Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

Lecture Note On Mathematical Statistics 2

B.Sc. in Mathematics

Fourth Stage

Assist. Prof. Dr. Feras Shaker Mahmood

Syllabus of Mathematical Statistics 2

Chapter 1: Bayesian Estimation

More Bayesian consepts, Some methods of Estimation, Informative and non informative prior.

Chapter 2: Interval Estimation:

Confidence Intervals for Means, Confidence Intervals for the Difference of Two Means, Confidence Intervals for Proportions sample size, Confidence intervals for Variance and for ration between two Variances, Confidence intervals for differences of probabilites

Chapter 3: Test Of Hypotheses:

General Concepts, Type of test of Hypothesis, Critical Region, Best of Critical region statistical test, Neywan – Pearson Theorem uniformly most powerful test, Likelihood ratio test, Sequential test.

References

- الاحصاء الرياضي، امير حنا ، الطبعة الاولى، جامعه الموصل، ١٩٩٥.
- مقدمة في الاحصاء الرياضي، صباح داود سليم، الطبعة الاولى، جامعة البصرة، ١٩٨٩.
- **Introduction to Mathematical Statistics, R. V. Hogg and A.T. Craig, (4, 5, 6) edition.**
- **Introduction to the Theory of Statistics, Alexander M.Mood, Franklin, A. Graybill, MC- Graw Hill, 1974.**
- **Mathematical Statistics with Applications, K. M. Ramachandran and C. P. Tsokos, 2009.**
- **Mathematical Statistics with Application, Richard L.Schaffer, Seventh Edition, Thomson Brooks, 2008.**
- **Probability and mathematical Statistics, Prasanna Sahoo, University of Louisville,, USA, 2008.**
- **An introduction to Mathematical Statistics and Its Applications, Ricard I. larson and Morris Luois, Fifth edition, Person Education, INC, 2012.**
- **Probability and Statistical Inference, Robert V. Hogg ,Elliot A. Tanis and ale L. Zimmerman, Ninth Edition, Pearson Education, USA,2015.**

Confidence Intervals

Parameter	Assumptions	Endpoints
μ	$N(\mu, \sigma^2)$ or n large, σ^2 known	$\overline{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
μ	$N(\mu, \sigma^2)$ σ^2 unknown	$\overline{x} \pm t_{\alpha/2}(n-1)\dfrac{s}{\sqrt{n}}$
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$
$\mu_X - \mu_Y$	Variances unknown, large samples	$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\dfrac{s_x^2}{n} + \dfrac{s_y^2}{m}}$

$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$\bar{x} - \bar{y} \pm t_{\alpha/2}(n+m-2) s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$, $s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$\mu_D = \mu_X - \mu_Y$	X and Y normal, but dependent	$\bar{d} \pm t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$
p	$b(n, p)$ n is large	$\frac{y}{n} \pm z_{\alpha/2} \sqrt{\frac{(y/n)[1-(y/n)]}{n}}$
$p_1 - p_2$	$b(n_1, p_1)$ $b(n_2, p_2)$	$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, $\hat{p}_1 = y_1/n_1, \ \hat{p}_2 = y_2/n_2$

Tests of Hypotheses

Hypotheses	Assumptions	Critical Region
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ or n large, σ^2 known	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$N(\mu, \sigma^2)$ σ^2 unknown	$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 known	$z = \dfrac{\overline{x} - \overline{y} - 0}{\sqrt{(\sigma_X^2/n) + (\sigma_Y^2/m)}} \geq z_\alpha$
$H_0: \mu_X - \mu_Y = 0$ $H_1: \mu_X - \mu_Y > 0$	Variances unknown, large samples	$z = \dfrac{\overline{x} - \overline{y} - 0}{\sqrt{(s_x^2/n) + (s_y^2/m)}} \geq z_\alpha$

$H_0\colon \mu_X - \mu_Y = 0$ $H_1\colon \mu_X - \mu_Y > 0$	$N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$, unknown	$t = \dfrac{\bar{x} - \bar{y} - 0}{s_p\sqrt{(1/n) + (1/m)}} \geq t_\alpha(n+m-2)$ $s_p = \sqrt{\dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$H_0\colon \mu_D = \mu_X - \mu_Y = 0$ $H_1\colon \mu_D = \mu_X - \mu_Y > 0$	X and Y normal, but dependent	$t = \dfrac{\bar{d} - 0}{s_d/\sqrt{n}} \geq t_\alpha(n-1)$
$H_0\colon p = p_0$ $H_1\colon p > p_0$	$b(n, p)$ n is large	$z = \dfrac{(y/n) - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$
$H_0\colon p_1 - p_2 = 0$ $H_1\colon p_1 - p_2 > 0$	$b(n_1, p_1)$ $b(n_2, p_2)$	$z = \dfrac{(y_1/n_1) - (y_2/n_2) - 0}{\sqrt{\left(\dfrac{y_1+y_2}{n_1+n_2}\right)\left(1 - \dfrac{y_1+y_2}{n_1+n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \geq z_\alpha$

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور فراس شاكر محمود

BAYESIAN ESTIMATION

We now describe another approach to estimation that is used by a group of Statisticians who call themselves Bayesians .To understand their approach Fully would require more text than we can allocate to this topic, but let us Begin this brief introduction by considering a simple application of the theorem of the Reverend Thomas Bayes.

Example:

Suppose we know that we are going to select an observation from a Poisson distribution with mean $\boldsymbol{\lambda}$ equal to 2 or 4. Moreover, prior to performing the experiment, we believe that $\lambda = 2$ has about four times as much chance of being the parameter as does $\lambda = 4$; that is the prior probabilities are

$P(\lambda = 2) = 0.8$ and $P(\lambda = 4) = 0.2$.

Solution:

The experiment is now performed and we observe that $x = 6$. At this point, our intuition tells us that $\lambda = 2$ seems less likely than before, as the observation $x = 6$ is much more probable with $\lambda = 4$ than with $\lambda = 2$, because, in an obvious notation,

$$P(X = 6/\lambda = 2) = 0.995 - 0.983 = 0.012$$

and

$$P(X = 6/\lambda = 4) = 0.889 - 0.785 = 0.104,$$

from Table .Our intuition can be supported by computing the conditional probability of $\lambda = 2$, given that $X = 6$:

$$P(\lambda = 2/X = 6) = \frac{P(\lambda = 2, X = 6)}{P(X = 6)}$$

$$= \frac{P(\lambda = 2)P(X = 6/\lambda = 2)}{P(\lambda = 2)P(X = 6/\lambda = 2) + P(\lambda = 4)P(X = 6/\lambda = 4)}$$

$$= \frac{(0.8)(0.012)}{(0.8)(0.012) + (0.2)(0.104)} = 0.316.$$

This conditional probability is called the posterior probability of $\lambda = 2$, given the single data point (here,$x = 6$).In a similar fashion, the posterior probability of $\lambda = 4$ is found to be 0.684 thus, we see that the probability of $\lambda = 2$ has decreased from 0.8 (the prior probability) to 0.316 (the posterior probability) with the observation of $x = 6$.

Example:

Suppose that Y has a binomial distribution with parameters n and $p = \theta$.

Then the pmf of Y, given $\boldsymbol{\theta}$, is

$$g(y/\theta) = \binom{n}{y} \theta^{y}(1-\theta)^{n-y}, \qquad y = 0,1,2,\dots,n$$

Solution:

Let us take the prior pdf of the parameter to be the beta pdf:-

$$h(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad 0 < \theta < 1.$$

Such a prior pdf provides a Bayesian a great deal of flexibility through the selection of the parameters α and β. Thus, the joint probabilities can be described by a product of a binomial pmf with parameters n and $\boldsymbol{\theta}$ and this beta pdf, namely,

$$k(y,\theta) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

On the support given by $y = 0,1,2,\dots,n$ and $0 < \theta < 1$. We find

$$k_1(y) = \int_0^1 k(y,\theta)d\theta$$

$$= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \frac{\Gamma(\alpha+y)\,\Gamma(n+\beta-y)}{\Gamma(n+\alpha+\beta)}$$

On the support $y = 0,1,2,\dots,n$ by comparing the integral with one involving a bete pdf with parameters $y+\alpha$ and $n-y+\beta$. Therefore,

$$k(\theta/y) = \frac{k(y,\theta)}{k_1(y)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\,\Gamma(n+\beta-y)} \theta^{y+\alpha-1}(1-\theta)^{n-y+b-1}, \qquad 0 < \theta < 1$$

Which is a beta pdf with parameters $y+\alpha$ and $n-y+\beta$. With the squared error loss function we must minimize, with respect to w(y), the integral

$$\int_0^1 [\theta - w(y)]^2\, k(\theta/y)\, d\theta,$$

to obtain the Bayes estimator. But, as noted earlier, if Z is a random variable with

A second moment, then $E[(Z-b)^2]$ is minimized by $b = E(Z)$. In the preceding integration, $\boldsymbol{\theta}$ is like the Z with pdf $k(\theta/y)$, and w(y) is like the b, so the minimization is accomplished by taking

$$w(y) = E(\theta/y) = \frac{\alpha+\beta}{\alpha+\beta+n}$$

Which is the mean of the beta distribution with parameters $y+\alpha$ and $n-y+\beta$.

It is instructive to note that this Bayes estimator can be written as

$$w(y) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

Which is a weighted average of the maximum likelihood estimate y/n of $\boldsymbol{\theta}$ and the mean $\alpha/(\alpha+\beta)$ of the prior pdf of the parameter. Moreover, the respective weights are $n/(\alpha+\beta+n)$ and $(\alpha+\beta)/(\alpha+\beta+n)$.Thus, we see that α and β should be selected so that not only is $\alpha/(\alpha+\beta)$ the desired prior mean, but also the sum $(\alpha+\beta)$ plays a role corresponding to a sample size. That is, if we want our prior opinion to have as much weight as a sample size of 20, we would take $\alpha+\beta=20$. So if our prior mean is 3/4, we select α=15 and β=5. That is, the prior pdf of $\boldsymbol{\theta}$ is beta (15, 5).If we observe n=40 and y=28, then the posterior pdf is beta (28+15=43, 12+5=17).

Example:

Let us consider again Example2, but now say that $X_1, X_2, \dots X_n$ is a random sample from the Bernoulli distribution with pmf

$$f(x/\theta)=\theta^{x}(1-\theta)^{1-x}, \quad x=0,1.$$

With the same prior pdf of $\boldsymbol{\theta}$, the joint distribution of $X_1, X_2, \dots X_n$ and $\boldsymbol{\theta}$

given by

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}, 0<\theta<1, x_i=0,1.$$

Of course, the posterior pdf of $\boldsymbol{\theta}$, given that $X_1=x_1, X_2=x_2, \dots, X_n=x_n$,

Is such that

$$k(\theta/x_1, x_2, \dots, x_n) \propto \theta^{\sum_{i=1}^{n}x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^{n}x_i+\beta-1}, 0<\theta<1,$$

Which is beta with $\alpha^{*}=\sum_{i=1}^{n}x_i+\alpha$, $\beta^{*}=n-\sum_{i=1}^{n}x_i+\beta$,the conditional mean of $\boldsymbol{\theta}$ is

$$\frac{\sum_{i=1}^{n}x_i+\alpha}{n+\alpha+\beta}=\left(\frac{n}{n+\alpha+\beta}\right)\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)+\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

Which, with $=\sum x_i$, is exactly the same result as that of Example 2.

MORE BAYESIAN CONCEPTS

Let $X_1, X_2, \dots X_n$ be a random sample from a distribution with pdf (pmf) $f(x/\theta)$, and let h ($\boldsymbol{\theta}$) be the prior pdf. Then the distribution associated with the marginal pdf of $X_1, X_2, \dots X_n$ namely,

$$k_1(x_1, x_2, \dots, x_n) = \int_{-\infty}^{\infty} f(x_1/\theta) f(x_2/\theta) \dots f(x_n/\theta) h(\theta) d\theta,$$

Is called the predictive distribution because it provides the best description of the Probabilities on $X_1, X_2, \dots X_n$. Often this creates some interesting distributions. For example, suppose there is only one X with the normal pdf

$$f(x/\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} e^{-(\theta x^2)/2}, \quad -\infty < x < \infty.$$

Here, $\theta = 1/\sigma^2$, the inverse of the variance, is called the precision of X. Say this precision has the gamma pdf

$$h(\theta) = \frac{1}{\Gamma(\alpha)\, \beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta}, 0 < \theta < \infty.$$

Then the predictive pdf is

$$k_1(x) = \int_0^{\infty} \frac{\theta^{\alpha+\frac{1}{2}-1} e^{-(\frac{x^2}{2}+\frac{1}{\beta})\theta}}{\Gamma(\alpha)\, \beta^\alpha \sqrt{2\pi}} d\theta$$

$$= \frac{\Gamma(\alpha + 1/2)}{\Gamma(\alpha)\, \beta^\alpha \sqrt{2\pi}} \frac{1}{(1/\beta + x^2/2)^{\alpha+1/2}}, \quad -\infty < x < \infty$$

Note that if $\alpha = r/2$ and $\beta = 2/r$, where r is a positive integer, then

$$k_1(x) \propto \frac{1}{(1 + x^2/r)^{(r+1)/2}}, \quad -\infty < x < \infty$$

Which is a t pdf with r degrees of freedom. So if the inverse of the variance-or precision $\boldsymbol{\theta}$-of a normal distribution varies as a gamma random variable , a generalization of a t distribution has been created that has heavier tails than the normal distribution. This mixture of normal (different from a mixed distribution) is attained by weighing with the gamma distribution in a process often called compounding.

Another illustration of compounding is given in the next example.

Example:

Suppose X has a gamma distribution with the two parameters k and θ^{-1}. (That is, the usual α is replaced by k and $\boldsymbol{\theta}$ by its reciprocal). Say h($\boldsymbol{\theta}$) is gamma with parameters α and β, so that

$$k_1(x) = \int_0^\infty \frac{\theta^k x^{k-1} e^{-\theta x}}{\Gamma(k)} \frac{1}{\Gamma(\alpha)\,\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta} d\theta$$

$$= \int_0^\infty \frac{x^{k-1}\theta^{k+\alpha-1} e^{-\theta(x+1/\beta)}}{\Gamma(k)\,\Gamma(\alpha)\,\beta^\alpha} d\theta$$

$$= \frac{\Gamma(k+\alpha) x^{k-1}}{\Gamma(k)\,\Gamma(\alpha)\beta^\alpha} \frac{1}{(x+1/\beta)^{k+\alpha}}$$

$$= \frac{\Gamma(k+x)\beta^k x^{k-1}}{\Gamma(k)\,\Gamma(\alpha)(1+\beta x)^{k+\alpha}} \ ,0 < x < \infty.$$

Of course, this is a generalization of the F distribution, which we obtain by letting

$$\alpha = r_2/2, \qquad k = r_1/2\ , \qquad and\ \ \beta = r_1/r_2.$$

Example:

(Berry, 1996) This example deals with predictive probabilities, and it concerns the breakage of glass panels in high-rise buildings. One such case involved 39 panels, and of the 39 panels that broke, it was known that 3 broke due to nickel sulfide (NiS) stones found in them. Loss of evidence prevented the causes of breakage of the other 36 panels from being known. So the court wanted to know whether the manufacturer of the panels or the builder was at fault for the breakage of these 36 panels.

From expert testimony, it was thought that usually about 5% breakage is caused By NiS stones. That is, if this value of p is selected from a beta distribution, we have

$$\frac{\alpha}{\alpha + \beta} = 0.05$$

Moreover, the expert thought that if two panels from the same lot break and one breakage was caused by NiS stones, then, due to the pervasive nature of the manufacturing process, the probability of the second panel breaking due to NiS stones increases to about 95%. Thus, the posterior estimate of p (see Example 2) with one "success" after one trial is

$$\frac{\alpha + 1}{\alpha + \beta + 1} = 0.95$$

Solving Equations 3 and 4 for α and β, we obtain

$\alpha = \frac{1}{360}$ and $\beta = \frac{19}{360}$

Now updating the posterior probability with 3 "success" out of 3 trials, we obtain the posterior estimate of p:

$$\frac{\alpha + 3}{\alpha + \beta + 3} = \frac{1/360 + 3}{20/360 + 3}$$

$$= \frac{1081}{1100} = 0.983.$$

Of course, the court that heard the case wanted to know the expert's opinion about the probability that all of the remaining 36 panels broke because of NiS stones. Using updated probabilities after the third break, then the fourth, and so on, we obtain the product

$$\left(\frac{1/360+3}{20/360+3}\right)\left(\frac{1/360+4}{20/360+4}\right)\left(\frac{1/360+5}{20/360+5}\right)\dots\left(\frac{1/360+38}{20/360+38}\right) = 0.8664.$$

That is, the expert held that the probability that all 36 breakages were caused by NiS stones was about 87%, which is the needed value in the court's decision.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Confidence Interval for Means μ

Given a random sample $X_1, X_2, \ldots, X_n$ from normal distribution $\boldsymbol{N(\mu, \sigma^2)}$, we shall now consider the closeness of $\bar{X}$, the unbiased estimator of **μ** , to the **unknown** mean **μ** . to do this , we use the error structure (distribution) of $\bar{X}$, namely , that $\bar{X}$ is $\boldsymbol{N(\mu, \frac{\sigma^2}{n})}$ to construct what is called a **confidence interval** for the unknown parameter **μ** when the variance $\boldsymbol{\sigma^2}$ is **known** . For the probability $1-\alpha$ we can find a number $\boldsymbol{z_{\alpha/2}}$ from table V in Appendix E such that

$$P\left(-\,z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$

For example , if $1-\alpha = 0.95$, then $z_{\alpha/2} = z_{0.05} = 1.645$ .Now recalling that $\boldsymbol{\sigma > 0}$, we see that the following inequalities are equivalent :

$$-\,\boldsymbol{z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}}$$

$$-\,\boldsymbol{z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \bar{X}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$\boldsymbol{-\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le -\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$\boldsymbol{\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \ge \mu \ge \bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Thus , since the probability of the first of these is $1-\alpha$, the probability of the last must also be $1-\alpha$, because the latter is true if an only if the former is true . that is , we have

$$\boldsymbol{P\left(\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha}$$

So the probability that the random interval

$$\boldsymbol{\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) , \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]}$$

Includes the **unknown** mean **μ** is $1-\alpha$

Once the sample is observed and the sample mean computed to equal $\overline{X}$, the interval [$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$, $\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$] becomes **known** . since the probability that the random interval covers **μ** before the sample is drawn is equal to $1-\alpha$, we now call the computed interval , $\overline{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ (for brevity) , a $100(1-\alpha)\%$ ***confidence interval for*** *the* ***unknown*** *mean* ***μ*** . Foe example $\overline{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is a 95% **confidence interval** for **μ** . The number $100(1-\alpha)\%$, or equivalently , $1-\alpha$ is called the ***confidence coefficient***.

Example : let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. if a random sample of $n = 27$ bulbs is tested until they burn out , yielding a sample mean of $\overline{X} = 1478$ hours , then a 95% **confidence interval for μ** is

$$\left[\overline{X} - z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X} + z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$= \left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right),\ 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right]$$

$$= [\,1478 - 13.58\,,\ 1478 + 13.58\,]$$

$$= [\,1464.42\,,\ 1491.58\,]$$

The next example will help to give a better intuitive feeling for the interpretation of a **confidence interval** .

Example: Let $\overline{X}$ be the observed sample mean of five observations of a random sample from the normal distribution $N(\mu, 16)$. A 90% **confidence interval** for the **unknown** mean **μ** is

$$\left[\overline{X} - 1.645\sqrt{\frac{16}{5}}\ , \overline{X} + 1.645\sqrt{\frac{16}{5}}\right]$$

Example : Let $X_1, X_2, \dots, X_{32}$ be a random sample of size 32 from a normal distribution $\mathbf{N(\mu, \sigma)}.^{2}$If $\overline{X} = 19.07$ and $S^2 = 10.60$, then what is the 95 % **confidence interval** for the
population mean **μ** ?

Solution : since $n = 32 \geq 30$, $z_{\alpha/2} = 1.96$ for 95% **confidence interval** ($\alpha/2 = 0.025$)

Hence , the **confidence interval for μ** at 95% **confidence level** is

$$19.07 - 1.96\sqrt{\frac{10.60}{32}} < \mu < 19.07 + 1.96\sqrt{\frac{10.60}{32}}$$

Thus 95% ***confidence interval*** *:* $17.94 < \mu < 20.20$

If the random sample arises from a normal distribution , we use the fact that

$$\boldsymbol{T = \frac{\overline{X} - \mu}{S/\sqrt{n}}}$$

has a t- distribution with $r = n - 1$ ***degrees of freedom*** , *where* S^2 *is the usual unbiased estimator of* σ^2. *Select* $t_{\alpha/2(n-1)}$ *so that* $\boldsymbol{P\left[T \geq t_{\alpha/2(n-1)}\right] = \alpha/2}$

$$\boldsymbol{1 - \alpha = P\left[-t_{\alpha/2(n-1)} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2(n-1)}\right]}$$

$$\mathbf{= P\left[-t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}} \leq \overline{X} - \mu \leq t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}}\right]}$$

$$\mathbf{= P\left[-\overline{X} - t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}} \leq -\mu \leq -\overline{X} + t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}}\right]}$$

$$= \mathbf{P}\left[\overline{\mathbf{X}} - \mathbf{t}_{\alpha/2(\mathbf{n-1})} \frac{\mathbf{S}}{\sqrt{\mathbf{n}}} \leq \boldsymbol{\mu} \leq \overline{\mathbf{X}} + \mathbf{t}_{\alpha/2(\mathbf{n-1})} \frac{\mathbf{S}}{\sqrt{\mathbf{n}}}\right]$$

Thus , the observations of a random sample provide $\overline{X}$ and S^2, and

$$\left[\overline{\mathbf{X}} - \mathbf{t}_{\alpha/2(\mathbf{n-1})} \frac{\mathbf{S}}{\sqrt{\mathbf{n}}} , \overline{\mathbf{X}} + \mathbf{t}_{\alpha/2(\mathbf{n-1})} \frac{\mathbf{S}}{\sqrt{\mathbf{n}}}\right]$$

is a $100(1-\alpha)\%$ **confidence interval for μ**

Example : Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves . Assume that the distribution of X is $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. To estimate μ , a farmer measured the butterfat production for $n = 20$ cows and obtained the following data

481	**537**	**513**	**583**	**453**	**510**	**570**	**500**	**457**	**555**
618	**327**	**350**	**643**	**499**	**421**	**505**	**637**	**599**	**392**

For these data , $\overline{X} = 507.50$ and $S = 89.75$. Thus , a point estimate of μ is $\overline{X} = 507.50$, since $t_{0.05}(19) = 1.729$. a 90% **confidence interval for μ** is

$507.50 \pm 1.729(\frac{89.75}{\sqrt{20}})$ or 507.50 ± 34.70

Or equivalently $[472.80 , 542.20]$

If we are not able to assume that the underlining distribution is normal , but μ and α are both **unknown** , approximate **confidence interval for μ** can still be constructed with the formula

$$\mathbf{T} = \frac{\overline{\mathbf{X}} - \boldsymbol{\mu}}{\mathbf{S}/\sqrt{\mathbf{n}}}$$

Which now only has an approximate t- distribution . Generally , this approximation is quite good (i.e., it is robust) for many not normal distribution ; in particular , it works will if the underlining distribution is symmetric , unimodal , and of the continuous type . However , if the distribution is highly skewed , there is great danger in using that approximation . in such a situation , it would be safer to use certain **nonparametric methods** for finding a **confidence interval** for the **median** of the distribution , one of which is given in this lecture. There is one other aspect of **confidence interval** that should be mentioned . so far , we have created only that are called **two- sided confidence interval for the mean μ** . sometimes , however , we might want only a **lower** (or **upper**) bound on **μ**. We proceed as follows .

Say $\overline{X}$ is the mean of a random sample of size n from the normal distribution $\mathbf{N(\mu, \sigma^2)}$, where , for the moment , assume that σ^2 is **known** .Then

$$P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_\alpha\right) = 1-\alpha$$

or equivalently

$$P\left(\overline{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu\right) = 1-\alpha$$

Once $\overline{X}$ is observed to be equal to $\overline{X}$, it follows that P[$\overline{X} - z_\sigma(\sigma/\sqrt{n})$, ∞) is a $100(1-\alpha)\%$ **one-sided confidence interval for μ** . That is , with the **confidence coefficient** $1-\alpha$, $\overline{X} - z_\sigma(\sigma/\sqrt{n})$, is **lower** bound for **μ** . similarly , ($-\infty$,$\overline{X} + z_\sigma(\sigma/\sqrt{n})$] is a **one-sided confidence interval for μ** and $\overline{X} + z_\sigma(\sigma/\sqrt{n})$ provides an **upper** bound for **μ** with the **confidence coefficient** $1-\alpha$. When σ is unknown , we will use $\mathbf{T} = (\overline{\mathbf{X}} - \boldsymbol{\mu}) \big/ (\mathbf{S}/\sqrt{\mathbf{n}})$

to find the corresponding **lower** or **upper** bounds for **μ** , namely

$$\overline{\mathbf{X}} - \mathbf{t}_\alpha(\mathbf{n}-\mathbf{1})(\mathbf{S}/\sqrt{\mathbf{n}}) \quad \textbf{and} \quad \overline{\mathbf{X}} + \mathbf{t}_\alpha(\mathbf{n}-\mathbf{1})(\mathbf{S}/\sqrt{\mathbf{n}})$$

CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO MEANS $\boldsymbol{\mu_x - \mu_y}$

Suppose that we are interested in comparing the means of two normal distributions. Let $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_m$ be, respectively, two independent random samples of sizes n and m from the two normal distributions $\mathbf{N(\mu_x, \sigma_x{}^2)}$ and $\mathbf{N(\mu_y, \sigma_y{}^2)}$. Suppose, for now, that $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are **known**. The random samples are independent; thus, the respective sample means $\overline{X}$ and $\overline{Y}$ are also independent and have distributions $\mathbf{N(\mu_x, \sigma_x{}^2)}$ and $\mathbf{N(\mu_y, \sigma_y{}^2)}$**.** Consequently, the distribution of $W = \overline{X} - \overline{Y}$ is $\mathbf{N\left(\mu_x - \mu_y , \frac{\sigma_x{}^2}{n} + \frac{\sigma_y{}^2}{m}\right)}$ and

$$\mathbf{P\left(-z_{\alpha/2} \leq \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x{}^2}{n}+\frac{\sigma_y{}^2}{m}}} \leq z_{\alpha/2}\right) = 1 - \alpha}$$

which can be rewritten as

$$\mathbf{P\left[\left(\overline{X} - \overline{Y}\right) - z_{\alpha/2}\sigma_W \leq \mu_x - \mu_y \leq \left(\overline{X} - \overline{Y}\right) + z_{\alpha/2}\sigma_W\right] = 1 - \alpha}$$

where $\boldsymbol{\sigma_W} = \sqrt{\mathbf{\frac{\sigma_x{}^2}{n} + \frac{\sigma_y{}^2}{m}}}$ is the **standard deviation** of $\mathbf{\overline{X} - \overline{Y}}$. Once the experiments have been performed and the means $\overline{X}$ and $\overline{Y}$ computed , the interval

$$\mathbf{[\,\overline{X} - \overline{Y} - z_{\alpha/2}\sigma_W\,,\ \overline{X} - \overline{Y} + z_{\alpha/2}\sigma_W\,]}$$

or, equivalently, $\left(\overline{X} - \overline{Y}\right) \pm z_{\alpha/2}\sigma_W$ provides a $100(1-a)\%$ **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** . Note that this interval is centered at the **point estimate** $\overline{X} - \overline{Y}$ of $\boldsymbol{\mu_x - \mu_y}$ and is completed by subtracting and adding the product of $z\alpha/_2$ and the **standard deviation** of the **point estimator**.

Example : In the preceding discussion, let $n = 15$, $m = 8$ $\overline{X} = 70.1$, $\overline{Y} = 75.3$, $\sigma_x{}^2 = 60$, $\sigma_y{}^2 = 40$ and $1 - \alpha = 0.90$. Thus , $1 - {}^{\alpha}/_{2} = 0.95 = \varphi(1.645)$. Hence ,

$$1.6450\sigma_W = 1.645\sqrt{\frac{60}{15} + \frac{40}{8}} = 4.935$$

and, since $\overline{X} - \overline{Y} = -5.2$, it follows that

$$[-5.2 - 4.935, -5.2 + 4.935] = [-10.135, -0.265]$$

is a 90% **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** . Because the **confidence interval** does not include zero, we suspect that $\boldsymbol{\mu_y}$ is greater than $\boldsymbol{\mu_x}$.

If the sample sizes are large and $\boldsymbol{\sigma_x}$ and $\boldsymbol{\sigma_y}$ are unknown, we can replace $\boldsymbol{\sigma_x{}^2}$ and $\boldsymbol{\sigma_y{}^2}$ with $S_x{}^2$and $S_y{}^2$, where $\mathbf{S_x{}^2}$and $\mathbf{S_y{}^2}$ are the values of the respective unbiased estimates of the variances. This means that

$$\overline{\mathbf{X}} - \overline{\mathbf{Y}} \pm \mathbf{z}{\alpha/_{2}}\sqrt{\frac{\mathbf{S_x{}^2}}{\mathbf{n}} + \frac{\mathbf{S_y{}^2}}{\mathbf{m}}}$$

serves as an approximate $100(1 - \alpha)\%$ **confidence interval for $\boldsymbol{\mu_x - \mu_y}$.**

Now consider the problem of constructing **confidence intervals** for the difference of the means of two normal distributions when the variances are **unknown** but the sample sizes are **small**. Let $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_m$ be two independent random samples from the distributions $\mathbf{N(\mu_X, \sigma_x{}^2)}$ and $\mathbf{N(\mu_y, \sigma_y{}^2)}$, respectively. If the sample sizes are not **large** (say, considerably smaller than 30), this problem can be a difficult one. However, even in these cases, if we can assume common, but **unknown**, **variances** (say, $\boldsymbol{\sigma_x{}^2 = \sigma_y{}^2 = \sigma^2}$) , there is a way out of our difficulty.

We know that

$$\mathbf{Z} = \frac{(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) - (\boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{y}})}{\sqrt{\boldsymbol{\sigma}^2/\mathbf{n} + \boldsymbol{\sigma}^2/\mathbf{m}}}$$

is $\mathbf{N(0,1)}$**.** Moreover, since the random samples are independent,

$$\mathbf{U} = \frac{(\mathbf{n}-\mathbf{1})\mathbf{S_x}^2}{\boldsymbol{\sigma}^2} + \frac{(\mathbf{n}-\mathbf{1})\mathbf{S_y}^2}{\boldsymbol{\sigma}^2}$$

is the sum of two independent **chi-square** random variables; thus, the distribution of U is $(n + m - 2)$. In addition, the independence of the sample means and sample variances implies that Z and U are independent. According to the definition of a T random variable,

$$\mathbf{T} = \frac{\mathbf{Z}}{\sqrt{\mathbf{U}/(\mathbf{n}+\mathbf{m}+\mathbf{2})}}$$

has a distribution with $n + m - 2$ **degrees of freedom**. That is,

$$\mathrm{T} = \frac{\dfrac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}})}{\sqrt{\frac{\boldsymbol{\sigma}_{\mathbf{x}}^2}{\mathbf{n}}+\frac{\boldsymbol{\sigma}_{\mathbf{y}}^2}{\mathbf{m}}}}}{\sqrt{\left[\frac{(\mathbf{n}-\mathbf{1})\mathbf{S_x}^2}{\boldsymbol{\sigma}^2}+\frac{(\mathbf{n}-\mathbf{1})\mathbf{S_y}^2}{\boldsymbol{\sigma}^2}\right]\Big/(\mathbf{n}+\mathbf{m}-\mathbf{2})}}$$

$$= \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}})}{\sqrt{\left[\frac{(\mathbf{n}-\mathbf{1})\mathbf{S_x}^2+(\mathbf{n}-\mathbf{1})\mathbf{S_y}^2}{\mathbf{n}+\mathbf{m}-\mathbf{2}}\right]\left[\frac{\mathbf{1}}{\mathbf{n}}+\frac{\mathbf{1}}{\mathbf{m}}\right]}}$$

degrees of freedom. Thus, with

has a t distribution with to $r = n + m - 2$ **degrees of freedom**. Thus, with $\mathbf{t_0} = \mathbf{t}_{\alpha/2}(\mathbf{n}+\mathbf{m}-\mathbf{2})$, we have

$$\mathbf{P}(-\mathbf{t_0} \le \mathbf{T} \le \mathbf{t_0}) = \mathbf{1} - \boldsymbol{\alpha}$$

solving the inequalities for $\boldsymbol{\mu_x - \mu_y}$, yields

$$\mathbf{P}\left(\overline{\mathbf{X}} - \overline{\mathbf{Y}} - \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}} \le \boldsymbol{\mu_x - \mu_y} \le \overline{\mathbf{X}} - \overline{\mathbf{Y}} + \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\right)$$

where the pooled estimator of the common **standard deviation** is

$$\mathbf{S_p} = \sqrt{\frac{(\mathbf{n-1})\mathbf{S_x}^2 + (\mathbf{m-1})\mathbf{S_y}^2}{\mathbf{n+m-2}}}$$

If $\overline{X}, \overline{Y}$, and S_p are the observed values of $\overline{X}$, $\overline{Y}$. and S_p , then

$$\left[\overline{\mathbf{X}} - \overline{\mathbf{Y}} - \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\ ,\ \overline{\mathbf{X}} - \overline{\mathbf{Y}} + \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\right]$$

is a $100(1-\alpha)\%$ **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** .

Example : Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $\mathbf{N}(\boldsymbol{\mu_X}, \boldsymbol{\sigma_x}^2)$ and $\mathbf{N}(\boldsymbol{\mu_y}, \boldsymbol{\sigma_y}^2)$, respectively, where $\boldsymbol{\sigma^2}$ is **unknown**. If a random sample of $n = 9$ students from large high schools yielded $\overline{X} = 81.31$, $\sigma_x^2 = 60.76$, and a random sample of $m = 15$ students from small high schools yielded $\overline{Y} = 78.61$, $\sigma_y^2 = 48.24$, then the endpoints for a 95% **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** are given by

$$\mathbf{81.31 - 78.61 \pm 2.074}\sqrt{\frac{\mathbf{8(60.76) + 14(48.24)}}{\mathbf{22}}}\sqrt{\frac{\mathbf{1}}{\mathbf{9}} + \frac{\mathbf{1}}{\mathbf{15}}}$$

because $t_{0.025}(22) = 2.074$. The 95% **confidence interval** is $[-3.65, 9.05]$.

REMARKS The assumption of equal variances, namely, $\sigma_x^2 = \sigma_y^2$. can be modified somewhat so that we are still able to find a confidence interval for $\mu_x - \mu_y$. That is, if we know the ratio $\boldsymbol{\sigma_x^2/\sigma_y^2}$ of the variances, we can still make this type of statistical inference by using a random variable with a t distribution. However, if we do not know the ratio of the variances and yet suspect that the unknown $\sigma_x{}^2$ and $\sigma_y{}^2$ differ by a great deal, what do we do? It is safest to return to

$$\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x{}^2}{n}+\frac{\sigma_y{}^2}{m}}}$$

for the inference about $\boldsymbol{\mu_x - \mu_y}$ but replacing $\sigma_x{}^2$ and $\sigma_y{}^2$ by their respective estimators $S_x{}^2$ and $S_y{}^2$. That is, consider

$$\mathbf{W} = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}}$$

What is the distribution of W? As before, we note that if n and m are large enough and the underlying distributions are close to normal (or at least not badly skewed), then W has an approximate normal distribution and a **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** can be found by considering

$$\mathbf{P}\left(-z_{\alpha/2} \le \mathbf{W} \le z_{\alpha/2}\right) \approx \mathbf{1} - \boldsymbol{\alpha}$$

However, for smaller n and m, Welch has proposed a Student's t distribution as the approximating one for W. Welch's proposal was later modified by Aspin. (See A. A. Aspin, "Tables for Use in Comparisons Whose Accuracy Involves Two Variances, Separately Estimated," Biometrika , 36 (1949), pp. 290-296, with an appendix by B. L. Welch in which he makes the suggestion used here.] The approximating Student's t distribution has r degrees of freedom, where

$$\frac{1}{r} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \quad \text{and} \quad \mathbf{c} = \frac{\frac{S_x{}^2}{n}}{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}$$

An equivalent formula for r is

$$\mathbf{r} = \frac{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_x^2}{n}\right)^2+\frac{1}{m-1}\left(\frac{S_y^2}{m}\right)^2}$$

In particular, the assignment of r by this rule provides protection in the case in which the smaller sample size is associated with the larger variance by greatly reducing the number of **degrees of freedom** from the usual $n + m - 2$. Of course, this reduction increases the value of $t_{\alpha/2}$. If r is not an integer, then use the greatest integer in r ; that is, use [r] as the number of degrees of freedom associated with the approximating Student's

t-distribution. An approximate $100(1-\alpha)\%$ **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$ is given by

$$\overline{\mathbf{X}} - \overline{\mathbf{Y}} \pm \mathbf{t}_{\alpha/2}(\mathbf{r})\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

It is interesting to consider the two-sample T in more detail. It is

$$\mathbf{T} = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

$$= \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\left[\frac{(n-1)S_x^2}{nm}+\frac{(m-1)S_y^2}{nm}\right]\left[\frac{n+m}{n+m-2}\right]}}$$

Now, since $(n-1)/n \approx 1$, $(m - 1)/m \approx 1$, and $(n + m)/(1 + m - 2) \approx 1$, we have

$$\mathbf{T} \approx \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}$$

We note that, in this form, each variance is divided by the wrong sample size! That is, if the sample sizes are large or the variances **known,** we would like

$$\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}} \quad \textbf{or} \quad \sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}$$

in the denominator; so T seems to change the sample sizes. Thus, using this T is particularly bad when the sample sizes and the variances are unequal; hence, caution must be taken in using that T to construct a **confidence interval for $\mu_x - \mu_y$** . That is, if $n < m$ and $\sigma_x^2 < \sigma_y^2$, then T does not have a t- distribution which is close to that of a Student t-distribution with

$n + m - 2$ degrees of freedom: Instead, its spread is much less than the Student t's as the term σ_y^2/n in the denominator is much larger than it should be. By contrast, if $m < n$ and $\sigma_x^2 < \sigma_y^2$, then $S_x^2/m + S_y^2/n$ is generally smaller than it should be and the distribution of T is spread out more than that of the Student t.

There is a way out of this difficulty, however: When the underlying distributions are close to normal, but the sample sizes and the variances are seemingly much different, we suggest the use of

$$\mathbf{W} = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}$$

where Welch proved that W has an approximate t distribution with [r] degrees of freedom, with the number of degrees of freedoms.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Interval Estimation

Definition. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population X with density $f(x;\theta)$ where θ is an unknown parameter. The interval estimator of θ is called a $100(1-\alpha)\%$ *confidence interval* for θ if $P(L \le \theta \le U) = 1-\alpha$

The random variable L is called the lower confidence limit and U is called the upper confidence limit. The number $(1-\alpha)$ is called the confidence coefficient or degree of confidence.

Thus, if we want a 95% confidence interval for, say , population mean μ , then $\alpha = 0.05$. Note that for the discrete random variables , we may not be able to find a lower confidence limit L and an upper confidence limit U such that the probability $P(L \le \theta \le U)$, is exactly $(1 - \alpha)$.

A Method of Finding the Confidence Interval

1- confidence interval for the population mean μ

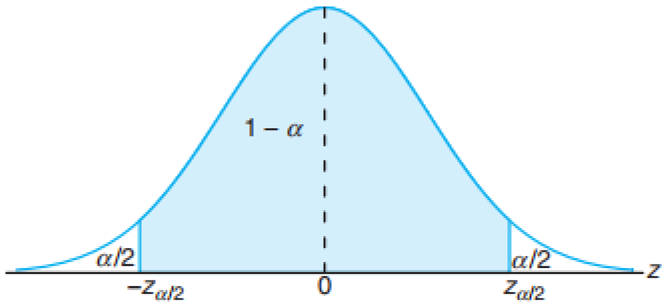
we first study the simpler but unrealistic case where we are trying to estimate μ

(a) : The Case of σ Known : The sampling distribution of $\overline{X}$ is centered at μ , Its variance is σ^2/n as we learned previously $z_{\alpha/2}$ is the value for which

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$$

Where $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Hence

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$



we learned previasly that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has stansard normal distribution for $n \geq 30$

$$P\left(\bar{X} - z_{\alpha/2}\ \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\ \sigma/\sqrt{n}\right) = 1 - \alpha$$

multiply by $\sigma/\sqrt{n}$, subtract $\bar{X}$, multiply by -1 to get this . Now we select a particuler sample of size n and get a spafic value of $\bar{X}$ then.

Confidence Interval on μ ; **σ Known**

If $\bar{X}$ is the mean of a random sample of size n from a population with known variance σ2 , a 100(1 - α)% confidence interval for μ is given by

$$\bar{X} - z_{\alpha/2}\ \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\ \sigma/\sqrt{n}$$

where $z_{\alpha/2}$ is the z-value from the stansard normal distribution leaving an area of $\alpha/2$ to the right .

note 1 : $L = \bar{X} - z_{\alpha/2}\ \sigma/\sqrt{n}$, $U = \bar{X} + z_{\alpha/2}\ \sigma/\sqrt{n}$

note 2 : The larger n , the tighter the confidence interval

note 3 : The smaller α , the wider the confidence interval

Example : A sample of 64 resistors from a population line are found to bake a mean resistance of 206 ohms. Find the 95% and 99% confidence intervals for the mean resistance of the population . Assume that the population standard deviation is 4 ohms .

Solution: The point estimate of μ is $\overline{X}$=206.

(a): 95% : $1-\alpha = 0.95$, $\alpha/2 = 0.025$

$z_{0.025} = 1.96$ (from normal table)

Hence , the confidence interval for μ at 95% confidence level is

$$206 - 1.96\frac{4}{\sqrt{64}} < \mu < 206 + 1.96\frac{4}{\sqrt{64}}$$

Thus 95% confidence interval : $205.02 < \mu < 206.98$

(b) : 99% : $1-\alpha = 0.99$, $\alpha/2 = 0.005$

$z_{0.005} = 2.575$ (from normal table)

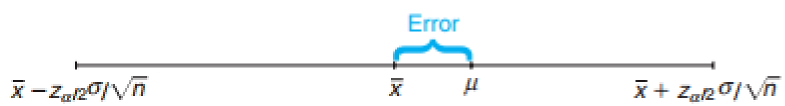
Hence , the confidence interval for μ at 99% confidence level is

$$206 - 2.575\frac{4}{\sqrt{64}} < \mu < 206 + 2.575\frac{4}{\sqrt{64}}$$

Thus 99% confidence interval : $204.71 < \mu < 207.29$

Theorem 1 : If $\overline{x}$ is used as an estimate of μ , we can be

100(1 - α)% confident that the error will not exceed $z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}$



Theorem 2 : If $\overline{x}$ is used as an estimate of μ, we can be

100(1 - α)% confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}}{e}\right)^2 \quad \text{rounded up}$$

Example 2: How large a sample size is required if in an previous example we want to be 95% confident that our estimate of μ (mean resistance of population) is off by less than 0.01?

Solution: The population standard deviation is $\sigma = 4$ $z_{\alpha/2} = 1.96$ for 95% confidence interval $(\alpha/2 = 0.025)$

So , $n = \left(\frac{1.96*4}{0.01}\right)^2 = 6146.56$, round up $n = 6146$.

(b): The Case of σ Unknown : Usually when we are trying to estimate μ , when **σ** is **unknown**

LARGE SAMPLE CONFIDENCE INTERVAL FOR μ

for a large sample of size n , let $\overline{X}$ be the sample mean. Then the large sample (1 - α)100% confidence interval for the population mean μ is

$$\overline{X} \pm z_{\alpha/2}\ {}^{\sigma}/_{\sqrt{n}} \approx \overline{X} \pm z_{\alpha/2}\ {}^{S}/_{\sqrt{n}}$$

where S is a point estimate of σ. That is

$$P\left(\overline{X} - z_{\alpha/2}\ {}^{S}/_{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\ {}^{S}/_{\sqrt{n}}\right) = 1 - \alpha$$

Example 3 : Let $X_1, X_2, \ldots, X_{32}$ be a random sample of size 32 from a normal distribution $N(\mu, \sigma^2)$. If $\overline{X} = 19.07$ and $S^2 = 10.60$, then what is the 95 % confidence interval for the population mean μ ?

Solution: since $n = 32 \geq 30$, $z_{\alpha/2} = 1.96$ for 95% confidence interval $(\alpha/2 = 0.025)$

Hence , the confidence interval for μ at 95% confidence level is

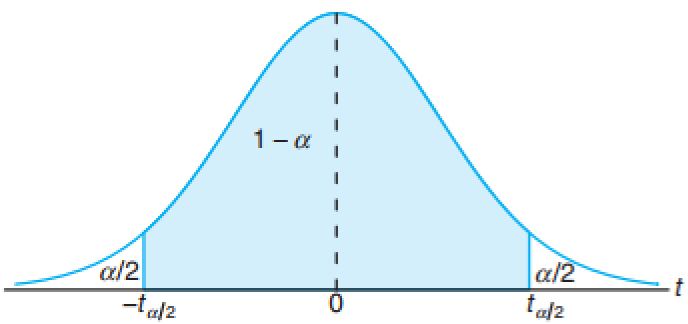
$$19.07 - 1.96\sqrt{\frac{10.60}{32}} < \mu < 19.07 + 1.96\sqrt{\frac{10.60}{32}}$$

Thus 95% confidence interval : $17.94 < \mu < 20.20$.

SMALL SAMPLE CONFIDENCE INTERVALS FOR μ

Let $X_1, X_2, \dots, X_n$ be a random sample from a normal distribution, then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student t-distribution with n - 1 degrees of freedom. σ (population standard deviation)is unknown ,but is replaced with S (sample standard deviation) ,Similar to before $P\left(-t_{\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}\right) = 1-\alpha$, multiply by $S/\sqrt{n}$, subtract $\bar{X}$, multiply by -1 to get

$$P\left(\bar{X} - t_{\alpha/2}\, S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2}\, S/\sqrt{n}\right) = 1-\alpha$$



With $t_{\alpha/2}$ being the t-value (from t-table) for $v = n-1$ degrees of freedom above which we can find an area of $\alpha/2$. The difference from before is the use of t-distribution table rather than the standard normal distribution .

Confidence Interval on μ , **σ Unknown** : If $\bar{X}$ and S are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a 100(1-α)% confidence interval for μ is

$$\bar{X} - t_{\alpha/2}\, S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2}\, S/\sqrt{n}$$

where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Some Examples and Definitions

The two principal areas of statistical inference are the areas of estimation of parameters and of tests of statistical hypotheses. The problem of estimation of parameters, both point and interval estimation, has been treated. In this chapter some aspects of statistical hypotheses and tests of statistical hypotheses will be considered. The subject will be introduced by way of example.

Example 1. Let it be known that the outcome X of a random experiment is $n(\theta, 100)$. For instance, X may denote a score on a test, which score we assume to be normally distributed with mean Band variance 100. Let us say that past experience with this random experiment indicates that $\theta = 75$. Suppose, owing possibly to some research in the area pertaining to this experiment, some changes are made in the method of performing this random experiment. It is then suspected that no longer does $\theta = 75$ but that now $\theta > 75$. There is as yet no formal experimental evidence that $\theta > 75$; hence the statement $\theta > 75$ is a conjecture or a statistical hypothesis. In admitting that the statistical hypothesis $\theta > 75$ may be false, we allow, in effect, the possibility that θ: s; 75. Thus there are actually two statistical hypotheses. First, that the unknown parameter θ: s; 75; that is, there has been no increase in θ. Second, that the unknown parameter $\theta > 75$. Accordingly, the parameter space is $\Omega = \{\theta; -\infty < \theta < \infty\}$. We denote the first of these hypotheses by the symbols H_0: θ:s; 75 and the second by the symbols $H_1 : \theta > 75$. Since the values $\theta > 75$ are alternatives to those where θ: s; 75, the hypothesis H_1: $\theta > 75$ is called the alternative hypothesis. Needless to say, Ho could be called the alternative H_1; however, the conjecture, here $\theta > 75$, that is made by the research worker is usually taken to be the alternative hypothesis. In any case the problem is to decide which of these hypotheses is to be

accepted. To reach a decision, the random experiment is to be repeated a number of independent times, say n, and the results observed. That IS, we consider a random sample Xl, X 2 , ... , Xn; from a distribution that is $n(\theta, 100)$, and we devise a rule that will tell us what decision t? make once the experimental values, say $x1, x2, \ldots, xn$, have been determined. Such a rule is called a test of the hypothesis $H_0 : \theta \leq 75$ against the alternative hypothesis $H_1 : \theta > 75$. There is no bound on the number of rules or tests that can be constructed. We shall consider three such tests. Our tests will be constructed around the following notion. We shall partition the sample space d into a subset e and its complement $C*$. If the experimental values of Xl, X 2 , ... , Xn, say $x1, x2, \ldots, xn$, are such that the point $(x1, x2, \ldots, xn) \in C$, we shall reject the hypothesis H_0 (accept the hypothesisH_1)' If we have$(x1, x2, \ldots , xn) \in C*$, we shall accept the hypothesis H_1 (reject the hypothesis H_1)

Test 1. Let$n = 25$. The sample space d is the set$\{(x1, x2, \ldots , x25); -\infty < xi < \infty, i = 1,2, \ldots , 25\}$. Let the subset C of the sample space be $C = \{(x1, x2, \ldots , x25); x1 + x2 + \ldots + x25 > (25)(75)\}$. We shall reject the hypothesis H_0 if and only if our 25 experimental values are such that $(x1, x2, \ldots , x25) \in C$. If $(x1, x2, \ldots , x25)$ is not an element of C, we shall accept the hypothesis H_0. This subset C of the sample space that leads to the rejection of the hypothesis $H0: \theta \le 75$ is called the critical region of Test 1. Now $\sum_1^{25} xi > (25)(75)$ if and only if $\bar{x} > 75$, where $\bar{x} = \sum_1^{25} xi/25$.

Thus we can much more conveniently say that we shall reject the hypothesis $H_0: \theta \le 75$and accept the hypothesis $H_1: \theta > 75$ if and only if the experimentally determined value of the sample mean $\bar{x}$ is greater than75. If $\bar{x} \le 75$, we accept the hypothesis$H_0: \theta \le 75$. Our test then amounts to this: We shall reject the hypothesis $H_0: \theta \le 75$ if the mean of the sample exceeds the maximum value of the mean of the distribution when the hypothesis H_0 is true.

It would help us to evaluate a test of a statistical hypothesis if we knew the probability of rejecting that hypothesis (and hence of accepting the alternative

$$Pr[(x_{1,\ldots,} x_{25}) \in C] = Pr(X > 75).$$

hypothesis). In our Test 1, this means that we want to compute the probability

Obviously, this probability is a function of the parameter 8 and we shall denote it by $K_1(\theta)$ The function $K_1(\theta) = Pr(\bar{X} > 75)$is called the power

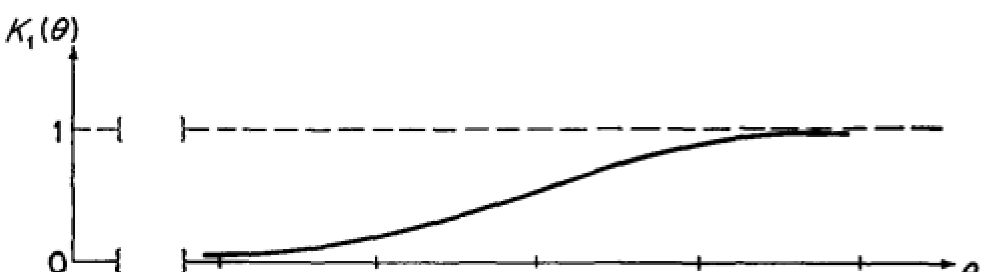


Figure 1.

function of Test 1, and the value of the power function at a parameter point is called the power of Test 1 at that point. Because $\overline{X}$ is $n(8, 4)$, we have

$$k_1\ (\theta) = \Pr\left(\frac{x-\theta}{2} > \frac{75-\theta}{2}\right) = 1 - N(\frac{75-\theta}{2})$$

So, for illustration, we have, by Table III of Appendix B, the power at $\theta = 75$ to be $K_1(75) = 0.500$. Other powers are $K_1(73) = 0.159, K_1(77) = 0.841$, and $K_1(79) = 0.977$. The graph of $K_1(8)$ of Test 1 is depicted in following Figure 1. Among other things, this means that, if $\theta = 75$, the probability of rejecting the hypothesis $H1: \theta \leq 75$ is $\frac{1}{2}$ That is, if $\theta = 75$ so that H_0 is true, the probability of rejecting this true hypothesis $H0$ is $\frac{1}{2}$. Many statisticians and research workers find it very undesirable to have such a high probability as $\frac{1}{2}$ assigned to this kind of mistake: namely the rejection of H_0 when H_0 is a true hypothesis. Thus Test 1 does not appear to be a very satisfactory test. Let us try to devise another test that does not have this objectionable feature. We shall do this by making it more difficult to reject the hypothesis H_0, with the hope that this will give a smaller probability of rejecting H_0 when that hypothesis is true.

Test 2. Let $n = 25$. We shall reject the hypothesis $H_0: \theta \leq 75$ and accept the hypothesis $H_1: \theta > 75$ if and only if $\bar{x} > 78$. Here the critical region is $C = \{(x_1, \ldots, x_{25});\ x_1 + \ldots + x_{25} > (25)(78)\}$. The power

$$K_2(\theta) = \Pr(X > 78\ ) = 1 - N\left(\frac{78-\theta}{2}\right)$$

function of Test 2 is, because X is $n(8, 4)$,Some values of the power function of Test 2 are $K_2(73) = 0.006, K_2(75) = 0.067,\ K_2(77) = 0.309$, and $K_2(79) = 0.691$. That is, if $\theta = 75$, the probability of rejecting $H0: \theta \leq 75$ is 0.067; this is much more desirable than the corresponding probability 1- that resulted from Test 1. However, if H_0 is false and, in fact, $\theta = 77$, the probability of rejecting $H0: \theta \leq 75$ (and hence of accepting $H1: \theta > 75$) is only 0.309. In certain instances, this low probability 0.309 of a correct decision (the acceptance of H_1 when H_1 is true) is objectionable. That is, Test 2 is not wholly satisfactory. Perhaps we can overcome the undesirable features of Tests 1 and 2 if we proceed as in Test 3.

Test 3. Let us first select a power function $K3(\theta)$ that has the features of a small value at $\theta = 75$ and a large value at$\theta = 77$. For instance, take $K3(75) = 0.159$ and$K3(77) = 0.841$. To determine a test with such a power function, let us reject $H0: \theta \leq 75$ if and only if the experimental value $\bar{x}$ of the mean of a random sample of size n is greater than some constant c. Thus the critical region is $C = \{(x_1, x_2, \dots, x_n) ;\ x_1 + x_2 + \dots + x_n > nc\}$. It should be noted that the sample size n and the

$$K_3(\theta) = \Pr(X > c) = 1 - N\left(\frac{c - \theta}{10\sqrt{n}}\right).$$

The conditions $K_3(75) = 0.159$ and $K_3(77) = 0.841$ require that

$$1 - N\left(\frac{c-75}{10/\sqrt{n}}\right) = 0.159, \qquad 1 - N\left(\frac{c-\theta}{10/\sqrt{n}}\right)$$

Equivalently , we have

$$\frac{c-75}{10/\sqrt{n}} = 1, \qquad \frac{c-77}{10/\sqrt{n}} = -1$$

constant c have not been determined as yet. However, since $\bar{X}$ is $n(\theta, 100/n)$, the power function is

The solution to these two equations in n and c is $n = 100$, $c = 76$. With these values of n and c, other powers of Test 3 are $K3(73) = 0.001$ and$K3(79) = 0.999$. It is important to observe that although Test 3 has a more desirable power function than those of Tests 1 and 2, a certain "price" has been paid-a sample size of $n = 100$ is required in Test 3, whereas we had $n = 25$ in the earlier tests.

Remark. Throughout the text we frequently say that we accept the hypothesis H_0 if we do not reject H_0 in favor of H_1. If this decision is made, it certainly does not mean that H_0 is true or that we even believe that it is true. All it means is, based upon the data at hand, that we are not convinced that the hypothesis H_0 is wrong. Accordingly, the statement "We accept H_0 " would possibly be better read as "We do not reject Hi;" However, because it is in fairly common use, we use the statement "We accept H_0," but read it with this remark in mind.

We have now illustrated the following concepts:

a) A statistical hypothesis.

b) A test of a hypothesis against an alternative hypothesis and the associated concept of the critical region of the test.
c) The power of a test.

These concepts will now be formally defined.

Definition : A statistical hypothesis is an assertion about the distribution of one or more random variables . if the statistical hypothesis completely specifies the distribution , it is called a simple statistical hypothesis ; if it does not, it is called composite statistical hypothesis. If we refer to example 1, we see that both $H0: \theta \leq 75$ and $H1 > 75$Are composite statistical hypothesis , since of them completely specifies The distribution. If there , instead of$H0: \theta \leq 75$ we had $H0: \theta = 75$, Then H_0 would have been a simple statistical hypothesis.

Definition:. A test of a statistical hypothesis is a rule which, when the experimental sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration.

Definition:. Let C be that subset of the sample space which, in accordance with a prescribed test, leads to the rejection of the hypothesis under consideration. Then C is called the critical region of the test.

Definition :. The power function of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is that function, defined for all distributions under consideration, which yields the probability that the sample point falls in the critical region C of the test, that is, a function that yields the probability of rejecting the hypothesis under consideration. The value of the power function at a parameter point is called the power of the test at that point.

Definition:. Let H_0 denote a hypothesis that is to be tested against an alternative hypothesis H_1 in accordance with a prescribed test. The significance level of the test (or the size of the critical region C) is the maximum value (actually supremum) of the power function of the test when H_0 is true.

Example: It is known that the random variable X has a P. d . f. of the from

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad 0 < x < \infty$$

$$= 0 \text{ elsewhere}$$

It is desired to test the simple hypothesis $H_0 : \theta$
$= 2$ against the alternative simple hypothesis$H_1 : \theta$
$= 4$. Thus Ω
$= \{\theta ; \theta = 2 , 4 \}$. A random simple X_1, X_2 of size n
$= 2$ will be used . The test to be used is defined by taking

If we refer again to Example 1, we see that the significance levels of Tests 1, 2, and 3 of that example are 0.500,0.067, and 0.159, respectively. An additional example may help clarify these definitions.

the critical region to be $C = \{(x_1\ x_2) ;\ 9.5 \leq x_1 + x_2 < \infty\}$. The power function of the test and the significance level of the test will be determined.

There are but two probability density functions under consideration, namely,$f(x;2)$ specified by $H_0$and $f(x;4)$ specified by H_1 . Thus the power function is defined at but two points $\theta = 2$ and $\theta = 4$. The power function of the test is given by $\Pr[(X_1, X_2) \in C]$. If H_0 is true, that is, $\theta = 2$, the joint p.d.f. of X_1 and X_2 is

$$f(x_1 ; 2)f(x_2 ; 2) = \frac{1}{4} e^{-(x_1+x_2)/2}, \qquad 0 < x_1 < \infty, \qquad 0 < x_2 < \infty$$

$$= 0 \text{ elsewher}$$

$$\Pr[(x_1, x_2) \in C = 1 - \Pr[(x_1, x_2) \in C^*]$$

$$= 1 - \int_0^{9.5} \int_0^{9.5-x_2} \frac{1}{4} e^{-(x_1+x_2)/2}\, dx_1\, dx_2$$

$$= 0.05 \text{ approximately .}$$

If H_1 is turn , that is , $\theta = 4$, the joint p. d. f. of x_1 and x_2 is

$$f(x_1 ; 4)f(x_2 ; 4) = \frac{1}{16} e^{-(x_1+x_2)/4} \quad 0 < x_1 < \infty, \qquad 0 < x_2 < \infty$$

$$= 0 \text{ elsewher}$$

and

$$\Pr[(x_1, x_2) \in C = 1 - \int_0^{9.5} \int_0^{9.5-x_2} \frac{1}{16} e^{-(x_1+x_2)/4}\, dx_1\, dx_2$$

$$= 0.31 \text{ approximately}$$

Thus the power of the test is given by 0.05 for $\theta = 2$ and by 0.31 for $\theta = 4$. That is, the probability of rejecting $H_0$when $H_0$is true is 0.05, and the probability of rejecting $H_0$when $H_0$is false is 0.31. Since the significance level of this test (or the size of the critical region) is the power of the test when $H_0$is true, the significance level of this test is 0.05. The fact that the power of this test, when $\theta = 4$, is only 0.31 immediately suggests that a search be made for another test which, with the same power when$\theta = 2$, would have a power greater than 0.31 when$\theta = 4$. However, Section 7.2 will make clear that such a search would be fruitless. That is, there is no test with a significance level of 0.05 and based on a random sample of size $n = 2$ that has a greater power at $\theta = 4$. The only manner in which the situation may be improved is to have recourse to a random sample of size n greater than 2.

Our computations of the powers of this test at the two points $\theta = 2$ and $\theta = 4$ were purposely done the hard way to focus attention on fundamental concepts. A procedure that is computationally simpler is the following. When the hypothesis H o is true, the random variable X is $X^2(2)$. Thus

$$\Pr(X \geq 9.5) = 1 - \Pr(X < 9.5) = 1 - 0.95 = 0.05$$

the random variable $X_1 + X_2 = Y$, say, is $X^2(4)$. Accordingly, the power of the test when H 0 is true is given by

from Table II of Appendix B. When the hypothesis H_1 is true, the random variable $X/2$ is $X^2(2)$; so the random variable $(X_1 + X_2)/2 = Z$, say, is $X^2(4)$. Accordingly, the power of the test when H_1 is true is given by

$$\Pr(X_1 + X_2 \geq 9.5) = \Pr(Z \geq 4.75)$$

$$\int_{4.75}^{\infty} \frac{1}{4} z e^{-z/2} dz,$$

Which is equal to 0.31 , approximately

Remark:. The rejection of the hypothesis $H_0$when that hypothesis is true is, of course, an incorrect decision or an error. This incorrect decision is often called a type I error; accordingly, the significance level of the test is the probability of committing an error of type 1. The acceptance of H_0 when

H_0 is false (H_1 is true) is called an error of type II. Thus the probability of a type II error is 1 minus the power of the test when H_1 is true. Frequently, it is disconcerting to the student to discover that there are so many names for the same thing. However, since all of them are used in the statistical literature, we feel obligated to point out that" significance level," "size of the critical region," "power of the test when H_0 is true," and" the probability of committing an error of type I" are all equivalent.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور فراس شاكر محمود

CONFIDENCE INTERVALFOR VARIANCE σ^2

In this topic we have two cases

(a): The case of μ is Unknown

Let $\mathbf{X^2 = \frac{(n-1)S^2}{\sigma^2}}$ chi-squared distribution with n - 1 degrees of freedom where **the sample variance** is given by

$$\mathbf{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

and S^2 is a point estimator of σ^2, then the confidence coefficient is

$$\mathbf{P\left(\chi^2{}_{1-\alpha/2} < X^2 < \chi^2{}_{\alpha/2}\right) = 1-\alpha}$$

$$\mathbf{P\left(\chi^2{}_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2{}_{\alpha/2}\right) = 1-\alpha}$$

$$\mathbf{P\left(\frac{(n-1)S^2}{\chi^2{}_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2{}_{1-\alpha/2}}\right) = 1-\alpha}$$

THEOREM : If S^2 is the variance of a random sample of size n from a normal population , a $100(1-\alpha)\%$ **confidence interval** for $\boldsymbol{\sigma^2}$ is

$$\mathbf{\frac{(n-1)S^2}{\chi^2{}_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2{}_{1-\alpha/2}}}$$

(b): The case of μ is Known

$$\mathbf{P\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2{}_{\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2{}_{1-\alpha/2}}\right) = 1-\alpha}$$

ESTIMATING THE RATIO OF TWO VARIANCES

The statistic $\frac{S_1{}^2}{S_2{}^2}$ is called an estimator of $\frac{\sigma_2{}^2}{\sigma_1{}^2}$.

<u>**THEOREM**</u> **:** If $\sigma_1{}^2$ and $\sigma_2{}^2$ are the variances of normal populations, we can establish an interval estimate of $\frac{\sigma_2{}^2}{\sigma_1{}^2}$ by using the statistic the random

variable F has an F-distribution with $r_1 = n - 1$ and $r_2 = m - 1$ degrees of freedom.

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

$$P\left(f_{1-\alpha/2}(r_1, r_2) < F < f_{\alpha/2}(r_1, r_2)\right) = 1 - \alpha$$

where $f_{1-\alpha/2}(r_1, r_2)$ and $f_{\alpha/2}(r_1, r_2)$ are the values of the F-distribution with r_1 and r_2 degrees of freedom, leaving areas of $1 - \alpha/2$ and $\alpha/2$, respectively

$$P\left(f_{1-\alpha/2}(r_1, r_2) < \frac{{\sigma_2}^2 {S_1}^2}{{\sigma_1}^2 {S_2}^2} < f_{\alpha/2}(r_1, r_2)\right) = 1 - \alpha$$

Multiply by $\frac{{S_2}^2}{{S_1}^2}$

$$P\left(\frac{{S_1}^2}{{S_2}^2} \frac{1}{f_{\alpha/2}(r_1, r_2)} < \frac{{\sigma_2}^2}{{\sigma_1}^2} < \frac{{S_1}^2}{{S_2}^2} \frac{1}{f_{1-\alpha/2}(r_1, r_2)}\right) = 1 - \alpha$$

replace the quantity $f_{1-\alpha/2}(r_1, r_2)$ by $\frac{1}{f_{1-\alpha/2}(r_1, r_2)}$ Therefore

$$P\left(\frac{{S_1}^2}{{S_2}^2} \frac{1}{f_{1-\alpha/2}(r_1, r_2)} < \frac{{\sigma_2}^2}{{\sigma_1}^2} < \frac{{S_1}^2}{{S_2}^2} f_{\alpha/2}(r_1, r_2)\right) = 1 - \alpha$$

Confidence Interval for $\frac{{\sigma_2}^2}{{\sigma_1}^2}$.

<u>THEOREM</u>: if ${S_1}^2$ and ${S_2}^2$ are the variances of independent samples of sizes n and m , respectively, from normal populations, then a $100(1 - \alpha)\%$ confidence interval for $\frac{{\sigma_2}^2}{{\sigma_1}^2}$ is

$$\left(\frac{{S_1}^2}{{S_2}^2} \frac{1}{f_{1-\alpha/2}(r_1, r_2)} < \frac{{\sigma_2}^2}{{\sigma_1}^2} < \frac{{S_1}^2}{{S_2}^2} f_{\alpha/2}(r_1, r_2)\right)$$

where $f_{\alpha/2}(r_1, r_2)$ is an f-value with $r_1 = n - 1$ and $r_2 = m - 1$ degrees of freedom, leaving an area of α/2 to the right, and $f_{\alpha/2}(r_1, r_2)$ is a similar f-value with $r_2 = m - 1$ and $r_1 = n - 1$ degrees of freedom.

Example: An optical form purchases for making lenses. A Scum that the refractive index of 20 pieces of glass have variance of 1.20×10^{-9} construct a 95% C I for the population variance.

Solution :

n= 20 , $s^2 = 10^{-9} \times 1.2$

n-1 = 19

1- α = 0.95 => α = 0.05

$\frac{\alpha}{2}$ = 0.05 => 1- $\frac{\alpha}{2}$ = 0.475

$\lambda^2_{\frac{\alpha}{2}} = 32.585$ n= 19 من الجدول حيث

$\lambda^2_{1-\frac{\alpha}{2}} = 8.9066$ n= 19 من الجدول حيث $\left(\frac{n\,s^2}{\lambda 2_{\frac{\alpha}{2}}} \;,\; \frac{n\,s^2}{\lambda 2_{1-\frac{\alpha}{2}}}\right)$

$= (7.304 \times 10^{-5} \;,\; 2.694 \times 10^{-5})$

Example: n = 12 taken from N (μ , σ^2) , $\bar{X}$= 10 , S^2 = 9. Find : a 90 % C.I. for σ^2.

Solution

n = 12 , $\bar{X}$= 10 , $S^2 = 9$

1 - α = 0.9 => α = 0.1 => $\frac{\alpha}{2}$ = 0.05

=> 1 - $\frac{\alpha}{2}$ = 0.025

$\lambda^2_{\frac{\alpha}{2}(0.05)} = 19.675$, $\lambda^2_{1-\frac{\alpha}{2}(0.025)} = 4.575$ from the table of chi square then

$\left(\frac{n\,s^2}{\lambda 2_{\frac{\alpha}{2}}} \;,\; \frac{n\,s^2}{\lambda 2_{1-\frac{\alpha}{2}}}\right) = (5.48,\ 23.606)$

Example: Construct a 95% C I for σ^2 with unknown mean using the following sample :

4.5 , 10.2 , 10.5 , 9.8 , 13.6 , 19.2 , 15.5 , 13.3 , 10.8 , 16.4 .

Solution

$S^2 = \frac{\sum (Xi - \bar{X})}{n-1}$

$1 - \alpha = 0.95$, $n = 10$

$\Rightarrow n - 1 = 9$

$\bar{X} = \frac{\sum Xi}{n} = \frac{123.2}{10}$

$\bar{X} = 12.23$

$S^2 = \frac{152.42}{9}$

$S^2 = 16.935$

$\frac{\alpha}{2} = 0.05$

$1 - \frac{\alpha}{2} = 0.95$

$\lambda^{2\,(\rho)}{}_{(0.05)} = 16.919$

$\lambda^{2\,(\rho)}{}_{(0.95)} = 3.325$

Xi	X - $\bar{X}$	(X - $\bar{X}$)^2
4.5	-7.73	59.75
10.2	-2.03	4.12
10.5	-1.73	2.99
9.8	-20.43	5.9
13	0.77	0.59
19.2	6.97	48.5
15.5	3.27	10.69
13.3	1.02	1.04
10.8	-1.43	2.04
16.4	4.17	17.3
123.2		152.92

$\left(\frac{n\,S^2}{\lambda 2_{\frac{\alpha}{2}}} , \frac{n\,S^2}{\lambda 2_{1-\frac{\alpha}{2}}} \right) = \left(\frac{10\,(16.935)}{16.919} , \frac{10\,(16.935)}{3.325} \right) = (9.038 , 45.90)$

اذا كانت μ معلوم

$\left(\frac{n\,S^2}{\lambda 2_{\frac{\alpha}{2}}} , \frac{n\,S^2}{\lambda 2_{1-\frac{\alpha}{2}}} \right)$

ملاحظة : نأخذ قيمة من الجدول بـ n

مثال : اذا علمت ان تباين عينة عشوائية ذات حجم ٢٥ مسحوبة من $N(10 , \sigma^2)$ وكان $S^2 = 9$ جد بمعامل ثقة ٠.٩٥ لتباين هذا المجتمع.

الحل //

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

$1 - \frac{\alpha}{2} = 0.975$

$\lambda^2{}_{\frac{25}{0.025}} = 13.1197$, $\lambda^2{}_{\frac{25}{0.975}} = 40.6465$

$(\frac{n\, s2}{\lambda 2_{\frac{\alpha}{2}}} \;,\; \frac{n\, s2}{\lambda 2_{1-\frac{\alpha}{2}}}) = (5.5355 \;,\; 17.1498)$

Example: A r.v of size 21 ~ N (μ , σ^2) with $S^2 = 9$. Determine 90% C.I. for σ^2

Solution

$n = 21 \quad , \; S^2 = 9$

$1 - \alpha = 0.90 \Rightarrow \alpha = 0.1 \Rightarrow \frac{\alpha}{2} = 0.05$

$1 - \frac{\alpha}{2} = 0.95$

$n = 21 \Rightarrow n\text{-}1 = 20$

$\lambda^2 \, \frac{20}{\frac{\alpha}{2}} = 31.410 \quad , \quad \lambda^2 \, \frac{20}{0.95} = 10.831$

$(\frac{n\, S^2}{\lambda 2_{\frac{\alpha}{2}}} \;,\; \frac{n\, S^2}{\lambda 2_{1-\frac{\alpha}{2}}}) = (\frac{21\,(9)}{31.410} \;,\; \frac{21\,(9)}{10.851}) = (6.017 \;,\; 17.4)$

تقدير ($\mu_1 - \mu_2$) عندما تكون σ^2 مشتركة وغير معلومة

$$((\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}}^{(n1+n2)-2} \; SP \sqrt{\frac{1}{n1} + \frac{1}{n2}}$$

حيث :

$$SP = \sqrt{\frac{(n1-1)\,(S1)^2 + (n2-1)(S2)^2}{(n1+n2)-2}}$$

Example: $n_1=32$, $\bar{X}_1 = 72$, $S_1=8$, $n_2 = 32$, $\bar{X}_2 = 72$, $S_2=6$. Constrict a 99% C. I. from the difference of mean (Assume S.D are equal)

Solution: $SP = \sqrt{\frac{(n1-1)\,(S1)^2 + (n2-1)(S2)^2}{(n1+n2)-2}} = \sqrt{\frac{(32)(64)+32(36)}{62}}$

$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \;,\; t_{\frac{\alpha}{2}}^{(n1+n2)-2} = 7.18$ where $\frac{\alpha}{2} = 0.005$

$= [(72\text{-}70) \pm (2.660)\,(7.18) \sqrt{\frac{1}{64}}] = [\, 2 \pm 19.0988\,(0.156) \,] = [\, 2 \pm 0.298 \,]$

$= [\, 1.702 \,,\, 2.298 \,]$

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Sample size

In statistical consulting, the first question frequently asked is, “How large should the sample size be to estimate a mean?” In order to convince the inquirer that the answer will depend on the variation associated with the random variable under observation, the statistician could correctly respond, “Only one observation is needed, provided that the standard deviation of the distribution is zero.” That is, if σ equals zero, then the value of that one observation would necessarily equal the unknown mean of the distribution. This, of course, is an extreme case and one that is not met in practice; however, it should help convince people that the smaller the variance, the smaller is the sample size needed to achieve a given degree of accuracy. This assertion will become clearer as we consider several examples. Let us begin with a problem that involves a statistical inference about the unknown mean of a distribution.

Example : A mathematics department wishes to evaluate a new method of teaching calculus with a computer. At the end of the course, the evaluation will be made on the basis of scores of the participating students on a standard test. There is particular interest in estimating μ, the mean score for students taking the course.

Thus, there is a desire to determine the number of students, n, who are to be selected at random from a larger group of students to take the course. Since new computing equipment must be purchased, the department cannot afford to let all of the school’s students take calculus the new way.

In addition, some of the staff question the value of this approach and hence do not want to expose every student to this new procedure. So, let us find the sample size n such that we are fairly confident that $x \pm 1$ contains the unknown test mean μ.

From past experience, it is believed that the standard deviation associated with this type of test is about 15. (The mean is also known when students take the standard calculus course.)

Accordingly, using the fact that the sample mean of the test scores, $\bar{X}$, is approximately $N(\mu, \frac{\sigma^2}{n})$, we see that the interval given by $\bar{X} \pm 1.96(15/\sqrt{n})$ will serve as an approximate 95% confidence interval for μ. That is, we want

$$1.96\left(\frac{15}{\sqrt{n}}\right) = 1$$

or, equivalently,

$$\sqrt{n} = 29.4, \qquad and\ thus \quad n \approx 864.36$$

or $n = 865$ because n must be an integer . It is quite likely that, in the preceding example, it had not been anticipated that as many as 865 students would be needed in this study.

If that is the case, the statistician must discuss with those involved in the experiment whether or not the accuracy and the confidence level could be relaxed some. For example, rather than requiring $x \pm 1$ to be a 95% confidence interval for μ, possibly $x \pm 2$ would be a satisfactory 80% one. If this modification is acceptable, we now have

$$1.282(\frac{15}{\sqrt{n}}) = 2$$

or, equivalently,

$$\sqrt{n} = 9.615, \qquad so\ that \quad n \approx 92.4$$

Since n must be an integer, we would probably use 93 in practice. Most likely, the persons involved in the project would find that a more reasonable sample size.

Of course, any sample size greater than 93 could be used. Then either the length of the confidence interval could be decreased from $x \pm 2$ or the confidence coefficient could be increased from 80%, or a combination of both approaches could be taken. Also, since there might be some question as to whether the standard deviation σ actually equals 15, the sample standard deviation s would no doubt be used in the construction of the interval. For instance, suppose that the sample characteristics observed are

$$n = 145, \qquad \bar{X} = 77.2, \qquad s = 13.2;$$

Then

$$\bar{X} \pm \frac{1.282s}{\sqrt{n}}, \quad or \quad 77.2 \pm 1.41.$$

provides an approximate 80% confidence interval for μ. In general, if we want the $100(1-\alpha)\%$ confidence interval for μ, $\bar{X} \pm Z_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$, to be no longer than that given by $\bar{X} \pm \varepsilon$, then the sample size n is the solution of

$$\varepsilon = Z_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$$

That

$$n = \frac{Z^2_{\frac{\alpha}{2}} \sigma^2}{\varepsilon^2},$$

where it is assumed that σ^2 is known. We sometimes call $\varepsilon = z_{\alpha/2}\,(\sigma/\sqrt{n})$ the **maximum error of the estimate**. If the experimenter has no idea about the value of σ^2, it may be necessary to first take a preliminary sample to estimate σ^2. The type of statistic we see most often in newspapers and magazines is an estimate of a proportion p.

We might, for example, want to know the percentage of the labor force that is unemployed or the percentage of voters favoring a certain candidate. Sometimes extremely important decisions are made on the basis of these estimates.

If this is the case, we would most certainly desire short confidence intervals for p with large confidence coefficients. We recognize that these conditions will require a large sample size. If, to the contrary, the fraction p being estimated is not too important, an estimate associated with a longer confidence interval with a smaller confidence coefficient is satisfactory, and in that case a smaller sample size can be used.

Example : Suppose we know that the unemployment rate has been about 8% (0.08). However, we wish to update our estimate in order to make an important decision about the national economic policy.

Accordingly, let us say we wish to be 99% confident that the new estimate of p is within 0.001 of the true p. If we assume Bernoulli trials (an assumption that might be questioned), the relative frequency y/n, based upon a large sample size n, provides the approximate 99% confidence interval:

$$\frac{y}{n} \pm 2.576\sqrt{\frac{\left(\frac{y}{n}\right)\left(1-\frac{y}{n}\right)}{n}}.$$

Although we do not know y/n exactly before sampling, since y/n will be near 0.08, we do know that

$$\frac{y}{n} \pm 2.576\sqrt{\frac{\left(\frac{y}{n}\right)\left(1-\frac{y}{n}\right)}{n}} \approx 2.576\sqrt{\frac{(0.08)(0.92)}{n}},$$

or, equivalently,

$$\sqrt{n} = 2576\sqrt{0.0736}, \qquad \text{and then} \qquad \text{n} \approx 488{,}394.$$

That is, under our assumptions, such a sample size is needed in order to achieve the reliability and the accuracy desired. Because *n* is so large, we would probably be willing to increase the error, say, to 0.01, and perhaps reduce the confidence level to 99%. In that case,

$$\sqrt{n} = \left(\frac{2.326}{0.01}\right)\sqrt{0.0736} \quad \textit{and} \quad n \approx 3{,}982\,,$$

which is a more reasonable sample size. From the preceding example, we hope that the student will recognize how important it is to know the sample size (or the length of the confidence interval and the confidence coefficient) before he or she can place much weight on a statement such as "Fifty-one percent of the voters seem to favor candidate A, 46% favor candidate B, and 3% are undecided." Is this statement based on a sample of 100 or 2000 or 10,000 voters? If we assume Bernoulli trials, the approximate 95%

confidence intervals for the fraction of voters favoring candidate A in these cases are, respectively, [0.41, 0.61], [0.49, 0.53], and [0.50, 0.52].

Quite obviously, the first interval, with n = 100, does not assure candidate A of the support of at least half the voters, whereas the interval with n = 10,000 is more convincing.

In general, to find the required sample size to estimate p, recall that the point estimate of p is $\hat{p} = y/n$ and an approximate $1 - \alpha$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Suppose we want an estimate of p that is within ε of the unknown p with $100(1-\alpha)\%$ confidence, where $\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the maximum error of the point estimate $\hat{p} = y/n$.

Since $\hat{p}$ is unknown before the experiment is run, we cannot use the value of $\hat{p}$ in our determination of n. However, if it is known that p is about equal to p^* , the necessary sample size n is the solution of

$$\varepsilon = \frac{z_{\frac{\alpha}{2}}\sqrt{p^*(1-p^*)}}{\sqrt{n}}.$$

That is ,

$$n = \frac{z_{\alpha/2}^2\, p^*(1-p^*)}{\varepsilon^2} \leq \frac{z_{\alpha/2}^2}{4\varepsilon^2}.$$

Often, however, we do not have a strong prior idea about p, as we did in Example 2 about the rate of unemployment. It is interesting to observe that no matter what value p takes between 0 and 1, it is always true that $p^*(1 - p^*) \le 1/4$. Hence,

$$n = \frac{z_{\frac{\alpha}{2}}^2 p^*(1 - P^*)}{\varepsilon^2} \le \frac{z_{\alpha/2}^2}{4\varepsilon^2}.$$

Thus, if we want them $100(1 - \alpha)\%$ confidence interval for p to be no longer than $y/n \pm \varepsilon$, a solution for n that provides this protection is

$$n = \frac{z_{\alpha/2}^2}{4\varepsilon^2}$$

Remark: Up to this point in the text, we have used the "hat"($\frown$) notation to **indicate** an estimator, as in $\widehat{p} = Y/n$ and $\widehat{\mu} = \bar{X}$

Note, however, that in the previous discussion we used $\widehat{p} = y/n$, an estimate of p. Occasionally, statisticians find it convenient to use the "hat" notation for an estimate as well as an estimator. It is usually clear from the context which is being used.

Example: A possible gubernatorial candidate wants to assess initial support among the voters before making an announcement about her candidacy. If the fraction p of voters who are favorable, without any advance publicity, is around 0.15, the candidate will enter the race. From a poll of n voters selected at random, the candidate would like the estimate y/n to be within 0.03 of p. That is, the decision will be based on a 95% confidence interval of the form $y/n \pm 0.03$. Since the candidate has no idea about the magnitude of p, a consulting statistician formulates the equation

$$n = \frac{(1.96)^2}{4(0.03)^2} = 1067.11 \ .$$

Thus, the sample size should be around 1068 to achieve the desired reliability and accuracy. Suppose that 1068 voters around the state were selected at random and interviewed and $y = 214$ express support for the candidate.

Then $\widehat{p} = \frac{214}{1068} = 0.20$ is a point estimate of p, and an approximate 95% confidence interval for p is

$$0.20 \pm 1.96\sqrt{(0.20)(0.80)/n}\, , or \ \ 0.20 \pm 0.024.$$

That is, we are 95% confident that p belongs to the interval $[0.176, 0.224]$. On the basis of this sample, the candidate decided to run for office. Note that, for a confidence coefficient of 95%, we found a sample size so that the maximum error of the estimate would be 0.03. From the data that were collected, the maximum error of the estimate is only 0.024. We ended up with a smaller error because we found the sample size assuming that $p = 0.50$, while, in fact, p is closer to 0.20.

Suppose that you want to estimate the proportion p of a student body that favors a new policy. How large should the sample be? If p is close to $1/2$ and you want to be 95% confident that the maximum error of the estimate is $\varepsilon = 0.02$, then

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401 \ .$$

Such a sample size makes sense at a large university. However, if you are a student at a small college, the entire enrollment could be less than 2401.

Thus, we now give a procedure that can be used to determine the sample size when the population is not so large relative to the desired sample size. Let N equal the size of a population, and assume that N_1 individuals in the population have a certain characteristic C.

Let $p = N_1/N$, the proportion with this characteristic. Then $1-p = 1-N_1/N$. If we take a sample of size n without replacement, then X, the number of observations with the characteristic C, has a hyper geometric distribution. The mean and variance of X are, respectively

$$\mu = n\left(\frac{N_1}{N}\right) = np\,.$$

and

$$\sigma^2 = n\left(\frac{N_1}{N}\right)\left(1-\frac{N_1}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right).$$

The mean and variance of X/n are, respectively,

$$E\left(\frac{X}{n}\right) = \frac{\mu}{n} = p$$

and

$$Var\left(\frac{X}{n}\right) = \frac{\sigma^2}{n^2} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

To find an approximate confidence interval for p, we can use the normal approximation :

$$P\left[-z_{\frac{\alpha}{2}} \le \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \le z_{\frac{\alpha}{2}}\right] \approx 1 - \alpha\,.$$

Thus,

$$1 - \alpha \approx P\left[\frac{X}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \le p \right.$$
$$\left. \le \frac{X}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\right].$$

Replacing p under the radical with $\widehat{p} = x/n$, we find that an approximate 1 – α confidence interval for p is

$$\widehat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}\left(\frac{N-n}{N-1}\right)}.$$

This is similar to the confidence interval for p when the distribution of X is $b(n, p)$. If N is large relative to n, then

$$\frac{N-n}{N-1} = \frac{1 - n/N}{1 - 1/N} \approx 1.$$

so in this case the two intervals are essentially equal.

Suppose now that we are interested in determining the sample size n that is required to have 1 - α confidence that the maximum error of the estimate of p is ε. We let

$$\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

and solve for n. After some simplification, we obtain

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

$$= \frac{z_{\frac{\alpha}{2}}^2 p(1-p)/\varepsilon^2}{\frac{(N-1)}{N} + \frac{z_{\frac{\alpha}{2}}^2 p(1-p)}{N}}$$

If we let

$$m = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2}$$

which is the n value given by Equation 2, then we choose

$$n = \frac{m}{1 + \frac{m-1}{N}}$$

for our sample size n .

If we know nothing about p, we set $p^* = 1/2$ to determine m. For example, if the size of the student body is $N = 4000$ and $1 - \alpha = 0.95$, $\varepsilon = 0.02$, and we let $p^* = 1/2$, then $m = 2401$ and $n = \frac{2401}{1+2400/4000} = 1501$, rounded up to the nearest integer. Thus, we would sample approximately 37.5% of the student body

Example: Suppose that a college of $N = 3000$ students is interested in assessing student support for a new form for teacher evaluation.

To estimate the proportion p in favor of the new form, how large a sample is required so that the maximum error of the estimate of p is $\varepsilon = 0.03$ with 95% confidence? If we assume that p is completely unknown, we use $p^* = 1/2$ to obtain

$$m = \frac{(1.96)^2}{4(0.03)^2} = 1068\,,$$

rounded up to the nearest integer. Thus, the desired sample size is

$$n = \frac{1068}{1 + 1067/3000} = 788,$$

rounded up to the nearest integer.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

Lecture Note On Mathematical Statistics 2

B.Sc. in Mathematics

Fourth Stage

Assist. Prof. Dr. Feras Shaker Mahmood

•

CONFIDENCE INTERVALS FOR PROPORTIONS

We have suggested that the histogram is a good description of how the observations of a random sample are distributed. We might naturally inquire about the accuracy of those relative frequencies (or percentages) associated with the various classes. To illustrate, concerning the weights of $n = 40$ candy bars, we found that the relative frequency of the class interval (22.25, 23.15) was $8/40 = 0.20$, or 20%. If we think of this collection of 40 weights as a random sample observed from a larger population of candy bar weights, how close is 20% to the true percentage (or 0.20 to the true proportion) of weights in that class interval for the entire population of weights for this type of candy bar?

In considering this problem, we generalize it somewhat by treating the class interval (22.25, 23.15) as "success." That is, there is some true probability of success, p—namely, the proportion of the population in that interval. Let Y equal the frequency of measurements in the interval out of the n observations, so that (under the assumptions of independence and constant probability p) Y has the binomial distribution $b(n, p)$. Thus, the problem is to determine the accuracy of the relative frequency Y/n as an estimator of p. We solve this problem by finding, for the unknown p, a confidence interval based on Y/n.

In general, when observing n Bernoulli trials with probability p of success on each trial, we shall find a confidence interval for p based on Y/n, where Y is the number of successes and Y/n is an unbiased point estimator for p.

In Section 5.7, we noted that

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution $N(0, 1)$, provided that n is large enough. This means that, for a given probability $1 - \alpha$, we can find a $z_{\alpha/2}$
such that

$$P\left[-z_{\alpha/2} \le \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right] \approx 1 - \alpha.$$

If we proceed as we did when we found a confidence interval for μ we would obtain

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1 - \alpha.$$

Unfortunately, the unknown parameter p appears in the endpoints of this inequality. There are two ways out of this dilemma. First, we could make an additional approximation, namely, replacing p with Y/n in $p(1-p)/n$ in the endpoints. That is, if n is large enough, it is still true that

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\right] \approx 1 - \alpha.$$

Thus, for large n, if the observed Y equals y, then the interval

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for p. Frequently, this interval is written as

$$\frac{y}{n} \pm z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}$$

for brevity. This formulation clearly notes, as does $\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ the reliability of the estimate y/n, namely, that we are $100(1-\alpha)\%$ confident that p is within $z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n}$ of $\widehat{p} = y/n$.

A second way to solve for p in the inequality in Equation 7.3-1 is to note that

$$\frac{|Y/n - p|}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}$$

is equivalent to

$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2\, p(1-p)}{n} \le 0.$$

But $H(p)$ is a quadratic expression in p. Thus, we can find those values of p for which $H(p) \le 0$ by finding the two zeros of $H(p)$. Letting $\widehat{p} = Y/n$ and $z_0 = z_{\alpha/2}$ in

$$H(p) = \left(1 + \frac{z_0^2}{n}\right)p^2 - \left(2\widehat{p} + \frac{z_0^2}{n}\right)p + \widehat{p}^{\,2}.$$

By the quadratic formula, the zeros of $H(p)$ are, after simplifications,

$$\frac{\widehat{p} + z_0^2/(2n) \pm z_0\sqrt{\widehat{p}(1-\widehat{p})/n + z_0^2/(4n^2)}}{1 + z_0^2/n},$$

and these zeros give the endpoints for an approximate $100(1-\alpha)\%$ confidence interval for p. If n is large, $z_0^2/(2n)$, $z_0^2/(4n^2)$, and z_0^2/n are small. Thus, the confidence intervals given by Equations 7.3-2 and 7.3-4 are approximately equal when n is large.

Example

Let us return to the example of the histogram of the candy bar weights, with $n = 40$ and $y/n = 8/40 = 0.20$. If $1-\alpha = 0.90$, so that $z_{\alpha/2} = 1.645$, then, using Equation 7.3-2, we find that the endpoints

$$0.20 \pm 1.645\sqrt{\frac{(0.20)(0.80)}{40}}$$

serve as an approximate 90% confidence interval for the true fraction p. That is, $[0.096, 0.304]$, which is the same as $[9.6\%, 30.4\%]$, is an approximate 90% confidence interval for the percentage of weights of the entire population in the interval $(22.25, 23.15)$. If we had used the endpoints , the confidence interval would be $[0.117, 0.321]$. Because of the small sample size, there is a nonnegligible difference in these intervals. If the sample size had been $n = 400$ and $y = 80$, so that $y/n = 80/400 = 0.20$, the two 90% confidence intervals would have been $[0.167, 0.233]$ and $[0.169, 0.235]$, respectively, which differ very little. ■

Example

A possible gubernatorial candidate wants to assess initial support among the voters before making an announcement about her candidacy. If the fraction p of voters who are favorable, without any advance publicity, is around 0.15, the candidate will enter the race. From a poll of n voters selected at random, the candidate would like the estimate y/n to be within 0.03 of p. That is, the decision will be based on a 95% confidence interval of the form $y/n \pm 0.03$. Since the candidate has no idea about the magnitude of p, a consulting statistician formulates the equation

$$n = \frac{(1.96)^2}{4(0.03)^2} = 1067.11.$$

Thus, the sample size should be around 1068 to achieve the desired reliability and accuracy. Suppose that 1068 voters around the state were selected at random and interviewed and $y = 214$ express support for the candidate. Then $\widehat{p} = 214/1068 = 0.20$ is a point estimate of p, and an approximate 95% confidence interval for p is

$$0.20 \pm 1.96\sqrt{(0.20)(0.80)/n}, \qquad \text{or} \qquad 0.20 \pm 0.024.$$

That is, we are 95% confident that p belongs to the interval $[0.176, 0.224]$. On the basis of this sample, the candidate decided to run for office. Note that, for a confidence coefficient of 95%, we found a sample size so that the maximum error of the estimate would be 0.03. From the data that were collected, the maximum error of the estimate is only 0.024. We ended up with a smaller error because we found the sample size assuming that $p = 0.50$, while, in fact, p is closer to 0.20. ■

Suppose that you want to estimate the proportion p of a student body that favors a new policy. How large should the sample be? If p is close to 1/2 and you want to be 95% confident that the maximum error of the estimate is $\varepsilon = 0.02$, then

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401.$$

Such a sample size makes sense at a large university. However, if you are a student at a small college, the entire enrollment could be less than 2401. Thus, we now give a procedure that can be used to determine the sample size when the population is not so large relative to the desired sample size.

Let N equal the size of a population, and assume that N_1 individuals in the population have a certain characteristic C (e.g., favor a new policy). Let $p = N_1/N$, the proportion with this characteristic. Then $1-p = 1-N_1/N$. If we take a sample of size n without replacement, then X, the number of observations with the characteristic C, has a hypergeometric distribution. The mean and variance of X are, respectively,

$$\mu = n\left(\frac{N_1}{N}\right) = np$$

and

$$\sigma^2 = n\left(\frac{N_1}{N}\right)\left(1 - \frac{N_1}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right).$$

The mean and variance of X/n are, respectively,

$$E\left(\frac{X}{n}\right) = \frac{\mu}{n} = p$$

and

$$\mathrm{Var}\left(\frac{X}{n}\right) = \frac{\sigma^2}{n^2} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

To find an approximate confidence interval for p, we can use the normal approximation:

$$P\left[-z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \le z_{\alpha/2}\right] \approx 1-\alpha.$$

Thus,

$$1-\alpha \approx P\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \le p \le \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\right].$$

Replacing p under the radical with $\widehat{p} = x/n$, we find that an approximate $1-\alpha$ confidence interval for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}\left(\frac{N-n}{N-1}\right)}.$$

This is similar to the confidence interval for p when the distribution of X is $b(n, p)$. If N is large relative to n, then

$$\frac{N-n}{N-1} = \frac{1-n/N}{1-1/N} \approx 1,$$

so in this case the two intervals are essentially equal.

Suppose now that we are interested in determining the sample size n that is required to have $1-\alpha$ confidence that the maximum error of the estimate of p is ε. We let

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

and solve for n. After some simplification, we obtain

$$\begin{aligned} n &= \frac{N z_{\alpha/2}^2 p(1-p)}{(N-1)\varepsilon^2 + z_{\alpha/2}^2 p(1-p)} \\ &= \frac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{\dfrac{N-1}{N} + \dfrac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{N}}. \end{aligned}$$

If we let

$$m = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2},$$

which is the n value ; , then we choose

$$n = \frac{m}{1 + \dfrac{m-1}{N}}$$

for our sample size n.

If we know nothing about p, we set $p^* = 1/2$ to determine m. For example, if the size of the student body is $N = 4000$ and $1 - \alpha = 0.95$, $\varepsilon = 0.02$, and we let $p^* = 1/2$, then $m = 2401$ and

$$n = \frac{2401}{1 + 2400/4000} = 1501,$$

rounded up to the nearest integer. Thus, we would sample approximately 37.5% of the student body.

Example

Suppose that a college of $N = 3000$ students is interested in assessing student support for a new form for teacher evaluation. To estimate the proportion p in favor of the new form, how large a sample is required so that the maximum error of the estimate of p is $\varepsilon = 0.03$ with 95% confidence? If we assume that p is completely unknown, we use $p^* = 1/2$ to obtain

$$m = \frac{(1.96)^2}{4(0.03)^2} = 1068,$$

rounded up to the nearest integer. Thus, the desired sample size is

$$n = \frac{1068}{1 + 1067/3000} = 788,$$

rounded up to the nearest integer.

Thanking for your Intention

Ass. Prof. Dr. Feras Shaker Mahmood

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Estimation of Proportion

1- Point Estimation of Proportion :- It makes sense to estimate the percentage of a phenomenon in a society by the percentage of the presence of that phenomenon in a random sample taken from that community. This supported the percentage estimate by placement as we studied it. For example, if you want to estimate the percentage of families who own a car, you can choose a random sample and calculate the percentage of the number of families that own a car and use the percentage in the sample as an estimate of the percentage in the community This introduction clearly shows us that if the pass rate for a binomial experiment is (p), it is possible to estimate (p) as follows: Take a random sample of size and assume that the number of successes in this sample (X) can be used $\overline{p} = \frac{x}{n}$

As an estimate of the success rate (p) and the point estimate of the parameter (p) is the success rate in the sample $\overline{p} = \frac{x}{n}$

Note that this phrase rewards our saying: Take a random sample from the Bernoulli community
Note that this phrase rewards our saying: Take a random sample from the Bernoulli community
b(1,p) Let it be the sample $x_1, x_2, \dots, x_n$. Then ($\overline{p}$) is the number of successes in the sample and is

$$\overline{p} = \frac{x}{n} = \frac{\sum x_i}{n}$$

Note that (X) is the statistic: the number of successes in the sample, but (x) is the value of (X) and we get it from the study of the sample, meaning that (x) is the value of (X) that we get from a specific sample. Examples of a binomial experiment are many, and examples of the need to estimate the success rate (p) are many as well. For example, you need to estimate the proportion of students who use eyeglasses in the tenth grade in a country. You need to estimate the percentage of students in the sixth grade who write in the left hand.

2- Interval Estimation of Proportion:- The estimate of the percentage in a period is to find a point estimate of the success rate in the community (p) and then find a distribution of that estimate and use this information to find a period with a certain confidence factor that limits the success rate (p) within it. If the unknown success rate (p) is not expected to be very close to zero or one and the sample size(n) is large, then you can use the theory that the distribution

$$z = \frac{p - \overline{p}}{\sqrt{\frac{p(1-p)}{n}}}$$

Approach the standard normal distribution if (n)is large. If these conditions are met, you can set the possibility phrase $p\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha$. Where

$$z = \frac{p - \overline{p}}{\sqrt{\frac{p(1-p)}{n}}}$$

Almost subject to standard normal distribution $x, \overline{p} = \frac{x}{n}$

The number of successes in the sample size(n) . It is difficult to use the previous probability statement as given to find a confidence interval for the ratio (p), because (p) is in the first place.

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

Not known , so we use

$$\overline{p} = \frac{x}{n}$$

Instead of (p) in the denominator we get the confidence interval
$100(1-\alpha)\%$

The approximate percentage (p):

$$\overline{p} - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} < p < \overline{p} + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

Ratio confidence interval (p) .

Theorem (4):

If it was $\overline{p} = \frac{x}{n}$ The success rate in a random sample of size (n) and (n) was large, the confidence period $100(1-\alpha)\%$. The approximate pass rate (p), (p) is a binomial parameter, the pass rate in society is:

$$\overline{p} - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} < p < \overline{p} + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

Where $z_{\frac{\alpha}{2}}$. It is the point on the normative natural axis to its left of space $\left(\frac{\alpha}{2}\right)$.

Example: Let us retum to the example of histogram of the candy bar weights examples with n=40, y=8 if [$1-\alpha$ =0.90] so that [$Z_{\frac{\alpha}{2}}$ =1.645] then using confidence intervals for proportions

Solution:-

$$\hat{p} = \frac{y}{n} \to \hat{p} = \frac{8}{40} \to \hat{p} = 0.20$$

$$1 - \hat{p} = 1 - 0.20 = 0.80$$

$$1 - \alpha = 0.90 \to \alpha = 1 - 0.90 \to \alpha = 0.1$$

$$Z_{\frac{\alpha}{2}} = z\frac{0.1}{2} = 0.05 \to Z_{\frac{\alpha}{2}} = 1.645$$

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.20 \pm 0.10403893502$$

$$= [0.20 - 0.10403893502 \le p \le 0.20 + 0.10403893502]$$

$$= [0.0959 \le p \le 0.3040]$$

If n=400 , y=80 , 90%

$$\hat{p} = \frac{80}{400} \to \hat{p} = 0.20 \to 1 - \hat{p} = 0.80$$

$$0.20 \mp (1.645)\sqrt{\frac{(0.20)(0.80)}{400}}$$

$$0.20 \mp 0.0329$$

$$[0.20 - 0.0329 \le p \le 0.20 + 0.0329]$$

Example: If the probability of success of a student studying a mathematics course is 0.9, a sample of 49 students will be taken from those who study this course find $p(\overline{p} \ge 0.8)$.

Solution:-

$$E(\overline{p}) = p = 0.9$$

$$1 - p = 1 - 0.9 = 0.1$$

$$\mathrm{E}(\overline{p}) = 0.9$$

$$\mathrm{var}(\overline{p}) = \frac{p(1-p)}{n} = \frac{(0.9)(0.1)}{49} = 0.0018$$

$$p(\overline{p} \ge 0.8) = p\left[\frac{\overline{p} - 0.9}{\sqrt{0.0018}} \ge \frac{0.8 - 0.9}{\sqrt{0.0018}}\right]$$

$$p[Z \ge -2.38] = 1 - p[Z \le -2.38]$$

$$1 - 0.0087 = 0.9918$$

Confidence Intervals for the difference between two means and the difference between two Proportion

We can use the previous method in building confidence intervals for the mean and for us to find confidence intervals for the difference between two media $(\mu_1 - \mu_2)$ The difference between two ratios and that, using the theories of sampling distributions, is the difference between two means and the difference between two ratios .

Theorem:

Let $x_1, x_2, \ldots, x_{n_1}$ random sample from normal distribution $N(\mu_1, \sigma_1^2)$and $y_1, y_2, \ldots, y_{n_2}$ random sample from normal distribution $N(\mu_2, \sigma_2^2)$ Independent of the first distribution and were σ_2^2, σ_1^2 Two facts, the confidence period $100(1-\alpha)\%$. The difference between the two media $(\mu_1 - \mu_2)$ is :

$$\left[(\bar{x} - \bar{y}) - z_{\frac{\alpha}{2}} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\,,\,(\bar{x} - \bar{y}) + z_{\frac{\alpha}{2}} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right]$$

Theorem:

Let $x_1, x_2, \ldots, x_{n_1}$ random sample from Bernoulli society $b(1, p_1)$ and $y_1, y_2, \ldots, y_{n_2}$ a random sample from an independent community $b(1, p_2)$. The confidence period $100(1-\alpha)\%$ The difference between the two ratios $(p_1 - p_2)$ she is :

$$p\left((\bar{p}_1 - \bar{p}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \le p_1 - p_2 \le (\bar{p}_1 - \bar{p}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}\right) = 1 - \alpha$$

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

Provided that n_1, and n_2 are large, where $\bar{p}_1$ and $\bar{p}_2$ the success rates in the two samples , respectively

Example: If $x \sim N(\mu, 4)$, $n=10$ observations of X.

55.95	56.54	57.58	55.13	57.48	56.06
59.93	58.30	52.57	58.46		

(a) Given a point estimate for μ

(b) Find a 95% confidence for μ

(c) less than 52 grams of candy ?

Solution:

(a) $\overline{x} = \frac{\sum x_i}{n} = \frac{55.95+56.54+52.58+55.13+57.48+\cdots}{10} = 56.8$

(b) $\overline{x} \pm Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma^2}{n}}\right)$

$1-\alpha = 0.95 \rightarrow \alpha = 1-0.95 \rightarrow \alpha = 0.05 \rightarrow \frac{Z_{0.05}}{2} = Z_{(0.025)}$ where $Z_{(0.025)} = 1.96$

$56.8 \pm (1.96)\sqrt{\frac{4}{10}}$

$[56.8 - 1.2396 \leq \mu \leq 56.8 + 1.2396]$

$[55.5603 \ldots \leq \mu \leq 58.0396 \ldots]$

(c) $p[x < 52] = p\left[\frac{x-\overline{x}}{\sigma} \leq \frac{52-56.8}{2}\right]$ then $p[z \leq -2.4] = 0.0082$

Example: Let $n_1=194$, $n_2=162$, $y_1=28$, $y_2=11$

(a) Given a point estimate of p_1, and p_2.

(b) Find 95% for p_1

(c) Given a point estimate p_1-p_2

(d) Find a 95% for p_1-p_2

Solution:

(a) $\overline{p}_1 = \frac{y_1}{n_1} = p_1 = \frac{28}{194} = 0.144$

$P_2 = \frac{y_2}{n_2} \rightarrow p_2 = \frac{11}{162} \rightarrow p_2 = 0.0679$

(b) $1-\alpha = 0.95 \rightarrow \alpha = 1-0.95 \rightarrow \alpha = 0.05 \rightarrow z_{\frac{\alpha}{2}} = \frac{z_{0.05}}{2} 0.025$

$Z_{(0.025)}=1.96$ then $\overline{p}_1 \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}}$

$0.144 \pm 1.96\left(\sqrt{\frac{0.144(1-0.144)}{194}}\right)$

$[0.09459 \ldots \leq p_1 \leq 0.19340 \ldots]$

(c) $p_1 - p_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2} = \frac{28}{194} - \frac{11}{162}$ then $p_1 - p_2 = 0.076$

(d) $\alpha = 0.05 \rightarrow Z_{\frac{\alpha}{2}} \rightarrow Z_{(0.025)} = 1.96$ then $(\bar{p}_1 - \bar{p}_2) \pm Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}\right)$

$$(0.144 - 0.0679) \pm 1.96\left(\sqrt{\frac{0.144(1-144)}{194} + \frac{0.0679(1-0.0679)}{162}}\right)$$

0.0261 ± 0.0543851 ... then $[0.130485 \ldots \leq p_1 - p_2 \leq 0.0218]$

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء 2

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

A Statistical Test of Hypothesis

A Statistical of hypothesis consists of five parts:

1. The null hypothesis denoted by $H_\circ$.
2. The alternative hypothesis denoted by H_1.
3. The test statistic and its P_value.
4. The rejection region.
5. The conclusion.

<u>Definition:</u>

(1) The Two competing hypothesis are the alternative hypothesis H_1 generally the hypothesis that the researcher wishes to support and the null hypothesis $H_\circ$ a contradiction of the alternative hypothesis.

(2) Test statistic :a single number calculated from the sample data .

(3) P_value :a probability calculated using the test statistic .

<u>Example</u> :. a random sample of 100 California carpenters , you wish to show that the average hourly witness of carpenters in the state of California is different from 14$,which is the national average .

<u>Solution</u> :.

(1) This is alternative hypothesis $H_1 : \mu \neq 14$

The null hypothesis $H_\circ : \mu = 14$

(2) Test statistic let $\bar{X} = 15$ lies

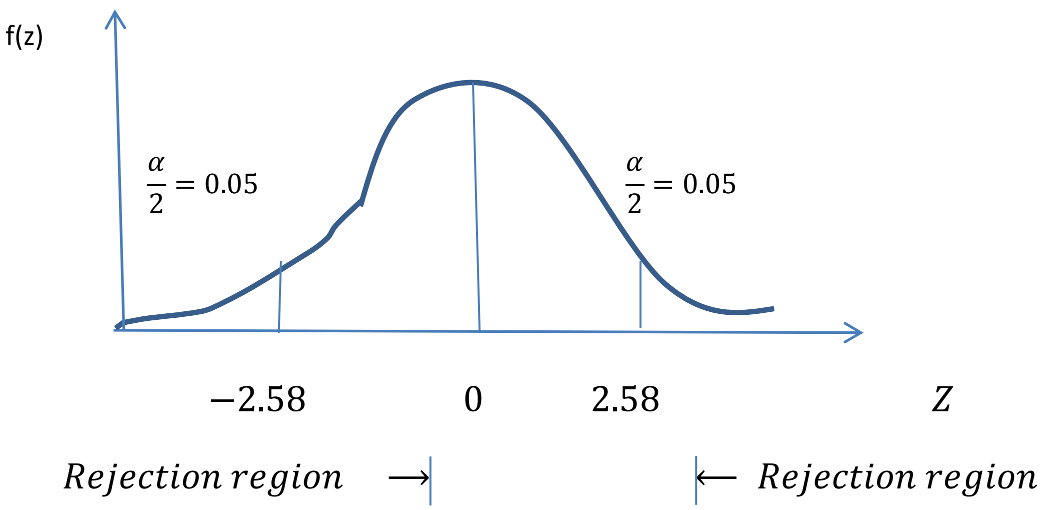
$$\text{S.E} = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{100}} \Longrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{15-14}{2/10} = 5$$

(3) P_value $= P(Z < 5) + P(Z > 5) \cong 0$

Small P_value Large P_value

Definition:

A Type I error for a statistical test is the error of rejecting the null hypothesis when it is true .The level of significance (significance level) for a statistical test of hypothesis is
α =p(Type I error)=p(falsely rejecting $H_\circ$)=p(rejecting $H_\circ$ when it is true)



Large sample statistical Test For μ

(1)Null hypothesis $H_\circ: \mu = \mu_\circ$
(2)Alternative hypothesis $H_1: \mu \neq \mu_\circ$ (Two Tailed test)
$H_1: \mu < \mu_\circ$ or $H_1: \mu > \mu_\circ$ (one Tailed test)

(3)Test statistic : Z= $\frac{\bar{x}-\mu_\circ}{\sigma/\sqrt{n}}$ estimated as Z= $\frac{\bar{x}-\mu_\circ}{s/\sqrt{n}}$
(4)Rejection region :Reject $H_\circ$ when(one Tail test) $Z > Z_\alpha$ or $Z < -Z_\alpha$ when alternative hypothesis $H_1: \mu < \mu_\circ$
(Two Tailed test)$Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$

Calculating the P_value

Definition: The P_value or observed significance level of a statistical test is the smallest value of α for which $H_\circ$ can be rejected . It is the actual risk of committing a Type I error ,if $H_\circ$ is rejected based on the observed value of the test statistic .The P_value measures the strength of the evidence against $H_\circ$.

Definition: If the P_value is less than or equal to α preassigned significance level α , then the null hypothesis can be rejected , and you can report the results are statistically significant at level α .

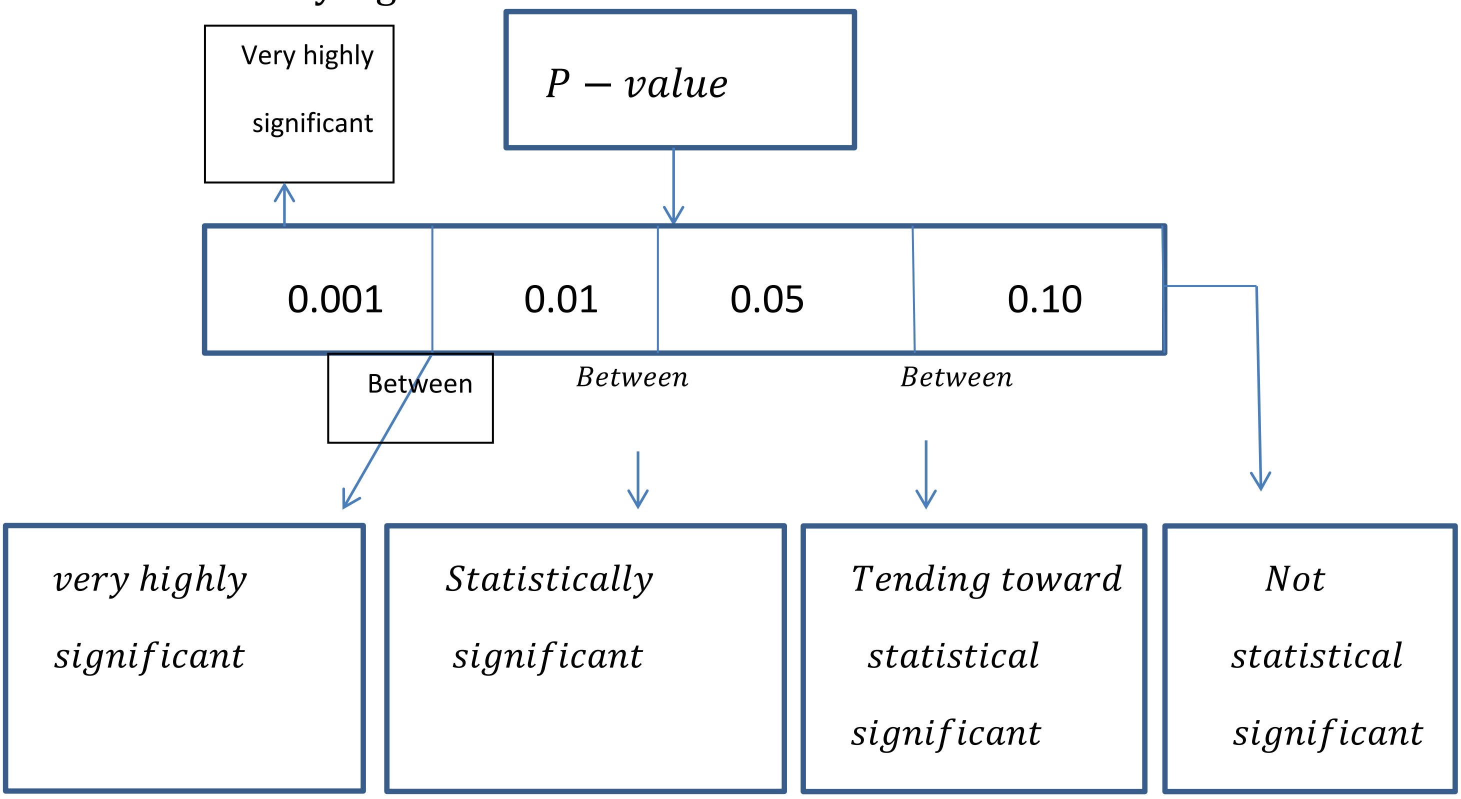
(1)If P_value is less than 0.01,$H_\circ$ is rejected or between 0.01 and 0.001 ,$H_\circ$ is rejected . The result are highly significant

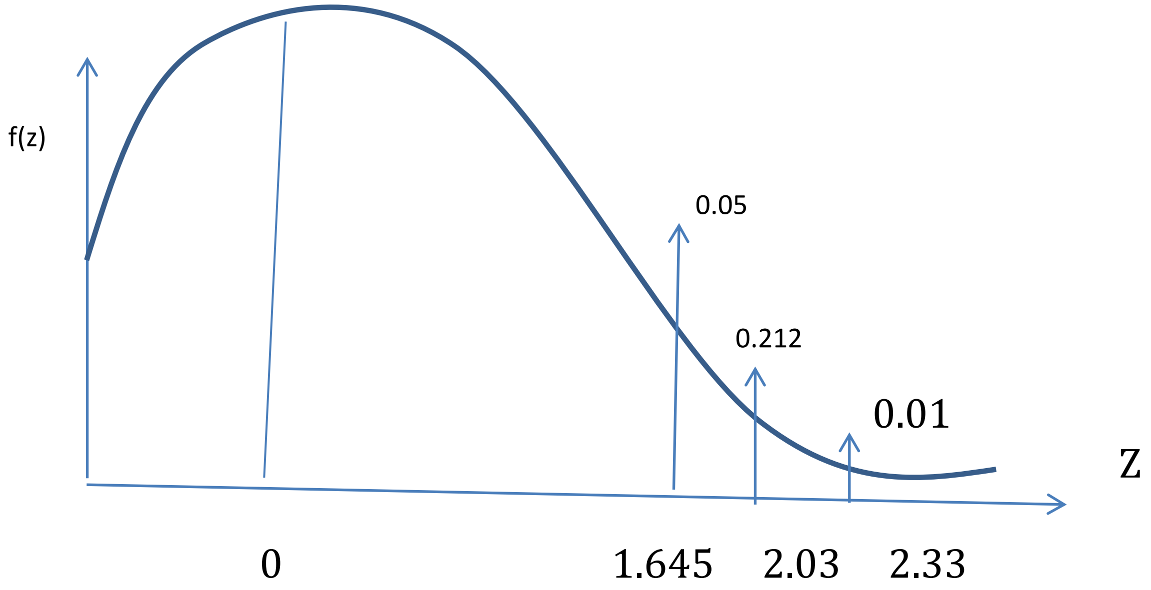
(2) If P_value between 0.01 and 0.05 ,$H_\circ$ is rejected .The result are statistically significant

(3)If P_value is greater than 0.001,the result are very highly significant

(4)If the P_value is between 0.05 and 0.10, $H_\circ$ is usually not rejected .The result are only tending toward statistical significance.

(5)If the P_value greater than 0.10 , $H_\circ$ is not rejected the result are not statistically significant.





Standard normal test Z

Two Type of Errors

Definition: :

(1)A Type I error for a statistical test is the error of rejecting $H_\circ$ when it is true. The probability of making a type I error is denoted by the symbol α .

(2)A Type II error for a statistical test is the error of accepting $H_\circ$ when it is false and some H_1 is true . The probability of making a Type II error is denoted by the symbol β.

(3)The power of a statistical test , given as $1-\beta$ =p(reject $H_\circ$ *when* H_1 *is true*) measures the ability of the test to perform as required .

Large Sample Statistical Test For$(\mu_1-\mu_2)$

(1)Null hypothesis :$H_\circ: \mu_1-\mu_2 = D_\circ$, where $D_\circ$ is some specified difference that you wish to test it .For many test $D_\circ = 0$ there is no difference between μ_1 *and* μ_2

(2)Alternative hypothesis:

(one Tailed Test)$H_1: \mu_1-\mu_2 > D_\circ$ *or* $H_1: \mu_1-\mu_2 < D_\circ$

(Two Tailed Test) $H_1: \mu_1-\mu_2 \neq D_\circ$

(3)Test statistic:

$$Z= \frac{(\overline{X}_1-\overline{X}_2)-D_\circ}{S.E} = \frac{(\overline{X}_1-\overline{X}_2)-D_\circ}{\frac{S_1}{\sqrt{n_1}}+\frac{S_2}{\sqrt{n_2}}}$$

(4)Rejection region : Reject $H_\circ$ when

(one Tailed Test)

$Z > Z_\alpha$ or $Z < -Z_\alpha$ when $H_1: (\mu_1-\mu_2 < D_\circ)$or when P_value $< \alpha$

(Two Tailed Test)

$Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$

Large Sample Test of Hypothesis for A Binomial Proportion

(1)Null hypothesis: $H_\circ : P = P_\circ$

(2)Alternative hypothesis:

(One Tailed Test):

$H_1 : P > P_\circ$ or $H_1 : P < P_\circ$

$(Two\ Tailed\ Test)$:

$H_1 : P \neq P_\circ$

(3)Test statistic :

$Z = \frac{\bar{P} - P_\circ}{S.E} = \frac{\bar{P} - P_\circ}{\sqrt{\frac{P_\circ(1-P_\circ)}{n}}}$ $with$ $\bar{P} = \frac{x}{n}$

Where x is the number of successes in n binomial trials .

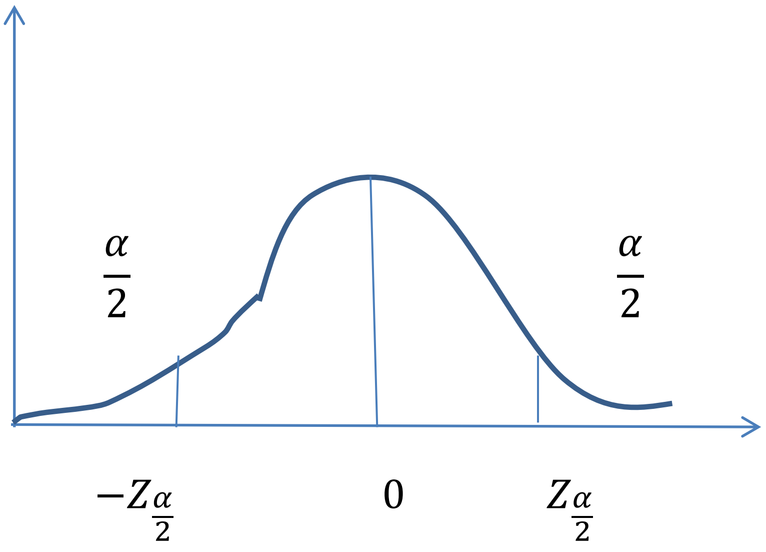
(4)Rejection region : Reject $H_\circ$ when

(One Tailed Test)

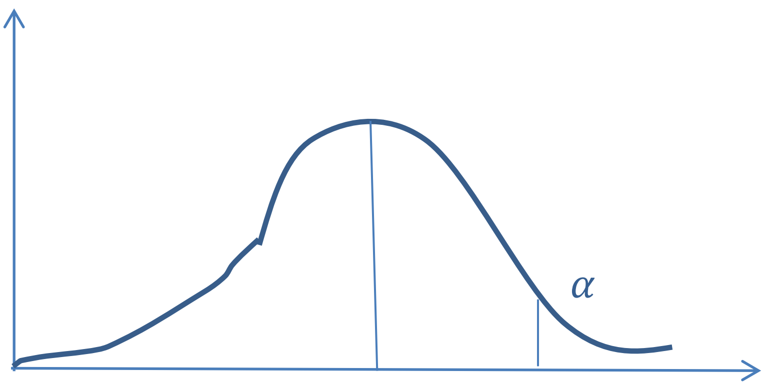
$Z > Z_\alpha$ or $Z < -Z_\alpha$ $when$ $H_1 : P < P_\circ$ or $when$ $P\_value < \alpha$

$(Two\ Tailed\ Test)$:

$Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$



)



Large Sample Statistical Test For $(P_1 - P_2)$

(1)Null hypothesis : $H_\circ: P_1 - P_2 = 0 \; or \; H_\circ: P_1 = P_2$

(2)Alternative hypothesis:

(One Tailed Test):

$H_1: P_1 - P_2 > 0 \; or \quad H_1: P_1 - P_2 < 0$

(Two Tailed Test):

$H_1: P_1 - P_2 \neq 0$

(3)Test Statistic:

$$Z = \frac{(\overline{P_1} - \overline{P_2}) - 0}{S.E} = \frac{(\overline{P_1} - \overline{P_2}) - 0}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$

where $\overline{P_1} = \frac{x_1}{n_1}$ and $\overline{P_2} = \frac{x_2}{n_2}$ $, since \; P_1 = P_2 = P$ used the s.e is unknown, it is estimated by $\hat{P} = \frac{x_1+x_2}{n_1+n_2}$ then the test statistic

$$Z = \frac{(\widehat{P_1} - \widehat{P_2}) - 0}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n_1} + \frac{\hat{P}(1-\hat{P})}{n_2}}} = \frac{(\widehat{P_1} - \widehat{P_2})}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

(4)Rejection Region : Reject $H_\circ$ when

(One Tailed Test):

$Z > Z_\alpha \; or \; Z < -Z_\alpha \; when \; H_1: P_1 - P_2 < 0 \quad or \; when \; P\_value \; < \alpha$

(Tow Tailed Test):

$Z > Z_{\frac{\alpha}{2}} \quad or \quad Z < -Z_{\frac{\alpha}{2}}$

Small Sample Hypothesis Test For μ

(1) Null hypothesis: $H_\circ : \mu = \mu_\circ$

(2) Alternative hypothesis:

(One Tailed Test):

$H_1 : \mu > \mu_\circ \quad or \quad H_1 : \mu < \mu_\circ$

(Two Tailed Test):

$H_1 : \mu \neq \mu_\circ$

(3)Test Statistic :

$$t = \frac{\bar{x} - \mu_\circ}{s/\sqrt{n}}$$

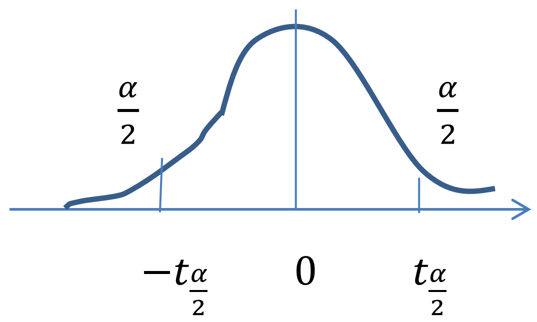
(4)Rejection Region: Reject , $H_\circ$ when

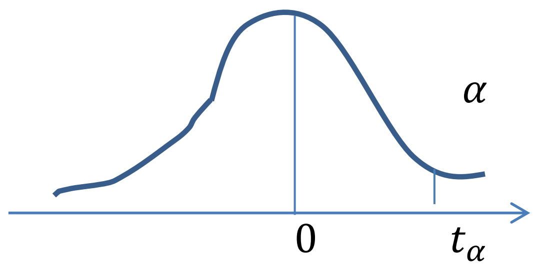
(One Tailed Test):

$t > t_\alpha \quad or \quad t < -t_\alpha \;\; when\; H_1 : \mu < \mu_\circ \quad or\; when\; P\_value < \alpha$

(Two Tailed Test):

$t > t_{\frac{\alpha}{2}} \quad or \quad t < -t_{\frac{\alpha}{2}}$





Test of Hypothesis Concerning The difference between Two means : Independent Random Samples.

(1)Null hypothesis :$H_\circ: \mu_1 - \mu_2 = D_\circ$, where $D_\circ$ is some specified difference that you wish to test it . For many test $D_\circ = 0$ there is no difference between μ_1 *and* μ_2

(2)Alternative hypothesis:

(one Tailed Test)$H_1: \mu_1 - \mu_2 > D_\circ$ *or* $H_1: \mu_1 - \mu_2 < D_\circ$

(Two Tailed Test) $H_1: \mu_1 - \mu_2 \neq D_\circ$

(3)Test Statistic :

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D_\circ}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $s^2 = \frac{(n_1-1){s^2}_1 + (n_2-1){s^2}_2}{n_1+n_2-2}$

(4)Rejection region : Reject $H_\circ$ when

(one Tailed Test):

$t > t_\alpha$ or $t < -t_\alpha$ when $H_1 : (\mu - \mu_\circ) < D_\circ$ or when P_value $< \alpha$

(Two Tailed Test):

$$t > t_{\frac{\alpha}{2}} \quad or \quad t < -t_{\frac{\alpha}{2}}$$

Test of hypothesis Concerning a population Variance

(1)Null hypothesis :

$H_\circ : \sigma^2 = {\sigma^2}_\circ$

(2)Alternative hypothesis :

(One Tailed Test):

$H_1: \sigma^2 > {\sigma^2}_\circ$ *or* $H_1: \sigma^2 < {\sigma^2}_\circ$

(Two Tailed Test):

$\sigma^2 \neq {\sigma^2}_\circ$

(3)Test statistic :

$$x^2 = \frac{(n-1)s^2}{{\sigma^2}_\circ}$$

(4)Rejection Region :Reject $H_\circ$ when

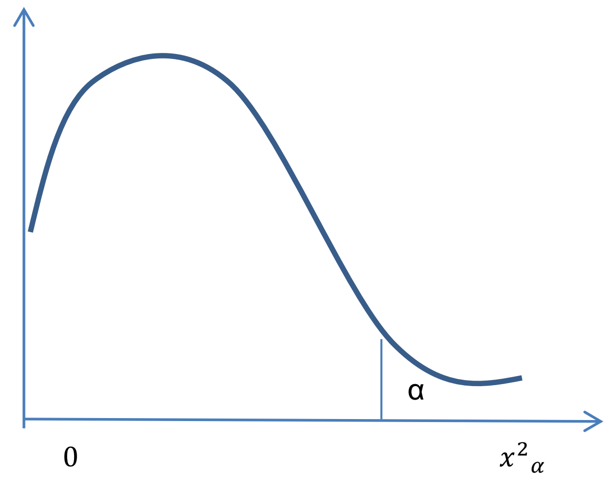
(One Tailed Test):

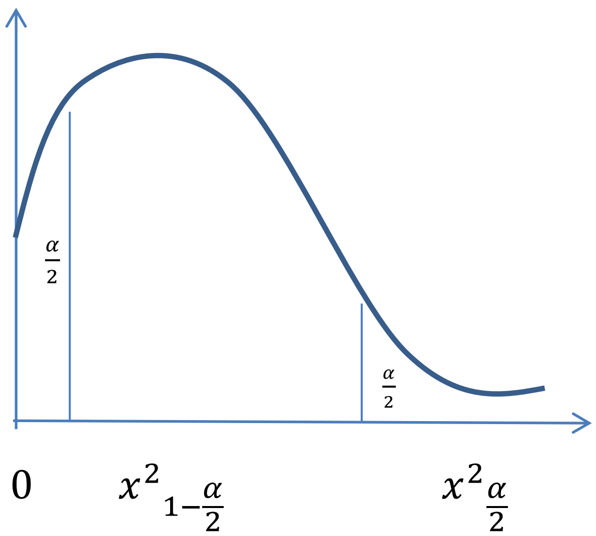
$x^2 > {x^2}_\alpha$ *or* $x^2 < {x^2}_{(1-\alpha)}$ *when* $H_1: \sigma^2 < {\sigma^2}_\circ$

(Two Tailed Test):

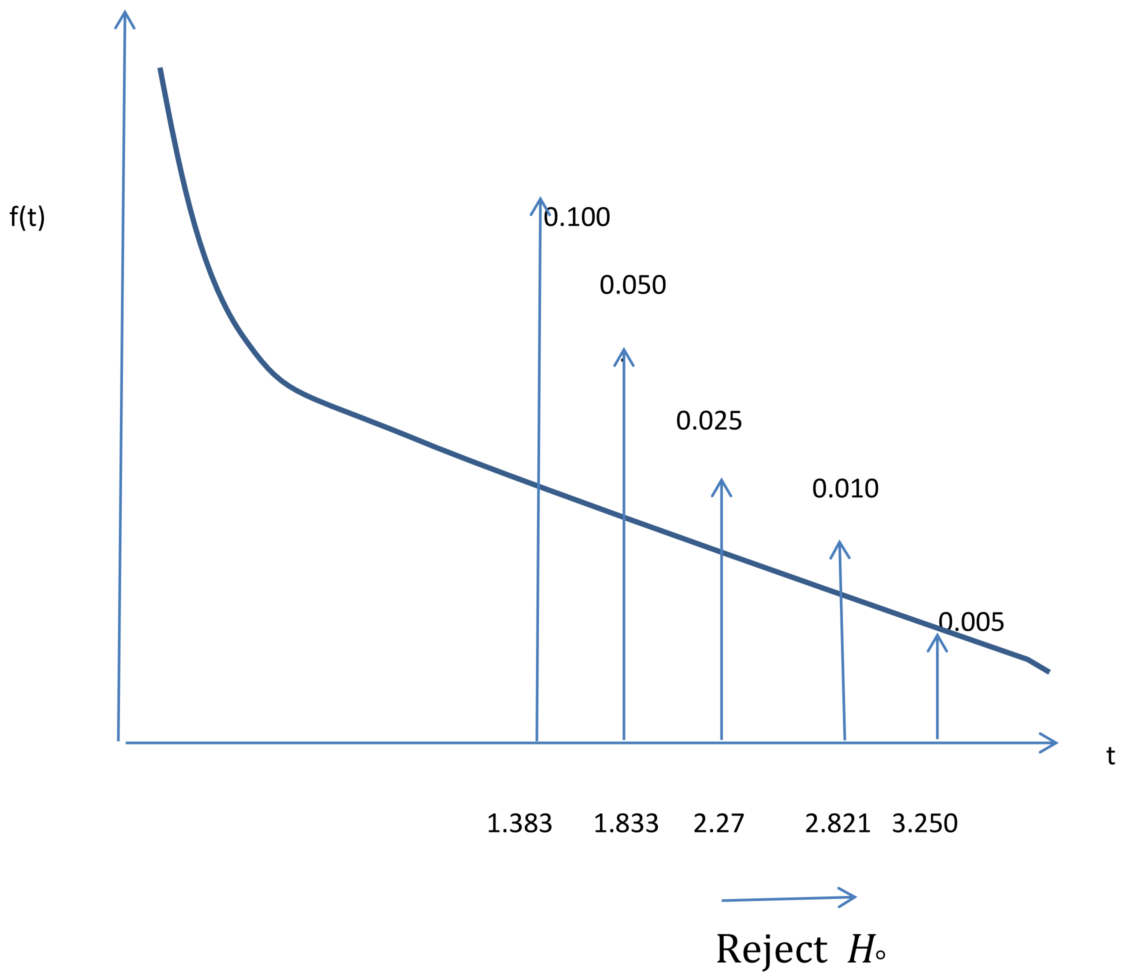
$x^2 > {x^2}_{\frac{\alpha}{2}}$ *or* $x^2 < {x^2}_{(1-\frac{\alpha}{2})}$

upper *lower*

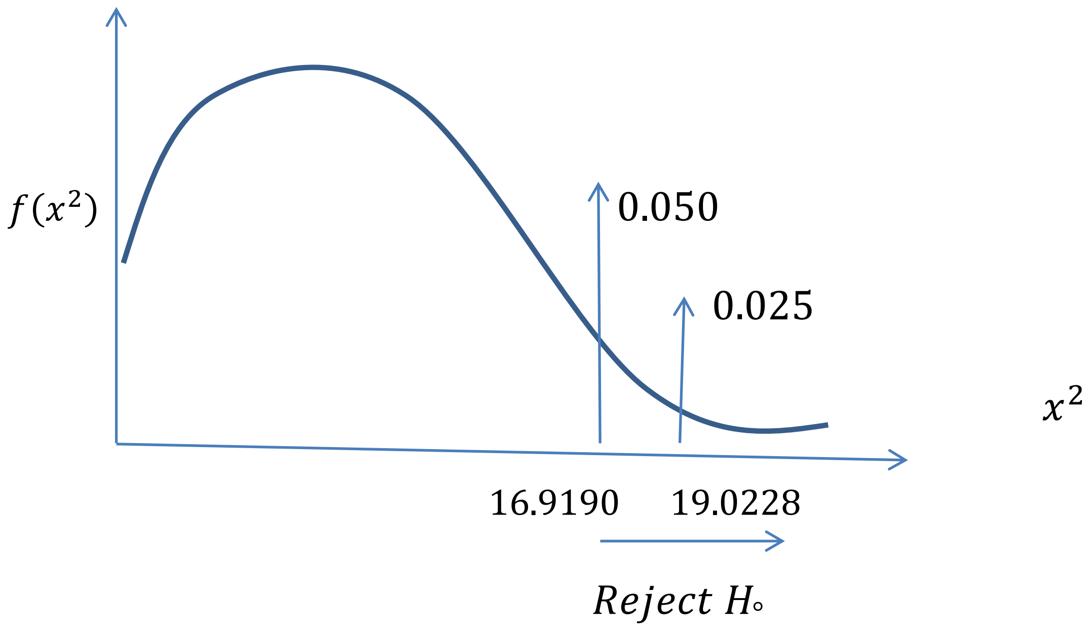




$t - student\ test$



$chi - square\ test$



Test of hypothesis Concerning the equality of two population variances

(1)Null hypothesis :

$H_\circ : \sigma^2{}_1 = \sigma^2{}_2$

(2)Alternative hypothesis :

(One Tailed Test):

$H_1: \sigma^2{}_1 > \sigma^2{}_2 \quad or \quad H_1: \sigma^2{}_1 < \sigma^2{}_2$

(Two Tailed Test):

$\sigma^2{}_1 \neq \sigma^2{}_2$

(3) Test Statistic :

(One Tailed Test):

$F_\propto = \frac{s^2{}_1}{s^2{}_2}$

(Two Tailed Test):

$F_{\frac{\propto}{2}} = \frac{s^2{}_1}{s^2{}_2}$

where $s^2{}_1$ is the larger sample variance .

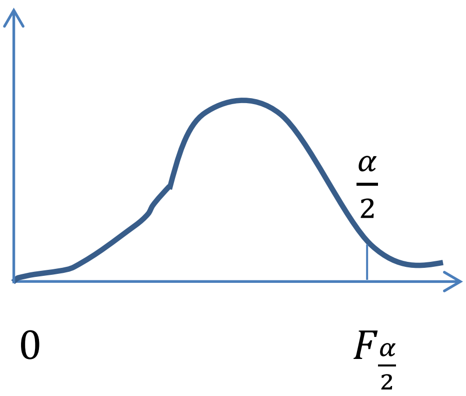
(4) Rejection Region :Reject $H_\circ$ when

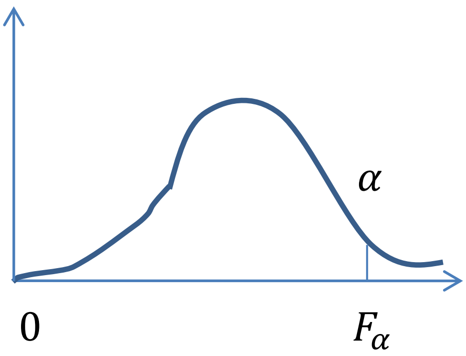
(One Tailed Test):

$F > F_\alpha$ or when P_value $< \alpha$

(Two Tailed Test):

$F > F_{\frac{\alpha}{2}}$





Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Large sample statistical test for μ

Example: The mean arithmetic for metal wires is 1800 km, and its standard deviation is 150 in a sample test of 40 wires, showing that the mean is 1840 km. What is the imposition in the significant level?

Solution:

1- Formulating the hypothesis on the two parties because it was not specified

$H_0 = \mu = 1800$

$H_1 = \mu \neq 1800$

2- The level of significance

$\alpha = 0.05$

3- Test function, including α known

$$Z_{\frac{\alpha}{2}}$$

$$Z = \frac{\overline{Z} - \mu}{\frac{\sigma}{\sqrt{n}}} \implies Z = \frac{1840 - 1800}{\frac{150}{\sqrt{40}}} \implies Z = 1.7$$

4- We calculate the statistical scale of the test since is the data on Z

$\pm Z_{\frac{\alpha}{2}} \implies \left[-Z_{\frac{0.05}{2}}, Z_{\frac{0.05}{2}}\right]$ = [-1.96 ,1.96]

5- Decision making

$$1.7 \epsilon [-1.96, 1.96]$$

Accept H_o

Small sample Hypothesis test for μ

Example : The director of a statistical studies company relies that the average monthly expenditure on food in specific homes equals 290 Iraqi dinars, so if he takes a random sample 10 from the houses we show that its mean $\overline{X} = 296$, standard deviation S = 5, then can this sample be used to confirm what the hypothesis of using confidence level 95%.

Solution :

Since σ is then unknown on the sample size n <30 permission on the t-distribution $\overline{X}$=296 $\mu = 290$ $, n = 10$

$$S = 5 \quad , \quad 1 - \alpha = 0.95$$

1- Formulation of hypotheses

$H_0 = \mu = 290$

$H_1 = \mu \neq 290$

2- We find the level of significance α

$1 - \alpha = 0.95 \Rightarrow \alpha = 1 - 0.95 \Rightarrow \alpha = 0.05$

2- We find the test function S

$$t = \frac{\overline{Z} - \mu}{\frac{S}{\sqrt{n}}} \Rightarrow t = \frac{296 - 290}{\frac{5}{\sqrt{10}}}$$

$t = 3.8$

4- We find the statistical scale for selection

Since σ is unknown n = 10≤30

$\pm t_{\frac{\alpha}{2}, n-1}$

$$\left[\left(\frac{-t_{(0.05)}}{2}, 9\right), \left(\frac{t_{(0.05)}}{2}, 9\right)\right]$$

$= [-t(0.025,9), t(0.025,9)]$

Find a table for t 9 with 0.025 and be 2.262

[-2.262 ,2.262]

5- Decision-making

3.8 ∉ [-2.262 ,2.262] Accept the alternative hypothesis H_1

Large sample statistical test for $(\mu_1 - \mu_2)$

Example : I took a random sample size of 100 from the marks of successful female students in the general secondary exam, and I gave an arithmetic mean $\overline{X}_1 = 68.5$ and the sample variance $S_1^2 = 100$ and took a sample size of 150 from the male students who passed the high school exams and was given an arithmetic mean $\overline{X}_2 = 66.9$ and variance $S_2^2 = 144$ .Test at the sample level α = 0.05.

Solution:

$n_1 = 100 \;, n_2 = 150 \,, s_1^2 = 100 \,, s_2^2 = 144$

1-Formulate the hypothesis on the two parties because it was not specified

$H_0: \mu_1 = \mu_2 \; or \; \mu_1 - \mu_2 = 0$

$H_1: \mu_1 \neq \mu_2$

2- level of significant

$\alpha = 0.05$

3-Test function

$$Z = \frac{\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$Z = \frac{68.5 - 66.9}{\sqrt{\frac{100}{100} + \frac{144}{150}}} \quad \Rightarrow \mathrm{Z} = 1.14$$

4- statistical scale test function

Since it is two parties

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{(0.025)}$

$\therefore Z_{(0.025)} = \pm 1.96 = [-1.96, 1.96]$

5- Decision making

$$Z = 1.14 \in [-1.96\,,1.96]$$

∴ Since the values of Z belong to the statistical scale, then the null hypothesis is accepted $\boldsymbol{H_0}$

We apply the law but formulate hypotheses $\mu_1 - \mu_2 = 1$

either in the case given

$$Z = \frac{68.5 - 66.9 - 1}{\sqrt{\frac{100}{100} + \frac{144}{150}}} \quad \Rightarrow Z = 0.43$$

∴ **Accept the null hypothesis H_0**

Small sample Hypothesis test for ($\boldsymbol{\mu_1 - \mu_2}$)Independent random sample

Example : If the weights of the male individuals follow a normal distribution, i.e. the same variance, where a sample of 10 people was chosen, their average weights were 70.19 kg and a variation of 8.7 kg, another random sample was tested 15, and their average weights were 68,58 kg, a variation of 12.56 kg. Find a heavier weight using individuals significant level α = 0.05.

Solution:

$n_1 = 10$, $n_2 = 15$, $\overline{X_1} = 70.19$ $\overline{X_2} = 68.58$, $S_1^2 = 8.71$, $S_2^2 = 12.56$, $\alpha = 0.05$

Since it is heavier than the one on the right side and the distribution t because σ is unknown and the sample flesh is less than 30

1- Formulate the hypothesis

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

2- The level of significance

$\alpha = 0.05$

3- Test function

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2}{n_1-1} + \frac{(n_2-1)S_2^2}{n_2-1}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$T = \frac{70.19 - 68.58}{\sqrt{\frac{(10-1)(8.71) + (15-1)(12.56)}{10+15-2}(\frac{1}{10} + \frac{1}{15})}}$$

4- The scale of the statistical function

Since it is one party

T=1.164

$t(\alpha, n_1 + n_2 - 2)$

t(0.05,10+15-2)= t(0.05, 23)=[-1.714, +1.714]

5- Decision making

$1.164 \in [-1.714, +1.1.714] \therefore$ Accept the null hypothesis H_0

Large sample test of Hypothsis for A Binominl proportion

Example : It is known that the percentage of seat belt users in cars before enacting the law of obligatory use is 0.80. A random sample of 200 drivers was studied after the enactment of the obligatory law. There were 170 of them using the belt. It was tested at the level of significance of 5% if the legislation increased the proportion of users of the seat belt?

Solution:

$$P_\circ = 0.80 \ , \ n = 200 \ , \ \propto = 0.05 \ , \ x = 170$$

$$\overline{P} = \frac{\text{Section } x}{\text{All } n} = \frac{170}{200} \rightarrow \overline{P} = 0.85$$

Since the amount has increased (unilateral)

1- Formulate the hypothesis

$H_\circ: P = P_\circ \ \& \ H_1 = P > P_\circ$

$$H_\circ: P = 0.80 \ \& \ H_1 = P > 0.80$$

2- The level of significance

$\propto = 0.05$

3- Test function

$$Z = \frac{\overline{P} - P_\circ}{\sqrt{\frac{P_\circ(1 - P_\circ)}{n}}} = \frac{0.85 - 0.80}{\sqrt{\frac{0.80(1 - 0.80)}{200}}}$$

$$Z = \frac{0.05}{\sqrt{\frac{0.16}{200}}} = \frac{0.05}{0.02828427}$$

Z= 1.7677669559

4- The scale of the statistical function

Since the function is one-sided right

$$+Z_\alpha = Z_{0.05} = 1.645$$

Then the confidence interval $[-1.645, 1.645]$

5 -Decision making

$$1.7677669559 \notin [-1.645, 1.645]$$

Reject H_0 and accept the alternative hypothesis H_1.

Large sample test of Hypothsis for A Binominl proportion

Example : If the ratio of the units used for a factory is equal to 30%, a sample of 1000 units was tested, and 350 of them were damaged. Do these results indicate that the percentage of damaged units exceeds 30% of the production of this factory? Use a significant level α= 0.01.

Solution :

$$P_\circ = 0.3 \;,\; n = 1000 \;,\; \propto = 0.01 \;,\; x = 350$$

$$\overline{P} = \frac{\text{Section } x}{All\ n} = \frac{350}{1000} \rightarrow \overline{P} = 0.35$$

Since the amount has increased (one-sided) right

1 -Formulate the hypothesis

$$H_\circ : P = 0.3 \quad \& \quad H_1 = P > 0.3 \qquad H_\circ : P = P_\circ \ \& \ H_1 = P > P_\circ$$

2-The level of significance

$\propto = 0.01$

3- Test function

$$Z = \frac{\bar{P} - P_\circ}{\sqrt{\frac{P_\circ(1-P_\circ)}{n}}} = \frac{0.35 - 0.3}{\sqrt{\frac{0.3(1-0.3)}{1000}}}$$

$$Z = \frac{0.05}{\sqrt{\frac{0.21}{1000}}} = \frac{0.05}{0.0144913767} = 3.4503278077$$

4-The scale of the statistical function

Since the function is one-sided right

$$+Z_\alpha = Z_{0.01} = 2.58$$

5- Decision making

$$3.4503278077 \notin (+2.58)$$

Reject H_0 and accept the alternative hypothesis H_1.

Large sample test of Hypothsis for A Binominl proportion

Example : If the percentage of blood disease in the city of Abba in 2008 is 28.8%, and in 2016 a sample of the population in this city tested a size of 1238 people, including 320 people who had blood pressure, then do these results indicate a decrease in the disease rate between 2008 and 2016? Morale level $\propto = 0.05$.

Solution

$$P_\circ = 0.288 \ , \ n = 1238 \ , \ \propto = 0.05 \ , \ x = 320$$

$$\bar{P} = \frac{\text{Section } x}{\text{All } n} = \frac{320}{1238} \rightarrow \bar{P} = 0.249$$

Since the magnitude has decreased (one-sided) left

1- Formulate the hypothesis

$$H_\circ: P = P_\circ \ \& \ H_1 = P > P_\circ$$

$$H_\circ: P = 0.288 \ \ \& \ \ H_1 = P > 0.288$$

2-The level of significance :

$\propto = 0.05$

3- Test function

$$Z = \frac{\bar{P} - P_\circ}{\sqrt{\frac{P_\circ(1-P_\circ)}{n}}} = \frac{0.249 - 0.288}{\sqrt{\frac{0.288(1-0.288)}{1238}}}$$

$$Z = \frac{-0.039}{\sqrt{\frac{0.205056}{1238}}} = \frac{-0.039}{0.0128699184}$$

Z=-3.0303222435

4- The scale of the statistical function

Since the function is one-sided right

$$-Z_\alpha = -Z_{0.05} = -01.645$$

5- Decision making

$-3.0303222435 \notin (-1..645)$

Reject H_0 and accept the alternative hypothesis H_1.

* $P_\circ$You should be $0 \le P_\circ \le 1$

Large sample statistical test for $(P_1 - P_2)$

Example: comparison of the percentage of smokers in the age group (18 – 25) with the category (26 – 30) years . Taking a random sample of 200 from the first category, 80 of them smoke and took a random sample independent of the first of the second age group of 100, so 52 of them smoke. Test the hypothesis $H_0\ p_1 = p_2$ versus $H_1: p_1 < p_2$ at the function level $\alpha = 0.05$.

Solution:

$$n_1 = 200 \ , n_2 = 100 \, , X_1 = 80 \, , X_2 = 52 \quad \alpha = 0.05$$

$$\overline{p}_1 = \frac{X_1}{n_1} = \frac{80}{200} \Rightarrow \overline{p}_1 = 0.4$$

$$\overline{p}_2 = \frac{X_2}{n_2} = \frac{52}{100} \Rightarrow \overline{p}_2 = 0.52$$

1- Formulate the hypothesis

$H_0: p_1 = p_2$

$$H_1: p_1 < p_2$$

2 -The* level *of significance

$$\alpha = 0.05$$

3-Test function

$$Z = \frac{\overline{p_1} - \overline{p_2}}{\sqrt{\overline{p}_1(1-\overline{p}_1)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$Z = \frac{0.4 - 0.52}{\sqrt{0.4(1-0.4)[\frac{1}{200} + \frac{1}{100}]}}$$

$Z = -2$

4 -The scale of the statistical function

$$-Z_\alpha = -Z_{0.05} = -1.645$$

5-Decision making

$$-2 \notin [-1.645]$$

Reject H_0 and accept the alternative hypothesis H_1.

Test of hypothesis conceming a population variance

Example: I took a random sample of 9 from a normal population, which varied 33 and gave it an arithmetic mean $\overline{X} = 56$ And contrast $\delta^2 = 30$. Test the connotations of significance$\alpha = 0.05$

Hypothesis $H_0: \sigma^2 = 33$ Opposite $H_1: \sigma^2 < 33$?

Solution:

$\mathbf{n = 9}$, $\sigma^2 = 33$, $\boldsymbol{\delta^2 = 30}$, $\boldsymbol{\alpha = 0.05}$

1- Formulate the hypothesis

$H_0: \sigma^2 = 33$,

$H_1: \sigma^2 < 33$

2- The level of significance

$\alpha = 0.05$

3- Test function

$$\boldsymbol{X^2 = \frac{(9-1)30}{33} \Rightarrow X^2 = 7.27}$$

4- The scale of the statistical function

$\boldsymbol{X^2(1-\alpha, n-1) \Rightarrow (1-0.05, 9-1)}$

$\boldsymbol{X^2(0.95, 8) = 2.733}$

We meet 8 and 0.95 in a table chi-sq

5- Decision making

$\boldsymbol{7.27 \notin [\,2.733]}$

Reject $\mathrm{H_0}$ and accept the alternative hypothesis $\mathrm{H_1}$.

Test of hypothesis conceming the equality of two population variances

Example: two independent samples were taken from normal societies and given

$$\overline{X}_1 = 27.4 \ ,\overline{X}_2 = 23.2 \ ,n_1 = 8 \ ,n_2 = 10\ ,S_1^2 = 16 \ ,S_2^2 = 12$$

Test hypothesis level 0.05 .$H_0: \sigma_2^2 = \sigma_1^2$ Opposite $H_1: \sigma_2^2 < \sigma_1^2$

Solution:

1- Formulate the hypothesis

$H_0: \sigma_2^2 = \sigma_1^2$

$H_1: \sigma_2^2 < \sigma_1^2$

2- The level of significance

$$\alpha = 0.05$$

3- Test function

$$\boldsymbol{F = \frac{S_1^2}{S_2^2} \quad \Rightarrow F = \frac{16}{12} \ \Rightarrow\ F = 1.333} \ldots$$

4- The scale of the statistical function

Since it is one-sided ((right))

$$(F_\alpha, n_1 - 1\ , n_2 - 1)$$

$$(F_{0.05}\ , 8 - 1, 10 - 1) = (F_{0.05}, 7,9) = 3.29$$

5- Decision making

$$\mathbf{1.33} \in [\mathbf{3.29}]$$

∴ Accept the null hypothesisH_0.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد

الدكتور فراس شاكر محمود

Critical Region

We begin this chapter on tests of statistical hypotheses with an application in which we define many of the terms associated with testing .

Example :- Let X equal the breaking strength of a steel bar . If the bar is manufactured by process I , X is N(50,36) i.e., X is normally distributed with $\mu = 50$ and $\sigma^2 = 36$. It is hoped that if process II (a new process) is used , X will be N(50,36) . Given a large number of steel bars manufactured by process II , how could we test whether the five-unit increase in the mean breaking strength was realized ?

In this problem , we are assuming that X is $N(\mu, 36)$ and μ is equal to 50 or 55 .We want to test the simple null hypothesis $H_o: \mu = 50$ against the simple alternative hypothesis $H_1: \mu = 55$. Note that each of these hypotheses completely specifies the distribution of X. That is , H_0 states that X is N(50,36) and H_1 states that X is N(55,36) . (If the alternative hypothesis had been $H_1: \mu > 50$, it would be a composite hypothesis , because it is composed of all normal distributions with =36 and means greater than 50) In order to test which of the two hypotheses , H_o or H_1 , is true , we shall set up a rule based on the breaking strengths $x_1, x_2, \dots, x_n$ of n bars (the observed values of a random sample of size n from this new normal distribution) . The rule leads to a decision to accept or reject H_o; hence , it is necessary to partition the sample space into two parts-say , C and C' - so that if $(x_1, x_2, \dots, x_n) \in C$, H_o is rejected , and if $(x_1, x_2, \dots, x_n) \in C'$, H_o is accepted (not reject) . The rejection region C for H_o is called the critical region for the test . Often , the partitioning of the sample space is specified in terms of the values of a statistic called the test statistic . In this example , we could let $\bar{x}$ be the test statistic and say ,take $C = \{(x_1, x_2, \dots, x_n): \bar{x} \geq 53\}$; that is we will reject H_o if $\bar{x} \geq 53$. If $(x_1, x_2, \dots, x_n) \in C$ when H_o is true , H_0 would be rejected when it is true , a Type I error . If $(x_1, x_2, \dots, x_n) \in C'$ when H_1 is true , H_o would be accepted (i.e. not rejected) when in fact H_1 is true , a Type II error . The probability of a Type I error is called the **significance level** of the test and is denoted by α .

That is , $\alpha = P[(x_1, x_2, \dots, x_n) \in C;\ H_o]$ is the probability that $(x_1, x_2, \dots, x_n)$ falls into C when H_o is true . The probability of a Type II error is denoted by β ; that is , $\beta = P[(x_1, x_2, \dots, x_n) \in C';\ H_1]$ is the probability of accepting (failing to reject) H_o when it is false.

As an illustration , suppose n=16 bars were tested and $C = \{\bar{x}: \bar{x} \geq 53\}$. Then $\bar{X}$ is N(50,36/16) when H_o is true and is N(55,36/16) when H_1 is true . Thus ,

$$\alpha = P(\bar{X} \geq 53; H_o) = P\left(\frac{\bar{X} - 50}{\frac{6}{4}} \geq \frac{53 - 50}{\frac{6}{4}} \; ; H_o\right)$$

$$= 1 - \emptyset(2) = 0.0228$$

And

$$\beta = P(\bar{X} \geq 53; H_1) = P\left(\frac{\bar{X} - 55}{\frac{6}{4}} \geq \frac{53 - 55}{\frac{6}{4}} \; ; H_1\right)$$

$= \emptyset\left(-\frac{4}{3}\right) = 1 - 0.9087 = 0.0913$.

Note that by changing the critical region , C , it is possible to decrease (increase) the size of α but this leads to an increase (decrease) in the size of β . Both α and β can be decreased if the sample size n is increased.

Example :- Assume that the underlying distribution is normal with unknown mean μ but known variance $\sigma^2 = 100$. Say we are testing the simple null hypotheses $H_o: \mu = 60$ against the composite alternative hypotheses $H_1;\ \mu > 60$ with a sample mean $\bar{X}$ based on n=52 observations . Suppose that we obtain the observed sample mean of $\bar{X} = 62.75$. If we compute the probability of obtaining an $\bar{X}$ of that value of 62.75 or greater when $\mu = 60$. then we obtain the P-value associated with $\bar{X} = 62.75$.

$$P - value = P(\bar{X} \geq 62.75 \, ; \mu = 60)$$

$$= P\left(\frac{\bar{X} - 60}{\frac{10}{\sqrt{52}}} \geq \frac{62.75 - 60}{\frac{10}{\sqrt{52}}} \; ; \; \mu = 60\right)$$

$$= 1 - \emptyset\left(\frac{62.75 - 60}{\frac{10}{\sqrt{52}}}\right) = 1 - \emptyset(1.983) = 0.0237.$$

If this p-value is small , we tend to reject the hypotheses $H_o\!: \mu = 60$. For example rejecting $H_o\!: \mu = 60$ if the p-value is less than or equal to $\alpha = 0.05$ is exactly the same as rejecting H_o if

$\bar{X} \geq 60 + (1.645)\left(\frac{10}{\sqrt{52}}\right) = 62.281$.

Here $\text{P} - \text{value} = 0.0237 < \alpha = 0.05$ and $\bar{\text{X}} = 62.75 > 62.281$

To help the reader keep the definition of p-value in mind , we note that it can be thought of as that tail-end probability , under H_o , of the distribution of the statistic (here $\bar{\text{X}}$) beyond the observed value of the statistic .

If the alternative were the two-sided $H_1 : \mu \neq 60$, then the p-value would have been double 0.0237 ; that is , then the p-value =2(0.0237)=0.0474 because we include both tails .

To test $H_0 : \mu = \mu_0$ against one of these three alternative hypotheses , a random sample is take from the distribution and an observed sample mean , $\bar{X}$, that is close to μ_0 supports H_o. The closeness of $\bar{X}\ to\ \mu_o$ is measured in terms of standard deviations of $\bar{X}\,, \alpha/\sqrt{n}$, when α is known , a measure that is sometimes called the standard error of the mean . Thus the test statistic could be defined by

H_o	H_1	Critical Region
$\mu = \mu_0$	$\mu > \mu_0$	$z \geq z_\alpha\ or\ \bar{X} \leq \mu_0 + z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$z \leq -z_\alpha\ or\ \bar{X} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z \rvert \geq z_{\alpha/2}\ or\ \lvert \bar{X} - \mu_0 \rvert \geq z_{\alpha/2} \sigma/\sqrt{n}$

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

And the critical regions , at a significance level α , for the three respective alternative hypotheses would be $(i) Z \geq Z_\alpha$, $(ii)\ Z \leq -Z_\alpha$, and $(iii) \lvert Z \rvert \geq Z_{\alpha/2}$. In terms of $\bar{X}$, these three critical regions become $(i)\bar{X} \geq \mu_0 + Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$, $(ii)\bar{X} \leq \mu_0 - Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$, and $(iii)\lvert \bar{X} - \mu_0 \rvert \geq Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$. The three tests and the distribution is $N(\mu\,, \sigma^2)$ and σ^2 is known. It is usually the case that the variance σ^2 is not known . Accordingly , we now take a more realistic position and assume that the variance is unknown . Suppose our null hypotheses is $H_o : \mu = \mu_0$ and the two-sided alternative hypotheses is $H_1 : \mu \neq \mu_0$. For a random sample $x_1, x_2, \dots, x_n$ taken a normal distribution $N(\mu, \sigma^2)$, a confidence interval for μ is based on

$$T = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

This suggests that T might be a good statistic to use for the test of $H_o : \mu = \mu_0$ with μ replaced by μ_0 . In addition , it is the natural statistic to use if we replace σ^2/n by its unbiased estimator s^2/n in $(\bar{X} - \mu_0)/\sqrt{\sigma^2/n}$. If $\mu = \mu_0$, we know that T has a t distribution with n-1 degrees of freedom . Thus , with $\mu = \mu_0$,

$$P\left[|T| \geq t_{\frac{\alpha}{2},(n-1)}\right] = P\left[\frac{|\bar{X} - \mu_0|}{\frac{s}{\sqrt{n}}} \geq t_{\frac{\alpha}{2},(n-1)}\right] = \alpha$$

Accordingly , if $\bar{X}$ and s are , respectively , the sample mean and sample standard deviation , then the rule that reject $H_o: \mu = \mu_o$ and accepts $H_1: \mu \neq \mu_o$ if and only if

$$|t| = \frac{|\bar{X} - \mu_0|}{\frac{s}{\sqrt{n}}} \geq t_{\frac{\alpha}{2}}(n-1)$$

Provides a test of this hypotheses with significance level α . Note that this rule is equivalent to rejecting $H_0: \mu = \mu_0$ if not the open $100(1-\alpha)\%$ confidence interval

$$\left(\bar{X} - t_{\frac{\alpha}{2},(n-1)}\left[\frac{s}{\sqrt{n}}\right], \bar{X} + t_{\frac{\alpha}{2},(n-1)}\left[\frac{s}{\sqrt{n}}\right]\right)$$

The following Table summarizes tests of hypotheses for a single mean along with the three possible alternative hypotheses , when the underlying distribution is $N(\mu, \sigma^2)$, σ^2 is unknown , $t = \frac{(\bar{X}-\mu_0)}{\frac{s}{\sqrt{n}}}$. and $n \leq 30$. If $n > 30$, we use the following Table for approximate tests , with σ replaced s .

H_o	H_1	$Critical\ Region$
$\mu = \mu_0$	$\mu > \mu_0$	$z \geq z_\alpha$ or $\bar{X} \leq \mu_0 + z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$z \leq -z_\alpha$ or $\bar{X} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\|z\| \geq z_{\alpha/2}$ or $\|\bar{X} - \mu_0\| \geq z_{\alpha/2}\sigma/\sqrt{n}$

Example : Let X (in millimeters) equal the growth in 15 days of a tumor induced in a mouse . Assume that the distribution of X is $N(\mu\,,\sigma^2)$. We shall test the null hypotheses $H_0: \mu = \mu_0 = 4.0\, mm$ aganst the two-sided alternative hypothesis $H_1: \mu \neq 4.0$ If we use n=9 observations and a significance level of $\alpha = 0.10$, the critical region is

$$|t| = \frac{|\bar{X} - 4.0|}{\frac{S}{\sqrt{9}}} \geq t_{\frac{\alpha}{2}}(8) = 1.860.$$

If we are given that $\mathrm{n} = 9\,, \bar{\mathrm{X}} = 4.3\,,$ and $\mathrm{S} = 1.2$ we see that

$$t = \frac{4.3 - 4.0}{\frac{1.2}{\sqrt{9}}} = \frac{0.3}{0.4} = 0.75.$$

Thus ,

$$|\mathrm{t}| = |0.75| < 1.860$$

And we accept (do not reject) $H_o: \mu = 4.0$ at the $\alpha = 10\%$ significance level. The p-value is the two-sided probability of $|T| \geq 0.75$, namely .

$$P - \text{value} = \mathrm{P}(|\mathrm{T}| \geq 0.75) = 2\mathrm{P}(\mathrm{T} \geq 0.75)$$

With our t tables with eight degrees of freedom , we cannot find this P-value exactly. It is about 0.50 , because

$$P(|T| \geq 0.706) = 2P(T \geq 0.706) = 0.50$$

Remark : In discussing the test of a statistical hypothesis , the word accept H_0 might better be replaced by do not reject H_o . That is , if , $\bar{X}$ is close enough to 4.0 so that we accept $\mu = 4.0$, we do not want that acceptance to imply that μ is actually equal to 4.0 , We want to say that the data do not deviate enough from $\mu = 4.0$ for us to reject that hypothesis : that is , we do not reject $\mu = 4.0$ with these observed data . With this understanding , we sometimes use accept . and sometimes fail to reject or do not reject , the null hypothesis .

Example : In attempting to control the strength of the wastes discharged into a nearby river , a paper firm has taken a number of measures. Members of the firm believe that they have reduced the oxgen-consuming power of their wastes from a previous mean μ of 500 (measured in parts per million of permanganate) . They plan to test $H_o: \mu = 500$ against $H_1: \mu < 500$, using readings taken on n=25 consecutive days. If these 25 values can be treated as a random sample , then the critical region , for a significance level of $\alpha = 0.01$, is

$$t = \frac{\bar{X} - 500}{\frac{s}{\sqrt{25}}} \leq -t_{0.01}(24) = -2.492$$

The observed values of the sample mean and sample standard deviation .

$\bar{X} = 308.8$ and $s = 115.15$. since

$$t = \frac{308.8 - 500}{\frac{115.15}{\sqrt{25}}} = -8.30 < -2.492$$

We clearly reject the null hypothesis and accept $H_1: \mu < 500$. Note , however , that although an improvement has been made , there stil might exist the question of whether the improvement is adequate . The one-sided 99% confidence interval for μ

$$\left[0,308.8 + 2.492\left(\frac{115.25}{\sqrt{25}}\right)\right] = [0,366.191.]$$

Provides an upper bound for μ and may help the company answer this question.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد

الدكتور فراس شاكر محمود

The Wilcoxon Tests

As mentioned earlier in the text, at times it is clear that the normality assumptions are not met and that other procedures, sometimes referred to as **nonparametric** or **distribution-free** methods, should be considered.

Example: Suppose some hypothesis, say, $H_0: m = m_0$, against $H_1: m \neq m_0$, is made about the unknown median, m, of a continuous-type distribution.

From the data, we could construct a $100(1-\alpha)\%$ confidence interval for m, and if m_0 is not in that interval, we would reject H_0 at the α significance level.

Now let X be a continuous-type random variable and let m denote the median of X. To test the hypothesis $H_0: m = m_0$ against an appropriate alternative hypothesis, we could also use a **sign test**. That is, if $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from this distribution, and if we let Y equal the number of negative differences among $X_1 - m_0, X_2 - m_0, \ldots, X_n - m_0$, then Y has the binomial distribution $b(n, 1/2)$ under H_0 and is the test statistic for the sign test. If Y is too large or too small, we reject $H_0: m = m_0$.

Example: Let X denote the length of time in seconds between two calls entering a call center.

Let m be the unique median of this continuous-type distribution. We test the null hypothesis $H_0: m = 6.2$ against the alternative hypothesis $H_1: m < 6.2$. If Y is the number of lengths of time between calls in a random sample of size 20 that are less than 6.2, then the critical region $C = \{y : y \geq 14\}$ has a significance level of $\alpha = 0.0577$.

A random sample of size 20 yielded the following data:

6.8	5.7	6.9	5.3	4.1	9.8	1.7	7.0
2.1	19.0	18.9	16.9	10.4	44.1	2.9	2.4
4.8	18.9	4.8	7.9				

Since $y = 9$, the null hypothesis is not rejected .

The sign test can also be used to test the hypothesis that two possibly dependent continuous-type random variables X and Y are such that $p = P(X > Y) = 1/2$.

To test the hypothesis $H_0 : p = 1/2$ against an appropriate alternative hypothesis, consider the independent pairs $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Let W denote the number of pairs for which $X_k - Y_k > 0$.When H_0 is true, W is $b(n, 1/2)$, and the test can be based upon the statistic W.

For example, say X is the length of the right foot of a person and Y the length of the corresponding left foot. Thus, there is a natural pairing, and here $H_0 : p = P(X > Y) = 1/2$ suggests that either foot of a particular individual is equally likely to be longer.

One major objection to the sign test is that it does not take into account the magnitude of the differences $X_1 - m_0, \dots, X_n - m_0$.

We now discuss a **test of Wilcoxon** that does take into account the magnitude of the differences $|X_k - m_0|, k = 1, 2, \dots, n$. However, in addition to assuming that the random variable X is of the continuous type, we must also assume that the pdf of X is symmetric about the median in order to find the distribution of this new statistic.

Because of the continuity assumption, we assume, in the discussion which follows, that no two observations are equal and that no observation is equal to the median.

We are interested in testing the hypothesis $H_0: m = m_0$, where $m0$ is some given constant. With our random sample $X_1, X_2, \ldots, X_n$, we rank the absolute values $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$ in ascending order according to magnitude.That is, for $k = 1, 2, \ldots, n$, we let R_k denote the rank of

$|X_k - m_0|$ among $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$.

Note that $R_1, R_2, \ldots, R_n$ is a permutation of the first n positive integers, $1, 2, \ldots, n$. Now, with each R_k, we associate the sign of the difference $X_k - m_0$; that is, if $X_k - m_0 > 0$, we use R_k, but if $X_k - m_0 < 0$, we use $-R_k$. The Wilcoxon statistic W is the sum of these n signed ranks, and therefore is often called the **Wilcoxon signed rank statistic**.

<u>Example</u>: Suppose the lengths of n = 10 sunfish are

$$xi: \ 5.0 \ \ 3.9 \ \ 5.2 \ \ 5.5 \ \ 2.8 \ \ 6.1 \ \ 6.4 \ \ 2.6 \ \ 1.7 \ \ 4.3$$

We shall test $H_0: m = 3.7$ against the alternative hypothesis $H_1: m > 3.7$. Thus, we have

$x_k - m_0$:	1.3,	0.2,	1.5,	1.8,	−0.9,	2.4,	2.7,	−1.1,	−2.0,	0.6
$\|xk - m0\|$:	1.3,	0.2,	1.5,	1.8,	0.9,	2.4,	2.7,	1.1,	2.0,	0.6
Ranks:	5,	1,	6,	7,	3,	9,	10,	4,	8,	2
Signed Ranks:	5,	1,	6,	7,	− 3,	9,	10,	− 4,	− 8,	2

Therefore, the Wilcoxon statistic is equal to

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25 .$$

Incidentally, the positive answer seems reasonable because the number of the 10 lengths that are less than 3.7 is 3, which is the statistic used in the sign test .

If the hypothesis $H_0: m = m_0$ is true, about one half of the differences would be negative and thus about one half of the signs would be negative.

Hence, it seems that the hypothesis $H_0: m = m_0$ is supported if the observed value of W is close to zero. If the alternative hypothesis is $H_1: m > m_0$, we would reject H_0 if the observed $W = w$ is too large, since, in this case, the larger deviations $|X_k - m_0|$ would usually be associated with observations for which $x_k - m_0 > 0$.

That is, the critical region would be of the form $\{w: w \geq c_1\}$.

If the alternative hypothesis is $H_1: m < m_0$, the critical region would be of the form $\{w: w \leq c_2\}$. Of course, the critical region would be of the form $\{w: w \leq c_3 \ \ or \ w \geq c_4\}$ for a two-sided alternative hypothesis $H_1: m \neq m_0$.

In order to find the values of c_1, c_2, c_3, and c_4 that yield desired significance levels, it is necessary to determine the distribution of W under H_0.

Accordingly, we consider certain characteristics of this distribution.

When $H_0: m = m_0$ is true,

$$P(X_k < m_0) = P(X_k > m_0) = \frac{1}{2}, \qquad k = 1,2,\ldots,n.$$

Hence, the probability is 1/2 that a negative sign is associated with the rank R_k of $|X_k - m_0|$.

Moreover, the assignments of these n signs are independent because $X_1, X_2, \ldots, X_n$ are mutually independent. In addition, W is a sum that contains the integers $1, 2, \ldots, n$, each with a positive or negative sign. Since the underlying distribution is symmetric, it seems intuitively obvious that W has the same distribution as the random variable

$$V = \sum_{k=1}^{n} V_k,$$

where $V_1, V_2, \ldots, V_n$ are independent and

$$P(V_k = k) = P(V_k = -k) = \frac{1}{2}, \quad k = 1,2, \ldots, n \,.$$

That is, V is a sum that contains the integers 1, 2, . . . , n, and these integers receive their algebraic signs by independent assignments. Since W and V have the same distribution, their means and variances are equal, and we can easily find those of V.

Now, the mean of V_k is

$$E(V_k) = -k\left(\frac{1}{2}\right) + k\left(\frac{1}{2}\right) = 0 \,;$$

Thus ,

$$E(W) = E(V) = \sum_{k=1}^{n} E(V_k) = 0.$$

The variance of V_k is

$$Var(V_k) = E(V_k^2) = (-k)^2 \left(\frac{1}{2}\right) + (k)^2 \left(\frac{1}{2}\right) = k^2 .$$

Hence,

$$Var(W) = Var(V) = \sum_{k=1}^{n} Var(V_k) = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} .$$

We shall not try to find the distribution of *W* in general, since that pmf does not have a convenient expression.

However, we demonstrate how we could find the distribution of *W* (or *V*) with enough patience and computer support.

Recall that the moment-generating function of V_i is

$$M_k(t) = e^{t(-k)} \left(\frac{1}{2}\right) + e^{t(+k)} \left(\frac{1}{2}\right) = \frac{e^{-kt} + e^{kt}}{2}, \quad k = 1,2, \dots, n$$

Let $n = 2$; then the moment-generating function of $V_1 + V_2$ is

$$M(t) = E\left[e^{t(V_1+V_2)}\right].$$

From the independence of V_1 and V_2, we obtain

$$M(t) = E(e^{tV_1})E(e^{tV_2})$$

$$= \left(\frac{e^{-t} + e^{t}}{2}\right)\left(\frac{e^{-2t} + e^{2t}}{2}\right)$$

$$= \frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4} .$$

This means that each of the points $-3, -1, 1, 3$ in the support of $V_1 + V_2$ has probability 1/4.

Next let $n = 3$;

then the moment-generating function of $V_1 + V_2 + V_3$ is

$$M(t) = E\left[e^{t(V_1+V_2+V_3)}\right]$$

$$= E\left[e^{t(V_1+V_2)}\right]E\left(e^{tV_3}\right)$$

$$= \left(\frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4}\right)\left(\frac{e^{-3t} + e^{3t}}{2}\right)$$

$$= \frac{e^{-6t} + e^{-4t} + e^{-2t} + 2e^{0} + e^{2t} + e^{4t} + e^{6t}}{8}.$$

Thus, the points $-6, -4, -2, 0, 2, 4,$ and 6 in the support of $V_1 + V_2 + V_3$ have the respective probabilities $1/8, 1/8, 1/8, 2/8, 1/8, 1/8,$ and $1/8$.

Obviously, this procedure can be continued for $n = 4, 5, 6, \ldots$, but it is rather tedious.

Fortunately, however, even though $V_1, V_2, \ldots, V_n$ are not identically distributed random variables, the sum V of them still has an approximate normal distribution for large samples.

To obtain this normal approximation for V (or W), a more general form of the central limit theorem, due to Liapounov, can be used which allows us to say that the standardized random variable

$$z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}}$$

is approximately $N(0, 1)$ when H_0 is true.

We accept this theorem without proof, so that we can use this normal distribution to approximate probabilities such as

$P(W \geq c; H_0) \approx P(Z \geq z_\alpha; H_0)$ when the sample size n is sufficiently large.

Example: The moment-generating function of W or of V is given by

$$M(t) = \prod_{i=1}^{n} \frac{e^{-kt} + e^{kt}}{2} \, .$$

Using a computer algebra system such as *Maple*, we can expand $M(t)$ and find the coefficients of e_{kt}, which is equal to $P(W = k)$.

In Figure, we have drawn a probability histogram for the distribution of W along with the approximating $N[0, n(n + 1)(2n + 1)/6]$ pdf for $n = 4$ (a poor approximation) and for $n = 10$. It is important to note that the widths of the rectangles in the probability histogram are equal to 2, so the "half-unit correction for continuity" mentioned in Section 5.7 now is equal to 1.

Example: Let m be the median of a symmetric distribution of the continuous type.

To test the hypothesis $H_0: m = 160$ against the alternative hypothesis $H_1: m > 160$, we take a random sample of size $n = 16$. For an approximate significance level of $\alpha = 0.05, H_0$ is rejected if the computed $W = w$ is such that

$$z = \frac{w}{\sqrt{\frac{16(17)(33)}{6}}} \geq 1.645 \, ,$$

or

$$w \geq 1.645\sqrt{\frac{16(17)(33)}{6}} = 63.626\,.$$

Say the observed values of a random sample are 176.9, 158.3, 152.1, 158.8, 172.4, 169.8, 159.7, 162.7, 156.6, 174.5, 184.4, 165.2, 147.8, 177.8, 160.1, and 160.5.

In Table 1, the magnitudes of the differences $|x_k - 160|$ have been ordered and ranked.Those differences $x_k - 160$ which were negative have been underlined, and the ranks are under the ordered values.

For this set of data,

$$w = 1 - 2 + 3 - 4 - 5 + 6 + \cdots + 16 = 60.$$

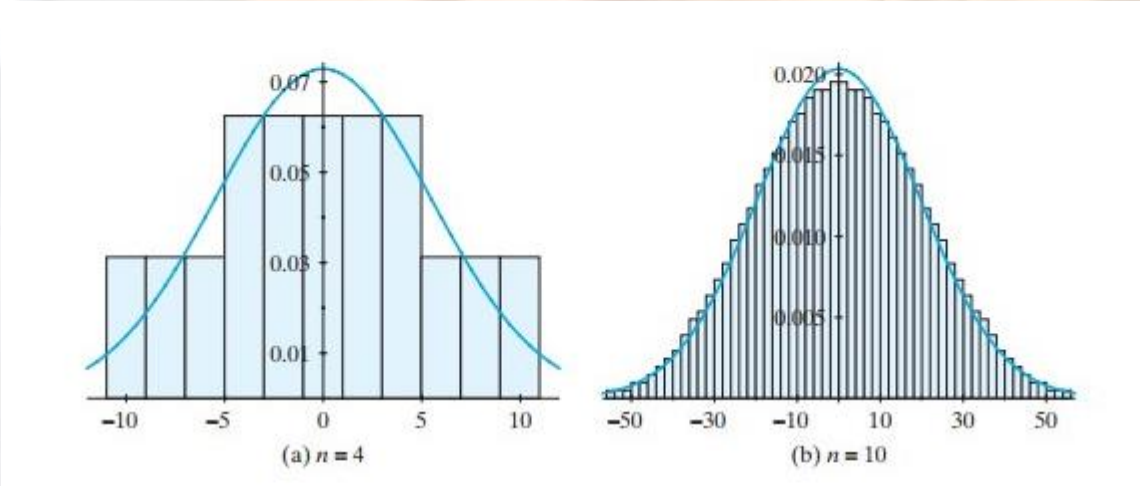


Figure 1

<u>Example</u>: The weights of the contents of $n_1 = 8$ and $n_2 = 8$ tins of cinnamon packaged by companies A and B, respectively, selected at random, yielded the following observations of *X* and *Y*:

x:	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
y:	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

The critical region for testing $H_0\colon m_X = m_Y$ against $H_1\colon m_X < m_Y$ is of the form $w \geq c$.

Since $n_1 = n_2 = 8$, at an approximate $\alpha = 0.05$ significance level H_0 is rejected if

$$z = \frac{w - 8(8+8+1)/2}{\sqrt{[(8)(8)(8+8+1)]/12}} \geq 1.645\,,$$

or

$$w \geq 1.645\sqrt{\frac{(8)(8)(17)}{12}} + 4(17) = 83.66\,.$$

To calculate the value of *W*, it is sometimes helpful to construct a **back-to-back stem-and-leaf display**. In such a display, the stems are put in the center and the leaves go to the left and the right.

(See Table 1.) Reading from this two-sided stem-and-leaf display, we show the combined sample in Table 2, with the Company B (*y*) weights underlined. The ranks are given beneath the values. From Table 2, the computed W is

$$w = 3 + 8 + 9 + 11 + 12 + 13 + 15 + 16 = 87 > 83.66.$$

Table 1 : Back-to-back stem-and-leaf diagram of weights of cinnamon

x Leaves		Stems	y Leaves	
	51	11*f*		
74	71	11*s*	72	
	95	11•		
19	13	12∗		
		12*t*	21	35
	45	12*f*	53	56
	78	12*s*	65	79
		12•	98	

Table 2: Combined ordered samples

115.1	117.1	117.2	117.4	119.5	121.3	121.9	122.1
1	2	3	4	5	6	7	8
123.5	124.5	125.3	125.6	126.5	127.8	127.9	129.8
9	10	11	12	13	14	15	16

Thus, H_0 is rejected.

Finally, making a half-unit correction for continuity, we see that the *p*-value of this test is

$$p - value = P(W \geq 87)$$
$$= P\left(\frac{w-68}{\sqrt{90.0667}} \geq \frac{86.5-68}{\sqrt{90.667}}\right)$$
$$\approx P(Z \geq 1.943) = 0.0260.$$

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Power of ASTATISTICAL TEST

We gave several tests of a fairly common statistical hypotheses in such a way that described the significance level α and the p-value of each . of course those tests were based on good sufficient) statics of the parameters, when the latter exist. In this lecture, we consider the probability of Making the other type of error accepting I the null hypothesis H_0 when the alternative a hypothesis $H_1$is true. This consideration leads to ways to find Most Powerful tests of the null hypothesis H_0against the alternation hypothesis H_1 . The first example introduces a new concept. using a test about P , the probability of success The Sample size is kept Small so that Table II in Appendix B Can be used to find probabilities. The application is one that you can actually Perform

Example:

Assume that when given a name lag, a person puts it on either the right or left Side Let f equal the probability that the name tag is placed on the right side. we shall test the null hypothesis, H_0:p=$1/2$ against the composite alternative the hypothesis H_1:p< $1/2$(Included with the null.

hypothesis are those values of P which are a greater than $1/2$,,that is, we could think of H_0 as H_0 :P≥112 we shall give name tags to a random sample n=20 people. denoting the placements of their name • tags with Bernoulli random variables, $X_1, X_2, \dots \dots \dots X_{20}$ where $X_{i=1}$ if a person places the name tag on the right and $X_i = 0$" if person places the name tag on the left, for our test statistic, we can then use Y= $\sum_{i=1}^{20} X_i$ which has the binomial distribution b (20,P). Say the critical region is defined by C={Y: Y≤6} or equivalently, by $\{(X_1, X_2, \dots, X_{20})$:$\sum_{i=1}^{20} X_i \le 6\}$ since y is b (20,$1/2$) if P=$1/2$ the significance level of the corresponding test is

б=P(Y≤6: P=$1/2$) = $\sum_{y=0}^{6} \binom{20}{y} \left(\frac{1}{2}\right)^{20}$ =0.0577 from Table II in Appendix

B. of course, the probability ß of a Type II error has different value, with different values of P select from the Composite alternative hypothesis $H_1: p < 1/2$ for example::, with P=$1/4$, β=P(7≤Y≤20; P=$1/4$) =

$\sum_{y=7}^{20}\binom{20}{y}\left(\frac{1}{14}\right)^{y}\left(\frac{3}{4}\right)^{20-y}$ = 0.2142 whereas with P=$^{1}/_{10}$, β=P(7≤Y≤20; P=$^{1}/_{10}$) =$\sum_{y=7}^{20}\binom{20}{y}\left(\frac{1}{10}\right)^{y}\left(\frac{a}{10}\right)^{20-y}$

=0.0024

Instead of considering the probability β of accepting H_0whenH_1, is true, we could compute the probability K of rejecting H_0 when $H_1$is true After all, β and K=1-β Provide the same information since k is a function of P, we denote this explicitly. by writing K(p). The probability

K(P) =$\sum_{Y=0}^{6}\binom{20}{Y} P^{Y}(1-P)^{20-Y}$,0< P≤$\frac{1}{2}$

is called the power function of the test of course, 6=K($^{1}/_{2}$) = 0.0577, 1-K($^{1}/_{4}$)=0.2142, and 1-k($^{1}/_{10}$) =0.0024. The value of the power function at a specified p. is called the power of the test at that point. For instance, K($^{1}/_{4}$)=0.7858 and K($^{1}/_{10}$) = 0.9976 are the powers at P=$^{1}/_{4}$ and P=$^{1}/_{10}$, respectively. An accepta able power function assumes Small Values when H_0 is true and larger values when P differs much from P=$^{1}/_{2}$.

Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة :الاستاذ المساعد

الدكتور فراس شاكر محمود

Best Critical Region

We consider the properties a satisfactory hypothesis test (or critical region) should possess. To introduce our investigation, we begin with a non-statistical example.

Example: Say that you have α dollars with which to buy books. Further, suppose that you are not interested in the books themselves, but only in filling as much of your book- shelves as possible. How do you decide which books to buy? Does the following approach seem reasonable? First of all, take all the available free books. Then start choosing those books for which the cost of filling an inch of bookshelf is smallest. That is, choose those books for which the ratio c/w is a minimum, where w is the width of the book in inches and c is the cost of the book. Continue choosing books this way until you have spent the α dollars.

Definition: uniformly most powerful critical region of size α .

A test defined by a critical region C of size α is a uniformly most powerful test if it is a most powerful test against each simple alternative in H_1 . The critical region C is called a uniformly most powerful critical region of size α.

LIKELHOOD RATIO TESTS

We consider a general test – construction method that is applicable when either of both of the null and alternative hypotheses – say . H_0 and H_1-are composite . We continue to assume that the functional form of the p. d. f. is Known . but that it depends on one or more unknown parameters . that is we assume that the p. d. f. of X is $f(X:\theta)$, where θ represent one or more inknown parameters . we let Ω denote the total parameters space – that is , the set of all possible values of the parameter θ given by either H_0or H_1 . these hypotheses will be stated as .

$H_0: \theta \in \omega, \qquad H_1: \theta \in \omega'$.

Where ω is a subset of Ω and ω′ is the complement of ω with respect to Ω . the test will be constructed with the use of a ratio of likelihood functions that have been maximized in ω and Ω , respectively . n a sense , this is natural generalization of the ratio appearing in the Neyman – Pearson lemma when the two hypotheses were simple .

Definition:

The likelihood ratio is the quotient

$$\lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})}$$

Where$L(\widehat{\omega})$ is the maximum of the likelihood function with respect to θ when $\theta \in \omega$ and $L(\widehat{\Omega})$ id the maximum of the likelihood function with respect to θ when $\theta \in \Omega$

Definition:

Consider the test of the simple null hypothesis $Ho: \theta = \theta o$ against the simple alternative hypothesis $H_1: \theta = \theta_1$ Let C be a critical region of size α: that is , $\alpha = P(C: \theta o)$. Then C is a best critical region of size α if. For every other critical region D of size $\alpha = P(D: \theta o)$. We have $P(C: \theta_1) \geq P(D: \theta_1)$

That is. When $H_{1:}\theta = \theta_1$is true . the probability of rejecting $H_0: \theta = \theta_0$ with the use of the critical region C is at least as great as the corresponding probability with the use of any other critical region D of size α .

Thus a best critical region of size α is the critical region that has great- est power among all critical region of size α . the Neyman- Person lemma gives sufficient conditions for a best critical region of size α

Theorem:**(Neyman- Person lemma)**

Let $x_1 . x_2 \dots \dots . x_n$ be a random sample of size n from a distribution with pdf or pmf $f\ (x: \theta)$. where θ_0 and θ_1 are two possible values of θ .

$$L\ (\theta) = L\ (\theta: x_1 . x_2 \dots . . x_n) = f(x_1: \theta)\ f\ (x_2: \theta) \dots f(x_{n:}\theta).$$

If there exist a positive constant k and a subset C of the sample space such that

(a) $P[(x_1 . x_2 \dots . . x_n) \in C: \theta_0] = \int_C L(\theta_0) dx = \alpha.$

(b) $\frac{L(\theta_0)}{L(\theta_1)} \leq k\ for\ (x_1 . x_2 \dots . . x_n) \in C. and$

(c) $\frac{L(\theta_0)}{L(\theta_1)} \geq k\ for\ (x_1 . x_2 \dots . . x_n) \in \bar{C}.$

Then C is a best critical region of size α for testing the simple null hypothesis $Ho: \theta = \theta o$ against the simple alternative hypothesis $H_1: \theta = \theta_1$

Proof: We prove the theorem when the random variables are the continuous type: for discrete – type random variables replace the integral signs by summation signs To simplify the exposition . we shall use the following notation :

$$\int_B L\ (\theta) = \int_B \dots \int\ L\ (\theta: x_1\ x_2 \dots \dots x_n)\ dx_1\ dx_2 \dots . dx_n.$$

Assume that there exists another critical region of size α- say . D, such that. in this new notation

$$a = \int_c L(\theta_0) = \int_D L(\theta_0).$$

Then we have

$$0 = \int_c L(\theta_0) - \int_D L(\theta_0)$$

$$= \int_{C\cap D} L(\theta_0) + \int_{C\cap D} L(\theta_0) - \int_{C\cap D} L(\theta_0) - \int_{C\cap D} L(\theta_0).$$

Hence

$$0 = \int_{C\cap D} L(\theta_0) - \int_{C\cap D} L(\theta_0)$$

By hypothesis (b). $kL(\theta_1) \geq L(\theta_0)$ at each point in C and therefore in $C\cap D$, thus .

$$k \int_{C\cap D} L(\theta_1) \geq \int_{C\cap D} L(\theta_0)$$

By hypothesis (b). $kL(\theta_1) \leq L(\theta_0)$ at each point in C and therefore in $C\cap D$: thus we obtain .

$$k \int_{C\cap D} L(\theta_1) \leq \int_{C\cap D} L(\theta_0)$$

Consequently .

$$0 = \int_{C\cap D} L(\theta_0) - \int_{C\cap D} L(\theta_0) \leq (k)\left\{\int_{C\cap D} L(\theta_1) - \int_{C\cap D} L(\theta_1)\right\}$$

That is .

$$0 \leq (k)\left\{\int_{C\cap D} L(\theta_1) + \int_{C\cap D} L(\theta_1) - \int_{C\cap D} L(\theta_1) - \int_{C\cap D} L(\theta_1)\right\}$$

Or equivalently .

$$0 \leq (k)\left\{\int_c L(\theta_1) - \int_D L(\theta_1)\right\}$$

Thus.

$$\int_c L(\theta_1) \geq \int_D L(\theta_1):$$

That is . $P(C:\theta_1) \geq P(D:\theta_1)$ since that is true for rvery critical region D of size α . C is a best critical region of size α.

Example:

Let $X_1, X_2 \ldots\ldots X_n$ denote a random sample of size n from a Poisson distribution with mean λ. A best critical region for $H_0: \lambda = 2$ against $H_1: \lambda = 5$ given by

$$\frac{L(2)}{L(5)} = \frac{2^{\sum_{i=1}^{n} x_i} e^{-2n}}{x_1!\, x_2! \ldots x_n!} \frac{x_1!\, x_2! \ldots x_n!}{5^{\sum_{i=1}^{n} x_i} e^{-5n}} \leq k.$$

This inequality can be written as

$\left(\frac{2}{5}\right)^{\sum_{i=1}^{n} x_i} e^{3n} \leq k$, or $(\sum_{i=1}^{n} x_i)\, In \left(\frac{2}{5}\right) + 3n \leq In\; k.$

Since In (2/5) < 0. The latter inequality is the same as

$$\sum_{i=1}^{n} x_i \geq \frac{Ink - 3n}{In\left(\frac{2}{5}\right)} = c.$$

If n = 4 and c = 13,then

$$\alpha = P\left(\sum_{i=1}^{4} x_i \geq 13; \lambda = 2\right) = 1 - 0.936 = 0.064.$$

Since $\sum_{i=1}^{4} x_i$ has a Poisson distribution with mean 8 when λ = 2.

Because λ is the quotient of nonnegative functions, $\lambda \geq 0$. In addition , since $\omega \subset \Omega$, it follows that $L(\widehat{\omega}) \leq L(\widehat{\Omega})$ and hence $\lambda \leq 1$. thus 0 $\leq \lambda \leq 1$. If the maximum of L in ω is much smaller than in Ω. It would seem that the data $x_1, x_2., x_n$ do not support the hypothesis $H_0: \theta \in \omega$. that is . a small value of the ratio $\lambda = L(\widehat{\omega})/ L(\widehat{\Omega})$ would lead to the rejection of H_0 . In contrast, a value of the ratio λ that is close to 1 would support the null hypothesis H_0 this reasoning leads us to the next definition.

Definition: To test $H_0: \theta \in \omega$ against $H_1: \theta \in \omega$ the critical region for the likelihood ratio test is the set of points in the sample space for which.

$\lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})} \leq k,$ Where $0 < k < 1$ and k is selected so that the test has a desired significance level α. The next example illustrates these definitions

Example:

Assume that weight X in ounces of a " 10- pound " bag of sugar is N (μ .5) . We shall test the hypothesis $H_0 : \mu = 162$ against the alternative hypothesis $H_1 : \mu \neq 162$. Thus, $\Omega = \{\mu : -\infty < \mu < \infty\}$ and $\omega = \{162\}$. To find the likelihood ratio, we need $L(\widehat{\omega})$ *and* $L(\widehat{\Omega})$ when H_0 is true , μ can take on only one value , namely $\mu = 162$. Hence $L(\widehat{\omega}) = L(162)$. To find $L(\widehat{\Omega})$ we must find the value of μ that maximizes $L(\mu)$. Recall that $\hat{\mu} = \hat{x}$ is the maximum likelihood estimate of μ. Then $L(\widehat{\Omega}) = L(\hat{x})$ and the likelihood ratio λ $L(\widehat{\omega})/L(\widehat{\Omega})$ is given by.

$$\lambda = \frac{(10_\mu)^{-\frac{n}{2}} exp\left[-\left(\frac{1}{10}\right)\sum_{i=1}^{n}(x_i - 162)^2\right]}{(10_\mu)^{-\frac{n}{2}} exp\left[-\left(\frac{1}{10}\right)\sum_{i=1}^{n}(x_i - \underline{x})^2\right]} = \frac{exp\left[-\left(\frac{1}{10}\right)\sum_{i=1}^{n}(x_i - \underline{x})^2 - \left(\frac{n}{10}\right)(\underline{x} - 162)^2\right]}{exp\left[-\left(\frac{1}{10}\right)\sum_{i=1}^{n}(x_i - \underline{x})^2\right]}$$

$$= exp\left[-\frac{n}{10}\left(\underline{x} - 162\right)^2\right]$$

On the one hand , a value of $\underline{x}$ close to 162 would tend to support H_0, and in that case λ is close to 1 . on the other hand , an $\underline{x}$ that differs from 162 by too much would tend to support H_1. (see Figure 1 for the graph of this likelihood ratio when n=5). A critical region for likelihood ratio is given by $\lambda \leq$ k , where k is selected so that the significance level of the test is α . Using this criterion and simplifying the

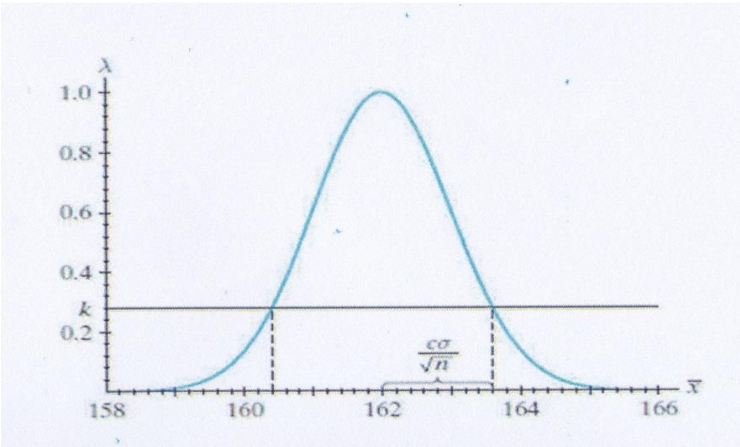


Figure 1. The likelihood ratio for testing $H_0 : \mu = 162$

Inequality as we do when we use the Neyman- Pearson lemma , we find that $\lambda \leq$ k is equivalent to cach of the following Inequalities:

$$-\left(\frac{n}{10}\right)\left(\underline{x} - 162\right)^2 \leq In\, k,$$

$$\left(\underline{x} - 162\right)^2 \geq -\left(\frac{10}{n}\right) In\, k$$

$$\frac{\{\underline{x} - 162\}}{\frac{\sqrt{5}}{\sqrt{n}}} \geq \frac{\sqrt{-\left(\frac{10}{n}\right) In\, k}}{\frac{\sqrt{5}}{\sqrt{n}}} = c$$

Since $z = \left(\underline{x} - 162\right) / \left(\frac{\sqrt{5}}{\sqrt{n}}\right)$ is N(0.1) when $H_0 : \mu = 162$ is true. Let $c = z_{\propto}/2$ Thus the critical region is .

$$c = \left\{ x : \frac{\left|\underline{x} - 162\right|}{\frac{\sqrt{5}}{\sqrt{n}}} \geq c = z_{\propto}/2 \right\}$$

To illustrate , if $\propto = 0.05$ then $z_{0.025} = 1.96$.

Neyman-Pearson Theorem:

تستخدم هذه النظرية لتحديد أفضل منطقة حرجة ولذلك تعد أهم نظرية في الاختبارات الإحصائية.

لاحظ أن النسبة $\frac{f(x;\theta_0)}{f(x;\theta_1)}$ في المثال السابق هي تمهيد لطريقة النظرية لتحديد BCR.

Let $x_1, x_2 \dots x_n$ be r.s having p.d.f $f(x;\theta)$, and the j.p.d.f is $L(\theta; x_1, x_2 \dots x_n) = f(x_1;\theta) f(x_2;\theta) \dots f(x_n;\theta)$. Let C be subset of Sample space such that:

1. $\frac{L(\theta_0; x_i)}{L(\theta_1; x_i)} \le K$ for each point $(x_1, x_2 \dots x_n) \in C$
2. $\alpha = P[(x_1, x_2 \dots x_n) \in C \mid H_0]$,

then C is B.C.R of size α to test $H_0: \theta = \theta_0$
$H_1: \theta = \theta_1$

Proof:

البرهان من الكتاب سيأتيكم مصور. وهو يعتمد على فكرة إذا كانت C هي فقط المنطقة التي حجمها α فالبرهان منتهٍ وأما إذا توجد منطقة أخرى مثل A حجمها α أيضاً فإن علينا إثبات أن القوة للمنطقة C أكبر أو تساوي القوة لـ A ، أي إثبات أن:

$$\int_C L(\theta_1) - \int_A L(\theta_1) \ge 0$$

($\int_C L(\theta_1)$: القوة لـ C ; $\int_A L(\theta_1)$: القوة لـ A)

---

ex: let $x_1, x_2 \dots x_n$ be r.s from $N(\theta, 1)$. find the B.C.R to test $H_0: \theta = 0$ against (ضد) $H_1: \theta = 1$, and find the power of the test.

Sol.: $f(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(x-\theta)^2}{2}\right]$, using Neyman-Pearson Theorem:

$$\frac{L(\theta_0; x_1, x_2 \dots x_n)}{L(\theta_1; x_1, x_2 \dots x_n)} \le K$$

$$\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{\sum(x_i - 0)^2}{2}\right]}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{\sum(x_i - 1)^2}{2}\right]} \le K$$

$$\exp\left[-\frac{\sum x_i^2}{2} + \frac{\sum x_i^2}{2} - \sum x_i + \frac{n}{2}\right] \le K$$

$$\exp\left[-\sum x_i + \frac{n}{2}\right] \le K$$

$$-\sum x_i + \frac{n}{2} \le \ln K$$

$$\sum x_i \ge \frac{n}{2} - \ln K = k_1$$

$$\therefore B.C.R = C = \{\sum x_i \ge k_1\}$$

we can write $C=\{\frac{\sum x_i}{n} > \frac{k_1}{n}\}$, $C=\{\bar{x} \geq k_2\}$

حيث ان $\bar{x}$ مقدر يمتلك صفات جيدة اعلى من $\sum x_i$ مما يجعل الاعتماد عليه في الاختبار الافضل ، ونحن نعرفها ا

$\alpha = P\{\bar{x} \geq k_2\}|H_0$, let $\alpha = 0.05$. نحددها مسبقاً

$0.05 = P\{\bar{x} \geq k_2\}$

ولو فرضنا ان حجم العينة $n=25$ فان :

$\bar{x} \sim N(\theta, \frac{1}{n})$

$0.05 = P\{\frac{\bar{x}-0}{\frac{1}{5}} \geq \frac{k_2 - 0}{\frac{1}{5}}\} = P\{Z \geq 5k_2\}$

$0.95 = P\{Z \leq 5k_2\} \rightarrow 5k_2 = 1.645$ من الجداول

$\therefore k_2 = \frac{1.645}{5} = 0.329$

then B.C.R is:

$C = \{\bar{x} \geq 0.329\}$

the power of the test is :

$p.o.t = P\{\bar{x} > 0.329\}|H_1 = P[\frac{\bar{x}-1}{\frac{1}{5}} > \frac{0.329-1}{\frac{1}{5}}] = P(Z > -3.$

$= P[Z > -3.355] = 0.999+$ قوة اختبار عالية جداً

<u>ملاحظة</u> : وجه آخر للنظرية ... ان تكون المنطقة تخص التوزيع كله وليس معلمة

H_0 : The j.p.d.f is $g(x_1, x_2, \ldots x_n)$

H_1 : " " = $h(x_1, x_2, \ldots x_n)$

$C = \{\frac{g(x_1, x_2 \ldots x_n)}{h(x_1, x_2 \ldots x_n)} \leq k\}$ وستكون منطقة الرفض

الواجب :

1 – Resolve the example using $H_0: \theta = 0$, $H_1: \theta = -1$

اعد حل المثال باستعمال هذه الفرضية

2 – let $x \sim exp(\theta)$, find B.C.R to test $H_0: \theta = 3$ against $H_1: \theta = 2$, if $n = 4$, $\alpha = 0.05$.