

Power Analysis for Chi-Squared Tests Adjusted for Misclassification Error

Sunyeong Heo

*Korean Educational Development Institute, 92-6, Umyun-Dong, Seocho-Gu
Seoul, 137-791, Korea
syheo@ns.kedi.re.kr*

John L. Eltinge

*U.S. Bureau of Labor Statistics
PSB 4915, 2 Massachusetts Avenue NE
Washington, DC 20212 U.S.A.
Eltinge_J@bls.gov*

1. Introduction: Chi-Squared Tests Under Misclassification Error

For data collected through a complex sample design, standard Pearson multinomial-distribution-based chi-squared tests generally do not achieve their nominal levels of type I error. Consequently, Scott and Rao (1981) considered adjustments of chi-squared test statistics to give asymptotically valid tests of homogeneity based on data from a complex sample design. See, e.g., Rao and Scott (1981, 1984, 1987), Fay (1985), Thomas and Rao (1987), Graubard and Korn (1993) and Thomas, Singh and Roberts (1996) for additional discussion of the effects of complex survey designs on quadratic-form test statistics and modifications thereof.

Another problem with the standard Pearson-type chi-squared test is that when misclassification errors exist, they can inflate the true type I error rate of the test, regardless of the sample design employed. For some general background on this issue, see, e.g., Mote and Anderson (1965), Tenenbein (1972), Hochberg (1977), Hochberg and Tenenbein (1983), Selen (1986) and references cited therein.

Rao and Thomas (1991) discussed methods to adjust chi-squared test statistics for goodness of fit with complex survey data subject to misclassification errors. When misclassification errors exist, the observed cell proportions are adjusted by a misclassification probability matrix, A_i , say, for the i -th population; and one must adjust variance estimators and test statistics accordingly. The present paper investigates chi-squared test statistics for homogeneity with complex survey data subject to misclassification errors. Principal emphasis is placed on trade-offs between test bias and power that arise as one considers limitations on the amount of information available on the underlying misclassification probabilities.

2. Quadratic Form Test Statistics

Suppose that we have two distinct populations labeled $i = 1, 2$, and suppose that we select samples of sizes n_1 and n_2 from these populations. In addition, suppose that each observation in the i -th sample is classified into one of J mutually exclusive and exhaustive classes. For each population i , let $\pi_{i+} = (\pi_{i1}, \dots, \pi_{iJ})'$ and $p_{i+} = (p_{i1}, \dots, p_{iJ})'$ be the vectors of the J true proportions and expected observed proportions, where the expectation is evaluated with respect to the distribution of misclassification errors. For example, in the application considered in the detailed version of this paper, the populations i are sets of persons or households within a state i in the United States, the

classes j are determined by the health or behavioral characteristics of these persons or households, π_{i+} is the corresponding vector of true prevalence rates within population i , and p_{i+} is the expected value of the corresponding vector of prevalence rates that would be reported if the fallible instrument were applied to all units in the population. Then a test of homogeneity of the true proportions will involve the null hypothesis $H_0 : \pi_1 = \pi_2 = \pi_0$ and the general alternative hypothesis $H_0 : \pi_1 \neq \pi_2$, where π_i is a vector containing the first $(J-1)$ elements of π_{i+} , $i = 0,1,2$; and π_0 is an unknown vector.

Now define Z to be an observed class, Y to be a true class, and $P_i(Z = k | Y = j)$ to be the probability that a unit reports membership in a class k conditional upon $Y = j$ for the i -th population. When the observed proportions are subject to misclassification error, customary design based estimators of the proportions of reported classifications will converge to

$$p_{i+} = A_i' \pi_{i+} \quad (1)$$

where $A_i = a_{i,jk}$ is a $J \times J$ dimensional matrix with (j,k) -th element $a_{i,jk} = P_i(Z = k | Y = j)$. Following Rao and Thomas (1991), consider the case in which one has consistent point estimators \hat{p}_{i+} and \hat{A}_i of p_{i+} and A_i , respectively. Then a simple estimator of π_i is

$$\hat{\pi}_{i+} = (\hat{A}_i')^{-1} \hat{p}_{i+} \quad (2)$$

where we assume that \hat{A}_i is invertible. Following, e.g., Scott and Rao (1981), we may consider testing the null hypothesis with a quadratic test statistic of the form,

$$(\hat{\pi}_1 - \hat{\pi}_2)' \hat{M} (\hat{\pi}_1 - \hat{\pi}_2) \quad (3)$$

where one could consider several possible symmetric matrices \hat{M} . For example, under appropriate conditions, including the approximate independence of $\hat{\pi}_1$ and $\hat{\pi}_2$, one could use the test statistic

$$(\hat{\pi}_1 - \hat{\pi}_2)' (n_1^{-1} \hat{V}_1 + n_2^{-1} \hat{V}_2)^{-1} (\hat{\pi}_1 - \hat{\pi}_2) \quad (4)$$

where $n_i^{-1} \hat{V}_i$ is an estimator of the variance of the approximate distribution of $\hat{\pi}_i$.

3. Assessment of the Power of Adjusted Tests

Practical applications of the test statistics (3) and (4) involve two important issues. First, in many cases direct design-based variance estimators $n_i^{-1} \hat{V}_i$ are not available to a secondary data analyst, or are relatively unstable. For such cases, following Rao and Scott (1981, 1984, 1987), one can develop first- and second-order Rao-Scott type versions of the general test statistic (3). Second, consider the performance of a test based on the misclassification-adjusted difference $\hat{\pi}_1 - \hat{\pi}_2$, relative to the performance of a similar test based on the unadjusted difference $\hat{p}_1 - \hat{p}_2$. If the misclassification probability matrices A_1 and A_2 are not equal, then the latter test generally will be biased. In general, this would lead one to prefer test statistics like (4) that depend on the misclassification-adjusted differences $\hat{\pi}_1 - \hat{\pi}_2$. However, in some cases the matrix of estimated

misclassification probabilities \hat{A}_i may be based on a subsample of moderate size, so that the sampling error $\hat{A}_i - A_i$ makes a nontrivial contribution to the overall variability of the misclassification-adjusted point estimator $\hat{\pi}_i$. Consequently, even if one adjusts the variance estimator $n_i^{-1}\hat{V}_i$ to account for this additional source of variability, the error $\hat{A}_i - A_i$ may substantially degrade the power of a test based on (4). As one considers a sequence of cases in which the sampling errors $\hat{A}_i - A_i$ become progressively larger and the differences $A_1 - A_2$ become progressively smaller, each relative to the sampling error of $\hat{p}_i - p_i$, one eventually reaches the point at which the adjusted test will have operating characteristics that are inferior to those of the corresponding unadjusted test based on $\hat{p}_1 - \hat{p}_2$. In formal terms, the relative degradation of these two tests can be evaluated through display of estimated forms of their respective power curves. The detailed version of this paper develops relatively simple methods for estimation of these power curves from available data and applies the proposed methods to data from a health survey.

Acknowledgements

The authors thank Drs. P. P. Biemer of the Research Triangle Institute and V. L. Parsons of the U.S. National Center for Health Statistics for providing the data from the Dual Frame NHIS/RDD Methodology and Field Test (Biemer, 1997) which motivated this work and is considered further in the detailed version of this paper. This paper is based on an unpublished Ph.D. dissertation (Heo, 1999) written by the first author under the direction of the second author in the Department of Statistics at Texas A&M University. This work was supported in part by the U.S. National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the U.S. National Center for Health Statistics, the U.S. Bureau of Labor Statistics or the Korean Educational Development Institute.

References

- Biemer, P.P. (1997). Dual frame NHIS/RDD methodology and field test. Analysis report prepared for the U.S. National Center for Health Statistics. Research Triangle Park, North Carolina: Research Triangle Institute.
- Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *J. Amer. Statist. Assoc.* **80**, 148-157.
- Graubard, B.I. and Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *J. Amer. Statist. Assoc.* **88**, 629-641.
- Heo, S. and Eltinge, J.L. (1999). The analysis of categorical data from a complex sample survey: Chi-squared tests for homogeneity subject to misclassification error. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 859-864.
- Heo, S. (1999). *Diagnostics for Survey Inference Accounting for Incomplete Data and Measurement Error*. Unpublished Ph.D. dissertation, Department of Statistics, Texas A&M University, College Station, Texas.
- Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *J. Amer. Statist. Assoc.* **72**, 914-921.

- Hochberg, Y. and Tenenbein, A. (1983). On triple sampling schemes for estimating from binomial data with misclassification errors. *Comm. Statist. A* **12**, 1523-1533.
- Mote, V.L. and Anderson, R.L. (1965). An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data. *Biometrika* **52**, 95-109.
- Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness-of-fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76**, 221-230.
- Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Ann. Statist.* **12**, 46-60.
- Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Ann. Statist.* **15**, 385-397.
- Rao, J.N.K. and Thomas, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, pp. 637-663. New York: Wiley.
- Scott, A.J. and Rao, J.N.K. (1981). Chi-squared tests for contingency with proportions estimated from survey data. In D. Krewski, R. Platek and J.N.K. Rao (eds.), *Current Topics in Survey Sampling*, pp. 247-265. New York: Academic Press.
- Selen, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *J. Amer. Statist. Assoc.* **81**, 75-81.
- Tenebein, A. (1972). A double sampling scheme for estimation from misclassified multinomial data with application to sampling inspection. *Technometrics* **14**, 187-202.
- Thomas, D.R. and Rao, J.N.K. (1987). Small-sample comparisons of level and power of simple goodness-of-fit statistics under cluster sampling. *J. Amer. Statist. Assoc.* **82**, 630-636.
- Thomas, D.R., Singh, A.C. and Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *Int. Statist. Rev.* **64**, 295-311.