SAMPLE SIZE DETERMINATION FOR LONGITUDINAL SURVEYS

Earl Bryant,Consultant
David R. Morganstein, Westat, Inc.,1650 Research Blvd., Rockville, MD 20850

1. Overview

This paper presents the results of research on a methodology for determining sample sizes needed for longitudinal surveys involving complex sampling plans seeking to detect differences among incidence rates. A paper by Schlesselman (1974) describes a solution to this problem for simple random sample designs where the sample sizes for the two comparison groups are the same. In an elaboration of Schlesselman's work, Walter (1977) provides extensive tables of the sample sizes needed. Lakatos (1986) proposes a Markoff chain model for determining sample size under time-dependent rate of losses and noncompliance, assuming simple random sampling and equal cohort sample sizes. In this paper the authors extend these efforts in several ways. First, the requirement for equal cohort sample sizes is removed. Second, a suggestion is made of a method for incorporating the effect of a complex sample design on sampling errors, thereby extending the work beyond simple random sample designs. Last, the incidence rate for a long-term condition is modeled as an exponential process based upon an annual incidence rate and not assumed as a multiple of an annual rate.

Definition of Terms

The following terms are used in the paper:

(1) ***Relative Risk (R)*** -- the ratio of two incidence rates; namely, $R = P_1/P_2$, where

P_1 = disease incidence rate for people exposed to risk during the specified time interval, and

P_2 = disease incidence rate for people not exposed to risk during the specified time interval.

(2) ***Detectable Relative Risk*** -- a relative risk ratio that is judged to be greater than unity based on a specified level of significance and statistical power.

Another way of expressing the "detectable" relative risk is that the value of R represents the smallest relative risk greater than 1 which can be detected by samples of size n with specified levels of size and power for a specified value of P_2.

(3) ***Effective Sample Size*** -- the number of people needed to detect a significant difference between incidence rates of exposed and nonexposed study groups assuming the sample design is a simple random sample and no mortality. The actual cohort size needed must take into account losses due to mortality and the effects of a complex sampling scheme.

To determine sample sizes necessary for a longitudinal study of a risk factor's association with a particular disease or condition, the following parameters must be specified:

(1) The actual or estimated annual incidence rate of the disease or condition in the population **not exposed** to the risk factor, P_2;

(2) The proportion of people exposed to the risk factor, E;

(3) The minimum detectable relative risk deemed important, R;

(4) The length of the followup period (number of years, t);

(5) A measure of the design effect of the sampling plan (δ);

(6) The maximum acceptable probability of a Type I error, α; and

(7) The minimum acceptable statistical power, $(1-\beta)$.

Statistical models are presented for use in assessing the required sample sizes for fulfilling specified conditions for the above factors. Conversely, these models can also be used to estimate the minimum relative risks detectable with an available sample size.

Results

The summary tables (Tables 1 and 2) presented below show the minimum detectable relative risk of getting a disease as a function of statistical significance (α= .05 one-sided); power of the test, ($1-\beta$ = .90 and .80); the disease rate for people not exposed to the risk factor, P_2; the total sample size for the cohort; and the percent of the total sample exposed to the risk factor. Three illustrative sample sizes (5,000, 1,000, 200), three estimates of the proportion in the high risk group (.5, .2, .1), and five varying periods of followup have been chosen for the example. The relative risk tables were calculated for detectable differences over the selected time periods for annual incidence rates ranging from 50 per 1,000 to 5 per 10,000. The rates chosen represent a reasonable range based on observed population rates for diseases such as diabetes, cancer, and coronary heart disease (CHD).

The computations assume that the sample size represents the effective number of persons followed after losses due to mortality and other factors. To determine the actual size of the cohort needed for a study, the effective sample sizes shown in the tables must be inflated to take into consideration design effect and losses to followup due to deaths, nonresponse, etc. A discussion of the effect of a complex sampling plan is presented later.

As can be seen in the tables, even for a cohort as small as 200 reasonably small relative risk (<2.0) can be detected after followup periods of five or more years provided that the annual disease incidence rate in the exposed proportion is significantly high, $P_2 \geq 50/1{,}000$, and that the proportion exposed is equal to the nonexposed. It can also be seen that as the percentage of the cohort in the high-risk group decreases, the minimum detectable relative risk increases. With the same incidence rate, but with only 10 percent of the cohort in the high risk category, the detectable relative risk increases to 2.52 also after 5 years of followup.

With disease incidence rates on the order of 5/10,000, large cohorts and long followup periods are required to detect reasonable relative risks. For example, with an effective cohort of 5,000 in which 50 percent is at high risk, a followup period of 15 years is necessary to detect a relative

risk of 2.20. After 25 years of followup, a relative risk of 1.88 can be detected with a one-sided $\alpha = .05$ and $\beta = .10$. Given losses from mortality, loss to followup and nonresponse, an initial cohort considerably larger would have to be entered into the study to achieve the desired outcome levels. With only 10 percent at high risk, a not uncommon risk factor prevalence rate, even after 25 years of followup, the minimum detectable relative risk would be 3.05. By decreasing the power of the test, a smaller statistically significant relative risk can be detected.

The tables presented allow an investigator to examine the level of relative risk possible to detect with the available cohort, or conversely, allow estimates of the sample sizes necessary to detect desired relative risks to be made.

2. Description of Analysis

2.1 Detectable Relative Risks

The basic model described in this paper is:

$$(P_1 - P_2)^2 = (Z_\alpha + Z_\beta)^2 \left[\frac{P_1Q_1}{En'} + \frac{P_2Q_2}{(1-E)n'}\right] \qquad (1)$$

where

- P_1 = incidence rate of a disease in the portion of the cohort **exposed** to the risk factor during the specified time period,
- P_2 = incidence rate of a disease in the portion of the cohort **not exposed** to the risk factor during the specified time period,
- Q_1 = $1-P_1$,
- Q_2 = $1-P_2$,
- n' = n/δ, the effective sample size. The actual sample needed, n, is the product of the effective sample size, n', and the design effect, δ.
- E = proportion of sample exposed to the risk factor
- $1-E$ = proportion of sample not exposed to the risk factor, and
- Z_α, Z_β = the points in the standard normal distribution defined by the Type I and Type II error rates, α and β respectively.

The effective sample size can be expressed in terms of relative risk, R, the ratio of P_1 to P_2, by replacing P_1 in equation (1) with RP_2. The result is:

$$R = \frac{(2En'P_2+K) \pm [(2En'P_2+K)^2 - 4(P_2)(En'+K)(En')(P_2 - \frac{Kq_2}{(1-E)n'})]^2}{2P_2(En'+K)} \qquad (2)$$

where $K = (Z_\alpha + Z_\beta)^2$.

The values of R shown in Tables 1 and 2 are based on Equation (2). The values of R represent the smallest significant relative risks greater than 1 that can be detected by samples of size n', at the α level of significance (one-sided) with probability $(1-\beta)$ for specified values of P_2. Equation (2) can also be used to determine the largest significant relative risks less than 1, depending on whether the sum or difference of the terms in the numerator is used. For the purposes of this paper, the former approach is used so the terms in the numerator are added.

In previous papers, the cumulative incidence rates, P_1 and P_2, were given as constants. Since our interest was to examine the power of sample sizes for studies of varying lengths, we needed to model incidence rate as varying as a function of time. We used an exponential model and computed the proportion of people expected to experience a disease in time t with an annual incidence rate of P as $1-e^{-Pt}$.

The values in the tables assume that a random sample has been selected (that is a design effect of 1) and that losses to followup due to deaths, nonresponse, etc., have already been taken into account. The next section discusses the effects of such losses to followup and indicates how to incorporate parameters such as, disease-specific incidence rates and design effects.

Table 1 and 2 demonstrate how the detectable relative risks vary with P_2, n, E, and the number of years of followup. As expected, the detectable relative risk decreases with an increase in n', P_2, and in the number of years of followup. That is, a smaller difference can be detected when the incidence rate is greater, the sample size is larger, or the study period is longer.

A sample size as small as 200 appears to be adequate for detecting significant differences when $P_2 \geq .05$ with long-term followup of 10 years or more, even if the exposure rate is low ($\leq .10$). If $P_2 \geq .025$, and exposure is .10 or less, cohorts of 1,000 are sufficient for detecting reasonable levels of risk. After only 5 or more years of followup, a relative risk of 2.1 can be detected as significant. Detectable relative risks less than 2.0 may be observed if P_2 is .01, for an effective cohort size of 5,000 persons with 10 percent at high risk if the cohort is followed for 5 years or more. If $P_2 = .001$, effective cohort sizes of 5,000 do not appear to be sufficient to detect relative risks on the order of 2.0 until a minimum of at least 10 years of followup with 50 percent of the cohort at high risk. With exposure levels less than this, extremely long periods of followup are required to detect a reasonable level of risk.

2.2 Determining Effective Sample Size

The sample sizes shown in the relative risk tables are the effective sample sizes. The effective sample is the expected size of the cohort for the selected followup period, assuming a simple random sampling procedure was used to select respondents. For example, if 5,000 subjects are required to detect a desired statistically significant difference, those 5,000 people represent the effective sample size. Adjustments in this number must be made to allow for expected losses due to death, nonresponse, and other factors. The effect of such losses will be discussed next.

The Effect of a Complex Design

Equations (1) and (2) assume a simple random sample. The sample size must be modified to account for a design effect for any complex sample selection involving clustering and stratification. The term n' which appears in these equations is the ratio of the actual sample size, n, divided by the effect of the sample design δ. To determine the actual sample size needed, multiply the effective sample size by the design effect.

The estimation of the design effect may have to rely upon previous studies or, when these are not available, good judgment. While certain disease conditions are likely to be heavily effected by clustering, others may not. For most chronic diseases, it seems likely that the incidence of disease is likely to be spread randomly in the population rather than

clustered within households or between larger sampling units such as segments of PSUs. The design effect for such conditions will be close to one. Other more virulent conditions may have large design effects since their presence is likely to be widespread within contiguous areas.

Losses Due to Mortality

For studies with durations of 5 or more years, losses due to mortality must also be considered as they can result in a substantially reduced effective sample size. The authors have used a life table approach to estimate the number of person-years a cohort can be expected to contribute to a longitudinal study over a specified period. Age-race and sex-specific death rates can be used to estimate the size of a cohort which will reach the end of the study. The expected loss rate can be used in a fashion similar to that of the design effect to inflate the effective sample size, n', to the cohort size needed for participation in the study.

Continual Monitoring vs. Single Long-Term Recontacts

The effective sample size is increased with continuous monitoring. If data are collected only after long intervals of time, the incidence rates will be based upon that portion of the initial sample present at the recontact. The effective sample size is reduced by the number of people who dropped out or were lost during the study period. By having continual or short periods between contact, an individual's disease history can be more completely and readily ascertained. If we only concern ourselves with those subjects who are alive at the end of followup and determine if each has had or has not had the disease of interest, then we ignore the experience of all those who did not survive the entire period of followup. In addition, frequent contacts and measurements will provide a substantial number of person-years of observations for cases that will ultimately be lost because of moves to nonsample counties or failure to trace. Another approach would be to ascertain for all subjects who entered into the study their disease experience (for the disease of interest) during the total followup period whether or not they survived the entire period. In this way, much more information is captured for the subjects since they are included in the study for the period of time prior to their deaths. The periods during which they are disease or condition free will contribute to the total person years of followup. The time during which their status is known also contributes to the total followup. Thus, the resultant effective sample size would be larger.

The approach taken would depend on the disease of interest and the amount of data collection effort desired. Diseases that are registered, for instance, would be better suited to the latter approach, as would diseases that are likely to be reported on death certificates or be reliably available from next-of-kin interviews. Diseases that require biological measurements or laboratory tests and that are not fatal or symptomatic, on the other hand, are better suited to the former approach. The subsequent discussion assumes the latter approach.

2.3 Procedures for Using the Tables

The basic objective for conducting the cohort study is to determine if people who are exposed to some disease risk factor actually get the disease more often than those not exposed. That is, $P_1 > P_2$, where P_1 is the incidence rate for the exposed group and P_2 is the incidence rate for the nonexposed group. To determine if $P_1 > P_2$, we want to test the hypothesis H_o: $P_1=P_2$ against the alternative H_a: $P_1 > P_2$. If $P_1 = P_2$, then $R = P_1/P_2 = 1$, and the question is how large must the estimated risk ratio be for the value to be statistically greater than unity.

To test the hypothesis, one must specify the level of significance (α) and the power of the test, ($1-\beta$). The values of α and ($1-\beta$) reflected in the accompanying relative risk tables are .05 (one-sided) .90 and .80 respectively. To estimate the sample sizes required to test these hypotheses, one must determine the value of P_2, the disease or condition incidence rate for the nonexposed group.

In order to use the tables, the effective sample size, n, and the proportion of the cohort exposed to risk, E, must be specified. This effective sample size must be inflated by the design effect, the losses due to mortality and the losses due to followup to determine the cohort sample size. For example, suppose that 10 percent of the cohort is lost because the sample persons cannot be located or they refuse to participate in the study. Further, suppose that the design effect is estimated to be about 1.5 and that 10 percent of the cohort is expected to die between data collection points. To determine the number of persons needed in the initial cohort, the values of n in the table would need to be inflated by 1.5/(.9 x.9) or about 85 percent.

As an illustration of how to use the tables, suppose $P_2 = .01$, n' = 5,000, and E = 0.1 (500 at high risk and 4,500 at normal risk), the minimum detectable relative risk would be 1.50 after 10 years of followup. Based on this, one would reject the null hypothesis that $P_1 = P_2$ and accept the alternative that $R > 1$. Thus, we would conclude that there is a 90 percent chance of detecting as significant at the .05 level a relative risk of 1.5.

3. Summary

This paper has discussed several problems which must be addressed in the assessment of sample sizes needed for longitudinal studies. Previously published papers assumed that simple random sampling procedures were used and that comparison groups were of equal size. This paper suggests a method for incorporating the effect of a complex sampling plan since many large-scale longitudinal studies are based upon stratified multistage cluster designs. In addition, the results are presented for comparison groups of different sizes. Also suggested are a way to handle time-varying incidence rates and a method for incorporating estimated losses due to mortality within the cohort.

References:

Lakatos, Edward, "Sample Size Determination in Clinical Trials with Time Dependent Rates of Losses and Noncompliance," **Controlled Clinical Trials**, Vol. 7, 1986.

Schlesselman, James J., "Sample Size Requirements in Cohort and Case Control Studies of Disease," **American Journal of Epidemiology**, Vol. 99, No. 6, June 1974.

Schlesselman, James J., **Case Control Studies, Design, Conduct, Analysis**, Oxford University Press, New York, 1982.

Walter, S.D., "Determination of Significant Relative Risks and Optimal Sampling Procedures in Prospective and Retrospective Comparative Studies of Various Sizes," **American Journal of Epidemiology**, 1977, Vol. 105, No. 4.

Table 1. Minimum detectable relative risks by years of followup for specified values of P_2, sample size percent of sample exposed to risk, E; alpha = .05, (one sided) and beta = .10

Annual incidence rate for nonexposed population P_2 =	Sample size n=	Percent of sample exposed to risk factor E=	Minimum detectable relative risks, R Years of followup				
			2	5	10	15	25
0.0500	5000	0.5	1.27	1.16	1.10	1.08	1.05
		0.2	1.36	1.20	1.13	1.10	1.06
		0.1	1.50	1.28	1.17	1.13	1.08
	1000	0.5	1.64	1.37	1.23	1.17	1.11
		0.2	1.89	1.48	1.29	1.21	1.13
		0.1	2.31	1.66	1.39	1.28	1.17
	200	0.5	2.61	1.85	1.51	1.37	1.23
		0.2	3.38	2.13	1.63	1.44	1.25
		0.1	4.47	2.52	1.80	1.53	1.29
0.0250	5000	0.5	1.40	1.24	1.16	1.12	1.09
		0.2	1.54	1.31	1.20	1.16	1.11
		0.1	1.78	1.43	1.28	1.21	1.15
	1000	0.5	1.98	1.56	1.37	1.28	1.20
		0.2	2.45	1.77	1.48	1.36	1.25
		0.1	3.23	2.11	1.66	1.49	1.32
	200	0.5	3.63	2.38	1.85	1.64	1.43
		0.2	5.27	2.98	2.13	1.81	1.52
		0.1	7.54	3.83	2.52	2.05	1.64
0.0100	5000	0.5	1.67	1.40	1.27	1.21	1.16
		0.2	1.96	1.54	1.36	1.28	1.20
		0.1	2.45	1.78	1.50	1.39	1.28
	1000	0.5	2.76	1.98	1.64	1.50	1.37
		0.2	3.88	2.45	1.89	1.68	1.48
		0.1	5.74	3.23	2.31	1.97	1.66
	200	0.5	6.30	3.63	2.61	2.21	1.85
		0.2	10.68	5.27	3.38	2.71	2.13
		0.1	16.60	7.54	4.47	3.41	2.52
0.0010	5000	0.5	3.88	2.55	2.01	1.79	1.59
		0.2	6.24	3.52	2.53	2.16	1.83
		0.1	10.39	5.23	3.43	2.79	2.24
	1000	0.5	11.07	5.76	3.84	3.14	2.52
		0.2	22.65	10.26	6.05	4.62	3.42
		0.1	41.41	17.71	9.77	7.09	4.92
	200	0.5	42.23	18.46	10.43	7.70	5.44
		0.2	90.14	37.22	19.55	13.65	8.89
		0.1	151.56	61.61	31.61	21.61	13.59
0.0005	5000	0.5	5.82	3.47	2.55	2.20	1.88
		0.2	10.59	5.36	3.52	2.87	2.31
		0.1	18.86	8.68	5.23	4.04	3.05
	1000	0.5	19.62	9.33	5.76	4.50	3.43
		0.2	43.21	18.53	10.26	7.47	5.20
		0.1	80.86	33.51	17.71	12.42	8.16
	200	0.5	81.72	34.33	18.46	13.13	8.80
		0.2	178.32	72.50	37.22	25.45	16.01
		0.1	301.47	121.58	61.61	41.62	25.61

Table 2. Minimum detectable relative risks by years of followup for specified values of P_2, sample size percent of sample exposed to risk, E; alpha = 0.5, (one sided) and beta = .20

Annual incidence rate for nonexposed population P_2 =	Sample size n=	Percent of sample exposed to risk factor E=	Minimum detectable relative risks, R Years of followup				
			2	5	10	15	25
0.0500	5000	0.5	1.23	1.13	1.09	1.07	1.04
		0.2	1.30	1.17	1.11	1.08	1.05
		0.1	1.41	1.23	1.15	1.11	1.07
	1000	0.5	1.54	1.31	1.20	1.15	1.10
		0.2	1.74	1.40	1.25	1.18	1.11
		0.1	2.07	1.56	1.33	1.24	1.14
	200	0.5	2.33	1.72	1.44	1.32	1.20
		0.2	2.95	1.95	1.55	1.38	1.22
		0.1	3.87	2.30	1.70	1.47	1.26
0.0250	5000	0.5	1.33	1.20	1.13	1.11	1.08
		0.2	1.45	1.26	1.17	1.13	1.09
		0.1	1.64	1.36	1.23	1.18	1.13
	1000	0.5	1.81	1.47	1.31	1.24	1.17
		0.2	2.18	1.64	1.40	1.31	1.21
		0.1	2.78	1.91	1.56	1.41	1.28
	200	0.5	3.14	2.14	1.72	1.54	1.37
		0.2	4.43	2.64	1.95	1.69	1.45
		0.1	6.33	3.37	2.30	1.91	1.57
0.0100	5000	0.5	1.56	1.33	1.23	1.18	1.13
		0.2	1.78	1.45	1.30	1.23	1.17
		0.1	2.17	1.64	1.41	1.32	1.23
	1000	0.5	2.43	1.81	1.54	1.42	1.31
		0.2	3.27	2.18	1.74	1.56	1.40
		0.1	4.69	2.78	2.07	1.80	1.56
	200	0.5	5.19	3.14	2.33	2.01	1.72
		0.2	8.58	4.43	2.95	2.42	1.95
		0.1	13.48	6.33	3.87	3.02	2.30
0.0010	5000	0.5	3.30	2.27	1.83	1.66	1.49
		0.2	5.01	2.99	2.23	1.94	1.68
		0.1	8.04	4.25	2.91	2.42	2.00
	1000	0.5	8.67	4.73	3.27	2.73	2.24
		0.2	17.09	8.01	4.90	3.83	2.92
		0.1	31.06	13.57	7.69	5.70	4.07
	200	0.5	31.87	14.29	8.31	6.25	4.54
		0.2	68.89	28.75	15.33	10.84	7.21
		0.1	119.70	48.92	25.31	17.43	11.11
0.0005	5000	0.5	4.77	2.98	2.27	1.99	1.73
		0.2	8.19	4.36	2.99	2.49	2.06
		0.1	14.21	6.80	4.25	3.37	2.62
	1000	0.5	14.93	7.39	4.73	3.78	2.95
		0.2	32.11	14.08	8.01	5.95	4.26
		0.1	60.16	25.24	13.57	9.66	6.50
	200	0.5	61.00	26.03	14.29	10.32	7.08
		0.2	135.76	55.51	28.75	19.81	12.64
		0.1	237.65	96.11	48.92	33.18	20.58