilon$$

Random Error



FaraDars.org

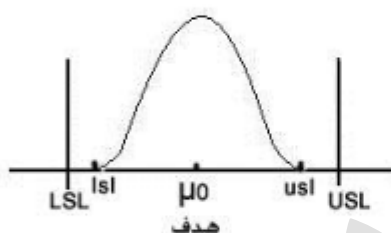
شاخص CP (شاخص بالقوه قابلیت فرآیند):

عبارت است از نسبت تفاوت حد بالا و پایین مشخصات فنی نسبت به تفاوت حد بالا و پایین فرآیند

$$(UCL - LCL = 6\hat{\sigma})$$

$$cp = \frac{USL - LSL}{6\hat{\sigma}}$$

اگر تفاوت USL-LSL از $6\hat{\sigma}$ بیشتر باشد، آنگاه:

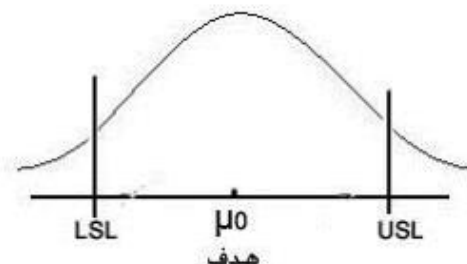


(خوب) $Cp > 1$

در این صورت فرآیند قادر به رعایت حدود مشخصات فنی می باشد.

اگر تفاوت USL-LSL از $6\hat{\sigma}$ کمتر باشد آنگاه:

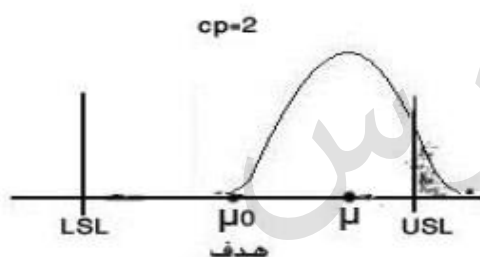
FaraDars.org



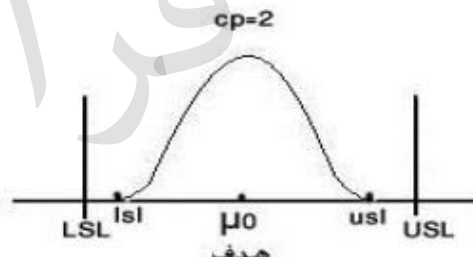
(بد) $C_p < 1$

در اینصورت فرآیند قادر به رعایت حدود مشخصات فنی نخواهد بود و ضایعات بالا می باشد.

شاخص C_p قابلیت فرآیند را بر حسب تغییرات $\hat{\sigma}$ بررسی می کند اما موقعیت فرآیند (یعنی جایگاه $\hat{\mu}$ میانگین فرآیند قرار میگیرد) در نظر نمی گیرد. به مثال های زیر توجه کنید:



شکل 2



شکل 1

در هر دو شکل $cp=2$ می باشد اما ضایعات در شکل 2 بسیار بالاست در حالیکه نسبت تفاوت

$USL - LSL$ نسبت به $6\hat{\sigma}$ در هر دو شکل یکی است اما همانطور که می بینیم میانگین فرآیند در شکل 1 و 2 متفاوت است اما شاخص cp نتوانسته موقعیت میانگین فرآیند $\hat{\mu}$ را از μ_0 (مقدار هدف) نشان دهد.

لذا برای حل این مشکل از شاخصی به نام cp_k (شاخص بالفعل قابلیت فرآیند) استفاده میکنند. که در آن k عاملی است که تفاوت میانگین فرآیند $\hat{\mu}$ را از μ_0 نمایش می دهد.

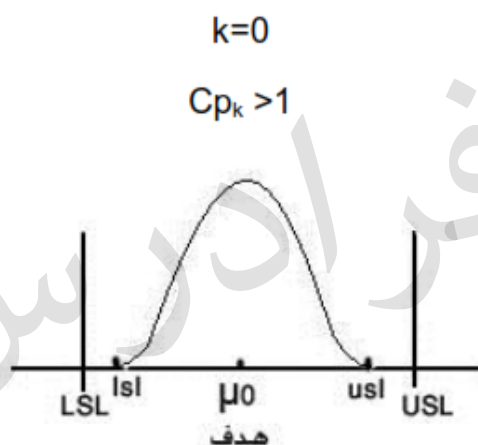
$$cp_k = \min \left\{ \frac{USL - \hat{\mu}}{3\hat{\sigma}} - \frac{\hat{\mu} - LSL}{3\hat{\sigma}} \right\} \quad \text{شاخص قابلیت فرآیند (بالفعل)}$$

$$cp_k = (1 - k)cp, \quad k = \frac{|\mu_0 - \hat{\mu}|}{\frac{USL - LSL}{2}}, \quad 0 \leq k \leq 1$$

IF	$k=0$	$\mu_0 = \hat{\mu}$	$\rightarrow$	$C_{pk}=C_p$
IF	$0 < k < 1$	$\mu_0 \neq \hat{\mu}$	$\rightarrow$	$C_{pk} < C_p$

شاخص cp_k نیز مانند cp اگر بزرگتر از يك باشد يعني فرآیند قادر به رعایت حدود مشخصات فني می باشد و بر عكس به مثال هاي زیر توجه کنید:

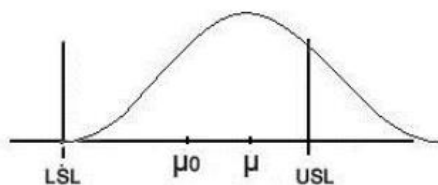
مثال (1)



مثال (2)

$$k \neq 0, \hat{\mu} > \mu_0$$

$$C_p, C_{pk} > 1$$



فلسفه استفاده از آزمون های فرض آماری

معمولا در هر جامعه آماری شاخص های واقعی که توصیف کننده جامعه باشند، مجهول می باشند. برای مثال شما اگر بخواهید میانگین واقعی قد آقایان را در یک جامعه به دست آورید باید قد هر یک از مردان جامعه را به دست آورده و میانگین واقعی ($E(X)$) آن را محاسبه کنید که این کار بسیار وقت گیر و هزینه بر است. بهترین راه نمونه گیری می باشد، شما با نمونه گیری از یک جامعه و استنباط آماری به وسیله ی نمونه ها می توانید در مورد کل جامعه نظر بدهید. یکی از بهترین و دقیق ترین روش های استنباط آماری، آزمون های فرض می باشد. در هر آزمون فرض آماری یک فرض در برابر خلاف آن فرض آزموده می شود. برای مثال مدیر یک مدرسه ادعا می کند که معدل دانش آموزان مدرسه اش برابر 18 می باشد حال شما با نمونه گیری به این نتیجه خواهید رسید که آیا مدیر مدرسه راست گفته است یا خیر؟ پس در هر آزمون فرض آماری، فرض صفر (H_0) که معمولا یک ادعا می باشد در برابر خلاف آن فرض (H_1) آزموده خواهد شد. در هر آزمون فرض به خاطر اینکه با استفاده از نمونه ها برای کل جامعه اظهار نظر می کنیم، دو نوع خطا وجود دارد که عبارتند از:

- خطای نوع I (α): احتمال اینکه فرض صفر رد شود، هنگامیکه صحیح است.
- خطای نوع II (β): احتمال اینکه فرض صفر قبول شود، هنگامیکه نادرست است.

مراحل انجام یک آزمون فرض

1. فرض صفر (H_0) و فرض یک (H_1) را تعیین کنید، در نظر داشته باشید فرض صفر همیشه شامل علامت مساوی می باشد مانند ($=$, $\leq$ و $\geq$)
2. نمونه های لازم را تهیه کرده و مقادیر آن را ثبت کنید. معمولا اندازه نمونه را با مشخص کردن احتمال کشف تغییر به دست می آورند که در مورد آن در فصل سوم و چهارم صحبت کردیم.
3. پس از به دست آوردن نمونه ها نوبت به آزمودن آن ها می باشد. روش های زیر از روش هایی می باشند که با استفاده از آن ها می توان تصمیم گرفت که فرض صفر را قبول یا رد کنیم.
- ا. یکی از این روش ها که در اکثر نرم افزارها نیز از آن استفاده می شود روش P-Value می باشد. P-Value یک احتمال از جنس α می باشد. در هر آزمون فرضی که P-Value بیشتر از α شود، فرض صفر قبول می شود در غیر این صورت فرض صفر رد می شود.

قدم سوم: در صورتیکه مشاهداتتان را در یک ستون وارد کردید Sample in columns را انتخاب کرده و ستون مورد نظرتان را وارد کنید، در این مثال ستون Values را وارد کنید و در قسمت Standard deviation انحراف معیار داده هایتان را وارد کنید که در این مثال برابر 0.2 می باشد. هم چنین اگر اطلاعات مسئله را در اختیار دارید گزینه Summarized data را انتخاب کرده و در قسمت Sample size اندازه نمونه و در قسمت Mean مقدار میانگین نمونه را وارد کنید. در قسمت Perform hypothesis test مقدار هدف یا مقدار میانگینی را که قصد دارید آن را آزمون کنید وارد کنید، در این مثال عدد 5 را در کادر مقابل Hypothesized mean وارد کنید. حال Options را انتخاب کرده و در قسمت Confidence level سطح اطمینان یا مقدار $1-\alpha$ را وارد کنید. در این مثال 90 را در کادر مقابل Confidence level وارد کنید و در قسمت Alternative نوع فرض یک را وارد کنید، در صورتیکه $H_1: \theta \neq \theta_0$ آنگاه Not equal و در صورتیکه $H_1: \theta < \theta_0$ آنگاه Less than و اگر $H_1: \theta > \theta_0$ آنگاه Greater than را انتخاب کنید در این مثال چون می خواهیم فرض $H_0: \mu = 5$ در برابر $H_1: \mu \neq 5$ بیازماییم پس Not equal را انتخاب کنید و OK را بزنید.

فرادرس
FaraDars.org

این فایل بر مبنای نکات مطرح شده در فرادرس
«تحلیل داده ها با استفاده از نرم افزار Minitab»
تهیه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر
مراجعه نمایید.

faradars.org/fvst94031