21
The size of investigations

Before embarking on an epidemiological study, it is important to ensure that the study will be large enough to answer the questions it addresses. Calculation of the required study size is often regarded as rather difficult, but in fact requires no new methods.

The problem is usually presented as if the scientist comes to the statistician with a clearly formulated hypothesis and the simple question 'How large should my study be?'. This is rarely the case. More usually the investigator has a very clear idea of the size of study proposed, this being determined by budgetary and logistic constraints, and requires an answer to the question 'Is my proposed study large enough?'. All too often calculations show the answer to be no! The investigator then needs to know how much larger the study needs to be.

This chapter will address the problem of study size from this standpoint. In addition to being more realistic, it follows more naturally from earlier chapters since the first stage of the calculation is to guess the results of the proposed study and analyse these. It will be convenient to develop the argument in the simplest case — the comparison of incidence in a cohort with that in a standard reference population. Generalization to other study designs is straightforward and will be discussed towards the end of the chapter.

21.1 The anticipated result

In order to answer the question 'Is my proposed study large enough?', we need to put ourselves in the position of having carried it out. To do this, it will be necessary to make some guesses about how things will turn out. A careful calculation of study size may involve a range of guesses. The most important thing to guess is the size of the effect of primary interest.

We shall take as an example a cohort study to investigate an occupational risk of lung cancer. In the proposed study, a cohort of industrial workers will be traced, and all deaths from lung cancer counted. This number will be compared with the expected number of deaths obtained by applying national age- and period-specific mortality rates to the table of person-time observation for the cohort. The first stage of the calculation will be to guess this person-time table, allowing for mortality in the cohort.

Let us assume that this has been done and that it leads to an expected number of lung cancer deaths of $E = 12.5$.

Exercise 21.1. What is the anticipated outcome of the study when θ, the rate ratio parameter for occupational exposure, is (a) 1.4, (b) 1.7, (c) 2.0, and (d) 5.0. In each case calculate the logarithm of θ and calculate the anticipated standard deviation for the log SMR (which estimates $\log(\theta)$). Is the study large enough to detect these rate ratios?

It is clear that the study would not be large enough to detect a rate ratio of 1.4, since the anticipated result would yield a 90% confidence interval which includes the null hypothesis $\theta = 1$ ($\log(\theta) = 0$). It should be equally clear that the study will almost certainly detect a rate ratio of 5, since in that case the size of effect is very large in comparison with its standard deviation. The two intermediate values for θ are more problematic and in such cases it is useful to further quantify the chances that the study will detect the effect.

21.2 Power

The *power* of a study is defined as the probability that it will yield a significant result when the true size of effect is as specified. The power is different for each size of effect considered, being greater for larger effects. Thus the power of a study is not a single number, but a whole range of values. The plot of power against size of effect is called a *power curve*. Two such curves for studies of different sizes are illustrated in Fig. 21.1. In practice it is rare for the entire power curve to be presented; more usually a few points in the range of effects are tabulated.

Exercise 21.2. Which curve corresponds to the larger study?

A significant result is defined as a result where the p-value for the null hypothesis is below a specified threshold (the *significance* level). Alternatively (and equivalently) it may be thought of as a result in which the null hypothesis falls outside a specified confidence interval. To calculate the power, it will be necessary to specify the significance (confidence) level to be used to categorize the result as significant. A study will have a higher power to detect a finding at the 5% level of statistical significance (95% confidence) than at the 1% level (99% confidence).

21.3 Calculating the power

It has already been stated that study size calculations require some guesswork. There is therefore little point in calculating power to a high order of accuracy. In this section we outline approximate power calculations which are accurate enough for all practical purposes.

Fig. 21.2 sets out our notation. The study aims to estimate an effect

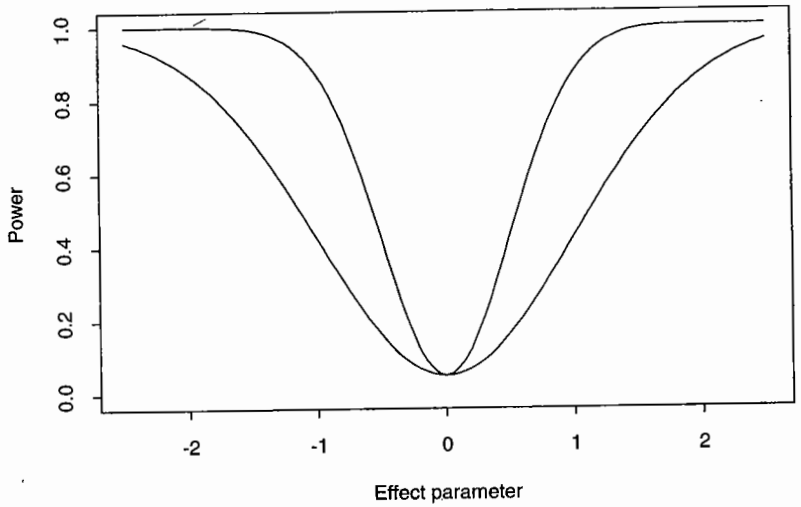

Fig. 21.1. Power curves for two studies.

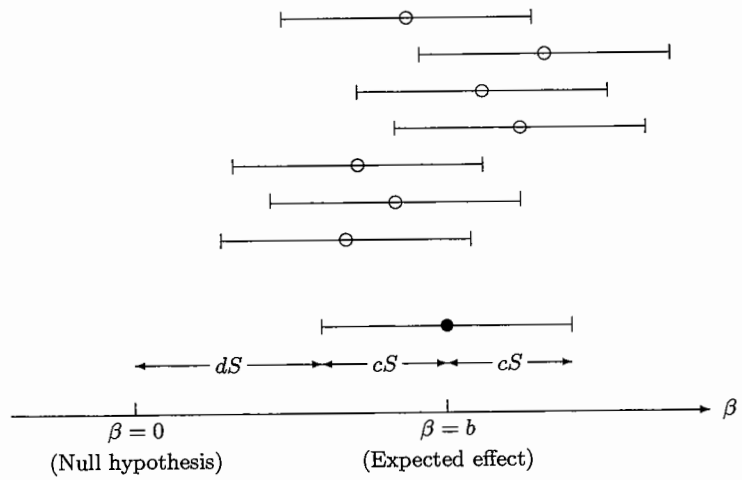

Fig. 21.2. Calculating the power of a study.

parameter, β,* and we assume that the log likelihood may be approximated by a Gaussian log likelihood with standard deviation S. To simplify notation, we also assume that the point $\beta = 0$ represents the null hypothesis (no effect). For example, β may be the log of a rate ratio or odds ratio. We wish to calculate the probability that the study will detect an effect of size $\beta = b$.

The lower part of the figure shows the anticipated result of the study. The black disc indicates the expected effect and the lines to either side indicate the expected confidence interval which would be calculated. The result will be taken as significant if the entire confidence interval lies to the right of the null hypothesis. The width of the interval depends upon the standard deviation S, and this in turn depends upon the size of the study. The interval also depends upon the significance or confidence level chosen. For example, for a 5% significance level we use the 95% confidence interval, which extends 1.96 standard deviations either side of the estimate, so $c = 1.96$.

If the expected value of the lower confidence limit lies above $\beta = 0$, the study would be expected to yield a positive result. However, it is not *guaranteed* to do so. If we imagine the study being repeated, the estimates obtained will vary from occasion to occasion. These estimates are indicated on the diagram by open circles.

The variation of estimates around the expected value is approximately Gaussian with standard deviation S. Ignoring the slight dependence of S upon the estimated value, the lower confidence limit will also vary around its expected value according to a Gaussian distribution with standard deviation S. The power of the study is the probability that this lower bound falls above zero. This depends upon the number of standard deviations between zero and the expected position of the lower bound. Referring to this number as d, the probability that the lower limit is above zero is then given by the probability that an observation in a standard Gaussian distribution exceeds $-d$. For example, if $d = 1.645$, the power is 0.95. When the expected location of the lower confidence limit is exactly at the null hypothesis, so that $d = 0$, the power is 0.50 and there is an even chance of obtaining a significant result. When the expected position is below zero $d < 0$, the power is less than 0.50. (Tables of the standard Gaussian distribution are widely available and are not included in this book.)

Exercise 21.3. For the study discussed in Exercise 21.1, calculate d for each value of the log rate ratio, assuming that a 5% significance level will be used (i.e. $c = 1.96$). Using tables of the Gaussian distribution, obtain the power in each case.

*We use this letter as it is the usual symbol for an effect parameter in regression models. It should not be confused with the 'type II error probability', for which it stands in some texts.

Table 21.1. Choice of c and d

Significance	c	Power	d
0.10	1.645	0.95	1.645
0.05	1.960	0.90	1.282
0.01	2.576	0.75	0.674

21.4 Increasing the power

If the results of the power calculations are disappointing, it will be necessary to increase the study size in some way. In this section we show how to determine by how much the study size must be increased to achieve the desired power.

Predetermining the significance level fixes the value of c. Similarly, predetermining the *power* fixes d. Since we require the distance $(c+d)S$ to equal the expected effect, b, we must choose the size of the study so that

$$S = \frac{b}{c+d}.$$

Table 21.1 lists some common requirements for significance and power. Note that, in each row of the Table, $(c+d)$ is between 3.2 and 3.3 so that these choices of significance and power suggest designing the study so that the expected effect, b, is just over 3 standard deviations.

Exercise 21.4. Calculate the value of the S which must be achieved if there were to be a power of 0.90 to detect a rate ratio $\theta = 1.7$ at the $p = 0.05$ significance level.

If the value of S required to achieve the desired power is smaller than that we expected to achieve with the study as originally proposed, then the study size must be increased. In general the factor by which the study size must be increased is

$$\left(\frac{\text{Current value of } S}{\text{Required value of } S}\right)^2$$

Exercise 21.5. Carrying on from the previous exercise, by what factor must the study be increased to achieve the required power? How could this be done in practice?

21.5 Application to other study designs

The extension of the above argument to different study designs introduces no serious new problems, although the first stage of the process — calculating the expected study result — may be more difficult.

COHORT STUDIES

When comparing exposed and unexposed groups in a cohort study, the standard deviation of the estimate of $\log\theta$ is

$$S = \sqrt{\frac{1}{D_0} + \frac{1}{D_1}}.$$

In order to predict the value of S, we need to be able to predict the values of D_0 and D_1. This can be done by using the total person-time of observation in the proposed cohort study, Y, and a guess for the disease rate in this population, λ. The total number of events we expect to observe is given by

$$D = \lambda Y.$$

If the proportion of the study cohort who will have been exposed is P, the person-time observed in the exposed and unexposed groups will be approximately PY and $(1-P)Y$ respectively. When the anticipated rate ratio is θ, the odds that a case was exposed will be

$$\theta\frac{PY}{(1-P)Y} = \theta\frac{P}{1-P},$$

and it follows that the D cases we anticipate are expected to split between exposed and unexposed as

$$D_1 = D\frac{\theta P}{1-P+\theta P}, \qquad D_0 = D\frac{1-P}{1-P+\theta P}.$$

The expected value of S for the estimated log rate ratio can then be calculated and the power calculated as before.

Exercise 21.6. You plan a cohort study of ischaemic heart disease in middle-aged men. The proposed size of the cohort is 10 000 men and a 5-year follow up period is envisaged. The estimated incidence rate in the study population is 10 per thousand person-years. What is the power of the study to detect a rate ratio of 1.5 for a risk factor to which 10% of the population is exposed?

CASE CONTROL STUDIES

Similar calculations are involved in the calculation of the power of a case control study. If it is planned to study D cases and H controls, and if the proportion of the population thought to be exposed to the factor of interest is P, we would expect the D cases to split between exposed and unexposed groups as above, and we expect the H controls to split as

$$H_1 = PH, \qquad H_0 = (1-P)H.$$

We are then in a position to calculate the expected standard deviation for the log odds ratio estimate, by the usual formula:

$$S = \sqrt{\frac{1}{D_0} + \frac{1}{D_1} + \frac{1}{H_0} + \frac{1}{H_1}}.$$

The calculation of the power follows as before.

Exercise 21.7. What is the power of a study of 100 cases and 200 controls to detect an odds ratio of 2.0 for an exposure present in 25% of the population?

STRATIFICATION AND MATCHING

Extension of these ideas to allow for stratification is straightforward in principle. In practice the difficulty is that the standard deviation of the effect of interest depends in a rather complicated way upon the strength of relationship between the exposure of interest and the stratifying variable(s). The same is true of matched case-control studies. It is particularly easy to see the difficulty in the case of the 1:1 design, since only case-control pairs which are discordant in exposure status contribute to the estimation of exposure effect. In such cases it will often be necessary to carry out a small pilot study, to provide estimates of the quantities necessary to calculate power.

DOSE-RESPONSE RELATIONSHIPS

If the level of exposure is graded, the log-linear model described in Chapter 20 allows an anticipated slope of a dose-response curve to be translated into a predicted increase in mean exposure of cases. If the standard deviation of the level of exposure in the study group is known, sample size calculations are then straightforward.

Solutions to the exercises

21.1 The anticipated number of deaths will be $D = \theta E$ and the corresponding standard deviation for the estimate of $\log\theta$ will be

$$\sqrt{\frac{1}{D}}.$$

For our four values of θ,

θ	1.4	1.7	2.0	5.0
D	17.5	21.25	25.0	62.5
$\log(\theta)$	0.336	0.531	0.693	1.609
S (estimated)	0.239	0.217	0.200	0.126

21.2 The larger study would correspond to the inner curve. For any size of effect, this curve predicts a higher probability of obtaining a significant result.

21.3 In each case, dS is obtained by subtracting $1.96S$ from the value of $\log(\theta)$. Thus, d is obtained by dividing this difference by S:

θ	d	Power
1.4	$(0.336 - 1.96 \times 0.239)/0.239 = -0.55$	0.29
1.7	$(0.531 - 1.96 \times 0.217)/0.217 = 0.49$	0.69
2.0	$(0.693 - 1.96 \times 0.200)/0.200 = 1.51$	0.93
5.0	$(1.609 - 1.96 \times 0.126)/0.126 = 10.81$	1.00

There is a slight chance of detecting a rate ratio of $\theta = 1.4$, quite a good chance for $\theta = 1.7$, a very good chance at $\theta = 2.0$ and the probability of failing to obtain a significant result at $\theta = 5.0$ is negligible.

21.4 The expected result at $\theta = 1.7$ is $b = 0.531$. By reference to Table 21.1 we see that $c = 1.960$ and $d = 1.282$ so that we need the standard deviation for the effect estimate to be:

$$S = \frac{0.531}{1.960 + 1.282} = 0.164.$$

21.5 The current standard deviation is 0.217 and it must be reduced to 0.164. The study must therefore be scaled up by a factor of

$$\left(\frac{0.217}{0.164}\right)^2 = 1.75.$$

The study must be increased so as to yield 75% more deaths. This can be achieved in practice either by increasing the size of the cohort or by extending the follow-up period.

21.6 The proposed study would accumulate $5 \times 10\,000 = 50\,000$ person-years of observations. At the anticipated incidence rate we would expect to observe $D = 10 \times 50 = 500$ disease events. If a proportion $P = 0.1$ of the total person-time is of exposed subjects and $(1 - P) = 0.9$ is of unexposed subjects, and if the rate ratio is $\theta = 1.5$, the expected number of exposed and unexposed cases is

$$\begin{aligned} D_1 &= 500 \times \frac{1.5 \times 0.1}{0.9 + 1.5 \times 0.1} \\ &= 71.4 \\ D_0 &= 500 \times \frac{0.9}{0.9 + 1.5 \times 0.1} \\ &= 428.6 \end{aligned}$$

The expected standard deviation for $\log(\theta)$ is

$$S = \sqrt{\frac{1}{71.4} + \frac{1}{428.6}} = 0.128$$

and $b = \log(1.5) = 0.405$. Thus, the number of standard deviations between expected result and null hypothesis, $(c + d)$, is $0.405/0.128 = 3.164$. For a 5% significance level, $c = 1.960$ so that $d = 3.164 - 1.960 = 1.204$. The power is the probability of exceeding -1.204 in a standard Gaussian distribution, given by tables as 0.885. The study has slightly less than 90% power to detect a rate ratio of 1.5.

21.7 Since the exposure is present in 25% of the population, we would expect the 200 controls to split as $H_1 = 50$ exposed, and $H_0 = 150$ unexposed. For $\theta = 2.0$, the expected split of the 100 cases is

$$\begin{aligned} D_1 &= 100 \times \frac{2.0 \times 0.25}{0.75 + 2.0 \times 0.25} \\ &= 40, \\ D_0 &= 60. \end{aligned}$$

The expected standard deviation of the estimate of $\log(\theta)$ is

$$S = \sqrt{\frac{1}{50} + \frac{1}{150} + \frac{1}{40} + \frac{1}{60}} = 0.261$$

and $b = \log(2.0) = 0.693$. The number of standard deviations between expected result and null hypothesis is 2.65. If a 5% significance level is to be used, $d = 2.65 - 1.96 = 0.69$. By referring -0.69 to the table of the standard Gaussian distribution, the power is 0.755 — just over 75%.