UNIT 14 SAMPLING DISTRIBUTIONS

Objectives

When you have successfully completed this unit, you should be able to:

- understand the meaning of sampling distribution of a sample statistic
- obtain the sampling distribution of the mean
- get an understanding of the sampling distribution of variance
- construct the sampling distribution of the proportion
- know the Central Limit Theorem and appreciate why it is used so extensively in practice
- develop confidence intervals for the population mean and the population proportion
- determine the sample size required while estimating the population mean or the population proportion.

Structure

14.1 INTRODUCTION

Having discussed the various methods available for picking up a sample from a population we would naturally be interested in drawing inferences about the population based on our observations made on the sample members. This could mean estimating the value of a population parameter, testing a statistical hypothesis about the population, comparing two or more populations, performing correlation and regression analysis on more than one variable measured on the sample members, and many other inferences. We shall discuss some of these problems in this and the subsequent units.

What is a Sampling Distribution?

Suppose we are interested in drawing some inference regarding the weight of containers produced by an automatic filling machine. Our population, therefore, consists of all the filled-containers produced in the past as well as those which are going to be produced in the future by the automatic filling machine. We pick up a sample of size n and take measurements regarding the characteristic we are interested in viz. the weight of the filled container on each of our sample members. We thus end up with n sample values x_i, x_2, ...x_n. As described in the previous unit, any quantity which can be determined as a function of the sample values x_i, x_2, ... , x_n is called a sample statistic.

Referring to our earlier discussion on the concept of a random variable, it is not difficult to see that any sample statistic is a random variable and, therefore, has a probability distribution or a probability density function. It is also known as the sampling distribution of the statistic. In practice, we refer to the sampling distributions of only the commonly used sampling statistics like the sample mean, sample variance, sample proportion, sample median etc., which have a role in making inferences about the population.

Why Study Sampling Distributions?

Sample statistics form the basis of all inferences drawn about populations. If we know the probability distribution of the sample statistic, then we can calculate the probability that the sample statistic assumes a particular value (if it is a discrete random variable) or has a value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences. We will demonstrate this by an example.

Suppose we know that 45% of the population of all users of talcum powder prefer our brand to the next competing brand. A "new improved" version of our brand has been developed and given to a random sample of 100 talcum powder users for use. If 60 of these prefer our "new improved" version to the next competing brand, what should we conclude? For an answer, we would like to know the probability that the sample proportion in a sample of size 100 is as large as 60% or higher when the true population proportion is only 45%, i.e. assuming that the new version is no better than the old. If this probability is quite large, say 0.5, we might conclude that the high sample proportion viz. 60% is perhaps because of sampling errors.and the new version is not really superior to the old. On the other hand, if this probability works out to a very small figure, say 0.001, then rather than concluding that we have observed a rare event we might conclude that the true population proportion is higher than 45%, i.e. the new version is actually superior to the old one as perceived by members of the population. To calculate this probability, we need to know the probability distribution of sample proportion or the sampling distribution of the proportion.

14.2 SAMPLING DISTRIBUTION OF THE MEAN

We shall first discuss the sampling distribution of the mean. We start by discussing the concept of the sample mean and then study its expected value and variance in the general case. We shall end this section by describing the sampling distribution of the mean in the special case when the population distribution is normal.

The Sample Mean

Suppose we have a simple random sample of size n picked up from a population. We take measurements on each sample member in the characteristic of our interest and denote the observation as x_1, x_2, . . . , x_n respectively. The sample mean for this sample, represented by x, is defined as

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

If we pick up another sample of size n from the same population, we might end tip a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance causes. The distribution of the sample mean is also referred to as the sampling distribution of the mean.

However, to observe the distribution of x empirically, we have to take many samples of size n and determine the value of x for each sample. Then, looking at the various observed values of z, it might be possible to get an idea of the nature of the distribution.

Sampling from Infinite Populations

We shall study the distribution of z in two cases-one when the population is finite and we are sampling without replacement; and the other when the population is infinitely large or when the sampling is done with replacement. We start with the latter.

We assume we have a population which is infinitely large and having a population mean of μ and a population variance of u^2. This implies that if x is a random variable denoting the measurement of the characteristic that we are interested in, on one element of the population picked up randomly, then

the expected value of x, E(x) = μ

and the variance of x, Var (x) = $_{6}2$

The sample mean, x, can be looked at as the sum of n random variables, viz x_1, x_2, ... , x_n, each being divided by (1/n). Here x_1, is a random variable representing the first observed value in the sample, x_2 is a random variable representing the second observed value and so on. Now, when the population is infinitely large, whatever be the value of x_l, the distribution of x_2 is not affected by it. This is true of any other pair of random variables as well.. In other words x_1, x_2, ... , x_n are independent random variables and all are picked up from the same population.

$$\therefore \quad E(x_1) = \mu \text{ and } Var\ (x_1) = \sigma^2$$
$$E(x_2) = \mu \text{ and } Var\ (x_2) = \sigma^2 \text{ and so on}$$

Finally,

$$E\ (\bar{x}) = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$
$$= \frac{1}{n}E\ (x_1) + \frac{1}{n}E\ (x_2) + \ldots + \frac{1}{n}E\ (x_n)$$
$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu$$
$$= \mu .$$

$$\text{and } Var\ (\bar{x}) = Var\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right]$$
$$= Var\left(\frac{x_1}{n}\right) + Var\left(\frac{x_2}{n}\right) + \ldots + Var\left(\frac{x_n}{n}\right)$$
$$= \frac{1}{n^2}Var\ (x_1) + \frac{1}{n^2}Var\ (x_2) + \ldots + \frac{1}{n^2}Var\ (x_n)$$
$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \ldots + \frac{1}{n^2}\sigma^2$$
$$= \frac{\sigma^2}{n}$$

We have arrived at two very important results for the case when the population is infinitely large, which we shall be using very often. The first says that the expected value of the sample mean is the same as the population mean while the second says that the variance of the sample mean is the variance of the population divided by the sample size.

If we take a large number of samples of size n, then the average value of the sample means tends to be close to the true population mean. On the other hand, if the sample size is increased then the variance of gets reduced and by selecting an appropriately large value of n, the variance of x can be made as small as desired.

Thee standard deviation of x is also called the standard error of the mean. Very often we estimate the population mean by the sample mean. The standard error of the mean indicates the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we are reasonably confident that whatever sample mean value we have observed cannot be very far away from the true value.

The standard error of the mean is represented by σ_x .

Sampling With Replacement

The above results have been obtained under the assumption that the random variables x_i, x_2, ... , x„ are independent. This assumption is valid when the population is infinitely large. It is also valid when the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. Hence, if the sampling is done with replacement, we would again have

$$E(\bar{x}) = \mu$$

$$\text{and } Var(\bar{x}) = \frac{\sigma^2}{n}$$

$$\text{i.e. } \sigma\bar{x} = \frac{\sigma}{\sqrt{n}}$$

Sampling Without Replacement from Finite Populations

When a sample is picked up without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on. As the n random variables representing the n sample members do not remain independent, the expression for the variance of x changes. We are only mentioning the results without deriving these.

$$E(\bar{x}) = \mu$$

$$\text{and } Var(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$\text{i.e. } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

By comparing these expressions with the ones derived above we find that the standard error of is the same but further multiplied by a factor

$\sqrt{(N\text{-}n)/(N\text{-}1)}$. This factor is, therefore, known as the finite population multiplier.

In practice, almost all the samples used picked up without replacement. Also, most populations are finite although they may be very large and so the standard error of the mean should theoretically be found by using the expression given above. However, if the population size (N) is large and consequently the sampling ratio (n/N) small, then the finite population multiplier is close to 1 and is not used, thus treating large finite populations as if they were infinitely large. For example, if N = 100,000 and n =100, the finite population multiplier

$$\sqrt{\frac{N-n}{N-1}} = \frac{100{,}000-100}{100{,}000-1}$$

$$= \frac{99{,}900}{99{,}999}$$

$$= .9995$$

Which is very close to 1 and the standard error of the mean would, for all practical purposes, be the same whether the population is treated as finite or infinite. As a rule of that, the finite population multiplier may not be used if the sampling ratio (n/N) is smaller than 0.05.

Sampling from Normal Populations

We have seen earlier that the normal distribution occurs very frequently among many natural phenomena. For example, heights or weights of individuals, the weights of filled-cans from an automatic machine, the hardness obtained by heat treatment, etc. are distributed normally.

We also know that the sum of two independent random variables will follow a normal distribution if each of the two random variables belongs to a normal population. The sample mean, as we have seen earlier is the sum of n random variables x_1, x_2,….. x_n each divided by n. Now, if each of these random variables is from the same normal population, it is not difficult to see that x would also be distributed normally.

Let $x \,\square\, N(\mu,\sigma^2)$ symbolically represent the fact that the random variable x is distributed normally with meanμ and variance σ^2. What we have said in the earlier paragraphs, amounts to the following:

If $x \,\square\, N(\mu,\sigma^2)$

then it follows that $x \,\square\, N(\mu,\frac{\sigma^2}{n})$

The normal distribution is a continuous distribution and so the population cannot be small and finite if it is distributed normally; that is why we have not used the finite population multiplier in the above expression. We shall now show by an example, how to make use of the above result.

Suppose the diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm?

Let x be a random variable representing the diameter of one component picked up at random.

We know that $x \,\square\, N(10, .01)$

Therefore, it follows that $x \,\square\, N(10, \frac{.01}{5})$

i.e. x will be distributed normally with a mean of 10 and a variance which is only 1/5 of the variance of the population, since the sample size is 5.

$$\Pr\ \{9.95 \leq \bar{x} \leq 10.05\} = 2 \times \Pr\ \{10 \leq \bar{x} \leq 10.05\}$$

$$= 2 \times \Pr\ \left\{\frac{10-\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq \frac{10.05-\mu}{\sigma/\sqrt{n}}\right\}$$

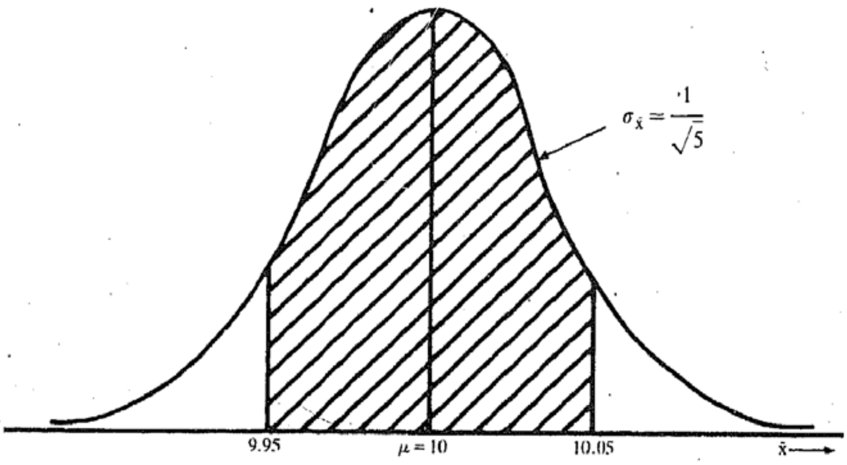
$$= 2 \times \Pr\ \left\{0 \leq z \leq \frac{10.05-10}{.1/\sqrt{5}}\right\}$$

$$= 2 \times \Pr\ \{0 \leq z \leq 1.12\}$$

$$= 2 \times 0.3686$$

$$= 0.7372$$

Figure I: Distribution of $\bar{x}$. The Shaded area represents the probability that the random variable $\bar{x}$ between 9.95 and 10.05.



We first make use of the symmetry of the normal distribution and then calculate the z value by subtracting the mean and then dividing it by the standard deviation of the random variable distributed normally, viz k. The probability of interest is also shown as the shaded area in Figure I above.

14.3 CENTRAL LIMIT THEOREM

In this section we shall discuss one of the most important results of applied statistics which is also known by the name of the central limit theorem.

If x_1, x_2, ... , x_n are n random variables which are independent and having the same distribution with mean p. and standard deviation σ, then if $n \to \infty$, the limiting distribution of the standardised mean

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ s the standard normal distribution.

In practice, if the sample size is sufficiently large, we need not know the population distribution because the central limit theorem assures us that the distribution of x can be approximated by a normal distribution. A sample size larger than 30 is generally considered to be large enough for this purposes.

Many practical samples are of size higher than 30. In all these cases, we know that the sampling distribution of the mean can be approximated by a normal distribution with an expected value equal to the population mean and a variance which is equal to the population variance divided by the sample size n.

We need to use the central limit theorem when the population distribution is either unknown or known to be non-normal. If the population distribution is known to be normal, then will also be distributed normally, as we have seen in section 14.2 above irrespective of the sample size.

Activity A

A sample of size 25 is picked up at random from a population which is normally distributed with a mean of 100 and a variance of 36. Calculate.

a) $P_r\{\bar{x} \leqslant 99\}$

b) $P_r\{98 \leqslant \bar{x} \leqslant 100\}$

Activity B

If in (i) above, the sample is increased to 36, recalculate the following

a) $P_r\{\bar{x} \leq 99\}$

b) $P_r\{98 \leq \bar{x} \leq 100\}$

Activity C

Refer to Table 2 in the previous unit where we have a population of size 5.

A,B,C,D and E are five members of a family with the following weights of each family member:

$$x_a = 70 \text{ kg}$$
$$x_b = 80 \text{ kg}$$
$$x_c = 50 \text{ kg}$$
$$x_d = 30 \text{ kg}$$
$$x_e = 10 \text{ kg}$$

Using the ten samples listed in Table 2, find the probability distribution of the sample mean and verify that

$$E(\bar{x}) = \mu = \frac{70 + 80 + 50 + 30 + 10}{5} = 48 \text{ kg.}$$

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

14.4 SAMPLING DISTRIBUTION OF THE VARIANCE

We shall now discuss the sampling distribution of the variance. We shall first introduce the concept of sample variance and then present the chi-square distribution which helps us in working out probabilities for the sample variance, when the population is distributed normally.

The Sample Variance

By now it is implicitly clear that we use the sample mean to estimate the population mean, when that parameter is unknown. Similarly; we use a sample statistic called the sample variance to estimate the population variance. The sample variance is usually denoted by s^2 and it again captures sc me kind of an average of the square of deviations of the sample values from the sample mean. Let us put it in an equation form

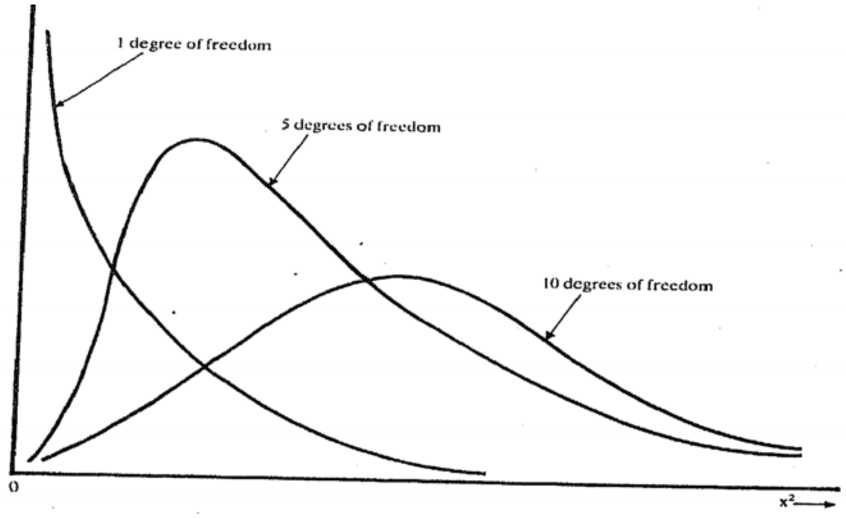
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

By comparing this expression with the corresponding expression for the population variance, we notice two differences. The deviations are measured from the sample mean and not from the population mean and secondly, the sum of squared deviations is divided by (n - 1) and not by n. Consequently, we can calculate the sample variance based only on the sample values without knowing the value of any population parameter. The division by (n - 1) is due to a technical reason to make the expected value of s^2 equal Q^2, which it is supposed to estimate.

The Chi-square Distribution

If the random variable x has the standard normal distribution, what would be the distribution of x^2? Intuitively speaking, it would be quite different from a normal distribution because now x^2, being a squared term, can assume only non-negative values. The probability density of x^2 will be the highest near 0, because most of the x value are close to 0 in a standard normal distribution. This distribution is called the chi-square distribution with 1 degree of freedom and is shown in Figure II below.

Figure II: Chi-square (x^2) distribution with different degrees of freedom



The chi-square distribution has only one parameter viz. the degrees of freedom and so there are many chi-square distributions each with its own degrees of freedom. In statistical tables, chi-square values for different areas under the right tail and the left tail of various chi-square distributions are tabulated.

If x_i, x_2, ... , x_n are independent random variables, each having a standard normal distribution, then (x_i + x_2 + ... + x_n) will have a chi-square distribution with n degrees of freedom.

If y_l and y_2 are independent random variables having chi-square distributions with y_l and y_2 degrees of freedom, then (y_1 + y_2) will have a chi-square distribution with y_1 +y_2 degrees of freedom.

We have stated some results above, without deriving them, to help us grasp the chi-square distribution intuitively. We shall state two more results in the same spirit.

If y_l and y_2 are independent random variables such that y_l has a chi-square distribution with y, degrees of freedom and (y_1 +y_2) has a chi-square distribution with y > y, degrees of freedom, then y_2 will have a chi-square distribution with (y - y_l) degrees of freedom.

Now, if x_1, x_2, ... , x_n, are n random va^{r}iables from a normal population with mean μ . and variance ..2,

i.e. $x_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$

it implies that $\frac{x_i - \mu}{\sigma} \sim N(0, 1)$

and so $\left(\frac{x_i - \mu}{\sigma}\right)^2$ will have a chi-square distribution with 1 degree of freedom. o-

Hence, $\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$ will have a chi-square distribution with n degrees of freedom.

We can break up this expression by measuring the deviation from x in place of μ.

We will then have

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu)]^2$$

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

$$=\frac{(n-1)s^2}{\sigma^2}+\left(\frac{\bar{x}-u}{\sigma/\sqrt{n}}\right)^2 \text{ since } \sum_{i=1}^{n}(x_i-\bar{x})=0.$$

Now, we know that the left hand side of the above equation is a random variable which has a chi-square distribution with n degrees of freedom. We also know that

$$\bar{x}\sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$$

$\therefore \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$ will have a chi-square distribution with 1 degree of freedom. Hence, if the two terms on the right hand side of the above equation are independent (which will be assumed as true here and you will have to refer to advanced texts on statistics for the proof of the same), then it follows that $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with (n -1) degrees of freedom. One degree of freedom is lost because the deviations are measured from z and not from μ.

Expected Value and Variance of s^2

In practice, therefore, we work with the distribution of $\frac{(n-1)s^2}{\sigma^2}$ and not with the distribution-of s^2 directly. The mean of a chi-square distribution is equal to its degrees of freedom and the variance is equal to twice the degrees of freedom. This can be used to find the expected value and the variance of s^2.

Since $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with (n-1) degrees of freedom,

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

$$\text{or } \frac{(n-1)}{\sigma^2}\cdot E(s^2)=n-1$$

$$\therefore\ E\ (s^2)=\sigma^2$$

$$\text{Also Var}\left[\frac{(n-1)s^2}{\sigma^2}\right]=2(n-1)$$

Using the definition of Variance, we get

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

$$\text{or, } E\left[\frac{(n-1)s^2}{\sigma^2}-(n-1)\right]^2=2(n-1)$$

$$\text{or, } \frac{(n-1)^2}{\sigma^4}\cdot E\ (s^2-\sigma^2)^2=2(n-1)$$

$$\therefore\ E(s^2-\sigma^2)^2=\frac{2\sigma^4}{n-1}$$

$$\text{i.e. } Var\left(s^2\right) = \frac{2\sigma^4}{n-1}$$

since the expected value of s^2 is equal to σ^2.

We therefore conclude that if we take a large number of samples, each with a sample size on n, from a normal population with mean and variance a^2, each sample will perhaps have a different value for its sample variance s^2. But the average of a large number of values of s^2 will be close to σ^2. Also, the variance of s^2 falls as the sample size increases.

Let us recall that in all our discussion about the sampling distribution of the variance, we have been assuming that the population is distributed normally. If the population does not have a normal distribution, then nothing can be said about the distribution of s^2.

14.5 THE STUDENT'S DISTRIBUTION

We studied the sampling distribution of the mean in section 14.2 above where we showed that if the population distribution is normal then the distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is the standard normal distribution. In actual practice, the value of the population standard deviation σ is often unknown which makes it necessary to replace this with an estimate, usually by s-the sample standard deviation. In such cases, we would like to know the exact sampling distribution of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ for random samples from normal populations and this is provided by the t distribution which is also known as the student's t distribution after the pen name adopted by its author.

The Concept of the t Statistic

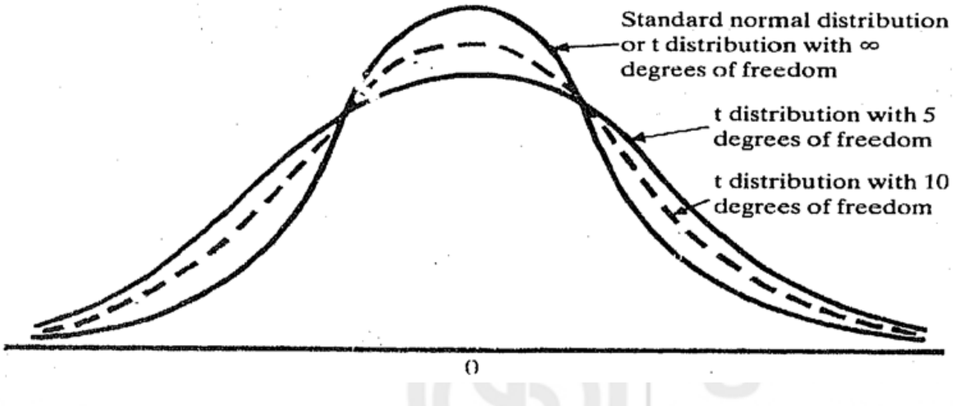
If x is a random variable having the standard normal distribution and y is a random variable having a chi-square distribution with v degrees of freedom and if x and y are independent, then the random variable

$$t = \frac{x}{\sqrt{y/\nu}}$$

has a distribution called the t distribution (or the Student's t distribution) with v degrees of freedom.

There are many t distributions, each with its degrees of freedom, which is the only parameter of this distribution. A t distribution is similar to the standard normal distribution as shown in Figure III below-only it is flatter and wider, thus having longer tails.

Figure III: The t distribution with different degrees of freedom



As the degrees of freedom increase, the t distribution comes closer to the standard normal distribution and when the degrees of freedom become infinitely large, the t distribution and the z distribution become indistinguishable.

The t Distribution in Practice

If we have a random sample of size n from a normal population with mean μ and variance σ^2, then we know that the sample mean will be distributed normally with mean μ and variance σ^2 / n. And so $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ will have a standard normal distribution.

We also know that in such a situation $\frac{(n-1)s^2}{\sigma^2}$ will have a chi-square distribution with (n -1) degrees of freedom. It has been shown in advanced texts that these two random variables are also independent and so

$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)s^2}{\sigma^2(n-1)}}$ will have a t distribution with (n - 1) degrees of freedom.

After simplification, we conclude that $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ would have a t distribution with (n - 1) degrees of freedom.

It is therefore, possible to know the sampling distribution of x even when σ is not known.

This result is really useful when the sample size is not very large. As we have seen earlier, if the sample size n is large, the t distribution with large degrees of freedom can be approximated by the z distribution. The t distribution is used when the degrees of freedom are not larger than 30; if the degrees of freedom are larger than 30, the t distribution is approximated by the standard normal or the z distribution.

The t distribution is again extensively tabulated because it is used quite frequently. As it is a symmetrical distribution, only one tail is generally tabulated and the other tail values can be worked out by using this property of symmetry.

14.6 SAMPLING DISTRIBUTION OF THE PROPORTION

Suppose we know that a proportion p of the population possesses a particular attribute that is of interest to us-e.g. a proportion p of the population prefer our product to the next competing brand. This also implies that a proportion (1 - p) of the population do not prefer our product as compared to the next competing brand. If we pick up one member of the population at random, the probability of success i.e. the probability that this person will prefer our product to the next competing brand is p.

If the population is large enough, then even if we make repeated trials, which are considered to be independent, each with a probability of success equal to p. In such a case, if we make n repeated trials to pick up a sample of size n, the probability of x success in the sample is given by a binomial probability distribution, viz.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

If there are x successes in the sample, the sample proportion of success p is given by

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

The expected value and the variance of x, i.e. the number of successes in a sample of size n is known to be:

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

We can, therefore, find the expected value and the variance of the sample proportion p, as below:

$$E(\bar{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}\cdot np$$

$$= p$$

$$\text{and } Var\ (\bar{p}) = Var\ \left(\frac{x}{n}\right) = \frac{1}{n^2}Var\ (x) = \frac{1}{n^2}.np\ (1-p)$$

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

Finally, if the sample size n is large enough, we can approximate the binomial probability distribution by a normal distribution with the same mean and variance. Thus, if n is sufficiently large,

$$\bar{p} \sim N\left(p,\ \frac{p\ (1-p)}{n}\right)$$

This approximation works quite well if n is sufficiently large so that both np and n(1-p) are at least as large as 5.

Activity D

A population is normally distributed with a mean of 100. A sample of size 15 is picked up at random from the population. If we know from t tables, that

$$P_r\ (t_{14} \geqslant 1.761) = 0.05$$

where t_{14} represents a t variable with 14 degrees of freedom, calculate

$$P_r\ (\bar{x} \geqslant 115)$$

If we know that the sample standard deviation is 33.

Activity E

In a Board examination this year, 85% of the students who appeared for the examination passed. 100 students appeared in the same examination from School Q. What is the probability that 90 or more of these students passed?

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………

14.7 INTERVAL ESTIMATION

Suppose we want to estimate the mean income of a population of households residing in a part of a city. We might proceed by picking up a random sample of 100 households from the population and calculate the sample mean i.e. the mean income of the 100 sample households. In the absence of any other information, the sample mean can be .used as a point estimate of the population mean.

However, if we also want to convey the precision involved in this estimation, we need Distributions to give the standard error of the mean. As we have seen in section 14.2 above, the standard error of the mean depends on the population variance and the sample size.

The lower the standard error of the mean, the greater is the confidence on the correctness of our estimation. This process is further refined in interval estimation, wherein we present our estimate as an interval and quantify our confidence so that the true population parameter is contained by the estimated interval.

The Confidence Level

As mentioned earlier, the sample mean is our estimate of the population mean. If we are asked to give an interval as our estimate, then we would add a range on the upper and the lower side of the sample mean and give that interval as our estimate. The larger the interval, the greater is our confidence that the interval does contain the true population mean. It is to be noted that the true population mean is a constant and is not a variable. On the other hand, the interval that we specify is a random interval whose position depends on the sample mean. For example if the sample mean is 50 and the standard error of the mean is 5, we may specify our interval estimate as (45,55) i.e. from 45 to 55 which spans one standard error of the mean on either side of the sample mean. On the other hand, if the interval estimate is specified as (40, 60) i.e. spanning two standard errors of the mean on either side of the sample mean, we are more confident that the latter interval contains the true population mean as compared to the former. However, if the confidence level is raised too high, the corresponding interval may become too wide to be of any practical use.

The confidence level, therefore, may be defined as the probability that the interval estimate will contain the true value of the population parameter that is being estimated. If we say that a 95% confidence interval for the population mean is obtained by spanning 1.96 times the standard error of the mean on either side of the sample mean, we mean that we take a large number of samples of size n, say 1000, and obtain the interval estimates from each of these 1000 samples and then 95% of these interval estimates would contain the true population mean.

Confidence Interval for the Population Mean

We shall now discuss how to obtain a confidence interval for the population mean. We shill assume that the population distribution is normal and that the population aflame is known. Later, we shall relax the second condition.

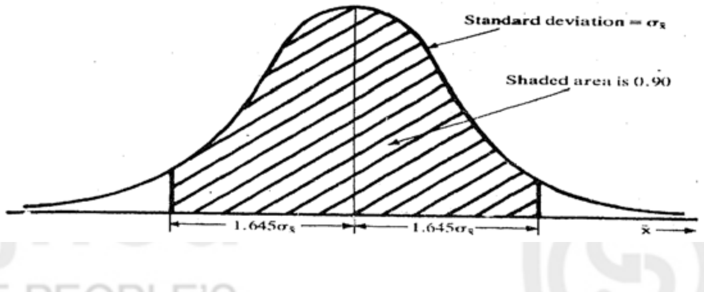
Suppose it is known that the weight of cement in packed bags is distributed normally with a standard deviation of 0.2 Kg. A sample of 25 bags is picked up at random and the mean weight of cement in these 25 bags is only 49.7 Kg. We want to find a 90% confidence interval for the mean weight of cement in filled bags.

Let x be a random variable representing the weight of cement in a bag picked up at random. We know that x is distributed normally with a standard deviation of 0.2 Kg.

The standard error of the mean can be easily calculated as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{25} = .04 \text{ Kg}$$

Figure IV: Distribution of $\bar{x}$



As shown in Figure IV above, we know that the sample mean is distributed normally with mean and standard deviation equal to 0.04 Kg. By referring to the normal table we can easily find that the probability that is between p. and (μ+ 1.645rr) is 0.45 and so the probability that z is between (p.- 1.645 (T) and (μ+ 1.645 *cr)* is 0.90. In other words, if we use an interval spanning from (X- 1.645 us) to (X+ 1.645az) then 9O% of the time this interval will contain p,

Hence, for a 90% confidence interval,

the lower limit $= \bar{x} - 1.645\sigma_{\bar{x}} = 49.7 - 1.645 \times 0.04$

$= 49.6342$

and the upper limit $= \bar{x} + 1.645\frac{\sigma}{x} = 49.7 + 1.645 \times 0.04$

$= 49.7658$

Therefore, we can state with 90% confidence level that the mean weight of cement in a filled hag lies between 49.6342 Kg and 49.7658 Kg.

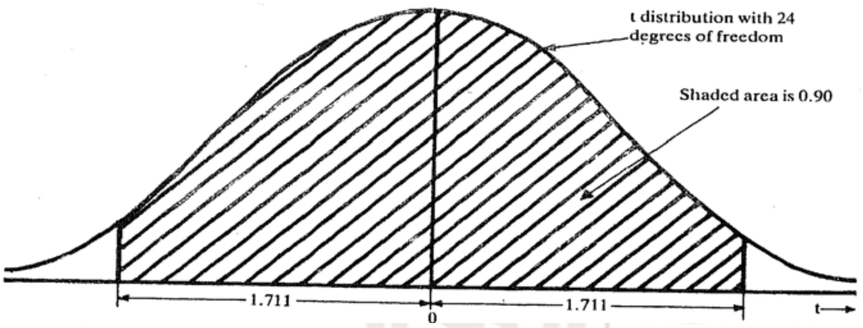
We can use the above approach when the population standard deviation is known or when the sample size is large n > 30 , in which case the sample standard deviation can he used as an estimate of the population standard deviation. However, if the sample size is not large, as in the example above, then one has to use the t distribution in place of the standard normal distribution to calculate the probabilities.

Let us assume that we are interested in developing a 90% confidence interval in the same situation as described earlier with the difference that the population standard deviation is now not known. However, the sample standard deviation has been calculated and is known to be O.2 Kg.

Since the sample size n = 25, we know that $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ follows a t distribution with 24

degrees of freedom. From t tables, we can see that the probability that a t statistic with 24 degrees of freedom lying between - 1.711 and 1.711 is 0.90-i.e. the probability that X lies between - 1.711 s/ Un and + 1.711 s/\ is 0.90. This is shown in Figure 5 below.

Figure V: Area under a t distribution



In other words, if we use an interval spanning from (X - 1.711s/V to (z + 1.711s/\In) then 90% of the time, this interval will contain μ. Hence, for a 90% confidence interval,

the lower limit $= \bar{x} - 1.711\frac{s}{\sqrt{n}} = 49.7 - 1.711\frac{0.2}{\sqrt{25}}$

$= 49.6316$

and the upper limit $= \bar{x} + 1.711\ \frac{s}{\sqrt{n}} = 49.7 + 1.711\ \frac{0.2}{\sqrt{25}}$

$= 49.7684$

In this case, we can state with 90% confidence level that the mean weight of cement in a filled hag lies between 49.6316 Kg and 49.7684 Kg.

14.8 THE SAMPLE SIZE

In section 14.7 above we have seen how the sampling distribution of a statistic helps us in developing a confidence interval for the corresponding population parameter. In this section we shall present another application of the sampling distributions. We have earlier referred to the fact that in some situations the sample size required can he determined on the basis of the precision of the estimates. We shall now demonstrate this process.

Sample Size for Estimating Population Mean

We assume that the population distribution is normal and the population standard deviation is known. In such a case the sample size required for a given confidence level and a required accuracy can he easily determined. We again take the help of an example.

Suppose we know that the weight of cement in filled bags is distributed normally with a standard deviation o of 0.2 Kg. We want to know how large a sample should he taken so that the mean weight of cement in a filled hag can be estimated within plus or minus 0.05 Kg of the true value with a confidence level of 90%.

We have seen in section 14.7 above that the interval $\left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}\right)$ to $\left(\bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right)$ contains the true value of the population mean 90% of the time. We also want that the interval (X-0.05) to (X+0.05) should give us a 90% confidence level.

Therefore, $1.645 \frac{\sigma}{\sqrt{n}} = 0.05$

and so n $= \left(\frac{1.645 \times 0.2}{0.05}\right)^2$

$= 43.3$

We must have a sample size of at least 44 so that the mean weight of cement in a filled bag can be estimated within plus or minus 0.05 Kg of the true value with a 90% confidence level.

It is to be noted that this approach does not work if the population standard deviation is not known because the sample standard deviation is known only after the sample has been analysed whereas the sample size decision is required before the sample is picked up.

Sample Size for Estimating Population Proportion

Suppose we want to estimate the proportion of consumers in the population who prefer our product to the next competing brand. How large a sample should be taken so that the population proportion can be estimated within plus or minus 0.05 with a 90% confidence level?

We shall use the sample proportion p to estimate the population proportion p. As mentioned in section 14.6 above, if n is sufficiently large, the distribution of p can be approximated by a normal distribution with mean p and variance p (1 - p)/n.

From normal tables, we can now say that the probability that p will lie between (p-1.645Vp(1-p)/n) and (p + 1.645Vp(l-p)/n) is 0.90. In other words, the

interval (p- *1.645Vp* (1-p)/n) to (p + 1.645Vp (1-p)/n) will contain p, 90% of the time.

We also want that the interval (p - 0.05) to (p + 0.05) should contain p, 90% of the time.

$$\text{Therefore, } 1.645\sqrt{\frac{p\,(1-p)}{n}} = 0.05$$

$$\text{or} \qquad \sqrt{\frac{p\,(1-p)}{n}} = \frac{0.05}{1.645} = 0.0304$$

$$\text{or} \qquad \frac{p\,(1-p)}{n} = 0.0009239$$

$$\therefore \qquad n = \frac{p(1-p)}{0.0009239}$$

But we do not know the value of p, so n cannot be calculated directly. However, whatever be the value of p, the highest value for the expression p (1 - p) is 0.25, which is the case when p = 0.5. Hence, in the worst case the highest possible value for p(1 -p) is 0.25. In that case 0.25

$$n = \frac{0.25}{0.0009239} = 270.6$$

Therefore, if we take a sample of size 271, then we are sure that our estimate of the population proportion would be within plus and minus 0.05 of the true value with a confidence level of 90% whatever he the value of p.

Activity F

100 Sodium Vapour Lamps were tested to estimate the life of such a lamp. The life of these 100 lamps exhibited a mean of 10,000 hours with a standard deviation of 500 hours. Construct a 90% confidence interval for the true mean life of a Sodium Vapour Lamp.

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………

Activity G

If the sample size in the previous situation had been 15 in place 100, what would be the confidence interval.

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………

Activity H

We want to estimate the proportion of employees who prefer the codification of rules and regulations. What should be the sample size if we want our estimate to he within plus or minus 0.05 with a 95% confidence level.

…………………………………………………………………………………………

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………

14.9 SUMMARY

We have introduced the concept of sampling distributions in this unit. We have discussed the sampling distributions of some commonly used statistics and also shown some applications of the same.

A sampling distribution of a sample statistic has been introduced as the probability distribution or the probability density function of the sample statistic. In the sampling distribution of the mean, we find that if the population distribution is normal, the sample mean is also distributed normally with the same mean but with a smaller standard deviation. In fact, the standard deviation of the sample mean, also known as the standard error of the mean, is found to be equal to the population standard deviation divided by the sample size.

We have also presented a very important result called the central limit theorem which assures us that if the sample size is large enough (greater than 30), the sampling distribution of the mean could be approximated by a corresponding normal distribution with the mean and standard deviation as given in the preceding paragraph.

We have then explored the sampling distribution of the variance and found that a related quantity viz. $\frac{(n-1)s^2}{\sigma^2}$ would have a chi-square distribution with (n -1) degrees of freedom. We have learnt that the chi-square distribution is tabulated extensively and so any probability calculations regarding s^2 could be easily made by referring to the tables for the chi-square distribution.

We have introduced one more distribution viz. the t distribution which is found to be applicable when the sampling distribution of the mean is of interest, but the population standard deviation is unknown. It is noticed that if the sample size is large enough (n>30), the t distribution is actually very close to the standard normal distribution.

We have also studied the sampling distribution of the proportion and then looked at two applications of the sampling distributions. One is in developing an interval estimate for a population parameter with a given confidence level, which is conceptualised as the probability that a random interval will contain the true value of the parameter. The second application is to determine the sample size required while estimating the population mean or the population proportion.

14.10 SELF-ASSESSMENT EXERCISES

1 What is the practical utility of the central limit theorem in applied statistics?

2 The daily wages of a random sample of farm labourers are:

14	17	14.5	22	27	16.5	.19.5	21	18	22.5

a) What is the best estimate of the mean daily wages of all farm labourers?

b) What is the standard error of the mean?

c) What is the 95% confidence interval for the population mean? Explain what it indicates and also any assumption you made before you could calculate the confidence interval.

3 An inspector wants to estimate the weight of detergent in packets filled by an automatic filling machine. She wants to be 95% confident that her estimate is not away from the true mean weight of detergent by more than 10 gms. What should the minimum sample size be if it is known that the standard deviation of the weight of detergent filled by that machine is 100 gms?

4 A steamer is certified to carry a load of 20,00() Kg. The weight of one person is distributed normally with a mean of 60 Kg and a standard deviation of 15 Kg.

a) What is the probability of exceeding the certified load if the steamer is carrying 340 persons?

b) What is the maximum number of persons that can travel by the steamer at any time if the probability of exceeding the certified load should not exceed 5%?

Indicate the most appropriate choice for each of the following situations:

5 The finite population multiplier is not used when dealing with large finite population because

a) when the population is large, the standard error of the mean approaches zero.

b) another formula is more appropriate in such cases.

c) the finite population multiplier approaches 1.

d) none of the above.

6 When sampling from a large population, if we want the standard error of the mean to be less than one-half the standard deviation of the population, how large would the sample have to be?

a) 3 b) 5 c) 4 d) none of these

7 A sampling ratio of 0.10 was used in a sample survey when the population size was 50. What should the finite population multiplier be?

a) 0.958

b) 0.10

c) 1.10

d) cannot be calculated from the given data.

8 As the sample size is increased, the standard error of the mean would

a) increase in magnitude

b) decrease in magnitude

c) remain unaltered

d) may either increase or decrease.

9 As the confidence level for a confidence interval increases, the width of the interval

a) increases

b) decreases

c) remains unaltered

d) may either increase or decrease.

14.11 FURTHER READINGS

Emory, L.W., 1976. *Business Research Methods,* Richard D. Irwin, Inc: Homewood.

Ferber, R.(ed.),1974. *Handbook of Marketing Research,* McGraw Hill Book Co.: New York.

Levin, R.I., 1987. *Statistics for Management,* Prentice Hall of India: New Delhi.

Mason, R.D., 1986. *Statistical Techniques in Business and Economics,* Richard D. Irwin, Inc: Homewood.

Mendenhall, W., R.L. Scheaffer and D.D. Wackerly, 1981. *Mathematical Statistics with Applications,* Dunbury Press: Boston.

Plane, D.R. and E.B. Oppermann, 1986. *Business and Economic Statistics,* Business Publications, Inc: Plano.