UNIT 4 SAMPLING DISTRIBUTIONS

Structure **Page No.**

4.1 INTRODUCTION

In the last block we emphasised the techniques used to describe data. To illustrate these techniques we organised the data into a frequency distribution and computed various averages and measures of dispersion. Measures such as mean and standard deviation were computed to describe the central tendency of the data and the extent of its spread. In this unit we start discussing the procedures of statistical inference which is the core of statistical analysis. We have already discussed the concepts of probability and probability distributions which lays a foundation stone for statistical inference.

As we have seen in Unit 1, in statistical inference we use the sample to make inference about the unknown population characteristic. As you know, it is a common practice for every body to draw inference about a large population by examining a sample from it. This technique is used in decision making based on statistical analysis. There if we want to make decision about a large population, may be finite or infinite, we take representative samples from the population and get results for these samples which is then used to make decisions about the population. But the results we obtain from the samples may vary from sample to sample. To make valid judgements about the population, it is necessary to study this variation. In this unit we introduce you to the concept of sampling distribution. Further we discuss how to construct a sampling distribution by selecting all samples of size, say, n from a population and how this is used to make inferences about the population. To start with we study the sampling distribution of means in Sec.4.3.

The fact that different samples are likely to give different results shows that there is a possibility of error which may affect the final decision. The sampling distribution allows us to calculate this error also.

Very often, it is not easy to determine the sampling distribution exactly. In such cases we make use of a fundamental theorem in statistics known as the Central Limit Theorem. We shall only state this theorem and discuss its utility. In Sec.4.3, we also discuss sampling distribution of proportions. You will see that this distribution is useful when we take samples from a binomial population.

In Sections 4.4, 4.5 and 4.6 we shall discuss three important sampling distributions, t, χ^2 and F.

The examples and exercises in this unit are focused on how sampling techniques can assist us in making decision about various real-life problems.

Here is a list of what you should be able to do by the end of this unit.

Objectives

After reading this unit, you should be able to

- explain the need to study the sampling distribution of a statistic,
- use sampling distribution of mean and proportion to draw inferences about the population mean and population proportion,
- calculate the error in sampling using sampling distribution,
- use central limit theorem to make inferences,
- use χ^2, F and t-distributions to solve some problems of statistical inference.

4.2 POPULATION AND SAMPLES

The term population in statistics is applied to sets or collection of objects, actual or conceptual, and mainly to sets of numbers, measurements, or observations. For example, if we collect data on the number of television sets for each household in Mumbai, then a listing of number of television sets per household constitutes the population.

In some cases, such as concerning the number of television sets per household, the population is **finite**; in other cases, such as the determination of some characteristic of all units past, present, and future, that might be manufactured by a given process, it is convenient to think of the population as **infinite**. Similarly, we look upon the results obtained in a series of flips of a coin as a sample from the hypothetically infinite population consisting of all conceivably possible flips of the coin

In the last block we have seen that populations are often described by the distributions of the values, and it is a common practice to refer to a population in terms of this distribution. (For finite populations, we are referring here to the actual distribution of its values; for infinite populations, we are referring to the corresponding probability distribution or probability density.). For example, we may refer to a number of flips of a coin as a sample from a "binomial population" or to certain measurements as a sample from a "normal population." Hereafter, when referring to a "population f(x)" we shall mean a population such that its elements have a frequency distribution, a probability distribution, or a density with values given by f(x).

If a population is infinite it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus, it is usually necessary to use a **sample**, a part of a population, and infer from it results pertaining to the entire population. Clearly, such results can be useful only if the sample is in some way "representative" of the population. It would be unreasonable, for instance, to expect useful generalisations about the population of 1984 family incomes in India on the basis of data pertaining to home owners only. Similarly, we can hardly expect reasonable generalisations about the performance of a tyre if it is tested only on smooth roads.

Here we notice that the elements in the population may not be always uniform in character. For example in the case of population of 1984, all of them may not be home owners. Like that all of them may not be having other characteristics like educated etc. Such population where the elements of the population are not having uniform

characteristic is called **heterogeneous population**. Otherwise the population is called **homogeneous**. So we have to take some care while taking samples from a heterogeneous population. To assure that a sample is representative of the population from which it is obtained, and to provide a framework for the application of probability theory to problems of sampling, we shall limit our discussion to what is called a random sample. For sampling from finite populations, they are defined as follows:

Definition 1 : A set of observations $x_1, x_2, \ldots, x_n$ constitutes a random sample of size n from a finite population of size N, if it is chosen so that each subset S containing n of the N elements of the population has the same probability of being selected.

Note that this definition of randomness pertains essentially to the manner in which the samples are selected. This holds also for the following definition of a random sample from a theoretical (possibly infinite) population:

Definition 2: A set of observations $x_1, x_2, \ldots, x_n$ constitutes a random sample of size n from a population with density or mass function f(x) if

1) each x_i is a realisation of a random variable whose frequency/ density function is given by f(x).

2) these n random variables are independent.

There are several ways of selecting a random sample. For relatively small sample, this can be done by drawing lots, or equivalently by using a table of "random numbers" specially constructed for such purposes. We shall discuss the utility of this table in and other selection procedures in detail in Unit 12, Block 4.

Before proceeding further, why don't you try some exercises now.

E1) Suppose we want to know the average age of female students enrolled in IGNOU BDP programme. What will be the population for this study? Is this population finite or infinite?

E2) Which of the following is an appropriate sample for studying the situation given in E1.

i) IGNOU female of BDP students selected from Delhi region.

ii) Randomly selected female students from the list of BDP students.

iii) Female students selected from two study centres of each regional centres.

iv) Randomly selected students from all students registered with IGNOU.

A population is completely determined (or characterised) by certain fixed quantities (often unknown). These fixed quantities are called **parameters** and the problem consists of "inferring" about these characteristics based on a sample data.

A function of a sample observation is called a **statistic**.

For example in the case of the problem determining the average of income daily wagers in a particular city, we find average income of a sample of workers chosen. Such measures (or quantities) calculated from a sample is called a statistic.

Try this exercise, now.

E3) Give an example to show the difference between a parameter and a statistic.

Now, suppose we want to study the population using the "mean" as given in the earlier example. When we are referring to the mean of a population, we call it **population mean** and when we are referring to the mean of a sample, we call it **sample mean**.

Population mean is usually denoted by μ and sample mean is denoted by $\bar{x}$

Similarly for other measures like standard deviation, proportion etc, we have population standard deviation, sample standard deviation and population proportion, sample proportion respectively.

To understand the terms parameters and statistic in a better way, let us consider an example.

Assume that we want to draw inferences regarding the accuracy of the quantity of Milk being packed by a leading milk processing company in Western Region of India, AMUL, in 500 ml. packets.

Here the population consists of all milk packets of 500 ml. packets by AMUL company in a day. By finding out the mean of the measurements obtained from all the packets of a day's production, we get population mean. We pick up a random sample of size say n, and take the measurements. We denote the measurements as $x_1, x_2, x_3, \ldots, x_n$. Note that these measurements vary with different samples. We use $X_1, X_2, \ldots, X_n$ to denote the random variables whose particular observations are $x_1, x_2, \ldots, x_n$.
To determine the average quantity of milk being packed, we calculate the mean of the random observations, $x_1, x_2, x_3, \ldots, x_n$. This is the sample mean for this sample. We denote this mean by $\bar{x}$. $\bar{x}$ is a particular value of $\bar{X}$, where $\bar{X}$ is

$$\bar{X} = t(X_1, X_2, \ldots, X_n)$$

From the above discussion, we observe that the sample statistic, like sample mean, is a random variable. Next we shall study these random variables in detail.

4.3 WHAT IS A SAMPLING DISTRIBUTION

Let us start with situation of 'quantity of milk' discussed in the above section.

Now to study the average quantity of milk, suppose we decide that we take a sample of 10 packets without replacement and observe the quantity of milk in each packets. So here the sample size is **n=10**. The observed values are given in Column 2 Table 1 in the next page.

The mean of these 10 observations $\bar{x}_1$ is 496.01 for example 1. That is the average of the quantities of milk obtained from the sample is 496.01. Now, if we are using this sample to make judgements about the population, then we say that, the sample mean value is 496.01 gives an estimate of the average quantity of milk for the whole population.

Table 1

No.	Quantity of Milk in mls	
	Sample 1	Sample 2
1	502	501
2	501	493.9
3	499.5	499.6
4	501.05	490.03
5	499.05	500.09
6	497.56	500
7	501.06	500
8	459.3	499.3
9	499.6	497.5
10	500	502.09

But, if we take another sample from the population and observe the values as given in Column 3 Table 1, the mean of these values is 498.35 for sample 2. This is different

from the mean of the first sample. Now, if we are using this sample to make inference about the population, then we get a different estimate of the average quantity of milk for the whole population. Like this we can take many different samples of size 10 and each case we get sample means which may or may not be distinct. From all these values we try to estimate the mean of the whole population. In this case it is the average amount of milk in any carton.

To give you a better understanding of generalising from sample statistic to the value of the parameter, let us look at the following example in which the size of the population is very small.

Example 1: Suppose we have a population of N=4 incomes of four business firms and we want to find the average return of these firms. The incomes (in Lakhs) are 100, 200, 300 and 400.

We first note that in this case the (population) mean income is 250 lakhs. Now we use this situation to illustrate how sample means differ from the population mean.

Suppose we select a sample of $n = 2$ observations in order to estimate the population mean μ. Now, there are $C(4, 2) = 6$ possible samples of size 2 and we will randomly be selecting one sample from this. We shall now calculate the means of these 6 different samples. These six different samples and their means are given in the following table.

Table 2

Sample	Sample elements X_i	Sample means $\overline{X}$
1	100,200	150
2	100,300	200
3	100,400	250
4	200,300	250
5	200,400	300
6	300,400	350

* * *

Now, from the table above, you can find that each sample has a different mean, with the exception of third and fourth samples. Therefore four of the six samples will result in some error in the estimation process. This sampling error is the difference between the population mean μ and the sample mean we use to estimate it.

Let us now consider the possible sample means and calculate with their probability. We assume that each sample is equally likely to be chosen. Then the probability of selecting a sample is $\frac{1}{6}$

Then we list every possible sample means and their respective possibilities in a table. (See Table 3 given below).

Table 3

Sample mean $\overline{X}$	Number of samples yielding $\overline{X}$	Probability $P(\overline{X})$
150	1	1/6
200	1	1/6
250	2	2/6
300	1	1/6
350	1	1/6
		Total 1

The table obtained above is called the sampling distribution of mean.

Definition 3: A list of all possible values for a sample statistic and the probability associated with each value is called a **sampling distribution of the statistic**.

In the above example if we take the sample mean $\bar{x}$ as the random variable, then Table 3 is nothing but the probability distribution of the means. That means, here the observed values are the means. Like any other list of numbers, these sample means have a mean. It is called **'mean of the sample means'** or the **grand mean**. The mean of the sample means is calculated in the usual fashion: the individual observations are summed, and the result is divided by the number of observation. Therefore if $\bar{\bar{X}}$ denotes this mean, then we have

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

where k is the number observation (samples). For the given situation of incomes, we can calculate $\bar{\bar{X}}$ from table as

$$\bar{\bar{X}} = \frac{150 + 200 + 250 + 250 + 300 + 350}{6}$$
$$= 250$$

Notice that this equals the population mean $\mu = 250$. This is no coincidence. The grand mean $\bar{\bar{X}}$ will always equal the population mean. Thus we have arrived at an important observation.

If we were to take every possible sample of size n from a population and calculate each sample mean, the mean of those sample means would equal the population mean. That means $\bar{\bar{X}}$ is the population mean.

Now, why don't you calculate the mean $\bar{\bar{X}}$ for the income problem discussed, in Example 1, by taking samples of size 3 (see E4).

E4) Construct the sampling distribution for the income problem discussed in Example 1, by taking samples of size 3. Calculate the grand mean and compare it with the population mean.

After doing E4, you must have noticed that in the case of samples of size 3 also, we get that the mean of the sample means is equal to the population mean.

Note: You should not confuse n, the number of observations in a single sample, with k, the number of possible samples. In the situation of 4 incomes, the sample size is $n = 2$, while the number of possible samples is $k = {}^4C_2 = 6$.

Let us do a problem now.

Problem 1: A Statistics professor has given five tests. A student scored 70, 75, 65, 80 and 95 respectively in five tests. The professor decides to determine his grade by randomly selecting a sample of three test scores. Construct the sampling distribution for this process. What observations might you make?

Solution: There are ${}^5C_3 = 10$ possible samples. The samples and their means are given in Table 4.

The sampling distribution is given in Table 5. Note that the population mean is $\mu = 77$. None of these samples gives the mean as 77. Five of the ten possible samples produce values of the sample mean $\bar{X}$ in excess of the population mean, while the other five samples underestimate it.

Table 4

Sample number	Sample elements X_i	Sample mean $\bar{X}$	Sample number	Sample elements X_i	Sample mean $\bar{X}$
1	70,75,65	70.0	6	70,80,95	81.7
2	70,75,80	75.0	7	75,65,80	73.3
3	70,75,95	80.0	8	75,65,95	78.3
4	70,65,80	71.7	9	75,80,95	83.3
5	70,65,95	76.7	10	65,80,95	80.0

Table 5

$\bar{X}$	$P(\bar{X})$
70.0	1/10
71.7	1/10
73.3	1/10
75.0	1/10
76.7	1/10
78.3	1/10
80.0	1/10
81.7	1/10
83.3	1/10
	1.00

You can try some exercises now.

E5) The ages of six executives of a company are

Name	Age
Mr. Ravi	54
Mrs. Veena	50
Mrs. Shanti	52
Mr. Suresh	48
Mr. Rajiv	50
Mr. Anil	52

i) How many samples of size 2 are possible?

ii) Construct the sampling distribution of means by taking samples of size 2 and organise the data.

iii) Calculate the mean of the sampling distribution and compare it with the population mean.

As we observed in the example above , it will be interesting to note how much these sample means vary from the population mean. From your knowledge of frequency distribution (see Unit 1, Block 1), you may think that this variance can be obtained by calculating the variance of the sample means. You are right. The sample means also have a variance. It measures the dispersion of the individual observations (sample means) around their population means. Furthermore this variance is calculated like any other variance. We can obtain this by performing the following.

1) the amount by which each of the observations (sample means) differs from the population mean.

2) squaring these deviations.

3) dividing the squared deviations by the number of sample means, k,

We denote by $\sigma_{\bar{X}}$, the standard deviation of the sampling distribution of sample means. Then we have

$$\sigma^2_{\bar{X}} = \frac{\sum(\bar{x}_i - \bar{\bar{X}})^2}{k} \quad (1)$$

Note that $\bar{\bar{X}}$ is the mean of the sampling distribution which is the same as the population mean.

Now that we have seen too many "standard deviations. One is the standard deviation of the entire population which we usually denote by σ. then the standard deviation of a single sample, and now we have the standard deviation $\sigma_{\bar{X}}$ of an entire set of sample means. **Since $\sigma_{\bar{X}}$ measures the dispersion of the sample means around μ, it gives a measure for the error in sampling. Because of this, $\sigma_{\bar{X}}$ is called sampling error or standard error. We formally define this now.**

Definition 3: The standard error is the standard deviation of the sample means around the population mean and is denoted by SE $(\bar{x})$.

Let us now calculate the standard errors of the sampling distribution of mean obtained in Problem 1.

Problem 2: Compute the standard error of the sampling distribution of sample scores obtained in Problem 1.

Solution: Here

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k} = 77$$

$$\sigma^2_{\bar{X}} = \frac{\sum_{i=1}^{k} \left(\bar{x}_i - \bar{\bar{X}}\right)^2}{k}$$

$$\sigma^2_{\bar{X}} = \frac{(70-77)^2 + (75-77)^2 + (80-77)^2 + (71.7-77)^2 + \ldots + (80-77)^2}{10}$$

$$= 17.6$$

$$\sigma_{\bar{X}} = \sqrt{17.6} = 4.20$$

This shows that the mean of all 10 possible sample means is 77. The square root of the average square deviation of the sample means from the population mean of 77 is 4.2.

———— × ————

You will study later in Unit 5 that the standard error of any statistic gives us an idea of how good a statistic is in estimating the parameters.

You may have realised that the computation of the standard deviation from the sampling distribution is a tedious process. There is an alternative method to compute standard error of the means, SE$(\bar{x})$, from a single sample if we know the population standard deviation. By this method, we have the following formula for obtaining the standard error of the mean for a finite population.

$$SE(\bar{x}) = \sqrt{\frac{(N-n)}{N-1}\frac{(\sigma^2)}{n}}$$

where **N** = population size = the total number of individuals, **n** = sample size = number of individuals selected in the random sample. σ = standard deviation of the individuals in the population.

We will not derive the formula here since the process is too technical for the scope of this course. The factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor. As a rule of thumb, when $\frac{n}{N}$ is less than 0.1, this correction factor can be ignored. We use the above formula for computing SE($\bar{x}$) when **N-n** is not very large. When **N** is large, relative to **n**, we use the formula,

$$SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This formula can also be used for calculating SE($\bar{x}$) for infinite population.

Let us see an example.

Problem 3: The U.S. Bureau of census wishes to estimate the birth rates per 1,00,000 people in the nation's largest cities. It is known that the standard deviation in the birth rates for these 100 urban centres is 12 births per 1,00,000 people. Then

(a) calculate the variance and standard error of the sampling distribution of i) $\mathbf{n} = 8$ cities ii) $\mathbf{n} = 15$ cities.

(b) compare the values obtained in both the cases.

Solution:

(a) i) Here $\mathbf{N} = 100$ and $n = 8$ and the population variance is 12. Therefore $\frac{n}{N}$ is less than 0.1. Then we use the formula for calculating the variance as

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{12^2}{8} = 18$$

and the standard error is $\sigma_{\bar{x}} = \sqrt{18} = 4.24$

ii) In this case $\mathbf{N} = 100$ and $\mathbf{n} = 15$ and therefore $\frac{n}{N}$ is greater than 0.1. Also $\sigma = 12$. Therefore we use the formula for variance as

$$\begin{aligned} \sigma^2_{\bar{x}} &= \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \\ &= \left(\frac{100-15}{100-1}\right)\left(\frac{12^2}{15}\right) \\ &= 8.24 \\ \text{Hence, } \sigma_{\bar{x}} &= \sqrt{8.24} = 2.87 \end{aligned}$$

(b) On comparing both the values we observe that, the larger sample has a smaller standard error and will tend to result in less sampling error in estimating the birth rates in the 100 cities.

———— ×————

Here is an exercise for you.

E6) From a population of 200 observations, a sample of n=50 is selected. Calculate the standard error if the population standard deviation equals 22.

Thus we have seen that we can compute the standard error of the sample means if we know the population standard deviation. Here you can note one thing. Usually, we do not know σ. But it is possible to estimate σ from the sample. We will talk about this later in the next section. Using that estimate we compute the standard error by the shortcut formula given earlier.

Now we summarise our discussion about mean and and the standard deviation of a sampling distribution. We state the following Theorem.

Theorem 1: If a random sample of size **n**, say $X_1, X_2, \ldots, X_n$ is taken from a population having the mean μ and variance σ^2, then $\overline{X}$ = the mean of $(X_1, X_2, \ldots X_n)$ is a random variable whose distribution has the mean μ. For samples from infinite populations the variance of the distribution is σ^2/n, for samples from finite population of size N, the variance is $\frac{\sigma^2}{n}\frac{N-n}{N-1}$

So far we have been discussing about the mean and standard deviation of the sampling distribution. Next we shall see some interesting properties regarding the shape of the sampling distribution. For that let us consider the following situation.

Suppose in an experiment 50 random samples of size **n** = 10 are taken from a population having discrete uniform distribution.

Sampling is with replacement, so that we are sampling from an infinite population. The sample means of these 50 samples are given below.

4.4	3.2	5.0	3.5	4.1	4.4	3.6	6.5	5.3	4.4
3.1	5.3	3.8	4.3	3.3	5.0	4.9	4.8	3.1	5.3
3.0	3.0	4.6	5.8	4.6	4.0	3.7	5.2	3.7	3.8
5.3	5.5	4.8	6.4	4.9	6.5	3.5	4.5	4.9	5.3
3.6	2.7	4.0	5.0	2.6	4.2	4.4	5.6	4.7	4.3

We group these means into a distribution with the classes 2.0-2.9, 3.0-3.9, ..., and 6.0-6.9, then we get the following table.

Table 6

$\bar{x}$	Frequency
2.0-2.9	2
3.0-3.9	14
4.0-4.9	19
5.0-5.9	12
6.0-6.9	3
	50

It is clear from this distribution as well as its histogram shown in Figure 1 that the distribution of the means is fairly **bell-shaped** even though the population itself has a discrete uniform distribution.

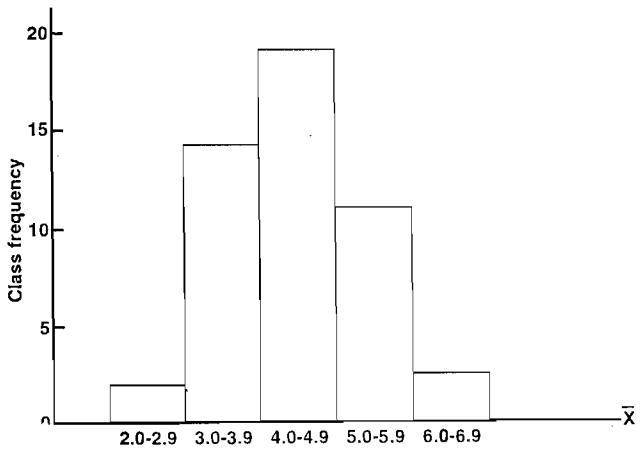


Fig. 1

This raises the question whether our result is typical of this population or any other!!

Let us consider the other situation which was discussed in the earlier section (where we want to estimate the average quantity of Milk).

Suppose in that experiment we are taking 10 random samples having size n=10 (In Block 4 we shall see how we can select random samples). The following table illustrates the 10 samples.

Table 7

1	2	3	4	5	6	7	8	9	10
502.0	501	493.9	497.8	502.09	502.09	502.9	493.9	459.3	497.3
501.0	493.9	499.6	499.3	499.6	501	501.9	493.09	497.56	493
499.5	499.6	502	499	490.9	500.09	501	499	493.0	499
501.05	490.03	501.3	501	496.8	500	500	499.6	499.05	499
499.05	500.09	500	502.9	500	500	500	499.9	499.5	499.3
497.56	500	499.9	501.9	503	499.3	499.3	500	499.6	500
501.06	500	493.99	500	502	499.6	499	500	500.0	500
459.3	499.3	500	499	501	493.9	499	501.3	501.0	501
499.6	497.5	502	500	493.9	497.5	493	502	501.03	501.9
500.0	502.09	499	493	503	490.03	497.8	502	501.06	502.9

If we plot these samples means, then we get the following graph.

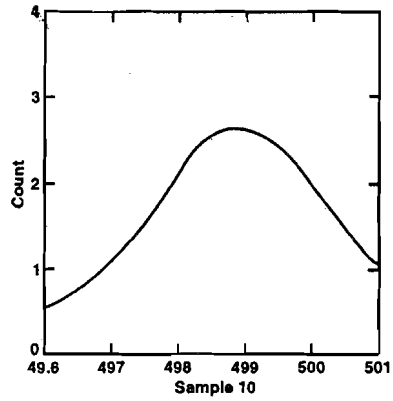


Fig.2

Here also you can observe that it is approximately bell-shaped.

Note that unlike in the earlier situation, in this case we do not know the distribution of the population.

Does the above observations give an indication that irrespective of the nature of the distribution of the population, the sampling distribution of means is approximately normal? This idea will be more clear to you if you look at the figure given at end of the unit (see Appendix 1). The figure shows graphs of the distribution function of 4 different populations see Fig. (a). The graphs in each of the figures in (b), (c) and (d) shows the sampling distribution for the sample sizes $n = 2, n = 5$ and $n = 30$ for the respective population. You note that the parent distribution for population 4 is normal and sampling distributions are also normal. This is not surprising. But the surprising fact is that for population 1, 2 and 3 parent distributions are not normal, still, as the sample size increases, the sampling distribution in each case approaches a normal distribution. The remarkable normality of the distribution of the sample mean is the substance of the famous central limit theorem. We state a simple version of the theorem.

Theorem 2: Central Limit Theorem : - $X_1, X_2, \ldots, X_n$ be n independent and identically distributed random variables having the same mean μ and variance σ. Let $S_n = X_1 + X_2 + \ldots + X_n$. Then $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ has approximately the standard normal distribution $(0, 1)$.

Now we shall illustrate this theorem in the case of sampling distribution of means $(\overline{X})$.

Suppose we take a sample. Let $X_1, X_2, \ldots X_n$, are independent random variables by the sampling scheme and also they are identically distributed. X_i's have the same mean $\mu < \infty$ and variance $\sigma^2 < \infty$ (also called the population mean and population variance) and therefore we can apply the central limit theorem to $X_1, X_2, \ldots, X_n$ and

conclude that $\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$ is distributed approximately as standard normal. But note that $\frac{X_1 + \ldots + X_n}{n}$ is nothing but the sample mean $\overline{X}$, therefore according to the central limit theorem we have $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ follows standard normal distribution $N(0, 1)$.

So far we have been discussing the sample means and their distribution. Our interest has been in some variable which might be measured and averaged. However there are many instances where we may need some other statistic, other than mean, to make inference about the population. In the next section we shall discuss such a statistic known as 'proportion' and discuss its sampling distribution.

Sampling Distribution of Proportion

There are many instances where we are interested in determining the number or proportion of observations falling in a particular category. For example, a doctor wants to know how many of his patients can survive after administering a particular drug. A politician may want to know how many voters will not vote for him. In all these situations there are two possible outcomes of the observation - whether an observation falls in a particular category or not. You might recall that such situations are related to the problems dealing with binomial distribution. The politician may not be interested in the actual number of people who are going to vote for him, she is interested on what percentage of the people will do so. In such a situation we deal with sample proportions instead of sample means. The population parameter in this case is the proportion. We denote this by π. In general, for a finite population, we define the population proportion π as

$$\pi = \frac{k}{n}$$

where **k** is the number of observations that fall in a particular category and **n** is the total number of observation. When the population is very large, we may take samples to study the population and for each sample we calculate the sample proportion, **p**, as

$$p = \frac{s}{n}$$

where **s** denotes the number of observations in the sample which meet the particular characteristic, under study and **n** is the sample size.

For example, assume that a politician surveys 1,000 voters and finds that only 350 are going to vote for him. Then

$$p = \frac{350}{1000} = 0.35$$

You note that, a different sample of n = 1,000 voters may yield a different p. If we calculate the possible sample proportions then a **list** of these observations is called the **sampling distribution of proportions**.

Let us consider an example. As, in the case of sample means, we are considering a simple situation where the population size is very small.

Example 2: In our discussion about the quantification of milk packets, let us put the condition that if the quantity in any packet measured is less than 495ml, it will be rejected by the consumer, hence call them as defective and other wise they are non-defective.

For the data given in Table 7 we consider the first five columns (five cartons) as population then the population has 50 milk packets.

From the population we get the proportion of defective as

$$\text{P(Defective)} = \pi = \frac{7}{50} = 0.14$$

and the proportion of non-defective is

$$\text{P(non-defective)} = 1 - \pi = 0.86$$

Let us try to use the sampling method to find this proportion. Assume that random samples of size 2 out of 5 (that is two cartons out of five cartons) have been selected and the possible samples and their defective proportions are as given below

Table 8

Sample	Sample proportions (p_i)
C_1, C_2	2/20
C_1, C_3	3/20
C_1, C_4	2/20
C_1, C_5	3/20
C_2, C_3	3/20
C_2, C_4	2/20
C_2, C_5	3/20
C_3, C_4	3/20
C_3, C_5	4/20
C_4, C_5	3/20

Let us calculate the mean of the sample proportion $\bar{p}$

$$\bar{p} = \frac{\Sigma p_i}{k}$$

$$= \frac{28}{20 \times 10}$$

$$= .14$$

This is same as the population proportion.

* * *

From the earlier example we conclude the following fact:

The mean of the sampling distribution of proportions is equal to the population proportion

Now the standard error of the sample proportions, by definition, is the standard deviation of this sampling distribution. As we have mentioned earlier, in the case of mean, the computation of standard deviation from the table is a tedious process. So we make use of a short cut formula by which we can compute the **standard error** which we **denote by SE(p)**. If we know the population proportion, population size and sample size, the formula is

$$SE(p) = \sqrt{\frac{(N-n)\pi(1-\pi)}{(N-1).n}}$$

where π is the population proportion and N and n are the population size and sample size, respectively. If the population in infinite or if the size of the population, relative to the sample size is extremely large, then

$$SE(p) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Let us now calculate the standard error for the situation in Example 3. In this case $\pi = 0.4, N = 5$ and n=2. Therefore,

$$\begin{aligned} SE(p) &= \sqrt{\frac{(0.14)(0.86)}{2}}\sqrt{\frac{5-2}{5-1}} \\ &= 0.2454 \times 0.866 \\ &= 0.2125. \end{aligned}$$

Here is an exercise for you.

E7) a) In a manufacturing process of electric bulbs, company gives an output of 250 electric bulbs per day. To test the defectiveness of the manufactured bulbs, a random sample of 50 has selected in a particular day's production. Inspection of the samples showed 4 as defective and the rest as non-defective. Calculate the standard error of sample proportion of defective.

b) If the process is considered to be a continuous one (observed for many days) and a random sample of 60 has been selected from the produced electric bulbs and observed that 7 are defective. What will be the standard error of sample proportion of defective.

So far we have discussed sampling distributions of two statistics $\bar{x}$ and $\bar{p}$. We also emphasised the role of Central limit theorem and the idea that distributions of $\bar{x}$ and p, will be approximately normal even when the data in the parent population are not distributed normally, provided the sample sizes are large.

But there are many situations where we may have to deal with small samples, may be due to limited availability of items or by other factors such as time of cost. We may also have to deal with certain statistics other than $\bar{x}$ and p. For example in some situations we may be interested in the distribution of $\bar{x}/s$, where s is the sample standard deviation. In some other situations, we may be interested in the distribution of $\sum_{i=1}^{n}(x_i - \bar{x})^2$. The important exact sampling distributions are chi-square, student t- and F-distributions. These distributions are widely used in statistical analysis. In the following sections we shall discuss these distributions.

4.4 t-DISTRIBUTION

William G. Gosset, a brew master for Guiness Brewarier, developed a family of distributions each of which corresponds to a parameter ν, (called nu-a positive integer). The density function of a random variable whose distribution belongs to this family is given by

$$f_\nu(y) = \frac{1}{\Gamma(\pi\nu)} \frac{\Gamma(\nu + 1/2)}{\Gamma(\nu/2)} \left(1 + \frac{y^2}{\nu}\right)^{-\nu/2}$$

where $\Gamma(x)$ for any $x > 0$ is called the gamma function. The values of $\Gamma(x)$ for different values of $x > 0$ are tabulated. The three members of the family of this distribution for $\nu = 9$, $\nu = 14$ and $\nu > 30$ are shown in the accompanying figure.(See Fig. 3 in the next page.)

Now we shall illustrate the importance of this distribution as distribution in the case of sampling distribution.

Suppose we take a sample from a population having normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. If $X_1, X_2, \ldots, X_n$ are n observations in the sample, then $X_1, \ldots, X_n$ are independently and identically distributed(i.i.d) random variables. and

we have $\frac{\overline{X} - \mu}{s/\sqrt{n}}$ has t–distribution with parameter $\nu = n - 1$, where s is the sample variance. The parameter ν is called the number of degrees of freedom (in short d.f) of the distribution.

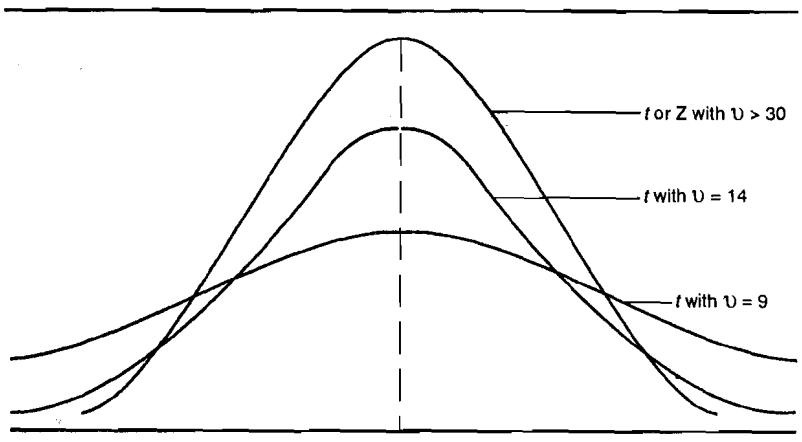


Fig. 3 t-distributions

Now that, you have been introduced to the distribution, we shall see how we use this distribution to make judgements about the population.

As we have seen in the case of normal distribution (see Unit 3, Block 1), there are tables available using which we can calculate probability for this distribution. One such table is given in the appendix (see Table 2 in appendix of this block). The table contains selected values of t_α for various values of ν, where t_α is such that the area under the t-distribution to its right is equal to α. In this table the left-hand column contains values of ν, the column headings are areas α in the right-hand tail of the t-distribution, and the entries are the values of t_α(See Fig. 4 below)

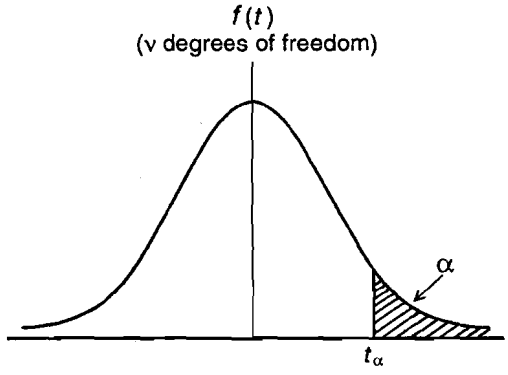


Fig. 4: Tabulated values of t

It is not necessary to tabulate the values of t_α for $\alpha > 0.50$, as it follows from the symmetry of the distribution that $t_{1-\alpha} = t_\alpha$; thus, the value of t that corresponds to a left-hand tail area of α is t_α. You note that the bottom row of the table the entries are the same as the values of the standard normal variate 2. For example $t_{0.025} = 1.96 = z_{0.025}$ for large values of ν i.e. $\nu \geq 30$.

Remark: Please note that the tables for the distribution can vary from book to book because of the definitions of parameters involved. Therefore, for the purposes of this course, please only refer to the table given at the end of the block.

Example 3: The graph of a t-distribution with 9 degrees of freedom is given below. Let us find the values of t_1, for which

i) the shaded area of the right = 0.05

ii) the total shaded area = 0.05

iii) the total unshaded area = 0.99

iv) the shaded area on the left = 0.01

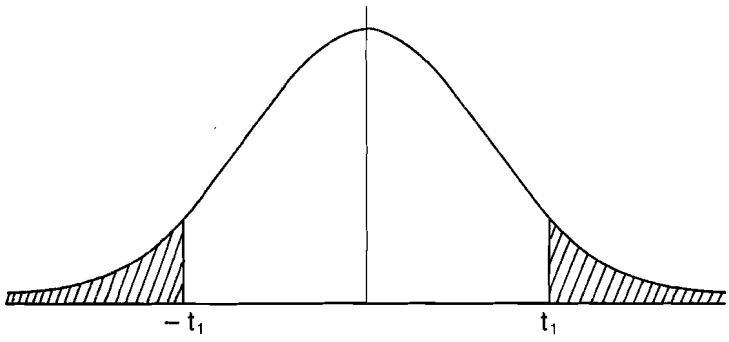


Fig. 5

Let us try (i) to (iv) one by one.

i) If the shaded area on the right is 0.05, comparing the figure above with Fig.4, we get that $\alpha = 0.05$. It is already given that $\nu = 9$. Referring to the table given in the Appendix, we proceed downward under column headed ν until entry 9 is reached. Then proceed right to the column headed 0.05. and get the required value of t as 1.833.

ii) Given that the total shaded area is 0.05. Then by symmetry, the shaded area on the right is $\frac{0.05}{2} = 0.025$ i.e. $\alpha = 0.025$. Also $\nu = 9$. Therefore from the table we get the required value as 2.262.

iii) If the total unshaded area is 0.99, then the total shaded area is $1 - 0.99 = 0.01$. Therefore the shaded area on the right of t_1 is $\frac{0.01}{2} = 0.005$. Then we get the required value from the table as 3.250.

iv) If the shaded area on the left is 0.01, then by symmetry the shaded area on the right is also 0.01. Then from the table we get the value as 2.821

* * *

Problem 4: For the given sample sizes and the t-values, find the corresponding probability α

i) $n = 26, \quad t = 2.485$

ii) $n = 14, \quad t = 1.771$

Solution:

i) Referring to the t-table given in the Appendix, we find the row corresponding to $\nu = 25$ and look for the entry $t = 2.485$. Then we observe that this t-value lies in the column headed $t_{0.01}$. This gives that the corresponding α-value is 0.01.

ii) Similarly you can see that α-value corresponding to $\nu = 13$ and $t = 1.771$ is 0.05.

———— ×————

Example 4: A manufacturer of fuses claims that with a 20 % overload, his fuses will blow in 12.40 minutes on the average. To test this claim, a sample of 20 of the fuses was subjected to a 20 % overload, and the times it took them to blow had a mean of 10.63 minutes and a standard deviation of 2.48 minutes. if it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim? Let us see how we can make use of the t-distribution to find out an answer for this.

We first note that the sample is taken from a normal population and the size is small, n = 20. Then as we have stated in the discussion above, the random variable $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ follows t-distribution with n − 1 degrees of freedom.

Here $\bar{x} = 10.63, \mu = 12.40$ and $s = 2.48$ and $n = 20$. Therefore,

$$t = \frac{10.63 - 12.40}{2.48/\sqrt{20}} = -3.19$$

Now we compare this t-value with the values of t given in the table for the parameter, $\nu = 20 - 1 = 19$. We look for the t-value which is nearest to 3.19 and less than 3.19. Then we find that the required t-value is 2.861. Also from the table, we find that the α-value corresponding to 2.861 is 0.005. That means the probability that t will exceed 2.861 is 0.005. Now we note that −3.19 is less than −2.861 therefore the α-value corresponding to −3.19 will be less than 0.005, which is very small (see Fig.6).

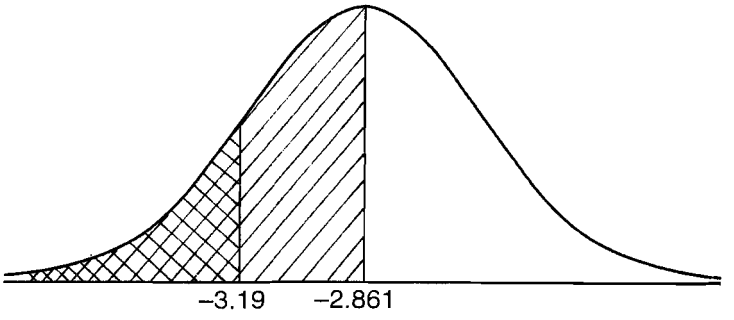


Fig. 6

Hence, we can conclude that the data tend to refute the manufacturer's claim.

* * *

Try these exercises now.

E8) Find the values of t for given values of ν and α

i) $\alpha = 0.01$ and $\nu = 19$

ii) $\alpha = 0.05$ and $\nu = 6$

iii) $\alpha = 0.025$ and $\nu = 23$

iv) $\alpha = 0.1$ and $\nu = 10$

v) $\alpha = 0.005$ and $\nu = 29$

E9) A process for making certain bearings is under control if the diameters of the bearings have a mean of 0.5 cm. What can we say about this process if a sample of 10 of these bearings has a mean diameter of 0.5060 cm. and a standard deviation of 0.0040 cm. Assume that the data constitute a random sample from a normal population.

E10) Find the values of t for which the area of the right-hand tail of the t-distribution is 0.05 if the number of degrees of freedom is equal to
(a) 16, (b) 27, (c) 200

Note that for applying t-distribution we made the assumption that the samples should come from a normal population and population standard deviation is not known. Studies have been shown later that the first assumption can be relaxed. It has been shown that the distribution of random variable with the values $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ is fairly close

to a t-distribution even for samples from certain non-normal population. In practice, it is necessary to make sure primarily that the population from which we are sampling is approximately bell-shaped and too skewed.

Next we shall consider another exact sampling distribution.

4.5 CHI-SQUARE DISTRIBUTION

Another exact sampling distribution which is very useful in statistical problems is the chi-square distribution. In this section we shall introduce you to this distribution.

We say that a random variable Y has χ^2 (called chi-square) distribution with ν degrees of freedom if the density function $f_Y(y)$ of Y is given by

$$f_Y(y) = \frac{y^{(\nu/2-1)}e^{-y}}{\Gamma(\nu/2)}, \text{if } y > 0$$
$$= 0, \text{if } y \leq 0$$

where $\Gamma(x)$ is the gamma function.

Note that this family of distributions is parametrised by ν, called the degrees of freedom. The graph of the density function for some members of this family ($\nu = 1, 3, 8, 10$) are shown in Fig.7 below

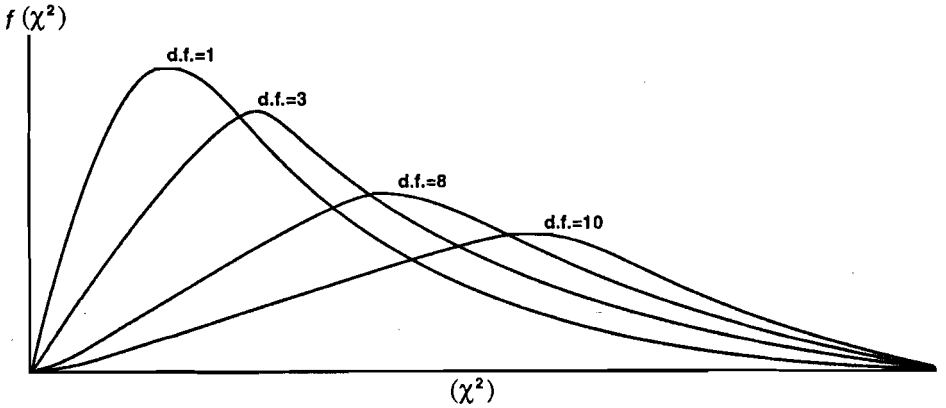


Fig.7 Various Chi-square distributions

As in the case of normal and t-distributions, a table containing selected values of χ^2_α for various values of υ, again called the number of degrees of freedom, is given in the Appendix (see Table 3). χ^2_α is the value such that the area under the chi-square distribution to its right is equal to α. (i.e. the probability that any value is greater than equal to or χ^2_α is α). In this table the left-hand column contains values of υ, the column headings are areas α in the right-hand tail of the chi-square distribution, and the entries are the values of χ^2_α. (see Fig.8 below).

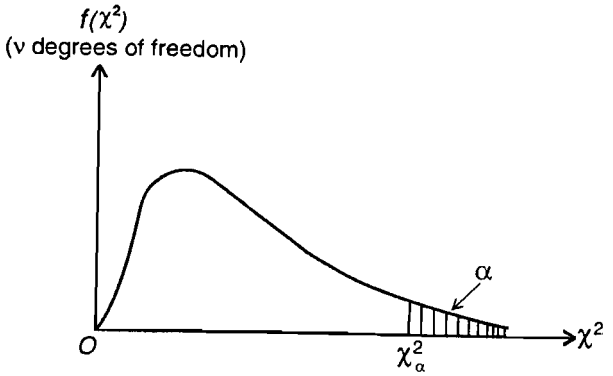


Fig.8

Note that unlike t-distribution, chi-square distribution is not symmetrical. Therefore, it is necessary to tabulate values of χ^2_α for $\alpha > 0.50$ also.

Let us see some examples

Example 5: Let us find the values of χ^2_α of the χ^2-distribution with 5 degrees of freedom for $\alpha = 0.05$ and 0.01.

Let us first consider $\alpha = 0.05$. To find the values of χ^2_α for 5 degrees of freedom and $\alpha = 0.05$, we refer to Table 3 in the Appendix. We proceed downward under column headed ν until entry 5 is reached, then proceed right to column headed $\alpha = 0.05$ which gives the required value as $\chi^2_\alpha = 11.070$.

Similarly for $\alpha = 0.01$, we get the required value as $\chi^2_\alpha = 15.056$.

* * *

Now we shall illustrate the importance of χ^2-distribution in the case of sampling distribution.

Suppose we take a random sample of size n when n is small ($n \leq 30$) from a normal population having the variance σ^2 and if s^2 is the variance of the random sample, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

is a value of a random variable having χ^2-distribution with $\nu = n - 1$.

Let us see how we use this result for solving the following problem.

Problem 5: An optical firm purchases glass to be ground into lenses, and it knows from past experience that the variance of the refractive index of this kind of glass is $1.26.10^{-4}$. As it is important that the various pieces of glass have nearly the same index of refraction, the firm rejects such a shipment if the sample variance of 20 pieces selected at random exceeds $2.00.10^{-4}$. Assuming that the sample values may be looked upon as a random sample from a normal population, what is the probability that a shipment will be rejected even though $\sigma^2 = 1.26.10^{-4}$?

Solution: We first note that the sample is taken from normal population and the size is n small, n = 20. Then as (we stated in the discussion above), the random variable $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ follows χ^2-distribution with $(n - 1)$ degrees of freedom.

Here $\sigma^2 = 1.26 \times 10^{-4}, s^2 = 2 \times 10^{-4}$ and n = 20. Therefore

$$\chi^2 = \frac{(10-1) \times 2 \times 10^{-4}}{1.26 \times 10^{-4}} = 30.2.$$

Now we compare this χ^2-value with the values of χ^2 given in the table for the parameter $\nu = 20 - 1 = 19$. Then we find that the χ^2-value 30.1 is close to the calculated value 30.2 and less than the calculated value. The corresponding α-value is 0.05. That means that the probability that χ^2-value exceeds 30.1 is 0.05. Therefore we can conclude that the probability that a good shipment will erroneously be rejected is less than 0.05.

———— ×————

Here is an exercise for you.

E11) Find the values of χ^2 for which the area of the right-hand tail of the χ^2-distribution is 0.5, if the number of degrees of freedom ν is equal to (a) 15, (b) 21.

E12) A random sample of 10 observations is taken from a normal population having the variance $\sigma^2 = 42.5$. Find the approximate probability of obtaining a sample standard deviation between 3.14 and 8.94.

In the next section we shall discuss another important distribution

4.6 F-DISTRIBUTION

R.A. Fisher, a British statistician, developed this distribution in the early 1920's. It is defined as follows.

Definition : If a random variable Y_1 has a χ^2-distribution with ν_1 degrees of freedom, and if a random variable Y_2 has a χ^2-distribution with ν_2 degrees of freedom and if Y_1 and Y_2 are independent, then $Y_1/\nu_1 \div Y_2/\nu_2$ has an F-distribution with ν_1 and ν_2 degrees of freedom.

In other words, the ratio of two independent χ^2 random variables, each divided by its number of degrees of freedom, is an F random variable.

Like the t and χ^2 distributions, the F distribution is in reality a family of probability distributions, each corresponding to certain numbers of degrees of freedom. But unlike the t and χ^2 distributions, the F distribution has two numbers of degrees of freedom, not one. Figure shows the F distribution with 2 and 9 degrees of freedom. As you can see, the F distribution is skewed to the right. However, as both numbers of degrees of freedom become very large, the F distribution tends toward normality. As in the case of the χ^2 distribution, the probability that an F random variable is negative is zero. This must be true since an F random variable is a ratio of two nonnegative numbers. (Y_1/ν_1 and Y_2/ν_2 are both nonnegative.) Once again, it should be emphasised that any F random variable has two numbers of degrees of freedom. Be careful to keep these numbers of degrees of freedom in the correct order, because an F distribution with ν_1 and ν_2 degrees of freedom is not the same as an F distribution with ν_2 and ν_1 degrees of freedom.

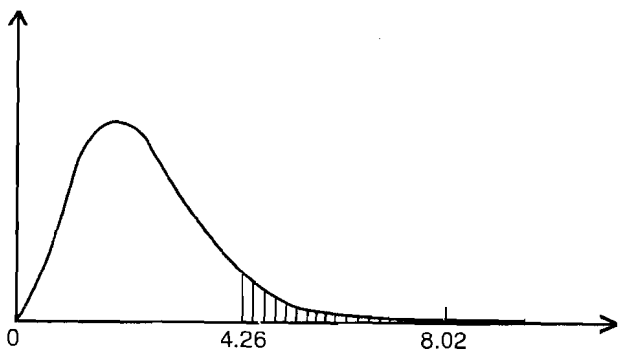


Fig.9

Tables are available which show the values of F that are exceeded with certain probabilities, such as .05 and .01. Appendix Table 9 shows, for various numbers of degrees of freedom, the value of F that is exceeded with probability equal to .05. For example, if the numbers of degrees of freedom are 2 and 9, the value of F that is exceeded with probability equal to .05 is 4.26.(See Fig.9.) Similarly, Appendix Table 10 shows, for various numbers of degrees of freedom, the value of F that is exceeded with probability equal to .01. For example, if the numbers of degrees of freedom are 2 and 9, the value of F exceeded with probability equal to .01 is 8.02.(See Fig.9.)

Let us see some examples.

Example 6: A random variable has F distribution with 40 and 30 degrees of freedom, let us find the probability that it will exceed a) 1.79 b) 2.30.

We first consider the value 1.79. We look at table of F-distribution, Table 4 for 0.05. We proceed downward under column headed ν_2 until entry 30 is reached, then proceed right to the column headed $\nu_1 = 40$ which shows the given value 1.79. Therefore we get that the probability is 0.05.

Since the value cannot be in Table 4 we look at Table 5 which is for 0.01. We proceed similarly and find the value 2.30 in the table. Therefore we get that the probability is 0.01.

* * *

We shall now illustrate the importance of this distribution in the case of sampling distribution.

Suppose that we have two independent random samples of sizes n_1 and n_2 respectively, taken from two normal population having the same variance. If s_1^2 and s_2^2 are sample variances, then the function

$$F = \frac{s_1^2}{s_2^2}$$

has F distribution with ν_1 and ν_2 degrees of freedom.

F-distribution is used to compare the equality of two sample variances.

Let us consider the following problem.

Problem 6: If two independent random samples of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population, what is the probability that the variance of the first sample will be at least three times as large as that of the second sample?

Solution: In this we have to compare the variances of the samples. Therefore we apply F-distribution. From Table we find that $F_{.05} = 3.00$ for $\upsilon_1 = 7 - 1 = 6$ and $\upsilon_2 = 13 - 1 = 12$; thus, the desired probability is 0.05.

———— ×————

Why don't you try these exercises now.

E13) A random variable has the F distribution with 15 and 12 degrees of freedom. What is the value of this random variable that is exceeded with a probability of 0.5 with a probability of .01?

E14) If a random variable F has 15 and 23 degrees of freedom, what is the probability that it will exceed 2.93.

E15) If two independent random samples of size $n_1 = 9$ and $n_2 = 16$ are taken from a normal population, what is the probability that the variance of the first sample will be at least four times as large as the variance of the second sample?

E16) If independent random samples of size $n_1 = n_2 = 8$ come from normal populations having the same variance, what is the probability that either sample variance will be at least seven times as large as the other?

In this unit you have only been introduced to these three specific distributions t, χ^2 and F. You will realise the utility of these distribution as you study the following units.

This brings us to the end of this unit. We have discussed the sampling distribution of a statistic at length. Let's now briefly recall the various concepts which we have covered here.

4.7 SUMMARY

In this unit we have explained the following concepts

1) i) Parameter - Values that describe the characteristic of a population.

ii) Statistic - Measures describing the characteristic of a sample.

iii) Random sample - A sample from a population in which all the items in the population have an equal chance of being in the sample.

iv) Sampling distribution of a statistic - For a given population, a probability distribution of all the possible values of a statistic may taken as for a given sample size.

v) Standard error - The standard deviation of the sampling distribution of a statistic.

2) We have discussed the central limit theorem.

3) We have introduced the following three specific distributions
   i) t-distribution
   ii) χ^2-distribution
   iii) F-distribution

   and explained the use of these distributions as sampling distribution.

4.8 SOLUTIONS/ANSWERS

E1) Female students enrolled in IGNOU BDP. Finite population

E2) (ii) and (iii)

E3) Suppose there are 500 employees working in a company and we are interested to find the average salary paid to them in a month, the average amount calculated from the enter employees is called parameter. On the other hand it is possible to estimate the average based on selected few employees. The average income calculated from sample is called statistic.

E4) The sampling distribution is given in the following table:

Table 9

Sample No.	Sample elements	Sample Means
1	100, 200, 300	200
2	100, 200, 400	233 (approximately)
3	100, 300, 400	267 (approximately)
4	200, 300, 400	300

The grand mean is 250 and is equal to the population mean.

E5) i) There are

$$
\begin{aligned}
&{}^6C_2 \\
&= \frac{6 \times 5}{1 \times 2} \\
&= \frac{30}{2} \\
&= 15
\end{aligned}
$$

ii) samples of size 2 are given in the following table.

Table 10

Sample	$\bar{X}_i$
54, 50	52
54, 52	53
54, 48	51
54, 50	52
54, 52	53

Sample	$\bar{X}_i$
50, 52	51
50, 48	49
50, 50	50
50, 52	51
52, 48	50

Sample	$\bar{X}_i$
52, 50	51
52, 52	52
48, 50	49
48, 52	50
50, 52	51

iii) The population mean, the grand mean,
$\mu = 51.00$
$\overline{X} = \frac{\Sigma \overline{x}_i}{k} = 51.000$
The grand mean is equal to the population mean.

E6) $N = 200$

$n = 50$

$\sigma = 22$

$SE(\overline{X}) = \sigma/\sqrt{n} \cdot \sqrt{\frac{N-n}{N-1}}$

$22/\sqrt{50}\sqrt{\frac{200-50}{199}}$

$= 2.7012$

E7) a) $SE(p) = \sqrt{\frac{250-50}{250-1} \times \frac{\frac{4}{50} \times \frac{46}{50}}{50}}$
$= 0.03438.$

b) $SE(p) = \sqrt{\frac{\frac{7}{60} \times \frac{53}{60}}{60}}$
$= 0.04144.$

E8) $t_{0.01} = 3.143$

$t_{0.025} = 2.069$

$t_{0.1} = 1.372$

$t_{0.005} = 2.756$

E9) We find the value of the random variable

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

where $\overline{x} = 0.5060, \mu = 0.5$ and $s = 0.0040$ and $n = 10$. Substituting, we get the value of t as t = 4.7434. From the table we get that the t-value for the parameter $\nu = 9$ and $\alpha = .05$ is 2.26 which is less than the calculate value t = 4.7434. Since the probability is less i.e. 0.05, we conclude that the process is out of control.

E10) From the table we find that corresponding to $\alpha = 0.5$ and the parameter
a) $\nu = 16$ the value is 1.75
b) $\nu = 27$ the value is 1.70
c) $\nu = 200$ is 1.645 (note that this value corresponds to the entry to the last row marked ∞).

E11) From the table in the appendix, we find that the value of χ^2_α for $\alpha = 0.5$ is
a) 24.996 for $\nu = 15$ and
b) 32.671 for $\nu = 21$.

E12) We make use of the formula

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

Note that we are given that $\sigma^2 = 42.5$ and n = 10. Let s_1 and s_2 denote the two sample standard deviations, then $s_1^2 = 3.14$ and $s_2^2 = 8.94$ corresponding χ^2 values are

$$\chi_1^2 = \frac{(n-1) \times 3.14}{42.5}$$

$$= \frac{9 \times 3.14}{42.5} = 0.664$$

$$\text{and } \chi_2^2 = \frac{9 \times 8.94}{42.5} = 1.893$$

To find the required approximate probability, it is enough to find the area α_0 given below.

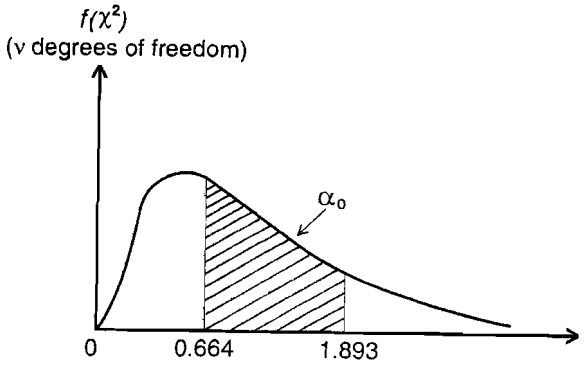


Fig. 10

From the table we find those values of χ^2 for 9 degrees freedom which are close to 0.664 and 1.893.

E13) Table 4 shows that the answer to the first question is 2.62 and Table 5 shows that the answer to the second question is 3.84.

E14) We first look at Table 4 and find that the given value is not there. Then it must be in Table 5 (check it!). Therefore the probability is 0.01.

E15) Solution: Here $n_1 = 9$ and $n_2 = 16 \therefore \nu_1 = 8$ and $\nu_2 = 15$. We have

$$F = \frac{s_1^2}{s_1^2} = 4$$

To find the probability we look for value F = 4 in both the tables and find that it is in Table 5. Therefore the probability is 0.01.

E16) $n_1 = n_2 = 8 \therefore \nu_1 = \nu_2 = 7$ and given that

$$F = \frac{s_1^2}{s_2^2} = 7$$

Proceeding similarly we find the value 6.99 in Table 5 which is almost equal to 7. Therefore the probability is 0.01.

UNIT 5 ESTIMATION

Structure **Page No.**

5.1 INTRODUCTION

In Units 2 and 3 you have seen that populations can be described by distributions which are fully determined with the help of their parameters. For example, in the case of binomial distribution, you need to know **n** and **p**; in a Poisson distribution, you need to know λ; and a normal distribution is determined by μ and σ. These quantities are called parameters. The problem with these parameters is that in real-life situations they are usually unknown. We have seen in Unit 4, that in such situations, we take a random sample from the population and compute a function of the sample values, called statistic. More precisely we try to estimate the population parameters by functions of sample values. In this unit we shall discuss certain methods by which we can estimate the population parameters. These processes are called estimation. As we have already stated, in estimation we expect that the sample value is 'reasonably close' to the population value. How do you judge this? Here we discuss some criteria which tell us how best the parameters can be estimated by sample values.

In this unit we discuss two methods of estimation - point estimation and interval estimation. In Sec. 5.2 we discuss point estimation. Point estimation concerns choosing a statistic, that is a single number calculated from the sample data. In contrast to this, we sometimes obtain an interval in which we can expect the parameter to lie with some degree of confidence. The method of constructing such intervals is called 'interval estimation'. In Sec.5.4, we illustrate construction of such an interval. There we first illustrate how such an interval is constructed for the population mean. We do this in different cases. First we consider the case when the population standard deviation is known and the sample size is large ($n > 30$). Then we take up the case where the standard deviation is unknown, both when the sample size is small and when it is large. After that we shall illustrate how interval estimates are constructed for the population proportion.

Objectives

After reading this unit, you should be able to

- choose an estimator corresponding to a particular situation under study,
- check whether an estimator is,

unbiased
or
efficient.

- construct confidence intervals for the population mean and proportion, using appropriate sampling distribution,
- distinguish between point estimation and interval estimation

5.2 POINT ESTIMATION

Imagine that you need to find the mean life-time of the bulbs produced by a company. Assume that the life of a bulb is distributed as normal with mean θ. Now to find the life-time of a bulb, you have to keep it on till it burns off, and note the time. So, it is a destructive process. If you do this for every bulb produced by the company, it will soon have to close down! The way out in this situation is to take a sample of the bulbs, and try to estimate the average life-time of the population on the basis of the life-time observations obtained from the sample. Of course, we cannot hope to get the exact value of the mean life-time. What we get from the sample is only an estimate. If $x_1, x_2, \ldots, x_n$ are the life-times of the bulbs which were chosen in a sample of size n, then we could take the sample mean, $\frac{x_1, x_2, + \ldots + x_n}{n}$ as an estimate of the population mean. Of course, this estimate will vary from sample to sample. You already have come across this concept in the previous unit.

But apart from the sample mean, there could be other ways of estimating the population mean from the sample. For example, we could take x_1 as an estimate, or we could take $\frac{x_{min} + x_{max}}{2}$ as an estimate where x_{min} is the minimum value and x_{max} is the maximum value.

$\frac{x_1 + x_2 + \ldots + x_n}{n}$ can be written as $\frac{\sum_{i=1}^{n} x_i}{n}$

In any case, the estimate is always based on some or all of the sample values. That is to say that we calculate some sample statistic and take it as an estimate of the population parameter. This sample statistic is called an **estimator**. The value of this estimator for our sample is the **estimate**.

Definition 1: An **estimator** is a function of the sample observations that is used to **estimate** an unknown parameter. A **point estimate** is a single value of an **estimator**. The process by which we choose an estimator and find the point estimate for estimating an unknown parameter is called **point estimation.**

For example, if a sample mean is used to estimate a population mean,and if the sample mean for a particular sample equals 10, then the estimator used is the sample mean, whereas the point estimate is 10.

To cite another example, suppose we are interested in finding the proportion of individuals in India preferring a given soft drink over another. Here the population parameter is proportion. If the sample proportion is used to estimate the population proportion and if the sample proportion for a particular sample equals 0.6, then the estimator used is the sample proportion and the point estimate is 0.6.

Why don't you try an exercise now.

E1) Write the estimator and estimate used in the following two situations.
   i) Suppose an organisation wants to have some information about the mileage for a whole fleet of used taxis, and for that they calculate the mean odometer reading (mileage) from a sample of used taxis and find it to be 98,000 miles.

ii) Suppose we want to find the proportion of teenagers who have criminal record and for that we take a sample of 50 teenagers and find that 2 % (or .02) have criminal record.

We can, in fact, have a number of estimators for a given parameter. Apart from the sample mean, the sample median or the average of the smallest and the largest observations in the sample could also be considered as estimators for the population mean. Since we have a variety of estimators for a parameter θ, we should choose the best of the lot to get a real good estimate. But what do we mean by the best? We'll see that in the next section.

5.3 CRITERIA FOR A GOOD ESTIMATOR

In this section we shall discuss some desirable properties of an estimator. You have already learnt in Unit 4 that an estimator takes different values for different samples. But the estimators such as sample mean, proportion have some nice properties. For example, the sample mean has the property that the means of repeated random sample values taken from a given population will centre on the population mean. You recall that in Unit 4 Sec.4.2, we stated this result that the mean of the sampling distribution of the means is equal to the population mean. This means that 'on the average' the estimator values (or estimates) will equal the parameter value. This property is considered to be one of the criteria for a good estimator. We have a definition here.

Definition 2: Suppose $\hat{\theta}$ (read theta hat) is an estimator of the population parameter, θ. The estimator $\hat{\theta}$ takes different values for different samples. If the mean of all these different estimates is the unknown parameter, θ, then we say that $\hat{\theta}$ is an unbiased estimator of θ. Otherwise, it is called a biased one. It follows from the definition of expectation of a r.v., that $\hat{\theta}$ is unbiased if and only if $E(\hat{\theta}) = \theta$. [Please see Sec.3.2, Unit 3, where we have discussed the expectation of a r.v.]

Let us now look at the estimator given in the following situation.

Let us consider some problems.

Problem 1: A Psychologist measures the reaction times of a sample of 6 individuals to certain stimulus. The measures are given by 0.53, 0.46, 0.50, 0.49, 0.52, 0.53 seconds. Determine an unbiased estimate of the population mean.

Solution: An unbiased estimate of the population mean is given by the sample mean,

$$\bar{x} = \frac{\sum x_i}{n}$$

Here $n = 6$ and $x_1 = 0.53, x_2 = 0.46, x_3 = 0.50, x_4 = 0.49, x_5 = 0.52, x_6 = 0.53$.

$$\bar{x} = \frac{0.53 + 0.46 + 0.50 + 0.49 + 0.52 + 0.53}{6}$$

$$= 0.51 \text{seconds}.$$

Then $\bar{x} = 0.5$ seconds. Therefore an unbiased estimate is 0.51 seconds and 0.51 seconds is a point estimate for the mean reaction time of individuals to the stimulus.

———— × ————

Problem 2: In a sample of 400 textile workers, 184 expressed dissatisfaction regarding a prospective plan to modify working conditions. The management felt that this is a strong negative reaction. So they want to know the proportion of total workers who have this feeling of dissatisfaction. Obtain an unbiased estimate of the population proportion.

Solution A point estimate of the population proportion is given by the sample proportion p, given as

$$p = \frac{s}{n}$$

where s denotes the number of observations in the sample which meet the particular characteristic, under study, and n is the sample size.

Here $s = 184$ and $n = 400$.

$$\therefore p = \frac{184}{400} = \frac{46}{100}$$

Therefore an unbiased estimate of the population proportion is $\frac{46}{100}$.

———— × ————

Here are some exercises for you.

E2) A law firm selects a random sample of 60 electronics stores in a particular area, and asks each of them to repair a compact disc player. In each case the law firm determines whether the store makes unnecessary repairs in order to inflate its bill. The law firm finds that 8 of the stores are guilty of this practice. Obtain a point estimate of the proportion of all such stores in the area that inflate bills in this way.

E3) A washing machine company chooses a random sample of 25 motors from those it receives from one of its suppliers. It determines the length of life of each of the motors. The results (expressed in thousands of hours) are as follows:

4.1	4.6	4.6	4.6	5.1
4.3	4.7	4.6	4.8	4.8
4.5	4.2	5.0	4.4	4.7
4.7	4.1	3.8	4.2	4.6
3.9	4.0	4.4	4.0	4.5

The firm's management is interested in estimating the mean length of life of the motors received from the supplier. Provide a point estimate of this population parameter.

We have seen that the sample mean and sample proportion are unbiased estimates for population mean and population proportion respectively. Does this indicate that the statistic or estimator corresponding to the population parameter is always unbiased? To find an answer to this, let us consider the following example.

Suppose we consider the parameter, 'standard deviation'. Then the sample statistic S given by the formula

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}} \qquad (1)$$

where $(x_1, x_2, \ldots, x_n)$ denote the sample observations, can be taken to be an estimator of the population standard deviation. It has been proved that the statistics has an expected value equal to $\sqrt{\left(\frac{n-1}{n}\right)}\sigma$ and not σ, this means that **S is not an unbiased estimator of σ**. Hence an **unbiased estimator of σ is obtained by the expression in (2)**

$$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}, \qquad (2)$$

instead of the expression in (1). For example, an unbiased estimate of the population standard deviation for the situation given in Problem 1 is

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$\frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[(00.53 - 0.51)^2 + (0.46 - 0.51)^2 + (0.50 - 0.51)^2 + (0.49 - 0.51)^2 + (0.52 - 0.51)^2 + (0.53 - 0.51)^2\right]$$

$$= 0.0006$$

$$\therefore S = \sqrt{0.0006} \text{ seconds.}$$

As we have seen in E1, in certain situations one can find more than one unbiased estimator for an unknown parameter θ. If we have to choose between two unbiased estimators for a fixed sample size, then we find the standard deviation (or variance) of the sampling distribution of these two estimators and choose that one with smaller standard deviation (or variance). An unbiased estimator T_1 of a parameter θ is said to be more efficient than another unbiased estimator T_2 of θ if $\text{Var}(T_1) \leq \text{Var}(T_2)$, and in such a case, the sampling distribution of T_1 has a smaller dispersion (spread) about θ than that of T_2 (See Fig.1).

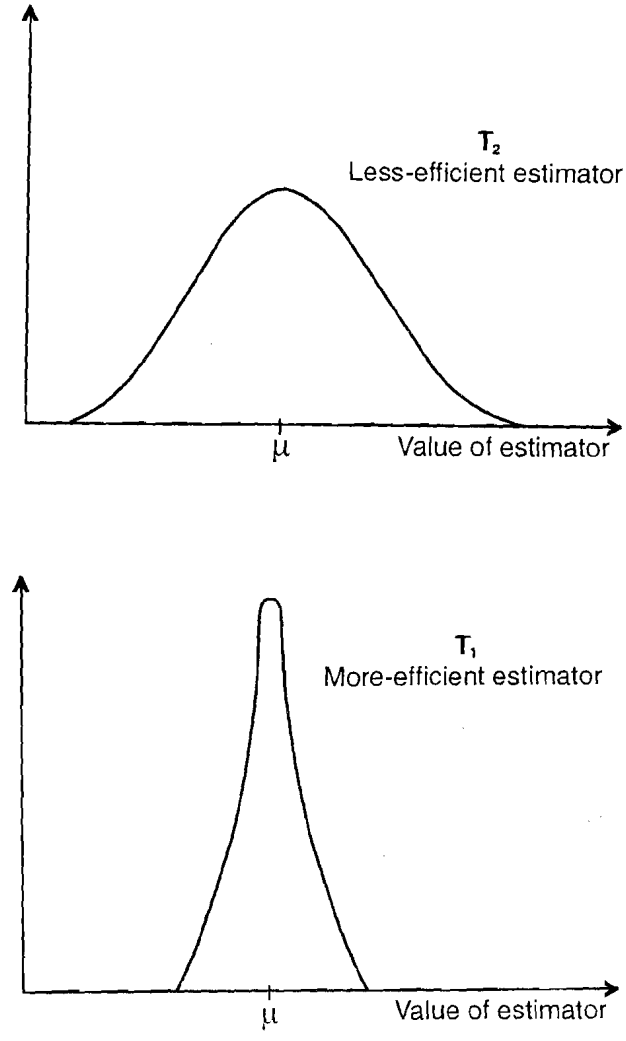


Fig.1:

As an example, let us take a random sample of size n from a normal population with mean μ and standard deviation σ and consider the sample mean and sample median as two estimators of μ. If we compare the sampling distributions of the mean and median for random samples of size n, we get that these two sampling distributions have the same mean but their variances differ. We have seen in Unit 4 that, the variance of the sampling distribution of the mean is σ^2/n, and it can be shown that for random samples of the same size from a normal population, the variance of the sampling distribution of the median is approximately $1.5708\frac{\sigma^2}{n}$.

Hence we get that both the mean and the median are unbiased estimators, but for a given sample size, the standard error for mean is less than that of median.

From what we have already observed now, we get that for random samples from normal populations the mean is more efficient than the median as an estimator of μ. This fact will be more clear to you when you try E4. In fact it can be shown that in most practical situations where we estimate a population mean μ, the variance of the sampling distribution of no other statistic is less than that of the sampling distribution of the mean. In other words, **in most practical situations the sample mean is the 'most acceptable' statistic for estimating a population mean** μ.

There exist several other criteria for assessing the "goodness" of estimators, but we shall not discuss them in this course.

Why don't you try this exercise now.

E4) To verify the claim that the mean is generally more efficient than the median a student conducted an experiment consisting of 12 tosses of three dice. The following are his results: 2,4, and 6; 5,3, and 5; 4, 5 and 3; 5,2 and 3; 6,1 and 5; 2,3 and 1; 3,1, and 4; 5,5 and 2; 3,3 and 4; 1,6 and 2; 3,3 and 3; and 4,5 and 3.

a) Calculate the 12 medians and the 12 means.

b) Group the medians and the means obtained in part (a) into separate distributions having the classes 1.5-2.5, 2.5-3.5, 3.5-4.5 and 4.5-5.5.

c) Draw histograms of the two distributions obtained in part (b) and explain how they illustrate the claim that the mean is generally more efficient than the median.

We can summarise our discussion up to now as follows:

- Population parameters are usually unknown and need to be estimated from a sample
- There could be a variety of estimators for the same parameter.
- "Unbiasedness" and "efficiency" are some of the desirable properties of a good estimator.

5.4 INTERVAL ESTIMATION

In the last section we have seen what a point estimate is. Sometimes it is difficult to evaluate the precision of a point estimator (as measured by its variance, say). Alternatively, we can think of giving an interval, computed on the basis of the sample values, which will contain the true parameter with a certain degree of confidence. This interval is called an interval estimator of the parameter. These intervals are also called confidence intervals. We shall first discuss confidence intervals for the population mean μ. We first consider the case when the variance σ is known.

5.4.1 Confidence Interval for the Mean with Known Variance

Suppose you have been suspecting that the 1 litre pack of milk that is delivered to your house every morning is not exactly 1 litre, but less. You feel that the filling machine which is supposed to fill each polypack with 1 litre of milk is not working properly. Of course, you are ready to admit that even though the machine is set for 1 litre, it has a certain variability and so there could be some packs which are less than 1 litre full while others which are more.

To end your doubts, you need to find the average volume of milk filled by the machine. Obviously, it would be impossible to do this except by taking a sample. Suppose you

measure the milk pack you get over a period of sixty days. That is, your sample size is 60. Suppose you find that the mean of your observations, which is the sample mean, is 950 ml. This is an estimate of the population mean. But you cannot immediately conclude that the machine is set for 950 ml. You must account for the variability of the sample means. For this you must also know the standard deviation, σ, or calculate it from the sample. Suppose we assume that $\sigma = 50$.

Now we shall construct an interval for the parameter μ the average amount of milk that the machine gives. For that we make use of the central limit theorem discussed in Unit 4. According to this Theorem, for sufficiently large sample size n the sample mean $\bar{x}$ is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Then we make use of the normal distribution table given in Appendix 2 at the end of this block and note that

$$P[-1.96 < Z < 1.96] = 0.95 \tag{3}$$

$$\text{where } Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ i.e. } Z\left(\sigma/\sqrt{n}\right) = \bar{x} - \mu$$

Now we rewrite Eqn.(3) using simple algebra as

$$P\left[-1.96\,\frac{\sigma}{\sqrt{n}} < Z\frac{\sigma}{\sqrt{n}} < 1.96\,\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\text{i.e.} P\left[-1.96\,\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < 1.96\,\frac{\sigma}{\sqrt{n}}\right].$$

Now we subtract $-\bar{x}$ from all the three terms inside the bracket. Then we get

$$P\left[-\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Now we multiply all the terms inside the bracket by -1 and (therefore the inequalities get reversed) and we get

$$P\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95 \tag{4}$$

If we multiply the terms in the inequality $y \geq 1$ by (-1), then the inequality gets reversed and we get $-y \leq -1$.

Thus corresponding to each sample mean $\bar{x}$, we got an interval given by

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{5}$$

which satisfies Equation (4). Let us now see what does Equation (4) implies. Let us, for example consider the sample value $\bar{x} = 950$ml. obtained for the problem regarding average volume of milk filled by the machine. Then the Equation (4) corresponding to $\bar{x} = 950$ml is

$$P[937.35 < \mu < 962.65] = 0.95$$

We interpret it in the way that we are 95 % confident that the interval (937.65, 962.65) contains the true value μ. This does not mean that "There is 95 % probability that μ lies in the interval (937.35, 962.65). This is a very common mis-interpretation of Equation (4) and it is incorrect. This is because the population mean μ is a fixed quantity and therefore μ either lies in the interval (937.35, 962.65) or it does not. Therefore the probability that μ lies in the interval is either 0 or 1. The 95 percent probability is assigned to our level of confidence that the interval contains μ. It is not assigned to the probability that μ lies in the interval.

Another interpretation of Equation (4) is based on the fact that we can construct a confidence interval for each sample mean $\bar{x}$. We will get different intervals for different values of sample means. So, in this case Equation (4) says that if all possible samples of size n are calculated, and the intervals $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ are calculated for each sample, then 95 % of all such intervals are expected to contain the population parameter μ. This does not mean that for a particular sample value $\bar{x}$, we can expect that the interval $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$ will contain μ.

The confidence intervals $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is also denoted as $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

That means if you select 100 samples and calculate the intervals about their sample means, then 95 of these will contain the population μ. Note that here we assume that σ is known. In the following figure we have illustrated this graphically, showing five such intervals

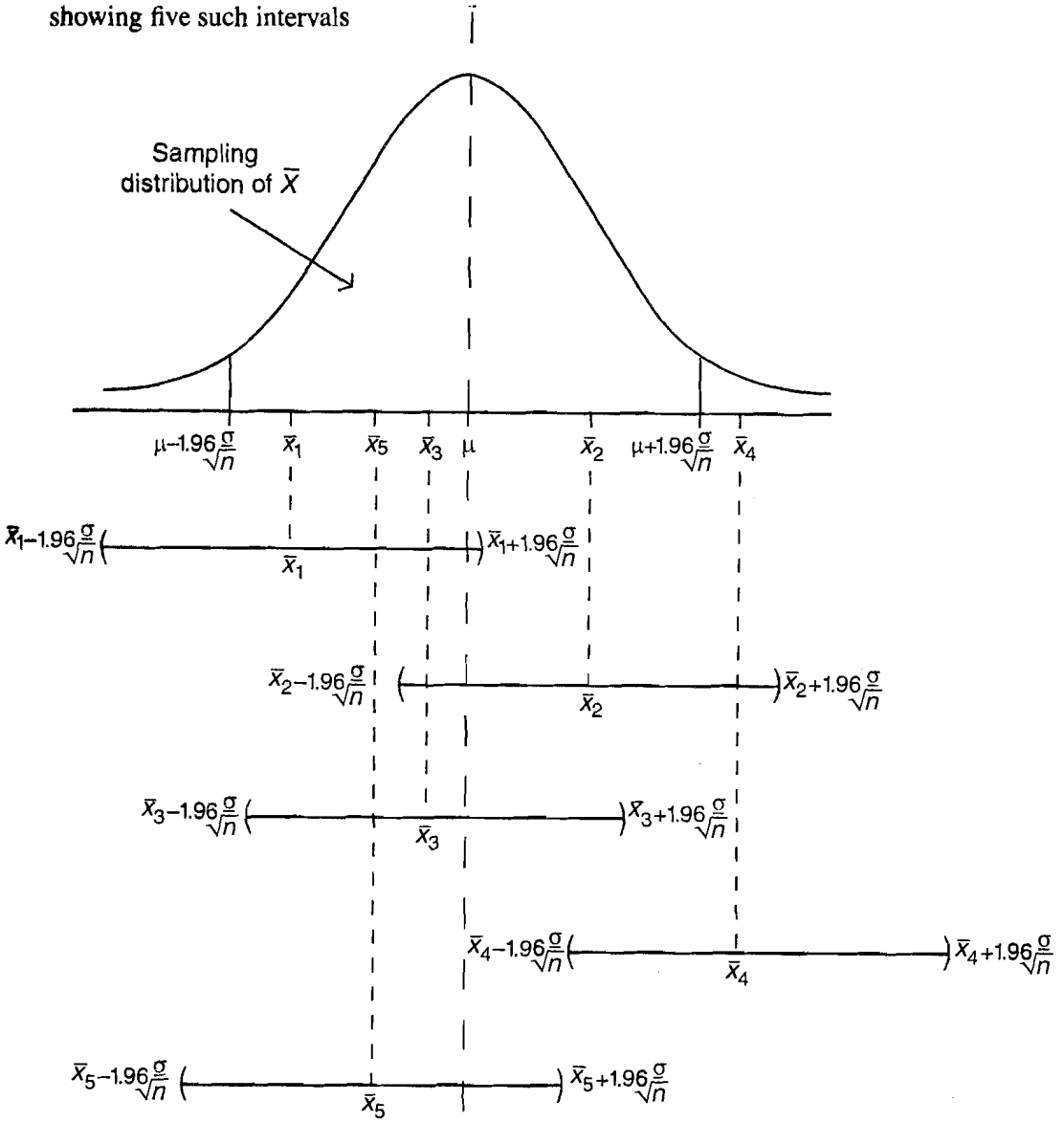


Fig.2: A number of intervals constructed around the population mean.

Only the interval constructed around the sample mean $\bar{x}_4$ does not contain the population mean.

The interval given by (5) is called a confidence interval.(C.I)

The value 0.95 (or 95%) attached with the confidence interval is called **confidence coefficient**. The left end point of the confidence interval is called **lower confidence limit (LCL)** and the right end point of the confidence interval is called upper **confidence limit** (UCL). The difference between the UCL and LCL is the width of the confidence interval. The width of the 95% confidence interval in the above example is

$$2 \times \left(1.96\frac{\sigma}{\sqrt{n}}\right)$$

Although 0.95 is frequently used as a confidence coefficient, we can have other values such as 0.90 or 0.99 as confidence coefficients. Using the normal distribution table, we can obtain the confidence interval for 0.90 (or 90%) as

$$\left(\bar{x} - 1.64\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.64\frac{\sigma}{\sqrt{n}}\right)$$

and for 0.99 (or 99%) as

$$\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}},\ \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right)$$

In the following problem we illustrate the use of confidence intervals.

Example 1: The Director of a marketing division wants to analyse the market value of business firms of a similar size. [Market value is defined as the number of common shares outstanding, multiplied by the share prize as listed on an organised exchange]. A sample of 600 firms revealed a mean market value of Rs.850 million. The earlier results reveals that the population is normally distributed with population standard deviation Rs.200 million. It is desired to set up a confidence interval for the (unknown) mean market value.

Given that the sample mean is $\bar{x} = 850$ million and the sample size $n = 600$ and $\sigma = 200$ million. Therefore can construct 95 % confident interval, which is given by

$$\left(850 - 1.96 \times \frac{200}{\sqrt{600}},\quad 850 + 1.96 \times \frac{200}{\sqrt{6000}}\right)$$

$$\text{i.e.}(850 - 15.99,\quad 850 - 15.99)$$

$$\text{i.e.}(834.01,\quad 865.99)$$

This shows that the director can be 95% confident that the interval (834.01, 865.99) contain the mean market value.

* * *

It is time to do some exercises.

E5) For each of the values given below, calculate the 95% confidence interval for the mean.

i) $\bar{x} = 0, \sigma = 10, n = 8$

ii) $\bar{x} = 550, \sigma = 40, n = 16.$

E6) If the mean length of hospitalisation of 140 patients was 11.4 days and the standard deviation of patient days is assumed to be 2.5 days, what is the 99% confidence interval for the average length of stay? Assume normality.

E7) Estimate the number of days between germination and the first pickable cucumbers using the following sample.

Date of germination	First Fruit
May 1	June 17
4	18
8	21
5	16
12	28
18	July 3
11	June 25
9	26

What is the 95% confidence interval assuming $\sigma = 2$ days?

Next we shall consider the case when σ is unknown.

5.4.2 Confidence Interval for Mean with Unknown Variance

In all the computations of the confidence interval for μ so far we have assumed that the population variance is known. Each time, the normal distribution was the appropriate sampling distribution used to determine the confidence intervals. However the normal

distribution is not appropriate when the population variance is unknown and the sample size is less than 30. In such situations we use t-distribution. As indicated in the previous unit, the sample standard deviation 's' is generally used as an estimator of the population standard deviation.

If the sample size is 30 or less and the population is normal (and large relative to the sample), a confidence interval for the population mean can be constructed by using the t-distribution in place of the standard normal distribution.

You are already familiar with t distribution from Unit 4. We now have to use the t distribution table given in Appendix to construct the confidence intervals corresponding to different levels of confidence, say 95% or 99%. Let us suppose that we want to find the confidence intervals at the 90% confidence level i.e. $\alpha = 0.1$ with a sample size of 14 similar to the ones we have given in Equation(5). Note that we don't know σ in this case. Therefore, as indicated in Unit 4, the sample standard deviations is used as an estimator of the population standard deviation. Thus if s is known, then 90% confidence interval is given as

$$\left(\bar{x} - t_{0.05}\frac{s}{\sqrt{n}}, \bar{x} - t_{0.05}\frac{s}{\sqrt{n}}\right)$$

where $t_{0.05}$ is the t-value corresponding to the value $\frac{\alpha}{2} = \frac{0.1}{2} = 0.05$ and for the parameter $\nu = n - 1$, where n is the sample size. Now to find the t-value we make use of the table 1 in the Appendix. For example, suppose that $n = 14$, then $\nu = 13$, then, from table 1 we get that the t-value is $t_{\alpha/2} = 1.771$ (See Fig. 3).

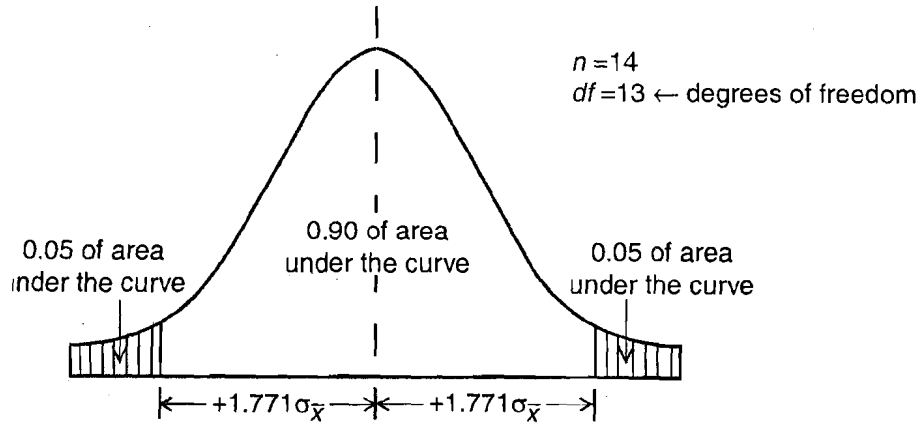


Fig.3: Confidence interval using the t-distribution

Like a z-value, the t-value 1.771 shows that if we mark off plus and minus $1.771\frac{s}{\sqrt{n}}$ on either side of the mean $\bar{x}$, the area under the curve between these two limits will be 90%, and the area outside these limits will be 10%.

Therefore the **90% confidence interval, for degrees of freedom 13 is.**

$$\left(\bar{x} - 1.771\frac{s}{\sqrt{n}}, \bar{x} + 1.771\frac{s}{\sqrt{n}}\right) \quad (6)$$

Similarly the **95% confidence interval for 20 degrees of freedom is**

$$\left(\bar{x} - 2.086\frac{s}{\sqrt{n}}, \bar{x} + 2.086\frac{s}{\sqrt{n}}\right) \quad (7)$$

In a similar way we can find), confidence intervals for different degrees of freedom. (see E8)

Let us consider some examples.

Example 2: A sample of 10 measurements of the diameter of a sphere has a mean $\bar{x} = 43.5$mm and $s^2 = 4$mm. Let us find the i) 95% and ii) 99% confidence intervals for the actual diameter.

i) Here n=10 and $\alpha = 0.5$. Therefore, we use t distribution with 9 d.f. From the table, we get $t_{\alpha/2} = t_{0.025} = 2.26$. So, the 95% confidence interval for μ is

$$\left(\bar{x} - t_{.025}\left(\frac{S}{\sqrt{n}}\right), \bar{x} + t_{0.025}\left(\frac{s}{\sqrt{n}}\right)\right) =$$
$$\left(43.5 - 2.26\left(\frac{2}{\sqrt{10}}\right),\ 43.5 + 2.26\left(\frac{2}{\sqrt{10}}\right)\right) = (42.07,\ 44.93).$$

So, we can be 95% confident that the true mean lies between 42.07 and 44.93

ii) Working similarly, the 99% confidence interval is
$\left(43.5 - 3.25\left(\frac{2}{\sqrt{10}}\right),\ 43.5 + 3.25\left(\frac{2}{\sqrt{10}}\right)\right) = (41.44,\ 45.55).$

* * *

The idea will be more clear to you if when you do the following exercises.

Problem 3: A manufacturer of light bulbs wants to estimate the mean length of life of a new type of bulb which is designed to be extremely durable. The firm's engineer tests nine of these bulbs and find that the length of life (in hours) of each is as follows:

5,000	5,100	5,400
5,200	5,400	5,000
5,300	5,200	5,200

Previous experience indicates that the lengths of life of individual bulbs of a particular type are normally distributed. Construct a 90 percent confidence interval for the mean length of life of all bulbs of this new type.

Solution: If x_i is the length of life of the ith light bulb in the sample, we find that

$$\sum_{i=1}^{9} x_i = 46,800$$
$$\bar{x} = 5,200$$
$$\sum_{i=1}^{9}(x_i - \bar{x})^2 = 1,80,000$$
$$\sum_{i=1}^{9}(x_i - \bar{x})^2/(n-1) = 22,500$$
$$\therefore s = \sqrt{22,500} = 150.$$

Since n=9, we make use of t-distribution. Because a 90 percent confidence interval is wanted, $t_{\alpha/2} = t_{0.05}$; and the number of degrees of freedom is $(n - 1) = 8$. Therefore, the t-distribution table given in the Appendix of Unit 4 shows that if there are 8 degree of freedom, $t_{.05} = 1.86$. Thus, the desired confidence interval is

$$5200 \pm 1.86\left(\frac{1.50}{\sqrt{9}}\right) \quad \left(5200 - 1.86\left(\frac{1.59}{\sqrt{9}}\right)\right),\ 5200 + 1.86\left(\frac{1.59}{\sqrt{9}}\right)$$

By simplifying, we get that the confidence interval is (5107, 5293).

———— ×————

Why don't you try these exercises now?

E8) Given the following sample sizes and confidence levels, find the appropriate $t_{\alpha/2}$ values for constructing confidence intervals.

i) n = 10; 99%

ii) n = 28; 95%

iii) n = 13; 90%

iv) $n = 25$; 99%

E9) A sample of 12 measurements of breaking strengths of cotton threads gave a mean of 0.738 N and a standard deviation of 0.124N. Find a 95% and 99% confidence intervals for the actual breaking strength.

E10) Five measurements of the reaction time of an individual to certain stimuli were recorded as: 0.28, 0.30, 0.27, 0.33 and 0.31 second. Find the 95% confidence interval for the actual reaction time.

E11) If you are given a sample of 20 candles from a large shipment of candles, and are asked to give an interval estimate of their average burning life, how would you proceed? What information would you need?

The above examples and exercises illustrate how we can use t-distribution to find the confidence intervals. **As we mentioned earlier, t-distribution can be used only if the population variance is unknown and the sample size is small.** Next we shall see how to construct the confidence intervals for large samples when the population variance is unknown.

Mathematicians have shown that if the sample size is large, we can simply substitute the sample standard deviation for the population standard deviation in the results obtained in the previous part of this section i.e. in Subsection 5.4.1. Thus, if we want to construct a 95% confidence interval — that is, a confidence interval with a confidence coefficient of 95 percent — we can substitute s for σ in Equation (5), the result being

$$p_r\left[\bar{x} - 1.96\frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{s}{\sqrt{n}}\right] = .95 \tag{8}$$

Consequently, the confidence interval is

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}},\ \bar{x} + 1.96\frac{s}{\sqrt{n}}\right) \tag{9}$$

Equation (8) is applicable only if the population is large relative to the sample.

The following example should make the above discussion more clear.

Problem 4: A random sample of 100 ball bearings made by a machine in 1 week was taken. The mean diameter was found to be 8.24 mm with a standard deviation of 0.42 mm. Find the 95% and 99% confidence intervals for the mean diameter of ball bearings produced by that machine.

Solution: Since the sample is large, from Equation(8), we get that the 95% confidence interval for μ is

$$\left(8.24 - 1.96\left(\frac{0.42}{\sqrt{100}}\right)\right),\ \left(8.24 + 1.96\left(\frac{0.42}{\sqrt{100}}\right)\right) = (8.16,\ 8.32)$$

Similarly, the 99% confidence interval is

$$\left(8.24 - 2.58\left(\frac{0.42}{\sqrt{100}}\right)\right), \left(8.24 + 2.58\left(\frac{0.42}{\sqrt{100}}\right)\right) = (8.13, 8.35)$$

———— × ————

See if you can solve these exercises:

E12) A random sample of marks obtained by 50 students in Mathematics showed a mean of 75 and a standard deviation of 10.

a) What are the 95% confidence limits for the mean marks in Mathematics?

b) With what degree of confidence can we say that the mean marks are between 74 and 76?

E13) A washing machine company's statistician says that 90% confidence interval for the mean length of motors received from Supplier II is 4,500 to 4,800 hours, based on a sample of 36 motors. The statistician also says that the standard deviation of the lengths of life of motors received from Supplier II is 500 hours. Is there any contradiction between the statements? If so, what is the contradiction?

Next we shall illustrate how confidence intervals are calculated for population proportions. We have talked in length about the estimation of population parameter μ. Another important population that we need to estimate is the population proportion, p. Let's see how to go about it.

5.4.3 Confidence Interval for Population Proportion

Let us start with a situation. In a random sample of 25 men from a city, 8 were found to be smokers. Can we estimate the proportion of smokers in the city?

Suppose the proportion of smokers in a city is π. Then $p = \frac{8}{25}$ is a point interval of π, obtained from this sample. Now we shall construct a confidence for estimate π. To do this we proceed similar to what we did for the population mean. We recall the result from Unit 4, that the sampling distribution of sample proportion p has mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. We also know from Unit 4, that if the **sample size is sufficiently large and if π is not very close to 0 or 1, the sampling distribution is approximately a normally distribution**. Then using the standard normal distribution table, we can find confidence intervals. If we want to construct 95% confidence intervals then that will be given by

$$\left(p - 1.96\sqrt{\frac{\pi(1-\pi}{n}},\quad p - 1.96\sqrt{\frac{\pi(1-\pi}{n}}\right) \tag{10}$$

so that

$$P\left[p - 1.96\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + 1.96\sqrt{\frac{\pi(1-\pi)}{n}}\right] = 0.95$$

The interval given by in (9) is called 95% confidence interval for π. Similarly, we can have 90% or 99% confidence intervals. The above intervals given in Equation (9) cannot be used as they involve the unknown, π.

However, if n is large, then π can be replaced by p without compromising accuracy. So that **for large samples, the 95% confidence interval for π will the**

$$\left(p - 1.96\sqrt{\frac{p(1-p}{n}},\ p + 1.96\sqrt{\frac{p(1-p}{n}}\right)$$

If we want to get a 99% confidence interval, we will have to replace 1.96 by 2.58, since

$$P\left[-2.58 \le \frac{P-\pi}{\sqrt{\frac{p(1-p)}{n}}} \le 2.58\right] = 0.99$$

We shall illustrate this with the problem given below.

Problem 5: : In a random sample of 75 parts produced by a machine, 12 have a surface finish which is rougher than the specification will allow. Find a i) 95% ii) 99% confidence interval for the proportion of rough parts produced by the machine.

Solution: Here $p = 12/75 = 0.16$. Then

a) 95% confidence interval is

$$\left(0.16 - 1.96\sqrt{\frac{0.16 \times (.84)}{75}},\ 0.16 + 1.96\sqrt{\frac{0.16 \times (.84)}{75}}\right) = (0.08, 0.24).$$

b) The 99% confidence interval for P is

$$\left(0.16 - 2.58\sqrt{\frac{.16 \times (.84)}{75}},\ 0.16 + 2.58\sqrt{\frac{.16 \times (.84)}{75}}\right) = (0.05,\ 0.27)$$

Here are some exercises for you.

E14) A random sample of 800 calculators contains 24 defective items. Compute a 99% confidence interval for the proportion of defective calculators.

E15) Of 1000 randomly selected lung cancer cases, 699 resulted in death. Construct a 95% confidence interval for the death rate from lung cancer.

E16) A student in a university wanted to decide whether or not a contest the election for the presidency of the students' union. Out of 50 students, 11 showed their willingness to vote for her. Find a 99% confidence interval for the true proportion of students voting for her.

We now summarise our discussion about interval estimation in the following table:

Table 1

Parameter	Point Estimator		Confidence Interval
μ	$\bar{x} \longrightarrow$	σ known	$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}},\ \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$
		σ unknown, large n	$\left(\bar{x} - z\frac{S}{\sqrt{n}},\ \bar{x} + z\frac{S}{\sqrt{n}}\right)$
		σ unknown, small n	$\left(\bar{x} - t\frac{S}{\sqrt{n}},\ \bar{x} + t\frac{S}{\sqrt{n}}\right)$
π	p	p is not too close to 0 or 1, large n	$\left(p - z\sqrt{\frac{p(1-p)}{n}},\ p + z\sqrt{\frac{p(1-p)}{n}}\right.$

Now we shall present a case study which shows how the techniques of the estimation discussed in this unit helps in tackling real-life problems.

Abrasion resistance of rubber is the extent to which rubber can withstand pressure against rubbing off or frictional action. For example, rubber of high abrasion resistance will have high road life.

Statistical Estimation in the Chemical Industry: A Case Study: A chemical firm called Imperial Chemical Industry (ICI) carried out the following experiment to estimate the effect of a chlorinating agent on the abrasion resistance of a certain type of rubber. Ten pieces of this type of rubber were cut in half, and one half-piece was treated the chlorinating agent, while the other half-piece was untreated. Then the abrasion resistance of each half was evaluated on a machine, and the difference between the abrasion resistance of the treated half-piece and the untreated half-piece was computed. Table below shows the 10 differences (1 corresponding to each of the pieces of rubber in the sample). Based on this experiment, ICI was interested in estimating the mean difference between the abrasion resistance of a treated and untreated half-piece of this type of rubber. In other words, if this experiment were performed again and again, an infinite population of such differences would result. ICI was interested in estimating the mean of this population, since the mean is a good measure to find the effect of the chlorinating agent on this type of rubber's abrasion resistance.

If you were a statistical consultant for ICI, how would you analyse these data? You would recognise that a good point estimate of the mean of this population is the sample mean, which is 1.27 as shown in Table below. Thus, your first step would be to advise ICI that if they want a single number as an estimate, 1.27 is a good number to use. Next you would point out that such a point estimate contains no indication of how much error it may contain, whereas a confidence interval does contain such information. Since the population standard deviation is unknown and the sample is small, expression ( ) should be used in this case to calculate a confidence interval. Assuming that the firm

wants a confidence coefficient of 95 percent, the confidence interval is (0.464 2.076), because $t_{.025} = 2.262, s = 1.1265$, and n=10. The chances are 95 out of 100 that such a confidence interval would include the population mean. (Note that this analysis assumes that the population is approximately normally distributed)

Place	Difference
1	2.6
2	3.1
3	-0.2
4	1.7
5	0.6
6	1.2
7	2.2
8	1.1
9	-0.2
10	0.6

$$\sum_{i=1}^{10} x_i = 12.7$$

$$\bar{x} = 1.27$$

$$s = 1.1265$$

The above analysis is, in fact, exactly how ICI's statisticians proceeded. Despite the fact that the sample consisted of only 10 observations, the evidence was very strong that the chlorinating agent had a positive effect on abrasion resistance. After all, the 95 percent confidence interval was that the mean difference between abrasion resistance of rubber with and without treatment was an increase of between 0.464 and 2.076. (For that matter, the statisticians found that the 98 percent confidence interval was that the mean difference was an increase of between0.265 and 2.275). The best estimate was that the chlorinating agent resulted in an increase of about 1.27 in abrasion resistance.

With the detailed example you have seen how several aspects covered in this unit has merged. In fact as you reflect on this case study you should check from the summary below how many points are actually covered in this case study.

With that we come to the end of this unit.

5.5 SUMMARY

In this unit we have seen that

1) When the population is large, its parameters, like mean, variance, proportion, need to be estimated from a sample.
2) There can be many different estimates of a parameter.
3) An estimator is unbiased if the mean of the estimates is the population parameter
4) Between two unbiased estimators we prefer the one with the smaller variance
5) Interval estimates are better than point estimates since we can easily specify the precision of our estimate.
6) The computation of confidence intervals is done by using the sampling distributions of the estimators.

5.6 SOLUTIONS/ANSWERS

E1) For (i) the estimator is the mean mileage of the sample of used taxis. The value 98,000 miles is an estimate.

For (ii) the estimator is the proportion and the value .02 is an estimate.

E2) An unbiased estimate of the population proportion is obtained by

$$p = \frac{8}{60} = \frac{2}{15}$$

E3) A point estimator of the population mean is obtained by calculating the sample mean.
The sample mean of 25 motors is 4.448 thousands of hours.

E4) a) The medians are 4,5,4,3,25,2,3,5,3,2,3 and 4; the means are 4,4.3, 4,3.3,2, 2.7,4,3.3,3 and 4.

b) The frequencies are2,4,3 and 3 for the medians and 1,5,6 and 0 for the means. Then obtain the frequency distribution.

c) The histograms of two distributions shows that the variance for the median is more than for the mean which illustrate the claim that the mean is generally more efficient that the median.

E5) 95% confidence interval for mean is $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$

i) Here $\bar{x} = 0$ and $\sigma = 10$ and $n = 8$. Therefore the interval is

$$\left(0 - 1.96\frac{10}{\sqrt{8}}, 0 + 1.96\frac{10}{\sqrt{8}}\right) = (-6.9296, 6.9296)$$

ii) Here $\bar{x} = 550, \sigma = 40$ and $n = 16$. Therefore the interval is

$$\left(550 - 1.96\frac{40}{\sqrt{16}}, 550 + 1.96\frac{40}{\sqrt{16}}\right) = (530.4, 569.6)$$

E6) Here $n = 140, \bar{x} = 11.4, \sigma = 2.5$
99% C.I. is given by
$\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right) = \left(11.4 - 2.58\frac{2.5}{\sqrt{140}}, 11.4 + 2.58\frac{2.5}{\sqrt{140}}\right) =$
$(10.855, 11.945)$

E7) Number of days are:
47,45,44,42,47,46,45,48
$\therefore \bar{x} = 5.5, n = 8, \sigma = 2$
There C.I. is given by $\left(5.5 - 1.96\frac{2}{\sqrt{8}}, 5.5 + 1.96\frac{2}{\sqrt{8}}\right) = (4.1141, 6.8859)$

E8) i) Note that here $t = n - 1 = 10 - 1 = 9$ and $\alpha = 0.005$. Therefore we look under column for 0.005 till we reach the row for 9. Then we get the value 3.250.

ii) Here $\alpha = 0.5$ the $t_{\alpha/2}$-value is 2.052

iii) Here $\alpha = 0.1$ the $t_{\alpha/2}$-value is 1.782

iv) Here $\alpha = 0.01$ the $t_{\alpha/2}$-value is 3.797

E9) $n = 12 \therefore$ d.f. $= 11$
The t value for 95% C.I. is 2.20 and that for 99% C.I. is 3.11.
$\therefore$ 95% C.I. : $0.738 \pm 2.20\left(\frac{0.124}{\sqrt{12}}\right) = (0.6592, 0.8167)$
99% C.I. : $0.738 \pm 3.11\left(\frac{0.124}{\sqrt{12}}\right) = (0.6267, 0.8493)$

E10) From the sample, $\bar{x} = 0.298$ and

$s = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} = 0.0213$ and d.f. = 4.

$\therefore$ The required value of t is 2.78

$\therefore$ 95% C.I. $= 0.298 \pm 2.78\left(\frac{0.0213}{\sqrt{5}}\right)$

$= (0.2715, 0.3245)$

E11) Light up the candles and measure the amount of time (life time) for which each candle burns. This data will have 20 observations. Find the mean ($\bar{x}$) and the standard deviation (s) of this data. The value of $\bar{x}$ is a point estimate. If we want 95% C.I., we find t = 2.09 for 19 d.f., since the sample size is 20. Then C.I. is $\bar{x} \pm \frac{s}{\sqrt{20}}$.

E12) a) $(72.2281, 77.7718)$

b) $P(74 \leq \mu \leq 76) = 2P(75 \leq \mu \leq 76)$

$= 2P\left(0 \leq Z \leq \frac{76.75}{10/\sqrt{5}}\right)$

$= 2P(0 \leq Z \leq 0.7071) = 0.5224$

$\therefore$ Degree of confidence is 52%

E13) C.I. is $p \pm 2.58\sqrt{\frac{p(1-p)}{n}}$

$p = \frac{24}{800} = 0.03, n = 800$

$\therefore$ C.I. $= (0.0144, 0.0456)$

E14) $p = 0.699$

C.I.is$(0.6706, 0.7274)$

E15) $p = \frac{11}{50} = 0.22$. Therefore the C.I is $(0.0688, 0.3711)$

UNIT 6 TESTS OF SIGNIFICANCE

Structure **Page No.**

6.1 INTRODUCTION

Statistical data are collected in many diverse fields. The main purpose of data collection is to arrive at certain decisions. For example, a company may want to decide whether or not to accept a consignment of ball bearings. A student may want to know whether or not to believe the claim of a coaching class owner that students in his class get 90% marks on the average in Maths. A doctor may need to test whether or not to prefer the new pain medicine to the old.

In all these situations one can identify a random variable whose distribution is not known to the decision maker, but will be useful for decision making. The company may identify a quality characteristic associated with each ball bearing and the company may want to know if the quality characteristic associated with a great majority of the ball bearings in the consignment satisfy the upper and lower specifications on it. The student wants to know if the expected value of the random variable, namely the score in Mathematics of a student of the coaching centre is 90% or not. The doctor wants to know whether the expected value of the random variable, which is the 'duration of relief' for the new pain medicine is more than that for the old medicine or not. In each of the above cases the decision has to be based on the information obtained from a random sample. In the previous unit you have seen how we can get a point and an interval estimates of parameters of a distribution by observing a finite number of realizations of the random variable. In this unit we shall see how a hunch or a claim or a hypothesis about the actual value of a parameter can be tested and a decision taken to reject or accept the hypothesis.

Let us consider again the random variables that were of interest in the above mentioned situations. For the company that has received a consignment of ball bearings, the quality characteristic of interest may be its thickness. The thickness of a ball bearing can be measured on a continuous scale and if $X_1, X_2, \ldots X_n$ denote n realizations (the thickness of n randomly chosen ball bearings, from a very large consignment),it is easy to see that they are a set of independent and identically distributed random variables. It is reasonable to assume that the common distribution is the normal distribution with some unknown mean μ and variance σ^2. The student can reasonably assume that the scores in Mathematics of a random sample of size n from among the alumni of the coaching centre are independent and identically distributed as normal with mean μ and variance σ^2. Similarly, the time durations of relief reported by n_1 randomly chosen test

patients using one dose of the new pain medicine can be modelled as n independent and identically distributed random variables distributed as normal(μ_1, σ_1) and those of another set of randomly chosen n_2 patients using one dose of the old medicine as i.i.d. normal(μ_2, σ_2^2). We shall deal with the problem of testing hypotheses about the mean and variance of normally distributed random variables in sections 6.3 and 6.4. In Sec. 6.5 we shall illustrate the problem of testing hypothesis about the proportion.

Objectives

After reading this unit, you should be able to

- formulate null and alternative hypothesis for a given problem,
- describe Type 1 and Type 2 errors,
- differentiate between one sided alternatives and 2-sided alternatives,
- test whether μ has the specified value μ_0 against either a one sided alternative or against a two sided alternative assuming that we have n independent observations on a random variable that is distributed as normal(μ, σ) where σ is either assumed known or is unknown,
- test whether the two normal populations have the same mean or not,
- test whether the population variance from a normally distributed population, has a specified value against either a one sided or two sided alternative,
- test if the population proportion is significantly different from the hypothetical population proportion,
- test if the difference between two population proportions is zero or not based on the observed sample proportions.

6.2 SOME BASIC CONCEPTS

Let us begin with a problem.

A company manufacturing Aluminium rods wants to find the average mean diameter of the length of rods manufactured in the company. Based on previous experience, the firm's statistician knew that, on the average, the rods should be 2 cm. in diameter. He also knew that the variance of the diameters of the rods produced by the process is $\sigma_0^2 = 0.25$ A test was needed for detecting any change in this mean diameter every day so that whatever factors were responsible for such a change could be corrected.

Note that in this situation the statistician is interested in making inference about the population parameter mean μ. However, she is not interested in estimating the value of μ, rather she is interested in testing a hypothesis about the value of μ. The hypothesis is that the mean diameter of the rods produced on a particular day is 2cm. (That is there has been no change in the mean diameter (μ) of all the rods produced in a day' from the established standard length.) Let us denote this hypothesis by H_0, and call it the null (no change) hypothesis. The statistician wishes to test this null hypothesis against the alternative that it is not true.

In statistics, the **null hypothesis is the hypothesis that is being tested**. A statistician has to be careful while formulating the null and alternative hypotheses since it is presumed that the null hypothesis holds unless there is a strong statistical evidence obtained from the sample of observations against it and in favour of the alternative hypothesis. The situation is somewhat similar to the modern system of justice. A person accused of a crime is presumed to be not guilty, unless the prosecution can produce strong evidence to the contrary. To make the concept clear, let me pose a question. A

strong evidence to the contrary. To make the concept clear, let me pose a question. A manufacturer of paints wants to use a new drying process. The present process requires a drying time of 15 minutes and it is claimed that the new process requires only 12 minutes of drying. The new process is to be tested on n sample pieces. What should be the null and alternative hypothesis if the manufacturer wants to continue to use the old process unless there is strong evidence for the claim regarding the new process? In this case you can see that the null hypothesis is $H_0 : \mu = \mu_0 = 15$ minutes and the alternative hypothesis is $H_1 : \mu = \mu_1 = 12$ minutes, where μ is the mean drying time of the new process. The new process is not accepted unless the null hypothesis is rejected. This is the case of a conservative manufacturer. A more risk taking manufacturer may decide that the new process may be accepted unless there is statistical evidence against the claim. In this case the null hypothesis is formulated as $H_0 : \mu = \mu_1 = 12$ minutes against the alternative hypothesis $H_1 : \mu = \mu_0 = 15$ minutes. Unless the null hypothesis in this formulation is rejected, the manufacturer goes for the new process.

We now continue with the case of the company manufacturing aluminium rods. The statistician is investigating if the diameter of a rod produced today is distributed as normal with a mean that is different from 2.0 cm. Unless there is strong statistical evidence for such a change, the statistician will not like to reject the hypothesis that $\mu = 2$ as that might entail stopping the process. If such a change is absent, the mean diameter is 2 cm. Thus the null hypothesis H_0 in this case is

H_0 : The mean diameter is 2 cm.

Now to make any decision, the company have to choose between this hypothesis and the alternative hypothesis, which we denote H_1, and this is stated as

H_1: The mean diameter of the given lot of rods is not 2 cm.

As the name suggests this hypothesis is alternate to H_0.

Thus the statistician has made a claim/hypothesis and she wants to test this claim against a suitable alternative hypothesis on the basis of a sample of observations.

Suppose $X_1, X_2 \ldots X_n$ are i.i.d. (independently and identically distributed) random variables having the common distribution function $F(\theta)$ where θ is a real valued parameter. The set of all possible values of θ is known as the parameter space and is denoted as Θ. We wish to test the hypothesis that $\theta \in \Theta_1$ against the alternative that $\theta \in \Theta_2$, where Θ_1 and Θ_2 are non-intersecting subsets of Θ. In the above example, the statistician will observe from the day's production a random sample of n rods. Let $X_1, \ldots X_n$ denote their diameters. Clearly, $X_1, X_2 \ldots X_n$ are i.i.d random variables. The statistician also assumes that their common distribution is normal with mean μ and variance σ_0^2, where σ_0^2 is assumed to be 0.25 The parameter space Θ may be taken as $(-\infty, \infty)$ although the mean diameter can never be negative. Also we take $\Theta_1 = \{2.0\}$ and $\Theta_2 = \Theta \setminus \{2.0\}$.

We call the null hypothesis *simple* if Θ_1 is a singleton set. Otherwise we call it a *composite hypothesis*. Similarly, the alternative hypothesis is called simple if the set Θ_2 is a singleton set; else it is called a composite hypothesis. The testing problem of the company manufacturing aluminium rods has a simple null hypothesis against a two sided (i.e., both $\mu > \mu_0$ and $\mu < \mu_0$) composite alternative hypothesis.

In any hypothesis testing problem such as the above you may easily recognise four possibilities, two for the true value of the hypothesis H_0 and two possibilities for the outcome of any test procedure. It is therefore clear that there can be two kinds of errors of judgement. First, one can reject the null hypothesis when it is true. Second, one can fail to reject the null hypothesis when it is false. These two kinds of error are called Type I and Type II errors and are defined as follows:

when it is true. A Type II error occurs if the null hypothesis is not rejected (or accepted) when it is false.

The four possibilities are described in the table below:

Table I

Two possible outcomes for H_0	Decision taken	
	Do Not reject H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_0 is not true	Type II error	Correct decision

To familiarise you more with the notions of Type I and Type II error, let us look at the problem of 'Milk packets' we discussed in the Unit 5. Suppose we state the hypothesis H_0 and H_1 as

H_0 : The machine is set for 1 litre. (i.e. the average quantity of milk is 1 litre)
H_1 : The average is not 1 litre.

Let us see what are the four possible situations.

Case-1: Suppose that the hypothesis H_0 is actually (really) true. Then if we decide to accept the hypothesis 'H_0 is true', then we have made the right decision. That means if the machine is really set for 1 litre, and if we accept this following our test procedure, then we have made a correct decision.

Case 2: Suppose that, H_0 is really true and on the basis of our procedure we reject it, then we have made a mistake. That means if the machine is really set for 1 litre and our decision is to reject this and confront the milk man, then we are making a **mistake/error**.

Case 3: Suppose that the hypothesis H_0 is actually false. If we accept such a hypothesis, then we have made a mistake. That means if the machine is not properly set for 1 litre, and based on the procedure we conclude that it is set for 1 litre, then we commit a **mistake/error**.

Case 4: Suppose that the H_0 is actually false. If we reject H_0 then our conclusion is correct. That means if the machine is not properly set, and our procedure also concludes so, then we would be fully justified in demanding an explanation from the milk man. Thus there are two situations in which an error occurs.

1) **Hypothesis is true, but we reject it. This is called Type 1 error.**

2) **Hypothesis is false, but we accept it. This is called Type 2 error.**

Now you can try some exercises to see how much you have followed.

E1) Suppose you have a cough. You open the medicine cupboard and find an unmarked bottle. You have a hunch that it is not a cough medicine, rather, it is some poison. If you are using hypothesis test to arrive at a decision, how will you state your null hypothesis and alternative hypothesis? What are the two situations that lead to Type 1 and Type II errors?

E2) If you test a hypothesis and reject the null hypothesis in favour of the alternative hypothesis, does your test prove that the alternative hypothesis is correct? Justify your answer.

Now that we are aware of the types of errors, our aim is to reduce the probability of their occurrence. It is, of course, not possible to eliminate both the errors at the same time.

Depending on the problem in hand, we'll have to choose the type of error which we may prefer to have. For this, we have to look at the consequences: We may have a

Depending on the problem in hand, we'll have to choose the type of error which we may prefer to have. For this, we have to look at the consequences: We may have a situation where, if we make a Type 1 error, we lose Rs.100, and if we make a Type 2 error, then we lose Rs.10,000. In this case we'll try to eliminate the Type 2 error, since it's more expensive.

Similarly let us look at the situation in E1. You open the medicine cupboard and find an unmarked bottle. You have a hunch that it is not a cough medicine, rather it's some poison! You are not sure though. If you reject this hunch and take it to be the cough medicine, when actually it is a poison, then you have made a Type 1 error. Of course, being dead, you would be beyond caring about errors by then! On the other hand, if it is really the cough medicine, but you accept your hunch and refuse to drink it, the only consequence will be that your cough would last a little longer. In this case you would surely opt for the Type 2 error.

Often we have to properly balance the two types of error.

We denote by α, the probability of committing an error of Type 1, and by β, the probability of an error of Type 2.

Let us look at another situation suppose our supplier has sent us 5000 pens. We want to decide whether to accept or reject this lot. We would like all the pens to be in working order, but also realise that there would be some defective ones. We are ready to allow 5% defective. It is not possible to test all the pens. So we decide to test a sample of 20. We now have to decide the criterion of acceptance or rejection of the lot. For example, we may decide to accept the lot if we find at most two defectives in the sample. So we reject the lot (equivalently, decide that the lot has more than 5 % defectives) if we find 2 or more defectives in the sample. If the hypothesis that there are 5 % defectives were true, we can find the probability of rejecting the lot, that is, P (2 or more pens are defective) by using a binomial probability distribution with $n = 20, p = 0.05$. This probability is α, the probability of committing an error of Type 1. The computation of β is not always possible. Because as in this example if $p \neq 0.05$, there are infinitely many alternatives for p. Now, in this example, as soon as we fix the criterion for rejection, α is fixed.

Thus, before start testing a hypothesis, we fix or specify the value of α, or equivalently the critical region. α is called the **level of significance** of the test.

So, it is in our hands to keep the probability of Type 1 error low.

So far we have discussed that, to carry out a test for making decision, we have to choose between the null hypothesis and the alternative. We also note that we have to be careful in making decision because it can lead to errors like Type 1 and Type 2 discussed earlier. So, the tests should be such that it should be possible to measure these errors and to some extent, at least be reduced.

Let us now go back to the problem of the company manufacturing aluminium rods. Let us see how the statistician conducts a test. Note that in this case the statistician actually wants to test whether there is any change in the mean diameter of the rods. Essentially he is testing the change in the parameter mean.

Note that the change can occur in both directions. Either the mean diameter can be very large or it can very small, and neither is desirable. This is why we formulated a two sided composite alternative.

Let us now see how the statistician proceeds. The first step is to state H_0 and H_1 which has been done earlier. We restate it as

$$H_0 : \mu = 2, \quad H_1 : \mu \neq 2$$

where μ is the population mean.

Since the random sample of observations $X_1, X_2 \ldots X_n$ are assumed to be i.i.d, with the common distribution function as Normal(μ, σ), the distribution of $\bar{X}$ is normal and if the null hypothesis is true (That is, if the population mean is 2cm.), the sampling distribution of the sample mean is as shown in Fig.1. We know from Unit 4 that the sampling distribution of mean is normally distributed and has a mean of 2 cm. and a standard deviation $\frac{\sigma}{\sqrt{n}}$ (where σ is the population standard deviation and n is the sample size.) How does this information help the company to decide on whether its hypothesis is acceptable or not? First, they have to choose a level of significance which is "reasonable" Let us say this is $\alpha = 0.05$ i.e. 5%. This α defines a region, which is the total shaded region in the figure below (See Fig.1 below). The unshaded portion in the figure below is called the critical region. Let us see how this region is calculated. Take $\frac{1-\alpha}{2} = \frac{0.95}{2} = 0.4750$. Look at the normal distribution table, and see where the value 0.4750 is given. You will find this listed in the row 1.9 and column 0.06. Thus it corresponds to z = 1.96. So, we consider the region below the curve, and bounded by $\mu_0 \pm 1.96\sigma/\sqrt{n}$ where μ_0 is the hypothetical mean (Here $\mu_0 = 2$). This region is the critical region shown in Fig. 1. The values z = 1.96 and z = −1.96 are called the **critical values.**

Now we calculate the sample mean and check whether this mean lies within the critical region or outside.

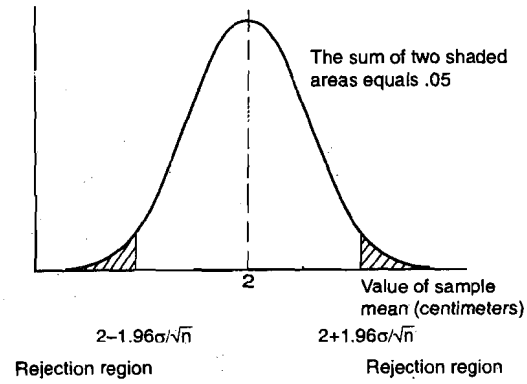


Fig.1

Critical region is calculated using the confidence level discussed in Unit 5. If it lies outside the critical region, then the decision is to reject H_0 in favour of H_1. That means the assumption that the mean diameter is 2 cm. is rejected. If it lies inside the critical region, then the conclusion is that we cannot reject H_0. That is, there is insufficient evidence to conclude that the population μ is not 2cm. Even if our sample statistic Fig.1 does fall in the unshaded region (the region that makes up 95% of the area under the curve), this does not prove that our null hypothesis (H_0) is true, it simply does not provide statistical evidence to reject it, why? Because the only way in which the hypothesis can be accepted with certainty is for us to know the population parameter, unfortunately this is not possible. Therefore, whenever we say that we accept null hypothesis, we actually mean that there is not sufficient statistical evidence to reject it.

Let us now summarise the important steps that we have discussed in the testing procedure of any hypothesis.

1) Formulate the null and alternate hypothesis .
2) Specify the significance level of the test (i.e. fix α).
3) Choose a test statistic
4) Find the critical region
5) Collect the sample and calculate the numerical value of the test statistic based on the sample of observations.

6) Conclusion:

i) If the numerical value of the test statistic falls in the rejection region, we reject the null hypothesis and conclude that the alternative hypothesis is true. We know that the hypothesis-testing process will lead to this conclusion incorrectly (Type I error) only $100\alpha\%$ of the time when H_0 is true.

ii) If the test statistic does not fall in the rejection region, we do not reject H_0. Thus, we reserve judgement about which hypothesis is true. We do not conclude that the null hypothesis is true, because we do not (in general) know the probability β that our test procedure will lead to an incorrect acceptance of H_0 (Type II error).

In the next section we shall illustrate these steps with many examples.

Let us look at another example.

Suppose a light bulb manufacturer believes and advertises that her bulbs have a mean life-time of 1500 hours, and she wants to test her belief. So She formulates the null hypothesis as $H_0 : \mu = 1500$ hours (h)and the alternative as $H_1 : \mu \neq 1500$h. However, she realises that her customers won't complain if the life of the bulbs they buy from her have a life span of more than 1500 hrs. So she reformulates her problem as

$$\text{Test } H_0 : \mu = 1500\text{h against } H_1 : \mu < 1500\text{h}.$$

So in this case we reject H_0 only if the mean life of the sampled bulbs is significantly below 1,000 hours. This situation is illustrated in Fig. 2.

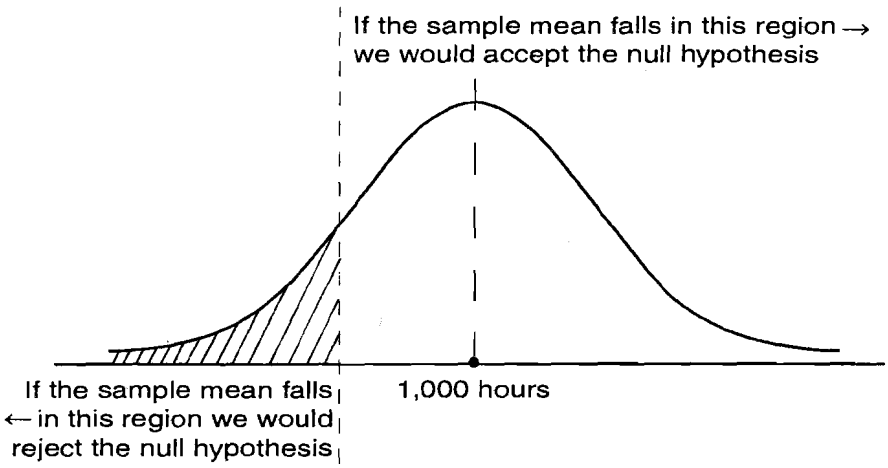


Fig. 2

Note that in this situation, the rejection region is in the left tail of the distribution of the sample mean, and so we call this test a left-tailed (or one-tailed) test.

A left-tailed test is one of the two kinds of one-tailed test. As you have probably guessed, the other kind of one-tailed test is right-tailed test which is used in the situations as given below:

A sales manager has asked her sales people to observe a limit on travelling expenses. The manager hopes to keep expenses to an average of Rs. 100 per salesperson per day. One month after the limit is imposed, a sample of submitted daily expenses is taken to see whether the limit is being observed. Here the null hypothesis is $H_0 = \mu = \text{Rs.}100$, but the manager is concerned only with excessively high expenses. Thus, the appropriate alternative hypothesis in this case is $H_1 : \mu > \text{Rs.}100$. So, in this case the null hypothesis is rejected (and corrective measures taken) only if the sample mean is significantly higher than Rs. 100. The situation is illustrated in Fig. 3.(See next page.) Note that in this situation, the rejection region is in the right-tail of the distribution of the sample means, and so we call this test a right-tailed test.

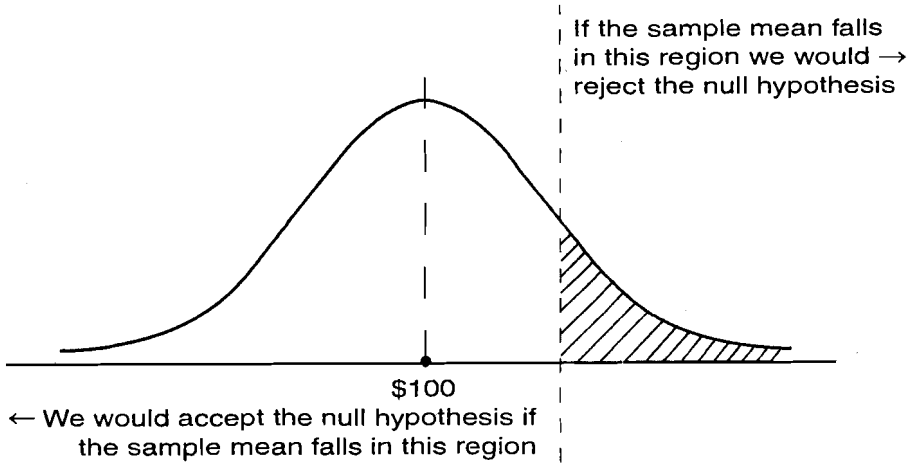


Fig. 3

Let us now summarise our discussion on the different types of tests.

The first case, where we test $H_0 : \mu = 1500$h against $H_1 : \mu \neq 1500$h, we need to use a **two tailed test**; in the second case where we test $H_0 : \mu = 1500$h against $H_1 : \mu < 1500$h, we use a **one-tailed test** (i.e. left-tailed test), and in the third case where we test μ =Rs. 100.00 against $H_1 : \mu >$ Rs. 100.00, we use a one-tailed test (i.e. a right-tailed test.).

So, in a 2-tailed test we are concerned about any difference from the hypothetical value of the parameter, whereas in a 1-tail test we are concerned with values which are either lower or higher than the hypothetical value. The difference would be clear through the examples. Try this exercise now.

E3) Radhika, a highway safety engineer, decides to test the load-bearing capacity of a bridge that is 20 years old. Considerable data are available from similar tests on the same type of bridge. Which is appropriate, a one-tailed or two-tailed test? If the minimum load-bearing capacity of the bridge must be 10 tons, what are the null and alternate hypothesis?

In the next section we shall now show how hypotheses about parameters are tested through some examples. In this unit we shall deal with only those cases, where the population is taken to be normally distributed.

We start with tests involving the parameter mean of the population.

6.3 TESTS ABOUT THE MEAN (z-test and t-test)

In this section we show how to test whether a sample is drawn from a population with a given mean. We'll also show how to test whether two samples belong to the same population or not.

6.3.1 Comparing the Mean of a Population and a Sample

We first illustrate the technique when the variance of the population is known.

Let us look at a problem.

Problem 1: The mean marks obtained by the students of a mathematics course in IGNOU is 54.5 with a standard deviation 8.0. At one of the study centres, where 100

students took the examination, the mean marks were 55.9. Are the students of this study centre significantly 1) different from 2) better than, the rest of the students of that course in IGNOU at 0.01 level?

Solution: We shall first solve (1). Let $X_1, X_2, \ldots X_{100}$ denote the random variables which are the marks obtained by the 100 students. We assume that these random variables are i.i.d as normal with mean μ and standard deviation $\sigma_0 = 8.0$ We want to test if $\mu = 54.5$ against the alternative that $\mu \neq 54.5$

Let us apply the six steps given in the earlier section.

1) Here $H_0 : \mu = 54.5, H_1 : \mu \neq 54.5$ and $\alpha = 0.01$. It is a two-tailed test. The test statistic is $\bar{x}$. Now we have to fix the critical region as discussed in the previous section. You know that if you take samples of size 100, then the sample means $\bar{x}$ are normally distributed with mean $\mu = 54.5$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{100}} = 0.8$. This means that $\frac{\bar{x} - 54.5}{0.8}$ follows a standard normal distribution. Now, since $\alpha = 0.01$ (1 % level), $\frac{1-\alpha}{2} = 0.4950$. So we find the z-value corresponding to 0.4950, which we denote by $z_{0.4950}$, such that

$$P\left[-z_{0.4950} < \frac{\bar{X} - 54.5}{0.8} < z_{0.4950}\right] = 1 - \alpha = 0.99.$$

Now we look at the normal distribution table given at the end of the block and see where the value 0.4950 is given. This is listed in the row for 2.5 and column for 0.08. That is $z_{0.4950} = 2.58$.

Therefore, we have $P\left[-2.58 < Z = \frac{\bar{X} - 54.5}{0.8} < 2.58\right] = 0.99$. Also 2.58 and -2.58 are the critical values and the accepted region is as shown in Fig. 4.

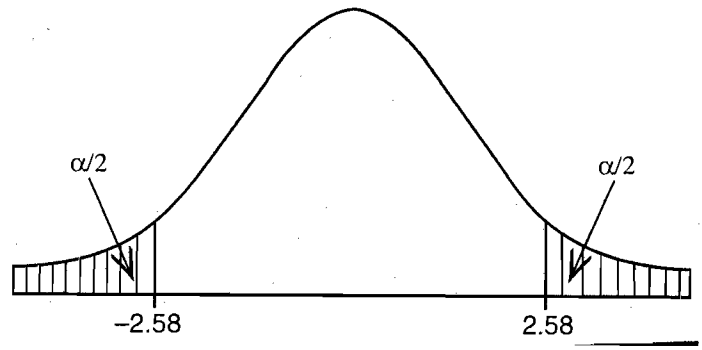


Fig.4: The total area of rejection (the shaded part) should be α

It is already given in the problem that the value of the sample statistic $\bar{x}$ is 55.9. The corresponding z-value is

$$z = \frac{55.9 - 54.5}{0.8} = 1.75$$

Since $z = 1.75$ is in the acceptance region, we accept the hypothesis, or more precisely, we fail to reject it.

This means the difference in the two mean marks is not significant enough to suggest that the students at that study centre are different from the rest of the IGNOU students.

2) Here we apply a one-tailed test.
$H_0 : \mu = 54.5, H_1 : \mu > 54.5$
Since we are interested in knowing whether the group of students is **better**, we need to look at only those values of $\bar{x}$ which are higher than 54.5. So we find the appropriate z-value such that the rejection area is the shaded portion α as shown Fig. 5 and the acceptance region is the unshaded portion. From the normal distribution table, we determine that the value of z for 40% of the area under the

curve is z = 2.33. Here we have only one critical value 2.33. Also we have

$$P\left[Z = \frac{\bar{X} - 54.5}{0.8} < 2.33\right] = 0.99.$$

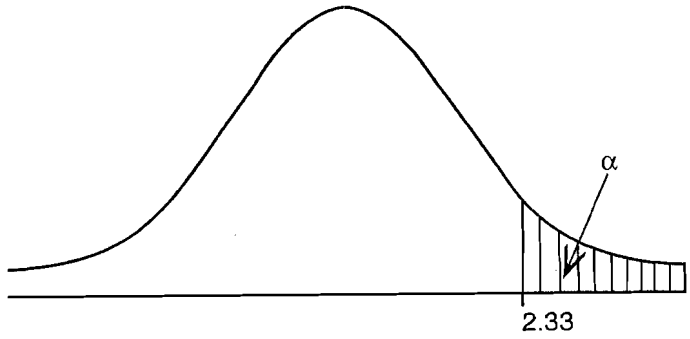


Fig.5

The shaded region is given in Fig.5. Again, the test statistic here is 1.75, which is less than 2.33 and therefore falls in the acceptance region and so we fail to reject H_0. That is, we conclude that the students from that study centre are not better than other students at 1 % level of significance.

———— × ————

You can see that the basic technique is to come up with an acceptance region by using the given level of significance (α), after noting whether it is a 1-tailed or a 2-tailed test. We then accept or reject the null hypothesis, depending on whether the given value of $\bar{x}$ falls in the region or not. You must have noticed here the similarity with the computation of confidence intervals discussed in Unit 5. There we find an interval around $\bar{x}$. Here we find an interval around μ_1, the hypothetical mean.

In the above example, α was given to be 0.01. If $\alpha = 0.05$, the only difference is in the value of z that we read from the tables. This value for a 2-tailed test is 1.96 and for a 1-tailed test is 1.64. You have been using these same values of z for calculating 95% and 99% confidence intervals in Unit 5.

In the following table, we give the critical values for both one-tailed and two-tailed tests at three significance levels, $\alpha = 0.1$, 0.01 and 0.05.

Table 2

level of significance α	0.1 (90%)	0.05 (95%)	0.01 (99%)
critical value for left-tailed test	−1.28	−1.64	−2.33
critical value for right-tailed test	1.28	1.64	2.33
critical values for two-tailed tests	−1.64 and 1.64	−1.96 and 1.96	−2.58 and 2.58

Note: - The test procedure applied in problems 1 and 2 is called **z-test**. Note that z-test is applicable when we assume that the random variables representing the sample of observations, $X_1, X_2, \ldots X_n$ are i.i.d normal(μ, σ^2) where σ^2 is assumed to be known. It is also valid if the sample size n is large i.e. $n > 30$ (by the central limit theorem) even if the common distribution is not normal or if σ^2 is unknown and has to be estimated from the sample.

Let us consider the following problem.

Problem 2: The manufacturer of an antacid claims that it relieves discomfort in 5 minutes (with a standard deviation of 2 minutes). Ten people volunteer to take it to test

the claim. The average time to get relief was 7.5 minutes. Do you accept the claim at a 10% level?

Solution: We shall assume that the time to get relief is distributed normally with unknown mean μ and known standard deviation 2. Here $H_0 : \mu = 5$mins, $H_1 : \mu \neq 5$ mins.

Here the test is two-tailed and from Table 1 we get that the critical values are 1.64 and -1.64. Also we have

$$P\left[-1.64 < \frac{\overline{X} - 5}{2/\sqrt{10}} < 1.64\right] = 0.9$$

Since $\frac{\overline{X} - 5}{2/\sqrt{10}} = \frac{7.5 - 5}{2/\sqrt{10}} = 3.95$ is greater than 1.64, does not lie in the acceptance region, so we reject H_0, and conclude that the claim is not justified at 10% level.

———— × ————

In the next problem we consider the situation where the sample size n is large i.e. $n \geq 30$ and σ is unknown whereas the sample standard deviation s is given. Note that in this situation we apply z-test by replacing σ by s.

Let us see an example.

Problem 3: : A consumer magazine, when comparing various brands of paints, stated that the drying time of one particular brand was found to be four hours. The manufacturer was not particularly pleased with this and consequently modified the paint to try to reduce the drying time. The paint was then tested by a random sample of 40 customers all of whom were decorating their living rooms. For this sample the mean drying time in hours was found to be 3.85 and the sample standard deviation was 0.55.

a) Analyse the sample data using the one-sided z-test.

b) Find a 95% confidence interval for the population mean of the drying times for the modified paint.

c) What can you conclude about the drying time of the modified paint?

Solution:

a) We want to test the hypothesis that the population mean (μ) of the drying times (in hours) of the modified paint is equal to 4. The appropriate null hypothesis is therefore

$$H_0 : \mu = 4$$

Since the manufacturer is looking for a reduction in the drying time (at least he does not expect that it should have increased!), the alternative hypothesis should be one-sided, giving

$$H_1 : \mu < 4.$$

The test statistic is

$$z = \frac{\overline{x} - 4}{\sigma}$$

Here we do not know the population variance. But we know that the sample standard deviation is 0.55. Therefore using the estimation, we can find an estimate of the population S.D as $\sigma = \frac{s}{\sqrt{n}} = \frac{0.55}{\sqrt{40}} = 0.0869626$.

Note that here the test is left-sided and $\alpha = 0.05$. From Table 1, we get the critical value as -1.64.

Since the sample mean $\bar{x}$ is 3.85, we get that $z = \dfrac{3.85 - 4}{0.869626} \sim -1.72$ (cutting to two decimal places).

Since the value of the test statistic is -1.72 is less than the critical value -1.64, we reject H_0 in favour of H_1 at the 5% significance level and conclude that it looks as though the population mean of the drying times for the modified paint is less than four hours as the manufacturer hoped. Of course, this result only applies to paint drying in living rooms.

b) A 95% confidence interval for μ is given by $(\bar{x} - 1.96\ s/\sqrt{n}, \bar{x} + 1.96\ s/\sqrt{n})$ which is

$$[3.85 - (1.96 \times 0.0869626), 3.85 + (1.96 \times 0.0869626)]$$

giving

$$[3.67,\ 4.02]$$

c)

Fig.6

As in the case of the earlier problem, it is always a good idea to follow up a hypothesis by finding a confidence interval since such an estimate provides useful extra information. Of course, with some hypothesis tests you may not be able to do this with one reason or other.

You can try some exercises now.

E4) The breaking strengths of cables made by a company had a mean of 1800 N. The company then adopted a new technique which is believed to increase the breaking strengths. 50 cables made by the new technique were tested to see if the belief is justified. or not. The mean breaking strength of these 50 is found to be 1850N with a standard deviation of 100N. Is the belief justified at a) 5% level b) 1% level.

E5) As part of a survey on drivers' reaction times for a driving magazine, 300 drivers were subjected to the following test : each driver was asked to press a lever with his/her foot in response to a flashing light. The reaction times (in seconds) were recorded and the sample mean was found to be 0.83. The sample standard deviation was 0.31. What can you conclude about drivers' reaction times?

So far we have seen that z-test can be used when σ is known whether the sample size is large or small. But when the sample size is small, z-test can not be applied when σ is not known and in the case we use a test based on t-distribution instead of normal distribution. In the following problem we illustrate the use of t-distribution.

Problem 4: A machine manufactures standard weights to be used in weighing scales. To check if the machine is working properly, a random sample of five 2-kg. weights was taken. Each 2-kg. weight was weighed on a special scale and the actual weights were found to have a mean of 1.962 kg. and a standard deviation of 0.038 kg. If $\alpha = 0.05$, can you say that the machine is in proper working order?

Solution: Assume that the random observations $X_1, X_2, \ldots X_5$ are distributed as i.i.d normal (μ, σ^2) Our null hypothesis is that the machine is in proper working order. That

is $H_0 : \mu = 2\text{kg}, H_1 : \mu \neq 2\text{kg}$.
Note that here we do not know the population standard deviation (This means that we are testing a composite hypothesis against a two sided composite alternative.)

Now we form the test statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.962 - 2}{0.038/\sqrt{5}} = -2.24$
This follows a t distribution with 4 d.f. Now this is a 2-tailed test. So we need to get $t_{.975}$ for 4 d.f. from the t- table. We see that $t_{.975} = 2.776$. So the acceptance region is $(-2.776,\ 2.776)$ and -2.24 falls in this region.

So, we conclude that at 5% level of significance, the mean of the weights manufactured by the machine is 2Kg. That is, we say that the machine is in proper working order. In other words, we say that we have not found sufficient evidence to suggest that the machine is not working properly.

———— × ————

The situation in the above example called for a 2-tailed test because we were interested in testing against a two sided alternative. whether the weights were any **different** from 2 kg. The next example presents the case for a 1-tailed test.

Problem 5: A management school claims that the starting salaries for its graduates average Rs.10,000 or more per month. A random sample of 7 students who had recently graduated, showed an average salary of Rs.9700 with a standard deviation of Rs.306. At a 5% level of significance would you accept the claim?

Solution: Here you would be interested in checking if the average salaries are **less than** claimed. (If the average is more than Rs.10,000, the claim is justified of course). So

$$H_0 : \mu = \text{Rs.}10,000,\ H_1 : \mu < \text{Rs.}10,000$$

Here $t = \frac{9700 - 10,000}{306/\sqrt{7}} = -2.59.$

Since we have to apply a 1-tail test to check the lower end, we have to get a lower limit of t corresponding to 5% level with 6 d.f. Now, to get $t_{0.05}$, we find $t_{.95}$ with 6 d.f. and take its negative. So, $t_{.05} = -1.943$. Since $t = -2.59$ falls in the rejection region or the critical region, we reject H_0. This means that the claim is not justified at 5% level.

———— × ————

See if you can do this exercise now.

E6) The specifications for the production of a certain alloy call for 23.2% copper. In 10 analyses, the mean copper content was found to be 23.5 % of 0.24%. Can we conclude that the product meets the specifications if $\alpha = 0.05$?

E7) The diameters of bolts manufactured by a machine are known to have a standard deviation of 0.0002 cm. A random sample of 10 bolts has an average diameter of 0.5046 cm. Test the hypothesis that the true mean diameter of bolts is 0.51 cm, using $\alpha = 0.01$.

In all the examples and exercises till now, we have been using a function of the sample mean to test hypothesis about the population mean. Suppose now we want to compare the effect of two medicines in providing relief from pain. The duration for which a medicine provides relief from pain can be observed and we may assume that the duration for medicine 1 is a random variable X that is distributed as normal with mean μ_X and standard deviation σ_X. Similarly the duration for which medicine 2 provides relief is a random variable Y which is assumed to be distributed as normal with mean

μ_Y and standard deviation σ_Y. We want to test the null hypothesis (the hypothesis of no difference) $H_0 : \mu_X = \mu_Y$ against the alternative that H_0 is not true. We discuss similar problems in the next sub-section.

6.3.2 Difference in the Means of Two Populations

Here we shall be dealing with two populations or two variables, X_1, X_2 assumed to be normally distributed with means μ_1, μ_2 and variances σ_1^2 and σ_1^2, respectively. We shall have to divide our discussion in two parts: 1) when σ_1, σ_2 are known, and 2) when σ_1, σ_2 are unknown. In Sec.6.3.1 you have seen that the test statistic for the problems discussed there follows a standard normal distribution when the standard deviation is known. You have also noted that it follows the t-distribution in case the standard deviation is not known.

We start with the first case now.

σ_1, σ_2 **known**:

We shall illustrate the method through an example.

Example 1: Suppose we want to investigate the following:

"Was there a difference in the performance of male and female students of IGNOU in the examination for Mathematics Elective courses in a particular year, say 2000? If so, what was the difference?

Let us see how we can use z-test to find an answer to this question.

We shall first set up the null and alternate hypothesis in terms of the population means as follows:

H_0: The mean examination mark for the population of all male students is the same as the mean examination mark for the population of all female students.

H_1: The mean examination mark for the population of all male students is not the same as the mean examination mark for the population of all female students.

We can now introduce some symbols that enable us to express these hypotheses more concisely. We let

μ_m denote the population mean of the marks of all male students

and

μ_f denote the population mean of the marks of all female students.

Then H_0 and H_1 become

$H_0 : \mu_m = \mu_f$
$H_1 : \mu_m \neq \mu_f$.

We also introduce now some other symbols that will be useful in our analysis of the sample data. Let

σ_m and σ_f denote the population standard deviations of the marks of all male and female students respectively;

$\overline{X}_m$ and $\overline{X}_f$ denote the two sample means;
s_m and s_f denote the two sample standard deviations;
n_m and n_f denote the two sample sizes.

Now suppose we take a sample of 150 male students and 100 female students. Then the table in the next page gives the sample means and sample s.d.'s for the samples taken. We may use $\frac{\overline{X}_m - \overline{X}_f}{SSD((X)_m - \overline{X}_f)}$ as the test statistic, where SSD denotes the estimate $\sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$.

	Sample size	Sample mean	Sample standard deviation
Male	$n_m = 150$	$\bar{x}_m = 57.88$	$s_m = 20.00$
Female	$n_f = 100$	$\bar{x}_f = 63.73$	$s_f = 22.10$

Here, we need to find the distribution of $\frac{\bar{X}_m - \bar{X}_f}{SSD((\bar{X}_m) - \bar{X}_f)}$. In our problem we have assumed that the marks obtained by the male or female students follow a normal distribution. Therefore the distributions of the sample means are also normal with appropriate means and standard deviations. Also $\bar{X}_m - \bar{X}_f$ will have a normal distribution with an appropriate mean and standard deviation. However the distribution of $\frac{\bar{X}_m - \bar{X}_f}{SSD((\bar{X}_m) - \bar{X}_f)}$ is unknown but can be taken approximately as standard normal by using the central limit theorem, provided the sample sizes n_m and n_f are large, say more than 30. From Unit 4 you know that the sampling distribution of the mean $\bar{X}_m$ is normal with mean μ_m and standard deviation $SE = \sigma_m/\sqrt{n_m}$. The standard deviation is given by

$$SE = \frac{\sigma_m}{\sqrt{n_m}} = \frac{\sigma_m}{\sqrt{150}}.$$

(See Fig.7)

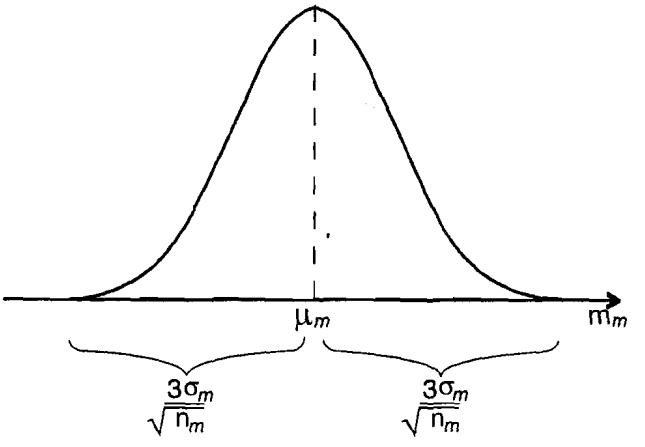


Fig. 7 Sampling distribution of $\bar{X}_m$

Similarly the sampling distribution of $\bar{X}_f$ is normal with mean μ_f and standard deviation

$$SE = \frac{\sigma_f}{\sqrt{n_f}} = \frac{\sigma_f}{\sqrt{100}}.$$

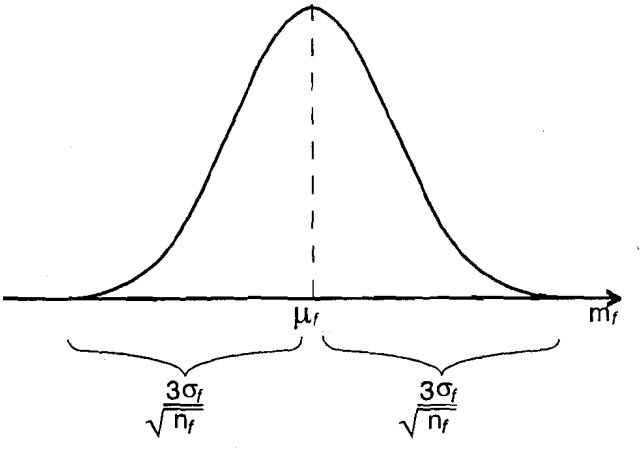


Fig. 8: Sampling distribution of $\bar{X}_f$

Now, consider all the possible pairs of samples of sizes 150 and 100 respectively that

we could select from the two populations of male and female students. For each of these pairs of samples (and there is a huge number of possible combinations) we can calculate the **difference between the sample means** $(\overline{X}_m - \overline{X}_f)$ and by considering all the combinations of such samples we obtain the **sampling distribution of** $\frac{\overline{X}_m - \overline{X}_f}{SSD((\overline{X}_m) - \overline{X}_f)}$. It turns out that the distribution of the statistic is also approximately Normal.

Moreover, the mean of this statistic is

$$\mu_m - \mu_f$$

and its standard deviation is given by

$$SE = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}$$

The sample estimate of the standard deviation is

$$SSD = \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$$

and it is this estimate that we shall use here. The sampling distribution of $(\overline{X}_m - \overline{X}_f)$ is given in Fig.9.

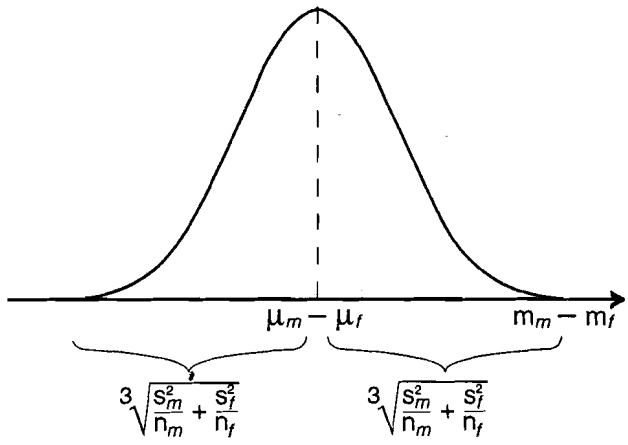


Fig.9 Sampling distribution of $\overline{X}_m - \overline{X}_f$

Before we proceed further we make a note.

Note: If we knew the standard deviations σ_m and σ_f then we didn't have to replace them by the sample estimates. In that case our test statistic will be given by

$$\frac{(\overline{X}_m - \overline{X}_f) - (\mu_m - \mu_f)}{\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}}$$

Now we apply the z-test based on the statistic $\overline{X}_m - \overline{X}_f$ to test the given null hypothesis. The hypotheses under consideration are

$$H_0 : \mu_m = \mu_f \text{ and } H_1 : \mu_m \neq \mu_f$$

and, because

$$\mu_m = \mu_f$$

can be rewritten as

$$\mu_m - \mu_f = 0,$$

we can express H_0 and H_1 in terms of the difference between μ_m and μ_f as follows

$$H_0 : \mu_m - \mu_f = 0$$
$$H_1 : \mu_m - \mu_f \neq 0.$$

Now we just apply the z-test discussed in the earlier section replacing μ by $\mu_m - \mu_f$ and $\overline{X}$ by $\overline{X}_m - \overline{X}_f$.
Hence the test statistic is

$$z = \frac{(\overline{x}_m - \overline{x}_f) - (\mu_m - \mu_f)}{SE}$$

and, since the null hypothesis is $\mu_m - \mu_f = 0$, this simplifies to

$$z = \frac{\overline{x}_m - \overline{x}_f}{SE}$$

where, $SE = \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$.

Now from the data given in Fig. we can calculate the z-value. We have

$$\begin{aligned} SE &= \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}} \\ &= \sqrt{\frac{(20.00)^2}{150} + \frac{(22.10)^2}{100}} \\ &= \sqrt{2.6666667 + 4.8841} \\ &= \sqrt{7.5507667} \\ &= 2.7478658 \\ \text{and } \overline{x}_m - \overline{x}_f &= 57.88 - 6373 = -5.85 \\ \text{so } z &= \frac{-5.85}{2.7478659} \sim -2.212 \end{aligned}$$

Hence the test statistic is $z = -2.12$ and now the procedure is exactly the same as it was in previous Section.

Since the test statistic -2.12 is less than the critical value -1.96, we reject H_0 in favour of H_1 at the 5% significance level and conclude that $\mu_m - \mu_f$ does not seem to be equal to zero. Indeed, because -2.12 is less than -1.96, it looks as though $\mu_m - \mu_f$ is less than zero and this means that μ_m seems to be less than μ_f. This suggests that there is a difference in examination performance between the sexes; it seems that the females perform better than the males.

* * *

We shall now summarise the steps used in the example above, for conducting the test.

Our aim here is to test whether two given normally distributed random variables have the same mean or not. For this, we first state H_0 and H_1 as

$$H_0 : \mu_1 = \mu_2 \text{ i.e. } (\mu_1 - \mu_2) = 0; \quad H_1 = \mu_1 \neq \mu_2.$$

Then take a random sample of size n_1 from the first population and a random sample of size n_2 from the second. We find the means of these samples: $\overline{X}_1$ and $\overline{X}_2$. Here The difference $\overline{X}_1 - \overline{X}_2$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
So, if H_0 is true, then the test statistic $Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is distributed as standard

normal.
So we find $Z_{1-\alpha/2}$ where α is the level of significance. We then reject H_0 if $Z > Z_{1-\alpha/2}$, or $Z < -Z_{1-\alpha/2}$.

If we have a situation where a one-tailed test is to be applied, say,

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2,$$

then we find $Z_{1-\alpha}$ and reject H_0 (that is, accept H_1) if $Z < -Z_{1-\alpha}$.

Study the following examples carefully now, so that you can solve the exercises which come later.

Problem 6: Two machines are used to fill cans with 200 ml. of a drink. The filling processes are assumed to be normal, with standard deviations $\sigma = 0.2$ and $\sigma = 0.25$. The quality control department wants to check if the two machines fill the same volume. A random sample is taken from the output of each machine:

M_1	M_2
200.3 200.1 200.4 198.9 200.5	200.2 2003 199.7 200.4 199.4
199.2 200.5 200.2 200.2 199.5	200.2 200.1 200.1 199.8 200

what is your conclusion? Use $\alpha = 0.05$.

Solution: Note that the standard deviations are assumed to be known. We state the hypothesis as

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

The test statistics is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

From the data given in the two samples we have
$\overline{x}_1 = 199.98$ and $\overline{x_2} = 200.02$

$$z = \frac{199.98 - 200.02}{\sqrt{\frac{0.2}{10} + \frac{0.25}{10}}} = 0.395$$

Now, for $\alpha = 0.5, z_{1-\alpha/2} = 1.96$. Since $z < z_{1-\alpha/2}$, we accept H_0.

That is, we have not found any evidence to suggest that the two machines fill different volumes of the drink.

Problem 7: Two different formulations of petrol are tested to study their road octane numbers. The variance for Formulation 1 is $\sigma_1^2 = 1.8$ and for Formulation 2 is $\sigma_2^2 = 1.2$. Two random samples of size $n_1 = 15$ and $n_2 = 20$ are taken. The mean road octane numbers are $\overline{x_1} = 89.6$ and $\overline{x_2} = 92.5$. Can you say that Formulation 2 produces a higher road octane number than Formulation 1 if 1) $\alpha = 0.05$, 2) $\alpha = 0.01$?
Solution: Here $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$.

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_1}}} = \frac{89.6 - 92.5}{\sqrt{\frac{1.8}{15} + \frac{1.2}{20}}} = -6.84$$

1) For $\alpha = 0.5, z < -z_{.95} = -1.64$. Therefore we reject H_0, that is accept H_1.

2) For $\alpha = 0.1, z < -z_{.99} = -2.33$. Therefore we again reject H_0, and accept H_1.

So, at both the levels we say that Formulation 2 produces a higher road octane number than Formulation 1.

———— × ————

If you have followed these examples, you would certainly be able to these exercises.

E8) A new teaching technique is to be tested. A group of 22 students were taught in the traditional way. Another group of 18 students was taught with the help of the new technique. The two groups were then given a standardised test which is known to have a standard deviation of 25. The mean score of the traditional group was 127 and that of the experimental group was 136. If $\alpha = 0.1$, do you think that the new technique is significantly better?

E9) A psychologist gave a test to decide if male students are as smart as female students. The sample of 40 female students had a mean score of 131 and the sample of 36 males had a mean score of 126. The test has a standard deviation of 16. Is there a difference at 0.01 level of significance?

We have been considering cases where σ_1 and σ_2 are known. If they are not known, they have to be estimated from the sample. If the samples are large, then these estimates are quite close to the real values and so we can use them in forming the test statistic Z. In the next exercise you see one such situation.

E10) A sample of 100 electric light bulbs produced by manufacturer A showed a mean life-time of 1190h and a standard of 90h. A sample of 75 bulbs produced by manufacturer B showed a mean life-time of 1230h and a standard deviation of 120h. a) Is there a difference between the two brands of bulbs at a significance level of 0.05? b) Are the bulbs of manufacturer B superior to those of manufacturer A at the same level?

Now we come to the second case. Here we shall see how to test for the difference between means when the population variance is not known and the sample sizes are small.

σ_1, σ_2 unknown

Before we go any further, we must make an additional assumption. The population variances are unknown here, but whatever they are, we are going to assume that they are equal. This is because, the situation becomes very complicated if the population variances are not equal, and in this course we are not yet ready to tackle it. The common but unknown population variance has to be then estimated from the samples. So, obviously, we have to compute s_1^2 and s_2^2 which are unbiased estimates of σ^2. Now, though we have assumed the populations variances to be equal. s_1^2 and s_2^2 may not be equal. Therefore, we use them to form a pooled variance, which we then take as a single estimate of σ^2.

We form $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_2 + n_2 - 2}$, and then compute the test statistic

$t = \dfrac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, which is distributed as student's t with $n_1 + n_2 - 2$ degrees of freedom. Here is an example to illustrate this procedure.

Problem 8: Suppose we have to choose between two types of paying surfaces to be used on a highway. One of the considerations is the stopping distance of cars. The

distance taken by cars travelling at 80 km. per hour to come to a complete stop was measured. The results (in meters) were as follows:

Surface A : $n_1 = 8, \bar{x}_1 = 42.3, s_1^2 = 38.8$

Surface B : $n_2 = 8, \bar{x}_2 = 43.2, s_2^2 = 51.$

Is there a difference in the stopping distance between the two surfaces at 0.05 level of significance?

Solution: $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$

Now, $s_p^2 = \dfrac{7(38.8 + 51)}{14} = 44.9$

Then $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{42.3 - 43.2}{\sqrt{44.9}\sqrt{\dfrac{1}{8} + \dfrac{1}{8}}} = -0.25$

The degrees of freedom are $8 + 8 - 2 = 14$. This is a two-tailed test. So, $t_{.975}$ for 14 d.f. is 2.14. Since t is within the limits, -2.14 and 2.14, we fail to reject H_0. Therefore, we conclude that at the given level, there is no difference between the two types of surfaces.

———— ×————

We have been testing for the equality of the means of two normally distributed populations by using the sample means. The samples that we had considered were independent samples. That is, the individuals in one sample were not matched or paired with those in the other in any way.

We now look at some situations where the individuals in the samples are paired or the samples are dependent.

For example, suppose the students of a particular class are given a crash course in speed reading. To test whether their reading speed has really increased or not, we need to take a sample and note the reading speeds of the students in this sample both before and after the course. In this case, the two data sets (comprising the before and after reading speeds) are not independent since each observation in the before set is matched with an observation in the after set. Here we compare the score of each student before he/she took the course to his/her score after the course, and try to find out if there is a pattern. If 60% of the students show an increased speed, is it reasonable to call the course a success? We can answer such questions with the help of t-distribution.

We start by calculating the difference for each of the pairs. In this particular in example? It means that we take the difference between the reading speeds of each student. We then assume that these values (differences) are normally distributed. Next, we calculate the mean and standard deviation from the sample and apply the t-test. See the following situation.

Problem 9: The following table gives the data on the speeds of 10 students:

Table 3

Before	9.4	10.3	8.4	6.8	7.8	9.8	9.2	11.2	9.4	9.0
After	9.3	10.6	8.8	7.0	7.7	10.0	9.8	11.7	9.7	9.0
Difference	-0.1	0.3	0.4	0.2	-0.1	0.2	0.6	0.5	0.3	0.0

Can you say that the reading course is a success at 0.05 level of significance?

We denote the mean difference in the speeds of the population by $\bar{D}$ and that of the sample by $\bar{d}$.

Solution: Suppose that $H_0 : \bar{D} = 0$ (i.e. there is not difference in the reading speeds

before and after). $H_1 : \overline{D} > 0$.

Here $\overline{d} = 0.23$, and the standard deviation for the differences (S_α) is 0.241. n = 10. The test statistic is:

$$t = \frac{\overline{d}}{S_d/\sqrt{n}}$$

$$= \frac{0.23}{0.241/\sqrt{10}} = 3.02.$$

This is a one-tailed test. Therefore $t_{0.05}$ for 9 degrees of freedom is 1.83. Since t=3.02 is more than this upper limit, we reject the null hypothesis and accept the alternative one. So, we can conclude that the speed reading course is helpful in increasing the reading speeds of students.

———— ×————

We are sure you can do these exercises now.

E11) We want to test the effect of a new fertiliser on wheat production. For this, 24 plots of land of equal area were chosen. Half of these were treated with the new fertiliser and the other half were treated with old one. With the new fertiliser, the mean yield was 48 kg. with a standard deviation of 4 kg. With the old fertiliser, the mean yield was 51 kg, with a standard deviation of 3.6 kg. Can we say at 5 % level of significance that there is an improvement in the yield because of the new fertiliser? What will be your conclusion at 1 % level?

E12) A botanist was interested in knowing if there was a difference in the time fruits matured on different parts of a plant, and recorded the day of the first fruit on the top and on the bottom for 15 plants. All the fruits came out during the same month.

Top	3	6	7	5	8	9	10	10	7	8	6	9	10	12	4
Bottom	7	9	5	8	8	10	11	12	6	9	7	13	8	13	8

Is there a significant difference in the time to mature at the 1% significance level?

E13) The pulse rates of 12 people were recorded before and after taking a new drug.

Before	68	71	84	93	67	74	82	77	71	83	62	66
After	71	70	81	97	73	80	90	76	80	79	80	67

Using 10% level, can you say that there is a significant increase in the pulse rate?

When you have solved E10, you might have found that you reject the hypothesis at 5 % level, but accept it at 1 % level. In such a case we say that the results are probably significant, that is, the new fertiliser is probably better than the old one. But we need to get more evidence to make a decision. After these numerous tests about the mean we now turn our attention to the variance of a population.

6.4 TEST ABOUT THE VARIANCE

You must have seen in Unit 4 that if the population is normally distributed with mean μ and variance σ^2, then the ratio $\frac{(n-1)s^2}{\sigma^2}$ is a χ^2 variable with (n-1) d.f. Now here we are again going to use the χ^2 distribution to test if a given sample could have come from a population with a given variance, σ^2. Of course, before applying this test, we should make sure that the population is normal. Actually, even when we discussed the

tests about the mean of a population, we had assumed the population to be normal. But we have to be especially careful in this case, because this test is particularly sensitive to the shape of the distribution. If we apply the test to the variance of a non-normal population, then chances are that we may be committing mistakes much more frequently than the α value indicates.

Suppose we have a sample of size n from a normal population. For this test our null hypothesis is that it comes from a population with a given σ^2. We form the test statistic, $\frac{(n-1)s^2}{\sigma^2}$. Now we compare the value of the test statistic with χ^2_α for a significance level of α. If the value of the test statistic is greater than or equal to χ^2_α, then we reject H_0. Otherwise we accept it.

Problem 10: : A machine is used to fill 2-kg packages of rice. It is known to have a standard deviation of 12.5 g. To check if the machine has become more erratic, a sample of 20 packages was taken. This showed a standard deviation of 16 g. Is the increase in variability significant at 1) 0.05 and 2) 0.01 levels?

Solution: $H_0 : \sigma = 12.5$g., $H_1 : \sigma > 12.5$g.

Now, $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19(16^2)}{(12.5)^2} = 31.13$

Here we have to use a one-tailed test.

1) $\chi^2_{.05}$ for 19 d.f. is 30.1. So, we reject H_0 at 0.05 level.

2) $\chi^2_{.01}$ for 19 d.f. is 36.2. So, we fail to reject H_0 at 0.01 level.

Because of the nature of conclusions in 1) and 2), we say that the variability of the machine has probability increased, and recommend that the machine should be examined. (Also see E)

———— × ————

We now ask you to do a few exercises.

E14) It has been found that the variability of driving speeds among drivers (and not speeding) is the main cause of accidents. It has also been found that the optimum standard deviation of speeds on a highway is 3 km. per hour. A sample of speeds of 16 cars was taken and its $ was found to be 14. Is this variability greater than the optimum at 5 % level of significance?

E15) The quality of printing paper depends on the variation in thickness. According to the specifications of a printer, the optimum variance of thickness is 0.0022cm. A sample of 13 readings of the thickness of the paper supplied showed a S^2 of 0.0031 cm. Should the printer reject the paper at 1 % level of significance?

We shall take up the last category of tests that we are going to discuss here: tests about proportions.

6.5 TESTS ABOUT THE POPULATION PROPORTION

In many applications we come across a binomial variable. These are the situations in which the population is divided in exactly two categories: male female, good bad, acceptable defective, and so on. Here we are interested in the proportion of one of the categories. Again, to estimate the proportion in the population we take the help of a random sample. You have already seen this in Unit 5. Even though the variable is binomial, p is still found to follow a normal distribution. This is of course true, when the sample size is large and p is not too close to either 0 or 1. We use this fact now to

test the hypotheses about the population proportion. You must have realised by now that tests of hypotheses run exactly parallel to the computation of confidence intervals. Before we illustrate the procedure through some examples, here are a few points to remember.

1) If π is the proportion of a certain category in the population, and p is that calculated from the sample, then under the above restrictions, $z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ is a standard normal variable.

2) If two samples of sizes n_1 and n_2 are drawn from the same population, and if p_1 and p_2 are the estimates of π obtained from them, then $z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ is a standard normal variable, where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Let us consider this problem.

Problem 11: : A private gallery purchased a rare painting, expecting that it would attract 75 % of its visitors. To verify this, a sample of 60 people was taken. It was found that out of these, 35 had looked at the painting. Do you think that the expectation was justified at 5% level?

Solution: $H_0 : \pi = 0.75, H_1 : \pi \neq 0.75$

Here $n = 60$ and $p = \frac{35}{60} = 0.58$

$$z = \frac{0.58 - 0.75}{\sqrt{\frac{(0.75)(0.25)}{60}}} = -3.036$$

Since this value is outside the two-tailed 5% limits, ± 1.96, we reject the hypothesis and conclude that the expectation was not correctly estimated.

———— × ————

Problem 12: Two types of computer systems are being tested for use in a new gun. The first system gave 250 hits out of 300 rounds, and the second one gave 182 hits out of 240 rounds. At 1% level can you say that the two systems differ?

Solution: $H_0 : \pi_1 = \pi_2, H_1 : \pi_1 \neq \pi_2$

Here $n_1 = 300, n_2 = 240, p_1 = \frac{250}{300}, 0.833, p_2 = \frac{182}{240} = 0.758$

Now, $p = \frac{250 + 182}{300 + 240} = 0.8$

The test statistic $z = \frac{0.833 - 0.758}{\sqrt{(0.8 \times 0.2)(\frac{1}{300} + \frac{1}{240})}} = 2.17$

The two-tailed limits at 1% level are ± 2.58. Since z lies within these limits, we cannot reject H_0. Therefore, we conclude that the two systems do not differ.

———— × ————

In the two examples above, we have shown how to test the hypotheses about proportions in two-tailed situations. We are sure you will be able to modify the procedure in case you have to deal with a one-tailed situation. You can check it out by doing these exercises now. We have included two-tailed as well as one-tailed situations here. In case you have a problem, you know you can always find the solution at the end of the unit.

E16) Two different machines are used to make plastic bottles. Two random samples,

each of size 500 are selected from the bottles produced by the machined. The sample from the first machine was found to contain 250 defective bottles, while that from the second machine contained 40 defective bottles. Is it reasonable to say that both the machines produce the same proportion of defectives if we use $\alpha = 0.05$?

E17) A random sample of size 1000 from machine 1 contained 20 defectives, and a random sample of size 1500 from machine 2 contained 40 defectives. If $\alpha = 0.05$, can you say that machine 1 is better than machine 2?

E18) A flue vaccine was given to 125 of a total of 200 employees of a firm. Thirty employees who had received the vaccine were down with flue, while 25 of those who did not, also were stricken. At 1% level of significance would you say that the vaccine was effective?

That brings us to the end of this unit. We now summarise our discussion.

6.6 SUMMARY

In this unit we have discussed

- some test procedures that are called significance test or hypothesis test for making inference about the population parameter using sample results. These tests are applied when we have a claim/hypothesis about the population parameter.
- null and alternate hypothesis, significance level, test statistic, rejection region as the basic elements of these tests.
- two types of errors - Type 1 and Type 2 errors
- when to use one-tailed or two-tailed tests.
- how to apply z-test and t-test.

6.7 SOLUTIONS/ANSWERS

E1) H_0: It is not a cough medicine.
H_1 : It is a cough medicine.

Suppose it is actually true that it is not a cough medicine i.e. H_0 is true, then if you are rejecting this hypothesis, i.e. you are making a decision that it is a cough medicine. Then this will lead to Type 1 error.

Suppose it is actually true that it is a cough medicine (i.e. H_0 is false) and you are accepting the hypothesis and decide not cough medicine then this will lead to Type II error.

E2) No.

E3) The engineer would be interested in whether a bridge of this age could withstand minimum load-bearing capacities necessary for safety purposes. She therefore wants its capacity to be above a certain minimum level, so a one-tailed test would be appropriate. The hypotheses are

$H_0 : \mu = 10$ tons, $H_1 = \mu > 10$ tons

E4) $H_0 : \mu = 1800$N
$H_1 : \mu > 1800$N

$|z| = \left|\frac{\bar{x} - \mu}{s/\sqrt{n}}\right| = \left|\frac{1850 - 1800}{100/\sqrt{50}}\right| = 3.5355$ The critical value for one-tailed test at

5% level is 1.64 and that at 1% level is 2.33. Since $|z|$ is greater than both these, we reject the hypothesis (H_0) and accept the alternative hypothesis (H_1) at both levels. So the new technique is effective.

E5) $\bar{x} = 0.83$ sec., $s = 0.31$ sec., $n = 300$

The 95% C.I. for μ is

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$$

$$= 0.83 \pm 1.96\frac{0.31}{\sqrt{300}} = 0.83 \pm 0.035$$

$$= (0.795, 0.865)$$

With 95% confidence we can say that the mean reaction time of drivers is between 0.795 and 0.865 seconds.

E6) d.f. =9 $H_0 : \mu = 23.2\%$

$H_1 : \mu \neq 23.2\%$.

The critical value of t with 9 d.f. at $\alpha = 0.05$ is 2.26

$$|t| = \left|\frac{\bar{x} - \mu}{s/\sqrt{n}}\right| = 3.9528$$

Therefore we reject H_0. The product does not meet the specifications.

E7) $\sigma = 0.0002$ cm., $n = 10$, $\bar{x} = 0.5046$, d.f. = 9 and $\alpha = 0.01$

$H_0 : \mu = 0.51$ cm.

$H_1 : \mu \neq 0.51$ cm.

This is a 2-tailed test.

$$|t| = \left|\frac{\bar{x} - \mu}{s/\sqrt{n}}\right| = \left|\frac{0.5046 - 0.51}{0.0002/\sqrt{10}}\right| = 85.3815$$

The critical value of t with 9 d.f. at $\alpha = 0.01$ is 2.821. Hence we reject H_0.

$\therefore$ the true mean diameter is not 0.51 cm.

E8) $n_1 = 22$ $n_2 = 18$, $\bar{x}_1 = 127$, $\bar{x}_2 = 136$ $\sigma = 25$ and $\alpha = 0.1$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_2 > \mu_1$

$$|z| = \left|\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| = \left|\frac{127 - 136}{25\sqrt{\frac{1}{22} + \frac{1}{18}}}\right| = 1.1327.$$

The critical value of z for $\alpha = 0.1$ is 1.64.

$\therefore$ we do not reject H_0.

$\therefore$ the new technique is not better.

E9) $n_1 = 40$ $n_2 = 36$ $\bar{x_1} = 131$, $\bar{x_2} = 126, \sigma = 16$ and $\alpha = 0.01$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$$|z| = \left|\frac{\bar{x_1} - \bar{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| = \left|\frac{131 - 126}{16\sqrt{\frac{1}{40} + \frac{1}{36}}}\right| = 1.36.$$

The critical value for $\alpha = 0.01$ for a 1-tail test is 2.33.

$\therefore$ We accept H_0

So male students are as smart as female students.

E10) $n_1 = 100$, $n_2 = 75$, $\overline{x_1} = 1190h$, $\overline{x_2} = 1230h$, $s_1 = 90h$ $s_2 = 120h$ and $\alpha = 0.05$

a) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| = \left| \frac{1190 - 1230}{\sqrt{\frac{8100}{100} + \frac{14400}{75}}} \right| = 2.421.$$

The critical value is 1.96 for a 2-tail test

$\therefore$ we reject H_0

$\therefore$ There is a significant difference.

b) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_2 > \mu_1$

$|z| = 2.421$. The critical value is 1.64 for a 1-tail test at $\alpha = 0.05$.

$\therefore$ we reject H_0 and accept H_1.

$\therefore$ The bulbs of manufacturer B are superior to those of manufacturer A.

E11) $n_1 = 12$, $n_2 = 12$, $\overline{x_1} = 48kg$, $\overline{x_2} = 51kg$, $s_1 = 4kg$, $s_2 = 3.6kg$, $\alpha = 0.05$ and d.f. $= 22$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Now,

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$s_p = 3.8052$ $|t| = 1.9312$

The critical value $t_{0.05} = 1.72$

$\therefore$ we reject H_0 and accept H_1.

$\therefore$ There is improvement.

The critical value $t_{0.01} = 2.51$

$\therefore$ At $\alpha = 0.01$, we fail to reject H_0, and conclude that there is no improvement.

E12)

Top	3	6	7	5	8	9	10	10	7	8	6	9	10	12	4
Bottom	7	9	5	8	8	10	11	12	6	9	7	13	8	13	8
Difference	−4	−3	2	−3	0	−1	−1	−2	1	−1	−1	−4	2	−1	−4

$n = 15$, $\bar{d} = -1.3333$ $s = 1.955$ d.f. $= 14$, $\alpha = 0.01$

$H_0 : \overline{D} = 0$

$H_1 : \overline{D} \neq 0$

$$|t| = \left| \frac{\bar{d}}{s/\sqrt{n}} \right| = \frac{1.3333}{1.955/\sqrt{15}} = 2.6412$$

This is a 2-tailed test. $t_{0.01}$ for 14 d.f. is 2.98 $\therefore$ we accept H_0.

$\therefore$ there is no significant difference.

E13)

Before	68	71	84	93	67	74	82	77	71	83	62	66
After	71	70	81	97	73	80	90	76	80	79	80	67
Difference	-3	1	3	-4	-6	-6	-8	1	-9	4	-18	-1

Leave space

make it bigger

$\bar{d} = 3.833 \quad s_d = 5.9 \quad n = 12 \quad$ d.f. $= 11$

$H_0 : \bar{d} = 0$

$H_A : \bar{d} < 0$

$$|t| = \left|\frac{\bar{d}}{s/\sqrt{n}}\right| = \frac{3.833}{5.9/\sqrt{12}} = 2.2505$$

This is a 2-tailed test. The critical value of t for $\alpha = 0.1$ for 11 d.f. is 1.36.

Hence we reject H_0 and accept H_1.

therefore there is a significant increase in the pulse rate.

E14) $\sigma = 3$km/hr. $\quad n = 16, \quad s = 14$km/hr. $\quad \alpha = 0.05$

$H_0 : \sigma = 3$km/hr.

$H_1 : \sigma > 3$km/hr.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{15(14)^2}{9} = 326.67.$$

This is a 1-tailed test.

$\chi^2_{.05}$ for 15 d.f. is 25

$\therefore$ we reject H_0 and accept H_1.

E15) $\sigma^2 = 0.0022$ cm. $\quad s^2 = 0.0031$ cm. $\quad n = 13 \quad \alpha = 0.01.$

$H_0 : \sigma^2 = 0.0022$ cm.

$H_A : \sigma^2 > 0.0022$ cm.

$$\chi^2 = \frac{12(0.0031)}{0.0022} = 16.91$$

$\chi^2_{.01}$ for 12 d.f. is 26.2.

$\therefore$ we accept H_0. The printer should not reject.

E16) $n_1 = n_2 = 500 \quad p_1 = \frac{250}{500} = 0.5, \quad p_2 \frac{40}{500} = 0.08.$

$$p = \frac{250 + 40}{1000} = 0.29$$

$H_0 : \pi_1 = \pi_2$

$H_A : \pi_1 \neq \pi_2$

$$|z| = \left|\frac{0.5 - 0.08}{\sqrt{0.29 \times 0.71\left(\frac{1}{500} + \frac{1}{500}\right)}}\right| = 14.635$$

The critical value for $\alpha = 0.05$ in 1.96. Hence we reject H_0.

E17) $n_1 = 1000, \quad n_2 = 1500, \quad \alpha = 0.05$

$$p_1 = \frac{20}{1000} = 0.02 \qquad p_2 = \frac{40}{1500} = 0.0267$$

$$p = \frac{60}{2500} = 0.024$$

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 < \pi_2$

$$|z| = \left|\frac{0.02 - 0.0267}{\sqrt{0.024 \times 0.976\left(\frac{1}{1000} + \frac{1}{1500}\right)}}\right| = 1.0723$$

1-tailed limit for $\alpha = 0.05$ is 1.64.

$\therefore$ we accept H_0.

E18) $n_1 = 125, \quad n_2 = 75, \quad \alpha = 0.01$

$$p_1 = \frac{30}{125} = 0.24 \qquad p_2 = \frac{25}{75} = 0.333$$

$$p = \frac{55}{200} = 0.275$$

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 < \pi_2$

$$|z| = \left| \frac{0.24 - 0.333}{\sqrt{0.275 \times 0.725 \left(\frac{1}{125} + \frac{1}{75} \right)}} \right| = 1.426$$

This is a 1-tailed test. The critical value of z for $\alpha = 0.01$ is 2.33.

$|z| < 2.33$.

$\therefore$ we accept H_0.

UNIT 7 APPLICATIONS OF CHI-SQUARE IN PROBLEMS WITH CATEGORICAL DATA

Structure **Page Nos.**

7.1 INTRODUCTION

In this block, you have already studied several problems of testing of hypotheses. The tests that you have studied so far relate to problems where the sample data have been obtained from a continuous distribution, for example, the normal distribution. In practice, however, one often obtains data in which the sampled "observations" are classified into classes according to one or more attributes. For example, a sample of flowers can be classified according to their colour — some of the flowers in the sample could be white, the others could be purple. Again, suppose it is claimed that a vaccine controls a disease. To 'verify' the truth of this claim, a sample of N individuals is taken and these individuals can be classified according to two attributes — inoculated or not inoculated, and affected or not affected by the disease.

When the sampled data are classified according to one or more attributes, we say that we have a set of **categorical data**. How do we tackle the inference problems arising out of categorical data? In this unit we shall discuss the use of one of the most widely used tests, the chi-square test, in this context.

To start with, in Sec.7.2, we shall consider the use of the chi-square test in "goodness-of-fit" problems. Then, in Sec.7.3, we shall see how the chi-square test can help us compare two features of a population to see if there is any relationship between them or not. In other words, we test to see if the features occur independent of each other or not.

While studying this unit, please keep comparing the situations in this unit and Units 5 and 6 to really understand the difference in the questions being asked and answered.

Objectives

After studying this unit, you should be able to

- define categorical data;
- identify inference problems associated with categorical data;
- use the chi-square test for solving some inference problems arising in categorical data.

7.2 GOODNESS-OF-FIT

Let us begin by trying to solve Ms.Dalta's problem. She is the marketing director of a company that sells four types of steel almirahs. As part of her duties, she has to make sure that there is no loss of sales due to less stock availability. So far she has been ordering new cupboards assuming that the demand for all four types is the same.

Recently, however, the stock inventories have become more difficult to control. Therefore, Ms.Dalta feels that she should check whether her hypothesis of uniform demand is valid or not.

Can you apply any of the methods you have studied so far for helping Ms.Dalta? There is no parameter that she is estimating and no assumption regarding the distribution of the population. So Ms.Dalta needs to look for some new tools. What she needs to do is to test the hypothesis :

H_0 is the null hypothesis, and H_A is the alternative hypothesis.

H_0 : The demand is uniform for all four types of almirahs

against

H_A : The demand is not uniform for all four types of almirahs.

For doing this, she selects a sample of 80 almirahs sold over the past few months. Ms.Dalta assumes that the demand is uniform. So the probability of an almirah of Type i being bought is the same, for i = 1, 2, 3, 4. If we denote this probability by p_i, then $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. So, if the demand is uniform, she can expect $80\left(\frac{1}{4}\right) = 20$ almirahs of each type to be sold. But the observed sales of each type are 23, 19, 18 and 20, respectively. Her problem is to see how well her hypothesis of uniform demand fits the observed sales. In other words, how can this data set be used for testing H_0?

More generally, suppose a sample of n individuals are classified into k classes. Suppose the number of individuals falling in the ith class is O_i (i = 1, ..., k). The problem of "**goodness-of-fit**" consists in testing the hypothesis, H_0, that the probability of an individual falling in the ith class is p_i (i = 1, ..., k), where $\sum_{i=1}^{k} p_i = 1$. In other words, the hypothesis H_0 to be tested is that the number of individuals (in a sample of size n) falling in the ith class is np_i (i = 1, ..., k). This is to be tested against the hypothesis H_A, that H_0 is not true.

Note that, $\chi^2_{\alpha,k-1}$ is denoted by 'χ^2_α **with k-1 degrees of freedom' in Unit 4.**

So, in this general situation, the "expected" number of individuals falling in the ith class is $E_i = np_i$ (i = 1, ..., k). Note that these "expected" numbers are known to us because to start with we assume H_0 and calculate them. That is, we assume that the probability of an individual falling in the ith class is p_i (i = 1, ..., k). Based on these numbers $O_1, O_2, ..., O_k, E_1, E_2, ..., E_k$, there is a way of testing the validity of H_0. Let us see what this method is.

Let $U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. This statistic U, under some mild conditions, is known to have an **approximate** chi-square distribution with **k – 1 degrees of freedom,** where k is the number of classes. If we want to test the hypothesis H_0 at the α level of significance, then we need to find $\chi^2_{\alpha,k-1}$, from the standard χ^2 distribution tables (given at the end of this block). If $U > \chi^2_{\alpha,k-1}$, we reject H_0. Otherwise we do not reject H_0.

U is also called the **sample χ^2 value**, or the **observed value of χ^2** for the data.

To see how this test works, let us consider Ms.Dalta's data, presented in Table 1.

Table 1

Type of almirah	Observed sales (O_i)	Expected sales (E_i) $= np_i = 80 \times \left(\frac{1}{4}\right)$
I	23	20
II	19	20
III	18	20
IV	20	20

Here $U = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$

$$= \frac{(23-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(20-20)^2}{20}$$

$$= \frac{9+1+4+0}{20} = 0.7$$

Here k = 4. If Ms.Dalta wants to test H_0 at a 5% level of significance, $\alpha = 0.05$. Now, $\chi^2_{0.05,3} = 7.815$. Since $U < \chi^2_{0.05,3}$, Ms.Dalta does not reject H_0. In other words, Ms.Dalta concludes that the demand for the four types of almirahs is uniform.

Another example may help you to see how this test works.

Example 1 (Experiment on the breeding of flowers of a certain species) : Jaswant is interested in breeding flowers of a certain species. The experimental breeding can result in four possible types of flowers :

(a) magenta flowers with a green stigma (MG),
(b) magenta flowers with a red stigma (MR),
(c) red flowers with a green stigma (RG),
(d) red flowers with a red stigma(RR).

According to the well-known Mendel's law, these four kinds of flowers should come out in the ratio 9 : 3 : 3 : 1. Jaswant found that under her experiment, out of 160 flowers that bloomed, the number of flowers with types MG, MR, RG and RR were 84, 35, 28 and 13, respectively. She wants to find out whether these data are compatible with Mendel's law or not.

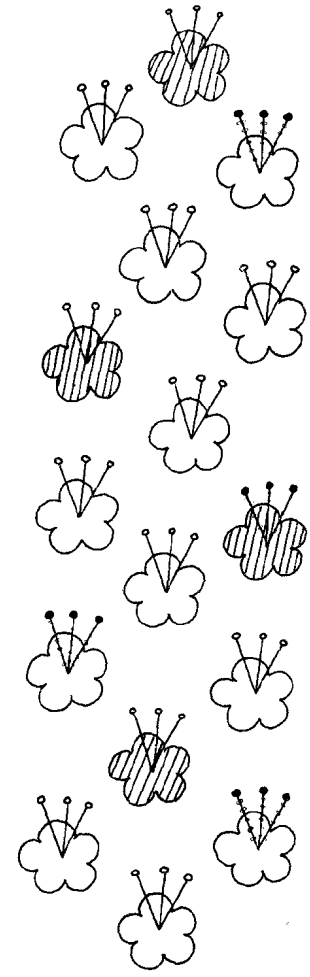


Fig.1

If they are compatible, then the probabilities of each of these types of blooming are $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, and $p_4 = \frac{1}{16}$. So Jaswant wants to test the hypothesis

H_0 : The distribution of the flower types is multinomial with $p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$.
against
H_A : H_0 is not true, that is, the distribution is not multinomial with the specified probabilities.

See the appendix to the unit for a brief introduction to the multinomial distribution.

Jaswant's data can be presented as shown in Table 2.

Table 2

Flower type	Observed number O_i	Expected number E_i (= np_i)
MG	84	90
MR	35	30
RG	28	30
RR	13	10
Total (n)	160	160

Here k = 4, and

$$U = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(84-90)^2}{90} + \frac{(35-30)^2}{30} + \frac{(28-30)^2}{30} + \frac{(13-10)^2}{10}$$

$$= 2.27$$

Jaswant needs to compare this value with the appropriate critical χ^2-value. She takes the significance level of the test as $\alpha = 0.05$. Also, in this case, since the number of classes is 4, the degrees of freedom are 4 – 1 = 3. So, she finds $\chi^2_{0.05,3}$, which is 7.81. Since U = 2.27 < 7.81 = $\chi^2_{0.05,3}$, she does not reject H_0. Thus, Jaswant concludes that her data is compatible with Mendel's law.

* * *

In the two situations above, the hypothetical probabilities p_1, p_2, ... were known to us from before because of the type of assumption H_0 was. However, in some problems, these probabilities may have to be estimated from the data itself. The following example illustrates this.

Example 2 : A consultant was employed by a city council to study the pattern of bus arrivals and departures at a very busy interstate bus terminus. Since many

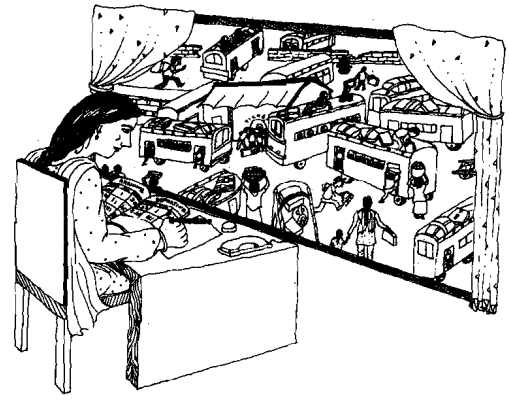


Fig.2

arrival processes fit the Poisson distribution, she decided to test the following hypothesis :

H_0 : The arrivals are distributed as a Poisson random variable,
against
H_A : The arrivals are not Poisson distributed.

She sampled the number of arrivals in 200 minutes. Then she grouped the arrivals into k = 6 categories, and noted her observations, as shown in Column 2 of Table 3 below.

However, since the parameter of the Poisson distribution is unspecified in the hypothesis, the consultant needed to estimate this from the data itself. For this she first computed the sample mean as

$$\bar{x} = \frac{(1\times 23)+(2\times 45)+\cdots+(5\times 41)}{200}$$

$$= 2.96$$

So, she estimated the parameter of the Poisson distribution as $\hat{\lambda} = 2.96$.

With this value of $\hat{\lambda}$, she computed the Poisson probabilities for the different classes from the tables (which are also provided at the end of this block). These are shown in Column 3 of the table below.

Table 3 : Arrivals at ISBT

No. of arrivals	Observed frequencies O_i	Prob. according to Poisson dist. p_i	Expected frequencies $E_i(=np_i)$
0	10	0.0524	10.48
1	23	0.1545	30.90
2	45	0.2277	45.54
3	49	0.2238	44.76
4	32	0.1651	33.02
5 or more	41	0.1765	35.30
Total	200	1.0000	200.00

According to her data,

$$U = \frac{(10-10.48)^2}{10.48} + \frac{(23-30.90)^2}{30.9} + \ldots + \frac{(41-35.3)^2}{35.3}$$

$$= 0.022 + 2.02 + 0.006 + 0.402 + 0.032 + 0.92$$

$$= 3.402$$

Here k = 6 but one parameter has been estimated. So, the degrees of freedom associated with the chi-square distribution is (k – 1) – 1 = k – 2 = 4. The critical value of chi-square at 4 degrees of freedom and 1 percent level of significance is 13.27. Since 3.402 < 13.27, the consultant did not reject the null hypothesis. In other words, she was in a position to conclude that the arrivals and departures at the bus terminus were Poisson distributed.

To find p_i for $\lambda = 2.96$, we take the average of the values in the columns corresponding to 2.9 and 3.0, respectively. Thus, $p_1 = \frac{0.055 + 0.0498}{2} = 0.0524$.

* * *

Let us now look at a problem involving normal distribution. While solving it, the following very important point about applying the χ^2-test will show up.

Remark 1 : If, corresponding to a category, say j, the expected value E_j is small, i.e., less than 5, then the chi-square approximation for the distribution of U will not be good. So, if the condition $E_i \geq 5$ is not satisfied for all i, then we should combine the category j with $E_j < 5$ with its adjacent categories j + 1, j + 2, ..., j + r, where $E_j + E_{j+1} + \ldots + E_{j+r} \geq 5$ but $E_j + E_{j+1} + \ldots + E_{j+r-1} < 5$. The number of classes, accordingly, gets reduced by r.

This remark will become more clear as you study the solution of Problem 1.

Problem 1 : A chemical company wants to know if its sales of a liquid chemical are normally distributed. This information will help them in planning and

controlling the inventory. The sales record for a random sample of 200 days is given in Table 4.

Table 4

Sales (in 1000 litres)	Number of days
Less than 34.0	0
34.0-35.5	13
35.5-37.0	20
37.0-38.5	35
38.5-40.0	43
40.0-41.5	51
41.5-43.0	27
43.0-44.5	10
44.5-46.0	1
46.0 or more	0
Total	200

We assume that the upper limit of a class shows that quantities less than that limit are in the class. So, for example, 35.5 will be included in the third class interval, not the second one.

At the 5% level of significance, test the hypothesis that the company's sales are normally distributed.

Solution : Let us start by clearly stating our hypotheses.

H_0 : The company's sales are normally distributed.
against
H_A : The company's sales are **not** normally distributed.

Now, we assume for just now that H_0 is valid. By methods known to us, we can calculate the sample mean and sample standard deviation $\bar{x}$ and s_x. You can check that these are :

$\bar{x}$ = 40,000 litres, s_x= 2.5 thousand litres.

Now, we need to find the expected frequencies E_i corresponding to each O_i. You know that $E_i = 200 \times p_i$, where p_i is the probability for each class in Table 4, computed under the assumption of normal distribution.

So, let us expand Table 4 to include all the class probabilities (Column 3), the expected frequencies (Column 4) and the corresponding values of $\frac{(O_i - E_i)^2}{E_i}$ (Column 5).

To get the first entry in Column 3, we compute $z = \frac{(x-\mu)}{\sigma}$ for x = 34. As you know, μ and σ are estimated by $\bar{x}$ and s_x, respectively. So, $z = \frac{(34-40)}{2.5} = -2.4$.

Now, from the table of normal probabilities in the Block Appendix, you know that
$P[-2.4 \le Z \le 0] = P[0 \le Z \le 2.4] = 0.4918$.
So, the probability we want is
$p_1 = 0.5 - P[-2.4 \le Z \le 0] = 0.5 - 0.4918 = 0.0082$.
Therefore, $E_1 = 200\ (0.0082) = 1.64$.

Similarly, you can compute the other expected frequencies and complete the 4th column of Table 5. You may wonder about the brackets in Columns 2, 3 and 4 of the table. This is because, as we have mentioned in Remark 1, **the χ^2 goodness-of-fit test is a good approximation only if the E_i are not very small**. This is why we have grouped the first two classes and the last two classes in Table 5.

To fill in the fifth column of Table 5, we treat the bracketed classes as a single class. So, $\frac{(O_1 - E_1)^2}{E_1} = \frac{(13-7.18)^2}{7.18} = 4.7176$. You can similarly calculate the other entries of Column 5 in the table below.

Table 5

Sales (in 1000 litres)	Observed frequency (O_i)	Class probability (p_i)	Expected frequency (E_i)	$\frac{(O_i - E_i)^2}{E_i}$
less than 34.0	0	0.0082	1.64	4.7176
34.0 – 35.5	13	0.0277	5.54 (7.18)	
35.5 – 37.0	20	0.0792	15.84	1.0925
37.0 – 38.5	35	0.1592	31.84	0.3136
38.5 – 40.0	43	0.2257	45.14	0.1015
40.0 – 41.5	51	0.2257	45.14	0.7607
41.5 – 43.0	27	0.1592	31.84	0.7357
43.0 – 44.5	10	0.0792	15.84	2.1531
44.5 – 46.0	1	0.0277	5.54	5.3193
greater than 46.0	0	0.0082	1.64 (7.18)	

Now, summing up the entries in the last column of Table 5, we get U = 15.194.

Next, to see whether we accept or reject H_0, we look up the value of χ^2 at the 5% level of significance and for the appropriate number of degrees of freedom. Note that, though we started with the data categorised into 10 classes, we needed to group two sets of 2 frequencies each. So, for purposes of the χ^2 test we now have 8 classes. Also, we have estimated two parameters, μ and σ. Therefore, the degrees of freedom are $(8 - 1) - 2 = 5$.

So, from the χ^2 table, we find $\chi^2_{0.05,5} = 11.07$.

Since $U > \chi^2_{0.05,5}$, we must reject H_0. That is, the normal distribution is not a good fit to the data.

* * *

Now try the following exercises.

E1) In Table 6 below you find the distribution of the heights for 100 college students. Estimate the mean and the standard deviation of the distribution. Check whether the sample is drawn from a normally distributed population at 5% level of significance.

Table 6

Class (cm)	Number of students (O_i)
Less than 161	4
161 - 164	11
164 - 167	16
167 - 170	19
170 - 173	25
173 - 176	18
176 - 179	4
179 - 182	2
182 or more	1
Total	100

E2) Test whether the observed frequencies, as given below, in 4 phenotypic classes AB, Ab, aB, ab are in agreement with the expected ratio 9: 3: 3: 1.

Class	AB	Ab	aB	ab
Frequency	102	25	28	5

E3) A die is rolled 1200 times with the following results :

No. that comes up	1	2	3	4	5	6
Frequency	205	279	217	257	133	109

Test if the die is unbiased.

In all the situations so far, the problem was related to data that were classified according to one attribute. Now let us see how the χ^2 test can be used to infer about situations in which the data are classified according to two or more attributes.

7.3 TEST OF INDEPENDENCE

Fig.3 :" Don't worry! You take this medicine, and you won't have any more colds in future."

In this section we shall look at inference problems like the following one.

Dr.Surya had recently developed a serum that she thought might be effective in preventing colds. But, she needed to verify its efficacy. For this purpose she carried out an experiment.

One thousand individuals were classified into two groups of the same size. The serum was administered to the members of the first group only. The number of individuals in each group who caught a cold zero times, or once, or more than once during some period after the treatment was noted. The data are shown in the following table having 2 rows and 3 columns.

Table 7 : Table showing the effect of serum

Category	The number catching a cold			Total
	Zero times	Once	More than once	
Group given serum	252	145	103	500
Untreated group	224	136	140	500
Total	476	281	243	1000

Dr.Surya's problem was to examine whether or not this serum is effective in preventing a cold. In other words, she wants to know if a person can catch a cold one or more times whether s/he has taken the serum or not. We can reword this as: is the treatment by the serum **independent** of the number of times of catching a cold?

So, Dr.Surya formulated the following null and alternative hypotheses to be tested:

H_0 : There is no interdependence between the serum treatment and the number of times of getting a cold.

H_A : H_0 is not true, i.e., the serum has some effect on preventing colds.

To test H_0 against H_A she planned to use the χ^2 test at the 5% significance level. As you know, to do so she needed to calculate the expected frequencies corresponding to each of the 6 entries in the 2×3 table, Table 7, assuming H_0, i..e., the independence of the number of times one gets a cold and of taking serum treatment.

Let us see how she obtained E_{11}. For this, she used the fact that out of the 1000 people, 476 had no cold. So, out of the 500 in the treatment group,

$\frac{476}{1000} \times 500 = 238$ were expected to not have any cold. Note that this is

$$\frac{\text{(sum of the first row entries)} \times \text{(sum of first column entries)}}{\text{(total sample size)}}$$

Similarly, she calculated the other expected frequencies :

$E_{12} = 500 \times \frac{281}{1000} = 140.5$, $E_{13} = 121.5$, $E_{21} = 238$, $E_{22} = 140.5$, $E_{23} = 121.5$.

Then , Surya calculated the sample statistic U as

$$U = \frac{(252-238)^2}{238} + \frac{(145-140.5)^2}{140.5} + \frac{(103-121.5)^2}{121.5} + \frac{(224-238)^2}{238} + \frac{(136-140.5)^2}{140.5} + \frac{(140-121.5)^2}{121.5} = 7.57$$

She took the significance level of the test as $\alpha = 0.05$. Also, in this case the number of degrees of freedom was $(2 - 1)(3 - 1) = 2$. So, comparing the value of U with $\chi^2_{0.05,2} = 5.99$, she found that $U > \chi^2_{0.05,2}$

For an m × n table, the number of degrees of freedom is (m−1) (n−1).

So, she rejected H_0, and concluded that the serum has some effect in preventing colds.

Let us look closely at the steps Dr. Surya went through for testing the independence of two features of the population under study.

Step 1 : She stated the hypothesis regarding the independence of two features of the sample.

Step 2 : As in the case of the goodness-of-fit tests, she noted the frequencies — how many of each type of person (treated or untreated) had which kind of feature (the number of times they catch a cold). These frequencies were written in a table, called a **contingency table**.

In this case, the contingency table had 2 rows and 3 columns, because corresponding to each of the two groups of people there were 3 possibilities about the cold they did or did not catch. In brief, we say that the table was a 2 × 3 contingency table.

Step 3 : Corresponding to each of the 6 cells of the contingency table, Dr.Surya calculated the expected frequency. She did this as follows :

E_{ij} = expected frequency for ith row and jth column

$$= \frac{\text{(sum of entries of ith row) (sum of entries of jth column)}}{\text{(total sample size)}}$$

where i = 1, 2 and j = 1, 2, 3.

Step 4 : Then the sample χ^2, U, was calculated by

$U = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where O_{ij} was the entry in the ith row and jth column.

Note that, more generally, if she had had an m × n contingency table, the value would be

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$, where m and n are natural numbers.

Step 5 : She compared this value with the value of $\chi^2_{\alpha,d}$, where α is the level of significance and $d = (m-1)(n-1)$ is the number of degrees of freedom. Then, as you saw in Sec. 7.2, if $U < \chi^2_{\alpha,d}$, H_0 is accepted. Otherwise H_0 is rejected.

Another example may help you to clarify your understanding regarding the process of testing for independence.

Example 4 : The Glorious Watch Company wants to find out if there is any relationship between the income of a person and the importance she attaches to the price of a brand name. Mr.Zafar, the Chief of the Marketing Division, wants to test the hypothesis

H_0 : Income of a person and importance to her of price attached are independent.
against
H_A : H_0 is not true.

Zafar does a survey among the customers. To analyse his results, he groups them into 3 income levels, and asks them to mark the level of importance they give on a 3-point scale — great, moderate or low. He noted the results in a contingency table (see Table 8). In this table, you will also find the expected frequency corresponding to each observed frequency written alongside. As you know, these will be calculated as follows :

$$E_{11} = \frac{(\text{sum of entries of first row})(\text{sum of entries of first column})}{(\text{total sample size})}$$

$$= \frac{170 \times 187}{500} = 63.58.$$

All the other E_{ij}s are calculated in the same way.

Table 8

Feature 1 (Importance Level)	Feature 2 (Income)						Total
	Low		Middle		High		
	O_{i1}	E_{i1}	O_{i2}	E_{i2}	O_{i3}	E_{i3}	
Great	79	63.58	58	61.2	33	45.22	170
Moderate	48	59.09	65	56.88	45	42.03	158
Low	60	64.33	57	61.92	55	45.75	172
Total	187		180		133		500

So, the sample χ^2 value that Zafar calculated was

$$U = \frac{(79-63.58)^2}{63.58} + \frac{(58-61.2)^2}{61.2} + \ldots + \frac{(57-61.92)^2}{61.92} + \frac{(55-45.75)^2}{45.75}$$

$$= 3.74 + 0.167 + 3.302 + 2.081 + 1.159 + 0.21 + 0.291 + 0.391 + 1.87$$

$$= 13.211$$

Then Zafar compared this with the value of χ^2 for (3–1) (3–1) = 4 degrees of freedom and at the 2% level of significance, which is $\chi^2_{0.02,4} = 11.668$.

He found U > $\chi^2_{0.02,4}$, which made him decide that he should reject H_0. In other words, Zafar is 98% certain that the level of income of a person is related to the importance she gives to the price of the brand of watches.

* * *

In the example above, it is interesting to note that if Zafar had chosen to be 99% certain, then $\chi^2_{0.01,4} = 13.277 > U$. So that, he would not have rejected H_0. What does this tell us about statistical analyses? Think about it.

And now here are some problems for you to solve.

E4) The data in the following table give mortality rates among vaccinated and non-vaccinated patients. Test if the vaccine has any effect in curing the disease.

Categories	Living	Dead	Total
Vaccinated	320	125	445
Non-vaccinated	98	230	328
Total	418	355	773

E5) Do the following data on sociability of soldiers recruited in cities and villages suggest that city soldiers are more sociable than village soldiers?

Sociability / Place	Sociable	Non-sociable
City	13	6
Village	7	14

E6) A group of 1650 school children were classified according to their performance in school tests and family economic level. Test if there is any association between these two attributes.

Performance / Economic level	Very Good	Good	Average	Poor	Total
Very Rich	4	7	16	25	52
Rich	13	37	79	73	202
Average	105	372	298	175	950
Poor	36	213	75	123	446
Total	157	629	468	396	1650

E7) In an experiment to study whether smoking affects health, the following data were collected. Test the hypothesis that smoking does not affect health.

	Light smoking	Moderate smoking	Heavy smoking
Health affected	16	29	35
Health not affected	36	23	17

In this section you have seen situations in which the population is tested to see if two or more common features of the population are related or not. This is as far as we intend to discuss the use of χ^2 for analysing categorical data. Let us end with a brief look at what we have covered in this unit.

7.4 SUMMARY

In this unit we have started with a look at data presented in the form of frequencies falling in different categories or classes. Based on such data we have undertaken different tests of hypotheses using the chi-squared distribution. We have considered two types of tests :

1) **Test of goodness-of-fit** : The hypotheses are given by
H_0 : The data fit a given distribution ; against
H_A : H_0 is not true, i.e., the data do not fit that distribution.

For testing whether H_0 is acceptable, we consider the observed and expected frequencies of the various categories in the data.

Suppose there are k categories with O_i as the observed frequency and E_i as the expected frequency of the ith category. Then the sample χ^2 value is

$$U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

If H_0 were acceptable, then this value should be less than $\chi^2_{\alpha, k-s-1}$ with 100α % significance level, where s is the number of parameters estimated in finding the expected frequencies.

So, if $U < \chi^2_{\alpha, k-s-1}$, then H_0 is not rejected. Otherwise, H_0 is rejected.

2) **Test of independence** : Suppose a population can be classified into r categories on the basis of feature A, and into c categories on the basis of feature B. The hypotheses are given by :

H_0 : There is no interdependence between the features A and B
H_A : H_0 is not true, that is, A has an effect on B.

The data is presented in the form of an r × c contingency table.
Let n_{ij} be the frequency in the ith row and jth column and let

$$n_{i0} = \sum_j n_{ij} \ , \ n_{0j} = \sum_i n_{ij} \ , \ n = \sum_i \sum_j n_{ij}$$

If the two classification criteria are mutually independent, the expected value E_{ij} for the ith row and jth column is given by

$$E_{ij} = \frac{n_{i0} \times n_{0j}}{n}$$

Then, the sample χ^2 value, $U = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

If this value is less than $\chi^2_{\alpha,(r-1)(c-1)}$, then H_0 is acceptable at the α level of significance.

And now you may like to check whether you have achieved the objectives of the unit listed in Sec.7.1. Also, while doing the exercises in this unit, you may have had some doubts. If so, please go through the following section also.

7.5 SOLUTIONS/ANSWERS

E1) Here $\bar{x} = 170$, $s^2 = 36$ and $n = 100$.

H_0 : The sample is drawn from a population with normal distribution $N(170.0, 6^2)$.
H_A : H_0 is not true.

In order to solve this problem by the same method as in Example 1, we consider the classes in Table 6 corresponding to categories of a multinomial distribution. Let O_i be the observed value for the ith category. Then, what is the expected value for the ith category in this case? Since the population distribution is completely specified as $N(170, 6^2)$ under the null hypothesis H_0, we can obtain the probability p_i with which the height of a student chosen randomly falls into the ith category. The expected value for the ith category is obtained by $E_i = np_i$. To compute the values p_i, the boundary points of the classes should be standardised by the population means and the standard deviation so as to make use of the table for a standard normal distribution. The standardised boundary points are given below in Table 11.

Table 11

Boundary Points of Class (x_i)	161	164	167	170	173	176	179	182
Standardised Boundary Points (z_i)	−1.5	−1.0	−0.5	0.0	0.5	1.0	1.5	2.0

Here, $z_i = \dfrac{x_i - 170.0}{6.0}$ The p_i and E_i values can be obtained as follows.

$p_1 = P[-\infty < Z < -1.5] = 1 - P[-\infty < Z < 1.5] = 0.0668$; $E_1 = 100 \times 0.0668 = 6.68$

$p_2 = P[-1.5 \leq Z < -1.0] = 0.0919$; $E_2 = 9.19$, and so on.

The values are all given in Table 12 below.

Table 12

Class (cm)	Number of students (O_i)	Probabilities (p_i)	Expected values E_i (=np_i)	
Less than 161	4	0.0668	6.68	
161 - 164	11	0.0919	9.19	
164 - 167	16	0.1498	14.98	
167 - 170	19	0.1915	19.15	
170 - 173	25	0.1915	19.15	
173 - 176	18	0.1498	14.98	
176 - 179	4	0.0919	9.19	
179 - 182	2	0.0440	4.40	6.68
182 or more	1	0.0228	2.28	
Total	100	1.0000	100	

Let us now test H_0 against H_A at the 5% significance level.

Now, from Table 12,

$$U = \frac{(4-6.68)^2}{6.68} + \frac{(11-9.19)^2}{9.19} + \frac{(16-14.98)^2}{14.98} + \frac{(19-19.15)^2}{19.15} + \frac{(25-19.15)^2}{19.15} + \frac{(18-14.98)^2}{14.98} + \frac{(4-9.19)^2}{9.19} + \frac{(3-6.68)^2}{6.68}$$

$= 8.86$

Taking the significance level of the test $\alpha = 0.05$, we have $\chi^2_{0.05,5} = 11.07$.

The degrees of freedom $8 - 3 = 5$, because the number of categories, after combining the last two categories is 8, and the number of parameters estimated is 2.

Since $U = 8.86 < 11.07 = \chi^2_{0.05,5}$, we conclude that there is good agreement between the observed frequencies and the fitted values. So H_0 is accepted.

E2) If the data are compatible with the given ratios, the expected frequencies are : AB: 90, Ab: 30, aB: 30, ab: 10.

The value of U is 5.07. This is less than $\chi^2_{0.05,3} = 7.815$. Hence, we accept the null hypothesis that the given data are in agreement with the expected ratios.

E3) Here H_0 : the expected frequency is 200 in each class.
H_A : H_0 is not true.

Therefore, $U = 112.87 > \chi^2_{0.05,5} = 11.070$. Hence, we conclude on the basis of the given data that we reject H_0. So the die is not unbiased.

E4) The hypothesis here is:

H_0 : There is no effect of the vaccine on mortality.
against
H_A : H_0 is not true.

The expected frequencies E_{ij} are given in the table below.

	Living	Dead	Total
Vaccinated	241	204	445
Non-Vaccinated	177	151	328
Total	418	355	773

The observed value of χ^2 is $U = 133.08$.
The number of degrees of freedom $= (2 - 1)(2 - 1) = 1$.
$\chi^2_{0.05,1} = 3.84 < U$.
Hence, we conclude that we cannot accept H_0. So, on the basis of the given data, we conclude that the vaccine has a definite effect on the mortality rate.

E5) H_0 : There is no interdependence between place and sociability level.
H_A : H_0 is not true.

The table of expected frequencies is

Sociability / Place	Social	Non-social	Total
City	9.5	9.5	19
Village	10.5	10.5	21
Total	20	20	40

So, $U = 12.25\left(\frac{2}{9.5} + \frac{2}{10.5}\right) = 4.9.$

The number of degrees of freedom = 1.

$\chi^2_{0.05,1} = 3.84 < U.$

Therefore, we reject H_0. So, the data suggests that the place a soldier comes from affects her/his sociability level.

E6) The expected frequencies are given below :

Performance / Economic level	Very Good	Good	Average	Poor
Very Rich	4.95	19.82	14.75	12.48
Rich	9.22	77.00	57.29	48.48
Average	90.39	362.15	269.45	228
Poor	42.44	170.02	126.50	107.04

The value of the sample χ^2 is $U = 127.61 > 25.0 = \chi^2_{0.05,15}$. Hence, the hypothesis of independence between the categories is rejected.

E7) Under the assumption that smoking does not affect health, the expected frequencies are given below.

	Light smoking	Moderate smoking	Heavy smoking
Health affected	26.67	26.67	26.67
Health not affected	25.33	25.33	25.33

The observed value of χ^2 is $U = 14.52 > \chi^2_{0.05,2} = 5.991$. Hence, it is concluded on the basis of the given data that smoking affects health.

APPENDIX : MULTINOMIAL DISTRIBUTION

This distribution is an extension of the binomial distribution that you studied in Unit 3. It shows up in the following situation :

There is an experiment which consists of n identical trials, which are independent. Each trial can have k possible outcomes. Suppose the probability of each of these outcomes is $p_1, p_2, \ldots, p_k$, with $p_1 + p_2 \ldots + p_k = 1$. These probabilities remain the same from trial to trial.

Mathematically, this situation is represented by considering k random variables $X_1, \ldots, X_k$ with probabilities $p_1, \ldots, p_k$ that $X_1 = x_1, \ldots, X_k = x_k$, respectively, where $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$, $p_i \neq 0 \forall i = 1, \cdots, k$. If the random vector $(X_1, \ldots, X_k)$ is multinomially distributed, the $P[X_1 = x_1, \cdots, X_k = x_k] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$.

So, if we are testing if a certain population is multinomially distributed, we will test

H_0 : The population of size n is multinomially distributed with probabilities $p_1, p_2, \ldots, p_k$ (known to us);
against
H_A : The population is not multinomially distributed.

As in all the other cases discussed in the unit, if the np_i are not very small, then the test statistic $U = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$ has approximately a chi-squared distribution with $(k - 1)$ degrees of freedom. The approximation is usually good for $E_i = np_i \geq 5$.

APPENDIX – 1

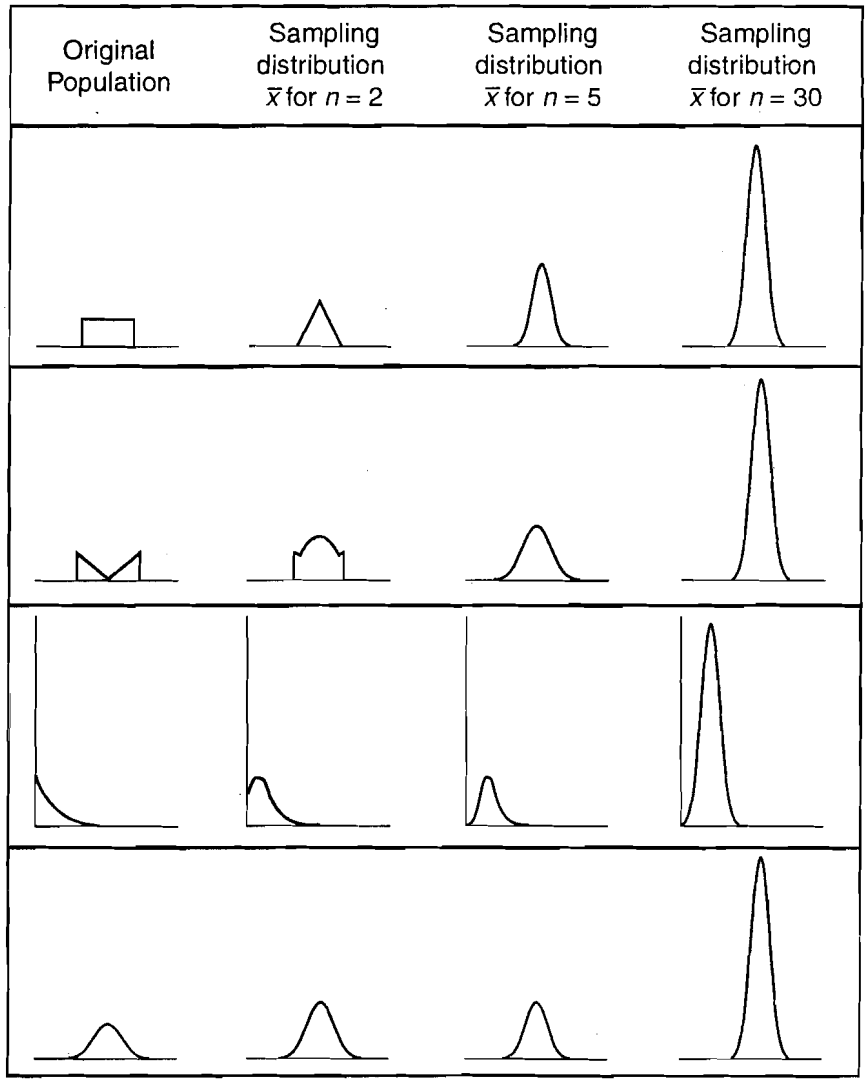


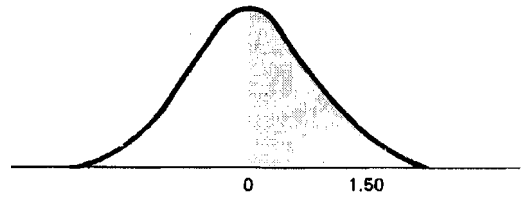
Fig. 1 Sampling distribution of $\bar{x}$ for different populations and different sample sizes.

APPENDIX-2

TABLE 1

AREAS UNDER THE STANDARD NORMAL CURVE

This table shows the area between zero (the mean of a standard normal variable) and z. For example, if $z = 1.50$, this is the shaded area shown below which equals .4332.

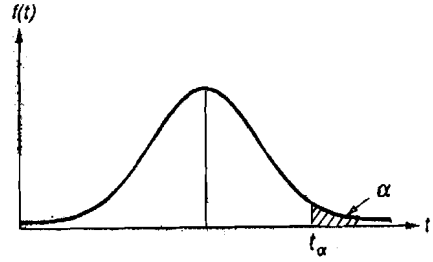


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: This table is adapted from National Bureau of Standards, *Tables of Normal Probability Functions*, Applied Mathematics Series 23, U.S. Department of Commerce, 1953.

TABLE-2

t-distribution



ν	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1 328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

TABLE-3
CHI-SQUARED DISTRIBUTION

ν	$\alpha = .995$	$\alpha = .990$	$\alpha = .975$	$\alpha = .950$	$\alpha = .900$	$\alpha = .750$	$\alpha = .500$	$\alpha = .250$	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .010$	$\alpha = .005$
1	0.0000393	0.000157	0.000982	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196

TABLE-4
F-distribution

$F_{0.05}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.50	19.00	19.20	19.20	19.30	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.50	19.50	19.50	19.50	19.50	19.50
3	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.38	2.30	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	3.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.93
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

TABLE-5
F-distribution

$F_{0.01}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.20	99.20	99.30	99.30	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.50	99.50	99.50	99.50	99.50	99.50
3	34.10	30.80	29.50	28.70	28.20	27.90	27.70	27.50	27.30	27.20	27.10	26.90	26.70	26.60	26.50	26.40	26.30	26.20	26.10
4	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.40	14.20	14.00	13.90	13.80	13.70	13.70	13.60	13.50
5	16.30	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.70	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.20	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.30	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.83
9	10.60	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.96	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00