

6

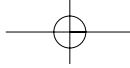
Statistical estimation using confidence intervals

In Chapter 2, the concept of the central nature and variability of data and the methods by which these two phenomena may be mathematically determined were described. As previously explained, one of the main purposes for calculating the central nature (e.g. the mean) and the variability (e.g. the standard deviation) of a set of sample data is to gain an understanding of the corresponding population statistics. In other words, the mean (\bar{X}) and standard deviation (s) of sample data are employed to estimate the true (population) mean and true (population) standard deviation. However, it is reasonable to ask, how reliable are the sample data at representing the population data? Is the sample data a good estimation of the population data?

6.1 The concept of confidence intervals

After calculating the mean and standard deviation of a sample, as is the normal approach in the pharmaceutical and related sciences, we need to provide an indication of the reliability of the data. For example, in a clinical trial ($N = 20$ patients) the volume of urine produced after the administration of a new diuretic drug has been calculated as 5.2 ± 1.9 L. In light of the small sample size, it is unreasonable to predict that the population mean and standard deviation will be identical to these observed sample values, as each sample will produce different mean and standard deviation values. Therefore, when reporting the mean of sample data it is good practice to present some indication of the reliability of the data, i.e. the quality of the estimation of the true mean from the sample mean. This is performed using *confidence intervals*. The confidence intervals are quoted as a mean and range (interval), the latter representing the probability of observing the true mean. The imposed probability is chosen by the person responsible for the statistical manipulation of data; however, the most frequently employed confidence intervals are 90%, 95% and 99%.

Confidence intervals are demonstrated graphically in Figures 6.1 and 6.2. In these, 10 samples, each containing 10 observations, were



136 Statistical estimation using confidence intervals

selected from a normal distribution, whose mean was known to be 50. For each of the 10 samples, the mean and 90% confidence intervals were calculated. (The method for doing this is shown in the following section and does not need to be explained at this stage.) An example of the outcome for each sample is shown in Figure 6.1. The central portion of the bar denotes the mean that has been calculated using the sample data. The regions that extend equally beyond the mean in either direction are the calculated 90% confidence intervals. As for any range, there are upper and lower limits. The calculated mean and (90%) confidence intervals for each sample may be plotted on the one graph, as shown in Figure 6.2. The mean of the population has been included in this graph as a dashed line.

Some features of Figure 6.2 require clarification to enable the reader to fully understand the basis of confidence intervals.

- First, it may be observed that, as expected, the means of all the different samples, although numerically close, are not equal to the population mean. Furthermore, the means of the individual samples are not numerically equal, once more reflecting sample variability.
- Secondly, the confidence intervals of the individual groups also differ in magnitude. These differences may also be logically explained. Confidence intervals are calculated using the standard deviation of the sample and therefore, if the variability of a sample is large, then the size of the confidence interval (i.e. the range) will be larger. Consequently, differences in the sizes of the confidence intervals reflect differences in the standard deviations of the individual samples.
- Finally, it may be observed from Figure 6.2 that, although the individual sample means do not exactly equal the population mean, the 90% confidence intervals in all but one case (denoted by an asterisk) do encompass the population mean. Therefore, the calculated 90% confidence intervals have included the population mean in 9 out of 10 observations, i.e. 90% of observations. If

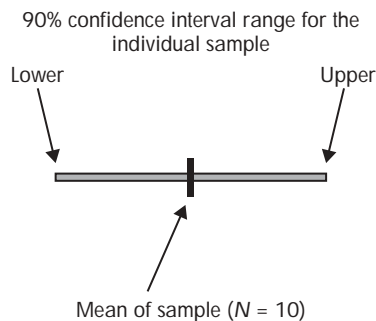
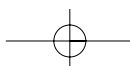


Figure 6.1 Mean and confidence intervals.



Confidence intervals for the population mean 137

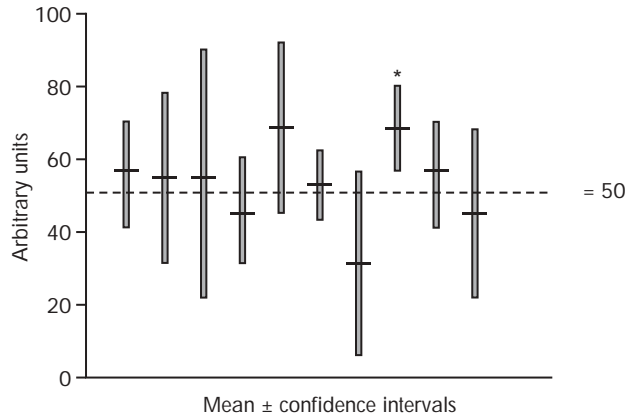


Figure 6.2 Mean and 90% confidence intervals derived from 10 samples (10 observations per sample group).

the 95% confidence intervals had been calculated for each sample and plotted as described in Figure 6.2, the confidence intervals would have encompassed the true mean in 19 out of 20 samples (95%), and so on.

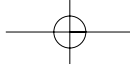
Confidence intervals are therefore calculated to provide the user with the probability that a single sample will contain the true mean (or indeed true standard deviation, etc). It is still possible that if a sample is removed and the mean \pm 95% confidence intervals calculated, the true mean will not fall within the range calculated. However, it is also true that if another 19 samples were removed and the mean \pm 95% confidence intervals calculated, the true mean would fall within the confidence intervals described by these 19 samples.

The calculation of confidence intervals is performed with reference to a probability distribution, e.g. the normal distribution, t distribution or χ^2 distribution. The mechanism for performing such calculations forms the basis of the remainder of this chapter.

6.2 Confidence intervals for the population mean and the normal distribution

One of the most frequently used methods for calculating confidence intervals involves the use of the normal distribution. If the data conforms to the normal distribution, the two-tailed confidence interval may be calculated using the following equation:

$$P\% = \bar{X} \pm \frac{z_{P\%} \sigma}{\sqrt{N}}$$



138 Statistical estimation using confidence intervals

where $P\%$ is the selected confidence interval, i.e. usually 90%, 95% or 99%, \bar{X} is the observed mean, σ is the population standard deviation and $z_{P\%}$ is the z value corresponding to the percentage confidence interval. Remember that the term σ/\sqrt{N} refers to the standard error of the mean, previously introduced in Chapter 2.

The z value is therefore an important parameter in the calculation of confidence intervals. Confidence limits are two-tailed events as there are two possible outcomes, i.e. one interval below and one interval above the mean. The z value chosen for inclusion in the equation or calculation of confidence is dependent on the selected value of the probability, as shown in Figure 6.3. Therefore, if we want to calculate the 95% confidence interval, we choose a z value of 1.96 as this value corresponds to the area under the standard normal distribution that encompasses 95% of all values. Similarly, 2.58 and 1.65 are chosen when 99% and 90% confidence intervals are to be calculated. The calculation of confidence intervals based on the standardised normal distribution is shown in the following example.

EXAMPLE 6.1 *A clinical trial ($N = 30$ patients) has been performed in which the volume of distribution of a new anti-diabetes drug has been calculated as 10.2 ± 1.9 L. Calculate the 95% and 99% confidence limits of the mean value (assuming that the data originated from a normal distribution).*

- *95% confidence interval.* From the standardised normal distribution, the z statistic associated with 95% probability (two-tailed) is 1.96. Therefore the 95% confidence interval is calculated as follows:

$$\begin{aligned} P_{95\%} &= \bar{X} \pm \frac{z_{95\%} \sigma}{\sqrt{N}} = 10.2 \pm \frac{1.96 \times 1.9}{\sqrt{30}} = 10.2 \pm 0.68 \text{ L} \\ &= 9.52\text{--}10.88 \text{ L} \end{aligned}$$

- *99% confidence interval.* Similarly, the 99% confidence interval is calculated using a z value of 2.58 (corresponding to 99% probability in a two-tailed outcome):

$$\begin{aligned} P_{99\%} &= \bar{X} \pm \frac{z_{99\%} \sigma}{\sqrt{N}} = 10.2 \pm \frac{2.58 \times 1.9}{\sqrt{30}} = 10.2 \pm 0.89 \text{ L} \\ &= 9.31\text{--}11.09 \text{ L} \end{aligned}$$

Having calculated these values, it is important at this point for the reader to fully comprehend the meaning of confidence intervals. In the clinical example described, the 95% confidence interval was 10.2 ± 0.68 L, i.e. 9.52–10.88 L. The observed mean value (10.2 L) is unlikely to be identical to the population mean value. However, the

Confidence intervals for the population mean 139

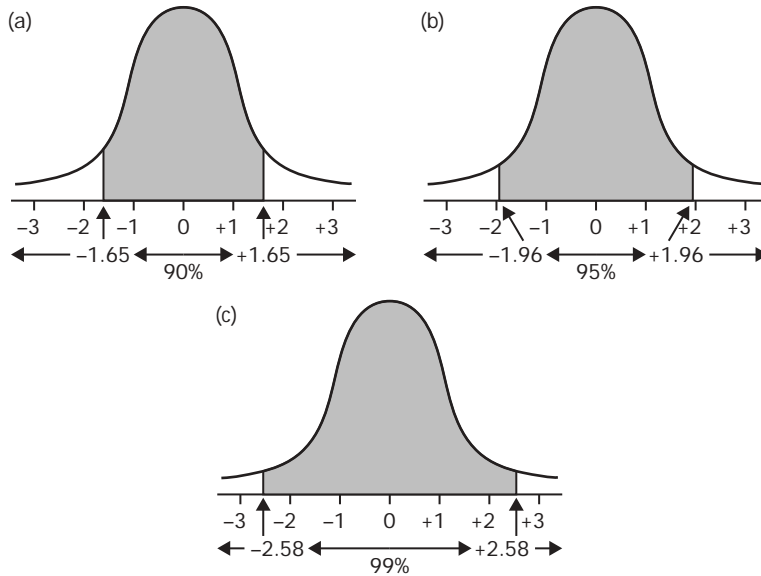
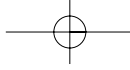


Figure 6.3 Probability functions within the standardised normal distribution.

calculated confidence interval provides an estimation of the reliability of the measured mean. Therefore, we are 95% certain that the true mean will lie within the range defined by the confidence intervals, i.e. 9.52–10.88 L. In other words, if 100 samples were selected and their means and confidence intervals calculated, it is likely that 95 such confidence intervals would contain the true mean. Similarly, in a 99% confidence interval, there is a 99% chance that the true mean will lie within the defined (calculated) range. In general laboratory practice, the mean and confidence intervals of one set of data (sample) are normally calculated, the use of confidence intervals providing sufficient information about the reliability of the mean.

Students often ask why the 99% confidence interval encompasses a greater range of values than the corresponding 95% interval. The answer is quite straightforward and concerns an appreciation of the role of probability in the calculation of confidence intervals. In the calculation of confidence intervals, reference is made to the appropriate probability distribution, the normal distribution in our case. As described previously, the total area under the normal distribution curve is 1 (equal to 100% probability). The area under the curve (and the z value) increases proportionally as the probability of an event increases. Hence the z value associated with 99% probability, i.e. 2.58, is greater than that for the 95% probability (1.96) and accordingly, to ensure 99%



140 Statistical estimation using confidence intervals

probability of an event, a greater spread of z values is required. Thus in the calculation of 99% confidence intervals a greater area under the normal curve is being considered than with 95% confidence intervals and hence the range of values within the 99% confidence interval exceeds that of the lower interval.

The calculation and applications of confidence intervals are further explored in the following examples.

EXAMPLE 6.2 *The aluminium content of 1000 1-L samples of intravenous fluids was analysed and the mean and standard deviation calculated as 50 ± 3 ppm. Calculate the 95% and 99% confidence limits for estimating the true mean aluminium concentration.*

- *95% confidence interval.* The sample size is large (1000), so it may be assumed that the aluminium concentration conforms to normality and therefore, the following equation may be employed to calculate confidence intervals:

$$P_{95\%} = \bar{X} \pm \frac{z_{95\%}\sigma}{\sqrt{N}}$$

where $Z_{95\%}$ denotes the z value corresponding to a probability of 95% (two-tailed), i.e. 1.96, $\sigma = 3$ ppm and $N = 1000$. Thus:

$$P_{95\%} = 50 \pm \frac{1.96 \times 3}{\sqrt{1000}} = 50 \pm 0.19 \text{ ppm}$$

The 95% confidence limits are 49.81–50.19 ppm, i.e. there is a 95% chance that the true population mean will be encompassed within this range.

- *99% confidence interval.* Once more, the following equation may be used:

$$P_{99\%} = \bar{X} \pm \frac{z_{99\%}\sigma}{\sqrt{N}}$$

where $z_{99\%}$ denotes the z value corresponding to a probability of 99% (two-tailed), i.e. 2.58, $\sigma = 3$ ppm and $N = 1000$. Thus:

$$P_{99\%} = 50 \pm \frac{2.58 \times 3}{\sqrt{1000}} = 50 \pm 0.24$$

The 99% confidence limits are 49.76–50.24 ppm, i.e. there is a 99% chance that the true population mean lies within this region.

EXAMPLE 6.3 *100 transdermal patches have been removed from a batch and the stability of the active agent determined at 25 °C. The observed mean (and standard deviation) degradation rate constant for the therapeutic agent was $0.09 \pm 0.01 \text{ day}^{-1}$. Calculate the 80% and 95% confidence intervals of the estimates of the mean, and the confidence that the mean degradation rate of the population is $0.09 \pm 0.0015 \text{ h}^{-1}$.*

Confidence intervals for differences between means 141

- *80% confidence interval*

$$P_{80\%} = \bar{X} \pm \frac{z_{80\%} \sigma}{\sqrt{N}}$$

From the standardised normal distribution, the z value associated with 80% probability is 1.28 (Appendix 1). Therefore

$$P_{80\%} = \bar{X} \pm \frac{z_{80\%} \sigma}{\sqrt{N}} = 0.09 \pm \frac{1.28 \times 0.01}{\sqrt{100}} = 0.09 \pm 0.001 \text{ h}^{-1}$$

The 80% confidence interval for the estimation of the population mean is 0.089–0.091 h^{-1} .

- *95% confidence interval*

$$P_{95\%} = \bar{X} \pm \frac{z_{95\%} \sigma}{\sqrt{N}} = 0.09 \pm \frac{1.96 \times 0.01}{\sqrt{100}} = 0.09 \pm 0.002 \text{ h}^{-1}$$

The 95% confidence interval for the estimation of the population mean is 0.088–0.092 h^{-1} .

- *Confidence that the mean degradation rate of the population is $0.09 \pm 0.0015 \text{ h}^{-1}$.* This question offers a different perspective on the use of confidence intervals. In the previous examples we have defined the probability of the confidence interval and then calculated the interval from this. In this example, the interval has been defined and the confidence of that interval is unknown and requires calculation. Once more the generic equation is employed:

$$P\% = \bar{X} \pm \frac{z_{P\%} \sigma}{\sqrt{N}}$$

The known parameters are inserted into this equation:

$$P\% = 0.09 \pm \frac{z_{P\%} \times 0.01}{\sqrt{100}} = 0.09 \pm (z_{P\%} \times 0.001)$$

$$0.09 \pm 0.0015 = 0.09 \pm 0.001 z_{P\%}$$

Therefore:

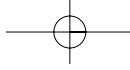
$$0.0015 = 0.001 Z$$

$$1.5 = z_{P\%}$$

The area under the standardised normal curve from $z = 0$ to $z = 1.5$ is 0.4332. Similarly the area under the standardised normal distribution from $z = 0$ to $z = -1.5$ is 0.4332. Therefore, the degree of confidence is 0.8664, i.e. 86.64%.

6.3 Confidence intervals for differences between means

Most frequently, confidence intervals are calculated to provide an estimation of the mean of a population, as described in section 6.2. However, in many cases, it is useful to describe the confidence intervals on the differences between means. Once more, the normal distribution



142 Statistical estimation using confidence intervals

may be employed for this purpose, i.e. it is assumed that the means are derived from a normal distribution. The equation used here is based on the one employed for the calculation of the confidence intervals of a single mean, with two exceptions:

- the difference between the means replaces the mean value
- the standard error of differences between means is used in place of the standard error of the mean (i.e. the square root of the sum of the squares of their standard errors)

The modified equation is:

$$P\% = (\bar{X}_1 - \bar{X}_2) \pm z_{P\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

where \bar{X}_1 and \bar{X}_2 are the means of samples 1 and 2, respectively; s_1^2 and s_2^2 are the variances of samples 1 and 2, respectively; $z_{P\%}$ is the z value relating to the chosen level of probability (e.g. 90%, 95%, 99%), N_1 and N_2 are the sample sizes in samples 1 and 2, respectively and

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

is the standard error of the difference between two independent means.

The use of this equation is illustrated in the following examples.

EXAMPLE 6.4 *A quality control laboratory wishes to examine and compare the mechanical properties of two wound dressings using tensile analysis. Accordingly, the mean (\pm standard deviation) tensile strength of a proprietary wound dressing was examined ($N = 250$) and recorded as 10.35 ± 0.57 MPa. The mean (\pm standard deviation) of a generic wound dressing was 8.99 ± 0.73 MPa ($N = 150$). Calculate the 90% and 95% confidence intervals of the difference between the tensile strengths of the two wound dressings.*

- *90% confidence interval.* From the information given, we know that \bar{X}_1 and \bar{X}_2 are 10.35 and 8.99 MPa, respectively; s_1 and s_2 are 0.57 and 0.73 MPa, respectively; $z_{P\%}$ is the z value relating to the chosen level of probability, i.e. 1.65 for a 90% confidence interval; and N_1 and N_2 are 250 and 150, respectively. Inserting this information into the relevant equation enables us to calculate the 90% confidence interval:

$$\begin{aligned} P_{90\%} &= (\bar{X}_1 - \bar{X}_2) \pm z_{90\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ &= (10.35 - 8.99) \pm 1.65 \sqrt{\frac{(0.57)^2}{250} + \frac{(0.73)^2}{150}} \\ &= 1.36 \pm 0.12 \text{ MPa} \end{aligned}$$

Confidence intervals for differences between means 143

Therefore, the 90% confidence interval for the difference between the mean tensile strength of the two wound dressings is 1.24–1.48 MPa. Therefore, there is a 90% probability that the true mean difference in tensile strengths between the two dressings may be found in the range 1.24–1.48 MPa.

- *95% confidence interval.* The various parameters in the equation below are identical to those described above with the exception of the z value, which now reflects 95% probability, i.e. 1.96:

$$\begin{aligned} P_{95\%} &= (\bar{X}_1 - \bar{X}_2) \pm z_{95\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ &= (10.35 - 8.99) \pm 1.96 \sqrt{\frac{(0.57)^2}{250} + \frac{(0.73)^2}{150}} \\ &= 1.36 \pm 0.14 \text{ MPa} \end{aligned}$$

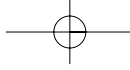
Therefore, the 95% confidence interval for the difference between the mean tensile strength of the two wound dressings is 1.22–1.50 MPa and there is a 95% probability that the true mean difference in tensile strengths between the two dressings may be found in this range.

EXAMPLE 6.5 *A new therapeutic agent has been developed to promote diuresis. In a clinical trial, the diuretic effects of this new agent and a commercial agent were assessed. In this the urine was collected over a 12-h period after administration of a single tablet. The mean (\pm standard deviation) volume of urine collected in the group of patients ($N = 65$) who received the new therapeutic agent was 48.8 ± 9.1 L, whereas in the control group (who received the commercially available preparation) the volume was 37.9 ± 4.6 L ($N = 95$). Calculate the 95% and 99% confidence intervals for the difference in urine volume induced by the two therapeutic agents.*

- *95% confidence interval.* Once more, the equation described at the start of section 6.3 is employed to calculate the required confidence interval, using $z_{95\%} = 1.96$:

$$\begin{aligned} P_{95\%} &= (\bar{X}_1 - \bar{X}_2) \pm z_{95\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ &= (48.8 - 37.9) \pm 1.96 \sqrt{\frac{(9.1)^2}{65} + \frac{(4.6)^2}{95}} \\ &= 10.9 \pm 2.40 \text{ L} \end{aligned}$$

Therefore, the 95% confidence interval for the difference in mean urine volume is 8.50–13.3 L.



144 Statistical estimation using confidence intervals

- *99% confidence interval.* Employing $z_{99\%} = 2.58$, the above equation may be written as:

$$\begin{aligned} P_{99\%} &= (\bar{X}_1 - \bar{X}_2) \pm z_{99\%} \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \\ &= (48.8 - 37.9) \pm 2.58 \sqrt{\frac{(9.1)^2}{65} + \frac{(4.6)^2}{95}} \\ &= 10.9 \pm 3.16 \text{ L} \end{aligned}$$

The confidence intervals for the differences between means is interpreted in the same way as other confidence intervals. Therefore, in this example, there is a 99% probability that the true difference between the two diuretics will fall within the range 7.74–14.06 L.

6.4 Confidence intervals for standard deviations

Determination of the confidence intervals for the standard deviation is most frequently employed to examine the variability of data, e.g. whenever there is unexpectedly high variation in a sample. For this purpose, the χ^2 distribution is used to calculate confidence intervals for the population standard deviation. From Appendix 3, χ^2 values may be obtained that provide information concerning the areas under the probability curve. Hence, to calculate the 95% confidence interval of the population standard deviation, the region under the χ^2 distribution that equates to 95% is found between $\chi_{0.025}^2$ and $\chi_{0.975}^2$. Therefore, 2.5% of observations lie both below $\chi_{0.025}^2$ and above $\chi_{0.975}^2$. With this in mind, the 95% confidence interval is calculated using the following equation:

$$\chi_{0.025}^2 < \frac{(N-1)s^2}{\sigma^2} < \chi_{0.975}^2$$

where $N-1$ is the number of degrees of freedom, i.e. the sample size minus 1, s is the sample standard deviation, σ is the population standard deviation, and $\chi_{0.025}^2$ and $\chi_{0.975}^2$ are the χ^2 statistics relating to probabilities of 2.5% and 97.5%, respectively, for $n-1$ degrees of freedom.

This equation may be more conveniently rearranged to give

$$\frac{s\sqrt{(N-1)}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.025}}$$

An example of the use of the χ^2 distribution to calculate the confidence

intervals associated with the population standard deviation is provided next.

EXAMPLE 6.6 *Before the start of a clinical trial, the weights of 20 healthy male volunteers (aged 21–30 years) were measured and the mean and standard deviation recorded as 76.1 ± 9.4 kg. Calculate the 90%, 95% and 99% confidence intervals of the population standard deviation.*

- *90% confidence intervals.* Using the equation described above,

$$\frac{s\sqrt{(N-1)}}{\chi_{0.95}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.05}}$$

where N is the number of volunteers (20); s is the sample standard deviation, i.e. 9.4 kg; $\chi_{0.95}$ for 19 degrees of freedom (i.e. 20 volunteers – 1) is $\sqrt{\chi_{0.95}^2}$, i.e. $\sqrt{30.1} = 5.49$; $\chi_{0.05}$ for 19 degrees of freedom is $\sqrt{\chi_{0.05}^2}$, i.e. $\sqrt{10.1} = 3.18$. Therefore:

$$\frac{s\sqrt{(N-1)}}{\chi_{0.95}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.05}} = \frac{9.4\sqrt{19}}{5.49} < \sigma < \frac{9.4\sqrt{19}}{3.18}$$

$$7.46 < \sigma < 12.88 \text{ kg}$$

Therefore, there is a 90% chance that the true population standard deviation (weights of healthy men aged 21–30 years) will reside within the above limits.

- *95% confidence intervals.* Using the appropriate equation:

$$\frac{s\sqrt{(N-1)}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.025}}$$

where N and s are as before; $\chi_{0.975}$ for 19 degrees of freedom is $\sqrt{\chi_{0.975}^2}$, i.e. $\sqrt{32.9} = 5.74$; and $\chi_{0.025}$ for 19 degrees of freedom is $\sqrt{\chi_{0.025}^2}$, i.e. $\sqrt{8.91} = 2.98$. Therefore:

$$\frac{s\sqrt{(N-1)}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.025}} = \frac{9.4\sqrt{19}}{5.74} < \sigma < \frac{9.4\sqrt{19}}{2.98}$$

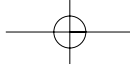
$$7.14 < \sigma < 13.75 \text{ kg}$$

and there is a 95% chance that the true population standard deviation falls within the above limits.

- *99% confidence intervals.* As before:

$$\frac{s\sqrt{(N-1)}}{\chi_{0.995}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.005}}$$

where N and s is the sample standard deviation; $\chi_{0.995}$ for 19 degrees of freedom is $\sqrt{\chi_{0.995}^2}$, i.e. $\sqrt{38.6} = 6.21$; and $\chi_{0.005}$ for 19 degrees of freedom is $\sqrt{\chi_{0.005}^2}$, i.e.



146 Statistical estimation using confidence intervals

$\sqrt{6.84} = 2.62$. Therefore:

$$\frac{s\sqrt{(N-1)}}{\chi_{0.995}} < \sigma < \frac{s\sqrt{(N-1)}}{\chi_{0.005}} = \frac{9.4\sqrt{19}}{6.21} < \sigma < \frac{9.4\sqrt{19}}{2.62}$$

$$6.60 < \sigma < 15.64 \text{ Kg}$$

and there is a 99% chance that the true population standard deviation (weights of healthy men aged 21–30 years) will lie between 6.60 and 15.64 kg.

It is worth pointing out that the confidence limits of the true (population) standard deviation from the sample standard deviation may be estimated using the normal distribution. In this it is assumed that the population from which the sample data has been derived is normally distributed. When the sample size is large, i.e. greater than 30, the mathematical function $\sqrt{2\chi^2} - \sqrt{2(N-1)} - 1 = z_p$ is appropriate and approximates to a normal distribution. Therefore, by rearrangement of the equation to produce $\chi^2 = \frac{1}{2}(z_p + \sqrt{2(N-1)} - 1)^2$, the χ^2 statistic may be calculated using the normal distribution. From this, the confidence intervals of the true standard deviation may be calculated from the sample standard deviation using the χ^2 statistic, as described previously.

Alternatively, the standard deviation may be directly calculated from the normal distribution using the following equation:

$$P\% = s \pm z_p \frac{s}{\sqrt{2N}}$$

where s is the sample standard deviation, N is the number of samples and z_p is the z value corresponding to the desired level of probability.

The uses of the different methods for calculating the confidence intervals of the standard deviation are described in the next example.

EXAMPLE 6.7 *A solid dosage form containing 100 mg of a therapeutic agent has been manufactured and stored for 1 year at 37 °C and 75% relative humidity. Analysis of the concentration of therapeutic agent in 101 tablets following this period of storage revealed that the mean (\pm standard deviation) drug content was 94.13 ± 4.56 mg. Calculate the 95% confidence interval of the population standard deviation of the concentration of therapeutic agent within the tablets.*

- Use of the χ^2 distribution

$$\frac{s\sqrt{(N)}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{(N)}}{\chi_{0.025}} = \frac{4.56\sqrt{101}}{\sqrt{129.6}} < \sigma < \frac{4.56\sqrt{101}}{\sqrt{74.2}}$$

$$4.03 < \sigma < 5.32 \text{ mg}$$

Therefore, the 95% confidence interval of the population standard deviation

Confidence intervals for proportions 147

of drug content in solid dosage forms is 4.03–5.32 mg (Note: N is used in place of $N - 1$ as the sample size is large).

- *Use of the normal distribution.* As described above,

$$\chi^2 = \frac{1}{2} (z_p + \sqrt{2(N-1) - 1})^2$$

Therefore:

$$\begin{aligned} \chi_{0.975}^2 &= \frac{1}{2} (z_p + \sqrt{2(N-1) - 1})^2 = \frac{1}{2} (1.96 + \sqrt{2(101) - 1})^2 \\ &= 130.21 \end{aligned}$$

$$\begin{aligned} \chi_{0.025}^2 &= \frac{1}{2} (z_p + \sqrt{2(N-1) - 1})^2 = \frac{1}{2} (-1.96 + \sqrt{2(101) - 1})^2 \\ &= 74.63 \end{aligned}$$

Note that the values of χ^2 calculated using the normal distribution are numerically similar to the values derived from the χ^2 distribution (Appendix 3).

To calculate the confidence intervals as requested, these χ^2 values are used, as described previously:

$$\begin{aligned} \frac{s\sqrt{N}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{N}}{\chi_{0.025}} = \frac{4.56\sqrt{101}}{\sqrt{130.21}} < \sigma < \frac{4.56\sqrt{101}}{\sqrt{74.63}} \\ 4.02 < \sigma < 5.30 \text{ mg} \end{aligned}$$

Results calculated using either method are similar.

6.5 Confidence intervals for proportions

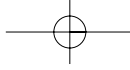
Confidence intervals of proportions may also be calculated using the approximation to the normal distribution and the sample standard deviation. The confidence interval for the population proportion may be calculated using the following equation:

$$P\% = p \pm z \sqrt{\frac{pq}{N}}$$

where p is the proportion of successes, q is the proportion of failures ($1 - p$), N is the sample size and z denotes the z value relating to a defined probability level.

The use of this equation is described in the following example.

EXAMPLE 6.8 *A healthcare company that specialises in the sterile packaging of disposable gloves has decided to perform a quality audit of the resistance of a batch of packaged disposable gloves to the ingress of unsterile air. Out of a batch of 2500 packages, 200 were removed and*



148 Statistical estimation using confidence intervals

their sterility examined using the British Pharmacopoeia sterility tests. The results showed that 35 packages failed the sterility tests. On the basis of these observations, calculate the 95% confidence limits for the proportion of defective gloves in the batch (population).

This problem may be solved using the equation described at the beginning of this section:

$$P_{95\%} = p \pm z_{95\%} \sqrt{\frac{pq}{N}}$$

where $p = 35/200 = 0.175$, $q = 1 - 0.175 = 0.825$, $N = 200$ and $z = 1.96$ (for a 95% probability, two-tailed). Thus:

$$\begin{aligned} P_{95\%} &= 0.175 \pm 1.96 \sqrt{\frac{(0.175) \times (0.825)}{200}} \\ &= 0.175 \pm 0.053 \end{aligned}$$

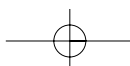
Therefore, the 95% confidence interval for the population proportion is 0.122–0.228, i.e. there is a 95% chance that the true proportion of defective gloves will fall within this range.

Remember that the standard error of the proportion, $\sqrt{(pq/N)}$, may only be used in the calculation of the confidence intervals of a proportion if the sample size is sufficiently large for the normal approximation to be valid. As outlined in Chapter 4, the validity of this assumption may be rapidly estimated by calculating the product of the probability of the event and the sample size. Normality is commonly observed whenever this product exceeds 6.

6.6 Confidence intervals for differences between proportions

The confidence intervals for differences in proportions may be estimated using a similar approach to that described in section 6.3, assuming that the proportions are obtained from a normal distribution. The main difference between the approach outlined in section 6.3 (for the calculation of confidence intervals for the difference between two means) and that required for the calculation of the confidence intervals for the differences in proportions is the mathematical definition of the standard error of the difference between means and proportions. Referring to section 6.3, it may be recalled that the standard error of the differences between two independent means is calculated using the following formula:

$$SE_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$



Confidence intervals for differences between proportions 149

Conversely, the standard error of the difference between proportions is mathematically defined as:

$$SE_{(p_1-p_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\left(\frac{p_1q_1}{n_1}\right) + \left(\frac{p_2q_2}{n_2}\right)}$$

Applying the same mathematical approach as that described in section 6.3, the confidence intervals for the difference between two proportions may be defined as:

$$\begin{aligned} P\% &= (p_1 - p_2) \pm (z_{\%} \times SE_{\text{diff}}) \\ &= (p_1 - p_2) \pm \left(z_{\%} \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right) \end{aligned}$$

where p_1 and p_2 are the two proportions under examination, $z_{\%}$ is the z value corresponding to the accepted level of probability and SE_{diff} is the standard error of the difference between proportions.

The use (and calculation) of the confidence intervals relating to the difference between two proportions is explained in the following example.

EXAMPLE 6.9 *The relative efficacies of two oral antibiotics for the treatment of bronchial pneumonia have been assessed in a clinical trial. Subjects presenting with bronchial pneumonia at a hospital were divided into two groups, each of which received one or other of the antibiotics. The results from the clinical trial are shown in Table 6.1. Calculate the 95% confidence interval for the difference between the proportions of cured patients in the two groups.*

The equation used to calculate the 95% confidence interval for the difference between the two proportions is

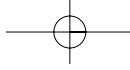
$$P_{95\%} = (p_1 - p_2) \pm \left(z_{95\%} \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$

where

$$p_1 = \frac{575}{(575 + 45)} = 0.927$$

Table 6.1 Efficacy of two antibiotics for the treatment of bronchial pneumonia

Cure	Number of patients	
	Antibiotic 1	Antibiotic 2
Yes	575	425
No	45	200



150 Statistical estimation using confidence intervals

i.e. the proportion of cures associated with antibiotic 1;

$$p_2 = \frac{425}{(425 + 200)} = 0.680$$

i.e. the proportion of cures associated with antibiotic 2; $z_{95\%}$ is 1.96 (95% probability, two-tailed outcome – see Appendix 1); $n_1 = (575 + 45) = 620$; and $n_2 = (424 + 200) = 625$. When these values are substituted into the above equation, the confidence interval for the differences between the proportions of cures resulting from treatment with the two antibiotics may be calculated:

$$\begin{aligned} P_{95\%} &= (0.927 - 0.680) \\ &\pm \left(1.96 \times \sqrt{\frac{0.927(1 - 0.927)}{620} + \frac{0.680(1 - 0.680)}{625}} \right) \\ P_{95\%} &= 0.247 \pm 0.042 \end{aligned}$$

Two further points should be mentioned at this stage.

- No units are associated with the calculated confidence interval of a difference in proportions, because proportions are, by definition, ratios. This is also true of the confidence intervals of a proportion.
- In this example, it has been assumed that the sample size is sufficiently large for the approximation to the normal distribution to apply. This may be easily confirmed by multiplying the number of observations in a particular group (e.g. n_1) by the probability (p_1). Values in excess of 6 confirm the validity of this assumption.

The next example also describes the calculation of the confidence intervals associated with the difference between proportions.

EXAMPLE 6.10 *The effects of treatment of blastospores of Candida albicans (stationary phase) with either polyhexamethylenebiguanide (PHMB, 0.1% v/v) or sterile water for 30 min on the number of blastospores adherent to buccal epithelial cells in vitro were examined. The results, expressed as the number of buccal epithelial cells with and without adherent blastospores, are presented in Table 6.2. Calculate the 95% confidence interval for the difference between the proportion of epithelial cells without adherent blastospores following treatment with either PHMB or sterile water.*

The relevant equation to calculate the 95% confidence intervals for the difference between the two proportions is as before:

$$P_{95\%} = (p_1 - p_2) \pm \left(z_{95\%} \times \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \right)$$

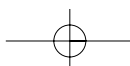


Table 6.2 Effects of treatment of blastospores of *Candida albicans* with either PHMB (0.1% v/v) or sterile water on the subsequent adherence to buccal epithelial cells *in vitro*

Outcome	Treatment	
	PHMB	Water
Cells without adherent blastospores	103	54
Cells with adherent blastospores	48	91

where $p_1 = 103/(103 + 48) = 0.682$, i.e. the proportion of epithelial cells devoid of adherent blastospores following treatment with PHMB; $p_2 = 54/(54 + 91) = 0.372$, i.e. the proportion of epithelial cells devoid of adherent blastospores following treatment with sterile water; $z_{95\%}$ is 1.96 (95% probability, two-tailed outcome, Appendix 1); $n_1 = (103 + 48) = 151$; and $n_2 = (54 + 91) = 145$. Substitution of these values into the above equation and computation yields the 95% confidence interval, i.e. $0.31 \pm 0.11 = 0.20-0.42$.

It may sometimes be difficult to fully appreciate the meaning of the calculated confidence interval for the difference between two proportions. Therefore, in medical research, it is commonplace to express the difference between proportions as ratios. In Example 6.10 it may be observed that treatment of blastospores with PHMB is more likely to inhibit adherence to buccal epithelial cells by a factor of $0.682/0.372 = 1.83$.

An alternative expression that is commonly employed to relate two ratios is the *odds ratio*. The odds of an event may be defined as the ratio of the probability of the occurrence of an event to that of the non-occurrence of the event. Reverting to Example 6.10, the odds of observing buccal epithelial cells devoid of adherent blastospores following treatment with PHMB is $(0.682/(1 - 0.682)) = 2.13$. Similarly, the odds of observing a buccal epithelial cell devoid of adherent blastospores following treatment with sterile water are $(0.373/(1 - 0.372))$. The ratio of these two odds is termed the *odds ratio* or *relative risk*, which for the example described, is $2.13/0.59 = 3.61$.

6.7 Confidence intervals and the t distribution

As highlighted in the introduction to this chapter, the z statistic is employed to calculate confidence intervals whenever the assumptions of the normal distribution are valid. One requirement is a large sample size. However, the inquisitive reader may ask, how are confidence intervals calculated for sample data? Under these circumstances, the t distribution is employed to calculate confidence intervals, using a similar

152 Statistical estimation using confidence intervals

strategy to that used to calculate the confidence intervals of the true mean for large samples. Thus:

$$P\% = \bar{X} \pm \frac{tS}{\sqrt{N}}$$

where $P\%$ is the chosen probability (usually 90%, 95% or 99%). \bar{X} is the sample mean, t is the t statistic corresponding to the appropriate probability level and degree of freedom and S is the sample standard deviation

The major difference between the above equation and the equation for the calculation of the confidence interval using the normal distribution is the nature of the test statistic. In the normal distribution the value of the z statistic depends only on the chosen probability level and is independent of sample size (the use of the normal distribution implies that the sample size is large). The t statistic, however, is dependent on two parameters: the chosen probability and, additionally, the number of degrees of freedom, which is itself dependent on the size of the sample ($df = N - 1$). It is important that the reader fully understands the nature of the differences between these two methods of calculating the confidence intervals of the population mean.

The use of the t distribution to calculate confidence intervals of the population mean is illustrated in the following examples.

EXAMPLE 6.11 *Table 6.3 shows the concentration of chlorhexidine (mg) in 15 1-mL samples that have been removed from a stock solution (5 L) and assayed by ultraviolet spectroscopy. Calculate the 95% confidence interval of the population mean.*

First, calculate the mean and standard deviation, i.e. 0.207 ± 0.01 mg/mL. As the sample size is small ($N = 15$), the confidence intervals must be calculated using the t distribution, i.e.

$$P_{95\%} = \bar{X} \pm \frac{tS}{\sqrt{N}}$$

The mean and standard deviation have been calculated using the sample data. The final unknown in the equation is therefore the t statistic. Two

Table 6.3 Concentration of chlorhexidine (mg/mL) in 15 samples removed from a stock solution

0.201	0.205	0.203	0.205	0.200
0.212	0.210	0.211	0.209	0.199
0.209	0.211	0.215	0.216	0.205

points should be considered before selecting this value from Appendix 2: the probability level (95%, two-tailed) and the number of degrees of freedom. The latter is calculated by subtracting 1 from the sample size, i.e. $15 - 1 = 14$ df . The t statistic corresponding to 14 df and 95% probability is 2.14 (Appendix 2).

The confidence intervals may then be calculated:

$$P_{95\%} = \bar{X} \pm \frac{tS}{\sqrt{N}} = 0.207 \pm \frac{2.14 \times 0.01}{\sqrt{15}} = 0.207 \pm 0.006 \text{ mg/mL}$$

The confidence interval is therefore 0.201–0.213 mg/mL.

A second example of this calculation is provided below.

EXAMPLE 6.12 *A chemist has developed a synthetic route for the preparation of a compound with antihistamine properties. A 10-kg batch has been prepared, from which 10 100-mg samples have been removed and their purity determined using an analytical method. The results are shown in Table 6.4. Calculate the 95% and 99% confidence intervals of the true purity of the compound.*

The first stage of the calculation involves the calculation of the mean and standard deviation of the data described in Table 6.4, i.e. $96.37 \pm 2.56\%$.

- 95% confidence interval

$$P_{95\%} = \bar{X} \pm \frac{tS}{\sqrt{N}}$$

where the mean and standard deviation have been determined as 96.37 and 2.56% and N , the sample size, is 10. The relevant value of the t statistic is derived from the table of the critical values of the t distribution (Appendix 2) with knowledge of the number of degrees of freedom ($10 - 1 = 9$) and the chosen level of probability (95%, two-tailed outcome). The t statistic is therefore 2.26.

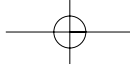
Consequently, the 95% confidence interval is

$$P_{95\%} = \bar{X} \pm \frac{tS}{\sqrt{N}} = 96.37 \pm \frac{(2.26 \times 2.56)}{\sqrt{10}} = 96.37 \pm 1.83\%$$

- 99% confidence interval. As the probability associated with the confidence interval has changed, the t statistic must be modified. From Appendix 2, the t

Table 6.4 Purities (%) of 10 aliquots of an antihistamine removed from a single batch of product

91.2	94.5	97.5	95.6	98.1
96.5	99.8	98.9	97.4	94.2



154 Statistical estimation using confidence intervals

statistic associated with 9 *df*, two-tailed outcome and 99% probability is 3.25. This knowledge allows calculation of the 99% confidence interval:

$$P_{95\%} = \bar{X} \pm \frac{tS}{\sqrt{N}} = 96.37 \pm \frac{(3.25 \times 2.56)}{\sqrt{10}} = 96.37 \pm 2.63\%$$

6.8 Conclusions

In this chapter the theory and applications of confidence intervals have been outlined. It is important to understand the concept of confidence intervals, because they are frequently employed to provide information about the magnitude of the true statistic from sample data. In particular, the confidence intervals associated with several descriptive statistics have been described, namely the mean, standard deviation, proportions and differences between these parameters. Furthermore, the uses of both the *z* and *t* statistics in the calculation of confidence intervals have been illustrated, and indeed this is the first example of the use of the *t* distribution for the statistical characterisation of small samples. Further examples of this application will be described in subsequent chapters.

