STATISTICS IN MEDICINE, VOL. 7, 759–764 (1988)

SAMPLE SIZE REQUIREMENTS FOR STUDIES ESTIMATING ODDS RATIOS OR RELATIVE RISKS

STANLEY LEMESHOW, DAVID W. HOSMER, JR., AND JANELLE KLAR
University of Massachusetts, Arnold House, Amherst, Massachusetts 01003, U.S.A.

SUMMARY

This paper presents formulae for determining the number of subjects necessary, in either a case-control or a cohort study, to estimate the odds ratio or relative risk, respectively, to within a selected percentage (ε) of the true population value with some specified probability. This approach differs somewhat from previous comparable work that estimated the *log* odds ratio within a stated fixed distance rather than as a percentage of the actual odds ratio. Comparable development for relative risk has not previously appeared in the literature. These formulae provide guidelines for determination of study size that does not depend on hypothesis testing considerations.

KEY WORDS Confidence interval Case-control study Cohort study Epidemiology

INTRODUCTION

Standard epidemiological methods textbooks[1,2] present formulae and/or tables to determine sample size requirements when odds ratios or relative risks are the parameters of interest. The basis for these formulae is a test of the hypothesis that the population odds ratio (or relative risk) is 1 versus some alternative hypothesis. The resulting sample size will assure that the investigator will reject the null hypothesis with some stated probability (power), when the true odds ratio (or relative risk) is a specified value different from one, and the investigator performs the test at some stated type I (or α) level.

There is, however, a growing feeling among many epidemiologists engaged in applied research that hypothesis tests may not provide appropriate solutions to research questions. They argue that the confidence interval contains more information for the population measure of effect than that obtained from the *p*-value of a significance test that results in a 'reject'/'fail-to-reject' decision. This makes it desirable to obtain point estimates close to the actual measure of effect in the population. This ability relates directly to sample size.

Along these lines, O'Neill[3] presented sample size requirements for case-control studies whose objective was the provision of confidence intervals for the log odds ratio of predetermined width. Extending this work, Smith *et al.*[4] considered sample size requirements for matched case-control studies. As in the O'Neill paper, interest focused on the width of the confidence interval for the log odds ratio. To date no comparable work has appeared for estimation of relative risks.

This paper develops formulae for sample sizes necessary to assure that an estimate of the odds ratio (rather than log odds ratio) for case-control studies, and the relative risk for cohort studies, is suitably close to the true population parameter. Although similar in appearance, the goal of our approach differs from that of confidence interval estimation where the width of the interval is a fixed distance rather than a percentage of the population parameter. One can easily evaluate these

0277–6715/88/070759–06$05.00

Received September 1987
Revised December 1987

formulae and thus readily construct tables for quick reference. We provide several examples to illustrate the use of the formulae and the resulting tables.

SAMPLE SIZE FOR ESTIMATING THE ODDS RATIO WITH STATED PRECISION 'ε'

Consider the 2 × 2 contingency table, Table I, that displays the results of a case-control study with a dichotomous exposure variable. An estimate of the 'effect' of exposure is the odds ratio

$$\hat{\text{OR}} = \frac{ad}{bc}.$$

Denote the true odds ratio in the population as OR. Suppose we wish to have $100(1-\alpha)$ per cent confidence that the estimate, $\hat{\text{OR}}$, is suitably close to the population parameter, OR. For very large sample sizes, we can accomplish this by dealing with the sampling distribution of $\hat{\text{OR}}$, which is approximately normal $N(\text{OR}, \text{var}(\hat{\text{OR}}))$. As a result, the probability of observing an estimated odds ratio somewhere in the following interval is $100(1-\alpha)$ per cent

$$\text{OR} - z_{1-\alpha/2}\sqrt{\{\text{var}\,(\hat{\text{OR}})\}} \leqslant \hat{\text{OR}} \leqslant \text{OR} + z_{1-\alpha/2}\sqrt{\text{var}\,(\hat{\text{OR}})}.$$

Here var $(\hat{\text{OR}})$ is the variance of the sampling distribution of $\hat{\text{OR}}$ and $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)$ per cent point of the standard normal distribution.

A problem arises with this approach since the normal approximation holds only for very large sample sizes. This is, in part, due to the fact that the distribution of OR ranges between 0 and ∞, with the 'no effect' or null value being 1. Fortunately, the sampling distribution of the $\ln(\hat{\text{OR}})$ is nearly normally distributed for much smaller sample sizes than that of $\hat{\text{OR}}$. As a result, we can calculate an interval based on the sampling distribution of $\ln(\hat{\text{OR}})$, and then transform the results to yield values for $\hat{\text{OR}}$ by exponentiation. However, after exponentiation, the resulting interval is asymmetric with the direction of the skew away from 1. This fact provides the rationale for the proposed approach, which is a two-step procedure. First we construct an interval containing $\ln(\hat{\text{OR}})$ with probability $100(1-\alpha)$ per cent as follows:

$$\ln \text{OR} - z_{1-\alpha/2}\sqrt{\text{var}\,(\ln \hat{\text{OR}})} \leqslant \ln \hat{\text{OR}} \leqslant \ln \text{OR} + z_{1-\alpha/2}\sqrt{\text{var}\,(\ln \hat{\text{OR}})}.$$

Second, we obtain the corresponding interval for $\hat{\text{OR}}$ by exponentiating the upper and lower bounds of the above interval.

Let P_1^* and P_2^* denote the true probabilities of exposure given disease presence or absence, respectively. The variance of the sampling distribution of $\ln(\hat{\text{OR}})$ is approximated (assuming $n_1 = n_2 = n$) as:

$$\text{var}\,(\ln \hat{\text{OR}}) \cong \frac{1}{nP_1^*} + \frac{1}{n(1-P_1^*)} + \frac{1}{nP_2^*} + \frac{1}{n(1-P_2^*)}.$$

Since this expression involves unknown population parameters, we need data from a pilot study, previous studies, or the literature, to obtain an estimate.

Figure 1 depicts the use of the ln (OR) and the relationship between the end points of the intervals as defined on two scales, when the point estimate is greater than 1·0. The notation 'SE' stands for the standard error of the estimated parameter (that is, SE = $\sqrt{\text{var}}$) and z stands for $z_{1-\alpha/2}$.

That is, based on the log scale, we know that, with probability $100(1-\alpha)$, the value of ln $(\hat{\text{OR}})$ will fall somewhere between $\ln(\text{OR})_u$ and $\ln(\text{OR})_L$ in Figure 1. Notice that whereas the interval

Table I. Contingency table for case-control study

	Exposed	Not exposed	
Case	a	b	n_1
Control	c	d	n_2
	m_1	m_2	N

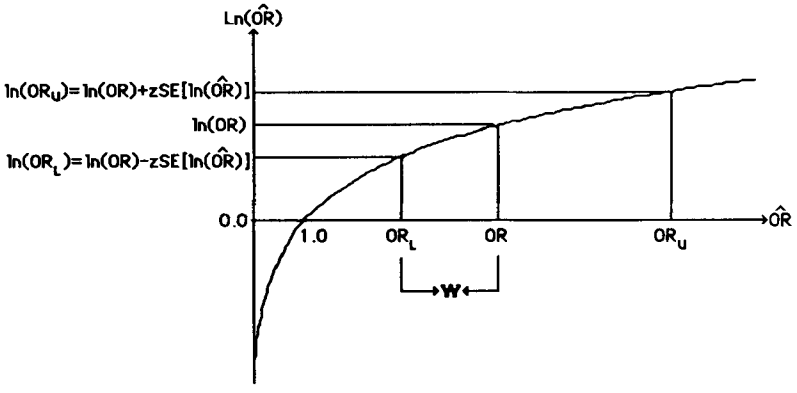

Figure 1. Plot of confidence interval for ln (OR) versus confidence interval for OR

established for ln (ÔR) is symmetric, the exponentiated values yield an interval on the ÔR scale which is asymmetric.

We wish to determine the number of study subjects, n, required in each of the case and control groups so that the width of the stated portion of the interval has length w.

For discussion purposes only, we make a number of assumptions at this point. First, we assume that the OR > 1. If this is not the case, we interchange the definition of 'exposed' and 'unexposed' without loss of generality in determination of sample size. Second, we wish to control the distance between OR_L and OR (w in Figure 1) since control of the distance between OR_U and OR would usually result in unrealistically large sample size requirements.

We find it much more informative to define the width w as a function of the odds ratio so we choose n to estimate the OR to within ε of its true value. We do this via the equation $w = \varepsilon OR$. It follows from Figure 1 that

$$w = \varepsilon OR = e^{\ln(OR)} - e^{\ln(OR) - zSE(\ln(\hat{OR}))}$$

$$\varepsilon OR = OR - OR\, e^{-zSE(\ln(\hat{OR}))}$$

$$\varepsilon = 1 - e^{-zSE(\ln(\hat{OR}))}$$

$$1 - \varepsilon = e^{-zSE(\ln(\hat{OR}))}$$

$$\ln(1-\varepsilon) = -zSE(\ln(\hat{OR}))$$

or

$$\ln(1-\varepsilon) = -z_{1-\alpha/2}\left[\frac{1}{n}\left\{\frac{1}{P_1^*(1-P_1^*)} + \frac{1}{P_2^*(1-P_2^*)}\right\}\right]^{1/2}$$

and solving for n,

$$n = \frac{z^2_{1-\alpha/2}\left\{\dfrac{1}{P_1^*(1-P_1^*)} + \dfrac{1}{P_2^*(1-P_2^*)}\right\}}{[\ln(1-\varepsilon)]^2}.$$

Note that there are three parameters (P_1^*, P_2^* and OR), but we need specify only two since any two determine the third. For example, with P_2^* and OR given,

$$P_1^* = \frac{\mathrm{OR}\,P_2^*}{\mathrm{OR}\,P_2^* + (1-P_2^*)}.$$

In practice the values of P_1^*, P_2^* and OR will be unknown. As a result, we must approximate them from sample data such as those shown in Table I or have some idea of their values from other sources. When we use sample data or other approximations, our statements change from ones involving probability to ones involving confidence.

To illustrate, consider the following example. Suppose we wish to know what sample size we need in each of two groups, cases and controls, to have 95 per cent confidence in estimation of the population odds ratio to within 10 per cent of the true value when we believe this true value is in the vicinity of 2. Furthermore, suppose that the exposure rate among the controls is approximately 0·30, (OR = 2, $P_2^* = 0{\cdot}30$). It follows that the proportion exposed among the cases is

$$P_1^* = \frac{2 \times 0{\cdot}3}{2 \times 0{\cdot}3 + 0{\cdot}7} = 0{\cdot}46.$$

Evaluating the required sample size from the above formula we have

$$n = \frac{1{\cdot}96^2\left\{\dfrac{1}{0{\cdot}46 \times 0{\cdot}54} + \dfrac{1}{0{\cdot}3 \times 0{\cdot}7}\right\}}{[\ln(1-0{\cdot}1)]^2} = 3041.$$

Hence, we need 3041 subjects in each of the case and control groups to estimate the odds ratio to within 10 per cent of its true value with 95 per cent confidence.

Now suppose we change ε to 0·5. Calculation for the sample size in this example is:

$$n = \frac{1{\cdot}96^2\left\{\dfrac{1}{0{\cdot}46 \times 0{\cdot}54} + \dfrac{1}{0{\cdot}3 \times 0{\cdot}7}\right\}}{[\ln(1-0{\cdot}5)]^2} = 70{\cdot}3.$$

Hence, only 71 subjects would be required in each of the two groups to estimate the OR to within 50 per cent of its true value.

The above example demonstrates that estimation of the odds ratio with moderate to high level of precision requires very large samples.

SAMPLE SIZE FOR ESTIMATING THE RELATIVE RISK WITH STATED PRECISION 'ε'

Consider now the cohort or follow-up study, with interest directed on the relative risk (or cumulative incidence ratio). Referring to Table I, we estimate the relative risk by

$$\hat{\mathrm{RR}} = \frac{a/m_1}{b/m_2}.$$

The normal approximation to the sampling distribution of the RR suffers from the same problems as that of the odds ratio. Hence we use the same two-step procedure; we begin with a transformation to the log scale and later transform the results by exponentiation.

An approximation to the variance of ln (RR) is

$$\text{var}[\ln \hat{\text{RR}}] \cong \frac{1-P_1}{mP_1} + \frac{1-P_2}{mP_2},$$

where we assume $m_1 = m_2 = m$ and P_1 and P_2 are the population proportions of exposed and unexposed individuals who develop the disease, respectively. We use ε in the same sense as in the previous section and assume that we wish to estimate the relative risk (RR) to within ε of the true population value. We must first, as was illustrated for the odds ratio, express the desired precision on the log scale. Figure 1 is again helpful here, with the understanding that we replace all references to OR with RR.

In this situation,

$$\begin{aligned} w = \varepsilon \text{RR} &= e^{\ln(\text{RR})} - e^{\ln(\text{RR}) - z\text{SE}(\ln(\hat{\text{RR}}))} \\ \varepsilon \text{RR} &= \text{RR}[1 - e^{-z\text{SE}(\ln(\hat{\text{RR}}))}] \\ 1 - \varepsilon &= e^{-z\text{SE}(\ln(\hat{\text{RR}}))} \\ \ln(1-\varepsilon) &= -z\text{SE}(\ln(\hat{\text{RR}})) \end{aligned}$$

or

$$\ln(1-\varepsilon) = -z_{1-\alpha/2}\left[\frac{1}{m}\left\{\frac{(1-P_1)}{P_1} + \frac{(1-P_2)}{P_2}\right\}\right]^{1/2}.$$

Thus, the necessary sample size is

$$m = \frac{z^2_{1-\alpha/2}\left\{\dfrac{(1-P_1)}{P_1} + \dfrac{(1-P_2)}{P_2}\right\}}{[\ln(1-\varepsilon)]^2}.$$

To illustrate the use of this formula, consider the planning of a cohort study where we expect the outcome will occur in 20 per cent of the unexposed group. What sample size do we need in each of the two groups to estimate the relative risk to within 10 per cent of the true value, which we believe is approximately 1·75, with 95 per cent confidence?

It follows from the given information that

$$\begin{aligned} P_2 &= 0{\cdot}2 \\ P_1 &= (\text{RR})P_2 = 0{\cdot}35 \end{aligned}$$

and

$$m = \frac{1{\cdot}96^2\left\{\dfrac{(0{\cdot}65)}{0{\cdot}35} + \dfrac{(0{\cdot}8)}{0{\cdot}2}\right\}}{[\ln(1-0{\cdot}1)]^2} = 2027.$$

Hence we would need 2027 in each of the two exposure groups. If we reduce our required precision to $\varepsilon = 0{\cdot}5$, the size is 47 subjects in each of the two exposure groups.

DISCUSSION

In this paper we have presented formulae for determining the number of subjects necessary to estimate the odds ratio in a case-control study, or the relative risk in a cohort study, to within a selected percentage (ε) of the true population value with specified probability ($1-\alpha$). These formulae provide guidelines for determination of study size that does not depend on hypothesis testing considerations. Clearly, the validity of the resulting sample size estimate depends upon our ability to accurately estimate var(ln ($\hat{OR}$)), which is a function of P_1^* and P_2^*. If we had precise knowledge of P_1^* and P_2^*, then performing the case-control study would be of dubious value. However, since this is rarely the case, it is our experience that sample size should be estimated several times, each time under a different possible combination of P_1^* and P_2^*. If our estimate of P_2^* is close to the true value and if the true odds ratio is as large or larger than the pre-specified value, then the proposed sample size formula will result in an estimate of OR which will be within ε of OR with probability at least $1-\alpha$. Final sample size should always be chosen as a compromise between anticipated population values and practical considerations. While the choice of ε is somewhat subjective, it should depend upon the anticipated population OR or RR and the goals of the study.

REFERENCES

1. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn., Wiley, New York, 1981.
2. Schlesselman, J. J. *Case-Control Studies: Design, Conduct, Analysis*, Oxford University Press, New York, 1982.
3. O'Neill, R. T. 'Sample sizes for estimation of the odds ratio in unmatched case-control studies', *American Journal of Epidemiology*, **120**, 145–153 (1984).
4. Smith, J., Connett, J., and McHugh, R. 'Planning the size of a matched case-control study for estimation of the odds ratio', *American Journal of Epidemiology*, **122**, 345–347 (1985).