sense to test the null hypothesis; rather the objective may be to estimate the magnitude of the effect and to do this with some acceptable specified precision.

In trials of new interventions, it is often not known whether there will be any impact at all of the intervention on the outcomes of interest, and what is required is 'proof of concept'. In these circumstances, it may be sufficient to ensure that there will be a good chance of obtaining a significant result if there is indeed an effect of some specified magnitude. It should be emphasized, however, that, if this course is adopted, the estimates obtained may be very imprecise. To illustrate this, suppose it is planned to compare two groups with respect to the mean of some variable, and suppose the true difference between the group means is D. If the trial size is chosen to give 90% power (of obtaining a significant difference with $p < 0.05$ on a two-sided test) if the difference is D, the 95% CI on D is expected to extend roughly from 0.4 D to 1.6 D. This is a wide range and implies that the estimate of the effect of intervention will be imprecise. In many situations, it may be more appropriate to choose the sample size by setting the width of the CI, rather than to rely on power calculations.

2.4 Trials with multiple outcomes

The discussion in Sections 2.1 to 2.3 concerns factors influencing the choice of trial size, with respect to a particular outcome measure. In most trials, several different outcomes are measured. For example, in a trial of the impact of insecticide-treated mosquito-nets on childhood malaria, there may be interest in the effects of the intervention on deaths, deaths attributable to malaria, episodes of clinical malaria, spleen sizes at the end of the malaria season, PCVs at the end of the malaria season, and possibly other measures.

Chapter 12, Section 2 highlights the importance of defining in advance the *primary* outcome and a limited number of *secondary* outcomes of a trial. In order to decide on the trial size, the investigator should first focus attention on the primary outcome, as results for this outcome will be given the most weight when reporting the trial findings, and it is essential that the trial is able to provide adequate results for this outcome. The methods of this chapter can then be used to calculate the required trial size for the primary outcome and each of the secondary outcomes.

Ideally, the outcome that results in the largest trial size would be used to determine the size, as then, for other outcomes, it would be known that better than the required precision or power would be achieved. It is often found, however, that one or more of the outcomes would require a trial too large for the resources that are likely to be available. For example, detecting changes in mortality, or cause-specific mortality, often requires very large trials. In these circumstances, it may be decided to design the trial to be able to detect an impact on morbidity and accept that it is unlikely to be able to generate conclusive findings about the effect on mortality. It is important to point out, however, that, if a trial shows that an intervention has an impact on morbidity, it may be regarded as unethical to undertake a further, larger trial to assess the impact on mortality. For this reason, it is generally advisable to ensure that trials are conducted at an early stage in which the outcome of greatest public health importance is the endpoint around which the trial is planned. This issue is discussed further in Chapter 6.

Sometimes, different trial sizes may be used for different outcomes. For example, it might be possible to design a trial in such a way that a large sample of participants are monitored for mortality, say by annual surveys, and only a proportion of participants are monitored for morbidity, say by weekly visits.

If it is not feasible to design the trial to achieve adequate power or precision for the primary outcome, the trial should either be abandoned or a different primary outcome should be adopted.

2.5 Practical constraints

In practice, statistical considerations are not the only factors that need to be taken into account in planning the size of an investigation. Resources, in terms of staff, vehicles, laboratory capacity, time, or money, may limit the potential size of a trial, and it is often necessary to compromise between the results of the trial size computations and what can be managed with the available resources. Trying to do a trial that is beyond the capacity of the available resources is likely to be unfruitful, as data quality is likely to suffer and the results may be subject to serious bias, or the trial may even collapse completely, wasting the effort and money that have already been expended. If calculations indicate that a trial of manageable size will yield power and/or precision that is unacceptably low, it is probably better not to conduct the trial at all.

A useful approach to examine the trade-off between trial size (and thus cost) and power is to construct *power curves* for one or two of the key outcome variables. Power curves show how power varies with trial size for different values of the effect measure. Figure 5.1 shows power curves for malaria deaths in the mosquito-net trial discussed in Section 2.4, assuming that equal numbers of children are to be allocated to the intervention and control groups and statistical significance is to be based on a two-sided test

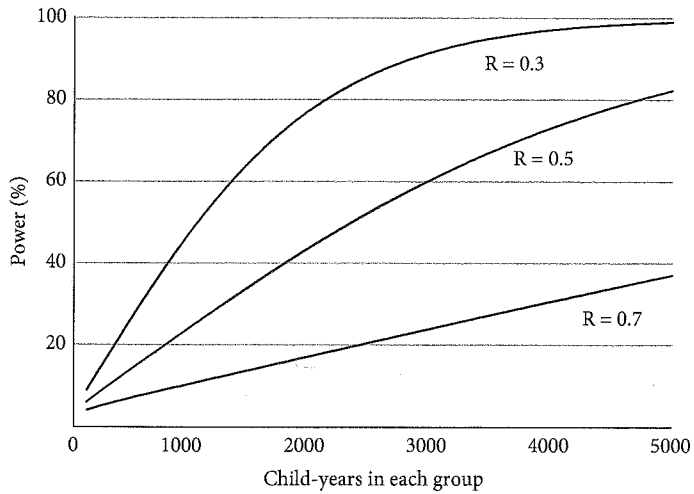

Figure 5.1 Power curves for a trial of the effect of mosquito-nets on malaria deaths.

Malaria death rate in the control group assumed to be 10/1000/year. R, relative rate in the intervention group. Assumes equal-sized groups, two-sided test, and significance $p < 0.05$.

Table 5.1 Relationship between z_2 and % power (numbers in the body of the table show power corresponding to each value of z_2)

	First decimal place of z_2									
z_2	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
−3.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−2.0	2.3	1.8	1.4	1.1	0.8	0.6	0.5	0.3	0.3	0.2
−1.0	15.9	13.6	11.5	9.7	8.1	6.7	5.5	4.5	3.6	2.9
−0.0	50.0	46.0	42.1	38.2	34.5	30.9	27.4	24.2	21.2	18.4
+0.0	50.0	54.0	57.9	61.8	65.5	69.1	72.6	75.8	78.8	81.6
+1.0	84.1	86.4	88.5	90.3	91.9	93.3	94.5	95.5	96.4	97.1
+2.0	97.7	98.2	98.6	98.9	99.2	99.4	99.5	99.7	99.7	99.8
+3.0	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: for example, $z_2 = -0.7$ corresponds to a power of 24.2%.

Having specified these values, the formulae or tables given in Sections 4.1 to 4.3 can be used to calculate the required trial size.

It is often useful, however, to proceed in the opposite direction, i.e. to explore the power that would be achieved for a range of possible trial sizes and for a range of possible values of the true difference D. This enables the construction of *power curves*, as illustrated in Figure 5.1. Formulae for this approach are also given in Sections 4.1 to 4.3.

4.1 Comparison of proportions

The trial size required in each group to detect a specified difference $D = p_1 - p_2$, with power specified by z_2 and significance level specified by z_1, is given by:

$$n = \left[(z_1 + z_2)^2 \, 2p(1-p)\right] / (p_1 - p_2)^2$$

where p is the average of p_1 and p_2.

For 90% power and significance at $p < 0.05$, this simplifies to:

$$n = \left[21p(1-p)\right] / (p_1 - p_2)^2.$$

Table 5.2 shows the required trial size for a range of values of p_1 and p_2 for 80%, 90%, or 95% power.

To calculate the power of a trial of specified size, calculate as follows, and refer the value of z_2 to Table 5.1.

$$z_2 = \left(\sqrt{\{n / [2p(1-p)]\}}\right)(|p_1 - p_2|) - z_1.$$

Example: assume that the spleen rate in the control group of the mosquito-net trial is around 40%. To have very high power (say 95%) of detecting a significant effect if the

Table 5.2 Sample size requirements for comparison of proportions

Smaller prop. p_1	Difference $D = p_2 - p_1$											
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
0.05	435	141	76	50	36	28	22	18	15	13	11	10
	583	189	102	67	48	37	30	25	21	18	15	13
	719	233	126	83	60	46	37	30	26	22	19	16
0.10	686	200	101	63	44	33	26	21	17	14	12	10
	919	268	135	84	59	44	34	28	23	19	16	14
	1134	330	166	104	72	54	42	34	28	24	20	17
0.15	906	251	122	74	50	37	28	22	18	15	13	10
	1212	336	163	98	67	49	38	30	24	20	17	14
	1497	415	201	122	83	60	46	37	30	25	21	18
0.20	1094	294	139	82	55	40	30	24	19	16	13	11
	1464	394	186	110	74	53	40	31	25	21	17	15
	1808	486	230	136	91	66	50	39	31	26	21	18
0.25	1250	329	153	89	59	42	31	24	19	16	13	11
	1674	441	205	119	79	56	42	32	26	21	17	14
	2067	544	253	147	97	69	52	40	32	26	21	18
0.30	1376	357	163	94	61	43	32	24	19	16	13	10
	1842	478	219	126	82	58	43	33	26	21	17	14
	2274	590	270	156	101	71	53	40	32	26	21	17
0.35	1470	376	170	97	63	44	32	24	19	15	12	10
	1968	504	228	130	84	58	43	32	25	20	16	13
	2430	622	282	160	103	72	53	40	31	25	20	16
0.40	1533	388	174	98	63	43	31	24	18	14	11	
	2052	520	233	131	84	58	42	31	24	19	15	
	2534	642	287	162	103	71	52	39	30	24	19	
0.45	1564	392	174	97	61	42	30	22	17	13		
	2094	525	233	130	82	56	40	30	23	18		
	2586	648	287	160	101	69	50	37	28	22		
0.50	1564	388	170	94	59	40	28	21	15			
	2094	520	228	126	79	53	38	28	21			
	2586	642	282	156	97	66	46	34	26			
0.55	1533	376	163	89	55	37	26	18				
	2052	504	219	119	74	49	34	25				
	2534	622	270	147	91	60	42	30				

Table 5.2 (continued) Sample size requirements for comparison of proportions

Smaller prop. p_1	Difference $D = p_2 - p_1$											
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
0.60	1470	357	153	82	50	33	22					
	1968	478	205	110	67	44	30					
	2430	590	253	136	83	54	37					
0.65	1376	329	139	73	44	28						
	1842	441	186	98	59	37						
	2274	544	230	121	72	46						
0.70	1250	294	122	63	36							
	1674	394	163	84	48							
	2067	486	201	104	60							
0.75	1094	251	101	50								
	1464	336	135	67								
	1808	415	166	83								
0.80	906	200	76									
	1212	268	102									
	1497	330	126									
0.85	686	141										
	919	189										
	1134	233										
0.90	435											
	583											
	719											

Shown in the body of the table are the sample sizes required in each group to give the specified power.*

* Upper figure: power, 80%; middle figure: power, 90%; lower figure: power, 95%. Using a two-sided significance test with $p<0.05$. The two groups are assumed to be of equal size.

intervention reduces the spleen rate to 30% (so that $p = 0.35$), the number of children required in each group is given by:

$$n=\left[(1.96+1.64)^2(2\times0.35\times0.65)\right]/(0.3-0.4)^2=590.$$

If the true risk ratio is R and we wish to power the trial, such that the lower confidence limit on the risk ratio will be greater than or equal to R_L, where R_L is the lowest acceptable efficacy (say, for whether or not to implement the intervention in a public health system, i.e. we need to be sure that the efficacy is at least R_L), the required sample size is:

$$n=(z_1+z_2)^2\left[(1-p_1)/(p_1)+(1-p_2)/(p_2)\right]/\left[\log_e(R/R_L)\right]^2.$$

4.2 Comparison of incidence rates

For a specified difference $D=r_1-r_2$ and values of z_1 and z_2, representing the required significance level and power, the required number of person-years in each group is given by:

$$y=\left[(z_1+z_2)^2(r_1+r_2)\right]/(r_1-r_2)^2$$

where r_1 and r_2 are the expected rates per person-year in the two groups. A rough estimate of the average of the two rates is therefore required, i.e. $\left[(r_1+r_2)/2\right]$. For 90% power and significance at $p < 0.05$, this formula simplifies to:

$$y=\left[10.5(r_1+r_2)\right]/(r_1-r_2)^2.$$

An alternative, but equivalent, formula gives the number of events required in group 2, the control group, in terms of the rate ratio R, for which the specified power is required:

$$e_2=\left[(z_1+z_2)^2(1+R)\right]/(1-R)^2.$$

This formula was used to construct Table 5.3, which shows the number of events needed in group 2 to detect a rate ratio of R with 80%, 90%, or 95% power. The total number of events needed in both groups can be calculated as $e_2(1+R)$. Since this can be computed without specifying the assumed rates in the two trial groups, this provides a particularly helpful approach when the rates are uncertain. Thus, in an *endpoint-driven trial*, we can specify the number of events that need to be observed to reach the required power, after which recruitment or follow-up may be terminated.

To calculate the power for a given trial size, compute:

$$z_2=\left\{\sqrt{\left[n/(r_1+r_2)\right]}\right\}\left(|r_1-r_2|\right)-z_1$$

where $|r_1-r_2|$ is the absolute value of the difference between the two rates.

Refer the resulting value of z_2 to Table 5.1 to determine the power of the trial.

Example: Assume, in the mosquito-net trial, that the death rate from malaria in the control group is 10/1000 child-years, so that $r_2=0.010$. Eighty per cent power is wanted to detect a significant effect if the true rate in children with bed-nets is reduced by 70% to $r_1=0.003$. The number of child-years of observation required in each group is given by:

$$y=\left[(1.96+0.84)^2(0.003+0.010)\right]/(-0.007)^2=2080.$$

The power curves shown in Figure 5.1 were constructed using the same assumption concerning the death rate in controls. For example, with $y = 2000$ and a rate ratio of $R=0.7$ (corresponding to a death rate of 7 per 1000 child-years in the intervention group), giving a power of 18% (Table 5.1):

$$z_2=\left\{\sqrt{\left[2000/(0.007+0.010)\right]}\right\}\left(|0.007-0.010|\right)-1.96=-0.93.$$

These formulae are used to ensure that there is a high probability of rejecting the

Table 5.3 Sample size requirements for comparison of rates

Relative rate *R**	Expected events in group 2 to give+		
	80% power	90% power	95% power
0.1	10.6	14.3	17.6
0.2	14.7	19.7	24.3
0.3	20.8	27.9	34.4
0.4	30.5	40.8	50.4
0.5	47.0	63.0	77.8
0.6	78.4	105.0	129.6
0.7	148.1	198.3	244.8
0.8	352.8	472.4	583.2
0.9	1489.6	1994.5	2462.4
1.1	1646.4	2204.5	2721.6
1.2	431.2	577.4	712.8
1.4	117.6	157.5	194.4
1.6	56.6	75.8	93.6
1.8	34.3	45.9	56.7
2.0	23.5	31.5	38.9
2.5	12.2	16.3	20.2
3.0	7.8	10.5	13.0
5.0	2.9	3.9	4.9
10.0	1.1	1.4	1.8

Numbers in the body of the table are expected number of events required in group 2 to give specified power if relative rate in group 1 is *R*.
* *R*, ratio of incidence rate in group 1 to incidence rate in group 2.
+ Using a two-sided significance test with P < 0.05. The two groups are assumed to be of equal size.

that the lower confidence limit for the effect size is close to the null, and this may provide insufficient evidence to recommend widespread adoption of the intervention. A larger sample size will be needed to ensure that the lower confidence limit exceeds a given value.

Suppose the assumed value of the rate ratio is R and that we wish to power the trial so that there is a high probability that the CI excludes a value R_L corresponding to the lower limit of efficacy desired. Then the required sample size is given by the formula:

$$y=(z_1+z_2)^2(1/r_1+1/r_2)/\left[\log_e(R/R_L)\right]^2.$$

Example: In the mosquito-net trial, we found that 2080 child-years were required in each trial group to reject the null hypothesis with 80% power if the true rate ratio R was

0.3, corresponding to an efficacy of 70%. Now suppose we wish to ensure that there is an 80% chance that the lower 95% CI for the efficacy exceeds 30%, corresponding to $R_L=0.7$. Applying the formula, we obtain the following, demonstrating the substantial increase in sample size that this would necessitate:

$$y=(1.96+0.84)^2(1/0.010+1/0.003)/\left[\log_e(0.3/0.7)\right]^2=4732.$$

4.3 Comparison of means

The trial size required in each group to detect a specified difference $D=\mu_1-\mu_2$, with power specified by z_2 and the significance level specified by z_1, is given by:

$$n=\left[(z_1+z_2)^2(\sigma_1^2+\sigma_2^2)\right]/(\mu_1-\mu_2)^2$$

where σ_1 and σ_2 are the standard deviations of the outcome variable in groups 1 and 2, respectively.

For 90% power and significance at $p<0.05$, this simplifies to:

$$n=10.5(\sigma_1^2+\sigma_2^2)/(\mu_1-\mu_2)^2.$$

To calculate the power of a trial of specified size, calculate the following, and refer the value of z_2 to Table 5.1:

$$z_2=\left\{\sqrt{\left[n/(\sigma_1^2+\sigma_2^2)\right]}\right\}(|\mu_1-\mu_2|)-z_1.$$

Estimates of σ_1 and σ_2 may be obtained from previous studies or from a pilot study. If appropriate values cannot be determined, an alternative is to dichotomize the continuous outcome variable and use the sample size formulae for comparison of proportions given in Section 4.1. This will give a conservative estimate of sample size, as it ignores some of the information, but will ensure an adequate sample size in the face of uncertainty regarding the standard deviations.

Example: In the mosquito-net trial, the mean PCV in the control group at the end of the trial is expected to be 33.0, with a standard deviation of 5.0. To have 90% power of detecting a significant effect if the intervention increases the mean PCV by 1.5, the number of children required in each group is given by:

$$n=\left[(1.96+1.28)^2(5.0^2+5.0^2)\right]/(1.5)^2=233.$$

Suppose it turns out that only 150 children are available for study in each group. The power in these circumstances is given by the following, corresponding to a power of about 74%:

$$z_2=\left\{\sqrt{\left[150/(5.0^2+5.0^2)\right]}\right\}(|1.5|)-1.96=0.64.$$

A summary of the various formulae that have been given for calculating the trial size requirements for the comparison of two groups of equal size is given in Table 5.4.

Table 5.4 Summary of formulae for calculating trial size requirements for comparison of two groups of equal size

Type of outcome	Formula	Notation	Section in text
A: Choosing trial size to achieve adequate precision			
Proportions:	$n = (1.96/\log_e f)^2 \{[(R+1)/(Rp_2)]-2\}$	n = number in each group R = prop. in group 1/prop. in group 2 Gives 95% CI from R/f to Rf	3.1
Rates:	$e_2 = (1.96/\log_e f)^2 [(R+1)/R]$	e_2 = expected events in group 2 R = rate in group 1/rate in group 2 Gives 95% CI from R/f to Rf	3.2
Means:	$n = (1.96/f)^2 (\sigma_1^2 + \sigma_2^2)$	n = number in each group σ_i = SD in group i D = mean in group 1 – mean in group 2 Gives 95% CI of $D \pm f$	3.3
B: Choosing trial size to achieve adequate power			
Proportions:	$n = [(z_1+z_2)^2 2p(1-p)]/(p_1-p_2)^2$	n = number in each group p_i = proportion. in group i p = average of p_1 and p_2	4.1
Rates:	$y = [(z_1+z_2)^2 (r_1+r_2)]/(r_1-r_2)^2$	y = person-years in each group r_i = rate in group i	4.2
Means:	$n = [(z_1+z_2)^2 (\sigma_1^2+\sigma_2^2)]/(\mu_1-\mu_2)^2$	n = number in each group σ_i = SD in group i μ_i = mean in group i	4.3

$z_1 = 1.96$ for significance at $p < 0.05$.
Power 80%, 90%, 95%
$z_2 = 0.84, 1.28, 1.64$.

5 More complex designs

5.1 Two groups of unequal size

Sections 3 and 4 considered the simplest situation where the two groups to be compared are of equal size. Sometimes, there may be reasons for wishing to allocate more individuals to one group than to the other. For example, if an experimental drug is very expensive, it may be desired to minimize the number of patients allocated to the

Table 5.5 Trial size necessary to achieve approximately the same power in a trial with two groups, one of which contains k times as many individuals as the other

k	n_1	n_2	n_1+n_2
1	n	n	$2n$
2	$0.75n$	$1.5n$	$2.25n$
3	$0.67n$	$2.0n$	$2.67n$
4	$0.62n$	$2.5n$	$3.12n$
5	$0.60n$	$3.0n$	$3.60n$
10	$0.55n$	$5.5n$	$6.05n$
100	$0.50n$	$50.0n$	$50.50n$

drug, and so the trial may be arranged so that there are two or three patients given the old drug for every patient given the new drug. In order to maintain the same power as in the equal allocation scheme, a larger total trial size will be needed, but the number given the new drug will be smaller. Conversely, in a trial of a new vaccine, it may be decided to allocate twice as many participants to the vaccinated group as are included in the placebo group, in order to increase the size of the safety database for the new vaccine, before it goes into public health programmes.

Let the size of the smaller of the two groups be n_1, and suppose the ratio of the two sample sizes to be k, so that there will be kn_1 individuals in the other group $(k>1)$. Then, to achieve approximately the same power and precision as in a trial with an equal number n in each group, n_1 should be chosen as:

$$n_1 = n(k+1)/(2k).$$

Examples are shown in Table 5.5 for various values of k. Notice that the number allocated to the smaller group can never be reduced below half the number required with equal groups. Little is gained by increasing k beyond 3 or 4, since, beyond this point, even a substantial increase in n_2 achieves only a small reduction in n_1.

5.2 Comparison of more than two groups

Field trials comparing two groups (for example, intervention and control, or treatment A and treatment B) are by far the commonest. However, in some trials, three or more groups may be compared. For example, in a trial of a new vaccine, there may be four trial groups receiving different doses of the vaccine. It is unusual for field trials to have more than four groups, because of logistical constraints or trial size limitations.

It is suggested that, in designing a trial with three or more groups, the investigator should decide which pair-wise comparisons between groups are of central interest. The methods of Sections 3 and 4 can then be used to decide on the trial size required in each group. Where there is one control group for comparison with several intervention groups, it is likely that the main pair-wise comparisons will be between each intervention group and the control group. Note, however, that direct comparisons between the