Chapter 4: Designing Comparative Studies

4.0 Introduction

Chapter 2 introduced the fundamental concepts of significance testing and confidence intervals; it did so in the context of analysing data from designed studies where the performance measures were continuous. Chapter 3 then applied these ideas and methods to the analysis of counted data. The design of such studies precedes, obviously, their analysis and this is the subject matter of this chapter.

Example 5 of Chapter 2 discussed a study that involved comparing the responses of a machine for two different machine speeds: 22 runs were carried out in randomised order. The need for randomisation will be discussed first. Exercise 2.2.1 studied the effect of cigarette smoking on platelet aggregation: blood was taken from 11 subjects before and after they smoked a cigarette and platelet aggregation was measured in each case. This involved two separate design concepts. The first concerns the need for a control in the study: platelet aggregation was studied not just under the ‘treatment’ (cigarette smoking’), but there was a direct comparison within the study to the aggregation properties of blood in the absence of the treatment. The control was organised in a special way: each subject acted as his/her own control, such that the difference in response under the two experimental conditions became the natural performance measure that was analysed. The design was ‘self-paired’ and the method of analysis was a paired t-test. Pairing is a special case of the more general concept of ‘blocking’: these ideas are discussed in the first section of this chapter.

Whenever a study is planned, the question arises as to how many observations are needed. This question is inherently statistical, since it requires balancing risks: the risk, in the face of chance variation, of incorrectly deciding that a systematic effect exists, when it does not, against the risk of failing to detect an effect which does exist. Larger sample sizes will tend to reduce the influence of chance variation, but at a cost. A structured approach to deciding on suitable study sizes will be introduced in Section 4.3. As an introduction to the ideas underlying the determination of sample size, Section 4.2 takes a closer look at the nature of significance tests and, in particular, discusses how errors arise (wrong decisions are made) when using these tests. This leads to a discussion of the concept of ‘statistical power’.

The final section of the chapter discusses the nature of replication – we typically measure several subjects/experimental units (the replicates) in a comparative study. Often though, these units exist in a nested or hierarchical structure and this can have an important influence on how the study should be organised. Such ideas form the basis of Section 4.4.
4.1 Arranging the conduct of the study: randomisation, control groups, blocking

Randomisation

In any study it is always possible that observed effects may be due, not to the condition under study, but to other factors which, unknown to those carrying out the work, influence the response of the system. For purely observational studies based on retrospective data analysis, there is no built-in protection against this happening. A thorough analysis of all opportunities for biases to occur is, therefore, essential (though, of course, this does not guarantee that all possible sources of bias will be identified). When inferences are made from such studies, a confirmatory experimental study (where this is possible) will often be appropriate in order to validate the conclusions. Experimental studies allow more control over the conditions under which data are collected. In particular, the opportunity to randomise the study sequence provides protection against unknown sources of bias. Before discussing how randomisation is carried out, the implications of two systematic orderings of the study sequence will be examined.

Suppose two system configurations A and B are to be compared (for example, two speeds for a spheroniser, as in Example 5 of Chapter 2) and that six runs using each configuration will be carried out. Figure 4.1.1 shows two possible systematic arrangements for the order in which the twelve results will be generated.

![Figure 4.1.1: Two possible experimental sequences](image)

Time order (1) appears a natural ordering for the study - a tidy-minded experimenter will naturally favour completing one set of runs, all carried out using the same configuration, before starting on the second set. This ordering, however, leaves the results of the study open to unknown sources of bias. Suppose that half-way through the study there is a once-off downwards shift in the responses of the system which is unrelated to the A/B 'treatments'. The shift will mean that all the B results will be reduced with respect to the A results, even if, fundamentally, there is no difference between the system responses when configurations A and B are used. Where an A, B difference does exist, the shift could either counterbalance or exaggerate this difference. If the shift occurs at a point other than half-way through the experimental sequence it will have a lesser effect, since it will affect all responses for one configuration and only some for the
other. Nevertheless, it will bias the results and, potentially, jeopardise the outcome of the study.

Suppose that, instead of a sudden shift, there is a downwards drift in results. This might, for example, be due to failure of a temperature controller, leading to a drift in the temperature at which a chemical reaction takes place. The effect on the system responses will be as before - although the responses for individual runs may be affected to a different extent, overall, configuration B results will all be reduced relative to the A results, leading to a bias in the analysis.

Time order (2) also appears to be a natural ordering for such a study, representing an attempt on the part of the experimenter to be 'fair' to both configurations. This ordering will protect against a once-off shift. Unless an A, B pair is split by the shift, A and B will be equally represented before and after any shift: this means that the effect of the shift on the averaged results will be eliminated when the difference between the averages for A and B is calculated. Splitting a single A, B pair will be unlikely to have a major impact on the statistical analysis. Time order (2) does not, however, protect the study results from the effects of a drift. An upwards drift in responses, for example, will mean that the B results for all pairs will be incremented more than the A results. Consequently, the average for B will be raised relative to A. Again, this could either exaggerate or reduce any real difference between the responses produced by the two system configurations. Time order (2) is also open to biases caused by a cyclical systematic effect.

In the preceding discussion it is the possibility of unknown sources of bias that is of concern when designing a study. If known about, they can either be designed out or, at least in principle, allowed for in the analysis.

Example 1: An industrial study

Hahn [1] described a process development study in the electronics industry. A statistician was approached by engineers who intended to carry out a study, as follows. They wanted to discover if a particular raw material could be substituted by a new one without increasing the percentage of non-conforming components produced by the process; if this could be done there would be a large economic gain as the new material was considerably cheaper. Prior to the process change, the daily percentage of defective units fluctuated appreciably, averaging about two percent. They intended running the process for a trial period with the new material, while sampling 50 components per day over a six-day week and classifying the components as conforming/non-conforming. They would then compare the rate of defective units to the rate for a comparable period using the standard material. The question for the statistician was “how many weeks were required for the study to achieve ‘statistically significant results’?”.
The intended study design corresponds to the first sequence of Figure 4.1.1. The statistician counselled against such an arrangement for the reasons discussed above. After discussion, the engineers agreed to alternate the sequences shown in Figure 4.1.2 from week to week. Hahn points out that “In other circumstances, it might be better to group days into pairs and randomly determine the days on which the new and old processes are to be run”.

![Figure 4.1.2: Two orderings for the study material](image)

The study was run for 8 weeks, which meant that 2400 components in total were sampled from the process, 1200 for each of the two materials. The overall study results are shown in Table 4.1.1.

<table>
<thead>
<tr>
<th>Old material</th>
<th>New material</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 4.1.1: Percentage non-conformances during the study

The study was successful: there was no evidence to suggest that the new material was any worse than the old material, since the observed numbers of non-conforming components was lower for the new material.

The statistician analysed the data in more detail to see what would have happened if the study had been carried out as originally proposed. Table 4.1.2 summarises the data split into the two successive time periods.

<table>
<thead>
<tr>
<th></th>
<th>First 4 weeks</th>
<th>Second 4 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old material</td>
<td>3.3</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>New material</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 4.1.2: Percentage non-conformances for the two time periods
It is clear that whichever material had been used first would have appeared worse, as approximately 3% non-conformances were obtained for the first 4 weeks and 1% for the second 4 weeks, irrespective of the type of material used in the process. The outcome of the study, as originally proposed, would have depended entirely on the order in which the materials were used in the process. Some other unknown influence determined that 3% non-conformances were obtained for the first 4 weeks while approximately 1% were obtained during the second 4 weeks.

Randomising the study sequence

As pointed out earlier, it is the possibility of unknown sources of bias that is of concern when designing a study. If known about, they can either be designed out (by blocking, for example; see later) or, at least in principle, allowed for in the analysis. Randomisation of the order in which the runs are carried out is the best protection against an unknown source of bias: it is unlikely that the randomised sequence will coincide with the source of bias. For two-sample comparisons of the type discussed above, the two groups/methods/treatments will tend to be affected equally by a shift or drift, under random allocation, and the effects of the bias will tend to cancel. From a purely technically statistical point of view, the process of randomisation also underpins the assumption of independent observations, which is embodied in the statistical models underlying significance tests and confidence intervals; it is, therefore, a good thing!

<table>
<thead>
<tr>
<th>Random Nos.</th>
<th>Labels</th>
<th>Sorted Nos.</th>
<th>Study Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.49100718</td>
<td>A</td>
<td>0.17139003</td>
<td>B</td>
</tr>
<tr>
<td>0.50394854</td>
<td>A</td>
<td>0.18515499</td>
<td>B</td>
</tr>
<tr>
<td>0.48992599</td>
<td>A</td>
<td>0.24861655</td>
<td>A</td>
</tr>
<tr>
<td>0.77446668</td>
<td>A</td>
<td>0.43278067</td>
<td>B</td>
</tr>
<tr>
<td>0.24861655</td>
<td>A</td>
<td>0.48992599</td>
<td>A</td>
</tr>
<tr>
<td>0.87444864</td>
<td>A</td>
<td>0.49100718</td>
<td>A</td>
</tr>
<tr>
<td>0.17139003</td>
<td>B</td>
<td>0.50394854</td>
<td>A</td>
</tr>
<tr>
<td>0.43278067</td>
<td>B</td>
<td>0.55030938</td>
<td>B</td>
</tr>
<tr>
<td>0.55030938</td>
<td>B</td>
<td>0.61849910</td>
<td>B</td>
</tr>
<tr>
<td>0.79971223</td>
<td>B</td>
<td>0.77446668</td>
<td>A</td>
</tr>
<tr>
<td>0.18515499</td>
<td>B</td>
<td>0.79971223</td>
<td>B</td>
</tr>
<tr>
<td>0.61849910</td>
<td>B</td>
<td>0.87444864</td>
<td>A</td>
</tr>
</tbody>
</table>

Table 4.1.3: Generating a randomised sequence for the study

To carry out the randomisation, the random number generator in either a statistical or spreadsheet package could be used. For example, column 2 of
Table 4.1.3 contains labels for the two groups/methods/treatments. Column 1 was generated using the random number generator in Excel. This generates (uniform) numbers between 0 and 1, with all numbers in that interval being equally likely. Using the SORT command, column 1 was sorted into ascending order and placed into column 3, while the labels in column 2 were carried with the corresponding values in column 1 and placed into column 4. Thus, 0.17139 is the smallest number in column 1 and, therefore, becomes the first value in column 3; the label B is carried with it and becomes the first label in column 4. Since the magnitudes of the numbers in column 1 are randomly ordered, sorting column 2 in association with column 1 generates a randomised ordering for the A, B labels. The study should then be carried out using the randomised order of the rows of column 4.

Randomisation is recommended virtually always for selection (e.g., selection of participants into a survey or experimental study) or allocation (e.g., allocation of patients to the treatment or control groups in a clinical trial), not just for time-ordered situations. Thus, random selection is implicitly assumed in carrying out statistical significance tests as part of acceptance sampling procedures of the type discussed in Chapter 2. It also underlies population estimates based on sample survey data (as discussed in Chapter 3), even though the selection of respondents is often not truly random. In experimental studies, unlike many sampling problems, the process of randomisation is usually relatively simple to implement, as shown above. It should be incorporated automatically into study protocols unless there are good reasons for not doing so (I am not sure what these might be!).

**Control Groups**

For studies of the type just described, it would be natural to ask why it is necessary to include the old material in the study: after all, it has been used for some time and the rate of non-conformances for the old material is well known. The answer lies in the outcome of the study. To use the historical information as a reference point would be to assume that nothing behaves differently during the course of the study – or to put it differently, this would be equivalent to running the old material first, to pre-date the beginning of the study to some point during recent production runs. Clearly, this would not have been a good idea in the current case. In general, it is advised that comparative studies should include control groups as well as ‘treatment’ groups of observations, i.e., whatever conditions are to be compared should be included in the study itself. Furthermore, all assignments should be randomised, whether these relate to time order, people or other experimental units (for example, where people are assigned to experimental conditions or drug treatments in psychological or medical studies).
Example 2: A meta analysis of coronary bypass surgery

Freedman et al. [2] cite a meta-analysis of studies of the efficacy of coronary bypass surgery [3] which illustrates the reason why randomised controlled trials are now the gold standard method for the evaluation of medical interventions. A meta-analysis is a study of studies – it draws together the accumulated evidence from the available studies on a particular treatment. Table 4.1.4 shows the conclusions drawn from 29 studies of coronary bypass surgery cross-classified by whether or not the study conclusions were positive or not on the use of the surgical procedure and by the nature of the controls used in the study, randomised or historical controls. Historical controls are patients with the same medical condition who were selected from medical records for the purposes of comparison with the experimentally treated patients.

<table>
<thead>
<tr>
<th>Conclusions</th>
<th>Randomised</th>
<th>Historical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Negative</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.1.4: Study conclusions classified by type of controls

There is a marked difference in the conclusions drawn from historically controlled studies 16/21 favourable as opposed to only 1/8 for the randomised studies. The three-year survival rates for the different groups are shown in Table 4.1.5.

<table>
<thead>
<tr>
<th>Study Group</th>
<th>Randomised</th>
<th>Historical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>Control</td>
<td>83</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 4.1.5: Three-year percentage survival rates for the different groups

The three-year survival rates of patients involved in randomised studies shows an increased survival rate of 5 percentage points for the surgery patients, where the non-randomised studies suggest an increased survival rate of around 20 percentage points. Clearly, use of historical controls has greatly exaggerated the possible efficacy of the surgical procedure (the 71% survival rate for the historical controls compared to the 83% rate for randomised controls calls their suitability as controls into question). The overall conclusion (applied to all the interventions considered in the meta-analysis, not just coronary bypass surgery) in the original paper [3] was “the data suggest that biases in patient selection may irretrievably weight the outcome of HCTs (i.e., historically controlled trials) in favour of new therapies”. Freedman et al. cite further similar examples.
**Pairing/Blocking**

**Pairing**

Example 1 of this chapter, the industrial study described by Hahn, is an example of a paired study. The pairing in this case was based on adjacent days. Historically, there had been substantial temporal variation in the proportions of non-conforming components produced by the process, and the arrangement of the study into pairs of days was designed to create uniform conditions for the comparison. Essentially, the new and old raw materials were compared within each two-day period and the differences were then averaged. Each pair of days would be more similar than more widely separated days (as in a fully randomised design) and so all other factors affecting the outcome of the study would be expected to be more closely matched, so that the difference within the two-day period would be less affected by extraneous factors than a similar comparison on two more widely separated days. If the chance variation in level of non-conformances between pairs of days was substantial, then the pairing would have resulted in greatly improved precision in the comparison of the two materials, over and above what would have been seen if the new and old materials had each been assigned randomly to 24 of the 48 days during which the study was conducted.

Consider again Exercise 2.2.1 of Chapter 2 (page 21) where the platelet aggregation of 11 individuals was measured on blood taken before and after smoking a cigarette. An alternative design for such a study would have been to take 22 people and randomly allocate 11 to be tested without being subjected to cigarette smoking, while the other 11 were tested after smoking a cigarette. This design would lead to a comparison of the mean responses using a two-sample t-test. The denominator of the two-sample t-test would contain a standard deviation that included person-to-person variation in the level of platelet aggregation. For both genetic and environmental/behavioural reasons person-to-person variation for all sorts of measures can be substantial – thus increasing the standard error in the denominator of the test statistic. The standard deviation that appears in the denominator of the paired t-test, on the other hand, includes only the variation from person to person of within-person differences (i.e., the between-person variation in platelet aggregation is largely excluded); accordingly, it would be expected to be smaller. The paired design should, therefore, give a more sensitive comparison. Note that, in a study like this, the 22 blood samples should be tested in a randomised order, to protect against the possibility of a time trend in the measurement process.

Pairing or matching is often used in psychological and medical studies. Pairing does not always mean self-pairing, as was the case in the platelet study and also in the retinol binding protein secretion study (pages 14-21, Chapter 2). Twin studies are frequently reported in the psychological literature and patient matching is common in the medical literature. Matched case-control studies, for
example, are used in investigations of possible antecedent factors for medical conditions, especially for rare diseases. A suitable number of patients with the disease is identified, and these are then matched individually with people not having the disease. Matching will be based on demographic characteristics such as age and sex, and possibly factors such as where the subjects live. The subjects will then be classified by potential antecedent factors to establish if there are statistically significant differences between the two groups. An example, of such a case-control study was discussed in Chapter 3 (pages 23-25). The study involved 317 patients, who were diagnosed as having endometrial carcinoma, being matched with other cancer patients. The controls all had one of cervical cancer, ovarian cancer or cancer of the vulva. Each case was matched by age at diagnosis (within 4 years) and year of diagnosis (within two years) to a corresponding control. The question of interest was whether or not there is an association between exposure to estrogen and the presence of endometrial cancer.

A numerical example

A numerical example, which will allow us to inspect some technical details, may help to clarify the implications of choosing a paired design over a two-independent-samples design in comparative studies. We will use the laboratory comparison study data of Chapter 2 (Example 4) to illustrate the discussion. This study was naturally paired: 40 batches were measured twice, once in each of two laboratories and the question at issue was whether or not the laboratories were biased relative to each other, i.e., whether in the long-run there was a difference in the mean results from the two laboratories. We will label a single measurement in Lab-1 as $X$ and one in Lab-2 as $Y$.

The test statistic for a $t$-test to compare the means of the two groups is the ratio of the sample mean difference and the estimated standard error of that difference:

$$ t = \frac{\bar{d}}{SE(\bar{d})} = \frac{(\bar{X} - \bar{Y}) - 0}{SE(\bar{X} - \bar{Y})} = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{Var(\bar{X} - \bar{Y})}} $$

Since the mean of the individual differences is numerically identical to the difference between the sample means, the numerator is the same for both the paired and two-sample versions of the $t$-test. The difference between the two $t$-tests is determined by how the denominator of this expression is calculated. In Chapter 1, where we discussed combining random quantities, we saw that the variance of a difference between two independent random quantities $X$, $Y$ was:

$$ Var(X - Y) = Var(X) + Var(Y) \quad (4.1) $$
The variances, in practice, have to be estimated from the sample data; we have seen that the variance of $X$ is estimated by:

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

with a similar expression for $S_Y^2$.

If the quantities of interest are NOT independent, then a modified expression is required:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X,Y) \quad (4.2)$$

Where Cov$(X,Y)$ is the covariance of $X$ and $Y$ – a measure of the extent to which they vary together. It is estimated by:

$$S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Note that if $Y$ is replaced by $X$ in this expression it reduces to $S_X^2$, which measures the extent of the variation in $X$.

The basis for this expression can be seen in Figure 4.1.3. When the means are subtracted to give the deviations $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, the origin from which variation is measured shifts from $(0,0)$ to $(\bar{X}, \bar{Y})$, as shown in Figure 4.1.3.
Points falling in quadrant I have \((X_i - \bar{X})\) and \((Y_i - \bar{Y})\) deviations which are positive and hence their product \((X_i - \bar{X})(Y_i - \bar{Y})\) will be positive. For points in quadrant III both deviations are negative and their product is again positive.

Points falling in quadrants II and IV have one positive and one negative deviation, so their product is negative.

If the data points fall mainly in quadrants I and III, then there is a ‘positive’ relationship (large positive covariance), i.e., the products \((X_i - \bar{X})(Y_i - \bar{Y})\) are mainly positive and \(X\) and \(Y\) tend to increase together; if they mainly fall into quadrants II and IV, the products are negative and the relationship is said to be ‘negative’ (large negative covariance) – as \(X\) increases \(Y\) decreases, and vice versa. When the points occupy all four quadrants, there is little or no relationship between \(X\) and \(Y\) and the covariance will be small (since the positive and negative products will cancel each other).

For given amounts of variation in \(X\) and \(Y\) the covariance will increase as the points fall closer and closer to a straight line: the more closely \(X\) can be used to predict \(Y\) (or vice versa) the greater the covariance. In terms of paired studies, this means the more closely matched the members of the pairs are (i.e., the smaller the chance variation within pairs) the greater is the covariance.

The denominator for a two-sample test statistic is based on expression (4.1), since the two sets of data are assumed independent. The denominator for a paired \(t\)-statistic (see below) is based on expression (4.2), which means that where the covariance is large and positive (i.e., where matching or pairing is effective) the standard error in the denominator will be smaller and, hence, for a given difference in the sample means, more likely to produce a statistically significant test result. Equivalently, it will produce a narrower confidence interval for the long-run mean difference.

Figure 4.1.4 shows measurements on the 40 batches of tablets from the two laboratories plotted against each other: there is a clear tendency for the larger measurements for Lab-1 to correspond to the larger measurements for Lab-2, and similarly for smaller measurements. Table 4.1.6 shows the sample variances (on the diagonal) and the covariance between the two sets of measurements.
Figure 4.1.4: A scatterplot of the paired measurements on the tablets.

Table 4.1.6: Variances and covariance for the tablet data.

If \( X \) represents a Lab-1 measurement and \( Y \) the corresponding value for Lab-2 then:

\[
\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{d}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2\text{Cov}(\bar{X}, \bar{Y})
\]

Which we estimate as:

\[
\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{d}) = \frac{S^2}{n} + \frac{S^2}{n} - 2 \frac{S_{XY}}{n}
\]

\[
= \frac{8.73836}{40} + \frac{9.22356}{40} - 2 \frac{7.69046}{40} = 0.064525
\]

the square root of which gives an estimated standard error of 0.254, precisely that shown in Table 4.1.7 which is a Minitab paired t-test analysis of the tablet data. This shows that the estimated standard error of the sample mean difference is identical whether calculated as above using expression (4.2) or as we did in Chapter 2 using \( S_d / \sqrt{n} \) (= 1.607 / \( \sqrt{40} \) as shown in Table 4.1.7).
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-1</td>
<td>40</td>
<td>40.040</td>
<td>2.956</td>
<td>0.467</td>
</tr>
<tr>
<td>Lab-2</td>
<td>40</td>
<td>41.595</td>
<td>3.037</td>
<td>0.480</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: (-2.069, -1.041)
T-Test of mean difference = 0 (vs not = 0): T-Value = -6.12  P-Value = 0.000

Table 4.1.7: A paired t-test analysis of the tablet data

If we ignore the pairing (which would be equivalent to assuming that 40 randomly selected batches were measured in Lab-1 and a different 40 randomly selected batches were measured in Lab-2) then:

\[
V_\text{ar}(\bar{X} - \bar{Y}) = \text{Var}(\bar{d}) = \frac{S_X^2}{40} + \frac{S_Y^2}{40}
\]

\[
= \frac{8.73836}{40} + \frac{9.22356}{40} = 0.4490
\]

the square root of which gives an estimated standard error of 0.670 – more than double that obtained when the covariance is taken into account. The corresponding two-sample t-test is shown in Table 4.1.8.

Two-sample T for Lab-1 vs 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-1</td>
<td>40</td>
<td>40.04</td>
<td>2.96</td>
<td>0.47</td>
</tr>
<tr>
<td>Lab-2</td>
<td>40</td>
<td>41.59</td>
<td>3.04</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Difference = mu (Lab-1) - mu (Lab-2)
Estimate for difference: -1.555
95% CI for difference: (-2.889, -0.221)
T-Test of difference = 0 (vs not =): T-Value = -2.32  P-Value = 0.023  DF = 78
Both use Pooled StDev = 2.9968

Table 4.1.8: A two-sample analysis of the tablet data.

Note that the t-statistic is very much smaller for the two-sample test -2.32 (i.e., -1.555/0.67) than for the paired t-test -6.12 (i.e., -1.555/0.254); the corresponding p-values are 0.023 versus 0.000, i.e., less than 0.0005.\(^1\)

\(^1\) If the sample size had been only 20 instead of 40, then almost certainly the paired t-test would have been statistically significant while the two-sample t-test would not – significance is strongly affected by sample size.
This example illustrates the benefits of matching/pairing. Creating uniform conditions within-pairs (i.e., resulting in high covariance) leads to a reduction in the standard error of the sample mean difference (through subtraction of the covariance term in (4.2)) and thus gives a more sensitive test (more powerful is the technical term – see Section 4.2) and a narrower confidence interval for the long-run or population difference (compare the widths of the confidence intervals in the two computer outputs: ±0.5 for the paired analysis versus ±1.3 for the two-sample analysis).

While the foregoing discussion was based on continuous data, the same ideas apply where the data are counts.

**Blocking**

Paired studies are a special case of ‘blocking’, which tries to eliminate the effect of influential factors external to the aims of the study and which cannot be held constant over the entire study. It does this by making comparisons between the treatments within blocks (i.e., within sets of experimental units for which these factors are likely to be constant) and by allowing these factors to vary between blocks. The term ‘block’ comes from agricultural experimentation, where it refers to adjacent plots in a field experiment, that are likely to be more similar in their fertility characteristics than are plots randomly selected over the entire experimental area.

![Figure 4.1.5: Two possible layouts for a field trial](image)

Figure 4.1.5 shows, schematically, two possible arrangements for allocating four treatments A-D (perhaps four varieties of a crop) to a total of 16 plots in a field experiment. In Figure 4.1.5 (a) the treatments have been allocated at random, while in (b) the four treatments were randomly allocated within blocks, where blocks are horizontal strips of plots. Suppose that there is a fertility gradient running from the top to the bottom of the picture. Clearly, in (b) each treatment has had an equal opportunity to demonstrate its yield at the different fertility levels – comparisons between them take place within-blocks – and the fertility
gradient will not affect the assessment of the comparisons. In Figure 4.1.5 (a), however, due to the outcome of the random allocation, C does not appear in the topmost strip and B does not appear in the bottom strip. Because of the random allocation, the fertility differences between the plots are treated as part of the chance variation in the study, but they serve to inflate this variation unnecessarily and, hence, make it more difficult to detect systematic treatment differences. The comparisons are, therefore, less sensitive.

The blocks, instead of being strips in a single field, might be entire fields in different geographical areas. The overall study gains generality in the process – any demonstrated differences will be shown to hold over a range of study conditions, not just in a single geographical setting. The process of blocking thus conforms to the scientific requirement that results (here differences) be shown to be generalisable and strengthens the conclusions that may be drawn from the study by showing that the results are reproducible over a range of experimental conditions.

In the conduct of the study it is desirable that experimental units within blocks should be treated as uniformly as possible. For example, in Figure 4.1.5 (b) it would make sense that any manipulation of the individual plots should coincide with the block structure. Thus, spraying or weeding of the individual plots should be done block by block, as should harvesting. In this way, if systematic differences arise (perhaps unknown to the experimenter) they will tend to coincide with the blocks and therefore, they will not affect comparisons within blocks. Similarly, in an industrial study, such as Hahn’s example, discussed earlier, where the blocks were pairs of consecutive days, if the assessment of the components involved a subjective element (cosmetic flaws can be important for some products) the same inspector should assess all the components of a two-day pair. In a medical or psychological study, the same assessor should assess both members of a matched pair. More generally, physical or (bio)chemical testing of samples of product should be carried out block by block, so that any (perhaps unknown) changes to the measurement systems will tend to coincide with blocks and will not, therefore, affect within-block treatment comparisons.

As a second example of blocking in an industrial context, suppose four different manufacturing process configurations are to be compared and ten replicates of each are to be run. Suppose that the raw material (perhaps biological) is variable from batch to batch and that one batch is sufficient for four runs of the process. A ‘completely randomised’ design would entail randomising the run order for all 40 batches. A ‘randomised block’ design would involve running each of the four configurations for each batch of raw material, while randomising the order of runs separately for each batch. The comparisons between the four process configurations would then be carried out within each batch (block), thus making the comparison more sensitive than that from a fully randomised design, assuming strong batch-to-batch variation in the raw material.
Where blocking is possible it should lead to greater precision (more powerful statistical tests - see later in this chapter for a discussion of power – and narrower confidence intervals), for a given sample size. Alternatively, it may also be used to reduce the sample size of the study, while maintaining the precision of the comparison between ‘treatments’. This can be important where the experimental material is costly, where the time taken to carry out individual experimental runs is long, or where the experimental material is scarce. For example, patients with rare diseases may only be available in small numbers for a study, and anything that will reduce the number required for a study will be beneficial.

Can blocking/pairing be bad in any way? Yes, if the pairing is inadequate. Recall that in a two-sample study with $2n$ observations, the degrees of freedom for the t-test are $2(n - 1)$. For the corresponding paired t-test, the degrees of freedom are $(n - 1)$ since there are $n$ differences. This reduction in degrees of freedom means that a larger t-value will be required for statistical significance. Assuming the sample mean differences are the same, this requires the denominator of the paired t-test to be smaller than that of the two-sample test to attain the same level of statistical significance. In other words, if the within-pairs chance variation is not substantially smaller than that between pairs, the pairing is not only pointless, it may make matters worse by causing a substantial loss in degrees of freedom. Put differently, if the covariance term in expression (4.2) is very small then the estimated standard error in the denominator of the t-test will be about the same as that given by (4.1), but we will have (needlessly) sacrificed $(n-1)$ degrees of freedom. Accordingly, pairing (or, more generally, blocking) should be used where it is clear that within pair/block chance variation is likely to be substantially less than between pair/block variation.

In summary, a general recipe for experimental design of comparative studies is to hold constant all influential factors not directly involved in the comparison, block those that cannot be controlled over the whole study, and randomise everything else

### 4.2 Statistical Power Analysis

The number of experimental units is a central issue in the design of comparative studies. Before addressing this question, we examine the errors that may arise (incorrect decisions) when statistical significance tests are carried out. This leads us to a consideration of statistical power, which is a fundamentally important concept underlying the choice of the number of experimental units for a study.
**Errors in Significance Testing**

When carrying out a statistical significance test two types of error may arise, as shown in Table 4.2.1. If we reject the null hypothesis $H_0$ when it is true, we make a Type I error – the probability of this occurring is usually labelled $\alpha$. If we accept the null hypothesis when, in fact, it is false, we make a Type II error – the probability of this happening is usually labelled $\beta$.

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ True</th>
<th>$H_0$ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept $H_0$</td>
<td>correct decision</td>
<td>Incorrect acceptance</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Incorrect rejection</td>
<td>Type II ($\beta$)</td>
</tr>
</tbody>
</table>

Table 4.2.1: Two types of error

**Practical Implications**

Consider an acceptance sampling situation: a statistical test is carried out on samples from a batch of material to decide whether or not to accept the batch. Type I error would mean that a good batch (i.e., one that conformed to the required specifications) would be incorrectly rejected and, therefore, returned to the supplier. Type II error would mean that the customer would incorrectly accept bad (i.e., out of specification) material.

Consider a study to test a newly developed cream designed to treat acne – the cream is applied to one cheek, a placebo (a cream without any active ingredient) is applied to the other cheek of each of $n$ patients, with a view to carrying out a paired t-test to assess the efficacy of the treatment. Type I error would mean incorrectly rejecting the null hypothesis of no difference, on average, thus deciding that the new cream works when it does not (with consequent manufacturing, marketing etc. costs, before its lack of efficacy is discovered, not to mention the distress to the users!). Type II error would mean that the cream was effective, but the statistical test failed to detect this (thus leading to abandoning the cream with the loss not only of the research and development costs, but incurring, also, the opportunity cost of not selling an efficacious product, not to mention loss of the opportunity to make many teenagers happier!).
Why do these errors arise?

For the purposes of illustrating how these errors arise, we will return to the acceptance sampling problem discussed in Chapter 2: we sample n containers from a consignment, measure some property of the material, such as its purity, and carry out a one-sided test to choose between:

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

We will assume (for simplicity) that the standard deviation (\( \sigma \)) is known; this means that instead of a t-test, we carry out a Z-test, using the standard Normal distribution. The test statistic is:

\[ Z = \frac{\bar{X} - 90}{\sigma / \sqrt{n}} \]

and the null hypothesis is rejected if Z lies below the critical value \(-Z_c\) (where \(-Z_c\) is determined by the significance level, \( \alpha \), chosen for the test; it will be \(-1.645\) if a significance level of 0.05 is chosen), as shown in Figure 4.2.1.

![Figure 4.2.1: The test criterion for Z](image)

For the purposes of our discussion it is convenient to recast the decision rule in terms of the value of \( \bar{X} \) itself, rather than using Z. Thus, a value of Z less than \(-Z_c\) is equivalent to a value of \( \bar{X} \) less than:

\[ C = 90 - Z_c \frac{\sigma}{\sqrt{n}} \]

as shown in Figure 4.2.2.
Thus, Figure 4.2.3 shows schematically our decision rule: if $\bar{X} < C$ we reject $H_0$ (and the consignment); if $\bar{X} > C$ we accept $H_0$ (and the consignment).

### Type I Error

When the null hypothesis is true (i.e., specifically when $\mu = \mu_0 = 90$) the sampling distribution of the sample mean, $\bar{X}$, is as shown in Figure 4.2.2. Since $\sigma$ is known, the sampling distribution is Normal with standard error $\sigma/\sqrt{n}$; this means that $C = 90 - 1.645 \cdot \frac{\sigma}{\sqrt{n}}$, if $\alpha$ is chosen as 0.05.

The significance testing procedure requires us to choose the critical value $C$ such that the probability that $\bar{X}$ is less than $C$ is $\alpha$, when the null hypothesis is true. Thus, when we specify the significance level to be $\alpha$, we automatically accept a probability of $\alpha$ of incorrectly rejecting the null hypothesis when it is true, i.e., we specify the probability of Type I error. This is generally true: for all statistical tests, the choice of the significance level, $\alpha$, determines the probability of rejecting the null hypothesis when it is true.
**Type II Error**

Suppose now that the null hypothesis is false and specifically that \( \mu=88 \) rather than 90. The sampling distribution of the sample mean, \( \bar{X} \), will then be centred on 88, as shown in Figure 4.2.4.

The area under the sampling distribution curve to the right of \( C \) is the probability that we will incorrectly accept the null hypothesis, i.e., that we will conclude that that \( \mu=90 \) (or better) **when, in fact, it is 88**. This is the probability of Type II error and is labelled \( \beta \) in Figure 4.2.4. The probability of Type II error depends on what the true batch mean is: as the mean moves to the left, i.e., \( \mu<88 \), then the area to the right of \( C \) (i.e., \( \beta \)) will decrease – this is as we would expect, since the further the mean is from 90, the less likely we are to conclude incorrectly that it is \( \geq 90 \). As the mean moves to the right (closer to 90), \( \beta \) will increase. Thus, for example, if the batch mean, \( \mu \), falls on \( C \), there will be a 50\% chance of \( \bar{X} \) falling on either side of \( C \) and, therefore, a 50\% chance either of rejecting \( H_0 \) or of accepting it.

**Power Curves**

If \( \beta \) is the probability of accepting \( H_0 \) then (1-\( \beta \)) is the probability of rejecting it. When \( H_0 \) is false (1-\( \beta \)) is known as the ‘power’ of the test. A powerful test is one which has a high probability of rejecting \( H_0 \) when it is false.
If we plot \((1-\beta)\), the probability of rejecting \(H_0\), as a function of the true mean, \(\mu\), we get Figure 4.2.5 – this is called the power curve of the test: it illustrates how good the significance test is at detecting when the null hypothesis should be rejected. As the batch mean, \(\mu\), moves further from the null hypothesis value of 90, the power of the test increases – we are more likely to reject \(H_0\) (correctly).

Drawing the power curve is just a calculation exercise of the type discussed in Chapter 1. For given values of \(\mu\), \(\sigma\) and \(\alpha\), use of the standard Normal curve will give the areas to the right and left of \(C\). Then, choosing a range of \(\mu\) values will result in a range of \((1-\beta)\) values, and plotting \((1-\beta)\) against \(\mu\) gives the power curve\(^2\).

Power is a fundamental property of statistical tests. If a test does not have sufficient power (and this should be investigated before a study is carried out) it is questionable whether the study should be undertaken at all. This is especially true where humans (especially sick ones) or animals are involved. Altman and Bland [4] (two British medical statisticians) make the following comments in a note on such matters “The sample size of controlled trials is generally inadequate, with a consequent lack of power to detect real, and clinically worthwhile, differences in treatment. Freiman et al found that only 30% of a sample of 71 trials published in the New England Journal of Medicine in 1978-9 with \(p\)-value >0.1 were large enough to have a 90% chance of detecting even a 50% difference in the effectiveness of the treatments being compared, and they found no improvement in a similar sample of trials published in 1988.” They refer to a study “carried out on a sample of only 100 despite a reported calculation that suggested 1800 patients were needed. This trial had only a 5% chance of getting a statistically significant result if the stated clinically worthwhile treatment difference truly existed”. The need for adequate sample sizes (our next topic) is obvious.

\(^2\) If we plot \(\beta\), the probability of accepting \(H_0\), instead of \((1-\beta)\), as a function of the true mean \(\mu\), we get a curve called the operating characteristic (OC) curve of the test: it illustrates how bad the significance test is at detecting when the null hypothesis should be rejected. As the batch mean gets closer to \(\mu_0=90\), the probability of (incorrectly) accepting that \(\mu=90\) increases.

The two curves give identical information, obviously. Power curves tend to be used in the statistical literature, OC curves are more commonly used in the engineering quality control literature.
4.3 Determining Sample Size

A major problem facing any researcher about to engage in a study is how many observations are required to get good results. Making too few measurements can jeopardise the objectives of the study; making too many is a waste of resources. This question is essentially statistical in nature: it involves balancing the risk of failing to detect important effects when they are present, with the risk of falsely concluding that effects are present when, in fact, they are not. A simple approach to deciding on study sizes is outlined below.

The Nature of the Problem

To provide a focus for the discussion we will consider the problem of studying the effect on process yield of two variants of a production process. These might be two machine speeds, as in Example 5 of Chapter 2, where a TCD research student wanted to optimise the performance of a spheroniser used in producing pellets for pharmaceutical capsules.

Because of the chance variation inherent in the manufacturing process we could:

• incorrectly decide that the two process settings result in different (long-run) yields when, in fact, they do not. This could lead us to modify the process (possibly incurring increased production costs) to no avail;
• incorrectly decide that the two settings give the same yield when, in fact, one is better. This could lead us to abandon a useful process modification with consequent opportunity costs.

We can progressively reduce the effects of chance variation by averaging, that is, by taking increasingly larger sample sizes. There is, however, a cost involved in doing this. In selecting a study size we must balance the cost of a large study against the costs that would follow on making either of the incorrect decisions outlined above.

*Specifying the Parameters of the Problem*

Since the presence of chance variation lies at the heart of the problem, some estimate of its magnitude is required before we can begin to think about the number of measurements required. We need, therefore, to specify the size of the standard deviation ($\sigma$) which we consider describes the expected chance variation. Given this, there are three other aspects of the problem which require consideration:

• since a statistical test will be carried out, the specification of the null hypothesis will have determined if a one or a two-sided test is required. We must then decide on the significance level ($\alpha$) that will be used: this defines the probability of incorrectly deciding a long-run difference exists when, in fact, the two process variants produce the same long-run yields.

• while, in principle, we would like to detect any difference between the two process variants, it is clear that it would require a very large sample size to detect a very small yield difference. We must, therefore, specify what difference ($\Delta=\mu_1-\mu_2$) we consider it would be important to detect;

• having specified $\Delta$, we must specify the risk ($\beta$) we are willing to take of failing to detect a difference of this size.

*Some Comments*

• The likely magnitude of the chance variation affecting the system under study, as measured by the standard deviation $\sigma$, is fundamental. A large standard deviation indicates large chance variation and, consequently, makes a large sample size necessary to detect small to moderate shifts in yield. The importance of knowing $\sigma$ at the planning stage of a sampling exercise or experimental study (or at least having a good estimate of its likely order of magnitude) cannot be overestimated. This points to the need to retain such historical information in an accessible form.
The choice of a significance level is usually a subjective one, determined by whether one considers a one in twenty or one in a hundred chance as small (the conventional values used for significance level are 0.05 and 0.01). However, since the significance level determines the probability of incorrectly deciding that there is a change in long-run yield where none is present, the consequences of such a decision may help in choosing an appropriate value.

Reducing the significance level automatically leads to larger sample sizes, so a balance needs to be struck between the cost of the study and the risk of incorrect rejection of the null hypothesis (bearing in mind the consequences of such an incorrect decision). A series of calculations, using different significant levels, will show the sensitivity of the required sample size, as the significance level varies.

The size of the shift in yield that we want to detect ($\Delta=\mu_1-\mu_2$) and the risk ($\beta$) we are willing to take of missing such a shift are obviously interconnected. While in principle we might like to detect a shift of 0.1%, in practice we would be unlikely to worry about missing such a shift. Accordingly, we can tolerate a high risk of failing to detect such a difference. On the other hand, a yield shift of 10% would, almost certainly, have serious financial implications and we might be unwilling to tolerate even a 0.01 chance of failing to detect such a shift. Since it will not be obvious how to specify values for $\Delta$ and $\beta$, the best strategy is to do a series of calculations using a range of plausible values and then think about the implied sample sizes in the light of the costs and risks involved.

Using a Sample Size Table

Table ST-5 gives sample sizes for two-sample studies which will be analysed using two-sample t-tests. It requires us to specify the various parameters discussed above. It is based on the assumption that equal sample sizes will be used for the two groups. Its use will be introduced via an example.

Suppose that in Example 5 of Chapter 2 the research student planned to compare two speeds for the spheroniser and that, while he hoped the higher speed would produce a higher yield, he could not be sure that this would be the case. Accordingly, the study would be planned with the intention of carrying out a two-sided test on the resulting data.

The economics of the process are such that a yield shift of $\Delta=\mu_1-\mu_2=5$ (here the units are % points) would be important. The magnitude of the likely chance variation is uncertain but earlier experimental runs suggest a value of $\sigma=3$ as a reasonable estimate. The research student decides that a risk of $\alpha=0.05$ of deciding a yield shift occurs when it does not and a risk of $\beta=0.05$ of missing a
yield shift of 5 units (here, units are expressed as percentage points), would be reasonable.

To use Table ST-5 we calculate:

\[
D = \frac{\Delta}{\sigma} = \frac{5}{3} = 1.67
\]

this determines which row of the table we enter. The column entry is determined by \(\alpha=0.05\) (two-sided test) and \(\beta=0.05\) and we find (using \(D=1.70\); in many cases interpolation will be required):

\[
N = 11.
\]

Note that 11 is the sample size for each speed, so that the full study size is 22.

A Sample Size Formula

Note that Table ST-5 is based on the assumption that the analysis will involve a \(t\)-test of the sample means, i.e., although we specify \(\sigma\) for determining the sample size, \(\sigma\) will be estimated from the experimental data for the purposes of carrying out the resulting statistical test, once the experimental data are available. If the standard deviation that determines the chance variation in the system under study really were known, then the \(t\)-test would become a \(Z\)-test and the more complicated analysis underlying the Table ST-5 would reduce to a simple explicit formula for sample size.

\[
n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\Delta^2}
\]

where \(Z_\alpha\) is the Z value that is given by our significance level and \(Z_\beta\) is the Z value that controls the Type II error (it is the Z value that has an area \((1-\beta)\) to its left under the standard Normal curve).

For our spheroniser example \(\sigma=3\) and \(\Delta=5\). For a two-sided test with \(\alpha=0.05\) \(Z_\alpha=1.96\). If we require a power of \(1-\beta=0.95\) when \(\Delta=5\), or equivalently a Type II error probability of \(\beta=0.05\), then \(Z_\beta=1.645\). This is the Z value corresponding to \(\beta=0.05\) in the upper tail and a power \((1-\beta)\) below the cut-off value. The cut-off value of Z corresponds to a one-sided area of 0.05 since we want a probability (power) of 0.95 of rejecting the null hypothesis of no shift, if the shift is as large as \(\Delta\) in either direction. The sample size formula gives a sample size of 9.4 which would be rounded up to 10. Table ST-5 gives a slightly larger value (between 11 and 12), but this is to be expected as the extra uncertainty in not assuming that \(\sigma\) is known (the formula underlying the Table assumes that \(\sigma\) will
be estimated from the study data) means that a greater sample size is required. For larger sample sizes the difference will be smaller or negligible.

In practice, the apparent greater exactness associated with the table (using the t-distribution rather than the Z-curve in the test) is spurious, as we never know $\sigma$ exactly and our sample size calculations are only as good as our estimates of the likely chance variation present in the system under study. The uncertainty in such estimates is likely to have a much greater influence on the calculated sample size than any difference between the table and the formula for a given value of the standard deviation. This applies equally to computer package sample size facilities, which will be based on the same underlying analysis as Table ST-5.

Inspection of the sample size formula brings out the issues that determine the sample size. Thus, if the standard deviation that describes chance variation in the system is increased by 33% from 3 to 4, the sample size is multiplied by $(4/3)^2=1.78$, i.e., the sample size increases to 20 (in each group).

Similarly, if the standard deviation were only 1 then the sample size would drop by a factor of 9.

If we wish to reduce the risks of error, we require Z values which are further out in the tails. Thus, choosing $\alpha=\beta=0.01$ instead of 0.05 would require $Z_\alpha=2.58$ and $Z_\beta=2.33$, so that the term $(Z_\alpha+Z_\beta)^2$ which was 13 for $\alpha=\beta=0.05$ becomes 24. This means that the sample size (in each group) increases by a factor of 24/13, resulting in an increase of 85% in the overall study size.

Since $\Delta$ is in the denominator, the smaller it is set, the larger the required sample size. Reducing it from 5 to 2 will multiply the sample size by a factor of $(5/2)^2=6.25$.

All of these calculations correspond to what we would expect intuitively: the larger the uncertainty, or the smaller the risks, or the smaller the difference we wish to detect, the greater the sample size that will be required. The formula simply quantifies the influence of each of these factors. Note that the important influence of the standard deviation on sample size calculations underlines the importance of maintaining good records regarding chance variation in different situations: this is key information in planning future studies.

**Power Calculations**

If, instead of finding an appropriate sample size, we want to investigate the implications of a given sample size (determined, perhaps, by available resources), we can carry out a power calculation. To do so, we simply re-cast the sample size formula to express the $Z_\beta$ value (which determines the power of the test) as a function of the other parameters.
The sample size formula:
\[
n = \frac{(Z_{\alpha} + Z_{\beta})^2 2\sigma^2}{\Delta^2}
\]
Can be re-cast to give:
\[
Z_{\beta} = \frac{n\Delta^2}{2\sigma^2} - Z_{\alpha} = \frac{\Delta}{\sigma} \sqrt{\frac{n}{2}} - Z_{\alpha}
\]
Recall that \(Z_{\beta}\) is the Z value that has an area \((1-\beta)\) to its left under the standard Normal curve, so that the power is \(P(Z< Z_{\beta})\).

For example, suppose we could only afford six runs of our speroniser at each speed, instead of the 10 indicated by the sample size formula; what would be the implications for the power of the test?

\[
Z_{\beta} = \frac{\Delta}{\sigma} \sqrt{\frac{n}{2}} - Z_{\alpha}
\]

\[
Z_{\beta} = \frac{5}{3} \sqrt{\frac{6}{2}} - 1.96 = 0.927
\]

\(P(Z<0.93) = 0.82\)

If we ran the study with a total sample size of 12 (instead of 20) then the probability of detecting a shift of \(\Delta=5\) would be only 0.82, instead of the desired power of 0.95.

For a fixed sample size, \(n\), a series of such calculations for different potential shifts \(\Delta=\mu_1-\mu_2\) would allow us to draw a power curve. Figure 4.3.1 shows three such curves for fixed sample sizes of \(n=6, 11, 20\) (in each group) and a significance level of \(\alpha=0.05\).

The curves show that, for any given sample size, the probability of rejecting the null hypothesis (i.e., the probability of detecting a real difference \(\Delta\)) increases as the size of the difference, \(\Delta\), increases. This is intuitively obvious, but the curves quantify the importance of the magnitude of the shift, \(\Delta\), which we are trying to detect.
The calculated power values for a shift of \( \Delta=5 \) are 0.74, 0.96 and 0.999 for sample sizes of 6, 11 and 20 in each group. The value of 0.74 is smaller than our calculated value of 0.82 as our value assumes \( \sigma=3 \) is known (and therefore contains no uncertainty), whereas the value calculated by Minitab assumes that when the t-test is carried out after the study the value of \( \sigma \) will be estimated from the sample data (which is subject to chance variation) and this implies extra uncertainty and lower power.

**Single sample studies**

The issues that arise in determining sample sizes for studies that involve only a single sample mean (e.g., the method validation study of Example 1 in Chapter 2, or the laboratory comparison study (Example 4), or the self-paired platelet aggregation study (Exercise 2.2.1)) are exactly the same as discussed above. They differ only in that in specifying the problem the difference of interest \( (\Delta=\mu-\mu_0) \) relates to the difference between a population mean \( \mu \) and some fixed constant, \( \mu_0 \) rather than the difference between two population means.

Consider the method validation study, where it was asked if the laboratory system was unbiased, i.e., was it varying around the certified value of \( \mu_0=27.1 \)? In determining an appropriate sample size we must specify the minimum difference from 27.1 (i.e., the bias) that we need to detect with a high probability (power). Suppose that the method development studies suggested that the day-to-day standard deviation for repeated measurements would be expected to be
between 0.3 and 0.5 and that it was required that there should be a power of
0.95 (i.e., a probability of Type II error of 0.05) if the bias was as much as
0.5 units (percentage fat) for a two-tail test using a significance level of \( \alpha = 0.05 \). For
\( D = \Delta / \sigma = 0.5 / 0.3 = 1.67 \) Table ST-5 (one-sample studies) gives a sample size of 7-8.
For \( D = \Delta / \sigma = 0.5 / 0.5 = 1 \) the sample size is 16.

The corresponding formula, based on the assumption that \( \sigma \) is known is given
below; note that the only difference from that for two-sample studies is the
absence of a 2 in the numerator – corresponding to the fact that we are
cconcerned with only one sample mean, not two\(^3\).

\[
n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\Delta^2}
\]

The term \( (Z_\alpha + Z_\beta)^2 \) is 13 for \( \alpha = \beta = 0.05 \), so, for \( \Delta = 0.5 \), the formula gives \( n = 4.7 \)
which is rounded up to 5 for \( \sigma = 0.3 \) and \( n = 13 \) for \( \sigma = 0.5 \). For studies requiring
larger sample sizes, the resulting sample sizes from the Table and the formula
would be expected to be closer. Re-arranging the formula to put \( Z_\beta \) on the left-
hand side allows power calculations to be carried out, as before.

**Paired Studies**

Paired t-tests are one-sample t-tests carried out on a set of differences.
Accordingly, the required number of **pairs** for a study can be determined from
Table ST-5 (one-sample studies) or the formula just discussed, provided we have
an estimate of the standard deviation of **differences**. Take the platelet study as
an example (Exercise 2.2.1). Suppose this had been a pilot study and a further
study which would have a power to detect a mean difference of \( \Delta = 5 \) was required
to have a power of 0.95 for a two-tail test using a significance level of \( \alpha = 0.05 \).
What number of self-pairs needs to be recruited?

The numbers specified imply a value of \( D = \Delta / \sigma = 5 / 8 = 0.625 \) (the observed
standard deviation of differences was 7.98). This gives a sample size of between
33 and 39. As always for sample size calculations, it makes sense to play
around with the assumptions before reaching a final view on the sample size that
will be used for the study.

---

\(^3\) The standard error of a single sample mean is \( \sqrt{\frac{\sigma^2}{n}} \) while the standard error of the difference
between two independent sample means is \( \sqrt{\frac{2\sigma^2}{n}} \), where each mean is based on \( n \) observations.
Sample Size for Counted Data

Consider a situation where two independent sample proportions \(p_1\) and \(p_2\) are to be compared using a Z-test. The null hypothesis is that the corresponding population proportions \(\pi_1\) and \(\pi_2\) are the same, i.e., \(\pi_1=\pi_2=\pi\). In most cases\(^4\) equal group sizes will be decided upon (for a given total sample size, \(N\), the Z test will be most powerful if \(n_1=n_2=N/2\)) and the following (approximate) formula is widely recommended for arriving at an appropriate value for \(n\), the number of subjects/experimental units in each group. Note that, apart from the part of the numerator between square brackets, the formula is identical to that for two-sample t-tests as discussed earlier (page 25).

\[
n = \frac{(Z_\alpha + Z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\Delta^2}
\]

where:

- \(Z_\alpha\) is the standard Normal critical value for a significance level of \(\alpha\). For a two-sided test with a significance level of \(\alpha=0.05\) this is 1.96, for a one-sided test it is 1.645.

- \(Z_\beta\) is the cut-off value in the right-hand tail of the standard Normal curve which gives an area to its left of \((1-\beta)\), corresponding to a power of \((1-\beta)\), or equivalently, a probability \(\beta\) of Type II error, (i.e., a probability \(\beta\) of accepting the null hypothesis when, in fact, it should be rejected). One-sided cut-off values are used here, since we want a power of \((1-\beta)\) irrespective of whether the difference \(\Delta\) is positive or negative. Thus, for \(\beta=0.05\), corresponding to a power of \((1-\beta)=0.95\) of rejecting \(H_0\), \(Z_\beta=1.645\) and for \(\beta=0.20\), corresponding to a power of 0.80, it is \(Z_\beta=0.84\).

- \(\Delta=\pi_1-\pi_2\) is the difference we are interested in detecting, with a power of \((1-\beta)\).

The specifications here correspond directly to those required for determining the sample size for two-sample t-tests, as discussed earlier.

Note that for continuous data we did not need to specify the population means \(\mu_1\) and \(\mu_2\), only the difference, \(\Delta=\mu_1-\mu_2\), we wished to detect. We did, however, have to specify the (assumed) common standard deviation, \(\sigma\). The standard error for the difference between two independent proportions depends on the

----

\(^4\) In clinical trials where there is strong prior evidence that one treatment is superior to the other, it may be decided on ethical grounds that more patients should be allocated to this treatment.
proportions themselves (see pages 9, 10 of Chapter 3); this is why we need to specify \( \pi_1 \) and \( \pi_2 \) in the formula. Two examples illustrate the use of this formula; the calculation of power for given sample size is then discussed.

**Anturan Trial**

Pocock [6] illustrates a discussion of sample size calculation by referring to the Anturan Reinfarction Trial (1978) [10]. This was a randomised double blind trial comparing the prescription of anturan versus a placebo to patients who have had a heart attack. Survival for one year after start of treatment was used as the simple criterion of success. They chose \( \pi_1=0.95 \) and \( \pi_2=0.90 \), \( \alpha=0.05 \) (two-sided test) and \( \beta=0.10 \). The required sample size, based on these assumptions, is therefore:

\[
 n = \frac{(Z_\alpha + Z_\beta)^2 [\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\Delta^2}
\]

\[
 n = \frac{(1.96+1.28)^2 [0.95(1-0.95) + 0.90(1-0.90)]}{0.05^2} = 578
\]

Pocock reports that the authors increased this figure to 750 per treatment to allow for dropouts and to increase power slightly (see later).

**Salk Polio Vaccine Trial**

Snedecor and Cochran (in an exercise in [7]) state that in planning the 1954 Salk polio vaccine trial sample size was critical, since it was unlikely that the trial could be repeated. Various estimates of sample size were made, in one of which it was assumed that the probability that an unprotected child would contract paralytic polio was 0.00030 (30 per 100,000). If the vaccine was 50\% effective, i.e., reduced this to 0.00015 (15 per 100,000), it was desired that there would be a 90\% chance of finding a significant difference using a two-tailed test with a significance level of \( \alpha=0.05 \). Thus, inserting \( \pi_1=0.00030 \), \( \pi_2=0.00015 \), \( Z_\alpha=1.96 \) and \( Z_\beta=1.28 \) into the formula:

\[
 n = \frac{(Z_\alpha + Z_\beta)^2 [\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\Delta^2}
\]

we find:

\[
 n = \frac{(1.96+1.28)^2 [0.00030(1-0.00030) + 0.00015(1-0.00015)]}{0.00015^2} = 210,096
\]
which means that about 200,000 children would be required for each of the two study groups, vaccinated and control.\(^5\)

**Power Calculations**

The sample size formula can be inverted to give an estimate of the power that will result from specifying \(\pi_1, \pi_2, \alpha,\) and the sample size \(n.\) This is useful especially where the calculated sample size is considered impractical and we wish to assess the implications of using a smaller sample. The sample size formula:

\[
n = \frac{(Z_\alpha + Z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\Delta^2}
\]

Is re-cast to give:

\[
Z_\beta = \frac{\Delta\sqrt{n}}{\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}} - Z_\alpha
\]

where \(\Delta = \pi_1 - \pi_2\) is the difference we are interested in detecting, with an adequate power. Since \(\Delta\) is not squared (as in the sample size formula), \(\pi_1\) should be the larger of the two proportions.

Recall that when we specified the cut-off value \(Z_\beta\) for the required power in our sample size formula, we said that \(Z_\beta\) was the standard Normal value that has an area \((1-\beta)\) to its left. Accordingly, the estimate of the power, here, is given by:

\[
P(Z<Z_\beta)
\]

which may be obtained from a standard Normal table.

---

\(^5\) A note from Wikipedia:

Salk’s vaccine was used in a test called the Francis Field Trial, led by Thomas Francis; the largest medical experiment in history. The test began with some 4,000 children at Franklin Sherman Elementary School in McLean, Virginia, and would eventually involve 1.8 million children, in 44 states from Maine to California. By the conclusion of the study, roughly 440,000 received one or more injections of the vaccine, about 210,000 children received a placebo, consisting of harmless culture media, and 1.2 million children received no vaccination and served as a control group, who would then be observed to see if any contracted polio. The results of the field trial were announced April 12, 1955 (the tenth anniversary of the death of Franklin Roosevelt). The Salk vaccine had been 60 - 70% effective against PV1 (poliovirus type 1), over 90% effective against PV2 and PV3, and 94% effective against the development of bulbar polio.
**Anturan Trial - revisited**

A randomised double blind trial comparing the prescription of anturan versus a placebo to patients who have had a heart attack was discussed above. The authors chose $\pi_1=0.95$ and $\pi_2=0.90$, $\alpha=0.05$ and $\beta=0.10$, corresponding to a power of 0.90. The required sample size, based on these assumptions, was 578 in each group. Pocock reported that the authors increased this figure to 750 per treatment to allow for dropouts and to increase power slightly. We can use the inverted formula to investigate the implications of the sample size increase.

The inverted formula is:

$$Z_\beta = \frac{\Delta \sqrt{n}}{\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}} - Z_\alpha$$

which gives:

$$Z_\beta = \frac{0.05 \sqrt{750}}{\sqrt{0.95(1-0.95) + 0.90(1-0.90)}} = 1.96 = 1.76$$

and a power of:

$$P(Z < Z_\beta) = P(Z < 1.76) = 0.96.$$ 

So, by increasing the sample size from 578 per group to 750 per group, they increased the power from 0.90 to 0.96. The increase of 30% in the sample size was probably decided upon more to allow for dropouts than to increase the power.

**Salk Polio Vaccine Trial - revisited**

The Salk vaccine trial clearly needed unusually high numbers of subjects because both the individual proportions being compared, and their difference, were very small. Suppose that instead of carrying out the sample size calculation, as described above, the trial planners had picked out of the air an ‘obviously large’ number of subjects, say (i) 50,000 or (ii) 100,000 for each group. What would have been the implications for the likely success of the trial, if the initial assumptions, that the probability that an unprotected child would contract paralytic polio was 0.00030 and that the vaccine might reduce this to 0.00015, were correct?

The expression for $Z_\beta$ is:

$$Z_\beta = \frac{\Delta \sqrt{n}}{\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}} - Z_\alpha$$
which for \( n = 50,000 \) gives:

\[
Z_\beta = \frac{0.00015\sqrt{50000}}{\sqrt{0.00030(1-0.00030) + 0.00015(1-0.00015)}} - 1.96 = -0.379
\]

and a power of:

\[
P(Z < Z_\beta) = P(Z < -0.379) = 0.352.
\]

Thus, with a sample size of **only** 50,000 for each group, there would be a 65% chance of Type II error, i.e., of failing to decide that the vaccine works if, in fact, it reduced the probability of contracting polio by 50%.

The equivalent calculation for \( n = 100,000 \) gives a power of 0.61, i.e., a 39% chance of failing to detect the effectiveness of the vaccine. These calculations show that even a very large sample size can be too small!

**Discussion**

The sample size formula on which our discussions have been based is a large sample approximation. If the numbers it produces are small, it may be worthwhile looking at other sources of information on sample size determination (both Machin and Campbell [8] and Fleiss [9] provide extensive tables and references to other work in this area).

While different statistical approaches will produce somewhat different sample sizes for a study, the technical differences are generally much less important than two practical considerations: the specification of the entries to the formula and the availability of resources. The formula requires us to specify \( \pi_1 \) and \( \pi_2 \) — if we knew these there would be no need for a study! Manipulation of the sample size formula shows that the resulting number \( n \), for each group, will be sensitive to the guesses used for these quantities and, in particular, to their difference \( \Delta = \pi_1 - \pi_2 \), which is squared in the denominator. The choices of the significance level, \( \alpha \), and the power, \( (1-\beta) \), are also quite arbitrary; while we might start out optimistically requiring a high power for a small \( \Delta = \pi_1 - \pi_2 \), it is quite likely that in many cases the resulting sample sizes will drive us towards more ‘realistic’ (or just plainly pragmatic!) values. The lack of available resources is, in practice, likely to be a major consideration, overshadowing the technical statistical considerations. This, however, does not diminish the value of the statistical analysis — on the contrary. The statistical analysis provides concrete answers to “what if?” questions and helps clarify the implications of the guesswork necessarily involved in planning a study where vital information is either absent or, at best, vague. In the extreme, when the power calculation shows that the available resources are insufficient to provide a reasonable chance of a
successful study, then, in general, the study should be abandoned. This would, very often, be mandatory for clinical trials, on ethical grounds.

Exercises

4.3.1 For the Timolol study of Chapter 3, suppose that the researchers had guessed that the control group would produce an angina-free rate of 10% and, if the drug were successful, that this would increase to (i) 20% or (ii) 25%. They intended carrying out a two-tailed test using a significance level of 0.05. If they required power values of (i) 90% or (ii) 95%, use the sample size formula to calculate the required sample sizes for all four sets of assumptions.

4.4 The Nature of Replicates

In a simple comparative study we are interested in comparing the responses of two groups A and B. If we had only a single observation from each group then we would ask if \((Y_A - Y_B)\) represented a systematic difference or the purely chance variation that is inevitably present in situations where statistical analysis is required. We cannot answer this question – there is no basis for distinguishing between systematic and chance variation. The way out of this is to use more than a single observation from each group. This allows separate estimates of the systematic effect \((\mu_A - \mu_B)\) and the size of the chance variation, as measured by \(\sigma\).

We typically observe several values (replicates) from each group and compare \(\bar{Y}_A\) and \(\bar{Y}_B\). We do this knowing that averaging reduces chance variability; we expect the chance component in an average to be smaller than that likely to be present in any one \(Y\) value and, hence, \((\bar{Y}_A - \bar{Y}_B)\) gives us a better estimate of \((\mu_A - \mu_B)\) than does \((Y_A - Y_B))\). In obtaining these replicates, though, care must be taken that they reflect properly the chance variation involved in the difference between single values from each group, \((Y_A - Y_B))\): this is the variation we wish to average out of the comparison. In many cases, what may appear to be replicates do not fully reflect the chance variation over which it is desired to average. A schematic example will illustrate the problem.

Suppose a researcher sets out to discover whether boys and girls at a given age perform differently on an academic performance test and does this by randomly selecting \(n\) pupils from a randomly selected class in a randomly selected large boys' school (i.e., one with several classes at the relevant age group) and \(n\) girls similarly. A model for the individual responses would be:

\[
\text{Performance (Y)} = \text{Gender effect} + \text{School effect} + \text{Class effect} + \text{Pupil effect}
\]

Some schools are more examination-focused than others and their pupils are more likely to perform better in examinations. Similarly, some classes within schools perform better in examinations than others, perhaps because they are
taught by better teachers, or because there is a stronger work ethic in the class. Pupils differ in their abilities and commitment to study.

If the response of one boy is $Y_B$ and that for one girl is $Y_G$ then the comparison will be between $\bar{Y}_B$ and $\bar{Y}_G$. The sample average $\bar{Y}_B$ is not, however, an appropriate estimate of the average response of all boys: it is an estimate of the average response of the boys in this class, in this school. The sample difference $\bar{Y}_B - \bar{Y}_G$ estimates the difference between the mean responses of the two classes in the study. If the boys’ class was a good one in a good school, and the girls’ class a poor one in a poor school (here ‘good’ and ‘poor’ refer to examination performance – not to be confused with real school quality!), then the difference $\bar{Y}_B - \bar{Y}_G$ would be a very poor estimate of the difference of interest, $\mu_B - \mu_G$, i.e., the mean difference between all boys and all girls. The problem here is that the replication averaged over the units at the bottom of a hierarchy.

![Figure 4.4.1: The hierarchical structure for school populations](image)

Figure 4.4.1 shows that pupils are nested under classes, which are in turn nested under schools. No averaging took place over either classes or schools. In order to average out the effects of different schools and different classes within schools, the sampling would need to take account of the nested structure and ensure that pupils from different classes within different schools were included in the study. Hierarchical structures are commonly encountered and account needs to be taken of the nesting when selecting samples for studies; otherwise, the kind of bias illustrated above can arise. A simulation study of the effect of inappropriate replication in a measurement context is outlined below.

The hypothetical school study talked of “randomly selecting n pupils from a randomly selected class in a randomly selected large boys’ school”. In practice, very many research samples are convenience samples (the researcher has a friend who is a teacher in a school…) rather than random samples from the relevant population. The effects of such non-random selection on the inferences drawn from a study are probably indeterminable (we would need full knowledge to assess the biases introduced). The blocking principle can help to some extent. If boys and girls from the same classes can be compared and this replicated over a range of schools of different types, with similar results, then we
can have more confidence in the results (though, obviously, we can say nothing about single sex schools). Where this cannot be done, then replication of the entire study will be required before conclusions can be accepted with confidence; if a number of non-randomised studies give broadly the same conclusions, it builds confidence in those conclusions. However, the experience of use of historical controls in clinical trials (see page 7) should make us aware of the dangers involved. Of course, study replication is always desirable to draw broader scientific conclusions, but it becomes even more important where random samples cannot be selected.

**A simulation study of a laboratory comparison**

In Example 2 of Section 1.7 we discussed the use of control charts in a laboratory context. Every time a batch of test materials is measured several measurements are made on control samples to monitor the stability of the measurement system. A batch of tests is referred to as an ‘analytical run’ or simply a ‘run’. Within a run everything is held as constant as possible – the same analyst carries out the measurements using the same equipment, same glassware, same reagents, over a short time period. Different runs, however, may take place on different days and involve different analysts, using different equipment, different glassware and different reagents, even though the same control material is being measured. The structure is, therefore, nested (with two levels) in the same way as the school example discussed above. The structure is shown in Figure 4.4.2.

![Figure 4.4.2: The structure of the control chart data](image)

Although the same material is measured all the time, experience shows that differences between individual test results will be greater if they come from different runs, than if they are obtained in the same run. This is allowed for when drawing control charts such as Figure 1.7.2 by averaging the within-run replicates and using the between-run standard deviation as the basis for calculating the control limits, see Mullins [5], Chapter 2. This nested structure can also have an important impact on experimental studies, as illustrated by a simulation study similar to the laboratory comparison of Example 4 of Chapter 2; for full details of the simulation and a detailed discussion of the issues involved, see Mullins [5], Chapter 4.
Suppose that the QC Director of a pharmaceutical company asks for a special inter-laboratory study to assess the magnitude, if any, of the relative bias between two of the group’s laboratories. Suppose, also, that it has been decided that each laboratory will assay the 20 most recent batches of the product that is supplied by one plant to the other and that the results will be compared using a paired t-test. How should the 20 measurements be made? Two possible study protocols suggest themselves. All \( n=20 \) test materials might be measured within one analytical run in each laboratory. Alternatively, each material might appear in a different analytical run. The first protocol would be the typical approach taken, since it would be preferable in terms of completing the study quickly and would appear to imply a more efficient use of the analysts’ time. However, statistical analysis and the results of a simulation study outlined below show that this approach is intrinsically flawed and that the second protocol, despite being logistically more difficult and perhaps more costly, is preferable.

A simulation study was carried out, based on the characteristics of real batches of pharmaceutical product, as follows. Twenty batches were ‘measured’ in each of two laboratories, that were assumed to have no relative bias. This was done twice, once with all measurements in the same analytical run and once with each batch measured in a different run, in each laboratory. This process was repeated 100,000 times. For each of the 200,000 simulations a paired t-test was carried out to compare the two laboratory means. Since the null hypothesis of no relative bias was true in all cases (by virtue of the design of the simulation) then, because a significance level of 5% was used for all the tests, the null hypothesis would be expected to be rejected 5% of the time, as discussed in Section 4.2. The study results are shown in Table 4.4.1.

<table>
<thead>
<tr>
<th>Protocol 1</th>
<th>Protocol 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-run design</td>
<td>20-run design</td>
</tr>
<tr>
<td>Signif. Results</td>
<td>57,749</td>
</tr>
<tr>
<td>Error rates</td>
<td>58%</td>
</tr>
</tbody>
</table>

Table 4.4.1: Results of the simulation study

Where the measurements were carried out in different runs, the simulation study performs exactly as expected – in other words, if a single real study were carried out in this manner, the statistical test would have the expected characteristics (i.e., the probability of incorrectly rejecting the null hypothesis would be \( \alpha=0.05 \)).
Table 4.4.1 shows that a real study that included all batches in single analytical runs in each laboratory would have a very high chance (58%) of falsely concluding that a bias existed, where in fact the laboratories were unbiased relative to each other.

The problem here is directly analogous to the schools example – if all pupils are selected from just two classes (one class from a boys’ school and one from a girls’ school) then the test compares the average scores for these two classes, rather than (as desired) for all boys and all girls in the population under study. In the measurement example, when all measurements are from only one analytical run the calculated standard deviation used in testing reflects only within-run chance variation. The test compares the two particular runs, one in each laboratory (but the denominator of the t-test takes no account of the chance variation from run to run), rather than the average performance of the two laboratories, which is what is of interest. Since the chance variation between runs may be substantial (compared to the within-run chance variation) the tests will very often be statistically significant (inappropriately, given the built in assumption of lack of bias).

The lesson is: if you have a nested structure be sure your data collection and statistical analysis take appropriate account of this structure. Chapters 4 and 8 of Mullins [5] discuss how to do this in a laboratory/measurement context, but the ideas are easily generalised.

References


4.3.1

The following sample sizes (for each group) were obtained using the sample size formula (page 27).

<table>
<thead>
<tr>
<th>Power</th>
<th>Increased proportion $\pi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>0.95</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Slightly larger values (shown below) were obtained using the sample size option in Minitab – the differences are negligible when we take into account the uncertainties in specifying the parameters of the problem.

<table>
<thead>
<tr>
<th>Power</th>
<th>Increased proportion $\pi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>0.95</td>
<td>0.25</td>
</tr>
</tbody>
</table>