UNIT 17 STATISTICAL INFERENCE

Structure

17.0 OBJECTIVES

After going through this Unit you should be in a position to:

- explain the concept of a hypothesis;
- explain the concept of statistical inference;
- test a hypothesis based on a single sample; and
- test the difference between two samples.

17.1 INTRODUCTION

As mentioned in Unit 6 of this course we undertake a sample survey instead of complete census of population because of certain constraints. These constraints could be availability of money, manpower and time. After collection of data through questionnaire, interview or participatory observation method we follow certain steps such as tabulation, presentation and analysis of data. We have discussed these issues in the earlier Units of this course. As you know, we can present data in the form of tables and graphs. Also data can be put to various statistical analyses. Thus we can find out i) measures of central tendency such as mean, median and mode, ii) measures of dispersion such as variance and standard deviation, and iii) correlation and regression coefficients.

Recall that the objective of our study is to analyse the behaviour of the population or the universe, not the sample. In order to make things feasible we are studying the sample and hence whatever results we have got are based on sample information. Naturally a question arises: Are the sample results valid for the population? In other words, can we draw inferences on the basis of sample results?

Let us take a concrete example from our daily life. You must have noticed that before election process starts or just before declaration of election results many newspapers and news channels conduct exit polls. The purpose is to predict election results before

the actual results are declared. At that point of time, it is not possible for the surveyors to ask all the voters about their preferences for electoral candidates - the time is too short, resources are scarce, manpower is not available, and a complete census before election defeats the very purpose of election!

The above is an example of statistical inference. The surveyor actually does not know the result, which is the outcome of votes cast by all the voters. Here all the voters taken together comprise the population. The surveyor has collected data from a representative sample of the population, not all the voters. On the basis of the information contained in the sample, (s)he is making forecast about the entire population.

17.2 CONCEPT OF STATISTICAL INFERENCE

As mentioned above, statistical inference deals with the methods of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Let us recall the example on reading habits of economics students in Sambalpur University. Suppose a question in the questionnaire is, 'How many hours do you study in a day'? We obtain the answer to this question from the students included in the sample (100 students), calculate arithmetic mean and find that 'average number of hours devoted to study by economics students in Sambalpur University is 9.5 hours'. The problem comes up because of two reasons:

i) Sample is a part of the population and there is no reason to expect that sample mean is equal to population mean (if it does, it is a rare coincidence!). In that case, what is the population mean?

ii) A number of samples can be drawn from the same population. Suppose we send two researchers to Sambalpur University on different days and ask them to administer the same questionnaire (on reading habits) on samples of 100 students each. Obviously both researchers would come out with different results (say 9.25 hours and 10.5 hours) as the sampling units are different. Which result do we take to be correct? Can we say that the difference between the studies is negligible?

Remember that population mean is not known to us. We know the sample mean only. We have posed two types of questions above. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose someone claims that the average number of hours devoted to study by economics students in Sambalpur University is 10 hours. On the basis of the sample information can we say that the population mean is not equal to 10 hours? This aspect of statistical inference is called hypothesis testing.

Thus statistical inference has two important aspects: statistical estimation and hypothesis testing (see Fig. 17.1). Estimation could be of two types: point estimation and interval estimation. In point estimation we estimate the value of population parameter as a single point. On the other hand, in the case of interval estimation we estimate lower and upper bounds around sample mean within which population mean is likely to remain.

Hypothesis as you know is an assertion or claim made about the population. It can be in the form of a null hypothesis and its counterpart, alternative hypothesis. We will explain these concepts along with examples below.

Fig. 17.1: Statistical Inference

17.3 STATISTICAL ESTIMATION

As mentioned earlier, we do not know the parameter value and want to guess it by using sample statistic. Obviously the best guess will be the value of the sample statistic. For example, if we do not know the population mean the best guess would be the sample mean. Here, in this case, we use a single value or point as 'estimate' of the parameter and this procedure is called 'point estimation'.

Interval Estimation

The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it. An alternative procedure is to give an interval, which would hold the parameter with certain probability. Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. We call the interval as 'confidence interval'. Here a question may be shaping up in your mind, 'How do we find out the confidence interval'? In order to estimate the confidence interval we have to specify the 'confidence coefficient'. If the confidence coefficient is 95 percent, we get a 95 percent confidence interval. If repeated samples are drawn from a population, a 95 percent confidence interval implies that in 95 out of 100 cases the population mean will remain within the confidence interval. Similarly, in the case of a 99 percent confidence coefficient, if repeated samples are drawn, the population mean will remain within the confidence interval in 99 out of 100 cases.

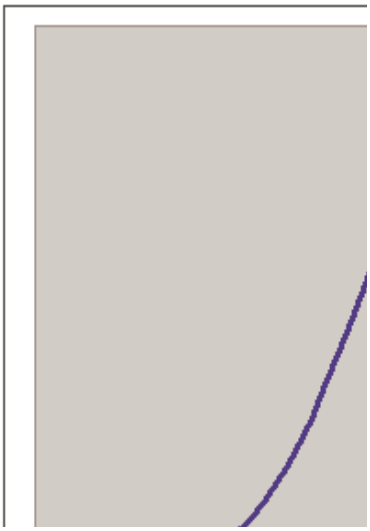
When we are estimating a 95 percent confidence interval, we expect population mean to remain within the interval 95 percent times. It implies that in 5 percent cases we are not sure whether population mean will remain within the interval or not. This 5 percent is called the 'level of significance' and is denoted by (Read as 'alpha'). We will use this concept later in the testing of hypothesis.

Statistical Background

Let us recapitulate some basic concepts from sampling theory. As mentioned earlier, we can draw many samples from a population. Suppose we could draw all possible samples from a population and calculated sample means ( ) from all the samples. We can arrange these sample means in the form of a relative frequency distribution (see Unit 7 for calculation of relative frequency), which is called 'sampling distribution'. If the sample size is large (n 30) then the sampling distribution will follow normal distribution, which looks like a bell-shaped curve when plotted on a graph paper. The sampling distribution has mean equal to population mean ( ) and standard deviation

equal to $\frac{}{\sqrt{n}}$ where is the standard deviation of the population from which the sample is drawn. Remember that the standard deviation of the sampling distribution is called the 'standard error'.

When we subtract the population mean from the sample mean divide it by the population standard deviation we obtain the standard normal variable z, which is equal to . The z variable has the properties that i) it is normally distributed (implies, it looks like a bell-shaped curve), ii) total area under the curve is =1, and iii) arithmetic mean of z is = 0. We plot the z variable in Fig. 17.2.



$\frac{\bar{x} \quad 0}{\sqrt{n}}$

Fig. 17.2: Standard Normal Curve

Estimation of Interval

Remember that z is defined in such a manner that $z \quad \frac{\bar{x}}{/\sqrt{n}}$. Therefore, when sample mean ($\bar{x}$) is equal to population mean ( ), we find that . When $\bar{x}$ is greater than , we obtain a positive value for z. Similarly, when is smaller than we obtain a negative value for z. Thus, as the value of z increases, the difference between sample mean ( ) and population mean ( ) increases.

In Fig. 17.2 we have shown that when z=1.96, the area covered under the curve is 95 percent. Therefore, if we add and subtract 1.96 from sample mean ( ) we obtain a 95 percent confidence interval. In symbols, lower limit and upper limit of the interval are $\bar{x} \quad 1.96\frac{}{\sqrt{n}}, \ \bar{x} \quad 1.96\frac{}{\sqrt{n}}$.

Similarly we obtain the 99 percent confidence interval as ,

since 99 percent area under the standard normal curve is covered when z = 2.58.

Confidence coefficient could take any value. We can ask for a confidence level of say 81 per cent or 97 per cent depending upon how precise our conclusions should be. However, conventionally two confidence levels are frequently used, namely, 95 per cent and 99 per cent.

Self Check Exercise

1) Define the following concepts:
   a) confidence coefficient
   b) confidence interval
   c) level of significance
   d) sampling distribution
   e) standard error
2) A sample of 50 employees was asked to provide the distance commuted by them to reach office. If sample mean was found to be 4.5 km. Find 95 percent confidence interval for the population. Assume that population is normally distributed with a variance of 0.36.
3) For a sample of 25 students in school the mean height was found to be 95 cm. with a standard deviation of 4 cm. Find the 99 percent confidence interval.

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

17.4 CONCEPT OF HYPOTHESIS TESTING

A hypothesis is a tentative statement about a characteristic of a population. It could be an assertion or a claim also. For example, official records for recent years show that female literacy in Orissa is 51 per cent. Since a statement or a claim about the rate of female literacy is being made, it could be considered as a hypothesis.

In hypothesis testing there are four important components: i) null hypothesis, ii) alternative hypothesis, iii) test statistic, and iv) interpretation of results. We discuss each of these below.

Usually statistical hypotheses are denoted by the alphabet H. There are two types of hypothesis: null hypothesis and alternative hypothesis. A null hypothesis is the statement that we consider to be true about the population and put to test by using a test statistic. Usually we denote null hypothesis by H_0. In the example on female literacy in Orissa our null hypothesis is

H_0 : 0.51 ...(17.1)

where is the parameter, in this case female literacy in Orissa.

There is a possibility that the null hypothesis that we intend to test is not true and female literacy is not equal to 51 per cent. Thus there is a need for an alternative hypothesis, which holds true in case the null hypothesis is not true. We denote alternative hypothesis by the symbol and formulate it as

H_A : 0.51 ...(17.2)

We have to keep in mind that null hypothesis and alternative hypothesis cannot be true simultaneously. Secondly, there cannot be a third possibility except for H_0 and H_A about the statement we make. For example, in the case of female literacy in Orissa, there are two possibilities - literacy rate is 51 per cent or it is not 51 per cent; a third possibility is not there.

In most cases we find a difference between sample mean ($\bar{x}$) and population mean ( ). Is the difference because of sampling fluctuation or is there a genuine difference between the sample and the population? In order to answer this question we need a test statistic to test the difference between the two. The result that we obtain by using the test statistic needs to be interpreted and a decision needs to be taken regarding the acceptance or rejection of the null hypothesis.

Let us go back to the standard normal curve given at Fig. 17.1. We mentioned that as the value of z increases, the difference between sample mean ( ) and population mean ( ) increases. Moreover, higher the difference between and , higher is the absolute value of z. Thus z-value measures the discrepancy between and , and therefore can be used as a test statistic for hypothesis testing. Note that we are concerned with the difference between and . Therefore, negative or positive sign of z does not matter much.

Our task is to find out a critical value of z beyond which the difference between and is significant. Hence, we take the absolute value of z (denoted by) and if it is less than the critical value we should not reject the null hypothesis. If the absolute value of z exceeds the critical value we should reject the null hypothesis and accept the alternative hypothesis.

Therefore, in the case of large samples z can be considered as a test statistic for hypothesis testing such that

$$z \quad \frac{|\bar{x} \quad |}{/\sqrt{n}} \qquad \ldots(17.3)$$

The above procedure is often called z-test. By applying sample values in the formula given at (17.3) above we obtain the observed value of z. We compare it with the critical value of z (to be discussed below). ...(17.4)

When the sample size is small, the sampling distribution does not follow normal distribution. Hence, we cannot apply z- test. In the case of small samples, however, we apply t-test, which again is bell-shaped, but has a larger variance compared to normal distribution. The test statistic for t-test is given by

$$t \quad \frac{|\bar{x} \quad |}{s/\sqrt{n}} \qquad \ldots(17.4)$$

A problem here is that the critical value of t depends upon the 'degree of freedom', defined as (n 1) where n is the sample size. For example, when sample size is 20, degree of freedom is 20 -1 =19. Thus the critical value of t varies according to two factors: i) degrees of freedom, and ii) requisite level of significance.

Fig. 17.3: Testing of Hypothesis

In Fig. 17. 3 we present the type of test to be applied in different situations. Some of the factors that guides us in deciding on the test statistic to be used are: i) whether population is normal or not, ii) whether sample size is small or large, and iii) whether population variance ( ) is known to us or not. You may wonder that since population mean is not known to us (our objective is to estimate it from the sample), how do we know population variance! However, we begin with the simpler case of known variance and later on consider the more realistic case of unknown population variance.

17.5 CRITICAL REGIONS AND TYPES OF ERRORS

In Section 17.3 we had mentioned that when z = 1.96, we have 5 percent level of significance (and 95 percent confidence level). An implication of the above is that in 95 out of 100 samples the parameter remains within the confidence interval. Therefore, in 5 out of 100 samples, the parameter will not remain within the confidence interval. The area on the extreme left and extreme right of the standard normal curve (see Fig. 17.2 above) belong to this 5 percent. This region of the curve is called the 'critical region' and the value of z beyond which the critical region starts is called 'critical value'.

The critical values for z depend upon the level of significance. In Table 17.1 these critical values for certain specified levels of significance ( ) are given for the tests to be conducted under the assumption of normal distribution.

Table 17.1: Critical Values for z-statistic

Significance Level ( )	0.10	0.05	0.01
Critical value of z	1.65	1.96	2.58

Type I and Type II Errors

In hypothesis testing we reject or do not reject a hypothesis with certain level of confidence. As mentioned above, a confidence level of 95 percent implies that in 95

out of 100 samples the parameter remains within the acceptance region and in 5 per cent cases parameter remains in the rejection region. Thus in 5 per cent cases the sample is drawn from the population but sample mean is too far away from the population mean. In such cases the sample belongs to the population but our test procedure rejects it. Obviously we commit an error such that H_0 is true but gets rejected. This is called 'Type I error'. Similarly there could be situations when the H_0 is not true, but on the basis of sample information we do not reject it. Such an error in decision-making is termed 'Type II error' (see Table 17.2).

Note that 'Type I error' specifies how much error we are in a position to tolerate. Type I error is equal to the level of significance, and is denoted by . Thus implies that we can tolerate 5 percent error in our decision-making. Remember that confidence level is equal to ( ).

Table 17.2: Type of Errors

	H_0 True	H_0 Not true
Reject H_0	Type I Error	Correct decision
Do not reject H_0	Correct decision	Type II Error

In the case of small samples we have to use t-test and thus critical values need to be decided on the basis of t-distribution. Application of t-test is a bit complex as we have to look for the i) degrees of freedom, and ii) the level of significance.

We will work out some examples based on z-test and t-test in the next Section.

As mentioned earlier the convention is to apply 5% or 1% level of significance. For these two levels of significance we present the critical values of t-destination for different degrees of freedom in Table 17.3 at the end of this unit.

1 0.05

Self Check Exercise

4) Distinguish between the following:

   a) Null hypothesis and Alternative hypothesis

   b) Confidence level and Level of significance

   c) Type I and Type II errors

5) Suppose a sample of 100 students has mean age of 12.5 years. Show through a diagram the critical region at 5 per cent level of significance to test hypothesis that the sample is equal to the population mean. Assume that population mean and standard deviation are 10 years and 2 years respectively.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

17.6 TESTING OF HYPOTHESIS FOR A SINGLE SAMPLE

We have so far explained the concepts of null and alternative hypotheses. Also we have learnt that in the case of large samples we apply z-test and in the case of small samples we apply t-test. In many situations we are asked to judge whether a sample is significantly different from a given population. For example, let us assume that we surveyed a sample of 400 households of Raigarh district of Chhatisgarh state and calculated the per capita income of these households. Subsequently, our task is to test the hypothesis that per capita income calculated from the sample is not different from the per capita income of the district.

In the above example we can have two different situations: i) population (in this case all the households of the district) variance is known, ii) population variance is not known to us. We explain the steps to be followed in each case below.

17.6.1 Population Variance is Known

The steps you should follow are:

1) Specify the null hypothesis.
2) Find out whether it requires one-tail or two-tail test. Accordingly identify your critical region. This will help in specification of alternative hypothesis.
3) Apply sample values to z-statistic.
4) Find out from z-table the critical value according to level of significance.
5) If you obtain a value lower than the tabulated value do not reject the null hypothesis.
6) If you obtain a value greater than the tabulated value reject the null hypothesis and accept the alternative hypothesis

Example 1

Let us consider the case that we know the per capita income of Raigarh district of Chhatisgarh as well as its variance. Suppose the data available in official records show that per capita income of Raigarh district is Rs. 10,000 and standard deviation of per capita income is Rs. 1,500. However, we did a sample survey of 400 households and found that their per capita income is Rs. 10,500. Do we accept the data provided in official records?

In this case = Rs. 10,000

= Rs. 1,500

= Rs. 10,500

n = 400

The sample size is large and variance of the population is known. As given in Fig.17.3 we apply z-test.

Our null hypothesis in this case is

$H_0 : \bar{x}$

The null hypothesis suggests that sample mean is equal to population mean. In other words, per capita income obtained from the sample is the same as the data provided in official records.

Our alternative hypothesis is

$H_A : \bar{x}$

By substituting values in the above we obtain

$$z \quad \frac{|10500 \quad 10000|}{1500/\sqrt{400}} \quad \frac{500}{500/20}$$

In the above case since z = 6.67, the sample lies in the critical region and we reject the hypothesis. Thus the per capita income obtained from the sample is significantly different from the per capita income provided in official records.

Example 2

Suppose the voltage generated by certain brand of battery is normally distributed. A random sample of 100 such batteries was tested and found to have a mean voltage of 1.4 volts. At 0.01 level of significance, does this indicate that these batteries have a general average voltage that is different from 1.5 volts? Assume that population standard deviation is 0.21 volts.

Here, H_0: $= 1.5$

Since average voltage of the sample can be different from average voltage of the population if it is either less than or more than 1.5 volts, our rejection region is on both sides of the normal curve. Thus it is a case of two-tail test and alternative hypothesis is

Since the population standard deviation s is known, the test statistic is

H_1 30 1.5

$$z \quad \frac{|\bar{x} \quad |}{\sqrt{n}} \quad \frac{|1.4 \quad 1.5|}{\frac{0.21}{\sqrt{100}}} \quad 4.8$$

From the Table 17.2 we find that the critical value at 1 per cent level significance is 2.58. Since the actual value of z is greater than 2.58 we reject the null hypothesis at 1% level and accept the alternative hypothesis that the average life of batteries is different from 1.5 volts.

17.6.2 Population Variance not Known

The assumption that population standard deviation ( ) is known to us is unrealistic, as we do not know the population mean itself. When is unknown we have to estimate it by sample standard deviation (s). In such situations there are two possibilities depending upon the sample size. If the sample size is large ( ) we apply z-statistic, that is,

$$z \quad \frac{|\bar{x} \quad |}{s/\sqrt{n}} \qquad \ldots(17.5)$$

In case the sample size is small ($n \quad 30$) we apply t-statistic with $n \quad 1$ degrees of freedom. The test statistic is

$$t \quad \frac{|\bar{x} \quad |}{s/\sqrt{n}} \qquad \ldots(17.6)$$

The steps you should follow are:

1) Specify the null hypothesis and alternative hypothesis.
2) Check whether sample size is large ($n \quad 30$) or small ($n \quad 30$).
3) In case $n \quad 30$, apply z-test (17.5).
4) Find out from z-table the critical value according to level of significance ( ).
5) In case , apply t-test (17.6).
6) Find out from t-table the critical value for $n \quad 1$ degrees of freedom and level of significance ( ).
7) If you obtain a value lower than the tabulated value do not reject the null hypothesis.
8) If you obtain a value greater than the tabulated value reject the null hypothesis and accept the alternative hypothesis

Example 3

A tablet is supposed to contain on an average 10 mg. of aspirin. A random sample of 100 tablets show a mean aspirin content of 10.2 mg. with a standard deviation of 1.4 mg. Can you conclude at the 0.05 level of significance that the mean aspirin content is indeed 10 mg.?

Here, the null hypothesis is

The rejection region is on both sides of 10 mg. Thus it requires a two-tail test and .

Also, the sample mean is $\bar{x} \quad 10.2$ and the sample size n = 100. Since population standard deviation is not known we estimate it by sample standard deviation s and our test statistic is . By applying relevant values from the sample we obtain

$$z \quad \frac{|\bar{x} \quad |}{\frac{s}{\sqrt{n}}} \quad \frac{|10.2 \quad 10|}{\frac{1.4}{\sqrt{100}}} \quad 1.43$$

At 5 per cent level of significance the critical value of z is 1.96. since the z value that we have obtained is less than 1.96, we do not reject the null hypothesis. Therefore the mean level of aspirin is 10 mg.

Example 4

The population of Haripura district has a mean life expectancy of 60 years. Certain health care measures are undertaken in the district. Subsequently, a random sample of 25 persons shows an average life expectancy of 60.5 years with a standard deviation of 2 years. Can we conclude at the 0.05 level of significance that the average life expectancy in the district has remained the same?

Here, H_0: $= 60$

We have to test for an increase in life expectancy. Thus it is a case of one-tail test and the rejection region will be on the right-hand tail of the standard normal curve.

Hence our alternative hypothesis is

Here population standard deviation s is not known and we estimate it by the sample standard deviation s. Here the sample size is small hence we have to apply t-statistic given at (17.6).

$$t \quad \frac{|\bar{x} \quad |}{\frac{s}{\sqrt{n}}} \quad \frac{|60.5 \quad 60|}{\frac{2}{\sqrt{25}}} \quad 1.25$$

Since sample size is 25, degrees of freedom are 25-1 = 24. From the t-table we find that for 24 degrees of freedom, 5 per cent level of significance.

Since t-value obtained above is less than the tabulated value we do not reject the hypothesis. Therefore, we accept the alternative hypothesis that life expectancy has not changed after the health care measures.

Self Check Exercise

6) A report claimed that in the 'School Leaving Examination', the average marks scored in Mathematics were 78 with a standard deviation of 16. However, a random sample of 37 students showed an average of 84 marks in Mathematics. In the light of this evidence, can we conclude that the average has remained unchanged? Use 0.05 level of significance.

7) A passenger car company claims that average fuel efficiency of cars is 35 kms per litre of petrol. A random sample of 50 cars shows an average of 32 kms per litre with a standard deviation of 1.2 km. Does this evidence falsify the claim of the passenger car company at 0.01 level of significance?

8) A random sample of 200 tins of coconut oil gave an average weight of 4.95 kg per tin with a standard deviation of 0.21 kg. Do we accept the hypothesis of net weight of 5 kg per tin at 0.01 level of significance?

9) According to a report, the national average annual income of the government employees during a recent year was Rs. 24,632 with a standard deviation of Rs. 1827. A random sample of 49 government employees during the same year showed an average annual income of Rs. 25,415. On the evidence of this sample, at 0.05 level of significance, Can we conclude that the national average annual income of government employees was indeed Rs. 24,632?

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

17.7 TEST FOR DIFFERENCE BETWEEN TWO SAMPLES

Many times we need to test for the difference between two samples. The objective may be to ascertain whether both samples are drawn from the same population or to

check whether a particular characteristic is the same in two populations. For example, we formulate a hypothesis that the production per worker in plant A is the same as the production per worker in plant B. We discuss below the procedure for testing of such a hypothesis.

Here again we deal with two different situations: whether variance of both the populations are known. Another consideration is sample size: large or small.

The null hypothesis is the statement that population means of both the populations are the same. In notations

$$H_0 : \quad _1 \quad _2 \qquad \ldots(17.7)$$

The alternative hypothesis is the statement that both the population means are different. In notations

$$H_A : \quad _1 \quad _2 \qquad \ldots(17.8)$$

Population Variance is known

When standard deviations (positive square root of variance) of both the populations are known we apply z statistic specified as follows:

$$z \quad \frac{\left|(\bar{x}_1 \quad \bar{x}_2) \quad (\quad _1 \quad _2)\right|}{\sqrt{\frac{^2_1}{n_1} \quad \frac{^2_2}{n_2}}} \qquad \ldots(17.9)$$

In (17.8) above, subscript 1 refers to the first sample and subscript 2 refers to the second sample. By applying relevant data in (17.9) we obtain the actual value of z and compare it with the tabulated value for specified level of significance.

Example 5:

A bank wants to find out the average savings of its customers in Delhi and Kolkata. A sample of 250 accounts in Delhi shows an average savings of Rs. 22500 while a sample of 200 accounts in Kolkata shows an average savings of Rs. 21500. It is known that standard deviation of savings in Delhi is Rs. 150 and that in Kolkata is Rs. 200. Can we conclude at 1 percent level of significance that banking pattern of customers in Delhi and Kolkata is the same?

In this case the null hypothesis is $H_0 : \quad _1 \quad _2$

and the alternative hypothesis is $H_A : \quad _1 \quad _2$

We are provided with the information that

$\bar{x}_1$ = Rs. 22500 $\qquad _1$ = Rs. 150

= Rs. 22400 $\qquad _2$ = Rs. 200

= 250 $\qquad n_2$ = 200

Since $_1$ and are known we apply z-test.

The test statistic is .

By applying the information provided above we obtain

$$z \quad \frac{|22500 \quad 22400|}{\sqrt{\frac{150^2}{250} \quad \frac{200^2}{200}}} \quad \frac{100}{\sqrt{90 \quad 20}}$$

We find that at 1 per cent level of significance the critical value obtained from Table 17.1 is 2.58.

Since the actual value is greater than the tabulated value the null hypothesis is rejected and the alternative hypothesis is accepted. Thus the banking pattern of customers in Delhi and Kolkata are different.

Population Variance is not known

When population variance (^{2}) is not known we estimate it by sample variance (s^2). If both samples are large in size (n>30) then we apply z statistic as follows:

$$z \quad \frac{|(\bar{x}_1 \quad \bar{x}_2) \quad (\ _1 \quad \ _2)|}{\sqrt{\frac{s_1^2}{n_1} \quad \frac{s_2^2}{n_2}}} \qquad \ldots(17.10)$$

On the other hand, if samples are small in size ($n \quad 30$) then we apply t-statistic as follows:

$$t \quad \frac{|(\bar{x}_1 \quad \bar{x}_2) \quad (\ _1 \quad \ _2)|}{\sqrt{\frac{s_1^2}{n_1} \quad \frac{s_2^2}{n_2}}} \qquad \ldots(17.11)$$

$t \quad ^2 \frac{|(\bar{x}_1 \quad \bar{x}_2) \quad (\ _1 \quad \ _2)|}{\sqrt{\frac{s_1^2}{n_1} \quad \frac{s_2^2}{n_2}}}$

Degrees freedom for t-test = $(n_1 \quad 1) \quad (n_2 \quad 1) \quad n_1 \quad n_2 \quad 2$

Example 6

A mathematics teacher wants to compare the performance of Class X students in two sections. She administers the same set of questions to 25 students in Section A and 20 students in Section B. she finds that Section A students have a mean score of 78 marks with standard deviation of 4 marks while Section B students have a mean score of 75 marks with standard deviation of 5 marks. Is the performance of students in both Sections different at 1 per cent level of significance?

In this case the null hypothesis is $H_0 : \ _1 \quad \ _2$

and the alternative hypothesis is $H_A : \ _1 \quad \ _2$

We are provided with the information that

$\bar{x}_1 = 78$ $\qquad$ $s_1 = 4$

$\bar{x}_2 = 75$ $\qquad$ $s_2 = 5$

$n_1 = 25$ $\qquad$ $n_2 = 20$

Since $\ _1$ and $\quad$ are not known and sample sizes are small we apply t-test.

$$= \frac{|78 \quad 75|}{\sqrt{\frac{4^2}{25} \quad \frac{5^2}{20}}} = \frac{3}{1.37} = 2.18$$

The degree of freedom in this case is 25+20-2 = 43.

We can find out from Table 17.3 that at the 1 per cent level of significance the t-value for 43 degrees of freedom is 2.69.

Since the tabulated value of t is less than actual value of t we reject the hypothesis. Therefore, students in Section A and Section B are different with respect to their performance in mathematics.

17.8 CONTINGENCY TABLE

The concepts of confidence interval and the procedures of testing a hypothesis discussed so far relate to numerical. In the case of qualitative data, however, we cannot undertake such tests, as we do not have parameters. In the qualitative data, therefore, we require to develop non-parametric tests.

There are many types of non-parametric tests depending upon our need. However, we confine ourselves to a common procedure, that is, chi-square (pronounced as kai-squared) test for the test of independence between variables. Recall that qualitative data can be arranged in categories (see Unit 6) and presented in the form of a two-way table.

In order to explain the application of chi-square test let us take a concrete example. Suppose we want to test a hypothesis that number of children in a family is independent of the occupation of father. We divide occupation of father into five categories - i) unemployed, ii) unskilled labour, iii) skilled labour, iv) self-employed, and v) professional. Similarly we divide families into five categories according the number of children - i) no child, ii) one child, iii) two children, iv) three children, and v) more than three children. For a sample of 650 families the data obtained is presented in Table 17.3.

Table 17.3: Occupation and Number of Children

Number of Children	Occupation					Total
	Unemployed	Unskilled Labour	Skilled Labour	Self-Employed	Professional	
	(1)	(2)	(3)	(4)	(5)	
0	10	15	10	12	11	58
1	35	25	17	18	25	120
2	22	33	45	40	43	183
3	11	40	48	58	30	187
4	11	33	30	19	9	102
Total89	146	150	147	118	650	

Table 17.3 is called contingency table, because we are trying to find whether the number of children is contingent upon the occupation of the father.

Our purpose is test for possible relationship between the number of children and the occupation of father. Thus the null hypothesis is specified as:

: Number of children and occupation of father are independent against the alternative hypothesis

H_A: Number of children and occupation of father are dependent

Expected Frequency

In Table 17.3 we have presented the observed frequency for each cell in the table. What should be the expected frequency when there is no relationship between the variables under consideration? We will answer this question below.

Expected frequency is calculated under the assumption that there is no relationship between number of children and occupation of father. For each cell in Table 17.2 the expected frequency is obtained by

$$E_{ij} = \frac{(\text{Row } i \text{ total})\ (\text{Column } j \text{ total}}{\text{Sample size}} \qquad \ldots(17.12)$$

Where E_{ij} is expected frequency for row 'i' and column j. For example, for row 2 and column 2 the expected frequency is

$$E_{22} \quad \frac{\text{row 2 total} \quad \text{colm. 2 to}}{\text{sample size}} \quad .$$

We find out the row and column totals for the data given in Table 17.3 and estimate the expected frequency for each cell. These are given in Table 17.4.

Table 17.4: Calculation of Expected Frequency for Each Cell

Number of Children		Occupation					Total
		Unemployed	Unskilled Labour	Skilled Labour	Self-Employed	Professional	
		c_1	c_2	c_3	c_4	c_5	
0	r_1	7.94	13.03	13.38	13.12	10.53	58.00
1	r_2	16.43	26.95	27.69	27.14	21.78	120.00
2	r_3	25.06	41.10	42.23	41.39	33.22	183.00
3	r_4	25.60	42.00	43.15	42.29	33.95	187.00
4		13.97	22.91	23.54	23.07	18.52	102.00
Total		89.00	146.00	150.00	147.00	118.00	650.00

r_5

The next step is to compare the observed frequency with the expected frequency. In order to compare the observed frequency with the expected frequency we construct the chi-square statistic, which is given by

$$_2 \quad \frac{(O_i \quad E_i)^2}{E_i} \qquad \ldots(17.3)$$

where O refers to observed frequency and E refers to expected frequency.

The chi-square statistic has degrees of freedom $(r \quad 1)(c \quad 1)$. For example, if there are 3 rows and 4 columns, then degrees of freedom is $(3 \quad 1)(4 \quad 1) \quad 6$.

Let us summarise the steps to be followed in chi-square test. These are:

1) specify the null and alternative hypotheses
2) calculate the expected frequency for each cell by using (20.5)

3) calculate the observed value of statistic by using (20.6)

4) determine degrees of freedom according to the formula $(r \quad 1)(c \quad 1)$

5) check the level of significance ( ) required

6) Find out the critical value of for and relevant degrees of freedom

7) compare the observed value of with the critical value of ^{2}

8) if observed value is less than critical value, then do not reject H_0

9) if observed value is greater than critical value, then reject H_0 and accept H_A

For the data given in Table 17.4 let us find out the observed value of ^{2}.

Table 17.5: Calculation of $\frac{(O_i \quad E_i)^2}{E_i}$ for each Cell

Number of Children	Occupation					Total
	Unemployed	Unskilled Labour	Skilled Labour	Self-Employed	Professional	
	(1)	(2)	(3)	(4)	(5)	
0	0.53	0.30	0.86	0.10	0.02	1.80
1	20.99	0.14	4.13	3.08	0.47	28.81
2	0.37	1.60	0.18	0.05	2.88	5.08
3	8.33	0.10	0.54	5.84	0.46	15.26
4	0.63	4.44	1.77	0.72	4.89	12.46
Total	30.85	6.58	7.48	9.77	8.72	63.41

Since there are 5 rows and 5 columns, the degrees of freedom is .
The critical values of x^2 for 5% & 1% level of significance for different degrees of freedom are given in Table 17.7 at the end of the Unit. We find from the table that for 16 d.f. the critical value of at 5 per cent level of significance is 26.30. The observed value of ^{2} to be 63.41. Since the observed value is greater than the critical value we reject the null hypothesis and accept the null hypothesis. Therefore, we conclude that the variables 'number of children' and 'occupation of father' are not independent.

Self Check Exercise

10) Explain the following concepts.

a) expected frequency

b) critical value of ^{2}

11) There are three brands (orange, cola and lemon) of soft drinks produced by a company. A survey of 160 persons in two states (one form north- Punjab and one from south- Tamil Nadu) provides the following information.

	Orange	Cola	Lemon
Punjab 33	26	31	
Tamil Nadu	17	24	29

Test the hypothesis that there is no preference for particular brand of soft drink on both the states (0.05).

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

17.9 SUMMARY

Drawing conclusions about a population on the basis of sample information is called statistical inference. Here we have basically two things to do: statistical estimation and hypothesis testing.

An estimate of an unknown parameter could be either a point or an interval. Sample mean is usually taken as a point estimate of population mean. On the other hand, in interval estimation we construct two limits (upper and lower) around the sample mean. We can say with stipulated level of confidence that the population mean, which we do not know, is likely to remain within the confidence interval. In order to construct confidence interval we need to know the population variance or its estimate. When we know population variance, we apply normal distribution to construct the confidence interval. In cases where population variance is not known, we use t distribution for the above purpose. Remember that when sample size is large (n>30) t-distribution approximates normal distribution. Thus for large samples, even if population variance is not known, we can use normal distribution for estimation of confidence interval on the basis of sample mean and sample variance.

Subsequently we discussed the methods of testing a hypothesis and drawing conclusions about the population. Hypothesis is a simple statement (assertion or claim) about the value assumed by the parameter. We test a hypothesis on the basis of sample information available to us. In this Unit we considered two situations: i) description of a single sample, and ii) comparison between two samples.

In the case of qualitative data we cannot have parametric values and hypothesis testing on the basis of z statistic or t-statistic cannot be performed. Chi-square test is applied to such situations. Chi-square test is a non-parametric test, where no assumption about population is required. There are various types of non-parametric tests beside chi-square test. Moreover, chi-square test can be applied to many situations. We learnt about a particular application of chi-square test - contingency table. In contingency table we test the null hypothesis that variables under consideration are independent against the alternative hypothesis that variables are related. We compare expected frequency with observed frequency and construct the chi-square statistic. If the observed value of chi-square exceeds the expected value of chi-square we reject the null hypothesis.

17.10 ANSWERS TO SELF CHECK EXERCISES

1) Go through the text and define these terms.

2) Since it is large sample we apply z-statistic. The confidence interval is

3) Since it is small sample and population variance is not given we apply t-statistics with degrees of freedom 24. The tabulated value of t at 99 per cent confidence level is 2.49. The confidence interval is .

4) Go through the text and define these terms.

5) It is large sample and is unknown. In order to show the rejection regions we use the standard normal curve. Accordingly draw the diagram.

6) Since it is large sample with known variance, we apply z-statistic. The alternative hypothesis is . The observed value of z is 2.28 and critical value of z at 5% level of significance is 1.96. Since the observed value is greater than the critical value we reject the null hypothesis. Therefore, we conclude that the average marks were different from 78.

7) It is a large sample with unknown variance. It requires two-tail test. The observed value of z is 17.68 and critical value of z at 1% level of significance is 2.58. Since the observed value is greater than the critical value, the null hypothesis is rejected.

8) It is a large sample with unknown variance. We test the null hypothesis with z-statistic. Observed value of z is 3.37. Null hypothesis is rejected.

9) Since it is large sample with known standard deviation, we apply z-statistic. Observed value of z is 3.00. Critical value of z at 5% level of significance is 2.58. Null hypothesis is rejected. Therefore, the national average of annual income of government employees was different from Rs. 24632.

10) Go through the text and define these terms.

11) The expected frequency are

	Orange	Cola	Lemon
Punjab	28.13	28.13	33.75
Tamil Nadu	21.88	21.88	26.25

The observed value of chi-square statistic is 2.98. Degrees of freedom is 2. The critical value of chi-square at 5 per cent level of significance at 2 degrees of freedom is 5.99. Hence null hypothesis is not rejected and soft drink consumption is independent of the region.

17.11 KEYWORDS

Confidence Level : It gives the percentage (probability) of samples where the population mean would remain within the confidence interval around the sample mean. If is the significance level the confidence level is (1-).

Contingency Table : A two-way table to present bivariate data. It is called contingency table because we try to find whether one variable is contingent upon the other variable.

Degrees of Freedom : It refers to the number of pieces of independent information that are required to compute some characteristic of a given set of observations.

Estimation : It is the method of prediction about parameter values on the basis of sample statistics.

Expected Frequency : It is the expected cell frequency under the assumption that both the variables are independent.

Nominal Variable : Such a variable takes qualitative values and do not have any ordering relationships among them. For example, gender is a nominal variable taking only the qualitative values, male and female; there is no ordering in 'male' and 'female' status. A nominal variable is also called an attribute.

Parameter : It is a measure of some characteristic of the population.

Population : It is the entire collection of units of a specified type in a given place and at a particular point of time.

Random Sampling : It is a procedure where every member of the population has a definite chance or probability of being selected in the sample. It is also called probability sampling. Random sampling could be of many types: simple random sampling, systematic random sampling and stratified random sampling.

Sample : It is a sub-set of the population. It can be drawn from the population in a scientific manner by applying the rules of probability so that personal bias is eliminated. Many samples can be drawn from a population and there are many methods of drawing a sample.

Sampling Distribution : It is the relative frequency or probability distribution of the values of a statistic when the number of samples tends to infinity.

Sampling Error : In the sampling method, we try to approximate some feature of a given population from a sample drawn from it. Now, since in the sample all the members of the population are not included, howsoever close the approximation is, it is not identical to the required population feature and some error is committed. This error is called the sampling error.

Significance Level : There may be certain samples where population mean would not remain within the confidence interval around sample mean. The percentage (probability) of such cases is called significance level. It is usually denoted by .

When = 0.05 (that is, 5 percent) we can say that in 5 per cent cases we are likely to reach an incorrect decision or commit Type I error. Level of significance could be at any level but it is usually taken at 5 percent or 1 percent level.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

17.12 REFERENCES AND FURTHER READINGS

Kiess, H.O. (1989). *Statistical Concepts for the Behavioral Sciences.* Boston: Allyn and Bacon.

IGNOU Course Material (2005). *EEC 13: Elementary Statistical Methods and Survey Techniques.* Block 7.