Chapter 2: Statistical Tests, Confidence Intervals and Comparative Studies

2.0 Introduction

This chapter introduces the fundamental tools of statistical inference, viz., statistical significance tests and confidence intervals. It also introduces comparative studies.

A statistical significance test asks if an observed result differs from some hypothesised value by more than would be expected from purely chance variation. If it does, the result is declared to be ‘statistically significant’, i.e., the observed difference suggests (‘signifies’) that a systematic difference from the hypothesised value exists. For example, if the mean of a set of measurements on a reference material differs from the certified value by more than would be expected from purely chance measurement error, the measurement system would be said to be ‘biased’. In a comparative study, the test might ask if the long-run mean responses for two groups were the same (the groups might be patients treated either with a new drug or a placebo); if we reject this hypothesis, then we would conclude that the difference between the sample mean responses was not due to chance variation – in the case of the clinical study the observed effect of the drug would be assumed to be reproducible.

Where sample means are statistically significantly different, it makes sense to measure the systematic effect (of the drug, for example). Obviously, the sample difference gives such a measure, but it will also be clear, from the chance variation in the study, that the sample difference would change in a repeat of the study. What we would like to know is what difference would be seen in the long-run, i.e., if the study were so large as to eliminate the effects of chance variation. Put differently, we would like error bounds around our sample difference, which would allow for the chance variation, and within which the long-run value might be expected to lie. The statistical approach to producing such error bounds is to calculate a confidence interval; this is the second important statistical method introduced in this chapter.

The principal focus will be on sample means. However, the ideas underlying the statistical tests and confidence intervals introduced in this chapter are quite general: they underpin the rest of the material discussed in the course.

Section 2.1 uses a measurement method validation study to introduce significance tests. Section 2.2 discusses paired comparative studies. Confidence intervals are the focus of Section 2.3. Comparative studies for independent groups are discussed in Section 2.4. Finally, Section 2.5 describes a case study which reviews all the ideas of the earlier sections.
2.1 Statistical Significance Tests

Statistical significance tests are widely used (and often abused!) in the research literature across most disciplines where empirical data are assessed. The concepts involved in all such tests are the same irrespective of the specific purpose of the test. The ideas will be introduced here in the context of tests on sample means. Our first example asks if a measurement system is unbiased, i.e., does it get the right answer, on average, when some property of a material is measured and the correct answer is known.

Prefatory remarks

Statistical significance tests are concerned with identifying when observed values may be considered unusually large or small when compared to expected or hypothesised values. Before engaging with these tests it may be worthwhile to recall a similar situation encountered in Chapter 1.

In Example 1 of Chapter 1 the idea of a medical reference interval was discussed. Figure 2.1.1 shows a Normal distribution for the heights (in cm) of a population of women ($\mu=162.4$, $\sigma=6.284$) with cut-off values 150.1 and 174.7 which bracket 95% of the population. The cut-off values were calculated as $\mu \pm 1.96\sigma$ since $Z=\pm 1.96$ gives the corresponding central 95% of the standard Normal curve. To decide whether or not a particular women is unusually tall or short (i.e., unusually far from the population mean), we can compare her height directly to (150.1, 174.7) or, alternatively, standardise her height and compare it to the standard Normal values of $Z=\pm 1.96$. Thus, a woman of height 178.1 cm has a standardised value of:

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

$$z = \frac{178.1 - 162.4}{6.284} = 2.5$$

which clearly points to her being unusually tall. The standardisation simply expresses how far her height deviates from the population mean in terms of standard deviation units.

Figure 2.1.1: 95% reference limits
The statistical test introduced below follows exactly the same logic: a standardised observed value is compared to cut-off values on a reference distribution. It differs in two respects. Firstly, instead of a single height value, a sample average is the focus of our attention. This means that the reference distribution is the sampling distribution of the (standardised) sample mean. Secondly, since we do not know the population standard deviation, $\sigma$, the sample value, $s$, is used in the standardisation. These differences result in technical differences in carrying out the test, but the essential logic of the procedure remains unchanged.

**Example 1: A method validation study**

Virtually all scientific and technological investigations involve making measurements. The quality of the measurement systems used to collect data is, clearly, of paramount importance, since bad systems result in bad measurements, which in turn lead to bad decisions. For this reason good practice requires that measurement systems be validated before use. Our first example is concerned with deciding whether or not an analytical chemistry method leads to results which are unbiased, ‘unbiased’ means that, on average, the correct answer is obtained. Put differently, unbiased means that individual measurements may be larger or smaller than the certified value for the quantity being measured, but the deviations are considered to be purely chance fluctuations. If enough measurements were made and averaged, the deviations would cancel and result in the ‘true value’.

Pendl et al. [1] reported a new gas chromatographic method for determining total fat in foods and feeds. Several standard reference materials (SRMs) were measured by two laboratories, laboratory (A), where the method was developed, and a peer laboratory (B). One of these was a baby food powder with a certified value of $\mu_0=27.1\%$ fat. Each laboratory analysed the SRM on ten different days and reported the results shown below.

<table>
<thead>
<tr>
<th>% Fat Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 26.75 26.73 26.40 26.68 26.90 27.11 27.18 27.21 27.0 27.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laboratories</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>26.943</td>
<td>26.040</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.294</td>
<td>0.441</td>
</tr>
<tr>
<td>Sample size</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.1.1: Test results and summary statistics for the baby food study
An important question for the laboratory managers is whether their analytical systems are unbiased, i.e., are their results varying at random around a long-run value of $\mu_o = 27.1\%$? The statistical test introduced below essentially asks if the observed mean result is consistent with just chance variation away from the certified or ‘true value’ – if the answer is no, then the analytical system is considered to be biased.

For laboratory A, the average result is 26.9% fat. While this is lower than the certified value of $\mu_o = 27.1\%$ the deviation might simply be a consequence of the chance day-to-day variation that is evidently present in both measurement systems. To assess whether or not the deviation, $\bar{x} - \mu_o$, is large compared to the likely chance variation in the average result, it is compared to the estimated standard error of the sample mean $s/\sqrt{n}$. The standard error is a measure of the variability of $\bar{x}$ values, just as the standard deviation, s, is a measure of the variability in individual x values. The ratio:

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} \quad \text{(2.1)}$$

measures by how many (estimated) standard errors the sample mean differs from the certified value.

Assume that the laboratory is unbiased, i.e., that the measurements are indeed varying at random around the certified value of $\mu_o = 27.1\%$. In such a case, if many thousands of samples of size n were obtained under the same conditions and $\bar{x}$ and s calculated for each sample and then used to calculate the t-statistic using equation (2.1), the histogram of the resulting t-values would approximate the smooth curve shown in Figure 2.1. This curve, which is called the ‘sampling distribution’ curve and can be derived mathematically, is known as Student’s t-distribution; the total area under the curve is 1. The concept of a sampling distribution is important: it describes the distribution of values of the test statistic (here the t-ratio, equation (2.1)) that would be obtained if very many samples were obtained under the same conditions. Accordingly, it describes the expected distribution of summary measures (such as $\bar{x}$, s or t) and allows us to identify unusual values.

If the resulting t-value is exceptional, by reference to this distribution, i.e., if it lies far out in the tails of the sampling distribution, this would call into question the hypothesis that the measurements are varying around 27.1%. A value in the body of the sampling distribution, on the other hand, would be much more likely if the analytical system is unbiased. Accordingly, such a result would be consistent with the assumption of an unbiased measurement system.
In order to carry out the test, cut-off values (called ‘critical values’) which will define what is considered exceptional, must be decided upon. If we consider as exceptional a value whose magnitude would have only a 5% chance of being exceeded if the laboratory is unbiased, then Figure 2.1.2 shows that ±2.26 are the critical values in this case. A t-value either less than −2.26 or greater than +2.26 would be considered exceptional, and would call into question the unbiasedness of the analytical system.

Figure 2.1.2: Student's t-distribution with 9 degrees of freedom

The calculated t-statistic for Laboratory A is:

\[
t = \frac{26.943 - 27.1}{0.294 / \sqrt{10}} = -1.69
\]

This lies close to the centre of the t-distribution, so we do not reject the hypothesis that the analytical system in Laboratory A is unbiased\(^1\).

Student's t-distribution is similar to the standard Normal distribution (in that it is a symmetrical bell-shaped distribution centred on zero), but it gives somewhat larger cut-off values than the Normal, when bracketing a given percentage of the area under the curve. Thus, for the standard Normal curve the values ±1.96 bracket 95% of the area, while ±2.26 were the cut-off values above. The required t-value can be found in a table (see Statistical Table ST-2) which gives values that enclose 95%, 99% etc. in the centre of the distribution. The table is indexed by the number of degrees of freedom, which is a measure of how much information is available in computing \(s\), the estimate of \(\sigma\), the long-run or ‘true’ standard deviation for individual measurements. The number of degrees of freedom is the denominator used in calculating \(s\), \(n-1\) in this case.

\(^1\) Note that failure to reject the hypothesis that the analytical system is producing results that vary at random around \(\mu_0=27.1\%\) fat is not the same as proving that this is the case. All that can be asserted is that the observed result is *consistent with* such an average. The data would also be consistent with other long-run values close to 27.1%. For this reason a confidence interval, which brackets the set of likely values for the long-run mean \(\mu\) is, generally, a better way to analyse and report the results of such a study; see Section 2.3.
Table 2.1.2 gives examples of cut-off values for enclosing 95% of the t-curve for various sample sizes.

<table>
<thead>
<tr>
<th>Sample size n</th>
<th>Degrees of freedom</th>
<th>central 95% t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>4.30</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2.78</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>2.26</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>2.09</td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>2.05</td>
</tr>
<tr>
<td>120</td>
<td>119</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Table 2.1.2: Selected t-values for central 95% of curve area

Note that, as the sample size increases, the t-distribution becomes more and more like a standard Normal distribution and the 95% cut-off values get closer and closer to ±1.96. If the degrees of freedom are more than about 30 many statisticians will use ±2 as the cut-off values.

The significance testing procedure

The approach, described above, to decide whether or not an observed difference is sufficiently large to warrant concluding the existence of a systematic difference from an hypothesised value, is the basis of all statistical tests. The various elements involved in this procedure will now be discussed and the standard statistical terminology will be introduced, using Laboratory B as an illustration. As we proceed, we will see that these same elements appear in all the statistical tests that are introduced in this and later chapters.

• Specify the hypothesis to be tested and the alternative that will be decided upon if this is rejected

The hypothesis to be tested is referred to as the 'null hypothesis' and is labelled $H_0$. The 'alternative hypothesis' is labelled $H_1$. Here we have:

$$H_0: \mu = 27.1\%$$
$$H_1: \mu \neq 27.1\%$$

where $\mu$ is the long-run mean for a very large number (in principle, an infinitely large number) of measurements on the reference material.

The null hypothesis is given special status: it is assumed to be true unless the measurement data clearly demonstrate otherwise. Thus, the significance testing procedure parallels the legal principle that the accused is considered
innocent until proven guilty: for conviction, the evidence must be such that the assumption of innocence is called into question.

• Specify a statistic which measures departure from the null hypothesis $\mu=\mu_0$.
  (here $\mu_0=27.1$)

  The test statistic

  $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  measures by how many (estimated) standard errors $\bar{x}$ deviates from $\mu_0$, i.e., it scales the deviation of $\bar{x}$ from $\mu_0$ in terms of standard error units. The t-distribution with $n-1$ degrees of freedom, see Figure 2.1, describes the frequency distribution of $t$-values that might be expected to occur if, in fact, the null hypothesis is true.

• Define what will be considered an exceptional outcome

  We consider a value of the test statistic to be exceptional if it has only a small chance of occurring when the null hypothesis is true. The probability chosen to define an exceptional outcome is called the ‘significance level’ of the test and is usually labelled $\alpha$; we choose $\alpha=0.05$, here$^2$. The cut-off values on the sampling distribution, defined by the significance level, are called ‘critical values’ and the regions of the curve beyond them are ‘critical’ or ‘rejection’ regions. Figure 2.1.1 shows that values either smaller than $-2.26$ or greater than $2.26$ occur with probability 0.05, where the test statistic has 9 degrees of freedom.

  Since the t-distribution describes what $t$-values might be expected when the null hypothesis is true, this implies that when we decide to reject the null hypothesis if the test statistic is in the rejection region, we automatically run a risk of $\alpha=0.05$ of rejecting the null hypothesis even though it is correct. This is a price we must pay when using the statistical significance testing procedure.

• Calculate the test statistic and compare it to the critical values

  For Laboratory B the test statistic is:

  $$t = \frac{26.04 - 27.10}{0.441/\sqrt{10}} = -7.60$$

$^2$ The implications of the choice of significance level are discussed more fully in Chapter 4; see Section 4.2.
This is much smaller than −2.26. Accordingly, we reject the null hypothesis that the laboratory measurements were varying around 27.1% and conclude that the analytical system in Laboratory B is biased downwards.

Note that the entire procedure and the criterion for rejecting the null hypothesis are defined before the calculations are carried out, and often before any measurements are made. This ensures that the decision criterion is not influenced by the data generated.

For Laboratory B the difference between the observed mean of 26.04 and the certified value of 27.10 is said to be ‘statistically significant’. The difference between the mean for Laboratory A and the certified value would be declared to be ‘not statistically significant’. ‘Statistically significant’ simply means ‘unlikely to be due to chance variation’ – it is evidence that a reproducible effect exists (remember that the root meaning of the verb ‘to signify’ is ‘to signal or indicate’ something). Thus, ‘statistically significant’ does not (in itself) mean ‘important’. Whether or not the observed difference is of any practical importance is an entirely different question. The answer will be context specific and will require an informed judgment on the part of someone with an understanding of the context.

Computer Analysis and p-values

An analysis of the Laboratory B data is shown in Table 2.1.3: this was generated using Minitab, but much the same output would be obtained from any statistical package.

**One-Sample T: Fat-%**

Test of mu = 27.1 vs not = 27.1

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat-%</td>
<td>10</td>
<td>26.040</td>
<td>0.441</td>
<td>0.139</td>
<td>(25.725, 26.355)</td>
<td>−7.61</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2.1.3: T-test and confidence interval for Laboratory B data

The t-value is −7.61 (note that the standard deviation is really 0.4406, hence the rounding difference when compared to our value of −7.60, which is based on 0.441) and this has an associated p-value of 0.000, i.e., <0.0005. The p-value is the area in the tails of the t-distribution (with 9 degrees of freedom), i.e., the sum of the areas to the right of 7.61 and to the left of −7.61; there is, in fact, an area of 0.000016 on each side, but Minitab reports only three decimal places. The t-distribution is the sampling distribution of the test statistic, **when the null hypothesis is true**. Hence, the tail areas give us the probability of seeing a more extreme t-value (in either direction) than the one obtained in the study, if the null hypothesis were true. The very small p-value indicates that our observed t-value is highly improbable under the hypothesis that the average around which
the measurement system is varying is indeed 27.1; this implies that the null hypothesis \( \mu_0 = 27.1 \) is implausible and should be rejected.

The use of the p-value as a basis for deciding whether or not to reject the null hypothesis is entirely consistent with our earlier approach of choosing a significance level, which in turn determines the critical values. If the p-value were exactly 0.05 it would mean that the observed test statistic fell exactly on one of the critical values. If the p-value is smaller than 0.05 then it means that the observed test statistic is further out into the tails than one of the critical values – and thus the null hypothesis should be rejected. For those who choose a significance level of \( \alpha = 0.05 \), a p-value less than 0.05 means that the result is statistically significant and the null hypothesis should be rejected. P-values greater than 0.05 then lead us to accept (at least provisionally) the null hypothesis. Thus, large p-values support the null hypothesis, while small p-values lead us to reject it. Our choice of significance level indicates what we consider large or small.

Since the choice of significance level is arbitrary (though \( \alpha = 0.05 \) is almost universally used, for no very good reason) many consider p-values as a more appropriate way of reporting the result of a significance test. P-values provide a measure of the strength of the evidence against the null hypothesis and, thus, allow the reader to form his or her own judgement of the significance of the result.

The use of p-values is, also, a simple and convenient way of reporting the results of statistical tests, without requiring the availability of statistical tables. Thus, in Chapter 1, Figure 1.3.2 shows a Normal plot of sample data used to explain the logic underlying such plots. The Figure shows, also, the results of an Anderson-Darling test of the hypothesis that the data come from a Normal distribution – it reports a p-value of 0.971, which supports the null hypothesis. Similarly, the Anderson-Darling test of Figure 1.3.6 gives a p-value of \( p < 0.005 \), which strongly rejects the assumption that the data (\( \beta \)-OHCS values) come from a Normal distribution. We do not need to resort to tables to understand what the test result indicates – indeed we do not even need to know how to calculate the test statistic! Simply knowing how to interpret the p-value allows us to interpret the result of the test reported by the computer analysis.

The computer output in Table 2.1.3 also contains a ‘confidence interval’ – confidence intervals are discussed in Section 2.3.

**Exercise**

2.1.1 A filling line has been in operation for some time and is considered stable. Whenever a product changeover takes place it is standard operating procedure to weigh the contents of 10 containers, as a process control check on the target setting. The most recent changeover was to a target value of \( \mu = 21 \)g. Table 2.1.4 shows summary results for the ten
measurements. Carry out t-tests to determine whether or not the target has been achieved. Use a significance level of $\alpha=0.05$.

<table>
<thead>
<tr>
<th></th>
<th>Head 1</th>
<th>Head 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>21.06</td>
<td>21.33</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.118</td>
<td>0.360</td>
</tr>
</tbody>
</table>

Table 2.1.4: Fill head data (g)

**Model Validation**

The statistical model that underlies the t-test requires that the data values are independent of each other and that they behave as if they come from a single Normal distribution. Since the material was measured on ten different days, the independence assumption is likely to hold. If we knew the time order in which the measurements were made then we could draw a time-series plot, to check that there were no trends, which would call into question the assumption of a single distribution. For example, a once-off shift or trend upwards or downwards in the plot would suggest changes in the mean level of the measurement system. A Normal probability plot assesses the Normality assumption, though with as few as ten observations only strong departures from Normality would be likely to be detected.

Figure 2.1.3 shows a Normal plot for the fat data: the data vary around a straight line and the p-value for the Anderson-Darling test is large: there is no reason to question the Normality of the data. Accordingly, we can have confidence in the conclusions drawn from our statistical test.
Exercise

2.1.2 In Chapters 1 and 2 of Mullins, 2003 [2], several examples from a Public Analyst’s laboratory of the analysis of sugars in orange juice were presented. Exercise 2.2 (page 49) gave 50 values of the recoveries of glucose from control samples used in monitoring the stability of the analytical system. The controls were either soft drinks or clarified orange juice, spiked with glucose. The summary statistics for the 50 results are: mean=98.443 and SD=2.451. Carry out a t-test, using a significance level of 0.05, of the hypothesis that the analytical system is unbiased, i.e., that the long-run average recovery rate is 100%.

One-sided tests

The tests described above are known as two-sided tests and the null hypothesis is rejected if either an unusually large or an unusually small value of the test statistic is obtained (they are also referred to as two-tailed tests – the alternative hypothesis is two-sided, leading to critical or rejection regions in two tails). In special circumstances, we would want to reject the hypothesis only if the observed average were large; for example, in a public health laboratory if we were measuring the level of trace metals in drinking water or pesticide residues in food to check if they exceeded regulatory limits. In other circumstances we would reject the hypothesis if the observed average were small; for example, if we were assaying a raw material for purity - our next example is based on the latter case.
Example 2: Acceptance sampling

Test portions are sampled from each of five randomly selected drums from a consignment containing a large number of drums of raw material; the test portions are then assayed for purity. The contract specified an average purity of at least 90%. If the average of five results is \( \bar{x} = 87.9\% \) and the standard deviation is \( s=2.4\% \), should the material be accepted?

To carry out a test we first specify null and alternative hypotheses:

\[
H_0: \mu \geq 90 \\
H_1: \mu < 90
\]

where \( \mu \) is the average purity of the entire consignment.

Here the hypotheses are directional. A result in the centre or in the right-hand tail of the sampling distribution curve would support the null hypothesis while one in the left-hand tail would support the alternative hypothesis. We specify the significance level to be \( \alpha=0.05 \). When the t-statistic:

\[
t = \frac{\bar{x} - 90}{s/\sqrt{n}}
\]

is calculated, it will be compared to the left-hand tail of a t-distribution with 4 degrees of freedom (see Figure 2.1.4). The null hypothesis will be rejected if the calculated t-statistic is less than the critical value. Statistical Table ST-2 gives a value of \(-2.13\) as the cut-off value which has an area of 0.05 below it, for a curve with 4 degrees of freedom.

![Figure 2.1.4 One-tail critical region for the t-distribution with 4 df.](image-url)
The test statistic is

$$t = \frac{87.9 - 90}{\frac{2.4}{\sqrt{5}}} = -1.96$$

And, since this does not lie in the critical region, we do not reject the null hypothesis that the average purity is at least 90%. Given the large variability involved in sampling and measurement error (s=2.4), a sample average value of 87.9, based on five measurements, is not sufficiently far below 90 to lead us to conclude that the average batch purity is below specification. Such a deficit could have arisen by chance in the sampling and measurement processes.

The null hypothesis being tested by the significance test is whether the results are varying at random around the lower specification bound of 90%. This might not be the case for two reasons, viz., the material could be out of specification or the analytical system could be biased. In carrying out the test as we did above, we implicitly assumed that the analytical system is unbiased. In practice, there is not much point in making measurements (or rather it is dangerous to do so!) without first validating the measurement system – this applies equally to social science measurements as it does to engineering or scientific measurements.

One or two-sided tests?

Note that if we had carried out a two-sided test using the same significance level of \( \alpha = 0.05 \) the critical values would have been \( \pm 2.78 \), which means that values between \(-2.13\) and \(-2.78\) would be statistically significant on a one-sided test but ‘non-significant’ on a two-sided test. In an academic context where \( \alpha = 0.05 \) is the conventionally used significance level there can be a temptation to switch from a two-sided to a one-sided test in order to get a ‘significant’ result – especially if this could make the difference between the work being published or not. Many statisticians would be strongly of the view that in scientific research a two-sided test would be natural; see quotation from Altman [3] below (Altman is a medical statistician) and the footnote on page 71.

There are several issues involved in this question. Suppose we obtain a test statistic corresponding to a p-value of 0.08 on a two-sided-test: this would be considered ‘non-significant, but if we switched to a one-sided test it would be

---

“In rare cases it is reasonable to consider that a real difference can occur in only one direction, so that an observed difference in the opposite direction must be due to chance. ....One sided tests are rarely appropriate. Even when we have strong prior expectations, for example that a new treatment cannot be worse than an old one, we cannot be sure that we are right. If we could be sure we would not need to do an experiment! If it is felt that a one-sided test really is appropriate, then this decision must be made before the data are analysed; it must not depend on what the results were. The small number of one-sided tests that I have seen reported in published papers have usually yielded p-values between 0.025 and 0.05, so that the result would have been non-significant with a two-sided test. I doubt that most of these were pre-planned one-sided tests.”
significant’, since the p-value would now be 0.04. However, the conventional significance level of α=0.05 is quite arbitrary – how many of us could tell the difference between a 0.08 chance and a 0.05 chance? Then, there is the question of having a rigid cut-off on one side of which (p=0.049) the result is ‘significant’ (and, therefore, worthy of notice) and on the other side of which (p=0.051) the result is ‘non-significant’ (and, therefore, not worthy of attention, or perhaps, publication). There is, clearly, no practical difference between p-values of 0.049 and 0.051, so why should we act differently on seeing one or the other as the result of a statistical test. The correct interpretation of a p-value of 0.08 (at least in my view!) is that there is evidence to suggest that the null hypothesis may be rejected, but that the evidence is weak. If the observed result is considered of practical importance (the most important aspect of any study outcome) then the study provides evidence that further work may be worthwhile.

2.2 Paired Comparisons

In many situations data are collected in such a way that two measured values have a special relationship to each other, and it is the difference between the two values that is of interest. A group of students is assessed on a competency test, given training, and then re-tested; interest then centres on the changes or improvements in the scores. A patient is measured on some health indicator (e.g., cholesterol level), put on a treatment, and then re-tested later; again, the change in the test result (a reduction in cholesterol level) is the quantity of interest. Twins have been extensively studied by psychologists to determine differences, e.g., twins separated at birth and raised in different environments have been compared to assess the contributions of ‘nature’ and ‘nurture’, i.e., genetic versus ‘environmental’ effects.

Industrial studies, also, are often designed in such a way that the resulting data may be considered matched or paired. Example 4 below is based on a comparison of two measurement systems, but the ideas are applicable to comparisons of production methods also. In all such comparisons, we focus on differences, thus reducing two sets of numbers to one: this brings us back to the same methods of analysis (one-sample t-test and the corresponding confidence interval, which will be discussed in Section 2.3) used for our validation study, Example 1. Although the test is identical to the one-sample t-test, it is generally referred to as a ‘paired t-test’, because the raw data come in the form of pairs.

Example 3: An animal fertility study

The endometrium functions as a lining for the uterus. Retinol-binding protein (RBP) is a major secretory product of the endometrium and is assumed to be of importance for early embryonic development. In a preliminary study, the concentrations of RBP in uterine fluid from both the ipsi and contra sides of the

---

4 I am grateful to Sinead Waters, a former student in the Postgraduate Diploma in Statistics and a researcher at Teagasc, for supplying me with the data.
uterus for 16 cows were measured by enzyme linked immunoassay to determine if there was a difference between them on the day of ovulation (day 15 of the oestrus cycle). The ipsi side of the uterus is the side where the fertilised egg implants and, thus, if it secreted more RBP than the contra side, this would suggest a direct role for RBP in implantation. The RBP data and some summary statistics (pg/μg protein) are shown in Table 2.2.1.

<table>
<thead>
<tr>
<th>Ipsi</th>
<th>Contra</th>
<th>Difference</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>8085</td>
<td>6644</td>
<td>1441</td>
<td>7364.5</td>
</tr>
<tr>
<td>8544</td>
<td>5818</td>
<td>2726</td>
<td>7181</td>
</tr>
<tr>
<td>9002</td>
<td>8942</td>
<td>60</td>
<td>8972</td>
</tr>
<tr>
<td>7786</td>
<td>6939</td>
<td>847</td>
<td>7362.5</td>
</tr>
<tr>
<td>9498</td>
<td>8594</td>
<td>904</td>
<td>9046</td>
</tr>
<tr>
<td>5906</td>
<td>5488</td>
<td>418</td>
<td>5697</td>
</tr>
<tr>
<td>7078</td>
<td>6124</td>
<td>954</td>
<td>6601</td>
</tr>
<tr>
<td>9766</td>
<td>8137</td>
<td>1629</td>
<td>8951.5</td>
</tr>
<tr>
<td>7109</td>
<td>6907</td>
<td>202</td>
<td>7008</td>
</tr>
<tr>
<td>7802</td>
<td>6154</td>
<td>1648</td>
<td>6978</td>
</tr>
<tr>
<td>8213</td>
<td>7709</td>
<td>504</td>
<td>7961</td>
</tr>
<tr>
<td>7184</td>
<td>8235</td>
<td>-1051</td>
<td>7709.5</td>
</tr>
<tr>
<td>9824</td>
<td>9711</td>
<td>113</td>
<td>9767.5</td>
</tr>
<tr>
<td>7136</td>
<td>6514</td>
<td>622</td>
<td>6825</td>
</tr>
<tr>
<td>7216</td>
<td>6907</td>
<td>309</td>
<td>7061.5</td>
</tr>
<tr>
<td>7708</td>
<td>5413</td>
<td>2295</td>
<td>6560.5</td>
</tr>
</tbody>
</table>

Table 2.2.1: RBP data and summary statistics

**Statistical Model**

It is clear from Table 2.2.1 that there is considerable variation both between animals (the mean RBP levels vary from 5697 to 9767) and within animals (the within-animal differences (ipsi – contra) vary from −1051 to 2726). The statistical task is to identify and measure a possible systematic ipsi-contra difference in the presence of this chance variation.

A simple statistical model for the within-animal differences provides a basis for our analysis. We will assume that a population of such animals would result in a Normal curve for the ipsi-contra differences, with some mean μ and some standard deviation σ, as shown in Figure 2.2.1. We assume also that the differences for the sixteen animals are independent of each other.
If the long-run mean of this curve, $\mu$, is zero, then there is no systematic ipsi-contra difference in RBP levels – what we see is simply chance variation from cow to cow. If, on the other hand, $\mu$ is non-zero, then there exists a systematic ipsi-contra RBP level difference. Of course, on top of this possible systematic difference there is an additional chance component which varies from cow to cow.

In this section we will carry out a $t$-test on the differences to address the statistical question “is $\mu$ equal to zero?”, following exactly the same procedure used for our earlier analytical validation example (Example 1). Later (see Section 2.3), we will address the issue of estimating the size of $\mu$.

**Model validation**

Our statistical model assumes that our sample differences are independent of each other and come from a single stable Normal distribution. Are these assumptions valid?

One possibility that would invalidate our model, would be a relationship between the magnitude of the differences and the average RBP levels for the cows. If this were the case it would suggest that we do not have a single underlying Normal curve, but several curves with different standard deviations, which depend on the average RBP levels. Figure 2.2.2 shows the differences plotted against the means of the RBP measurements for the 16 cows.
There is no suggestion of any kind of systematic relationship between the differences and the means. If, for example, the data were fanning outwards as we move from the LHS to the RHS of the picture (i.e., larger positive or negative values for larger means than for smaller ones) it would undermine our assumption of a single underlying distribution with a constant standard deviation. This does not happen here.

Is it reasonable to assume that this underlying distribution is Normal? Figure 2.2.3 shows a Normal plot of the differences. The plotted points vary around a straight line and the Anderson-Darling test statistic has a p-value of $p=0.53$, which supports the Normality assumption. Note that in small datasets (even those generated from a Normal distribution) it is not uncommon that there will be one or two values at the edges of the data that appear somewhat out of line. We should not, therefore, be too concerned at the low value in Figure 2.2.3.
We do not have the necessary information to investigate independence, so we will assume it holds in this case. Possible reasons for non-independence might be genetic relations between the cows in the study or measurement problems. If we had the time order in which the measurements were made, it would be a good idea to draw a time-series plot – drifts or shifts in this might suggest problems with the measurements, which could carry implications for the independence of the resulting data.

Having established the reasonableness of the underlying model, we can proceed with our formal statistical test.

**Significance Test**

To carry out the test we first specify the competing hypotheses as:

\[
H_0: \mu = 0 \\
H_1: \mu \neq 0
\]

And choose a significance level of \(\alpha=0.05\). For a two-sided test with degrees of freedom \(n-1=15\), we use critical values of \(t_c=\pm2.13\), as shown in Figure 2.2.4.
Figure 2.2.4: The reference distribution for the t-test

We now calculate the test statistic and compare it to the critical values.

The test-statistic\(^5\) is

\[
t = \frac{\bar{d} - \mu_0}{s / \sqrt{n}}
\]

\[
t = \frac{851 - 0}{933 / \sqrt{16}} = 3.65
\]

Since the test statistic \( t=3.65 \) greatly exceeds the critical value of 2.13, we reject the null hypothesis: the long-run mean \( \mu \) does not appear to be zero. Since the sample difference is positive, we conclude that, on average, the ipsi side of the uterus secretes higher concentrations of RBP than does the contra side.

Computer Analysis

The computer output shown in Table 2.2.2 gives summaries of the data. Note that the line for “differences” gives the mean of the differences, which is arithmetically the same as the difference between the means for ipsi and contra. It also gives the standard deviation of differences. Note, however, that this cannot be calculated directly from the standard deviations for ipsi and contra (a covariance term would be required). The standard error is just the standard deviation divided by the square root of the sample size, \( n=16 \).

\(^5\) Zero has been inserted explicitly in the formula to emphasise that what is being examined is the distance from \( \bar{d} \) to the hypothesised mean of zero.
Paired T for Ipsi - Contra

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ipsi</td>
<td>16</td>
<td>7991</td>
<td>1104</td>
<td>276</td>
</tr>
<tr>
<td>Contra</td>
<td>16</td>
<td>7140</td>
<td>1284</td>
<td>321</td>
</tr>
<tr>
<td>Difference</td>
<td>16</td>
<td>851</td>
<td>933</td>
<td>233</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (354, 1349)
T-Test of mean difference = 0 (vs not = 0): T-Value = 3.65  P-Value = 0.002

Table 2.2.2: A Minitab analysis of the RBP data

The p-value of 0.002 shown in the table is the probability of obtaining, by chance, a more extreme test statistic than that calculated for our study (t=3.65), if the null hypothesis were true. Figure 2.2.5 shows the sampling distribution of the test statistic when the null hypothesis is true – this is Student’s t-distribution with 15 degrees of freedom. The areas to the right of 3.65 and to the left of −3.65 are 0.001, respectively. This means that the probability of obtaining a t-value further from zero than these two values is 0.002, the p-value shown in the table.

![Diagram showing t-distribution and critical values](image.png)

Figure 2.2.5: A schematic representation of the calculation of the p-value (3.65 would be further out into the tail)

The small p-value implies that the null hypothesis is implausible and leads us to reject it. While simply inspecting Figure 2.2.2 strongly suggests that the observed differences do not vary at random around zero (as required by the null hypothesis) the t-test provides an objective measure (via the p-value) of how unlikely our observed values would be if the null hypothesis were true.

Calculating tail areas (p-values) for t-distributions is, conceptually, the same as calculating tail areas for the standard Normal curve, as was done in Chapter 1.
The mathematics, however, is much more difficult – fortunately we have computers to take care of the calculations!

Our analysis of the RBP data leads us to conclude that the ipsi side of the uterus secretes more RBP than does the contra side – this has scientific implications as outlined in the introduction. In Section 2.3 we will return to this dataset and obtain an estimate, together with error bounds, for the magnitude of this difference in the population of cows.

**Exercise**

2.2.1 Rice [4] presents data from a study by Levine on the effect of cigarette smoking on platelet aggregation. Blood was taken from 11 subjects before and after they smoked a cigarette. The background to the study is that it is known that smokers suffer from disorders involving blood clots more often than non-smokers, and platelets are involved in the formation of blood clots. Table 2.2.3 gives a measure of platelet aggregation (where larger values mean more aggregation; units are the percentages of platelets that aggregated) for each subject before and after exposure to the stimulus.

Is there evidence that smoking even one cigarette affects the ability of platelets to aggregate? Note: the mean of the differences is 10.27 and the standard deviation is 7.98.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>27</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>44</td>
<td>56</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>67</td>
<td>82</td>
<td>15</td>
</tr>
<tr>
<td>53</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>53</td>
<td>80</td>
<td>27</td>
</tr>
<tr>
<td>52</td>
<td>61</td>
<td>9</td>
</tr>
<tr>
<td>60</td>
<td>59</td>
<td>-1</td>
</tr>
<tr>
<td>28</td>
<td>43</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 2.2.3: Platelet aggregation data before and after smoking (percentages)
**Example 4: A Laboratory Comparison Study**

Measurement systems are of fundamental importance in science, industry and commerce – they provide the data on which we make decisions or build our theories. I have, several times, encountered situations where the same materials (typically batches of raw materials) were measured by different laboratories and the measurements did not agree. Obviously, such disagreements can have serious commercial implications. The example discussed below relates to batches of tablets made in Ireland and exported to Asia. The data are taken from a more extensive case study which is described in Mullins (2003 [2], pages 154 – 159).

Modern tablets are manufactured in such a way that, depending on requirements, their active ingredients may be programmed to be released into the stomach over many hours. Consequently, the quality control laboratories of manufacturers carry out dissolution tests in which tablets are placed in flasks containing a liquid which mimics the contents of the stomach. The flasks are held at human stomach temperature in a water bath and test samples are extracted at pre-determined time intervals and the percentage dissolution of the active ingredient is measured over the planned release profile of the tablet.

Forty consecutive batches of tablets were measured both in the manufacturer’s laboratory and that of a major customer. Table 2.2.4 shows the percentage of the active ingredient that was measured by each of the laboratories as having dissolved after one hour; the laboratories are arbitrarily labelled as A and B.

<table>
<thead>
<tr>
<th>Batch</th>
<th>A</th>
<th>B</th>
<th>Batch</th>
<th>A</th>
<th>B</th>
<th>Batch</th>
<th>A</th>
<th>B</th>
<th>Batch</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.7</td>
<td>40.6</td>
<td>11</td>
<td>40.9</td>
<td>41.4</td>
<td>21</td>
<td>34.1</td>
<td>36.2</td>
<td>31</td>
<td>38.0</td>
<td>39.6</td>
</tr>
<tr>
<td>2</td>
<td>42.1</td>
<td>43.6</td>
<td>12</td>
<td>41.3</td>
<td>44.9</td>
<td>22</td>
<td>39.3</td>
<td>40.6</td>
<td>32</td>
<td>42.9</td>
<td>43.2</td>
</tr>
<tr>
<td>3</td>
<td>37.0</td>
<td>39.0</td>
<td>13</td>
<td>40.5</td>
<td>41.9</td>
<td>23</td>
<td>37.3</td>
<td>40.9</td>
<td>33</td>
<td>37.3</td>
<td>38.7</td>
</tr>
<tr>
<td>4</td>
<td>37.6</td>
<td>39.0</td>
<td>14</td>
<td>37.8</td>
<td>37.3</td>
<td>24</td>
<td>38.2</td>
<td>38.8</td>
<td>34</td>
<td>42.8</td>
<td>42.6</td>
</tr>
<tr>
<td>5</td>
<td>35.2</td>
<td>38.3</td>
<td>15</td>
<td>39.6</td>
<td>39.6</td>
<td>25</td>
<td>37.7</td>
<td>39.8</td>
<td>35</td>
<td>40.2</td>
<td>38.2</td>
</tr>
<tr>
<td>6</td>
<td>43.9</td>
<td>48.9</td>
<td>16</td>
<td>39.9</td>
<td>44.4</td>
<td>26</td>
<td>42.0</td>
<td>44.7</td>
<td>36</td>
<td>49.2</td>
<td>51.3</td>
</tr>
<tr>
<td>7</td>
<td>39.0</td>
<td>40.5</td>
<td>17</td>
<td>39.2</td>
<td>41.2</td>
<td>27</td>
<td>36.3</td>
<td>39.4</td>
<td>37</td>
<td>40.3</td>
<td>43.0</td>
</tr>
<tr>
<td>8</td>
<td>39.2</td>
<td>40.8</td>
<td>18</td>
<td>39.4</td>
<td>40.1</td>
<td>28</td>
<td>39.8</td>
<td>42.7</td>
<td>38</td>
<td>41.7</td>
<td>41.8</td>
</tr>
<tr>
<td>9</td>
<td>37.8</td>
<td>40.0</td>
<td>19</td>
<td>40.2</td>
<td>41.6</td>
<td>29</td>
<td>45.8</td>
<td>44.3</td>
<td>39</td>
<td>40.7</td>
<td>41.0</td>
</tr>
<tr>
<td>10</td>
<td>46.3</td>
<td>46.1</td>
<td>20</td>
<td>37.8</td>
<td>39.3</td>
<td>30</td>
<td>41.6</td>
<td>43.7</td>
<td>40</td>
<td>40.0</td>
<td>44.8</td>
</tr>
</tbody>
</table>

*Table 2.2.4: The percentage one-hour dissolutions for the 40 batches*

If we were to study the 40 measurements from either laboratory (or their averages) we would learn about batch-to-batch variation in dissolution rate. This would, of course, be of interest to the manufacturing engineers, but it would not tell us anything about the possible difference (relative bias) between the two
measurement systems. The set of 40 differences, on the other hand, will allow us to investigate if one laboratory consistently measures higher than the other. Here, we will carry out a paired t-test to establish whether or not there is a relative bias between the laboratories. In Section 2.3 we will calculate a confidence interval to measure the size of any such bias. Note that without a ‘gold standard’ we cannot establish which (if either) laboratory produces correct results – we can only establish that they do or do not agree.

Before carrying out the formal statistical test we will, as before, investigate the assumptions that underlie the test. Plotting the data in time order, in the form of a control chart (see Chapter 1, Section 1.6 for a discussion of control charts) does not show any time-related variation in the differences which might call the assumption of data independence into question – the plotted points appear to vary at random around the centre line (which is the average of the 40 observations). Figure 2.2.6 suggests we have a stable distribution of differences. It also suggests that the differences do not vary about a long-run mean of zero. We will, nevertheless, carry out a significance test of this hypothesis.

Figure 2.2.7 in which the laboratory differences are plotted against the average of the two measurements for each batch shows no tendency for batches with higher dissolution rates to have either greater or smaller variation in differences than batches with lower dissolution rates. Any systematic relationship between the variability of the differences and the batch means would suggest that the assumption of a single stable distribution (constant mean and standard deviation) of differences was false. This assumption appears to be reasonable based on the evidence of Figures 2.2.6 and 2.2.7.

Figure 2.2.6: Individuals chart for the laboratory differences
Figure 2.2.7: Laboratory differences versus laboratory means for the 40 pairs of data

Figure 2.2.8: Normal plot of laboratory differences

Figure 2.2.8 shows a Normal plot of the differences. The scatterplot of the 40 differences versus their corresponding Normal scores produces a fairly straight line, as would be expected if the data came from a Normal distribution. The Anderson-Darling test statistic has a corresponding p-value of 0.558 which supports the assumption of Normality. Figures 2.2.6-2.2.8 taken together provide reasonable assurance that the assumptions underlying the t-test (as shown graphically in Figure 2.2.1) are valid in this case.
**Significance Test**

The summary statistics for the 40 batches are given below in Table 2.2.5

<table>
<thead>
<tr>
<th></th>
<th>Lab-B</th>
<th>Lab-A</th>
<th>Individual differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>41.60</td>
<td>40.04</td>
<td>1.555</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>3.04</td>
<td>2.96</td>
<td>1.607</td>
</tr>
</tbody>
</table>

Table 2.2.5: Summary statistics for the laboratory study

To carry out the test we first specify the competing hypotheses as:

\[ H_0: \mu = 0 \]
\[ H_1: \mu \neq 0 \]

and choose a significance level of \( \alpha=0.05 \). Here, \( \mu \) is the relative bias, i.e., the long-run mean difference between the results that would be produced by the two laboratories if a very large number of batches (in principle, an infinite number) were analysed by both laboratories. For a two-sided test with degrees of freedom \( n-1=39 \), we use critical values of \( t_c=\pm 2.02 \). The test-statistic is

\[
t = \frac{\bar{d} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{1.555 - 0}{1.607} = \frac{6.12}{\sqrt{40}}
\]

Since the test statistic \( t=6.12 \) is much greater than the critical value of 2.02, we reject the null hypothesis: the long-run mean \( \mu \) does not appear to be zero – we conclude that there is a relative bias between the two laboratories. Laboratory B gives higher results, on average. In the next section we will put error bounds around the sample mean difference of 1.56 in order to have a more trustworthy estimate of the long-run difference (the relative bias) between the two laboratories.
Exercise

2.2.2 Disagreements arose regarding the purity of material being supplied by one plant to a sister plant in a multinational corporation [2]. The quality control (QC) laboratories at the two plants routinely measured the purity of each batch of the material. The results (units are percentage purity) from the six most recent batches as measured by each laboratory are presented in Table 2.2.6 below. Carry out a paired t-test to determine if there is a relative bias between the laboratories.

<table>
<thead>
<tr>
<th>Batch</th>
<th>Lab-1</th>
<th>Lab-2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.01</td>
<td>90.45</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>89.34</td>
<td>89.94</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>89.32</td>
<td>90.05</td>
<td>0.73</td>
</tr>
<tr>
<td>4</td>
<td>89.11</td>
<td>89.62</td>
<td>0.51</td>
</tr>
<tr>
<td>5</td>
<td>89.52</td>
<td>89.99</td>
<td>0.47</td>
</tr>
<tr>
<td>6</td>
<td>88.33</td>
<td>89.40</td>
<td>1.07</td>
</tr>
</tbody>
</table>

mean 89.272, 89.908, SD 0.553, 0.364

<table>
<thead>
<tr>
<th>Mean</th>
<th>89.272</th>
<th>89.908</th>
<th>0.637</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>0.553</td>
<td>0.364</td>
<td>0.237</td>
</tr>
</tbody>
</table>

Table 2.2.6: Purity measurements on 6 batches made by two laboratories

2.2.3. The QC manager who carried out the analysis of the laboratory comparison data in the preceding example noticed that the difference between the results for the last batch was by far the largest difference in the dataset. Since the result for this batch from laboratory 1 was the only value in the dataset which was less than 89.0, she wondered if it was correct and whether the large difference for batch 6 was responsible for the strong statistical significance of the difference between the laboratory means. Before returning to the laboratory notebooks to investigate this result, she decided to exclude this batch from the dataset and re-analyse the remaining data. The data with summaries are shown in Table 2.2.7. Carry out the calculations and draw the appropriate conclusions. Compare your results to those obtained from the full dataset.

<table>
<thead>
<tr>
<th>Batch</th>
<th>Lab-1</th>
<th>Lab-2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.01</td>
<td>90.45</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>89.34</td>
<td>89.94</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>89.32</td>
<td>90.05</td>
<td>0.73</td>
</tr>
<tr>
<td>4</td>
<td>89.11</td>
<td>89.62</td>
<td>0.51</td>
</tr>
<tr>
<td>5</td>
<td>89.52</td>
<td>89.99</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Mean 89.460, 90.010, SD 0.340, 0.297

Table 2.2.7 Reduced dataset for the two laboratories
Summary – Significance Tests

The statistical significance testing procedure discussed in Sections 2.1 and 2.2 is quite straightforward. It may be summarised in the following steps:

• Frame the question of interest in terms of a numerical hypothesis

• Calculate a statistic which measures departure from this hypothesis

• Compare the calculated value of the statistic to the sampling distribution of values that might have been found if the hypothesis were true

• If the calculated value of the test statistic is exceptional in relation to this distribution, reject the hypothesis

• Exceptional values of a test statistic are those that have only a small chance of being observed. The small probability that defines ‘exceptional’ is called the significance level of the test.

These steps underlie all the significance tests presented in this text (and hundreds of other tests also). Anyone who understands this simple logical procedure, and who assesses the statistical assumptions required for the relevant test, should be able to apply effectively any statistical test without knowledge of the underlying mathematical details (especially if validated statistical software is available).

2.3 Confidence intervals

Statistical significance tests address the question “is the sample difference likely to reflect an underlying long-run systematic difference?”. A statistically significant test result means that the answer to this question is “yes”. A natural follow-up question is “how big is the difference?”.

Confidence intervals provide the statistical answer to this question – they put error bounds around the observed sample result in such a way as to provide not just a ‘point estimate’ (the observed sample value) but an interval estimate of the long-run value. Providing a measure of the size of the systematic quantity being studied is better, clearly, than simply saying that we can, or cannot, rule out a particular long-run value.

Confidence intervals will be introduced by re-considering the one-sample analytical validation data of Example 1 and they will then be applied to the paired comparative studies of Examples 3 and 4.
One-sample Studies

We return now to our validation study of Example 1, but instead of asking whether or not the analytical system in Laboratory B is biased (we already know this) we ask how big is the bias? The bias is the difference between the certified true value of 27.1 and the long-run average around which the analytical system in Laboratory B is varying. What is this?

Our sample mean of 26.04 is an estimate of that value. It is clear, though, that if another ten measurements were made, the result almost certainly would be other than 26.04, since our sample average is affected by chance measurement error. It makes sense, therefore, to attach error bounds to our sample average to reflect this uncertainty. The statistical approach to obtaining error bounds is to calculate a confidence interval.

A confidence interval simply re-arranges the elements used in carrying out the statistical test to produce error bounds around the sample mean. Thus, a 95% confidence interval for the long-run mean result that would be obtained if very many measurements were made on the baby food in Laboratory B is given by:

\[
\bar{x} \pm t_{c} \frac{s}{\sqrt{n}}
\]

\[
26.04 \pm 2.26 \frac{0.441}{\sqrt{10}}
\]

\[
26.04 \pm 0.32.
\]

We estimate that the long-run mean is somewhere between 25.72 and 26.36 units (percentage fat). Note that this interval does not include the certified value of 27.1. It will always be the case that where a confidence interval does not contain the null hypothesis value for the corresponding t-test, the test result will be statistically significant, i.e., the null hypothesis will be rejected. Conversely, if the test fails to reject the null hypothesis, then the hypothesised mean will be included in the confidence interval. Thus, the confidence interval is the set of possible long-run values which would not be rejected by a t-test.

The logic behind this interval is most easily seen by focusing on the Normal curve. The t-distribution coincides with the Normal curve when \(\sigma\) is known and it is simpler to focus on the Normal, which we have studied extensively in Chapter 1. We have seen that if a quantity \(X\) is Normally distributed, then 95% of the \(X\) values lie in the interval \(\mu \pm 1.96\sigma\), as shown in Figure 2.3.1(a). Similarly, as shown in Figure 2.3.1(b), 95% of means based on samples of size \(n=4\) sampled from this distribution will lie in the interval \(\bar{x} \pm 1.96\sigma/\sqrt{n}\). Thus, if we take a single sample of size \(n\) and calculate \(\bar{X}\), there is a probability of 0.95 that it will be within a distance of \(1.96\sigma/\sqrt{n}\) of \(\mu\).

\[\text{JF Mathematics (SH and TSM)}\]

28
We now consider the distance of $\mu$ from $\bar{X}$ instead of the distance from $\bar{X}$ to $\mu$: the same quantity, obviously, but the focus changes. Thus, rephrasing our conclusion above, we can state that the interval $\bar{X} \pm 1.96 \sigma / \sqrt{n}$ has a probability of 0.95 of covering $\mu$ (the unknown long-run mean).

Probability statements refer to random variables ("what might happen if..."). Once we replace the symbols in the expression $\bar{X} \pm 1.96 \sigma / \sqrt{n}$ by the observed numbers, we are no longer in the probability domain. We cannot say that there is a probability of 0.95 that the interval 25.72 to 26.36 units covers the unknown long-run mean for the analytical system in Laboratory B – it either does or it does not (it is no longer a question of what might happen!).

So what status does our interval 25.72 to 26.36 units have? From Figure 2.3.1(b) we deduce that if we calculate intervals in this way, then 95% of the intervals so calculated will cover the unknown long-run mean. This gives us confidence in the particular interval we have just calculated, and we say we are 95% confident that the interval covers $\mu$. Note that the 95% confidence is a property of the method used in calculating the interval, rather than a property of the pair of numbers we calculate.

By changing the multiplier 1.96, we can get different confidence levels; thus, using 2.58 gives 99% confidence, as this multiplier corresponds to the cut-off
values that bracket 99% of the area under a standard Normal curve. Note, though, that the interval will be correspondingly wider: there is a trade-off between having greater confidence at the expense of a wider (and perhaps less useful) confidence interval. The calculated interval gives us bounds for the long-run average around which the measurements vary: to obtain an interval estimate of the laboratory bias we simply subtract 27.1 (the certified value) from the two endpoint values.

A simulation exercise

The properties of this method of attaching error bounds to sample means, based as it is on the idea of repeated sampling, can be illustrated usefully by computer simulation of repeated sampling. The simulation described below is designed to help the reader to understand the implications of the confidence bounds and of the associated confidence level.

Fifty samples of size n=4 were generated randomly, as if from a process with known mean, μ, and standard deviation, σ. From each sample of four test results the calculated sample mean, \( \bar{x} \), and the corresponding interval were then calculated; the intervals are shown as horizontal lines in Figure 2.3.2.

\[
\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}
\]

were then calculated; the intervals are shown as horizontal lines in Figure 2.3.2.

The centre of each horizontal line is the observed mean, \( \bar{x} \), its endpoints are the bounds given by adding or subtracting \( 1.96 \sigma / \sqrt{n} \) from \( \bar{x} \). The vertical line represents the known mean, μ. The flat curve represents the frequency distribution of individual values; the narrow curve the sampling distribution of averages of 4 values. Only two of the fifty intervals do not cover the true mean, μ. Some intervals have the true mean almost in the centre of the interval (i.e., \( \bar{x} \) is very close to μ), but some just about cover the true mean, which since is relatively far from μ. Theory suggests that 5% of all such intervals would fail to cover the true mean; the results of the simulation exercise are consistent with this.

In practice, of course, only a single set of n measurements will be made and a single interval calculated, and we will not know whether or not the calculated interval covers the unknown true mean, μ. However, the theory guarantees that if intervals are calculated repeatedly in this way, 95% of them will cover the true mean.

---

6 Note the use of a small x to identify an observed mean, where a capital X was used to label the ‘random variable’; this difference in notation is a refinement we will not worry about in general in the text. Note, also, that where σ is known the t value will be the same as the Standard Normal value ±1.96. This was used here to simplify the drawing of the figure: it meant all the bars were the same length. If σ is not known (as is usually the case) then the sample estimates, s, will vary randomly from sample to sample and the widths of the intervals will vary accordingly.
mean. This gives us confidence in the particular interval we have calculated. Accordingly, the calculated interval is described as 'a 95% confidence interval' for \( \mu \). Note, again, that the 95% confidence level refers to the method we use, rather than the particular interval we have calculated.

![Figure 2.3.2: 50 simulated confidence intervals based on samples of size n=4](image)

**Exercises**

2.3.1. For the baby food data of Table 2.1.1 (page 3) calculate a 95% confidence interval for the long-run average around which the results in Laboratory A were varying. Since the t-test did not reject the null hypothesis that the long-run mean was 27.1, we would expect this interval to contain 27.1. Does it?

2.3.2. For the fill head data of Table 2.1.4 calculate 95% confidence intervals for the long-run average fill levels for the two heads. Are your intervals consistent with the results of Exercise 2.1.1 (pages 9, 10) where t-tests of the hypothesis that the target value of \( \mu_0 = 21 \) g was being achieved were carried out?
Paired Studies

Example 3: An animal fertility study – revisited

Our earlier analysis showed that the ipsi side of the uterus secretes more RBP than does the contra side. In order to measure the size of this difference we calculate a 95% confidence interval for the long-run mean difference (i.e., the value that would be obtained in a population of cows). This merely requires us to apply the same formula as above, but our calculations are now based on a sample of differences, rather than on individual raw measurements.

A 95% confidence interval for the long-run mean is:

\[ \bar{d} \pm t_c \frac{s}{\sqrt{n}} \]

where \( s \) is the standard deviation of the sample of 16 differences. Thus, we are 95% confident that the long-run mean ipsi-contra RBP difference lies between 354 and 1348 pg/µg protein. Note the width of this interval: this is due to the quite high estimate (933) for the standard deviation of the ipsi-contra differences (meaning that the difference can be much larger in some cows than in others) and the small number of animals (16) in the study.

Exercise

2.3.3 Return to Exercise 2.2.1 (Platelet aggregation study of cigarette smoking; page 21) and calculate a 95% confidence interval for the long-run mean shift in the measure of platelet aggregation after smoking one cigarette.

Example 4: A Laboratory Comparison Study – revisited

We return now to the laboratory comparison study of Example 4. Table 2.3.1 shows a Minitab analysis of the 40 laboratory differences.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab-B</td>
<td>40</td>
<td>41.595</td>
<td>3.037</td>
<td>0.480</td>
</tr>
<tr>
<td>Lab-A</td>
<td>40</td>
<td>40.040</td>
<td>2.956</td>
<td>0.467</td>
</tr>
<tr>
<td>Difference</td>
<td>40</td>
<td>1.555</td>
<td>1.607</td>
<td>0.254</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (1.041, 2.069)
T-Test of mean difference = 0 (vs not = 0): T-Value = 6.12  P-Value = 0.000

Table 2.3.1: Paired t-test and confidence interval for the laboratory data
The null hypothesis for the t-test is that the long-run mean difference is zero – the very large t-value (with correspondingly small p-value \( p<0.0005 \) clearly rejects this hypothesis, as we established earlier. The conclusion is that the laboratories are biased relative to each other. The 95% confidence interval, based on the same formula as for the RBP study above, estimates the bias to be between 1.041 and 2.069 percentage point units. I would report the bias as being between 1 and 2 percentage point units.

**Exercise**

2.3.4 Return to Exercises 2.2.2 and 2.2.3 (page 26) and calculate 95% confidence intervals for the relative bias between the two laboratories, based on the two sample sizes (\( n=6 \) for Exercise 2.2.2 and \( n=5 \) for exercise 2.2.3). Comment on their relative widths. What is the relation between the widths of these intervals and the sizes of the t-values calculated for the significance tests in the earlier exercises?

Note: In discussing confidence intervals we have focused entirely on means. Confidence intervals may also be calculated for standard deviations (see, Mullins (2003) [2]) and for other model parameters – thus, in Chapter 6 we will meet confidence intervals for the slope and intercept of a regression line.

### 2.4 Comparative Studies for Independent Groups

**Introduction**

Simple comparative studies involve comparing responses (performance measures) from two data samples. We have already considered a special case of such studies, where the data were paired, in Sections 2.2 and 2.3. Here, we consider the more commonly encountered case where sets of data are obtained under two conditions, but there is no special relationship between the individual values from the two sets. The analysis will involve using a t-test to test for a systematic difference in mean responses (described as a ‘two-sample’ or ‘independent groups’ t-test) and calculating the corresponding confidence interval to measure the size of the possible long-run difference.

Two-sample data may arise from an observational study; for example, we might want to compare the average junior certificate points achieved by two randomly selected samples of boys and girls. The study design may, alternatively, be experimental; for example, a group of students might be randomly split into two, one group might be asked to learn a body of material using a textbook as a learning tool, while the second group uses an interactive teaching package: the resulting scores on a common test will then be compared. In both cases, we will be concerned with differences that might arise in suitably defined populations rather than with the observed differences between the groups of subjects selected for study.
More complicated comparative studies arise where more than two groups are compared, or where, for example, the groups are divided in different ways. For example, if the learning study involved separate groups of boys and girls, each split into textbook/software-package groups, we would have a more complicated (but more interesting) dataset. Here, we will be interested in the simplest two-group comparisons.

The analysis of the data resulting from such studies typically involves formal comparison of the means of the two sets of results, using statistical significance tests and confidence intervals; these will be discussed first. These inferential methods pre-suppose a particular underlying statistical model. Graphical methods for validating that model are then discussed, as is a formal test to compare standard deviations. Study design questions, such as randomisation of the study sequence for experimental studies and use of control groups, will be discussed in Chapter 4. A key question that arises before the study is undertaken is “what sample size is required?”. This, also, will be addressed in Chapter 4.

**Example 5: A typical process development study**

A research student in the School of Pharmacy in Trinity College, Dublin carried out a study which was concerned with optimising the yield of the very small pellets (approximately 1mm in diameter) that are used to fill capsules when making pharmaceutical products. The study involved charging a spheroniser (essentially a centrifuge) with a dough-like material and running it at two speeds (labelled A and B, hereafter) to determine which gives the higher yield. The yield is the percentage of the material that ends up as usable pellets (the pellets are sieved to separate out and discard pellets that are either too large or too small). Eleven runs were made under each condition; the run order for the 22 runs was randomised. The results are shown in Table 2.4.1, followed by graphical and numerical summaries, in Figure 2.4.1 and Table 2.4.2

<table>
<thead>
<tr>
<th>Speed-B</th>
<th>Speed-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.2</td>
<td>73.5</td>
</tr>
<tr>
<td>81.3</td>
<td>77.0</td>
</tr>
<tr>
<td>77.0</td>
<td>74.8</td>
</tr>
<tr>
<td>79.9</td>
<td>72.7</td>
</tr>
<tr>
<td>76.4</td>
<td>75.4</td>
</tr>
<tr>
<td>76.2</td>
<td>77.1</td>
</tr>
<tr>
<td>77.6</td>
<td>76.1</td>
</tr>
<tr>
<td>80.5</td>
<td>74.4</td>
</tr>
<tr>
<td>81.5</td>
<td>78.1</td>
</tr>
<tr>
<td>77.3</td>
<td>76.5</td>
</tr>
<tr>
<td>78.2</td>
<td>75.0</td>
</tr>
</tbody>
</table>

Table 2.4.1: Percentage yields for the two speeds
It is clear from the dotplots and the summary statistics that in this study Speed-B gave a higher average percentage yield. Several questions arise:

- given that there is quite a lot of chance variability present in the data (thus, the yields for Speed-A vary from 72.7 to 78.1%, while those for Speed-B range from 76.2 to 81.5%, a range of approximately 5 units in each case), could the difference of 2.9% between the two averages be a consequence of the chance variation?

- if there is a yield difference, what is the best estimate of its long-run average value?

- does the variability of the results depend on which speed is used?

The first of these questions will be addressed by carrying out a t-test of the difference between the means. The second by obtaining a confidence interval for the long-run mean difference between results produced by the two process configurations, and the third by using a different significance test (the F-test) for comparing the standard deviations. Initially, the analyses will assume that the
answer to the third question is that the chance variability is the same for both speeds, but this assumption will be investigated later. All three analyses will be based on the assumption of data Normality; this assumption will also be investigated later.

**Statistical Model**

A simple statistical model for the data is that all the observations may be regarded as being independently generated from Normal distributions with standard deviation $\sigma$, common to both process configurations. The values generated using Speed-A have a long-run mean $\mu_1$, while those from Speed-B have a long-run mean $\mu_2$.

The underlying model is illustrated schematically by Figure 2.4.2. The two long-run means, $\mu_1$ and $\mu_2$, could be the same or different, but (in the simplest case) the standard deviations are assumed to be the same. Because the standard deviations are assumed the same, the shapes of the two Normal curves are identical. These curves describe the properties of the populations from which the two samples are drawn. The question as to whether there is a long-run difference between the yields may then be posed in terms of the difference (if any) between $\mu_1$ and $\mu_2$. The question is addressed directly, using a statistical significance test.

Figure 2.4.2: Two Normal curves: (possibly) different means but the same standard deviation, $\sigma$
**A statistical significance test for comparing means**

In Example 1, we carried out a t-test of the statistical significance of the difference between a single sample mean \( \bar{y} \) and a certified value \( \mu_0 \) by dividing by the estimated standard error of \( \bar{y} \):

\[
t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}
\]

and comparing the resulting t-value to the appropriate t-distribution. We will do essentially the same here: we will ask if the difference between the sample means is far from zero, by comparing the difference to the estimated standard error of the difference. This immediately points to the need to estimate the standard error of the difference between two sample means.

For a single sample mean, \( \bar{y} \), the standard error is \( \sqrt{\frac{s^2}{n}} \) - the variance is \( \frac{s^2}{n} \). The variance of the difference between two independent sample means is the sum of their variances, i.e., \( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{2\sigma^2}{n} \) for common standard deviation \( \sigma \) and sample size \( n \). It follows from this result that the standard error of the difference between two independent sample means is \( \sqrt{\frac{2\sigma^2}{n}} \). We have two estimates of \( \sigma \), one from each speed, and these can be combined into a single estimate of \( \sigma^2 \); this is done by averaging the two sample variances.

\[
s^2 = \frac{s_1^2 + s_2^2}{2} = \frac{(1.635)^2 + (2.054)^2}{2} = 3.446
\]

This estimate has 20 degrees of freedom, since each of the sample standard deviations has 11–1=10 degrees of freedom.

The means are compared formally by specifying the null hypothesis of no difference and the alternative hypothesis that denies this:

\[
H_0: \mu_1 - \mu_2 = 0 \\
H_1: \mu_1 - \mu_2 \neq 0
\]

\[\text{Note that the variances add, even though the means are subtracted. If this seems odd, ask yourself if you would expect the combination of two uncertain quantities to be more or less uncertain than the two quantities being combined. Refer back to Chapter 1, Section 1.4 where we discussed combining random quantities.}\]
and then carrying out a t-test. An obvious estimator of the difference between the long-run means, $\mu_2 - \mu_1$, is the difference between the sample means, $\bar{y}_2 - \bar{y}_1$. If this sample difference is far from zero, it would be evidence that $H_0$ is false. In order to assess the magnitude of the observed difference, given the presence of chance variation, the distance from $\bar{y}_2 - \bar{y}_1$ to zero is re-scaled by dividing by its estimated standard error, and the result, $t$, is compared to a Student's t-distribution with 20 degrees of freedom:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{\sqrt{\frac{2s^2}{n}}}$$

As for one-sample t-tests, a definition is required for what constitute exceptionally large or small values of $t$, the test statistic. These are specified by choosing the significance level for the test: the significance level is a small probability which is used to determine cut-off points (the critical values) on the tails of the distribution curve beyond which observed t-values are considered exceptional. Thus, if we choose a significance level of $\alpha=0.05$ and if $H_0$ is true, the t-value from the study will be expected to lie between $-2.09$ and $2.09$, since 95% of the area under the curve lies between these values, see Figure 2.4.3. If, in fact, there is no long-run yield difference, a t-value beyond these critical values will only occur, by chance, with probability 0.05.

The test statistic can fall into the tails for one of two reasons: just by chance when there is no long-run difference, or because the two long-run means are not the same. In using statistical significance tests, the latter is always assumed to be the reason for an exceptional test-statistic. Accordingly, an unusually large or small test-statistic will result in the null hypothesis being rejected.
The calculated t-value is:

\[ t = \frac{78.37 - 75.51}{\sqrt{\frac{2(3.446)}{11}}} = \frac{2.86}{0.7915} = 3.61 \]

and as this value falls outside the critical values (-2.09, +2.09) \( H_0 \) is rejected. The two sample means are said to be statistically significantly different. Speed-B is considered to give a higher yield, on average.

If you compare the steps involved in carrying out the two-sample t-test described above with those previously described for a single-sample (or paired) t-test you will see that the procedures are virtually identical. The differences are purely technical: they simply allow for the fact that in the two-sample case there are two sample means each subject to chance variation, whereas there is only one for the one-sample or paired t-test.

**Exercises**

2.4.1. Samuels [5] reports a study of a pain-killing drug for treating uterine cramping pain after childbirth (She describes the data as fictitious, but realistic, which presumably means the numbers are based on experience of similar studies). Fifty women were randomly assigned into two groups of 25, one of which received the drug and the other a placebo. A pain-relief score, based on hourly interviews throughout the day, which varied from 0 (no relief) to 56 (complete relief for 8 hours), was assigned to each study participant. The data summaries are shown in Table 2.4.3.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>31.96</td>
<td>12.05</td>
</tr>
<tr>
<td>Placebo</td>
<td>25.32</td>
<td>13.78</td>
</tr>
</tbody>
</table>

Table 2.4.3: Summary results for the pain relief study

Does the study suggest that the drug works? Carry out a test to determine if there is a statistically significant difference between the sample means.

2.4.2 Samuels [5] presents data from an experiment designed to see if wounding a tomato plant would induce changes that improve its defence against insect attack. Larvae of the tobacco hornworm (Manduca Sexta) were grown on 17 wounded and 17 control plants\(^8\). Summary statistics for the weights of the larvae (mg) after 7 days of growth are shown in Table 2.4.4. Analyse the study results.

---

\(^8\) Note: the actual numbers were 16 and 18, but to avoid the complication of taking a weighted average of the standard deviations, I have made the sample sizes the same – it does not affect the results in any substantive way.
Estimating the long-run difference in yields

For the pellet process development study, the difference between the two sample means, \( \bar{y}_2 - \bar{y}_1 \), is a point estimate of the difference, \( \mu_2 - \mu_1 \), between the long-run means for the two process speeds. This observed difference is obviously subject to chance variation, so we might ask what it tells us about the long-run difference. Following the previously discussed one-sample case, where long-run means were estimated from sample means (in Example 1 we estimated the long-run average value around which the laboratory was varying in measuring the percentage fat in baby food and in Example 4 we estimated the long-run mean difference between two laboratories), a natural approach to answering this question is to calculate a confidence interval for \( \mu_2 - \mu_1 \).

In the one-sample case, a 95% confidence interval for a single mean, \( \mu \), was given by:

\[
\bar{y} \pm t_c \frac{s}{\sqrt{n}}
\]

or equivalently:

\[
\bar{y} \pm t_c \sqrt{\frac{s^2}{n}}
\]

where \( \sqrt{\frac{s^2}{n}} \) is the estimated standard error of \( \bar{y} \). By extension, a confidence interval for the difference between two long-run means \( \mu_2 - \mu_1 \), is:

\[
(\bar{y}_2 - \bar{y}_1) \pm t_c \sqrt{\frac{s^2}{n_2} + \frac{s^2}{n_1}}
\]

which in this case reduces to:

\[
(\bar{y}_2 - \bar{y}_1) \pm t_c \sqrt{\frac{2s^2}{n}}
\]
Here, $\sqrt{\frac{2s^2}{n}}$ is the estimated standard error of the difference between the two sample means $\bar{y}_2 - \bar{y}_1$, each based on $n$ values, and $s$ is the combined estimate of the common standard deviation $\sigma$, which was calculated above.

The calculated confidence interval is:

$$(78.37 - 75.51) \pm 2.09 \sqrt{\frac{2(3.446)}{11}}$$

$$2.86 \pm 1.65$$

Although the study showed an average difference of 2.86 units, the confidence interval estimates that the difference in long-run yields is somewhere between 1.2 and 4.5 percentage points, with 95% confidence.

A confidence interval that covers zero

Suppose that the confidence interval had turned out as 2.86±5.00, i.e., ranging from −2.14 to +7.86. How would this result be interpreted?

Such a result could mean that Speed-B gives a long-run yield that is greater by as much as 7.86 units; it could, alternatively, mean that Speed-A gives results higher by 2.14 units, on average. In other words, the data cannot tell unambiguously which speed gives higher results, on average. In such a case the sample means are said to be not statistically significantly different from each other, i.e. the observed difference could have resulted from chance variation.

The relationship between confidence intervals and significance tests is the same for two-sample studies as it is for single means (discussed earlier). If the confidence interval covers zero the null hypothesis that $(\mu_2 - \mu_1)=0$ will not be rejected by the significance test. If the interval does not contain zero then the null hypothesis will be rejected and the two sample means will be declared to be ‘statistically significantly different’.

Exercises

2.4.3. Refer to Exercises 2.4.1 and 2.4.2 and calculate 95% confidence intervals for the long-run mean difference between the two treatments in each case. Interpret your intervals.
A typical computer analysis

The analysis shown in Table 2.4.5 was specified under the assumption of equal long-run standard deviations for both methods.

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

Two-sample T for Speed-B vs Speed-A

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed-B</td>
<td>11</td>
<td>78.37</td>
<td>2.05</td>
<td>0.62</td>
</tr>
<tr>
<td>Speed-A</td>
<td>11</td>
<td>75.51</td>
<td>1.63</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Difference = mu (Speed-B) - mu (Speed-A)
Estimate for difference:  2.86364
95% CI for difference:  (1.21237, 4.51490)
T-Test of difference = 0 (vs not =): T-Value = 3.62  P-Value = 0.002  DF = 20
Both use Pooled StDev = 1.8565

Table 2.4.5: A Minitab analysis of the pellet process development data

The numerical results are the same as those presented above for the test statistic and conference interval (apart from some small rounding differences), but, in addition, the output gives us a p-value associated with the test statistic.

Interpretation of p-value

It is common that statistical packages show p-values as indicators of the statistical significance, or otherwise, of the results. This was discussed earlier in the context of a paired t-test. In the current context, the p-value is the answer to the question "In a repeat of the study, what would be the probability of obtaining a more extreme t-value than that observed, if there really were no difference between the long-run means?" As Figure 2.4.4 shows, the probability of obtaining a t-value either less than −3.62 or greater than +3.62 is 0.002.
Both tails are taken into account since a t-value of −3.62 would be considered as indicating a statistically significant difference, also, i.e., we have a two-sided alternative hypothesis. The test is often called 'two-tailed' for this reason - one-sided hypotheses with corresponding one-tailed critical or rejection regions were encountered for acceptance sampling decisions, earlier.

If a significance level of 0.05 is chosen for the significance test, then a p-value less than 0.05 is taken as indicating that the observed difference between the means is statistically significant. If the p-value is greater than 0.05 then the result is not statistically significant, i.e., the observed difference is considered consistent with only chance variation away from zero. The advantage of quoting p-values is that it is immediately obvious how extreme the t-value is, i.e., how unlikely such a value is under the hypothesis of no long-run difference. It is worthwhile pointing out again that 'significant' means 'likely to be due to other than chance variation' – it may or may not be the case that the observed difference is 'important' – this would depend on considerations that have nothing to do with statistics and everything to do with the study domain.

Validating the model assumptions

As discussed earlier, and illustrated by Figure 2.4.2, the model underlying our statistical analysis assumes equal long-run standard deviations and data Normality within the two populations from which the data are considered to be randomly selected; it also assumes independence. The validity of the independence assumption follows from the randomisation of the order in which the runs were carried out.
It is obvious from Figure 2.4.1 that the variability is about the same for the two samples (suggesting that the constant long-run standard deviation assumption holds). Sometimes, it is helpful to subtract the respective group means from the sample data before plotting, as shown in Figure 2.4.5; subtracting the group means results in both samples varying around zero and this can make it (a little) easier to compare the scatter, as the data will now be lined up opposite each other. When the means are subtracted, the resulting deviations are called ‘residuals’, as they are what remain after the systematic components in the responses (the group means here, but we will see other possibilities, later) are removed.

To assess whether the pellet process development data are Normal, separate Normal plots might be drawn for the eleven results from each process speed. However, the sample size would be small in both cases. If the long-run standard deviations can be assumed equal for the two groups (as we have seen, this appears a reasonable assumption here), the two sets of residuals may be combined into one, since they have the same mean of zero and a common, but unknown, standard deviation. We can then draw a Normal plot to determine if the residual variation is consistent with a single Normal distribution.

Figure 2.4.6 shows a Normal plot of these residuals. The scatter of the plotted points is close to a straight line, as would be expected if they come from a Normal distribution. The p-value indicates that 22 observations selected randomly from a truly Normal distribution would have a probability of $p=0.324$ of showing stronger departure from a straight-line relationship than that observed in this study. Accordingly, there is no reason to reject the hypothesis that the data come from a Normal distribution.
Note that the assumption of Normality relates to within-group variation. If the long-run means were different and all the raw data were combined into a single sample, then what we would have would be a mixture of two distributions, which could be very different from a single Normal distribution, even if the two separate distributions were individually Normal. Accordingly, residuals rather than the raw data should be combined for investigating Normality.

Formal Comparison of standard deviations

In carrying out the t-test and in calculating the confidence interval it was assumed that the long-run standard deviations were the same for the two spheroniser speeds. Clearly, the sample standard deviations are close in this case, as can be seen both from Figure 2.4.5 and the summary statistics of Table 2.4.2, but a formal statistical test of the equality of the long-run values may be desired in other cases. Such a test (the F-test\(^9\)) will be described below. However, the value of this widely used test is open to question on two grounds. Moore [6] argues against it on the grounds that it is highly sensitive to the assumption of data Normality. It is also known not to be powerful (see e.g., Mullins [2], pp. 166-168): this means that large sample sizes are required in order to be reasonably confident of detecting even moderately large differences between the standard deviations of the populations being compared. For the modest sample sizes often encountered in research studies the test will not be powerful. This means

---

\(^9\) The test is named F for Fisher, in honour of Sir Ronald Fisher who made major contributions to the theory and practice of statistics (and genetics) in the first half of the 20th century. The test is commonly described as a test for equal variances, but since the square of the standard deviation is the variance, the two descriptions are equivalent. Standard deviations are measured in the same units are the original measurements; for this reason they appear to me to be a more natural way to describe the variability of the data.
that there could well be a substantial difference between the standard deviations but the test will fail to detect this difference. Accordingly, careful study of routinely produced quality control data in an industrial context, or of previously published data in academic studies, together with professional experience in the relevant discipline may well have to replace, or at least supplement, statistical significance testing in making such comparisons.

To carry out the test, the null hypothesis of equal standard deviations is specified; the alternative hypothesis denies their equality:

\[
\begin{align*}
H_0 : \sigma_1 &= \sigma_2 \\
H_1 : \sigma_1 &\neq \sigma_2
\end{align*}
\]

The test statistic is the ratio of the standard deviations squared, or equivalently, the ratio of the sample variances:

\[
F = \left(\frac{s_2}{s_1}\right)^2
\]

If the two long-run standard deviations are equal, i.e., \( \sigma_1 = \sigma_2 \), this test statistic follows an F-distribution with \( n_2 - 1 \) numerator and \( n_1 - 1 \) denominator degrees of freedom, associated with \( s_2 \) and \( s_1 \), respectively. Here, the F-distribution has 10 degrees of freedom for both numerator and denominator. Figure 2.4.7 shows that if a significance level of \( \alpha = 0.05 \) is selected the critical values are 0.27 and 3.72; this means that if the long-run standard deviations are indeed equal, there are tail probabilities of 0.025 of observing an F-ratio either greater than 3.72 or less than 0.27 (Note ST-3 gives right-hand tail critical values for the F distribution). Either very large or very small F-ratios result in the null hypothesis being rejected; large F ratios (F>3.72) suggest that \( \sigma_2 > \sigma_1 \), while small F ratios (F<0.27) suggest that \( \sigma_1 < \sigma_2 \).

Figure 2.4.7: F-distribution with degrees of freedom (10,10)
For the process development study the F-ratio is:

\[ F = \left( \frac{2.054}{1.635} \right)^2 = 1.58 \]

Since the calculated F statistic lies in the body of the distribution the test provides no basis for rejecting the assumption of equal long-run standard deviations, i.e., the variability appears to be about the same for the two process speeds.

Note that the F-distribution is not symmetrical and, therefore, for a two-sided test critical values are required which are different in magnitude for the two tails. Statistical tables usually give the right-tail critical values only. To avoid having to calculate the left-hand critical value, it is conventional to use the larger sample standard deviation as the numerator of the F-ratio. If this is done, then the test asks if the sample ratio is too large to be consistent with random fluctuation away from F=1.0, which is what the result would be if the long-run standard deviations were known and equal. In doing this, it is important that a table which gives the critical value for 0.025 in the right-tail should be consulted, even though a significance level of \( \alpha = 0.05 \) is chosen for the test, since a left-tail critical value is, in principle, also applied; the left-tail critical value is not explicitly specified because we have arranged for the sample ratio to be bigger than one and, consequently, what we wish to determine is if it is unusually large.

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th>StDev</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>2.054</td>
<td>4.219</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>1.635</td>
<td>2.673</td>
</tr>
</tbody>
</table>

Ratio of standard deviations = 1.256
Ratio of variances = 1.578

95% Confidence Intervals

<table>
<thead>
<tr>
<th>Distribution of Data</th>
<th>CI for StDev Ratio</th>
<th>CI for Variance Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>(0.652, 2.422)</td>
<td>(0.425, 5.866)</td>
</tr>
</tbody>
</table>

Tests

<table>
<thead>
<tr>
<th>Method</th>
<th>DF1</th>
<th>DF2</th>
<th>Test Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>F Test (normal)</td>
<td>10</td>
<td>10</td>
<td>1.58</td>
<td>0.483</td>
</tr>
</tbody>
</table>

Table 2.4.6: A Minitab analysis of the pellet data standard deviations

---

10 The left-tail value \( F_{0.025,a,b} \) (the value that leaves 0.025 in the left-hand tail where the degrees of freedom are ‘a’ for the numerator and ‘b’ for the denominator) is given by the reciprocal of \( F_{0.975,b,a} \) (the value that leaves 0.025 in the right-tail; note the reversal of degrees of freedom). Thus, for example, Table ST-3 gives \( F_{0.975,3,6} = 6.6 \), so, \( F_{0.025,6,3} = 1/6.6 = 0.15 \). The p-value shown in the Table is twice the area to the right of F=1.578, since the area to the left of 1/1.578 is equal to that to the right of 1.578 for an \( F_{10,10} \) distribution.
The Minitab analysis gives us the F-ratio of 1.58 (as calculated above) and also gives us a p-value of 0.48 which is consistent with our not rejecting the hypothesis of equal long-run standard deviations. The 95% confidence interval for the ratio of the standard deviations is 0.65 to 2.42; the fact that this covers 1 is consistent with the test not rejecting the equality of the long-run standard deviations.

**Exercise**

2.4.4. Carry out F-tests to check the equality of the long-run standard deviations for the data in Exercises 2.4.1 and 2.4.2

2.4.5. Pendl et al. [1] describe a fast, easy and reliable gas chromatographic method for determining total fat in foods and animal feeds. The data below (% Fat) represent the results of replicate measurements on a margarine, which were made on ten different days by two laboratories A, B. Verify that an F-test, with a significance level of 0.05, will reject the hypothesis of equal precision (equal standard deviations for repeated measurements) in the two laboratories.

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>79.63</td>
<td>79.64</td>
<td>78.86</td>
<td>78.63</td>
<td>78.92</td>
<td>79.19</td>
<td>79.66</td>
<td>79.37</td>
<td>79.42</td>
<td>79.60</td>
</tr>
<tr>
<td></td>
<td>S_A</td>
<td>0.374</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>74.96</td>
<td>74.81</td>
<td>76.91</td>
<td>78.41</td>
<td>77.95</td>
<td>79.17</td>
<td>79.82</td>
<td>79.31</td>
<td>77.65</td>
<td>78.36</td>
</tr>
<tr>
<td></td>
<td>S_B</td>
<td>1.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4.6: Replicate %Fat measurements on a margarine from two laboratories

**Unequal Population Standard Deviations**

In some cases the long-run or population standard deviations will not be equal, $\sigma_1 \neq \sigma_2$. This is more likely to arise in observational studies in which groups with potentially markedly different characteristics are to be compared. In randomised experimental studies the long-run standard deviations are more likely to be the same, since they will be determined by the very many factors not controlled in the study and are, thus, likely to be similar for two groups formed by random allocation. Of course, there is always the possibility that the experimental treatment (e.g., changing the speed of a machine in an engineering study) will affect not only the mean response, but also the variability of responses.

Where the group sizes are reasonably large (say greater than 30 in each case), the inequality of the standard deviations is unlikely to cause a problem. The standard error of the difference between the sample means can be estimated using the formula:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
and the resulting test statistic can be compared to critical values determined by the standard Normal distribution, or simply to ±2 for a significance level of $\alpha = 0.05$.

For small sample sizes, a modified version of the t-test is available – the test statistic is calculated using the standard error formula given above, but the degrees of freedom are obtained using a rather complicated formula (this is given in Mullins [2] or Moore and McCabe [7]); statistical packages will automatically take care of the messy calculations. For small sample sizes it is virtually impossible to compare, usefully, two sample standard deviations – the power of the test will be very low (see preceding discussion). Consequently, as indicated above, it will very often be the case that professional judgement (of the study domain, not statistical) will be required to decide whether or not the assumption of equal population standard deviations is, or is not, appropriate. See Chapter 7 of Moore and McCabe [7] for further discussion.

2.5 Review

The main purpose of this chapter was to introduce the ideas of significance testing, confidence intervals and comparative studies. These ideas will now be reviewed in a final example.

Example 6: Use of a neuroscience test for studying ADHD$^{11}$

The Sustained Attention to Response Task (SART) is a neuroscience test developed for the study of people with traumatic brain injuries (Robertson et al, [8]). It involves subjects viewing a computer screen on which the digits 1-9 are presented in a randomised order at regular intervals (every 1.15 sec) over a fixed period of time. One of the digits is designated as a target digit. The subjects are required to press the mouse when any one of the other eight digits appears, but not to do so when the designated target digit appears. Performance on the test may be evaluated in terms of the numbers of errors in pressing (or not pressing) and also in terms of the average response times.

One part of an investigation of the possible use of SART in studying Attention Deficit Hyperactivity Disorder (ADHD) involved a comparison of the response times of a group of 65 children suffering from ADHD with those of a control group of 65 children. The mean response time (ms) for each child was measured for the first part (P1) and the second part (P2) of the SART test and the difference in

---

$^{11}$ I am grateful to Dr. Jane Sanders, a former student in the TCD Postgraduate Diploma in Statistics, for permission to use her data.
mean response time was calculated for each individual child. Summary statistics for the two groups of children are shown in Table 2.5.1.

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>Individual Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD children</td>
<td>Mean</td>
<td>485</td>
<td>507</td>
</tr>
<tr>
<td></td>
<td>St. Dev.</td>
<td>112</td>
<td>120</td>
</tr>
<tr>
<td>Control children</td>
<td>Mean</td>
<td>502</td>
<td>493</td>
</tr>
<tr>
<td></td>
<td>St. Dev.</td>
<td>113</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 2.5.1: Summary statistics for the response times (ms)

It is of interest to know whether or not differences exist between the two groups of children. We might guess that for the first part of the test any difference might be small or even zero, but that as the test progresses, requiring more sustained attention, the difference, if any exists, might grow. This would arise if the performance within one or both groups changed over the test duration. Our analysis will attempt to throw light on these questions.

Data Collection

The structure of the SART dataset is a little more complex than the comparative studies discussed earlier in the chapter. Our first example of a comparative study involved one group being measured twice, with the objective of detecting a shift in the long-run mean using a paired t-test (for example, RBP measurements on two sides of the uterus for a sample of cows, or batches of tablets being measured in two laboratories). Subsequently, we had two independent groups of results (e.g., process yields for 11 runs of a spheroniser under two different speeds); here the objective was to make inferences about possible differences between the long-run mean yields, using a two-sample t-test. The SART dataset combines both types of comparison; it involves two independent groups of 65 subjects each measured twice. In the psychology literature this would be referred to as a ‘repeated measures design’; obviously, if there is only one group, the subjects of which are each measured twice, the repeated measures design reduces to the paired design discussed in Sections 2.2 and 2.3.

In general, each subject in a repeated measures design is measured several times (not just twice) and there may be more than two groups of subjects involved in the study. The analysis of the more complex structures is normally carried out using the statistical method called ‘Analysis of Variance (ANOVA)’ which will be introduced in Chapter 5, though only at the most elementary level. Here, we have a special case, involving only two groups each measured twice, which allows us to use the simpler and more transparent t-tests as our methods of analysis. We will carry out two-sample t-tests to make comparisons between the ADHD and Control groups (which are independent of each other) and paired
t-tests to make comparisons within the two groups, between the response times for the two parts of the SART study.

**Graphical Analysis**

Figure 2.5.1 shows dotplots of the raw data on which Table 2.5.1 is based.

![Dotplots for the response times (ms) of the two groups on the two parts of SART](image)

Two aspects of the figure are noteworthy: the within-group variability for all four datasets is about the same and is considerably greater than any between-group mean differences. However, because the sample sizes are relatively large, the standard errors of the sample means will be much reduced, so it makes statistical sense to carry out significance tests for possibly small between-dataset mean differences. It is clear, though, from Figure 2.5.1 that any differences between the ADHD and control groups are small. Possible differences within-subject, between the first and second parts of the test, cannot be discerned from Figure 2.5.1, which displays between-child variation.

The two sample t-test discussed in Section 2.4 assumes constant standard deviation (this is supported by Figure 2.5.1) and Normality. Figures 2.5.2-2.5.5 show Normal plots for the response times for the four datasets.

In all cases we get approximately straight lines and Anderson-Darling test statistics which have associated p-values of 0.5 or higher, both of which support the assumption of underlying Normal distributions. Since we are dealing with 130 different children in a context where there is not likely to be any dependency between their test results, the independence assumption also appears safe.
Accordingly, we can confidently carry out two-sample t-tests to compare the two groups of children, separately for each part of the SART.

Figure 2.5.2: Normal plot of response times (ms) for the ADHD children for P1

Figure 2.5.3: Normal plot of response times (ms) for the ADHD children for P2
Figure 2.5.4: Normal plot of response times (ms) for the Control children for P1

Figure 2.5.5: Normal plot of response times (ms) for the Control children for P2
Figures 2.5.6 and 2.5.7 show Normal plots of the within-child differences for the two parts of the SART.

Although the scatterplots approximate reasonably straight lines, in both cases the Anderson-Darling test statistics have associated p-values which are on the borderline of being statistically significant, thus calling the Normality assumption
into question. However, the t-test is considered a ‘robust’ test, i.e., not very sensitive to the Normality assumption – the test is based on sample means and, as we saw in Chapter 1, when we average a reasonable number of values the mean tends to become Normal, even though the underlying distribution may be quite non-Normal. In the current case our sample sizes are moderately large – 65 in each case – and the plots are fairly straight, so the underlying distributions are not highly non-Normal: thus, it appears reasonable to carry out paired t-tests on these data.

Formal Tests and Confidence Intervals

Between-group differences

First, we will carry out statistical tests to ask if the sample mean differences between the two groups of children suggest long-run mean differences; we will do this separately for the two parts of the SART. As discussed in the introduction, we might expect that while there might or might not be a difference on the first part, as the test progresses (requiring more sustained attention) a difference might develop or increase with time.

For the first part of the SART (P1) we ask if a difference exists, on average, between the two populations of children. The hypotheses are specified as:

\[ H_0: \mu_1 - \mu_2 = 0 \]
\[ H_1: \mu_1 - \mu_2 \neq 0 \]

where \( \mu_1 \) refers to the control children and \( \mu_2 \) refers to the ADHD children. We choose a significance level of \( \alpha = 0.05 \).

The combined sample estimate of the population variance is:

\[ s^2 = \frac{22^2 + 112^2}{2(65)} = \frac{12656.5}{65} \]

This has 128 degrees of freedom, since the standard deviation for each group has \( n-1=64 \) degrees of freedom. For a two-sided test with degrees of freedom 2n−2=128, the critical values are \( t_c = \pm 1.98 \); with such high degrees of freedom many statisticians would simply use \( \pm 2 \) as the cut-off values, corresponding to the standard Normal cutoff values of \( \pm 1.96 \), which rounds to 2. The test statistic is:

\[ t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{2s^2}{n}}} = \frac{(502 - 485) - 0}{\sqrt{\frac{2(12656.5)}{65}}} = 0.86 \]
This value falls in the body of the reference distribution and so we do not reject the null hypothesis: we have not demonstrated a statistically significant difference between the mean response times of the two groups on part 1 of the SART.

Table 2.5.2 shows the corresponding Minitab analysis to compare the sample means of the two groups on the second part of the SART.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD-P2</td>
<td>65</td>
<td>506</td>
<td>120</td>
<td>15</td>
</tr>
<tr>
<td>Control-P2</td>
<td>65</td>
<td>493</td>
<td>106</td>
<td>13</td>
</tr>
</tbody>
</table>

Difference = mu (ADHD-P2) - mu (Control-P2)
Estimate for difference: 13.8
95% CI for difference: (-25.6, 53.1)
T-Test of difference = 0 (vs not =): T-Value = 0.69
P-Value = 0.490  DF = 128
Both use Pooled StDev = 113.2759

Table 2.5.2: Minitab analysis of the data for the second part of the test (P2)

The test statistic is 0.69 which has a corresponding p-value of 0.49: the sample difference of 13.8 msec between the mean response times for the two groups is not statistically significant. Note that the confidence interval for μ1 - μ2 covers zero, as is required when the test is not statistically significant.

The large child-to-child variability in response times (see standard deviations in Table 2.5.1 and the dotplots of Figure 2.5.1) is such that the observed mean differences between the two groups (17 and 14 msec for parts 1 and 2, respectively) are consistent with chance variation. The lack of statistical significance with such relatively large samples strongly supports the null hypothesis of no systematic difference in mean response times between the two populations of children.

Within-group differences

We now turn to within-group comparisons of the average response times for the two parts of the SART. For the ADHD children, first, we ask if the mean difference over time (P2-P1) is statistically significantly different from zero:

\[ H_0: \mu = 0 \]
\[ H_1: \mu \neq 0 \]

We choose a significance level of \( \alpha=0.05 \). For a two-sided test with degrees of freedom \( n-1=64 \), we use critical values of \( t_c=\pm2 \). The test statistic is:
This exceeds the critical value of 2 so we reject the null hypothesis and conclude that our observed mean difference of 22 reflects an underlying non-zero population difference.

A 95% confidence interval for the population mean difference is given by:

\[
\bar{d} \pm t_c \frac{s_d}{\sqrt{n}}
\]

which is 3.7 to 39.9. We are 95% confident that a population of such ADHD children would show a mean difference of between 4 and 40 msec in response times on the two parts of the SART. This represents a slowing of their responses to the stimuli over the course of the test. As always, the confidence level refers to the method used rather than the particular numbers quoted above.

The corresponding analyses for the control group were carried out in Minitab; Table 2.5.3 displays the relevant output.

**Paired T-Test and CI: Control-P2, Control-P1**

Paired T for Control-P2 - Control-P1

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control-P2</td>
<td>65</td>
<td>492.7</td>
<td>105.7</td>
<td>13.1</td>
</tr>
<tr>
<td>Control-P1</td>
<td>65</td>
<td>501.7</td>
<td>113.2</td>
<td>14.0</td>
</tr>
<tr>
<td>Difference</td>
<td>65</td>
<td>-8.99</td>
<td>61.41</td>
<td>7.62</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (-24.20, 6.23)  
T-Test of mean difference = 0 (vs not = 0): T-Value = -1.18  
P-Value = 0.242

Table 2.5.3: Minitab analysis of the P2-P1 differences for the control group
The paired t-test comparing the mean response times for the two parts of the test does not show a statistically significant difference. Thus, for the control group of children there is no evidence of a change in mean response time over the course of the SART. The confidence interval covers zero (−24 to +6) as expected, given that the test statistic is not statistically significant.

The two-sample t-tests suggested no mean differences between the two groups of children for the two parts of the test, when they were compared separately. The paired t-test on the control children detected no difference in response times over the two parts of the study, while that on the ADHD children pointed to an increased response time in the second part of the SART. Why the apparent contradiction?

Essentially, the paired tests are more sensitive (powerful) – they have a greater capacity to detect small differences. This is so because the measure of chance variation (s_d) that appears in the denominator of the paired test statistic is based on the variation between children of the within-child difference (which for ADHD is 73), while that in the two-sample t-tests measures between-child variation of individual response times. Figure 2.5.1 shows the between-child variation in response times to be large (thus, the standard deviation for ADHD children on P1 of the test was 112). Consequently, although the between-group difference for P1 (502−485=17) is not markedly different from that between the two parts of the SART for the ADHD children (507−485=22), the two-sample t-test between groups is not able to separate the sample means statistically, whereas the paired t-test on the mean difference for the ADHD children is statistically significant (p=0.02).

Concluding Remarks

As mentioned in the introduction, the usual starting point for the analysis of data obtained from a repeated measures study design would be ANOVA. This method provides a single statistical test (called a test for ‘interaction’) which shows that the difference in mean response times between P1 and P2 is different for the two groups of children. We carried out four tests in arriving at this conclusion. The undesirability of multiplying the number of tests carried out in any analysis will be discussed in Chapter 5. Here, though, our rather crude approach provided us with an opportunity to review the two different variants of the t-test which formed the substance of this chapter.
References


Text © Eamonn Mullins, 2014; data, see references

Note that some portions of the text are based on Mullins (2003), the copyright owner is the Royal Society of Chemistry.
Fill head Data

We wish to test the hypothesis that each of the fill heads delivers $\mu_o = 21g$ into the containers, on average. We carry out a t-test for each head. We specify the null and alternative hypotheses as:

$$H_0: \mu = 21.00$$
$$H_1: \mu \neq 21.00$$

where $\mu$ is the long-run average quantity delivered into the containers. We calculate the test statistics using:

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

For a significance level of $\alpha=0.05$ we have critical values of $\pm 2.26$, since the standard deviations each have $(n-1)=9$ degrees of freedom and this means the appropriate reference distribution is a Student's t-distribution with 9 degrees of freedom; this is shown in Figure 2.1.1.1

![Figure 2.1.1.1: Student's t-distribution with 9 degrees of freedom](image)

For head 1 we get:

$$t = \frac{21.06 - 21.00}{0.118/\sqrt{10}} = 1.61$$
For head 2 we get:

\[ t = \frac{21.33 - 21.00}{0.360/\sqrt{10}} = 2.90 \]

The test statistic for head 1 falls between the critical values and so we do not reject the null hypothesis; the sample data are consistent with fill head 1 being on-target. The test statistic for head 2 is outside the right-hand critical value and leads us to reject the null hypothesis; we do not accept that head 2 is on-target – it appears to be overfilling, on average.

2.1.2

**Glucose Recovery Data**

H0: \( \mu = 100\% \)

H1: \( \mu \neq 100\% \).

The test statistic is:

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{98.443 - 100}{2.4511/\sqrt{50}} = -4.4 \]

The critical values for a significance level of 0.05 and 49 degrees of freedom are ± 2.01. Accordingly, we reject H0 and conclude that the system gives a recovery rate lower than 100%.

2.2.1 (and 2.3.3)

**Platelet aggregation data**

The experimental design for this study involved self-pairing of the subjects; the natural approach to analysing the resulting data is, consequently, to carry out a paired t-test and calculate the corresponding confidence interval.

If we wish to use a paired t-test and the corresponding confidence interval in analysing the data then we are reliant on the underlying model which asserts that the differences behave as if randomly generated (this implies that they are independent of each other) from a single Normal distribution. Plotting the differences against the means is often useful as it can indicate a systematic relationship, which would violate the model assumptions. For example, it could be that the effect of smoking might be to multiply or divide the response studied.
by some factor, instead of simply shifting it a constant amount (apart from chance variation). If this were the case we might expect to see a trend upwards or downwards when the differences are plotted against the means.

In Figure 2.2.1.1, the largest difference (27 corresponding to a mean of 66.5) draws our eyes upwards to the right, suggesting the possibility of an upwards trend. When this is removed (Figure 2.2.1.2), however, the remaining points show no upwards trend. With such a small dataset we cannot be sure that no such trend exists, but it seems reasonable to work with the simplest option of a constant shift, apart from chance variation. There are some odd patterns in the data but we will ignore them – this underlines the difference between carrying out a data analysis in collaboration with the data owners and simply using other people’s data without intimate knowledge of the conditions under which the data were generated.

If we had the time order in which the measurements were made then a time series plot would be a good idea, as it could indicate a drift with time of the measurements. This could, for example, be the result of a drift in the temperature of a water bath in which the blood samples were stored, or some other time related measurement problem. Any such drift would affect the independence assumption built into the statistical test.

![Figure 2.2.1.1: Differences plotted against before-after means for the 11 subjects](image-url)
There is also an assumption, usually unstated, of homogeneity of the materials/subjects under study, which needs to be considered carefully at the design stage of any study. In the current context there is an assumption that the differences are comparable, in the sense that the different subjects are assumed to react in the same way to the experimental treatment (smoking). Suppose, for example, that previous exposure to a blood related disease (e.g., malaria) affected the response to the treatment. This would mean that subjects belonging to the two different sub-groups (exposed and not exposed) would respond differently and it would be meaningless to discuss the average response in the study for the group as a whole, without taking account of sub-group differences. An extreme example may help to clarify the issue. Suppose one sub-group responds strongly in the positive direction, and the other strongly in the negative direction, then the average change will be close to zero and the statistical test will almost certainly be non-significant. It would, however, be entirely wrong to conclude that “the treatment had no effect”. When extra information is available on the subjects (e.g., age, sex, previous medical history etc.) then it usually makes sense when analysing the results to check that there are no indications of other influences on the responses. It is, of course, better to consider such possibilities at the study design stage. Since we do not have further information here, we will assume that the subjects are drawn from a homogeneous population and proceed with our analysis.
The Normal plot of differences, shown in Figure 2.2.1.3, and the associated Anderson-Darling test (p-value=0.62) do not raise doubts about the Normality assumption.

Paired T-Test and CI: After, Before

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>11</td>
<td>52.4545</td>
<td>18.2995</td>
<td>5.5175</td>
</tr>
<tr>
<td>Before</td>
<td>11</td>
<td>42.1818</td>
<td>15.6129</td>
<td>4.7075</td>
</tr>
<tr>
<td>Difference</td>
<td>11</td>
<td>10.2727</td>
<td>7.9761</td>
<td>2.4049</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (4.9143, 15.6311)
T-Test of mean difference = 0 (vs not = 0): T-Value = 4.27
P-Value = 0.002

Table 2.2.1.1: Minitab paired analysis of the platelet data

Table 2.2.1.1 shows a Minitab paired t-test analysis of the platelet data. The null hypothesis for the test is that the mean of the population of differences is zero: there are 10 degrees of freedom, hence, the critical value for a two-tailed test with a significance level of $\alpha=0.05$ is 2.228 – the observed test statistic of $t=4.27$ (which is $10.27/2.40$) is highly statistically significant; it has a corresponding p-value of 0.002. The interpretation of the test result is that the observed mean difference of 10.3 units is unlikely to be due to the inevitable chance biological variation and measurement error in the study. It suggests an underlying systematic effect. The magnitude of this effect is estimated by the 95% confidence interval in Table 2.2.1.1:
which gives an increase of between 4.9 and 15.6 units as our estimate of the population increase in our measure of platelet aggregation due to smoking one cigarette.

Note the correspondence between the statistical test and the confidence interval – the test rejects the hypothesis that the long-run mean is zero, the interval does not include zero as a possible value for the long-run mean. This agreement between the two methods is necessary, as they embody the same information. However, the confidence interval is more useful, in that it not only tells us that the long-run mean is unlikely to be zero, but also provides a range within which we are confident that the mean difference lies.

In summary, based on this small study, it appears that smoking one cigarette increases the aggregation ability of platelets. Of course, the study raises many more questions: e.g., is this only a short-term effect or does it persist?; if many cigarettes are smoked does the effect increase?; if yes, is the increase additive or multiplicative, does it tail off as the number of cigarettes increases?; is there a cumulative effect of chronic smoking?, and so on.

In this small study 10/11 of the responses increased after smoking a cigarette, so smoking even one cigarette does have an effect. The analysis based on the measured differences would be more informative (it indicates the size of the effect) if we could have confidence in it. However, the two plots (Figures 2.2.1.1/2) raise doubts about the data which could only be resolved in discussion with the researchers and this is not an option for us. An important lesson to be drawn from this example, then, is that simple graphical analyses may be more informative than apparently more sophisticated statistical tests!

2.2.2 (and 2.3.4)

**Laboratory comparison data**

A simple statistical model for the data asserts that the differences are independent and, if we had a very large sample (in principle, an infinite sample) the values would be Normally distributed about a mean, \( \mu \), with a standard deviation \( \sigma \). The question of the existence of a relative bias can then be addressed by asking if the long-run mean difference between the results produced by the two laboratories, \( \mu \), is zero. If the significance test rejects this
hypothesis, then the magnitude of the bias, \( \mu \) may be estimated using a confidence interval. We will assume that the model holds in this case (such a decision could be based on prior analyses of similar but more extensive data) – with so few differences it is hardly worthwhile carrying out graphical analyses.

*Significance Test*

If the proposed statistical model is appropriate for the data, then a paired t-test of the hypothesis that \( \mu=0 \) may be carried out. To carry out the test a significance level, say \( \alpha=0.05 \), is chosen, the null and alternative hypotheses are specified as:

\[
H_0: \mu = 0 \\
H_1: \mu \neq 0
\]

and the test-statistic \( t \):

\[
t = \frac{\bar{d} - 0}{s / \sqrt{n}}
\]

\[
t = \frac{0.637 - 0}{0.237 / \sqrt{6}} = 6.58
\]

is calculated. The resulting value is then compared to Student's t-distribution with \( n-1=5 \) degrees of freedom; this is shown in Figure 2.2.2.1.

![Figure 2.2.2.1: Student's t-distribution with 5 degrees of freedom](image)

This distribution has 95% of its area between +2.57 and −2.57. Since the calculated t-value lies far out in the right-hand tail, the data strongly suggest that the laboratories do not, on average, obtain the same purity results when measuring the same material.
Confidence Interval

For the six batches of product included in the study the average difference was 0.637. However, if a different set of six batches was measured (or if these same batches were measured again) we would almost certainly obtain a different result, because a different set of sampling and measurement errors would be included in our results. Accordingly, rather than report 0.637 as the long-run average difference (relative bias) between the laboratories, it makes sense to attach error bounds to this point estimate; we obtain these by calculating a confidence interval.

As discussed in Section 2.3, a confidence interval simply re-arranges the elements used in carrying out the significance test to produce error bounds around the sample mean difference. Thus, a 95% confidence interval for the long-run mean difference (relative bias) in purity results obtained by the two laboratories is:

\[
\bar{d} \pm t_{c} \frac{s}{\sqrt{n}}
\]

\[
0.637 \pm 2.57 \frac{0.237}{\sqrt{6}}
\]

\[
0.637 \pm 0.249.
\]

The estimate of the relative bias between the laboratories indicates that it is somewhere between 0.39 and 0.89 units.

Computer Analysis

Table 2.2.2.1 shows a Minitab analysis of the paired laboratory data.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab-2</td>
<td>6</td>
<td>89.908</td>
<td>0.364</td>
<td>0.149</td>
</tr>
<tr>
<td>Lab-1</td>
<td>6</td>
<td>89.272</td>
<td>0.553</td>
<td>0.226</td>
</tr>
<tr>
<td>Difference</td>
<td>6</td>
<td>0.6367</td>
<td>0.2368</td>
<td>0.0967</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (0.3882, 0.8852)
T-Test of mean difference = 0 (vs not = 0): T-Value = 6.59 P-Value = 0.001

Table 2.2.2.1: A Minitab analysis of the laboratory data.

The output gives a confidence interval and a test of the hypothesis that the long-run mean difference is zero; the test statistic (slightly different from our value of
6.58 due to our rounding the standard deviation to three decimal places) of 6.59 has an associated p-value of 0.001. The curve in Figure 2.2.2.2 represents the sampling distribution of the test statistic, when the null hypothesis is true, i.e., no long-run mean difference. The p-value (the tail areas) gives the probability of observing a more extreme value of t than was obtained in our study, when this is the case: if there really were no long-run mean difference, then the chance measurement errors in the study could combine to produce a value of t as big as, or bigger, than 6.59 (in either direction). However, the chances of this happening would be of the order of 0.001. Because this is very small, we regard the null hypothesis as implausible, and therefore we reject it, and conclude that a long-run difference exists – there is a relative bias between the two laboratories.

![Figure 2.2.2.2: A schematic representation of the calculation of the p-value](image)

We have identified a relative bias between the laboratories. How important this is would depend on the economics of the product – it is not a statistical question. The confidence interval estimates the size of the bias – this would allow the scientists to evaluate the importance of the difference.

2.2.3 (and 2.3.4)

**Reduced Laboratory Comparison Data**

The Minitab analysis of the reduced dataset is shown below.

**Paired T for Lab-1 - Lab-2**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab-2</td>
<td>5</td>
<td>90.010</td>
<td>0.297</td>
<td>0.133</td>
</tr>
<tr>
<td>Lab-1</td>
<td>5</td>
<td>89.460</td>
<td>0.340</td>
<td>0.152</td>
</tr>
<tr>
<td>Difference</td>
<td>5</td>
<td>0.5500</td>
<td>0.1173</td>
<td>0.0524</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (0.4044, 0.6956)

T-Test of mean difference = 0 (vs not = 0): T-Value = 10.49  P-Value = 0.000

Table 2.2.3.1: Minitab analysis of reduced laboratory comparison data
Note that the t-statistic increased markedly from 6.58 to 10.49 when the sixth difference was dropped from the analysis. The reason for this is that although deleting the large difference, corresponding to batch 6, reduces the average difference from 0.637 to 0.550, it has a marked effect on the standard deviation of differences; this changes from 0.237 to 0.117 – it more than halves. The standard deviation appears in the denominator of the t-statistic and, hence, the reduced size of the standard deviation leads to a larger t-statistic.

Similarly, the confidence interval narrows with the reduced standard deviation; the full width of the interval is 0.497 when 6 differences are used in the analysis, but this reduces to 0.291 when only 5 differences are used. These changes are complementary.

This example shows the marked impact even a single unusually large or small value can have on a statistical analysis. This underlines the need to examine data carefully rather than simply loading them into a computer and requesting a standard statistical test. **Data processing is not the same as data analysis!**

### 2.3.1 Baby food data

A 95% confidence interval for the long-run mean result that would be obtained if very many measurements were made on the baby food in Laboratory A is given by:

$$
\bar{x} \pm t_c \frac{s}{\sqrt{n}}
$$

26.943 ± 2.26 \( \frac{0.294}{\sqrt{10}} \)


We estimate that the long-run mean is somewhere between 26.733 and 27.153 units (percentage fat). This interval covers the certified value of 27.100. This was expected as the statistical test did not reject the certified value. The critical value of 2.26 used in the calculation is based on 9 degrees of freedom, since the sample size was 10.

### 2.3.2 Fill heads data
Tables 2.3.2.1 and 2.3.2.2 give Minitab output which contains both the t-tests (see Exercise 2.2.1) and the corresponding confidence intervals for the long-run average fill levels for the two fill heads.

Test of \( \mu = 21 \) vs not \( \neq 21 \)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>21.0600</td>
<td>0.1180</td>
<td>0.0373</td>
<td>(20.9756, 21.1444)</td>
<td>1.61</td>
<td>0.142</td>
</tr>
</tbody>
</table>

Table 2.3.2.1 One-sample analysis for head 1

Test of \( \mu = 21 \) vs not \( \neq 21 \)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>21.330</td>
<td>0.360</td>
<td>0.114</td>
<td>(21.072, 21.588)</td>
<td>2.90</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 2.3.2.2 One-sample analysis for head 2

Note that for head 1 the confidence interval contains 21.00, corresponding to the null hypothesis value of 21.00 not being rejected by the corresponding t-test, while the interval for head 2 does not contain 21.00, corresponding to the rejection of such an hypothesised value.

2.3.3 and 2.3.4

Note the confidence intervals for 2.3.3 and 2.3.4 are discussed together with the t-tests in solutions 2.2.2 and 2.2.3, respectively.

2.4.1 (and 2.4.3)

**Pain-relief Study**

Since we do not have the raw data we will have to assume that the data are sufficiently close to Normality to permit us to carry out a t-test. Samuels uses the data in an exercise on t-tests so, presumably, this is a reasonable assumption. In any case, the sample sizes are reasonably large so that the distribution of the test-statistic would not be badly affected by minor departures from this assumption.

A two-sample t-test addresses the hypothesis that the population means are the same. In a case like this, obviously, we do not have a random sample from a fixed population. If we can regard the subjects in the study as being representative of mothers who have just delivered a baby, or at least some well-specified subset of mothers, then we can hope that our inferences generalise beyond the set of women being studied. The means in question are then the
(population) means that would be obtained if very large numbers (instead of just 25) were measured in each of the two treatment groups.

The t-statistic is given by:

\[
t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{2s^2}{n}}}
\]

where the combined \( s^2 \) is given by:

\[
s^2 = \frac{s_1^2 + s_2^2}{2} = \frac{12.05^2 + 13.78^2}{2} = 167.54
\]

The calculated t-statistic is:

\[
t = \frac{(31.96 - 25.32) - 0}{\sqrt{\frac{2(167.54)}{25}}} = 1.81
\]

The sampling distribution of the test-statistic has 48 degrees of freedom, since the \( s^2 \) used in the calculation has 48 degrees of freedom – it takes (25-1) from each of the standard deviations from the two groups. The critical value for a two-sided test using a significance level of \( \alpha = 0.05 \) is 2.01, so our t-statistic is not quite statistically significant according to this criterion.

The Minitab analysis shown in Table 2.4.1.1 gives a p-value of 0.076; this applies for a two-tail test. If a one-tail test were carried out (as suggested by Samuels\(^{12}\)) then the p-value would be half of this (0.038) and the results would be considered statistically significant, when compared to the conventional significance level of 0.05. The one-tailed critical value for a test with a significance level of \( \alpha = 0.05 \) is 1.68, and our observed t-statistic exceeds this – which is exactly equivalent to the p-value being less than 0.05.

\(^{12}\) In our discussion of one versus two-sided tests (page 13-14) it was pointed out that many statisticians would discourage one-sided tests in a research context. So why might Samuels look for a one-sided test here? In this study a drug is being compared to a placebo (rather than to an established treatment). In this case any placebo effect will apply equally to the drug and the placebo, so it is reasonable to look for an increased pain relief score in the drug group.
Two-Sample T-Test and CI

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>25</td>
<td>32.0</td>
<td>12.1</td>
<td>2.4</td>
</tr>
<tr>
<td>Placebo</td>
<td>25</td>
<td>25.3</td>
<td>13.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Difference = mu (1) - mu (2)
Estimate for difference: 6.64000

95% CI for difference: (–0.72113, 14.00113)
T-Test of difference = 0 (vs not =): T-Value = 1.81
P-Value = 0.076 DF = 48
Both use Pooled StDev = 12.9439

Table 2.4.1.1: Minitab two-sample analysis of the pain relief study data

It is important to review results like this one and ask why the difference (on a two-sided test) is ‘not statistically significant’. The implicit assumption for those who simply declare the result ‘not statistically significant’ is that there really is no long-run difference between the population means, i.e., the treatment has no effect. However, if we examine the test-statistic, it becomes clear that it may be small because the sample difference $(\bar{y}_1 - \bar{y}_2)$ is small (perhaps because there is no treatment effect), or because the denominator is large, i.e., because the sampling variability as measured by the common standard deviation, s, is large, or because the sample size in each group, n, is small. If the sampling variability is large then in designing the study a sample size, n, which reflects the large sampling variability (here the woman-to-woman variation in pain relief scores), should have been chosen. If this was not done then the reason that the result was not statistically significant could simply be that the sample size was too small to counteract the large sampling variability. Technically, this is described by saying that the power of the test to detect the treatment effect was too low. This immediately raises the question as to how many observations are needed in studies of this type – sample size determination will be discussed in Chapter 4.

An alternative approach to reporting the outcome of studies of this type is to present a confidence interval for the population mean difference. This is given by:

$$(\bar{y}_1 - \bar{y}_2) \pm t_c \sqrt{\frac{2s^2}{n}}$$
using a 95% confidence coefficient. As shown in Table 2.4.1.1, this interval stretches from –0.72 to +14.0. The interpretation of the interval is that, with a confidence coefficient of 0.95, the drug could give pain relief scores that are up to 14 units higher on average (in the population) than does the placebo. It also suggests that the placebo could give average relief scores higher by up to 0.72 units. Thus, there is not clear evidence the treatment works (the fact that the confidence interval covers zero, is equivalent to obtaining a non-significant result from the t-test). The advantage of the confidence interval over the t-test (where the t-test is interpreted simply as ‘significant’ or not) is that it shows that there could be quite a large effect (presuming that values of up to 14 are of practical interest). It also explicitly shows a measure of the likely error involved in taking the observed difference of 6.64 at face value – thus focusing attention on the large chance variation that affects our analysis.

2.4.2: (and 2.4.3)

**Tobacco Plant Study**

This question requires exactly the same calculations as the previous one, so we will take these as given and examine the Minitab analysis as shown in Table 2.4.2.1.

\[
(31.96 - 25.32) \pm 2.01 \sqrt{\frac{2(12.94)^2}{25}} \approx 6.64 \pm 7.36
\]

The t-statistic of \( t = 2.68 \) is highly statistically significant (\( p \)-value is 0.012) and so we reject the null hypothesis that in a very large study (in principle, infinitely
large) the average weight of the larvae would be the same for wounded and control plants. The wounded plants clearly give smaller weights; it appears that wounding the plant does inhibit larvae growth. The critical values for a two-tailed test, using a 5% significance level are ±2.04, since the t-statistic will have 32 degrees of freedom. This value is used in calculating the confidence interval, also. As shown in Table 2.21, the 95% confidence interval is between 2.2 and 16.3 (the fact that the interval does not cover zero is equivalent to the test result being statistically significant). We interpret the confidence interval as meaning that if the experiment were repeated under identical conditions, but with very large numbers of plants (an infinite number, in principle) the reduction in the weight of larvae on the wounded versus the control plants would be between 2.2 and 16.3 mg, on average.

2.4.3

The confidence interval for the pain-relief study is given in Solution 2.4.1, while that for the tobacco plant study is given in Solution 2.4.2.

2.4.4

Pain relief and Tobacco plant studies

For the pain relief data of Exercise 2.4.1, the F ratio is $F=\left(\frac{13.78}{12.05}\right)^2=1.31$. The critical value is $F_{c}=F_{0.975,24,24}=2.27$ for a significance level of $\alpha=0.05$ – there is no reason to reject the assumption of underlying standard deviations which are the same.

For the tobacco plant data of Exercise 2.4.2, the F ratio is $F=\left(\frac{11.14}{9.02}\right)^2=1.53$. The critical value is $F_{c}=F_{0.975,16,16}=2.76$ for a significance level of $\alpha=0.05$. – again, there is no reason to reject the assumption of equal underlying standard deviations.
2.4.5:

**Margarine Fat Data**

The standard deviations are 0.374 and 1.723 for Laboratories A and B, respectively. These give an F-ratio of 21.18, which is highly statistically significant when compared to the critical value of $F_{c} = F_{0.975,9,9} = 4.03$, for a significance level of $\alpha = 0.05$.

Minitab gives the following output for a t-test on the data – note that equal standard deviations are not assumed by this test. Note, also, that the degrees of freedom of the t-test are shown as 9 rather than 18, as would be expected if the population standard deviations were assumed equal and the sample values were combined for the test (see comments on page 49).

**Two-sample T for A vs B**

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>79.292</td>
<td>0.374</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>77.73</td>
<td>1.72</td>
</tr>
</tbody>
</table>

Difference = mu A - mu B
Estimate for difference: 1.557
95% CI for difference: (0.296, 2.818)
T-Test of difference = 0 (vs not =): T-Value = 2.79  P-Value = 0.021  DF = 9

Table 2.4.5.1: Comparing groups where the population standard deviations are different.