

01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

Determining the Sample Size

Hain't we got all the fools in town on our side? and aint that a big enough majority in any town?
Mark Twain, *Huckleberry Finn*

Nothing comes of nothing.
Shakespeare, *King Lear*

13.1 BACKGROUND

Clinical trials are expensive, whether the cost is counted in money or in human suffering, but they are capable of providing results which are extremely valuable, whether the value is measured in drug company profits or successful treatment of future patients. Balancing potential value against actual cost is thus an extremely important and delicate matter and since, other things being equal, both cost and value increase the more patients are recruited, determining the number needed is an important aspect of planning any trial. It is hardly surprising, therefore, that calculating the sample size is regarded as being an important duty of the medical statistician working in drug development. This was touched on in Chapter 5 and some related matters were also considered in Chapter 6. My opinion is that sample size issues are sometimes over-stressed at the expense of others in clinical trials. Nevertheless, they are important and this chapter will contain a longer than usual and also rather more technical background discussion in order to be able to introduce them properly.

All scientists have to pay some attention to the precision of the instruments with which they work: is the assay sensitive enough? is the telescope powerful enough? and so on are questions which have to be addressed. In a clinical trial many factors affect precision of the final conclusion: the variability of the basic measurements, the sensitivity of the statistical technique, the size of the effect one is trying to detect, the probability with which one wishes to detect it if present (power), the risk one is prepared to take in declaring it is present when it is not (the so-called 'size' of the test, significance level or type I error rate) and the number of patients recruited. If it be admitted that the variability of the basic measurements has been controlled as far as is practically possible, that the statistical technique chosen is appropriately sensitive, that the magnitude of the effect one is trying to detect is an external 'given' and that a conventional type I error rate and power are to be used, then the only factor which is left for the trialist to

Statistical Issues in Drug Development/2nd Edition Stephen Senn
© 2007 John Wiley & Sons, Ltd

manipulate is the sample size. Hence, the usual point of view is that the sample size is the determined function of variability, statistical method, power and difference sought. In practice, however, there is a (usually undesirable) tendency to 'adjust' other factors, in particular the difference sought and sometimes the power, in the light of 'practical' requirements for sample size.

In what follows we shall assume that the sample size is going to be determined as a function of the other factors. We shall take the example of a two-arm parallel-group trial comparing an active treatment with a placebo for which the outcome measure of interest is continuous and will be assumed to be Normally distributed. It is assumed that analysis will take place using a frequentist approach and via the two independent-samples t -test. A formula for sample size determination will be presented. No attempt will be made to derive it. Instead we shall show that it behaves in an intuitively reasonable manner.

We shall present an approximate formula for sample size determination. An exact formula introduces complications which need not concern us. In discussing the sample size requirements we shall use the following conventions:

- α : the probability of a type I error, given that the null hypothesis is true.
- β : the probability of a type II error, given that the alternative hypothesis is true.
- Δ : the difference sought. (In most cases one speaks of the 'clinically relevant difference' and this in turn is defined 'as the difference one would not like to miss'. The idea behind it is as follows. If a trial ends without concluding that the treatment is effective, there is a possibility that that treatment will *never* be investigated again and will be lost both to the sponsor and to mankind. If the treatment effect is indeed zero, or very small, this scarcely matters. At some magnitude or other of the true treatment effect, we should, however, be disturbed to lose the treatment. This magnitude is the difference we should not care to miss.)
- σ : the presumed standard deviation of the outcome. (The anticipated value of the measure of the variability of the outcomes from the trial.)
- n : the number of patients in each arm of the trial. (Thus the total number is $2n$.)

The first four basic factors above constitute the *primitive* inputs required to determine the fifth. In the formula for sample size, n is a function of α , β , Δ and σ , that is to say, given the values of these four factors, the value of n is determined. The function is, however, rather complicated if expressed in terms of these four primitive inputs and involves the solution of two integral equations. These equations may be solved using statistical tables (or computer programs) and the formula may be expressed in terms of these two solutions. This makes it much more manageable. In order to do this we need to define two further terms as follows.

$Z_{\alpha/2}$: this is the value of the Normal distribution which cuts off an upper tail probability of $\alpha/2$. (For example if $\alpha = 0.05$ then $Z_{\alpha/2} = 1.96$.)

Z_{β} : this is the value of the Normal distribution which cuts off an upper tail probability of β . (For example, if $\beta = 0.2$, then $Z_{\beta} = 0.84$.)

We are now in a position to consider the (approximate) formula for sample size, which is

$$n = 2(Z_{\alpha/2} + Z_{\beta})^2 \sigma^2 / \Delta^2. \quad (13.1)$$

(N.B. This is the formula which is appropriate for a two-sided test of size α . See chapter 12 for a discussion of the issues.)

Power: That which statisticians are always calculating but never have.

Example 13.1

It is desired to run a placebo-controlled parallel group trial in asthma. The target variable is forced expiratory volume in one second (FEV₁). The clinically relevant difference is presumed to be 200 ml and the standard deviation 450 ml. A two-sided significance level of 0.05 (or 5%) is to be used and the power should be 0.8 (or 80%). What should the sample size be?

Solution: We have $\Delta = 200$ ml, $\sigma = 450$ ml, $\alpha = 0.05$ so that $Z_{\alpha/2} = 1.96$ and $\beta = 1 - 0.8 = 0.2$ and $Z_{\beta} = 0.84$. Substituting in equation (13.1) we have $n = 2(450 \text{ ml})^2(1.96 + 0.84)^2 / (200 \text{ ml})^2 = 79.38$. Hence, about 80 completing patients per treatment arm are required.

It is useful to note some properties of the formula. First, n is an *increasing* function of the standard deviation σ , which is to say that if the value of σ is increased so must n be. This is as it should be, since if the variability of a trial increases, then, other things being equal, we ought to need more patients in order to come to a reasonable conclusion. Second, we may note that n is a *decreasing* function of Δ : as Δ increases n decreases. Again this is reasonable, since if we seek a bigger difference we ought to be able to find it with fewer patients. Finally, what is not so immediately obvious is that if either α or β decreases n will increase. The technical reason that this is so is that the smaller the value of α , the higher the value of $Z_{\alpha/2}$ and similarly the smaller the value of β , the higher the value of Z_{β} . In common-sense terms this is also reasonable, since if we wish to reduce either of the two probabilities of making a mistake, then, other things being equal, it would seem reasonable to suppose that we shall have to acquire more information, which in turn means studying more patients.

In fact, we can express (13.1) as being proportional to the product of two factors, writing it as $n = 2F_1F_2$. The first factor, $F_1 = (Z_{\alpha/2} + Z_{\beta})^2$ depends on the error rates one is prepared to tolerate and may be referred to as *decision precision*. For a trial with 10% size and 80% power, this figure is about 6. (This is low decision precision). For 1% size and 95% power, it is about 18. (This would be high decision precision.) Thus a range of about 3 to 1 covers the usual values of this factor. The second factor, $F_2 = \sigma^2 / \Delta^2$, is specific to the particular disease and may be referred to as *application ambiguity*. If this factor is high, it indicates that the natural variability from patient to patient is high compared to the sort of treatment effect which is considered important. It is difficult to say what sort of values this might have, since it is quite different from indication to indication, but a value in excess of 9 would be unusual (this means the standard deviation is 3 times the clinically relevant difference) and the factor is not usually less than 1. Putting these two together suggests that the typical parallel-group trial using continuous outcomes should have somewhere between $2 \times 6 \times 1 = 12$ and $2 \times 18 \times 9 \approx 325$ patients per arm. This is a big range. Hence the importance of deciding what is indicated in a given case.

In practice there are, of course, many different formulae for sample size determination. If the trial is not a simple parallel-group trial, if there are more than two treatments, if the outcomes are not continuous (for example, binary outcomes, or length of survival

01 or frequency of events), if prognostic information will be used in analysis, or if the object
02 is to prove equivalence, different formulae will be needed. It is also usually necessary to
03 make an allowance for drop-outs. Nevertheless, the general features of the above hold.

04 A helpful tutorial on sample size issues is the paper by Steven Julious in *Statistics*
05 *in Medicine* (Julious, 2004); a classic text is that of Desu and Raghavarao (1990).
06 Nowadays, the use of specialist software for sample size determination such as NQuery,
07 PASS or Power and Precision is common.

08 We now consider the issues.

09

10

11 13.2 ISSUES

12

13 13.2.1 In practice such formulae cannot be used

14

15 The simple formula above is adequate for giving a basic impression of the calculations
16 required to establish a sample size. In practice there are many complicating factors
17 which have to be considered before such a formula can be used. Some of them present
18 severe practical difficulties. Thus a cynic might say that there is a considerable disparity
19 between the apparent precision of sample size formulae and our ability to apply them.

20 The first complication is that the formula is only approximate. It is based on the
21 assumption that the test of significance will be carried out using a *known* standard
22 deviation. In practice we do not know the standard deviation and the tests which we
23 employ are based upon using an estimate obtained from the sample under study. For
24 large sample sizes, however, the formula is fairly accurate. In any case, using the correct,
25 rather than the approximate, formula causes no particular difficulties in practice.

26 Nevertheless, although in practice we are able to substitute a sample estimate for
27 our standard deviation for the purpose of carrying out statistical *tests*, and although
28 we have a formula for the sample size calculation which *does* take account of this
29 sort of uncertainty, we have a particular practical difficulty to overcome. The problem
30 is that we do not know what the sample standard deviation will be until we have
31 run the trial but we need to plan the trial before we can run it. Thus we have to
32 make some sort of guess as to what the true standard deviation is for the purpose of
33 planning, even if for the purpose of analysis this guess is not needed. (In fact, a further
34 complication is that even if we knew what the sample standard deviation would be for
35 sure, the formula for the power calculation depends upon the unknown 'true' standard
36 deviation.) This introduces a further source of uncertainty into sample size calculation
37 which is *not* usually taken account of by any formulae commonly employed. In practice
38 the statistician tries to obtain a reasonable estimate of the likely standard deviation
39 by looking at previous trials. This estimate is then used for planning. If he is cautious
40 he will attempt to incorporate this further source of uncertainty into his sample size
41 calculation either formally or informally. One approach is to use a range of reasonable
42 plausible values for the standard deviation and see how the sample size changes. Another
43 approach is to use the sample information from a given trial to construct a Bayesian
44 posterior distribution for the population variance. By integrating the conditional power
45 (given the population variance) over this distribution for the population variance, an
46 unconditional (on the population variance) power can be produced from which a sample
47 size statement can be derived. This approach has been investigated in great detail by
48 Steven Julious (Julious, 2006). It still does not allow, however, for differences from trial

01 to trial in the true population variance. But it at least takes account of pure sampling
 02 variation in the trial used for estimating the population standard deviation (or variance)
 03 and this is an improvement over conventional approaches.

04 The third complication is that there is usually no agreed standard for a clinically
 05 relevant difference. In practice some compromise is usually reached between 'true'
 06 clinical requirements and practical sample size requirements. (See below for a more
 07 detailed discussion of this point.)

08 Fourth, the levels of α and β are themselves arbitrary. Frequently the values chosen
 09 in our example (0.05 and 0.20) are the ones employed. In some cases one might
 10 consider that the value of β ought to be much lower. In some diseases, where there are
 11 severe ethical constraints on the numbers which may be recruited, a very low value
 12 of β might not be acceptable. In other cases, it might be appropriate to have a lower
 13 α . In particular it might be questioned whether trials in which β is lower than α are
 14 justifiable. Note, however, that β is a theoretical value used for planning, whereas α is
 15 an actual value used in determining significance at analysis.

16 It may be a requirement that the results be robust to a number of alternative analyses.
 17 The problem that this raises is frequently ignored. However, where this requirement
 18 applies, unless the sample size is increased to take account of it, the power will be
 19 reduced. (If power, in this context, is taken to be the probability that *all* required tests
 20 will be significant if the clinically relevant difference applies.) This issue is discussed in
 21 section 13.2.12 below.

22
 23

24 **13.2.2 By adjusting the sample size we can fix our probability of being** 25 **successful**

26
 27
 28
 29
 30

This statement is not correct. It must be understood that the fact that a sample size
 has been chosen which appears to provide 80% power does not imply that there is an
 80% chance that the trial will be successful, because even if the planning has been
 appropriate and the calculations are correct:

31
 32
 33
 34
 35
 36
 37
 38
 39

- (i) The drug may not work. (Actually, strictly speaking, if the drug doesn't work we wish to conclude this, so that failure to find a difference is a form of success.)
- (ii) If it works it may not produce a clinically relevant difference.
- (iii) The drug might be better than planned for, in which case the power should be higher than planned.
- (iv) The power (sample size) calculation covers the influence of random variation *on the assumption* that the trial is run competently. It does not allow for 'acts of God' or dishonest or incompetent investigators.

40
 41
 42
 43

Thus although we can *affect* the probability of success by adjusting the sample size, we cannot *fix* it.

44
 45
 46

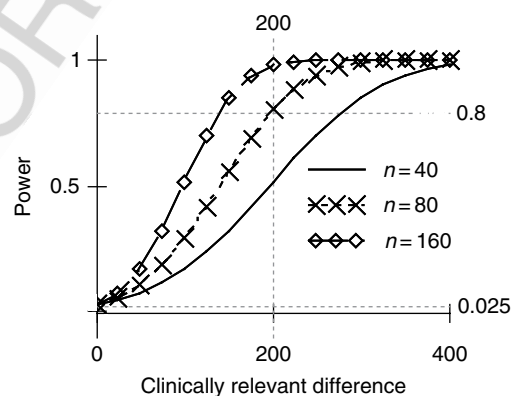
47 **13.2.3 The sample size calculation is an excuse for a sample size and** 48 **not a reason**

47
 48

There are two justifications for this view. First, usually when we have sufficient back-ground information for the purpose of planning a clinical trial, we already have a good

01 idea what size of trial is indicated. For example, so many trials now have been conducted
 02 in hypertension that any trialist worth her salt (if one may be forgiven for mentioning
 03 salt in this context) will already know what size the standard trial is. A calculation is
 04 hardly necessary. It is a brave statistician, however, who writes in her trial protocol,
 05 'a sample size of 200 was chosen because this is commonly found to be appropriate
 06 in trials of hypertension'. Instead she will usually feel pressured to quote a standard
 07 deviation, a significance level, a clinically relevant difference and a power and apply
 08 them in an appropriate formula.

09 The second reason is that this calculation may be the final result of several hidden
 10 iterations. At the first pass, for example, it may be discovered that the sample size is
 11 higher than desirable, so the clinically relevant difference is adjusted upwards to justify
 12 the sample size. This is not usually a desirable procedure. In discussing this one should,
 13 however, free oneself of moralizing cant. If the only trial which circumstances permit
 14 one to run is a small one, then the choice is between a small trial or no trial at all. It is
 15 not always the case that under such circumstances the best choice, whether taken in
 16 the interest of drug development or of future patients, is no trial at all. It may be useful,
 17 however, to calculate the sort of difference which the trial is capable of detecting so
 18 that one is clear at the outset about what is possible. Under such circumstances, the
 19 value of Δ can be the determined function of α , β , σ and n and is then not so much the
 20 *clinically relevant* as the *detectable* difference. In fact there is a case for simply plotting
 21 for any trial the power function: that is to say, the power at each possible value of the
 22 clinically relevant difference. A number of such functions are plotted in Figure 13.1 for
 23 the trial in asthma considered in Example 13.1. (For the purposes of calculating the
 24 power in the graph, it has been assumed that a one-sided test at the 2.5% level will
 25 be carried out. For high values of the clinically relevant difference, this gives the same
 26 answer as carrying out a two-sided test at the 5% level. For lower values it is preferable
 27 anyway.)



45 **Figure 13.1** Power as a function of clinically relevant difference for a two-parallel-group trial
 46 in asthma. The outcome variable is FEV_1 , the standard deviation is assumed to be 450 ml, and n
 47 is the number of patients per group. If the clinically relevant difference is 200 ml, 80 patients per
 48 group are needed for 80% power.

01 **13.2.4 If we have performed a power calculation, then upon rejecting**
 02 **the null hypothesis, not only may we conclude that the**
 03 **treatment is effective but also that it has a clinically relevant**
 04 **effect**
 05

06 This is a surprisingly widespread piece of nonsense which has even made its way into
 07 one book on drug industry trials. Consider, for example, the case of a two parallel
 08 group trial to compare an experimental treatment with a placebo. Conventionally we
 09 would use a two-sided test to examine the efficacy of the treatment. (See Chapter 12
 10 for a discussion. The essence of the argument which follows, however, is unaffected by
 11 whether one-sided or two-sided tests are used.) Let τ be the true difference (experimental
 12 treatment–placebo). We then write the two hypotheses,

$$14 \quad H_0 : \tau = 0 \quad H_1 : \tau \neq 0. \quad (13.2)$$

15
 16 Now, if we reject H_0 , the hypothesis which we assert is H_1 , which simply states that
 17 the treatment difference is *not* zero or, in other words, that there is a difference between
 18 the experimental treatment and placebo. This is not a very exciting conclusion but it
 19 happens to be the conclusion to which significance in a conventional hypothesis test
 20 leads. As we saw in Chapter 12, however (see section 13.2.3), by observing the sign
 21 of the treatment difference, we are also justified in taking the further step of deciding
 22 whether the treatment is superior or inferior to placebo. A power calculation, however,
 23 merely takes a particular value, Δ , within the range of possible values of τ given by
 24 H_1 and poses the question: ‘if this particular value happens to obtain, what is the
 25 probability of coming to the correct conclusion that there is a difference?’ This does not
 26 at all justify our writing in place of (13.2),

$$27 \quad H_0 : \tau = 0 \quad H_1 : \tau = \Delta, \quad (13.3)$$

28
 29 or even

$$31 \quad H_0 : \tau = 0 \quad H_1 : \tau \geq \Delta. \quad (13.4)$$

32
 33 In fact, (13.4) would imply that we knew, before conducting the trial, that the treatment
 34 effect is either zero or at least equal to the clinically relevant difference. But where
 35 we are unsure whether a drug works or not, it would be ludicrous to maintain that
 36 it cannot have an effect which, while greater than nothing, is less than the clinically
 37 relevant difference.

38 If we wish to say something about the difference which obtains, then it is better to
 39 quote a so-called ‘point estimate’ of the true treatment effect, together with associated
 40 confidence limits. The point estimate (which in the simplest case would be the difference
 41 between the two sample means) gives a value of the treatment effect supported by the
 42 observed data in the absence of any other information. It does not, of course, have
 43 to obtain. The upper and lower $1 - \alpha$ confidence limits define an interval of values
 44 which, were we to adopt them as the null hypothesis for the treatment effect, would
 45 not be rejected by a hypothesis test of size α . If we accept the general Neyman–Pearson
 46 framework and if we wish to claim any single value as the proven treatment effect, then
 47 it is the lower confidence limit, rather than any value used in the power calculation,
 48 which fulfills this role. (See Chapter 4.)

13.2.5 We should power trials so as to be able to prove that a clinically relevant difference obtains

Suppose that we compare a new treatment to a control, which might be a placebo or a standard treatment. We could set up a hypothesis test as follows:

$$H_0 : \tau < \Delta \quad H_1 : \tau \geq \Delta. \quad (13.5)$$

H_0 asserts that the treatment effect is less than clinically relevant and H_1 that it is at least clinically relevant. If we reject H_0 using this framework, then, using the logic of hypothesis testing, we decide that a clinically relevant difference obtains. It has been suggested that this framework ought to be adopted since we are interested in treatments which have a clinically relevant effect.

Using this framework requires a redefinition of the clinically relevant difference. It is no longer 'the difference we should not like to miss' but instead becomes 'the difference we should like to prove obtains'. Sometimes this is referred to as the 'clinically irrelevant difference'. For example, as Cairns and Ruberg point out (Cairns and Ruberg, 1996; Ruberg and Cairns, 1998), the CPMP guidelines for chronic arterial occlusive disease require that, 'an irrelevant difference (to be specified in the study protocol) between placebo and active treatment can be excluded' (Committee for Proprietary Medicinal Products, 1995). In fact, if we wish to prove that an effect equal to Δ obtains, then unless for the purpose of a power calculation we are able to assume an alternative hypothesis in which τ is greater than Δ , the maximum power obtainable (for an infinite sample size) would be 50%. This is because, in general, if our null hypothesis is that $\tau < \Delta$, and the alternative is that $\tau \geq \Delta$, the critical value for the observed treatment difference must be greater than Δ . The larger the sample size the closer the critical value will be to Δ , but it can never be less than Δ . On the other hand, if the true treatment difference is Δ , then the observed treatment difference will be less than Δ in approximately 50% of all trials. Therefore, the probability that it is less than the critical value must be greater than 50%. Hence the power, which is the probability under the alternative hypothesis that the observed difference is greater than the critical value, must be less than 50%.

The argument in favour of this approach is clear. The conventional approach to hypothesis testing lacks ambition. Simply proving that there is a difference between treatments is not enough: one needs to show that it is important. There are, however, several arguments against using this approach. The first concerns active controlled studies. Here it might be claimed that all that is necessary is to show that the treatment is at least as good as some standard. Furthermore, in a serious disease in which patients have only two choices for therapy, the standard and the new, it is only necessary to establish which of the two is better, not by how much it is better, in order to treat patients optimally. Any attempt to prove more must involve treating some patients suboptimally and this, in the context, would be unacceptable.

A further argument is that a nonsignificant result will often mean the end of the road for a treatment. It will be lost for ever. However, a treatment which shows a 'significant' effect will be studied further. We thus have the opportunity to learn more about its effects. Therefore, there is no need to be able to claim on the basis of a single trial that a treatment effect is clinically relevant.

13.2.6 Most trials are unethical because they are too large

The argument is related to one in Section 13.2.5. If we insist on 'proving' that a new treatment is superior to a standard we shall study more patients than are necessary to obtain some sort of belief, even it is only a mere suspicion, that one or the other treatment is superior. Hence doctors will be prescribing contrary to their beliefs and this is unethical.

I think that for less serious non-life-threatening and chronic diseases this argument is difficult to sustain. Here the patients studied may themselves become the future beneficiaries of the research to which they contribute and, given informed consent, there is thus no absolute requirement for a doctor to be in equipoise. For serious diseases the argument must be taken more seriously and, indeed, sequential trials and monitoring committees are an attempt to deal with it. The following must be understood, however. (1) Whatever a given set of trialists conclude about the merits of a new treatment, most physicians will continue to use the standard for many years. (2) In the context of drug development, a physician who refuses to enter patients on a clinical trial because she or he is firmly convinced that the experimental treatment is superior to the standard treatment condemns *all* her or his patients to receive the standard. (3) if a trial stops before providing reasonably strong evidence of the efficacy of a treatment, then even if it looks promising, it is likely that collaborating physicians will have considerable difficulties in prescribing the treatment to future patients.

It thus follows that, on a purely logical basis at least, a physician is justified in continuing on a trial, even where she or he believes that the experimental treatment is superior. It is not necessary to start in equipoise (Senn, 2001a, 2002). The trial may then be regarded as continuing either to the point where evidence has overcome initial enthusiasm for the new treatment, so that the physician no longer believes in its efficacy, or to the point at which sceptical colleagues can be convinced that the treatment works. Looked at in these terms, few conventional trials would be too large.

13.2.7 Small trials are unethical

The argument here is that one should not ask patients to enter a clinical trial unless one has a reasonable chance of finding something useful. Hence small or 'inadequately powered' trials are unethical.

There is something in this argument. I do not agree, however, that small trials are uninterpretable and, as was explained in Section 13.2.2, sometimes only a small trial can be run. It can be argued that if a treatment will be lost anyway if the trial is not run, then it should be run, even if it is only capable of 'proving' efficacy where the treatment effect is considerable. Part of the problem with small trials is, to use Altman and Bland's memorable phrase, that 'absence of evidence is not evidence of absence' (Altman and Bland, 1995) and there is a tendency to misinterpret a nonsignificant effect as an indication that a treatment is not effective rather than as a failure to prove that it is effective. However, if this is the case, it is an argument for improving medical education, rather than one for abandoning small trials. The rise of meta-analysis has also meant that small trials are becoming valuable for the part which they are able to

contribute to the whole. As Edwards *et al.* have argued eloquently, some evidence is better than none (Edwards *et al.*, 1997).

Clinically relevant difference: Used in the theory of clinical trials as opposed to *cynically* relevant difference, which is used in practice.

13.2.8 A significant result is more meaningful if obtained from a large trial

It is possible to show, with an application of Bayes' theorem, that if we allow a certain prior probability that a product is effective, then the posterior probability of the effectiveness of the product, given a significant result, is an increasing function of the power of the test. Suppose, for example, that the prior odds for a given alternative hypothesis against the null hypothesis are pr_1/pr_0 . Let L_1 be the likelihood of observing a given piece of evidence under H_1 and L_0 be the likelihood under H_0 . Let po_1/po_0 be the posterior odds. Then Bayes' theorem implies (see Chapter 4) that

$$\frac{po_1}{po_0} = \frac{pr_1 L_1}{pr_0 L_0}. \quad (13.6)$$

If the evidence is that a result is significant at level α and the power for the given alternative is $1 - \beta$, then these are the two likelihoods associated with the evidence and we may write

$$\frac{po_1}{po_0} = \frac{pr_1 (1 - \beta)}{pr_0 \alpha} \quad (13.7)$$

Now, from (13.7) for given prior odds and fixed α , the posterior odds are greater the smaller the value of β , which is to say the greater the power of the test. But the power increases with sample size. Hence, other things being equal, significant results are more indicative of efficacy if obtained from large trials rather than small trials.

Although it is technically correct, one should be *extremely careful* in interpreting this statement, as will be shown below.

13.2.9 A given significant P -value is more indicative of the efficacy of a treatment if obtained from a small trial

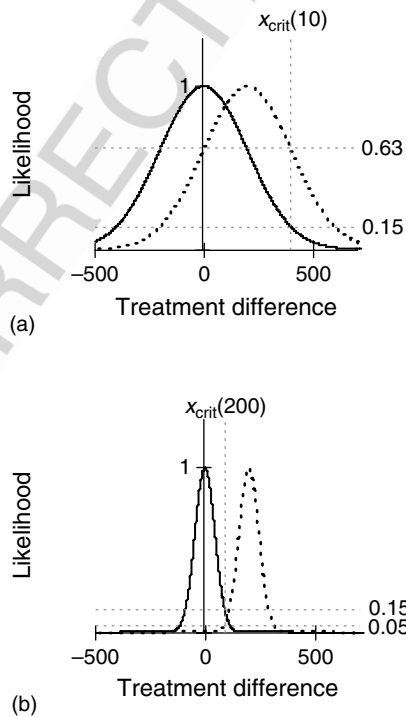
The surprising thing is that this statement can also be shown to be true given suitable assumptions (Royall, 1986). We talked above about the power of the alternative hypothesis, but typically this hypothesis includes all sorts of values of τ , the treatment effect. One argument is that if the sample size is increased, not only is the power of finding a clinically relevant difference, Δ , increased, but the power also of finding lesser differences. Another argument is as follows.

In general, we do not merely observe that a trial is significant or not significant. To do so is to throw information away. We shall observe an exact P -value. For example

01 we might say that the results were significant at the 5% level but in fact know that
 02 $P=0.028$. Hence, although (13.6) gives a general expression for the posterior odds, the
 03 particular application of it given by (13.7) is irrelevant. Instead of using the likelihood
 04 of significance we need to use the likelihoods $L_1(P)$ and $L_0(P)$ associated with the
 05 observed P -value. Hence we may write

$$\frac{po_1}{po_o} = \frac{pr_1 L_1(P)}{pr_o L_0(P)} \quad (13.8)$$

06
 07
 08
 09
 10 Now, whereas $(1 - \beta)/\alpha$ is a monotonically increasing function of the sample size
 11 n , $L_1(P)/L_0(P)$ is not. It may increase at first, but eventually it will decline. The situation
 12 is illustrated in Figure 13.2, which takes the particular example of the trial is asthma
 13 considered in section 13.1 above and shows the likelihood for the null hypothesis (scaled
 14 to equal 1 in the case where the observed treatment difference is zero) for all possible
 15 observed treatment differences and also for the alternative hypothesis where the true
 16 treatment effect is equal to the clinically relevant difference. The situations for trials
 17 with 10 and 200 patients per group are illustrated. The critical value of the observed
 18 treatment difference for a two-sided test at the 5% level is marked in each case. For the
 19 smaller trial, a larger difference is required for significance. In the larger trial, a smaller
 20 difference is adequate.



21
 22
 23
 24
 25
 26
 27
 28
 29
 30
 31
 32
 33
 34
 35
 36
 37
 38
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
Figure 13.2 Scaled likelihood for null and alternative hypotheses for trials with (a) 10 and (b) 200 patients per group. Under the alternative hypothesis the clinically relevant difference obtains.

01 Note that the scaled likelihood for H_0 is the same in both cases and approximately
 02 equal to 0.15. However, in the first case the likelihood is more than four times as high
 03 under H_1 , being equal to 0.63, and in the second case only a third as high under H_1 ,
 04 being equal to 0.05. Hence a given P -value of exactly 0.05 would provide evidence
 05 in favour of the alternative hypothesis in the first case and against it in the second.
 06 Thus, on this interpretation, moderate significance from a large trial would actually
 07 be evidence for the null hypothesis. (Note that if one is told only that the result is
 08 significant, it is the area under the curve to the right of the critical value which is
 09 relevant, and this is much higher in the larger trial. This was the situation considered
 10 in section 13.2.8 but is not the situation here.)

11 The fact that a conventionally significant result can give evidence in favour of the
 12 null hypothesis is known as the Jeffreys–Lindley paradox (Bartlett, 1957; Lindley, 1957;
 13 Senn, 2001b). In practice, such a situation would hardly ever arise. This is because
 14 a significant P -value is not very likely given the null hypothesis (this is the basic
 15 idea behind the significance test) and it is even less likely given the sort of alternative
 16 hypothesis illustrated. The point is rather that, given that this unusual situation has
 17 occurred, it actually gives more evidence in favour of the null hypothesis. It is one
 18 argument, once a trial has reached a certain power, to reduce the level of significance
 19 required.

20 Again, however, one has to be very careful in interpreting this result. The diagram
 21 only illustrates the alternative hypothesis corresponding to the clinically relevant
 22 difference. But if this is the difference we should not like to miss, it does not
 23 follow that it is the only difference we should like to find. There may be lower
 24 values of the treatment effect which are of interest and these will produce higher
 25 likelihoods.

26 27 28 **13.2.10 For a given P -value, the evidence against the null hypothesis is 29 the same whatever the size of the trial** 30

31 This has been referred to as the *alpha postulate*. There is a sense in which this is
 32 true also! Look at Figure 13.3. Here, instead of displaying the scaled likelihood for
 33 that alternative hypothesis which corresponds to the clinically relevant difference, the
 34 alternative hypothesis for which the true treatment effect corresponds to the critical
 35 value is illustrated. For $P = 0.05$, the ratio of this likelihood is the same for the trial
 36 with 10 patients per group as for the trial with 200 as, indeed, it is for any trial
 37 whatsoever. But we do not know which value of the alternative hypothesis is true if
 38 the null hypothesis is false. It thus follows that for a P -value of exactly 0.05, there is
 39 always one value of the alternative hypothesis for which the likelihood is $1/0.15$ or
 40 more than six times as high as for the null hypothesis. The evidence in favour of *this*
 41 alternative hypothesis (which hypothesis changes according to the size of the trial) is
 42 always the same.

43 This is one interpretation of the significance test. It is the sceptic's concession to
 44 the gullible. The sceptic asserts the null hypothesis, because he doesn't believe in the
 45 efficacy of treatment. The gullible believes exactly whatever the data tell him. He thus
 46 adopts as his presumed treatment effect whatever the observed mean difference is. Since
 47 the ratio of likelihoods in favour of this hypothesis will always be greater than 1, he
 48 will always obtain evidence in favour of his hypothesis. All that the sceptic can do is

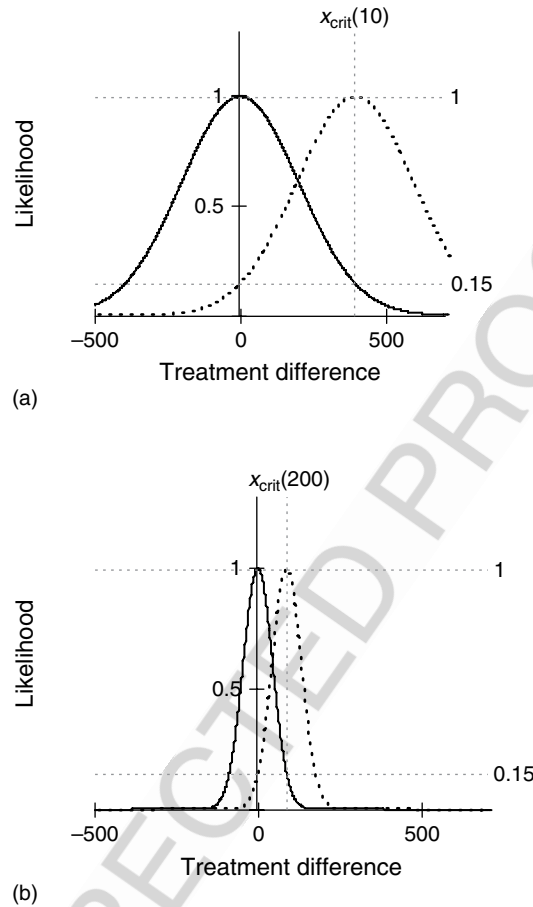


Figure 13.3 Scaled likelihood for null and alternative hypotheses for (a) 10 and (b) 200 patients per group. Under the alternative hypothesis the difference corresponding to the critical value of the test obtains.

counter this by explaining what the consequences of this sort of behaviour are. ‘If you regard odds of 6 to 1 for the best supported difference compared to the null hypothesis as being evidence in favour of a treatment effect, then in at least 1 in 20 trials where the treatment is useless you will conclude it is effective.’ On this interpretation, a conventionally significant result then becomes a *minimal* requirement for concluding efficacy of a treatment.

In my opinion, this is, the way that drug development usually works. Treatments are not registered unless a significant result is obtained. This is, however, generally a minimal requirement. Other information (treatment estimates, confidence intervals, the analyses of other outcomes, the results of further trials) is presented and unless this is favourable, the drug will not be registered.

The above discussions should serve as a warning, however: the evidential interpretations of *P*-values is a delicate matter (Senn, 2001b).

13.2.11 The effect of the two-trials rule on sample size

If we are required to prove significance in two trials, as is usually believed to be necessary in phase III for a successful NDA application, then from the practical point of view it may be the power of the combined requirement which is important, rather than that for individual trials. On the assumption that the trials are competent and that the background planning has been carried out appropriately and that trial by treatment interactions may be dismissed, then the power to detect the clinically relevant difference in both of two trials, each with power 80%, is 64% since $0.8 \times 0.8 = 0.64$. This means that $1 - 0.64 = 0.36$, or more than one-third, of *such* drug development programmes would fail in phase III for failure of one or both clinical trials on this basis. (This does not mean that one-third of drug developments will fail in phase III, since many drugs which survive that far may, indeed, have an effect which is superior to the clinically relevant difference. On the other hand, there may be other reasons for failure.) If it is desired to have 80% power overall, then it is necessary to run each trial with 90% power since $0.9 \times 0.9 \cong 0.80$.

As explained in Chapter 12, however, the two-trials rule is not particularly logical. The alternative pooled-trials rule would actually require only 4/5 of the number of patients of the two-trials rule for an overall power of 80% and an overall size of 1/1600, which is what, effectively, the two-trials rule implies.

Clinically relevant difference: That which is used to justify the sample size but will be claimed to have been used to find it.

13.2.12 The effect of multiple requirements

Cairns and Ruberg point out that the FDA requires that for registration of treatments for senile dementia, both global cognitive function and overall assessment by physician must be significantly superior to placebo (Cairns and Ruberg, 1996). In Europe, a third additional requirement for 'activities of daily living' is made. This means that to plan a trial of adequate sample size to meet these requirements one would need to use two (three in Europe) clinically relevant differences. One would also have to know the correlation between the repeat measures. A conservative approach would be to assume independence. In that case a sample size determination could be made for each measure for 90% power (USA) or 93% power (Europe) and the largest of all of these determinations taken. (The European requirement would arise because $0.93 \times 0.93 \times 0.93 \cong 0.80$). Comparing the resulting sample size to the largest that would be required for any one of the outcomes taken singly with 80% power, this leads to a 31% increase for the USA and a 46% increase for Europe. Of course, this is extreme. In practice these outcomes will be correlated and better approaches can be used. Nevertheless there will be an increase in sample size required. This issue is discussed in some detail by Kieser *et al.* (2004).

Most statisticians are, of course, aware of the point and will be concerned about requirements for multiple outcomes. What is sometimes overlooked, however, is that a similar point arises if conclusions are required to be robust to a number of

01 analyses: the Mann–Whitney–Wilcoxon test as well as the t -test, for example, or
 02 including or omitting various covariates. Such multiple requirements also lead to an
 03 increase in the sample size needed. (Although, at a guess, the problem will not be as
 04 severe as for multiple outcomes.)

05 Such multiple requirements seem from one point of view to be a good thing. A stronger
 06 standards of evidence overall is implicit and in the end we shall have greater confidence
 07 in the value of a registered drug. By the same token, however, the ethical problems
 08 which arise through sample size requirements can be aggravated by such requirements.
 09 In any case, the practical consideration is that the pharmaceutical statistician has to
 10 become aware of the requirements of **conjunctive power**, that is to say the need for
 11 test 1 *and* test 2 *and* test 3, etc. (hence conjunctive) to be significant.

12 There is also one technical matter which is worth comment. Suppose we require
 13 that a t -test and a Mann–Whitney–Wilcoxon test both be significant at the 5% level.
 14 Let us call this the two-test procedure. These two tests are certainly highly correlated
 15 so that the overall type I error rate will be somewhere between 0.05 (the value were
 16 they perfectly correlated) and 0.0025 (the value were they independent). Suppose, for
 17 argument's sake, that the value is 0.01. I suspect that for moderately sized trials the
 18 loss in power as a result of the double requirement, where Normality applies and either
 19 test could be used, will be small. Nevertheless, there will be some loss. Of course, there
 20 is the gain that a higher standard of evidence of efficacy is required. However, a higher
 21 standard of efficacy could be obtained simply by requiring that a t -test on its own be
 22 significant at the 1% level. It would be interesting to see this requirement compared
 23 formally with that of the two-test procedure.

24 For further discussion of the effect of multiple outcomes on power see Chuang-Stein
 25 *et al.* (2007), Offen *et al.* (2007) and Senn and Bretz (2007).

28 **13.2.13 In order to interpret a trial it is necessary to know its power**

30 This is a rather silly point of view that nevertheless continues to attract adherents. A
 31 power calculation is used for planning trials and is effectively superseded once the data
 32 are in. For an impression of the precision of the result, one is best looking at confidence
 33 intervals or standard errors. If the result is significant, then to the extent that one
 34 accepts the logic of significance tests, there is no point arguing about the result. An
 35 analogy may be made. In determining to cross the Atlantic it is important to consider
 36 what size of boat it is prudent to employ. If one sets sail from Plymouth and several
 37 days later sees the Statue of Liberty and the Empire State Building, the fact that the
 38 boat employed was rather small is scarcely relevant to deciding whether the Atlantic
 39 was crossed.

40 Retrospective power calculations are sometimes encountered for so-called failed trials,
 41 but this seems particularly pointless. The clinically relevant difference does not change
 42 as a result of having run the trial, in which case the power is just a function of the
 43 observed variance. It says nothing about the effect of treatment.

44 Some check-lists for clinical trials seem to require evidence that a power calculation
 45 was performed, but this surely can have very little relevance to the interpretation of the
 46 final result. At best it can be a very weak indicator of the quality of the study.

47 For a trenchant criticism of the use of retrospective power calculations see Hoenig
 48 and Heisey (2001).

01 **13.2.14 The dimension of cost**

02

03 A very unsatisfactory feature of conventional approaches to sample size calculation is
04 that there is no mention of cost. This means that for any two quite different indications
05 with the same effect size, that is to say the same ratio of clinically relevant difference to
06 standard deviation, the sample size would be the same whatever the cost or difficulty of
07 recruiting and treating patients. This is clearly illogical and trialists probably manage
08 this issue informally by manipulating the clinically relevant difference in way discussed
09 in Section 13.2.3. Clearly, it would be better to include the cost explicitly, and this
10 suggests decision-analytic approaches to sample size determination. There are various
11 Bayesian suggestions and these will be discussed in the next section.

12

13 **13.2.15 Bayesian approaches to sample size determination**

14

15 A Bayesian approach to sample size determination may be discussed in terms of a rather
16 artificial and not very realistic case, namely that where a single phase III study must be
17 run in a nonsequential manner and on the basis of this the treatment will be registered
18 or not. Now consider a trial of a given sample size and suppose that a decision will be
19 made on the basis of a statistic from such a trial. Suppose that it is known for every
20 value of the statistic whether the regulator will register the drug or not and that a loss
21 may be associated with failing to register. Of course, the regulator's decision might itself
22 be based on a suitable loss function and prior distribution, but from the sponsor's point
23 of view this does not matter provided only that the decision that will be taken is known.

24 Now, given a suitable prior distribution (which does not have to be the same as the
25 regulator's), the sponsor can calculate the predictive distribution for the test statistic and
26 hence the expected loss for any given sample size, including the cost of experimentation.
27 The optimal sample size is then the one with the smallest expected loss. This is a double
28 optimization procedure: the optimal decision (minimum loss) for a given value of the
29 test statistic and sample size must be determined and then the sample size which yields
30 the smallest minimum loss is chosen (Lindley, 1997).

31 Various Bayesian approaches that have been suggested are variants of this. For
32 example, Lindley's approach (Lindley, 1997) involves a sophisticated use of loss func-
33 tions and is fully Bayesian but applies when regulator and sponsor have the same beliefs
34 and values. That of Gittins and Pezeshk is a hybrid Bayes-frequentist system in which
35 it is supposed that eventual sales of a pharmaceutical are a function of how impres-
36 sive the trial results are (measured by conventional significance) but that in planning,
37 prior distributions and a Bayesian approach are used (Gittins and Pezeshk, 2000a,b;
38 Pezeshk and Gittins, 2002). For an implementation in SAS of a simpler approach with
39 costs per patient but a single reward for a significant trial, see Burman *et al.* (2007).
40 This uses a prior distribution on the treatment effect. Application of this in a hybrid
41 Bayesian/frequentist approach is discussed in Section 13.2.16.

42

43 **13.2.16 An appropriate approach to sample size determination is to
44 calculate assurance**

45

46 Even closer to a frequentist approach than the methods of Section 13.2.15 is that of
47 calculating what O'Hagan *et al.* call **assurance** (O'Hagan *et al.*, 2005). This is the
48

01 Bayesian probability of a clinical trial yielding a significant result. Rather than being
 02 conditional on a particular posited clinically relevant difference, it uses a prior distri-
 03 bution for the treatment effect and integrates the conventional frequentist power over
 04 this distribution to obtain an unconditional expected power, assurance. If this approach
 05 alone is used for sample size determination it means that the clinically relevant differ-
 06 ence has no role whatsoever in determining how large the trial should be. As suggested
 07 in Section 13.2.14, conventional power calculations are already unsatisfactory in that
 08 they treat identically two indications with differing costs but identical effect sizes. Assur-
 09 ance, however, seems to take this one stage further. Provided that the prior distributions
 10 for treatment effects for two indications are the same and the precision is the same,
 11 then if assurance is to be the guide, the sample sizes will be the same, even if in terms
 12 of practical interest the prior distributions are quite different.

13 For this and other reasons, I am not particularly keen on the use of assurance as
 14 the primary criterion for designing a clinical trial, although it may well be useful to
 15 calculate it in addition. My own view is that if I am going to be Bayesian about sample
 16 size calculation I would rather be hanged for a sheep than a lamb and use the methods
 17 of section 13.2.15.

18 19 20 References

- 21 Altman DG, Bland JM (1995) Absence of evidence is not evidence of absence. *British Medical*
 22 *Journal* **311**: 485.
- 23 Bartlett MS (1957) A comment on D.V. Lindley's statistical paradox. *Biometrika* **44**: 533–534.
- 24 Burman C-F, Grieve AP, Senn S (2007) Decision analysis in drug development. In: Dmitrienko
 25 A, Chuang-Stein C, Agostino R (eds), *Pharmaceutical Statistics Using SAS: A Practical Guide*. SAS
 26 Institute, Cary, pp. 385–428.
- 27 Cairns V, Ruberg S (1996) The confirmatory package of trials–design. In: Jones B, Teather B,
 28 Teather D (eds), *Proceedings of Statistical Issues in Biopharmaceutical Environments: USA and*
 29 *European Perspectives*, De Monfort University, Leicester.
- 30 Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W (2007) Challenge of multiple co-primary
 31 endpoints: a new approach. *Statistics in Medicine* **26**: 1181–1192.
- 32 Committee for Proprietary Medicinal Products (1995) *Note for guidance on the clinical investigation*
 33 *of medicinal products in the treatment of chronic peripheral arterial occlusive disease*. EMEA, London.
 34 <http://www.emea.europa.eu/pdfs/human/ewp/071498eu.pdf>.
- 35 Desu MM, Raghavarao D (1990) *Sample Size Methodology*. Academic Press, Boston.
- 36 Edwards SJ, Lilford RJ, Braunholtz D, Jackson J (1997) Why 'underpowered' trials are not neces-
 37 sarily unethical [see comments]. *Lancet* **350**: 804–807.
- 38 Gittins J, Pezeshk H (2000a) A behavioral Bayes method for determining the size of a clinical
 39 trial. *Drug Information Journal* **34**: 355–363.
- 40 Gittins J, Pezeshk H (2000b) How large should a clinical trial be? *Journal of the Royal Statistical*
 41 *Society Series D–The Statistician* **49**: 177–187.
- 42 Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations
 43 for data analysis. *American Statistician* **55**: 19–24.
- 44 Julious SA (2004) Tutorial in biostatistics–Sample sizes for clinical trials with normal data.
 45 *Statistics in Medicine* **23**: 1921–1986.
- 46 Julious SA (2006) *Designing clinical trials with uncertain estimates of variability*. PhD, University
 47 College London, London.
- 48 Kieser M, Röhm J, Friede T (2004) Power and sample size determination when assessing the
 clinical relevance of trial results by 'responder analyses'. *Statistics in Medicine* **23**: 3287–3305.
- Lindley DV (1957) A statistical paradox. *Biometrika* **44**: 187–192.

212 **Determining the Sample Size**

01 Lindley DV (1997) The choice of sample size. *Statistician* **46**: 129–138.
02 O’Hagan A, Stevens JW, Campbell MJ (2005) Assurance in clinical trial design. *Pharmaceutical*
03 *Statistics* **4**: 186–201.
04 Offen W, Chuang-Stein C, Dmitrienko A, *et al.* (2007) Multiple co-primary endpoints: medical and
05 statistical solutions—A report from the Multiple Endpoints Expert Team of the Pharmaceutical
06 Research and Manufacturers of America. *Drug Information Journal* **41**: 31–46.
07 Pezeshk H, Gittins J (2002) A fully Bayesian approach to calculating sample sizes for clinical
08 trials with binary responses. *Drug Information Journal* **36**: 143–150.
09 Royall RM (1986) The effect of sample size on the meaning of significance tests. *The American*
10 *Statistician* **40**: 313–315.
11 Ruberg S, Cairns V (1998) Providing evidence of efficacy for a new drug. *Statistics in Medicine*
12 **17**: 1813–1823.
13 Senn SJ (2001a) The misunderstood placebo. *Applied Clinical Trials* **10**: 40–46.
14 Senn SJ (2001b) Two cheers for *P*-values. *Journal of Epidemiology and Biostatistics* **6**: 193–204.
15 Senn SJ (2002) Ethical considerations concerning treatment allocation in drug development trials.
16 *Statistical Methods in Medical Research* **11**: 403–411.
17 Senn SJ, Bretz F (2007) Power and sample size when multiple endpoints are considered. *Pharma-*
18 *ceutical Statistics*. [In press].
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48