If we compare the groups who actually received the treatment they were intended to have, we find that the 2-year mortality on surgery (4.1%, 15 deaths out of 369 patients) compares favorably with that on medical treatment (8.4%, 27 deaths out of 323 patients). Applying a χ^2 test gives $P = 0.018$, suggesting that the difference has not arisen by chance.

On the other hand if we compare the groups as they were formed by randomization, then surgical treatment does not appear to do so well.

$$\text{Mortality for those allocated to surgical treatment is } \frac{6+15}{26+369} = 5.3\%$$

$$\text{Mortality for those allocated to medical treatment is } \frac{2+27}{48+323} = 7.8\%$$

Applying a χ^2 test gives $P = 0.16$, so chance now becomes a plausible explanation of the observed difference between the treatments.

A clue about what might be going on comes from the high mortality (23.1%) attached to patients randomized to surgery but who received medical treatment. It is a strong possibility that these patients were sufficiently ill that the surgeon did not believe that an operation was in their best interests — their chances of surviving it were too low to proceed — so they were given the medical option. Similar patients allocated to medical treatment would simply proceed with their allotted treatment. Hence, in the comparison between those who received the treatments to which they were allocated, a group of high-risk patients has been removed from the surgery group. It follows that this comparison is biased and potentially misleading.

A safer option is to compare the groups as formed by randomization, notwithstanding the fact that some of those allocated to surgery got medicine and vice versa. A rationale for this, both mathematical and clinical, is given in the next section.

10.3 Analysis by Intention-to-Treat

10.3.1 Informal Description

When faced with all manner of deviations from the protocol, not just the preceding descriptions, there is no entirely satisfactory solution. However, there is a guiding rule that should only be broken in the full knowledge of the potential consequences. The principle is that you should compare the groups as they were formed by randomization, regardless of what has subsequently happened to the patients. You are not necessarily analyzing the patients according to how they were treated, but according to how you

intended to treat them. Because of this, the principle is often referred to as the dictum of *analysis by intention-to-treat*.

The justification of the approach is that any other way of grouping the patients cannot be guaranteed to have been comparable at the start of the study. Admittedly, you will often compare groups that are "contaminated" in some way — in the example in Subsection 10.2.2, each allocated treatment group contains a few patients who are given the other treatment.

Although there are no fully satisfactory solutions, comparing apparently strange groups can often be viewed in ways that make the comparison seem less strange. In the example in Subsection 10.2.1, if surgery turned out to be the treatment of choice in the future, then patients found to have inoperable tumors only at surgery would continue to arise. We would need to decide how to treat such patients; in the MRC trial they were offered radiotherapy. In other words, there will never be a time in clinical practice when surgery is appropriate for all patients, and the provision of alternatives for those whose tumors cannot be excised is implicit in the allocation of surgery. Consequently, a more realistic approach is to view the trial as comparing "the policy of offering surgery but accepting that some will need to resort to radiotherapy with the policy of offering radiotherapy."

The comparison of the groups as randomized is then an appropriate way to compare these policies.

10.3.2 Theoretical Description

A more mathematical description can be given. It is rather simplistic but illustrates the issues quite clearly. For definiteness, we will use the example of the surgery vs. radiotherapy trial, although the principle applies more widely.

The outcome, survival time, X, is supposed to have different means for different types of patients and these are listed below.

1. Operable tumors allocated to radiotherapy have mean, say, $\mu_O + \tau_R = \mathrm{E}(X \mid O, R)$
2. Operable tumors allocated to surgery have mean $\mu_O + \tau_S = \mathrm{E}(X \mid O, S)$
3. Inoperable tumors allocated to radiotherapy have mean $\mu_I + \tau_R = \mathrm{E}(X \mid I, R)$
4. Inoperable tumors allocated to surgery have mean $\mu_I + \tau_R = \mathrm{E}(X \mid I, S)$

The terms μ_O, μ_I represent the mean survival time for the patients with, respectively, operable and inoperable tumors (and it is likely that $\mu_O > \mu_I$). The change in mean survival time that surgery and radiotherapy confer are, respectively, τ_S, τ_R, and the aim of the trial is to estimate $\tau_R - \tau_S$. Note that

the mean in group 4 contains the term τ_R rather than τ_S because patients in the surgery group with inoperable tumors receive radiotherapy.

Suppose that the proportion of patients who have inoperable tumors is λ, which, because of randomization, we expect to be the same in the two groups. We write $\Pr(I) = \lambda$ for the probability that a patient has an inoperable tumor, and $\Pr(O)$ for the complementary probability.

The mean survival time in the radiotherapy group is

$$\begin{aligned} E(X \mid R) &= \Pr(I)E(X \mid I, R) + \Pr(O)E(X \mid O, R) \\ &= \lambda(\mu_I + \tau_R) + (1-\lambda)(\mu_O + \tau_R) \\ &= \lambda\mu_I + (1-\lambda)\mu_O + \tau_R \end{aligned}$$

If the radiotherapy group were compared with the group of patients who actually received surgery, then we would be comparing this mean with a group that has mean $\mu_O + \tau_S$. Consequently, the difference in treatment means would have expectation:

$$\begin{aligned} E(X \mid R) - E(X \mid O, S) &= \lambda\mu_I + (1-\lambda)\mu_O + \tau_R - (\mu_O + \tau_S) \\ &= \lambda(\mu_I - \mu_O) + \tau_R - \tau_S \end{aligned}$$

As we do not expect μ_O to equal μ_I, the comparison of these groups provides a biased estimator. Moreover, the presence of the μ parameters means that we have little idea how this quantity relates to the quantity of interest $\tau_R - \tau_S$.

The mean survival time for the group allocated to surgery, regardless of which treatment they actually received, is

$$\begin{aligned} E(X \mid S) &= \Pr(I)E(X \mid I, S) + \Pr(O)E(X \mid O, S) \\ &= \lambda(\mu_I + \tau_R) + (1-\lambda)(\mu_O + \tau_S) \\ &= \lambda\mu_I + (1-\lambda)\mu_O + \lambda(\tau_R - \tau_S) + \tau_S \end{aligned}$$

and the difference between the means of the groups as randomized is, therefore:

$$\begin{aligned} E(X \mid R) - E(X \mid S) &= [\lambda\mu_I + (1-\lambda)\mu_O + \tau_R] - [\lambda\mu_I + (1-\lambda)\mu_O + \lambda(\tau_R - \tau_S) + \tau_S] \\ &= (1-\lambda)(\tau_R - \tau_S) \end{aligned}$$

Again we get a biased result, but this time, the bias only depends on the quantity of interest, $\tau_R - \tau_S$ and λ. We know that $(1-\lambda)(\tau_R - \tau_S)$ is an attenuated version of $\tau_R - \tau_S$, as $0 < \lambda < 1$. Also, we see that the comparison is unbiased when $\lambda = 0$, i.e., when there are no patients with inoperable tumors (so there would not have been a problem in the first place). The groups are identical if $\lambda = 1$, i.e., if all patients have inoperable tumors so no one ends up with surgery and the comparison is then vacuous.

The preceding derivation makes many dubious assumptions, so it would be unwise to suppose that in these (and related) circumstances, the comparison of randomized groups necessarily leads to an attenuated estimate of the treatment effect. Nevertheless, the results are not without an aspect of realism and do serve to illustrate the nature of the difficult problem that the intention-to-treat dictum attempts to address.

Because comparisons of the groups as randomized is the only comparison that is based on comparable groups, it should always be presented in the report of the trial. Other analyses, such as the comparison of the groups of patients who were actually treated as specified in the protocol (usually called the *per protocol* analysis) can be presented but should be interpreted cautiously.

Exercises

1. Pain from muscle strains generally lasts about two weeks. A trial was performed to compare a fortnight of treatment with one of two types of painkiller for the relief of pain from this condition. One treatment was paracetamol (acetaminophen) with codeine (C) and the other was indomethacin (I). About 20% of the patients randomized to I complained of stomach pains within 2 d of starting treatment and had to stop taking the treatment. Your clinical colleague suggests comparing those allocated to C with those who completed two weeks taking I. What is wrong with this strategy? What comparison should be made? This would be an instance of what dictum?

11

Some Special Designs: Crossovers, Equivalence, and Clusters

The trials considered so far have all been of a simple kind, often referred to as *parallel group designs*, in which individual patients are allocated to one of the treatments under investigation, and the intention is to assess if the treatments used differ. However, there are many other kinds of trials, and three more specialized designs will be described in this chapter. In crossover trials, patients may receive, in turn, several of the treatments under investigation. In cluster randomized trials, it is not individual patients but groups of patients or other subjects who are randomized to different treatments. Equivalence trials seek to establish equivalence, rather than difference, between treatments.

11.1 Crossover Trials

In all the trials considered so far in this book, each patient has received just one of the treatments being compared. This is natural for a majority of diseases and conditions. Investigation of a new material for use in the construction of plasters for fractures, new approaches to removing the appendix and antithrombolytic treatment (preventing blood clots) following heart attacks are examples of trials in which each patient will have the opportunity of only one treatment. However, what about conditions such as asthma and diabetes, which cannot be cured? What about comparing different dialyzer membranes for patients having kidney dialysis thrice a week? Such patients could be given several treatments.

Trials in which the aim is not to cure a condition present the possibility of giving more than one treatment to each patient. Such trials are known as *crossover trials*.

The main advantage of using a crossover trial is that the outcome of a patient when given treatment A is not compared to the outcome from some different patients given treatment B but to the outcome from the same patient

when given B. Crossover trials therefore seem to offer the possibility of obtaining more precise treatment comparisons.

11.2 The AB/BA Design

For two treatments, the simplest form of crossover design would be to give each patient treatment A and then follow it with B. However, the results would be ambiguous for a reason that has two forms.

1. If A appeared worse than B, it may be that the treatment given second does better, whatever it is, and had we given the treatments in the opposite order, it would have been B that fared worse. This is not entirely fanciful; if blood pressures are measured serially, they are commonly found to be higher the first few times they are measured.
2. If all the patients start the trial at the same time, there may be a trend that affects the outcomes. One instance might be if all the readings from the laboratory were higher on Monday than Tuesday.

To overcome this, the simplest form of crossover trial that is used randomly allocates patients to two groups. Patients in group 1 receive the treatments in the order AB, whereas those in group 2 receive them in the opposite order. The times when treatments are given are referred to as *treatment periods* or simply, *periods*. The design is represented schematically in the following table.

Treatment Allocations in the AB/BA Design

	Period 1	Period 2
Group 1	A	B
Group 2	B	A

This design overcomes the problems described in 1 and 2. If A appears to do worse than B in group 1, this can only be ascribed to an order or period effect if B appears to do worse than A in group 2.

11.3 Analysis of AB/BA Design for Continuous Outcomes

11.3.1 The Theory

An analysis needs to take account of several features.

1. It should ensure that pairs of measurements from a single patient be kept together, in some sense.
2. There may be systematic differences between the treatment periods as well as treatment effects.

In the following it is assumed that the numbers of patients allocated to groups 1 and 2 are n_1 and n_2, respectively. It will also be assumed that the outcome has a normal distribution.

The analysis starts from a model for the outcome. In group 1, the outcome on patient i ($i = 1, \ldots, n_1$) in period j ($j = 1,2$) is assumed to be x_{ij}. For $i = 1, \ldots, n_1$,

$$x_{i1} = \mu + \pi_1 + \tau_A + \xi_i + \varepsilon_{i1} \text{ (period 1) and } x_{i2} = \mu + \pi_2 + \tau_B + \xi_i + \varepsilon_{i2} \text{ (period 2)}$$

Here, π_j ($j = 1,2$) is the systematic effect of period j, τ_A, τ_B are the systematic effects of treatment, and μ is a general mean. The term ξ_i represents the effect of the ith patient. Patients may have a tendency to always give a high or low response, and we model this by taking ξ_i to be a normal random variable with mean 0 and variance σ_B^2. Note that the same realization of ξ_i appears in the outcome for both periods. The ε terms are independent error terms with zero mean and variance $\tilde{\sigma}^2$.

For group 2, an identical argument gives, for $i = n_1 + 1, \ldots, n_1 + n_2$,

$$x_{i1} = \mu + \pi_1 + \tau_B + \xi_i + \varepsilon_{i1} \text{ (period 1) and } x_{i2} = \mu + \pi_2 + \tau_A + \xi_i + \varepsilon_{i2} \text{ (period 2)}$$

The term ξ_i represents the variability that exists between patients. However, this should not affect our analysis as the adoption of a crossover design has effectively eliminated this source of variation. This gives the clue on how to proceed most simply. We can remove ξ_i from the analysis by taking differences within each patient. The differences in group 1 are:

$$d_i = x_{i1} - x_{i2} = \pi + \tau + \eta_i \quad i = 1, \ldots, n_1$$

where $\pi = \pi_1 - \pi_2$, $\tau = \tau_A - \tau_B$, and $\eta_i = \varepsilon_{i1} - \varepsilon_{i2}$. The first of these parameters measures the difference between the treatment periods, the second is what we are interested in, namely, the difference between the treatments, and the third is another error term, with zero mean and variance $\sigma^2 = 2\tilde{\sigma}^2$. In group 2, the differences are:

$$d_i = x_{i1} - x_{i2} = \pi - \tau + \eta_i \quad i = n_1 + 1, \ldots, n_1 + n_2$$

Therefore, the expected value of the differences in group 1 is $\pi+\tau$ and in group 2 is $\pi-\tau$. If there is no treatment difference, then $\tau=0$, and the two sets of differences have the same expectation. In other words, the hypothesis of no treatment difference in the AB/BA design is tested simply by using a two-sample t-test to compare the two sets of within-patient differences.

If the mean sample difference in group k is $\bar{d}_k$, $k=1,2$ then $E(\bar{d}_1-\bar{d}_2)=2\tau$. So, an estimate of the treatment difference τ is $\frac{1}{2}(\bar{d}_1-\bar{d}_2)$. A confidence interval for τ can be found by dividing the ends of the usual confidence interval for the difference between the means of the two sets of differences by 2. This procedure is illustrated in the following text.

Note that the term ξ_i was eliminated from the analysis by taking differences. Hence, the precision of the results depends only on the variance of the εs, $\bar{\sigma}^2$ and not on the variance of ξ_i, σ_B^2. For outcomes that differ much more between individuals than within individuals, the elimination of a relatively large variance component σ_B^2 from analysis is valuable. This is a mathematical expression of the intuitive observations that it is better to use a patient as his or her own control.

11.3.2 An Application

Children suffering from enuresis (bed wetting) are given a drug to alleviate their problem. The drug is given for a fortnight, and the outcome is the number of dry nights (out of 14) observed. The control drug is a placebo. Patients are allocated to group 1, in which the drug is given for a fortnight. At the end of this period, a placebo is administered for a fortnight and the same outcome variable is recorded. In group 2, the placebo is given first followed by the drug. The data for 29 patients are shown in Table 11.1.

The mean differences in the two groups are

$$\text{Group 1: } \bar{d}_1 = 2.824 \quad \text{Group 2: } \bar{d}_2 = -1.25$$

The mean in group 1 is positive, and as this is the mean of differences (drug-placebo), it suggests that the number of dry nights is larger on the drug than the placebo. The mean in group 2 is negative, and as this is the mean of differences (placebo-drug), it again suggests that the number of dry nights is larger on the drug than the placebo.

Does this stand up to more careful scrutiny? To do this we compare the two sets of differences using a two-sample t-test. Doing this in Minitab, with group 1 differences in a column named *Group 1*, etc., we obtain the following output:

```
Two-Sample T for Group 1 vs. Group 2
N          Mean    StDev      SE   Mean
Group 1      17     2.82    3.47   0.84
Group 2      12    -1.25    2.99   0.86
```


TABLE 11.1

Data from Enuresis Trial: Number of Dry Nights out of 14

	Group 1 Drug > Placebo			Group 2 Placebo > Drug			
Patient	Period 1	Period 2	Difference	Patient	Period 1	Period 2	Difference
1	8	5	3	18	12	11	1
2	14	10	4	19	6	8	–2
3	8	0	8	20	13	9	4
4	9	7	2	21	8	8	0
5	11	6	5	22	8	9	–1
6	3	5	–2	23	4	8	–4
7	13	12	1	24	8	14	–6
8	10	2	8	25	2	4	–2
9	6	0	6	26	8	13	–5
10	0	0	0	27	9	7	2
11	7	5	2	28	7	10	–3
12	13	13	0	29	7	6	–1
13	8	10	–2				
14	7	7	0				
15	9	0	9				
16	10	6	4				
17	2	2	0				

Source: Data from Armitage P., Hills, M. (1982), The two-period crossover trial, *The Statistician*, 31, 119–131.

```
95% CI for mu Group 1 - mu Group 2: (1.54, 6.61)
T-Test mu Group 1 = mu Group 2 (vs. not =): T = 3.29
P = 0.0028 DF = 27
Both use Pooled StDev = 3.28
```

We see that the test of the null hypothesis that the means in the two groups are the same, i.e., $\tau=0$, yields $P=0.0028$, indicating that the drug does appear to alleviate the problem.

The difference in means, $\bar{d}_1-\bar{d}_2=4.074$, and the preceding output shows that an associated 95% confidence interval is (1.54, 6.61).

At the end of Subsection 11.3.1, The Theory, it was shown that $E(\bar{d}_1-\bar{d}_2)=2\tau$, i.e., $\bar{d}_1-\bar{d}_2$ is an unbiased estimator not of the quantity of interest but of twice that quantity. Consequently, the estimator of τ is $\frac{1}{2}(\bar{d}_1-\bar{d}_2)=2.037$ nights. A 95% confidence interval for τ is ($^1/_2\times 1.54$, $^1/_2\times 6.61$) = (0.77, 3.31) nights.

11.4 The Issue of Carryover

An obvious potential problem with a crossover trial is that the effects of the treatment given in period 1 may still persist during period 2. Such a persis-

tence of a treatment effect is known as a *carryover effect*. What might be the effect if such a phenomenon were present?

A way to investigate this is to adapt the model presented in Subsection 11.3.1. This can be done as follows, using the notation of that subsection.

In group 1, so for $i = 1, \ldots, n_1$, we leave the model for period 1 unchanged (carryover cannot affect the first period), but add a term γ_A to the model for period 2.

$$x_{i1} = \mu + \pi_1 + \tau_A + \xi_i + \varepsilon_{i1} \qquad \text{(period 1) and}$$

$$x_{i2} = \mu + \pi_2 + \tau_B + \gamma_A + \xi_i + \varepsilon_{i2} \quad \text{(period 2)}$$

The response in period 2 might be affected by the persistent effect of treatment A given in period 1, and γ_A is a parameter that represents this effect. Similarly for γ_B, so the model for responses in group 2, i.e., for $i = n_1 + 1, \ldots, n_1 + n_2$, becomes

$$x_{i1} = \mu + \pi_1 + \tau_B + \xi_i + \varepsilon_{i1} \qquad \text{(period 1) and}$$

$$x_{i2} = \mu + \pi_2 + \tau_A + \gamma_B + \xi_i + \varepsilon_{i2} \quad \text{(period 2)}$$

If we did not take any notice of this amended form of the model and decided to estimate treatment effect as before, i.e., using $\frac{1}{2}(\bar{d}_1 - \bar{d}_2)$, what would we actually be estimating?

From the amended model, it follows that $E(\bar{d}_1) = \pi + \tau - \gamma_A$ and $E(\bar{d}_2) = \pi - \tau - \gamma_B$, and hence, $E[\frac{1}{2}(\bar{d}_1 - \bar{d}_2)] = \tau - \frac{1}{2}\gamma$, where $\gamma = \gamma_A - \gamma_B$. Therefore, if there is a carryover effect, i.e., $\gamma \neq 0$, and it is ignored in the analysis, the estimator of τ is biased.

What can be done about this? One proposal that was widely followed for many years was to perform a preliminary analysis to test the null hypothesis that $\gamma = 0$. Such a hypothesis test is simple to implement: it is a two-sample t-test comparing the sums $s_i = x_{i1} + x_{i2}$ between groups 1 and 2. The procedure continued as follows:

1. If the test rejected the null hypothesis $\gamma = 0$ then the data from period 1 only were compared using the usual t-test for a parallel group trial (to which this study has now been reduced).
2. If the test could not discredit $\gamma = 0$ then the procedure described in Section 11.3 of this chapter is followed.

This approach is not recommended now because there are several problems. The most transparent one is that ξ_i has not been eliminated from the test of $\gamma = 0$. So this test is affected by between-patient variation. This is likely

to be large, and as the size of the trial would be determined by a sample size calculation based on the smaller variance $\tilde{\sigma}$, it is likely that the test of $\gamma = 0$ has poor power. Consequently, the decision to follow the procedure in Section 11.3 of this chapter may well be taken even in the presence of a substantial nonzero value of γ.

The recommended approach is not to use this particular crossover design when there is a possibility of a carryover effect. You should try to use nonstatistical arguments, perhaps based on the half-lives of drugs, etc., to decide how long treatment effects are likely to persist. The AB/BA design can then be used if the treatment periods are separated by "washout periods" whose duration is sufficient to ensure that carryover cannot occur.

11.5 Equivalence Trials

11.5.1 General Remarks

There are circumstances when the aim of a trial is not to detect differences between the treatments under study but to establish that, for all practical purposes, the efficacy of two treatments is equivalent. It may be that one treatment might be thought to be safer than another, one might be cheaper, or there may be advantages in terms of convenience.

A fundamental feature of an equivalence trial is that the usual hypothesis test is of little value. Failing to establish that one treatment is superior to the other is not the same as establishing their equivalence; see Subsection 3.2.1. On the other hand, a difference that is detected may have little importance and could well correspond to clinical equivalence.

The usual method when determining the equivalence of two treatments, A and B, is to compute a 95% confidence interval or, in general, a $100(1-\alpha)\%$ interval for the difference in the treatment means, μ_A and μ_B. For the purposes of illustration, it will be assumed that the standard deviation of the outcomes, σ, is known. If n_A and n_B patients are recruited to each group, then the confidence interval is

$$(\bar{d} - z_{1-\frac{1}{2}\alpha}\sigma\lambda, \bar{d} + z_{1-\frac{1}{2}\alpha}\sigma\lambda) \qquad (11.1)$$

where $\bar{d}$ is the difference in the sample means, z_ξ is such that $\Pr(Z > z_\xi) = \xi$, where Z is a standard normal variable and, as in Chapter 3, $\lambda = \sqrt{n_A^{-1} + n_B^{-1}}$.

A commonly used method is to consider the treatments to be equivalent if both ends of the interval in Equation 11.1 lie within the prespecified interval of equivalence $(-\delta, \delta)$. If this does not occur, then equivalence has not been established. There is a more general formulation using an interval (δ_L, δ_U), which can be helpful when comparing a new treatment with a

standard, and there is less concern if the new treatment is better than the standard. This refinement of the methodology will not be pursued here. The specification of δ must be made in close collaboration with clinical experts and is, in some ways, analogous to specifying τ_M in a conventional parallel groups trial (cf. Chapter 3).

The methods explained in Chapter 3 for determining the size of a conventional RCT do not apply directly to equivalence studies, because they are focused on a test of the hypothesis $\tau = \mu_A - \mu_B = 0$, which is inappropriate in this setting. The calculations and associated error probabilities required for equivalence trials are set out in the following subsection.

11.5.2 Sample Sizes for Equivalence Trials with Normally Distributed Outcomes

It should be recalled that the null hypothesis can only be discredited, it cannot be shown to be true. In the context of an equivalence trial, it is therefore useful to think in terms of the null hypothesis representing difference and the alternative being equivalence. These hypotheses might be written as

$$H_0 : |\tau| > \delta \text{ vs. } H_1 : |\tau| \leq \delta$$

It should be noted that there is no point in trying to establish exact equivalence, $\delta = 0$, as there will always be some uncertainty in our estimates, and such a null hypothesis would never be discredited. As with conventional trials, two kinds of mistakes can be made when conducting an equivalence trial: it can be concluded that the treatments are equivalent when they are not, or it can be concluded that genuinely equivalent treatments are not equivalent. In terms of the preceding hypotheses, these are the type I and type II errors, respectively. As with conventional trials, the sample size is set to place acceptable values on these error probabilities.

Two treatments will be deemed equivalent if the interval in Equation 11.1 lies within $(-\delta, \delta)$. This amounts to requiring that $\bar{d} \in (-\varsigma, \varsigma)$ where $\varsigma = \delta - z_{1\alpha}\sigma\lambda$. Note that it would be impossible to assert equivalence if $\delta < z_{1\alpha}\sigma\lambda$. The distribution of $\bar{d}$ is normal with mean τ and standard deviation $\sigma\lambda$, so

$$\Pr(\bar{d} \in (-\varsigma, \varsigma)) = \Phi\left(\frac{\varsigma - \tau}{\lambda\sigma}\right) - \Phi\left(\frac{-\varsigma - \tau}{\lambda\sigma}\right) = \Phi\left(\frac{\delta - \tau}{\lambda\sigma} - z_{1\alpha}\right) - \Phi\left(\frac{-\delta - \tau}{\lambda\sigma} + z_{1\alpha}\right) \quad (11.2)$$

The chance of asserting equivalence when the treatments are not equivalent (which in terms of parameters we take to mean $|\tau| \geq \delta$), i.e., the type I error rate varies with τ, reaching a maximum over the region of difference,

$|\tau| \geq \delta$, when $|\tau| = \delta$. The type I error rate is taken to be the value of Equation 11.2 when $\tau = \delta$, which is

$$\Phi(-z_{1\alpha}) - \Phi\left(z_{1\alpha} - \frac{2\delta}{\lambda\sigma}\right) = \tfrac{1}{2}\alpha - \Phi\left(z_{1\alpha} - \frac{2\delta}{\lambda\sigma}\right) \quad (11.3)$$

The power of the trial, $1 - \beta$, is the probability of declaring the treatments are equivalent when they really are equivalent. For this purpose, the power is defined as the value of Equation 11.2 when there is exact equivalence, i.e., $\tau = 0$. Hence

$$1 - \beta = 2\Phi\left(\frac{\delta}{\lambda\sigma} - z_{1\alpha}\right) - 1$$

and, therefore,

$$1 - \tfrac{1}{2}\beta = \Phi(z_{1\beta}) = \Phi\left(\frac{\delta}{\lambda\sigma} - z_{1\alpha}\right)$$

It follows that

$$\delta/(\lambda\sigma) = (z_{1\alpha} + z_{1\beta}) \quad (11.4)$$

so if the two treatment groups have the same size, n, we have

$$n = 2\frac{\sigma^2}{\delta^2}(z_{1\alpha} + z_{1\beta})^2 \quad (11.5)$$

Substituting $\delta/(\lambda\sigma) = (z_{1\alpha} + z_{1\beta})$ into Equation 11.3 gives the type I error rate as

$$= \tfrac{1}{2}\alpha - \Phi(-z_{1\alpha} - 2z_{1\beta}) \approx \tfrac{1}{2}\alpha$$

assuming a reasonable power, say, >70%. Therefore, if equivalence is based on a $100(1 - \alpha)\%$ confidence interval, then the type I error rate is $100(1/2\alpha)\%$, so a 95% confidence interval has type I error 2.5%.

11.5.3 Comparison of Conventional and Equivalence Trials

The roles of τ_M, the minimal clinically important difference, and δ are similar, especially in the way they appear in the formulae in Subsection 11.5.2 of this chapter. However, it will often be prudent to use a value of δ substantially smaller than a value of τ_M used in a related but conventional trial. In a trial looking for a difference, a clinician may only be interested in changing treatments if the new therapy offers a substantial advantage. In an equivalence trial, evidence is sought to support the interchangeability of the two treatments, and it may then be appropriate to demand a closer agreement in their mean response. A further reason why equivalence trials are often larger than conventional trials can be found from comparing formulae in Equation 3.3 of Chapter 3 and Equation 11.4. The term z_β in Equation 3.3 of Chapter 3 becomes $z_{2\beta}$ in Equation 11.4 and as $z_{2\beta} > z_\beta$, this will increase the required sample size.

There are many less technical issues surrounding equivalence trials but these will not be discussed in depth. A general comment that is often made concerns the standard of execution of these studies. In a conventional trial, sloppy conduct is not at all in the interests of the investigator: e.g., poor data recording and checking will increase σ and make it more difficult to find genuine differences. In an equivalence study, sloppiness tends to increase the chances that the investigator will be unable to discover a difference, which is now the aim of the study. This is certainly true if part of the sloppiness includes an inappropriate analysis. However, if the analysis is based on comparing the confidence interval in Equation 11.1 with a prespecified interval of equivalence, then poor technique will tend to widen Equation 11.1, thereby reducing the chance it will be contained within a properly chosen interval of equivalence. It is, nevertheless, worth reinforcing that poor technique, such as inadequate attention to blindness, can cause problems for an equivalence study that are every bit as severe as those caused to conventional trials.

11.6 Cluster Randomized Trials

11.6.1 Introduction and Rationale

The trials described up to this time have allocated each patient to a treatment, or in the case of crossover trials, a sequence of treatments. As RCTs have become more widely accepted as the method of choice for the assessment of treatment efficacy, there has been an increase in the areas in which investigators have wanted to use this methodology. However, application in areas different from the traditional use in clinical medicine, which essentially deals with the health of individual patients, often gives rise to a particular difficulty. This is that it is no longer appropriate to randomize individual patients

but whole clusters, or groups of patients must be allocated en bloc to a given treatment. These are known as *cluster randomized* or *group randomized trials*.

One example is provided by a trial designed to assess the impact of improved methods for the treatment of sexually transmitted diseases (STDs) on the incidence of HIV infection in a rural region of Tanzania, near Mwanza (Grosskurth et al., 1995). The treatment comprised a program that among other things, involved the training of health center staff to manage STDs better, the provision of better laboratory facilities in Mwanza, and the provision of a reliable supply of effective drugs for the treatment of STDs. Twelve large communities, each being the catchment area of a health center, were involved in the study. The RCT had to apply the improved methods and compare these with the current methods. Individuals cannot be randomized as the methods are not applied to the patients but to the staff of the health centers, each serving many hundreds of patients. Therefore, the community and its health center were randomly allocated to receive the improved treatment scheme or the current scheme. With only twelve communities in the trial, if the communities exhibit substantial initial differences in their HIV incidences, it is clear that the randomization could lead to substantial differences between the two treatment groups. This was indeed the case: communities near to major roads or the shore of Lake Victoria did exhibit higher incidence of HIV infection than did more remote communities. To overcome this, the investigators formed six pairs from the communities, matched with respect to their location and a number of other factors. One member of each pair was randomly allocated to the new treatment scheme and the other received the standard treatment.

This illustrates a general feature of cluster randomized trials: the number of clusters is generally much lower than the number of patients in the usual individual-patient parallel group study. Often, small numbers of clusters arise when each cluster comprises many patients and, in these cases, the clusters may not be particularly heterogeneous, so important imbalances might not arise. However, as the preceding example shows, this is by no means always the case and, in these instances, paired designs such as that just described are used.

Another example is provided by a study that is similar insofar as it is aimed at assessing the effect of an intervention in primary care. The study is designed to assess whether additional training of nurses and GPs in a general practice would improve the care of patients with newly diagnosed type II diabetes mellitus (Kinmonth et al., 1998). Forty-one practices in the south of England were randomized to the status quo or to receive additional training for their staff. The outcomes were measures of quality of life and of diabetic control. In this trial there was no pairing of practices.

The main statistical problem that arises with this sort of study is that you cannot analyze the data as if the patients themselves had been individually randomized to treatment. It cannot be assumed, *ab initio*, that responses on patients that are from the same cluster are independent. Such a correlation could arise because of similarities in the way certain measurements are taken

by the practice nurse, or that methods for sending samples to the laboratory might differ more between practices than they do within a practice. For variables such as measures of quality of life, less tangible aspects, such as the atmosphere within a practice, could have a bearing. If these responses have a positive correlation, then they will be more similar than would be expected if they were independent. If the method used for the analysis assumes that the individual responses are independent then the estimated variance will be too small. This can be seen more formally as follows.

Suppose the outcome of the jth individual in the ith cluster is represented by a continuous random variable X_{ij} with mean μ_A, μ_B in treatment groups A and B, respectively, and variance σ^2. Suppose also that responses within a cluster have correlation ρ and that responses in different clusters are independent. The estimate of variance, computed from one of the treatment groups (so the mean of each response is the same), but without regard to the presence of clusters is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2}{N-1} \qquad (11.6)$$

the ith cluster has size n_i and $N = \sum_{i=1}^{K} n_i$, where there are K clusters. The mean is computed without taking account of the clustering, so $\bar{X} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} X_{ij} / N$.

Expanding the numerator of Equation 11.6 and taking expectations we obtain:

$$E([N-1]\hat{\sigma}^2) = N\sigma^2 - N^{-1}E\left(\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij} - E(X_{ij})]\right\}^2\right)$$

$$= N\sigma^2 - N^{-1}\,\mathrm{var}\left(\sum_{i=1}^{K} T_i\right)$$

$$= N\sigma^2 - \frac{1}{N}\sum_{i=1}^{K}\mathrm{var}(T_i)$$

as clusters are independent and T_i is the sum of responses in cluster i. This variance can be computed as

$$\mathrm{var}(T_i) = \sum_{j=1}^{n_i}\mathrm{var}(X_{ij}) + \sum_{j\neq t}\mathrm{cov}(X_{ij}, X_{it})$$

$$= n_i\sigma^2 + n_i(n_i-1)\rho\sigma^2$$

so the expectation of Equation 11.6 is $\sigma^2(1 - C\rho)$ where

$$C = \frac{\sum n_i(n_i-1)}{N(N-1)}$$

Thus if the analyst supposed that Equation 11.6 was a valid estimate of error, then the analysis would be biased with too small a value used for the estimate of standard deviation, with a consequent exaggeration of the significance of treatment effects.

11.6.2 Methods of Analysis for Cluster-Randomized Trials

A valid method of analysis is to use the simple treatment means, $\bar{X}_A, \bar{X}_B$, found ignoring clustering and to adapt the preceding calculations so that a legitimate standard error for $\bar{X}_A - \bar{X}_B$ is used. The variance of a treatment mean is, in the notation of the previous subsection,

$$N^{-2}\,\mathrm{var}\left(\sum_{i=1}^{K} T_i\right) = N^{-2}\sum_{i=1}^{K}\mathrm{var}(T_i) = N^{-2}\sum_{i=1}^{K}(n_i\sigma^2 + n_i(n_i-1)\rho\sigma^2)$$

$$= \frac{\sigma^2}{N}\left[1 + \rho\left(\frac{\sum n_i^2}{N} - 1\right)\right]$$

Thus the effect of clustering is to increase the variance of the sample mean by a factor $[1 + \rho(\Sigma n_i^2 / N - 1)]$. If the clusters all have the same size, n, then this factor becomes $[1 + \rho(n-1)]$. Thus, the variance, V, of $\bar{X}_A - \bar{X}_B$ can be found as the sum of this expression for the two treatment groups and a test of the null hypothesis of no treatment difference can be made by referring $(\bar{X}_A - \bar{X}_B)/\sqrt{V}$ to a standard normal distribution.

There remains the problem of how to estimate the variance of an individual response, σ^2, and the within-cluster correlation ρ. Naïve methods for σ^2 have been shown to be misleading.

Maximum likelihood methods could be employed but a simpler approach can be used. In this a model is postulated for the outcomes from a cluster-

randomized trial. The model has some similarities to the model proposed in Subsection 11.3.1 for a crossover trial. The idea is that the response on the jth individual in the ith cluster is modeled by

$$X_{ij} = \mu_T + G_i + \varepsilon_{ij} \tag{11.7}$$

where T is either A or B, depending on which treatment was applied in the ith cluster. The terms ε_{ij} are independent random variables with zero mean and common variance σ_W^2. The term G_i is a random variable, also with zero mean, independent of ε_{ij}, which measures the effect of the ith cluster. As with the random variable measuring the patient effect in a crossover trial, ξ_i, the same realization of G_i is applied to each member of the cluster. The variance of G_i is σ_G^2. Consequently, responses in a cluster that has a larger value of G_i will all tend to be higher, and it is this feature of the model that induces the within-cluster correlation needed to make the model reasonable. More specifically, the variance of any member of a cluster is $\sigma^2 = \sigma_G^2 + \sigma_W^2$, and the covariance of any two members of the same cluster is, from Equation 11.7, for $j \neq \ell$:

$$E\left[(X_{ij} - \mu_T)(X_{i\ell} - \mu_T)\right] = E\left[(G_i + \varepsilon_{ij})(G_i + \varepsilon_{i\ell})\right] = \sigma_G^2$$

It follows that the within-cluster correlation $\rho = \sigma_G^2/\sigma^2$.

Estimates of σ_G^2 and σ_W^2 can be found as the standard between- and within-group variance components applied in each treatment group, and the results averaged appropriately across the treatments. Alternatively, the mixed model in Equation 11.7 can be fitted directly to all the data. Estimates of σ^2 and ρ can be found from their relationships with σ_G^2 and σ_W^2.

An alternative and simpler analysis is to compute the mean response within each cluster, $\bar{X}_i$, and use these means as if they were raw data in a t-test. The clusters are independent, so the analysis is free from the difficulties due to within-cluster dependency. A potential criticism is that a t-test assumes each number used in the test has the same variance. As the variance of the mean of the ith cluster is $\sigma_G^2 + \sigma_W^2 / n_i$, this is not true unless the clusters all have the same size. However, the analysis will lose only a little efficiency if either the clusters have similar sizes or if σ_G^2 is substantially larger than σ_W^2, in which case the term varying with cluster size is relatively unimportant.

11.6.3 Sample Size Estimation for Continuous Outcomes from Cluster-Randomized Trials

If the method of analysis is affected by the presence of clustering, so too is the manner in which sample sizes are estimated. In the absence of clustering, the formula for the size of each group is given by Equation 3.4 of Chapter 3

$$N = \frac{2\sigma^2(z_\beta + z_{1-\frac{\alpha}{2}})^2}{\tau_M^2}$$

where α is the type I error rate, $1 - \beta$ is the power to detect a difference of τ_M, and σ^2 is the variance of the response of an individual patient. The formula provides a link with N because the variance of each treatment mean is σ^2/N.

For a cluster randomized trial, the variance of the treatment mean is not σ^2/N but $(\sigma^2/N)[1+\rho(\Sigma n_i^2 / N - 1)]$. A difficulty with cluster-randomized trials is that the sizes of the clusters may well not be known when the trial is planned. In this case, it is essential to have some idea about this quantity, perhaps through an estimate of the likely average cluster size, n_a. If this is available, the variance of the mean is taken to be $(\sigma^2/N)[1+\rho(n_a - 1)]$, and it is $\sigma^2[1+\rho(n_a - 1)]$ that is used in place of σ^2 in the sample size formula; that is, the total number of patients in the clusters receiving each treatment should be

$$N = \frac{2\sigma^2(1+\rho(n_a - 1))(z_\beta + z_{1-\frac{\alpha}{2}})^2}{\tau_M^2} \tag{11.8}$$

As can be seen, the presence of clustering means that the planning of a cluster-randomized trial requires knowledge not only of all the quantities needed for Equation 3.4 of Chapter 3 but additional information about the size and effect of clustering, through n_a and ρ.

An alternative approach is to base the calculation of sample size on the cluster means, giving a formula for the number of clusters that should receive each treatment. If the outcome follows the model in Equation 11.7, then the variance of the sample mean should be used in Equation 3.4 of Chapter 3 giving the number of clusters receiving each treatment as

$$\frac{2(\sigma_G^2 + \sigma_W^2 / n_a)(z_\beta + z_{1-\frac{\alpha}{2}})^2}{\tau_M^2} \tag{11.9}$$

11.6.4 General Remarks about Cluster-Randomized Trials

Cluster-randomized trials raise many complicated practical issues. The sample size estimates in the previous subsection are just one example. They require the specification of not only the usual quantities but of the cluster size and the intraclass correlation, which is likely to need extensive experience of the area of application before the trial can be planned adequately. Sensible assessment of the sensitivity of the required number of patients will

generally require the statistician to investigate the effect of a range of values for ρ in Equation 11.8.

Other issues, such as the way withdrawals and dropouts are handled, and the meaning of informed consent, present problems that are absent from conventional trials. On a more technical level, the foregoing discussion has concentrated on continuous outcomes for good reason. The technicalities presented by binary outcomes are rather more formidable. Recent developments in hierarchical data-analysis do, however, mean that the efficient analysis of cluster-randomized trials is becoming easier.

Exercises

1. The following data are from an AB/BA crossover trial in which patients are treated with two bronchodilators (a widely used form of inhaled drug designed to help patients with asthma), namely salbutamol (S) and formoterol (F). The outcome presented in the following table is the peak expiratory flow (PEF) in liters per minute (l/min). Patients were randomized to receive the drug in the order F then S or S then F (Data from Senn and Auclair, *Statistics in Medicine*, 1990, 9, 1287–1302).

Patient	PEF in Period 1 (l/min)	PEF in Period 2 (l/min)	Order (1 = FS, 2 = SF)
1	310	270	1
2	370	385	2
3	310	400	2
4	310	260	1
5	380	410	2
6	370	300	1
7	410	390	1
8	290	320	2
9	250	210	1
10	380	350	1
11	260	340	2
12	90	220	2
13	330	365	1

Analyze the preceding data, assuming that there is no carryover effect of treatment; be careful to define the statistical model that you use. Make sure that your analysis includes a test of the null hypothesis that there is no difference in the mean treatment effect when treated with formoterol or salbutamol. You should also provide point and interval (95%) estimates of the treatment effect, making sure that you define clearly what you mean by this term. Compute similar

quantities using the parallel group trial formed from the data in the first period and comment.

2. Suppose that the assumption of no carryover cannot be sustained on nonstatistical grounds and it is decided to try to use the data to assess if carryover is present. Suppose that the model for the responses is now:

$$x_{i1} = \mu + \pi_1 + \tau_F + \xi_i + \varepsilon_{i1} \quad \text{(period 1) and}$$

$$x_{i2} = \mu + \pi_2 + \tau_S + \gamma_F + \xi_i + \varepsilon_{i2} \quad \text{(period 2)}$$

in the F then S group and in which the terms are as defined in Section 11.4 of this chapter but with A and B replaced by F and S to conform with the present application.

(a) What is the expectation of $S_i = x_{i1} + x_{i2}$? What is the corresponding expectation in the S and F group?

(b) What null hypothesis does a two-sample t-test between the two samples of S_is test? Perform this test using the data from question 1. What conclusion can you draw?

(c) If you proceed with the analysis used in question 1 but the true model is that shown in this question, what is the expectation of the treatment estimator? What is a 95% confidence interval for the bias term?

3. The model for the outcomes on the ith patient from an AB/BA crossover trial is as follows:

Sequence	Period 1	Period 2
AB ($i = 1, \ldots, n$)	$x_{i1} = \mu + \pi_1 + \tau_A + \xi_i + \varepsilon_{i1}$	$x_{i2} = \mu + \pi_2 + \tau_B + \xi_i + \varepsilon_{i2}$
BA ($i = n+1, \ldots, 2n$)	$x_{i1} = \mu + \pi_1 + \tau_B + \xi_i + \varepsilon_{i1}$	$x_{i2} = \mu + \pi_2 + \tau_A + \xi_i + \varepsilon_{i2}$

where μ is the general mean, π_j is the effect of period j (=1,2), the treatment effect of interest is $\tau = \tau_A - \tau_B$, and ξ_i, ε_{ij} are independent residuals with zero mean and variances σ_B^2, σ^2, respectively.

(a) Define $d_i = x_{i1} - x_{i2}$ and let the mean of these in sequence AB be $\bar{d}_{AB}$ and similarly for $\bar{d}_{BA}$. Also let $\bar{x}_{1AB}, \bar{x}_{1BA}$ be the mean responses in period 1 for patients allocated to sequences AB and BA, respectively. Show that $\frac{1}{2}(\bar{d}_{AB} - \bar{d}_{BA})$ has the same expectation as $\bar{x}_{1AB} - \bar{x}_{1BA}$ and identify this quantity.

(b) Find the variance of $\frac{1}{2}(\bar{d}_{AB} - \bar{d}_{BA})$ and of $\bar{x}_{1AB} - \bar{x}_{1BA}$ and the ratio R of these quantities.

(c) If $\sigma_B^2 = 6\sigma^2$ evaluate R and comment on the implication of this value when deciding whether to use a crossover design or a parallel group design.