

Symmetric association measures in effect-control sampling

Riccardo Borgoni, Donata Marasini and Piero Quatto

Abstract In order to measure the association between an exposure variable X and an outcome variable Y , we introduce the effect-control sampling design and we consider the family of symmetric association measures. Focusing on the case of binary exposure and outcome variables, a general estimator of such measures is proposed and its asymptotic properties are also discussed. We define an allocation procedure for a stratified effect-control design, which is optimal in terms of the variance of such an estimator. Finally, small sample behaviour is investigated by Monte Carlo simulation for a measure belonging to the family, which we believe particularly interesting as it possess the appealing property of being normalized.

Keywords: Sample design and estimation, association measures, optimal sampling

1 Introduction

In survey sampling theory, the control sample has been introduced to measure the association between two dichotomous variables, denoted by X and Y (Borgoni et al., 2009). In this context, X may be the indicator variable for the exposure to a possible cause of an effect, represented by an outcome variable Y . The aim is to measure the strength of the assumed causal relation.

¹ Riccardo Borgoni, Department of Statistics, University of Milano-Bicocca;
riccardo.borgoni@unimib.it

² Donata Marasini, Department of Statistics, University of Milano-Bicocca;
donata.marasini@unimib.it

³ Piero Quatto, Department of Statistics, University of Milano-Bicocca;
piero.quatto@unimib.it

To exemplify, suppose that one wants to identify potential causes of students drop out of the University. If one restricts his attention only on a sample of those who dropped out, it may happen that a high percentage of students interviewed says that they withdrew because the University studies were very difficult. This suggests the study difficulty as a cause of the drop out. However, if a control sample of students having continued their studies is considered together with the drop out sample, one may find a similar proportion of students believing that the university studies are very difficult among the controls as well. In this case, the study difficulty can be hardly considered a cause of drop out. Clearly, many other variables can have an impact on individuals' decision to withdraw such as being a student-worker, the school of provenience and so on. For these variables, i.e. the potential causes, we need to calculate the association with the effect under study. As there may be other variables that affect the association between the effect and its potential cause, one may want to control for their influence in order to prevent confounding. In this sense, such variables can be considered as auxiliary variables and can be used to divide the population into suitable strata (e.g. gender and age).

In this paper we introduce a family of association measure which can be used in effect-control sampling. Effect-control designs are introduced in Section 2. The proposed family of association measures is described in Section 3 along with a class of estimators. In Section 4 we focus on the estimation procedure of a normalized measure of the family which finite sample behaviour is evaluated by Monte Carlo simulation in Section 5.

2 Effect-control sampling

Let Y be the outcome variable, i.e. Y is the dichotomous variable taking value 1 if the effect is present and 0 otherwise. The target population U is assumed to be finite and it is divided into two sub-populations according to Y . Let N and M be the number of units for which $Y = 1$ and $Y = 0$, respectively. The N units for which Y equals 1, identify the effect group, whereas the M for which Y equals 0 identify the control group.

The two groups are partitioned in H strata according to the values of a set of auxiliary variables identifying homogeneous profiles. Let N_h and M_h be the size of the stratum h ($h = 1, 2, \dots, H$) in the two groups with $N = \sum_{h=1}^H N_h$ and $M = \sum_{h=1}^H M_h$.

The effect and control samples, say e and c , are drawn in two steps. First a sample, say e_h , is taken within the units in stratum h of the effect group; then a sample c_h of size m_h is taken from stratum h of the control group, in order to guarantee that effect and control units are similar in terms of the stratification variables. Clearly, we obtain that

$e = \bigcup_{h=1}^H e_h$ and $c = \bigcup_{h=1}^H c_h$ with size $n = \sum_{h=1}^H n_h$ and $m = \sum_{h=1}^H m_h$, respectively. The

design allocates a positive probability to the samples of size $n_h + m_h$ that can be obtained from the $N_h + M_h$ units in stratum h , assuming a random sampling without replacement. Every $n_h + m_h$ -tuple is supposed to have the same probability

$1 / \binom{N_h}{n_h} \binom{M_h}{m_h}$ $h=1, \dots, H$ to occur. It can be noted that the effect-control sampling

design may be considered as a particular stratified sampling design in which the population is partitioned into $2H$ strata according to the H values of a set of auxiliary variables and to the two values of the dichotomous outcome variable.

3 Symmetric association measures

As mentioned above, let X be a dichotomous exposure variable and let Y be a dichotomous outcome variable possibly associated with X . We consider the population U consisting of $N + M$ units of which N are characterized by $Y = 1$ and M are characterized by $Y = 0$.

Assume that N' units out of N and M' units out of M are exposed ($X = 1$). Thus, the odds ratio (OR) parameter is defined by

$$OR = \frac{N'}{N - N'} \frac{M - M'}{M'} . \quad (1)$$

Unlike other measures of association for binary data (such as the relative risk), the odds ratio treats the two variables X and Y symmetrically (Edwards, 1963).

In particular, an odds ratio of 1 indicates that X and Y are independent, an odds ratio greater than 1 indicates positive association and an odds ratio less than 1 indicates negative association.

With respect to the stratification, the stratum specific version of (1) is given by

$$\theta_h = \frac{N'_h}{N_h - N'_h} \frac{M_h - M'_h}{M'_h} , \quad (2)$$

assuming that N'_h units out of N_h and M'_h units out of M_h are exposed in the stratum $h = 1, 2, \dots, H$.

According to Edwards (1963), we consider the symmetric association measures defined by $f(OR)$, where f is a strictly increasing function.

So, the well known homogeneity hypothesis

$$\theta_1 = \dots = \theta_h = \theta \quad (3)$$

(Jewell, 2004; Lachin, 2000) becomes

$$f(\theta_1) = \dots = f(\theta_h) = f(\theta) .$$

Under such a hypothesis, the parameter $f(\theta)$ can be estimated by a suitable weighted mean

$$\sum_{h=1}^H w_h f(\hat{\theta}_h) , \quad (4)$$

where

$$\hat{\theta}_h = \frac{n'_h}{n_h - n'_h} \frac{m_h - m'_h}{m'_h}$$

is the usual estimate of (2), assuming that in the sample of n_h units characterized by $S = h$ and $Y = 1$, n'_h are exposed, and in the sample of m_h units characterized by $S = h$ and $Y = 0$, m'_h are exposed.

It can be proved, by applying the delta method (Lehmann, 2001), that the estimator corresponding to (4) is approximately normal with mean $f(\theta)$ and variance given by

$$\sum_{h=1}^H [w_h f'(\theta_h) \theta_h]^2 \left[\frac{M_h^2}{m_h M'_h (M_h - M'_h)} + \frac{N_h^2}{n_h N'_h (N_h - N'_h)} \right], \quad (5)$$

where f' denotes the derivative of f .

Then, from Slutsky's theorem (Lehmann, 2001), it follows that (5) can be consistently estimated by

$$\sum_{h=1}^H [w_h f'(\hat{\theta}_h) \hat{\theta}_h]^2 \left(\frac{1}{m'_h} + \frac{1}{m_h - m'_h} + \frac{1}{n'_h} + \frac{1}{n_h - n'_h} \right). \quad (6)$$

Moreover, optimal weights can be obtained by minimizing (5) under the constraint

$$\sum_{h=1}^H w_h = 1.$$

For this purpose, the method of Lagrange multipliers provides the solution of the optimization problem

$$w_h = \frac{\frac{1}{\sigma_h^2}}{\sum_{h=1}^H \frac{1}{\sigma_h^2}}, \quad (7)$$

where

$$\sigma_h^2 = [f'(\theta_h) \theta_h]^2 \left[\frac{M_h^2}{m_h M'_h (M_h - M'_h)} + \frac{N_h^2}{n_h N'_h (N_h - N'_h)} \right] \quad (8)$$

($h = 1, 2, \dots, H$).

It should be noted that, under hypothesis (3), the optimal weights (7) turn out to be independent of the choice of transformation f .

Hence, the optimal weights (7) can be approximated by

$$\hat{w}_h = \frac{\frac{1}{\hat{\sigma}_h^2}}{\sum_{k=1}^H \frac{1}{\hat{\sigma}_k^2}}, \quad (9)$$

where the consistent estimate

$$\hat{\sigma}_h^2 = \frac{1}{m'_h} + \frac{1}{m_h - m'_h} + \frac{1}{n'_h} + \frac{1}{n_h - n'_h} \quad (10)$$

corresponds to the variance stabilizing transformation

$$f(OR) = \log(OR).$$

In addition, we observe that such a choice of f leads to the classical estimate suggested by Woolf (1955)

$$\sum_{h=1}^H \hat{w}_h \log(\hat{\theta}_h).$$

Finally, optimal allocation can be obtained by finding the minimum of (5) subject to the constraint

$$\sum_{h=1}^H (n_h + m_h) = t,$$

where t represents the total sample size.

Under homogeneity hypothesis (3), the method of Lagrange multipliers provides the optimal sample sizes which turn out to be independent of f .

$$\left\{ \begin{array}{l} n_h = \frac{w_h/\eta_h}{\sum_{h=1}^H w_h \left(\frac{1}{\eta_h} + \frac{1}{\gamma_h} \right)} t \\ m_h = \frac{w_h/\gamma_h}{\sum_{h=1}^H w_h \left(\frac{1}{\eta_h} + \frac{1}{\gamma_h} \right)} t \end{array} \right., \quad (11)$$

where

$$\left\{ \begin{array}{l} \gamma_h = \sqrt{\frac{M'_h(M_h - M'_h)}{M_h^2}} \\ \eta_h = \sqrt{\frac{N'_h(N_h - N'_h)}{N_h^2}} \end{array} \right., \quad (12)$$

($h = 1, 2, \dots, H$).

In order to realize the optimal allocation (11), it is necessary to conduct a pilot survey for estimating the unknown parameters (12).

4 A normalized symmetric measure of association

In this section we consider the normalized OR suggested by Landenna (1994),

$$f(OR) = \frac{OR}{1 + OR} = \frac{(M - M')N'}{M'(N - N') + (M - M')N'} \in [0, 1] \quad (13)$$

In particular, the value taken on by (13) is equal to 0.5 when the variables X and Y are independent, is greater than 0.5 in presence of positive association and is less than 0.5 in presence of negative association.

Putting

$$\lambda_h = f(\theta_h) = \frac{\theta_h}{1 + \theta_h} = \frac{(M_h - M'_h)N'_h}{M'_h(N_h - N'_h) + (M_h - M'_h)N'_h}$$

($h = 1, 2, \dots, H$), the homogeneity hypothesis (3) becomes

$$\lambda_1 = \dots = \lambda_h = \lambda.$$

So, the estimate (4), the variance (5) and its consistent estimate (6) turn out to be, respectively,

$$\sum_{h=1}^H w_h \hat{\lambda}_h, \quad (14)$$

$$\sum_{h=1}^H [w_h \lambda_h (1 - \lambda_h)]^2 \left[\frac{M_h^2}{m_h M'_h (M_h - M'_h)} + \frac{N_h^2}{n_h N'_h (N_h - N'_h)} \right] \quad (15)$$

and

$$\sum_{h=1}^H [w_h \hat{\lambda}_h (1 - \hat{\lambda}_h)]^2 \left(\frac{1}{m'_h} + \frac{1}{m_h - m'_h} + \frac{1}{n'_h} + \frac{1}{n_h - n'_h} \right),$$

where

$$\hat{\lambda}_h = \frac{(m_h - m'_h)n'_h}{m'_h(n_h - n'_h) + (m_h - m'_h)n'_h}$$

($h = 1, 2, \dots, H$).

Under homogeneity hypothesis (3), the parameter λ can be estimated by means of (14), in which the optimal weights specified by (7) and (8) can be approximated using (9) and (10), and the sample sizes are provided by the proportional allocation

$$\begin{cases} n_h = \frac{N_h}{N + M} t \\ m_h = \frac{M_h}{N + M} t \end{cases} \quad (16)$$

($h = 1, 2, \dots, H$; t being the total sample size).

Finally, the variance (15) could further be reduced by the optimal allocation (11), in which the parameters (12) may be estimated on the basis of the preceding sample survey.

5 A simulation study

In order to evaluate the small sample behaviour of the estimate (14) with the optimal coefficients (7), an extensive simulation study have been performed. Results are reported in Table 1.

The simulation design is as follows. A finite population consisting of 4000 units stratified in two groups has been considered, assuming a grid of predetermined values of λ . The population was sampled without replacement 1000 times for each value of λ using both the proportional and the optimal allocation for the strata size. Calculating the optimal allocation (11) requires a preliminary estimate of the parameters (12), a problem that can be addressed by a pilot survey in real situations. In order to implement the optimal design in the simulation study presented here, we firstly extracted a pilot sample from the simulated population. This sample was used to estimate the unknown parameters (12) which, in turn, permitted to calculate the optimal sample size of each stratum. Then a second sample was selected without replacement from the population, which is finally employed to estimate λ .

The simulation has been repeated for different sampling fractions (p) in order to evaluate whether the performance of the estimator is influenced by the overall sample size and the velocity at which its distribution converges to normality. The Monte Carlo (MC) bias has been reported in Table 1 for the two schemes together with the ratio between the MC MSE obtained by the optimal allocation (B) and the MC MSE

obtained by the proportional allocation (A). Hence, a value smaller than 1 of this ratio suggests a better performance of the optimal allocation.

It can be observed that the MC MSE in the case of the optimal allocation is always smaller than the MC MSE obtained using a proportional allocation and quite often much smaller. Broadly speaking, a bigger advantage was found when the sampling fraction is small for a value of λ bigger than 0.5. For λ smaller than 0.5 the optimal allocation produces, roughly, a constant reduction of the MSE in the case of the proportional allocation whatever the sampling fraction is.

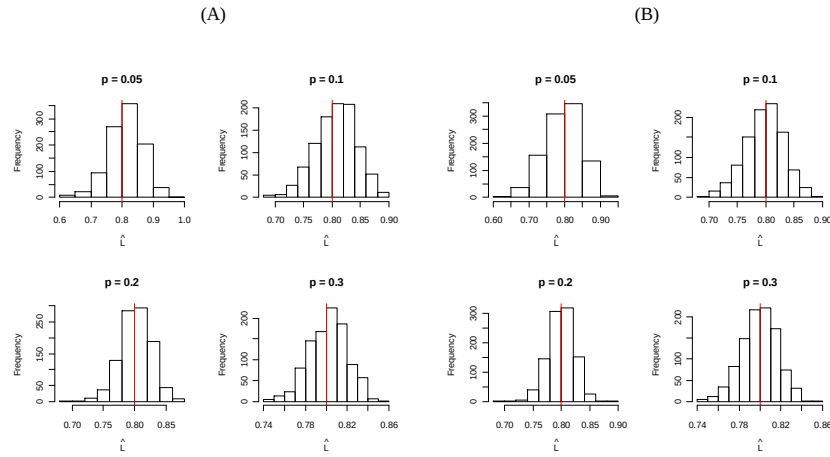
As far as the bias is concerned results are less clear as both designs tend to have a good performance. This was somehow expected as the estimator is asymptotically unbiased under both allocation schemes. However, the optimal allocation has a bias which is generally greater than the proportional allocation, suggesting that this scheme seems preferable also in this respect.

Table 1: Results of the Monte Carlo study in the case A of proportional allocation and in the case B of optimal allocation

OR	λ	p	A bias	B bias	B mse / A mse
4	0.8	0.05	0.011	-0.005	0.87
		0.1	0.007	-0.003	0.90
		0.2	0.002	-0.001	0.83
		0.3	0.002	-0.001	0.99
2	0.66	0.05	0.007	-0.002	0.76
		0.1	0.002	0.000	0.80
		0.2	-0.001	0.000	0.83
		0.3	0.001	-0.001	0.86
1	0.5	0.05	-0.004	-0.002	0.67
		0.1	0.001	-0.002	0.77
		0.2	0.001	0.001	0.75
		0.3	0.000	0.001	0.73
0.5	0.33	0.05	-0.008	0.002	0.69
		0.1	-0.007	0.000	0.71
		0.2	-0.003	0.002	0.77
		0.3	-0.001	0.001	0.62
0.25	0.2	0.05	-0.014	0.006	0.80
		0.1	-0.008	0.003	0.71
		0.2	-0.003	0.001	0.71
		0.3	-0.001	0.001	0.75

Finally, Figure 1 shows that the distribution of the considered estimator approaches normality quickly under both designs as it resembles the Gaussian density, also in the presence of small samples. The graphs are drawn using samples simulated under a values of λ equal to 0.8. Similar results were found for all the other values of λ considered in the simulation.

Figure 1: Monte Carlo sampling distribution of $\hat{\lambda}$ under proportional (A) and optimal (B) sample size allocation for different sampling fractions. Vertical line represents the “true” value of λ used to simulate samples



References

1. Borgoni, R., Marasini, D., Quatto, P.: Effect-control sampling and odds ratio, ITACOSM 2009, Siena (2009)
2. Edwards, A.W.F.: The measure of association in 2×2 table. JRSS. Series A, **126**, 1, 109-114 (1963)
3. Jewell, N.P.: Statistics for Epidemiology, Chapman&Hall/CRC (2004)
4. Lachin, J.M.: Biostatistical Methods: The Assessment of Relative Risks, John Wiley and Sons (2000)
5. Landenna, G.: Fondamenti di statistica descrittiva, il Mulino, Bologna (1994)
6. Lehmann E.L.: Elements of Large-Sample Theory, Springer, New York (2001)
7. Woolf, B.: On estimating the relation between blood group and disease. Annals of Human Genetics, **19**, 251-253 (1955)