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,

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

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

$$\begin{aligned}\overline{x} &= \frac{0.53 + 0.46 + 0.50 + 0.49 + 0.52 + 0.53}{6} \\ &= 0.51 \text{seconds}.\end{aligned}$$

Then $\overline{x} = 0.5$ 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

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

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)**

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

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

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

$$
\begin{aligned}
\frac{\sum(x_i - \bar{x})^2}{n-1} &= \frac{1}{n-1}\Big[(00.53 - 0.51)^2 + (0.46 - 0.51)^2 + (0.50 - 0.51)^2 + \\
&\qquad (0.49 - 0.51)^2 + (0.52 - 0.51)^2 + (0.53 - 0.51)^2\Big] \\
&= 0.0006 \\
\therefore S &= \sqrt{0.0006} \text{ seconds.}
\end{aligned}
$$

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).

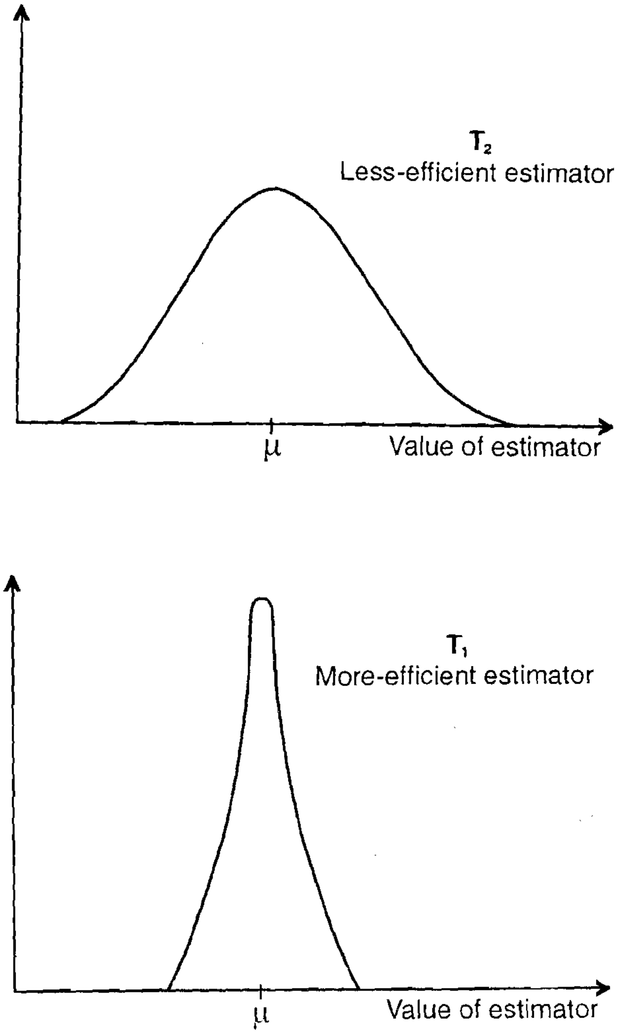


Fig.1:

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

$$P[-1.96 < Z < 1.96] = 0.95 \qquad (3)$$

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

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

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

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

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

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

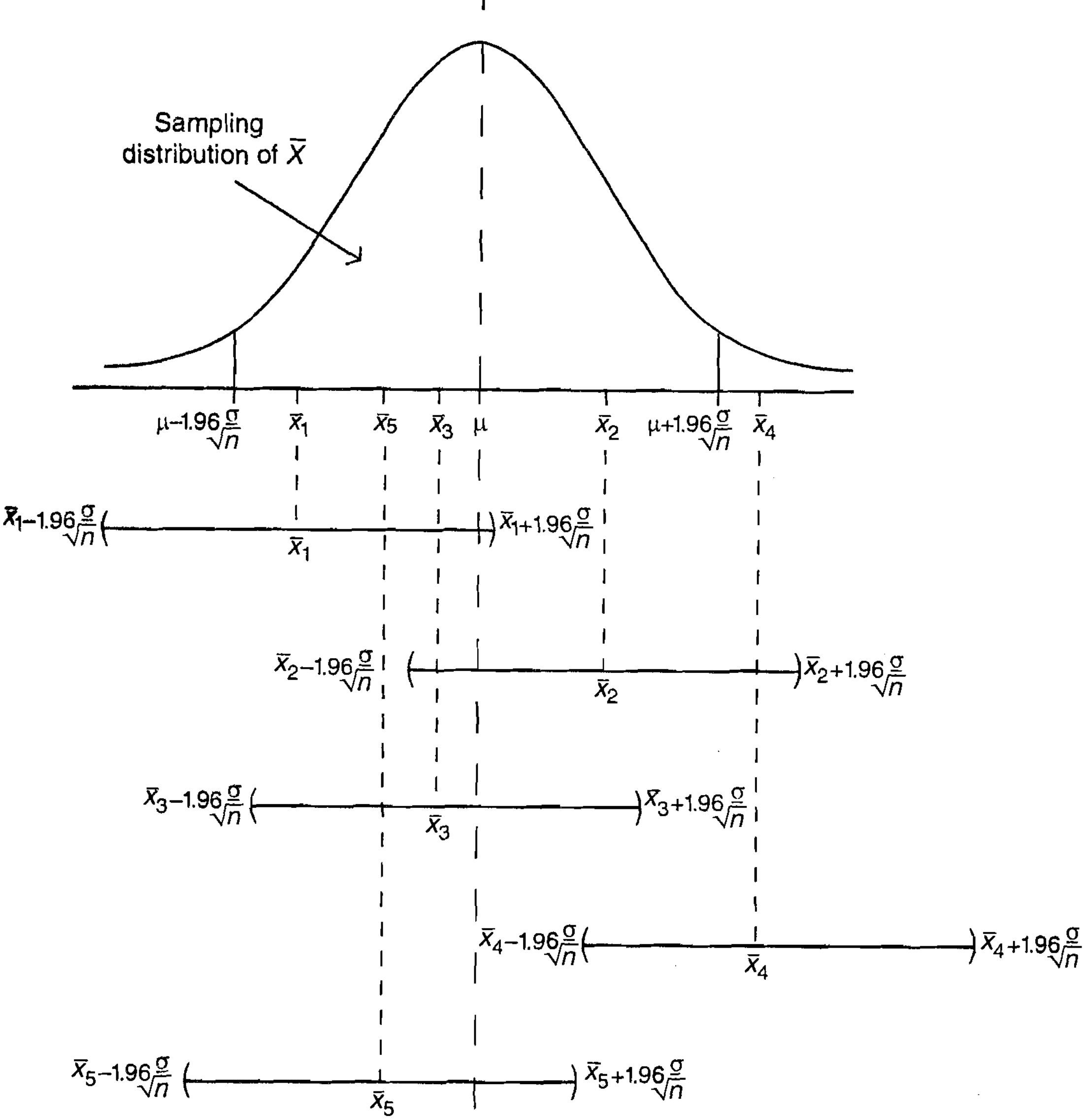
$$P[937.35 < \mu < 962.65] = 0.95$$

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



Sampling distribution of $\bar{X}$
$\mu-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_1$
$\bar{x}_5$
$\bar{x}_3$
μ
$\bar{x}_2$
$\mu+1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_4$
$\bar{x}_1-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_1$
$\bar{x}_1+1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_2-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_2$
$\bar{x}_2+1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_3-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_3$
$\bar{x}_3+1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_4-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_4$
$\bar{x}_4+1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_5-1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}_5$
$\bar{x}_5+1.96\frac{\sigma}{\sqrt{n}}$

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

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

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

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

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).

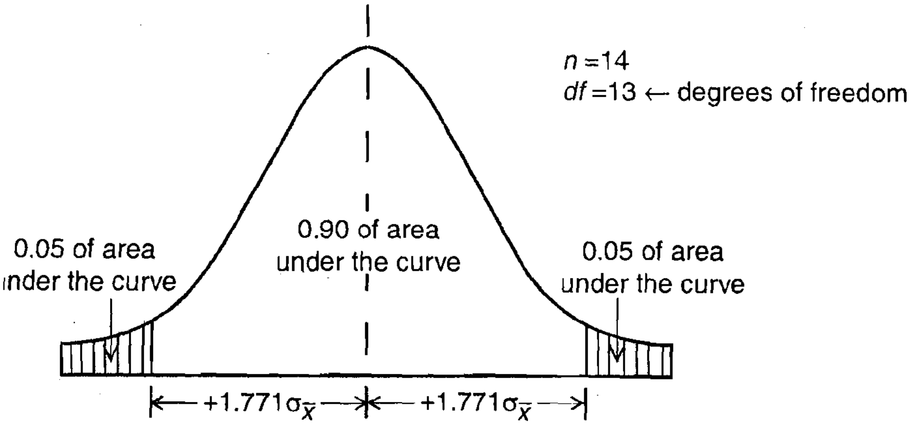


Fig.3: Confidence interval using the t-distribution

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.**

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

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

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

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

$$\left(\bar{x} - t_{.025}(\frac{S}{\sqrt{n}}), \bar{x} + t_{0.025}(\frac{s}{\sqrt{n}}\right) =$$

$$\left(43.5 - 2.26\left(\frac{2}{\sqrt{10}}\right), \ 43.5 + 2.26\left(\frac{2}{\sqrt{10}}\right)\right) = (42.07, \ 44.93).$$

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

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

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

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

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

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

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
		σ known	$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}},\ \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$
μ	$\bar{x} \longrightarrow$	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 $\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$

i) Here $\overline{x} = 0$ 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 $\overline{x} = 550, \sigma = 40$ 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 $n = 140, \overline{x} = 11.4, \sigma = 2.5$
99% C.I. is given by
$\left(\overline{x} - 2.58\frac{\sigma}{\sqrt{n}};\ \overline{x} + 2.58\frac{\sigma}{\sqrt{n}}\right) = \left(11.4 - 2.58\frac{2.5}{\sqrt{140}},\ 11.4 + 2.58\frac{2.5}{\sqrt{140}}\right) =$
$(10.855, 11.945)$

E7) Number of days are:
47,45,44,42,47,46,45,48
$\therefore \overline{x} = 5.5, n = 8, \sigma = 2$
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)
$$\begin{aligned} P(74 \leq \mu \leq 76) &= 2P(75 \leq \mu \leq 76) \\ &= 2P\left(0 \leq Z \leq \frac{76.75}{10/\sqrt{5}}\right) \\ &= 2P(0 \leq Z \leq 0.7071) = 0.5224 \end{aligned}$$

$\therefore$ 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 $(0.0688, 0.3711)$