STATISTICS

STATISTICAL INFERENCE

Complementary course

for B.Sc. MATHEMATICS

III Semester

(2014 Admission)

CU-CBCSS

UNIVERSITY SCHOOL OF DISTANCE EDUCATION **OF CALICUT**

Calicut University P.O. Malappuram, Kerala, India 673 635

984

School of Distance Education

UNIVERSITY OF CALICUT
SCHOOL OF DISTANCE EDUCATION

STUDY MATERIAL

III Semester

Complementary Course
for B Sc. Mathematics
STATISTICS : STATISTICAL INFERENCE

Prepared and Scrutinised by:

Dr. K.X.Joseph,
Director,
Academic Staff College,
University of Calicut.

Layout & Settings
Computer Section, SDE

SYLLABUS
SEMESTER III

COMPLIMENTARY COURSE III

STATISTICAL INFERENCE

Module 1. Sampling Distributions: Random sample from a population distribution, Sampling distribution of a statistic, Standard error, Sampling from a normal population, Sampling distributions of the sample mean and variance. Chi-square, Student's t and F distributions - derivations, simple properties and inter relationships. 25 hours

Module 2. Theory of Estimation: Point estimation, Desirable properties of a good estimator, unbiasedness, consistency, sufficiency, statement of Fisher Neyman factorization criterion, efficiency. Methods of estimation, method of moments, Method of maximum likelihood-Properties estimators obtained by these methods

25 hours

Module 3. Interval Estimation: Interval estimates of mean, difference of means, variance, proportions and difference of proportions, Large and small sample cases.

10 hours

Module 4. Testing of Hypotheses: Concept of testing hypotheses, simple and composite hypotheses, null and alternative hypotheses, type I and type II errors, critical region, level of significance and power of a test. Neymann-Pearson approach-Large sample tests concerning mean, equality of means, proportions, equality of proportions. Small sample tests based on t distribution for mean, equality ot means and paired mean for paired data. Tests based on F distribution for ratio of variances. Test based on chi square-distribution for variance, goodness of fit and for independence of attributes. 30 hours

MODULE I

SAMPLING DISTRIBUTIONS

Here we are interested with the study of population characteristics based on a sample taken from the population. The process of making inferences about the population based on samples taken from it is called *statistical inference* or inferential statistics. We have already discussed the sampling theory which deals with the methods of selecting samples from the given population. The sample selected should be such that it is capable of exhibiting the characteristics of the population. Here we focus our attention on characteristics calculated from simple random samples.

If $X_1, X_2,... X_n$ are independent and identically distributed r.v.s., we say that they constitute a random sample from the population given by their common distribution. According to this definition a random sample $x_1, x_2,... x_n$ of size 'n' is a collection of random variables $(X_1, X_2, ...X_n)$ such that the variables X_i are all independent but identically distributed as the population random variable X. That means observations of a random variable from the repetitions of a random experiment can be treated as i.i.d. random variables. We can justify this interpretation by means of suitable examples.

Parameter and Statistics

Any measure calculated on the basis of population values is called a 'parameter'. For example, population mean μ population standard deviation σ, population variance σ^2, population correlation coefficient ρ etc. For example, λ is the parameter of a Poisson distribution, μ and σ are the parameters of normal distribution. Statistical inferences are usually based on 'Statistics', that is, on random variables $X_1, X_2, X_3...X_n$ constituting a random sample. In other words, *any measure computed on the basis of sample values is called a statistic.* For example, sample mean $\bar{x}$, sample standard deviation s, sample variance s^2, sample correlation coefficient r etc.

Sampling Distributions

In the case of random sampling the nature of the sampling distribution of a statistic can be deduced theoretically, provided the nature of the population is given, from considerations of probability theory. Let x_1, x_2,...x_n be a random sample taken from the population under investigation. We can consider the random observations as independent random variables X_1, X_2,....X_n following the same distribution of the population. Let $t = g(X_1, X_2 .... X_n)$ being a function of these r.v.s, is also a r.v. That is t is a r.v. The probability distribution of $t = g(X_1, X_2 ... X_n)$ is called sampling distribution of t. In other words, *by a sampling distribution we mean the distribution of a statistic*. If t is a statistic, its sampling distribution is usually denoted as f(t). The sampling distribution of one sample differs from the sampling distribution of another even if both are defined on the same sample. The determination of the sampling distribution of a statistic depends on the selection procedure of the sample, the size of the sample, and the distribution of the population.

Standard error

The standard deviation of the sampling distribution of a statistics is called standard error of the statistic. If t is a statistic with sampling distribution f(t)l the standard error (SE) of t is given by

SE of $t = \sqrt{V(t)}$ where $V(t) = E(t^2) - \{E(t)\}^2$

Uses of Standard Error

Standard Error plays a very important role in large sample theory and forms the basis of testing of hypothesis

1. Since SE is inversely proportional to the sample size n it is very helpful in the determination of the proper size of a sample to be taken to estimate the parameters.
2. It is used for testing a given hypothesis.
3. SE gives an idea about the reliability of a sample. The reciprocal of SE is a measure of reliability of the sample.
4. SE can be used to determine the confidence limits of population parameters.

Here we discuss sampling distributions under two headings.

A. Sampling distribution of small samples drawn from normal population.

B. Sampling distribution of large samples drawn from any large population.

Conventionally by a small sample we mean a sample of size less than 30 where as a sample of size greater than or equal to 30 is treated as a large sample.

A. Sampling distribution of small samples

The probability distribution of statistics computed from small samples drawn from a normal population are discussed here. Being a small sample we get exact probability distribution of these statistics or exact sampling distributions.

Sampling distribution of sample mean

Let x_1, x_2,....x_n be a random samples of size n drawn from $N(\mu, \sigma)$. Let $\bar{x}$ be the sample mean. To find the probability distribution of $\bar{x}$ we shall use the mgf technique. We can consider the random observations as independent and normally distributed r.v.s each having the same normal law $N(\mu, \sigma)$.

$$\therefore M_{x_i}(t) = e^{\mu t + \frac{1}{2}t^2\sigma^2}, \quad i = 1, 2, 3 .... n$$

$$\therefore M_{\bar{x}}(t) = M_{\frac{\sum x_i}{n}}(t) = M_{\sum x_i}(t/n) = \prod_{i=1}^{n} M_{x_i}(t/n)$$

$$= \prod_{i=1}^{n} \left(e^{\frac{\mu t}{n} + \frac{1}{2}\frac{t^2}{n^2}\sigma^2} \right) = \left(e^{\frac{\mu t}{n} + \frac{1}{2}\frac{t^2}{n^2}\sigma^2} \right)^n$$

$$= e^{\mu t + \frac{1}{2}t^2\frac{\sigma^2}{n}} = \text{mgf of } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\therefore \bar{x} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Note:

1.When $\bar{x}$ is the mean of sample of size n drawn from a population which is not normal, the sampling distribution of $\bar{x}$ can be approximated as normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ using central limit theorem, provided n is sufficiently large.

2.In the case of normal population, the distribution of $\bar{x}$ is normal $N(\mu, \sigma/\sqrt{n})$ for any sample size n.

3.The above results show that E($\bar{x}$) = μ, and V($\bar{x}$) = $\frac{\sigma^2}{n}$

$\therefore$ SE of $\bar{x}$ = $\frac{\sigma}{\sqrt{n}}$

The pdf of the random variable $\bar{x}$ is given by

$$f(\bar{x}) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{\frac{-n(\bar{x}-\mu)^2}{2\sigma^2}}, -\infty \le x \le \infty$$

Chi square Distribution

Karl Pearson in about 1900 described the well known probability χ^2. (Square of greek letter chi) χ^2 is a random variable used as a test statistic. Let a random sample $X_1, X_2, \ldots X_n$ be taken from a normal population with mean μ and variance σ^2.

ie.$X_i \to N(\mu, \sigma^2)$. We define χ^2 statistic as the sum of the squares of standard normal variates.

$$\text{ie., } \chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

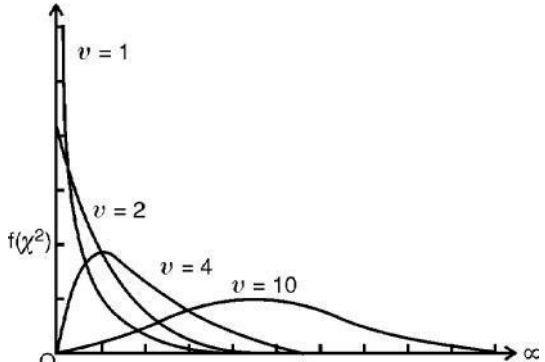
Definition

A continuous r.v. χ^2 assuming values from 0 to ∞ , is said to follow a chi square distribution with n degrees of freedom if its pdf is giving by

$$f(\chi^2) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\left[\frac{n}{2}\right]} e^{-\chi^2/2} (\chi^2)^{\frac{n}{2}-1}, 0 \le \chi^2 \le \infty$$

= 0 , otherwise

Here n is the parameter of the distribution and we write this as $\chi^2 \to \chi^2$ (n) df.



χ^2 distribution curves

The shape of χ^2 curve depends on the value of n. For small n, the curve is positively skewed. As n, the degrees of freedom increases, the curve approaches to symmetry rapidly. For large n the χ^2 is approximately normally distributed. The distribution is unimodal and continuous.

Degrees of freedom

The phrase 'degrees of freedom' can be explained in the following intutive way.

Let us take a sample of size n = 3 whose average is 5. Therefore the sum of the observations must be equal to 15. That means $X_1 + X_2 + X_3 = 15$

We have complete freedom in assigning values to any two of the three observations. After assigning values to any two, we have no freedom in finding the other value, because the latter is already determined. If we

 assign values to $X_1 = 3$, $X_2 = 8$, then $X_3 = 4$.

Given the mean, we have only n – 1 degrees of freedom to compute the mean of a set of n observations. Here we have complete freedom in assigning values to any n – 1 observations. ie., they are independent observations. The complete freedom is called the degrees of freedom. Usually it is denoted by ν . Accordingly the sample variance has n – 1 degrees of freedom. But the sum of squared deviations from the population mean μ ie., $\sum_{i=1}^{n}(x_i - \mu)^2$ has n degrees of freedom. Thus by degrees of freedom we mean the number of independent observations in a distribution or a set.

Moments of χ^2

Mean

$E(\chi^2) = n$

Variance

$V(\chi^2) = E(\chi^2)^2 - [E(\chi^2)]^2$

Moment generating function

$$M_{\chi^2}(t) = E(e^{t\chi^2}) = (1-2t)^{-n}$$

Sampling Distribution of Sample Variance s^2

Let $x_1, x_2, \ldots x_n$ be a random sample drawn from a normal population $N(\mu, \sigma)$. Let $\bar{x}$ be the sample mean and s^2 be its variance. We can consider the random observations as independent and normally distributed r.v.s. with mean μ and variance σ^2.

ie., $X_i \to N(\mu, \sigma).i = 1, 2, \ldots. n$

$$\therefore \frac{X_i - \mu}{\sigma} \to N(0,1)$$

$$\text{Then} \left(\frac{X_i - \mu}{\sigma}\right)^2 \to \chi^2 \quad (1)\ df$$

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \to \chi^2 \quad (n)\ df,\ \text{by additive property}$$

Also we know that $\bar{x} \to N(\mu, \sigma/\sqrt{n})$

$$\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right) \to N(0,1)$$

$$\therefore \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2 \to \chi^2(1)$$

d.f.

Now consider the expression

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\bar{x} - \mu)^2$$

$$= ns^2 + 2(\bar{x} - \mu) \times 0 + n(\bar{x} - \mu)^2$$

ie., $\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{ns^2}{\sigma^2} + \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2$, by dividing each term term by σ^2

ie., $\chi^2(n) = \frac{ns^2}{\sigma^2} + \chi^2(1)$

Here we have seen that the LHS follows χ_2 (n). So by additive property of χ^2 distribution we have.

$$\frac{ns^2}{\sigma^2} \to \chi^2 \quad (n-1)$$

Therefore the pdf of χ^2 distribution with (n – 1) df is given by

$$f\left(\frac{ns^2}{\sigma^2}\right) = \frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\left[\frac{(n-1)}{2}\right]} e^{\frac{-ns^2}{2\sigma^2}} \left(\frac{ns^2}{\sigma^2}\right)^{\frac{n-1}{2}-1}, 0 \le \frac{ns^2}{\sigma^2} \le \infty$$

Put $u = \frac{ns^2}{\sigma^2}$, $\frac{du}{ds^2} = \frac{n}{\sigma^2}$

$f(s^2) = f(u) \left|\frac{du}{ds^2}\right|$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\left[\frac{(n-1)}{2}\right]} e^{\frac{-ns^2}{2\sigma^2}} \left(\frac{ns^2}{\sigma^2}\right)^{\frac{n-1}{2}-1} \times \frac{n}{\sigma^2}$$

$$= \frac{\left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}}}{\left[\frac{(n-1)}{2}\right]} e^{\frac{-ns^2}{2\sigma^2}} (s^2)^{\frac{n-1}{2}-1}, 0 \le s^2 \le \infty$$

This is the sampling distribution of s^2. It was discovered by a German mathematician. Helmert in 1876. We can determine the mean and variance of s^2 similar to the case of χ^2 distribution.

Thus $\mathbf{E(s^2)} = \frac{n-1}{n}\sigma^2$

$\mathbf{V(s^2)} = \frac{n-1}{2.n^2}\sigma^4$

SD of $s^2 = \sqrt{\frac{2\sigma^4(n-1)}{n^2}}$

Definition

A continuous random variable t assuming values from $-\infty$ to $+\infty$ with the pdf given by

$$f(t) = \frac{1}{\sqrt{n}\,\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, -\infty \le t \le \infty$$

is said to follow a student's t distribution with n degrees of freedom. The t distribution depends only on 'n' which is the parameter of the distribution.

For n = 1, the above pdf reduces to $f(t) = \frac{1}{\pi(1+t^2)}$, which is known as the Cauchy pdf.

Note: $\beta(m, n) = \frac{\Gamma m\,\Gamma n}{\Gamma(m+n)}$

Definition of 't' statistic

If the random variables $Z \to N(0, 1)$ and $Y \to \chi^2(n)$ and if Z and Y are independent then the statistic defined by

$t = \frac{Z}{\sqrt{Y/n}}$ follows the Student's 't' distribution with n df.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \to t_{(n-1)} \quad df$$

Later we can see that the above r.v. will play an important role in building schemes for inference on μ.

$$\text{SE of } t = \frac{s}{\sqrt{n-1}}$$

ii. *Distribution of t for comparison of two samples means:*

We now introduce one more t statistic of great importance in applications.

Let $\bar{x}_1$ and $\bar{x}_2$ be the means of the samples of size n_1 and n_2 respectively drawn from the normal populations with means μ_1 and μ_2 and with the same unknown variance σ^2. Let s_1^2 and s_2^2 be the variances of the samples.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightarrow t(n_1 + n_2 - 2) df$$

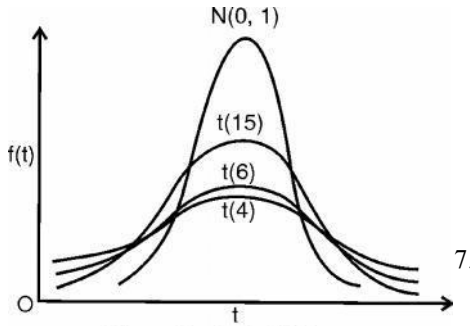
Characteristics of t distribution

1. The curve representing t distribution is symmetrical about t = 0
2. The t curve is unimodal
3. The mode of t curve is 0 ie, at t = 0, it has the maximum probability
4. The mean of the t distribution is 0 ie., E(t) = 0
5. All odd central moments are 0.
6. The even central moments are given by

$$\mu_{2r} = \mu'_{2r} = \frac{(2r-1)(2r-3)....3.1}{(v-2)(v-4)...(v-2r)} v^{2r}$$, υ denotes df

Putting r = 1, $\mu_2 = \frac{v}{v-2}$, for υ > 2

Hence variance $\sigma^2 = \frac{v}{v-2}$, it is slightly greater than 1. Therefore, it has a greater dispersion than normal distribution. The exact shape of the t distribution depends on the sample size or the degrees of freedom υ. If υ is small, the curve is more spread out in the tails and flatter around the centre. As υ increases, t distribution approaches the normal distribution.



Three Student's 't' Curves and a Standard Normal Curve

7. For each different number of degrees of freedom, there is a distribution of t. Accordingly t distribution is not a single distribution, but a family of distributions.
8. The mgf does not exist for t distribution.
9. The probabilities of the t distribution have been tabulated. (See tables) The table provides

$$P\{|t| > t_0\} = \int_{-\infty}^{-t_0} f(t)\,dt + \int_{t_0}^{\infty} f(t)\,dt$$

$$= 2\int_{t_0}^{\infty} f(t)\,dt \text{, because of symmetry.}$$

Values of t_0 have been tabulated by Fisher for different probabilities say 0.10, 0.05. 0.02 and 0.01. This table is prepared for t curves for υ = 1, 2, 3,... 60.

Snedecor's F Distribution

Another important distribution is F distribution named in honour of Sir. Ronald. A. Fisher. The F distribution, which we shall later find to be of considerable practical interest, is the distribution of the ratio of two independent chi square random variables divided by their respective degrees of freedom.

If U and V are independently distributed with Chi square distributions with n_1 and n_2 degrees of freedom.

$$\frac{U/n_1}{V/n_2}$$

then, $F = \frac{U/n_1}{V/n_2}$ is a random variable following an F distribution with (n_1, n_2) degrees of freedom.

Definition

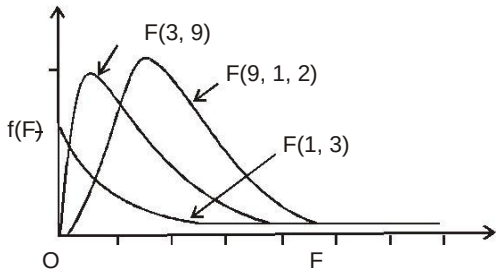
A continuous random variable F, assuming values from 0 to ∞ and having the pdf given by

$$f(F) = \frac{\frac{n_1}{2}\ \frac{n_2}{2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \times \frac{F^{\frac{n_1}{2}-1}}{(n_1 F + n_2)^{\frac{n_1+n_2}{2}}}, 0 \le F \le \infty$$

is said to follow an F distribution with (n_1, n_2) degrees of freedom. The credit for its invention goes to G.W. Snedecor. He chose the letter F to designate the random variable in honour of R.A. Fisher.

The F distributions has two parameters, n_1 and n_2 corresponding to the degrees of freedom of two χ^2 random variables in the ratio; the degrees of freedom for the numerator random variable is listed first and the ordering makes a difference, if $n_1 \neq n_2$. The reciprocal of the F random variable $\left(ie., \frac{1}{F} = \frac{V/n_2}{U/n_1}\right)$ again is the ratio of two independent χ^2 r.v.s. each divided by its degrees of freedom, so it again has the F distribution, now with n_2 and n_1 degrees of freedom.

In view of its importance, the F distribution has been tabulated extensively. See table at the end of this book. This contain values of $F_{(\alpha; n_1, n_2)}$ for $\alpha = 0.05$ and 0.01, and for various values of n_1 and n_2, where $F_{(\alpha; n_1, n_2)}$ is such that the area to its right under the curve of the F distribution with(n_1, n_2) degrees of freedom is equal to α. That is $F_{(\alpha; n_1, n_2)}$ is such that $P(F \ge F_{(\alpha; n_1, n_2)}) = \alpha$



Applications of F distribution arise in problems in which we are interested in comparing the variances σ_1^2 and σ_2^2 of two normal populations. Let us have two independent r.v.s. X_1 and X_2 such that $X_1 \rightarrow N(\mu_1, \sigma_1^2)$ and $X_2 \rightarrow N(\mu_2, \sigma_2^2)$. The random samples of sizes n_1 and n_2 are taken from the above population. The sample variances s_1^2 and s_2^2 are computed. Then we can observe that the statistic $\frac{n_1 s_1^2}{n_1 - 1} \Big/ \frac{n_2 s_2^2}{n_2 - 1}$ has an F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

Characteristics of F distribution

1. The mean of F distribution is $\frac{n_2}{n_2 - 2}$

   ie., E (F) = $\frac{n_2}{n_2 - 2}$, No mean exist for $n_2 \le 2$.
2. The variance of the F distribution is

   $V(F) = \frac{2n_2^2 (n_1 + n_2 - 2)}{n_1 (n_2 - 2)^2 (n_2 - 4)}$, No variance exist if $n_2 \le 4$.
3. The distribution of F is independent of the population variance σ^2.
4. The shape of the curve depends on n_1 and n_2 only. Generally it is non symmetric and skewed to the right. However when one or both

parameters increase, the F distribution tends to become more and more symmetrical.

5. If $F \to F(n_1, n_2)$ df, then $\frac{1}{F} \to F(n_2, n_1)\ df$

   It is called reciprocal property of F distribution

6. Th two important uses of F distribution are (a) to test the equality of two normal population variances and (b) to test the equality of three or more population means.

7. Since the applications of F distribution are based on the ratio of sample variances, the F distribution is also known as variance - ratio distribution.

Inter relationship between t, χ^2 and F distributions;

1. The square of t variate with n df is F(1, n)

Let $x_1, x_2 .... x_n$ be a random sample drawn from $N(\mu, \sigma)$. We can consider the random observations as i.i.d. r.v.s.

Then $\bar{x} = \frac{1}{n}\Sigma x_i$, $s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

We know that $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$

$$Y = \frac{ns^2}{\sigma^2} \to \chi^2(n-1)$$

Define

$$t = \frac{X}{\sqrt{\frac{y}{n-1}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}$$

$$\therefore t^2 = \frac{Square of\ N(0,1)}{\chi^2(n-1)/(n-1)} = \frac{\chi^2(1)/1}{\chi^2(n-1)/(n-1)} \to F(1, n-1)$$

ie., the square of a variate with n –1 df is F(1, n –1).

So the square of at variate with n df is F(1, n)

2. **F is the ratio of two χ^2**

$$\text{Let } F(n_1 - 1, n_2 - 1) = \frac{n_1 S_1^2}{n_1 - 1} \Big/ \frac{n_2 S_2^2}{n_2 - 1}$$

$$= \frac{\frac{n_1 S_1^2}{\sigma_1^2}/(n_1 - 1)}{\frac{n_2 S_2^2}{\sigma_2^2}/(n_2 - 1)}$$

$$= \frac{\chi^2(n_1 - 1)/(n_1 - 1)}{\chi^2(n_2 - 1)/(n_2 - 1)}$$

Hence the result.

SOLVED PROBLEMS

Example l

Let X be N (100, $10^{2)}$ distributed. Find the sample size n so as to have P ($\bar{x} \geq$ 101.645) = 0.05

Solution

Given $X \to N(100, 10^2)$. Take a random sample of size n from the population. Let $\bar{x}$ be its mean, then $\bar{x} \to N(100, \frac{10}{\sqrt{n}})$

We have to find n such that

$$p(\bar{x} \geq 101.645) = 0.05$$

$$\text{ie } P\left(\frac{\bar{x} - 100}{10/\sqrt{n}} \geq \frac{101.645 - 100}{10/\sqrt{n}}\right) = 0.05$$

$$\text{ie, } P\left(Z \geq \frac{1.645\sqrt{n}}{10}\right) = 0.05$$

ie, $P\left(Z \geq 0.1645\sqrt{n}\right) = 0.05$

ie, $P\left(0 \leq Z \leq 0.1645\sqrt{n}\right) = 0.45$

From normal table, $0.1645\sqrt{n} = 1.645$

$\therefore \sqrt{n} = 10$, ie, **n = 100**

Example 2

If X_1, X_2, X_3 and X_4 are independent observations from a univariate normal population with mean zero and unit variance, state giving reasons the sampling distribution of the following.

(i) $u = \dfrac{\sqrt{2}\,X_3}{\sqrt{X_1^2 + X_2^2}}$ (ii) $v = \dfrac{3X_4^2}{X_1^2 + X_2^2 + X_3^2}$

Solution

Given $X_i \to N(0, 1)$, $i = 1, 2, 3, 4$

$\therefore X_i^2 \to \chi^2(1)$, $i = 1, 2, 3, 4$

$\therefore X_i^2 + X_2^2 \to \chi^2(2)$ and $X_i^2 + X_2^2 + X_3^2 \to \chi^2(3)$

(i) $u = \dfrac{\sqrt{2}\,X_3}{\sqrt{X_1^2 + X_2^2}}$

$= \dfrac{X_3}{\sqrt{\dfrac{X_1^2 + X_2^2}{2}}} \approx \dfrac{N(0,1)}{\sqrt{\dfrac{\chi^2(2)}{2}}} \to$ **t(2)** *df*

(ii) $v = \dfrac{3X_4^2}{X_1^2 + X_2^2 + X_3^2}$

$= \dfrac{\dfrac{X_4^2}{X_1^2 + X_2^2 + X_3^2}}{3} = \dfrac{\chi^2(1)}{\chi^2(3)/3}$

$= \dfrac{\chi^2(1)/1}{\chi^2(3)/3} =$ **F(1, 3)**

Example 3

If X_1, and X_2, are independent χ^2 r.v.s. each with one degree of freedom. Find λ such that $P(X_1 + X_2 > \lambda) = \dfrac{1}{2}$

Solution

Given $X_1 \to \chi^2(1)$

$X_2 \to \chi^2(1)$

$\therefore Y = X_1 + X_2 \to \chi_2(2)$, by additive property

We have to find λ such that

$P(X_1 + X_2 > \lambda) = \dfrac{1}{2}$

ie, $P(Y > \lambda) = \dfrac{1}{2}$

ie, $\int_\lambda^\infty f(y)\,dy = \dfrac{1}{2}$

ie, $\int_\lambda^\infty \dfrac{\left(\frac{1}{2}\right)^{\frac{2}{2}}}{\left|\frac{2}{2}\right|} e^{-\frac{y}{2}} y^{\frac{2}{2}-1}\,dy = \dfrac{1}{2}$

$$\frac{1}{2}\int_{\lambda}^{\infty} e^{-\frac{y}{2}}\, dy = \frac{1}{2}$$

$$\left[\frac{e^{-y/2}}{-1/2}\right]_{\lambda}^{\infty} = 1, \qquad -2\left[0 - e^{-\frac{\lambda}{2}}\right] = 1$$

$$2e^{-\frac{\lambda}{2}} = 1, \qquad e^{-\frac{\lambda}{2}} = \frac{1}{2}$$

$$\therefore \; e^{\frac{\lambda}{2}} = 2, \quad \text{ie} \quad \frac{\lambda}{2} = \log_e 2$$

$$\therefore \; \lambda = \mathbf{2 \log_e 2}$$

EXERCISES

Multiple Choice Question

1 Simple random sample can be drawn with the help of
a. random number tables b. Chit method
c. roulette wheel d. all the above

1 Formula for standard error of sample mean $\bar{x}$ based on sample of n^2, when population consisting of N items is
a. s/n b. $s/\sqrt{n-1}$ c. $s/\sqrt{N-1}$ d. $s/\sqrt{n}$

1 Which of following statement is true
a. more the SE, better it is
b. less the SE, better it is
c. SE in always zero
d. SE is always unity

1 Student's 't' distribution was discovered by
a. G.W. Snedecor b. R.A. Fisher
c. W.Z. Gosset c. Karl Pearson

1 Student's t distribution curve is symmetrical about mean, it means that
a. odd order moments are zero
b. even order moments are zero
c. both (a) and (b)d. none of (a) and (b)

1 If $X \to N(0, 1)$ and $Y \to \chi^2$ (n), the distribution of the variate $x/\sqrt{\chi^2/n}$ follows
a. Cauchy's distribution
b. Fisher's t distribution
c. Student's t distribution
d. none of the above

1 The degrees of freedom for student's 't' based on a random sample of size n is
a. n – 1 b. n c. n – 2 d. (n – 1)/2

1 The relation between the mean and variance of χ^2 with n df is
a. mean = 2 variance b. 2 mean = variance
c. mean = variance d. none of the above

1 Chi square distribution curve is
a. negatively skewed b. symmetrical
c. positively skewed d. None of the above

1 Mgf of chi square distribution with n df is
a. $(1 - 2t)^{n/2}$ b. $(1 - 2it)^{n/2}$
c. $(1 - 2t)^{-n/2}$ d. $(1 - 2it)^{-n/2}$

1 F distribution was invented by
a. R.A. Fisher b. G.W. Snedecor
c. W.Z. Gosset d. J. Neymann

1 The range of F - variate is
a. $-\infty$ to $+\infty$ b. 0 to 1
c. 0 to ∞ d. $-\infty$ to 0

1 The relation between student's t and F distribution is
a. $F_{1,1} = t_n^2$ b. $F_{n,1} = t_1^2$
c. $t_\infty = F_{1,n}$ d. none of the above

ı Student's t curve is symmetric about
a. t = 0b. t = μ c. t = 1d. t = n

Fill in the blanks

ı If the number of units in a population are limited, it is known as ..... population.

ı Any population constant is called a .....

ı Another name of population is .....

ı The index of precision of an estimator is indicated by its .....

ı In the above case, the distribution of $\frac{\chi_1^2/n_1}{\chi_2^2/n_2}$ is

ı The mean of the χ^2 distribution is of its variance

ı If the df is for Chi square distribution is large, the chi-square distribution tends to .....

ı t distribution with 1 df reduces to .....

ı The ratio of two sample variances is distributed as .....

ı The relation between Fisher's Z and Snedecor's F is .....

ı The square of any standard normal variate follows ..... distribution.

Very Short Answer Questions

ı What is a random sample? ı
Define the term 'statistic'.
ı Define the term 'parameter'. ı What
is sampling distribution? ı Define
standard error.

ı What is the relationship between SE and sample size.

ı Define χ^2 distribution with n df. ı
Define student's t distribution.

ı Define F distribution.

ı Give an example of a t statistic.

ı Give an example of an F statistic. ı
Define sampling error.
ı Give four examples of statistics. ı Give
four examples of parameters

ı What is the relationship between t and F.

ı What are the importance of standard error? ı
What are the mean and variance of s^2

Short Essay Questions

ı Explain the terms (i) parameter (ii) statistic (iii) sampling distribution. ı
What is a sampling distribution? Why does one consider it?

ı Explain the meaning of sampling distribution of a statistic T and the standard error of T. Illustrate with the sample proportion.

ı Explain the terms (i) statistic (ii) standard error and (iii) sampling distributions giving suitable examples.

ı Define sampling distribution and give an example.

ı Derive the sampling distribution of mean of samples from a normal population.

Long Essay Questions

ı State the distribution of the sample varience from a normal population

ı Define χ^2 and obtain its mean and mode.

ı Define χ^2 statistic. Write its density and establish the additive property.

ı Give the important properties of χ^2 distribution and examine its relationship with the normal distribution.

ı Define a χ^2 variate and give its sampling distribution. Show that its variance is twice its mean.

ı Define the F statistic, Relate F to the t statistic and $F_{n,m}$ to $F_{m,n}$

MODULE II
THEORY OF ESTIMATION

The Theory of estimation was expounded by Prof: R.A. Fisher in his research research papers round about 1930. Suppose we are given a random sample from a population, the distribution of which has a known mathematical form but involves a certain number of unknown parameters. The technique of coming to conclusion regarding the values of the unknown parameters based on the information provided by a sample is known as the problem of 'Estimation'. This estimation can be made in two ways.

i. Point Estimation ii. Interval Estimation

Point Estimation

If from the observations in a sample, a single value is calculated as an estimate of the unknown parameter, the procedure is referred to as point estimation and we refer to the value of the statistic as a point estimate. For example, if we use a value of $\bar{x}$ to estimate the mean μ of a population we are using a point estimate of μ. Correspondingly, we, refer to the statistic $\bar{x}$ as point estimator. That is, the term 'estimator' represents a rule or method of estimating the population parameter and the estimate represents the value produced by the estimator.

An estimator is a random variable being a function of random observations which are themselves random variables. An estimate can be counted only as one of the possible values of the random variable. So estimators are statistics and to study properties of estimators, it is desirable to look at their distributions.

Properties of Estimators

There are four criteria commonly used for finding a good estimator. They are:

1. Unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

1. Unbiasedness

An unbiased estimator is a statistic that has an expected value equal to the unknown true value of the population parameter being estimated. An estimator not having this property is said to be biased.

Let X be random variable having the pdf f(x , θ), where θ may be unknown. Let $X_1, X_2....X_n$ be a random sample taken from the population represented by X. Let

$t_n = t(X_1, X_2....X_n)$ be an estimator of the parameter θ.

If $E(t_n) = \theta$ for every n, then estimator t_n is called unbiased estimator.

2. Consistency

One of the basic properties of a good estimator is that it provides increasingly more precise information about the parameter θ with the increase of the sample size n. Accordingly we introduce the following definition.

Definition

The estimator $t_n = t(X_1, X_2....X_n)$ of parameter θ is called consistent if t_n converges to θ in probability. That is, for ε > 0

$$\lim_{n\to\infty} P\left(|t_n - \theta| \le \varepsilon\right) = 1 \text{ or } \lim_{n\to\infty} P\left(|t_n - \theta| \ge \varepsilon\right) = 0$$

The estimators satisfying the above condition are called weakly consistent estimators.

The following theories gives a sufficient set of conditions for the consistency of an estimator.

Theorem

An estimator t_n, is such that $E(t_n) = \theta_n \to \theta$ and $V(t_n) \to 0$ as n $\to \infty$, the estimator t_n is said to be consistent for θ.

3. Efficiency

Let t_1 and t_2 be two unbiased estimators of a parameter θ. To choose between different unbiased estimators, one would reasonably consider their variances, ie., If $V(t_1)$ is less than $V(t_2)$ then t_1 is said to be more efficient than t_2. That is as variance of an estimator decreases its efficiency

increases. $V(t_1)/V(t_2)$ is called the relative efficiency of t_2 with respect to t_1 and we can use this to compare the efficiencies of estimators.

4. Sufficiency

An estimator t is said to be sufficient if it provides all information contained in the sample in respect of estimating the parameter θ. In other words, an estimator t is called sufficient for θ, if the conditional distribution of any other statistic for given t is independent of θ.

Factorisation Theorem

Let $x_1, x_2....x_n$ be a random sample of size n from a population with density functions f(x; θ) where θ denotes the parameter, which may be unknown. Then a statistic $t = t(x_1, x_2....x_n)$ is sufficient if and only if the joint probability density function of $x_1, x_2....x_n$ (known as likelyhood of the sample) is capable of being expressed in the form

$$L(x_1, x_2....x_n; \theta) = L_1(t, \theta).\ L_2(x_1, x_2....x_n)$$

where the function $L_2(x_1, x_2....x_n)$ is non negative and does not involve the parameter θ and the function $L_1(t, \theta)$ is non negative and depending on the parameter θ.

Method of Moments

This is the oldest method of estimation introduced by Karl Pearson. According to it to estimate k parameters of a population, we equate in general, the first k moments of the sample to the first k moments of the population. Solving these k equations we get the k estimators.

Let X be a random variable with the probability density function f(x, θ). Let μ_r' be the r-th moment about O. $\mu_r' = E(X^r)$. In general, μ_r' will be a known function of θ and we write $\mu_r' = \mu_r'(\theta)$. Let $x_1, x_2....x_n$ be a random sample of size n drawn from the population with density function f(x, θ). Then r-th sample moment will be $m_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$. Form the equation $m_r' = \mu_r'(\theta)$ and solve for θ. Let $\hat{\theta}$ be the solution of θ. Then $\hat{\theta}$ is the estimator of θ obtained by the method of moments.

Properties

i. Moment estimators are asymptotically unbiased.

ii. They are consistent estimators.

iii. Under fairly general conditions, the distribution of moment estimators are asymptotically normal.

Method of Maximum Likelyhood

The method of moments is one procedure for generating estimators of unknown parameters, it provides an attractive rationale and is generally quite easy to employ. In 1921 Sir. R. A. Fisher proposed a different rationale for estimating parameters and pointed out a number of reasons that it might be preferable. The procedure proposed by Fisher is called method of Maximum likelyhood and is generally acknowledged to be superior to the method of moments. In order to define maximum likelyhood estimators, we shall first define the likelyhood function.

Likelyhood function

The likelyhood function of n random variables $X_1, X_2....X_n$ is defined to be the joint probability density function of the n random variables, say $f(x_1, x_2....x_n; \theta)$ which is considerd to be a function of θ. In particular suppose that X is a random variable and $X_1, X_2....X_n$ is a random sample of X having the density f(x, θ). Also $x_1, x_2....x_n$ are the observed sample values. Then the likelyhood function is defined as

$$L(x_1, x_2.....x_n;\ \theta) = f(x_1, x_2\ .....x_n; \theta)$$

$$= f(x_1; \theta)\ f(x_2; \theta)..... f\ (x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i, \theta)$$

The likelyhood function can also be denoted as L(X; θ) or L(θ). The likelyhood function $L(x_1, x_2 ..... x_n; \theta)$ give the likelyhood that the random variables assume a particular value $x_1, x_2 ..... x_n$.

The principle of maximum likelyhood consists in finding an estimator of the parameter which maximises L for variations in the parameter. Thus the problem of finding a maximum likelyhood estimator is the problem of finding the value of θ that maximises L(θ). Thus if there exists a function t = t($x_1, x_2 .....x_n$) of the sample values which maximises L for variations in θ, then t is called Maximum likelyhood Estimator of θ. (MLE).

Thus t is a solution if any of

$$\frac{\partial L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 L}{\partial \theta^2} < 0$$

Also L and log L have maximum at the same value of θ we can take log L instead of L which is usually more convenient in computations.

Thus MLE is the solution of the equations $\frac{\partial \log L}{\partial \theta} = 0$, provided

$$\frac{\partial^2 \log L}{\partial \theta^2} < 0$$

The maximum likelyhood estimator can also be used for the simultaneous estimation of several parameters of a given population. In that case we must find the values of the parameters that maximise the likelyhood function.

Properties of Maximum likelyhood estimators.

Under certain very general conditions (called regularity conditions) the maximum likelyhood estimators possess several nice properties.

1. Maximum likelyhood estimators are consistent
2. The distribution of maximum likelyhood estimators tends to normality for large samples.
3. Maximum likelyhood estimators are most efficient.
4. Maximum likelyhood estimators are sufficient if sufficient estimators exists
5. Maximum likelyhood estimators are not necessarily unbiased.
6. Maximum likelyhood estimators have invariance property, (ie. if t is the m.l.e. of θ , then g(t) is also the MLE of g (θ), g being a single valued function of θ with a unique inverse).

SOLVED PROBLEMS

Example l

Let x_1, x_2, x_3 .... x_n be a random sample drawn from a given population with mean μ and variance σ^2. Show that the sample mean $\bar{x}$ is an unbiased estimator of population mean μ.

Solution

We know that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Taking expected value, we get

$$E(\bar{x}) = E\left\{\frac{1}{n}\sum_{i=1}^{n} x_i\right\} = \frac{1}{n}E\left\{\sum_{i=1}^{n} x_i\right\}$$

$$= \frac{1}{n}E(x_1 + x_2 + ... + x_n)$$

$$= \frac{1}{n}\{E(x_1) + E(x_2) + ... + E(x_n)\}$$

Now $E(x_i) = \mu$ (given)

$$\therefore E(\bar{x}) = \frac{1}{n}\{\mu + \mu + ... + \mu\} = \frac{n\mu}{n} = \mu$$

Therefore sample mean is an unbiased estimator of population mean.

Example 2

Let $x_1, x_2, x_3...., x_n$ is a random sample from a normal distribution N(μ, 1) show that

$t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiased estimator of $\mu^2 + 1$

Solution

We are given that

$E(x_i) = \mu$

$V(x_i) = 1$ for every i = 1, 2, 3,...n

Now $V(x_i) = E(x_i^2) - [E(x_i)]^2$

$\therefore E(x_i^2) = \mu^2 + 1$

$$E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\mu^2+1) = \frac{1}{n}n(\mu^2+1)$$

$= \mu^2 + 1$

Example 3

If T is an unbiased estimator of θ, show that T^2 and $\sqrt{T}$ are the biased estimator of θ^2 and $\sqrt{\theta}$ respectively.

Solution

Given $E(T) = \theta$

Now $var(T) = E[T - E(T)]^2 \neq 0$ as var > 0

or $E\{T^2 - 2T\theta + \theta^2\} = E(T)^2 - 2\theta E(T) + \theta^2 \neq 0$

$E\{T^2) - 2\theta^2 + \theta^2 \neq 0$

or $E(T)^2 \neq \theta^2$, ie., T^2 is biased

Alsovar $(\sqrt{T}) = E[\sqrt{T} - E(\sqrt{T})]^2 \neq 0$

$= E(T) - \{E(\sqrt{T})\}^2 \neq 0$

or $E(T) = \theta \neq \{E(\sqrt{T})\}^2$

$E(\sqrt{T}) \neq \sqrt{\theta}$.

Hence the result. ie., $\sqrt{T}$ is not an unbiased estimator of $\sqrt{\theta}$.

Example 4

$x_1, x_2, ..., x_n$ is a random sample from a population following Poisson distribution with parameter λ. Suggest any three unbiased estimators of λ.

Solution

Since x_i is a random observation from a Poisson population with parameter λ, $E(x_i) = \lambda$, i = 1, 2, ... n

$$\therefore t_1 = x_1, t_2 = \frac{x_1 + x_2}{2}, t_n = \frac{x_1 + x_2 + ... + x_n}{n}$$

are unbiased estimators of λ. It may be noted that .

$E(t_1) = E(x_i) = \lambda$

$E(t_2) = \frac{1}{2}[E(x_1) + E(x_2)] = \frac{1}{2}[\lambda + \lambda] = \lambda$

$E(t_n) = \frac{1}{n}E(x_1 + x_2 + ... x_n)$

$= \frac{1}{n}[E(x_1) + E(x_2) + ...]$

$= \frac{1}{n}[\lambda + \lambda + ... + \lambda] = \frac{n\lambda}{n} = \lambda$

$\therefore$ t_1, t_2 and t_n are unbiased estimators of λ.

Example 5

Show that sample variance is a consistent estimator of the population variance in the case of normal population N(μ, σ).

Solution

Let $x_1, x_2, ..., x_n$ be a random sample from $N(\mu, \sigma^2)$. Let $\bar{x}$ be the mean and s^2 is its variance. From the sampling distribution of s^2, we have

$$E(s^2) = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2$$

But $V(s^2) = 2.\frac{n-1}{n^2}\sigma^4 \to 0 \quad as \quad n \to \infty$

Thus the sufficient conditions are satisfied.

^{2} is consistent for σ^2

Example 6

Give an example of estimators which are

(a) Unbiased and efficient,

(b) Unbiased and inefficient,

(c) Biased and inefficient.

(a) The sample mean $\bar{x}$ and modified sample variance $S^2 = \frac{n}{n-1}s^2$ are two such examples.

1

(b) The sample median, and the sample statistic $\frac{1}{2}[Q_1 + Q_3]$ where Q_1 and Q_3 are the lower and upper sample quartiles, are two such examples. Both statistics are unbiased estimators of the population mean, since the mean of their sampling distribution is the population mean.

(c) The sample standard deviation s, the modified standard deviation $\bar{s}$, the mean deviation and the semi-in-terquartile range are four such examples.

Example 7

For the rectangular distribution oven an interval (α, β); $\alpha < \beta$. Find the maximum likelihood estimates of α and β.

Solution

For the rectangular distribution over (α, β), the p.d.f. of X is given by

$f(x) = \frac{1}{\beta - \alpha}, \alpha \le x \le \beta$

Take a random sample $x_1, x_2, \ldots x_n$ from (α, β)

Then $L(x_1, x_2, \ldots x_n ; \alpha, \beta) = \frac{1}{\beta-\alpha}.\frac{1}{\beta-\alpha}\ldots\frac{1}{\beta-\alpha} = \left(\frac{1}{\beta-\alpha}\right)^n$

L is maximum when $(\beta - \alpha)$ is maximum i.e. when β is minimum and α is maximum. If the sample observations are arranged in ascending order, we have

$\alpha \le x_1 \le x_2 \le x_3 \ldots \le x_n \le \beta$

Here the minimum value of β consistent with the sample is x_n and maximum value of α is x_1. Thus the M.L.E.'s of α and β are

$\hat{\alpha} = x_1, \hat{\beta} = x_n$

EXERCISES

Multiple Choice Questions

1 An estimator is a function of
a. population observations
b. sample observations
c. Mean and variance of population
d. None of the above

1 Estimate and estimator are
a. synonyms b. different
c. related to population d. none of the above

1 The type of estimates are
a. point estimate
b. interval estimate
c. estimates of confidence region
d. all the above

1 The estimator $\bar{x}$ of population mean is
a. an unbiased estimator b. a consistant estimator
c. both (a) and (b) d. neither (a) nor (b)

1 Factorasation theorem for sufficiency is known as
a. Rao - Blackwell theorem

b. Cramer Rao theorem

c. Chapman Robins theorem

d. Fisher - Neymman theorem

1 If t is a consistent estimator for θ, then

a. t is also a consistent estimator for θ^2

b. t^2 is also consistent estimator for θ

c. t^2 is also consistent estimator for θ^2

d. none of the above

1 The credit of inventing the method of moments for estimating parameters goes to

a. R.A. Fisher b. J. Neymann

c. Laplace d. Karl Pearson

1 Generally the estimators obtained by the method of moments as compared to MLE are

a. Less efficient b. more efficient

c. equally efficient d. none of these

Fill in the blanks

1 An estimator is itself a .....

1 A sample constant representing a population parameter is known as .....

1 A value of an estimator is called an .....

1 A single value of an estimator for a population parameter θ is called its ..... estimate

1 The difference between the expected value of an estimator and the value of the corresponding parameter is known as .....

1 The joint probability density function of sample variates is called .....

1 A value of a parameter θ which maximises the likelyhood function is known as ..... estimate of θ

1 An unbiased estimator is not necessarily .....

1 Consistent estimators are not necessarily .....

1 As estimator with smaller variance than that of another estimator is .....

1 The credit of factorisation theorem for sufficiency goes to .....

Very Short Answer Questions

1 Distinguish between an estimate and estimator. 1

What is a point estimate?

1 Define unbiasedness of an estimator 1

Define consistency of an estimator. 1 Define

efficiency of an estimator.

1 Define sufficiency of an estimator.

1 State the desirable properties of a good estimator.

1 Give one example of an unbiased estimator which is not consistent. 1

Give an example of a consistent estimator which is not unbiased. 1 Give the names of various methods of estimation of a parameter. 1 What is a maximum likelyhood estimator?

1 Discuss method of moments estimation. 1

What are the properties of MLE?

1 Show that sample mean is more efficient than sample median as an estimator of population mean.

1 State the necessary and sufficient condition for consistency of an estimator.

Short Essay Questions

1 Distinguish between Point estimation and Interval estimation.

1 Define the following terms and give an example for each: (a) Unbiased statistic; (b) Consistent statistic; and (c) Sufficient statistic,

1 Describe the desirable properties of a good estimator.

1 Explain the properties of a good estimator. Give an example to show that a consistent estimate need not be unbiased.

1 Define consistency of an estimator. State a set of sufficient conditions

 for the consistency of an estimate and establish it.

- In a N(μ, 1), show that the sample mean is a sufficient estimator of μ.
- Describe any one method used in estimation of population parameter.
- Explain method of moments and method of maximum likelihood.
- Explain the method of moments for estimation and comment on such estimates.
- Explain the maximum likelihood method of estimation. State some important properties of maximum likelihood estimate.
- State the properties of a maximum likelihood estimator. Find the maximum likelihood estimator for θ based on n observations for the frequency function

  $f(x, \theta) = (1 + \theta)\ x^{\theta}$; $\theta > 0$, $0 < x < \theta$
  $= 0$ elsewhere.
- Given a random sample of size n from

  $f(x ; \theta) = \theta e^{-\theta x}$, $x > 0$; $\theta > 0$.
  find the maximum likelihood estimator of θ. Obtain the variance of the estimator.

MODULE III
INTERVAL ESTIMATION

Thus far we have dealt only with point estimation. A point estimator is used to produce a single number, hopefully close to the unknown parameter. The estimators thus obtained do not, in general, coincide with true value of the parameters. We are therefore interested in finding, for any population parameter, an interval called 'confidence interval' within which the population parameter may be expected to lie with a certain degree of confidence, say α. In other words, given a random sample of n independent values x_1, x_2 ....x_n of a random variable X having the probability density f(x ; θ), θ being the parameter, we wish to find t_1 *and* t_2 the function of x_1, x_2 ....x_n such that $p(t_1 < \theta < t_2) = 1 - \alpha$.

This leads to our saying we are $100(1-\alpha)$% confident that our single interval contains the true parameter value. The interval (t_1, t_2) is called *confidence interval* or *fiducial interval* and $1 - \alpha$ is called 'confidence coefficient' of the interval (t_1, t_2). The limits t_1 and t_2 are called 'confidence limits'.

For instance if we take $\alpha = 0.05$, the 95% confidence possesses the meaning that if 100 intervals are constructed based on 100 different samples (of the same size) from the population, 95 of them will include the true value of the parameter. By accepting 95% confidence interval for the parameter the frequency of wrong estimates is approximately equal to 5%. The notion of confidence interval was introduced and developed by Prof: J. Neyman in a series of papers.

Now we discuss the construction of confidence interval of various parameters of a population or distribution under different conditions.

Confidence interval for the mean of a Normal population N(μ, σ)

Case (i) when σ is known.

To estimate μ, let us draw a random sample x_1, x_2 ....x_n of size n from the normal population.

Let $\bar{x}$ be the mean of a random sample of size n drawn from the normal population N(μ, σ).

Then $\bar{x} \to N(\mu, \sigma/\sqrt{n})$; $\therefore\ Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$

From the area property of standard normal distribution, we get

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

ie. $$P\left\{-z_{\alpha/2} \le Z \le z_{\alpha/2}\right\} = 1-\alpha$$

ie. $$P\left\{-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1-\alpha$$

ie. $$P\left\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

ie. $$P\left\{-\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

ie. $$P\left\{\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

ie. $$P\left\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Here the interval $\left\{\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right\}$ is called $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population.

Here $z_{\alpha/2}$ is obtained from the table showing the 'area under a standard normal curve' in such a way that the area under the normal curve to its right is equal to $\alpha/2$.

Note:

1. If $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, so the 95% confidence interval for μ is
$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

2. If $\alpha = 0.02$, $z_{\alpha/2} = 2.326$, so the 98% confidence interval for μ is
$$\left[\bar{x} - 2.326\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.326\frac{\sigma}{\sqrt{n}}\right]$$

3. If $\alpha = 0.01$, $z_{\alpha/2} = 2.58$, so the 99% confidence interval for μ is
$$\left[\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right]$$

4. If $\alpha = 0.10$, $z_{\alpha/2} = 1.645$, so the 90% confidence interval for μ is
$$\left[\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right]$$

Case (ii) when σ is unknown, n is large (n≥30)

When the sample is drawn from a normal population or not, by central limit theorem,

$$Z = \frac{\bar{x}-\mu}{s/\sqrt{n}} \to N(0,1) \text{ as } n \to \infty$$

Here we know that $P\{|Z| \le z_{\alpha/2}\} = 1-\alpha$

Proceeding as above, we get

$$P\left\{\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right\} = 1-\alpha$$

Thus the $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Case (iii) when σ is unknown, n is small (n<30)

Let $X_2, X_2, ...X_n$ be a random sample drawn from $N(\mu, \sigma)$ where σ is unknown. Let $\bar{x}$ be the sample mean and s^2 be its sample variance. Here we know that the statistic.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \rightarrow t_{(n-1)}\ df$$

Hence $100(1-\alpha)\%$ confidence interval for μ is constructed as follows.

Let $P\{|t| \le t_{\alpha/2}\} = 1-\alpha$

$$\Rightarrow \quad P\{-t_{\alpha/2} \le t \le t_{\alpha/2}\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{-t_{\alpha/2} \ge \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \le t_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{-t_{\alpha/2}\frac{s}{\sqrt{n-1}} \le \bar{x} - \mu \le t_{\alpha/2}\frac{s}{\sqrt{n-1}}\right\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{-\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n-1}} \le -\mu \le \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n-1}}\right\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n-1}} \ge \mu \ge \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n-1}}\right\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n-1}} \le \mu \le \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n-1}}\right\} = 1-\alpha$$

Thus the $100(1-\alpha)\%$ confidence interval for μ is

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

where $t_{\alpha/2}$ is obtained by referring the Student's t table for (n–1) d.f. and probability α.

Interval Estimate of the Difference of two population means

Case (i) When σ_1, σ_2 known

Let $\bar{x}_1$ be the mean of a sample of size n_1 taken from a population with mean μ_1 and *SD* σ_1.

Then $\bar{x}_1 \rightarrow N(\mu_1, \sigma_1/\sqrt{n_1})$

Let $\bar{x}_2$ be the mean of a sample of size n_2 taken from a population with mean μ_2 and *SD* σ_2.

Then $\bar{x}_2 \rightarrow N(\mu_2, \sigma_2/\sqrt{n_2})$
Then by additive property,

$$\bar{x}_1 - \bar{x}_2 \rightarrow N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\therefore Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

By the area property of ND, we know that

$$P\left[|Z| \le Z_{\alpha/2}\right] = 1-\alpha$$

$$\therefore\ P\left[-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right] = 1-\alpha$$

$$\text{i.e. } P\left[-Z_{\alpha/2} \le \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z_{\alpha/2}\right] = 1-\alpha$$

On simplification as in the case of one sample, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left((\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\mu_1 - \mu_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

where the value of $Z_{\alpha/2}$ can be determined from the Normal table.

When $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$. So, $100(1-\alpha)\% = 95\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left((\bar{x}_1 - \bar{x}_2) - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\mu_1 - \mu_2) + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

When $\alpha = 0.01$, $Z_{\alpha/2} = 2.58$. So, the 99% eq for $\mu_1 - \mu_2$ is

$$\left((\bar{x}_1 - \bar{x}_2) - 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\mu_1 - \mu_2) + 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Case (ii) When σ_1, σ_2 unknown, n_1, n_2 large

In this case we replace σ_1 and σ_2 respectively by their estimates s_1 and s_2.

So 95% CI for $\mu_1 - \mu_2$ is

$$\left((\bar{x}_1 - \bar{x}_2) - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\mu_1 - \mu_2) + 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

Similarly we can find 98% and 99% confidence intervals replacing 1.96 respectively by 2.326 and 2.58.

Case (iii) When $\sigma_1 = \sigma_2 = \sigma$ unknown, n_1, n_2 small

Here $t = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] / \sqrt{\sigma^2/n_1 + \sigma^2/n_2} \rightarrow$ students 't' distribution with $\upsilon = (n_1 + n_2 - 2)$ d.f.

Where $\sigma^2 = \left(n_1 s_1^2 + n_2 s_1^2\right) / (n_1 + n_2 - 2)$

Refer the 't' curve for $\upsilon = (n_1 + n_2 - 2)$ d.f. and probability level P = α

The table value of t is $t_{\alpha/2}$

Then we have $P\{|t| > t_{\alpha/2}\} = \alpha$

ie. $P\{|t| \le t_{\alpha/2}\} = 1-\alpha$

ie. $P = \{-t_{\alpha/2} \le t \le +t_{\alpha/2}\} = 1-\alpha$.

Substituting t and simplifying we get the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ as

$$\left\{(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{(\sigma^2/n_1) + (\sigma^2/n_2)},\ (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\sigma^2/n_1 + \sigma^2/n_2}\right\}$$

where $t_{\alpha/2}$ is obtained by referring the t table for $n_1 + n_2 - 2$ df and probability α.

School of Distance Education

Confidence interval for the variance of a Normal population

N (μ,σ

Let s^2 be the variance of a sample of size n(n<30) drawn from $N(\mu, \sigma)$. We know that the statistic

$$\chi^2 = \frac{ns_2}{\sigma^2} \rightarrow \chi^2 (n-1)\ d.f.$$

Now by referring the χ^2 table we can find a $\chi 1^2_{-\alpha/2}$ and $\chi\alpha^2_{/2}$ such that

$$P\left\{ \chi^2_{1-\alpha/2} \leq \chi 2 \leq \chi^2_{\alpha/2} \right\} = 1-\alpha$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are obtained by referring the table for n–1 d.f. and probabilities $1-\alpha/2$ and $\alpha/2$ respectively**.**

ie. $$P\left\{ \chi^2_{1-\alpha/2} \leq \frac{ns2}{\sigma 2} \leq \chi 2_{\alpha/2} \right\} = 1-\alpha$$

ie. $$P\left\{ \frac{1}{\chi 2_{1-\alpha/2}} \geq \frac{\sigma^2}{ns^2} \geq \frac{1}{\chi 2_{\alpha/2}} \right\} = 1-\alpha$$

ie. $$P\left\{ \frac{ns^2}{\chi^2_{1-\alpha/2}} \geq \sigma^2 \geq \frac{ns^2}{\chi^2_{\alpha/2}} \right\} = 1-\alpha$$

ie. $$P\left\{ \frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{1-\alpha/2}} \right\} = 1-\alpha$$

Thus the $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left[\frac{ns^2}{\chi 2_{\alpha/2}}, \frac{ns^2}{\chi 2_{1-\alpha/2}} \right]$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are obtained by referring the χ^2 table for n–1 d.f. and probabilities $1-\alpha/2$ and $\alpha/2$ respectively**.**

Confidence interval for the proportion of success of a binomial population

$\left(\frac{x}{n}\right)$ be the proportion of success of a sample of size n drawn

from a binomial population with parameters n and p where p is unknown and n is assumed to be known. Then we know that

$$Z = \frac{p'-p}{\sqrt{\frac{pq}{n}}} \rightarrow N(0,1) \quad \text{for large n}$$

From normal tables we get,

$$P\left\{ |Z| \leq z_{\alpha/2} \right\} = 1-\alpha$$

ie. $$P\left\{ -z_{\alpha/2} \leq Z \leq z_{\alpha/2} \right\} = 1-\alpha$$

ie. $$P\left\{ -z_{\alpha/2} \leq \frac{p'-p}{\sqrt{\frac{pq}{n}}} \leq z_{\alpha/2} \right\} = 1-\alpha$$

As in the previous cases, on simplification we get

ie. $$P\left\{ p' - z_{\alpha/2}\sqrt{\frac{pq}{n}} \leq p \leq p' + z_{\alpha/2}\sqrt{n} \right\} = 1-\alpha$$

So, the $100(1-\alpha)\%$ confidence interval for p is

ie. $$\left\{ p' - z_{\alpha/2}\sqrt{\frac{pq}{n}}, p' + z_{\alpha/2}\sqrt{n} \right\}$$

But, since p is unknown, we can replace p and q by their unbiased estimators p′ and q′. Thus the $100(1-\alpha)\%$ confidence interval for p is

$$\left[p' - z_{\alpha/2}\sqrt{\frac{p'q'}{n}},\ p' + z_{\alpha/2}\sqrt{\frac{p'q'}{n}}\right]$$

where $z_{\alpha/2}$ can be determined from the normal tables for a given α.

Note

When $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, so the 95% C.I. for p is

$$\left[p' - 1.96\sqrt{\frac{p'q'}{n}},\ p' + 1.96\sqrt{\frac{p'q'}{n}}\right]$$

when $\alpha = 0.02$, $z_{\alpha/2} = 2.326$, so the 98% C.I. for p is

$$\left[p' - 2.326\sqrt{\frac{p'q'}{n}},\ p' + 2.326\sqrt{\frac{p'q'}{n}}\right]$$

when $\alpha = 0.01$, $z_{\alpha/2} = 2.58$, so the 99% C.I. for p is

$$\left[p' - 2.58\sqrt{\frac{p'q'}{n}},\ p' + 2.58\sqrt{\frac{p'q'}{n}}\right]$$

Interval Estimate of the difference of proportions of two binomial populations:

From the study of sampling distribution it is known that the difference of proportions obtained from two samples

ie. $p_1' - p_2' \to N\left(p_1 - p_2,\ \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$

$$\therefore\quad z = \frac{(p_1' - p_2') - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \to N(0,1)$$

From normal tables we have

$P(|z| \le +1.96) = 0.95$

ie. $P(-1.96 \le z \le +1.96) = 0.95$

From this result we can write the 95% confidence interval for $p_1 - p_2$ as

$$(p_1' - p_2') \pm 1.96\sqrt{\frac{p_1}{n}q_1 + \frac{p_2}{n}q_2},$$

$$(p_1' - p_2') - 1.96\sqrt{\frac{p_1}{n}q_1 + \frac{p_2}{n}q_2}$$

Since p_1, q_1 and p_2, q_2 are unknown, they are estimated as $p_1 = p_1'$, $q_1 = q_1'$, $p_2 = p_2'$ and $q_2 = q_2'$.

The 95% confidence interval for $(p_1 - p_2)$ is

$$\left\{(p_1' - p_2') - 1.96\sqrt{\frac{p_1'q_1'}{n_1} + \frac{p_2'q_2'}{n_2}},\ (p_1' - p_2') + 1.96\sqrt{\frac{p_1'q_1'}{n_1} + \frac{p_2'q_2'}{n_2}}\right\}$$

Note: To construct 98% and 99% confidence intervals for $p_1 - p_2$ we have to replace 1.96 by 2.326 and 2.58 respectively.

SOLVED PROBLEMS

Example 1

Obtain the 95% confidence interval for the mean (when σ known) of a normal population $N(\mu, \sigma)$.

Solution

Let $x_1, x_2, x_2 \ldots, x_n$ be a random sample of size n drawn from $N(\mu, \sigma)$. Let $\bar{x}$ be its mean and σ^2 is its variance. Then we know that

$\bar{x} \to N(\mu, \sigma/\sqrt{n})$

$$\therefore Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

From normal tables, we will get

$$P\{|Z| \le 1.96\} = 0.95$$

ie. $$P\{-1.96 \le Z \le 1.96\} = 0.95$$

ie. $$P\left\{-1.96 \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le 1.96\right\} = 0.95$$

ie. $$P\left\{-1.96 \cdot \frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

ie. $$P\left\{-\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

ie. $$P\left\{\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

ie. $$P\left\{\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

Thus the 95% CI for μ is $\left[\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}\right]$

Example 2

Obtain the 95% confidence interval for the variance of a normal population $N(\mu, \sigma)$

Solution

To obtain the CI for σ^2 let us use the result

$$\chi^2 = \frac{ns^2}{\sigma^2} \to \chi^2_{(n-1)}$$

From χ^2 table, we can find a $\chi^2_{0.975}$ and $\chi^2_{0.025}$ such that

$$P\left\{\chi^2_{0.975} \le \chi^2 \le \chi^2_{0.025}\right\} = 0.95$$

ie $$P\left\{\chi^2_{0.975} \le \frac{ns^2}{\sigma^2} \le \chi^2_{0.025}\right\} = 0.95$$

ie. $$P\left\{\frac{ns^2}{\chi^2_{0.975}} \ge \sigma^2 \ge \frac{ns^2}{\chi^2_{0.025}}\right\} = 0.95$$

ie. $$P\left\{\frac{ns^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{0.975}}\right\} = 0.95$$

Thus the 95% CI for σ^2 is $\left[\frac{ns^2}{\chi^2_{0.025}}, \frac{ns^2}{\chi^2_{0.975}}\right]$ where $\chi^2_{0.975}$ and $\chi^2_{0.025}$ are obtained by referring the χ^2 table for $n-1$ df and probabilities 0.975 and 0.025 respectively.

Example 3

Obtain the 99% CI for the difference of means of two normal populations $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ when (i) σ_1, σ_2 known (ii) σ_1, σ_2 unknown.

Solution

Case (i) when σ_1, σ_2 known.

Let X_1 and X_2 be two independently normally distributed random variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively.

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of n_1 and n_2 observations.

since $X_1 \to N(\mu_1, \sigma_1^2)$, $X_2 \to N(\mu_2, \sigma_2^2)$

then $\bar{x}_1 \to N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$, $\bar{x}_2 \to N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

therefore $\bar{x}_1 - \bar{x}_2 \to N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

$$\therefore Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1)$$

From normal tables we can write

$P(-2.58 \le Z \le 2.58) = 0.99$

Substituting Z and simplifying, we get

$$P\left\{(\bar{x}_1 - \bar{x}_2) - 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 0.99$$

Thus the 99% CI for $\mu_1 - \mu_2$ is $\left\{(\bar{x}_1 - \bar{x}_2) \pm 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}$

Example 4

Obtain the 99% confidence interval for the difference of means of two normal populations $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ when σ_1 *and* σ_2 unknown, by drawing small samples.

Solution

Since σ_1^2 and σ_2^2 are unknown, they are estimated from samples, as $\sigma_1^2 = \sigma_2^2 = \sigma^{*2}$, where

$$\sigma^{*2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

Then the statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \to t(n_1 + n_2 - 2) df$$

Thus we have

$$P\left\{-t_{\alpha/2} \le \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \le t_{\alpha/2}\right\} = 1 - \alpha$$

where $t_{\alpha/2}$ is obtained by referring the t table for $n_1 + n_2 - 2$ *df* and probability $\alpha = 0.01$.

This gives $P\left\{(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \times \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \le\right.$

School of Distance Education

$$\left\{(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \times \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right\} = 1 - \alpha$$

Thus the 99% CI for $\mu_1 - \mu_2$ is

$$\left\{(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right\}$$

Example 5

If the mean age at death of 64 men engaged in an occupation is 52.4 years with standard deviation of 10.2 years. What are the 98% confidence limits for the mean age of all men in that occupation?

Solution

Here n = 64, $\bar{x} = 52.4$, $s = 10.2$

98% CI for the population mean is $\left\{\bar{x} \pm 2.326 \frac{s}{\sqrt{n}}\right\}$

ie. $\left\{52.4 \pm 2.326 \times \frac{10.2}{\sqrt{64}}\right\} = \{49.435, 55.365\}$

Example 6

A random sample of size 15 from a normal population gives $\bar{x} = 3.2$ and $s^2 = 4.24$. Determine the 90% confidence limits for σ^2.

Solution

The 90% CI for σ^2 is $\left\{\frac{ns^2}{\chi^2_{0.05}}, \frac{ns^2}{\chi^2_{0.95}}\right\}$

n = 15, $s^2 = 4.24$

From table, $\chi^2_{14,0.05} = 23.68$

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

Thus 98% CI for σ^2 is $\left\{\frac{15 \times 4.24}{23.68}, \frac{15 \times 4.24}{6.571}\right\}$

= **(2.685, 9.678)**

Example 7

A medical study showed 57 of 300 persons failed to recover from a particular disease. Find 95% confidence interval for the mortality rate of the disease.

Solution

n = 300, x = 57

$p' = \frac{x}{n} = \frac{57}{300} = 0.19$

$\therefore q' = 1 - p' = 1 - 0.19 = 0.81$

The 95% CI for the mortality rate is

$$\left\{p' - 1.96\sqrt{\frac{p'q'}{n}}, p' + 1.96\sqrt{\frac{p'q'}{n}}\right\}$$

ie. $\left\{0.19 - 1.96\sqrt{\frac{0.19 \times 0.81}{300}}, 0.19 + 1.96\sqrt{\frac{0.19 \times 0.81}{300}}\right\}$

ie. **{0.146, 0.234}**

Example 8

A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Obtain the 95% and 99% confidence interval for the population mean.

Solution

i. Here n = 16, $\bar{x} = 41.5$, $\sum(x - \bar{x})^2 = ns^2 = 135$

The 95% CI for μ is $\left\{ \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n-1}} \right\}$

ie. from table, $t_{15,0.05} = 2.131$

So the required confidence interval is $\left\{ 41.5 \pm 2.131 \times \frac{3}{4} \right\}$

ie. **{39.902, 43.098}**

ii. For 99% CI, $t_{15,0.01} = 2.947$

$\therefore$ 99% CI is $\left\{ 41.5 \pm 2.947 \times \frac{3}{4} \right\}$ = **{39.29, 43.71}**

Example 9

A certain psychological test was given to two groups of Army prisoners (a) first offenders and (b) recidivists. The sample statistics were as follows.

Population	Sample size	Sample mean	Sample S.D.
a) first offenders	580	34.45	8.83
b) recidivists	786	28.02	8.81

Construct 95% confidence limits of the difference of the means $(\mu_1 - \mu_2)$ of the two populations.

Solution

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$\left\{ \bar{x}_1 - \bar{x}_2 - 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

Since σ_1 and σ_2 are unknown, we shall replace them respectively by s_1 and s_2.

The lower limit $= (\bar{x}_1 - \bar{x}_2) - 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$= (34.45 - 28.02) - 1.96 \sqrt{\frac{(8.83)^2}{580} + \frac{(8.81)^2}{786}}$

$= 6.43 - .95 = 5.48.$

The upper limit $= (\bar{x}_1 - \bar{x}_2) + 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$= (34.45 - 28.02) + 1.96 \sqrt{\frac{(8.83)^2}{580} + \frac{(8.81)^2}{786}}$

$= 6.43 + .95 = 7.38.$

$\therefore$ The 95% confidence interval for $(\mu_1 - \mu_2)$ is (5.48, 7.38)

EXERCISES

Multiple Choice Questions

1 The notion of confidence interval was introduced and developed by

a. R.A. Fisher b. J. Neymann

c . Karl Pearson d. Gauss

1 The 95% confidence interval for mean of a normal population $N(\mu, \sigma)$ is

a. $\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$ b. $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

c . $\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$ d. $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$

1 The $100(1 - \alpha)\%$ confidence interval for μ of $N(\mu, \sigma)$ when σ unknown, using a sample of size less than 30 is

a. $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n-1}}$ b. $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

c. $\bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n-1}}$ d. $\bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n}}$

1 A random sample of 16 housewives has an average body weight of 52kg and a standard deviation of 3.6kg. 99% confidence limits for body weight in general are

a. (54.66, 49.345) b. (52.66, 51.34)

c . 55.28, 48.72) d. none of the above

1 Formula for the confidence interval for the ratio of variances of the two normal population involves

a. χ^2 distribution b. F distribution

c . t distribution d. none of the above

Fill in the blanks

1 The notion of confidence interval was introduced and developed by _____

1 The confidence interval is also called _____ interval

1 An interval estimate is determined in terms of _____

1 An interval estimate with _____ interval is best

1 Confidence interval is specified by the _____ limits

1 Confidence interval is always specified with a certain _____

1 To determine the confidence interval for the variance of a normal distribution _____ distribution is used

Very Short Answer Questions

1 What is an interval estimate?

1 Explain interval estimation

1 State the 95% confidence interval for the mean of a normal distribution $N(\mu, \sigma)$ when σ is known

1 Give the 95% CI for the variance of a normal population

1 Give the formula for obtaining confidence limits for the difference between the mean of two normal populations

1 Why interval estimate is preferred to point estimate for estimating an unknown parameter.

1 What do you mean by confidence level?

Short Essay Questions

1 Distinguish between point estimation and interval estimation. Explain how you will construct $100(1-\alpha)$% confidence interval for normal population mean when population S.D. is (i) known and (ii) unknown.

1 Explain how you would find interval estimates for the mean and variance of a normal population.

1 What do you mean by interval estimation? Obtain 99% confidence limits for θ of the normal distribution N(θ, σ^2), with the help of a random sample of size n.

1 Explain the idea of interval estimation. Obtain a $100(1-\alpha)$% confidence interval for the mean of a normal distribution whose variance is also unknown.

1 Obtain 95% confidence interval for the mean of a normal population with unknown variance on the basis of a small sample of size n taken from the population. What happens when n is large?

Long Essay Questions

1 A random sample of 20 bullets produced by a machine shows an average diameter of 3.5 mm and a s.d. of 0.2 mm. Assuming that the diameter measurement follows $N(\mu, \sigma)$ obtain a 95% interval estimate for the mean and a 99% interval estimate for the true variance.

1 The mean and s.d. of a sample of size 60 are found to be 145 and 40. Construct 95% confidence interval for the population mean.

1 Two independent random samples each of size 10 from two

independent normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ yield $\bar{x}_1 = 4.8$, $s_1^2 = 8.6$, $\bar{x}_2 = 5.6$ and $s_2^2 = 7.9$. Find 95% confidence interval for $\mu_1 - \mu_2$.

1 Two random samples of sizes 10 and 12 from normal populations having the same variance gave $\bar{x}_1 = 20$, $s_1^2 = 25$ and $\bar{x}_2 = 24$, $s_2^2 = 36$. Find 90% confidence limits for $(\mu_1 - \mu_2)$.

1. In a sample of 532 individuals selected at random from a population, 89 have been found to have Rh-ve blood. Find an interval estimate of the proportion of individuals in the population with Rh-ve blood with 95% confidence.

1 Of 250 insects treated with a certain insecticides, 180 were killed. Set approximate 95% confidence interval to the value of p, the proportion of insects likely to be killed by this insecticides in future use.

1 Suppose a sample of 500 people were interviewed and 200 of them stated they were in favour of a certain candidate as president. Obtain the 98% confidence limits for the population proportion in favour of the said candidate.

1 150 heads and 250 tails resulted from 400 tosses of a coin. Find 90% confidence interval for the prob: of a head.

1 A random sample of 500 apples was taken from a large consignment and of these 65 were bad. Estimate the proportion of bad apples by a 90% confidence interval.

1 A sample poll of 100 voters in a given district indicated that 55% of them were in favour of a particular candidate. Find 95% and 99% confidence limits for the proportion.

MODULE 4
TESTING OF HYPOTHESIS

Most tests of statistical hypothesis concern the parameters of distributions, but sometimes they also concern the type, or nature of the distributions, themselves. If a statistical hypothesis completely specifies the distribution, it is referred to as a *simple hypothesis* if not, it is referred to as a *composite hypothesis*.

The statistical hypothesis that X follows normal with mean 15 is a composite hypothesis since it does not specify the standard deviation of the normal population. The statement that X follows a poison distribution with parameter $\lambda = 2$ is a simple hypothesis since it specifies the population completely.

A statistical hypothesis which refers only to the numerical values of unknown parameters of a random variable is called a parametric hypothesis. Eg. In a normal population if we test that whether $\mu = 10$ or not is a parametric hypothesis. A hypothesis which refers to the form of an unknown distribution is called a non parametric hypothesis. eg. The form of the density function in a population is normal.

Definition of terms

The following are definitions of some terms which are frequently used in this module.

Test of Hypothesis

Rules or procedures which enable us to decide whether to accept or reject the hypothesis or to determine whether observed samples differ significantly from expected results are called *tests of hypothesis or tests of significance*.

In our subsequent discussions we are concerned with hypothesis about only one parameter.

Null Hypothesis

The hypothesis to be tested is usually referred to as the 'Null hypothesis' and is denoted by the symbol H_0. Thus a hypothesis which is set up with the possibility of its being rejected at some defined probability value is called a *null hypothesis*. For instance, if we want to show that students of College A have a higher average IQ than the students of College B, then we might formulate the hypothesis that there is no difference viz, $H_0 : \mu_A = \mu_B$

Alternative Hypothesis

In the testing process H_0 is either 'rejected' or 'not rejected'. If H_0 is not rejected, it means that the data on which the test is based do not provide sufficient evidence to cause rejection. But if H_0 is rejected it means that the data on hand are not compatible with some other hypothesis. This other hypothesis is known as '*alternative hypothesis*', denoted by H_1. The rejection or 'non rejection' of H_0 is meaningful when it is being tested against a rival hypothesis H_1 .

Type I and Type II errors

Research requires testing of hypothesis. In this process two types of wrong inferences can be drawn. These are called type I and type II errors.

Rejecting a null hypothesis H_0 when it is actually true is called *type I error* or *error of the first kind*.

Accepting a null hypothesis H_0 when it is false is called *type II error* or *error of the second kind*.

These can be schematically shown as below.

Action	H_0 true	H_0 false
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

Any test of H_0 will tell us either to accept H_0 or reject H_0, based on the observed sample values. Thus is not possible to commit both errors simultaneously.

We will define

α = P(Type I error)
= P(rejecting H_0 given H_0 is true)
= P(rejecting $H_0 \mid H_0$)

β = P(Type II error)
= P(accepting H_0 given H_1 is true)
= P(Accepting $H_0 \mid H_1$)

Every test of H_0 has values for the pair (α, β) associated with it. It would seem ideal if we could find the test that simultaneously minimises both α and β, but this is not possible. Since each of α and β is a probability we know that $\alpha \geq 0$ and $\beta \geq 0$; that is 0 is the minimum value for each. No matter whatH_0 and H_1 state and what observed values occur in the sample, we could use the test: Accept H_0. With this test we would never commit a type I error, since we would not reject H_0 no matter what the sample values were. Thus for this test $\alpha = 0$ implies $\beta = 1$. The converse of this test, which would always reject H_0 given $\beta = 0$, $\alpha = 1$. Neither of this test is desirable, because they maximise one of the two probabilities of error while minimising the other. Now our objective is to choose the decision rule that will lead to probabilities of these errors being as small as possible.

Test statistic

The testing of a statistical hypothesis is the application of an explicit set of rules for deciding whether to accept the null hypothesis or to reject it in favour of the alternative hypothesis, consistent with the results obtained from the random sample taken from the population. As the sample itself is set of observations, usually an appropriate function of the sample observation is chosen and the decision either to accept a reject the hypothesis is taken based on the value of this function. This function is called '*test statistic*' or '*test criterion*', in order to distinguish it from an ordinary descriptive statistic or estimator such as $\bar{x}$ or s^2.

We can note that a test statistic is a random variable, being a measurable function of random observations which are themselves random variables. The test procedure, therefore, partitions the possible values of the test statistic into two subsets: an 'acceptance region for H_0 and a rejection region for H_0.

Critical Region

The basis of testing the hypothesis is the partition of the sample space into two exclusive regions, namely, the region of acceptance and region of rejection. If the sample point falls in the region of rejection, H_0 is rejected. The region of rejection is called '*critical region*'. Thus critical region is

 the set of those values of the test statistic which leads to the rejection of the null hypothesis. Critical region is denoted by ω. Acceptance regions is the set of those values of the test statistic for which we are accepting the null hypothesis. Every test is identified with a critical region *w* and we are facing embarrassing richness of potential tests. Here we want to find best critical region (BCR) *w*, guided only by the principle of minimising the probabilities of errors of type I and II.

Level of significance

The validity of H_0 against that of H_1 can be tested at a certain 'level of significance. The *level of significance* is defined as the probability of rejecting the null hypothesis H_0 when it is true or probability of type I error. Actually this is the probability of the test statistic falling in the critical region when the null hypothesis is true. So significance level is also called '*size of the critical region*', '*size of the test*' or *producer's risk*. It is denoted by α. α is usually expressed as a percentage such as 1%, 2%, 5% and 10%.

ie., $\alpha = P(\text{Rejecting } H_0 \mid H_0) = P(x \in w \mid H_0)$

For instance, if the hypothesis is accepted at 5% level, the statistician in the long run, will be making wrong decisions in 5 out of 100 cases. If the hypothesis is rejected at the same level, he runs the risk of rejecting a true hypothesis about 5% of the time.

The best test for a Simple Hypothesis

Often the test statistic is to be determined by controlling α and β. The ideal thing is to minimise α and β simultaneously but in practice when α is minimised, β becomes large and vice versa. Hence the attempt is to minimise β for a fixed α and if there exists such a test statistic, it is called the 'best test'.

Power of a test

The probability of rejecting the null hypothesis H_0 when it is actually not true is called *power of a test* and it is given by 1 - β. Power of a test is also called power of the critical region.

ie., Power = $P(\text{Rejecting } H_0 \mid H_1 \text{ is true})$

$= 1 - P(\text{accepting } H_0 \mid H_1)$

$= 1 - P(\text{accepting } H_0 \mid H_0 \text{ is not true})$

$= 1 - P(\text{Type II error}) = 1 - \beta$

The larger the value of $1 - \beta$ for fixed α, the better is the test in general. We define sensitiveness of a test as its ability to ascertain the correctness of the alternative hypothesis when it is true for fixed α. Thus, power is a measure of the sensitiveness of the test. Therefore if other things are identical the comparison of two tests is the comparison of their respective powers.

Critical value

The value of test statistic which separates the critical region and acceptance region is called 'critical value'. The critical value is usually referred to as Z_α or t_α depending on the sampling distribution of the test statistic and level of significance used.

We now summarise the steps involved in testing a statistical hypothesis. **Step 1.** State the null hypothesis H_0 and the alternative hypothesis H_1. **Step 2.** Choose the level of significance α

Step 3. Determine the test statistic

Step 4. Determine the probability distribution of the test statistic

Step 5. Determine the Best Critical Region

Step 6. Calculate the value of the test statistic.

Step 7.Decision: If the calculated value of the test statistic falls in the critical region, reject the null hypothesis H_0, otherwise accept it. ie., if the calculated value exceeds the table value, reject H_0, otherwise accept H_0.

Neymann Pearson Theory of testing Hypothesis

The conceptual structure of the theory is as follows. To test the simple $H_0 : \theta = \theta\theta$ versus the simple $H_1 : \theta = \theta_1$, based on a random sample of size n, the solution is given by the celebrated Neymann-Pearson Lemma. This specifies the explicit form of the test (ie., critical region) which has pre assigned probability of error of type I and a minimal probability of error of type II. This is same as maximising power subject to the condition that the type I error is a constant. This process is equivalent to choosing a critical region of size α which has at least the same power as any other critical region of the same size. Such a critical region is called the *best critical region*, abbreviated as BCR. The test based on BCR is called Most Powerful Test.

SOLVED PROBLEMS

Example l

Let X follows B(10, p). Consider the following test for testing H_0 : p = 1/2 against H_1 : p = 1/4: "Reject H_0 if X≤2". Find the significance level and power of the test.

Solution

Given $X \to B(10, p)$

Significance level = P(Rejecting H_0/H_0) = $P(X \le 2 \mid p = 1/2)$

$= P(X = 0) + P(X = 1) + P(X = 2)$ when p = 1/2

$$= {}^{10}C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10} + {}^{10}C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9 + {}^{10}C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^8$$

$$= \left(\frac{1}{2}\right)^{10}[1 + 10 + 45] = \frac{56}{1024}$$

Power of the test = P(Accepting H_0/H_0)

$= P(\text{Rejecting } H_0/H_1)$

$= P(X \le 2 \mid p = 1/4)$

$= P(X = 0) + P(X = 1) + P(X = 2)$ when p = 1/4

$$= {}^{10}C_0\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{10} + {}^{10}C_1\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^9 + {}^{10}C_2\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^8$$

$$= (3/4)^8\left[\frac{9}{16} + 10 \times \frac{3}{16} + 45 \times \frac{1}{16}\right]$$

$$= (3/4)^8\left[\frac{84}{16}\right] = \mathbf{5.25 \times (3/4)^8}$$

Example 2

If $X \ge 1$ is the critical region for testing H_0 : $\theta = 2$ against

H_1 : $\theta = 1$ on the basis of a single observation from

$f(x, \theta) = \theta e^{-\theta x}$, $x \ge 0$, obtain the probabilities of type I and type II errors.

Solution

Given $f(x, \theta) = \theta e^{-\theta x}$, $x \ge 0$

P(Type I error) = P(Rejecting $H_0|H_0$)

$= P(X \ge 1 \mid \theta = 2)$

$$= \int_1^\infty f(x)dx \quad \text{when } \theta = 2$$

$$= \int_1^\infty 2.e^{-2x}dx = 2\left[\frac{e^{-2x}}{-2}\right]_1$$

$$= -(0 - e^{-2}) = \mathbf{e^{-2}}$$

P(Type I error) = P(Accepting $H_0|H_1$)

$= P(X \le 1 \mid \theta = 1)$

$$= \int_0^1 \theta e^{-\theta x}dx \quad \textit{when } \theta = 1$$

$$= \int_0^1 e^{-x}dx \qquad \left[\frac{e^{-x}}{-1}\right]_0^1$$

$$= -\left(e^{-1} - 1\right) = 1 - e^{-1}$$

Example 3

Given a binomial distribution

$$f(x, p) = \begin{cases} {}^{n}C_x\, p^x q^{n-x}, & x = 0,1,2,3,4 \\ 0, & \text{otherwise} \end{cases},$$

It is desired to test $H_0 : p = 1/3$ against $H_1 : p = 1/2$ by agreeing to accept H_0 if $x \le 2$ in four trials and to reject otherwise. What are the probabilities of committing.

(a) type I error, (b) type II error

Solution

(a) α = P(type I error)

= P(reject H_0 /H_0 is true)

= P(X > 2/p = 1/3)

$$= \sum_{x=3}^{4} {}^4C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{4-x}$$

$$= {}^4C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{4-3} + {}^4C_4 \left(\frac{1}{3}\right)^4$$

$$= 4 \times \frac{2}{3^4} + \frac{1}{3^4} = \frac{1}{3^2} = \mathbf{\frac{1}{9}}$$

(b) β = P(type II error)

= P(accept H_0 /when H_1 is true) or

= P(X ≤ 2/p = 1/2)

$$= \sum_{x=0}^{2} {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}$$

$$= {}^4C_0 \left(\frac{1}{2}\right)^4 + {}^4C_1 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$$

$$= \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^4 + 6\left(\frac{1}{2}\right)^4 = \mathbf{\frac{11}{2^4}}$$

Example 4

The hypothesis $H_0 : \theta = 2$ is accepted against $H_1 : \theta = 5$ if $X \le 3$ when X has an exponential distribution with mean θ. Find type I and type II error probabilities of the test.

Solution

Given X has exponential distribution with mean θ, that is X follows an exponential distribution with parameter θ.

ie., $f(x) = \frac{1}{\theta} e^{-x/\theta}, x \ge 0, \theta > 0$

P(Type I error) = P(Rejecting $H_0 \mid H_0$)

= P(X > 3|θ = 2)

$$= \int_3^\infty f(x)\,dx \text{ when } \theta = 2$$

$$= \int_3^\infty \frac{1}{\theta} e^{-x/\theta} \text{ when } \theta = 2$$

$$= \int_3^\infty \frac{1}{2} e^{-x/2} dx$$

$$= \frac{1}{2} \left[\frac{e^{-x/2}}{-1/2}\right]_3^\infty$$

$$= -[0 - e^{-3/2}] = \mathbf{e^{-3/2}}$$

P(Type II error) = P(Accept H_0 /H_1)

= P(X ≤ 3 | θ = 5)

$$= \int_0^3 \frac{1}{5} e^{-x/5}\,dx = \frac{1}{5}\left[\frac{e^{-x/5}}{-1/5}\right]_0^3$$

$$= -\left[e^{-3/5} - 1\right] = \mathbf{1 - e^{-3/5}}$$

Example 5

Find the probability of type I error of the test which rejects H_0 if $X > 1 - \alpha$ in favour of H_1 if X has pdf

$f(x) = \theta x^{\theta-1}, 0 < x < 1$ with $H_0 : \theta = 1$ and $H_1 : \theta = 2$. Find the power of the test.

Solution

Given $f(x) = \theta x^{\theta-1}, 0 < x < 1$

P(Type I error)= P(Rejecting H_0 | H_0 true)

$= P(X > 1 - \alpha \mid \theta = 1)$

$= \int_{1-\alpha}^{1} \theta x^{\theta-1} dx$ when $\theta = 1$

$= \int_{1-\alpha}^{1} dx = (x)_{1-\alpha}^{1}$

$= 1 - (1 - \alpha) = \alpha$

Power of the test = P(Accepting H_0 | H_0 true)

= P(Rejecting H_0 | H_1 true)

$= P(X > 1 - \alpha \mid \theta = 2)$

$= \int_{1-\alpha}^{1} \theta x^{\theta-1} dx$ when $\theta = 2$

$= \int_{1-\alpha}^{1} 2x\, dx$

$= \left(x^2\right)_{1-\alpha}^{1} = 1 - (1 - \alpha)^2$

$= 2\alpha - \alpha^2$

EXERCISES

Multiple Choice Questions

1 An 'hypothesis' means
a. assumption b. a testable preposition
c. theory d. supposition

1 A hypothesis may be classified as
a. simple b. composite
c. null d. all the above

1 A wrong decision about H_0 leads to
a. One kind of error b. Two kinds of error
c. Three kinds of error d. Four kinds of error

1 Area of critical region depends on
a. Size of type I error b. Size of type II error
c. Value of the statistic d. Number of observations

1 The idea of testing of hypothesis was first set forth by
a. R.A. Fisher b. J. Neymann
c. E.L. Lehman d. A. Wald

1 The hypothesis under test is a
a. simple hypothesis b. alternative hypothesis
c. null hypothesis d. none of the above.

1 Rejecting H_0 when H_0 is true is
a. Type I error b. Standard error
c. Sampling error d. Type II error

1 H_1 is accepted when the test statistic falls in the
a. critical region b. probability space
c. acceptance region d. None of the above

1 Power of a test is related to
a. Type I error b. Type II error
c. both (a) and (b) d. neither (a) nor (b)

- Level of significance is the probability of
  a. type I error b. type II error
  c. not committing error d. any of these
- Level of significance is also called
  a. size of the test b. size of the critical region
  c. producer's risk d. all the above

Fill in the blanks

- A hypothesis is a testable .....
- The parametric testing of hypothesis was originated by ..... and .....
- The hypothesis which is under test for possible rejection is called .....
- ..... error is not severe than ..... error.
- Probability of type I error is called .....
- Rejecting H_0 when H_0 is true is called .....
- Accepting H_0 when H_0 is flase is called .....

Very Short Answer Questions

- Define the term 'test of hypothesis'
- Define simple and composite hypothesis
- Define null and alternative hypothesis
- Define type I and type II errors
- Define level of significance
- Define critical region.
- Define power of a test
- Define test statistic
- State Neymann Pearson lemma
- Define a parametric test of hypothesis
- What is a statistical hypothesis.
- Define size of the test.
- Which is the first step in testing a statistical hypothesis?

Short Essay Questions

- What do you mean by a statistical hypothesis? What are the two types of errors? Outline the Neyman-Pearson approach.
- Explain the following with reference to testing of hypothesis: (i) Type I and II errors; (ii) Critical region; and (iii) Null and alternate hypothesis.
- Distinguish between Simple and Composite hypotheses. Give one example each.
- Explain the terms (i) Errors of the first and second kind; (ii) Critical region; (iii) Power of a test; and (iv) Significance level in test of hypothesis.
- Explain with illustrative examples the terms : two types or error, critical region and significance level.
- Explain the terms (1) Null hypothesis; (2) Level of significance; and (3) Critical region.
- Explain the terms (i) statistical hypothesis; (ii) critical region (iii) power of a test.

Long Essay Questions

- To test the hypothesis $H_0 : p = 1/2$ against $H_1 : p > 1/2$, where p is the probability of head turning up when a coin is tossed, the coin was tossed 8 times. It was decided to reject H_0 in case more than 6 heads turned up. Find the significance level of the test and its power if $H_1 : p = .7$.
- X_1 and X_2 are independent Bernoulli r.v.s. such that $P(X_1 = 1) = \theta = P(X_2 = 1)$. To test $\theta = 1/3$ against $\theta = 2/3$ the test suggested is to reject if $X_1 + 2X_2 > 1$. Find the power of this test.
- Consider the problem of testing the hypothesis H_0 : X has uniform distribution over (0, 3) against H_1 : X has uniform distribution over (5, 7). If (5, 7) is the critical region. find the probabilities of two kinds of errors.
- Let $X_1,... X_n$ be a r.s. from $N(\theta, 4)$. Obtain Best critical region for testing $H_0 : \theta = 10$ against $H_1 : \theta = 15$ with a sample of size n = 16 and with level of significance 0.05.

1 To test H : $\theta = 1$ against the alternative $\theta = 2$ based on X which has the p.d.f. f(x) = $\frac{1}{\theta} e^{-x/\theta}, x > 0$; = 0. otherwise, the test proposed is to reject if X > log4. Compute probabilities of committing type I and II errors if this test is used.

LARGE SAMPLE TESTS

The statistical tests may be grouped into two. (a) *Large sample tests* (b) *Small sample tests*. For small sample tests the exact sampling distribution of the test statistic will be known. In large sample tests the normal distribution plays the key role and the justification for it is found in the famous central limit theorem. That is when the sample size is large most of the statistics are normally or atleast approximately normally distributed. Let Y be a statistic satisfying the conditions of the CLT, then the statistic given by

$$Z = \frac{Y - E(Y)}{\sqrt{V(Y)}} \to N(0,1), \text{ for large n.}$$

Here $\sqrt{V(Y)}$ is called the Standard Error of Y.

$$\therefore Z = \frac{Y - E(Y)}{SE of\ Y} \to N(0,1)$$

If Z is chosen as the test statistic, the critical region for a given significance level can be determined from normal tables. The test based on normal distribution is called 'normal test' or 'Z test'.

To explain the terminology, let us consider a situation in which we want to test the null hypothesis $H_0 : \theta = \theta_0$ against the two sided alternative hypothesis $H_1 : \theta \neq \theta_0$. Since it appears reasonable to accept the null hypothesis when our point estimate $\hat{\theta}$ of θ is close to θ_0 and to reject it when $\hat{\theta}$ is much larger or much smaller than θ_0 , it would be logical to let the critical region consists of both tails of the sampling distribution of our

test statistic. Such a test is referred to as a *two-tailed test or two-sided test*.

On the other hand, if we are testing the null hypothesis $H_0 : \theta = \theta_0$ against the one sided alternative $H_1 : \theta < \theta_0$, it would seem reasonable to reject H_0 only if when $\hat{\theta}$ is much smaller than θ_0 . Therefore, in this case it would be logical to let the critical region consist only of the left hand tail of the sampling distribution of the test statistic. Likewise, in testing $H_0 : \theta = \theta_0$ against the one sided alternative $H_1 : \theta > \theta_0$ we reject H_0 only for larger values of $\hat{\theta}$ and the critical region consists only of the right tail of the sampling distribution of the test statistic. Any test where the critical region consists only one tail of the sampling distribution of the test statistic is called a one tailed test, particularly they are called *left tailed* and *right tailed test* respectively.

Best critical regions of z-test

To test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$

$$: \theta > \theta_0$$

$$: \theta \neq \theta_0$$

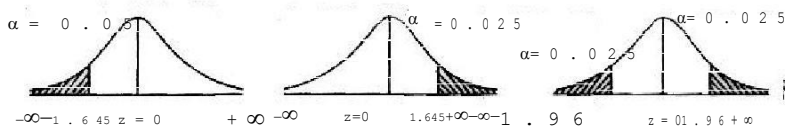
for the significance level α the best critical regions are respectively.

$\omega \equiv Z < -Z_\alpha$ $\omega \equiv Z > Z_\alpha$ and, $\omega \equiv |Z| \geq Z_{\alpha/2}$

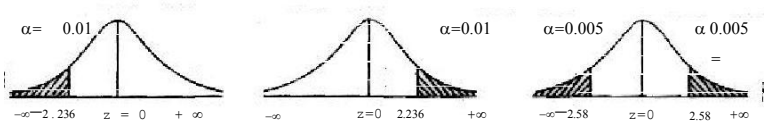
For example, when

$\alpha = 0.05$, the best critical regions are respectively

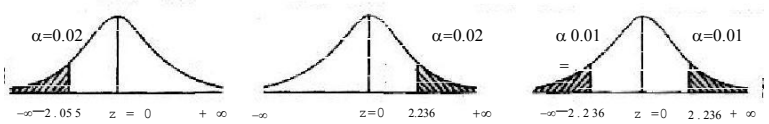
$\omega \equiv Z < -1.645$, $\omega \equiv Z > 1.645$ and $\omega \equiv |Z| \geq 1.96$



When $\alpha = 0.01$, the best critical regions are respectively



When $\alpha = 0.02$, the best critical regions are respectively



Testing mean of a Population

By testing the mean of population we are actually testing the significant difference between population mean and sample mean. In other words we are deciding whether the given sample is drawn from the population having the mean given by H_0 .

Suppose we want to test the null hypothesis $H_0 : \mu = \mu_0$ against one of the alternatives $H_1 : \mu < \mu_0$; $H_1 : \mu > \mu_0$ or $H_1 : \mu \neq \mu_0$ on the basis of a random sample of size n from a normal population with known variance σ^2 .

For the significance level α , the best critical regions are respectively, $\omega \equiv Z < -Z_\alpha$, $\omega \equiv Z > Z_\alpha$ and $\omega \equiv |Z| \geq Z_{\alpha/2}$.

The test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Using the sample data, calculate the value of Z. If it lies in the criticl region, reject H_0 otherwise accept. (Usually we will accept H_0 if the calculated value is less than the table value)

Note:

(i) For $\alpha = 0.05$, $\alpha = 0.02$ or $\alpha = 0.01$ we can fix the critical regions by determining the critical values using normal area table. (Refer best critical regions)

(ii) If the population is given to be normal, the test procedure is valid even for small samples, provided σ is known.

(iii) When σ is unknown and n is large, in the statistic we have to replace σ by its estimate s.

Example 1

A sample of 25 items were taken from a population with standard deviation 10 and the sample mean is found to be 65. Can it be regarded as a sample from a normal population with $\mu = 60$.

Solution

Given n = 25, $\sigma = 10$, $\bar{x} = 65$, $\mu_0 = 60$

We have to test $H_0 : \mu = 60$ against $H_1 : \mu \neq 60$.

Let $\alpha = 0.05$. The best critical region is $\omega \equiv |Z| \geq 1.96$. Here the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{65 - 60}{10/\sqrt{25}} = \frac{25}{10} = 2.5$$

$$\therefore \quad |Z| = 2.5 > 1.96$$

Since Z lies in the critical region, H_0 is rejected.

That is, the sample cannot be regarded as drawn from a normal population with $\mu = 60$

Example 2

A news stereo needle was introduced into the market claiming that it has an average life of 200 hours with a standard deviation of 21 hours. This claim came under severe criticism from dissatisfied customers. A customer group tested 49 needles and found that they have an average life of 191 hours. Is the claim of the manufacturer justified?

Solution

Given $\mu_0 = 200$, $\sigma = 21$, n = 49, $\bar{x} = 191$

We have to test

$H_0 : \mu = 200$ against $H_1 : \mu < 200$

Let $\alpha = 0.05$. The BCR is $\omega \equiv Z < -1.645$

The test statistics is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{191 - 200}{21/\sqrt{49}} = \frac{-63}{21} = -3$$

Since Z lies in the critical region. H_0 is rejected.

That is, the claim of the manufacturer is not justified.

Testing the Equality of two population Means

By testing the equality of two population means we are actually testing the significant difference between two sample means. In other words we are deciding whether the two samples have come from populations having the same mean.

In applied research, there are many problems in which we are interested in hypothesis concerning difference between the means of two populations.

Suppose we want to test the null hypothesis.

$H_0 : \mu_1 - \mu_2 = 0$ (or $H_0 : \mu_1 = \mu_2$) against one of the alternatives.

$H_1 : \mu_1 - \mu_2 < 0$, $H_1 : \mu_1 - \mu_2 > 0$ or $H_1 : \mu_1 - \mu_2 \neq 0$, based on independent random samples of sizes n_1 and n_2 from two populations having the means μ_1 and μ_2 and the known variances σ_1^2 and σ_2^2.

For the significance level α, the critical regions are respectively

$\omega \equiv Z < -Z_\alpha$, $\omega \equiv Z > Z_\alpha$ and $\omega \equiv |Z| \geq Z_{\alpha/2}$

The test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Calculate the value of Z using the sample information, and if it lies in the critical region reject H_0, otherwise accept it.

Note:

(i) When we deal with independent random samples from populations with unknown variances which may not even be normal we can still use the test which we have just described with s_1 substituted for σ_1 and s_2 substituted for σ_2 so long as both samples are large enough to invoke the central limit theorem.

(ii) To test $H_0 : \mu_1 - \mu_2 = \delta$ against

$H_1 : \mu_1 - \mu_2 < \delta$, $H_1 : \mu_1 - \mu_2 > \delta$, $H_1 : \mu_1 - \mu_2 \neq \delta$

the procedure is exactly the same as in the case of equality of two population means. In this case the test statistic is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Example 1

Suppose that 64 senior girls from College A and 81 senior girls from College B had mean statures of 68.2" and 67.3" respectively. If the standard deviation for statures of all senior girls is 2.43, is the difference between the two groups significant?

Solution

Given $n_1 = 64$, $n_2 = 81$, $\bar{x}_1 = 68.2$, $\bar{x}_2 = 67.3$

$\sigma_1 = \sigma_2 = \sigma = 2.43$

We have to test

$H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$

Let $\alpha = 0.05$. The BCR is $\omega \equiv |Z| \geq 1.96$

The test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

$$= \frac{68.2 - 67.3}{\sqrt{\dfrac{2.43^2}{64} + \dfrac{2.43^2}{81}}} = 2.21$$

$\therefore \quad |Z| = 2.21 > 1.96$

Since Z lies in the critical region, we reject H_0 .

That is, the two groups are significantly different with reference to their mean statures.

Example 2

A random sample of 1000 workers from factory A shows that the mean wages were Rs. 47 per week with a standard deviation of Rs. 23. A random sample of 1500 workers from factory B gives a mean wage of Rs. 49 per week with a standard deviation of Rs. 30. Is there any significant difference between their mean level of wages?

Solution

Given $n_1 = 1000$, $n_2 = 1500$, $\bar{x}_1 = 47$, $\bar{x}_2 = 49$

$s_1 = 23$ $s_2 = 30$

We have to test

$H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$

Let $\alpha = 0.02$. The BCR is $\omega \equiv |Z| \geq 2.326$

The test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

$$= \frac{47 - 49}{\sqrt{\dfrac{23^2}{1000} + \dfrac{30^2}{1500}}} = \frac{-2}{\sqrt{\dfrac{529}{1000} + \dfrac{900}{1500}}}$$

$$= \frac{-2}{\sqrt{.529 + 0.6}} = \frac{-2}{\sqrt{1.129}} = 1.882$$

$\therefore \quad |Z| = 1.882 < 2.326$

Since Z lies in the critical region, H_0 is accepted. That is, there is no significant difference between the samples.

Testing the proportion of success of a population

By testing population proportion of success we mean the testing of the significant difference between population proportion of success and the sample proportion of success.

Now let us familiarise the following notations.

p : population proportion of success (unknown) p_0

: the assumed value of p (given)

p' : $\dfrac{x}{n}$; the proportion of success of a sample

x : the number of successes

n : sample size

Suppose we want to test the null hypothesis

$H_0 : p = p_0$ against one of the alternatives

$H_1 : p < p_0$ or $H_1 : p > p_0$ or $H_1 : p \neq p_0$ based on a large sample of size n whose proportion of success is p' .

For the significance level α , critical regions are respectively.

$\omega \equiv Z < -Z_\alpha$, $\omega \equiv Z > Z_\alpha$ and $\omega \equiv |Z| \geq Z_{\alpha/2}$

The test statistic is $Z = \dfrac{p' - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$

Calculate the value of Z and if it lies in the critical region reject H_0, otherwise accept it.

Example 1

In a survey of 70 business firms it was found that 45 are planning to expand their capacities next year. Does the sample information contradict the hypothesis that 70% the firms are planning to expand next year.

Solution

Here we have $p' = \dfrac{x}{n} = \dfrac{45}{70} = 0.643$

p_0 =70% = 0.70, n = 70

Here we are testing

$H_0 : p = 0.70$ against $H_1 : p < 0.70$

Let $\alpha = 0.05$. The BCR is $\omega \equiv Z < -1.645$

The test statistic is $Z = \dfrac{p' - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$

$$= \frac{0.643 - 0.70}{\sqrt{\dfrac{0.7 \times 0.3}{70}}} = 1.04$$

Since Z lies in the acceptance region, H_0 is accepted. That is, 70% of the firms are planning to expand their capacities next year.

Testing difference of two population proportions

By testing the difference of two population proportions we are testing the equality of two population proportions or the significance difference between two sample proportions. In other words we are deciding whether the two samples have come from populations having the same proportions of success.

Let us consider the following notations.

p_1 : proportion of success of the first population

p_2 : porportion of success of the second population.

x_1 : number of successes in the first sample

x_2 : number of successes in the second sample

n_1 : first sample size

n_2 : second sample size

p_1' : porportion of success of the first sample = x_1 / n_1

p_2' : porportion of success of the second sample = x_2 / n_2

Suppose we want to test the null hypothesis

$H_0 : p_1 - p_2 = 0$ against one of the alternatives

$H_1 : p_1 - p_2 < 0$ or $H_1 : p_1 - p_2 > 0$ or $H_1 : p_1 - p_2 \neq 0$ based on two independent large samples of sizes n_1 and n_2 with proportions of success p_1' and p_2' respectively.

For the significance level α, the critical regions are respectively.

$\omega \equiv Z < -Z_\alpha$, $\omega \equiv Z > Z_\alpha$ and $\omega \equiv |Z| \geq Z_\alpha$

The test statistic is $Z = \dfrac{p_1' - p_2'}{\sqrt{p^* q^* \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

where $p^* = \dfrac{n_1 p_1' + n_2 p_2'}{n_1 + n_2}$

and $q^* = 1 - p^*$

Calculate Z and if it lies in the critical region, reject H_0, otherwise accept it.

Example 1

Before an increase in excise duty on tea 800 persons out of a sample 1000 perons were found to be tea drinkers. After an increase in duty 800 people were tea drinkers in a sample of 1200 people. Test whether there is significant decrease in the consumption of tea after the increase in duty.

Solution

We have $p_1 = \frac{800}{1000} = 0.8$ $p_2 = \frac{800}{1200} = 0.67$

Here we have to test

$H_0 : p_1 - p_2 = 0$ against $H_1 : p_1 - p_2 > 0$

Let $\alpha = 0.05$. The BCR is $\omega \equiv Z \geq 1.645$

The test statistic is $Z = \frac{p_1 - p_2}{\sqrt{p^* q^* \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$p^* = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{1600}{2200} = 0.727$

Now $p^* = \frac{n_1 + n_2}{1000 + 1200} = 0.727$

$q^* = 1 - p^* = 1 - 0.727 = 0.273$

$$Z = \frac{0.80 - 0.67}{\sqrt{0.727 \times 0.273 \left(\frac{1}{1000} + \frac{1}{1200}\right)}} = 6.816$$

$\therefore \quad |Z| = 6.816 > 1.645$

Since Z lies in the critical region, H_0 is rejected.

That is, there is a significant decrease in the consumption of tea after the increase in duty.

Example 2

In a sample of 600 men from city A, 450 are found to be smokers. Out of 900 from city B, 450 are smokers. Do the data indicate that the cities are significantly different with respect to prevalence of smoking.

Solution

Here $p_1 = \frac{450}{600} = 0.75$ $p_2 = \frac{450}{900} = 0.50$

We are testing

$H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$

Let $\alpha = 0.01$. The BCR is $\omega \equiv |Z| \geq 2.58$

The test statistic is $Z = \frac{p_1 - p_2}{\sqrt{p^* q^* \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$p^* = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{450 + 450}{600 + 900} = \frac{900}{1500} = 0.6$

Now $p^* = \frac{n_1 + n_2}{600 + 900} = 0.6$

$q^* = 1 - p^* = 1 - 0.6 = 0.4$

$$Z = \frac{0.75 - 0.50}{\sqrt{0.6 \times 0.4 \left(\frac{1}{600} + \frac{1}{900}\right)}} = 9.68$$

$\therefore \quad |Z| = 9.68 > 2.58$

Since Z lies in the critical region, H_0 is rejected.

That is, the two cities are significantly different w.r.t. prevalence of smoking.

EXERCISES

Multiple Choice Questions

1 Large sample tests are conventionally meant for a sample size

a. n = 20 b. n < 30 c. n ≥ 30 d. n = 100

1 A parametric test is performed as a large sample test using

a. central limit theorem b. Techebysheff inequality

c. Weak law of large numbers d. none of these

1 For a two failed test with $\alpha = 0.05$, the best critical region of a Z test is

a. $Z < 2.58$ b. $Z > 2.58$

c. $|Z| \geq 1.96$ d. $|Z| \geq 2.58$

1 To test $H_0 : \mu = \mu_0$ against $H_0 : \mu > \mu_0$ when σ is known, the appropriate test is

a. t-test b. Z-test

c. F-test d. none of these

1 To test $H_0 : \mu = 500$ against $H_0 : \mu < 500$, we use

a. one sided left tailed test

b. one sided right tailed test

c. two-tailed test d. all the above

1 Testing $H_0 : \mu = 200$ against $H_0 : \mu \neq 500$ leads to

a. left tailed test b. right tailed test

c. two-tailed test d. none of these

1 To test an hypothesis about proportions of success in a class, the usual test is

a. t-test b. F-test c. Z-test d. None of these

Fill in the blanks

1 A test based on the outcome of tosing of a coin is a ..... test.

1 When σ is known, the hypothesis about population mean is tested by .....

1 If the smple drawn from a population is large, then the hypothesis about μ can be tested by .....

1 A large population of heights of person is distributed with mean 66 inches and SD = 10 inches. A sample of 400 persons had the mean height = 62 inches. The data ..... the hypothesis $H_0 : \mu = 66$ inches.

1 The critical value of one sided left tailed Z test, for $\alpha = 0.05$ is .....

1 A two sided test is used for testing a null hypothesis $H_0 : \mu = \mu_0$ against .....

Very Short Answer Questions

1 Distinguish between large sample and small sample tests. 1 How will you decide the best critical regions of a Z test?

1 Give the Z statistic to test the mean of a population when σ is known.

1 State the test statistic for testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

Short Essay Questions

1 Distinguish between large sample and small sample tests illustrating with suitable examples.

1 Explain the importance of normal distribution in large sample tests.

1 Discuss the use of standard error in large sample tests. Illustrate with an example.

1 Explain the method of testing the significance of the difference between a large sample mean and population mean.

Long Essay Questions

1 An electrical firm manufactures light bulbs that have a length of life that is aporoximatey normally distributed with a mean of 800 hours and a standard deviation of 40 hours. Test $H_0 : \mu = 800$ hours, against the alternative $H_1 : \mu \neq 800$ hours if a random sample of 30 bulbs has an average life of 788 hours.

1 A random sample of 36 drinks from a soft-drink machine has an average content of 7.4 ounces with a standard deviation of 0.48 ounces. Test the hypothesis $H_0 : \mu = 7.5$ ounces against $H_1 : \mu < 7.5$ at $\alpha = 0.5$

1 A random sample of 625 items from a normal population of unknown mean has $\bar{x} = 10$ and Standard Deviation = 1.5. At a later stage it was claimed that the population mean is 9. Test the truth of this claim.

1 A random sample of 900 members is found to have a mean of 3.4 cms. Could it come from a large population with mean $\mu = 3.25$ cms. and $\sigma = 2.61$cms?

1 A sample of 200 boys who passed the S.S.L.c. Examination were found to have an average of 50 marks with S.D. = 5. The average marks of 100 girls was found to be 48 with S.D. = 4. Does it indicate any significant difference between the performance of boys and girls. $\alpha = 0.5$.

1 Test the significance of the difference between the means of the samples from the following data: $\alpha = .01$.

	Size	Mean	S.D.
Sample A :	100	61	4
Sample B :	200	63	6

1 Two samples of people consisting of 400 and 500 individuals have mean heights 171.3 and 165.3 cms. with variances 40 and 37.5 respectively. Examine whether the populations from which the samples are taken have the same mean.

1 A coin is tossed 10,000 times and it turns up head 5195 times. Is it reasonable to think that the coin is unbaised. $\alpha = .05$.

1 In 324 throws of a six faced die, odd points appeared 181 times. Would yu say that the die is not fair. $\alpha = .01$.

1 In a hospital 480 females and 520 males were born in a week. Do these figures confirm the belief that males and females are born equal numbers. $\alpha = .05$.

1 In a sample of 628 men from town A, 379 are found to be smokers. In another sample of 943 from town B, 415 are smokers. Do the data indicate that the two towns are significantly different with respect to the prevalence of smoking among men. Use two tailed test and $\alpha = .05$.

SMALL SAMPLE TESTS

In the above section we discussed some large sample tests, where central limit theorem and normal distribution plays an important role, to shape the sampling distribution of the test statistic as normal or approximately normal. But in small sample tests, when the sample size n is less than 30 the exact sampling distribution of the test statistic can be determined. Based on the sampling distribution of the test statistic, the tests can be classified as Z - test, t - test, F-test and chi-square test. We have already discussed Z-test in detail as a large sample test. Here we discuss other small sample tests.

The student's 't' test

The test of hypothesis based on the Student's 't' distribution is called t-test. The t-test is a very powerful test procedure in statistics. The t-test is used as a test of significance, in the following cases.

1. To test the significance of the mean of a small sample from a normal population.
2. To test the significance of the difference between the means of two independant samples taken from a normal population.
3. To test the significance of the difference between the means of two dependant samples taken from a normal population.
4. To test the significance of an observed correlation coefficient.
5. To test the significance of an observed regression coefficient.

We now discuss some of these tests in some detail to emphasize the importance of 't' - distribution.

1. t-test for population mean

To test the mean of a population using Student's t-test, the following assumptions must be made.

i. The parent population from which the sample is drawn is normal.
ii. The sample observations are independent and random.
iii. The sample should be small ($n < 30$)
iv. The population standard deviation σ is unknown.

By testing the mean of a normal population, we are actually testing the