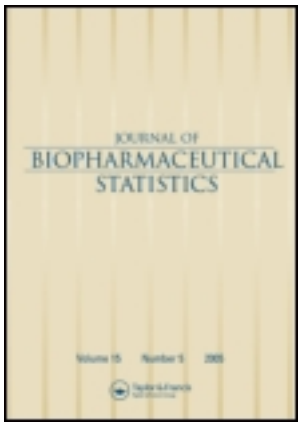

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

ESTIMATING SAMPLE SIZES FOR CONTINUOUS, BINARY, AND ORDINAL OUTCOMES IN PAIRED COMPARISONS: PRACTICAL HINTS

S. A. Julious [a] , M. J. Campbell [b] & D. G. Altman [c]

[a] Glaxo Wellcome, Clinical Pharmacology Data Sciences, Greenford Road, Greenford, London, UK

[b] Institute of General Practice and Primary Care, School of Health and Related Research, University of Sheffield, Community Sciences Centre, Northern General Hospital, Sheffield, S5 7AU, UK

[c] ICRF Medical Statistics Laboratory, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford, OX3 7LF, UK

Version of record first published: 02 Feb 2007.

Journal of Biopharmaceutical Statistics, 9(2), 241–251 (1999)

ESTIMATING SAMPLE SIZES FOR CONTINUOUS, BINARY, AND ORDINAL OUTCOMES IN PAIRED COMPARISONS: PRACTICAL HINTS

S. A. Julious,[1] *M. J. Campbell,*[2] *and D. G. Altman*[3]

[1]Clinical Pharmacology Data Sciences
Glaxo Wellcome
Greenford Road
Greenford, London, UK

[2]Institute of General Practice and Primary Care
School of Health and Related Research
University of Sheffield
Community Sciences Centre
Northern General Hospital
Sheffield S5 7AU, UK

[3]ICRF Medical Statistics Laboratory
Centre for Statistics in Medicine
Institute of Health Sciences
Old Road
Headington
Oxford OX3 7LF, UK

Key words. Sample size; Power; Paired data; Matched data; Crossover trials; Ordinal data

Abstract

Paired data occur in crossover trials and matched case-control studies, and it is rare to find studies reporting sample size calculations associated with these types of studies, despite recommendations from editors that sample size calculations should be justified. In this article we describe some simple formulas and strategies for calculating the number of patients that should be entered into a matched or paired study when the outcome measures are continuous, binary, or ordinal.

1. Introduction

Campbell et al. (1) and Phillips and Campbell (2) have described simple sample size calculations for the comparison of two or more groups. They assume that the data for one group are independent of the data for the other; that is, if the results in one group were changed they would not affect the results of the other. Such data arise from parallel group randomized trials and from unmatched case-control studies. However, a common situation is when the data are paired in some way and so the assumption of independence no longer holds. Paired data arise in a number of situations: Subjects may have their knowledge about a medical condition measured before and after an educational initiative; treatments may be evaluated on the same patients by using a crossover design; subjects with a disease may be individually matched with controls in a case-control study, for example by age, sex, and area of residence [Bland and Altman (3)]. The purpose of matching is to reduce or eliminate variability in the data and to provide a greater chance of detecting the effect of interest. Sample size calculations for independent groups are not appropriate. The purpose of this article is to describe methods for calculating sample sizes for studies that yield paired data. We consider outcomes that are either continuous, binary, or ordinal.

2. Paired Continuous Data

For the analysis of paired continuous data we are interested in the case in which a paired *t*-test is the appropriate test, so that the differences in measurements are plausibly sampled from a normal distribution.

The underlying principles of sample size calculations were explained by Campbell et al. (1). For sample size calculations we have to specify a significance level α, an effect size δ, and power $1 - \beta$. As in the two-sample case, we must also specify the standard deviation, σ, of the data. For studies with paired data, one must specify the expected standard deviation of the difference of the outcome variable measured on the two occasions, or the difference between the measurement on the case and on the matched control. Thus, if x_1 is the first measurement and x_2 the second measurement, we require the standard deviation of the differences $x_1 - x_2$, denoted σ_d. The sample size calculation uses the standardized effect size, which is defined as $d = \delta/\sigma_d$.

3. Hints on Estimating the Within-Subject Standard Deviation

It is important to distinguish here between the within- (intra-) subject standard deviation σ_w and the between- (inter-) subject standard deviation σ_b. The within-subject standard deviation quantifies the expected variation among repeated

measurements on the same individual. It is a compound of true variation in the individual and measurement error. If only one measurement is made per individual it is impossible to estimate σ_w, and one obtains σ_b, the between-subject standard deviation, which quantifies the expected variation of single measurements from different individuals. (Note that σ_b contains contributions from the within-subject variability as well, because the measurement error will affect the estimate of σ_b. An estimate of the true between-subject variability, with σ_w removed, is known as the between-subject component of variance, but we will not discuss its estimation here.) The focus of paired designs is the difference $x_1 - x_2$ within individuals or matched pairs of subjects. The standard deviation of the differences is often not known. However, σ_d can be estimated in several ways. If we have a value for σ_w (perhaps from a published study), then we can calculate σ_d as $\sigma_w\sqrt{2}$. This expression does not contain σ_b because the effect of differencing is to take out the between-subject standard deviation. Note that it is not always clear in the literature whether the statistic given is σ_d or σ_w, and so care is required. For certain clinical chemistry measurements, Fraser et al. (4) give tables for the within-subject coefficient of variation (CV), which is defined as the within-subject standard deviation divided by the mean ($\bar{x}$) multiplied by 100. Thus to obtain the within-subject standard deviation we calculate $\sigma_w = \text{CV} \times \bar{x}/100$.

If the between-subject standard deviation, σ_b, based on a single observation per individual, is known, then the standard deviation of $x_1 - x_2$ is given by $\sigma_d = \sigma_b \sqrt{2(1 - \rho)}$, where ρ is the Pearson correlation coefficient between the values of the outcome measure on the two occasions. An exploratory approach is to try out various plausible values of ρ to see what effect this will have on the sample size. It is unlikely that ρ will be negative because paired studies are designed to exploit the positive association within pairs, and so when $\rho = 0$ we obtain a conservative estimate of the required sample size. In a crossover study this is the same size sample as would be required in one arm of a two-sample study, and so the sample size for the crossover study in this case is half that of the corresponding two-sample trial. If one can be sure of a positive correlation, then the sample size will be reduced.

In none of these approaches is possible, a pragmatic way to estimate σ_d is to postulate the plausible range of values for the treatment differences among individuals and divide this range by 4.

A formula given by Guenther (5) and Machin et al. (6) is similar to that used to calculate sample sizes in single-sample studies and is given as equation (1) in the Appendix. Table 1 gives the sample sizes required for different values of d at various powers and two-sided 5% significance level, which is the usual choice.

A simple formula for the calculation of the number of pairs required (for a two-sided significance level of 5% and power of 80%) is 8 divided by the standardized difference squared plus 2, or $n = (8/d^2) + 2$ (and $n = (10.5/d^2) + 2$

Table 1. Sample Sizes Required at the Two-Sided 5% Significance Level for Different Values of the Standardized Difference (d) and Various Powers

	Power(%)				
d	99	95	90	80	50
0.1	1840	1302	1053	787	387
0.2	462	327	265	199	98
0.3	207	147	119	90	45
0.4	117	84	68	51	26
0.5	76	54	44	34	18
0.6	53	39	32	24	13
0.7	40	29	24	18	10
0.8	31	23	19	15	8
0.9	25	18	15	12	7
1.0	21	15	13	10	6
1.1	18	13	11	9	6
1.2	15	11	10	8	5
1.3	13	10	9	7	5
1.4	12	9	8	6	4
1.5	11	8	7	6	4

for a power of 90%). This formula can be used for quick approximate calculations.

Worked Example

Fraser et al. (4) give the within-subject coefficient of variation of serum cholesterol measured on 10 occasions at 14-day intervals in subjects over 70 years old as 5.8%, with a mean of 6.3 mmol/L. Then an estimate of $\sigma_w = 5.8 \times 6.3/100 = 0.3654$ and $\sigma_d = \sqrt{2} \times 0.3654 = 0.57$ mmol/L. Suppose we wished to design a crossover trial of a cholesterol lowering agent against a placebo, and we assume that there will be an adequate washout period, so that there is no carryover and no period effect. If the trial is designed to detect an effect size δ of about 0.5 mmol/L, then the standardized effect $d = 0.5/0.57 = 0.9$, and from Table 1 we would require 15 patients using a 5% significance level and 90% power, or 12 subjects for 80% power. Alternatively, for 90% power the simple formula gives $n = 10.5/0.9^2 + 2 = 15$ patients, and for 80% power the simple formula gives $n = 8/0.9^2 + 2 = 12$ patients.

4. Paired Binary Data

Similar conditions apply for binary outcomes. Again some difficulty is likely in gaining prior insight into paired responses. Suppose we wish to design a trial in

Table 2. Expected Proportions for a Binary Paired Study

First response	Second response 0	1	
0	r	s	p_1
1	t	u	$1 - p_1$
	p_2	$1 - p_2$	1

which patients are asked about symptom relief while receiving one of two medications. To calculate the required sample size we need to specify the number who will answer ''yes'' on one treatment and ''no'' on the other. The notation, given in Table 2, is taken from Gardner and Altman (7).

Here, for example, r is the proportion whose expected response is ''no'' on both treatments and p_1 and p_2 are the proportions who answer ''no'' on each occasion; they are known as marginal probabilities. The usual test to analyze these data is McNemar's test. In order to calculate the required sample size, we need to specify s and t or, equivalently, the total proportion of discordant pairs $p_{\text{dis}} = s + t$, and the odds-ratio defined as $OR = s/t$. The odds ratio OR is a measure of how much more likely it is that a patient will respond ''yes'' on the first response and ''no'' on the second as opposed to ''no'' on the first and ''yes'' on the second. The total $s + t$ is the proportion of discordant pairs. The formulas for the number of discordant pairs n_{dis} and the total number of pairs, n_{tot}, reported by Connett et al. (8), are given as equations (2) and (3) in the Appendix. A good approximation to equation (3) is given by simply dividing n_{dis} from equation (2) by p_{dis}. Table 3 gives sample sizes required for given $s + t$ and OR, for a two sided 5% significance level and power of 80%. For crossover trials Table 3 shows the number of patients; for matched case-control studies it shows the number of pairs—that is, the number of cases or the number of controls.

An investigator may find difficulty in prescribing s and t but may be able to specify p_1 and p_2, the marginal probabilities. As stated before, it is most unlikely that the outcome variable measured on each pair of subjects is negatively correlated, and so the maximum sample size will be obtained if the distributions are independent. In this case Royston (9) has suggested we can estimate s by $p_1 \times (1 - p_2)$ and t by $p_2 \times (1 - p_1)$.

Worked Example

Table 4 is taken from a paper by Morrison et al. (10) using 40 cases and controls. They wished to identify the reasons some children received more out-of-hours visits by general practitioners than others. The cases were children aged under 10

Table 3. Total Sample Sizes Required to Obtain Two-Sided 5% Significance with 80% Power for Connett's Formula from Equation (3) with Various Proportions of Discordant Pairs

	Odds ratio for discordant pairs ($OR = s/t$)							
$s + t$	2	3	4	5	10	50	100	∞
0.05	1411	626	434	351	233	168	162	155
0.10	705	312	216	175	115	83	80	77
0.15	469	207	143	116	76	55	53	50
0.20	351	155	107	86	57	41	39	37
0.25	281	124	85	69	45	32	31	29
0.30	234	103	71	57	37	26	25	24
0.35	200	88	60	49	32	22	21	20
0.40	175	77	53	42	27	19	18	18
0.45	155	68	47	37	24	17	16	15
0.50	139	61	42	33	21	15	14	14
0.55	127	55	38	30	19	13	13	12
0.60	116	50	34	27	17	12	11	11
0.65	107	46	32	25	16	11	10	10
0.70	99	43	29	23	15	10	9	9
0.75	92	40	27	22	13	9	9	8
0.80	86	37	25	20	12	8	8	7
0.85	81	35	24	19	12	8	7	7
0.90	77	33	22	18	11	7	6	6
0.95	72	31	21	17	10	6	6	5
1.00	69	29	20	16	9	6	5	4

who were identified as high out-of-hours users, and the controls, who were not high out-of-hours users, were matched by age and sex. The estimated odds ratio for a single/divorced mother was 12, which is large, and the difference between cases and controls for this example is highly significant (McNemar $\chi^2 = 9.31$, degrees of freedom (df) $= 1$, $p = 0.002$).

Suppose a researcher wished to undertake a study similar to that carried out

Table 4. Case/Control Status by Whether a Child's Mother Was Single/Divorced or Married/Cohabiting [from Morrison et al. (10)]

Cases (high out-of-hours attendance)	Controls (low out-of-hours attendance)		
	Single/divorced	Married/cohabiting	Total
Single/divorced	3	12	15
Married/cohabiting	1	24	25
Total	4	36	40

$p_1 = 15/40$, $p_2 = 4/40$, $OR = 12$.

by Morrison et al. (10) in another area of the country, but the researcher expected around 40% of cases to differ from their controls (i.e., $s + t = 0.4$). The researcher thinks an odds ratio risk of 3 would be worth detecting; that is, a single/divorced mother may be 3 times more likely to be a high out-of-hours user than a married/cohabiting one. From Table 3 the sample size to obtain two-sided 5% significance with 80% power, for $s + t = 0.4$ and $OR = 3$, is 77 cases and hence 77 controls. Equation (2) in the Appendix gives the number of discordant pairs as 29, and so the approximation described earlier gives an approximate sample size as $29/0.4 = 72.5$ or 73 cases or controls.

Suppose one were unable to specify s and t but believed that about 10% of controls would be single/divorced mothers compared to 30% of cases. Then $p_1 = 0.1$ and $p_2 = 0.3$. Assuming independence we estimate $s = 0.1 \times 0.7 = 0.07$ and $t = 0.3 \times 0.9 = 0.27$, $s + t = 0.34$, and $OR = 3.86$. (Note, here we choose $OR = t/s$ because the sample size is the same for OR and $1/OR$, and so Table 3 gives only values of OR greater than 1.) If we approximate 0.34 by 0.35 and 3.86 by 4, we can use Table 3, which shows that we need 60 pairs of cases and controls (using the exact values and equation (3) indicates 65 patients are required).

4.1 Greater Than 1:1 Matching

A case-control study can be designed with more than one control matched to each case. This can be a useful way to increase power when cases are rare. Parker and Bregman (11) show that the sample sizes should be adjusted by an allocation ratio given as equation (4) in the Appendix. The total number of cases and controls increases as the number of controls per case increases. In general, there is no benefit in having more than four controls for each case, although if the exposure variable is likely to be missing in a large number of controls, then a greater number should be sought (12).

Worked Example

Suppose the researcher wishing to repeat the results of Morrison et al. (10) believed it possible to obtain four controls for every case in the study. This would lead to a modified sample size of $n = 77 \times 5/8 = 49$ high out-of-hours users and $4 \times 49 = 196$ low out-of-hours users.

5. Ordered Categorical Data

Sometimes the outcome of interest has more than two categories. Such data can arise, for example, when patients are classified as improved, no change, or worse, or when values of a scale are divided into subgroups. Often quality of life measures

form an ordinal scale. Such ordinal data provide more information than a dichotomization and should be analyzed appropriately. Bull and Campbell (13) compared the Hospital Anxiety and Depression Scale (HADS) scores of a group of healthy women before and after they had had a mammogram. The original depression scale has 21 values, but these can be grouped into three predefined categories: normal, borderline, and abnormal. When the outcome is in the form of an ordered categorical (ordinal) scale, each observation contains more information than one measured on a nominal scale, but less information than a continuous scale, and so intuitively one would expect smaller sample sizes when the outcome is ordinal than is required for binary outcomes but larger sample sizes than required for continuous outcomes. To calculate the sample size for an outcome that is ordered categorical, the investigator has to specify not only the odds ratio but also the likely distribution of the outcome under one of the treatments. Julious and Campbell (14) have suggested that the important factor is whether the paired differences are likely to be positive, zero, or negative, and it is less important to consider the actual size of the differences. Thus for the HADS one might combine the borderline and abnormal categories and postulate how many women become depressed, stay the same, or recover from depression after mammography. This will give a larger sample size than one would expect from the ordered categorical scale, the proportionate increase being greater when a larger number of categories are split. The increase will depend on the distribution of the original scale, and Julious and Campbell (14) have shown, using simulation, that except in extreme cases when all categories are equally represented, the increase is not great. However, it is now required to estimate the total sample size. This is done by dividing the required discordant sample size by the chance of a discordant pair. The chance of a discordant pair occurring at random clearly increases as the number of categories increases; the chance of two coins showing different faces when tossed is much less than the chance of two dice showing different numbers when rolled. Thus the correction factor gets smaller as the number of categories increases.

We suggest that a rule of thumb when the number of categories exceeds two is to estimate the total sample size as the number of discordant pairs required from the formula for the two-category situation [equation (2) in the Appendix]. All that is required then is to specify the odds ratio. The increased sample size required because the ordinal nature of the data is ignored will help compensate for the fact that the sample size is not to be adjusted for discordant pairing. This approach will be progressively more conservative as the number of categories increases, and an alternative when there are a large number of categories is to treat the data as normally distributed. Sometimes the distribution of the data shows a ''floor'' or ''ceiling'' effect, particularly for quality of life measures. This occurs when one category predominates, such as the majority of people being ''healthy.'' In this case the use of neither odds ratios nor normal approximations suffices, and our suggestion is to dichotomize the data into the predominant category and others,

such as ''healthy'' and ''sick,'' and use equations (2) and (3). Note that even when such simplification is used in sample size calculations, the ordinal nature of the data should be preserved when the data are analyzed.

Worked Example

Suppose we wish to design a study that could detect a 20% increase in the risk of depression in women after they have had a mammogram ($OR = 1.2$) with 80% power and 5% (two-sided) significance; from equation (2) the number of discordant pairs required $(1.96 \times 2.2 + \sqrt{2 \times 0.8416 \times 1.2})^2/0.2^2 = 948$ patients. Thus if we use the HADS, the rule of thumb would suggest about 950 women should be entered into the study.

6. Discussion

The attraction of a crossover trial is that it requires fewer patients than an equivalent parallel group study. As has been emphasized before, however, the apparent gain in efficiency of two-period crossover trials is severely compromised by the possibility of carryover effects, or treatment by period interactions as discussed by Senn (15). One common strategy is to perform a test for carryover and, if this is significant, evaluate the treatment only on the first period, as if it were a parallel group design. Senn (15) has shown this method to be seriously biased. Thus two-period crossover designs should be used only when there is a minimal chance of carryover. To get around this one can increase the number of periods, and this enables separate estimates of the carryover effect and the treatment by period interactions. In general parallel group designs are to be preferred, but there can be occasions, when for example the disease is stable and subjects vary a great deal in their baseline values, that the efficiency of crossover trials makes them attractive. They are particularly useful in phase I bioequivalence studies.

In many matched case-control designs, the rationale for matching is to provide guidance as to the choice of suitable controls. If this is the case, and the matching criteria are only weakly linked to both exposure and outcome, then when the study is analyzed when matching is taken into account, it should yield similar results to those when matching is ignored. Thus a starting point for the size of a matched case-control study would be the corresponding size of the unmatched study.

Sample size calculations are rarely described in papers reporting studies for paired designs. The methods we have outlined make such calculations possible and allow researchers to choose an appropriate size for their studies. However, the calculations require information on the within-subject variation of the outcome measures, and we encourage authors to include such information, or similar information such as confidence intervals based on within-subject standard errors, in their reports.

Acknowledgment

We are grateful to Professor D. Machin for comments on an early manuscript.

References

1. Campbell MJ, Julious SA, Altman DG: Sample sizes for binary, ordered categorical and continuous outcomes in two group comparisons. *Br Med J* 311:1145–1148, 1995.
2. Phillips A, Campbell M: Using aspects of study design in sample size estimation. *J Biopharm Stat* 7:215–226, 1997.
3. Bland JM, Altman DG: Matching. *Br Med J* 309:1128, 1994.
4. Fraser CG, Cummings ST, Wilkinson SP, et al.: Biological variability of 26 clinical chemistry analytes in elderly people. *Clin Chem* 35:783–786, 1989.
5. Guenther WC: Sample size formulas for normal theory *t* tests. *Am Stat* 35:243–244, 1981.
6. Machin D, Campbell MJ, Fayers P, Pinol A: *Statistical Tables for the Design of Clinical Studies*. 2nd ed. Blackwell Scientific Publications, Oxford, 1997.
7. Gardner MJ, Altman DG (Eds): *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. BMJ Publications, London, 1989, p. 31.
8. Connett JE, Smith JA, McHugh RB: Sample size and power for pair-matched case-control studies. *Stat Med* 6:53–59, 1987.
9. Royston P: Exact conditional and unconditional sample size for pair-matched studies with binary outcome: A practical guide. *Stat Med* 12:699–712, 1993.
10. Morrison JM, Gilmour H, Sullivan F: Children seen frequently out of hours in one general practice. Br Med J 303:1111–1114, 1991.
11. Parker RA, Bregman DJ: Sample size for individually matched case-control studies. *Biometrics* 42:919–926, 1986.
12. Miettinen OS: Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* 25:339–355, 1969.
13. Bull AR, Campbell MJ: Assessment of the psychological impact of a breast screening programme. *Br J Radiol* 64:510–515, 1991.
14. Julious SA, Campbell MJ: Sample sizes for paired or matched ordinal data. *Stat Med* 17: 1635–1642, 1998.
15. Senn SJ: *The Design and Analysis of Cross-Over Trials*. Wiley, Chichester, 1992.

Appendix

A.1 Continuous Data

To calculate the sample size required to detect a difference d at the 100 α% level with power of 100 $(1 - \beta)$% (5):

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2} + \frac{z_{1-\alpha/2}^2}{2}. \quad (1)$$

Here $d = \delta/\sigma_d$, where δ is the effect size and σ_d is taken as the standard deviation of $x_1 - x_2$ (where x_1 and x_2 are the values of the outcome measure on each of the

pairs of measurements); $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the normal distribution corresponding to a two-sided test at the 100 α% level and power 100 $(1 - \beta)$%. For 80% (90%) power and two-sided 5% significance level, n is approximately $8/d^2 + 2$ $(10.5/d^2 + 2)$.

A.2 Binary Outcome

To calculate the number of discordant pairs, n_{dis}, required for a specified odds ratio $OR = s/t$ at two-sided 100 α% level with power 100 $(1 - \beta)$% (8):

$$n_{\text{dis}} = \frac{[z_{1-\alpha/2}(OR + 1) + 2z_{1-\beta}\sqrt{OR}]^2}{(OR - 1)^2}. \tag{2}$$

To calculate the total sample size, that is, the number of pairs n_{tot} required for a given proportion of discordant pairs $p_{\text{dis}} = s + t$ and specified odds ratio $OR = s/t$ at the two-sided 100 α% level with power 100 $(1 - \beta)$% (8),

$$n_{\text{tot}} = \frac{[z_{1-\alpha/2}(OR + 1) + z_{1-\beta}\{(OR + 1)^2 - (OR - 1)^2 p_{\text{dis}}\}^{1/2}]^2}{(OR - 1)^2 p_{\text{dis}}}. \tag{3}$$

A.3 Greater than 1:1 Matching

Given n (calculated assuming an equal number of controls to cases for a binary outcome), let n' be the number of cases in the study and qn' the number of controls. Then n' is given by

$$n' = \frac{n(q + 1)}{2q}, \tag{4}$$

where q is the ratio of controls to cases.