STATISTICS IN MEDICINE, VOL. 12, 2257–2271 (1993)

SAMPLE SIZE CALCULATIONS FOR ORDERED CATEGORICAL DATA

JOHN WHITEHEAD

Department of Applied Statistics, University of Reading, Earley Gate, P.O. Box 238, Reading RG6 2AL, U.K.

SUMMARY

Many clinical trials yield data on an ordered categorical scale such as *very good, good, moderate, poor*. Under the assumption of proportional odds, such data can be analysed using techniques of logistic regression. In simple comparisons of two treatments this approach becomes equivalent to the Mann–Whitney test. In this paper sample size formulae consistent with an eventual logistic regression analysis are derived. The influence on efficiency of the number and breadth of categories will be examined. Effects of misclassification and of stratification are discussed, and examples of the calculations are given.

1. INTRODUCTION

An important part of any clinical trial protocol is the statement of the number of patients to be recruited. Usually this is calculated from an appropriate sample size formula, to fulfil a requirement concerning the power of the study. Adjustments for possible withdrawals may then be made.

In this paper, attention will be restricted to clinical trials comparing a single experimental treatment regimen with a control, to demonstrate superior efficacy. It will be assumed that a primary patient response, reflecting efficacy, can be identified for use in sample size determination. That response will be one of a limited number of categories, ordered in terms of their desirability. A proportional odds model will be assumed in the derivation.

Sample size formulae for use with binary responses are given by Pocock[1] and Cassagrande *et al.*[2] For survival data, Schoenfeld[3] provides a formula, Schoenfeld and Richter[4] a nomogram, Freedman[5] tables and Dupont and Plummer[6] a computer program. Tables for sample size calculations in a wide variety of settings are presented in the book by Machin and Campbell.[7] Little appears to have been published about sample size calculation for ordinal data.

In Section 2, a sample size formula will be presented and justified, together with an example of its use. Like formulae developed for other types of response, it is based on a Normal approximation, and is accurate only when it generates moderate to large sample sizes. Consequences for trial design are explored in Section 3. The consequences of misclassification in categories are explored in Section 4, and the influence of prognostic factors is studied in Section 5.

2. THE SAMPLE SIZE FORMULA

To plan a clinical trial, it is necessary to look ahead and imagine the form of the data which will eventually be collected. With an ordered categorical response essential data on treatment efficacy

0277–6715/93/242257–15$12.50

Table I. Ordered categorical data available at the end of a clinical trial

Treatment	C_1	C_2	...	C_k	Total
Experimental	n_{1E}	n_{2E}	...	n_{kE}	n_E
Control	n_{1C}	n_{2C}	...	n_{kC}	n_C
Total	n_1	n_2	...	n_k	n

drawn from a total of n patients can be summarized as shown in Table I. Here and elsewhere, the subscripts E and C indicate the experimental and control groups, respectively. The possible categories of response are labelled $C_1, \ldots, C_k$, with C_i being more desirable than C_j if $i < j$. Thus C_1 is the best outcome and C_k is the worst. Lower and upper cumulative totals will be denoted as follows:

$$L_{iE} = n_{1E} + \cdots + n_{(i-1)E}, \qquad i = 2, \ldots, k,$$

$$U_{iE} = n_{(i+1)E} + \cdots + n_{kE}, \qquad i = 1, \ldots, k-1,$$

with $L_{1E} = U_{kE} = 0$. Similar definitions apply to the control group.

A model for the eventual data can be constructed. Let p_{iE} denote the probability that an individual receiving the experimental treatment gives a response in category C_i, and Q_{iE} be the probability of the outcome C_i or *better*:

$$Q_{iE} = p_{1E} + \cdots + p_{iE}, \qquad i = 1, \ldots, k,$$

($Q_{kE} = 1$). Let p_{iC} and Q_{iC} be defined similarly for the control group. The parameter θ_i, defined by

$$\theta_i = \log\left\{\frac{Q_{iE}(1-Q_{iC})}{Q_{iC}(1-Q_{iE})}\right\}, \qquad i = 1, \ldots, k-1, \tag{1}$$

is the log-odds-ratio of the outcome C_i or better for an experimental subject relative to a control subject. It is a measure of the advantage of the experimental treatment over the control, and a positive value indicates superiority of the experimental treatment. We assume that $\theta_1 = \ldots = \theta_{k-1}$, and denote their common value by θ. This is the proportional odds model described by McCullagh.[8] We suppose that the two treatment groups are homogeneous sets of patients satisfying this model; the effects of prognostic factors on sample size determination will be investigated in Section 4.

A general approach to sample size calculations can be based on the statistics Z and V, where Z is the efficient score given by the first derivative of the log-likelihood and V is Fisher's information given by minus the second derivative, both evaluated at $\theta = 0$. When nuisance parameters are present a profile likelihood, conditional likelihood or marginal likelihood can be used in place of the full likelihood.

For the model described above a marginal likelihood based on the ordering of the data is used, and the efficient score Z and Fisher's information V are given by

$$Z = \frac{1}{n+1}\sum_{i=1}^{k} n_{iC}(L_{iC} - U_{iC}) \tag{2}$$

and

$$V = \frac{n_E n_C n}{3(n+1)^2}\left\{1 - \sum_{i=1}^{k}\left[\frac{n_i}{n}\right]^3\right\}. \tag{3}$$

These statistics are derived by Jones and Whitehead:[9] equations (2.8) of that paper can be shown to be equivalent to equations (2) and (3) here, with some large sample approximations being used to simplify the expression for V. (A correction[10] to Reference 9 concerned the case of censored observations only; the original is correct for uncensored observations with ties.) The equations above also appear in Section 3 of Whitehead.[11] Siegel[12] presents a version of Mann and Whitney's[13] non-parametric test statistic M, together with an expression for its null variance, $\text{var}_0 M$. It can be shown that $M = (n + 1)Z$ and $\text{var}_0 M = (n + 1)^2 V$. Essentially we are planning an eventual Mann–Whitney test allowing for ties.

For sample size calculation, the approximate Normal distribution of Z, with mean θV and variance V can be used when θ is small and sample sizes are large. To test the null hypothesis H_0: $\theta = 0$ (the treatments are equivalent) against the alternative H_1: $\theta \neq 0$ (the treatments differ), Z will be calculated, and H_0 will be rejected if $|Z| > c$ for a suitably chosen critical value c.

A power requirement will be imposed on the trial: if $\theta = \theta_R (> 0)$, then H_0 should be rejected at significance level α with probability $1 - \beta$. The value θ_R can be referred to as the *reference improvement*. It represents the advantage of experimental treatment over control which, if present, should be detected. It is a clinically relevant difference. Taking c to be the critical value corresponding to the significance level α, we require

$$P(|Z| > c; \theta = 0) = \alpha \tag{4}$$

and

$$P(|Z| > c; \theta = \theta_R) = 1 - \beta. \tag{5}$$

According to its approximate Normal form, the distribution of Z is symmetric, so that (4) implies

$$P(Z > c; \theta = 0) = \alpha/2.$$

We denote the standard Normal distribution function by Φ, and its 100 γth upper percentage point by u_γ, so that $1 - \Phi(u_\gamma) = \gamma$, for $0 < \gamma < 1$. Then as Z has mean θV, and variance V,

$$1 - \Phi(c/\sqrt{V}) = \alpha/2,$$

and

$$c = u_{\alpha/2}\sqrt{V}. \tag{6}$$

When $\theta = \theta_R$, the probability that $Z < -c$ is negligible. Hence equation (5) implies that

$$P(Z > c; \theta = \theta_R) = 1 - \beta,$$

from which it follows that

$$-c + \theta_R V = u_\beta \sqrt{V}. \tag{7}$$

Adding (6) and (7) gives

$$V = \left[\frac{u_{\alpha/2} + u_\beta}{\theta_R}\right]^2. \tag{8}$$

Equation (8) can be used to calculate the amount of information V needed, given the values of α, β and θ_R specified by the power requirement. When the trial is completed, V will be calculated from equation (3). At the planning stage, certain components of equation (3) have to be anticipated. Assuming that recruitment to the experimental and control groups proceeds in the planned ratio $1:A$, then approximately

$$n_E = \frac{1}{A+1}n \quad \text{and} \quad n_C = \frac{A}{A+1}n.$$

The quantity A can be referred to as the *allocation ratio* (although the allocation ratio R of Reference 11 is $R = 1/A$). The overall proportions of patients who will fall into each category of response need to be anticipated: $\bar{p}_i$ will denote the anticipated proportion for category C_i, $i = 1, \ldots, k$. Then, using $n/(n+1) \approx 1$, equation (3) gives

$$V = \frac{An}{3(A+1)^2}\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right], \tag{9}$$

and combining this with equation (8) the following sample size formula is obtained:

$$n = \frac{3(A+1)^2(u_{\alpha/2} + u_\beta)^2}{A\theta_R^2\left(1 - \sum_{i=1}^{k} \bar{p}_i^3\right)}. \tag{10}$$

Example 1

Consider a comparative clinical trial in which the primary patient response is an assessment of progress after three months of treatment, made by a doctor, and classified as *very good*, *good*, *moderate* or *poor*. These categories are denoted by $C_1, \ldots, C_4$, respectively. The probabilities that a patient in the control group will give a response in category C_i (p_{iC}) and a response in category C_i *or better* (Q_{iC}), $i = 1, \ldots, 4$, can be anticipated from previous experience and the special circumstances and entry criteria for the trial. Suppose that the following values are considered appropriate:

	Response			
	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
p_{iC}	0·2	0·5	0·2	0·1
Q_{iC}	0·2	0·7	0·9	1

Thus, the control probability of a *good* or *very good* outcome is anticipated to be $Q_{2C} = 0{\cdot}7$. An increase due to the experimental treatment to $Q_{2ER} = 0{\cdot}85$ is felt to be both desirable and achievable. This is chosen as the reference improvement, to be detected as significant at the 5 per cent level with power 0·9. The extra 'R' in the subscript of Q denotes the reference improvement. The value of θ_R can be deduced:

$$\theta_R = \log\left\{\frac{Q_{2ER}(1 - Q_{2C})}{Q_{2C}(1 - Q_{2ER})}\right\}$$
$$= 0{\cdot}887.$$

Since

$$\theta_R = \log\left\{\frac{Q_{iER}(1 - Q_{iC})}{Q_{iC}(1 - Q_{iER})}\right\}$$

for $i = 1, 2$ and 3, it follows that

$$Q_{iER} = \frac{Q_{iC}}{Q_{iC} + (1 - Q_{iC})\exp(-\theta_R)}, \tag{11}$$

and so the reference improvement corresponds to the following probabilities in the experimental group:

	Response			
	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
p_{iER}	0·378	0·472	0·106	0·044
Q_{iER}	0·378	0·85	0·956	1

The category probabilities and cumulative probabilities are displayed in Figure 1. These probabilities and diagrams should be displayed to investigators while the protocol is being drafted. The anticipated effect of the experimental treatment, under the constraint of proportional odds, is systematically to shift patients into better categories.

Although investigators might consider the value of $Q_{2ER} = 0{\cdot}85$ realistic, they may now feel that $Q_{3ER} = 0{\cdot}956$ would be impossible. It might be thought that reduction of the proportion of *poor* outcomes is unlikely, and perhaps that these might rise from some side effect of the drug. Figure 1 provides a valuable way of communicating the meaning of proportional odds to investigators, and allows them to assess whether this model is appropriate. Here, we shall proceed assuming that the proportional odds model is satisfactory.

With $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}1$, the required Normal percentage points are $u_{\alpha/2} = 1{\cdot}960$ and $u_\beta = 1{\cdot}282$. Suppose that patients are evenly divided between the two treatment groups, so that the allocation ratio $A = 1$. If the reference improvement is achieved, then the overall category probabilities will be as follows:

	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
$\bar{p}_i$:	0·289	0·486	0·153	0·072

being derived from $\bar{p}_i = \frac{1}{2}(p_{iC} + p_{iER})$, $i = 1, \ldots, 4$. Now

$$1 - \sum_{i=1}^{k} \bar{p}_i^3 = 1 - 0{\cdot}143 = 0{\cdot}857,$$

and so equation (10) gives

$$n = \frac{3 \times 4(1{\cdot}960 + 1{\cdot}282)^2}{0{\cdot}887^2 \times 0{\cdot}857}$$

$$= 187.$$

Hence 94 patients are required in each group. In practice, this would usually be rounded up to 100.

In the calculation above, overall category probabilities $\bar{p}_i$ were calculated assuming that the reference improvement will be achieved. Alternatively, we could assume no treatment effect, so that $\bar{p}_i = p_{iC}$, $i = 1, \ldots, 4$. This gives $1 - \Sigma\bar{p}_i^3 = 1 - 0{\cdot}142 = 0{\cdot}858$ so that n is still equal to 187.

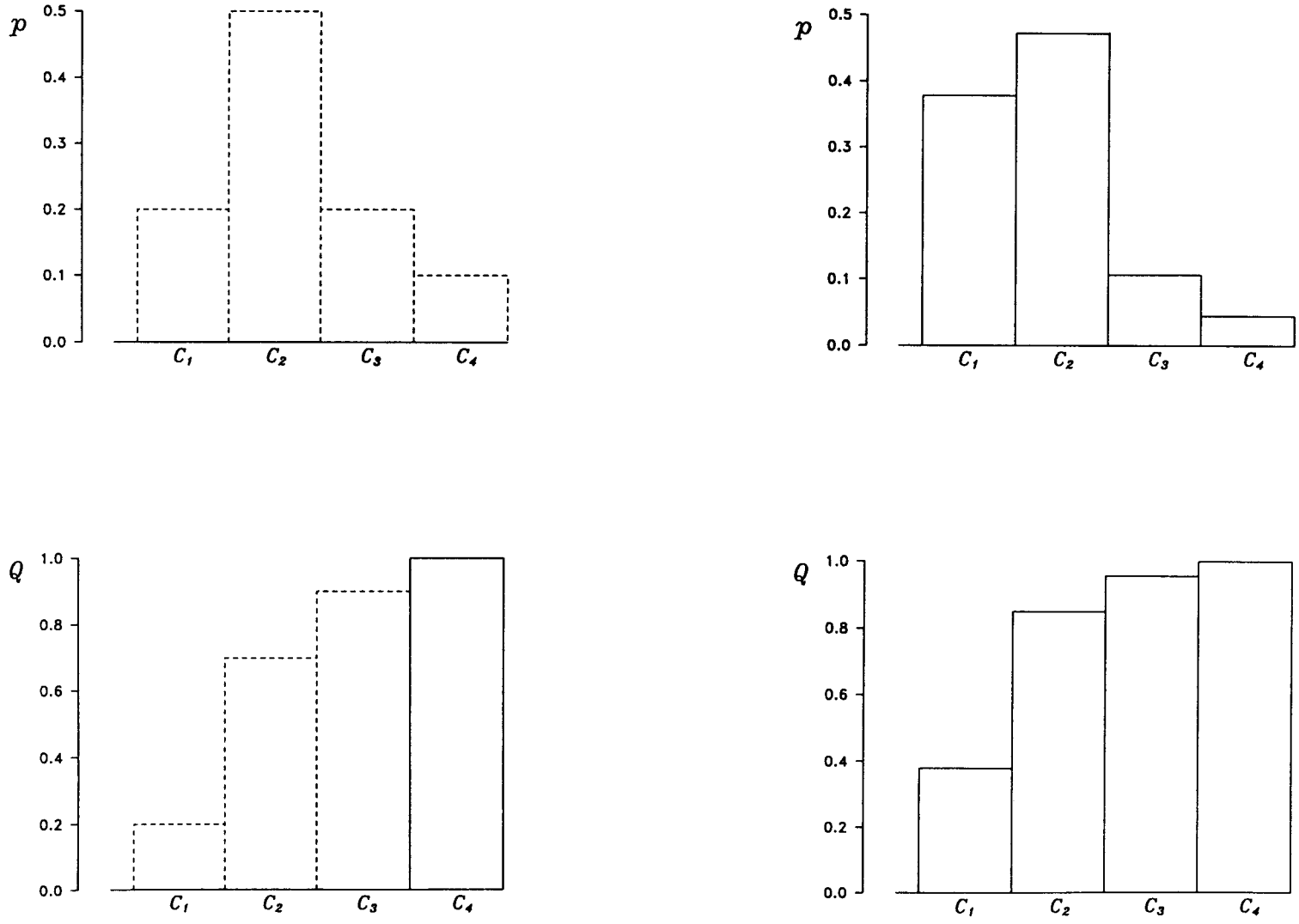

Figure 1. Absolute and cumulative category probabilities for Example 1
---- control group —— experimental group under reference improvement

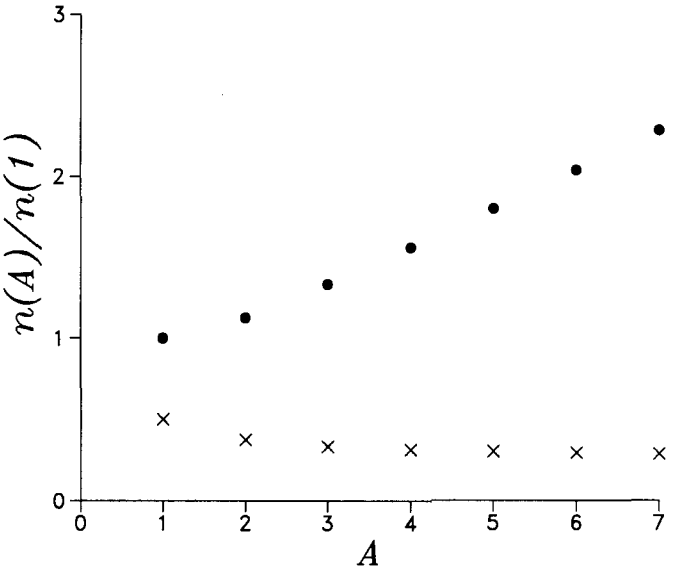

Figure 2. Relationship between sample size and allocation ratio.
• $n(A)$, the total number of subjects required for an allocation ratio of A, as a proportion of $n(1)$
× $n_E(A)$, the number of experimental subjects required for an allocation ratio of A, as a proportion of $n(1)$

When sample sizes are moderate to large, and none of the $\bar{p}_i$ are close to 1, then the value of n given by equation (10) is generally robust to small changes in the $\bar{p}_i$.

To assess the accuracy of this sample size calculation, 10,000 simulations of a trial with 187 patients (93 in the experimental group, 93 in the control group and 1 randomly allocated) were performed in each of two situations. In the first there was no difference between treatments, each having category probabilities of 0·2, 0·5, 0·2 and 0·1. The value of $|Z|/\sqrt{V}$ exceeded 1·96 to indicate rejection at the 5 per cent level in 5·05 per cent of these simulations (standard error 0·22 per cent). In the second situation the control group had category probabilities of 0·2, 0·5, 0·2 and 0·1, whereas for the experimental group these were 0·378, 0·472, 0·106. 0·044. This corresponds to the reference improvement under the proportional odds model with $\theta_R = 0{\cdot}887$. In this case, the value of $|Z|/\sqrt{V}$ exceeded 1·96 in 89·45 per cent of simulations (standard error 0·31 per cent). This compares well with the intended power of 90 per cent. The accuracy of the method is adequate in this case.

3. DESIGN CONSIDERATIONS

The dependence of the total number of patients, n, on the allocation ratio A is illustrated in Figure 2. It is assumed that the values of α, $1 - \beta$ and θ_R specified by the power requirement are fixed, and that various allocation ratios are being considered for the trial. If $n(A)$ denotes the sample size required when the allocation ratio is A, then it follows from equation (10) that

$$n(A) = \frac{(A+1)^2}{4A} n(1). \tag{12}$$

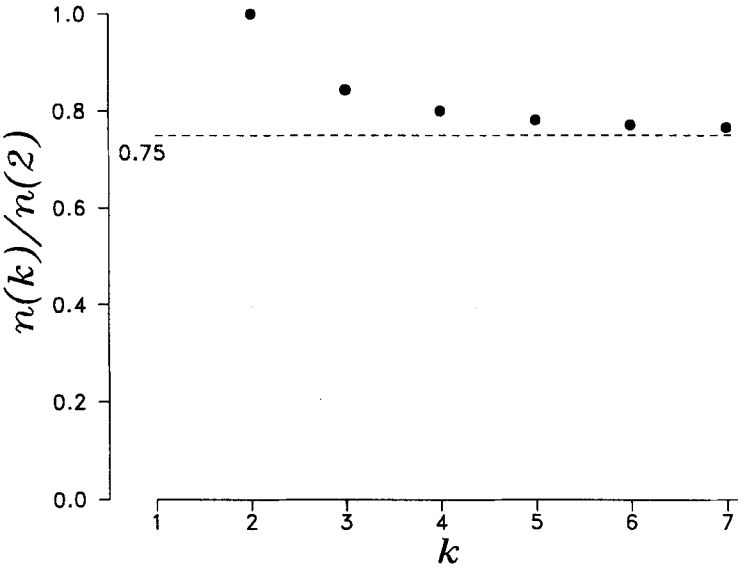

Figure 3. Relationship between sample size and number of categories, k, when $\bar{p}_1 = \ldots = \bar{p}_k$.
• $n(k)$, the total number of subjects required for k equally probable categories, as a proportion of $n(2)$
---- $n(\infty)$

Figure 2 shows how the overall sample size rises as A is increased above 1; as $A \to \infty$, $n(A) \to \infty$. Values of A greater than 1 may be justified to reduce numbers allocated to the experimental treatment, $n_E(A)$, which may be expensive or in short supply. Now

$$n_E(A) = \frac{1}{A+1} n(A) = \frac{A+1}{2A} n_E(1), \tag{13}$$

and as $A \to \infty$, $n_E(A) \to \frac{1}{2} n_E(1)$. It is evident from Figure 2 that $n_E(A)$ rapidly approaches this lower limit. In view of the steadily increasing number of controls required, allocation ratios in excess of 4 would seldom be justified.

The sample size formulae, equations (12) and (13), are valid for various types of response, including binary, survival and Normally distributed endpoints. The influence of allocation ratio is discussed by Pocock[1] (see his Figure 5.1) in terms of loss of power. In view of the large increases in sample size required to achieve only small improvements in power, it is perhaps fairer to compare sample sizes for different allocation ratios under fixed power, as has been done here.

The dependence of n on the number of categories k is illustrated in Figure 3. It is assumed that categories are equally probable ($\bar{p}_1 = \ldots = \bar{p}_k$ for all k). Denoting by $n(k)$ the sample size required when there are k equally probable categories, and keeping A, θ_R, α and β constant, it follows from equation (10) that

$$n(k) = \frac{0{\cdot}75}{1 - 1/k^2} n(2). \tag{14}$$

In the limit as $k \to \infty$, $n(k) \to 0{\cdot}75n(2)$. The limiting case is approached in large samples in which a full ranking of patient outcomes is achieved, as envisaged in the Mann–Whitney test. A full ranking is equivalent to a categorization with only one patient in each category. For $k = 5$, $n(5) = 0{\cdot}78\, n(2)$, and consequently it is of little value to use more than five categories; the

Table II. Sample sizes for different overall category probabilities

Case	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$	n/n_0
0	0·25	0·25	0·25	0·25	1
1	0·10	0·20	0·30	0·40	1·04
2	0·10	0·30	0·30	0·30	1·14
3	0·05	0·05	0·45	0·45	1·15
4	0·10	0·10	0·10	0·70	1·43

Sample size (n) with given distributions; n_0 under equally probable categories

hypothesis test will be (0·75/0·78) 100 = 96 per cent efficient relative to the use of a full ranking. When data are truly Normally distributed, the full Mann–Whitney test is in turn 94 per cent efficient relative to a t-test (Section 2.4 of Lehmann[14]). Thus the five equally probable categories design is 90 per cent efficient relative to the t-test, when the data are truly Normally distributed. These relative efficiencies are all asymptotic, and are only valid for moderate to large sample sizes.

Table II illustrates the effect of category probabilities. Again, A, θ_R, α and β are kept constant, and k is fixed at 4. The sample size required for equally probable categories is denoted by n_0. Four other possible sets of overall categories are compared with the equally probable case, which achieves the minimum possible sample size. Case 1 represents a considerable departure from equal probabilities, but there is little increase in sample size. Cases 2 and 3 depart further, and the increase in sample size becomes more important. Case 4 illustrates that sample sizes begin to increase substantially when there is a single dominant category. Because equation (10) is symmetric in the $\bar{p}_i$ this effect will occur whichever category is dominant. This pattern is already familiar from experience with binary data.

Investigators will not always have control over the number and nature of categories. If a well known and validated categorical scale is used, then it is best not to alter it. Even if many categories are involved, they should be retained, unless some pooling is required to overcome computational limitations. However, sometimes a new scale has to be devised. In other cases, patient responses may be continuous but non-Normal, and incapable of satisfactory transformation to Normality, so that a grouping into ordered categories is desirable. In such cases it will be sensible to create up to five categories in a way which leads to them having similar probabilities of occurrence. Consideration of prognostic factors, to be described in Section 5, may motivate creation of a greater number of categories.

4. EFFECTS OF MISCLASSIFICATION

Many ordered categorical scales are summaries of subjective judgements of a patient's condition. Clinicians may have to rate the patient's general well-being, or score an X-ray according to a predetermined grading system. Whenever such judgements are used, inconsistencies and errors are likely to occur. The reference improvement θ_R to be detected with specified power, will usually be set with the true categories of patients in mind. When categories are recorded with error, then differences between treatments will be diluted, and the observed value of θ will be smaller than the true value. The purpose of this section is to point out the dangers of misclassification and to quantify their effects. Two examples will be given.

Example 2

In Example 1, Section 2 a sample size was sought for a trial designed to detect an improvement of category probabilities from

	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
p_{iC}:	0·2	0·5	0·2	0·1

to

	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
p_{iER}:	0·378	0·472	0·106	0·044

The reference improvement was $\theta_R = 0{\cdot}887$, and the choices $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}10$ led to a total sample size requirement of $n = 187$.

Now suppose that the category probabilities above refer to the true state of the patient, but that the grading of the patient is done with error. A simple model might be that 20 per cent of patients in category C_i are actually classified as being in C_{i-1} (if $i > 1$) and 20 per cent are classified as being in C_{i+1} (if $i < 4$). For example, 0·5 of the control group is likely to be in C_2. Twenty per cent of these, 0·1 of the total, will be classified as C_1, and 0·1 of the total will be classified as C_3. Thus 0·3 of the control group will be true C_2's, classified as such. In addition $0{\cdot}2 \times 0{\cdot}2 = 0{\cdot}04$ of the control group will be C_1's classified as C_2, and 0·04 will be C_3's classified as C_2. The observed proportion of control patients in C_2 will be 0·38. The true category probabilities given at the beginning of this example will be transformed by misclassification into the observed probabilities below:

	Very good (C_1)	Good (C_2)	Moderate (C_3)	Poor (C_4)
$p_{iC}^{(0)} =$	0·26	0·38	0·24	0·12
$p_{iER}^{(0)} =$	0·397	0·380	0·167	0·056

The superscript (0) denotes observed probabilities.

The two sets of observed probabilities no longer satisfy the proportional odds assumption. The log-odds-ratios for being in category C_1, C_1 or C_2, and C_1, C_2 or C_3 are 0·627, 0·672 and 0·832, respectively. A rough overall value for $\theta_R^{(0)}$ may be found by averaging these, weighting by $\bar{Q}_i^{(0)}(1 - \bar{Q}_i^{(0)})$ to allow for the different variances of their estimates, where

$$\bar{Q}_i^{(0)} = \tfrac{1}{2}(Q_{iC}^{(0)} + Q_{iER}^{(0)}), \qquad Q_{iC}^{(0)} = p_{1C}^{(0)} + \cdots + p_{iC}^{(0)},$$

and similarly for $Q_{iER}^{(0)}$. This weighted averaging procedure gives $\theta_R^{(0)} = 0{\cdot}678$. A recalculation of sample size now gives $n = 305$; a considerable increase.

To summarize, a true improvement in patients' responses corresponding to $\theta_R = 0{\cdot}887$, will be observed through the obscuring effect of misclassification as a more modest improvement of

$\theta_R^{(0)} = 0{\cdot}678$. The sample size needs to be set to detect observed rather than true differences, and so 305 patients are required. An impression of misclassification probabilities might be gained from the results of a consistency study in which all clinicians score a common set of patients. Rather than increasing sample size, it may be concluded that efforts to improve consistency will be more cost-effective. The example above is only meant to illustrate how a rough idea of the effects of misclassification might be gained: it is not intended as an accurate alternative method of sample size determination.

Example 3

Suppose that clinicians traditionally rate patients' outcomes as *success* or *failure*. An improvement from 50 per cent success on the control to 70 per cent success on the experimental treatment is sought. This corresponds to a log-odds-ratio of $\theta_R = 0{\cdot}847$. With $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}1$ the method of Section 2 gives a sample size of $n = 244$. Notice that when $k = 2$

$$1 - \sum_{i=1}^{2} \bar{p}_1^3 = 3\bar{p}_1(1 - \bar{p}_1)$$

so that when $A = 1$, equation (10) becomes

$$n = \frac{4(u_{\alpha/2} + u_\beta)^2}{\theta_R^2 \bar{p}_1(1 - \bar{p}_1)} \tag{15}$$

which is perhaps a more familiar expression for the binary case.

In order to reduce sample size, it is proposed to subdivide the *success* category into *complete success* (C_1) and *partial success* (C_2). The *failure* category will also be subdivided into *partial failure* (C_3) and *complete failure* (C_4). Control and experimental category probabilities, with $\theta_R = 0{\cdot}847$, would be anticipated as follows:

	Success category			
	Complete success (C_1)	Partial success (C_2)	Partial failure (C_3)	Complete failure (C_4)
p_{iC}:	0·2	0·3	0·3	0·2
p_{iER}:	0·368	0·332	0·203	0·097

The method of Section 2 gives a sample size of $n = 190$. This is a substantial reduction from the sample size of 244 needed when the classification is binary.

Next, suppose that there is 20 per cent misclassification between C_1 and C_2 and between C_3 and C_4. The overall judgement of *success* and *failure* is assumed reliable. The following observed category probabilities would result:

	Success category			
	Complete success (C_1)	Partial success (C_2)	Partial failure (C_3)	Complete failure (C_4)
$p_{iC}^{(0)}$:	0·22	0·28	0·28	0·22
$p_{iER}^{(0)}$:	0·361	0·339	0·182	0·118

The three different log-odds-ratios for the better outcome are 0·695, 0·847 and 0·746, with a weighted average of $\theta_R^{(0)} = 0{\cdot}769$ (following the approach of Example 2). The recalculated sample size is $n = 230$. Comparing this with 190 for no misclassification and 244 for binary outcomes, it can be seen that misclassification has removed most of the advantage gained from subdividing the two main categories. Such considerations must temper the advice given in Section 3 concerning the advantages of using four to five categories when possible.

5. THE INFLUENCE OF PROGNOSTIC FACTORS

Often it will be important to adjust the primary ordered categorical analysis of a clinical trial for prognostic factors. The adjustment could take the form of a stratified analysis, rather like the meta-analysis procedure described by Whitehead and Whitehead.[15] Alternatively, linear modelling could be implemented using PROC LOGISTIC in SAS,[16] program PR of BMDP,[17] or library procedure 'ordinallogistic' of Genstat 5.[18]

The main objective of the primary analysis is to compare the outcomes of control and experimental patients, for whom all prognostic factors other than treatment were the same.

It is well known that adjustment for prognostic factors improves the power of analyses of Normally distributed observations. For survival data, adjustment has little effect on power (Schoenfeld[19]). Robinson and Jewell[20] have pointed out that covariate adjustment in the logistic regression analysis of binary data can lead to an apparent loss of power. The same is true in the case of ordered categorical data. To preserve power, it will be necessary to increase sample size.

The calculations presented here will be for stratification. Even when the final analysis will involve linear modelling, sample size calculations based on an approximate stratified model will usually be appropriate. Suppose that patients in the trial belong to one of m strata, $S_1, \ldots, S_m$. The anticipated proportion of patients in S_h will be denoted by s_h, $h = 1, \ldots, m$. The overall probability (taking both treatment groups together) of a patient in stratum S_h having outcome category C_i will be denoted by $\bar{p}_{hi}$. The allocation ratio will be A in all strata. For each stratum, the information contributed by the ns_h patients recruited will be given by equation (9):

$$V_h = \frac{An\,s_h}{3(A+1)^2}\left[1 - \sum_{i=1}^{k} \bar{p}_{hi}^3\right].$$

It follows that a generalisation of the formula for n is given by

$$n = \frac{3(A+1)^2(u_{\alpha/2} + u_\beta)^2}{A(\theta_R^*)^2 \sum_{h=1}^{m} s_h\left[1 - \sum_{i=1}^{k} \bar{p}_{hi}^3\right]}. \tag{16}$$

The reference improvement θ_R^* is now the log-odds-ratio comparing an experimental and a control patient *within the same stratum*; it is assumed to be constant in all strata. The term $(1 - \Sigma \bar{p}_{hi}^3)$ is minimized when $\bar{p}_{h1} = \ldots = \bar{p}_{hk}$. Recognition of stratification when the sample size is calculated will result in setting a power requirement in terms of θ_R^* rather than θ_R. As strata probabilities will be more variable than overall probabilities, unless a larger target is set for θ_R^* than would have been set for θ_R, there will be a consequent increase in sample size. The use of formula (16) requires the anticipation of the proportions s_h and individual strata probability distributions $\bar{p}_{h1}, \ldots, \bar{p}_{hk}$. It is unlikely that these quantities will be foreseeable with any confidence: a better way of using (16) might be to try out various scenarios to evaluate the robustness of a proposed design.

Table III. Anticipated distributions of outcome categories for each stratum of Example 4

Baseline category (stratum)	Proportion in each stratum (s_h)	Category probabilities following treatment ($\bar{p}_{hi}$)					
		C_1	C_2	C_3	C_4	C_5	C_6
C_3	0·4	0·3	0·5	0·2	0	0	0
C_4	0·3	0	0	0·6	0·4	0	0
C_5	0·2	0	0	0·2	0·3	0·5	0
C_6	0·1	0	0	0	0·2	0·2	0·6
Overall		0·12	0·20	0·30	0·20	0·12	0·06

Example 4

For a six category response, overall trial probabilities are anticipated to be

	(C_1)	(C_2)	(C_3)	(C_4)	(C_5)	(C_6)
$\bar{p}_i$:	0·12	0·20	0·30	0·20	0·12	0·06.

Now $(1 - \Sigma\bar{p}_i^3) = 0{\cdot}953$ and so from equation (10) we see that

$$n = \frac{3(A+1)^2(u_{\alpha/2} + u_\beta)^2}{0{\cdot}953\, A\, \theta_R^2}. \tag{17}$$

Now the baseline category of the patient, representing condition at time of commencement of treatment, is an important prognostic factor. Only patients with response categories $C_3, \ldots, C_6$ will be admitted to the trial, as the others do not require such intensive therapy. Table III shows the proportions in each of these strata likely to achieve the six outcome categories after treatment. Also shown are the proportions of the total patient population likely to fall into each baseline category. The overall outcome category probabilities are the same as the $\bar{p}_i$ given above.

In recalculating the sample size to allow for stratification, the values of α, β and A will remain unchanged. From Table III

$$\sum_{h=3}^{6} s_h \left(1 - \sum_{i=1}^{6} \bar{p}_{hi}^3\right) = 0{\cdot}797.$$

Hence, equation (16) gives

$$n_S = \frac{3(A+1)^2(u_{\alpha/2} + u_\beta)^2}{0{\cdot}797\, A(\theta_R^*)^2}, \tag{18}$$

where n_S denotes the sample size calculated allowing for stratification. Assuming that the same value would be chosen for θ_R^* if equation (18) were used as for θ_R in equation (17), the ratio n_S/n is equal to 1·196. Thus a 20 per cent increase in sample size is necessary to allow for stratification. The example shows how extreme the effect of stratification has to be to affect sample size materially. More modest variations in outcome category distributions, such as may arise in different treatment centres, are unlikely to have a major impact on power.

6. CONCLUSIONS

In this paper a sample size formula has been presented for use when the primary patient response is measured on an ordered categorical scale. The proportional odds model underlies the derivation. When investigators have control over the number and definition of categories, then sample sizes can be reduced by increasing the number above two, but there is little gain from using more than five. Categories which are equally likely to occur, lead to greatest efficiency. Efficiency falls appreciably when one category becomes dominant. Misclassification can, in practice, cause the loss of apparent gains in efficiency due to the creation of extra categories of outcome. When the outcome response is heavily dependent on membership of some stratum of the patient population, this should be allowed for both in design and analysis. The consequence will be an increase in sample size.

The sample size calculations described are implemented within the computer package PEST3,[21] (in which the term *allocation ratio* is taken to mean $R = 1/A$). The principal purpose of PEST3 is the design and analysis of *sequential* clinical trials, including those with ordered categorical outcomes. The sequential approach ensures that the trial lasts long enough to secure the power required even when there is little prior information with which to anticipate the values of such quantities as $\bar{p}_i$, $i = 1, \ldots, k$ and in the stratified case S_h and $\bar{p}_{hi}$, $h = 1, \ldots, m$; $i = 1, \ldots, k$. In PEST3 a simulation facility allows the user to check the accuracy of the error rates of both fixed sample and sequential designs based on ordered categorical responses. This facility was used in the simulations of Example 1. Simple Mann–Whitney analyses and stratified analyses are also be available.

An alternative to the proportional odds model is provided by the continuation ratio model (McCullagh and Nelder[22]). The difference between the two is only one of detail, only in large datasets could it be evident that one is a good fit while the other is not. The principles discussed here can be used to obtain a sample size formula from the continuation ratio model.

Although there does not seem to be any published papers concerning sample size calculations for ordered categorical data, there are two accounts of new procedures in press at the time of writing. These are by Hilton and Mehta[23] and by Lesaffre *et al.*[24] Both take quite different approaches from that presented here.

REFERENCES

1. Pocock, S. J. *Clinical Trials: A Practical Approach*, Wiley, Chichester, 1983.
2. Cassagrande, J. T., Pike, M. C. and Smith, P. G. 'The power function of the exact test for comparing two binomial distributions', *Applied Statistics*, **27**, 176–180 (1978).
3. Schoenfeld, D. A. 'The asymptotic properties of nonparametric tests for comparing survival distributions', *Biometrika*, **68**, 316–319 (1981).
4. Schoenfeld, D. A. and Richter, J. R. 'Nomograms for calculating the number of, patients needed for a clinical trial with survival as an end point', *Biometrics*, **38**, 163–170 (1982).
5. Freedman, L. S. 'Table of the number of patients required in clinical trials using the logrank test', *Statistics in Medicine*, **1**, 121–130 (1982).
6. Dupont, W. D. and Plummer, W. D. 'Power and sample size calculations: a review and computer program', *Controlled Clinical Trials*, **11**, 116–128 (1990).
7. Machin, D. and Campbell, M. J. *Statistical Tables for the Design of Clinical Trials*, Blackwell, Oxford, 1987.
8. McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society, Series B*, **43**, 109–142 (1980).
9. Jones, D. R. and Whitehead, J. 'Sequential forms of the log rank and modified Wilcoxon tests for censored data', *Biometrika*, **66**, 105–113 (1979).
10. Correction to above. *Biometrika*, **68**, 576 (1981).

11. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, (2nd edn), Ellis Horwood, Chichester, 1992.
12. Siegei, S. *Nonparametric Statistics for the Behavioural Sciences*, McGraw-Hill, New York, 1956.
13. Mann, H. B. and Whitney, D. R. 'On a test of whether one of two variables is stochastically larger', *Annals of Mathematical Statistics*, **33**, 498–512 (1947).
14. Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.
15. Whitehead, A. and Whitehead, J. 'A general parametric approach to the meta-analysis of randomised clinical trials', *Statistics in Medicine*, **10**, 1665–1677 (1991).
16. SAS Institute Inc. *SAS Technical Report P-200, SAS/STAT Software: CALIS and LOGISTIC Procedures, Release 6.04*, SAS Institute Inc, Cary, NC, 1990.
17. Dixon, W. J. (Ed.) *BMDP Statistical Software Manual, Volume 2*, University of California Press, Berkeley, 1990.
18. Payne, R. W. and Arnold, G. M. (Eds.). *Genstat 5, Procedure Library Manual, Release 1.3(2)*, NAG Ltd., Oxford, 1989.
19. Schoenfeld, D. A. 'Sample size formulae for the proportional hazards regression model', *Biometrics*, **39**, 499–503 (1983).
20. Robinson, L. D. and Jewell, N. P. 'Some surprising results about covariate adjustment in logistic regression models', *International Statistical Review*, **59**, 227–240 (1991).
21. Brunier, H. and Whitehead, J. *PEST3: Operating Manual*, University of Reading, 1993.
22. McCullagh, P. and Nelder, J. *Generalised Linear Models*, (2nd edn), Chapman and Hall, London, 1989.
23. Hilton, J. and Mehta, C. R. 'Power and sample size calculations for exact conditional tests with ordered categorical data', *Biometrics* (in press).
24. Lesaffre, E., Scheys, I., Fröhlich, J. and Bluhmki, E. Calculation of power and sample size with bounded outcome scores', *Statistics in Medicine*, **12**, 1063–1078 (1993).