Block 4. Statistical inference

SOLVED EXAMPLES

Estimation

Example 4.1. You need to find out the distribution function for the response times for requests in a control centre. You know that the exponential distribution describes this situation accurately, but you do not know its parameter. Determine the moment estimator that provides the value of the distribution parameter.

Solution.

Take a random sample, $X_1, X_2, \cdots, X_n$, of response times from n requests received at the centre. As the exponential distribution has λ as its only parameter, you have to estimate just one parameter.

Equalise the first population moment $E(X)$ and the sample moment $\bar{X}$. Then, if you take into account that for the exponential distribution $E(X) = \frac{1}{\lambda}$; then, $\frac{1}{\lambda} = \bar{X}$ or $\lambda = \frac{1}{\bar{X}}$. The moment estimator of λ is $\hat{\lambda} = \frac{1}{\bar{X}}$.

Notice that the moment estimator of the parameter is determined with the inverse of the mean sample value. So, in this problem, you only need a set of sample values and then you calculate their arithmetic mean and inverse. You take the estimated value as the population parameter and, this way, determine the exponential distribution.

Example 4.2. A service company wants to analyse the time, in hours, the weekly equipment maintenance takes. The experts suggest establishing a probability model, because the maintenance times are random. They take a random sample of the times $X_1, X_2, \cdots, X_n$. They notice that the distribution has an asymmetrical bell shape, but they do not know its parameters. Determine the moment estimators that provide the values of the distribution parameters.

Solution.

As the distribution has an asymmetrical bell shape, you can assume a gamma distribution. If this assumption is true, you have to find the parameters α and β that characterise this distribution.

Use the following procedure to find the moment estimators:

1. Equalise the first population moment $E(X)$ and the second sample moment $\bar{X}$. If you take into account that for the gamma distribution it is true that $E(X) = \alpha\beta$; then, $\bar{X} = \alpha\beta$.
2. Equalise the second population moment $E(X^2)$ and the second sample moment $\frac{1}{n}\sum X_i^2$. If you take into account that for the gamma distribution it is true that $E(X^2) = \beta^2 \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \beta^2(\alpha+1)\alpha$; then, $\frac{1}{n}\sum X_i^2 = \beta^2(\alpha+1)\alpha$.
3. Solve the equations in steps 2 and 3:. $\bar{X}^2 = \alpha^2\beta^2$ and $\frac{1}{n}\sum X_i^2 = \bar{X}^2 + \beta^2\alpha$. Divide this equation by the penultimate one. You get: $\hat{\beta} = \frac{\frac{1}{n}\sum X_i^2 - \bar{X}^2}{\bar{X}}$. Then, $\hat{\alpha} = \frac{\bar{X}^2}{\frac{1}{n}\sum X_i^2 - \bar{X}^2}$.

Notice that the moment estimators of the parameters are determined with the expressions given by $\hat{\alpha}$ and $\hat{\beta}$, expressed in terms of sample values. As in Example 4.1, if you have numerical values in the problem, the estimated values are taken as the population parameters and you use these to determine the gamma distribution.

Example 4.3. You want to find out the distribution function of processes that obey the exponential distribution, as in Example 4.1, but where you do not know the parameter. Determine the maximum likelihood estimator that provides the value of the distribution parameter.

Solution.

There is another way to calculate the population parameter. Take a random sample $X_1, X_2, \cdots, X_n$ of an exponential distribution with the parameter λ. The likelihood function is a product of the probability density function,

$$f(x_1, x_2, \cdots, x_n; \lambda) = \left(\lambda e^{-\lambda x_1}\right)\left(\lambda e^{-\lambda x_2}\right)\cdots\left(\lambda e^{-\lambda x_n}\right) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

$$\ln[f(x_1, x_2, \cdots, x_n; \lambda)] = n\ln(\lambda) - \lambda \sum_{i=1}^n x_i.$$

Now derive the logarithm of λ and equalise this to 0, and you get:

$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$ or $\lambda = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}$.

The maximum probability estimator is $\hat{\lambda} = \frac{1}{\bar{X}}$. This method is identical to the method for moments; however, it is not an unbiased estimator $\left(E\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{E(\bar{X})}\right)$.

Example 4.4. In order to find the population distribution function in a given process, take a random sample $X_1, X_2, \cdots, X_n$ that has a normal distribution, but for which you do not know the parameters. Determine the maximum likelihood estimators that provide the values of the distribution parameters.

Solution.

The maximum likelihood function is expressed as follows:

$$f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}} e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}}$$

In this way,

$$\ln[f(x_1, x_2, \cdots, x_n; \mu, \sigma^2)] = \frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}.$$

Take the partial derivatives of $\ln f$ for μ and σ^2 to find the values μ and σ^2 that maximise the likelihood function. Then, equalise them to zero and solve the resulting equations. The maximum likelihood estimators are $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

The maximum likelihood estimator of σ^2 is not an unbiased estimator. Therefore, two different principles for estimating (maximum likelihood estimator and unbiased) provide two different estimators.

Example 4.5. To find the population distribution function in a given process, take a random sample $X_1, X_2, \cdots, X_n$ that has a Weibull distribution, but for which you do not know the parameters. Determine the maximum likelihood estimators that provide the values of the distribution parameters.

Solution.

The probability density function of the Weibull distribution is given by

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x\geq 0\\ 0, & x<0\end{cases}.$$

You find the likelihood and then the logarithm of that. For the latter, calculate the partial derivatives with respect to α and β, and equalise these to zero. You get:

$$\alpha=\left[\frac{\sum_{i=1}^{n}x_i^{\alpha}\cdot\ln(x_i)}{\sum_{i=1}^{n}x_i^{\alpha}}-\frac{\sum_{i=1}^{n}\ln(x_i)}{n}\right] \text{ and } \beta=\left(\frac{\sum_{i=1}^{n}x_i^{\alpha}}{n}\right)^{\frac{1}{\alpha}}.$$

The equations cannot be solved explicitly to provide the general formulae for the maximum likelihood estimations $\hat{\alpha}$ and $\hat{\beta}$. However, for each sample $x_1, x_2, \cdots, x_n$, the equations are solved by means of an iterative numerical procedure. The procedures for determining the moment pairs of α and β are complex.

Example 4.6. In Example 4.4, for the normal distribution, you determined the maximum likelihood estimator of μ and σ^2. These general expressions are given by $\hat{\mu}=\bar{X}$ and $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$. Determine the maximum likelihood estimator of the function $h(\mu,\sigma^2)=\sqrt{\sigma^2}=\sigma$.

Solution.

In order to get the maximum likelihood estimator of the function $h(\mu,\sigma^2)=\sqrt{\sigma^2}=\sigma$, replace the maximum likelihood estimator in the function : $\hat{\sigma}=\sqrt{\sigma^2}=\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]^{\frac{1}{2}}$.

The maximum likelihood estimator of σ is not the standard deviation of the sample S, although they are close, unless n is very low.

Example 4.7. The mean value of the random variable X with a Weibull distribution is $\mu=\beta\cdot\Gamma\left(1+\frac{1}{\alpha}\right)$,. What is the maximum likelihood estimation?

Solution.

If $\mu = \beta \cdot \Gamma\left(1 + \frac{1}{\alpha}\right)$, the maximum likelihood estimator is $\hat{\mu} = \hat{\beta}\,\Gamma\left(1 + \frac{1}{\hat{\alpha}}\right)$. Where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators of α and β.

$\bar{X}$ is not the maximum likelihood estimator of μ, although it is an unbiased estimator. For a high n, $\hat{\mu}$ is a better estimator than $\bar{X}$.

Example 4.8. Let X be a normal random variable with a mean of μ (with an unknown value) and a standard deviation of $\sigma = 2$. From a random sample (with replacement) of 25 values of X, you get a sample mean of $\bar{x} = 10$. Find the margin of error E for a 95% confidence interval for μ and determine the corresponding confidence interval. Interpret the result.

Solution.

You know that the random variable of the population X is normal, but one of its parameters, μ, is unknown. You have to estimate this parameter through a confidence interval inference procedure. Use the real sample observations, $x_1, x_2, \cdots, x_{25}$, which are the result of a random sample $X_1, X_2, \cdots, X_{25}$ of a normal distribution.

As X is normal, $\bar{X}$ is also normal. Therefore, in the expression $P(\mu - E \leq \bar{X} \leq \mu + E) = 0.95$, you standardise the random variable $\bar{X}$, because it is not the normal standard distribution.

Determining the confidence interval means that you know the margin of error, E, which satisfies the equation $P(\mu - E \leq \bar{X} \leq \mu + E) = 0.95$. As X is normal, $\bar{X}$ is also normal. As $\bar{X}$ is not the normal standard distribution, standardise it:

$$P(\mu - E \leq \bar{X} \leq \mu + E) = P\left(\frac{\mu - E - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu + E - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(\frac{-E}{\frac{2}{5}} \leq Z \leq \frac{E}{\frac{2}{5}}\right) = P\left(\frac{-E}{0.4} \leq Z \leq \frac{E}{0.4}\right) = 0.95.$$

As $\bar{X}$ is normal, Z is also normal. As Z is a standardised random variable, from the tables, you find that

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \Leftrightarrow P(0 \leq Z \leq 1.96) = \frac{0.95}{2} = 0.475.$$

The value 1.96 is the critical value of Z corresponding to a 0.95 probability. Therefore, $\frac{E}{0.4} = 1.96 \Rightarrow E = 0.4 \times 1.96 = 0.784$.

The confidence interval of 95% is $[\bar{x} - E, \bar{x} + E] = [10 - 0.784, 10 + 0.784] = [9.216, 10.784]$.

There is 95% confidence that the mean μ of X is a value within that interval; this means that, as $\bar{x}$ takes all the possible values of $\bar{X}$, 95% of all the intervals $[\bar{x} - 0.784, \bar{x} + 0.784]$will contain μ. Although the different random samples of size 25 can give different values of $\bar{x}$, the value of E is the same for each sample.

Example 4.9. A specialist company measures the zinc concentration in a river. A sample of the zinc concentration measurements, taken in 36 different locations, gives a mean concentration of $2.6\, g/ml$. If the population standard deviation is 0.3. Determine the 95% and 99% confidence intervals for the mean concentration of zinc in the river.

Solution.

The population random variable X has an unknown mean μ and a known standard deviation ($\sigma = 0.3$). As the sample size is $n = 36 \geq 30$, then, the distribution of $\bar{X}$ is approximately a normal distribution. With the 95% confidence level, and taking into account the fact that the point estimation of μ gives $\bar{x} = 26$, proceed as follows:

1. As the confidence interval is 95%, $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 0.95 = 1 - \alpha$. Therefore, $\alpha = 0.5$, and the critical value $z_{\alpha/2} = 1.96$ leaves an area 0.025 to its right (find this with the help of a table or software).
2. $E = \frac{z_{\alpha/2}\ \sigma}{\sqrt{n}} = \frac{1.96\ \ 0.3}{\sqrt{36}} = 0.098$.

3. The confidence interval is $[2.6 - 0.098, 2.6 + 0.098] = [2.50, 2.70]$.

Therefore, the 95% interval is the one whose population mean is in the interval $2.50 \leq \mu \leq 2.70$. Hence, 95% of the samples taken by the company would have a mean zinc concentration in the reported interval.

Note. You must analyse the confidence interval for a 99% confidence level. In that case, $2.47 \leq \mu \leq 2.73$.

Example 4.10. A sample of the temperature values determined at several points of a property provides the following results: $10.4, 9.6, 10.2, 10.2, 9.8, 10.0$ and 9.8℃. Find a 95% confidence interval for the mean of all the points on the property, assuming an approximately normal distribution.

Solution.

You must find a confidence interval for μ, where σ is unknown. $\bar{X}$ is distributed approximately as a normal distribution.

Let $x_1 = 10.4$, $x_2 = 9.6$, $x_3 = 10.2$, $x_4 = 10.2$, $x_5 = 9.8$, $x_6 = 10.0$, $x_7 = 9.8$. If you calculate the statistics: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}\sum_{i=1}^{7} x_i = 10$ and $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{7-1}\sum_{i=1}^{7}(x_i - 10)^2} = 0.283$. Then, apply the following procedure:

1. Find the critical value of the random variable t with 6 degrees of freedom that meets $P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 0.95$, using a table or software program. As $\alpha = 0.05$, the critical value $t_{\alpha/2} = 2.447 \approx 2.45$ leaves a 0.025 area to its right.

In the same way $P(0 \leq t \leq t_{\alpha/2}) = \frac{0.95}{2} = 0.475$. If you look in a table for the probability corresponding to 6 degrees of freedom $t^* \approx 2.45$ $t_{\alpha/2} \approx 2.45$.

2. The error, $E = \frac{t_{\alpha/2}\ s}{\sqrt{n}} = \frac{2.45\ \ 0{,}283}{\sqrt{7}} = 0.26$

3. The confidence interval: $[\bar{x} - E, \bar{x} + E] = [10 - 0.26, 10 + 0.26\ ] = [9.74, 10.26]$.

The 95% confidence interval for the mean temperature is $9.74 \leq \mu \leq\ 10.26$.

Therefore, 95% of the samples, where the temperature was measured, would show a temperature in the reported interval.

Example 4.11. In a random sample of 1000 businessmen in a region, 60% of them prefer to develop a particular technological line to implement in the production of certain materials. Determine a confidence interval for the proportion of all businessmen that prefer the use of that technological line with a 90% confidence level.

Solution.

The group of businessmen is divided into those who want to develop the technological line and those who do not. The proportion p of successes in the population is unknown. As the size of the sample is high ($1000 \geq 30$) and it is necessary that $\hat{p} = 0.60$, apply the following procedure:

1. As the significance level is 90%, find the critical value that satisfies $P(0 \leq Z \leq z_{\alpha/2}) = \frac{0.90}{2} = 0.45$. In this case, $z_{\alpha/2} = 1.65$.

2. $E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.65\sqrt{\frac{0.60\ (1-0.60)}{1000}} \approx 0.03$

3. $[\hat{p} - E, \hat{p} + E] = [0.60 - 0.03, 0.60 + 0.03] = [0.57, 0.63]$

The confidence interval for the proportion of all businessmen that prefer to develop the technological line with a 90% confidence level is $0.57 \leq p \leq\ 0.63$. What does this mean?

Example 4.12. Steel rods are cut randomly. The final lengths represent a random variable with a normal distribution, and the values taken from a random sample are: $55, 65, 82, 48, 55, 75, 70, 62$. Calculate the confidence interval where the variability of the rod lengths is found with a 90% probability.

Solution.

Let X be the random variable for the final lengths. X is distributed as a normal distribution, $N(\mu, \sigma)$. The parameter μ is unknown and you have to estimate the variance of X. In this situation, the recommended statistic is $\chi^2_{n-1} = \frac{(n-1)\,S^2}{\sigma^2}$. The curve that describes the Chi-squared random variable is not symmetric and the confidence interval does not have the usual shape $[s^2 - E, s^2 + E\ ]$.

You obtain the values of X, $x_1, x_2, \cdots, x_n$, from a random sample with a size of n, i.e., $x_1 = 55$, $x_2 = 65$, $x_3 = 82$, $x_4 = 48$, $x_5 = 55$, $x_6 = 75$, $x_7 = 70$, $x_8 = 62$. You get the value of the sample mean from $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}\sum_{i=1}^{8} x_i = 64$, and the sample variance from $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - 64)^2 = 129.14$. Then, you must apply the following procedure:

You find the critical values of χ^2_{n-1} in a Chi-square table, for 7 degrees of freedom, or using a software program. With a significance level of 90% for the variance of X, $P(\chi^2_{n-1} \leq a) = \frac{1-0.90}{2} = 0.05$ for $a = 2.17$ and $P(\chi^2_{n-1} \leq b) = \frac{1+0.90}{2} = 0.95$ for $b = 14.1$.

The confidence interval is $\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right] = \left[\frac{(8-1)\ 129.14}{14.1}, \frac{(8-1)\ 129.14}{2.17}\right] = [64.1, 416.6]$. What does the result mean?

Problems with two populations

Example 4.13. The weights of two mouse species are known to be normally distributed. X-type mice have a mean weight of $28\ g$ and a standard deviation of $\sigma_X = 3\ g$; Y-type mice have a mean weight of $28\ g$ and a standard deviation of $\sigma_Y = 2\ g$. The aim is to increase the mean weight of each species by means of a diet. The mean weights of a sample of the X-type mice are: 29, 28, 30, 31, 26, 32, 25, and 34. And for the Y-type are: 27, 31, 30, 28, 29, 25, 31, 30, 29, and 26. Find the 90% confidence interval for $\mu_X - \mu_Y$ corresponding to the mice with the new diet.

Solution.

X and Y are independent random variables with known σ_X and σ_Y. $\bar{X}$ and $\bar{Y}$ are distributed approximately as a normal distribution. This case corresponds to the confidence interval of the difference of the means with known standard deviations and is based on the statistic Z, whose expression is $\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}} \to N(0.1)$.

You obtain the values of X and Y from random samples with sizes m and n, respectively, i.e., x_1, $x_2, \cdots, x_m$ and $y_1, y_2, \cdots, y_n$. The respective sample values are: $x_1 = 29$, $x_2 = 28$, $x_3 = 30$, $x_4 = 31$, $x_5 = 26$, $x_6 = 32$, $x_7 = 25$, $x_8 = 34$ and $y_1 = 27$, $y_2 = 31$, $y_3 = 30$, $y_4 = 28$, $y_5 = 29$, $y_6 = 25$, $y_7 = 31$, $y_8 = 30$, $y_9 = 29$, $y_{10} = 26$. You obtain the values of the sample means from $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \frac{1}{8}\sum_{i=1}^{8} x_i = 29.38$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}\sum_{i=1}^{10} y_i = 28.60$. Then, you must apply the following procedure:

1. Using a table of the standardised normal distribution (or a software program), find the critical value $z_{\alpha/2}$ of the standardised normal random variable Z for which $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 0.9$. The value found is $z_{\alpha/2} = 1.65$.

2. The error: $E = z^* \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} = 1.65 \sqrt{\frac{9}{8} + \frac{4}{10}} \cong 2.04$

3. Confidence interval: $[\bar{x} - \bar{y} - E, \bar{x} - \bar{y} + E] = [29.38 - 28.60 - 2.04, 29.38 - 28.60 + 2.04] = [-1.26,\ 2.82]$.

As 0 is within the confidence interval, you do not have enough evidence at 90% that each mean weight is higher with the new diet.

Example 4.14. Consider that, in Example 4.13, the random weights X and Y of the two mouse species would have a standard deviation of $2.5\ oz$ after the new diet. Find the confidence interval for $\mu_X - \mu_Y$ with a 90% confidence level, assuming that the new standard deviations are unknown but equal.

Solution.

X and Y are independent random variables with unknown σ_X and σ_Y. $\bar{X}$ and $\bar{Y}$ are distributed approximately as a normal distribution. This case corresponds to the confidence interval of the difference of the means with unknown but equal standard deviations. The characteristic statistic t, whose expression is $\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \rightarrow t_{m+n-2}$; where $s_p = \sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}$.

You obtain the values of X and $Y$from random samples with sizes m and n, respectively. The sample values $x_1, x_2, \cdots, x_m$ and $y_1, y_2, \cdots, y_n$ are given in Example 4.13. The values of the sample means are $\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 29.38$ and $\bar{y}\frac{1}{10}\sum_{i=1}^{10} y_i = 28.60$ (calculated in Example 4.13). You obtain the sample variances from $s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - 29.38)^2 = 9.125$ and $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{10-1}\sum_{i=1}^{10}(y_i - 29.38)^2 = 4.267$. The joint estimator $s_p = \sqrt{\frac{(8-1)\ 9.125 + (10-1)\ 4.267}{8+10-2}} = 2.53$. Then, you must apply the following procedure:

1. With the help of a distribution table t, with 16 degrees of freedom, (or a software program), find the critical value $t_{\alpha/2}$ of the random variable t for which $P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 0.9$. The value found is $t_{\alpha/2} = 1.75$.

2. The error: $E = t_{\alpha/2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 1.75 \sqrt{\frac{1}{8} + \frac{1}{10}} = 2.10$

3. Confidence interval: $[\bar{x}-\bar{y}-E, \bar{x}-\bar{y}+E] = [29.38-28.60-2.10, 29.38-28.60+2.10] = [-1.32,\ 2.88]$.

Note. You must analyse the results.

Example 4.15. Consider Example 4.13. Find the confidence interval for $\mu_X - \mu_Y$ with a 90% confidence level, assuming that $\tau_k = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{s_X^2}{8}+\frac{s_Y^2}{10}}}$ is a random variable t.

Solution

The given statistic corresponds to the difference of the means with unknown standard deviations. As $m \geq 5$, $n \geq 5$, $\bar{X}$, and $\bar{Y}$ are normally distributed, these can be approximated by means of a random variable t. Then, calculate the number of degrees of freedom by means of

$$k = \frac{\left(\frac{s_X^2}{m}+\frac{s_Y^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_X^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{s_Y^2}{n}\right)^2} = \frac{\left(\frac{9.125}{8}+\frac{2.476}{10}\right)^2}{\frac{1}{8-1}\left(\frac{9.125}{8}\right)^2+\frac{1}{10-1}\left(\frac{2.476}{10}\right)^2} = 11.93$$

The number of degrees of freedom is the higher integer, $[k]$, that satisfies the condition $[k] \leq k$. Then, τ has 11 degrees of freedom, i.e., τ_{11}.

Apply the following procedure:

1. With the help of a distribution table t, with 11 degrees of freedom, (or a software program), find the critical value $t_{\alpha/2}$ of the random variable t for which $P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 0.9$. The value found is $t_{\alpha/2} = 1.80$.

2. The error: $E = t_{\alpha/2} s_p \sqrt{\frac{1}{m}+\frac{1}{n}} = 1.80\sqrt{\frac{9.125}{8}+\frac{4.276}{10}} = 2.25$

3. Confidence interval: $[\bar{x}-\bar{y}-E, \bar{x}-\bar{y}+E] = [29.38-28.60-2.25, 29.38-28.60+2.25] = [-1{,}147,\ 3.03]$.

Note. You must analyse the results.

Example 4.16. In a random sample of 50 teachers from the Spanish region of Aragon, 40 support the implementation a pedagogical model; and 25 out of 48 from the Navarra region agree to this, too. Find the 95% confidence interval for $p_X - p_Y$, where p_X and p_Y are the proportions that supportthe implementation of the pedagogical model in Aragon and Navarra, respectively.

Solution

The size of the teacher samples from Aragon and Navarra are associated to n_x and n_y, respectively. The samples satisfy $n_x \geq 30$ and $n_y \geq 30$, and you obtain a value $\hat{p}_i$ of $\hat{P}_i$ (where $i = x, y$) from a random sample with a size $n_i \geq 30$. This case corresponds to the confidence interval of the difference of proportions, $\hat{p}_x - \hat{p}_y$. The sample proportions are: $\hat{p}_x = \frac{40}{50} = 0.8$ and $\hat{p}_y = \frac{25}{48} = 0.52$. The standard deviation of the sample is $\sigma_{\hat{p}_x - \hat{p}_y} = \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = \sqrt{\frac{0.8\,(1-0.8)}{50} + \frac{0.52(1-0.52)}{48}} = 0.09$. Next, apply the following steps:

1. This gives the critical value $z_{\alpha/2}$ of the standardised normal random variable Z, where $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = \gamma \Leftrightarrow P(0 \leq Z \leq z_{\alpha/2}) = \frac{\gamma}{2}$. With the help of a table of the standardised normal distribution or software, find the critical value $z_{\alpha/2}$: such as $P(0 \leq t \leq z_{\alpha/2}) = \frac{0.95}{2} = 0.475$, $z_{\alpha/2} = 1.96$.

2. The error: $E = z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = 1.96 \times 0.09 = 0.18$

3. Confidence interval: $[\hat{p}_x - \hat{p}_y - E, \hat{p}_x - \hat{p}_y + E] = [0.8 - 0.52 - 0.18,\ 0.8 - 0.52 + 0.18] = [0.10,\ 0.46]$.

As zero is not within the confidence interval, the sample evidences that $p_x > p_y$, i.e., the proportion of teachers who favour the pedagogical model is higher in Aragon than in Navarra.

Example 4.17. In a study, two random samples have been selected independently to study the variability in the measurement of a thermal characteristic in a process carried out in two factories. In the first sample, the values 32, 40, 25, 31, 24, 28 are recorded; in the second, the values are 15, 14, 18, 12, 20, 16, 17, 16. It is assumed that both samples are normally distributed. Find the 90% confidence interval for σ_X/σ_Y.

Solution.

In this case, it corresponds to a confidence interval for the ratios of the variances with unknown population means. Let X and Y be the random variables associated to the first and second samples, respectively. You know that both random variables are independent and their sample values are: $x_1 = 32, x_2 = 40, x_3 = 25, x_4 = 31, x_5 = 24, x_6 = 28$ and $y_1 = 15, y_2 = 14, y_3 = 18, y_4 = 12, y_5 = 20, y_6 = 16, y_7 = 17, y_8 = 16$.

The sample measurements are $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \frac{1}{6}\sum_{i=1}^{6} x_i = 30$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{8}\sum_{i=1}^{8} y_i = 16$. The sample variances are: $s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 = \frac{1}{6-1}\sum_{i=1}^{6}(x_i - 30)^2 = 34$ and $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{8-1}\sum_{i=1}^{8}(y_i - 16)^2 = 6$. You must use the following steps:

1. You find the critical values of F, F_1^*, and F_2^*, which meet $P[F(m-1, n-1) \leq F_1^*] = \frac{1+\gamma}{2}$ and $P[F(n-1, m-1) \leq F_2^*] = \frac{1+\gamma}{2}$. With the help of the distribution table F from the appendix (or a software program), you find that, for $P[F(6-1, 8-1) \leq F_1^*] = \frac{1+0.90}{2}$, you get $P[F(5,7) \leq F_1^*] = 0.95$, and the critical value is $F_1^* = 3.97$. Similarly, for $P[F(7,5) \leq F_2^*] = 0.95$, the critical value is $F_2^* = 4.88$.

2. The confidence interval of the variance ratios is $\left[\frac{1}{F_1^*}\frac{s_X^2}{s_Y^2}, F_2^*\frac{s_X^2}{s_Y^2}\right] = \left[\frac{1}{3.97}\frac{34}{6}, 4.88\frac{34}{6}\right] = [1.43, 27.65]$. Then, the confidence interval of the standard deviation ratios is $\left[\sqrt{1.43}, \sqrt{27.65}\right] = [1.20, 5.26]$.

As the confidence interval does not contain the value one, you cannot accept at 90% confidence that $\sigma_X/\sigma_Y = 1$. This means that the variabilities of the thermal characteristic are different when measurements are taken in the two factories; i.e., $\sigma_X \neq \sigma_Y$.

Contrasting a population

Example 4.18. (Statement modified from: César Pérez López, "Estadística (Problemas resueltos y aplicaciones", chapter 8, Ed. Pearson Prentice Hall, 2003**)**. A technician suspects that a communication switchboard receives at least 42 calls per day. To verify this, he makes a check. Over 10 days selected at random, he notices that there is a mean of 40 calls with 3.5% spread. The technician assumes normality with a variance of 16 in the distribution of calls and performs a comparison with a significance level of $\alpha = 0.05$. Is the initial assumption correct?

Solution.

In this problem, you want to verify an assumption about the switchboard receiving at least 42 calls per day (the number of calls is a random variable and is represented by X). This fact allows you to use the assumption as a conjecture and a hypothesis test to verify the initial assumption. So, proceed as follows:

You use a one-tailed test for the population mean μ_X. The reason for doing this is that it says "at least", and this slightly modifies the expression of H_0, as shown in the table. Then, you pose the **null hypothesis** and the **alternative hypothesis**:

H_0: $\mu_X \geq \mu_0 = 0.42$ (Null hypothesis)

H_a: $\mu_X < \mu_0 = 0.42$ (Alternative hypothesis)

In this case, you want to compare μ_X with a known σ_X^2. Therefore, you use the statistic $Z = \frac{\bar{X}-\mu_X}{\frac{\sigma_X}{\sqrt{n}}} \rightarrow N(0.1)$.

Next find the **value of the statistic** Z, under H_0. This is $z_0 = \frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}} = \frac{0.4-0.42}{\frac{4}{\sqrt{10}}} = -0.0158$.

With the help of a standardised normal distribution table or a software program, find the **critical region** corresponding to H_a with the given significance level of α. The critical region is delimited by the **critical value** z_α of $N(0.1)$, which under the given significance level has the value $z_{0.05} = 1.64 \implies$. The critical region of the comparison is $Z < -1.64$.

As the value of Z under H_0 is outside the critical region $(-0.0158 > -1.64)$, accept the H_0 that the switchboard receives at least 42 calls per day.

Example 4.19. (Statement modified from: César Pérez López, "Estadística (Problemas resueltos y aplicaciones", chapter 8, Ed. Pearson Prentice Hall, 2003**)**. In a metallurgical company, there is a quality control that checks the hardness of the metal sheets they manufacture. In the control process, there are two procedures, with different data and which have different goals. For the first, 100 measurements of the hardness are taken and it gives: a mean of 10 and a quasi standard deviation of one. Can you compare, with a 95% confidence level, that the quality hardness is normally distributed with a mean of 10.3? For the second, 20 measurements are

taken, giving a mean of 10 Can you assure with a 95% confidence level that the hardness quality is normally distributed with a mean of 14 and a variance of 9?

Solution.

For quality control purposes, in this problem you want to confirm if the hardness quality of the metal sheets is normally distributed. The measurements taken behave randomly, and you need a hypothesis test to make any conjecture about the hardness quality being normal.

In the **first procedure**, you use a two-tailed test for the population mean μ_X. The reason for doing this is that you want the mean to be 10.3.

H_0: $\mu = \mu_0 = 10.3$

H_a: $\mu \neq \mu_0 = 10.3$

You assume that the hardness quality is a random variable X that is distributed according to a normal distribution, but with an unknown variance σ_X^2. If $X \equiv N(\mu_0, \sigma_X)$, the standard deviation is unknown and H_0 is true; then, $t = \frac{\bar{X}-\mu_X}{S/\sqrt{n}} \rightarrow t_{n-1}$, i.e., the comparison estimator under the null hypothesis is distributed as a t of Student's t-test with $n-1$ degrees of freedom. Its value, based on the sample under H_0, is $t_0 = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{10-10.3}{\frac{1}{\sqrt{100}}} = -3$. Here, μ_0 is considered the value of the population mean.

The critical value $t_{\frac{\alpha}{2}, n-1}$ of the Student's t-test is $t_{0.025,\,99} = 1.984 \implies$. The critical region of the comparison is $|t| > 1.984$.

As for the value of t under H_0 it is true that $|t_0| = 3$, and this is inside the critical region $(3 > 1.984)$, so you reject the H_0 that the quality hardness is distributed with $N(10.3)$.

In the **second procedure**, you use a two-tailed test for the population mean μ_X. Now, you know the variance σ_X^2. You assume that there is normality and pose the two hypothesis:

H_0: $\mu = \mu_0 = 14$

H_a: $\mu \neq \mu_0 = 14$

You know that the statistic for this case is $Z = \frac{\bar{X}-\mu_X}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0.1)$. The value of the statistic Z under H_0, $z_0 = \frac{10-14}{\frac{3}{\sqrt{20}}} = -4.22$.

The critical value $z_{\frac{\alpha}{2}}$ of $N(0.1)$, with $\alpha = 0.05$, is $z_{0.025} = 1.96 \Longrightarrow$; the critical region of the comparison is $|Z| > 1.96$.

As the value of $|Z|$, under H_0, is outside the critical region $(|Z| = 4.22 > 1.96)$, you reject the H_0 that the quality hardness is distributed with $N(14.9)$.

Example 4.20. The turnover of a company is normally distributed with $N(\mu_X, \sigma_X) = N(75, 8)$. Over the last period, the turnover has shown signs of decreasing and greater variability. They decide to check the situation and take a sample of 41 values of the turnover, with $\bar{x} = 73$ and $s = 9.6$. If you select a significance level of 0.05 to compare the situations, should the company be worried?

Solution.

In this problem, you want to verify whether the signs of decrease and greater variability are true. To check this, compare the data available in the statement. To do this, proceed as follows:

The turnover is a random variable X distributed with a normal distribution. As the turnover shows signs of decreasing, take the values of the mean and the standard deviation of the sample to make the comparison. Therefore, for the one-tailed test of σ^2 with a significance level of $\alpha = 0.05$, pose the null hypothesis and the alternative hypothesis:

$H_0: \sigma_X^2 = \sigma_0^2 = 64$

$H_a: \sigma_X^2 > \sigma_0^2 = 64$

If H_0 is true, the comparison statistic is a Chi squared random variable $\frac{(n-1)\,S^2}{\sigma_X^2} \rightarrow \chi_{n-1}^2$ with $n - 1$ degrees of freedom. In this case, you have 40 degrees of freedom. The value of the statistic based on the sample under the null hypothesis is $\chi_0^2 \equiv \frac{(n-1)\,s^2}{\sigma_0^2} = \frac{(41-1)\,(9.6)^2}{64} = 57.6$.

For $H_a: \sigma_X^2 > 64$ with $\alpha = 0.05$, the critical region is comprised of the values $\chi_0^2 > \chi_{\alpha,n-1}^2$, where $\chi_{\alpha,n-1}^2 > 0$ is a value that satisfies $P\left(\chi_{\alpha,n-1}^2 > \chi_0^2\right) = \alpha$. In this problem, using a Chi-square table (or a software program), you find that $\chi_{\alpha,41-1}^2 = 55.8$.

The comparison value $\chi_0^2 = 57.6 > 55.8 = \chi_{\alpha,40}^2$ and is inside the critical region. Therefore, the testing is significant at a level of 0.05 and $H_0: \sigma_X^2 = 64$ is rejected at this level.

Contrasting two populations

Example 4.21. The weights of two mouse species are known to be normally distributed. X-type mice have a mean weight of $28\ g$ and a standard deviation of $\sigma_X = 3\ g$; Y-type mice have a mean weight of $28\ g$ and a standard deviation of $\sigma_Y = 2\ g$. The aim is to increase the mean weight of each species by means of a diet. The mean weights of a sample of the X-type mice are: 29, 28, 30, 31, 26, 32, 25, and 34. And for the Y-type are: 27, 31, 30, 28, 29, 25, 31, 30, 29, and 26. Are there enough criteria to say that, with the new diet, the mean weights are equal with a significance level of 0.1?

Solution.

In this problem, you want to verify an assumption that the mean weights of both mouse species are the same with the new diet (the weights are random variables and are represented with X and Y). Then, you can make an assumption and use a hypothesis test to verify the equality of the weights. So, proceed as follows:

Use a two-tailed test for the population mean $\mu_X - \mu_Y$. The reason for doing this is that you want to find if the mean weights are equal. Then, you pose the **null hypothesis** and the **alternative hypothesis**:

$H_0: \mu_X = \mu_Y$

$H_a: \mu_X \neq \mu_Y$

X and Y are independent random variables with known σ_X and σ_Y. $\bar{X}$ and $\bar{Y}$ are distributed approximately as a normal distribution. If H_0 is true, the testing statistic $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}$ is approximately the standardised normal random variable, whose testing value is $z_0 = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}$. In Example 4.13, you calculated the values of the sample means, and their results are $\bar{x} = 29.38$ and $\bar{y} = 28.60$. So, $z_0 = \frac{(29.38-28.60)}{\sqrt{\frac{3^2}{8}+\frac{2^2}{10}}} = 0.63$.

For$H_a: \mu_X \neq \mu_Y$, the critical region is comprised of the points that satisfy the condition $P(Z \leq -z_{\alpha/2}) + P(Z \geq z_{\alpha/2}) = \alpha$ or, equivalently, $P(Z \geq z_{\alpha/2}) = \alpha/2$. In this case, $P(Z \geq z_{\alpha/2}) = 0.1/2 = 0.05$. With the help of a standardised normal distribution table or software,

determine the critical value; this value is $z_{\alpha/2} = 1.65$. The comparison value z_0 must be less than the critical value $z_{\alpha/2}$, i.e., $|0.63| < 1.65$.

As $|z_0| < z_{\alpha/2}$, the H_0 is not rejected at the significance level of 0.1. With the new diet, the mean weights are not different for the two populations.

Note. To verify whether H_0 is rejected or not, you can determine the confidence interval for $\mu_X - \mu_Y$. In Example 4.13, you determined the confidence interval in a similar problem. As the value 0 is within the estimated interval, i.e., confidence interval $\mu_X - \mu_Y = 0 \in$, there is no evidence for rejecting H_0 for the 90% confidence level. So, use a hypothesis test with a significance level of 0.1 to see whether H_0 is accepted or not. However, if the value of the confidence interval is $0 \notin$, then you can reject H_0. This note must not be ignored.

Example 4.22. Consider that, in Example 8.4, the random weights X and Y of the two mouse species would have a standard deviation of $2.5\ oz$ after the new diet. Are there enough criteria to say that, with the new diet, the mean weights are the same with a significance level of 0.1?

Solution.

You want to verify an assumption that the mean weights of the two mouse species are the same with the new diet (the weights are random variables and are represented with X and Y). Make a conjecture and use a hypothesis test to verify the equality of the weights.

Use a two-tailed test for the population mean $\mu_X - \mu_Y$. The **null hypothesis** and the **alternative hypothesis** are:

$H_0: \mu_X = \mu_Y$

$H_a: \mu_X \neq \mu_Y$

X and Y are independent random variables with unknown σ_X and σ_Y. $\bar{X}$ and $\bar{Y}$ are distributed approximately as a normal distribution. If H_0 is true; the comparison statistic $t = \frac{(\bar{X}-\bar{Y})}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$ with $m + n - 2$ is approximately the standardised normal random variable, whose testing value is $t_0 = \frac{(\bar{x}-\bar{y})}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$; where $s_p = \sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}$. In Example 4.14, you calculated the values of the sample means and variances. The results are $\bar{x} = 29.38$, $\bar{y} = 28.60$, $s_X^2 = 9.125$, and $s_Y^2 =$

4.267. The joint estimator is $s_p = \sqrt{\frac{(8-1)\ 9.125 + (10-1)\ 4.267}{8+10-2}} = 2.53$. Therefore, $t_0 = \frac{(29.38-28.60)}{2.53\sqrt{\frac{1}{8}+\frac{1}{10}}} = 0.65$.

For$H_a: \mu_X \neq \mu_Y$, the critical region is comprised of the points that satisfy the condition $P(t \leq -t_{\alpha/2}) + P(t \geq t_{\alpha/2}) = \alpha$ or, equivalently, $P(t \geq t_{\alpha/2}) = \alpha/2$. In this case, $P(t \geq t_{\alpha/2}) = 0.1/2 = 0.05$. With the help of a Student's t-test table or software, determine the critical value; the value is $t_{\alpha/2} = 1.75$. The comparison value t_0 is less than the critical value $t_{\alpha/2}$, i.e., $|0.65| < 1.75$. The H_0 is not rejected at the significance level of 0.1.

With the new diet, the mean weights are not different for the two populations. The H_0 is not rejected at the significance level of 0.1. With the new diet, the mean weights are not different for the two populations.

Example 4.23. The oxide layers on semiconductor sheets are etched in a gas mixture. Sheet thickness variability is a critical feature. Two different gas mixtures are studied to find if one is better than the other for reducing the variability of the oxide thickness. Twenty sheets are etched with each gas, and the standard deviations of samples of oxide thickness are $s_1 = 1.96$Å and $s_2 = 2.13$Å, respectively. Is there any evidence that one of the gases is better at achieving less variability than the other at a significance level of 0.05?

Solution.

You want to find if there is any evidence that one of the gases is better than the other (the weights are random variables and are represented with X and Y). Make a conjecture and use a hypothesis test to verify the equality of the weights.

Use a two-tailed test for the population variance σ_X^2/σ_Y^2. The **null hypothesis** and the **alternative hypothesis** are:

$H_0: \sigma_X^2 = \sigma_Y^2$

$H_a: \sigma_X^2 \neq \sigma_Y^2$

X and Y are normal independent random variables with unknown σ_X and σ_Y. If H_0 is true, the comparison statistic $F_{m-1,n-1} = \frac{S_X^2}{S_Y^2}$ is approximately the standardised normal random variable, whose testing value is $F_0 = \frac{s_X^2}{s_Y^2} = \frac{1.96^2}{2.13^2} = 0.85$.

For $H_a: \sigma_X^2 \neq \sigma_Y^2$, the critical region comprises the sample values $\frac{S_Y^2}{S_X^2} = F_{n-1,m-1} \geq F_0$ or $\frac{S_X^2}{S_Y^2} = F_{m-1,n-1} \geq F_0$ that satisfy the conditions $P\left(F_{n-1,m-1} \leq F_{\alpha/2}\right) = \alpha/2$ and $P\left(F_{m-1,n-1} \leq F_{1-\alpha/2}\right) = \alpha/2$. As $n = m$, $P\left(F_{20-1,20-1} \leq F_{\alpha/2}\right) = 0.05/2$. With the help of an F distribution table or software, determine the critical value; this is $F_{\alpha/2} = 1.75$. The comparison value t_0 is less than the critical value $t_{\alpha/2}$, i.e., $|0.65| < 1.75$. The H_0 is not rejected at the significance level of 0.05.

Example 4.24. In a city, the aim is to prove that the proportion of families that have completely paid for their home is 20%. To do this, they take a sample of 800 families and notice that the proportion of families that have completely paid for their home is 18%. Is the hypothesis to be tested consistent with the result obtained from the sample with a confidence margin of 95%?

Solution.

In this problem, you want to verify whether the proportion of families that own a fully paid for home is 20%. You can use a hypothesis test for the proportions ($\hat{p} = \frac{x}{n}$ is the estimated proportion of the number of times that a Bernoulli event happens out of n repetitions of an experiment –x is the number of families that own fully paid for homes).

Use a one-tailed test for the proportion. Then, pose the **null hypothesis** and the **alternative hypothesis**:

H_0: $p = 20$ (fully paid for homes)

H_a: $p > 20$

In this case, you want a comparison between the proportion and the statistic $Z = \frac{\hat{P}_x - \hat{P}_y}{\sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n_x} + \frac{\hat{P}_y(1-\hat{P}_y)}{n_y}}} \equiv N(0.1)$.

Next find the **value of the statistic** Z, under H_0. This is $z_0 = \frac{0.18-0.20}{\sqrt{\frac{0.18(1-0.18)}{800}}} = -1.472$.

You find the **critical region** corresponding to H_a with a $\alpha = 0.05$ significance level. Use a standardised normal distribution table or software to find the **critical value** $z_{\alpha/2} = z_{0.025} = 1.96$. The critical region of the comparison is $|Z| > 1.96$.

As the value of Z under H_0, $|z_0| = 1.472 < 1.96$, is outside the critical region, accept the H_0. This means that 20% of the families own fully paid for homes.

Example 4.25. In a city, they want to prove the effectiveness of an additive to reduce CO_2 emissions in vehicles. The additive is used in a fleet of 100 vehicles and is compared to a control fleet of 100 vehicles previously rated as having highCO_2 emissions (i.e., highly polluting). Of the vehicles that receive the additive, 8 maintain their rating. Of those that did not receive it, 25 also maintained their rating. Can you be sure that the additive is effective so that vehicles are not rated as highly polluting if you consider a significance level of 0.05?

Solution.

In this problem, you want to verify the efficiency of an additive so that vehicles are not rated as highly polluting. The fact that the additive can be effective or not refers to the fact that you have a dichotomy; this is, therefore, a Bernoulli event. The comparison with a reference group suggests that you must use the difference between two populations; specifically, the difference between two population proportions. So, you define the random variables as follows:

$X =$ "$number\ of\ polluting\ vehicles\ that\ receive\ the\ additive$"

$Y =$ "$number\ of\ polluting\ vehicles\ that\ do\ not\ receive\ the\ additive$"

These are Bernoulli random variables, i.e., $X \equiv B(p_x, n_x)$ and $Y \equiv B(p_y, n_y)$.

In the sampling, you find that $\hat{p}_x = \frac{8}{100} = 0.08$ and $\hat{p}_y = \frac{25}{100} = 0.25$. On the other hand, $n_x = n_y = 100 \geq 30$, i.e., you can approximate both distributions to normal distributions, $N\left(np, \sqrt{np(1-p)}\right)$.

You can use a hypothesis test to verify the effectiveness of the additive. Use a one-tailed test for the difference between proportions $p_X - p_Y$. Then, pose the **null hypothesis** and the **alternative hypothesis**:

H_0: $p_X \geq p_Y$

H_a: $p_X < p_Y$

In this case, you want to compare the difference between proportions, and the statistic is $Z = \frac{\hat{P}_X - \hat{P}_Y}{\sqrt{\frac{\bar{P}_X(1-\bar{P}_X)}{n_X} + \frac{\bar{P}_Y(1-\bar{P}_Y)}{n_Y}}} \equiv N(0.1)$.

Next find the **value of the statistic** Z, under H_0. This is $z_0 = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}} = \frac{0.08 - 0.25}{\sqrt{\frac{0.08(1-0.08)}{100} + \frac{0.25(1-0.25)}{100}}} = -3.4$.

You find the **critical region** corresponding to H_a. The critical region, with $p_X < p_Y$, comprises the values $Z \leq z_{\alpha/2}$, where $z_{\alpha/2} < 0$ is the value that satisfies $P(Z \leq z_{\alpha/2}) = \alpha$. In this example, with a significance level of $\alpha = 0.05$, $P(Z \leq z_{\alpha/2}) = 0.05$. Using a standardised normal distribution table or software, find the **critical value**, $z_{\alpha/2} = -1.65$. The critical region of the comparison is $|Z| > 1.65$.

As the value of Z under H_0, $|z_0| = 3.4 > 1.65$, is within the critical region, reject the H_0. This means that the additive is efficient, i.e., $p_X - p_Y < 0$ indicates that the proportion of vehicles that receive the additive and still pollute has decreased in comparison with the proportion in the control group.

Alternatively, you get the same result if you use the $\boldsymbol{p - value}$ ($\alpha_p = P(reject\ the\ statistic | H_0\ is\ true)$). Here, the first steps, up to the determination of the comparison statistic, are common to the critical region procedure. With the help of a standardised normal distribution table or software, find the comparison value of P corresponding to H_a. $\alpha_p = P(z > 3.4) = 0.00337$. As $\alpha_p = 0.00337 < \alpha = 0.05$, reject H_0, i.e., z and $\hat{p}_X - \hat{p}_Y$ are statistically meaningful at the level of α.

Example 4.26. In an experimental procedure, a value of 4 is determined for a Poisson random variable X. For this value, the conjecture is that its distribution parameter has a value of 2. If you work with a significance level of 0.05, can you say that the parameter is 2?

Solution.

You must solve the conjecture using a hypothesis test. As the conjectured value is 2, you use a one-tailed test for the Poisson distribution parameter λ:

$H_0: \lambda = 2$

$H_a: \lambda > 2$

You do not know if the condition $\lambda n > 25$ is satisfied. For this reason, you cannot use the value of the statistic that appears in the second row of Table 8.5. However, in the example, an experimental value is provided that can be assumed to be the value of the statistic under H_0.

The critical region is to the right of the acceptance region. To find the critical value $x_{\alpha/2}$ that separates the two regions, use the following condition: $P(X \geq x_{\alpha/2}) \leq \alpha = 0.05$, i.e., $\sum_{i=x_{\alpha/2}}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} \leq 0.05$. Using a Poisson distribution table or software, you find that, for $\lambda = 2$, the highest $x_{\alpha/2}$ that satisfies the imposed condition is $x_{\alpha/2} = 6$. Next, determine the critical region for $X \geq 6$.

As $X \geq 6$, 4 is not within the critical region and H_0 is not rejected at a significance level of 0.05, meaning that you can consider $\lambda = 2$ to be the Poisson distribution parameter of the random variable X.