The Exposure-Stratified Retrospective Study: Application to High-Incidence Diseases

Peng T. Liu and Debra A. Street
Division of Public Health and Biostatistics, CFSAN, FDA
5100 Paint Branch Pkwy, College Park, MD 20274

Abstract

For measuring the degree of association between a disease and a risk factor, we have developed the "exposure-stratified retrospective study" that can serve as an alternative to the conventional case-control study. When the disease rate is greater than the exposure rate, this study design is superior in terms of the required sample size and the efficiency of the estimation.

Key Words: Retrospective study, stratified sampling,, sample size, efficiency of estimate, odds ratio

1. Introduction

The case-control study or retrospective study (Cornfield,1951) is a "disease-stratified sampling". It divides the population into diseased and non-diseased strata, draws a proper number of samples from each stratum, surveys their exposure status and estimates the proportion of exposure for each group. In a rare-disease situation, the odds ratio estimated from this study is a good approximation of the relative risk.

The "retrospective" approach is a "backward" approach to survey the respondent's past disease-exposure status. In addition to disease-stratified sampling, simple random sampling and exposure-stratified sampling are also available. However, these two designs are rarely discussed in the literature due to their typically low efficiency of estimation. Hence, disease-stratified sampling has become the sole design involving a retrospective approach.

Recently, we reviewed the literature related to diabetes and obesity in order to identify possible associations with food consumption. Since we are dealing with high-incidence diseases, the "rare disease" assumption is clearly inappropriate and the odds ratio is no longer a good approximation of relative risk. Moreover, it turns out that the case-control study is no longer the most effective design under certain conditions.

The objectives of this article are to: 1) develop procedures for estimating the sample size with fixed margins, 2) compare sample sizes and relative efficiencies between two different stratified sampling designs, and 3) define the inequality of sample size and inequality of design efficiency between two stratified designs and determine the feasible regions for each design.

2. Sample size with fixed margins

The sample size for detecting a significant difference between two independent proportions is shown in the following equation (Snedecor, 1989; Fleiss, 1981; Schlesselman, 1974).

$$n = \{Z_{\alpha/2}\,[2\bar{p}\,(1-\bar{p})]^{1/2} + Z_{\beta}\,[p_1(1-p_1) + p_2(1-p_2)]^{1/2}\}^2 / (p_1 - p_2)^2$$

Where, $\bar{p} = (p_1 + p_2)/2$,

$Z_{\alpha/2}$ = standardized normal value associated with two-tailed type-I error α.

Z_{β} = standardized normal value associated with one-tailed type-II error β.

In the case-control study, $p_1 = P(E|\overline{D})$ represents the probability of exposure given a non-diseased case, and $p_2 = P(E|D)$ represents the probability of exposure given a diseased case. (Note: the terms rate, proportion and probability are used synonymously).

Usually, the exposure rate in the non-diseased group p_1 is assumed to be known or capable of being reasonably approximated, and the exposure rate in the exposed group p_2 depends on p_1 and r, where r is the minimum relative risk we desire to detect. (Schlesselman; 1974).

In most situations, the disease rate P(D) and exposure rates P(E) in population are available but their association is unknown and to be detected by the proposed study. In this article, we utilize P(D) and P(E) directly in sample size estimation, avoid some improper assumptions such as $P(D \mid \overline{E}) = P(D)$ and $P(E \mid \overline{D}) = P(E)$, and improve the accuracy of sample size estimation. The developed procedures are illustrated by the following two examples.

2.1 Example 1

Suppose the prevalence rate of diabetes (disease) is 5.0% and the prevalence rate of obesity (risk factor) is 21%. What sample size is needed to be 90% sure of detecting a relative risk $r = 2.0$ at the 5% significance level, if we propose a case-control study or an exposure-stratified study?

Solution:

The exposure-disease status in a population can be expressed by a 2 x 2 table. The two marginal distributions are (0.21, 0.79), and (0.05, 0.095) respectively. To satisfy these two margin-restrictions and to have a relative risk $r = A(C+D)/C(A+B) = 2.0$, simple algebraic operations yield the population distribution shown in Table 1.

Table 1.The obesity-diabetes distribution under the hypothesis $r = 2.0$.

	Diabetes	Non-Non-diabetes	
Obesity	A= 0.0174 *	B= 0.1926	0.2100
Non-obesity	C= 0.0326	D= 0.7574	0.7900
	0.0500	0.95000	1.0000

* If $P(D) = \delta$ and $P(E) = \varepsilon$ and $r = A(1-\varepsilon) / C\varepsilon$ (where $C = \delta - A$), then $A = r\delta\varepsilon / (1-\varepsilon + r\varepsilon)$.

2.1.1 Sample size for the case-control study

The case-control study divides the population into diabetes and non-diabetes strata. The exposure distribution for each stratum is shown in Table 2.

Table 2. The distribution of obesity in the diabetes and non-diabetes groups

	Diabetes	Non-diabetes
Obesity	0.3480	0.2070
Non-obesity	0.6520	0.7973
	1.00000	1.00000

The probability of obesity given a diabetes case, $p_1 = P(O \mid D) = 0.3480$ and the probability of obesity given a non diabetes case, $p_2 = P(O \mid \overline{D}) = 0.2027$. Substituting $p_1 = 0.3480$, $p_2 = 0.2070$, $\bar{p} = (p_1 + p_2)/2 = 0.2775$, $Z\alpha_{/2} = Z_{0.05/2} = 1.960$ and $Z_\beta = Z_{0.10} = 1.282$ into equation (1), yields n = 210 (per group).

2.1.2 Sample size for the exposure-stratified study

The exposure-stratified study divides the population into obesity and non-obesity strata.
The disease distribution for each group is shown in Table 3.

Table 3. The distribution of diabetes in the obesity and non-obesity groups.

	Diabetes	Non-diabetes	
Obesity	0.0826	0.9176	1.00000
Non-obesity	0.0413	0.9587	1.00000

The probability of diabetes given an obesity case, $p_1 = P(D \mid O) = 0.0826$ and the probability of diabetes given a non obesity case, $p_2 = P(D \mid \overline{O}) = 0.0413$. Substituting $p_1 = 0.0826$, $p_2 = 0.0413$, $\overline{p} = (p_1 + p_2)/2 = 0.0619$, $Z\alpha_{/2} = Z_{0.05/2} = 1.960$ and $Z_\beta = Z_{0.10} = 1.282$ into equation (1) yields n = 714 (per group). Thus, in this example, the case-control study actually reduces the required sample size by 70.6% ((714-210)/714). In other words, the exposure-stratified study requires 3.39 times as many samples as the case-control study.

2.2. Example 2

Suppose the prevalence rate of obesity (disease) is 21% and the prevalence rate of vegetarianism (anti-risk factor) is 3%. What sample size is needed to be 90% sure of detecting an odds ratio $o = 2.0$ at the 5% significance level if we propose a case-control study and an exposure-stratified study?

Solution:

The prevalence rate of obesity is 21% and prevalence rate of vegetarianism is 3%. The marginal distributions are (0.21, 0.79) and (0.97, 0.03), respectively. To satisfy these two margins and to have an odds ratio $o = AD/BC = 2.0$, simple algebraic operations yield the population distribution shown in Table 4.

Table 4.The population distribution under the hypothesis $o = 2.0$.

	Obesity	Non-obesity	
Non-vegetarian	A= 0.2064	B = 0.7636	0.9700
Vegetarian	C = 0.0036	D= 0.0264	0.0300
	0.2100	0.7900	1.0000

2.2.1 Sample size for the case-control study

The case-control study divides the population into obesity and non-obesity strata. The exposure distribution for each stratum is shown in Table 5:

Table 5. Sample distribution for obesity and non-obesity groups

	Obesity	Non-obesity
Non-vegetarian	0.9828	0.9665
Vegetarian	0.0172	0.0335
	1.0000	1.0000

The probability of being vegetarian given an obesity case, $p_1 = P(V \mid O) = 0.0172$ and the probability of being vegetarian given a non-obesity case, $p_2 = P(V \mid \overline{O}) = 0.0335$.

Substituting $p_1 = 0.0172$, $p_2 = 0.0335$, $\overline{p} = 0.0254$, $Z\alpha/2 = Z0.025 = 1.960$, and $Z_\beta = Z_{0.10} = 1.282$ into equation (1) yield n = 2,170 (per group).

2.2.2 Sample size for the exposure-stratified study

The exposure-stratified study divides the population into vegetarian and non-vegetarian strata.
The disease distribution for each strata is shown in Table 6.

Table 6. Obesity distribution for vegetarian and non-vegetarian groups

	Obesity	Non-obesity	
Non-vegetarian	0.2127	0.7873	1.0000
Vegetarian	0.1200	0.8800	1.0000

The probability of obesity given a vegetarian, $p_1 = P(O \mid V) = 0.1200$ and the probability of obesity given a non-vegetarian, $p_2 = P(O \mid \overline{V}) = 0.2127$. Substituting $p_1 = 0.1200$, $p_2 = 0.2127$, $\overline{p} = 0.1664$, $Z_{\alpha/2} = Z_{0.025} = 1.960$, and $Z_\beta = Z_{0.10} = 1.282$ into equation (1) yield n = 337 (per group).

Thus, in this worked example, the exposure-stratified study could reduce the required samples size by as much as 84.4% ((2170-337)/2,170).

3. Inequality of sample sizes between two stratified studies

The above two examples are sufficient to illustrate the sample size properties for detecting the desired level of relative risk or odds ratio subject to the constraints of the arbitrarily chosen Type I and Type II error rates. Example 1 shows a situation where the case-control study requires a smaller sample size than the exposure-stratified study. By contrast, example 2 shows the situation where the case-control study requires a greater sample size.

When the disease rate is equal to the exposure rate, the population distribution is characterized by a symmetric matrix. The exposure rate in the diseased group and in the non-diseased group derived from the disease-stratified study are the same as the corresponding disease rate in the exposed group and in the non-exposed group, respectively. Hence, the sample size for the two stratified designs becomes identical. In general, the "inequality of sample size between two stratified designs" may be stated as follows:

If the disease rate is less than the exposure rate, then the sample size for the case-control study is less the exposure-stratified study.
If the disease rate is equal to the exposure rate, then the sample size for the case-control study is equal to the sample size for the exposure-stratified study.
If the disease rate is greater than the exposure rate, then the sample size for the case-control study is greater than the sample size for the exposure stratified study.

These inequalities could be used as a guideline for design selection. For example, suppose we want to study the degree of association between diabetes (with 5% prevalence rate) and a vegan diet (with 3% prevalence rate). As seen from our guideline, the exposure-stratified study requires a smaller sample size than the case-control study.

4. Inequality of efficiencies between two stratified studies

The Pitman relative efficiency of estimation (Zacks, 1985) is expressed by the reciprocal ratio of two corresponding standard deviations. The odds ratios estimated from the three types of retrospective studies are asymptotically unbiased but with different efficiencies (i.e., variances). The relative efficiency of the odds ratio estimate can be simplified as the ratio of two corresponding k values. (where $k = 1/a + 1/b + 1/c + 1/d$, and a, b, c, and d reflect the sample distribution and are the entries in the matrix).

In example 1, the relative efficiency of the odds ratio estimate as determined from the exposure stratified design compared to that for the case-control study is $R.E._{(ESS:CCS)} = \{(1/0.0826 + 1/0.9176 + 1/0.0413 + 1/0.9587)/ (1/0.3480 + 1/0.2070 + 1/0.6520 + 0.7973)\}^{½} = 0.54$ (54%). In example 2, the corresponding result is $R.E._{(ESS:CCS)} = \{(1/0.2127 +1/0.7873+0.1210+1/0.8800)/ (1/0.9828 + 1/0.0172 + 1/0.9665 + 1/0.0355)\}^{½} = 6.06$ (606%).

Analogously to the results for the sample size, the "inequality of two design efficiencies" can be stated as follows:

If the exposure rate is less than the disease rate, then the efficiency of the odds ratio estimate from the case-control study is less than the efficiency of the odds ratio estimate from the exposure-stratified study.

If the exposure rate is equal to the disease rate, then the efficiency of the odds ratio estimate from the case-control study is equal to the efficiency of the odds ratio estimate from the case-control study.

If the exposure rate is greater than the disease rate, then the efficiency of the odds ratio estimate from the case-control study control is greater than the efficiency of the odds ratio estimate from the exposure-stratified study.

The "inequality of sample size" statements and the "inequality of design efficiencies" statements for the two designs are parallel and consistent with one another.

5. Discussion

The procedures developed here –which may be called "reverse weighting procedures" - call for i) decomposing either the disease rate in the population into disease rate in the exposed group and disease rate in the non-exposed group; or for ii)decomposing the exposure rate in the population into the exposure rate in the diseased group and the exposure rate in the non-diseased group, depending on the proposed design. These procedures could avoid certain improper assumptions, and improve the accuracy of estimation.

This approach can also evaluate the functional relation between the odds ratio and relative risk, and examine the degree of assumption violation. In example 1, the sample size for detecting relative risk $r = 2.0$ is equivalent to detecting the odds ratio $o = 2.1$. The bias is 5% ((229-210/210) x100% x). In example 2, the sample size for detecting the odds ratio $o = 2.0$ is equivalent to detecting a relative risk $r = 1.75$. The bias is 14.5% ((2.0-1.75)/1.75) x 100%).

The sample size for detecting a desired relative risk in the case control study has been discussed by Schlesselman in 1974. His sample size table developed in the same year has been widely used. Since the "relative risk" cannot be estimated from a case-control study, some approximations should be adopted in his estimation. Therefore, his table turns out to be suited only for detecting a desired odds ratio instead of a relative risk. The discrepancy between the sample size for detecting the same value of odds ratio or relative risk could be greater than 20 % under certain conditions. Based on the "inequality of odds ratio and relative risk", when the odds ratio is greater than 1, the odds ratio is greater than the relative risk (Liu and Street; 2004). To treat a relative risk as an odds ratio could substantially underestimate the true required sample size.

References:

Cornfield J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of lung, breast, and cervix. J. Natl. Cancer Inst. Vol. 11: 1269-1275.

Fleiss , J.L. et al. (2003) *Statistical Methods for Rates and Proportions*, John Wiley & Sons, Inc. New York.

Liu, P. T. and D. A Street (2004) The Relative Risk and Odds Ratio Inequalities and Conversions. The 2004 Proceeding of Joint Statistical Meeting, Statistics in Epidemiology Section, American Statistical Association.

Schlesselman, J. J. (1974). Sample Size Requirements in Cohort and Case-Control Studies of Disease", Am. J. of Epidemiology, 99:381.

Schlesselman, J. J. (1974). Tables of the Sample Size Requirement for Cohort and Case-Control Studies of Disease, NICHD, NIH, Bethesda, Md.

Snedecor, G.W. and W.G.Cochran. (1989). *Statistical Methods*. (8^{th} edition), Iowa State Univ. Press . Ames, IA.

Zacks, S. (1985). Pitman efficiency in Encyclopedia of Statistical Sciences , Vol. 6, S. Kotz & N.L. Johnson. Eds. Wiley, NewYork. pp: 731-735.

.