Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons

M J Campbell, S A Julious, D G Altman

Sample size calculations are now mandatory for many research protocols, but the ones useful in common situations are not all easily accessible. This paper outlines the ways of calculating sample sizes in two group studies for binary, ordered categorical, and continuous outcomes. Formulas and worked examples are given. Maximum power is usually achieved by having equal numbers in the two groups. However, this is not always possible and calculations for unequal group sizes are given.

A sample size calculation is now almost mandatory in research protocols and to justify the size of clinical trials in papers.[1] Nevertheless, one of the most common faults in papers reporting clinical trials is in fact a lack of justification of the sample size, and it is a major concern that important therapeutic effects are being missed because of inadequately sized studies.[2] A recent paper has concluded "the reporting of statistical power and sample size needs to be improved."[3] Recent articles in the *BMJ* have described the basis of sample size calculations,[4][5] and explained the fundamental concepts of statistical significance (α), effect size (δ), and power ($1-\beta$). A nomogram for sample size calculations for continuous data is also available.[6] However, there have been some recent developments in the theory of sample size calculations, which are likely to prove useful, and the purpose of this paper is to make available a collection of formulas and examples for a variety of situations likely to be encountered in practice. In particular, situations not dealt with in previous articles are two group comparisons with unequal sample sizes, and sample sizes for ordered categorical outcomes (for example categories better, same, or worse). The paper describes sample size calculations, and provides tables, for studies comparing two groups of individuals that have outcome variables that are binary (yes/no), ordered categorical, or continuous. A further paper will consider studies when the data are paired. Further examples are given by Machin and Campbell.[7]

Parameter definition

Of all the parameters that have to be specified before the sample size can be determined the most critical is the effect size. Reducing the effect size by half will quadruple the required sample size. The effect size can be interpreted as a "clinically important difference," but this is often difficult to quantify. A valuable attempt at classification was made by Burnand et al, who reviewed three major medical journals and looked for words such as "impressive difference," "important difference," "dramatic increase" and then calculated a standardised effect size.[8] This provided a guide to the size of effect regarded as important by other authors. There are several ways of eliciting useful sample sizes: a Bayesian perspective has been given recently,[9] along with an economic approach,[10] and one based on patients' rather than clinicians' perceptions of benefit.[11]

In statistical significance tests one sets up a null hypothesis and, given the observed difference of interest, calculates the probability of observing the difference (or a more extreme one) under the null hypothesis. This yields the P value. If the P value is less than some prespecified level then we reject the null hypothesis. This level is known as the significance level α. If we reject the null hypothesis when it is true we make a type I error, and we set α, the significance level, to control the probability of doing this. If the null hypothesis is in fact false but we fail to reject it, we make a type II error, and the probability of a type II error is denoted as β. The probability of rejecting the null hypothesis when it is false is termed the power and is defined as $1-\beta$.

Unequal numbers in each group

For a given total sample size the maximum power is achieved by having equal numbers of subjects in the two groups. Often, however, in observational studies an equal number is not expected in each group since the incidence of a particular factor may be higher in one group than in another. In clinical trials, also, the numbers of subjects taking one treatment may have to be limited, so to achieve the necessary power one has to allocate more patients to the other treatment. In this case the sample sizes should be adjusted by a factor dependent on the allocation ratio,[12] given as equation 1 in the Appendix.

If one were to maintain the same sample size as calculated for a 1:1 ratio but then allocated in the ratio 2:1 the loss in power would be quite small (around 5%). However, if the allocation ratio is allowed to exceed 2:1 with the same total sample size the power falls very quickly (a loss of around 25% in power for a ratio of 5:1) and consequently a considerably larger total sample size is required to maintain a fixed power with an imbalanced study than with a balanced one.

Continuous data

In a two group comparative study where the outcome measure is a continuous variable which is plausibly normally distributed, such as blood pressure, a two sample t test would be the statistical test used in the final analysis.

To calculate a sample size, in addition to the parameters discussed above, an estimate of the population standard deviation σ must be given. The sample size formula[7] is given as equation 2 in the Appendix, and table I gives the sample size required for different values of the standardised difference d, defined as $d=\delta/\sigma$, at various levels of power at the two sided 5% significance level.

Alternatively, Lehr gives a quick formula for calculating these sample sizes.[13] For a two sided significance level of 5% and power of 80%, the number required in each group is given approximately by $m=16/d^2$. This formula overestimates the sample sizes a little for small values of d; otherwise it gives close approximations to the sample size.

WORKED EXAMPLE

In a recent paper, Godfrey et al[14] found that 46 people who had no whorls on their fingers had a mean systolic blood pressure of 136 mm Hg compared with

Medical Statistics and Computing, University of Southampton, Southampton General Hospital, Southampton SO16 6YD
M J Campbell, *reader in medical statistics*
S A Julious, *statistician programmer*

Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX
D G Altman, *head*

Correspondence to:
Dr Campbell.

93 patients with at least one whorl for whom the mean blood pressure was 144 mm Hg.

Suppose an experimenter wished to confirm these findings but suspected that the mean difference would be less than that observed, with 5 mm Hg being the clinically minimum difference accepted. The overall standard deviation of blood pressure in each group is assumed to be 17 mm Hg, the same as that published. We find $d=5/17=0.294$, which is about 0·3, and so from table I the sample size required to detect this difference with a two sided significance level of 5% and with 80% power would be 176 subjects in each group and so 352 subjects in total. Alternatively, from Lehr's

TABLE I—*Sample sizes required per group at the two sided 5% significance level for different values of d and power (d=expected mean difference/standard deviation)*

d	Power (1-β)				
	99	95	90	80	50
0·10	3676	2600	2103	1571	770
0·20	920	651	527	394	194
0·30	410	290	235	176	87
0·40	231	164	133	100	49
0·50	148	105	86	64	32
0·60	104	74	60	45	23
0·70	76	54	44	33	17
0·80	59	42	34	26	13
0·90	47	34	27	21	11
1·00	38	27	22	17	9
1·10	32	23	19	14	8
1·20	27	20	16	12	7
1·30	23	17	14	11	6
1·40	20	15	12	9	5
1·50	18	13	11	8	5

quick formula we get $m=16/0.294^2=185$ patients per group. Suppose, like Godfrey et al, we would expect to recruit two people with whorls for every one person with no whorls. With $r=2$ from equation 1 we find that $m'=3\times176/4=132$ and so $rm'=264$, giving a modified total sample size of 396. The overall sample size is larger if the groups were unequal because the design has less power than a design of the same size with equal numbers in the two groups.

Binary data

A binary outcome is a response which has just two categories. These categories may be of the form yes/no or presence/absence in relation to a given factor, for example alive/dead. Often an experimenter may wish to compare treatments by testing whether the difference in proportions responding on each treatment could be due to chance. In this case the effect size can be formulated as $\delta=p_A-p_B$, where p_A and p_B are the proportions expected in the two treatment groups. The statistical test used to test for the association between two binary variables is the Pearson χ^2 test.

To calculate the number of patients required in each arm of a binary trial use equation 3 in the Appendix. For proportions greater than about 0·1 this simplifies to equation 4. Table II gives the sample sizes required for various values of p_A and p_B for two sided signifi-

cance level α and power 1-β. Note, however, that for p_A in the table only values up to 0·5 are given. This is because having a success rate of 65%, say, is identical to a failure rate of 35% and so the sample sizes for comparing p_A to p_B are the same as those for comparing $1-p_A$ and $1-p_B$.

An approximate result similar to Lehr's formula[13] for 80% power and two sided 5% significance level is that $m=16\bar{p}(1-\bar{p})/(p_A-p_B)^2$, where $\bar{p}=(p_A+p_B)/2$. Like Lehr's equation given earlier, this overestimates the sample size a little.

Observational surveys such as case control studies are often summarised by an odds ratio or relative risk, rather than a difference in proportions. If p_A is the proportion of cases exposed to a risk factor and p_B is the proportion of controls exposed to the same risk factor, then the odds ratio of being a case given the risk factor is odds ratio=$p_A(1-p_B)/\{p_B(1-p_A)\}$. An approximate sample size formula using the odds ratio (OR) is given by equation 5 in the Appendix.

WORKED EXAMPLE

Tovey and Bonell stated that 52 (19%) out of 281 men found condoms too tight.[15] Of these 68% had experienced their condom splitting compared with only 26% of men whose condoms were not tight. Suppose from anecdotal evidence a researcher suspected that the prevalence of reported splitting was nearer 50% in the group finding condoms too tight and wished to conduct a study to show this prevalence still to be significantly higher than in the other group.

The expectation is that the observed ratio of the frequencies of "not tight" (A) to "tight" (B) would be 4:1. Here $p_A=0.5$, $p_B=0.25$ and $r=4$. From table II the sample size required with equal allocation in each group would be 58, and using equation 2 one derives a modified sample size of just 37 subjects in the group who found condoms too tight and 148 in the other group, giving a total of 185. In the unlikely event of equal group sizes a total of 116 subjects would be required, yielding a saving of 69 subjects. Again, this arises because the equal groups case is more efficient. Note that Lehr's formula for equal sized groups gives approximately 60 per group or a total of 120 subjects required. If we specified the effect size as an odds ratio, then the postulated odds of splitting when the condom is too tight are three times that when it is not. From equation 5, we find in this case that for equal allocation we require 55 subjects per group.

Ordered categorical data

A study may be undertaken where the outcome measure of interest is an ordered scale, such as a Likert scale (strongly disagree, disagree, agree and strongly agree) or a rating scale (better, same, worse). The statistical test used in this instance is the Mann-Whitney U test, with allowance for ties.[16] The calculation of sample sizes when the data are ordered is not immediately straightforward. The problem becomes considerably easier, however, if one considers

TABLE II—*Sample sizes to detect a difference in two proportions, p_A and p_B, at a 5% significance level with 80% power*

p_A	p_B																			
	0·05	0·10	0·15	0·20	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70	0·75	0·80	0·85	0·90	0·95	1·00
0·00	152	74	48	35	27	22	18	15	13	11	10	8	7	6	6	5	4	4	3	2
0·05		435	141	76	49	36	27	22	18	15	12	11	9	8	7	6	5	4	4	3
0·10			686	199	100	62	43	32	25	20	16	14	11	10	8	7	6	5	4	4
0·15				906	250	121	73	49	36	27	22	17	14	12	10	8	7	6	5	4
0·20					1094	294	138	82	54	39	29	23	18	15	12	10	8	7	6	5
0·25						1251	329	152	89	58	41	31	24	19	15	12	10	8	7	6
0·30							1377	356	163	93	61	42	31	24	19	15	12	10	8	6
0·35								1471	376	170	96	62	43	31	24	18	14	11	9	7
0·40									1534	388	173	97	62	42	31	23	17	14	11	8
0·45										1565	392	173	96	61	41	29	22	16	12	10

TABLE III—*For ordered categorical data, values for 6 $(z_{1-\alpha/2}+z_{1-\beta})^2/$ log OR)^2 for various values of the odds ratio (OR) and power (1-β) at two sided 5% significance*

Odds ratio	Power (1-β)				
	99	95	90	80	50
0·75	1331·97	942·09	761·77	569·03	278·50
1·25	2213·86	1565·85	1266·13	945·78	562·89
1·50	670·52	474·25	383·48	286·45	140·20
1·75	352·00	248·96	201·31	150·38	73·60
2	229·44	162·60	131·22	98·02	47·97
3	91·33	64·60	52·23	39·02	19·10
4	57·36	40·97	32·80	24·50	11·99
5	42·56	30·10	24·34	18·18	8·90
10	20·79	14·71	11·89	8·88	4·35

a number of pragmatic steps which will be described later in this section.

As before, we need to specify an effect size, and here it turns out to be easier to use the odds ratio. We must also specify the proportion of subjects expected in each category of the scale for one of the groups. Suppose we have t categories, with the higher ordered categories indicating worse prognosis, and the proportions expected in group A are $p_{A1}, p_{A2}, ... p_{At}$ (where $p_{A1}+p_{A2}+... +p_{At}=1$) with similar notation for group B. Let $c_{A1}, c_{A2}, ... c_{At}$, be the cumulative probabilities, so $c_{A1}=p_{A1}$, $c_{A2}=p_{A1}+p_{A2}$, etc. The odds ratio is the chance of a subject being in a given category or lower in one group compared with the other. For category 1 it is given by $OR_1=\{c_{A1}/(1-c_{A1})\}/\{c_{B1}/(1-c_{B1})\}$ and similarly OR_2 for category 2, up to category t-1. As will be shown later, the odds ratio may not necessarily be too difficult to estimate, as the proportions expected for one group may already be known through a pilot study or from previous research. The experimenter may postulate that on the new treatment a patient is only half as likely to have a score above a given level than on the old treatment and so the odds ratio would be estimated as 0·5. Alternatively, an experimenter may know the expected proportions in each category for one group and speculate that, if a proportion, p, were in a particular category or better, then a clinically significant difference would be for the corresponding proportion to be about 20% higher in the other group. From this information an odds ratio can be calculated and hence the other expected proportions and the sample size.

Equation 6 in the Appendix gives the formula for sample size calculations for ordered categorical data. It assumes that the odds ratio is constant for each pair of adjacent categories, that is $OR_1=OR_2=... OR_{t-1}$, and this assumption means that the Mann-Whitney U test is the best test to use. It also means that one can estimate the odds ratio from any cumulative proportion from each group. To aid the calculations table III gives values of the numerator from equation 6 for different values of odds ratio and power.

If the number of categories is large it is difficult to postulate the proportion of people who would fall in a given category. However, Whitehead has shown that there is little increase in power (and hence saving in

TABLE IV—*Correction factor to be used with table III when the number of categories is ≤5*

No of categories	Correction factor
2	1·333
3	1·125
4	1·067
5	1·042

TABLE V—*Playfulness in children*

Category	Numbers*		Proportions		Cumulative proportions		
	A	B	A p_{Ai}	B p_{Bi}	A c_{Ai}	B c_{Bi}	Odds ratios
Normal	3	6	0·14	0·27	0·14	0·27	0·440
Slightly listless	5	9	0·24	0·41	0·38	0·68	0·287
Moderately listless	5	5	0·24	0·23	0·62	0·91	0·1625
Very listless	8	2	0·38	0·09	1·00	1·00	—
Total	21	22	1·00	1·00			

*A=control, B=paracetamol.

number of subjects recruited) to be gained by increasing the number of groups beyond five.[17]

WORKED EXAMPLE

In a randomised controlled trial of paracetamol for the treatment of feverish children, Kinmonth *et al* categorised playfulness as normal or slightly, moderately, or very listless.[18] The results for the 43 replies are given in table V, together with the proportions and the cumulative proportions. The first odds ratio in the table is calculated from $\{0·14/(1-0·14)/(0·27/(1-0·27))\}$ $=0·44$, and in a similar way we get 0·287 and 0·1625 for the other two pairs. The average is about 0·3.

Suppose a new study was planned in which we wished to replicate these results. The distribution of children in the control group (group A) was expected to be about the same as was found previously and shall be used in the calculation of the sample sizes. If an odds ratio of 0·33 in favour of paracetamol (or equivalently an odds ratio of about 3 against the control) was expected, then from the definition of the odds ratio we can calculate the expected cumulative proportions in the treatment group (group B) from the formula $c_{Bi}=c_{Ai}/(c_{ai}+OR (1-C_{ai}))$. Thus the proportion expected in the first category of group B is 0·14/(0·14+ 0·33 (1-0·14))=0·33 and so on. The cumulative proportions expected in group B are 0·33, 0·65, 0·83, and 1·00, and so the actual proportions expected are 0·33, 0·32=(0·65-0·33), 0·18=(0·83-0·65), and 0·17=(1·00-0·83). The average proportions $\bar{p}$ are given by 0·235, 0·280, 0·210, and 0·275. Thus $(1-\Sigma\bar{p}^3)=$ 0·935. For 80% power and 5% significance level, from table III, the numerator is 39·02, and so the sample size is 39·02/0·935=41·7, or about 42 patients per group.

The formula is quite complicated and we have a number of suggestions to simplify matters. If the mean proportions ($\bar{p}_i$'s) in each category are roughly equal then the denominator in equation 6 is constant for a given number of categories, and if the number of categories exceeds five it is approximately unity. Thus for 80% power and a two sided significance of 5%, an estimate of the sample size can be obtained from m= 47/(log OR).[2] If the number of categories is less than or equal to five then multiply this sample size estimate by a correction factor given in table IV. From this table, in the situation of approximately equal proportions, it is evident that having only two categories in your data for analysis may require you to recruit a third more patients than if the data were kept continuous. For our example, the correction factor from table IV is 1·067 and so n=1·067×47/(log 0·33)^2=40·8, or 41 patients.

Another simplification occurs if the proportion of subjects in one category for both groups is expected to be large. We can combine categories until there are only two left and use the formula and table given previously for binary data. Combining categories reduces the amount of information available, so one would expect the required sample size to increase.

In the worked example if we had pooled those scoring 1-2 and those scoring 3-4, we would compare proportion $p_A=0·38$ to $p_B=0·65$. Formula 4 shows that this study would require 49·9, or about 50 patients per group. Thus, use of all four categories, rather than simply two, yields a reduction of 16% in the study size, and this might outweigh the benefit of an easier sample size calculation.

Comment

From the equations in the Appendix it is clear that the sample size, significance level, power, and effect size are all interlinked. Given any three parameters, in principle the equations can be solved for the fourth. Thus, if the sample size were limited by resources, and the significance level fixed in advance, one could

arbitrarily increase the power of the study by postulating larger effect sizes. In practice, however, the estimate of the effect of an intervention often proves too optimistic, resulting in many trials which are too small. The need for sample size calculations provides an excellent opportunity to involve a statistician early in the planning of a study and not just when the analysis is required. This paper has covered only a limited range of designs, and a statistician could advise on other designs. These include comparison of more than two groups,[19] comparison of survival curves,[7 20 21] and studies to demonstrate bioequivalence.[22] Computer software is available for some of the sample size calculations discussed here,[23-26] and other reviews have been given.[27 28]

Appendix

In each of the following m is the number of subjects required in each group for a two sided significance α and power $1-\beta$, and $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard Normal distribution for the $100(1-\alpha/2)$ and $100(1-\beta)$ percentiles respectively. Some useful values are the following: for two sided $\alpha=0.05$, $z_{1-\alpha/2}=1.96$; for two sided $\alpha=0.01$, $z_{1-\alpha/2}=2.58$; for $\beta=0.2$, $z_{1-\beta}=0.84$; and for $\beta=0.1$, $z_{1-\beta}=1.28$.

UNEQUAL ALLOCATION

Given m, calculated assuming equal sized groups, let m' be the sample size in the first group and rm' the sample size in the second group. Then m' is given by

$$m'=\frac{r+1}{2r}\times m, \quad (1)$$

where r is the allocation ratio.

CONTINUOUS DATA

To detect a difference δ we require[7]:

$$m=\frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{d^2}+\frac{z^2_{1-\alpha/2}}{4} \quad (2)$$

where $d=\delta/\sigma$ and σ is the standard deviation of the measurements. The last term in the equation is a correction factor to enable Normal tables rather than t tables to be used and can be ignored except for very small sample sizes. For a 5% two sided significance level it increases the sample size by 1. Table I gives the sample size required for different values of d and power from 50% to 99%.

BINARY OUTCOME

Suppose the expected proportions in groups A and B were p_A and p_B.

$$m=\frac{[z_{1-\alpha/2}\sqrt{\{2\bar{p}(1-\bar{p})\}}+z_{1-\beta}\sqrt{\{p_A(1-p_A)+p_B(1-p_B)\}}]^2}{\delta^2} \quad (3)$$

where $\delta=p_A-p_B$, and $\bar{p}=(p_A+p_B)/2$.

An approximate, simpler formula, is:

$$m=\frac{(z_{1-\alpha/2}+z_{1-\beta})^2[p_A(1-p_A)+p_B(1-p_B)]}{\delta^2} \quad (4)$$

which is sufficiently accurate except when p_A, p_B are small (say <0.05). Table II gives the sample size required per group at 5% significance level and 80% power for values of p_A between 0 and 0.45 and p_B between 0.05 and 1.00.

If the effect size is specified as an odds ratio

$$OR=\frac{p_A(1-p_B)}{p_B(1-p_A)},$$

then an approximate formula is given by

$$m=\frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{\log(OR)^2\,\bar{p}(1-\bar{p})} \quad (5)$$

Ordered categorical data

$$m=\frac{6(z_{1-\alpha/2}+z_{1-\beta})^2/(\log OR)^2}{[1-\sum_{i=1}^{k}\bar{p}_i^{-3}]}, \quad (6)$$

where OR is the odds ratio of a patient being in category i or less for one treatment compared to the other, k is the number of categories and $\bar{p}_i$ is the mean proportion expected in category i—that is, $\bar{p}_i=(p_{Ai}+p_{Bi})/2$ where p_{Ai} and p_{Bi} are the proportions expected in category i for the two groups A and B respectively.

1 Gardner MJ, Machin D and Campbell MJ. Use of checklists in assessing the statistical content of medical studies. BMJ 1986;292:810-12.
2 Freiman JA, Thomas AB, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978;299: 690-4.
3 Moher D, Dulberg CS, Wells GA. Statistical power, sample size and their reporting in randomised controlled trials. JAMA 1994;272:122-4.
4 Florey C du V. Sample sizes for beginners. BMJ 1993;306:1181-4.
5 Daly L. Confidence intervals and sample sizes: don't throw out all your old sample size tables. BMJ 1991;302:333-6.
6 Gore SM, Altman DG. Statistics in practice. BMA: London, 1982.
7 Machin D, Campbell MJ. Statistical tables for the design of clinical trials. Oxford: Blackwell Scientific, 1987.
8 Burnand B, Kernan WN, Feinstein AR. Indexes and boundaries for "quantitative significance" in statistical decisions. J Clin Epidemiol 1990;43: 1273-84.
9 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. J R Statistic Soc A 1994;157:357-416.
10 Drummond M, O'Brian B. Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes. Health Economics 1993;2:205-12.
11 Naylor CD, Llewellyn-Thomas HA. Can there be a more patients-centred approach to determining clinically important effect size for randomised treatments? J Clin Epidemiol 1994;47:787-95.
12 Woodward M. Formulas for sample-size, power and minimum detectable relative risk in medical studies. Statistician 1992;41:185-96.
13 Lehr R. Sixteen s squared over d squared: a relation for crude sample size estimates. Statistics in Medicine 1992;11:1099-1102.
14 Godfrey KM, Barker DJP, Peace J, Cloke J, Osmond C. Relationship of fingerprints and shape of the palm to fetal growth and adult blood pressure. BMJ 1993;307:405-9.
15 Tovey SJ, Bonell CP. Condoms: a wider range needed. BMJ 1993;307:987.
16 Conover WJ. Practical nonparametric statistics. 2nd ed. New York: John Wiley, 1980.
17 Whitehead J. Sample size calculations for ordered categorical data. Statistics in Medicine 1993;12:2257-72.
18 Kinmonth A-L, Fulton Y, Campbell MJ. Management of feverish children at home. BMJ 1992;305:1134-6.
19 Day SJ, Graham DF. Sample size estimation for comparing two or more groups. Statistics in Medicine 1991;10:33-43.
20 Borenstein M. Planning for precision in survival studies. J Clin Epidemiol 1994;47:1277-85.
21 Fayers PM, Machin D. Sample size: How many patients are necessary? Br J Cancer 1995;72:1-9.
22 Diletti E, Hauschke D, Steinijans VW. Sample size determination for bioequivalence assessment by means of confidence intervals. Int J Clin Pharm Therapy Toxicol 1991;29:1-8.
23 Goldstein R. Power and sample size via MS/PC-DOS computers. American Statistician 1989;43:253-60.
24 Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. Controlled Clinical Trials 1990;11:116-28.
25 Rahlfs V. N Handbook. Munchen: IDV-Datenanalyse und Versuchplanung, 1987.
26 Pinol A. SAMPLE. Geneva: HRP, World Health Organisation, 1995.
27 Lachin JM. Introduction to sample size determination and power analysis in clinical trials. Controlled Clinical Trials 1981;2:93-113.
28 Donner A. Approaches to sample size estimation in the design of clinical trials —a review. Statistics in Medicine 1984;3:199-214.

(Accepted 21 July 1995)