CHAPTER 5

Estimation

Inferential statistics is the part of statistics that helps us to make decisions about some characteristics of a population based on sample information. *Estimation* is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample.

In inferential statistics, μ is called the *true population mean* and p is called the *true population proportion*. The value(s) assigned to a population parameter based on the value of a sample statistic is called an *estimate*. The sample statistic used to estimate a population parameter is called an *estimator*.

If the mean of the sampling distribution of a statistic equals to the corresponding population parameter, the statistic is called an *unbiased estimator* of the parameter, otherwise it is called a *biased estimator*. The corresponding values of such statistics are called *unbiased* or *biased estimates* respectively. If the sampling distributions of two statistics have the same mean (or expectation), the statistic with the smaller variance is called an *efficient estimator* of the mean while the other statistic is called an *inefficient estimator*. The corresponding values of the statistic are called *efficient* or *inefficient estimates* respectively.

Statistical inference is divided into two major categories: (1) Parameter estimation and (2) Hypothesis testing (see Chapter 6). Parameter estimation is further classified into two major types: (1) Point estimation and (2) Interval estimation.

5.1 POINT ESTIMATION

Estimation is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample. In inferential statistics, μ is called the *true population mean* and p is called the *true population proportion.* The value(s) assigned to a population parameter based on the value of a simple statistic is called an *estimate* of the population parameter.

If the population is small, we can ordinarily determine μ exactly by first taking a census and then computing μ from the population data. But if the population is large, as it often is in practice, taking a census is generally impractical, extremely expensive, or impossible. Nonetheless, we can usually obtain sufficiently accurate information about μ by taking a sample from the population.

An estimate of a population parameter given by a single number is called *point estimate* of parameter. An estimate of a population parameter is given by two numbers between which the parameter may be considered to lie is called an *interval estimator* of the parameter. The sample statistic used to estimate a population parameter is called an *estimator*. Hence, the sample mean, $\bar{X}$ is an estimator of the population mean, μ and the sample population, $\hat{p}$ is an estimator of the population proportion, *p*.

The desired estimators depend on the distribution used. These estimators are not always given by $\bar{X}$ and S^2. The common distributions and their estimators are given below.

Distribution	Parameter	Suggested Estimator(s)
Poisson	λ	$\hat{\lambda} = \bar{X}$
Binominal	p	$\hat{p} = \bar{X}$
Normal	μ, σ^2	$\hat{\sigma}^2 = S^2$ (denominator is $n - 1$)

After a sample is taken and the data are analysed, the result is an *estimate*.

The estimation procedure involves the following steps:

1. Select a sample
2. Collect the required information from the members of the sample
3. Calculate the value of the sample statistic
4. Assign value(s) to the corresponding population parameter

If we select a sample and compute the value of the sample statistic for this sample, the value gives the *point estimate* of the corresponding population parameter.

5.2 INTERVAL ESTIMATION

An interval estimate for a population parameter is called a *confidence interval*. In *interval estimation*, instead of assigning a single value to a population parameter, an interval is constructed around the point estimate and then a probabilistic statement that this interval contains the corresponding population parameter is made. This probabilistic statement is given by the *confidence level*. An interval that is constructed based on the confidence level is called a *confidence interval*. The confidence level associated with a confidence interval states how much confidence we have that the interval contains the true population parameter. The confidence level is denoted by $(1 - \alpha)100\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $(1 - \alpha)$. Note that α is called the *significance level*.

A *tolerance interval* is another important type of interval estimate. For a normal distribution, we know that 95% of the distribution is in the interval $\mu - 1.96\ \sigma, \mu + 1.96\ \sigma$.

Confidence and tolerance intervals bound unknown elements of a distribution. A *prediction interval* provides bounds on one (or more) future observations from the population.

Summarising, the purpose of the 3 types of interval estimates are:

A confidence interval bounds population or distribution parameter (such as the mean weight of a person).

A tolerance interval bounds a selected proportion of a distribution.

A prediction interval bounds future observations from the population or distribution.

A point estimate of a parameter consists of a single value with no indication of the accuracy of the estimate. A confidence interval consists of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the interval.

A confidence interval estimate of a parameter consists of an interval of number obtained from the point estimate of the parameter together with a 'confidence level' that specifies how confident we are that the interval contains the parameter. This is superior to a point estimate because it provides some information about the accuracy of the estimate whereas a point estimate does not.

Interval estimates indicate the precision or accuracy of an estimate and are therefore preferable to point estimates. A statement of the error of precision of an estimate is often called its *reliability*. An interval estimate consists of two sample statistics, L and U. The probability that the interval formed by L and U contains the true value of a parameter is $1 - \alpha$. For instance, to construct an interval estimate on the parameter θ, we find two statistics, L and U, such that

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{5.1}$$

The resulting interval $[L \leq \theta \leq U]$ is called a $100(1 - \alpha)\%$ confidence interval on unknown parameter θ. If a large number of independent samples were taken from the population being considered, approximately $100(1 - \alpha)\%$ of the intervals formed would be expected to contain the true value of θ. Confidence intervals that contain both an L, or lower side, and a U, or upper side, are called *two-sided confidence intervals*. Confidence intervals that contain only a lower or an upper side are called *one-sided confidence intervals*. A one-sided $100(1 - \alpha)\%$ confidence interval on θ is given by

$$L \leq \theta$$

with the probability property

$$P(L \leq \theta) = 1 - \alpha \tag{5.2}$$

Similarly, the one-sided upper $100(1 - \alpha)\%$ confidence interval on θ is given by $\theta \leq U$ with the probability property.

$$P(\theta \leq U) = 1 - \alpha \tag{5.3}$$

The length of the observed confidence interval is an important measure of the quality of the information obtained from the sample. The half-interval length $\theta - L$ or $U - \theta$ is called the *accuracy* of estimator. The longer the confidence interval, the more confident we are that interval actually contains the true value of θ and the less information we have about the true value of θ.

Example E5.1

The following readings give the weights of 6 persons (in kg) picked at random from college students in a particular college: 50, 52, 55, 60, 65 and 70. Find estimates of the following:

(a) the true (population) mean weight of all the students

(b) the true variance of weights of all the students

(c) the true standard deviation of weights of all the students.

SOLUTION:

$$n = 6$$

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{352}{6} = 58.6667 \text{ kg}$$

$$\Sigma(X_i - \bar{X})^2 = (50 - 58.6667)^2 + (52 - 58.6667)^2 + \cdots = 303.3333$$

$$\text{Variance} = \frac{\Sigma(X_i - X)^2}{n-1} = \frac{303.3333}{6-1} = 60.6667 \text{ kg}^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{60.6667} = 7.7889 \text{ kg}$$

(*a*) The estimate of the true mean weight is $\bar{X}$, that is, 58.6667 kg

(*b*) The estimate (in kg^2) of the true variance is

$$\frac{\Sigma(X_i - \bar{X})^2}{n-1} = 60.6667$$

(*c*) The estimate (in kg) of the true standard deviation is $\sqrt{60.6667} = 7.7889$ kg

Example E5.2

A simple random sample of 44 engineers in New York city yielded the following data on their monthly income from employment (in thousands of dollars):

8	14	8	11	7	4	5	7	5
12	6	7	7	4	8	12	7	9
12	6	9	6	5	7	3	10	10
12	8	8	10	5	6	3	8	11
4	5	7	6	11	7	6	8	

(*a*) use the data to obtain a point estimate for the mean monthly income of all engineers working in New York city

(*b*) is your point estimate in part *(a)* likely to equal μ exactly?

Solution:

(*a*) $\Sigma X_i = 334$, and $n = 44$

Hence $\bar{X} = 334/44 = 7.6$

(*b*) It is not likely that $\bar{X}$ is exactly equal to μ (population mean). Some sampling error is to be expected.

5.3 CONFIDENCE INTERVAL ON MEAN, VARIANCE KNOWN

This section presents methods for using sample data to find a point estimate and confidence interval estimate of a population mean. The key requirement in this section is that in addition to having sample data, we also know σ, the standard deviation of the population.

For obtaining a confidence interval for μ, in this section, we make the following assumptions:

1. The sample consists of n independent observations, that is, any one observation does not influence any other.
2. The underlying population is normally distributed with mean μ, which is, of course, unknown ($n > 30$).
3. The population standard deviation σ is known.

Let X be a random variable with unknown mean μ and known variance σ^2, and suppose that a random sample of size n, $X_1, X_2, \ldots, X_n$ is taken. A 100(1 – α)% confidence interval on μ can be obtained by considering the sampling distribution of the sample mean $\bar{X}$. The mean of $\bar{X}$ is μ and the variance σ^2/n. Therefore, the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is taken to be a standard normal distribution.

The distribution of $Z = \left(\bar{X} - \mu\right)\Big/\left(\sigma/\sqrt{n}\right)$ is shown in Fig. 5.1. From this figure we note that

$$P\{-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\} = 1 - \alpha$$

or $$P\left\{-Z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right\} = 1 - \alpha$$

This can be arranged as

$$P\left\{\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right\} = 1 - \alpha \tag{5.4}$$

From Eqs. (5.1) and (5.4), we note that the 100(1 – α)% two-sided confidence interval on μ is

$$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \tag{5.5}$$

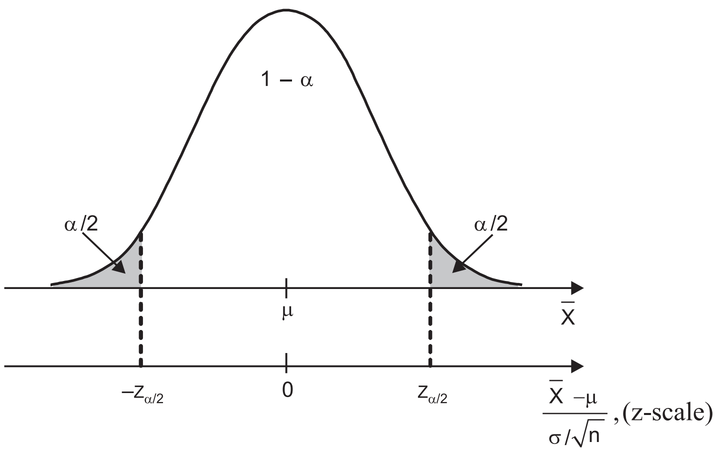

Fig. 5.1: The distribution of z

From Eq. (5.4), we observe that $\bar{X}$ will differ from μ by at most $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ with probability (1 – α). The quantity $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is, therefore, called the *maximum error of estimate* of μ at the (1 – α)100 per cent level. The maximum error of estimate of μ is commonly called the *margin of error* or the *sampling error*.

Hence the margin of error of estimate of μ at the (1 – α)100 per cent level is the maximum error in estimating μ and is given by

$$\text{margin of error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The interval $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is a random interval and the probability that it covers the mean μ is (1 – α). The end points

$$L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and
$$U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

of the interval are random variables. Their values depend on the value of $\bar{X}$ which, in turn, depends on the sample values.

Therefore, a confidence interval for μ is often given simply by specifying the limits as

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean based on n observations.

The above construction of the confidence interval for μ was based on the assumption that the population is normally distributed. This assumption was necessary since it permitted us to proceed by stating that $\bar{X}\left(\text{and } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)$ has a normal distribution. If n is large (at least 30), the assumption of a normal distribution is crucial, because the central limit theorem allows us to proceed by stating that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has approximately a normal distribution (we still get the same confidence limits given above, but now they are approximate limits). When we say that the confidence interval is *exact*, we mean that the true confidence level equals 1 – α; similarly, when we say that the confidence interval is *approximately correct*, we mean that the true confidence level only approximately equals 1 – α.

The procedure to find a confidence interval for a population mean when the standard deviation is known is shown in Table 5.1.

Example E5.3

Refer to Example E5.1. Construct a 90% confidence interval for μ, the true mean weight of the students.

Solution:

From Example E5.1, we have

$$n = 6,\ \bar{X} = 58.6667 \text{ kg and } \sigma = 7.7889 \text{ kg.}$$

For a 90% confidence interval, $\alpha = 0.10$. Thus, $z_{\alpha/2} = z_{0.05} = 1.645$ (from the table in Appendix-*E*). The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$58.6667 - 1.645\frac{7.7889}{\sqrt{6}} \le \mu \le 58.6667 + 1.645\frac{7.7889}{\sqrt{6}}$$

or

$$58.6667 - 5.2308 \le \mu \le 58.6667 + 5.2308$$

$$53.4359 \le \mu \le 63.8975$$

which can be simplified to $53.44 < \mu < 63.90$, giving the confidence interval (53.44, 63.90). Hence, with 90% confidence we estimate that the mean weight μ of the students in that college is between 53.44 and 63.90 kg.

The term *normal population* is used here as an abbreviation for "the variable under consideration is normally distributed". The z-interval procedure works reasonably well even when the variable is not normally distributed and the sample size is small, or moderate, provided the variable is not too far from being normally distributed. For large samples, say, of size 30 or more ($n \ge 30$), the z-procedure can be used essentially without restriction. For samples of moderate size ($15 \le n \le 30$), the z-interval procedure can be used unless the data contains outliers or the variable under consideration is far from being normally distributed. For small samples, say, of size less than 15 ($n \le 15$), the z-interval procedure should be used only when the variable under consideration is normally distributed or very close to being so. If outliers are present then their removal is justified in a data set in order to use the z-interval procedure.

Table 5.1: Procedure to find a confidence interval for a population mean when σ is known

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$.

Step 2: The confidence interval for μ is from

$$\bar{X} - z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\bar{X}$ is computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.4

Refer to Example E5.2. Assume that the recent monthly income of engineers from employment in New York city are normally distributed with a standard deviation of $8100.

(a) determine a 95.44% confidence interval for the mean cost, μ, of all recent engineers in New York city

(b) interpret your result in part *(a)*

(c) does the mean monthly income from employment of all engineers in New York city lie in the confidence interval obtained in part *(a)*?

Explain your answer.

SOLUTION:

(a) The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $x_{\alpha/2} = 2.0$ (from the table in Appendix-*E*).

Therefore $7.6 - 2\frac{(2.4)}{\sqrt{44}}$ to $7.6 + 2\frac{(2.4)}{\sqrt{44}}$

or 6.9 to 8.3

(b) Since we know that 95.44% of all samples of 44 engineers employment monthly income have the property that the interval from $\bar{X}$ – 0.7 to $\bar{X}$ + 0.7 contains μ, we can be 95.44% confident that the interval from 6.9 to 8.3 contains μ.

(c) The confidence interval in part *(a)* would be exact if the population of engineers' monthly employment income were exactly a normal distribution. However, since engineers' monthly employment income is a discrete random variable and the normal distribution is continuous, monthly employment income cannot follow a normal distribution exactly.

Example E5.5

It has been established that in a certain packaging process, the packages are of an average weight of μ but that μ changes over time as the process adjustment changes through wear. It was also found that the variance of the package weight is a constant at 16 kg even though the mean changes overtime. A sample of 36 packages has revealed their mean weight $\bar{X}$ of 100 kg. Assuming that the weight of the individual packages are normally distributed around μ, *(a)* find the 90% confidence limits for μ, *(b)* find the 95% confidence limits for μ.

SOLUTION:

(a) Here $1 - \alpha = 0.90$ or $\alpha = 0.10$ and $\alpha/2 = 0.05$.

$z_{0.05} = 1.645$ (from the table in Appendix-*E*), $\alpha = \sqrt{16} = 4$ kg

By computation,

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.645(4)}{\sqrt{36}} = 100 - 1.0967 = 98.9033$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.645(4)}{\sqrt{36}} = 100 + 1.0967 = 101.0967$$

Hence, we are 90% confident that m lies between 98.9033 and 101.0967 kg.

(b) Here $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*)

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.96(4)}{\sqrt{36}} = 100 - 1.3067 = 98.6933$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.96(4)}{\sqrt{36}} = 100 + 1.3067 = 101.3067$$

Therefore, we are 95% confident that μ lies between 98.6933 and 101.3067 kg.

One-sided Confidence Intervals

One-sided confidence intervals for μ are obtained by setting either $L = -\infty$ or $U = \infty$ and replacing $z_{\alpha/2}$ by z_α. The 100(1 – α)% upper-confidence interval for μ is

$$\mu \le \bar{X} + z_\alpha \sigma/\sqrt{n} \tag{5.6}$$

and the 100(1 – α)% lower-confidence interval for μ is

$$\bar{X} - z_\alpha \sigma/\sqrt{n} \le \mu \tag{5.7}$$

It is most preferable to obtain and pinpoint the parameter with 100% certainty. Due to the variability inherent in the population, it is hard to accomplish this. One can try to achieve the turn goal of a high level of confidence and a narrow interval.

Length of the confidence interval for μ: The length of any interval from *a* to *b* is (*b* – *a*). Hence, the length of the confidence interval is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Therefore, the length of the confidence interval for μ

$$= 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2 \text{ (margin of error)}$$

The length of the confidence interval is also called its *width*.

Hence, the length of the confidence interval is twice the margin of error or the margin of error is one-half the length of the confidence interval.

We also note that the length of the confidence interval for μ does not depend on $\bar{X}$, but it depends on $z_{\alpha/2}$, σ, and *n*. Hence, in order to achieve a high accuracy in estimating μ (that is, a narrow confidence interval with high degree of confidence) then one way to accomplish this goal is to pick an appropriately large sample. The margin of error is the standard error of the mean multiplied by $z_{\alpha/2}$. The length of a confidence interval, and thus the precision with which $\bar{X}$ estimates μ, is determined by the margin of error. Increasing the confidence level while keeping the sample size the same will increase the value of $z_{\alpha/2}$ and hence the length of the confidence interval (decrease the precision of the estimate). Increasing the sample size while keeping the same confidence level will decrease the margin of error and the length of the confidence interval (increase the precision).

Example E5.6

A random sample of 18 venture-capital investments in the fiber optics sector yields the following data, in crores of rupees:

2.04	5.48	5.60	5.96	6.27	10.51
4.13	5.58	5.74	5.95	6.67	8.63
4.21	4.98	6.66	7.71	8.64	9.21

(a) determine a 95% confidence interval for the mean amount, μ, of all venture, capital investments in the fiber optics sector. Assume that the population standard deviation is 2.04 crores of rupees

(b) interpret your finding in part *(a)*

(c) find a 99% confidence interval for μ

(d) why is the confidence interval found in part (*c*) longer than the one in part *(a)*?

(e) which confidence interval yields a more precise estimate of μ? Explain your answer.

SOLUTION:

(a) $n = 18$ and $\Sigma X = 113.97$ crores

$$\bar{X} = \frac{\Sigma X}{n} = \frac{113.97}{18} = 6.33 \text{ crores}$$

The 95% confidence interval for μ is

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*),

$$6.33 - 1.96\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 1.96\frac{(2.04)}{\sqrt{18}}$$

or 5.39 to 7.27 crores.

(b) We can be 95% confident that the interval from 5.39 crores to 7.27 crores contains the population mean venture capital investment in the fiber-optics sector.

(c) $n = 18$, $\bar{X} = 6.33$ and $\sigma = 2.04$

$\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.575$ (from the table in Appendix-*E*),

Hence
$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$6.33 - 2.575\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 2.575\frac{(2.04)}{\sqrt{18}}$$

or 5.09 to 7.57 crores.

(*d*) The confidence interval in (*c*) is longer than the one part (*a*) because we have changed the confidence level from 95% in part (*a*) to 99% in part (*c*). We notice that increasing the confidence level from 95% to 99% increases the $z_{\alpha/2}$ value from 1.96 to 2.575. The larger z-value, in turn, results in a longer interval. In order to accomplish a higher level of confidence that the interval contains the population mean, we need a longer interval.

(*e*) See Fig. E5.6.

The 95% confidence interval is shorter and therefore provides a more precise estimate of μ.

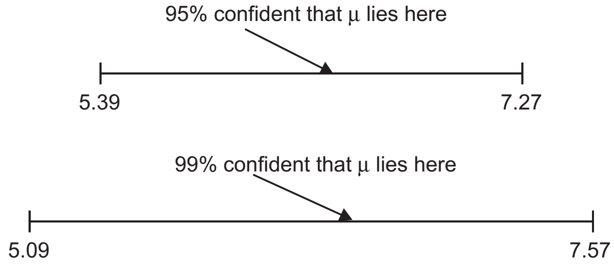

Fig. E5.6

5.4 CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE UNKNOWN

The method presented here is valid for any arbitrating sample size but is particularly important when the sample size is small. For setting a confidence interval for μ, we make the following assumptions:

1. The sample is a simple random sample.
2. The observations are picked from the population under study and are independent.
3. The population has a normal distribution.

Here we find a confidence interval on the mean of a distribution when the variance is unknown. Suppose a random sample of size n, $X_1, X_2, \ldots, X_n$ is available and $\bar{x}$ and S^2 are the sample mean and sample variance, respectively. In order to use a valid confidence interval when the sample size small, we make an assumption that the underlying population is normally distributed. This leads to confidence intervals based on the t-distribution. Therefore, let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We know the sampling distribution of the statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is the t-distribution with $n - 1$ degrees of freedom.

The distribution of $t = \left(\bar{X} - \mu\right)\Big/\left(S/\sqrt{n}\right)$ is shown in Fig. 5.2. Letting $t_{\alpha/2,\, n-1}$ be the upper $\alpha/2$ percentage point of the t-distribution with $n - 1$ degree of freedom, we note from Fig. 5.2, that

$$P\left\{-t_{\alpha/2,\, n-1} \le t \le t_{\alpha/2,\, n-1}\right\} = 1 - \alpha$$

or $$P\left\{-t_{\alpha/2,\,n-1} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2,\,n-1}\right\} = 1-\alpha \tag{5.8}$$

Rearranging Eq. (5.8)

$$P\left\{\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n}\right\} = 1-\alpha \tag{5.9}$$

Comparing Eqs. (5.4) and (5.9), we see that a 100 (1 – α)% two-sided confidence interval on μ is

$$\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n} \tag{5.10}$$

A 100 (1 – α) % lower-confidence interval on μ is given by

$$\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \tag{5.11}$$

and a 100 (1 – α) % upper-confidence interval on μ is given by

$$\mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n} \tag{5.12}$$

The above method assumes that the samplings are from a normal population.

The procedure to find a confidence interval for a population mean when the standard deviation σ is unknown, is given in Table 5.2. Properties and guidelines for use of the *t*-interval procedure are the same as those for the *z*-interval procedure. The *t*-interval procedure is robust to moderate violations of the normality assumptions.

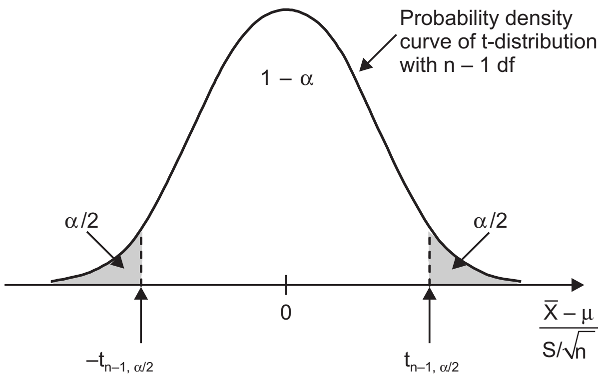

Fig. 5.2: t-values such that there is an area α/2 in the right tail and an area α/2 in the left tail of the distribution

Table 5.2: Procedure to find a confidence interval for a population mean when σ is unknown

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1: For a confidence level of $1 - \alpha$, use the Table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n - 1$, where n is the sample size

Step 2: The confidence interval for μ is from

$$\bar{X} - t_{\alpha/2} \bullet \frac{S}{\sqrt{n}} \text{ to } \bar{X} + t_{\alpha/2} \bullet \frac{S}{\sqrt{n}}$$

where $t_{\alpha/2}$ is found in step 1 and $\bar{X}$ and S are computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.7

Find the following confidence intervals for μ_d assuming that the populations of paired differences are normally distributed.

(a) $n = 9$, $\bar{X} = 25$, $S = 13$, confidence level 99%.

(b) $n = 26$, $\bar{X} = 13$, $S = 5$, confidence level 95%.

(c) $n = 12$, $\bar{X} = 35$, $S = 12$, confidence level 90%.

Solution:

(a) The *t*-value for $\nu = n - 1 = 9 - 1 = 8$ d.o.f. and 0.005 area in the right tail is 3.355 (from the table in Appendix-*G*). The 99% confidence interval with $t_{0.005,8} = 3.355$ is

$$\bar{X} - t_{\alpha/2,\nu} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\nu} S/\sqrt{n}$$

$$25 - 3.355(13)/\sqrt{9} \le \mu \le 25 + 3.355(13)\sqrt{9}$$

$$25 - 14.538 \le \mu \le 25 + 14.538$$

$$10.462 \le \mu \le 39.538$$

The complete statement of the confidence interval, with the associated probability is

$$P(10.462 \le \mu \le 39.538) = 0.99$$

(b) Using the data given, a 95% confidence interval with $t_{0.025,25} = 2.060$ from the table in Appendix-*G*, is given

$$13 - 2.060(5)/\sqrt{26} \le \mu \le 13 + 2.060(5)/\sqrt{26}$$

$$13 - 2.02 \le \mu \le 13 + 2.02$$

$$10.98 \le \mu \le 15.02$$

The complete statement of the confidence interval with the associated probability is

$P(10.98 \le \mu \le 15.02) = 0.95$

(c) Using the data given, a 90% confidence interval with $t_{0.05,11}$ = 1.796 from the table in Appendix-*G*, is given by

$$35 - 1.796(12)/\sqrt{12} \le \mu \le 35 + 1.796(12)/\sqrt{12}$$

$$35 - 0.16 \le \mu \le 35 + 0.16$$

$$34.84 \le \mu \le 35.16$$

The complete statement of the confidence interval with the associated probability is

$P(34.84 \le \mu \le 35.16) = 0.90$

Example E5.8

Measurements on the percentage of enrichment of 12 fuel rods were reported as follows:

2.94	3	2.9	2.75	3	2.95
2.75	2.95	2.92	2.81	3.05	2.8

Find a 99% two-sided confidence interval on the mean percentage of enrichment of fuel rods. Can we state the mean percentage of enrichment is 2.95%? Why?

SOLUTION:

Here $n = 12$ and $\Sigma X = 34.82$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{34.82}{12} = 2.9017$$

Also $\Sigma(X_i - \bar{X})^2 = 0.112567$

Hence $S = \sqrt{\dfrac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\dfrac{0.112567}{12-1}} = 0.10116$

The 99% two-sided confidence interval on the mean per cent enrichment is found from

$$\bar{X} - t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right) \le \mu \le \bar{X} + t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right)$$

For $\alpha = 0.01$ and $n = 12$, $t_{\alpha/2,\, n-1} = t_{0.005,11} = 3.106$ (from the table in Appendix-*G*).

Therefore $2.9017 - 3.106\left(\dfrac{0.10116}{\sqrt{12}}\right) \le \mu \le 2.9017 + 3.106\left(\dfrac{0.10116}{\sqrt{12}}\right)$

$$2.9017 - 0.896625$$

or $2.005 \le \mu \le 3.798325$

We can state that the mean percentage of enrichment 2.95 is included in the 99% two-sided confidence interval.

5.5 CONFIDENCE INTERVAL ON THE VARIANCE OF A NORMAL DISTRIBUTION

Here, we accept S^2 as providing a good estimate for σ^2. For setting a confidence interval for σ^2, we make the following assumptions:

1. The observations are independent.
2. The parent population has a normal distribution.
3. The sample is a simple random sample.

Suppose that X is normally distributed with unknown mean μ and unknown variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n, let S^2 be the sample variance. We note that the sampling distribution of

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

is chi-square with $n - 1$ degrees of freedom. This distribution is shown in Fig. 5.3.

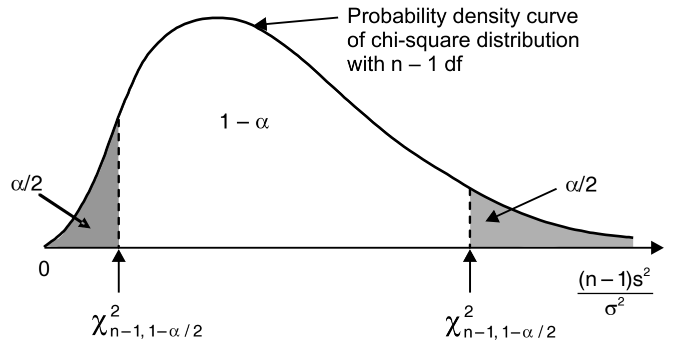

Fig. 5.3: Chi-square values such that areas 1 – α/2 and α/2 are to their right

We observe from Fig. 5.3 that,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \chi^2 \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

or
$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha \tag{5.13}$$

Equation (5.13) can be rearranged to give

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha \tag{5.14}$$

Comparing Eqs. (5.14) and (5.4), we see that a 100 (1 – α) % two-sided confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{5.15}$$

when S^2 is the sample variance based on n observations from a normal population.

In order to find a 100 (1 – α)% lower-confidence interval on σ^2, set $U = \infty$ and replace $\chi^2_{\alpha/2,n-1}$ with $\chi^2_{\alpha,n-1}$, giving

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \tag{5.16}$$

The 100 (1 – α)% upper-confidence interval is found by setting $L = 0$ and replacing $\chi^2_{1-\alpha/2,n-1}$ with $\chi^2_{1-\alpha,n-1}$, resulting in

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \tag{5.17}$$

In all the above Eqs. (5.13 to 5.17), $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ represent values from the chi-square distribution with $n - 1$ degrees of freedom such that they leave, respectively, area of $1 - \alpha/2$ and $\alpha/2$ to their right, as shown in Fig. 5.3. The procedure to find a confidence interval for a population standard deviation is given in Table 5.3.

Table 5.3: Procedure to find a confidence interval for a population standard deviation, σ

Assumptions:	
1.	Simple random sample
2.	Normal population
Step 1:	For a confidence level of $1 - \alpha$, use the table in Appendix-*F* to find $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ with $df = n - 1$
Step 2:	The confidence interval for σ is from $\sqrt{\frac{(n-1)}{\chi^2_{\alpha/2}}}S$ to $\sqrt{\frac{(n-1)}{\chi^2_{1-\alpha/2}}}S$ where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are found in step 1, n is the sample size and S is computed from the sample data obtained
Step 3:	Interpret the confidence interval.

Example E5.9

Data are collected on the driving time required to reach place *A* from place *B*. The driving time is assumed to follow a normal distribution. Twenty-one driving times are collected and the sample variance is calculated to be 2 hours. A 99% confidence interval on the true variance is desired.

SOLUTION:

The two-sided confidence interval is given by (Appendix-*F*)

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu}}$$

where $\nu = n - 1$.

From the table in Appendix-*F* with $\nu = 21 - 1 = 20$ and $\alpha/2 = 0.005$, we find that $\chi^2_{0.005,20} = 39.997 = \chi^2_{0.995,20} = 7.434$.

Therefore,

$$\frac{(21-1)(2^2)}{39.997} \le \sigma^2 \le \frac{(21-1)(2^2)}{7.434} \text{ or } 2 \le \sigma^2 \le 10.761$$

The complete statement of the confidence interval with the associated probability is

$P(2 \text{ hrs} \le \sigma^2 \le 10.761 \text{ hrs}) = 0.99$

Example E5.10

The capability of a specific gauge can be studied by measuring the weight of paper. The data for repeated measurements of one sheet of paper are given in Table E5.10. Construct a 95% one-sided, upper confidence interval for the standard deviation of these measurements.

Table E5.10

3.481	3.449	3.484	3.476	3.473
3.478	3.473	3.465	3.475	3.471
3.472	3.471	3.478	3.473	3.475

SOLUTION:

A 95% upper-confidence interval is found from Eq. (5.17) as follows:

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{0.95,14}}$$

$n = 15$, $S = 0.001354$ and $\chi^2_{0.95,14} = 6.57$ (from the table in Appendix-*F*).

$$\Sigma X_i = 52.094\,,\ \Sigma(X_i - \bar{X})^2 = 0.000901$$

$$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = 0.001354 \text{ and } S^2 = 1.83 \times 10^{-6}$$

$$\sigma^2 = \frac{14(1.83 \times 10^{-6})}{6.57}$$

$$\sigma^2 \le 3.9073 \times 10^{-6}$$

The above can be converted into a confidence interval on the standard deviation σ by taking the square root of both sides, resulting in $\sigma = 0.00197667$.

Therefore, at the 95% level of confidence, the data indicates that the process standard deviation could be as large as 0.00197667.

5.6 CONFIDENCE INTERVAL ON A POPULATION PROPORTION

We will now give an interval estimate for population proportion under the following assumptions:

1. The sample consists of n independent observations.
2. The sample size is large.
3. The population proportion is not too close to 0 or 1.

If a random sample size n has been taken from a large (possibly infinite) population, and $X(\leq n)$ observations in this sample belong to a class of interest, then $\hat{p} = X / n$ is the point estimator of the proportion of the population belonging. n and p are the parameters of binomial distribution. The sampling distribution $\hat{p}$ is approximately normal mean p and variance $p(1 - p)/n$, if p is not too close to either 0 or 1, and if n is relatively large. Therefore, the distribution of

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

is approximately standard normal distribution.

We also observe that

$$P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\} = 1 - \alpha$$

or
$$P\left\{-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \leq Z_{\alpha/2}\right\} = 1 - \alpha \tag{5.18}$$

Rearranging Eq. (5.18)

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right\} = 1 - \alpha \tag{5.19}$$

The quantity of $\sqrt{p(1-p)/n}$ is called the *standard error of the point estimator* $\hat{p}$. Replacing p by $\hat{p}$ in the standard error, giving an estimated standard error.

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} = 1 - \alpha \tag{5.20}$$

The above formula, Eq. (5.20), is applicable only if the observed sample proportion is not too close to 0 or 1 and the observed number of successes X and the observed number of failures $(n - X)$ both exceed 5. If these conditions are not met, the procedure is not recommended.

The above procedure depends on the adequacy of the normal approximation to the binomial. Conservatively speaking, this requires that np and $n(1 - p)$ be greater than or equal to 5. In situations where the approximation is inappropriate (especially when n is small) other methods must be used. Tables of binomial distribution could be used to obtain a confidence interval for p.

We may find approximate one-sided confidence bounds on p by a simple modification of Eq. (5.20).

The approximate $100(1 - \alpha)\%$ lower and upper confidence bounds are

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \quad \text{and} \quad p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

respectively.

The procedure to find a confidence interval for a population proportion, p is given in Table 5.4.

Table 5.4: Procedure to find a confidence interval for a population standard proportion, p

Assumptions:
1. Simple random sample
2. The number of successes, x, and the number of failures, $n - x$ are both 5 or greater

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$

Step 2: The confidence interval for p is from

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ to } \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\hat{p} = x/n$ is the sample proportion

Step 3: Interpret the confidence interval.

Example E5.11

In a random sample of 300 automobile crankshaft bearings, 12 have a surface finish that is rougher than the specifications allow.

(*a*) Calculate a 95% two-sided confidence interval on the fraction of defective crankshafts produced by this particular process.

(*b*) Calculate a 95% upper confidence bound on the fraction of defective crankshafts.

Solution:

(*a*) 95% confidence interval on the fraction defectives produced with the process.

$$\hat{p} = \frac{12}{300} = 0.04,\ n = 30,\ z_{\alpha/2} = 1.96 \text{ (from the table in Appendix-}E\text{).}$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.04 - 1.96\sqrt{\frac{0.04(1-0.04)}{300}} \le p \le 0.04 + 1.96\sqrt{\frac{0.04(1-0.04)}{300}}$$

$$0.01 - 0.022175 \le p \le 0.04 + 0.02275$$

$$0.017825 \le p \le 0.062175$$

(*b*) 95% upper confidence bound:

$z_{\alpha/2} = z_{0.05} = 1.65$ (from Appendix-*E*)

$$p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$p \le 0.04 + 1.65\sqrt{\frac{0.04(1-0.04)}{300}}$$

or $\le 0.04 + 0.018668$

Hence $p \le 0.058668$

Example E5.12

A sample of 300 of electronic components produced showed 36 non-conforming ones. Construct a 95% confidence interval for this. Assume binomial distribution.

Solution:

A two-sided confidence interval on p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Since n is large and the underlying distribution is binomial, $\hat{p}$ = 36/300 = 0.12

A 95% confidence interval, with $Z_{0.025} = 1.96$ (from the table Appendix-*E*) is given by

$$0.12 - 1.96\sqrt{\frac{(0.12)(0.88)}{300}} \le p \le 0.12 + 1.96\sqrt{\frac{(0.12)(0.88)}{300}}$$

or

$$0.12 - 1.96\,(0.01876) \le p \le 0.12 + 1.96\,(0.01876)$$

$$0.12 - 0.0368 \le p \le 0.12 + 0.0368$$

$$0.0832 \le p \le 0.1568$$

The complete statement of the confidence interval with the associated probability is

$$P(0.0832 \le p \le 0.1568) = 0.95$$

Margin of Error:

The margin of error from Eq. (5.20) is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

It can be shown that the maximum value of $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is $z_{\alpha/2}\sqrt{\frac{1}{4n}}$ and occurs when $\hat{p} = \frac{1}{2}$. This called the *safe* or *conservative margin of error*, whatever the actual value of $\hat{p}$ in the population.

Hence, conservative margin of error = $z_{\alpha/2}\sqrt{\frac{1}{4n}}$.

Example E5.13

Determine the margin of error in Example E5.12.

Solution:

The margin of error is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*), $\hat{p}$ = 36/300 = 0.12

Hence, margin of error = $1.96\sqrt{\dfrac{0.12(1-0.12)}{300}}$ = 0.0368

5.7 CONFIDENCE INTERVAL ON THE DIFFERENCE IN TWO MEANS, VARIANCE KNOWN

The procedure presented in this section for finding a confidence interval is valid if the following assumptions are justified:

1. The two populations are normally distributed. This assumption is not very important if both sample sizes are at least 30 (large sample size).
2. The standard deviations of the two populations are known.
3. Two random samples are picked, one from each population. They are independent, that is, any outcome in one sample does not influence any outcome in the other sample.
4. Within each sample the outcomes are independent.

Consider two independent random variables X_1 with unknown mean μ_1 and known variance σ_1^2 and X_2 with unknown mean μ_2 and variance σ_2^2. We wish to find a 100 (1 – α)% confidence interval on the difference in means μ_1, μ_2. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample of n_1 observations from X_1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample of n_2 observations from X_2. If $\bar{X}_1$ and $\bar{X}_2$ are the sample means, the statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is standard normal if X_1 and X_2 are normal or approximately standard normal if the conditions of the central limit theorem apply, respectively. Also, σ_1 and σ_2 are the population standard deviations. From Fig. 5.2 it is clear that

$$P\{-z_{\alpha/2} \le z \le z_{\alpha/2}\} = 1 - \alpha$$

or

$$P\left\{-z_{\alpha/2} \le \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Rearranging, we have

$$P\left\{\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 1 - \alpha \qquad (5.21)$$

Comparing Eqs. (5.4) and (5.21), we note that 100(1 – α)% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad (5.22)$$

One-sided confidence intervals on $\mu_1 - \mu_2$ may also be obtained. A 100 $(1 - \alpha)$% upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad (5.23)$$

and a $100(1 - \alpha)$ % lower-confidence interval is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \quad (5.24)$$

The confidence level $(1 - \alpha)$ is exact when the populations are normal. For non-normal populations, the confidence level is approximately valid for large sample sizes.

The margin of error E is

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad (5.24a)$$

Example E5.14

The mean hourly wage for male workers was \$15 and \$13 for female workers in a particular manufacturing sector in the year 2000. These two estimations were based on random samples of 1000 and 1200 workers taken, respectively from two independent populations. The standard deviations of the two populations are known to be \$2 and \$1.50 respectively. Construct a 95% confidence interval for the difference between the mean hourly wages of the two populations.

Solution:

For the 95% confidence level $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*).

The 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$15 - 13 \pm 1.96\sqrt{\frac{2^2}{1000}+\frac{1.5^2}{1200}} = 2 \pm 1.96\,(0.2046) = 2 \pm 0.4011 = \$1.60 \text{ to } \$2.40$$

We have obtained (\$1.60, \$2.40) as an 95% confidence interval for $\mu_1 - \mu_2$.

Example E5.15

Two different formulations of oxygenated motor fuel are being tested to study their road octane numbers. The variances of road octane number for formulation *A* is $\sigma_1^2 = 1.5$ and for formulation *B* it is $\sigma_2^2 = 1.2$. Two random samples of sizes $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane numbers observed are

$\bar{X}_1 = 90$ and $\bar{X}_2 = 93$. Assume normality and calculate a 95% confidence interval on the difference in means.

SOLUTION:

95% confidence interval: $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ (from the table in Appendix-*E*).

$$\left(\bar{X}_1 - \bar{X}_2\right) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$90 - 93 - 1.96\sqrt{\frac{1.5}{15} + \frac{1.2}{20}} \le \mu_1 - \mu_2 \le 90 - 93 + 1.96\sqrt{\frac{1.5}{15} + \frac{1.2}{20}}$$

$$-3 - 0.784 \le \mu_1 - \mu_2 \le -3 + 0.784$$

or $$-3.784 \le \mu_1 - \mu_2 \le -2.216$$

With 95% confidence, we believe the mean road octane number for formulation *B* exceeds that of formulation *A* since 0 is not included in the confidence interval.

5.8 CONFIDENCE INTERVAL ON THE DIFFERENCE IN MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES UNKNOWN

We make the following assumptions:

1. The two populations are normally distributed.
2. The population variances σ_1^2 and σ_2^2, though unknown are the same, say, each equal to σ^2.
3. Two random and independent samples are picked, one from each population.

Consider two independent normal random variables, say X_1 with mean μ_1 and variance σ_1^2, and X_2 with mean μ_2 and variance σ_2^2. Both the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown. Assume that both variances are equal; that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to find a 100 $(1 - \alpha)$% confidence interval on the difference in means $\mu_1 - \mu_2$.

Random samples of size n_1 and n_2 are taken on X_1 and X_2, respectively; let the sample means be denoted by $\bar{X}_1$ and $\bar{X}_2$, and the sample variances be denoted S_1^2 and S_2^2. Since both S_1^2 and S_2^2 are estimates of the common variance σ^2, we may obtain a combined (or "pooled") estimator of σ^2 as the pooled estimate of the common population standard deviation, S_p, is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{5.25}$$

To find the confidence interval for $\mu_1 - \mu_2$, note that the distribution of the statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Hence,

$$P\left\{-t_{\alpha/2, n_1+n_2-2} \le t \le t_{\alpha/2, n_1+n_2-2}\right\} = 1 - \alpha$$

or
$$P\left\{-t_{\alpha/2, n_1+n_2-2} \le \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \le t_{\alpha/2, n_1+n_2-2}\right\} = 1 - \alpha$$

Here $t_{\alpha/2, n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Rearranging,

$$P\left\{\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\} = 1 - \alpha \tag{5.26}$$

Hence, a 100 (1 – α)% two-sided confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.27}$$

A one-sided 100 (1 – α)% lower-confidence interval on $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \tag{5.28}$$

and a one-sided 100 (1 – α)% upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.29}$$

The margin of error E is given by

$$E = z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Since σ is not known, we replace it with its pooled estimate S_p and thereby we cannot use z-values from the normal table. Here, we use values from the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. This is especially important if the degrees of freedom are less than 30 because then the t-values will be markedly different from the z-values. Hence, the marginal of error is

$$E = t_{n_1+n_2-2, \alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where
$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\Sigma(X_{1i} - \bar{X}_1)^2 + \Sigma(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, based, respectively on n_1 and n_2 observations, and S_p is the pooled estimate of σ. The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.5.

Table 5.5: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:
1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n_1 + n_2 - 2$

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.16

The overall distance travelled by a golf ball is tested by striking the ball with a specially made mechanical golfer. Ten randomly selected balls of two different brands *A* and *B* are tested and the overall distance measured as shown below:

Brand *A*	263	267	271	273	275	276	279	283	286	287
Brand *B*	244	258	261	263	265	268	270	271	273	280

Construct a 95% two-sided confidence interval on the mean difference in overall distance between the two brands of golf balls.

Solution:

Here $n_1 = 10, n_2 = 10, \Sigma X_{1i} = 2760, \Sigma X_{2i} = 2653$

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{2760}{10} = 276\ , \quad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{2653}{10} = 265.3$$

$$\Sigma(X_{1i} - \bar{X}_1)^2 = 564 \text{ and } \Sigma(X_{2i} - \bar{X}_2)^2 = 868.1$$

$$S_1^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = 62.6667 \text{ and } S_1 = 7.9162$$

$$S_2^2 = \frac{\Sigma(X_{2i} - \bar{X}_2)^2}{n_2 - 1} = 96.4556 \text{ and } S_2 = 9.8212$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)(62.6667)+(10-1)(96.4556)}{10+10-2}}$$

or $S_p = 8.9197$

95% two-sided confidence interval: $t_{\alpha/2, n_1+n_2-2} = t_{0.025,18} = 2.101$ (from the table in Appendix-*G*).

$$\left(\bar{X}_1 - \bar{X}_2\right) - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$(276 - 265.3) - 2.101(8.9197)\sqrt{\frac{1}{10}+\frac{1}{10}} \le \mu_1 - \mu_2 \le (276 - 265.3) + 2.101(8.9197)\sqrt{\frac{1}{10}+\frac{1}{10}}$$

$$2.3191 \le \mu_1 - \mu_2 \le 19.0809.$$

Example E5.17

To compare the tensile strength of two materials, 12 test specimens of each material were tested. The order was decided at random. The higher the test value, the better the strength of these materials. The data obtained are summarised in Table E5.20.

Table E5.20

Material A	12.5	11.7	9.70	9.60	10.5	9.60	9.40	11.5	8.70	11.5	10.6	9.7
Material B	9.40	8.40	11.6	7.20	9.70	7.00	10.4	8.20	7.30	9.20	6.90	12.7

Solution:

For the both materials

Material *A*	Material *B*
$n = 12$	$n = 12$
$\Sigma X_{1i} = 124.8$	$\Sigma X_{2i} = 108$
$\bar{X}_1 = \frac{\Sigma X_{1i}}{n} = \frac{124.8}{12} = 10.4$	$\bar{X}_2 = \frac{\Sigma X_{2i}}{n} = \frac{108}{12} = 9$
$\Sigma X_{1i}^2 = 1312$	$\Sigma X_{2i}^2 = 1010.64$
$\left(\frac{1}{n_1}\right)\left(\Sigma X_{1i}\right)^2 = 1297.92$	$\left(\frac{1}{n_2}\right)\left(\Sigma X_{2i}\right)^2 = 972$
$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 = 14.08$	$\Sigma\left(X_{2i} - \bar{X}_2\right)^2 = 38.64$

Hence, the summary table is

Material	n	Degree of freedom	$\bar{X}$	Sum of squares
A	12	11	10.4	14.08
B	12	11	9	38.64

Total: 52.72

Therefore, $d = \bar{X}_1 - \bar{X}_2 = 1.4$

The pooled estimate of variance is $S^2 = 52.72/22 = 2.4$

The estimated variance of the difference between the means of

$$S_d^2 = S^2\left(\frac{1}{n}+\frac{1}{n}\right) = 2.4\left(\frac{2}{12}\right) = 0.4$$

The estimated standard deviation of d is

$$S_d = \sqrt{0.4} = 0.63$$

Hence, if we desire a 95% confidence interval, we must obtain $t_{0.025,22} = 2.074$ (from the table in Appendix-*G*). The 95% confidence interval now be formed using the relationship

$$L = d - t_{\alpha/2, n_1+n_2-2}\ S_d = 1.40 - (2.074)(0.63) = 0.09$$

$$U = d + t_{\alpha/2, n_1+n_2-2}\ S_d = 1.40 + (2.074)(0.63) = 2.71$$

We can be a 95% confidence that the difference between the actual means is between 0.09 and 2.71.

When we cannot assume $\sigma_1^2 = \sigma_2^2$, we can find a $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ using the fact that the statistic

$$t^* = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

distributed approximately as t with degrees of freedom given by

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1-1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2-1}} \tag{5.30}$$

Therefore, an approximate $100\,(1-\alpha)\%$ two-sided confidence interval on $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{5.31}$$

Upper (lower) one-sided confidence limits may be found by replacing the lower (upper) confidence limit with $-\infty$ (∞) and changing $\alpha/2$ to α.

The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.6.

Table 5.6: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:
1. Simple random samples
2. Independent samples
3. Normal populations or large samples

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = \nu$, where

$$\nu = \frac{\left[\left(S_1^2/n_1\right)+\left(S_2^2/n_2\right)\right]^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1}+\frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

rounded down to the nearest integer.

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Step 3: Interpret the confidence interval.

Example E5.18

The administration of a college wants to test if the mean GPAs (grade point averages) of all male and female students who actively participate in sports are different. A random sample of 28 male students and 24 female students who are actively participated in sports. The mean GPAs of the two groups were found to be 2.00 and 2.12 respectively, with the corresponding standard deviations equal to 0.43 and 0.38. Assume the GPAs of all male and female students who participated in sports activities have a normal distribution with unequal standard deviations. Construct a 90% confidence interval for the difference between the two population means.

Solution:

Given $S_1 = 0.43, S_2 = 0.38, n_1 = 28, n_2 = 24$

The desired confidence interval is given by

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

where the degrees of freedom

$$\nu = \frac{\left[\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left[\frac{(0.43)^2}{28}+\frac{(0.38)^2}{24}\right]^2}{\frac{\left(\frac{0.43^2}{28}\right)^2}{28-1}+\frac{\left(\frac{0.38^2}{24}\right)^2}{24-1}} = 49 \text{ (rounded)}$$

Area in each tail of the t-distribution curve = $0.5 - \frac{0.90}{2}$ = 0.05. The t-value for df = 49 and 0.05 area in the right tail is 1.677 (from the table in Appendix-G).

The 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$(2.0 - 2.12) \pm 1.677\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

or $$-0.12 \pm 1.677\sqrt{\frac{(0.43)^2}{28} + \frac{(0.38)^2}{24}}$$

$$-0.12 \pm 1.677(0.11234) = -0.12 \pm 0.19 = -0.31 \text{ to } 0.07.$$

Example E5.19

Assuming that two populations are normally distributed with unequal and unknown population standard deviations, construct a 99% confidence interval for $\mu_1 - \mu_2$ for the following data:

$$n_1 = 15,\ \bar{X}_1 = 50,\ S_1 = 3.0,\ n_2 = 19,\ \bar{X}_2 = 40,\ S_2 = 6.0$$

Solution:

From the table in Appendix-G, $t_{0.005,22}$ is 2.756. The desired confidence interval is given by

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where $$\nu = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Substituting the values for S_1, S_2, n_1 and n_2 and simplifying, we get n = 29. Thus, the desired confidence interval is given by:

$$50 - 40 - 2.756\sqrt{\frac{3^2}{15} + \frac{6^2}{19}} \le \mu_1 - \mu_2 \le 50 - 40 + 2.756\sqrt{\frac{3^2}{15} + \frac{6^2}{19}}$$

$$10 - 2.756\,(1.579) \le \mu_1 - \mu_2 \le 10 + 2.756\,(1.579)$$

$$10 - 4.353 \le \mu_1 - \mu_2 \le 10 + 4.353$$

$$5.647 \le \mu_1 - \mu_2 \le 14.353$$

The complete statement of the confidence level with the associated probability is

$$P(5.647 \le \mu \le 14.353) = 99$$

5.9 CONFIDENCE INTERVAL ON $\mu_1 - \mu_2$ FOR PAIRED OBSERVATIONS

In general, suppose that the data consists of n pairs (X_{11}, X_{21}), (X_{12}, X_{22}), …, (X_{1n}, X_{2n}). Both X_1 and X_2 are assumed to be normally distributed with mean μ_1 and μ_2, respectively. The random variables within different pairs are independent. Since, there are two measurements on the same experimental unit, the two measurements within the same pair may not be independent. Consider the n differences $D_1 = X_{11} - X_{21}$, $D_2 = X_{12} - X_{22}$, …, $D_n = X_{1n} - X_{2n}$. Now the mean of the differences D, say μ_D, is

$$\mu_D = E(D) = D(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

because the expected value of $X_1 - X_2$ is the difference in expected values regardless of whether X_1 and X_2 are independent. Hence, we can construct a confidence interval on $\mu_1 - \mu_2$ by just finding a confidence interval on μ_D. To construct the confidence interval for $\mu_D = \mu_1 - \mu_2$, note that

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}}$$

follows a t-distribution with $n - 1$ degrees of freedom. Then, $P(-t_{\alpha/2,\, n-1} \le T \le t_{\alpha/2,\, n-1}) = 1 - \alpha$, we can substitute for T in the above expression and perform the necessary steps to isolate $\mu_D = \mu_1 - \mu_2$ between the inequalities. This leads to the following 100 $(1 - \alpha)$% confidence interval on $\mu_1 - \mu_2$. The differences D_i are normally and independently distributed, we can apply the t-distribution procedure described earlier to find the confidence interval on μ_D. By analogy with Eq. (5.10), the 100 $(1 - \alpha)$% confidence interval on $\mu_D = \mu_1 - \mu_2$ is

$$\bar{D} - t_{\alpha/2,n-1} s_D / \sqrt{n} \le \mu_D \le \bar{D} + t_{\alpha/2,n-1} s_D / \sqrt{n} \tag{5.32}$$

where $\bar{D}$ and s_D are the sample mean and sample standard deviation of the differences D_i, respectively. This confidence interval is valid for the case where $\sigma_1^2 \ne \sigma_2^2$, because s_D^2 estimates $\sigma_D^2 = V(X_1 - X_2)$ and for large samples ($n \ge 30$ pairs) and the assumption of normality is not necessary.

Summarising, if $\bar{d}$ and s_d are the sample mean and standard deviation of the difference of n random pairs of normally distributed measurements, a $100(1 - \alpha)$% confidence interval on the difference in means $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2,n-1} s_d / \sqrt{n} \le \mu_D \le \bar{d} + t_{\alpha/2,n-1} s_d / \sqrt{n}$$

where $t_{\alpha/2,\, n-1}$ is the upper $\alpha/2$% of point of the t-distribution with $n - 1$ degrees of freedom.

This confidence interval is also valid for the case where $\sigma_1^2 \ne \sigma_2^2$, since $s_D^2 = V(X_1 - X_2)$. Also, for large samples (say, $n \ge 30$), the explicit assumption of normality is unnecessary because of the central limit theorem.

The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.7.

Table 5.7: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:
1. Simple random paired samples
2. Normal differences or large sample

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n - 1$,

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal differences and is approximately correct for large samples and normal differences.

Example E5.20

A company wanted to find out if attending a special course on "how to be a successful salesperson" can increase the average sales of its employees. The company sent 10 of its salespersons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended this course.

Before	11	18	25	10	14	16	17	18	19	21
After	17	25	26	15	20	21	22	23	26	24

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the sales before and after attending the special course. Assume that the population of paired differences is approximately normally distributed.

SOLUTION:

$n = 10$

	Before	After	Difference, d	d^2
	11	17	–6	36
	18	25	–7	49
	25	26	–1	1
	10	15	–5	25
	14	20	–6	36
	16	21	–5	25
	17	22	–5	25
	18	23	–5	25
	19	26	–7	49
	21	24	–3	9
Σ	169	219	–50	280

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-50}{10} = -5$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{280 - \frac{(-50)^2}{10}}{10-1}} = 1.8257$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{1.8257}{\sqrt{10}} = 0.5774$$

$$df = n - 1 = 10 - 1 = 9;\ \alpha/2 = 0.025$$

From the table in Appendix-G, $t_{0.025,9} = 2.262$.

The 95% confidence interval for μ_d is

$$\bar{d} \pm t\, s_{\bar{d}} = -5 \pm 2.262(0.5774)$$

or $$-6.3060 \text{ to } -3.6940.$$

Example E5.21

A college claims that the Math tutoring service offers significantly increases the test scores of students in mathematics. The following table gives the scores per 120 of 12 students before and after they took the tutoring help.

Before	81	75	88	90	67	71	92	69	73	74	75	79
After	96	74	93	79	79	73	111	76	75	78	79	84

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the score before attending the tutoring service minus the score after attending the tutoring service. Assume that the population of paired differences is approximately normally distributed.

SOLUTION:

$n = 12$

Before	After	Difference, d	d^2
81	96	–15	225
75	74	1	1
88	93	–5	25
90	79	11	121
67	79	–12	144
71	73	–2	4
92	111	–19	361
69	76	–7	49
73	75	–2	4
74	78	–4	16
75	79	–4	16
79	84	–5	25
934	997	–63	991

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-63}{12} = -5.25$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{991 - \frac{(-63)^2}{12}}{12-1}} = 7.7474$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{7.7474}{\sqrt{12}} = 2.2365$$

$$df = n - 1 = 12 - 1 = 11; \ \alpha/2 = 0.05 - \frac{0.95}{2} = 0.025$$

From the table in Appendix-G, $t_{11,0.025} = 2.201$

The 95% confidence interval for μ_d is

$$\pm t = -5.25 + (2.201)(2.2365)$$

or $$-10.1725 \text{ to } -0.3275$$

5.10 CONFIDENCE INTERVAL ON THE RATIO OF VARIANCE OF TWO NORMAL DISTRIBUTIONS

Let X_1 and X_2 are independent normal random variables with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2, respectively. We wish to find a 100 (1 – α)% confidence interval on the ratio of σ_1^2 / σ_2^2. Let two random samples of sizes n_1 and n_2 are taken on X_1 and X_2, and let S_1^2 and S_2^2 denote the sample variances. The sampling distribution of

$$F = \frac{S_2^2 / \sigma_2^2}{S_1^2 / \sigma_1^2}$$

is F with $n_2 - 1$ and $n_1 - 1$ degrees of freedom. This distribution is shown in Fig. 5.4.

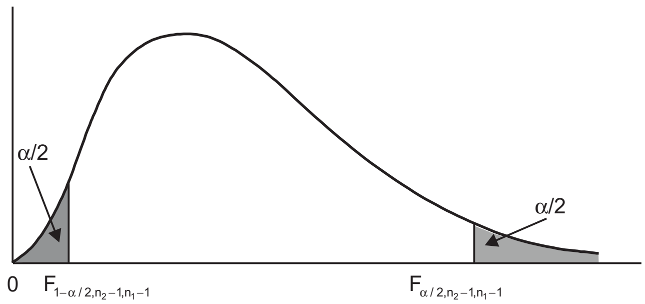

Fig. 5.4: The distribution of F_{n_2-1, n_1-1}

From Fig. 5.4, we observe that

$$P\left\{F_{1-\alpha/2, n_2-1, n_1-1} \le F \le F_{\alpha/2, n_2-1, n_1-1}\right\} = 1 - \alpha$$

or $$P\left\{F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \le F_{\alpha/2,n_2-1,n_1-1}\right\} = 1-\alpha$$

Hence $$P\left\{\frac{S_1^2}{S_2^2}F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1}\right\} = 1-\alpha \tag{5.33}$$

Comparing Eqs. (5.33) and (5.4), we see that a 100 (1 – α)% two-sided confidence interval on σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2}F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1} \tag{5.34}$$

where the lower 1 – α/2 tail-point of the F_{n_2-1,n_1-1} distribution is given by

$$F_{1-\alpha/2,n_2-1,n_1-1} = \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} \tag{5.35}$$

We can also construct one-sided confidence intervals. A 100 (1 – α)% lower-confidence limit on σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2}F_{1-\alpha,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \tag{5.36}$$

while a 100(1 – α)% upper-confidence interval on σ_1^2/σ_2^2 is

$$\frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha,n_2-1,n_1-1} \tag{5.37}$$

where $F_{\alpha/2,n_2-1,n_1-1}$ and $F_{1-\alpha/2,n_2-1,n_1-1}$ are the upper and lower α/2 percentage points of the *F*-distribution with $n_2 - 1$ numerator and $n_1 - 1$ denominator degrees of freedom respectively. A confidence interval on the ratio of the standard deviations can be obtained by taking the square roots in Eq. (5.34).

Example E5.22

Two samples are drawn from two independent populations, which are normally distributed. A comparison of the samples is to be made by developing a 99% confidence interval on the ratio of their variances. The following information is known: $n_1 = n_2 = 21$, $S_1^2 = 5$ and $S_2^2 = 8$. Determine the complete statement of the confidence interval, with the associated probability.

SOLUTION:

A 100 (1 – α)% two-sided confidence interval on σ_1^2/σ_2^2 is given by

$$\frac{S_1^2}{S_2^2}F_{1-\alpha/2,\nu_2,\nu_1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,\nu_2,\nu_1}$$

where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$

Therefore, $F_{0.025,20,20} = 2.46$ and $F_{0.975,19,19} = 1/2.46 = 0.4065$ (from the table in Appendix-*H*).

Confidence interval is given by

$$\frac{5}{8}(0.4065) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{5}{8}(2.46)$$

or
$$0.254 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.538$$

The complete statement of the confidence interval with the associated probability is

$$P\left(0.254 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.538\right).$$

Example E5.23

The diameter of steel rods manufactured on two different machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected and the sample means and sample variations are: $\bar{X}_1 = 9$, $S_1^2 = 0.3$, $\bar{X}_2 = 8.95$ and $S_2^2 = 0.40$, respectively. Assume that the data is drawn from a normal distribution. Construct the following:

(a) a 90% two-sided confidence interval on σ_1/σ_2

(b) a 95% two-sided confidence interval on σ_1/σ_2. Comment on the comparison of the width of this interval with width of the interval in part *(a)*

(c) a 90% lower-confidence bound on σ_1/σ_2.

Solution:

(a) $n_1 = 15, n_2 = 17$

$f_{1-\alpha/2,\, n_1-1,\, n_2-1} = 0.412$

$f_{\alpha/2,\, n_1-1,\, n_2-1} = 2.33$

90% confidence interval for the ratio of variances:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2, n_1-1, n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{S_1^2}{S_2^2}\right) f_{\alpha/2, n_1-1, n_2-1}$$

$f_{0.05,14,17} = 2.33$ and $f_{0.95,14,16} = 0.412$ (from the table in Appendix-*H*)

$$\left(\frac{0.3}{0.4}\right)0.412 \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{0.3}{0.4}\right)2.33$$

$$0.309 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.7475$$

$$0.55588 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.32193$$

(b) 95% confidence interval:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2, n_1-1, n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{S_1^2}{S_2^2}\right) f_{\alpha/2, n_1-1, n_2-1}$$

$f_{0.025,14,16} = 2.82$ and $f_{0.975,14,16} = 0.342$ (from the table in Appendix-H)

$$\left(\frac{0.3}{0.4}\right) 0.342 \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{0.3}{0.4}\right) 2.82$$

$$0.2565 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.115$$

$$0.50646 \le \frac{\sigma_1}{\sigma_2} \le 1.454304$$

The 95% confidence interval is wider than the 90% confidence interval.

(c) 90% lower-sided confidence interval:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2, n_1-1, n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2}$$

$f_{0.1,14,16} = 0.5$ (from the table in Appendix-H)

$$\left(\frac{0.3}{0.4}\right) 0.5 \le \frac{\sigma_1^2}{\sigma_2^2}$$

$$0.375 \le \frac{\sigma_1^2}{\sigma_2^2}$$

or $$0.61237 \le \frac{\sigma_1}{\sigma_2}.$$

5.11 CONFIDENCE INTERVAL ON THE DIFFERENCE IN TWO PROPORTIONS

Here, we construct a confidence interval for $p_1 - p_2$ under the following assumptions:

1. The population proportions $\hat{p}_1$ and $\hat{p}_2$ are not too close to 0 or 1.
2. Two random samples are taken, one from each population, and the two samples are independent.
3. Sample sizes n_1 and n_2 are large.

If two independent samples of size n_1 and n_2 are taken from infinite populations so that X_1 and X_2 are independent, binomial random variables with parameters (n_1, p_1) and (n_2, p_2), respectively, where X_1 represents the number of sample observations from the first population that belongs to a class of interest and X_2 represents the number of sample observations from the second population that belongs to a class of interest, then $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$ are independent estimator of p_1 and p_2 respectively. Considering the fact that the normal approximation to the binomial applies, the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is distributed approximately as standard normal. An approximate 100 (1 – α)% two-sided confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad (5.38)$$

An approximate 100 (1 – α)% lower-confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2$$

and an approximate 100 (1 – α)% upper-confidence interval for $p_1 - p_2$ is

$$p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution. The procedure to find a confidence interval for the difference between two population proportions, p_1 and p_2 is given in Table 5.8. The margin of error is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Table 5.8: Procedure to find a confidence interval for the difference between two population proportions p_1 and p_2

Assumptions:
1. Simple random samples
2. Independent samples
3. x_1, $n_1 - X_1$, x_2 and $n_2 - X_2$ are all 5 or greater

Step 1: For a confidence level of 1 – α, use the table in Appendix-*E* to find $z_{\alpha/2}$

Step 2: The end points of the confidence interval for $p_1 - p_2$ are

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution.

Step 3: Interpret the confidence interval.

Summary of confidence intervals for single population and two populations are given in Tables 5.9 and 5.10 respectively.

Example E5.24

An investigation on body mass index (bmi – a measure of body fat based on height and weight) for adults, a bmi of greater than 25 indicates an above health weight (i.e., overweight or obese). Of 800 randomly selected adults whose highest degree is bachelors, 412 have an above healthy weight; and of 1000 randomly selected adults with a graduate degree, 474 have an above healthy weight.

(*a*) determine a 90% confidence interval for the difference between the percentages of adults in the two degree categories who have an above healthy weight.

(*b*) repeat part (*a*) for an 80% confidence interval.

Solution:

(*a*) Population 1: Bachelors degree, $\hat{p}_1 = \frac{412}{800} = 0.515$

Population 2: Graduate degree, $\hat{p}_2 = \frac{474}{1000} = 0.474$

$z_{\alpha/2} = 1.645$ (from the table in Appendix-*E*)

90% confidence interval is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.515 - 0.474) \pm 1.645\sqrt{\frac{0.515(1-0.515)}{800} + \frac{0.474(1-0.474)}{1000}}$$

$$(0.515 - 0.474) \pm 1.645\sqrt{9.752813\times10^{-4} + 2.49324\times10^{-4}}$$

0.041 ± 0.057566 or -0.016566 to 0.098566

(*b*) The 80% confidence interval is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\ ;\ z_{\alpha/2} = 1.28 \text{ (from Appendix-}E\text{)}$$

$$(0.515 - 0.474) \pm 1.28\sqrt{\frac{0.515(1-0.515)}{800} + \frac{0.474(1-0.474)}{1000}}$$

0.041 ± 0.04793 or 0.0037929 to 0.08893.

Table 5.9: Summary of confidential intervals (single population)

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1-\alpha$
1.	Quantitative data; standard deviation σ is known; Population normal	μ, the population mean	Draw a sample of size n and compute the value of $\bar{X}$, the estimate of μ	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

2.	Quantitative data; standard deviation σ is not known; population normal; important when sample size is small ($n < 30$)	μ, the population mean	Draw a sample of size n and compute $\bar{X}$ and $S = \sqrt{\frac{1}{n-1}\Sigma(X_i - \bar{X})^2}$	$\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$ $t_{n-1,\alpha/2}$ is the value obtained from the t-distribution with $n - 1$ degrees of freedom
3.	Quantitative data; standard deviation σ is not known; population not necessarily normal; sample size is large ($n \geq 30$)	μ, the population mean	Draw a sample of size n and compute $\bar{X}$ and S	$\bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$ Confidence interval is approximate
4.	Quantitative data; binomial case	p, the probability of success (the population proportion)	Draw a sample of size n and note X, the number of successes; obtain X/n, the estimate of p. Here n is assumed large	$\frac{X}{n} \pm Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{X}{n}\right)}{n}}$ Confidence interval is based on the central limit theorem, thus, is approximate
5.	Quantitative data; population normal	σ^2, the population variance	Draw a sample of size n and compute s^2	$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$

Table 6.10: Summary of confidence intervals (two populations)

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1 - \alpha$
1.	Quantitative data; variances σ_1^2 and σ_2^2 are known; both populations are normally distributed	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes m and n from the two populations, get the respective means $\bar{X}_1$ and $\bar{X}_2$ and find the value of $\bar{X}_1 - \bar{X}_2$, an estimate of $\mu_1 - \mu_2$	$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
2.	Quantitative data; σ_1^2, σ_2^2 are not known but assumed equal; both populations are normally distributed	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes n_1 and n_2 from the two populations, compute $\bar{X}_1 - \bar{X}_2$ and $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{m+n-2}}$	$(\bar{X}_1 - \bar{X}_2) \pm t_{m+n-2,\alpha/2}S_p\sqrt{\frac{1}{m} + \frac{1}{n}}$

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1-\alpha$
3.	Quantitative data; σ_1^2 and σ_2^2 are not known and not assumed equal; populations may not be normal; sample sizes m and n are large	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes n_1 and n_2 compute $\bar{X}_1 - \bar{X}_2$; compute S_1^2 and S_2^2	$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ Confidence interval is approximate
4.	Quantitative data μ_1 and μ_2 are the means of populations 1 and 2 respectively. Standard deviation of differences are unknown	$\mu_D = \mu_1 - \mu_2$ Difference in means of two normal distributions in a paired analysis	Paired analysis calculates differences and uses one sample method for inference on the mean difference	$\bar{d} - t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}} \le \mu_D \le \bar{d} + t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$
5.	Quantitative data; σ_1^2 and σ_2^2 are the variances of population 1 and 2; σ_1^2 and σ_2^2 known	σ_1^2/σ_2^2 Ratio of variances of two normal distribution	Draw 2 samples, one from each population, compute S_1^2 and S_2^2	$\frac{S_1^2}{S_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f_{\alpha/2,n_2-1,n_1-1}$
6.	Quantitative data; p_1 is the proportion in population 1 and p_2 in population 2; sample sizes n_1 and n_2 are assumed large	$p_1 - p_2$, the difference of the population proportions	Draw two samples, one from each population, of sizes n_1 and n_2; find $\frac{\chi_1}{n_1}$ and $\frac{\chi_2}{n_2}$, the estimates of p_1 and p_2 respectively; then get $\frac{\chi_1}{n_1} - \frac{\chi_2}{n_2}$	$\frac{\chi_1}{n_1} - \frac{\chi_2}{n_2} \pm Z_{\alpha/2}\sqrt{\frac{\frac{\chi_1}{n_1}\left(1-\frac{\chi_1}{n_1}\right)}{n_1} + \frac{\frac{\chi_2}{n_2}\left(1-\frac{\chi_2}{n_2}\right)}{n_2}}$ Confidence interval is based on the central limit theorem, thus is approximate

5.12 SAMPLE SIZE SELECTION

5.12.1 Sample Size Selection for Estimating Population Mean

Each level is constructed with regard to a given *confidence level* and is called a *confidence interval*. The confidence level associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by 100 (1 – α)%. When expressed as probability, it is called the *confidence coefficient* and is denoted by (1 – α). α is called the *significance level*.

The confidence interval for μ for large samples is given as follows:

The 100 (1 – α)% confidence interval for μ is

$$\bar{x} \pm z\sigma_x \qquad \text{if } \sigma \text{ is known}$$

$$\bar{x} \pm zs_x \qquad \text{if } \sigma \text{ is unknown} \tag{5.39}$$

where $\sigma_x = \sigma\sqrt{n}$ and $s_x = s/\sqrt{n}$

The value of z used here is obtained from standard normal distribution table for the given confidence level. The quantity $z\sigma_x$ (or zs_x when σ is not known) in the confidence interval formula is called the *maximum error of estimate* and is denoted by E.

Thus, $E = zs_x$ or $E = z\sigma_x$

The standard deviation of the sample mean is $\sigma/\sqrt{n}$

$$E = z\frac{\sigma}{\sqrt{n}} \tag{5.40}$$

Given the confidence level and the standard deviation of the population, the sample size that will produce a predetermined error E of the confidence interval estimate of μ is

$$n = \frac{z^2\sigma^2}{E^2} \tag{5.41}$$

where n = required sample size.

We note that

1. As the desired length of the interval $2E$ decreases, the required sample size n increases for a fixed value of σ and specified confidence.
2. As σ increases, the required sample size n increases for a fixed desired length $2E$ and specified confidence.
3. As the level of confidence increases, the required sample size n increases for fixed desired length $2E$ and standard deviation σ.

Example E5.25

At a 99% confidence level, the standard deviation σ for the population is 0.7. How large a sample should be selected if we want the estimate to be within 0.01 of the population mean?

Solution:

The mean size is $\bar{x} \pm 0.01$

Therefore, the maximum size of the error of the estimate is to be 0.01, that is $E = 0.01$. The value of z for a 99% confidence level is 2.58. The value of σ is given to be 0.7. Hence, substituting the values and simplifying Eq. (5.41), we get

$$n = \frac{z^2\sigma^2}{E^2} = \frac{(2.58)^2(0.7)^2}{(0.01)^2} = 32,616$$

Hence, the required sample size is 32, 616.

5.12.2 Sample Size for the Estimation of Proportion

The maximum error, E, of the interval estimation of the population proportion is

$$E = z\sigma_{\hat{p}} = z\sqrt{\frac{pq}{n}}$$

The above expression can be manipulated algebraically to write it in terms of E, p, q and z and the final result is given here. Given the confidence level and the values of p and q, the sample size that will produce a predetermined maximum error E of the confidence interval estimate p is given by

$$n = \frac{z^2 pq}{E^2}$$

Another approach to choosing n uses the fact that the sample size from the above equation will always be a maximum for $p = 0.5$ ($q = 0.5$) or $p(1 - p) \le 0.25$ with equality for $p = 0.5$. This can be used to obtain an upper bound on n. We are at least 100 $(1 - \alpha)$% confident that the error in estimating p by $\hat{p}$ is less than E if the sample size is

$$n = \frac{Z^2}{E^2}(0.25)$$

Example E5.26

(a) How large a sample should be selected so that the maximum error of estimated for a 99% confidence interval for p is 0.04 when the value of the sample proportion obtained from a preliminary sample is 0.2?

(b) Find the most conservative sample size that will produce the maximum error for a 99% confidence interval for p equal to 0.04.

Solution:

(a) Here $E = 0.04$, $\hat{p} = 0.2$, $\hat{q} = 1 - 0.2 = 0.8$

$z = 2.58$ for 99% confidence level (from the table in Appendix-E)

Hence $n = \dfrac{z^2 \hat{p}\hat{q}}{E^2} = \dfrac{2.58^2(0.2)(0.8)}{(0.04)^2} = 66.564 \approx 67$

(*b*) $E = 0.04$, $z = 2.58$, and $p = q = 0.50$ for most conservative sample size.

Thus, $n = \dfrac{z^2 pq}{E^2} = \dfrac{2.58^2(0.5)(0.5)}{(0.04)^2} = 1040.062 \approx 1040$

5.13 SUMMARY

In this chapter, we introduced that part of statistics called *inferential statistics*. In Chapter 1, inferential statistics was described as the part of statistics that help us to make decisions about some characteristics of a population based on sample information. Inferential statistics makes use of the sample results to make decisions and draw conclusions about the population from which the sample is drawn. *Estimation* is the first topic to be covered in our presentation of inferential statistics. *Estimation* and *hypothesis testing* (presented in Chapter 6) topics taken together are usually referred to as *making inferences*. This chapter examined how to estimate the population mean, population proportion for a single population. We have extended our discussion of estimation to the difference between two population means and the difference between two population proportions.

PROBLEMS

P5.1 A manufacturing company tested 25 printed circuit boards and the findings are as follows: Mean length of life for 25 boards = 3,566 hours. Standard deviation of life length in sample = 150 hours. Construct a 90% confidence interval for the mean length of life for the new printed circuit boards.

P5.2 The mean tensile strength of a sample of 25 high quality steel specimens equals to 50,000 MPa. Find a 95% confidence interval on the mean tensile strength if the standard deviation is known to be 500 MPa.

P5.3 (*a*) The standard deviation of a normally distribution manufacturing process is 3g. Determine the sample size for a 90% confidence interval so that the estimation of the mean process is within 1g of the true but unknown mean yield.

(*b*) An electrical light bulb manufacturing company likes to estimate the light bulb's mean life. Assume a normal distribution of the life of a light bulb. Assuming the standard deviation is 30 hours, find the how many bulbs should be tested so as to be: *(i)* 95% confident that the estimate $\bar{x}$ will not differ from the true mean life by more than 10 hours. *(ii)* 98% confident to accomplish *(i)*.

P5.4 A college administrator wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 21 years. From past studies, the standard deviation is known to be 1.5 years. Construct a 90% confidence interval of the population mean age.

P5.5 In order to estimate the amount of time (in minutes) that a teller spends on a customer, a bank manager decided to observe 64 customers picked at random. The amount of time the teller spent on each customer was recorded. It was found that a sample mean was 4 minutes will standard deviation 1.2 minutes. Find a 98% confidence interval for the mean amount of time μ.

P5.6 When 25 cigarettes of a particular brand were tested in a laboratory for the amount of tar content, it was found that their mean content was 20 milligrams with $S = 2$ milligrams. Set a 90% confidence interval for the mean tar content μ in the population of cigarettes of the brand. Assume that the amount of tar in a cigarette is normally distributed.

P5.7 The breaking strength of a machine part is required to be at least 75 MPa. Based on previous test data, the standard deviation of breaking strength is 6 MPa. A random sample of 16 specimens is tested and the average value is found to be 73 MPa. Construct a 95% confidence interval on the mean breaking strength.

P5.8 The diameters of copper shafts produced by a certain manufacturing process should have a mean diameter of 0.51 mm. The diameter is known to have a standard deviation of 0.0002 mm. A random sample 40 shafts has an average diameter of 0.509 mm. Construct a 95% confidence interval on the mean shaft diameter.

P5.9 A normally distributed random variable has a known variance of $\sigma^2 = 36$ and unknown mean. Construct a 95% confidence interval on the mean that has a total width of 2.0.

P5.10 In an experimental lab, 10 students are selected at random from a class and were assigned to set up an experiment. Their time to complete the experimental set up are recorded (in minutes) as 51, 55, 54, 50, 56, 59, 57, 59, 58 and 60. Use a 95% confidence interval to estimate μ, the mean time for the student to set up the experiment. Assume that the population is normally distributed.

P5.11 A study claims that students in a particular college spend an average of 18 hours per week on leisure activities. The administration wanted to test this claim. A sample of 10 students selected at random gave their responses as follows (in hours):

14, 16, 19, 22, 23, 25, 26, 33, 38 and 41

Assume that the time spent on leisure activities by all the students is normally distributed. Use a 95% confidence interval activities.

P5.12 The time required to change flat tires in a garage is a normal distributed random variable. The tire changing times for 15 automobiles selected at random are as follows:

5, 9, 7, 6, 24, 11, 4, 13, 10, 9, 20, 8, 19, 17 and 25

Use a 95% confidence interval to estimate μ, the mean time required to change flat tires in that garage.

P5.13 A sample of 26 observations selected from a normally distributed population produced a sample variance of 35. Construct a confidence interval for σ^2 for each of the following confidence levels.

(a) $1 - \alpha = 0.99$

(b) $1 - \alpha = 0.95$

(c) $1 - \alpha = 0.90$

P5.14 Refer to Problem 5.31. For the two steel specimen types in Prob. 5.31, determine a two-sided 95% confidence interval on the true variance.

P5.15 In a random sample of 30 machine turned parts, the sample standard deviation of the diameter of the machined parts was found to be 10 mm. Compute a 90% upper confidence interval on the variance of the diameter of the turned parts.

P5.16 The following data represent the amount of corrosion coating (in ounces) need to paint 6 plates:

8.2, 8.8, 7.7, 7.9, 8.4 and 7.9

Determine the following:

(a) a point estimate of σ^2, the variance of the amount of corrosion coating needed to paint a plate

(b) confidence intervals for σ^2 and σ with the following confidence levels, assuming that the amount of corrosion coating per plate is normally distributed. *(i)* 95%, *(ii)* 98%.

P5.17 Suppose 30 samples of an allergic medicine was randomly selected and weighed. The sample standard deviation is 1.3 milligrams. Assuming the weights are normally distributed, construct a 99% confidence intervals for the population variance and standard deviation.

P5.18 The following measurements (in kg) were obtained on the weights of 6 manufactured metal parts picked at random: 12.1, 8.1, 7.2, 12.3, 14.05 and 13. Determine the confidence intervals for σ^2 and σ with 95% confidence level.

P5.19 The manufacturer of a new line of automotive transmission is estimating the fraction of non-conforming units produced. Find a 95% two-sided confidence interval for the following:

(a) 150 transmissions in the sample with 15 non-conforming

(b) 1500 transmissions in the sample with 150 non-conforming

(c) 15,000 transmissions in the sample with 1500 non-conforming

P5.20 With a random sample of 1000 people, 400 are women who turn up to see a baseball game. Find a 95% confidence interval for the percentage of women at the game.

P5.21 A sample of 600 observations selected from a population produced a sample proportion equal to 0.7.

(a) make a 90% confidence interval for p.

(b) construct a 95% confidence interval for p.

(c) make a 99% confidence interval for p.

(d) does the width of the confidence intervals constructed in part *(a)* through *(c)* increase as the confidence level increases?

P5.22 In a survey, 90% of drivers rated their driving as excellent or good. Suppose that this percentage was based on a random sample of 400 drivers,

(a) what is the point estimate of this corresponding population proportion?

(b) find a 95% confidence interval for the corresponding population proportion.

P5.23 A mail order company guarantees its customers that the products ordered will be mailed within 72 hours after an order is receives. The quality control department took a sample of 50 orders and found that 40 of them were mailed within 72 hours of the placement of the orders.

(a) construct a 98% confidence interval for the percentage of all orders that are mailed within 72 hours of their placement

(b) if the confidence interval found in part *(a)* is too wide, suggest a way to reduce the width of the interval.

P5.24 In a poll of 2000 adults, 60% of adults said that public education needs to be improved. Construct a 95% confidence interval for the proportion of all adults who hold this opinion.

P5.25 If a random sample of 60 non-smokers have a mean life of 75 years with a standard deviation of 7 years, and a random sample of 50 smokers live 60 years with a standard deviation of 8 years, find a 95% confidence interval for the difference of mean life time of non-smokers and smokers.

P5.26 The Chairman of the Mechanical Engineering Department is concerned about grade inflation in that department. In a random sample of 100 students in Mechanical Engineering courses this year, 80 received *A* or *B*; in a sample of 120 grades in Mechanical Engineering 5 years ago, 65 were *A* or *B*. Find a 90% confidence interval for the difference in the percentage of *A* and *B* grades now and 5 years ago.

P5.27 Machine systems for a manufacturing company are supplied by two suppliers. Machines from supplier 1 have a standard deviation of 25 units and supplier 2 have a standard deviation of 30 units. A random sample of 10 machines from supplier 1 have an average of 155 units and a similar sample of supplier 2 machines have an average of 137.4 units. Construct a 90% confidence interval on the true mean difference in the means of two machines.

P5.28 At the end of crash dieting program administered to 49 men and 36 women, the following information was obtained about the loss of weight (in kg).

	Men	Women
Mean loss of weight	$\bar{X}_1 = 16$	$\bar{X}_2 = 12$
S, sample standard deviation	$S_1 = 5$	$S_2 = 4$

Find an approximate 85% confidence interval for the difference in the mean losses.

P5.29 A consulting agency was asked by a large health insurance company to investigate if business majors were better sales persons. A sample of 40 sales persons with a business degree showed that they sold an average of 10 insurance policies per week with a standard deviations of 1.80 policies. Another sample of 45 sales persons with a degree other than business showed that they sold an average of 8.5 insurance policies per week with a standard deviation of 1.35 policies. Construct a 99% confidence interval for the difference between the two population means.

P5.30 An administrator wants to test if the mean GPAs (grade point averages) of all male and female college students who actively participate in sports are different. A random sample of 28 male and 24 female students who actively participate in sports was selected. The mean GPAs of the two groups are found to be 2.62 and 2.74 out of 4.0 respectively, with the corresponding standard deviations equal to 0.43 and 0.38. Assume that the GPAs of all male and female students have a normal distribution with the same standard deviation. Construct a 90% confidence interval for the difference between the two population means.

P5.31 Two slightly different steel specimen types of bar samples are being tested for tensile strength. Ten samples of each specimen type were selected one at random from ten different lots. The tensile tests on these bars resulted in

$$\bar{X}_1 = 25 \text{ MPa},\ s_1^2 = 2.2 \text{ (MPa)}^2,$$

$$\bar{X}_2 = 25.5 \text{ MPa and } s_2^2 = 2.7 \text{ (MPa)}^2.$$

Assume the tensile strengths are normally distributed.

(*a*) find a 99% confidence interval on the difference in means

(*b*) if only 5 samples were selected for the second specimen type, how does this affect the confidence interval? Assume the value $\bar{X}_2$ of s_2^2 and remain unchanged.

P5.32 The outputs of two manufacturing firms over one year period with random days picked gave the following data:

$$n_1 = 20 \qquad \bar{X}_1 = 250 \qquad s_1^2 = 35$$

$$n_2 = 25 \qquad \bar{X}_2 = 200 \qquad s_2^2 = 40$$

Assuming the two outputs are normally distributed with the same variance, find a 95% two-sided confidence interval on the difference in their means.

P5.33 Random sample of weights were taken from two packages in a packaging process. The first sample consisted of n_1 = 150 and the second consisted of n_2 = 250 packages. The two sample means and standard deviations are as follows:

$$\bar{X}_1 = 10 \text{ kg} \qquad \bar{X}_2 = 9.8 \text{ kg}$$

$$s_1 = 0.25 \text{ kg} \qquad s_2 = 0.35 \text{ kg}$$

Determine the interval within which we can be 90% sure that $\mu_1 - \mu_2$ lies.

P5.34 Two diets, diet *A* and diet *B*, were fed to two groups of randomly picked cats of a certain bread. Diet *A* was fed to 10 cats. Their mean weight $\bar{X}_1$ was 10 kg, and the sample standard deviation S_1 was 3 kg. Diet *B* was fed to 14 cats. Their mean weight $\bar{X}_2$ was 12 kg and the sample standard deviation S_2 was 4 kg. Find a confidence interval for the difference in means with a 90% confidence level. Assume the weights of the cats given diet *A* and the weights of the cats given diet *B* are normally distributed with the same variance.

P5.35 Ten samples of standard cement had an average weight per cent calcium of $\bar{X}_1$ = 80, with a sample standard deviation of S_1 = 5, while 15 samples of lead-doped cement has an average weight per cent calcium of $\bar{X}_2$ = 77 with a sample standard deviation of S_2 = 4. Assume that the weight per cent calcium is normally distributed. Find a 95% confidence interval on the difference in means, $\mu_1 - \mu_2$, for two types of cement. Further, assume that both normal populations have the same standard deviation.

P5.36 A new voltage-regulating device is installed in an electromechanical system. Before the installation, a random sample yielded the following information about the percentage of breakdowns of the system: $\bar{X}_1$ = 8, S_1^2 = 36 and n_1 = 8. After installation, a random sample yields $\bar{X}_2$ = 7, S_2^2 = 25 and n_2 = 9. Find a confidence interval for the difference of the means with a 90% confidence level.

P5.37 Table P5.37 gives the systolic blood pressure of 7 adults before and after the completion of a special dietary plan based on a special dietary plan for 3 months. Construct a 95% confidence interval for μ_d. Assume that the population of paired differences, μ_d, is (approximately) normally distributed.

Table P5.37

Before	209	179	195	221	232	199	223
After	192	185	186	223	221	182	232

P5.38 A college claims that the Math tutoring service offers significantly increases the test scores of students in mathematics. The following table gives the scores per 120 of 8 students before and after they took the tutoring help.

Before	82	75	89	91	66	70	91	69
After	97	72	94	111	80	72	117	76

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the score before attending the tutoring service minus the score after attending the tutoring service. Assume that the population of paired differences is approximately normally distributed.

P5.39 A medical agency measured the corneal thickness of 8 patients who had glaucoma in one eye but not in the other. The following are the data on corneal thickness in microns.

Patient	Normal	Glaucoma
1	484	488
2	479	479
3	493	481
4	445	427
5	437	441
6	399	409
7	464	459
8	477	460

Assume that the population of paired differences is approximately normally distributed. Make a 95% confidence interval for the mean μ_d of the population difference, where a paired difference is equal to the difference in corneal thickness for normal and Glaucoma in microns.

P5.40 A company wanted to find out if attending a special course on "How to be a successful salesperson" can increase sales of its employees. The company sent 6 of its sales persons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended this course.

Before	11	18	25	10	14	16
After	17	25	26	15	20	21

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the sales before and after attending the special course. Assume that the population of paired differences is approximately normally distributed.

P5.41 A company claims that its 12-week special exercise program significantly reduces weight. A random sample of 6 persons was selected, and these persons were put on this exercise program for 12 weeks. The following table gives the weight (in kg) of those 6 persons before and after the program.

Before	80	85	65	100	103	97
After	79	81	63	97	101	95

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the weight before joining this exercise program minus the weight at the end of the 12-week program.

P5.42 A manufacturing company sent 10 of its employees for a special training course to increase their productivity. The following table gives the one-week production of these employees before and after they attend this course.

Before	12	10	18	23	25	9	14	16	7	8
After	18	13	24	22	24	14	19	20	10	11

Make a 99% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the production before taking this special course minus the production after taking the course. Assume the population of paired differences having a normal distribution.

P5.43 The soldering machine has a supposed life of 6 years. Two machines are under consideration for acquisition by the manufacturer. Ten samples of machine A were subjected to simulated operating conditions, and the sample variance of machine life was computed as $S_A^2 = 4$ years^2. For machine B, $S_B^2 = 8$ years^2 for $n_B = 12$. Determine a 95% two-sided confidence interval on the ratio of σ_A^2/σ_B^2 assuming that the machine lives are normally distributed.

P5.44 The length of steel bars manufactured on two different machines is being investigated. Two random samples of sizes $n_1 = 16$ and $n_2 = 18$ are selected and the sample means and sample variances are $\bar{X}_1 = 10$, $S_1^2 = 0.4$, $\bar{X}_2 = 9.6$ and $S_2^2 = 0.45$, respectively. Assuming σ_1^2/σ_2^2, construct a 95% two-sided confidence interval on ratio of the population variance σ_1^2/σ_2^2. Is it reasonable to conclude that the two variances are equal?

P5.45 A study has been made to compare the aluminium contents of two brands of machine parts. Ten parts of brand A had an average aluminium content of 3 grams with a standard deviation of 0.4 grams, while 8 parts of brand B had an average aluminium content of 0.3 grams with a standard deviation of 0.5 grams. Assuming that the two sets of data are independent random samples from normal populations, with equal variance find a 98% confidence interval for σ_1^2/σ_2^2.

P5.46 Two different grinding processes are used to finish certain machine part. Both processes can produce parts at identical mean surface roughness. The company would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $S_1 = 5$ micro millimeters, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $S_2 = 4$ micro millimeters. Find a 90% confidence interval on the ratio of the two standard deviations, σ_1/σ_2.

P5.47 A study has been conducted to study the capability of a gauge by measuring the weights of 2 sheets of paper. The data are shown below:

Paper 1 $S_1 = 0.091159$ $n_1 = 15$

Paper 2 $S_2 = 0.084499$ $n_2 = 15$

Use $\alpha = 0.05$ and find a confidence interval.

P5.48 The diameter of brass rods manufactured on two different machines is being studied. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are

$\bar{X}_1 = 9.75$, $S_1^2 = 0.3$, $\bar{X}_2 = 7.65$ and $S_2^2 = 0.4$ respectively. Assume that the data are drawn from a normal distribution. Construct the following:

(a) a 90% two-sided confidence interval on σ_1/σ_2

(b) a 95% two-sided confidence interval on σ_1/σ_2. Compare the width of this interval with the width of the interval in part (*a*).

(c) a 90% lower confidence bound on σ_1/σ_2

P5.49 A survey was conducted involving 500 adult whites and 300 adult blacks 18 years or older. The following question was asked: "Do you think that police officers treat criminal suspects differently in low-income neighbourhoods than in middle- or high-income neighbourhoods?". The survey resulted in 300 whites responding yes and 250 blacks responding yes.

(a) find a point estimate of $p_1 - p_2$, the difference in the proportions, where p_1 is the proportion among all white adults who would respond yes and p_2 is the proportion among all black adults who would respond yes

(b) obtain an approximate 95% confidence interval for $p_1 - p_2$.

P5.50 A company wanted to estimate the difference between the percentages of users of two electric shavers who will never switch to another brand. In a sample of 500 users of electric shaver *A* taken, 100 said that they will never switch to another brand. In another sample of 400 users of brand *B* taken, 70 said they will never switch to another brand.

(a) find the point estimate of $p_1 - p_2$, where p_1 and p_2 are the proportion of all users of brands *A* and *B* respectively.

(b) construct a 97% confidence interval for the difference between the proportions of all users of the two brands who will never switch.

P5.51 In a random sample of 90 turbine blades, 10 have a surface finish that is rougher than the specifications allow. Suppose that a modification is made in the surface roughness finishing process and that, subsequently, a second random sample of 90 turbine blades is obtained. The number of defective blades in this second sample is 8. Find an approximate 95% confidence interval on the difference in the proportion of defective turbine blades produced under the two processes.

P5.52 In a random sample of 50 people from state *A*, 40 said they favoured death penalty and 24 out of 48 from state *B* were in favour of death penalty. Find a 95% confidence interval for $p_1 - p_2$, where p_1 is the proportion of those in state *A* favouring death penalty and p_2 is the proportion of those in state *B* favouring death penalty.

P5.53 An independent health agency investigated the health of independent random samples of white and African-American elderly (aged 70 or older). Of the 5989 elderly surveyed, 629 had at least one stroke, whereas 203 of 1006 African-American elderly surveyed reported at least one stroke. Find a 95% confidence interval for the difference between the stroke incidences of white and African-American elderly.

P5.54 In a survey of 1000 drivers 25–34 years old, 27% said they buckle up, whereas 350 of 1000 drivers 45–64 years old said that they did. Find a 95% confidence interval for the difference between the proportion of seat-belt users for drivers in the age group 25–34 years and 45–64 years.

P5.55 *(a)* The standard deviation of the lifetimes of a particular brand of light bulbs is 120 hours. How large a sample must be so that we are 95% confident that the error in the estimated mean life of the lights bulbs is less than 8 hours?

(b) If the standard deviation of the lifetimes of the light bulbs is not known, how large the samples must be tested to be 95% confident that the error in the estimated mean life of all the light bulbs is less than 0.1σ?

P5.56 A certain manufacturing firm that manufactures a certain type of electronic part wants to estimate its mean life. Assuming that the standard deviation σ is 50 hours, find how many bulbs should be tested so as to be:

(a) 95% confident that the estimate $\bar{x}$ will not differ from the true mean life μ more than 10 hours

(b) 98% confident to accomplish accuracy in part *(a)*.

Assume a normal distribution of the life of the electronic part.

P5.57 A government transportation agency wants to estimate at a 95% confidence level the mean speed for all automobiles travelling on a specific national highway. From a previous study, the agency knows that the standard deviation of speeds of automobiles travelling on this highway is 6 kilometers per hour, what sample size should the agency choose for the estimate to be within 2 kilometers per hour of the population mean?

P5.58 The school administration wants to determine a 95% confidence interval for the mean number of hours that the school students spend doing homework per week. It was known that the standard deviation for hours spent per week by all school students during homework is 10 hours. How large a sample should the school administration select so that the estimate will be within 1 hour of the population mean?

P5.59 The government would like to estimate the mean family size for all families in a particular state at a 99% confidence level. It is known that the standard deviation σ for all sizes of all families in that state is 0.7. How large a sample should the government select if it wants its estimate to be within 1% of the population mean?

P5.60 *(a)* How large a sample should be selected so that the maximum error of estimate for a 99% confidence interval for p is 0.04 when the value of the sample proportion obtained from a preliminary sample is 0.6?

(b) Find the most conservation sample size that will produce the maximum error for a 98% confidence interval for p equal to 0.04?

P5.61 A hardware store company guarantees all hardware deliveries within 30 minutes of the placement of orders on the plane. An agency wants to estimate the proportion of all hardware delivered within 30 minutes by the company. What is the most conservative estimate of the sample size that would limit the maximum error to within 0.03 of the population proportion for a 99% confidence interval?

P5.62 A transportation safety agency wants to estimate the proportion of all drivers who wear seat belts while driving. Assume that a preliminary study has shown that 85% of drivers wear seat belts while driving. How large should the sample size be so that the 90% confidence interval for the population proportion has a maximum error of 0.02?

P5.63 *(a)* A company wants to estimate the proportion of parts made by a particular machine that are defective. The company wants this estimate to be within 0.01 of the population proportion for a 95% confidence level. What is the most conservative estimate of the sample size that will limit the maximum error to within 0.01 of the population proportion?

(b) Refer to Problem P5.63*(a)*. Suppose a preliminary sample of 300 parts produced by that machine showed that 6% of them are defective. How large a sample should the company select so that the 95% confidence interval for p is within 0.01 of the population proportion?

REVIEW QUESTIONS

1. Explain the meaning of the following terms used in estimation.
   (a) estimator and estimate *(b)* point estimate and interval estimate
2. What is the point estimator of the population mean, μ? Describe how the margin of error for a point estimate of μ is calculated?
3. What are the various alternatives for decreasing the width of a confidence interval?
4. Explain the following concepts:
   (a) How the width of confidence interval decreases with an increase in the sample size?
   (b) How the width of confidence interval decreases with a decrease in the confidence level?
5. Explain the difference between a confidence level and confidence interval.
6. What is the maximum error of estimate for μ for a large sample? How is it obtained?
7. What are the similarities and the differences between the standard normal distribution and the *t*-distribution?
8. Describe the parameters of a normal distribution and the *t*-distribution.
9. Explain the meaning of the degrees of freedom for a *t*-distribution.
10. What assumptions are made in using the *t*-distribution to make a confidence interval for μ?
11. What assumptions are made in using the normal distribution to make a confidence interval for the proportion population, *p*?
12. What is the point estimator of the population proportion, *p*? How is the margin of error for a point estimate of *p* calculated?
13. Describe the following terms:
   (a) estimation *(b)* estimator
   (c) maximum error of estimate *(d)* *t*-distribution

STATE TRUE OR FALSE

1. A point estimate of a parameter is the value of a statistic used to estimate the parameter. (True/False)
2. Confidence interval is an interval of numbers obtained from a point estimate of a parameter. (True/False)
3. Confidence level is the confidence we have that the parameter lies in the confidence interval (i.e., that the confidence interval contains the parameter). (True/False)
4. For a fixed sample size, increasing the confidence level improves the precision, and vice versa. (True/False)
5. A statistical procedure is said to be robust if it is sensitive to departures from the assumptions on which it is based. (True/False)
6. The margin of error equals the standard error of the mean multiplied by $z_{\alpha/2}$. (True/False)
7. The margin of error can be determined if we know only the confidence level. (True/False)
8. The confidence level can be determined if we know only the margin of error. (True/False)

9. The margin of error can be determined if we know only the confidence level, population standard deviation, and sample size. (True/False)
10. The confidence level can be determined if we know only the margin of error, population standard deviation and sample size. (True/False)
11. For samples of size n, the variable $t = (\bar{x} - \mu)/(s/\sqrt{n})$ has the t-distribution with n degrees of freedom. (True/False)
12. When the population standard deviation is unknown, the confidence interval procedure for a population mean cannot be based on the standardised version of $\bar{x}$. (True/False)
13. A confidence-interval estimate of a parameter consists of an interval of numbers obtained from a point estimate of the parameter and a percentage that specifies how confident we are that the parameter lies in the interval. (True/False)
14. Independent samples means that the sample selected from one of the populations has no effect or bearing on the sample selected from the other population. (True/False)
15. The pooled t-procedures require equal population standard deviation. (True/False)
16. The non-pooled t-procedures do not require equal population standard deviations. (True/False)
17. The shape of a normal distribution is determined by its standard deviation. (True/False)
18. By using a paired sample, extraneous sources of variation cannot be removed. (True/False)
19. For very small sample sizes, the normality assumption is essential for both t-procedures (pooled and non-pooled). (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **True** 3. **True** 4. **False** 5. **False** 6. **True** 7. **False** 8. **False** 9. **True** 10. **True**
11. **False** 12. **True** 13. **True** 14. **True** 15. **True** 16. **True** 17. **True** 18. **False** 19. **True**