IDHAYA COLLEGE FOR WOMEN, KUMBAKONAM

DEPARTMENT OF MATHEMATICS

CLASS : I B.Sc., MATHS
SUBJECT NAME : MATHEMATICAL STATISTICS - III
SUBJECT CODE : 16SACMS2
SEMESTER : II
UNIT : IV
FACULTY NAME : Mrs. RUBEELA MARY

UNIT IV

LARGE SAMPLE

Population

A statistical population is the set of all possible measurements on data corresponding to the entire collection of units for which an inference is to be made.

Sample

A sample is a part of the statistical population (i.e) it is a subset which is collected to draw an inference about the population.

Parameter and Statistic

Earlier we learnt how compute the arithmetic mean, median, mode, standard deviation etc. from the data contained in a sample. These are called some characterizations of a statistical distribution. These characteristics are called parameters if they are calculated for a population and are called statistics if they are calculated for a sample. For example the mean of a population is called a parameter and the mean of a sample is called a statistic.

Sampling Distribution

We now study how these statistics vary from sample to sample if repeated random samples of the same size are draw from a statistical population. The probability distribution of a such a statistic is called the sampling distribution. Thus we have the sampling distribution of mean, the sampling distribution of proportion, the sampling distribution of correlation coefficient and so on.These sampling distributions are the foundations of statistical inference and we consider some of these sampling distributions here.

Test of hypothesis

Hypothesis testing addresses the important question of how to choose among alternative proportions while controlling and minimizing the risk of wrong decisions. Before going into hypothesis testing, to understand it more clearly, let us consider the following non-statistical hypothesis testing procedure. Suppose an accused individual is judged in a court of law. The person before the bar is assumed to be innocent. We want to test the hypothesis that he is innocent and take this hypothesis as H_0 .Then there exists an alternative hypothesis that he guilty which we denote by H_1 .In the analysis of this case, the following possibilities arise. The first two correspond to the case " H_0 is true" and the last two corresponds to the case " H_0 is False".

- The person is innocent (H_0 is true) and the judge finds that he is innocent (accepts H_0)
- The person is innocent (H_0 is true) and the judge finds that he is innocent (reject H_0)
- The person is guilty (H_0 is false) and the judge finds that he is innocent (accepts H_0)
- The person is guilty (H_0 is false) and the judge finds that he is innocent (rejects H_0)

In case (1) and (3) the judge takes the correct decision. In case (2) and (4) the judge makes an error. In case (2) the hypothesis H_0 is rejected when it is true and this type of error is called Type I. In case (4) the hypothesis H_0 is accepted when it is false and it is called type II error. Type I error is considered to be more serious than type II error. If we take the null hypothesis as the person is guilty then type I and Type II errors would have been reversed. We give below the case in the following table:

	H_0 is true	H_0 is false
AcceptH_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Thus in statistical hypothesis test if an hypothesis (H_0) is rejected when it is true it is called type I error and if an hypothesis H_0 is accepted when it is false it is called type II error.

Generally a hypothesis which is tested for possible rejection is called null hypothesis and the hypothesis H_1 is designated as the alternative hypothesis.

Hypothesis Testing Procedure- One tail test and Two tail test

There are two basic types of decision problems that can be considered in hypothesis testing procedure.

- Whether a population parameter has changed from or differs from a particular value.
- i) Whether the sample has come (i) from the population that has a parameter value less than a hypothesis value or

ii) from the population that has a parameter value more than the hypothesized value. In these situations the attention is focused on the direction of the deviation-either on both sides or only on one side.

Suppose the average life of the company bulbs has been estimated as 2000 hours. We take a sample from the population and suppose the sample mean is 1950 hours. We want to decide whether the mean life time has changed. Here the null hypothesis is H_0: Mean life is 2000 hours and alternate hypothesis is H_1: Mean life is different from 2000 hours. We want to test whether the difference between the sample mean and the population mean is more than what could be attributed as errors due to sampling in which case we reject H_0 . If this difference is less enough

to attribute it as due to sampling then we accept H_0 . This difference which may be attributed to sampling error is the essence of hypothesis testing.

The hypothesis testing procedure is a decision rule that specifies for every possible value of a statistics observable in a simple random sample of size n, whether the null hypothesis is to be accepted or rejected. The set of all possible values of the sample statistics is referred to as the sample space. The test procedure divides the sample space into two mutually exclusive parts called acceptance region and rejection region (critical region).

In the first type (two tailed test) in order to determine whether the mean of the life time has changed from 2000 hours, we determine two values C_1 and C_2 which set the limits on the amount of sampling variation consistent with null hypothesis H_0

If the sample life time lies between C_1 and C_2 we can accept H_0 .A test in which we want to conclude whether a population parameter has changed regardless of its direction is referred to as two tail test.

The second type of hypothesis is one where we have to decide the change with direction. For this, we consider the example that the mean life of bulbs to be at least 2000 hours and the sample mean is 1950 hours. In this case the null and alternative hypothesis may be started as follows .

H_0=$\mu \geq 2000$, $H_1 = \mu < 2000$

In this case the decision rule is as follows:

If the sample mean $\bar{x}$ is less than a number (say) we reject the null hypothesis.

This is represented in the following diagram

Significance Level

The null hypothesis must be stated in such a way that the probability of type I errors can be calculated. This is the reaction for including the equality sign in the statement of null hypothesis (for example $\mu = 2000$)

Is the deviation between μ =2000 hours and the sample mean of 1950 hours so great that we will be unwilling to accept such a difference as due to chance errors of sampling?. If we conclude that such a difference is large then we say that there is a significant difference between the two means. An observed significant difference between a statistic and a parameter rejects the null hypothesis.

I One tail test

The possible sets of null and alternate hypothesis are:

- $H_0: \mu = \mu_0$
  $H_1: \mu > \mu_0$
- $H_0: \mu = \mu_0$
  $H_1: \mu < \mu_0$

- $H_0: \mu \geq \mu_0$

$H_1: \mu < \mu_0$

- $H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$

Where μ_0 is the hypothesized value.

II Two tail test

The null and alternative hypothesis for two tail test are stated as follows:

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

Where μ_0 is the hypothesized value.

Procedure for Test of Hypothesis

Step:1 State the null hypothesis

Step:2 Decide the alternative hypothesis (say one tail or two tail test to be applied)

Step:3 Choose the significance level α

Step:4 Calculate the test statistic $z = \frac{X - E(X)}{S.E\ of\ X}$

Step:5 Find the table of z and decide whether the sample statistics falls within the region of acceptance or rejection.

Large Sample test

We now consider the following tests under large sample test.

- Test for a specified mean
- Test for the equality of two means
- Test for specified proportion
- Test for the equality of two proportions

Test for a specified mean

A random sample of size n ($n \geq 30$) is drawn from a population .We want to test that the population mean has a specified value μ_0

Procedure for Testing (For Two tail test)

The null hypothesis is $H_0: \mu = \mu_0$. The alternative hypothesis $H_1: \mu \neq \mu_0$

Since n is larger the sampling distribution of $\bar{x}$ is approximately normal.

(On the assumption that H_0 is true the statistic $z=\frac{\bar{x}-\mu}{S.E\ of\ \bar{x}}$ is approximately N(0,1).We take the level of significance as α).

Inference

For a significance level α=0.05 (5% level)

- if $|z| < 1.96$, H_0 is accepted at 5% level.
- if $|z| > 1.96$, H_0 is rejected at 5% level.

For α=0.01 (1% level)

- if $|z| < 2.58$, H_0 is accepted at 1% level.
- if $|z| > 2.58$, H_0 is rejected at 1% level.

Procedure for one tail test

i)One tail test (left tail)

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$ (left tail)

At α=0.05, the critical value of $|z|$=1.645

If z <-1.645, H_0 is rejected

If z>-1.645 , H_0 is accepted.

At α=0.01 , the critical value of z=2.33

ii)One tail test (right tail)

If z <1.645, H_0 is accepted

If z>1.645 , H_0 is rejected.

Examples:

A Manufacturer of ball pens claims that a certain pen he manufactures has a mean writing life of 400 pages with a standard deviation of 20 pages. A purchasing agent selects a sample of 100 pens and puts them for test. The mean writing life for the sample was 390 pages. Should the purchasing agent reject the manufactures claim at 5% level?.Table value of z at 5% level is 1.96 for two tail test and 1.64 approximately for one tail test.

Solution:

Given Population mean is μ=400

Population SD σ=20

Sample size n=100

Sample mean $\bar{x} = 390$

Since the sample is a large sample we have to apply z test for testing the sample mean.

H_0: μ=400 (i.e. the sample belongs to a population with mean μ)
$H_1: \mu \neq 400$

The test statistics is z= $\frac{\bar{x}-\mu}{(\frac{\sigma}{\sqrt{n}})}$

$=\frac{390-400}{\frac{20}{\sqrt{100}}}$

=-5

Table value of z at 5% level =1.96

Conclusion

H_o is rejected at 5% level

Since the calculated value of $|z|$ is greater than the tabulated value .

Therefore $\mu \neq 400$ and the manufacturer's claim is rejected at 5% level of significance.

Test for Equality of Two means

Suppose two independent large samples of sizes n_1, and n_2 are drawn from two populations with and means μ_1 and μ_2 and the standard deviations σ_1 and σ_2 .We want to test whether the means are equal.

The Procedure for testing is given below:

$H_0: \mu_1 = \mu_2$. $H_1: \mu_1 \neq \mu_2$ (two tail test)

If the null hypothesis is true, then the sampling distribution of $(\overline{x_1} - \overline{x_2})$ has mean $0(\mu_1 - \mu_2 = 0)$

S.E of $(\overline{x_1} - \overline{x_2})$= $\sqrt{\frac{{\sigma_1}^1}{n_1} + \frac{{\sigma_2}^1}{n_2}}$

Z=$\frac{\overline{x_1}-\overline{x_2}}{S.E\ of\ \overline{x_1}-\overline{x_2}}$ is approximately normal.

Inference:

if $|z| < 1.96$, H_0 is accepted at 5% level.

if $|z| > 1.96$, H_0 is rejected at 5% level.

if $|z| < 2.58$, H_0 is accepted at 1% level.

if $|z| > 2.58$, H_0 is rejected at 1% level.

Example:

Random samples drawn from two places gave the following data relating to the heights of adult males.

	Place A	Place B
Mean height (inches)	68.50	68.58
SD of Heights	2.5	3.0
Sample of sizes	1200	1500

Test at 5% level that the mean height is the same for adults in the two places. (Table value of z at 5% level for two tailed test is 1.96)

Solution:

Given Mean of the first sample=$\overline{x_1}$= 68.50

S.D of the first sample S_1 = 2.5

Size of the first sample n_1 =1200

Mean of the second sample= $\overline{x_2}$ =68.50

S.D of the first sample S_2 = 3.0

Size of the first sample n_2 = 1500

$H_0 : \mu_1 = \mu_2$ (Mean height is the same in the two places)

$H_1 : \mu_1 \neq \mu_2$

The test statistics is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{68.5 - 68.58}{\sqrt{\frac{2.5^2}{1200} + \frac{3^2}{1500}}}$$

$|z|$=-0.76

Conclusion:

H_0 is accepted at 5% level.

Since the calculated value of z is less than the table value of z.

Hence the mean height is the same.

Test For a specified proportion

A random sample of size n>30 with sample proportion p is drawn from a large population. We want to test the hypothesis that the population proportion p has a specified value p_o .

Procedure for testing

$H_0 = p = p_o$

$H_1 = p \neq p_o$ (two tail test)

For a large n , the sampling distribution is approximately normal and the test statistics is given by

$$Z=\frac{p-p_o}{S.E\ \ of\ p}$$

The S.E of p is given by $\sqrt{\frac{P_oQ_o}{n}}$ when H_o is true.

Inference:

If the calculated value of

if $|z| < 1.96$, H_0 is accepted at 5% level.

if $|z| > 1.96$, H_0 is rejected at 5% level.

if $|z| < 2.58$, H_0 is accepted at 1% level.

if $|z| > 2.58$, H_0 is rejected at 1% level.

Example:

A person threw 10 dice 500 times and obtained 2560 times 4,5 or 6. Can this be attributed to fluctuations in sampling?

Solution:

Given

Sample size =10*500 =5000

P= Sample proportion for getting 4,5,6

$=\frac{2560}{5000}$

P= Sample proportion for getting 4,5,6

$P = \frac{3}{6} = \frac{1}{2}$

$Q=1-P=\frac{1}{2}$

H_o: $P=\frac{1}{2}$,

H_o: $P \neq \frac{1}{2}$

The test statistics is $z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$

$$=\frac{\frac{2560}{5000}-\frac{1}{2}}{\sqrt{\frac{1}{2}*\frac{1}{2}*\frac{1}{5000}}}$$

$$=\frac{60}{5000}=\sqrt[2]{5000}=1.697$$

The table value of z at 5% level =1.96

Test For a specified proportion

Given two samples of sizes n_1 and n_2 from the two populations with the sample proportions p_1 and p_2 we would like to test whether the proportions P_1 and P_2 of the two populations are equal.

The following procedure is adopted for this test

$H_0 = p_1 = p_2$

$H_1 = p_1 \neq p_2$ (two tail test)

On the assumption H_0 is true , the sampling distribution of $p_1 - p_2$ is approximately normal with mean 0 . SE of $(p_1 - p_2)$= $\sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}$

Where $P=\frac{n_1p_1+n_2p_2}{n_{1+}n_2}$

The test statistics in this case is

$$Z= \frac{p_1 - P_2}{\sqrt{PQ(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Where Q=1-P

Inference:

If the calculated value of

if $|z| < 1.96$, H_0 is accepted at 5% level.

if $|z| > 1.96$, H_0 is rejected at 5% level of significance.

if $|z| < 2.58$, H_0 is accepted at 1% level significance.

if $|z| > 2.58$, H_0 is rejected at 1% level Significance.

Example:

In a sample of 600 men from a certain city, 450 men are found to be smokers. In a sample of 900 from another city 450 are found to be smokers. Do the data indicate that the two cities are significantly different with respect to prevalence of smoking habit among men?

Solution:

Given n_1=600 , n_2=900

p_1=proportional smokers in the I city =$\frac{450}{600}$

p_2=proportional smokers in the I city =$\frac{450}{900}$

H_o:P_1=P_2 (Proportional smokers in the two cities are equal)

$H_1: p_1 \neq p_2$

The best estimate of the population proportion P is given by

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$= \frac{600 * \frac{450}{600} + 900 * \frac{450}{900}}{600+900}$$

$$P = \frac{450+450}{1500} = \frac{900}{1500} = 0.6$$

The test statistics is

$$\mathbf{z} = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.75 - 0.5}{\sqrt{0.6 * 0.4\left(\frac{1}{600} + \frac{1}{900}\right)}}$$

$$z = \frac{0.25}{\sqrt{0.24\left(\frac{5}{1800}\right)}}$$

Conclusion:

The table value of z at 5% level is 2.58 and the calculated value of z > the table value .

Therefore H_o is rejected at 1% level.

Therefore is a significant difference between the two cities w.r.to the prevalence of smoking habit among men.