SAMPLE SIZE DETERMINATION FOR CASE-CONTROL STUDIES

by

Hal Morgenstern

Center for Health Studies
Yale University, New Haven, CT

and

Roger C. Grimson and Deborah M. Winn*

Department of Biostatistics and Department of Epidemiology*
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1290

JUNE 1980

SAMPLE SIZE DETERMINATION
FOR CASE-CONTROL STUDIES*

by

Hal Morgenstern, Ph.D.[1, 2]

Roger C. Grimson, Ph.D.[3]

Deborah M. Winn, M.S.P.H.[4]

Revised March 1979

1. Department of Epidemiology and Public Health, School of Medicine, Yale University, New Haven, CT 06510

2. Center for Health Studies, Institution for Social and Policy Studies, Yale University, New Haven, CT 06520

3. Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27514

4. Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC 27514

* This work was partly supported by three grants from: NIH (#1-T32-ES07036), NCI (#1-P01-CA-16359-05), and the Kaiser Foundation.

The authors wish to thank Jim Fletcher for programming assistance.

Address reprint requests to: Dr. Hal Morgenstern, Center for Health Studies, Yale University, 89 Trumbull Street, New Haven, CT 06520.

Running head: Sample size determination

SUMMARY

A simple method is presented for combining both statistical and practical considerations in the determination of sample size requirements for case-control studies involving binary exposure and disease variables. The basic method consists of a two-step procedure: (1) calculate the "optimal" sampling ratio of noncases to cases independently of sample size, so as to maximize the "cost efficiency" of the sampling strategy; and (2) calculate the "optimal" number of subjects for a given amount of statistical error (α and β). This basic procedure can then be repeated assuming a different set of parameter specifications, including a non-optimal sampling ratio, to trade off statistical requirements with practical constraints. Unlike other procedures for determining sample size, the proposed method allows the user to deal separately and explicitly with two practical constraints: a unit cost difference between studying cases and non-cases, and a limited number of available cases. While described for case-control studies, the method is also appropriate for any type of observational or experimental design.

Selected "optimal" solutions are presented in tabular form, and the method is illustrated with a specific example. It is found that the proposed technique for optimizing the sampling ratio is consistent with previous formulations by other authors. Typical results conform very closely to those of another, more complicated method involving an iterative solution. Finally, the issues of control variables in the analysis and matching in the selection of subjects are discussed with regard to the proposed method.

This paper will describe a simple method for combining both statistical and practical considerations in the determination of sample size requirements for case-control studies. Before presenting the proposed method, a brief review of relevant literature will be undertaken to outline the current "state of the art." It will be assumed throughout this paper that the "exposure" (or study factor) is dichotomized and that the odds ratio is used to measure the association between the exposure and disease.

The common asymptotic formula for total sample size, assuming equal size groups, is based on the statistical test for the difference between two proportions and consequently applies to both case-control and follow-up studies (Halperin, et al., 1968). The required sample size is easily obtained by specifying four parameters: the proportion of subjects in each of the compared groups which has (or develops) the outcome variable, the probability of making a type I error (α), and the probability of making a type II error (β). Schlesselman (1974) pointed out how these parameters differ for case-control studies in which the exposure is the outcome variable and for follow-up studies in which the disease is the outcome. He revised the well known formula so that sample size would be a function of the appropriate parameters in each study type. For case-control studies, the required number of subjects is found by specifying the following parameters: the proportion of the noncases that are exposed, the minimum odds ratio regarded by the investigator as important to detect, α, and β. In the situation where the disease is "rare", the proportion exposed among noncases can be estimated from the total population.

Since the maximum number of available cases is often limited, Walter (1977a) suggested that the investigator would generally prefer to know the smallest odds ratio (or risk ratio in a follow-up study) that would be statistically significant with a given sample size. Using an approximation to simplify the computation he solved Schlesselman's formula for the least significant odds ratio which can be calculated by specifying four parameters: the number of available cases (equal to the number of noncases), the proportion exposed among the noncases, α, and β.

In the same paper, Walter (1977a) also considered changing the sampling ratio of noncases to cases. A statistically optimal sampling ratio is found by minimizing the variance of the log of the odds ratio for a fixed total sample size. While this optimal ratio can differ substantially from one, Walter recommends using equal sample sizes. This advice rests on the fact that little precision (in the odds ratio estimation) is lost by using an equal number of cases and noncases, unless the minimum odds ratio worth detecting is set high. However, he later notes that unequal sample sizes may be preferred when "substantially different costs of sampling within the comparison groups are experienced" (Walter, 1977b). Alternatively, a limited number of available cases may force the investigator to increase precision or power by increasing the relative number of noncases. It has been shown, however, that incremental gains in power diminish rapidly with an increasing sampling ratio and that ratios greater than four are seldom worthwhile (Gail, et al., 1976).

Relative sampling costs were considered by Miettinen (1969) who found that the total cost of doing a matched case-control study is minimized when the ratio of noncases to cases is equal to the square root of the unit cost ratio of cases to noncases. Thus the optimization of the sampling ratio with respect to total sampling costs is found to be independent of other statistical parameters (e.g., the odds ratio, proportion of exposed noncases, α, and β). The same conclusion - known as the "square root rule" - holds for unmatched case-control and follow-up studies when the variance of the outcome variable is identical for the two compared groups (Gail, et al. 1976). A more generalized version of the "square root rule" is that the preferred sampling ratio is equal to the square root of the product of the unit cost ratio and the ratio of outcome variances - e.g., the exposure variance for cases divided by the variance for noncases (ibid). Unfortunately, in the planning of epidemiologic studies, it is difficult to specify a priori any deviation from equal variances.

Meydrech and Kupper (1978) have suggested another method for combining statistical and cost considerations in follow-up and case-control studies. An optimal number of cases and noncases is found by solving two equations simultaneously: a power (1-β) function for unequal group sizes and an expression for the total study cost as a function of the two unit costs. Since this procedure does not always permit a solution, they propose two alternative approaches. One maximizes the power for a fixed total cost, and the other minimizes the total cost for a specified power. All three procedures involve iterative solutions and therefore require computer programs. More recently, Pike and Casagrande (1979) have suggested a non-iterative method for obtaining approximately "optimal" sample sizes by extending the equal sample size solution of Schlesselman (1974), incorporating a consideration of sampling costs. They show that their method compares favorably to the results obtained by Meydrech and Kupper (1978).

Both of the above methods result in solutions of nearly equal group sizes whenever the unit cost ratio is assumed to be one. Yet, a relative shortage of cases is not necessarily reflected by a unit cost ratio greater than one. On the contrary, it is typical in the planning of case-control studies to be faced with a limited number of diagnosed cases for which the cost of obtaining exposure information is the same as for an equal number of noncases. The papers by Meydrech and Kupper (1978) and by Pike and Casagrande (1979) do not take this constraint into consideration. Furthermore, these methods do not readily allow the user to trade-off between power and cost criteria.

The proposed method involves a two-step procedure that combines both statistical and practical considerations through the quantification of "cost efficiency." In addition to computing mathematically "optimal" group sizes, the method allows for the comparison of alternative sampling strategies without the need for a computer.

METHOD

A major advantage of the case-control design is that it is well suited for studying "rare" diseases. Nevertheless, the infrequency of the disease imposes practical constraints on the study design. Specifically, the cost of studying a case will exceed the cost of studying a noncase, and/or the relative availability of cases will be limited.

The two steps of the basic method are: (1) determine the "optimal" sampling ratio independently of the sample size, and (2) determine the "optimal" number of subjects based on the previously determined ratio. The procedure can then be repeated assuming a different set of parameters, including a non-optimal sampling ratio, to trade off statistical requirements with practical constraints. It is assumed that matching is not done in the selection of subjects and that no other variables will be controlled in the analysis. We will return to these issues again in the discussion section.

Let the data be represented as in Table 1 and let the sampling ratio of noncases to cases be r. The odds ratio (OR) is then equal to ad/bc, and the number of noncases (m_0) is rm_1. Let the proportion of exposed noncases (b/m_0) be p_0 and the proportion of exposed cases (a/m_1) be p_1. We can then express the cell frequencies as functions of estimates of the other parameters: $a = m_1 p_0 OR/[p_0(OR-1) + 1]$; $b = p_0 r m_1$; $c = m_1 - m_1 p_0 OR/[p_0(OR - 1) + 1]$; and $d = rm_1(1 - p_0)$.

[Table 1]

The optimization of r is done by maximizing a measure of "cost efficiency" which is defined as the relative amount of statistical information gained toward rejecting the null hypothesis per unit of cost spent on studying all subjects. Cost efficiency (CE) is calculated by dividing the inverse of the variance of the natural log of the odds ratio (1/V(lnOR)) as a measure of precision (PC) by the relative cost (RC), in arbitrary units, of studying m_1 cases and m_0 noncases. Thus,

$$CE = PC/RC = [RC \cdot V(\ln OR)]^{-1}. \tag{1}$$

Letting the unit cost ratio of cases to noncases be C, the relative cost (RC) can be expressed as $m_0 + m_1 C = m_1(r + C)$. The Taylor series approximation for the variance of lnOR is $1/a + 1/b + 1/c + 1/d$. Substituting the above values for the cell frequencies into the expressions for PC and RC, we get the following formula for CE;

$$CE = \frac{p_0(1-p_0) r OR/(r + C)}{(1 - p_0)\ [OR + rp_0(OR - 1) + r] + p_0[OR + rp_0 OR(OR - 1) + rOR]}. \tag{2}$$

It should be noted that m_1 has dropped out of this expression, making it possible to optimize r independently of the sample size in the asymptotic case. This is done by taking the first derivative of CE with respect to r, setting it equal to zero, and solving for the optimal ratio (r^*) for which CE is maximized (i.e., has a zero slope). With some algebra, the solution reduces to,

$$r^* = \sqrt{C(OR)}\ /\ [p_0(OR - 1) + 1] \tag{3}$$

where $OR \geq 1$. It can be shown that PC has a finite upper limit as r approaches infinity. Since RC does become infinite for a fixed number of cases as r approaches infinity, CE approaches zero as r gets very large. Thus, the maximization of CE must have a unique finite solution for r^* which is greater than zero.

The second step of the method involves using the value of r^* to compute an optimal number of cases and noncases. We begin with the power function for unequal group sizes in a case-control study (Meydrech and Kupper, 1978).

$$z_{1-\beta} = \frac{z_\alpha\sqrt{(m_0 + m_1)\bar{p}(1 - \bar{p})} - \sqrt{m_0 m_1}(p_1 - p_0)}{\sqrt{m_0 p_1(1 - p_1) + m_1 p_0(1 - p_0)}}$$

where $p_1 = p_0 OR/[p_0(OR - 1) + 1]$, $\bar{p} = (p_1 + rp_0)/(r + 1)$, and z_α = the standard normal deviate associated with a given α. Noting that $z_{1-\beta} = - z_\beta$, we can solve the above equation for m_0.

$$m_0 = rm_1 = \frac{(z_\alpha\sqrt{(r + 1)\bar{p}(1 - \bar{p})} + z_\beta\sqrt{p_0(1 - p_0) + rp_1(1 - p_1)})^2}{(p_1 - p_0)^2} \qquad (4)$$

Solving equation (4) with r equal to r^* yields an optimal number of subjects in each group such that cost efficiency is maximized. Table 2 presents this optimal solution for selected conditions in which p_0, C, and OR vary, α is set at .05, and β is set at .10.

[Table 2]

It is quite possible for the "optimal" solution to require more cases than are readily available in the time allotted for subject selection. In this situation, the investigator may proceed in three ways. He may sacrifice some cost efficiency by determining the ratio (r) necessary to detect the same OR for a given number of cases. Since equation (4) cannot be solved algebraically for r as a function of m_1, it is easiest to approximate the desired sampling ratio by repeated applications of equation (4) with different values of r. This option will increase the required number of noncases and the total sample size so that the number of cases is reduced, compared to the optimal solution. Secondly, the power of the test may be sacrificed by using the available number of cases without changing the sampling ratio (r^*). This solution does not sacrifice any cost efficiency and may be solved directly by rearranging equation (4).

$$z'_\beta = \frac{(p_1 - p_0)\sqrt{r^* m'_1} - z_\alpha\sqrt{(r^* + 1)\bar{p}(1 - \bar{p})}}{\sqrt{p_0(1 - p_0) + r^* p_1(1 - p_1)}} \qquad (5)$$

where m'_1 is the number of available cases, and z'_β is the standard normal deviate associated with the larger β. Finally, the investigator may be willing to accept a larger OR worth detecting in his study. This option requires a new optimal solution since the optimal sampling ratio depends on OR, demanding that the entire procedure be repeated.

ILLUSTRATION

Consider a case-control study with the following specifications: $p_0 = .3$, OR = 2, C = 2, $\alpha = .05$, and $\beta = .10$. Using a one-sided test of the null hypothesis (OR = 1), z_α and z_β are 1.645 and 1.282, respectively. The optimal sampling ratio from equation (3) is found to be 1.54. Equation (4) is then used to compute the optimal number of subjects: 126 cases and 194 noncases. Suppose that only 100 cases are available to the investigator, making the optimal sampling scheme impractical. Cost efficiency may be sacrificed by increasing the sampling ratio as shown in Table 3. It is seen that a ratio of three

(m_1 = 101 and m_0 = 303) corresponds approximately to the number of available cases (m_1' = 100). Using equation (2) to compare cost efficiencies, this option results in a reduction in cost efficiency of 10.2 percent without any loss in power. The loss in cost efficiency is reflected by an increase in the total sample size from 320 to 404 subjects.

[Table 3]

If the optimal sampling ratio is maintained, equation (5) can be used to compute z_β' when only 100 cases are available, thereby sacrificing power. In the example, z_β' is found to be .965, corresponding to a β of .17 or a power ($1 - \beta$) of .83 (compared to the original power of .90). Since the optimal ratio (1.54) is maintained, this alternative calls for 154 noncases.

The last option is to increase the value of OR, thereby decreasing the required number of total subjects. Table 4 presents the solutions for two values of OR in addition to the original specification of OR = 2. It is found that 100 cases with an optimal ratio of 1.54 (and therefore 154 noncases) are required to detect an OR of 2.17 with the original specifications for α and β. Although this solution requires an identical sampling scheme as the second option in which power was sacrificed, this will not always be the case since the optimal sampling ratio may change when OR is changed.

[Table 4]

The appropriate choice of the above alternatives is likely to differ for various studies, depending on how much is known about the etiology of the disease and the degree of flexibility with regard to study costs and time factors. If the availability of cases is severely limited, relative to the optimal solution, it may be desirable to sacrifice some combination of cost efficiency, power, and the minimum effect regarded as clinically significant.

DISCUSSION

The proposed method may be done with a calculator[1] and combines both statistical and practical considerations. The basic two-step procedure makes it possible to compute an optimal sampling ratio and sample size separately, thereby obviating the need for an iterative solution. The method easily affords the user an opportunity to trade off design specifications and constraints, which enables him to select a sampling scheme that is suited to a particular study.

Because the odds ratio represents a comparable measure of association for both case-control and follow-up studies in the absence of bias (Fleiss, 1973), the proposed method can be adapted to sample size determination in follow-up studies involving a binary exposure variable. This is done by redefining the parameters as follows: m_1 = number of exposed (or treated) subjects; m_0 = number of unexposed (or control) subjects; $r = m_1/m_0$; p_1 = proportion of exposed who are expected to develop the disease or other outcome event during follow-up; p_0 = proportion of unexposed who are expected to develop the outcome event; and C = unit cost ratio of exposed to unexposed subjects.

Examination of equation (3) reveals that the solution for the optimal sampling ratio is consistent with previous formulations. If the ratio of exposure variances for cases versus noncases (VR) is assumed to be approximately $p_1(1 - p_1)/p_0(1 - p_0)$, which is actually correct only when r = 1, equation (3) becomes equivalent to the "generalized square root rule" of Gail, et al. (1976); that is, $\sqrt{C \cdot VR} \cong \sqrt{Cp_1(1 - p_1)/p_0(1 - p_0)} = r^*$. If the unit cost ratio (C) equals one, r^* becomes $\sqrt{OR}/[p_0(OR - 1) + 1]$ which is equivalent to the expression derived by Walter (1977a) in which cost was not considered. It is also the solution that minimizes the total number of required subjects (Gail, et al., 1976

[1] Programs for Texas Instruments (TI-58/59) and Hewlett Packard (HP-67/97) calculators have been written to perform the proposed method and are available from the authors.

Furthermore, if OR is set equal to one, r^* becomes $\sqrt{C}$, which is the solution derived by Miettinen (1969) for matched studies and by Gail, et al. (1976) for unmatched studies. This is also the solution for which the total sampling cost is minimized (ibid.). In the situation where both C and OR are equal to one, the optimal ratio is found to be one. In other words, when the unit cost of cases and noncases is equal and it is important to detect a very small effect, the optimal sampling scheme is to select an equal number of cases and noncases. This is consistent with the general recommendation of Walter (1977b), who tends to ignore the constraint imposed by having only a limited number of cases available for investigation. Moreover, when there is already adequate power (i.e., $\beta < .10$) with a limited number of available cases and $r = 1$, little power can be gained by adding noncases to the design (Gail, et al., 1976).

Since the proposed method involves two operationally separate criteria - i.e., cost efficiency in step 1 and statistical power in step 2, it is of interest to compare results for selected design situations with the results obtained by the one-step procedure of Meydrech and Kupper (1978). Table 5 presents optimal sample size requirements for both methods in addition to the equal group solution ($r = 1$) advocated by Schlesselman (1974) and Walter (1977a; 1977b). The latter sampling scheme can be determined from equation (4) by setting r equal to one. Table 5 corresponds to Table 4 in Meydrech and Kupper's paper for which the same set of conditions is analyzed and found to differ substantially from the solution assuming equal group sizes. It is observed here (in Table 5) that the optimal selection strategies of the proposed method are very close to the sample size requirements found by Meydrech and Kupper.

[Table 5]

Earlier it was stated that this method assumed no variables would be controlled in the analysis. This strict limitation was imposed because the control of other factors in the analysis of case-control studies usually results in a loss of statistical efficiency, although such control is often needed to correct for confounding effects. How this added requirement should be handled in the planning of epidemiologic studies is not clear and certainly has not received much attention in the literature. Because sampling in case-control studies is generally done from two populations, the degree of confounding in the data is difficult to predict.

Another stated assumption of this method is that individual matching is not used in the selection of noncases. Since matching requires a larger pool of eligible noncases than does independent sampling, the net effect may be to reduce the unit cost ratio (C), possibly to a value less than one. The problem arises, however, in predicting the effect on precision. If matching is done on "known" predictors of the disease, there is likely to be a gain in precision relative to an unmatched design. Unfortunately, without making additional specifications that would be extremely tenuous, the amount of gain in precision remains unknown. Therefore, it is recommended that the issue of matching be ignored in establishing sample size requirements - of course, subject to the restriction that r be an integer.

REFERENCES

Fleiss, J. L. (1973). Statistical Methods for Rates and Proportions. John Wiley & Sons: New York.

Gail, M., Williams, R., Byar, D. P., and Brown, C. (1976). How many controls? Journal of Chronic Diseases, **29**, 723-731.

Halperin, M., Rogot, E., Gurian, J., and Ederer, F. (1968). Sample size for medical trials with special reference to long-term therapy. Journal of Chronic Diseases, **21**, 13-24.

Meydrech, E. F. and Kupper, L. L. (1978). Cost considerations and sample size requirements in cohort and case-control studies. American Journal of Epidemiology, **197**, 201-205.

Miettinen, O. S. (1969). Individual matching with multiple controls in the case of all-or-none responses. Biometrics, **25**, 339-355.

Pike, M. C. and Casagrande, J. T. (1979). Re: "cost considerations and sample size requirements in cohort and case-control studies" (letter to the editor). American Journal of Epidemiology, **110**, 100-102.

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

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

Walter, S. D. (1977b). Optimal sampling ratios for prospective studies - the author replies. American Journal of Epidemiology **106**, 437-438.

TABLE 1

Data layout by disease status and exposure

	case	non-case
exposed	a	b
unexposed	c	d
total	m_1	m_0

TABLE 2[1]

Optimal sample size requirements for selected specifications in case-control studies ($\alpha = .05$, $\beta = .10$) (solutions are presented for one and two-sided tests, respectively)

			OR		
p_0	C		1.5	2.0	5.0
		r^*	1.17	1.29	1.60
.1	1	m_1^*	921/1398	273/414	35/53
		m_0^*	1074/1630	351/532	56/85
		r^*	1.65	1.82	2.26
.1	2	m_1^*	794/1206	236/359	31/47
		m_0^*	1309/1989	429/653	69/105
		r^*	2.33	2.57	3.19
.1	4	m_1^*	703/1069	210/320	27/42
		m_0^*	1641/2494	540/823	87/133
		r^*	1.11	1.18	1.24
.2	1	m_1^*	553/839	172/261	27/41
		m_0^*	615/934	203/308	33/51
		r^*	1.57	1.67	1.76
.2	2	m_1^*	475/721	148/225	23/35
		m_0^*	748/1135	247/376	41/62

Table 2 (cont.)

p_0	C		1.5	2.0	5.0
.2	4	r^*	2.23	2.36	2.48
		m_1^*	420/638	131/200	20/31
		m_0^*	935/1420	310/471	51/77
.5	1	r^*	.98	.94	.75
		m_1^*	427/647	153/231	38/57
		m_0^*	418/634	144/218	28/43
.5	2	r^*	1.39	1.33	1.05
		m_1^*	364/551	130/196	32/48
		m_0^*	504/764	173/262	33/50
.5	4	r^*	1.96	1.89	1.49
		m_1^*	319/484	113/172	27/41
		m_0^*	625/948	214/324	40/61

[1] Values of m_1^* and m_0^* are rounded off to the nearest integers.

TABLE 3[1]

Number of cases (m_1), noncases (m_0), total sample size (n) and cost efficiency (CE) by allocation ratio (r) of noncases to cases ($p_0 = .3$, $OR = 2$, $C = 2$, $z_\alpha = 1.645$, and $z_\beta = 1.282$)

	Optimal	Non-optimal	
	$r^* = 1.54$	$r = 2$	$r = 3$
m_1	126	114	101
m_0	194	228	303
n	320	342	404
CE x 10^3	39.70	39.03	35.64
% reduction in CE	-	1.7	10.2

[1] Values of m_1 and m_0 have been rounded off to the nearest integers.

TABLE 4[1]

Number of cases (m_1^*), noncases (m_0^*), and total sample size (n^*) by the minimum odds ratio (OR) regarded as important to detect ($p_0 = .3$, $C = 2$, $z_\alpha = 1.645$, $z_\beta = 1.282$)

	OR=2	OR=2.17	OR=3
	$r^* = 1.54$	$r^* = 1.54$	$r^* = 1.53$
m_1^*	126	100	49
m_0^*	194	154	75
n^*	320	254	124

[1] Values of m_1^* and m_0^* have been rounded off to the nearest integers.

TABLE 5[1]

Comparison of selected sample size results for three approaches: equal group sizes; the method of Meydrech and Kupper (1978); and the proposed method using "optimal" solutions (p_0 = .3, α = .025, β = .10, assuming one-sided tests)

		Equal Sizes		Meydrich, Kupper			Proposed method		
OR	C	m_1=m_0	n	m_1	m_0	n	m_1^*	m_0^*	n^*
2	1	188	376	188	188	376	180	196	376
3	1/7	73	146	129	50	179	125	51	176
3	1/5	73	146	119	52	171	111	54	165
3	1/3	73	146	102	56	158	95	59	154
3	1/2	73	146	88	62	150	84	64	148
3	1	73	146	72	73	145	70	76	146
3	2	73	146	61	89	150	60	92	152
3	3	73	146	56	101	157	55	104	159
3	5	73	146	51	120	171	51	123	174
3	7	73	146	48	136	184	48	138	186
4	1	45	90	45	45	90	44	47	91
5	1	34	68	34	34	68	34	34	68
7	1	24	48	24	24	48	24	23	47
10	1	18	36	18	18	36	19	16	35

[1] Values of m_1 and m_0 are rounded off to the nearest integers.