The Statistician (1992) 41, pp. 185–196

Formulae for sample size, power and minimum detectable relative risk in medical studies

MARK WOODWARD

Department of Applied Statistics, Reading University, Whiteknights, Reading RG6 2AN, UK

Abstract. The power calculation approach to the determination of sample size for comparing two independent means and two independent proportions in cross-sectional surveys, cohort studies, case-control studies, clinical trials and community trials is developed. These formulae are inverted to produce expressions for power given sample size and minimum detectable relative risk (or in the case of means, minimum detectable difference) given power and sample size. Although the basic methodology of the approach is well known it is shown that various approximations are necessary in the case of proportions, and the particular approximation used determines the formula obtained. Case-control studies require special formulae of their own. Allowance for unequal sample sizes in the groups to be compared is used throughout. A discussion of the procedures used by the statistical package INSTAT to calculate sample size is included.

1 Introduction

An earlier report in this journal (Woodward, 1989) described a set of commands for calculating sample size in medical studies which are encapsulated within a statistical software package INSTAT (Bailey, 1988). Correspondence received by the author concerning this report has shown that there is considerable confusion about the lack of agreement between published accounts of sample size requirements in the introductory statistics books typically used by medical researchers (see, for example, Pocock, 1983; Armitage & Berry, 1987; Woodward & Francis, 1988; Everitt, 1989; Kahn & Sempos, 1989; Campbell & Machin, 1990; Altman, 1991), despite each using the well-known power calculation approach. Confusion arises because of differences in assumptions, approximations and notation. Another source of confusion comes in the matching of study design to the appropriate sample size formula. Confusion arises here because some approaches are study-specific (e.g. only considering clinical trials) and some are not specific enough (e.g. failing to mention that a given formula is inappropriate for case-control studies). Furthermore, several introductory accounts fail to consider the very common situation in which two groups (e.g. cases and controls) of unequal size are to be compared.

The present article will address each of these issues by deriving sample size formulae for the comparison of two means and two proportions for a range of medical studies. The derivation will be from first principles, allowing different possible assumptions and approximations to be clear. The sample size formulae will be inverted to give expressions for power given sample size and minimum detectable difference or relative risk for a given power and sample size. The particular procedures used by the INSTAT commands will be indicated.

2 Comparison of two means

This section describes the sample size calculations where the problem is to compare the means of two independent samples, and would be suitable primarily for cross-sectional surveys, clinical trials and community trials. The formulae are also appropriate to cohort

studies, although in these the major end-point of interest is usually a proportion (e.g. a disease incidence) rather than a mean. In each case the units of study (usually people) are to be sampled at random with respect to the medical end-point of interest. This is not the situation in case-control studies, and hence these formulae are not appropriate there.

Sampling to obtain the data may progress in one of two ways. First, the sample may be a random selection from one single patient population, which is subsequently stratified into two groups according to the factor being studied. For instance in a clinical trial of a new treatment for hypertension (high blood pressure) patients for study should be drawn at random from the population of hypertensives who consent to be studied. These selected individuals are then randomly allocated to receive either the new treatment or a placebo (say). The factor of interest here is the treatment received and the end-point might be the diastolic blood pressure after one month of treatment. The second method of sampling is when two groups are sampled at random separately, that is stratified sampling is used with strata defined by the factor of interest. Thus in a cross-sectional study the strata might be distinguished by sex (the factor) where men and women are subsequently to be compared for mean body mass index (the end point).

Consider, first, a one-sided significance test for comparing the means μ_1 and μ_2 of X_1 and X_2, respectively, where X_1 and X_2 are random variables representing the end-points in the two groups. To test

$$H_0\colon \mu_1 = \mu_2$$

$$H_1\colon \mu_1 = \mu_2 + \theta \text{ (for } \theta \neq 0)$$

Let n_1 be the sample size in group 1, n_2 be the sample size in group 2, N be the total sample size ($N = n_1 + n_2$) and $r = n_1/n_2$ (i.e. ratio of sample sizes in first and second samples). Also, let σ be the standard deviation of X_1 and X_2 (assumed common), and z_γ be the value from tables of the cumulative normal distribution, i.e. $P(Z < z_\gamma) = 1 - \gamma$ for a standard normal, Z

$$\begin{aligned}
\alpha = P(\text{type I error}) &= P(\text{reject } H_0 | H_0 \text{ true}) \\
&= \text{size of test} \\
\beta = P(\text{type II error}) &= P(\text{fail to reject } H_0 | H_1 \text{ true}) \\
&= 1 - P(\text{reject } H_0 | H_1 \text{ true}) \\
&= 1 - (\text{power of test})
\end{aligned}$$

Assuming that X_1 and X_2 are normally distributed (or n_1 and n_2 are big enough for the Central Limit Theorem) the well-known test proceeds by rejecting H_0 whenever

$$T > z_\alpha$$

where T is the test statistic chosen so that

$$P(T > z_\alpha | H_0 \text{ true}) = \alpha$$

for a test of size α. Here it turns out that

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\Delta}$$

where

$$\Delta = \sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}$$

Notice that σ is assumed to be known, even though it may have to be inferred from previous studies, so that the normal distribution can be used. An alternative is to replace σ

by its usual sample-based estimate and adopt an iterative solution based on the t distribution (Snedecor & Cochran, 1980).

If the test is chosen to have power $1-\beta$ when the alternative hypothesis, H_1, is true then

$$P(T > z_\alpha | H_1 \text{ true}) = 1-\beta \tag{1}$$

Let

$$S = (\bar{X}_1 - \bar{X}_2 - \theta)/\Delta$$

Then

$$P(T > z_\alpha | H_1 \text{ true}) = P(S > z_\alpha - \theta/\Delta | H_1 \text{ true}) = 1-\beta$$

Now when H_1 is true $S \sim N(0,1)$ and so

$$z_\alpha - \theta/\Delta = z_{1-\beta} = -z_\beta$$

$$\Rightarrow \theta/\Delta = z_\alpha + z_\beta$$

$$\Rightarrow N = \frac{(r+1)^2(z_\alpha + z_\beta)^2\sigma^2}{r\theta^2} \tag{2}$$

A two-sided test would proceed the same way except that H_0 is rejected whenever

$$|T| > z_{\alpha/2}$$

Hence

$$P(T < -z_{\alpha/2} | H_1 \text{ true}) + P(T > z_{\alpha/2} | H_1 \text{ true}) = 1-\beta \tag{3}$$

Now

$$P(T < -z_{\alpha/2} | H_1 \text{ true}) = P(S < -z_{\alpha/2} - \theta/\Delta | H_1 \text{ true})$$

$$< P(Z < -z_{\alpha/2}) = \alpha/2$$

which is negligible for small α (e.g. for a 5% test, $\alpha/2 = 0{\cdot}025$).

Thus, approximately, (3) becomes

$$P(T > z_{\alpha/2} | H_1 \text{ true}) = 1-\beta$$

which is the same as equation (1) except that α is replaced by $\alpha/2$.

Making this same alteration to equation (2) gives

$$N = \frac{(r+1)^2(z_{\alpha/2} + z_\beta)^2\sigma^2}{r\theta^2} \tag{4}$$

This last case is the one used in many introductory texts. It rejects H_0 whenever the difference between means is large in *either* direction.

To find a sample size the investigator must specify the difference between means, θ, which he or she wishes to detect with a power of $1-\beta$ given that a significance test of size α is to be employed. He/she must also be able to give at least a reasonable estimate of the standard deviation, σ.

As an example of the use of the formulae consider a comparative study of two cities in which sample surveys, involving the measurement of several aspects of health, will be undertaken in each city. A real-life example is the comparison of cardiovascular risk factors in Edinburgh and North Glasgow reported by Smith *et al.* (1990). Suppose that the primary interest lies in comparing systolic blood pressure between the two cities and, hence, this comparison will determine the value of N (an alternative approach would be to calculate N for each variable in the survey and take the maximum of these as the sample size requirement). Assume that simple random sampling from among 40–44-year-old men

is to be used in each city with twice as many sampled from City 1 as from City 2 (for economic or other reasons), so that $r=2$. Systolic blood pressure is to be compared using a one-sided 5% significance test (i.e. $\alpha=0{\cdot}05$). The medical investigators wish to be 95% sure of detecting when the average blood pressure in City 1 exceeds that in City 2 by 3 mm Hg (i.e. $\beta=0{\cdot}05$ and $\theta=3{\cdot}0$). From published literature (Smith *et al.* 1989) the standard deviation of systolic blood pressure is likely to be 15·6 mm Hg (i.e. $\sigma=15{\cdot}6$). Using equation (2), the sample size required is

$$N=\frac{(2+1)^2(1{\cdot}64+1{\cdot}64)^2(15{\cdot}6)^2}{2(3{\cdot}0)^2}=1309{\cdot}1$$

The INSTAT commands described in Woodward (1989) give the answer 1316·8, which will be more accurate since more decimal places are used for the standard normal deviates. Notice that, although sample sizes are given here (and in other places in the article) to decimal point accuracy, sample size determination is not as accurate as this implies, since the various assumptions made (such as known variance) may be inaccurate and the specifications required (such as the required power) may be only vaguely known. In practice, 1316·8 needs to be rounded upwards to produce whole numbers, that is 878 should be sampled from City 1 and 439 from City 2.

The basic formulae (2) or (4) may be easily inverted to give an expression for θ, that is the difference between means which it is possible to detect with the specified power (and size) of test (or, more usefully, the smallest difference detectable with at least the given power). This gives

$$\theta=\frac{(r+1)(z_\alpha+z_\beta)\sigma}{(rN)^{1/2}}$$

in the case of a one-sided test. Similarly a value for the power, β, given N and θ comes from

$$z_\beta=\frac{\theta(rN)^{1/2}}{\sigma(r+1)}-z_\alpha$$

Although most introductory medical statistics books will not provide as much detail as given here, there is general agreement about the formulae for N, θ and z_β. Differences in results obtained are likely to be merely due to differences in the use of one or two-sided tests.

3 Comparison of two proportions

When two independent samples have been collected in the way described in the last section, it may be that, instead of comparing means, the object is to compare proportions. For instance a clinical trial of an anti-hypertensive drug might seek to compare the proportion of patients whose diastolic blood pressure has fallen more than 5 mm Hg after treatment with the drug against the same proportion amongst patients treated with a placebo. A comparison of proportions is usually the major aim of a cohort study, such as a study of smokers against non-smokers where the proportions with lung cancer some years after the start of the study are to be compared.

Let Π_1 and Π_2 be the proportions with the end-point of interest in the two populations. The null hypothesis to be tested is then

$$H_0\colon\ \Pi_1=\Pi_2$$

The alternative could be expressed as $\Pi_1=\Pi_2+\theta$, but it is more usual to measure differences between proportions on a multiplicative scale; in particular disease incidence is

usually compared through the *relative risk*, Π_1/Π_2. Hence the alternative hypothesis will be

$$H_1: \Pi_1/\Pi_2 = R \quad (\text{for } R \neq 1)$$

The well-known test statistic for this case (arising from a normal approximation to a binomial) is

$$T = \frac{P_1 - P_2}{\Delta}$$

where P_1 and P_2 are random variables representing the two proportions

$$\Delta = \left\{ p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{1/2}$$

and

$$p_c = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad (\text{the 'combined sample proportion'})$$

For a one-sided test of size α and power $1-\beta$ reject H_0 if

$$T > z_\alpha$$

such that

$$P(T > z_\alpha | H_1 \text{ true}) = 1-\beta$$

(notation introduced for means in the last section is re-used where appropriate). Now when H_1 is true

$$\frac{P_1 - P_2 - (R-1)\Pi}{\Delta_1} \sim N(0,1)$$

where $\Pi = \Pi_2 =$ true proportion in group 2 (assumed known) and

$$\Delta_1 = \left\{ \frac{R\Pi(1-R\Pi)}{n_1} + \frac{\Pi(1-\Pi)}{n_2}\right\}^{1/2}$$

$$= \left[\frac{r+1}{rN} \{ R\Pi(1-R\Pi) + r\Pi(1-\Pi)\} \right]^{1/2}$$

(using the fact that $\Pi_1 = R\Pi$ under H_1).

Now

$$P(T > z_\alpha | H_1 \text{ true}) = P\left\{ \frac{P_1 - P_2 - (R-1)\Pi}{\Delta_1} > z_\alpha \frac{\Delta}{\Delta_1} - \frac{(R-1)\Pi}{\Delta_1} | H_1 \text{ true} \right\}$$

$$= P\left\{ Z > \frac{z_\alpha \Delta - (R-1)\Pi}{\Delta_1} \right\} = 1-\beta$$

$$\Rightarrow z_\alpha \Delta - (R-1)\Pi = z_{1-\beta}\Delta_1 = -z_\beta \Delta_1$$

$$\Rightarrow z_\alpha \Delta + z_\beta \Delta_1 = (R-1)\Pi$$

Substituting for Δ and Δ_1 and rearranging gives

$$N = \frac{r+1}{r(R-1)^2\Pi^2} [z_\alpha \{(r+1)p_c(1-p_c)\}^{1/2} + z_\beta \{R\Pi(1-R\Pi) + r\Pi(1-\Pi)\}^{1/2}]^2 \qquad (5)$$

Now p_c is unknown in advance, but a reasonable approximation may be obtained by assuming that p_1 and p_2 take their population values under H_1, i.e. $R\Pi$ and Π, respectively. Thus

$$p_c \simeq \frac{n_1 R\Pi + n_2 \Pi}{n_1 + n_2} = \frac{\Pi(rR+1)}{r+1} \tag{6}$$

INSTAT uses this value for p_c and equation (5). The same formula is given by Armitage & Berry (1987) and Woodward & Francis (1988). Simpler formulae may be obtained if it is reasonable to assume that p_1 and p_2 are not very different from their weighted average, i.e. p_c as given by equation (6). With the new assumption

$$(r+1)p_c(1-p_c) \simeq R\Pi(1-R\Pi) + r\Pi(1-\Pi)$$

This can be used to simplify equation (5) in two ways, either to arrive at the formula

$$N \simeq \frac{r+1}{r(R-1)^2\Pi^2}\{R\Pi(1-R\Pi) + r\Pi(1-\Pi)\}(z_\alpha + z_\beta)^2 \tag{7}$$

or the formula

$$N \simeq \frac{(r+1)^2 p_c(1-p_c)(z_\alpha + z_\beta)^2}{r(R-1)^2\Pi^2} \tag{8}$$

where p_c is given by equation (6). In the case of equal sample sizes in the two groups, a formula equivalent to equation (7) is given by Pocock (1983), Everitt (1989) and Campbell & Machin (1990), while a formula equivalent to equation (8) is given by Altman (1991) and Kahn & Sempos (1989). The latter also give an alternative formula based on a slight variation of this general theme.

To be able to use equations (5), (7) or (8) the investigator must specify the relative risk R that is to be detected with a power of $1-\beta$ given a test of size α. A value for Π, the proportion with the end-point of interest in the reference group, must also be specified. As before, two-sided tests are dealt with by changing α to $\alpha/2$ in equation (5), (7) or (8).

Taking part of Example 2 in Woodward (1989), derived from Pocock (1983), suppose that mortality rates, amongst patients with myocardial infarction treated with either an active drug, anturan, or a placebo, are to be compared. The investigators wish to be 90% sure of detecting when anturan halves the mortality rate using a two-sided 5% significance test (i.e. $\beta = 0{\cdot}1$, $R = 0{\cdot}5$ and $\alpha/2 = 0{\cdot}025$). It is expected that 10% of patients will die after treatment with the placebo (i.e. $\Pi = 0{\cdot}1$). Suppose that twice as many patients are given anturan as are given placebo (i.e. $r = 2$).

From equation (6)

$$p_c = \frac{0{\cdot}1(2 \times 0{\cdot}5 + 1)}{2+1} = 0{\cdot}0667$$

Then, using equation (5), $N = 1275{\cdot}6$, as reported in Woodward (1989). Using the approximate formulae (7) and (8) gives answers of 1434·3 and 1176·8 respectively. Since none of these numbers for N is exactly divisible by 3, rounding upwards will be necessary to arrive at sample sizes for each treatment group.

In the example above, the results using the approximate formulae (7) and (8) are quite different from the result using equation (5). Table 1 shows the sample size requirements obtained using each of the three formulae for a range of values of Π, R and r such as are typically found in medical research. As would be expected (by any of the formulae), sample size increases as r moves away from unity, as R moves towards unity and as Π decreases. Approximate formulae (7) and (8) may under- or overestimate N depending upon the combinations of the other parameters; the 'true' result, from equation (5), is always between the other two. The approximations tend to worsen, in absolute terms, as sample

Table 1. Total sample size requirements when comparing two proportions for various values of Π (the proportion in the reference group), R (the relative risk) and r (the sample size ratio): requirements shown are calculated from formulae (5), (7) and (8) in the text for one-sided 5% tests with 90% power (formula (5) is the preferred formula)

	Π=0·2			Π=0·3			Π=0·4		
	(7)	(5)	(8)	(7)	(5)	(8)	(7)	(5)	(8)
$R=0{\cdot}5$									
$r=0{\cdot}25$	557	683	790	343	411	469	236	275	308
$r=0{\cdot}5$	437	491	535	265	296	321	180	199	214
$r=1$	428	433	437	257	262	265	171	176	180
$r=2$	527	480	445	313	291	274	206	196	188
$r=4$	781	656	565	460	397	351	300	268	244
$R=0{\cdot}9$									
$r=0{\cdot}25$	20082	20644	21086	11875	12138	12344	7772	7885	7973
$r=0{\cdot}5$	14618	14846	15025	8624	8732	8817	5626	5675	5712
$r=1$	13171	13176	13180	7747	7752	7756	5036	5040	5044
$r=2$	15017	14797	14627	8808	8709	8632	5704	5665	5635
$r=4$	21078	20520	20091	12336	12081	11884	7964	7861	7780
$R=1{\cdot}5$									
$r=0{\cdot}25$	1071	984	918	571	545	526	321	326	330
$r=0{\cdot}5$	745	714	689	403	395	390	231	236	240
$r=1$	634	639	642	348	353	357	206	210	214
$r=2$	681	721	754	381	398	411	231	236	240
$r=4$	910	1003	1079	517	552	579	321	326	330
$R=2$									
$r=0{\cdot}25$	300	268	244	139	138	137	58	72	83
$r=0{\cdot}5$	206	196	188	98	101	103	45	53	60
$r=1$	171	176	180	86	90	94	43	48	51
$r=2$	180	199	214	94	101	107	51	53	54
$r=4$	236	275	308	128	139	148	75	71	67

size increases, and can be considerably different to the result from equation (5) when R is close to unity. Notice that the approximations are very good when $r=1$.

Formula (5) is easily inverted to enable β to be calculated from N and R

$$z_\beta = \frac{\Pi(|R-1|)(Nr)^{1/2} - z_\alpha(r+1)\{p_c(1-p_c)\}^{1/2}}{[(r+1)\{R\Pi(1-R\Pi)+r\Pi(1-\Pi)\}]^{1/2}}$$

It is not, however, possible to solve equation (5) for R. Various iterative techniques may be applied to produce a value for R given N and β. A starting value for the iterations could be the appropriate solution for R from equation (7)

$$R \simeq \frac{1}{2a}[b \pm (b^2-4ac)^{1/2}]$$

where

$$a = rN\Pi^2 + \Pi^2(r+1)(z_\alpha+z_\beta)^2$$

$$b = 2rN\Pi^2 + \Pi(r+1)(z_\alpha+z_\beta)^2$$

$$c = rN\Pi^2 - \Pi(1-\Pi)r(r+1)(z_\alpha+z_\beta)^2$$

Alternatively the starting value could be the appropriate solution from equation (8)

$$R \simeq \frac{1}{2a}[b \pm (b^2-4ac)^{1/2}]$$

where

$$a = \Pi^2 N + r\Pi^2(z_\alpha + z_\beta)^2$$

$$b = 2\Pi^2 N + \Pi(r + 1 - 2\Pi)(z_\alpha + z_\beta)^2$$

$$c = \Pi^2 N - \frac{\Pi(r + 1 - \Pi)}{r}(z_\alpha + z_\beta)^2$$

Table 1 suggests that a better starting value would be the arithmetic mean of these two approximations. The original version of the INSTAT sample size commands, described in Woodward (1989), used the second approximate result. A revised version uses an iterative calculation.

As an illustration of the use of the two approximations, consider the anturan example used earlier. Taking $z_{\alpha/2} = 1{\cdot}96$, $z_\beta = 1{\cdot}28$, $\Pi = 0{\cdot}1$, $r = 2$ and the sample size (N) of 1275·6, the relative risk should be 0·5. The first approximation produces a result for R of 0·473 and the second approximation gives $R = 0{\cdot}516$. The average approximation, 0·495, is very accurate.

4 Case-control studies

The formulae derived in this section are appropriate only to case-control studies, i.e. medical studies wherein the cases, a sample of people with the end-point of interest (usually some specific disease) are compared to the controls, a sample without. Notice that sampling is now stratified according to the end-point, which is the essential difference from other studies, and makes special sample size formulae necessary.

As in cohort studies, the usual objective of a case-control study is to compare disease rates. Formula (5) and its associates cannot, however, be used directly since the data to be collected cannot test the hypothesis which generates equation (5). This is a consequence of the particular sampling method used, and leads to the calculation of an odds ratio, as an approximate relative risk, in case-control studies (see Woodward & Francis, 1988). Equation (5) can, however, be used to find N when the objective is to test

$$H_0\colon \Pi_1^* = \Pi_2^*$$

against

$$H_1\colon \Pi_1^* = R^*\Pi_2^* \quad \text{for } R^* \neq 1$$

where

$$\Pi_1^* = P(\mathrm{E}|\mathrm{D})$$

$$\Pi_2^* = P(\mathrm{E}|\bar{\mathrm{D}})$$

for the events E = 'person has been exposed to the factor' and D = 'person has the disease (i.e. the end-point)'. Then from equations (5) and (6)

$$N = \frac{r+1}{r(R^* - 1)^2\Pi^{*2}}[z_\alpha\{(r+1)p_c^*(1-p_c^*)\}^{1/2} + z_\beta\{R^*\Pi^*(1 - R^*\Pi^*) + r\Pi^*(1-\Pi^*)\}^{1/2}]^2 \tag{9}$$

where

$$\Pi^* = \Pi_2^*$$

and

$$p_c^* = \frac{\Pi^*(rR^*+1)}{r+1}$$

This is not useful as it stands since the investigator would wish to test hypotheses about $R = P(\mathrm{D}|\mathrm{E})/P(\mathrm{D}|\bar{\mathrm{E}})$ rather than R^*. However, using Bayes theorem

$$\Pi_1^* = P(\mathrm{E}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{E})P(\mathrm{E})}{P(\mathrm{D}|\mathrm{E})P(\mathrm{E}) + P(\mathrm{D}|\bar{\mathrm{E}})P(\bar{\mathrm{E}})} = \frac{RP(\mathrm{E})}{1+(R-1)P(\mathrm{E})}$$

$$\Pi_2^* = P(\mathrm{E}|\bar{\mathrm{D}}) = \frac{P(\bar{\mathrm{D}}|\mathrm{E})P(\mathrm{E})}{P(\bar{\mathrm{D}})} = P(\mathrm{E})\frac{(1-P(\mathrm{D}|\mathrm{E}))}{1-P(\mathrm{D})} \simeq P(\mathrm{E})$$

The above approximation is exact if $R=1$ and is very good if the disease is rare even among the 'factor' group. Hence $\Pi_2^* \simeq p$ and $R^* \simeq R/\{1+(R-1)p\}$ where $p = P(\mathrm{E})$. Substituting into equation (9) gives

$$N = \frac{(r+1)[1+(R-1)p]^2}{rp^2(p-1)^2(R-1)^2}\left[z_\alpha\{(r+1)p_c^*(1-p_c^*)\}^{1/2} + z_\beta\left\{\frac{Rp(1-p)}{[1+(R-1)p]^2} + rp(1-p)\right\}^{1/2}\right]^2 \tag{10}$$

where

$$p_c^* = \frac{p}{r+1}\left(\frac{rR}{1+(R-1)p}+1\right) \tag{11}$$

Formula (10) is also given, for the case of $r=1$, by Schlesselman (1974) and is used by INSTAT. An alternative derivation by Lemeshow *et al.* (1990) assumes that the null hypothesis will be specified for the odds ratio rather than the relative risk. In this case the odds ratio replaces R and Π_2^* replaces p in (10). In practice there is unlikely to be much difference between the two approaches (see also Schlesselman, 1982).

Approximate formulae are easily derived from equation (10) if it is assumed that

$$(r+1)p_c^*(1-p_c^*) \simeq R^*\Pi^*(1-R^*\Pi^*) + r\Pi^*(1-\Pi^*)$$

just as was done in the last section.

To be able to use equation (10), the investigator must specify p, the rate of exposure to the factor of interest which is anticipated among the entire population, as well as the relative risk, R, which it is required to detect (through the odds ratio) with power $1-\beta$ using a test of size α and a case : control ratio of $r:1$ subjects. Two-sided tests would be dealt with as before.

An example from Schlesselman (1982) will illustrate the calculations. This is a case-control study of the relationship between the use of oral contraceptives around the time of conception and congenital heart disease in the subsequent offspring. Using a two-sided 5% test it is required to detect a relative risk of 4 with 90% power. It is estimated that 30% of women of childbearing age are exposed to oral contraceptives within 3 months of conception. Twice as many controls as cases will be sampled. Here $z_{\alpha/2} = 1{\cdot}96$, $z_\beta = 1{\cdot}28$, $R=4$, $p=0{\cdot}3$ and $r=0{\cdot}5$.

Using equation (11)

$$p_c^* = \frac{0{\cdot}3}{1{\cdot}5}\left(\frac{2}{1{\cdot}9}+1\right) = 0{\cdot}4105$$

Substituting this into equation (10) gives $N = 101{\cdot}1$. Thus, with rounding, 34 cases and 68 controls are needed. INSTAT returns the same result.

Inverting equation (10) gives the result

$$z_\beta = \frac{\frac{p(Nr)^{1/2}}{(r+1)^{1/2}}\left|\frac{(R-1)(p-1)}{1+(R-1)p}\right| - z_\alpha\{(r+1)p_c^*(1-p_c^*)\}^{1/2}}{\left[\frac{Rp(1-p)}{\{1+(R-1)p\}^2}+rp(1-p)\right]^{1/2}}$$

As before, a simple expression for R cannot be derived from equation (10), although approximate R may be found by inverting approximate formulae for N (as shown by Walter, 1977). Using the approximate formula analogous to equation (7) gives

$$R \simeq 1 + \frac{-b \pm (b^2-4ac)^{1/2}}{2a}$$

where

$$a = rp^2 - \frac{Nrp(1-p)}{(z_\alpha+z_\beta)^2(r+1)}$$

$$b = 1+2rp$$

$$c = r+1$$

whereas using the approximate formula analogous to equation (8) gives

$$R \simeq 1 + \frac{-b \pm (b^2-4ac)^{1/2}}{2a}$$

where

$$a = p(r+p) - \frac{Nrp(1-p)}{(z_\alpha+z_\beta)^2}$$

$$b = (r+1)(r+2p)$$

$$c = (r+1)^2$$

Once again, a better approximation will be to take the arithmetic mean of these two and this can be used as the starting value for an iterative determination of R. The INSTAT procedure described in Woodward (1989) used a different approximation for R given in Schlesselman (1982). This has now been replaced by the more accurate iterative procedure.

When the two approximations are applied to the oral contraceptives example with $N = 101{\cdot}1$ the values of R returned are 3·891 and 4·085 respectively. The average approximation, 3·988, is very close to the true value of 4.

5 Conclusions

Although the power calculation approach to sample size determination is well understood and straightforward when comparing two independent means, approximations are necessary when two independent proportions are compared (using any type of study design). Much of the confusion over differences between different formulae arises because the nature of these approximations is not understood. Case-control studies require their own special sample size formula because of the sampling method used. Although several authors only consider equal sized groups, the formulae for comparing unequal sized groups are only slightly more complex.

There are, of course, many possible extensions and/or alternatives to the approach used in this article, and it is appropriate to mention a few in conclusion. The most basic

alternative is to derive N from considerations of estimation rather than hypothesis testing. In this situation N is chosen so as to determine (stochastically) the width of a confidence interval. Basic theory using normal distributions is developed by Lemeshow *et al.* (1990) and Woodward & Francis (1988). Both of these books also consider sample size calculation in one-sample situations, which would be appropriate to use when the data are paired rather than independent (as assumed in this article). Korn (1986) describes a method for calculating sample size so as to bound the upper limit of a confidence interval using exact binomial theory. A number of alternative methods for calculating N when comparing two independent proportions have been suggested, including use of the continuity correction (Fleiss, 1973), angular transformation (Cochran & Cox, 1957), Fisher's exact test (Casagrande *et al.*, 1978) an exact unconditional test (Suissa & Shuster, 1985) and an improved approximation to that given in equation (7) by Fleiss *et al.* (1980) with further comment in Campbell (1982). Extensions to the range of topics considered here include sample size for analysis of variance (Bratcher *et al.*, 1970) and matched case-control studies (Schlesselman, 1982). Machin & Campbell (1987) give tables of sample size for a range of situations in the context of clinical trials (although also useful elsewhere), including determining correlation coefficients and comparing survival curves. Lemeshow *et al.* (1990) provide an extensive set of other references. Finally, there is the approach of sequential analysis where N is not fixed in advance, but instead sampling ceases once the true hypothesis is apparent. The mean or median sample size is often considerably less than the fixed sample size alternative for the same power etc. Armitage (1975) and Whitehead (1992) show how sequential analysis may be useful in clinical trials where there are important ethical advantages.

References

ALTMAN, D. G. (1991) *Practical Statistics for Medical Research* (London, Chapman & Hall).

ARMITAGE, P. (1975) *Sequential Medical Trials* (Oxford, Blackwell).

ARMITAGE, P. & BERRY, G. (1987) *Statistical Methods in Medical Research*, 2nd edn (Oxford, Blackwell).

BAILEY, T. (1988) Statistical software review, *Applied Statistics*, 37, pp. 273–277.

BRATCHER, T. L., MORAN, M. A. & ZIMMER, W. J. (1970) Tables of sample size in the analysis of variance, *Journal of Quality Technology*, 2, pp. 156–164.

CAMPBELL, M. J. (1982) The choice of relative group sizes for the comparison of independent proportions (correspondence with reply), *Biometrics*, 38, pp. 1093–1094.

CAMPBELL, M. J. & MACHIN, D. (1990) *Medical Statistics. A Commonsense Approach* (Chichester, Wiley).

CASAGRANDE, J. T., PIKE, M. C. & SMITH, P. G. (1978) The power function of the exact test for comparing two binomial distributions, *Applied Statistics*, 27, pp. 176–180.

COCHRAN, W. G. & COX, G. M. (1957) *Experimental Designs*, 2nd edn (London, Wiley).

EVERITT, B. S. (1989) *Statistical Methods for Medical Investigations* (London, Edward Arnold).

FLEISS, J. L. (1973) *Statistical Methods for Rates and Proportions* (New York, Wiley).

FLEISS, J. L., TYTUN, A. & URY, H. K. (1980) A simple approximation for calculating sample sizes for comparing independent proportions, *Biometrics*, 36, pp. 343–346.

KAHN, H. A. & SEMPOS, C. T. (1989) *Statistical Methods in Epidemiology* (New York, Oxford University Press).

KORN, E. L. (1986) Sample size tables for bounding small proportions, *Biometrics*, 42, pp. 213–216.

LEMESHOW, S., HOSMER, D. W., KLAR, J. & LWANGA, S. K. (1990) *Adequacy of Sample Size in Health Studies* (Chichester, Wiley).

MACHIN, D. & CAMPBELL, M. J. (1987) *Statistical Tables for the Design of Clinical Trials* (Oxford, Blackwell).

POCOCK, S. J. (1983) *Clinical Trials. A Practical Approach* (Chichester, Wiley).

SCHLESSELMAN, J. J. (1974) Sample size requirements in cohort and case-control studies of disease, *American Journal of Epidemiology*, 99, pp. 381–384.

SCHLESSELMAN, J. J. (1982) *Case-Control Studies: Design, Conduct and Analysis* (New York, Oxford University Press).

SMITH, W. C. S., TUNSTALL-PEDOE, H., CROMBIE, I. K. & TAVENDALE, R. (1989) Concomitants of excess coronary deaths: Major risk findings from 10,359 men and women in the Scottish Heart Health Study, *Scottish Medical Journal*, 34, pp. 550–555.

SMITH, W. C. S., SHEWRY, M. C., TUNSTALL-PEDOE, H., CROMBIE, I. K. & TAVENDALE, R. (1990) Cardiovascular disease in Edinburgh and North Glasgow—A tale of two cities, *Journal of Clinical Epidemiology*, 43, pp. 637–643.

SNEDECOR, G. W. & COCHRAN, W. G. (1980) *Statistical Methods*, 7th edn (Ames, IA, Iowa State University Press).

SUISSA, S. & SHUSTER, J. J. (1985) Exact unconditional sample sizes for the 2×2 binomial trial, *Journal of the Royal Statistical Society, Series A*, 148, pp. 317–327.

WALTER, S. D. (1977) Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes, *American Journal of Epidemiology*, 105, pp. 387–397.

WHITEHEAD, J. (1992) *The Design and Analysis of Sequential Clinical Trials*, 2nd edn (Chichester, Ellis Horwood).

WOODWARD, M. (1989) A computer-based method for determining optimal sample size with medical applications, *The Statistician*, 38, pp. 221–226.

WOODWARD, M. & FRANCIS, L. M. A. (1988) *Statistics for Health Management and Research* (London, Edward Arnold).