7

Process experimentation with a single factor

Experiment, and it will lead you to the light. (Cole Porter, 'Experiment' from *Nymph Errant*, 1933)

Overview

This chapter deals with statistical tools that are relevant to the improve phase of Six Sigma projects. Having made changes to a process, how do we formally assess data from the modified process for evidence of improvement? Statistical inference techniques may be use to address questions such as:

- Has the change of supplier of the lens coating fluid led to a reduction in the proportion of nonconforming lenses?
- Has the appointment of specialist nurses, empowered to administer thrombolytic treatment to acute myocardial infarction patients on admission to hospital, led to a reduction in the mean door to needle time?

Estimation techniques provide point estimates of the population proportion and the population mean, respectively, for the modified processes in the above scenarios, i.e. of population parameters that are of interest. Estimation techniques provide intervals in which we can have confidence that the values of the parameters are located.

Some of the techniques are based on the normal distribution while others make no such assumption. Minitab is well equipped to deal with both classes of technique.

Six Sigma Quality Improvement with Minitab, Second Edition. G. Robin Henderson.

7.1 Fundamentals of hypothesis testing

In Chapter 1, the description of the improve phase in a Six Sigma project given by Roger Hoerl included the statement 'determine how to intervene in the process to significantly reduce the defect levels' (Hoerl, 1998, p. 36). Process experimentation may be thought of as a formal approach to the question of determining how to intervene in the process in order to improve it. Wheeler (1993, p. 21) writes:

- Before one can improve any system one must listen to the voice of the system (the voice of the process).
- Then one must understand how the inputs affect the outputs of the system.
- Finally, one must be able to change the inputs (and possibly the system) in order to achieve the desired results.
- This will require sustained effort, constancy of purpose, and an environment where continual improvement is the operating philosophy.

Wheeler (2007, p. 7) distinguishes between *observational* and *experimental* studies. The routine collection of data in order to monitor, using control charts, a process running under normal conditions is an example of an observational study. The collection of data on a process when it is being run under special conditions, with a view to learning how the process might be improved, is an example of an experimental study. He states: 'when we analyze experimental data we are looking for differences that we have paid good money to create and that we believe are contained within the data'.

Consider the process of administering thrombolytic treatment to acute myocardial infarction patients at a hospital. Records show that the process has been behaving in a stable, predictable manner, with door to needle time (DTN) being adequately modelled by the normal distribution with mean 19 minutes and standard deviation 6 minutes. An experiment was conducted over a period of 1 month during which one of a team of specialist nurses, empowered to administer the thrombylotic treatment, was on duty at all times in the accident and emergency department of the hospital. The DTN times for the 25 patients treated during the experimental period are shown in Table 7.1 and are available in the worksheet DTNTime.MTW. The mean DTN for acute myocardial infarction patients during the experimental period was 16.28 minutes. Does this sample mean represent a 'real' improvement to the process in the sense of a reduction of the population mean DTN for acute myocardial infarction patients from 19 minutes to a new, lower population mean? Can we infer from the data that the regular deployment of the specialist nurses would ensure process improvement?

Table 7.1 DTN (minutes) during experimental period.

26	7	24	3	12
17	24	4	5	16
16	22	14	15	14
19	21	18	14	20
29	20	17	9	21

The various steps involved in performing the appropriate statistical inference will now be detailed under the headings Hypotheses, Experimentation, Statistical model and Conclusion.

Hypotheses. Denoting the *null hypothesis* by H_0 and the *alternative hypothesis* by H_1, our hypotheses are:

$$H_0 : \mu = 19, \quad H_1 : \mu < 19.$$

The null hypothesis represents 'no change' – were the introduction of the specialist nurses to have no impact on DTN, the population mean would remain at 19 minutes. Thus μ represents the population mean DTN with the specialist nurses deployed. The alternative hypothesis represents what might be referred to as the *experimental hypothesis* – the objective of the experiment is to determine whether or not there is evidence that the specialist nurse input improves the process by leading to a mean DTN time which is less than 19 minutes.

Experimentation. During an experimental period of 1 month, with specialist nurse input, the mean DTN for the 25 patients treated was 16.28 minutes.

Statistical model. Three assumptions are made:

1. The variability of DTN time is assumed to be unaffected by the process change, i.e. it is assumed that the standard deviation continues to be $\sigma = 6$ minutes.

2. The null hypothesis is assumed true, i.e. it is assumed that the process mean continues to be $\mu = 19$ minutes. (This is analogous to the situation whereby the defendant in a trial in a court of law is considered innocent until there is evidence to the contrary.)

3. The sample of 25 DTNs, obtained during the experimental period, is regarded as a random sample from the population of normally distributed times with mean 19 and standard deviation 6 minutes.

The statistical model is detailed in Box 7.1.

The question asked of the statistical model is ' What is the probability of observing a sample mean for 25 patients which is 16.28 minute or less?' Minitab readily provides the answer using **Calc** > **Probability Distributions** > **Normal** to obtain the Session window output in Panel 7.1. **Cumulative probability** must be selected, with **Mean:** 19, **Standard deviation:** 1.2 and **Input constant:** 16.28 specified. Thus, if the null hypothesis was true, i.e. if the deployment of the specialist nurses had no impact on mean DTN, then the probability of observing a mean for a sample of 25 patients as low as 16.28, or lower, would be 0.012 (to three decimal places). This probability of 0.012 is the *P*-value for testing the hypotheses specified above.

Conclusion. It is conventional in applied statistics to state that a *P*-value less than 0.05 provides evidence for rejection of the null hypothesis, in favour of the alternative hypothesis, at

Door to needle time, Y, is normally distributed with mean 19 and standard deviation 6, i.e. $Y \sim N(19, 6^2)$. Mean DTN for samples of $n = 25$ patients, $\bar{Y}$, will be normally distrinbuted with mean $\mu_{\bar{Y}} = \mu = 19$ and standard deviation $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 6/\sqrt{25} = 1.2$, i.e. $\bar{Y} \sim N(19, 1.2^2)$.

Box 7.1 Statistical model.

Cumulative Distribution Function

```
Normal with mean = 19 and standard deviation = 1.2

    x  P( X <= x )
16.28    0.0117053
```

Panel 7.1 Probability that sample mean is 16.28 minutes or less.

the 5% level of significance. A *P*-value less than 0.01 would be said to provide evidence for rejection of the null hypothesis, in favour of the alternative hypothesis, at the 1% level of significance. (The value 0.001, corresponding to the 0.1% level of significance, is also widely used and 0.1, corresponding to the 10% level of significance is sometimes used.) In addition to the highly technical statement that 'the *P*-value of 0.012 provides evidence for rejection of the null hypothesis, in favour of the alternative hypothesis, at the 5% level of significance', it is important to state that 'the experiment provides evidence that mean DTN is significantly reduced through the deployment of the specialist nurses' and that 'a point estimate of the new mean DTN is a little over 16 minutes'. (It should be noted that although the result of the experiment is *statistically* significant, a reduction of the mean DTN of around 3 minutes might not be of any *practical* significance from a medical point of view. Statistical significance does not equate to practical significance.)

From now on the use of the word 'evidence' will imply that the evidence is convincing, where the word 'convincing' can be further qualified by the significance level. In teaching this topic the author has taught his students to think in terms of a *P*-value less than 0.05 providing evidence for rejection of the null hypothesis, a *P*-value less than 0.01 providing strong evidence and a *P*-value less than 0.001 providing very strong evidence. A *P*-value less than 0.1 might also be regarded as providing slight evidence for rejection of the null hypothesis.

Of course the mean DTN might have remained at 19 minutes with the introduction of the specialist nurses and the experimenters might have been unlucky enough to obtain a sample of times with a low enough mean to provide evidence, in the sense discussed above, of a reduction in the population mean time. Two types of error can occur in the performance of a test of hypotheses as indicated in Table 7.2.

The probability of a Type I error is denoted by the Greek letter α (alpha) and is the significance level of the test. Thus, if one decides to perform a test of hypotheses at the 5% level of significance, $\alpha = 0.05$ and there is a probability of 0.05 that the null hypothesis will be rejected when it is in fact true. In the case of the DTN scenario this means that, were the

Table 7.2 Possible errors in testing hypotheses.

Errors possible in testing hypotheses		True state	
		H_0 true	H_0 false
Conclusion reached	Accept H_0	Correct decision	Type II error probability $= \beta$
	Reject H_0	Type I error probability $= \alpha$	Correct decision

```
Inverse Cumulative Distribution Function

Normal with mean = 19 and standard deviation = 1.2

P( X <= x )         x
       0.05   17.0262
```

Panel 7.2 Determining the cut-off mean DTN.

introduction of the specialist nurses to have no impact whatsoever on the mean of 19 minutes, there is a probability of 0.05 that the conclusion would be the erroneous one that there was evidence of a decrease. The lower the significance level selected then the lower the risk of committing a Type I error.

Having decided, say, on a significance level of $\alpha = 0.05$ for the DTN experiment, one can use **Calc** > **Probability Distributions** > **Normal...** to complete the final stage of reaching a conclusion in a different way. The statistical model indicates that the means of samples of 25 times follow the $N(19,1.2^2)$ distribution. Panel 7.2 displays the Session window output obtained using the **Inverse cumulative probability** function for this normal distribution via **Calc** > **Probability Distributions** > **Normal...** and specifying **Input constant:** 0.05. Thus the cut-off between acceptance and rejection of the null hypothesis occurs at the value 17.026 2 for the sample mean. The conclusion would therefore be:

- Do not reject H_0 if the sample mean is greater than 17.026 2.
- Reject H_0 if the sample mean is less than or equal to 17.026 2.

From the data it was established that the sample mean was 16.28. Since this is less than 17.026 2 the conclusion reached was to reject the null hypothesis at the significance level of $\alpha = 0.05$.

A graphical representation of the test is displayed in Figure 7.1. The use of the 5% significance level corresponds to an area of 0.05 in the shaded left-hand tail of the normal curve

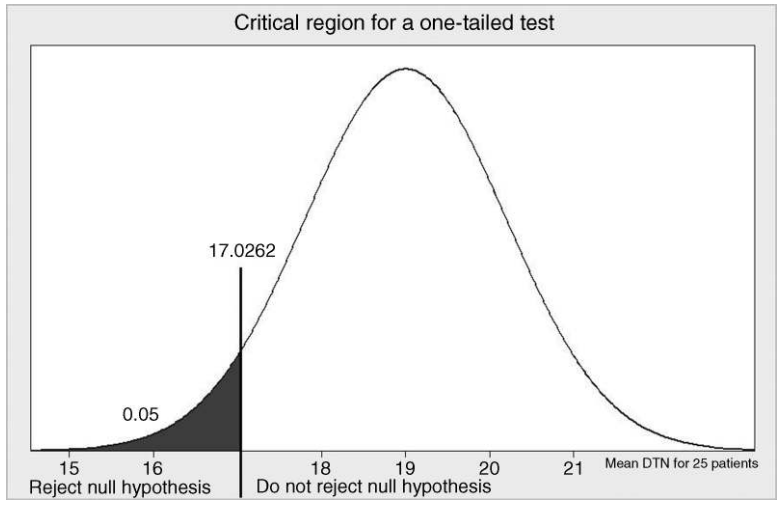


Figure 7.1 Critical region for the statistical test.

```
Cumulative Distribution Function

Normal with mean = 16 and standard deviation = 1.2

      x  P( X <= x )
17.0262     0.803771
```

Panel 7.3 Probability of rejecting null hypothesis when new population mean is 16.

that specifies the statistical model for the distribution of sample means in this case. As only one tail of the distribution is involved, this test may be referred to as a *one-tailed test*. Values of the sample mean DTN less than 17.026 2 comprise the critical region for the test.

Suppose now that the deployment of the specialist nurses had actually led to a reduction in the mean of the population of DTNs from 19 to 16 minutes. What is the probability that, once data are available for a random sample of 25 patients, the conclusion reached – to reject H_0 in favour of H_1 – will be the correct one? We require the probability of observing a sample mean of 17.0262 or less when the population mean is actually 16. Again **Calc** > **Probability Distributions** > **Normal...** provides the answer – see Panel 7.3.

Thus there is a probability of 0.80 (to two decimal places) of the test of hypotheses providing evidence of a three-minute reduction in the population mean DTN. This probability is the power of the statistical test to detect the change from a population mean of 19 minutes to a population mean of 16 minutes. The other side of the coin is that there is probability $1 - 0.80 = 0.20$ of the experiment failing to provide evidence of the reduction, i.e. there is probability $\beta = 0.20$ of committing a Type II error with a test based on a sample of 25 patients when the mean actually drops from 19 to 16 minutes. (The Greek letter β is beta.)

Figure 7.2 show a plot of the power of the test against the population mean after the process change. The greater the drop in the population mean the more likely it is that the test will provide evidence of the change, i.e. the more powerful is the test. For example, were the mean to drop by 1 minute, so that the population mean after the process change became 18 minutes,

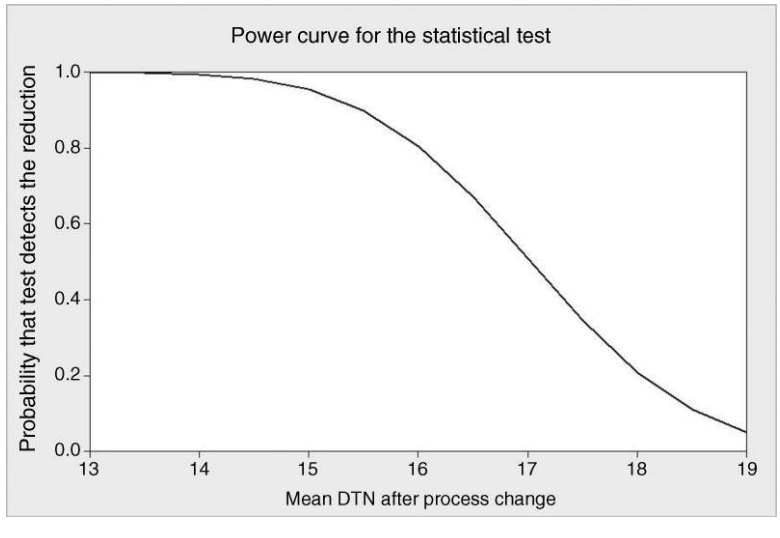


Figure 7.2 Power curve for the statistical test.

then the probability is 0.21 that the conclusion would be to reject the null hypothesis. On the other hand, were the mean to drop by 5 minutes, so that the population mean after the process change became 14 minutes, then the probability is 0.99 that the conclusion would be to reject the null hypothesis. The power of a test may be increased through use of a larger sample.

Now consider the lens coating process referred to in the previous chapter. Records show that prior to the change of coating fluid supplier the process was yielding 4.5% nonconforming lenses. In an experimental run with the new coating fluid there were 80 nonconforming lenses in a batch of 2400, i.e. 3.3% nonconforming. Can we infer from these data that there has been process improvement?

Hypotheses. In this case, using p to represent the population proportion of nonconforming lenses,

$$H_0 : p = 0.045, \quad H_1 : p < 0.045.$$

The null hypothesis represents 'no change' – were the switch to a new supplier of the coating fluid to have no impact on nonconformance, the proportion of nonconforming lenses would remain at 0.045. The alternative hypothesis represents the experimental hypothesis – the objective of the experiment is to determine whether or not there is evidence that the change of supplier improves the process by leading to a reduction in the proportion of nonconforming lenses.

Experimentation. From an experimental run, during which 2400 lenses were processed, 80 were found to be nonconforming.

Statistical model. Three assumptions are made:

1. The null hypothesis is assumed to be true, i.e. that the proportion of nonconforming lenses is assumed to be unaffected by the process change and remains at 0.045.
2. The conditions for the binomial distribution apply, i.e. that there is constant probability of 0.045 that a lens is nonconforming and that the status of each lens is independent of that of all other lenses.
3. The set of 2400 lenses, manufactured using coating fluid from the new supplier, may be considered as a random sample from a population of lenses in which the proportion 0.045 is nonconforming.

The statistical model is detailed in Box 7.2.

The question asked of the statistical model is: 'What is the probability of observing 80 or fewer nonconforming lenses in a batch of 2400?' Minitab readily provides the answer using **Calc > Probability Distributions > Binomial...** to obtain the output in Panel 7.4. **Cumulative probability** must be selected, with **Number of trials:** 2400, **Event probability:** 0.045 and **Input constant:** 80 specified. Thus, if the null hypothesis were true, i.e. if the change of supplier had no impact on the proportion of nonconforming lenses, then the probability of

> The number of nonconforming lenses, Y, in a batch of 2400 will have the binomial distribution with parameters $n = 2400$ and $p = 0.045$, i.e. $Y \sim B(2400, 0.045)$.

Box 7.2 Statistical model.

```
Cumulative Distribution Function

Binomial with n = 2400 and p = 0.045

 x  P( X <= x )
80    0.0024365
```

Panel 7.4 Probability that sample includes 80 or fewer nonconforming lenses.

observing 80 or fewer nonconforming lenses in a batch of 2400 would be 0.0024. Thus 0.0024 is the *P*-value for testing the hypotheses specified above.

Conclusion. Since the *P*-value is less than $0.01 = 1\%$ the null hypothesis would be rejected in favour of the alternative hypothesis at the 1% significance level. Thus the data from the experiment provide strong evidence that the change of supplier has led to a significant reduction in the proportion of nonconforming lenses from the previous level of 4.5%. A point estimate of the new proportion of nonconforming lenses is $80/2400 = 0.033 = 3.3\%$.

Having decided, say, on a significance level of $\alpha = 0.01 = 1\%$ for the lens coating experiment, one can use **Calc** > **Probability Distributions** > **Binomial...** to look at the final stage of reaching a conclusion in a different way. The statistical model indicates that the number of nonconforming lenses in a batch will have the $B(2400, 0.045)$ distribution. Panel 7.5 displays the Session window output obtained using the **Inverse cumulative probability** function for this binomial distribution via **Calc** > **Probability Distributions** > **Binomial...** and specifying **Input constant:** 0.01. Thus the cut-off between acceptance and rejection of the null hypothesis occurs at the value 84 for the number of nonconforming lenses in the batch when the significance level is $\alpha = 0.01 = 1\%$. With the discrete binomial distribution it has not been possible to determine a value x such that $P(X \leq x)$ is precisely 0.01, so the value 84 is used since $P(X \leq 84)$ is closest to, but less than, 0.01. The conclusion would therefore be:

- Do not reject H_0 if the number nonconforming is greater than 84.
- Reject H_0 if the number nonconforming is less than or equal to 84.

It was established that there were 80 nonconforming lenses in the batch. Since this is less than 84, the conclusion reached was to reject the null hypothesis at the significance level of $\alpha = 0.01 = 1\%$.

Suppose now that change of supplier had actually led to a reduction in the proportion of nonconforming lenses from 0.045 to 0.030. What is the probability that, once the nonconforming lenses had been counted in a batch of 2400, the conclusion reached – to reject H_0 in favour of H_1 – will be the correct one? Again **Calc** > **Probability Distributions** > **Binomial...** provides the answer – see Panel 7.6.

```
Inverse Cumulative Distribution Function

Binomial with n = 2400 and p = 0.045

 x  P( X <= x )        x  P( X <= x )
84    0.0085010       85    0.0112890
```

Panel 7.5 Determining the cut-off number of nonconforming lenses.

```
Cumulative Distribution Function

Binomial with n = 2400 and p = 0.03

 x  P( X <= x )
84     0.929871
```

Panel 7.6 Probability of rejecting null hypothesis when new population proportion is 0.030.

Thus there is a probability of 0.93 (to two decimal places) of the test of hypotheses providing the evidence of a reduction from 0.045 to 0.030 in the population proportion of nonconforming lenses. This probability is the power of the statistical test to detect the change from a population proportion of 0.045 to a population proportion of 0.030. The other side of the coin is that there is probability $1 - 0.93 = 0.07$ of the experiment failing to provide evidence of the reduction, i.e. there is probability $\beta = 0.07$ of committing a Type II error with a test based on a sample of 2400 lenses.

Figure 7.3 shows a plot of the power of the test against the population proportion nonconforming after the process change. Note that the larger the drop in the proportion the more likely it is that the test will provide evidence of the change. For example, were the population proportion of nonconforming lenses to drop by 0.02, so that the population proportion after the process change was 0.025, then the probability is 0.999 that the conclusion would be to reject the null hypothesis, i.e. it is virtually certain that the test would lead to making the correct decision.

In discussing the above two tests of hypotheses the statistical models used employed specific probability distributions, the normal distribution in the case of the DTNs and the binomial distribution in the case of the lens coating process. Other tests of hypotheses are available which do not require use of specific probability distributions. These are referred to as *distribution-free* or *nonparametric* tests. Specific cases of such tests will be introduced later in the chapter.

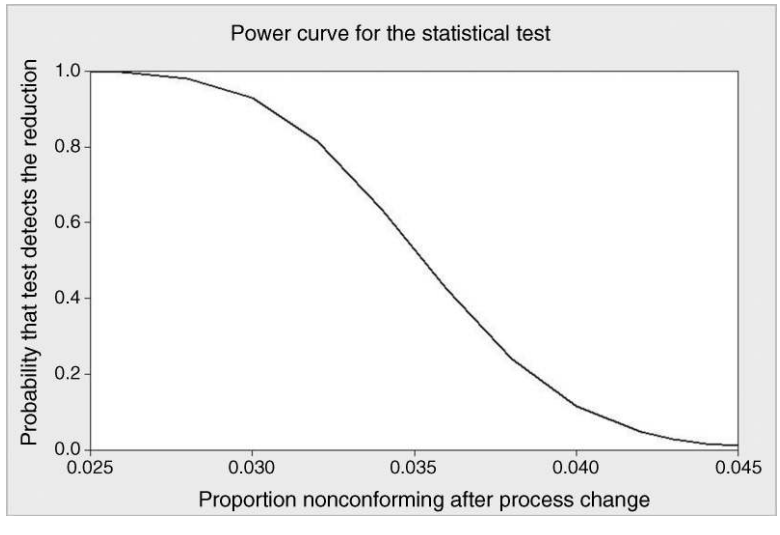


Figure 7.3 Power curve for the statistical test.

Both the above tests of hypotheses were performed from first principles. In the next section we will see how Minitab can be used to streamline performance of the tests.

7.2 Tests and confidence intervals for the comparison of means and proportions with a standard

7.2.1 Tests based on the standard normal distribution – z-tests

Consider again the thrombolytic treatment example and the DTN data in Table 7.1 and worksheet DTNTime.MTW. The standard DTN could be thought of as the mean $\mu = 19$ minutes of the normal distribution of DTNs. When the data for the 25 patients treated with the specialist nurses available are to hand we wish to compare these data with the standard via the formal test of the hypotheses:

$$H_0 : \mu = 19, \quad H_1 : \mu < 19.$$

Recall, too, that the variability was assumed to remain unchanged, with the standard deviation being 6 minutes. The dialog involved in performing the test in Minitab using **Stat** > **Basic Statistics** > **1-Sample Z...** is shown in Figure 7.4.

Here the sample of DTNs is available in column C1. **Standard deviation:** 6 is specified and, with **Perform hypothesis test** checked, **Hypothesized mean:** 19 indicates the null hypothesis. Under **Options...** the alternative hypothesis is specified by use of the scroll arrow to select **less than** in the **Alternative:** window. Finally, under **Graphs...** one can select to display the data in the form either of a histogram, an individual values plot or a boxplot; in this case the **Histogram of data** option was selected. The Session Window output is shown in Panel 7.7 and the graphical output is shown in Figure 7.5.

The Session window output includes the following:

- a statement of the hypotheses under test in the first line;
- the value of the standard deviation assumed to apply (6 in this case);
- the sample size, sample mean and sample standard deviation;

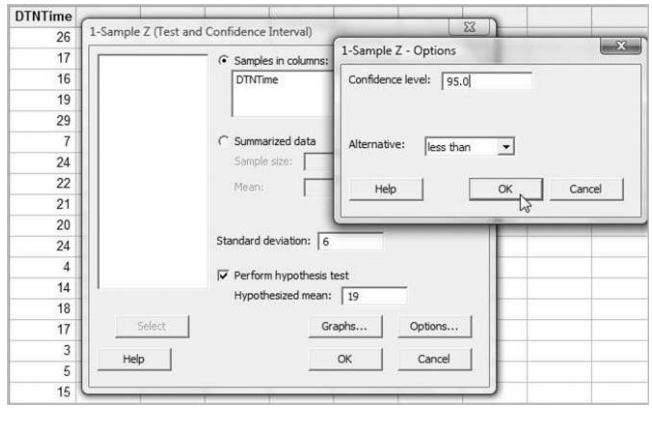


Figure 7.4 Dialog for performing a one-sample z-test.

One-Sample Z: DTNTime

```
Test of mu = 19 vs < 19
The assumed standard deviation = 6

                                                95% Upper
Variable    N    Mean   StDev  SE Mean          Bound       Z       P
DTNTime    25   16.28    6.83     1.20          18.25   -2.27   0.012
```

Panel 7.7 Session window output for z-test.

- the standard error of the mean which is the standard deviation of the sample mean given by $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 6/\sqrt{25} = 1.2$;
- a 95% upper bound of 18.25 which will be explained later in this section;
- a Z-value of -2.27, which is explained in Box 7.3;
- the P-value of 0.012 for the test.

The P-value of 0.012 was calculated in the previous section in the 'grass-roots' version of the test. Since it is less than 0.05 one can immediately conclude that the null hypothesis $H_0 : \mu = 19$ would be rejected at the $\alpha = 0.05$ significance level in favour of the alternative hypothesis $H_1 : \mu < 19$. In other words, the experiment provides evidence of a significant reduction in the population mean DTN at the 5% level of significance.

In the previous section it was noted that the cut-off between acceptance and rejection of the null hypothesis, when using the 5% significance level, occurs at the value of 17.026 2 for the sample mean. The value of Z corresponding to 17.026 2 is -1.64. In terms of the standardized variable Z the conclusion would therefore be:

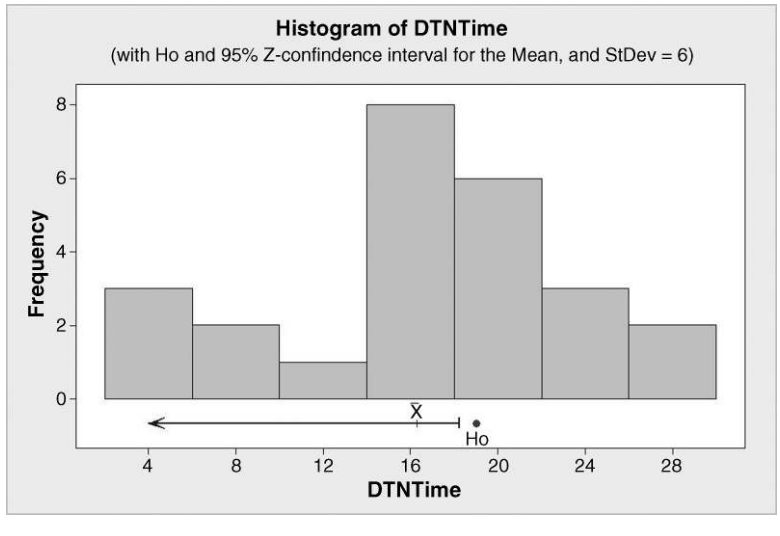


Figure 7.5 Histogram of door to needle times with z-test annotation.

If V is a random variable with mean μ_V and standard deviation σ_V, then the random variable $Z = (V - \mu_V)/\sigma_V$ is the standardized variable which has mean 0 and standard deviation 1. If V is normally distributed then so is Z. In the case of door to needle time, Y, in this example the sample mean, $\bar{Y}$, is, under the null hypothesis, normally distributed with mean 19 and standard deviation 1.2, so the corresponding standardised variable is given by $Z = (\bar{Y} - 19)/1.2$. The mean of the sample of 25 times was $\bar{y} = 16.28$ with corresponding $z = (16.28 - 19)/1.2 = -2.27$ This is the value of Z given in the Session window output.

Box 7.3 Calculation of the z-statistic given in Session window output.

- Do not reject H_0 if Z is greater than -1.64.
- Reject the null hypothesis H_0 if Z is less than or equal to -1.64.

From the data it has been established that the value of Z corresponding to the sample mean of 16.28 was -2.27. Since this is less than -1.64 the conclusion reached would be to reject the null hypothesis at the significance level of $\alpha = 0.05$.

Prior to the availability of statistical software packages, such as Minitab, tests of this type were typically conducted by calculating Z using the formula

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

The conclusion regarding acceptance or rejection of the null hypothesis would then be made by reference to tables of critical values of Z, taking into account the significance level of interest and the nature of the alternative hypothesis. (The creation of such a table was set as Exercise 7 in Chapter 4.) The key involvement of Z in such tests of hypotheses involving a single sample gives rise to the nomenclature *one-sample Z-test*.

Finally, using the formula for Z we can answer the question: 'What null hypotheses would be acceptable at the 5% significance level?' The mathematical manipulations are given in Box 7.4 for the interested reader. Others may skip over the mathematics to the interpretation that follows.

For the null hypothesis to be accepted we require

$$\begin{aligned} Z > -1.64 &\Rightarrow \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} > -1.64 \\ &\Rightarrow \frac{16.28 - \mu}{1.2} > -1.64 \\ &\Rightarrow 16.28 - \mu > -1.64 \times 1.2 \\ &\Rightarrow \mu < 16.28 + 1.968 \\ &\Rightarrow \mu < 18.25. \end{aligned}$$

Box 7.4 Calculation of range of acceptable population means.

The value of 18.25 is given in the Session window output as the 95% upper bound. In applied statistics 95% confidence level goes hand in hand with the 5% significance level. Had the null hypothesis been $H_0 : \mu = 17$, with alternative $H_1 : \mu < 17$, then one can deduce immediately that the null hypothesis would not be rejected at the 5% level of significance since 17 is less than 18.25, the 95% confidence upper bound. Similarly, $H_0 : \mu = 19$, with alternative $H_1 : \mu < 19$ would be rejected at the 5% level of significance (as we have already seen) since 19 is greater than 18.25. Thus one could report the results of the experiment by stating that 'on the basis of the data collected, the population mean DTN was estimated to be 16.28 minutes and that, with 95% confidence, it could be claimed that the true population mean was at most 18.25 minutes'.

The histogram created as the graphical output in Figure 7.5 is annotated with a point labelled H_0 indicating the value 19 specified in the null hypothesis. The arrowed line segment has a tick mark labelled $\bar{X}$ on it, indicating the sample mean. The segment extends from the value 18.25 downwards and indicates the 95% confidence interval for the population mean following the process change. The fact that the point corresponding to H_0 does not lie on the line segment indicates rejection of the null hypothesis $H_0 : \mu = 19$ in favour of the alternative $H_1 : \mu < 19$ at the 5% level of significance.

Had the data been presented in summary form, i.e. that a sample of 25 times following the process change had mean 16.28, then the test could still be performed using **Stat > Basic Statistics > 1-Sample Z...** by checking **Summarized data** and entering **Sample size:** 25, **Mean:** 16.28 and **Standard deviation:** 6. With **Perform hypothesis test** checked, **Hypothesized mean:** 19 indicates the null hypothesis. Under **Options...** the alternative hypothesis is specified by use of the scroll arrow to select **less than** in the **Alternative:** window. Without the raw data no graphical output is possible. The reader is invited to check that the output in Panel 7.8 results and that it is identical to that obtained by performing the test using the column of raw data, except that it is not possible to deduce the standard deviation of the sample from the summary information provided to the software.

As a second example, consider a type of glass bottle for which burst strength (psi) could be adequately modelled by the normal distribution with mean 480 and standard deviation 64. During a Six Sigma project with the aim of increasing the burst strength of the bottles, a new glass formulation was used in a large production run of the bottle. The burst strength data for a random sample from the batch are given in Table 7.3 and also as a single column in the worksheet Burst_Strength.MTW. Does the data provide evidence of increased mean burst strength?

One-Sample Z

```
* NOTE * Graphs cannot be made with summarized data.

Test of mu = 19 vs < 19
The assumed standard deviation = 6

                        95% Upper
 N   Mean   SE Mean      Bound      Z      P
25  16.28     1.20      18.25  -2.27  0.012
```

Panel 7.8 Session window output for z-test using summarized data.

Table 7.3 Burst strength (psi) for a sample of 50 bottles.

535	476	439	541	526	523	465	476	468	524
449	444	431	582	580	447	503	498	488	467
545	528	538	570	453	700	535	454	403	498
573	558	442	490	503	476	609	483	484	443
535	476	439	541	526	523	465	476	468	524

Here the hypotheses are:

$$H_0 : \mu = 480, \quad H_1 : \mu > 480.$$

Proceeding as in the previous case, use of **Stat > Basic Statistics > 1-Sample Z...** is required with the **Standard deviation** of 64 specified, **Perform hypothesis test** checked and **Hypothesized mean:** 480 entered. Under **Options...** the alternative hypothesis is specified by use of the scroll arrow to select **greater than** in the **Alternative:** window. Under **Graphs:** the **Boxplot of data** option was selected. The Session window output is shown in Panel 7.9 and the graphical output is shown in Figure 7.6.

One-Sample Z: Burst Strength

```
Test of mu = 480 vs > 480
The assumed standard deviation = 64

                                                    95% Lower
Variable          N    Mean   StDev  SE Mean       Bound      Z      P
Burst Strength   50  504.08   55.55     9.05      489.19   2.66  0.004
```

Panel 7.9 Session window output for z-test on burst strength data.

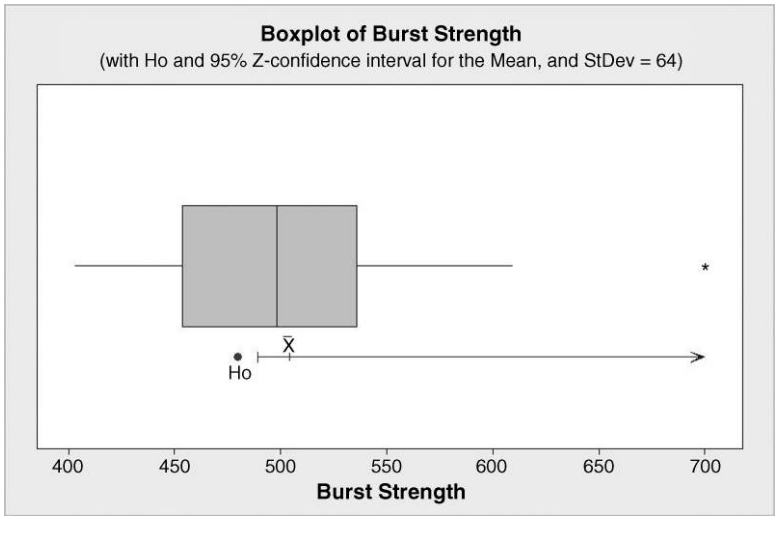


Figure 7.6 Boxplot of burst strength data with z-test annotation.

The conclusion would be that the data provide evidence that the new glass formulation has led to an increase in population mean burst strength – the null hypothesis $H_0 : \mu = 480$ being rejected in favour of the alternative $H_1 : \mu > 480$ at the significance level $\alpha = 0.01$. The P-value for the test is 0.004. The data also enable one to state with 95% confidence that, with the new glass formulation, the population mean burst strength will be at least 489 psi. (Of course a decision as to whether or not to change to the new glass formulation would be likely to involve cost and other considerations.)

As a third example, consider the following scenario. Before revising staffing arrangements at a busy city branch, a major bank determined that the service time (seconds) for business customers could be adequately modelled by a normal distribution with mean 453 and standard deviation 38 seconds. Following implementation of the revision, the mean service time for a random sample of 62 business customers was 447 seconds. The data are stored in Service. MTW. Do these data provide any evidence of a change (either an increase or a decrease) in the mean service time for business customers?

Proceeding as in the previous two cases, but with **Alternative:** set to **not equal**, we can test the null hypothesis $H_0 : \mu = 453$ against the alternative hypothesis $H_1 : \mu \neq 453$. The third available graphical option of an **Individual value plot** (dotplot) was selected in this case. The Session window output is shown in Panel 7.10 and the graphical output in Figure 7.7.

One-Sample Z: Service Time

```
Test of mu = 453 vs not = 453
The assumed standard deviation = 38


Variable       N    Mean  StDev  SE Mean        95% CI            Z      P
Service Time  62  447.00  36.54     4.83  (437.54, 456.46)  -1.24  0.214
```

Panel 7.10 Session window output for z-test on service time data.

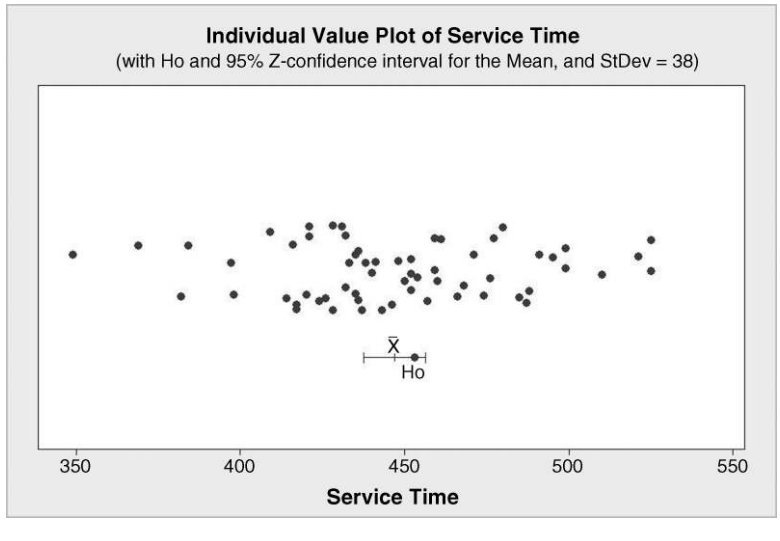


Figure 7.7 Individual value display of service time data with z-test annotation.

(On obtaining the graph the author double-clicked on a data point to access the **Edit Individual Symbols** menu. Under the **Identical Points** tab the **Jitter** option was selected. On clicking **OK** the display shown was obtained. Use of jitter reveals overlapping points.)

The conclusion would be that the data provide no evidence that the revised staffing arrangements have led to a change in population mean service time since the *P*-value for the test is 0.214. Since the *P*-value exceeds 0.05 the null hypothesis cannot be rejected at the significance level $\alpha = 0.05$. The data also enable one to state with 95% confidence that, following the revision of staffing arrangements, the population mean service time lies in the interval 438 to 456 seconds, rounded to the nearest integer. This 95% confidence interval (CI) for the population mean service time has been taken from the Session window output and rounded. The fact that the 95% confidence interval includes the mean of 453 specified in the null hypothesis indicates that $H_0 : \mu = 453$ cannot be rejected at the 5% level of significance. This is evident in the graphical display also, as the point representing the value of the population mean specified in the null hypothesis lies on the line segment that represents the 95% confidence interval.

The first and second examples involved one-tailed tests. In terms of z, the criterion for rejection of the null hypothesis at the 5% level of significance was z less than or equal to -1.64 in the first case, z greater than or equal to 1.64 in the second case, and in the third case either z less than -1.96 or z greater than 1.96. (The fact that 1.96 rounds to 2.00 may explain why the 5% level is the most widely used significance level in applied statistics – for a null hypothesis to be rejected at the 5% level in a two-tailed z-test, it is easy to remember that z must exceed 2 in magnitude.)

The test could have been carried out in the third case by obtaining the values of the mean service time for a random sample of 62 customers, on the assumption that the null hypothesis is true, that correspond to the z-values of -1.96 and 1.96. The values are 443.5 and 462.5, respectively. The situation is illustrated in Figure 7.8. Values of the sample mean either less than 443.5 or greater than 462.5 comprise the critical region for the *two-tailed test* here.

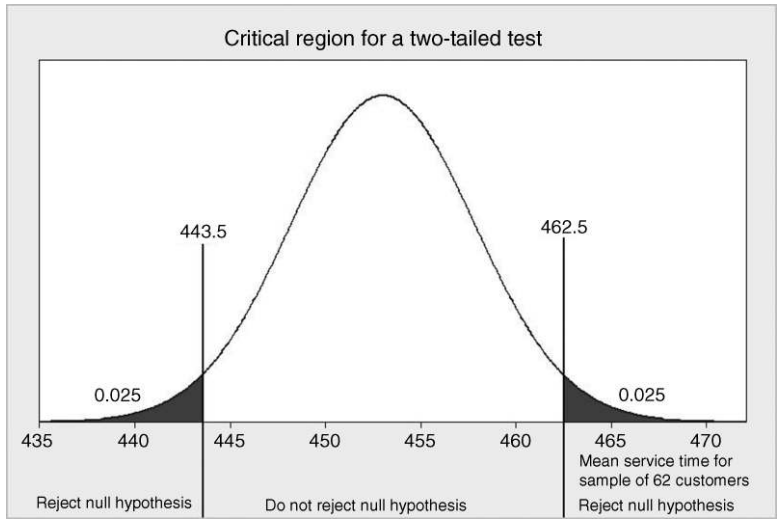


Figure 7.8 Critical region for a two-tailed test.

The sample mean obtained was 447.0, which is not in the critical region – hence the decision not to reject the null hypothesis at the 5% level of significance. Thus there is no evidence from the data to suggest that the revised staffing arrangements have had any impact on the mean time taken to serve business customers.

7.2.1.1 Some comments on tests of hypotheses and P-values

Vickers (2010) states that 'the p-value is the probability that the data would be at least as extreme as those observed, if the null hypothesis were true'. His book gives much sound advice on hypothesis testing via both light-hearted contexts and real applications. Ronald Fisher, who played a key role in the development of statistical inference, wrote on *P*-values in *Statistical Methods for Research Workers*, first published in 1925: 'We shall not often be astray if we draw a conventional line at 0.05' (Fisher, 1954, p. 80) This comment will undoubtedly have contributed to the level of significance 0.05 becoming the most widely used in applied statistics.

Statistical tests of hypothesis are controversial. Sterne and Davey Smith (2001, p. 226) make the following summary points:

> *P* values ... measure the strength of the evidence against the null hypothesis; the smaller the *P* value, the stronger the evidence against the null hypothesis.
>
> An arbitrary division of results, into 'significant' or 'nonsignificant' according to the *P* value, was not the intention of the founders of statistical inference.
>
> A *P* value of 0.05 need not provide strong evidence against the null hypothesis, but it is reasonable to say that $P < 0.001$ does. In the results ... the precise *P* value should be presented, without reference to arbitrary thresholds.
>
> Confidence intervals should always be included in reporting the results of statistical analyses, with emphasis on the implications of the ranges of values in the intervals.

The value of z quoted in the output is known as the test statistic. Montgomery (2009, p. 117) comments:

> It is customary to call the test statistic (and the data) significant when the null hypothesis H_0 is rejected; therefore, we may think of the *P*-value as the smallest level α at which the data are significant. Once the *P*-value is known, the decision maker can determine for himself or herself how significant the data are without the data analyst formally imposing a pre-selected level of significance.

This author strongly recommends that a display of the data should be incorporated into the reporting whenever it is possible to do so.

7.2.1.2 Some comments on confidence intervals

In order to aid the reader's understanding of confidence intervals a series of 80 random samples of size 62 was generated from the $N(453,38^2)$ distribution using Minitab. (The sample size and parameters chosen relate to the earlier example on customer service time in a bank, but one

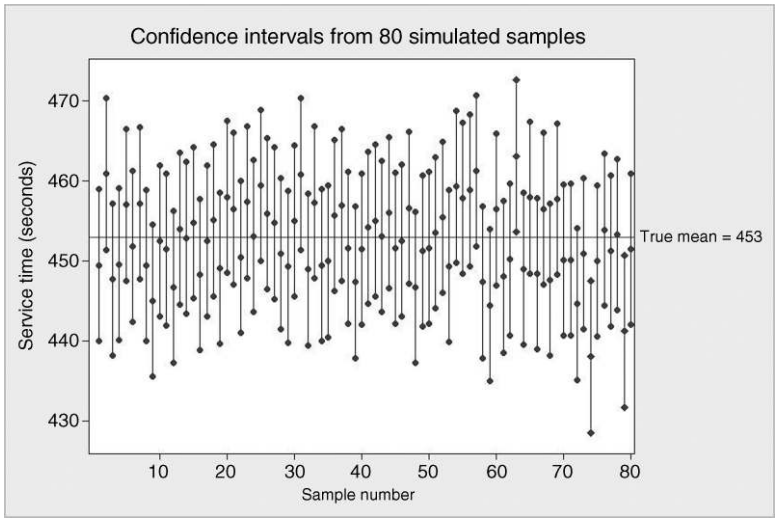


Figure 7.9 Display of 95% confidence intervals from simulated samples.

could choose arbitrary values.) On the assumption that the standard deviation of the population was known to be 38, two-sided 95% confidence intervals for the population mean were computed and displayed using the **1-Sample Z...** facility in Minitab – see Figure 7.9.

The reference line indicates the true population mean of 453. The first vertical line segment represents the 95% confidence interval based on the first simulated sample which was (440.1,459.0) and captures within it the true population mean of 453 – capture is indicated by the line segment crossing the reference line. The 63rd line segment represents the 95% confidence interval based on the 63rd sample which was (453.7, 472.6) and fails to capture within it the true population mean of 453 – failure to capture is indicated by the segment not crossing the reference line. Of the 80 confidence intervals, the 63rd, 74th and 79th fail to 'capture' the true population mean. This corresponds to 77 captures from 80 attempts, which is equivalent to 96.2%. This capture rate is close to the long-term capture rate for such intervals of 95%.

The formula for calculating a 95% two-sided confidence interval for the mean of a normally distributed population, with known standard deviation σ, based on a sample of size n, is

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

Suppose that a machine for filling jars with instant coffee granules delivers amounts that are normally distributed with standard deviation 0.3 g. The process manager wishes to know how big a sample of jars is required in order to estimate, with 95% confidence, the mean amount delivered to within 0.1 g of its true value. This means that we require

$$1.96 \frac{0.3}{\sqrt{n}} < 0.1 \Rightarrow n > \left(\frac{1.96 \times 0.3}{0.1} \right)^2 = 34.5.$$

Thus a random sample of at least 35 jars would be required.

The value 0.1 g may be referred to as the margin of error, E. In general, in order to estimate, with 95% confidence, the mean of a population to within E of its true value requires a sample size n of at least $3.84\sigma^2/E^2$. If the standard deviation, σ, of the population is unknown then an estimate of the standard deviation from a pilot sample may be used in the calculation. If the pilot sample is small then caution should be exercised in applying the formula. For estimation with 99% confidence, the factor of 3.84 is replaced by 6.63 in the above formula, and for 99.9% confidence it is replaced by 10.83. A point to note is that halving the margin of error quadruples the sample size required.

7.2.2 Tests based on the Student *t*-distribution – *t*-tests

At the core of the tests we have considered so far in this section is the test statistic given by

$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}.$$

In each of the examples considered it was assumed that process variability was unaffected by the process changes so that the standard deviation, σ, was known.

Hence the tests are called z-tests.

What do we do if this assumption is suspect? If we have modified a process. might not the changes made affect variability as well as location? A natural thing to do is to calculate the test-statistic value:

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

This formula may be obtained from the previous one by using the sample standard deviation s in place of the population standard deviation, σ. If the underlying random variable Y has a normal distribution then the random variable T has a Student's t-distribution. The distribution is named in honour of William S. Gosset who was appointed to a post in the Guinness brewery in Dublin in 1899 and made a major contribution to the development of applied statistics. He developed the t-test to deal with small samples used for quality control in brewing. He wrote under the pseudonym 'Student' because his company had a policy against work done for the company being made public. A sample of n values has $\nu = \mathrm{n} - 1$ degrees of freedom. (The Greek letter ν is nu.) There is a separate t-distribution for each number of degrees of freedom 1, 2, 3, ...

As an example suppose that initially assembly of P87 modules took on average 48.0 minutes with standard deviation 3.7 minutes. At a later date the following sample of eight assembly times was obtained:

46 48 45 48 46 47 43 48.

We wish to evaluate the evidence from this sample of assembly times for a reduction in the population mean assembly time.

In order to perform the t-test of the null hypothesis $H_0 : \mu = 48$ against the alternative hypothesis $H_1 : \mu < 48$, first set up the data in a column named Assembly time. Use of **Stat > Basic Statistics > 1-Sample t...** is required with 'Assembly time' selected under

One-Sample T: Assembly time

```
Test of mu = 48 vs < 48

                                                  95% Upper
Variable        N     Mean   StDev   SE Mean       Bound      T      P
Assembly time   8   46.375   1.768     0.625      47.559  -2.60  0.018
```

Panel 7.11 Session window output for *t*-test on assembly time data.

Samples in columns:, **Perform hypothesis test** checked, **Hypothesized mean:** 48 entered and **less than** specified via **Options...** under **Alternative:**. With such a small sample an **Individual value plot**, rather than a histogram or boxplot, is recommended under **Graphs...**. The Session window output is shown in Panel 7.11 and the graphical output in Figure 7.10.

The output follows the same pattern as for a *z*-test. A statement of the hypotheses of interest is followed by summary statistics for the sample. Finally, the confidence interval, test statistic and *P*-value are given. The data provide evidence via the *t*-test of a reduction in the population mean assembly time, the null hypothesis that the mean is 48 minutes being rejected at the 5% significance level (P-value $= 0.018$). A point estimate of the new population mean is 46.4 minutes and, with 95% confidence, it can be stated that the new population mean is at most 47.6 minutes. Note how, in Figure 7.10, the point representing the value of the mean specified in the null hypothesis does not lie on the line segment that represents the 95% confidence interval. This provides visual confirmation of the rejection of the null hypothesis at the 5% significance level.

The *t*-test requires the random variable of interest to be normally distributed. The normal probability plot in Figure 7.11 provides no evidence of nonnormality of the data (P-value $= 0.237$).

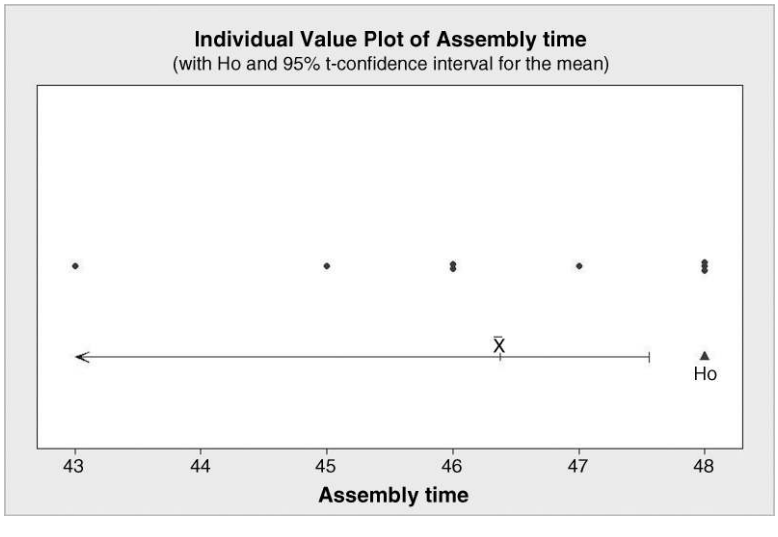


Figure 7.10 Individual value display of assembly time with *t*-test annotation.

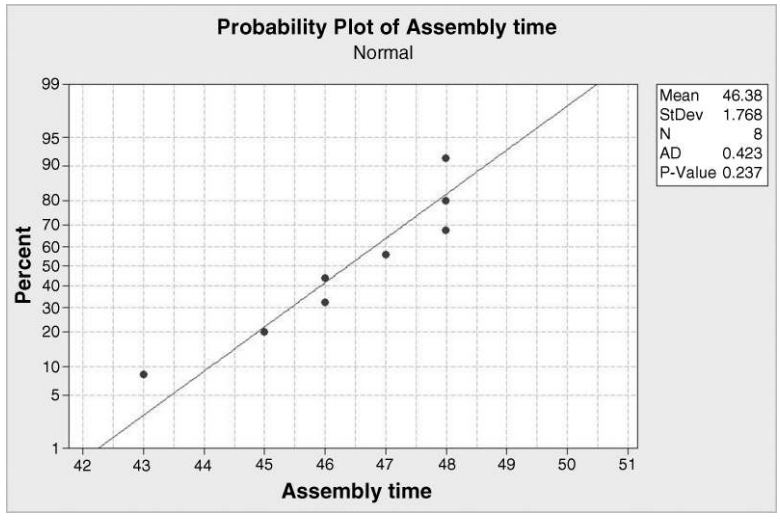


Figure 7.11 Probability plot of assembly time data.

If the normality test provided strong evidence of nonnormality of the distribution of the random variable of interest, then one possible approach would be to seek an appropriate transformation of the data. Box–Cox transformation may be explored via **Stat > Control Charts > Box-Cox Transformation...**. An alternative approach would be to perform a nonparametric test that will be introduced later in the chapter.

The probability density functions for the standard normal distribution, i.e. the N(0,1) distribution, and the t distribution with parameter $\nu = 7$ degrees of freedom are displayed in Figure 7.12. One of Gosset's major contributions was the creation of tables of critical values of to enable tests such as the above to be performed. In Panel 7.12 the critical values for the above

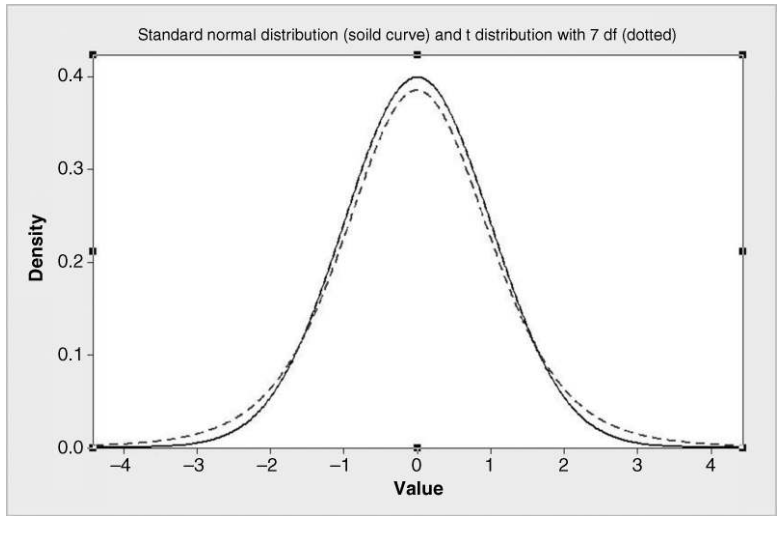


Figure 7.12 Probability density functions for the $N(0,1)$ and t_7 distributions.

```
Inverse Cumulative Distribution Function

Student's t distribution with 7 DF

P( X <= x )          x
       0.05  -1.89458

Student's t distribution with 7 DF

P( X <= x )          x
       0.01  -2.99795
```

Panel 7.12 Critical values of t with 7 degrees of freedom.

test at both the 5% and 1% levels of significance are displayed, i.e. -1.89 and -3.00 respectively, rounded to two decimal places. The calculated value of the test statistic, t, was given as -2.60 in Panel 7.11. It follows that, since -2.60 is less than -1.89, the null hypothesis would be rejected at the 5% level. Since -2.60 is greater than -3.00, the null hypothesis cannot be rejected at the 1% level. This also indicates that the P-value must lie between 0.05 and 0.01. Minitab provided us with the precise P-value of 0.018.

7.2.2.1 Power and sample size

Recall that, in the DTN example, the population mean time before the process change was 19 minutes with population standard deviation 6 minutes. Suppose that the project team decided that, were the population mean time to decrease by 3 minutes due to the introduction of the specialist nurses, then they would like to be 95% certain to detect the decrease, using a significance level of $\alpha = 0.05$. Thus they would be specifying power 0.95 for the z-test to detect the change from population mean 19 to population mean 16. The question to be answered is therefore that of how big a sample should be taken. Minitab provides the answer via **Stat > Power and Sample Size > 1-Sample Z...**. The required dialog is shown in Figure 7.13.

The change of interest, from 19 to 16, is specified as **Differences:** -3; the population standard deviation is assumed to remain unchanged so **Standard deviation:** 6 is entered. The Session window output is shown in the Panel 7.13. In addition, a power curve similar to that displayed in Figure 7.2 is created by default.

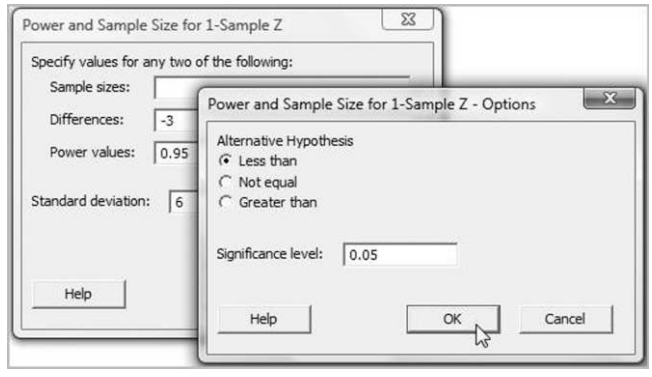


Figure 7.13 Dialog for sample size calculation.

```
Power and Sample Size

1-Sample Z Test

Testing mean = null (versus < null)
Calculating power for mean = null + difference
Alpha = 0.05  Assumed standard deviation = 6

            Sample  Target
Difference    Size   Power  Actual Power
        -3      44    0.95      0.952715
```

Panel 7.13 Session Window output for z-test on Burst Strength data.

Panel 7.13 shows that the required sample size is 44 patients. (Note the actual power is 0.953 – Minitab computes the lowest sample size that gives power greater than the target power specified by the user.) Minitab also provides power and sample size calculations for t-tests. An exercise will be provided.

7.2.3 Tests for proportions

In the lens coating example the nonconformance rate was 4.5% prior to the introduction of the new supplier. In a trial run with the coating fluid from the new supplier there were 80 nonconforming lenses in a batch of 2400. Here the test is of null hypothesis $H_0 : p = 0.045$ versus the alternative hypothesis $H_1 : p < 0.045$. To perform the test directly via Minitab, use **Stat > Basic Statistics > 1 Proportion...** with the dialog shown in Figure 7.14.

The Session window output is given in Panel 7.14. The P-value of 0.002 was obtained in Section 7.1 using grass-roots computation and the binomial distribution – see Panel 7.4. Thus the data provide evidence of a reduction in the proportion of nonconforming lenses at the 1% significance level since the P-value is less than 0.01. The estimated proportion nonconforming following the process change is 3.3%. The process owner can be 95% confident that the proportion nonconforming is at worst 4.0% (the upper bound expressed as a percentage and rounded to one decimal place) following the process change. Of course a decision on whether or not to continue with the new supplier would typically involve consideration of the costs involved.

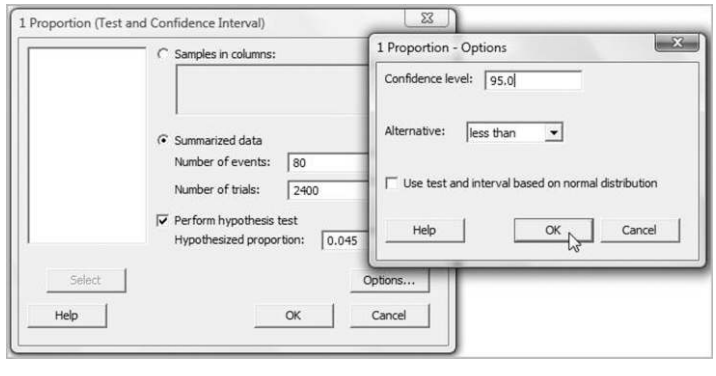


Figure 7.14 Dialog for testing a proportion.

Test and CI for One Proportion

```
Test of p = 0.045 vs p < 0.045


                                95% Upper    Exact
Sample    X     N  Sample p       Bound  P-Value
1        80  2400  0.033333    0.040008    0.002
```

Panel 7.14 Session window output for test of proportion nonconforming.

Suppose that the process team had wished to detect a reduction in the proportion of nonconforming lenses from 4.5% to 3.0% with power 0.99, using a significance level of 0.01. Use of **Stat > Power and Sample Size > 1 Proportion...** with the dialog shown in Figure 7.15 provides the sample size required in the Session window output shown in Panel 7.15. In addition, a power curve similar to that displayed in Figure 7.3 is created by default. The size of sample required is 3435. Thus in order to ensure a probability of 0.99 of detecting a reduction in the proportion of nonconforming lenses from 4.5% to 3.0%, using a significance level of 0.01, the process team would require a sample size of the order of 3500.

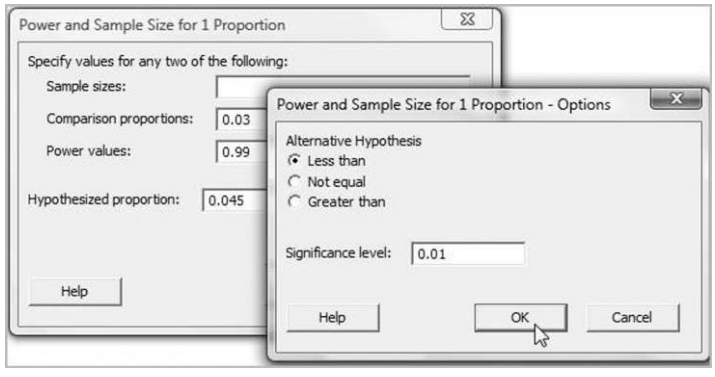


Figure 7.15 Dialog for sample size calculation.

Power and Sample Size

```
Test for One Proportion

Testing p = 0.045 (versus < 0.045)
Alpha = 0.01


               Sample  Target
Comparison p     Size   Power  Actual Power
        0.03     3435    0.99      0.990004
```

Panel 7.15 Session window output for sample size computation.

Table 7.4 Waiting time (seconds) data.

471	36	55	185	88
188	58	72	18	47
42	34	41	67	28

7.2.4 Nonparametric sign and Wilcoxon tests

Following alterations to a checkout facility at a supermarket, a random sample of customers was observed and the time they waited for service recorded. The data are displayed in Table 7.4 and available in Wait.MTW.

Prior to the alterations extensive monitoring had shown that the median waiting time was 120 seconds. For the symmetrical normal distribution the mean and median are identical, so were we to evaluate the evidence for a reduction in the median following the changes by performing a t-test of null hypothesis $H_0: \mu = 120$ versus the alternative hypothesis $H_1: \mu < 120$ then the output in Panel 7.16 would be obtained from Minitab.

With a P-value of 0.212 there would appear to be no evidence of a reduction in the mean, and by implication, in the median. However, a normal probability plot of the data provides strong evidence of the distribution of waiting time being nonnormal. As an alternative to seeking a suitable transformation of waiting time one may proceed as follows. Order the data and record a negative sign if an observed value is less than 120; record a positive sign if an observed value is greater than 120. (Any values of exactly 120 would be excluded from the analysis.) The results are shown in Table 7.5.

By definition, the probability that an observed value is less than the median of a population is 0.5; likewise, the probability that it is greater than the median of the population is 0.5. For the above data we have observed a value less than the hypothesized median of 120 for 12 observations from a total of 15 observations (corresponding to the 12 negative signs from a total of 15 signs). We therefore ask what is the probability of observing this pattern, or a more extreme one. This is equivalent to asking what is the probability of obtaining 12 or more tails from 15 tosses of a coin. Consider therefore the random variable X having the $B(15, 0.5)$ distribution. We require

One-Sample T: Wait

```
Test of mu = 120 vs < 120

                                                    95%
                                                  Upper
Variable   N     Mean     StDev  SE Mean     Bound      T      P
Wait      15  95.3333  115.9136  29.9288  148.0471  -0.82  0.212
```

Panel 7.16 Session window output for t-test on waiting time data.

Table 7.5 Coded data for waiting time.

Wait	18	28	34	36	41	42	47	55	58	67	72	88	185	188	471
Sign	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+

```
Cumulative Distribution Function

Binomial with n = 15 and p = 0.5

 x  P( X <= x )
11     0.982422
```

Panel 7.17 Calculation of probability of 11, or fewer, + signs.

$P(X \geq 12) = 1 - P(X \leq 11)$. Use of **Calc** > **Probability Distributions** > **Binomial...** yields the result in Panel 7.17. Hence, $P(X \geq 12) = 1 - P(X \leq 11) = 1 - 0.982\,422 = 0.017\,578$. Since this probability is less than 0.05 we have evidence of a reduction in the median waiting time. This test of hypotheses has made no appeal to any particular probability distribution of the waiting time. It is known as the *sign test* and is an example of a distribution-free or nonparametric test.

Use of **Stat** > **Nonparametrics** > **1-Sample Sign...** enables the test to be performed directly. The completed dialog is shown in Figure 7.16. The Session window output is shown in Panel 7.18. The output includes a statement of the hypotheses under test together with counts of the numbers of observations in the sample that are less than, equal to and greater than the

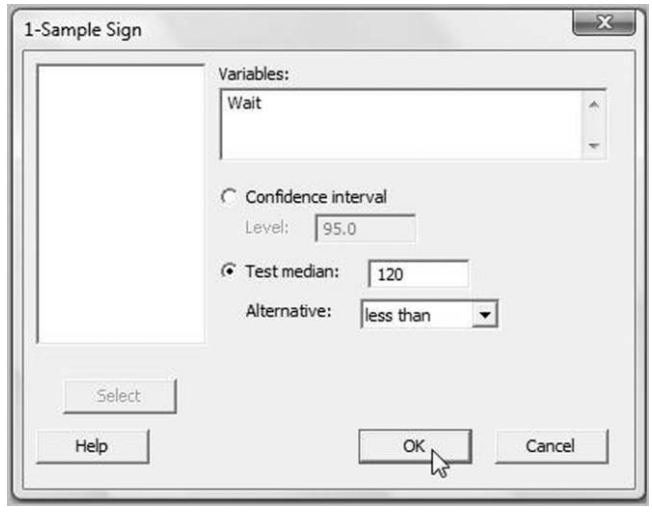


Figure 7.16 Dialog for sign test.

```
Sign Test for Median: Wait

Sign test of median =  120.0 versus < 120.0

       N  Below  Equal  Above       P  Median
Wait  15     12      0      3  0.0176   55.00
```

Panel 7.18 Session window output for sign test.

Sign CI: Wait

```
Sign confidence interval for median

                                 Confidence
                       Achieved    Interval
       N  Median  Confidence  Lower  Upper  Position
Wait  15   55.00      0.8815  41.00  72.00         5
                      0.9500  37.87  82.02       NLI
                      0.9648  36.00  88.00         4
```

Panel 7.19 Session window output giving confidence interval for median.

median specified in the null hypothesis. The *P*-value, computed as above using the binomial distribution, and the median of the sample are also given. In situations where observations are equal to the median specified in the null hypothesis these are discounted from the calculations of the *P*-value for the test.

The sign test procedure in Minitab offers as default the option to obtain a two-sided confidence interval for the median instead of performing a test of hypotheses. The Session window output for the waiting time data is shown in Panel 7.19. The key element is the 95% confidence interval (37.87, 82.02) for the population median waiting time following the alterations to the checkout facility. The fact that this interval does not include 120 formally provides evidence at the 5% level of significance of a change in the population median result of the alterations. (The reader will recall that 95% confidence intervals go hand in hand with 5% significance.)

Another nonparametric test that may be used in the type of scenario discussed in this chapter, where comparison is being made with a standard, is the one-sample Wilcoxon signed-rank test. With **Stat** > **Nonparametrics** > **1-Sample Wilcoxon...** one can test hypotheses concerning the median or obtain the corresponding point estimate and confidence interval. As with the sign test no appeal is made to any particular underlying distribution for the random variable of interest, but the assumption that the data constitute a random sample from a continuous, symmetric distribution is required. The reader is invited to verify that this test yields a *P*-value of 0.035 so the conclusion, at the 5% level of significance, is the same as that from the sign test. The median of the sample of 15 values of waiting time is 55 seconds. The output from the Wilcoxon test includes an estimate of 58.75 for the population median. As with the sign test, the default output from the Wilcoxon procedure is computation of a two-sided 95% confidence interval for the population median.

In all the scenarios discussed in this section we have been concerned with processes as depicted in Figure 1.3. Typically there has been a well-established current parameter for a process performance measure, *Y*, of interest. The performance measure has been usually either a mean or a proportion. The author has chosen to refer to the current level of performance as a standard in the heading for section 7.2. The methods introduced are useful for assessing the impact, if any, of a change to an input, *X*, or factor on *Y*. In the next section we will look at techniques which can be applied in situations where two choices are available for the levels of the factor, e.g. where we wish to compare two processes for dealing with the administration of thrombolytic drugs in a hospital accident and emergency department or to compare two potential suppliers of lens coating fluid where there has are no well-established current parameters for the process performance measure, *Y*, of interest.

7.3 Tests and confidence intervals for the comparison of two means or two proportions

7.3.1 Two-sample *t*-tests

An assembly operation requires a 6-week training period for a new employee to reach maximum efficiency. A new method of training was proposed and an experiment was carried out to compare the new method with the standard method. A group of 18 new employees was split into two groups at random. Each group was trained for 6 weeks, one group using the standard method and the other the new method. The time (in minutes) required for each employee to assemble a device was recorded at the end of the training period. Here the X is the training method and the Y is the assembly time. The two levels of the factor training method are 'standard' and 'new'.

If the data can be considered as independent random samples from normal distributions with means μ_1 and μ_2 with common variance σ^2, then a two-sample t-test is available via **Stat** > **Basic Statistics** > **2-Sample t...**. The completed dialog is shown in Figure 7.17. The data are available in Assembly.MTW.

Training method is indicated in the text column named Method, with entries New and Standard. Minitab treats these identifying labels, New and Standard, in alphabetical order so that it takes μ_1 to refer to the new method and μ_2 to refer to the standard method of training. The null hypothesis is $H_0: \mu_1 = \mu_2$, i.e. that there is no difference in mean assembly time for employees trained by the new and the standard method. The alternative hypothesis is $H_1: \mu_1 < \mu_2$, since it was of interest to determine whether or not there was evidence that new training method led to a reduction in the mean assembly time. Thus under **Options...** less than has to be selected as **Alternative:**. **Assume equal variances** was checked. One should always choose one of the display options under **Graphs...**. **Boxplots of data** was selected in

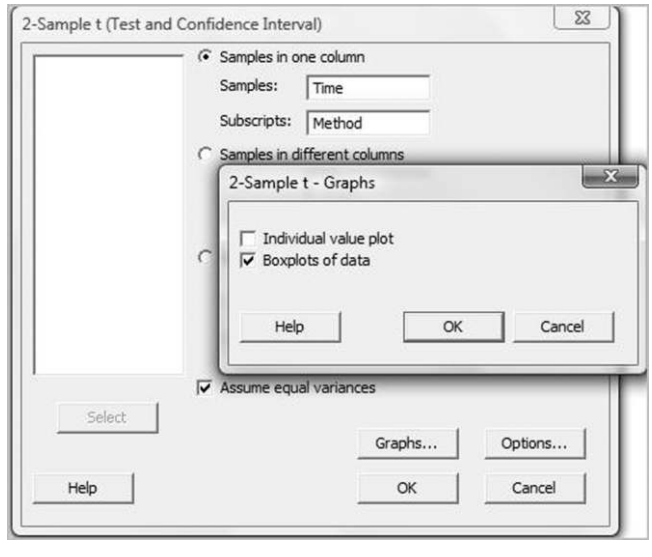


Figure 7.17 Dialog for two-sample *t*-test.

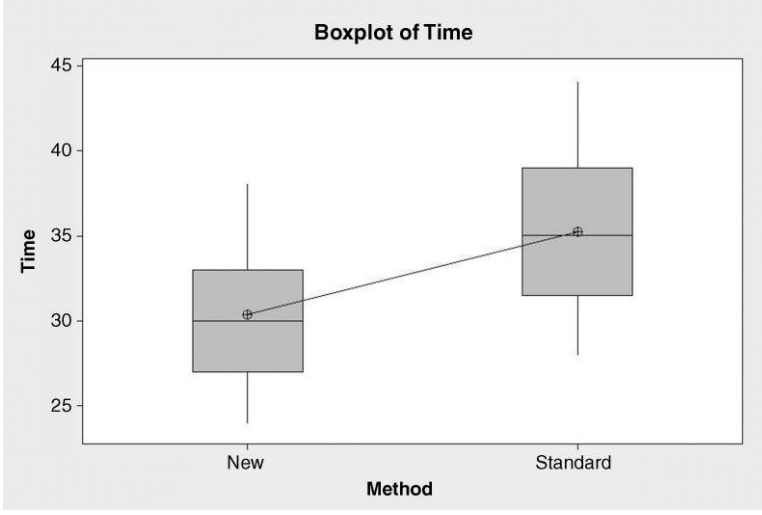


Figure 7.18 Boxplots of assembly time data.

this case. The graphical output is shown in Figure 7.18. The Session window output is shown in Panel 7.20.

The box sections of the boxplots are of similar length, indicating that the assumption of equal variances for the two populations is reasonable. Note that the sample means are also displayed and connected by a line segment. Normal probability plots of the two samples provide no evidence of nonnormality. Thus a two-sample t-test, with the assumption of equal variances, would appear to be a sound method of analysis.

The Session window output gives the summary statistics sample size, mean, standard deviation and standard error of the mean for each sample. The null and alternative hypotheses,

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 < \mu_2,$$

may be written in terms of the *difference* between the population means as

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_1 : \mu_1 - \mu_2 < 0.$$

Two-Sample T-Test and CI: Time, Method

```
Two-sample T for Time

Method     N   Mean  StDev  SE Mean
New        9  30.33   4.15      1.4
Standard   9  35.22   4.94      1.6


Difference = mu (New) - mu (Standard)
Estimate for difference:  -4.89
95% upper bound for difference:  -1.13
T-Test of difference = 0 (vs <): T-Value = -2.27  P-Value = 0.019  DF = 16
Both use Pooled StDev = 4.5659
```

Panel 7.20 Session window output for two-sample t-test.

Table 7.6 Magnesium assay data.

Method A	3.6	3.5	3.4	3.5				
Method B	2.6	2.9	3.5	2.7	3.8	3.2	2.8	3.3

Minitab states the hypotheses in this format in the Session window output in the rather cryptic shorthand 'difference = 0 (vs <)'. The *P*-value is 0.019 so, since this is less than 0.05, there is evidence, at the 5% level of significance, that the population mean assembly time for operators trained by the new method is lower than that for operators trained by the standard method. The statement 'Estimate for difference: −4.89' indicates that the estimated reduction in the mean assembly time is 4.89 minutes. The statement '95% upper bound for difference: −1.13' indicates that, with 95% confidence, it may be stated that the reduction in the mean is at least 1.13 minutes.

Each sample of nine observations has 8 degrees of freedom, yielding a total of 16 degrees of freedom, indicated by 'DF = 16' in the output. A common variance was assumed for the two populations; the final component of the output is an estimate of this common variance. Further detail may be found in Montgomery (2009, pp. 132–134) or Hogg and Ledolter (1992, p. 236).

As a second example consider the data in Table 7.6 on determinations of the percentage of magnesium in a batch of ore by two chemical assay procedures. Performance of a two-sample *t*-test of the hypotheses $H_0: \mu_A = \mu_B$ and $H_1: \mu_A \neq \mu_B$, with equal variances for the two populations of determinations assumed, yields the Session window output shown in Panel 7.21. The data are available in Magnesium.MTW in two separate columns named A and B. In this case, as the data appear in separate columns, the **Samples in different columns** option is required with **First:** A and **Second:** B specified. **Assume equal variances** was checked. **Individual value plot** was selected under **Graphs...**.

The null hypothesis $H_0: \mu_A = \mu_B$ may be stated in the form $H_0: \mu_A - \mu_B = 0$, i.e. that the difference in the population means is zero. The alternative hypothesis $H_1: \mu_A \neq \mu_B$ may be stated in the form $H_1: \mu_A - \mu_B \neq 0$, i.e. that the difference in the population means is nonzero. In the Session window output the hypotheses are indicated by 'difference = 0 (vs not =)'. The 95% confidence interval (−0.084, 0.884) for $\mu_A - \mu_B$ includes the value 0, which indicates that the null hypothesis cannot be rejected at the 5% significance level. This conclusion is confirmed by the *P*-value of 0.096 being in excess of 0.05. This analysis suggests that both methods of determining the magnesium content of the ore would yield the same mean value from many repeated assays.

Two-Sample T-Test and CI: A, B

```
Two-sample T for A vs B

   N     Mean    StDev  SE Mean
A  4   3.5000   0.0816    0.041
B  8    3.100    0.421     0.15


Difference = mu (A) - mu (B)
Estimate for difference:  0.400
95% CI for difference:  (-0.084, 0.884)
T-Test of difference = 0 (vs not =): T-Value = 1.84  P-Value = 0.096  DF = 10
```

Panel 7.21 Session window output for two-sample *t*-test – equal variances assumed.

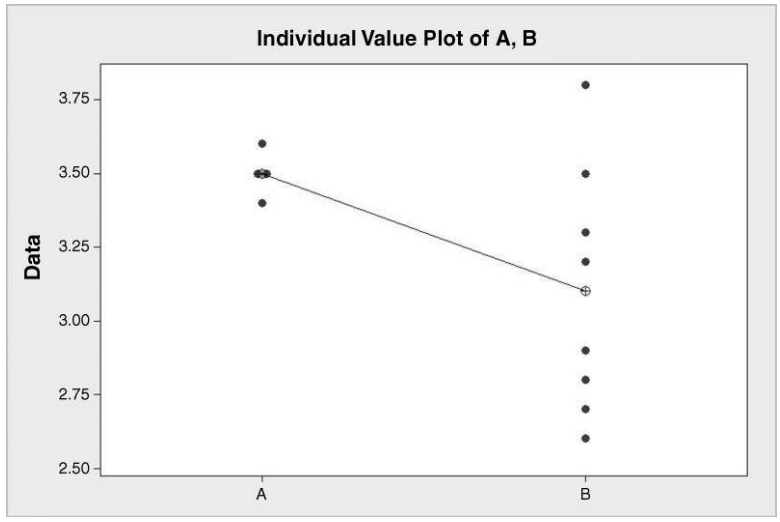


Figure 7.19 Dotplots of magnesium assay data.

Although normal probability plots provide no evidence of nonnormality, the individual value plots of the data displayed in Figure 7.19 cast doubt on the assumption of equal variances, the spread of the method A data being much greater than that of the method B data. Performing the test again, with **Assume equal variances** unchecked, yields the Session window output in Panel 7.22. This version of the test provides evidence, at the 5% level of significance, that the population means differ for the two methods of assay. Later in the chapter we will look at tests of hypotheses concerning variances.

In the examples considered the null hypothesis has been that the two population means are equal, i.e. that the difference between the population means is 0. It is possible to test the null hypothesis that the difference between the population means is some value other than 0. For example, if it is claimed that the use of high octane fuel would improve the fuel consumption of a type of vehicle by 5 mpg on average over that obtained with regular octane fuel then the null hypothesis would be

$$H_0 : \mu_{\text{High}} = \mu_{\text{Low}} + 5, \quad \text{i.e. } \mu_{\text{High}} - \mu_{\text{Low}} = 5.$$

Two-Sample T-Test and CI: A, B

```
Two-sample T for A vs B

   N     Mean    StDev  SE Mean
A  4   3.5000   0.0816    0.041
B  8    3.100    0.421     0.15


Difference = mu (A) - mu (B)
Estimate for difference:  0.400000
95% CI for difference:  (0.035131, 0.764869)
T-Test of difference = 0 (vs not =): T-Value = 2.59  P-Value = 0.036  DF = 7
```

Panel 7.22 Session window output for two-sample *t*-test – equal variances not assumed.

7.3.2 Tests for two proportions

A manufacturer of laptop computers claims that a higher proportion of his machines will be operating without any hardware faults after 1 year than those of a competitor. A multinational company which had purchased a large number of machines from both manufacturers established that, of a sample of 200 machines from the manufacturer making the claim, 13 had experienced hardware faults during the first year while, of a sample of 150 produced by the rival manufacturer, 19 had experienced hardware faults during the first year. In order to put the manufacturer's claim to the test formally one can proceed as follows.

Let p_1 represent the proportion of the manufacturer's machines that develop hardware faults during the first year and let p_2 represent the proportion for the competitor. Our hypotheses are as follows:

$$H_0 : p_1 = p_2 \quad \text{or} \quad p_1 - p_2 = 0,$$
$$H_1 : p_1 < p_2 \quad \text{or} \quad p_1 - p_2 < 0.$$

The test may be performed using **Stat > Basic Statistics > 2 Proportions...**. The dialog required is shown in Figure 7.20. It is recommended that the pooled estimate of a common proportion be used for the test (Montgomery, 2009, p. 139; Hogg and Ledolter, 1992, p. 242). Thus **Use pooled estimate of p for test** should be checked under **Options...**, together with **Alternative: less than**.

The Session Window output is shown in Panel 7.23. A summary of the data provided is given indicating that, to two decimal places, 6.5% of the manufacturer's machines developed hardware faults within a year while 12.7 % of the competitor's machines developed hardware faults within a year. The *P*-value of 0.024 provides evidence, at the 5% level of significance, that the manufacturer's claim is true. The point estimate of the difference is 6.1% fewer machines developing hardware faults within a year for the manufacturer compared with the competitor. The confidence interval for the difference indicates that at least 0.9% fewer of the manufacturer's machines developed hardware faults within a year (after rounding –0.008 587 15 to three significant figures, converting to a percentage and

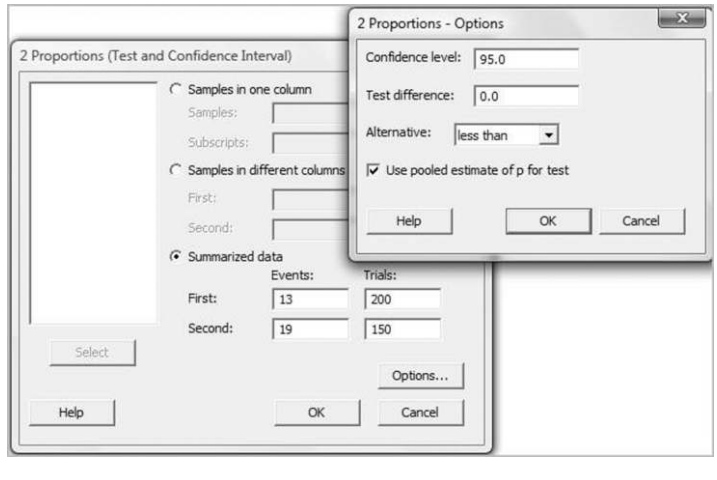


Figure 7.20 Dialog for hypothesis test of two proportions.

Test and CI for Two Proportions

```
Sample   X    N  Sample p
1       13  200  0.065000
2       19  150  0.126667

Difference = p (1) - p (2)
Estimate for difference:  -0.0616667
95% upper bound for difference:  -0.00858715
Test for difference = 0 (vs < 0):  Z = -1.98  P-Value = 0.024

Fisher's exact test: P-Value = 0.037
```

Panel 7.23 Session window output for test of proportions.

interpreting the negative to imply fewer). The z-value quoted is the test statistic used. The P-value of 0.037 results from application of the alternative test of the hypotheses provided by Fisher's exact test and leads to the same conclusion.

7.3.2.1 Power and sample size

Minitab enables power and sample size calculations to be performed for the two-sample t-test and for tests concerning two proportions. Suppose that yields from a batch chemical process are known to be normally distributed and to vary with a standard deviation of the order of 5 kg under a wide variety of operating conditions. Suppose also that discussions with the process team reveal that a switch to a new catalyst would be viable from an economic point of view if the mean yield per batch were to increase by 4 kg. Using **Stat > Power and Sample Size > 2-Sample t-test...**, once can determine the sample size required to perform a two-sample t-test of the null hypothesis $H_0: \mu_{\text{Standard}} = \mu_{\text{New}}$, i.e. $\mu_{\text{Standard}} - \mu_{\text{New}} = 0$, at significance level $\alpha = 0.05$ and with power = 0.9. The dialog is shown in Figure 7.21. Note that the difference is specified as -4 and that the alternative hypothesis is $H_1: \mu_{\text{Standard}} < \mu_{\text{New}}$, i.e. $\mu_{\text{Standard}} - \mu_{\text{New}} < 0$, that **Less than** is checked.

The Session window output is shown in Panel 7.24. The calculated sample size, for each group, is 28. Thus a sample of 28 yields from the process run with the standard catalyst and a sample of 28 yields from the process run with the new catalyst would be required.

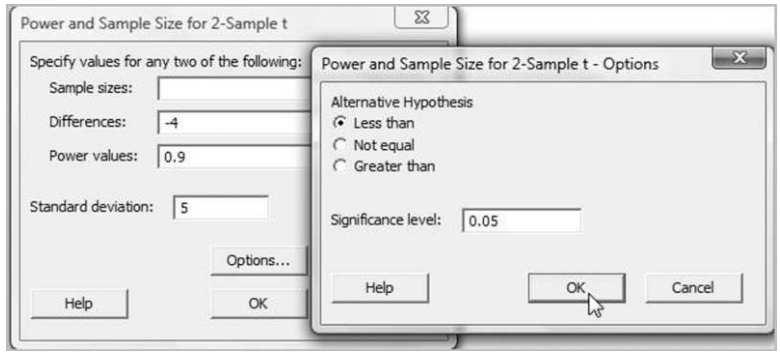


Figure 7.21 Dialog for sample size calculation.

Power and Sample Size

```
2-Sample t Test

Testing mean 1 = mean 2 (versus <)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Assumed standard deviation = 5

            Sample  Target
Difference    Size   Power  Actual Power
        -4      28     0.9      0.905010
```

Panel 7.24 Sample size calculation for two-sample *t*-test.

Suppose that a procurement manager wishes to ascertain whether there is evidence that the proportion of nonconforming items from supplier A is 2% lower than that obtained from supplier B, for whom records indicate that about 8% of items are nonconforming. Suppose that random samples of 500 items from each supplier were to be checked. The manager would like to know the power of a test performed at the 5% level of significance, based on these sample sizes, to detect superior performance by supplier A, by 2%, in the proportion of nonconforming items. Use of **Stat > Power and Sample Size > 2 Proportions...** provides the answer. The dialog is displayed in Figure 7.22.

The Session window output is shown in Panel 7.25. It gives the power of the test to detect superiority of supplier A, by 2%, as 0.34. This means that the probability of committing a Type II error is $1 - 0.34 = 0.66$, which indicates that if supplier A is actually operating with a nonconformance rate of 6%, as compared with 8% for supplier B, then there is probability of approximately 0.66 that the test would fail to provide evidence of the difference. If you decide that you would like the power to be 0.9 then, by specifying this value in the dialog shown in Figure 7.22 and clearing the box **Sample sizes:**, the procedure returns a sample size of 2786. The reader is invited to verify this as an exercise. People involved in quality improvement often fail to realize the size of sample required to formally detect changes or differences of a magnitude that is of practical significance to their organization.

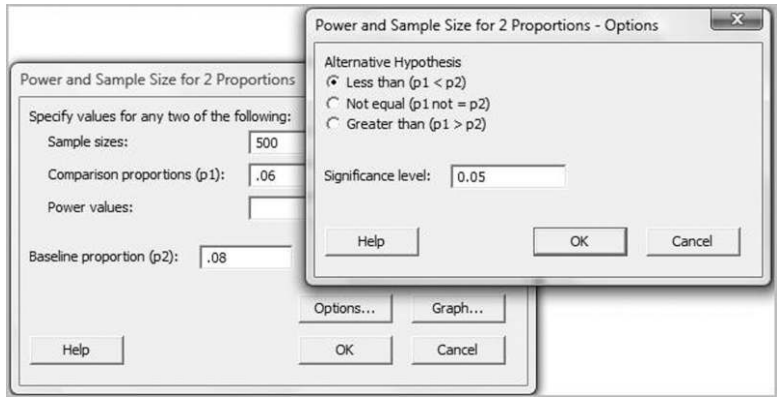


Figure 7.22 Dialog for power calculation.

Power and Sample Size

```
Test for Two Proportions

Testing proportion 1 = proportion 2 (versus <)
Calculating power for proportion 2 = 0.08
Alpha = 0.05

               Sample
Proportion 1     Size     Power
        0.06      500  0.342455

The sample size is for each group.
```

Panel 7.25 Sample size calculation for test of two proportions.

7.3.3 Nonparametric Mann–Whitney test

This nonparametric test provides an alternative to the two-sample *t*-test. It is based on allocating ranks to the combined data from both samples. It is also referred to as the two-sample rank test or the two-sample Wilcoxon rank-sum test. The null hypothesis is that the two population medians are equal. The assumptions required are that the data are independent random samples from two distributions that have the same shape.

Consider the data in Table 7.7 on the length of drive (in metres) achieved on striking golf balls of two different types with a mechanical club device used by a golf manufacturer for product testing purposes. The manufacturer's quality manager wishes to know if the data provide evidence that the median length of drive achieved with type A is less than that achieved with type B.

The test involves ranking the combined data is shown in Table 7.8. In this case the actual observed rank sum for type A is $W = 1 + 2 + 3 + 4 + 8 = 18$. Had all the type A distances been less than all the type B distances then the rank sum for type A would have been

Table 7.7 Distance driven (m) for two types of golf ball.

Type A	181	183	176	221	180
Type B	215	197	229	222	195

Table 7.8 Ranked data for distance.

Distance	Type	Rank
176	Type A	1
180	Type A	2
181	Type A	3
183	Type A	4
195	Type B	5
197	Type B	6
215	Type B	7
221	Type A	8
222	Type B	9
229	Type B	10

$W = 1 + 2 + 3 + 4 + 5 = 15$. In this case one might feel that there was no need for a formal test of hypotheses! The reader might wish to take a few moments to convince him/herself that the possible ranks for type A that would give rise to a rank sum of 18 or less are as follows:

$$W = 1+2+3+4+5 = 15,$$
$$W = 1+2+3+4+6 = 16,$$
$$W = 1+2+3+4+7 = 17,$$
$$W = 1+2+3+5+6 = 17,$$
$$W = 1+2+4+5+6 = 18,$$
$$W = 1+2+3+4+8 = 18,$$
$$W = 1+2+3+5+7 = 18.$$

Thus there are seven possible rankings for type A yielding a rank sum of 18 or less. Combinatorial mathematics indicates that there are 252 possible rankings for type A that can arise when two samples of size 5 are tested. Were the null hypothesis true then each ranking for type A would have equal probability of occurring in the experiment and $P(W \leq 18) = 7/252 = 0.0278$. Since this probability is less than 0.05, the null hypothesis that the medians are equal would be rejected in favour of the alternative hypothesis that the median for type A is less than that for type B. (The Greek letter η (eta) may be used to denote a population median.)

With the data for each type in separate columns in a Minitab worksheet the test may be performed using **Stat** > **Nonparametrics** > **Mann-Whitney...** with **Alternative: less than** selected. The Session window output is shown in Panel 7.26. It begins by giving the sample sizes and sample medians by way of data summary. Next follows the point estimate for the difference in the population medians – note that this is not the difference between the sample medians. (Interested readers will find details of the estimation procedure employed in Minitab via the Help facility.) Even though a one-sided alternative hypothesis was specified here, Minitab gives an approximate two-sided 96.3% confidence interval. (Since it includes zero we have an indication that the null hypothesis of equal population medians would not be rejected in favour of the alternative hypothesis of unequal population medians at the $(100 - 96.3)\% = 3.7\%$ significance level.) Then the rank sum $W = 18$ found earlier is stated. (If there are two or more equal values in the combined data set then the mean of the associated ranks is allocated to these equal values. Thus noninteger values of W may occur.) Finally, the null and alternative hypotheses are stated and the P-value is given as 0.0301 rather than 0.0278,

Mann-Whitney Test and CI: A, B

```
   N  Median
A  5  181.00
B  5  215.00


Point estimate for ETA1-ETA2 is -21.00
96.3 Percent CI for ETA1-ETA2 is (-48.00,5.98)
W = 18.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0301
```

Panel 7.26 Session Window output for Mann-Whitney test.

the value calculated above. This is because Minitab uses an approximate method, based on the normal distribution, to compute the probability, and not because the author has made an error!

In discussing the Mann–Whitney test Daly *et al.* (1995, pp. 372–373) write:

> The idea of using ranks instead of the data values is an appealing one. Furthermore, it has an obvious extension to testing two groups of data when a two-sample *t*-test may not be applicable because of lack of normality. The test itself was first proposed by H.B. Mann and D.R. Whitney in 1947, and modified by Wilcoxon; it turns out to be very nearly as powerful as the two-sample *t*-test, which tests for equal means. However it is nevertheless a test of the equality of locations of the two groups and using it as an alternative to the two-sample *t*-test is an approximation often made in practice.

7.4 The analysis of paired data – *t*-tests and sign tests

A finance company gave a group of employees a test before and after a refresher course on tax legislation. The scores obtained are displayed in Table 7.9 and are available in the worksheet TaxTest.MTW.

In order to evaluate the evidence for an improvement in the knowledge of the employees there are two approaches. The first approach is to form the differences obtained by subtracting, for each employee, the score obtained before the course from the score obtained after the course. If the differences may reasonably be regarded as a random sample from the normal distribution $N(\mu, \sigma^2)$ then the evidence may be evaluated by using a one-sample *t*-test of the null hypothesis $H_0 : \mu = 0$ versus the alternative hypothesis $H_1 : \mu > 0$. Given the two columns of before and after scores, **Calc > Calculator** may be used to form the differences and subsequently **Stat > Basic Statistics > 1-Sample t** ... may be used to perform the *t*-test. However, Minitab provides **Stat > Basic Statistics > Paired t** ... to enable the test to be performed via a single dialog as displayed in Figure 7.23.

Table 7.9 Test scores before and after refresher course.

Employee number	Score before	Score after	Difference x	Sign
1	48	58	10	+
2	87	91	4	+
3	82	81	−1	−
4	44	55	11	+
5	56	60	4	+
6	71	68	−3	−
7	60	66	6	+
8	66	82	16	+
9	84	89	5	+
10	48	55	7	+
11	63	73	10	+
12	48	49	1	+

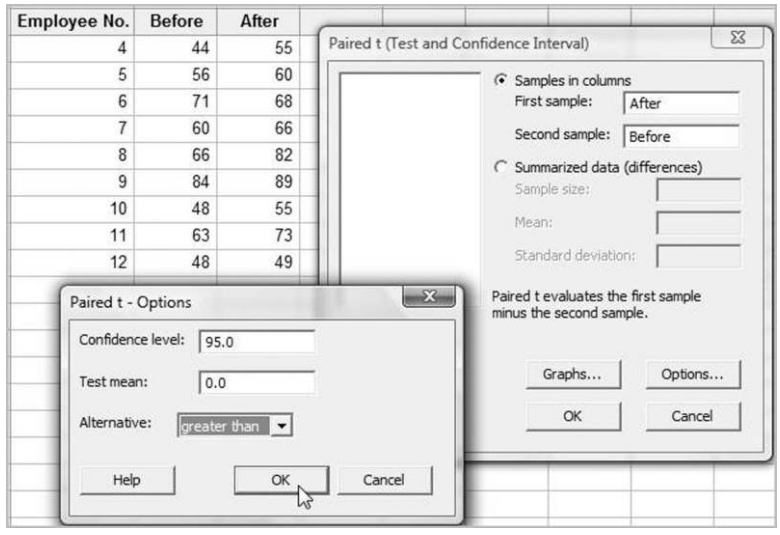


Figure 7.23 Dialog for paired *t*-test.

It is important to note the statement in the main dialog box: **Paired t evaluates the first sample minus the second sample**. Thus care has to be taken in specifying which column is deemed to contain the first sample data and which deemed to contain the second sample data and in specifying the hypothesis of interest in relation to that choice. It could be argued that the natural thing to do would be to specify the pre-course scores as the first sample; the reason the author chose the reverse was that positive differences then correspond to improvement. The alternative hypothesis is specified in the usual way using the **Options...** subdialog box and, as ever, creation of some form of display of the data using **Graphs...** is recommended. Here an individual value plot was selected. (The author edited the symbols so that an employee whose score after was higher that his/her score before is represented by an upward pointing triangle and so that an employee whose score after was lower that his/her score before is represented by an downward pointing triangle. To edit all graph symbols click on a symbol, pause and double-click to access the **Edit Individual Symbols** menu. To edit a single graph symbol click on it, pause, click again, pause and double-click to access the **Edit Individual Symbols** menu.)

The individual value plot of the differences in Figure 7.24 indicates that the scores increased for 10 of the 12 employees but decreased for the remaining two. The fact that the value of 0, specified under the null hypothesis H_0, lies outwith the line segment representing the one-sided 95% confidence interval for the mean of the population of differences indicates that the null hypothesis would be rejected in favour of the alternative at the 5% level of significance.

The Session Window output is shown in Panel 7.27. Summary statistics are given for the two sets of scores and for the differences. The *P*-value of 0.002 indicates that the null hypothesis would be rejected in favour of the alternative at the 1% level of significance. Rounded to the nearest whole number, the point estimate of the mean increase in score is 6 and, with 95% confidence, it can be stated that the mean increase in score is at least 3 points. A normal probability plot of the difference data provides no evidence of nonnormality, so the *t*-test is an appropriate method of analysis.

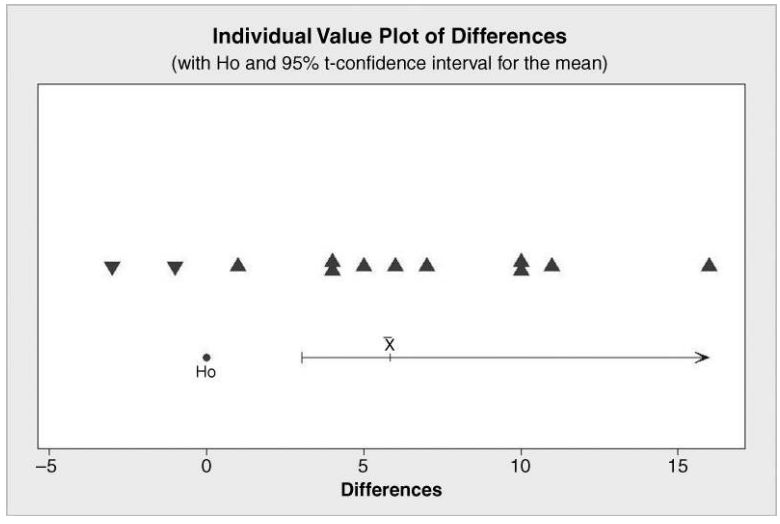


Figure 7.24 Display of differences with *t*-test annotation.

Paired T-Test and CI: After, Before

```
Paired T for After - Before

             N   Mean  StDev  SE Mean
After       12  68.92  14.20     4.10
Before      12  63.08  15.21     4.39
Difference  12   5.83   5.41     1.56

95% lower bound for mean difference: 3.03
T-Test of mean difference = 0 (vs > 0): T-Value = 3.74  P-Value = 0.002
```

Panel 7.27 Session window output for paired *t*-test.

Alternatively, the nonparametric sign test may be used to test the null hypothesis that the median of the population of differences is 0 against the alternative hypothesis that the median is greater than 0. The Session window output from this test is shown in Panel 7.28. With *P*-value 0.0193 the null hypothesis that the median population difference is 0 would be rejected in favour of the alternative hypothesis that the median is greater than 0 at the 5% level of significance. Thus the sign test also provides evidence that the refresher course has improved the employees' knowledge of the tax legislation. The sign test is a less powerful test than the

Sign Test for Median: Difference

```
Sign test of median =  0.00000 versus > 0.00000

             N  Below  Equal  Above       P  Median
Difference  12      2      0     10  0.0193   5.500
```

Panel 7.28 Session window output for sign test.

paired t-test. If one is concerned about normality of the distribution of differences then the sign test is available as a nonparametric, but less powerful, alternative to the paired t-test.

Had one erroneously analysed the data using the two-sample t-test then no evidence of a significant impact of the refresher course would have been found. This emphasizes the fact that the two sets of scores do not constitute independent random samples from two normal populations and also the need for care in the selection of methods for the analysis of data. Paired experiments of the type discussed here are a special case of the use of blocking in the design of experiments. This powerful technique will be discussed in detail later in the chapter.

7.5 Experiments with a single factor having more than two levels

We will now look at situations where the factor of interest, X, has more than two levels. In some cases the effects of the factor are *fixed*. For example, suppose there are only three adhesives available on the market that may be used to bond components to a substrate in the fabrication of a particular type of electronic circuit. For an experiment in which bond strength, Y, was measured for 10 components bonded to substrate with each available adhesive, the factor adhesive would be said to be fixed. In some cases the effects of the factor are said to be *random*. Were there many adhesives available then, for an experiment in which bond strength, Y, was measured for 10 components bonded to substrate with each adhesive from a sample of three adhesives, selected at random from the available adhesives, the factor adhesive would be said to be random.

In analysing data from fixed effect scenarios, interest centres on testing hypotheses concerning means and making comparisons between means. Thus in the case of there being only three adhesives, all involved in the experiment, the questions being addressed would be:

- Is there evidence that the population mean bond strengths differ for the three available adhesives?
- If the answer to the first question is an affirmative, then what is the extent of the differences?

In analysing data from random effects scenarios, interest centres on partitioning the variation observed into components. Thus in the case of there being a random sample of three adhesives involved in the experiment, the questions being addressed would be:

- How much of the variation observed is attributable to real differences between means in the population of adhesives from which the three used in the experiment were selected?
- How much of the variation observed is attributable to random variation about these population means?

7.5.1 Design and analysis of a single-factor experiment

In order to introduce key concepts and techniques data for a fictitious golfer, Lynx Irons, will be used. Lynx is interested in improving the process of driving a golf ball from the tee at holes

where she needs to use a driving club in order to achieve maximum distance. She wishes to determine the effect of ball type on the length of her drives. She will hit a number of drives with each of a *fixed* set of ball types she is prepared to use – Exoset, Flyer and Gutty (E, F and G for short).

Coleman *et al.* (1996, pp. 137–141) refer to requirements of good experimentation including:

- reliable measurement,
- randomization,
- replication.

Let us assume that we can reliably measure the response, Y, of interest – the length of drive in metres. Suppose that it has been decided to incorporate replication by having Lynx hit five balls of each type. (Were she to hit only a single ball of each type there is a risk that the distance achieved with one particular type might be atypically low and lead to failure to identify an opportunity for process improvement.) Finally, let us assume that all balls of a particular type are absolutely identical – unrealistic, of course, but necessary to make the example tractable. To achieve randomization the 15 balls could be put into a bag, given a thorough mix and a ball selected in turn for each drive. Were Lynx asked to hit all five balls of type E first, then all five of type F second and finally all five of type G then, for example, fatigue might lead to lower drive length for type G than might otherwise be obtained.

Minitab may also be used to carry out the randomization. Having decided to hit five balls of each type, this can be achieved by setting up columns as shown in the background in Figure 7.25. **Calc** > **Make Patterned Data** > **Simple Set of Numbers...** and **Calc** > **Make**

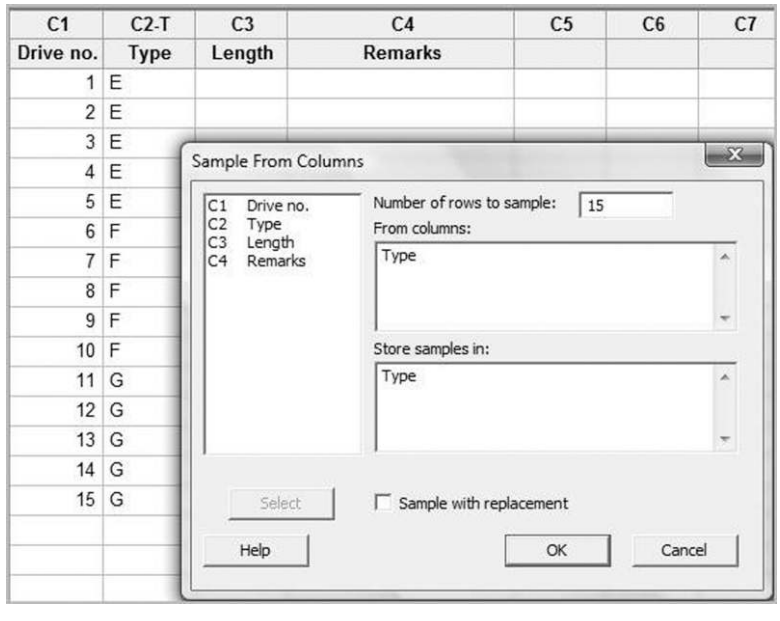


Figure 7.25 Initial worksheet for golf ball experiment with dialog for randomization.

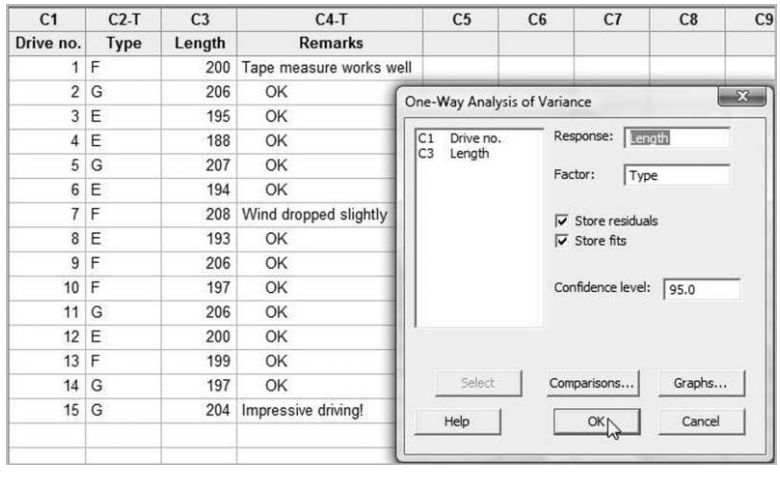
Patterned Data > **Text Values...** may be used to set up the first two. (It is a simple matter to make the required entries via the keyboard, but experience of using the facilities for creating patterned data is worth having. It should be noted that in creating a column named Drive no. the entry 'Drive no.' is required in the **Store patterned data in:** window.) The third will be use to record the length of each drive. The final column may be used to note any unusual occurrences during the conduct of the experiment or any information that might be relevant.

To achieve randomization, use may be made of **Calc** > **Random Data** > **Sample From Columns...** via the dialog shown in Figure 7.25. By sampling, without replacement, 15 values at random from the entries in column 2 (**Sample with replacement** must not be checked) and using the same column to store the results, the original entries in column 2 are rearranged into random order. (If the reader tries this for her/himself it is unlikely, but possible, that the same sequence will be obtained as that obtained by the author in Figure 7.26!)

The resulting worksheet may then be stored in a project file and also printed off as a pro forma for the recording of the length of the 15 drives at the golf range where the experiment is to be performed. The data are displayed in Figure 7.26 and are available in Types.MTW.

We could analyse these data formally using three two-sample *t*-tests - one to compare E with F, one to compare F with G, and a final one to compare G with E.

Had Lynx wished to investigate seven ball types this approach would have required 21 two-sample *t*-tests in total. Apart from the tedium, there is a problem with this approach. When employing a 5% significance level there is a 5%, or 1 in 20, probability of a Type I error, i.e. of rejecting a null hypothesis when in fact that hypothesis is true. Thus with seven ball types, which in reality have identical mean distances for Lynx, we would expect the *t*-test approach to throw up spurious evidence of a significant difference between two of the ball types. We are now going to look at a *single* analysis that will seek evidence from the data of a real difference between ball types as far as mean length for Lynx is concerned. The technique is *analysis of variance* (ANOVA).



C1	C2-T	C3	C4-T	C5	C6	C7	C8	C9
Drive no.	Type	Length	Remarks					
1	F	200	Tape measure works well					
2	G	206	OK					
3	E	195	OK					
4	E	188	OK					
5	G	207	OK					
6	E	194	OK					
7	F	208	Wind dropped slightly					
8	E	193	OK					
9	F	206	OK					
10	F	197	OK					
11	G	206	OK					
12	E	200	OK					
13	F	199	OK					
14	G	197	OK					
15	G	204	Impressive driving!					

Figure 7.26 Data from golf ball experiment and dialog for ANOVA.

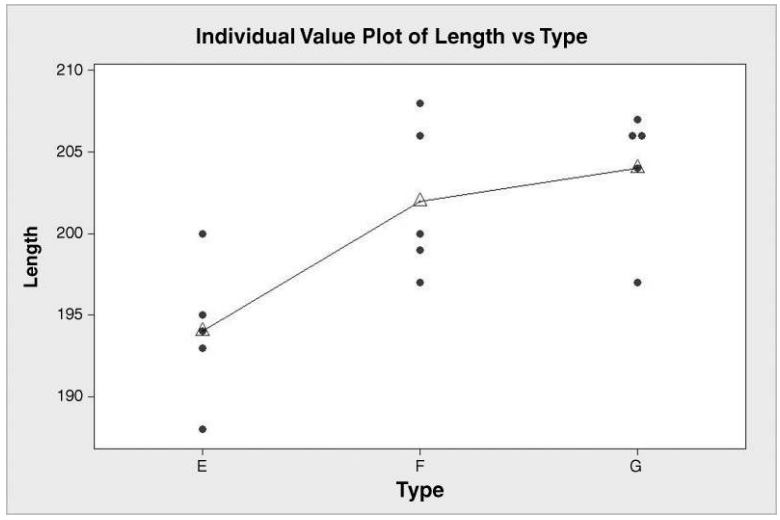


Figure 7.27 Individual value plot of length by type.

The null hypothesis is that the population means are equal and the alternative is that not all the means are identical:

$$H_0 : \mu_E = \mu_F = \mu_G,$$
$$H_1 : \text{Not all } \mu\text{s are identical.}$$

The ANOVA can be performed using **Stat** > **ANOVA** > **One-Way...**, the 'one-way' indicating that there is only a single factor of interest in this experiment. The dialog required is also displayed in Figure 7.26. The **Response:** (Y) is length and the **Factor:** (X) is type of ball. **Store residuals** and **Store fits** have both been checked. The default **Confidence level:** of 95% has been accepted. In order to display the data **Individual value plot** was selected under **Graphs...** together with **Four in one** for the **Residual plots**. Finally under **Comparisons... Tukey's, family error rate:** 5 (the default 5%), was selected. The output will now be discussed step by step.

The individual value plot of the data is shown in Figure 7.27. The plot suggests that ball types F and G are on a par (no pun intended!) as far as length performance for Lynx Irons is concerned, while E is inferior to both. The triangular symbols denote the mean lengths of the five drives with each ball type. The author edited the crossed circle symbols obtained by default. The mean value for both E and G equals an observed value.

The Session window output enables a formal test of the null hypothesis of equal population means to be tested against the alternative specified above. The relevant section of the output is presented in Panel 7.29. This section of the output is the ANOVA table. Its construction will

One-way ANOVA: Length versus Type

Source	DF	SS	MS	F	P
Type	2	280.0	140.0	7.30	0.008
Error	12	230.0	19.2		
Total	14	510.0			

Panel 7.29 ANOVA table for golf ball experiment.

be explained later in the chapter. The test statistic is the value 7.30 under the heading F and the associated *P*-value of 0.008 is given in the next column. Since the *P*-value is less than 0.01 null hypothesis would be rejected in favour of the alternative hypothesis at the 1% level of significance. Thus the experiment provides strong evidence that Lynx does not achieve the same mean drive length with the three types of golf ball. In view of the appearance of the individual value plot, this conclusion is not surprising.

The theory underlying the test requires that the three populations of length, for the three ball types driven by Lynx, have normal distributions with equal variances. If these requirements are met and the null hypothesis is true then the test statistic has an *F*-distribution that may be used to compute the *P*-value. The distribution is named in honour of Ronald Fisher, a pioneer in developing the application of statistical methods to experimentation. Since the spreads of the points in the individual value plot are similar for each ball type, the assumption of equal variances would appear to be a reasonable one. Some descriptive statistics for length and for length by type are shown in Panel 7.30. The overall mean length for all 15 drives was 200 m.

Were we to examine 95% confidence intervals for all the differences between population means, based on two-sample *t*-tests, then we would encounter a similar problem to that which would arise were we to compare means using a series of two-sample *t*-tests. If we require a statement with 95% confidence for *all* possible differences between population means then Tukey's multiple comparison provides this. The corresponding section of the Session window output is shown in Panel 7.31. Thus with overall 95% confidence we can state the following:

- The population mean length achieved with type F is greater than that achieved with type E by 8 m, with confidence interval (0.6, 15.4)
- The population mean length achieved with type G is greater than that achieved with type E by 10 m, with confidence interval (2.6, 17.4)
- The population mean length achieved with type G is greater than that achieved with type F by 2 m, with confidence interval (−5.4, 9.4).

The first two confidence intervals do not include 0, while the third does. Thus, with overall 95% confidence, we can state that there is evidence from the experiment that F is superior to E and that G is superior to E, as far as the mean length of drive for Lynx is concerned. (We can make this statement since the corresponding confidence intervals do not include the value 0 and the intervals cover ranges of positive values.) There is no evidence of any difference between F and G in this respect. (The corresponding confidence interval includes 0.) Thus we use ANOVA to

Descriptive Statistics: Length

Variable	Mean	StDev
Length	200.00	6.04

Descriptive Statistics: Length

Variable	Type	Mean	StDev
Length	E	194.00	4.30
	F	202.00	4.74
	G	204.00	4.06

Panel 7.30 Descriptive statistics for length and for length by ball type.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Type

Individual confidence level = 97.94%

Type = E subtracted from:

Type  Lower  Center   Upper  ---+---------+---------+---------+------
F     0.619   8.000  15.381                (---------*----------)
G     2.619  10.000  17.381                   (---------*----------)
                            ---+---------+---------+---------+------
                            -7.0        0.0       7.0      14.0

Type = F subtracted from:

Type   Lower  Center  Upper  ---+---------+---------+---------+------
G     -5.381   2.000  9.381       (----------*---------)
                            ---+---------+---------+---------+------
                            -7.0        0.0       7.0      14.0
```

Panel 7.31 Session window output for Multiple Comparison procedure.

look for evidence that the factor of interest (type of ball) influences the response (length). If such evidence is found then the follow-up using multiple comparisons establishes evidence of where the differences lie.

These conclusions are summarized by the software in the component of the Session window output displayed in Panel 7.32. The rows for types F and G have the letter A in common, indicating the earlier conclusion of no difference in mean length for these two types. The rows for E and F and the rows for E and G have no letters in common, indicating significant differences between E and F on the one hand and between E and G on the other. In essence group A consists of types F and G and group B consists of type E on its own.

The overall mean for the experiment was 200. The mean for type E was 194, so we can think of the effect for type E as being -6, i.e. using type E reduced mean length by 6 m from the overall mean. The mean for type F was 202, so we can think of the effect for type F as being 2, i.e. using type F increased mean length by 2 m from the overall mean. Similarly, for type G the effect was 4. Note that the three effects sum to 0. In fitting a statistical model to data it is usual to write

$$\text{Observed data value} = \text{Value fitted by model} + \text{Residual},$$

or, more succinctly,

$$\text{Data} = \text{Fit} + \text{Residual}.$$

```
Grouping Information Using Tukey Method

Type  N     Mean  Grouping
G     5  204.000  A
F     5  202.000  A
E     5  194.000    B

Means that do not share a letter are significantly different.
```

Panel 7.32 Session window summary of Tukey Multiple Comparisons procedure.

In this case we take

$$\text{Fit} = \text{Overall mean} + \text{Effect}.$$

Thus we have:

$$\begin{aligned}
&\text{type E,} \quad \text{Fit} = 200 + (-6) = 194;\\
&\text{type F,} \quad \text{Fit} = 200 + 2 = 202;\\
&\text{type G,} \quad \text{Fit} = 200 + 4 = 204.
\end{aligned}$$

Thus knowing the data and fit values we can compute the residual values as the difference Data – Fit.

The fit and residual values were computed by Minitab by checking **Store fits** and **Store residuals** in the ANOVA dialog. They are displayed in Figure 7.28 – Minitab assigns the names FITS1 and RESI1 to the columns containing the fitted values and residuals, respectively. The first drive was of length 200 with a ball of type F for which the fit is 202. Hence, Residual = Data − Fit = 200 − 202 = −2. The reader is invited to check the remaining residual values displayed in Figure 7.28.

The **Four in one** plot facility under **Graphs...** yields four plots that will be discussed in turn. They are displayed in Figure 7.29. Theoretically the ANOVA methods used in this chapter require the assumptions of independence, random samples and normal distributions with equal variances. The four plots can often indicate when these assumptions are suspect.

1. *Normal probability plot of residuals*. In this case the plot is reasonably linear so the normality assumption underlying the valid use of the *F*-distribution for testing the null hypothesis appears reasonable.

C1	C2-T	C3	C4-T	C5	C6
Drive no.	Type	Length	Remarks	RESI1	FITS1
1	F	200	Tape measure works well	-2	202
2	G	206	OK	2	204
3	E	195	OK	1	194
4	E	188	OK	-6	194
5	G	207	OK	3	204
6	E	194	OK	0	194
7	F	208	Wind dropped slightly	6	202
8	E	193	OK	-1	194
9	F	206	OK	4	202
10	F	197	OK	-5	202
11	G	206	OK	2	204
12	E	200	OK	6	194
13	F	199	OK	-3	202
14	G	197	OK	-7	204
15	G	204	Impressive driving!	0	204

Figure 7.28 Worksheet with columns of fits and residuals Columns.

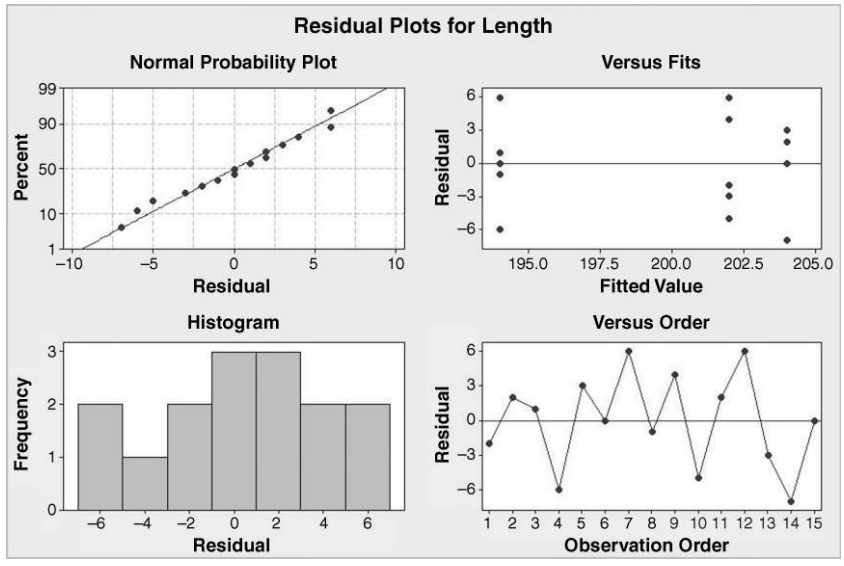


Figure 7.29 Four in one residual plots.

2. *Histogram of residuals*. A histogram of residuals that does not exhibit reasonable symmetry would suggest that the assumption of normality was suspect.

3. *Residuals versus the fitted values*. In addition to normality, the valid use of the F-distribution requires that the populations of lengths obtained with each type of ball have equal variances. Support for the assumption of equal variances is provided by similar vertical spread in the three sets of points.

4. *Residuals versus the order of the data*. In this case the data are in time order, so any unusual features in this run chart of the residuals could alert the experimenters to some factor other than type that might be having an influence on length.

These checks of the assumptions underlying valid use of the F-distribution are often referred to as *diagnostic checks*. Having the fits and residuals stored in the worksheet means that one can carry out one's own diagnostic checks, e.g. one could use **Stat > Basic Statistics > Normality Test...** in order to obtain a normal probability plot of the residuals with associated P-value. However visual scrutiny of the four plots discussed above will often be sufficient. The use of the F-test is fairly robust to relatively minor departures from the assumptions of normality and equal variances. However, if there are major concerns from scrutiny of the diagnostic plots one can either seek to transform the data or carry out a nonparametric test.

The portion of the Session window output in Panel 7.33 will now be considered. The number $s = 4.378$ is an estimate of the common standard deviation, assumed to apply to all three ball types. It is used to obtain, using the appropriate t-distribution, the 95% confidence intervals for the population means for the three ball types that are displayed to the right of the summary statistics. The R-sq (R^2) value of 54.9% is the coefficient of determination for length and fit expressed as a percentage. The reader may readily verify that the correlation between length and fitted value is 0.741 yielding $r^2 = 0.549 = 54.9\%$. It indicates that the model fitted to

```
S = 4.378   R-Sq = 54.90%   R-Sq(adj) = 47.39%

                          Individual 95% CIs For Mean Based on Pooled StDev
Level  N    Mean  StDev     -+---------+---------+---------+--------
E      5  194.00   4.30     (--------*--------)
F      5  202.00   4.74                   (--------*--------)
G      5  204.00   4.06                       (--------*--------)
                          -+---------+---------+---------+--------
                         190.0     195.0     200.0     205.0

Pooled StDev = 4.38
```

Panel 7.33 Portion of the Session window output from one-way ANOVA.

the data explains just over half the variation in length observed. The R-sq (adj) value will be discussed later in the book.

To sum up, the experiment has provided evidence that ball type influences length of drive for Lynx. The follow-up analysis indicates that, if the greatest achievable length is desirable, then she should use either type F or type G but not type E.

As a second example, consider a company involved in telesales that was running a Six Sigma project in order to improve the sales performance of its staff. A group of 40 new employees with similar educational backgrounds was split at random into four equal sized groups. The first group received the standard in-house training, while the other three groups were each trained by a different external training provider. The three external training providers comprise the list of accredited trainers for the company. Numbers of sales made by each employee during their first month of telephone contacts with prospective customers was recorded. The data are displayed in Table 7.10 and are available in the worksheet Sales.MTW.

The ANOVA for unstacked data in this form can be carried out using **Stat** > **ANOVA** > **One-way (Unstacked)...**. The dialog required is shown in Figure 7.30. Here the factor (X) of interest is the training with in-house, trainer P, trainer Q and trainer R as levels. The in-house trained sales staff may be regarded as a control group so, clicking on **Comparisons...**, **Dunnett's** multiple comparison method was selected. As there is no run order in this case Minitab offers, by clicking on **Graphs...**, a **Three in one** plotting option for the residuals from the fitted model, which was selected together with **Boxplots of data**. (These plots are not given

Table 7.10 Telesales data.

In-house	Trainer P	Trainer Q	Trainer R
71	62	55	71
71	62	67	72
59	82	67	90
67	82	71	94
66	64	62	80
45	70	71	80
58	71	72	77
58	71	69	76
55	82	58	75
67	93	53	80

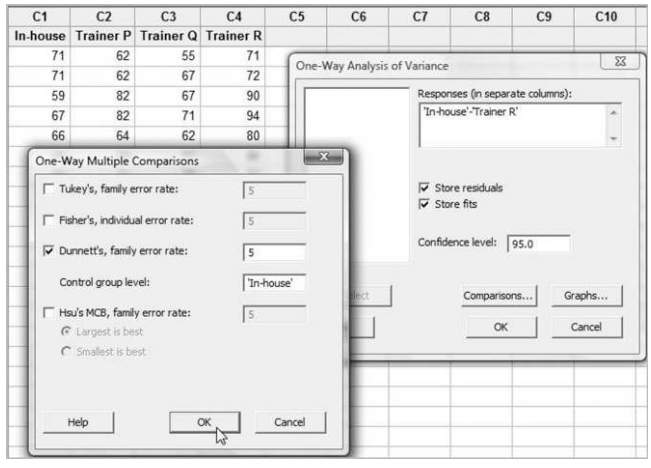


Figure 7.30 Dialog for ANOVA with unstacked data.

in the text.) Note that the default family error rate of 5% has been accepted and **Control group level:** 'In-house' indicates that employees trained in-house comprise the control group.

The reader is invited to verify that a P-value of 0.000 is obtained, which indicates that the actual P-value is less than 0.0005 (in fact it is 0.000 075). Thus the experiment provides strong evidence for rejection of the null hypothesis of equal mean sales performance for populations of employees trained by the four methods. The residual plots were deemed satisfactory. It should be noted that the data here are actually discrete but the residual plots indicate that the assumption of approximate normal distributions with equal variances is reasonable.

The Session window output for Dunnet's multiple comparison procedure is shown in Panel 7.34. The family error rate of $\alpha = 0.05$ means that there is an overall probability of 5% of a Type I error, which means in turn that we can have 95% confidence in the group of three confidence intervals provided. This can be interpreted to mean that, were we to repeat the experiment over and over again, on 5 occasions out of 100 in the long term the three confidence intervals would fail to capture all three true differences between the trainer means and the in-house mean. Conversely, on 95 occasions out of 100 in the long term the three

```
Dunnett's comparisons with a control

Family error rate = 0.05
Individual error rate = 0.0192

Critical value = 2.45

Control = In-house

Intervals for treatment mean minus control mean

Level       Lower  Center   Upper  ------+---------+---------+---------+---
Trainer P   3.036  12.200  21.364           (--------*--------)
Trainer Q  -6.364   2.800  11.964  (--------*--------)
Trainer R   8.636  17.800  26.964                   (--------*--------)
                                   ------+---------+---------+---------+---
                                         0         10        20        30
```

Panel 7.34 Dunnett's multiple comparisons for training experiment.

confidence intervals would capture all three true differences between the trainer means and the in-house mean. The individual error rate of $\alpha = 0.0192$ means that, were we to repeat the experiment over and over again, on 192 occasions out of 10 000 in the long term an individual confidence interval would fail to capture the true difference between the trainer mean and the in-house mean. Conversely, on 9808 occasions out of 10 000 in the long term an individual confidence interval would capture the true difference between the trainer mean and the in-house mean.

Thus from the experiment we estimate (rounded to the nearest integer) that the trainer P population would achieve 12 more sales in a month on average than the in-house population, with confidence interval (3, 21), the trainer Q population would achieve 3 more on average than the in-house population, with confidence interval (−6, 12), and the trainer Q population would achieve 18 more on average than the in-house population, with confidence interval (9, 27). Since the confidence interval for trainer Q includes 0, this indicates that there is insufficient evidence to conclude that employees trained by trainer Q will perform any better than those trained in-house. Since the confidence intervals for trainers P and R do not include 0 and cover positive ranges of values, this indicates that there is evidence that employees trained by trainer P and R will perform better than those trained in-house. This information is of potential value to the Six Sigma project team. The summary provided by the software is displayed in Panel 7.35.

7.5.2 The fixed effects model

Consider again the golf ball experiment where there was a *fixed* number, $a = 3$, of ball types of interest. Thus the factor, X, of interest had $a = 3$ levels. The response, Y, of interest was the length of drive achieved. There were $n = 5$ replications, i.e. the response was measured for five drives with a ball of each type. The underlying statistical model assumed was that, for each type, the response Y was normally distributed with the same variance for all three types. Thus the model states that:

$$\begin{aligned} &\text{for type E,} \quad Y \sim N(\mu_{\mathrm{E}}, \sigma^2); \\ &\text{for type F,} \quad Y \sim N(\mu_{\mathrm{F}}, \sigma^2); \\ &\text{for type G,} \quad Y \sim N(\mu_{\mathrm{G}}, \sigma^2). \end{aligned}$$

With this formulation of the model the null and alternative hypotheses are stated as follows:

$$H_0 : \mu_{\mathrm{E}} = \mu_{\mathrm{F}} = \mu_{\mathrm{G}}, \quad H_1 : \text{Not all } \mu\text{s are identical.}$$

```
Grouping Information Using Dunnett Method

Level                 N    Mean  Grouping
In-house (control)   10  61.700  A
Trainer R            10  79.500
Trainer P            10  73.900
Trainer Q            10  64.500  A

Means not labeled with letter A are significantly different from control level mean.
```

Panel 7.35 Summary of conclusions from experiment.

Table 7.11 ANOVA table for golf ball experiment with expected mean squares (fixed effects).

Source of variation	Degrees of freedom (DF)	Sum of squares (SS)	Mean square (MS)	Expected mean square (EMS)
Type	2	280	140	$\sigma^2 + \frac{n\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Error	12	230	19.2	σ^2
Total	14	510		

It is customary to write μ_E as $\mu + \alpha_1$, μ_F as $\mu + \alpha_2$ and μ_G as $\mu + \alpha_3$, where μ is referred to as the *overall mean* and the *effects* α_1, α_2, and α_3 are such that $\alpha_1 + \alpha_2 + \alpha_3 = 0$. Thus the model states that:

$$\begin{aligned} &\text{for type E,} \quad Y \sim N(\mu + \alpha_1, \sigma^2) \\ &\text{for type F,} \quad Y \sim N(\mu + \alpha_2, \sigma^2) \\ &\text{for type G,} \quad Y \sim N(\mu + \alpha_3, \sigma^2). \end{aligned}$$

With this formulation of the model the null and alternative hypotheses are stated as follows:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0, \quad H_1 : \text{Not all } \alpha\text{s are zero.}$$

Table 7.11 shows the ANOVA table for the golf ball experiment together with the expected values of the mean squares. The analysis of variance partitions the total variation as represented by the *total* sum of squares (510 with 14 degrees of freedom) into a component attributable to the source *type* of ball (280 with 2 degrees of freedom) and a component attributable to the source random variation or random *error* (230 with 12 degrees of freedom). The component degrees of freedom and the component sums of squares add up to the corresponding totals ($2 + 12 = 14$ and $280 + 230 = 510$). In total the experiment yielded a sample of 15 values of length and a sample of 15 values has 14 degrees of freedom. We can think also of a sample of three means corresponding to the data for the three ball types involved in the experiment and a sample of three has two degrees of freedom. The mean square corresponding to the type and error components is obtained by dividing the sum of squares by degrees of freedom. Montgomery (2009, pp. 142–146) gives general formulae for the calculation of degrees of freedom and sums of squares.

The test statistic is the ratio of the mean squares, i.e. $140/19.2 = 7.30$. If the null hypothesis is true then all the αs would be zero and the expected value of both mean squares would be σ^2. Thus, if the null hypothesis is true the test statistic would be expected to have a value around 1. If the null hypothesis is false then not all the αs are zero and the expected value of the numerator of the ratio yielding the test statistic would be greater than the expected value of the denominator. Thus if the null hypothesis is false the test statistic would be expected to have a value greater than 1. When the null hypothesis is true, with the model specified above, the test statistic has the F distribution with parameters 2 and 12, i.e. the degrees of freedom for type and error respectively. **Calc > Probability > F...** may be used to confirm the P-value for the test.

Cumulative Distribution Function

```
F distribution with 2 DF in numerator and 12 DF in denominator

  x P (X <= x)
7.3    0.991571
```

Panel 7.36 Calculation of the *P*-value for the golf ball experiment.

The Session window output in Panel 7.36 indicates that the probability of obtaining a value for F of 7.3 or greater is $1 - 0.991\,571 = 0.008\,429$ so the P-value for the test is 0.008, to three decimal places, as displayed in the Session window output in Panel 7.29.

The fixed effects model may also be specified as detailed in Box 7.5. The number of levels of the factor is a and the number of replicates is n. With $i = 2$ and $j = 3$ we have, for example,

$$Y_{23} = \mu + \alpha_2 + \varepsilon_{23}.$$

In terms of the golf ball experiment, where there were just three levels of the factor type of interest, this equation states that the length, Y, for level 2 of the factor (ball type F) with drive 3 is made up of the overall mean, μ, plus the effect, α_2, for level 2 (ball type F) plus a random error, ε_{23} (a value from the normal distribution with mean 0 and variance σ^2).

7.5.3 The random effects model

Let us examine the data from the golf ball experiment again, but now with one major difference. Instead of a fixed set of three ball types of interest, let us consider the three types used in the experiment to have been a *random sample* of types from the myriad available on the market. In this scenario a random effects model is appropriate. The questions to be addressed by the analysis would thus be: How much of the variation observed is attributable to real differences between mean length, in the population of types from which the three used in the experiment were selected? How much of the variation observed is attributable to random variation about these population means?

The *random* effects model may be specified as detailed in Box 7.6. As in the case of the fixed effects model, the number of levels of the factor is a and the number of replicates is n. With $i = 2$ and $j = 3$ we have, for example,

$$Y_{23} = \mu + \alpha_2 + \varepsilon_{23}.$$

Observed data value = Overall mean + Effect + Random Error,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, a, \qquad j = 1, 2, \ldots, n$$

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Box 7.5 Fixed effects model.

In terms of the golf ball experiment, where there were just three levels of the factor type of interest, this equation states that the length, Y, for level 2 of the factor (ball type F) with drive 3 is made up of the overall mean, μ, plus the effect, α_2 (a value from the normal distribution with mean 0 and variance σ_a^2) for level 2 (ball type F) plus a random error, ε_{23} (a value from the normal distribution with mean 0 and variance σ^2).

The ANOVA table is exactly as for the fixed effects model in the case where there is a single factor of interest. However, the null and alternative hypotheses are:

$$H_0 : \sigma_a^2 = 0, \quad H_1 : \sigma_a^2 \neq 0.$$

In order to be able to make a full analysis in the random effects case we will obtain the ANOVA table using **Stat > ANOVA > Balanced ANOVA...** as indicated in the dialog in Figure 7.31.

The design or plan used for the experiment was such that there were equal numbers of drives made with each level of the factor type. This makes the design a balanced one. The following points should be noted concerning the dialog:

- Under **Graphs...** here there is no facility to create boxplots or an individual plot as in Figure 7.27 by way of initial display of the data, but these may – and, in the author's view, one or other should – always be created separately using the **Graphs** menu.
- Here with a single factor, type, involved the model is:

  $$\text{Observed data value} = \text{Overall mean} + \text{Effect of type} + \text{Random error}.$$

  This information is communicated to the software by inserting or selecting Type under **Model:**. There is always an overall mean and a random error term on the right-hand side of the equation that specifies models of the sort employed here, so in this case of a single factor this entry conveys the key information.
- The information that the factor Type is random is communicated by inserting Type under **Random factors:**.
- Under **Graphs...** the **Four in one** option for **Residual Plots** is strongly recommended.
- Finally, under **Results...** the option **Display expected mean squares and variance components** should be checked and **Display means corresponding to the terms:** should specify the factor Type.

Observed data value = Overall mean + Effect + Random error,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, a, \qquad j = 1, 2, \ldots, n$$

$$\alpha_i \sim N(0, \sigma_a^2), \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

(the random variables α_i and ε_{ij} are independent)

Box 7.6 Random effects model.

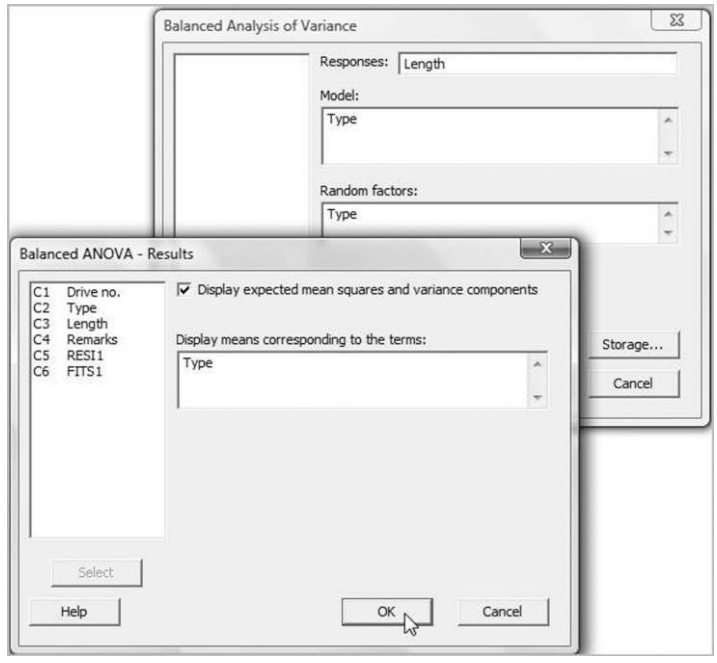


Figure 7.31 Dialog for balanced ANOVA.

The individual value plot in Figure 7.27 suggests that there is a component of variation in the response length of drive (Y) that may be attributed to factor ball type (X). Formal confirmation is obtained from the portion of the Session window output displayed in Panel 7.37. The null hypothesis $H_0: \sigma_a^2 = 0$ would be rejected in favour of the alternative hypothesis $H_1: \sigma_a^2 \neq 0$ at the 1% level of significance, since the P-value is 0.008. Thus the experiment provides strong evidence that σ_a^2, the component of variance in drive length attributable to ball type, is nonzero.

Table 7.12 gives the ANOVA table, together with the expected values of the mean squares for a random effects scenario. Here the number of replicates, n, was 5. We can take the observed mean squares as estimates of the corresponding expected mean squares. Thus σ^2 is estimated by 19.17 so σ is estimated by the square root of 19.17, which is 4.378. Also $\sigma^2 + 5\sigma_a^2$

ANOVA: Length versus Type

```
Factor  Type    Levels  Values
Type    random       3  E, F, G

Analysis of Variance for Length

Source  DF      SS      MS     F      P
Type     2  280.00  140.00  7.30  0.008
Error   12  230.00   19.17
Total   14  510.00
```

Panel 7.37 ANOVA table for golf ball experiment.

Table 7.12 ANOVA table for golf ball experiment with expected mean squares (random effects).

Source of variation	Degrees of freedom (DF)	Sum of squares (SS)	Mean square (MS)	Expected mean square (EMS)
Type	2	280	140	$\sigma^2 + n\sigma_a^2$
Error	12	230	19.17	σ^2
Total	14	510		

is estimated by 140, so $5\sigma_a^2$ is estimated by the difference $140 - 19.17 = 120.83$. Hence, σ_a^2 is estimated by $120.83/5 = 24.17$. The two components of variance are both given in the annotated section of the Session window output shown in Panel 7.38. The shorthand (2) represents the component of variance due to ball type i.e. to σ_a^2. The shorthand (1) represents the random error variance σ^2. Thus Minitab does all the calculations of the components of variance. The value s in the top left corner is the estimate of σ. The residuals and fitted values are the same as in the case of the fixed effects model so the R-sq value is as before.

The final portion of the Session window output gives the means for the three ball types that were selected at random from the population of available types. The fits and residuals are exactly as before and therefore the diagnostic plots are as before.

Imagine that Lynx goes to a driving range and selects a bucket of golf balls that constitute a random sample from the large population of golf ball types used in the random effects experiment. We can use the model to predict the distribution of length that will be achieved:

$$\text{Observed data value} = \text{Overall mean} + \text{Effect of type} + \text{Random error}.$$

Since the observed data value is a constant plus the sum of two independent random variables the result in Box 4.2 in Section 4.3.1 may be applied to calculate the mean and variance of the random variable length as detailed in Box 7.7. (A constant may be considered as a random variable with variance zero!) Thus the estimated total variance is 43.34 and the estimated proportion of total variance accounted for by ball type is $24.17/43.34 = 55.8\%$. Since a sum of independent normally distributed random variables is also normally distributed we can finally predict that the distribution of length on the driving range would be $N(200, 6.58^2)$. Were Lynx to opt, for example, to use a balls of type G only then the distribution of length would be estimated to be $N(204, 4.38^2)$.

```
S = 4.37798   R-Sq = 54.90%   R-Sq(adj) = 47.39%


                                Expected Mean
                                Square for Each
                                Term (using
              Variance  Error   unrestricted
   Source   component    term   model)
1  Type         24.17       2   (2) + 5 (1)     [corresponding to σ²+nσ²a]
2  Error        19.17           (2)             [corresponding to σ²]
```

Panel 7.38 Components of variance for golf ball experiment.

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \alpha_i \sim N(0, \sigma_a^2); \varepsilon_i \sim N(0, \sigma^2)$.

Mean of $Y_{ij} = \mu + 0 + 0 = \mu$ which is estimated as 200, the overall mean for the experiment.

Variance of $Y_{ij} = 0 + \sigma_a^2 + \sigma^2$ which is estimated by $0 + 24.17 + 19.17 = 43.34 = 6.58^2$.

Box 7.7 Calculation of the mean and variance of length.

Montgomery (2005a, p. 487) gives an industrial example involving components of variance. A textile company weaves a fabric on a large number of looms. Interest centred on loom-to-loom variability in the tensile strength of the fabric. Four looms were selected at random and four random samples of fabric from each loom were tested, yielding the data in Table 7.13. The data are available in stacked form in the worksheet Looms.MTW and are reproduced by permission of John Wiley & Sons, Inc., New York.

Initial analysis of the data was carried out using **Stat** > **ANOVA** > **One-Way...** with the **Individual value plot** option selected under **Graphs...** together with **Normal plot of residuals** and **Residuals versus fits**. Scrutiny of the individual value plot suggests that there is variation attributable to the factor loom; this is confirmed by a *P*-value of 0.000 (to three decimal places) which indicates very strong evidence of such significant variation. The normal probability plot was reasonably straight and the vertical spreads of residuals similar in the plot of residuals versus fits. Thus one can be satisfied that a random effects model of the form used in the previous example is appropriate.

Having established a significant loom effect, **Stat** > **ANOVA** > **Balanced ANOVA...** was used to obtain the components of variance shown in Panel 7.39. Hence the mean and variance of Tensile Strength may be estimated as detailed in Box 7.8. Thus the estimated total

Table 7.13 Tensile strength data.

Loom	Tensile strength (psi)			
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95
4	95	96	99	98

```
S = 1.37689   R-Sq = 79.68%   R-Sq(adj) = 74.60%

                                Expected Mean
                                Square for Each
                                Term (using
             Variance  Error    unrestricted
   Source  component   term    model)
1  Loom        6.958      2  (2) + 4 (1)
2  Error       1.896         (2)
```

Panel 7.39 Components of variance for looms experiment.

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \alpha_i \sim N(0, \sigma_a^2);\ \varepsilon_i \sim N(0, \sigma^2)$.

Mean of $Y_{ij} = \mu$ which is estimated as 95.438, the overall mean for the experiment.

Variance of $Y_{ij} = \sigma_a^2 + \sigma^2$ which is estimated by $6.958 + 1.896 = 8.854 = 2.976^2$.

Box 7.8 Calculation of the mean and variance of tensile strength.

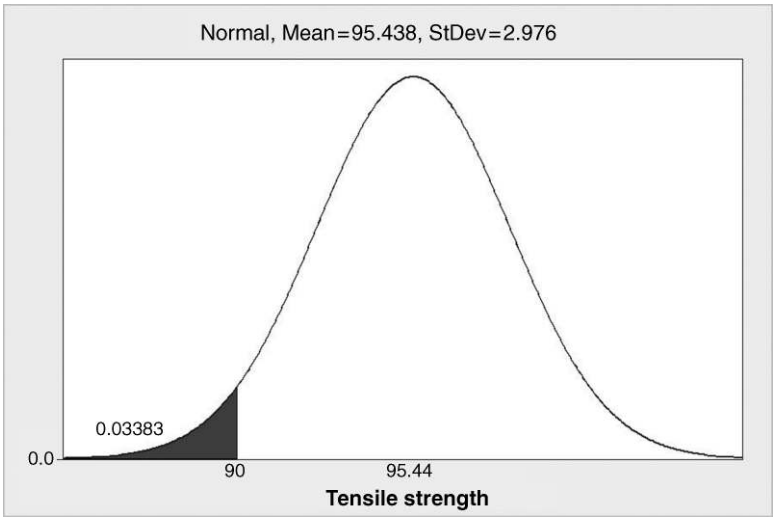


Figure 7.32 Estimated distribution of tensile strength.

variance is 8.854 and the estimated proportion of total variance accounted for by the factor loom is 6.958/8.854 = 78.6%. The overall mean for the experiment was 95.438, so we can estimate that the distribution of tensile strength for fabric produced on the population of looms would be $N(95.438, 2.976^2)$. This model, together with a reference line indicating the lower specification limit of 90 psi for tensile strength, is shown in Figure 7.32. The process is therefore operating with a C_{pk} of the order of 0.6 (sigma quality level of around 3.3).

Montgomery comments:

> A substantial proportion of the production is fallout. This fallout is directly related to the excess variability resulting from differences between looms. Variability in loom performance can be caused by faulty set-up, poor maintenance, inadequate supervision, poorly trained operators and so forth. The engineer or manager responsible for quality improvement must remove these sources of variability from the process.

7.5.4 The nonparametric Kruskal–Wallis test

Consider again the telesales data displayed in Table 7.10 and available in worksheet Sales.MTW. The sales figures are actually counts, so it could be argued that an analysis of variance,

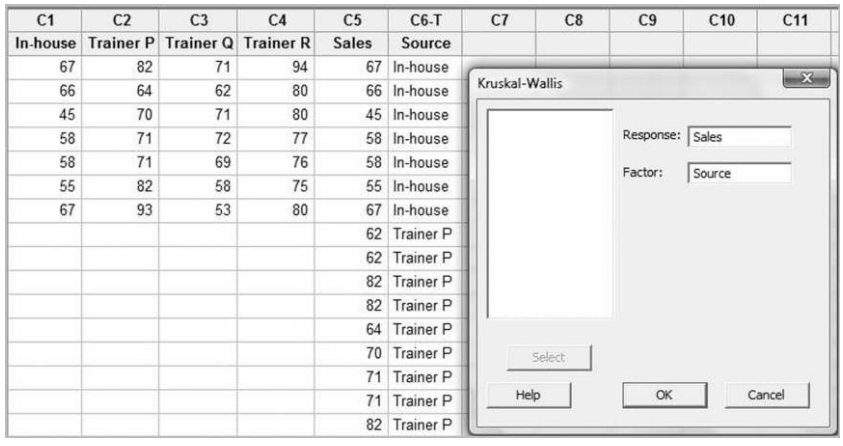


Figure 7.33 Dialog for Kruskal–Wallis test.

with the underlying assumption of normality, is inappropriate. Minitab provides two non-parametric tests for experiments involving a single factor with more than two levels – the Kruskal–Wallis test and Mood's median test.

In order to perform a Kruskal–Wallis test with Minitab the response data and the factor levels must appear in two columns. This can readily be arranged using **Data** > **Stack** > **Columns...**. All four columns are entered into the **Stack the following columns:** window. With **Column of current worksheet:** checked and Sales entered, **Store subscripts in**: Source specified and **Use variable names in subscript column** checked, the stacked data are stored in a column named Sales and the levels of the factor are stored in a column named Source. A portion of the stacked data can be seen in Figure 7.33. Then **Stat** > **Nonparametrics** > **Kruskal-Wallis...** leads to the dialog also shown in Figure 7.33.

The Session window output is shown in Panel 7.40. The sample size and median are given for each level of Source. The test is a generalization of the Mann–Whitney test and is based on ranks. The average rank for each level of the factor is given together with a corresponding z-value. The null hypothesis is that the samples are from identical populations. The test-statistic is denoted by the letter H. The corresponding P value is given and in this case indicates very strong evidence that the populations sampled are not identical, i.e. that the training methods are not equally effective. The procedure does not

Kruskal-Wallis Test: Sales versus Source

```
Kruskal-Wallis Test on Sales

Source      N  Median  Ave Rank      Z
In-house   10   62.50      11.4  -2.83
Trainer P  10   71.00      24.4   1.22
Trainer Q  10   67.00      14.4  -1.91
Trainer R  10   78.50      31.8   3.51
Overall    40              20.5

H = 19.09  DF = 3  P = 0.000
H = 19.24  DF = 3  P = 0.000  (adjusted for ties)
```

Panel 7.40 Session window output for Kruskal–Wallis test.

provide an option to display the data, nor does it provide any facility for comparisons. When used in a situation in which ANOVA could legitimately be used it provides a less powerful test than ANOVA.

7.6 Blocking in single-factor experiments

To introduce the idea of blocking, we will consider an experiment where the single factor of interest is variety of potato and the response of interest is yield in tonnes per hectare (t/ha). Denote the three levels of the variety factor by A, B and C. Suppose that 12 plots, numbered 1 to 12, are available for the experiment, as shown in Figure 7.34, and that random allocation of varieties to the plots led to the design indicated.

Imagine that there is a wood to the west of the plots and a river to the east. This could conceivably lead to a fertility gradient in the direction of the arrow due to greater amounts of both moisture and nutrients in the soil, the further plots are from the wood. A concern with this completely randomized design is that variety A might appear to perform well in terms of yield not because it was superior to the other varieties but because the plots planted with A were favourably placed in terms of a possible fertility gradient.

A superior experimental design in this situation would be achieved through the use of blocking. Each strip of three plots running in a north–south direction would be designated as a block, yielding four blocks as indicated in Figure 7.35. Subsequently the three varieties would be allocated at random within each block. Suppose that the arrangement shown in Figure 7.36 arose. This is a randomized complete block design. The numbers in brackets are the yield values. (As with the golf ball experiment, fictitious integer data have been used in order to make the arithmetic simple when introducing key concepts.)

The term 'block' is a legacy from the early application of designed experiments to agricultural research. 'Block' meant a block of land as in this introductory example. In experimental design it now refers to groups of experimental units that are homogeneous. If the factor of interest has, say, four levels then a block might consist of four plots of land adjacent to each other, four test cubes of concrete from the same batch, four pigs from the same litter etc.

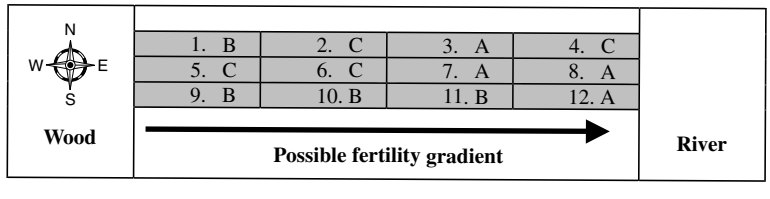


Figure 7.34 Completely randomized design.

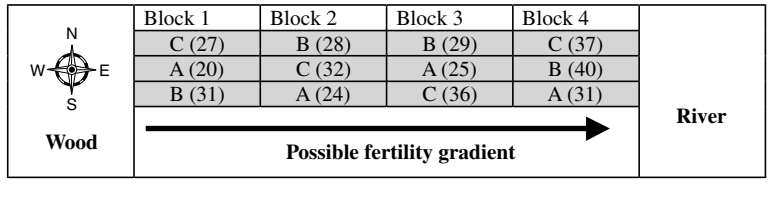


Figure 7.35 Randomized complete block design.

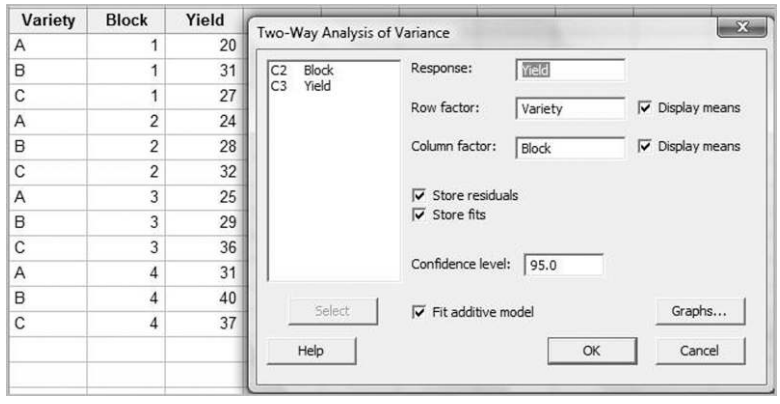


Figure 7.36 Dialog for a two-way analysis of variance.

The data from the experiment are tabulated in Table 7.14 and available in Potato.MTW. In order to analyse the data via Minitab they have to be arranged into three columns specifying variety, block and yield, respectively. Yield is the response and we have two factors, variety and block. Thus we may use **Stat** > **ANOVA** > **Two-Way...** to perform an analysis of variance. The dialog is shown in Figure 7.36.

Yield is entered as the **Response:**, Variety as the **Row factor:** and Block as the **Column factor:**. (The levels of variety correspond to the rows, and the levels of block correspond to the columns of Table 7.14.) The **Display means** option was checked for both factors. **Store residuals** and **Store fits** were checked. It is essential that **Fit additive model** be checked when using this procedure to analyse data from a randomized complete block design where there is a single factor of interest (variety in this case).

Under **Graphs...**, the option to display the data using an **Individual value plot** was selected together with **Normal plot of residuals**. The individual value plot given in Figure 7.37 suggests that both variety B and variety C give heavier yield than does variety A. It also suggests that the use of blocking may have been wise since yield generally increases across the blocks from west to east.

The ANOVA table from the Session window output is displayed in Panel 7.41. The *P*-values for both variety and block are less than 0.05 so the experiment provides evidence, at the 5% level of significance, that variety has a significant influence on the response yield and evidence of block effect. Thus it would appear that the experimenters were justified in using blocking.

The overall mean yield for the experiment was 30. Having checked the **Display means** option for both the row and column factors, the means for variety and block are displayed below the ANOVA table in the Session window output and are given in Table 7.15.

Table 7.14 Data from randomized complete block design.

	Block 1	Block 2	Block 3	Block 4
Variety A	20	24	25	31
Variety B	31	28	29	40
Variety C	27	32	36	37

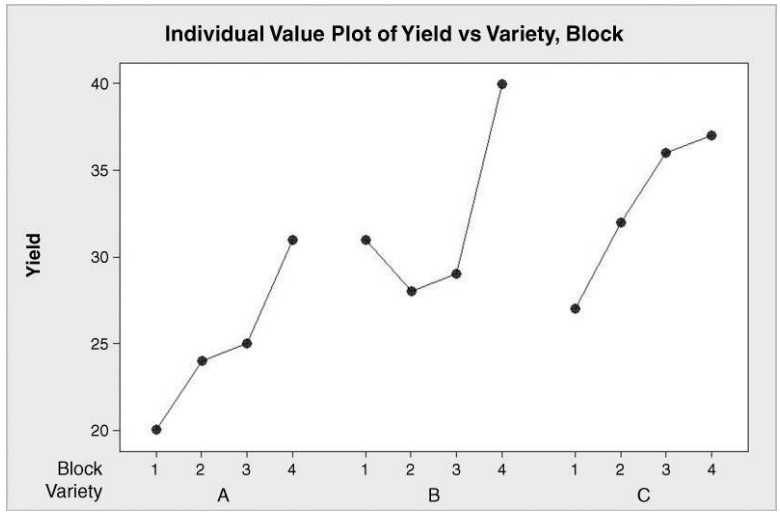


Figure 7.37 Individual value plot of Yield.

Two-way ANOVA: Yield versus Variety, Block

Source	DF	SS	MS	F	P
Variety	2	152	76.0000	9.91	0.013
Block	3	168	56.0000	7.30	0.020
Error	6	46	7.6667		
Total	11	366			

S = 2.769 R-Sq = 87.43% R-Sq(adj) = 76.96%

Panel 7.41 ANOVA for potato experiment.

The mean for variety A was 25. One can therefore think of the effect for variety A as being −5, i.e. variety A reduced mean yield by 5 t/ha from the overall mean of 30 t/ha. The mean for variety B was 32 so we can think of the effect for variety B as being 2, i.e. variety B increased mean yield by 2 t/ha from the overall mean of 30. Similarly, for variety C the effect was 3. Note that the three variety effects sum to 0. The mean for block 1 was 26. One can think of the effect for block 1 as being –4, i.e. block 1 reduced mean yield by 4 t/ha from the overall mean. Similarly,the effects for blocks 2, 3 and 4 were −2, 0 and 6, respectively. Note that the four block effects sum to 0.

Table 7.15 Data with means from randomized complete block experiment.

	Block 1	Block 2	Block 3	Block 4	Mean
Variety A	20	24	25	31	25
Variety B	31	28	29	40	32
Variety C	27	32	36	37	33
Mean	26	28	30	36	30

We have already seen the general form of model:

$$\text{Data} = \text{Fit} + \text{Residual}.$$

In this case we take

$$\text{Fit} = \text{Overall mean} + \text{Variety effect} + \text{Block effect}.$$

Thus we have:

$$\begin{aligned}
&\text{variety A in block 1} : \text{Fit} = 30 + (-5) + (-4) = 21,\\
&\text{variety B in block 1} : \text{Fit} = 30 + 2 + (-4) = 28,\\
&\text{variety C in block 1} : \text{Fit} = 30 + 3 + (-4) = 29,\\
&\text{variety A in block 2} : \text{Fit} = 30 + (-5) + (-2) = 23,\\
&\text{variety B in block 2} : \text{Fit} = 30 + 2 + (-2) = 30,\\
&\text{variety C in block 2} : \text{Fit} = 30 + 3 + (-2) = 31,\\
&\text{variety A in block 3} : \text{Fit} = 30 + (-5) + 0 = 25,\\
&\text{variety B in block 3} : \text{Fit} = 30 + 2 + 0 = 32,\\
&\text{variety C in block 3} : \text{Fit} = 30 + 3 + 0 = 33,\\
&\text{variety A in block 4} : \text{Fit} = 30 + (-5) + 6 = 31,\\
&\text{variety B in block 4} : \text{Fit} = 30 + 2 + 6 = 38,\\
&\text{variety C in block 4} : \text{Fit} = 30 + 3 + 6 = 39.
\end{aligned}$$

We can now compute the residual values as the differences Data – Fit. The reader is invited to check the calculations in the first few rows of Table 7.16 and to observe that, having checked both **Store residuals** and **Store fits**, both residuals and fits are displayed in the worksheet.

The fixed effects model may also be specified as detailed in Box 7.9. For the potato experiment the number of levels of the factor of interest, variety, is $a = 3$ and the number of

Table 7.16 Fitted values and residuals for potato experiment.

Variety	Block	Data (yield)	Overall mean	Variety effect	Block effect	Fit	Residual
A	1	20	30	−5	−4	21	−1
B	1	31	30	2	−4	28	3
C	1	27	30	3	−4	29	−2
A	2	24	30	−5	−2	23	1
B	2	28	30	2	−2	30	−2
C	2	32	30	3	−2	31	1
A	3	25	30	−5	0	25	0
B	3	29	30	2	0	32	−3
C	3	36	30	3	0	33	3
A	4	31	30	−5	6	31	0
B	4	40	30	2	6	38	2
C	4	37	30	3	6	39	−2

Observed data value = Overall mean + Factor effect + Block effect + Random error

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b,$$

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{i=1}^{b} \beta_i = 0, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Box 7.9 The fixed effects model

levels of what some refer to as a 'nuisance' factor block is $b = 4$. With $i = 2$ and $j = 3$, for example, we have the specific relationship

$$Y_{23} = \mu + \alpha_2 + \beta_3 + \varepsilon_{23}.$$

This equation therefore states that the yield for the plot planted with variety 2 (B) in block 3 is made up of the overall mean, μ, plus the effect for variety 2, α_2, plus the effect for block 3, β_3, plus a random error (value) from the normal distribution with mean 0 and variance σ^2. At the core of this model we have the addition of the two effects α_2 and β_3, one for variety and one for block – hence the need to check **Fit additive model** in the dialog.

There are two null hypotheses to be tested against alternatives. The first states that all the variety effects are zero, the second that all the block effects are zero. Formally they are stated as follows:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0, \quad H_1 : \text{Not all } \alpha\text{s are zero};$$
$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{Not all } \beta\text{s are zero}.$$

The P-values corresponding to these were 0.013 and 0.020 respectively, so both null hypotheses would be rejected at the 5% level of significance.

Having obtained evidence of variety having leverage in determining yield, it is useful to be able to carry out follow-up analysis using multiple comparisons. This is not available via **Stat > ANOVA > Two-Way...** but is available via **Stat > ANOVA > General Linear Model...**. The dialog is shown in Figure 7.38.

In **Model:** we are communicating to Minitab the nature of the model we are using, i.e. $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$. Generally such models always include an overall mean and the random error term so by entering Variety and Block in **Model:** we are indicating the expression $\alpha_i + \beta_j$ at the core of the equation that defines the model in this scenario.

The subdialog for **Comparisons:** is also shown in Figure 7.38. Here **Pairwise comparisons** were selected, by the **Tukey** method with **Terms:** Variety. **Grouping information** and **Confidence interval**, with default **Confidence level:** 95.0, were checked. The latter part of corresponding section of the Session window output is shown in Panel 7.42.

The interpretation of this is as follows:

- On average the yield of variety B is 7 t/ha more than that for variety A, with confidence interval 1 to 13 t/ha (to the nearest integer).
- On average the yield of variety C is 8 t/ha more than that for variety A, with confidence interval 2 to 14 t/ha (to the nearest integer).

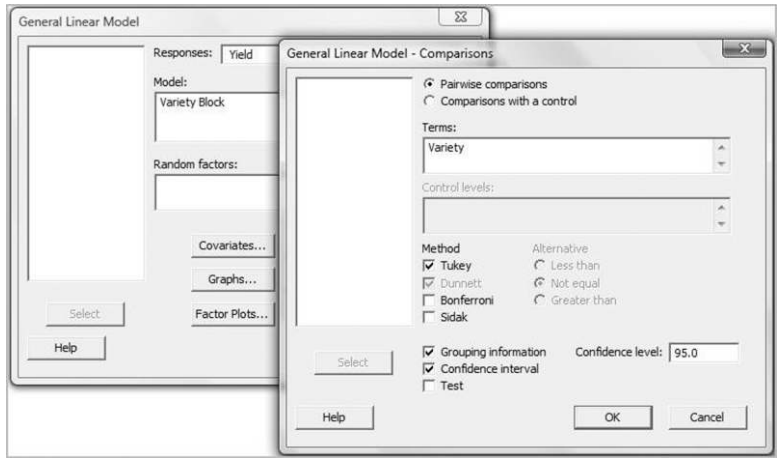


Figure 7.38 General Linear Model dialog.

- On average the yield of variety C is 1 t/ha more than that for variety B, with confidence interval –5 to 7 t/ha (to the nearest integer).

Since the first two confidence intervals do not include 0 we have evidence that the yield of both varieties B and C is superior to that of variety A. The fact that the third confidence interval includes 0 means that we have no evidence of a difference in mean yield for varieties B and C. Thus the experimentation has provided evidence that both varieties B and C give significantly greater mean yield than does variety A. However, it does not provide any evidence of a difference in yield for B and C.

The earlier part of the Session window output from Comparisons is displayed in Panel 7.43. This summarizes the conclusions that stem from scrutiny of the confidence intervals, i.e. that

```
Tukey 95.0% Simultaneous Confidence Intervals
Response Variable Yield
All Pairwise Comparisons among Levels of Variety
Variety = A  subtracted from:

Variety   Lower  Center  Upper  --------+---------+---------+--------
B        0.9915   7.000  13.01            (---------*---------)
C        1.9915   8.000  14.01             (---------*---------)
                               --------+---------+---------+--------
                                     0.0       6.0      12.0


Variety = B  subtracted from:

Variety   Lower  Center  Upper  --------+---------+---------+--------
C        -5.008   1.000  7.008  (---------*---------)
                               --------+---------+---------+--------
                                     0.0       6.0      12.0
```

Panel 7.42 Session window output from Comparisons.

```
Grouping Information Using Tukey Method and 95.0% Confidence

Variety  N  Mean  Grouping
C        4  33.0  A
B        4  32.0  A
A        4  25.0     B

Means that do not share a letter are significantly different.
```

Panel 7.43 Session window output from Comparisons.

the means for varieties B and C do not differ significantly whereas the means for both A and B and A and C do.

Iman and Conover (1989, p. 631) give an example of a purchasing agent seeking to obtain word-processing software with which operators have the best production rate. Three candidate packages for purchase were assessed in a randomized complete block experiment in which six operators were treated as blocks. The response was the time taken (minutes) to input a standard document. The data are reproduced by permission of the authors in Table 7.17 and are available, in stacked form, in the worksheet Packages.MTW.

It is left as an exercise for the reader (remember to check **Fit additive model**!) to verify that there is evidence of differences between packages (P-value 0.045) and very strong evidence of differences between operators (P-value 0.000 to three decimal places). The normal probability plot of the residuals is satisfactory. The Session window output for **Multiple Comparisons** obtained via **Stat** > **ANOVA** > **General Linear Model...** using the **Tukey method** is shown in Panel 7.44. Note that in the **Comparisons...** subdialog box **Terms:** Package is required.

For example, the point estimate of the population mean time using package 1 was 47.3 minutes, while that for package 3 was 50.3 minutes. The point estimate of the difference in the means is 3 minutes with confidence interval (0.2, 5.8) minutes (rounded to one decimal place). The fact that this confidence interval does not include 0 indicates that the document can be produced significantly faster with package 1 than with package 3. Since the other confidence intervals both include 0 we cannot claim a significant difference between package 1 and 2 and we cannot claim a significant difference between package 2 and 3. These three confidence intervals together have overall confidence level of 95%.

Table 7.17 Word-processing package assessment data.

	Word-processing package		
Operator	1	2	3
1	42	45	45
2	37	36	40
3	53	56	55
4	68	73	75
5	48	45	47
6	36	39	40

```
Grouping Information Using Tukey Method and 95.0% Confidence

Package  N  Mean  Grouping
3        6  50.3  A
2        6  49.0  A B
1        6  47.3    B

Means that do not share a letter are significantly different.


Tukey 95.0% Simultaneous Confidence Intervals
Response Variable Time
All Pairwise Comparisons among Levels of Package
Package = 1  subtracted from:

Package   Lower  Center  Upper  -------+---------+---------+---------
2        -1.147   1.667  4.480   (-------------*-------------)
3         0.186   3.000  5.814           (-------------*-------------)
                                -------+---------+---------+---------
                                     0.0       2.0       4.0


Package = 2  subtracted from:

Package   Lower  Center  Upper  -------+---------+---------+---------
3        -1.480   1.333  4.147  (-------------*-------------)
                                -------+---------+---------+---------
                                     0.0       2.0       4.0
```

Panel 7.44 Session window output from Comparisons.

Some report the grouping information provided at the top of Panel 7.44 by listing the levels of the factor of interest and underlining those for which the mean responses do not differ significantly:

Package 1 Package 2 Package 3

Iman and Conover (1989, p. 638) comment that 'the data are not sufficiently strong to indicate a difference between software package 2 and the other word processing software packages, but they are significantly strong to declare a difference between software packages 1 and 3'.

Some authors refer to experimental designs that involve blocking as 'noise-reducing' experimental designs. 'Designing for noise reduction is based on the single principle of making all comparisons of treatments within relatively homogeneous groups of experimental units. The more homogeneous the experimental units the easier it is to detect a difference in treatments' (Mendenhall *et al*., 1986, p. 525). The reader is invited to verify that, were we to ignore the blocking and do a one-way ANOVA then a *P*-value of 0.923 would be obtained. Thus naïve analysis of the data, which fails to take operator into account as a blocking factor, provides no evidence of any package effect.

The Friedman test is a nonparametric test that may be used to analyse data from a randomized complete block experiment. It may be implemented using **Stat** > **Nonparametrics** > **Friedman. . .**. For the data from the potato experiment the Session window output is shown in Panel 7.45. The reader is invited to check it as an exercise. In some cases where

```
Friedman Test: Yield versus Variety blocked by Block
S = 6.00  DF = 2  P = 0.050

                          Sum of
Variety  N  Est Median    Ranks
A        4       24.500     4.0
B        4       31.167    10.0
C        4       31.833    10.0

Grand median = 29.167
```

Panel 7.45 Session window output for Friedman test.

identical values or ties occur amongst the values of the response two *P*-values are given, the second taking ties into account.

In the dialog for this analysis Minitab refers to 'Treatment' rather than 'Factor'. The *P*-value quoted is for a test of the null hypothesis H_0 : all treatment effects are zero, versus the alternative hypothesis H_1 : not all treatment effects are zero. It is on the brink of being significant at the 5% level. The earlier analysis based on the assumption of underlying normal distributions gave a corresponding *P*-value of 0.013.

The Friedman test is another non-parametric procedure that is based on ranks. Residuals and fits may be computed using Minitab. Multiple comparisons based on ranks may be made, but this facility is not provided by Minitab – technical details are given in Iman and Conover (1989, p. 658).

7.7 Experiments with a single factor, with more than two levels, where the response is a proportion

A glass bottle manufacturer had been receiving complaints from customers concerning tears in and imperfect sealing of the shrinkwrap used on pallets of bottles. A Six Sigma project team carried out a single-factor experiment in which shrinkwrap from each of three suppliers A, B and C was used to seal 1200 pallets of bottles. Following shipment to a customer all 3600 pallets were checked and the numbers of nonconforming pallets recorded. The data are summarized in Table 7.18.

The null hypothesis of interest here is $H_0 : p_1 = p_2 = p_3$ and the alternative is H_1 : Not all p_i are identical ($i = 1, 2, 3$), where p_1, p_2 and p_3 represent the population proportion of nonconforming pallets sealed with shrinkwrap from suppliers A, B and C, respectively. We could analyse the above data formally using three tests for equality of two proportions – one to compare A with B, one to compare B with C, and a final one to compare C with A. However, the

Table 7.18 Nonconfoming pallet data.

	Supplier			
Status	A	B	C	Total
Nonconforming	34	57	29	120
Conforming	1166	1143	1171	3480
Total	1200	1200	1200	3600
% Nonconforming	2.8	4.8	2.4	3.3

problem of the increased risk of a Type I error with this approach has already been discussed in Section 7.5.

If the null hypothesis is true then the proportions are homogeneous across suppliers – hence the test to be used is referred to as a test of homogeneity. It is available using **Stat > Tables > Chi-Square Test (Two-Way Table in Worksheet)...**. The dialog is shown in Figure 7.39. The table required can be seen in the worksheet in the figure and consists of the shaded portion of Table 7.18. Note that the first row of each column gives the number of nonconforming pallets for the supplier and the second row gives the number of conforming pallets for the supplier.

The Session window output is displayed in Panel 7.46. Out of a total of 3600 pallets, 120 were nonconforming. If the null hypothesis is true then $120/3600 = 1/30$ provides an estimate of the common proportion of nonconforming pallets for shrinkwrap from all three suppliers. Thus for shrinkwrap from each supplier we would expect to find one in 30 of the 1200, i.e. 40 pallets, to be nonconforming and the remaining 1160 to be conforming. These expected counts, E_i, have been computed and displayed below the observed counts, O_i, in the table in the output.

The test statistic involves the differences between the observed and expected counts and is calculated as shown in Box 7.10. The central involvement of $O_i - E_i$ in the formula for the chi-square test statistic means that its value is relatively low when there is good agreement between observed and expected counts. Poor agreement arises when the null hypothesis is false, leading to a relatively large value of the test statistic. The P-value may be confirmed using **Calc > Probability Distributions > Chi-square...**. The Session window output is given in

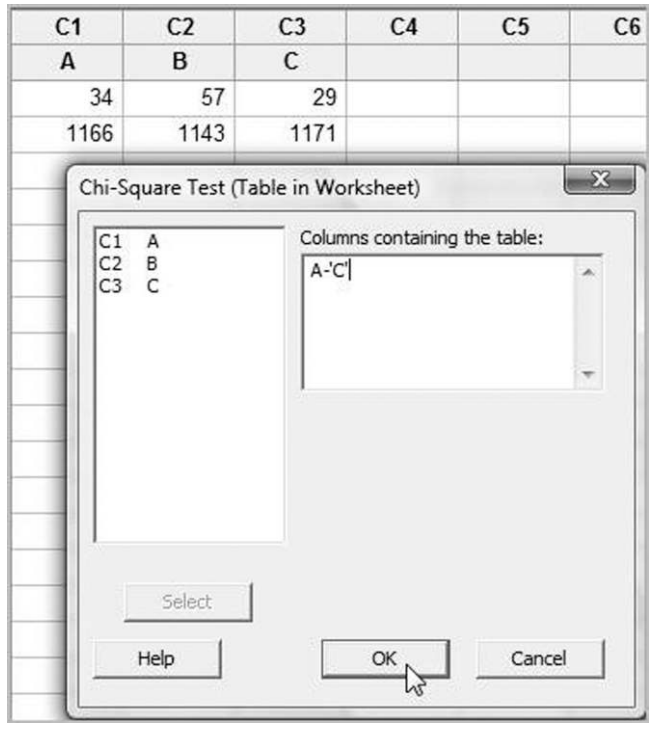


Figure 7.39 Dialog for test of homogeneity of proportions.

Chi-Square Test: A, B, C

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

              A         B         C  Total
    1        34        57        29    120
          40.00     40.00     40.00
          0.900     7.225     3.025

    2      1166      1143      1171   3480
       1160.00   1160.00   1160.00
         0.031     0.249     0.104

Total      1200      1200      1200   3600

Chi-Sq = 11.534, DF = 2, P-Value = 0.003
```

Panel 7.46 Session window output for chi-square test.

Panel 7.47. It indicates that the probability of obtaining a chi-square value of 11.534 or greater, were the null hypothesis true, would be $1 - 0.996\,871 = 0.003\,129$, or 0.003 to three decimal places, as stated in Panel 7.46. Thus the data from the experiment provide evidence, at the 1% level of significance, that the suppliers perform differently in terms of proportion of non-conforming. Formally, the null hypothesis $H_0 : p_1 = p_2 = p_3$ would be rejected in favour of the alternative H_1 : Not all p_i are identical ($i = 1, 2, 3$) at the 1% level of significance. It therefore appears that supplier B performs significantly worse than the other two (4.8% nonconforming compared with 2.8% and 2.4%, respectively).

The chi-square test statistic is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where k is the number of cells in the table. Thus

$$\begin{aligned}\chi^2 &= \frac{(34-40)^2}{40} + \frac{(57-40)^2}{40} + \frac{(29-40)^2}{40} \\ &\quad + \frac{(1166-1160)^2}{1160} + \frac{(1143-1160)^2}{1160} + \frac{(1171-1160)^2}{1160} \\ &= 0.900 + 7.225 + 3.025 + 0.031 + 0.249 + 0.104 \\ &= 11.534\end{aligned}$$

(note how these six contributions to the test statistic appear in the output). The only parameter of a chi-square distribution is its number of degrees of freedom, which in this case is $a - 1$ where a is the number of levels of the supplier factor, i.e. $3 - 1 = 2$.

Box 7.10 Calculation of the chi-square test statistic.

```
Cumulative Distribution Function

Chi-Square with 2 DF

     x  P( X <= x )
11.534     0.996871
```

Panel 7.47 Calculation of the *P*-value.

7.8 Tests for equality of variances

Two-sample *t*-tests and one-way ANOVA tests that required the assumption of equal population variances were discussed earlier in the chapter. Support for the assumption may be obtained from scrutiny of displays of the data in the form of individual value plots or boxplots. Minitab provides formal tests of the null hypothesis of equal variances. Consider again the magnesium assay data in Table 7.6, stored in Magnesium.MTW. Use of **Stat** > **Basic Statistics** > **2 Variances...** provides a test of the null hypothesis that the two population variances are the same, i.e. the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, versus the alternative hypothesis that they are not, $H_1 : \sigma_1^2 \neq \sigma_2^2$. The null hypothesis is equivalent to H_0: $\sigma_1 = \sigma_2$, which in turn is equivalent to $H_0 : \sigma_1/\sigma_2 = 1$. The dialog is shown in Figure 7.40. The alternative hypothesis is equivalent to $H_1 : \sigma_1 \neq \sigma_2$, which in turn is equivalent to $H_1 : \sigma_1/\sigma_2 \neq 1$. In completing the dialog the reader is urged to use the **Graphs...** button to select either an individual value plot or boxplot of the data.

Key components of the Session window output are shown in Panel 7.48. The *F*-test for equality of variances yields a *P*-value 0.021, so the null hypothesis of equal variances would be rejected at the 5% level of significance. Valid application of this test requires the distributions to be normal. Levene's test for equality of variances yields a *P*-value 0.016, so the null

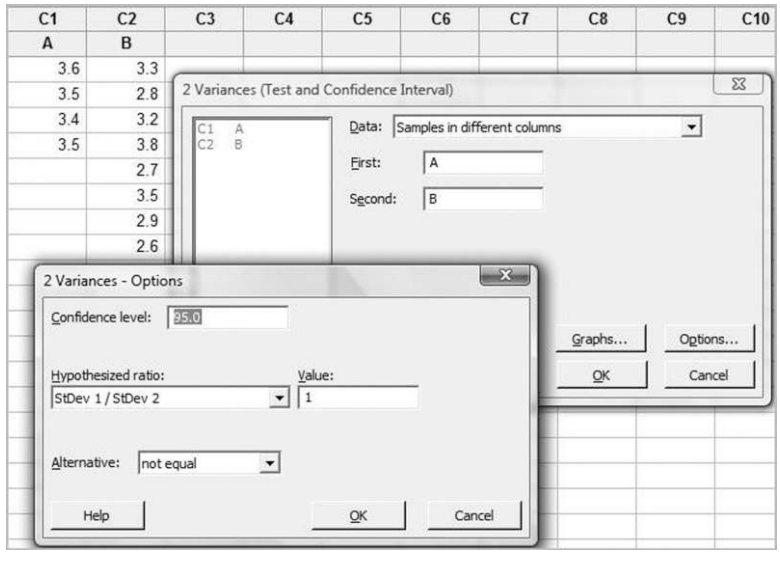


Figure 7.40 Dialog for test of equality of variances.

Test and CI for Two Variances: A, B

Method

Null hypothesis Sigma(A) / Sigma(B) = 1
Alternative hypothesis Sigma(A) / Sigma(B) not = 1
Significance level Alpha = 0.05

Tests

Method	DF1	DF2	Test Statistic	P-Value
F Test (normal)	3	7	0.04	0.021
Levene's Test (any continuous)	1	10	8.28	0.016

Panel 7.48 Session window output for test for equal variances.

hypothesis of equal variances would be rejected at the 5% level of significance. Valid application of this test simply requires the distributions to be continuous.

As a second example, consider the drive length data in Figure 7.26, also available in Types.MTW. Here, in the fixed effects scenario, there were three populations of interest corresponding to the three ball types. Use of **Stat > ANOVA > Test for Equal Variances...** provides a test of the equality of two or more variances – in this case of the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ versus the alternative hypothesis H_1: not all variances are equal. The dialog is shown in Figure 7.41.

The Session Widow output is shown in Panel 7.49. Graphical output is provided but is not reproduced here. With *P*-values of 0.956 and 0.851 there is no reason to doubt the null hypothesis of equal variances for the three types. Bonferroni confidence intervals are given for the population standard deviations. As with the Tukey procedure for multiple comparisons, the Bonferroni procedure is designed to yield a set of confidence intervals with an overall joint confidence level – the default of 95% was specified in this example. The user may change the

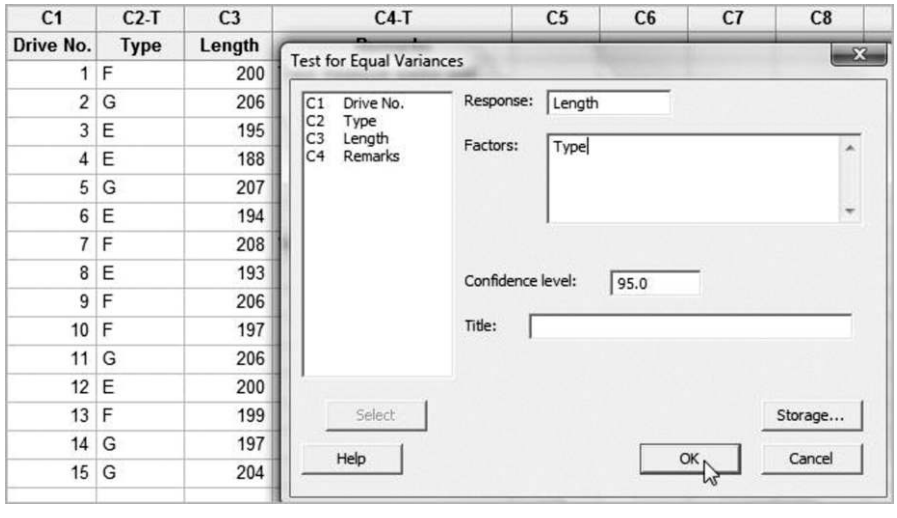


Figure 7.41 Dialog for test of equality of variances.

```
Test for Equal Variances: Length versus Type

95% Bonferroni confidence intervals for standard deviations

Type  N    Lower    StDev    Upper
   E  5  2.32449  4.30116  16.5547
   F  5  2.56350  4.74342  18.2569
   G  5  2.19525  4.06202  15.6343

Bartlett's Test (Normal Distribution)
Test statistic = 0.09, p-value = 0.956

Levene's Test (Any Continuous Distribution)
Test statistic = 0.16, p-value = 0.851
```

Panel 7.49 Session window output for test for equal variances.

confidence level via **Options:**. The assumption required for the valid use of each test is stated in brackets in the Session window output after the name of each test.

In addition to the use of these tests concerning the variability of populations for checking the validity of assumptions underlying the use of tests concerning location, they provide a means of analysing data from an experiment with a single factor where variability is the response of interest.

7.9 Exercises and follow-up activities

1. The output from a manufacturing operation over many weeks was, on average, 2000 units per day with standard deviation 400 units. Following changes to plant configuration the outputs for a sample of 25 days were as displayed in Table 7.19.

   The data are available in the supplied file Output.xls.

   (i) Display the data.

   (ii) Do you think that there has been a 'real' increase in mean daily output?

   (iii) State null and alternative hypotheses.

   (iv) Carry out a formal test of these hypotheses.

   (v) State your conclusion both formally and in plain English.

   (vi) What assumption(s) have you made in order to carry out the test?

2. The United States Golf Association (2008) has a requirement concerning golf balls which states: 'The combined carry and roll of the ball, when tested on apparatus

Table 7.19 Sample of daily outputs.

2325	2756	2358	2239	2660
1711	2176	1840	2082	2008
1760	2473	1998	2304	2215
2145	1900	1779	1709	2145
1596	2278	2482	2426	2160

Table 7.20 Sample of distances (yards) achieved.

316	315	315	317	333	328	330	317
313	329	339	302	324	323	311	296

approved by the United States Golf Association, must not exceed the distance specified under the conditions set forth in the Overall Distance Standard for golf balls on file with the United States Golf Association.' The QC Manager at Acme tested a random sample of 16 balls of a particular brand using similar equipment on the company's range, with the results displayed in Table 7.20. (Experience has shown that a standard deviation of 12 yards is typical for balls manufactured by Acme.)

(i) Display the data.

(ii) Do you think that the data for the brand of ball suggest a mean value that differs from the current Overall Distance Standard of 317 yards?

(iii) Explain why a two-tailed test is appropriate here.

(iv) State null and alternative hypotheses and carry out a formal test of these hypotheses.

(v) State your conclusion both formally and in plain English.

(vi) What assumption(s) have you made in order to carry out the test?

3. Suppose that initially assembly of P86 modules took on average 50.0 minutes with standard deviation 4.2 minutes. At a later date a sample of nine assembly times (minutes) was 44, 48, 45, 45, 46, 49, 48, 51 and 47.

   Evaluate the evidence for a reduction in the mean assembly time using a z-test, a t-test, a sign test and a Wilcoxon test. State any assumptions required for each test and whether or not they are reasonable.

   Estimate the size of sample required to detect evidence, at significance level 0.01 and with power 0.99, of a reduction in the mean of 2 minutes assuming that the standard deviation has remained at 4.2 and also without making this assumption.

4. An accountant believes that a company's cash flow problems are due to outstanding accounts receivable. She claims that 70% of the current accounts receivable are over 3 months old. A sample of 120 accounts receivable revealed 78 over 3 months old. Verify that the accountant's claim cannot be rejected at the at the 5% level of significance. Estimate the size of sample required to provide evidence, at significance level 0.05 and with power 0.9, of a reduction in the proportion from 70% to 60%.

5. A supplier claims that at least 95% of the parts it supplies meet the product specifications. In a sample of 500 parts received over the last 6 months, 36 were defective. Test the supplier's claim at the 5% level of significance.

6. In discussion of the Minitab Pulse.MTW data set earlier in the book it was noted that 35 students from a class of 92 ran on the spot, the decision whether or not to run

supposedly based on the outcome of the flip of a coin by the student. Do these data provide any evidence of 'cheating'?

7. Under the Weights and Measures (Packaged Goods) Regulations 1986, display of the text '330 ml e' on a bottle of beer with nominal content volume of 330 ml means that not more than 2.5% of bottles may be deficient in content volume by more than the tolerable negative error (TNE) specified for the nominal quantity (Trading Standards Net). The TNE for nominal volume of 330 ml is 3% of the nominal volume.

   A brewery has bottling machines which deliver amounts that are normally distributed with known standard deviation 6 ml and which are capable of filling bottles with nominal capacities of both 330 and 500 ml.

   (i) Explain why, when filling 330 ml bottles, the brewery should aim to set up the filling process to operate with a mean of 332 ml (to the nearest ml).

   (ii) Following set-up for a run of 330 ml bottles, after a run of 500 ml bottles, a sample of 24 bottles was checked and found to have the content volumes stored in Beer.xls. Use hypothesis testing to investigate whether or not the machine has been set up correctly.

8. This exercise has been created as an aid to understanding the concept of a confidence interval. The worksheet Strength.MTW contains data for the tensile strengths (N/mm^2) of a sample of 16 components. Assume that the production process for the components is known to operate with a standard deviation of 1.9 N/mm^2.

   Verify, using Minitab, that in a z-test of $H_0: \mu = 50.0$ versus $H_1: \mu \neq 50.0$ the decision would be to accept H_0. Repeat the test for all the other null hypotheses listed in Table 7.21 and record your decisions. Note the range of values for the population mean, μ, that would be accepted. By further changing of the mean value specified in the null hypothesis, determine, to 2 decimal places, the range of mean values that would be accepted. Check that the formula $\bar{x} \pm 1.96\sigma/\sqrt{n}$ gives the same results, apart form a small rounding error.

   The range of mean values that would be accepted is a *95% confidence interval for the mean* tensile strength. Check that the 95% confidence interval provided by Minitab confirms your calculations.

   Obtain a 99% confidence interval for the mean tensile strength. Note that with greater confidence we now have a wider range of mean values that would be accepted.

Table 7.21 Decisions from tests of hypotheses.

Null hypothesis H_0	Decision at 5% significance level (two-tailed z-test)
$\mu = 48.5$	
$\mu = 49.0$	
$\mu = 49.5$	
$\mu = 50.0$	Accept H_0
$\mu = 50.5$	
$\mu = 51.0$	
$\mu = 51.5$	

9. Obtain a 95% confidence interval for the mean content volume from the bottle data in Beer.xls. How does your answer confirm your earlier conclusion regarding set-up in Exercise 6?

10. Suppose that a quality manager claims that 70% of units pass final inspection first time and that you check a sample of 40 units from the database and find that 23 of them passed final inspection first time.

    Calculate the percentage of the sample that passed final inspection first time and note your 'gut feeling' concerning the manager's claim. Use Minitab to obtain a 95% confidence interval for the proportion of the population of units which pass final inspection first time and state whether the result lends support to the manager's claim or otherwise. Was your gut feeling supported by the statistical analysis?

    Investigate the situation where checking a sample of 400 units revealed 230 failures.

11. The workbook Verify.xls contains information on whether or not a series of units passed verification first time. Are the data consistent with the manufacturing operation achieving a first-time pass rate of 80%?

12. A manufacturer of automatic teller machines introduced changes to the procedure for installing the printer in the carcass as part of a process improvement initiative. Random samples of installation times for a technician before and after the changes are tabulated in Table 7.22 and available in Printer.MTW.

    Investigate the evidence for a reduction in the mean installation time using both parametric and nonparametric tests. Whenever possible check any assumptions required.

    As an exercise, perform the two-sample *t*-tests using the data as presented in the worksheet, using the data in stacked form, and finally using the summarized data. These three methods of presenting the data to Minitab correspond to **Samples in one column**, **Samples in different columns** and **Summarized data** in the dialog box for the two-sample *t*-test.

13. A process improvement project on the manufacture of light bulbs was carried out in order to compare two different types of lead wire. Misfed lead wires require operator intervention. Data on the average hourly number of misfed leads for 12 production runs with the standard type of wire and for 12 production runs with a modified type of wire are given in Misfeeds.MTW. Investigate the evidence for a process improvement using both parametric and nonparametric tests. Whenever possible, check any assumptions required.

14. The PCS-12 is a generic measure of physical health status. The measure has been devised in such a way that in the general population of people in good health it has mean 50 and standard deviation 10. Table 7.23 gives PCS-12 scores for a random sample of patients who had hip joint replacement operations carried out, both

Table 7.22 Before and after samples of installation times.

Before	64	39	59	31	42	52	43
After	19	41	29	45	37	35	39

Table 7.23 Before and after PCS12 scores for a sample of 12 patients.

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Pre	36	45	30	63	48	52	44	44	45	51	39	44
Post	39	42	33	70	53	51	48	51	51	51	42	50

immediately prior to the operation (Pre) and 6 months later (Post). The data are available in PCS12.MTW.

(a) Perform both parametric and nonparametric tests of hypotheses, investigating any assumptions required where possible, to investigate whether or not the data provide evidence of improvement in the physical health status of patients following the operation.

(b) Investigate whether or not the post-operative data are consistent with the mean for the general population.

15. Table 7.24 gives wear resistance data for four fabrics, obtained from a completely randomized single-factor experiment in which four samples of each one of a set of four fabrics of interest were tested. Wear was assessed by measuring the weight loss after a specific number of cycles in the wear-testing machine. The data, available in Fabrics.MTW, are from p.63 of *Fundamental Concepts in the Design of Experiments*, 5th edition, by Charles R. Hicks and Kenneth V. Turner, Jr, copyright © 1964, 1973, 1982, 1993, 1999 and used by permission of Oxford University Press, Inc., New York. Carry out a one-way ANOVA using Minitab, and perform diagnostic checks of assumptions and follow-up analysis if appropriate. Summarize your findings.

16. Sample sizes may be unequal in an experiment with a single factor. Box *et al.* (2005, p. 134) give an example on coagulation time for samples of blood from animals fed on a fixed set four diets A, B, C and D which were of interest. (Box *et al.*, 1978, p. 166). The data are available in Coagulation.MTW. (Reproduced by permission of John Wiley & Sons, Inc., New York.) Carry out an analysis of variance. The shorter the coagulation time is, the better from an animal health point of view. What recommendations would you make on the basis of the experiment?

17. Analyse the data in Exercises 15 and 16 using the Kruskal–Wallis procedure.

18. Analyse the data in Exercise 12 using ANOVA and verify that the P-value is exactly the same as that obtained from a two-sample t-test (assuming equal variances). In this situation the two tests are mathematically equivalent.

Table 7.24 Wear data for four fabrics.

Fabric			
A	B	C	D
1.93	2.55	2.40	2.33
2.38	2.72	2.68	2.40
2.20	2.75	2.31	2.28
2.25	2.70	2.28	2.25

Table 7.25 Yield data for five batches.

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
74	68	75	72	79
76	71	77	74	81
75	72	77	73	79

19. A company supplies a customer with many batches of raw material in a year. The customer is interested in high yields of usable chemical in the batches. For quality control of incoming material purposes three sample determinations of yield are made for each batch.

    The data, available in Batches.MTW and displayed in Table 7.25, are from p.78 of *Fundamental Concepts in the Design of Experiments*, 5th edition, by Charles R. Hicks and Kenneth V. Turner, Jr, copyright © 1964, 1973, 1982, 1993, 1999 and used by permission of Oxford University Press, Inc., New York. Show that about 87% of the variation in yield is due to batch-to-batch variation with the remaining 13% of variation being due to variation within batches.

20. If four brands of car tyre A, B, C and D were to be tested using four tyres of each type and four cars, explain why design 1 displayed in Table 7.26 would be unsatisfactory.

    Table 7.27 gives the design and the results for the experiment actually carried out. State the type of design used. Set up the data in Minitab, analyse them and report your findings.

21. Table 7.28 gives weekly revenue (£000) for three city restaurants of the same size belonging to a restaurant chain. The weeks were a random sample of weeks during

Table 7.26 Design 1.

Design 1	Car			
	P	Q	R	S
	A	B	C	D
Tyre Brand	A	B	C	D
	A	B	C	D
	A	B	C	D

Table 7.27 Design 2.

Design 2	Car			
	P	Q	R	S
	B(1.9)	D(1.6)	A(1.8)	C(1.4)
Tyre Brand and	C(1.7)	C(1.7)	B(1.8)	D(1.4)
Wear (mm)	A(2.2)	B(1.9)	D(1.6)	B(1.4)
	D(1.8)	A(1.9)	C(1.5)	A(1.8)

Table 7.28 Weekly turnover data for three restaurants.

	Restaurant		
Week	1	2	3
1	8.3	7.4	9.2
2	10.7	10.0	12.8
3	9.5	8.5	14.6
4	3.2	3.9	7.2
5	12.7	12.6	13.2

2004. Analyse the data, stored in Restaurants.MTW, and report your findings. Is there evidence from the data of a clear winner of 'Restaurant of the Year' from the point of view of revenue?

22. Set up the data from Exercise 14 as data from a randomized complete block experiment with the patients as blocks. Carry out an ANOVA and verify that the *P*-value for testing the effect of the operation is the same as was obtained from the paired *t*-test. Paired data experiments are a special case of randomized complete block designs.