

Sampling and estimation theories

66.1 Introduction

The concepts of elementary sampling theory and estimation theories introduced in this chapter will provide the basis for a more detailed study of inspection, control and quality control techniques used in industry. Such theories can be quite complicated; in this chapter a full treatment of the theories and the derivation of formulae have been omitted for clarity—basic concepts only have been developed.

66.2 Sampling distributions

In statistics, it is not always possible to take into account all the members of a set and in these circumstances, a **sample**, or many samples, are drawn from a population. Usually when the word sample is used, it means that a **random sample** is taken. If each member of a population has the same chance of being selected, then a sample taken from that population is called random. A sample that is not random is said to be **biased** and this usually occurs when some influence affects the selection.

When it is necessary to make predictions about a population based on random sampling, often many samples of, say, N members are taken, before the predictions are made. If the mean value and standard deviation of each of the samples is calculated, it is found that the results vary from sample to sample, even though the samples are all taken from the same population. In the theories introduced in the following sections, it is important to know whether the differences in the values obtained are due to chance or whether the differences obtained are related in some way. If M samples of N members are drawn at random from a population, the mean values for

the M samples together form a set of data. Similarly, the standard deviations of the M samples collectively form a set of data. Sets of data based on many samples drawn from a population are called **sampling distributions**. They are often used to describe the chance fluctuations of mean values and standard deviations based on random sampling.

66.3 The sampling distribution of the means

Suppose that it is required to obtain a sample of two items from a set containing five items. If the set is the five letters A, B, C, D and E , then the different samples that are possible are:

$$AB, AC, AD, AE, BC, BD, BE, CD, CE \text{ and } DE,$$

that is, ten different samples. The number of possible different samples in this case is given by ${}^5C_2 = \frac{5!}{2!3!} = 10$, from combinations on pages 120 and 358. Similarly, the number of different ways in which a sample of three items can be drawn from a set having ten members, ${}^{10}C_3 = \frac{10!}{3!7!} = 120$. It follows that when a small sample is drawn from a large population, there are very many different combinations of members possible. With so many different samples possible, quite a large variation can occur in the mean values of various samples taken from the same population.

Usually, the greater the number of members in a sample, the closer will be the mean value of the sample to that of the population. Consider the set of numbers 3, 4, 5, 6 and 7. For a sample of 2 members, the lowest

value of the mean is $\frac{3+4}{2}$, i.e. 3.5; the highest is $\frac{6+7}{2}$, i.e. 6.5, giving a range of mean values of $6.5 - 3.5 = 3$. For a sample of 3 members, the range is, $\frac{3+4+5}{3}$ to $\frac{5+6+7}{3}$ that is, 2. As the number in the sample increases, the range decreases until, in the limit, if the sample contains all the members of the set, the range of mean values is zero. When many samples are drawn from a population and a sample distribution of the mean values is small provided the number in the sample is large. Because the range is small it follows that the standard deviation of all the mean values will also be small, since it depends on the distance of the mean values from the distribution mean. The relationship between the standard deviation of the mean values of a sampling distribution and the number in each sample can be expressed as follows:

Theorem 1

'If all possible samples of size N are drawn from a finite population, N_p , without replacement, and the standard deviation of the mean values of the sampling distribution of means is determined, then:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}}$$

where $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of means and σ is the standard deviation of the population'

The standard deviation of a sampling distribution of mean values is called the **standard error of the means**, thus

standard error of the means,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \tag{1}$$

Equation (1) is used for a finite population of size N_p and/or for sampling without replacement. The word 'error' in the 'standard error of the means' does not mean that a mistake has been made but rather that there is a degree of uncertainty in predicting the mean value of a population based on the mean values of the samples. The formula for the standard error of the means is true for all values of the number in the sample, N . When N_p is very large compared with N or when the population is infinite (this can be considered to be the case when sampling is

done with replacement), the correction factor $\sqrt{\frac{N_p - N}{N_p - 1}}$ approaches unity and equation (1) becomes

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \tag{2}$$

Equation (2) is used for an infinite population and/or for sampling with replacement.

Theorem 2

'If all possible samples of size N are drawn from a population of size N_p and the mean value of the sampling distribution of means $\mu_{\bar{x}}$ is determined then

$$\mu_{\bar{x}} = \mu \tag{3}$$

where μ is the mean value of the population'

In practice, all possible samples of size N are not drawn from the population. However, if the sample size is large (usually taken as 30 or more), then the relationship between the mean of the sampling distribution of means and the mean of the population is very near to that shown in equation (3). Similarly, the relationship between the standard error of the means and the standard deviation of the population is very near to that shown in equation (2).

Another important property of a sampling distribution is that when the sample size, N , is large, **the sampling distribution of means approximates to a normal distribution**, of mean value $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$. This is true for all normally distributed populations and also for populations that are not normally distributed provided the population size is at least twice as large as the sample size. This property of normality of a sampling distribution is based on a special case of the 'central limit theorem', an important theorem relating to sampling theory. Because the sampling distribution of means and standard deviations is normally distributed, the table of the partial areas under the standardised normal curve (shown in Table 41.1 on page 367) can be used to determine the probabilities of a particular sample lying between, say, ± 1 standard deviation, and so on. This point is expanded in Problem 3.

Problem 1. The heights of 3000 people are normally distributed with a mean of 175 cm, and a standard deviation of 8 cm. If random samples are taken of 40 people. predict the standard deviation and the mean of the sampling distribution of means if sampling is done (a) with replacement, and (b) without replacement

For the population: number of members, $N_p = 3000$; standard deviation, $\sigma = 8$ cm; mean, $\mu = 175$ cm
For the samples: number in each sample, $N = 40$

- (a) When sampling is done **with replacement**, the total number of possible samples (two or more can be the same) is infinite. Hence, from equation (2) the **standard error of the mean (i.e. the standard deviation of the sampling distribution of means)**.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{8}{\sqrt{40}} = \mathbf{1.265 \text{ cm}}$$

From equation (3), **the mean of the sampling distribution, $\mu_{\bar{x}} = \mu = 175$ cm.**

- (b) When sampling is done **without replacement**, the total number of possible samples is finite and hence equation (1) applies. Thus **the standard error of the means**

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \\ &= \frac{8}{\sqrt{40}} \sqrt{\frac{3000 - 40}{3000 - 1}} \\ &= (1.265)(0.9935) = \mathbf{1.257 \text{ cm}}\end{aligned}$$

As stated, following equation (3), provided the sample size is large, the mean of the sampling distribution of means is the same for both finite and infinite populations. Hence, from equation (3),

$$\mu_{\bar{x}} = \mathbf{175 \text{ cm}}$$

Problem 2. 1500 ingots of a metal have a mean mass of 6.5 kg and a standard deviation of 0.5 kg. Find the probability that a sample of 60 ingots chosen at random from the group, without replacement, will have a combined mass of (a) between 378 and 396 kg, and (b) more than 399 kg

For the population: numbers of members, $N_p = 1500$; standard deviation, $\sigma = 0.5$ kg; mean $\mu = 6.5$ kg

For the sample: number in sample, $N = 60$

If many samples of 60 ingots had been drawn from the group, then the mean of the sampling distribution of means, $\mu_{\bar{x}}$ would be equal to the mean of the population. Also, the standard error of means is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}}$$

In addition, the sample distribution would have been approximately normal. Assume that the sample given in the problem is one of many samples. For many (theoretical) samples:

the mean of the sampling distribution

$$\text{of means, } \mu_{\bar{x}} = \mu = 6.5 \text{ kg}$$

Also, the standard error of the means,

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \\ &= \frac{0.5}{\sqrt{60}} \sqrt{\frac{1500 - 60}{1500 - 1}} \\ &= 0.0633 \text{ kg}\end{aligned}$$

Thus, the sample under consideration is part of a normal distribution of mean value 6.5 kg and a standard error of the means of 0.0633 kg.

- (a) If the combined mass of 60 ingots is between 378 and 396 kg, then the mean mass of each of the 60 ingots lies between $\frac{378}{60}$ and $\frac{396}{60}$ kg, i.e. between 6.3 kg and 6.6 kg.

Since the masses are normally distributed, it is possible to use the techniques of the normal distribution to determine the probability of the mean mass lying between 6.3 and 6.6 kg. The normal standard variate value, z , is given by

$$z = \frac{x - \bar{x}}{\sigma}$$

hence for the sampling distribution of means, this becomes,

$$z = \frac{x - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Thus, 6.3 kg corresponds to a z -value of $\frac{6.3 - 6.5}{0.0633} = -3.16$ standard deviations.

Similarly, 6.6 kg corresponds to a z -value of $\frac{6.6 - 6.5}{0.0633} = 1.58$ standard deviations.

Using Table 41.1 (page 367), the areas corresponding to these values of standard deviations are 0.4992 and 0.4430 respectively. Hence **the probability of the mean lying between 6.3 kg and 6.6 kg is $0.4992 + 0.4430 = 0.9422$** . (This means that if 10 000 samples are drawn, 9422 of these samples

will have a combined mass of between 378 and 396 kg.)

- (b) If the combined mass of 60 ingots is 399 kg, the mean mass of each ingot is $\frac{399}{60}$, that is 6.65 kg.

The z -value for 6.65 kg is $\frac{6.65 - 6.5}{0.0633}$, i.e. 2.37 standard deviations. From Table 41.1 (page 367), the area corresponding to this z -value is 0.4911. But this is the area between the ordinate $z = 0$ and ordinate $z = 2.37$. The ‘more than’ value required is the total area to the right of the $z = 0$ ordinate, less the value between $z = 0$ and $z = 2.37$, i.e. $0.5000 - 0.4911$.

Thus, since areas are proportional to probabilities for the standardised normal curve, **the probability of the mean mass being more than 6.65 kg** is $0.5000 - 0.4911$, i.e. **0.0089**. (This means that only 89 samples in 10 000, for example, will have a combined mass exceeding 399 kg.)

Now try the following exercise

Exercise 223 Further problems on the sampling distribution of means

- The lengths of 1500 bolts are normally distributed with a mean of 22.4 cm and a standard deviation of 0.0438 cm. If 30 samples are drawn at random from this population, each sample being 36 bolts, determine the mean of the sampling distribution and standard error of the means when sampling is done with replacement.
[$\mu_{\bar{x}} = 22.4$ cm, $\sigma_{\bar{x}} = 0.0080$ cm]
- Determine the standard error of the means in Problem 1, if sampling is done without replacement, correct to four decimal places.
[$\sigma_{\bar{x}} = 0.0079$ cm]
- A power punch produces 1800 washers per hour. The mean inside diameter of the washers is 1.70 cm and the standard deviation is 0.013 cm. Random samples of 20 washers are drawn every 5 minutes. Determine the mean of the sampling distribution of means and the standard error of the means for one hour’s output from the punch, (a) with replacement

and (b) without replacement, correct to three significant figures.

- | | |
|-----|---|
| (a) | $\mu_{\bar{x}} = 1.70$ cm, |
| | $\sigma_{\bar{x}} = 2.91 \times 10^{-3}$ cm |
| (b) | $\mu_{\bar{x}} = 1.70$ cm, |
| | $\sigma_{\bar{x}} = 2.89 \times 10^{-3}$ cm |

A large batch of electric light bulbs have a mean time to failure of 800 hours and the standard deviation of the batch is 60 hours. Use this data and also Table 41.1 on page 367 to solve Problems 4 to 6.

- If a random sample of 64 light bulbs is drawn from the batch, determine the probability that the mean time to failure will be less than 785 hours, correct to three decimal places.
[0.023]
- Determine the probability that the mean time to failure of a random sample of 16 light bulbs will be between 790 hours and 810 hours, correct to three decimal places.
[0.497]
- For a random sample of 64 light bulbs, determine the probability that the mean time to failure will exceed 820 hours, correct to two significant figures.
[0.0038]

66.4 The estimation of population parameters based on a large sample size

When a population is large, it is not practical to determine its mean and standard deviation by using the basic formulae for these parameters. In fact, when a population is infinite, it is impossible to determine these values. For large and infinite populations the values of the mean and standard deviation may be estimated by using the data obtained from samples drawn from the population.

Point and interval estimates

An estimate of a population parameter, such as mean or standard deviation, based on a single number is called a **point estimate**. An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an **interval estimate**. Thus if an estimate is made of the length of an object and the result is quoted as 150 cm, this is a point estimate.

If the result is quoted as 150 ± 10 cm, this is an interval estimate and indicates that the length lies between 140 and 160 cm. Generally, a point estimate does not indicate how close the value is to the true value of the quantity and should be accompanied by additional information on which its merits may be judged. A statement of the error or the precision of an estimate is often called its **reliability**. In statistics, when estimates are made of population parameters based on samples, usually interval estimates are used. The word estimate does not suggest that we adopt the approach 'let's guess that the mean value is about..', but rather that a value is carefully selected and the degree of confidence which can be placed in the estimate is given in addition.

Confidence intervals

It is stated in Section 66.3 that when samples are taken from a population, the mean values of these samples are approximately normally distributed, that is, the mean values forming the sampling distribution of means is approximately normally distributed. It is also true that if the standard deviation of each of the samples is found, then the standard deviations of all the samples are approximately normally distributed, that is, the standard deviations of the sampling distribution of standard deviations are approximately normally distributed. Parameters such as the mean or the standard deviation of a sampling distribution are called **sampling statistics**, S . Let μ_S be the mean value of a sampling statistic of the sampling distribution, that is, the mean value of the means of the samples or the mean value of the standard deviations of the samples. Also let σ_S be the standard deviation of a sampling statistic of the sampling distribution, that is, the standard deviation of the means of the samples or the standard deviation of the standard deviations of the samples. Because the sampling distribution of the means and of the standard deviations are normally distributed, it is possible to predict the probability of the sampling statistic lying in the intervals:

mean ± 1 standard deviation,

mean ± 2 standard deviations,

or mean ± 3 standard deviations,

by using tables of the partial areas under the standardised normal curve given in Table 41.1 on page 367. From this table, the area corresponding to a z -value of +1 standard deviation is 0.3413, thus the area corresponding to +1 standard deviation is 2×0.3413 , that is, 0.6826. Thus the percentage probability of a sampling statistic lying between the mean ± 1 standard deviation is

68.26%. Similarly, the probability of a sampling statistic lying between the mean ± 2 standard deviations is 95.44% and of lying between the mean ± 3 standard deviations is 99.74%

The values 68.26%, 95.44% and 99.74% are called the **confidence levels** for estimating a sampling statistic. A confidence level of 68.26% is associated with two distinct values, these being, $S - (1 \text{ standard deviation})$, i.e. $S - \sigma_S$ and $S + (1 \text{ standard deviation})$, i.e. $S + \sigma_S$. These two values are called the **confidence limits** of the estimate and the distance between the confidence limits is called the **confidence interval**. A confidence interval indicates the expectation or confidence of finding an estimate of the population statistic in that interval, based on a sampling statistic. The list in Table 66.1 is based on values given in Table 41.1, and gives some of the confidence levels used in practice and their associated z -values; (some of the values given are based on interpolation). When the table is used in this context, z -values are usually indicated by ' z_C ' and are called the **confidence coefficients**.

Table 66.1

Confidence level, %	Confidence coefficient, z_C
99	2.58
98	2.33
96	2.05
95	1.96
90	1.645
80	1.28
50	0.6745

Any other values of confidence levels and their associated confidence coefficients can be obtained using Table 41.1.

Problem 3. Determine the confidence coefficient corresponding to a confidence level of 98.5%

98.5% is equivalent to a per unit value of 0.985%. This indicates that the area under the standardised normal curve between $-z_C$ and $+z_C$, i.e. corresponding to $2z_C$, 0.9850 of the total area. Hence the area between the

mean value and z_C is $\frac{0.9850}{2}$ i.e. 0.4925 of the total area. The z-value corresponding to a partial area of 0.4925 is 2.43 standard deviations from Table 41.1. Thus, **the confidence coefficient corresponding to a confidence limit of 98.5% is 2.43**

(a) Estimating the mean of a population when the standard deviation of the population is known

When a sample is drawn from a large population whose standard deviation is known, the mean value of the sample, \bar{x} , can be determined. This mean value can be used to make an estimate of the mean value of the population, μ . When this is done, the estimated mean value of the population is given as lying between two values, that is, lying in the confidence interval between the confidence limits. If a high level of confidence is required in the estimated value of μ , then the range of the confidence interval will be large. For example, if the required confidence level is 96%, then from Table 66.1 the confidence interval is from $-z_C$ to $+z_C$, that is, $2 \times 2.05 = 4.10$ standard deviations wide. Conversely, a low level of confidence has a narrow confidence interval and a confidence level of, say, 50%, has a confidence interval of 2×0.6745 , that is 1.3490 standard deviations. The 68.26% confidence level for an estimate of the population mean is given by estimating that the population mean, μ , is equal to the same mean, \bar{x} , and then stating the confidence interval of the estimate. Since the 68.26% confidence level is associated with ‘ ± 1 standard deviation of the means of the sampling distribution’, then the 68.26% confidence level for the estimate of the population mean is given by:

$$\bar{x} \pm 1\sigma_{\bar{x}}$$

In general, any particular confidence level can be obtained in the estimate, by using $\bar{x} + z_C\sigma_{\bar{x}}$, where z_C is the confidence coefficient corresponding to the particular confidence level required. Thus for a 96% confidence level, the confidence limits of the population mean are given by $\bar{x} + 2.05\sigma_{\bar{x}}$. Since only one sample has been drawn, the standard error of the means, $\sigma_{\bar{x}}$, is not known. However, it is shown in Section 66.3 that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}}$$

Thus, **the confidence limits of the mean of the population are:**

$$\bar{x} \pm \frac{z_C\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \tag{4}$$

for a finite population of size N_p

The **confidence limits for the mean of the population are:**

$$\bar{x} \pm \frac{z_C\sigma}{\sqrt{N}} \tag{5}$$

for an infinite population.

Thus for a sample of size N and mean \bar{x} , drawn from an infinite population having a standard deviation of σ , the mean value of the population is estimated to be, for example,

$$\bar{x} + \frac{2.33\sigma}{\sqrt{N}}$$

for a confidence level of 98%. This indicates that the mean value of the population lies between

$$\bar{x} - \frac{2.33\sigma}{\sqrt{N}} \quad \text{and} \quad \bar{x} + \frac{2.33\sigma}{\sqrt{N}}$$

with 98% confidence in this prediction.

Problem 4. It is found that the standard deviation of the diameters of rivets produced by a certain machine over a long period of time is 0.018 cm. The diameters of a random sample of 100 rivets produced by this machine in a day have a mean value of 0.476 cm. If the machine produces 2500 rivets a day, determine (a) the 90% confidence limits, and (b) the 97% confidence limits for an estimate of the mean diameter of all the rivets produced by the machine in a day

For the population:

standard deviation, $\sigma = 0.018$ cm

number in the population, $N_p = 2500$

For the sample:

number in the sample, $N = 100$

mean, $\bar{x} = 0.476$ cm

There is a finite population and the standard deviation of the population is known, hence expression (4) is used for determining an estimate of the confidence limits of the population mean, i.e.

$$\bar{x} \pm \frac{z_C\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}}$$

- (a) For a 90% confidence level, the value of z_C , the confidence coefficient, is 1.645 from Table 66.1.

Hence, the estimate of the confidence limits of the population mean,

$$\begin{aligned}\mu &= 0.476 \\ &\pm \left(\frac{(1.645)(0.018)}{\sqrt{100}} \right) \sqrt{\frac{2500 - 100}{2500 - 1}} \\ &= 0.476 \pm (0.00296)(0.9800) \\ &= 0.476 \pm 0.0029 \text{ cm}\end{aligned}$$

Thus, the **90% confidence limits are 0.473 cm and 0.479 cm.**

This indicates that if the mean diameter of a sample of 100 rivets is 0.476 cm, then it is predicted that the mean diameter of all the rivets will be between 0.473 cm and 0.479 cm and this prediction is made with confidence that it will be correct nine times out of ten.

- (b) For a 97% confidence level, the value of z_C has to be determined from a table of partial areas under the standardised normal curve given in Table 41.1, as it is not one of the values given in Table 66.1. The total area between ordinates drawn at $-z_C$ and $+z_C$ has to be 0.9700. Because the area is $\frac{0.9700}{2}$, i.e. 0.4850. From Table 41.1 an area of 0.4850 corresponds to a z_C value of 2.17. Hence, the estimated value of the confidence limits of the population mean is between

$$\begin{aligned}\bar{x} &\pm \frac{z_C \sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \\ &= 0.476 \\ &\pm \left(\frac{(2.17)(0.018)}{\sqrt{100}} \right) \sqrt{\frac{2500 - 100}{2500 - 1}} \\ &= 0.476 \pm (0.0039)(0.9800) \\ &= 0.476 \pm 0.0038\end{aligned}$$

Thus, the **97% confidence limits are 0.472 cm and 0.480 cm.**

It can be seen that the higher value of confidence level required in part (b) results in a larger confidence interval.

Problem 5. The mean diameter of a long length of wire is to be determined. The diameter of the wire is measured in 25 places selected at random throughout its length and the mean of these values

is 0.425 mm. If the standard deviation of the diameter of the wire is given by the manufacturers as 0.030 mm, determine (a) the 80% confidence interval of the estimated mean diameter of the wire, and (b) with what degree of confidence it can be said that 'the mean diameter is 0.425 ± 0.012 mm'

For the population: $\sigma = 0.030$ mm

For the sample: $N = 25$, $\bar{x} = 0.425$ mm

Since an infinite number of measurements can be obtained for the diameter of the wire, the population is infinite and the estimated value of the confidence interval of the population mean is given by expression (5).

- (a) For an 80% confidence level, the value of z_C is obtained from Table 66.1 and is 1.28.

The 80% confidence level estimate of the confidence interval of

$$\begin{aligned}\mu &= \bar{x} \pm \frac{z_C \sigma}{\sqrt{N}} \\ &= 0.425 \pm \frac{(1.28)(0.030)}{\sqrt{25}} \\ &= 0.425 \pm 0.0077 \text{ mm}\end{aligned}$$

i.e. the **80% confidence interval is from 0.417 mm to 0.433 mm.**

This indicates that the estimated mean diameter of the wire is between 0.417 mm and 0.433 mm and that this prediction is likely to be correct 80 times out of 100

- (b) To determine the confidence level, the given data is equated to expression (5), giving:

$$0.425 \pm 0.012 = \bar{x} \pm z_C \frac{\sigma}{\sqrt{N}}$$

But $\bar{x} = 0.425$, therefore

$$\pm z_C \frac{\sigma}{\sqrt{N}} = \pm 0.012$$

$$\begin{aligned}\text{i.e. } z_C &= \frac{0.012 \sqrt{N}}{\sigma} \\ &= \pm \frac{(0.012)(5)}{0.030} = \pm 2\end{aligned}$$

Using Table 41.1 of partial areas under the standardised normal curve, a z_C value of 2 standard deviations corresponds to an area of 0.4772 between the mean value ($z_C = 0$) and $+2$ standard deviations. Because the standardised normal curve

is symmetrical, the area between the mean and ± 2 standard deviations is 0.4772×2 , i.e. 0.9544

Thus the confidence level corresponding to 0.425 ± 0.012 mm is 95.44%.

(b) Estimating the mean and standard deviation of a population from sample data

The standard deviation of a large population is not known and, in this case, several samples are drawn from the population. The mean of the sampling distribution of means, $\mu_{\bar{x}}$ and the standard deviation of the sampling distribution of mean (i.e. the standard error of the means), $\sigma_{\bar{x}}$, may be determined. The confidence limits of the mean value of the population, μ , are given by:

$$\mu_{\bar{x}} \pm z_C \sigma_{\bar{x}} \quad (6)$$

where z_C is the confidence coefficient corresponding to the confidence level required.

To make an estimate of the standard deviation, σ , of a normally distributed population:

- (i) a sampling distribution of the standard deviations of the samples is formed, and
- (ii) the standard deviation of the sampling distribution is determined by using the basic standard deviation formula.

This standard deviation is called the standard error of the standard deviations and is usually signified by σ_S . If s is the standard deviation of a sample, then the confidence limits of the standard deviation of the population are given by:

$$s \pm z_C \sigma_S \quad (7)$$

where z_C is the confidence coefficient corresponding to the required confidence level.

Problem 6. Several samples of 50 fuses selected at random from a large batch are tested when operating at a 10% overload current and the mean time of the sampling distribution before the fuses failed is 16.50 minutes. The standard error of the means is 1.4 minutes. Determine the estimated mean time to failure of the batch of fuses for a confidence level of 90%

For the sampling distribution: the mean, $\mu_{\bar{x}} = 16.50$, the standard error of the means, $\sigma_{\bar{x}} = 1.4$

The estimated mean of the population is based on sampling distribution data only and so expression (6) is used, i.e. the confidence limits of the estimated mean of the population are $\mu_{\bar{x}} \pm z_C \sigma_{\bar{x}}$

For an 90% confidence level, $z_C = 1.645$ (from Table 66.1), thus

$$\begin{aligned} \mu_{\bar{x}} \pm z_C \sigma_{\bar{x}} &= 16.50 \pm (1.645)(1.4) \\ &= 16.50 \pm 2.30 \text{ minutes} \end{aligned}$$

Thus, the 90% confidence level of the mean time to failure is from 14.20 minutes to 18.80 minutes.

Problem 7. The sampling distribution of random samples of capacitors drawn from a large batch is found to have a standard error of the standard deviations of $0.12 \mu\text{F}$. Determine the 92% confidence interval for the estimate of the standard deviation of the whole batch, if in a particular sample, the standard deviation is $0.60 \mu\text{F}$. It can be assumed that the values of capacitance of the batch are normally distributed

For the sample: the standard deviation, $s = 0.60 \mu\text{F}$
For the sampling distribution: the standard error of the standard deviations,

$$\sigma_S = 0.12 \mu\text{F}$$

When the confidence level is 92%, then by using Table 41.1 of partial areas under the standardised normal curve,

$$\text{area} = \frac{0.9200}{2} = 0.4600,$$

giving z_C as ± 1.751 standard deviations (by interpolation)

Since the population is normally distributed, the confidence limits of the standard deviation of the population may be estimated by using expression (7),

$$\begin{aligned} \text{i.e. } s \pm z_C \sigma_S &= 0.60 \pm (1.751)(0.12) \\ &= 0.60 \pm 0.21 \mu\text{F} \end{aligned}$$

Thus, the 92% confidence interval for the estimate of the standard deviation for the batch is from $0.39 \mu\text{F}$ to $0.81 \mu\text{F}$.

Now try the following exercise

Exercise 224 Further problems on the estimation of population parameters based on a large sample size

1. Measurements are made on a random sample of 100 components drawn from a population of size 1546 and having a standard deviation of 2.93 mm. The mean measurement of the components in the sample is 67.45 mm. Determine the 95% and 99% confidence limits for an estimate of the mean of the population.

$$\left[\begin{array}{l} 66.89 \text{ and } 68.01 \text{ mm,} \\ 66.72 \text{ and } 68.18 \text{ mm} \end{array} \right]$$

2. The standard deviation of the masses of 500 blocks is 150 kg. A random sample of 40 blocks has a mean mass of 2.40 Mg.

- (a) Determine the 95% and 99% confidence intervals for estimating the mean mass of the remaining 460 blocks.
- (b) With what degree of confidence can it be said that the mean mass of the remaining 460 blocks is 2.40 ± 0.035 Mg?

$$\left[\begin{array}{l} \text{(a) } 2.355 \text{ Mg to } 2.445 \text{ Mg;} \\ \quad 2.341 \text{ Mg to } 2.459 \text{ Mg} \\ \text{(b) } 86\% \end{array} \right]$$

3. In order to estimate the thermal expansion of a metal, measurements of the change of length for a known change of temperature are taken by a group of students. The sampling distribution of the results has a mean of $12.81 \times 10^{-4} \text{ m } ^\circ\text{C}^{-1}$ and a standard error of the means of $0.04 \times 10^{-4} \text{ m } ^\circ\text{C}^{-1}$. Determine the 95% confidence interval for an estimate of the true value of the thermal expansion of the metal, correct to two decimal places.

$$\left[\begin{array}{l} 12.73 \times 10^{-4} \text{ m } ^\circ\text{C}^{-1} \text{ to} \\ 12.89 \times 10^{-4} \text{ m } ^\circ\text{C}^{-1} \end{array} \right]$$

4. The standard deviation of the time to failure of an electronic component is estimated as 100 hours. Determine how large a sample of these components must be, in order to be 90% confident that the error in the estimated time

to failure will not exceed (a) 20 hours, and (b) 10 hours. [(a) at least 68 (b) at least 271]

5. The time taken to assemble a servomechanism is measured for 40 operatives and the mean time is 14.63 minutes with a standard deviation of 2.45 minutes. Determine the maximum error in estimating the true mean time to assemble the servo-mechanism for all operatives, based on a 95% confidence level. [45.6 seconds]

66.5 Estimating the mean of a population based on a small sample size

The methods used in Section 66.4 to estimate the population mean and standard deviation rely on a relatively large sample size, usually taken as 30 or more. This is because when the sample size is large the sampling distribution of a parameter is approximately normally distributed. When the sample size is small, usually taken as less than 30, the techniques used for estimating the population parameters in Section 66.4 become more and more inaccurate as the sample size becomes smaller, since the sampling distribution no longer approximates to a normal distribution. Investigations were carried out into the effect of small sample sizes on the estimation theory by W. S. Gosset in the early twentieth century and, as a result of his work, tables are available which enable a realistic estimate to be made, when sample sizes are small. In these tables, the t -value is determined from the relationship

$$t = \frac{(\bar{x} - \mu)}{s} \sqrt{N - 1}$$

where \bar{x} is the mean value of a sample, μ is the mean value of the population from which the sample is drawn, s is the standard deviation of the sample and N is the number of independent observations in the sample. He published his findings under the pen name of 'Student', and these tables are often referred to as the '**Student's t distribution**'.

The confidence limits of the mean value of a population based on a small sample drawn at random from the population are given by:

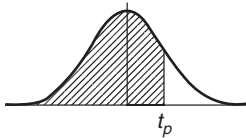
$$\bar{x} \pm \frac{tcs}{\sqrt{N - 1}} \quad (8)$$

In this estimate, t_C is called the confidence coefficient for small samples, analogous to z_C for large samples, s is the standard deviation of the sample, \bar{x} is the mean value of the sample and N is the number of members in the sample. Table 66.2 is called ‘percentile values for Student’s t distribution’. The columns are headed t_p where p is equal to 0.995, 0.99, 0.975, . . . , 0.55. For a confidence level of, say, 95%, the column headed $t_{0.95}$ is selected and so on. The rows are

headed with the Greek letter ‘nu’, ν , and are numbered from 1 to 30 in steps of 1, together with the numbers 40, 60, 120 and ∞ . These numbers represent a quantity called the **degrees of freedom**, which is defined as follows:

‘the sample number, N , minus the number of population parameters which must be estimated for the sample’.

Table 66.2 Percentile values (t_p) for Student’s t distribution with ν degrees of freedom (shaded area = p)



ν	$t_{0.995}$	$t_{0.99}$	$t_{0.975}$	$t_{0.95}$	$t_{0.90}$	$t_{0.80}$	$t_{0.75}$	$t_{0.70}$	$t_{0.60}$	$t_{0.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	0.727	0.325	0.158
2	9.92	6.96	4.30	2.92	1.89	1.061	0.816	0.617	0.289	0.142
3	5.84	4.54	3.18	2.35	1.64	0.978	0.765	0.584	0.277	0.137
4	4.60	3.75	2.78	2.13	1.53	0.941	0.741	0.569	0.271	0.134
5	4.03	3.36	2.57	2.02	1.48	0.920	0.727	0.559	0.267	0.132
6	3.71	3.14	2.45	1.94	1.44	0.906	0.718	0.553	0.265	0.131
7	3.50	3.00	2.36	1.90	1.42	0.896	0.711	0.549	0.263	0.130
8	3.36	2.90	2.31	1.86	1.40	0.889	0.706	0.546	0.262	0.130
9	3.25	2.82	2.26	1.83	1.38	0.883	0.703	0.543	0.261	0.129
10	3.17	2.76	2.23	1.81	1.37	0.879	0.700	0.542	0.260	0.129
11	3.11	2.72	2.20	1.80	1.36	0.876	0.697	0.540	0.260	0.129
12	3.06	2.68	2.18	1.78	1.36	0.873	0.695	0.539	0.259	0.128
13	3.01	2.65	2.16	1.77	1.35	0.870	0.694	0.538	0.259	0.128
14	2.98	2.62	2.14	1.76	1.34	0.868	0.692	0.537	0.258	0.128
15	2.95	2.60	2.13	1.75	1.34	0.866	0.691	0.536	0.258	0.128
16	2.92	2.58	2.12	1.75	1.34	0.865	0.690	0.535	0.258	0.128
17	2.90	2.57	2.11	1.74	1.33	0.863	0.689	0.534	0.257	0.128
18	2.88	2.55	2.10	1.73	1.33	0.862	0.688	0.534	0.257	0.127

Table 66.2 (Continued)

ν	$t_{0.995}$	$t_{0.99}$	$t_{0.975}$	$t_{0.95}$	$t_{0.90}$	$t_{0.80}$	$t_{0.75}$	$t_{0.70}$	$t_{0.60}$	$t_{0.55}$
19	2.86	2.54	2.09	1.73	1.33	0.861	0.688	0.533	0.257	0.127
20	2.84	2.53	2.09	1.72	1.32	0.860	0.687	0.533	0.257	0.127
21	2.83	2.52	2.08	1.72	1.32	0.859	0.686	0.532	0.257	0.127
22	2.82	2.51	2.07	1.72	1.32	0.858	0.686	0.532	0.256	0.127
23	2.81	2.50	2.07	1.71	1.32	0.858	0.685	0.532	0.256	0.127
24	2.80	2.49	2.06	1.71	1.32	0.857	0.685	0.531	0.256	0.127
25	2.79	2.48	2.06	1.71	1.32	0.856	0.684	0.531	0.256	0.127
26	2.78	2.48	2.06	1.71	1.32	0.856	0.684	0.531	0.256	0.127
27	2.77	2.47	2.05	1.70	1.31	0.855	0.684	0.531	0.256	0.127
28	2.76	2.47	2.05	1.70	1.31	0.855	0.683	0.530	0.256	0.127
29	2.76	2.46	2.04	1.70	1.31	0.854	0.683	0.530	0.256	0.127
30	2.75	2.46	2.04	1.70	1.31	0.854	0.683	0.530	0.256	0.127
40	2.70	2.42	2.02	1.68	1.30	0.851	0.681	0.529	0.255	0.126
60	2.66	2.39	2.00	1.67	1.30	0.848	0.679	0.527	0.254	0.126
120	2.62	2.36	1.98	1.66	1.29	0.845	0.677	0.526	0.254	0.126
∞	2.58	2.33	1.96	1.645	1.28	0.842	0.674	0.524	0.253	0.126

When determining the t -value, given by

$$t = \frac{(\bar{x} - \mu)}{s} \sqrt{N - 1}$$

it is necessary to know the sample parameters \bar{x} and s and the population parameter μ . \bar{x} and s can be calculated for the sample, but usually an estimate has to be made of the population mean μ , based on the sample mean value. The number of degrees of freedom, ν , is given by the number of independent observations in the sample, N , minus the number of population parameters which have to be estimated, k , i.e. $\nu = N - k$. For the equation

$$t = \frac{(\bar{x} - \mu)}{s} \sqrt{N - 1}$$

only μ has to be estimated, hence $k = 1$, and $\nu = N - 1$.

When determining the mean of a population based on a small sample size, only one population parameter is to be estimated, and hence ν can always be taken as $(N - 1)$. The method used to estimate the mean of a population based on a small sample is shown in Problems 8 to 10.

Problem 8. A sample of 12 measurements of the diameter of a bar is made and the mean of the sample is 1.850 cm. The standard deviation of the samples is 0.16 mm. Determine (a) the 90% confidence limits, and (b) the 70% confidence limits for an estimate of the actual diameter of the bar

For the sample: the sample size, $N = 12$;
mean, $\bar{x} = 1.850$ cm;
standard deviation, $s = 0.16$ mm = 0.016 cm

Since the sample number is less than 30, the small sample estimate as given in expression (8) must be used. The number of degrees of freedom, i.e. sample size minus the number of estimations of population parameters to be made, is $12 - 1$, i.e. 11.

- (a) The percentile value corresponding to a confidence coefficient value of $t_{0.90}$ and a degree of freedom value of $\nu = 11$ can be found by using Table 66.2, and is 1.36, that is, $t_C = 1.36$. The estimated value of the population is given by

$$\begin{aligned}\bar{x} \pm \frac{t_C s}{\sqrt{N-1}} \\ &= 1.850 \pm \frac{(1.36)(0.016)}{\sqrt{11}} \\ &= 1.850 \pm 0.0066 \text{ cm.}\end{aligned}$$

Thus, **the 90% confidence limits are 1.843 cm and 1.857 cm.**

This indicates that the actual diameter is likely to lie between 1.843 cm and 1.857 cm and that this prediction stands a 90% chance of being correct.

- (b) The percentile value corresponding to $t_{0.70}$ and to $\nu = 11$ is obtained from Table 66.2, and is 0.540, that is, $t_C = 0.540$

The estimated value of the 70% confidence limits is given by:

$$\begin{aligned}\bar{x} \pm \frac{t_C s}{\sqrt{N-1}} \\ &= 1.850 \pm \frac{(0.540)(0.016)}{\sqrt{11}} \\ &= 1.850 \pm 0.0026 \text{ cm}\end{aligned}$$

Thus, **the 70% confidence limits are 1.847 cm and 1.853 cm**, i.e. the actual diameter of the bar is between 1.847 cm and 1.853 cm and this result has an 70% probability of being correct.

Problem 9. A sample of 9 electric lamps are selected randomly from a large batch and are tested until they fail. The mean and standard deviations of the time to failure are 1210 hours and 26 hours respectively. Determine the confidence level based on an estimated failure time of 1210 ± 6.5 hours

For the sample: sample size, $N = 9$; standard deviation, $s = 26$ hours; mean, $\bar{x} = 1210$ hours. The confidence

limits are given by:

$$\bar{x} \pm \frac{t_C s}{\sqrt{N-1}}$$

and these are equal to 1210 ± 6.5

Since $\bar{x} = 1210$ hours, then

$$\pm \frac{t_C s}{\sqrt{N-1}} = \pm 6.5$$

$$\begin{aligned}\text{i.e. } t_C &= \pm \frac{65\sqrt{N-1}}{s} \\ &= \pm \frac{(65)\sqrt{8}}{26} = \pm 0.707\end{aligned}$$

From Table 66.2, a t_C value of 0.707, having a value of $N - 1$, i.e. 8, gives a t_p value of $t_{0.75}$

Hence, **the confidence level of an estimated failure time of 1210 ± 6.5 hours is 75%**, i.e. it is likely that 75% of all of the lamps will fail between 1203.5 and 1216.5 hours.

Problem 10. The specific resistance of some copper wire of nominal diameter 1 mm is estimated by determining the resistance of 6 samples of the wire. The resistance values found (in ohms per metre) were:

2.16, 2.14, 2.17, 2.15, 2.16 and 2.18

Determine the 95% confidence interval for the true specific resistance of the wire

For the sample: sample size, $N = 6$ mean,

$$\begin{aligned}\bar{x} &= \frac{2.16 + 2.14 + 2.17 + 2.15 + 2.16 + 2.18}{6} \\ &= 2.16 \Omega \text{ m}^{-1}\end{aligned}$$

standard deviation,

$$\begin{aligned}s &= \sqrt{\frac{(2.16 - 2.16)^2 + (2.14 - 2.16)^2 \\ &\quad + (2.17 - 2.16)^2 + (2.15 - 2.16)^2 \\ &\quad + (2.16 - 2.16)^2 + (2.18 - 2.16)^2}{6}} \\ &= \sqrt{\frac{0.001}{6}} = 0.0129 \Omega \text{ m}^{-1}\end{aligned}$$

The percentile value corresponding to a confidence coefficient value of $t_{0.95}$ and a degree of freedom value of $N - 1$, i.e. $6 - 1 = 5$ is 2.02 from Table 66.2. The estimated value of the 95% confidence limits is given by:

$$\begin{aligned}\bar{x} \pm \frac{t_{CS}}{\sqrt{N-1}} &= 2.16 \pm \frac{(2.02)(0.0129)}{\sqrt{5}} \\ &= 2.16 \pm 0.01165 \Omega \text{ m}^{-1}\end{aligned}$$

Thus, the 95% confidence limits are $2.148 \Omega \text{ m}^{-1}$ and $2.172 \Omega \text{ m}^{-1}$ which indicates that there is a 95% chance that the true specific resistance of the wire lies between $2.148 \Omega \text{ m}^{-1}$ and $2.172 \Omega \text{ m}^{-1}$.

Now try the following exercise

Exercise 225 Further problems on estimating the mean of a population based on a small sample size

- The value of the ultimate tensile strength of a material is determined by measurements on 10 samples of the material. The mean and standard deviation of the results are found to be 5.17 MPa and 0.06 MPa respectively. Determine the 95% confidence interval for the mean of the ultimate tensile strength of the material.
[5.133 MPa to 5.207 MPa]
- Use the data given in Problem 1 above to determine the 97.5% confidence interval for the mean of the ultimate tensile strength of the material.
[5.125 MPa to 5.215 MPa]
- The specific resistance of a reel of German silver wire of nominal diameter 0.5 mm is estimated by determining the resistance of 7 samples of the wire. These were found to have resistance values (in ohms per metre) of:

1.12 1.15 1.10 1.14 1.15 1.10 and 1.11

Determine the 99% confidence interval for the true specific resistance of the reel of wire.
[1.10 $\Omega \text{ m}^{-1}$ to 1.15 $\Omega \text{ m}^{-1}$]
- In determining the melting point of a metal, five determinations of the melting point are made. The mean and standard deviation of the five results are 132.27°C and 0.742°C . Calculate the confidence with which the prediction 'the melting point of the metal is between 131.48°C and 133.06°C ' can be made.
[95%]