The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 853-864

Chi-squared Tests for Homogeneity based on Complex Sample Survey Data Subject to Misclassification Error

Sunyeong Heo[1)]

Abstract

In the analysis of categorical data subject to misclassification errors, the observed cell proportions are adjusted by a misclassification probabilities and estimates of variances are adjusted accordingly. In this case, it is important to determine the extent to which misclassification probabilities are homogeneous within a population. This paper considers methods to evaluate the power of chi-squared tests for homogeneity with complex survey data subject to misclassification errors. Two cases are considered: adjustment with homogeneous misclassification probabilities; adjustment with heterogeneous misclassification probabilities. To estimate misclassification probabilities, logistic regression method is considered.

Keywords : Heterogeneous misclassification probabilities, Logistic regression, Measurement error, Stratified multistage sample survey

1. Introduction

In the analysis of categorical data, if misclassification errors exist, then estimated cell probabilities may biased and standard Pearson chi-squared tests may have inflated true type I error rates. For some general background on the analysis of categorical data subject to misclassification, see e.g., Mote and Anderson (1965), Tenenbein (1972), Hochberg and Tenenbein (1983) and Selen (1986). For specific work with misclassification problems in the analysis of stratified multistage sample survey data, see, e.g., Rao and Thomas (1991).

Rao and Thomas (1991) discussed methods to adjust chi-squared test statistics for goodness-of-fit with complex survey data subject to misclassification probabilities are equal across all units in a specified population.

This paper considers extensions of the Rao and Thomas (1991) method to tests of homogeneity, following Scott and Rao (1981). In addition, this paper examines cases in which misclassification probabilities may be heterogeneous within populations. For the latter case, I

1) Full-time Lecturer, Department of Statistics, Changwon National University, Changwon, 641-773, Korea.
E-mail : syheo@sarim.changwon.ac.kr

use estimated power curves to examine the extent to which heterogeneous misclassification probabilities may have a serious impact on inference. The proposed methods are applied to the data from the Dual frame National Health Interview Survey (NHIS) / Random-Digit-Dialing (RDD) Methodology and Field Test Project. Research Triangle Institute (RTI) designed this study and carried out data collection and initial data analysis.

2. Notation

Suppose that there are two independent populations and that two independent samples of sizes n_1 and n_2, respectively, are taken from these populations. In addition, suppose that there is a categorical variable with J mutually exclusive and exhaustive classes. Define $\pi_{i+}=(\pi_{i1},\pi_{i2},\cdots,\pi_{iJ})'$ and $p_{i+}=(p_{i1},p_{i2},\cdots,p_{iJ})'$ to be the vectors of J true and observed proportions, respectively, corresponding to the J classes for populations $i=1,2$. The hypothesis of homogeneity of the two populations is H_0: $\pi_1=\pi_2=\pi_0$ against H_1: $\pi_1\neq\pi_2$, where π_i are vectors with the first $(J-1)$ elements of π_{i+}, $i=1,2$; and π_0 is an unknown vector. In addition, define Z be an observed class, Y the true class, X a predictor, S a population label. Let $P(Z=k\mid Y=j,X=x,S=i)$ equal the probability that a unit reports membership in class k conditional upon $Y=j$, $X=x$, $S=i$. For convenience, I use the notation $P(Z=k\mid Y=j,X=x,S=i)$ and $P_i(Z=k\mid Y=j,X=x)$ interchangeably.

When misclassification errors exist, it can be important to determine the extent to which misclassification probabilities are homogeneous within specified groups. For this paper, I will say that misclassification probabilities are homogeneous within a population i if, for a given vector of explanatory variables x, $P_i(Z=k\mid Y=j,X=x)$ does not depend on x. In addition, I will say that misclassification errors of a population i are homogeneous when the population has homogeneous misclassification probabilities.

When misclassification probabilities are homogeneous, customary design based estimators of the proportions of reported classifications will converge to

$$p_{i+}=A_i'\,\pi_{i+} \tag{2.1}$$

where $A_i=[a_{i,jk}]$ is a $J\times J$ matrix with (j,k)th element $a_{i,jk}$. The (j,k)th element of matrix A_i is the probability, denoted $P_i(Z=k\mid Y=j)$, of a unit being classified into the kth class when its true class is j.

Suppose now that there are categorical explanatory variables and that the intersection of all of the explanatory variable categories partitions the population i into C groups. Then, for group c and population i,

$$p_{ic+}=A_{ic}'\,\pi_{ic+} \tag{2.2}$$

where p_{ic+} is a vector of proportions of observed classification rates for group c in population i, π_{ic+} a vector of true proportions and A_{ic} is the associated misclassification matrix. More specifically, define $A_{ic}=[a_{ic,jk}]$ to be a $J\times J$ matrix with (j,k)th element $a_{ic,jk}$, where $a_{ic,jk}=P_{ic}(Z=k \mid Y=j)$ for group c and population i. The vector p_{ic+} is defined as

$$p_{ic+}=(M_{ic})^{-1}(\sum_{t\in U_{ic}} I_{t1},\cdots,\sum_{t\in U_{ic}} I_{tJ})'$$

where U_{ic} is the subpopulation of persons in group c and population i, M_{ic} is the size of U_{ic} and I_{tk} is a dummy variable that equals one if a person gives answer k and zero otherwise. Similarly, the vector π_{ic+} is

$$\pi_{ic+}=(M_{ic})^{-1}(\sum_{t\in U_{ic}} \delta_{t1},\cdots,\sum_{t\in U_{ic}} \delta_{tJ})'$$

where δ_{tj} equals one if a person's true category is j and zero otherwise. By this definition, the combined vector of observed proportions for population i is

$$p_{i+}=\sum_{c=1}^{C} R_{ic}A'_{ic}\pi_{ic+} \tag{2.3}$$

where $R_{ic}=M_i^{-1}M_{ic}$ and M_i is the number of units in population i. When $A_{i1}=\cdots=A_{iC}=A_i$, expression (2.3) is equal to $A_i'\pi_{i+}$ where $\pi_{i+}=M_i^{-1}(\sum_{t\in U_i}\delta_{t1},\cdots,\sum_{t\in U_i}\delta_{tJ})'$ and U_i is the population i.

Assume now that all A_{ic} are all nonsingular matrices and that are not all equal. Let $B_{ic}=(A_{ic}')^{-1}$. Then from expressions (2.2) and (2.3)

$$\pi_{i+}=\sum_{c=1}^{C} R_{ic}B_{ic}p_{ic+} \tag{2.4}$$

and $B_{ic}=[b_{ic,jk}]$. When all A_{ic} are equal, expression (2.4) simplifies to $\pi_{i+}=(A_i')^{-1}p_{i+}$ where $p_{i+}=M_i^{-1}(\sum_{t\in U_i} I_{t1},\cdots,\sum_{t\in U_i} I_{tJ})'$.

3. Estimation of Cell Probabilities with Heterogeneous Misclassification Rates

3.1 Point Estimation

For population i, I assume the following design condition, quoted with minor modifications from Shao (1996, p. 205-206).

(D.1) The population has been stratified into L strata with N_h clusters in the hth stratum. For the hth stratum, $n_h \geq 2$ clusters are selected independently across the strata. These first-stage clusters are selected with unequal probabilities p_{hi} and with replacement. Within the ith first-stage cluster in the hth stratum, $n_{hi} \geq 1$ ultimate units are sampled according to some sampling methods with selection probabilities p_{hij} from N_{hi} units, $j=1,\cdots,n_{hi}$, $i=1,\cdots,n_h$, $h=1,\cdots,L$. The total number of ultimate units in the population is $N=\sum_{h=1}^{L}\sum_{i=1}^{N_h} N_{hi}$ and in the sample is $n=\sum_{h=1}^{L}\sum_{i=1}^{n_h} n_{hi}$.

For convenience, I will replace the triple subscript (hij) with the single subscript t in the following expressions if it is not necessary to specify strata, clusters and ultimate units. Under the design (D.1), let w_t be a unit-level survey weight. Then I have standard estimators of R_{ic} and p_{ic},

$$\hat{R}_{ic} = \hat{M}_i^{-1}\hat{M}_{ic} \tag{3.1}$$

where $\hat{M}_i = \sum_{t\in s_i} w_t$ and s_i is the set of sample units in population i; $\hat{M}_{ic} = \sum_{t\in s_{ic}} w_t$ and s_{ic} is the set of sample units in group c within population i; and

$$\hat{p}_{ic+} = \hat{M}_{ic}^{-1}\left(\sum_{t\in s_{ic}} w_t I_{t1},\cdots,\sum_{t\in s_{ic}} w_t I_{tJ}\right)'. \tag{3.2}$$

Thus from expressions (3.1) and (3.2),

$$\hat{R}_{ic}\,\hat{p}_{ic+} = \hat{M}_i^{-1}\left(\sum_{t\in s_{ic}} w_t I_{t1},\cdots,\sum_{t\in s_{ic}} w_t I_{tJ}\right)' = \hat{e}_{ic}, \tag{3.3}$$

say. In addition, from expression (2.4) I have

$$\hat{\pi}_{i+} = \sum_{c=1}^{C} B_{ic}\,\hat{e}_{ic} \tag{3.4}$$

and the jth element of $\hat{\pi}_{i+}$ equals $\hat{\pi}_{ij} = \sum_{c=1}^{C} B_{icj\cdot}\,\hat{e}_{ic}$ where $B_{icj\cdot} = (b_{ic,j1},\cdots,b_{ic,jJ})$ is the jth row of $J\times J$ matrix B_{ic}.

3.2 Variance Estimation

Assume that the matrices A_{ic} and thus B_{ic} are known. Define a $CJ\times 1$ vector $\hat{e}_{i\cdot} = (\hat{e}_{i1}',\cdots,\hat{e}_{iC}')'$. Note that $\hat{e}_{i\cdot}$ is a customary vector of sample ratios. Consequently, I can use standard methods (as in, e.g., Shao, 1996) to compute a design-based estimator of the variance of the approximate distribution of $\hat{e}_{i\cdot}$, $\hat{V}(\hat{e}_{i\cdot})$, say.

Also, note that expression (3.4) can be written as

$$\hat{\pi}_{i+} = B_{i\cdots}\ \hat{e}_{i\cdot} \tag{3.5}$$

where $B_{i\cdots}$ is a $J \times CJ$ matrix with jth row equal to a $1 \times CJ$ vector $B_{i\cdot j\cdot} = (B_{i1j\cdot}, \cdots, B_{iCj\cdot})$. Thus, an estimator of the variance of the approximate distribution of $\hat{\pi}_{i+}$ is

$$\hat{V}(\hat{\pi}_{i+}) = B_{i\cdots}\ \hat{V}(\hat{e}_{i\cdot})B_{i\cdots}' \tag{3.6}$$

with jth diagonal element $\hat{V}(\hat{\pi}_{ij}) = B_{i\cdot j\cdot}\ \hat{V}(\hat{e}_{i\cdot})B_{i\cdot j\cdot}'$.

4. Logistic Regression-Based Estimation of Misclassification Matrices

Let X be a vector of explanatory variables and let Z be a binary random variable with success probability $\pi(x)$ when X takes value x. Then a logistic regression model is

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

where (β_0, β_1) is a fixed vector of coefficients.

For an empirical analysis in Section 6, I consider that there are only two classes, $J=2$. Therefore, to estimate $a_{ic,jk}$ the logistic regression method can be considered. In this case, a logistic regression model is expressed by

$$g_i(x, D_j) = \beta_0 + \beta_1 D_j + \beta x \tag{4.1}$$

where $(\beta_0, \beta_1, \beta)$ is a fixed vector of coefficients, x is a vector of demographic or other auxiliary variables and D_j is an indicator variable indicating true category membership, and equals one when a unit's true category is j and 0 otherwise. In addition, $g_i(x, D_j) = \ln[P_i(Z=k \mid Y=j, X=x)/\{1-P_i(Z=k \mid Y=j, X=x)\}]$.

When all x are categorical variables and they partition each population into C groups, model (4.1) indicates that the probability of misclassification $a_{ic,jk}$ of a unit that truly belongs to class j can be estimated by

$$\hat{P}_{ic}(Z=k \mid Y=j, X=x) = [1+\exp\{\hat{g}_i(x, D_j)\}]^{-1}\exp\{\hat{g}_i(x, D_j)\},$$

where $\hat{g}_i(x, D_j) = \hat{\beta}_0 + \hat{\beta}_1 D_j + \hat{\beta} x$ and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta})$ is a consistent estimator of the vector $(\beta_0, \beta_1, \beta)$. Note that within group c and population i, all units in the sample have the same vector x. Thus, an estimator of A_{ic} is given by $\hat{A}_{ic} = [\hat{a}_{ic,jk}]$ where $\hat{a}_{ic,jk} = \hat{P}_{ic}(Z=k \mid Y=j, X=x)$.

5. Effect of Heterogeneous Misclassification Probabilities

When one considers heterogeneity of misclassification probabilities, the variances of adjusted estimators of cell proportions may be inflated due to the variability of A_{ic} within population i. If the bias of a biased test from incorrectly assuming their homogeneity is small relative to the amount of inflation in the point variance estimator that arise from accounting for heterogeneous misclassification probabilities, then one arguably might prefer the slightly biased test for some alternative hypothesis values. In this section, I will examine powers from unbiased and biased tests that do and do not account of heterogeneity of A_{ic}.

5.1 Asymptotic Distribution

Before examining powers, I first look at the asymptotic properties of $\hat{\pi}_{i+}$. The $\hat{\pi}_{i+}$ in expression (3.5) is a linear function of $\hat{e}_{i\cdot}$ and $\hat{e}_{i\cdot}$ is a vector of sample ratios. Therefore, if there are some conditions available under which $\hat{e}_{i\cdot}$ is a consistent estimator of $R_{ic}=M_i^{-1}M_{ic}$ and its asymptotic distribution follows normal distribution, then under the same conditions $\hat{\pi}_{i+}$ is a consistent estimator of π_{i+} and $n_i^{1/2}(\hat{\pi}_i-\pi_i)$ converge in distribution to $N_{J-1}(0, V_{\pi i})$, a $(J-1)$-variate normal distribution with mean 0 and covariance matrix $V_{\pi i}$, where π_i and $\hat{\pi}_i$ are vectors with the first $(J-1)$ elements of π_{i+} and $\hat{\pi}_{i+}$ in expression (2.4) and (3.5).

Under design (D.1), Shao (1996) gives conditions for consistency and asymptotic normality for design-based estimators of nonlinear functions of population totals, e.g., design-based sample ratio for estimating population ratio.

For a Wald-type test, I will now add the following condition.

(C.1) The matrix $n_i\{\hat{V}(\hat{\pi}_i)\}$ is a consistent estimator of $V_{\pi i}$ where $\hat{V}(\hat{\pi}_i)$ is the upper $(J-1)\times(J-1)$ submatrix of $\hat{V}(\hat{\pi}_{i+})$ in (3.6).

Then under design (D.1), condition (C.1) and additional regularity conditions, the Wald test statistics for homogeneity, $H_0: \pi_1=\pi_2=\pi_0$,

$$X^2_{he}=(\hat{\pi}_1-\hat{\pi}_2)'\,\hat{V}^{-1}(\hat{\pi}_1-\hat{\pi}_2), \tag{5.1}$$

where $\hat{V}=\hat{V}(\hat{\pi}_1)+\hat{V}(\hat{\pi}_2)$, is asymptotically distributed as χ^2_{J-1}, a chi-square random variable on $(J-1)$ degrees of freedom under $H_0: \pi_1=\pi_2=\pi_0$ for sufficiently large n_i, $i=1.2$.

For methods to obtain design-based consistent estimators of variances, see e.g., Krewski and Rao (1981) and Shao (1996).

5.2 Power Evaluation

For any nonzero $D_\pi = \pi_1 - \pi_2$, the test statistic X^2_{he} is distributed asymptotically as $\chi^2_{J-1}(\lambda)$, a chi-square random variable on $(J-1)$ degrees of freedom with noncentrality parameter λ, where $\lambda = D_\pi' \ V^{-1} D_\pi/2$ and $V = V(\hat{\pi}_1) + V(\hat{\pi}_2)$. Thus the power of the Wald test in (5.1) is

$$1-\beta_{he} = \Pr(X^2_{he} > \chi^2_{J-1,\alpha} \mid D_\pi) \doteq \Pr(W_{J-1} > \chi^2_{J-1,\alpha} \mid D_\pi)$$

where W_{J-1} is distributed as $\chi^2_{J-1}(\lambda)$ and $\chi^2_{J-1,\alpha}$ is the upper αth quantile of χ^2_{J-1}. When $H_0: \pi_1 = \pi_2 = \pi_0$ is true, the Wald test in (5.1) achieves the nominal type I error rate α.

Now assume $A_{i1} = \cdots = A_{iC} = A_i$ and assume that A_i are known. Then for known A_i, the estimator of π_{i+} is

$$\hat{\pi}^*_{i+} = (A_i' \ )^{-1} \hat{p}_{i+} \tag{5.2}$$

where $\hat{p}_{i+}$ are observed proportions. Its variance is estimated by

$$\hat{V}(\hat{\pi}^*_{i+}) = (A_i' \ )^{-1} \hat{V}(\hat{p}_{i+}) A_i^{-1}.$$

From a sample obtained by design (D.1), $\hat{p}_{i+} = \hat{M}_i^{-1} (\sum_{t \in s_i} w_t I_{t1}, \cdots, \sum_{t \in s_i} w_t I_{tJ})'$ for ith population, $i = 1, 2$. As with $\hat{e}_{i\cdot}$, $\hat{p}_{i+}$ is a vector of sample ratios and $\hat{V}(\hat{p}_{i+})$ is obtained by the same methods as $\hat{V}(\hat{e}_{i\cdot})$. Then the Wald test statistic for homogeneity, $H_0: \pi_1 = \pi_2 = \pi_0$, is

$$X^2_{ho} = (\hat{\pi}^*_1 - \hat{\pi}^*_2)' \ \hat{V}^{*-1} (\hat{\pi}^*_1 - \hat{\pi}^*_2) \tag{5.3}$$

where $\hat{V}^* = \hat{V}(\hat{\pi}^*_1) + \hat{V}(\hat{\pi}^*_2)$; $\hat{\pi}^*_i$ is a vector with the first $(J-1)$ elements of $\hat{\pi}^*_{i+}$; and $\hat{V}(\hat{\pi}^*_i)$ is a upper $(J-1) \times (J-1)$ submatrix of $\hat{V}(\hat{\pi}^*_{i+})$. The power of the Wald test in (5.3) is

$$1-\beta_{ho} = \Pr(X^2_{ho} > \chi^2_{J-1,\alpha} \mid D_\pi) \doteq \Pr(W^*_{J-1} > \chi^2_{J-1,\alpha} \mid D_\pi)$$

where W^*_{J-1} is distributed as $\chi^2_{J-1}(\lambda^*)$; $\lambda^* = (D_\pi + B)' (V^*)^{-1} (D_\pi + B)/2$; $V^* = V(\hat{\pi}^*_1) + V(\hat{\pi}^*_2)$; $B = b_1 - b_2$; and $b_i = \mathrm{E}(\hat{\pi}^*_i) - \pi_i$. Here E denotes expectation operator with respect to design (D.1).

When misclassification probabilities are heterogeneous, b_i is not zero. Due to this bias in $\hat{\pi}^*_i$, the power $1-\beta_{ho}$ under $H_0: \pi_1 = \pi_2 = \pi_0$ may be different from the nominal type I

error rate α and the Wald test statistic in (5.3) gives a biased test.

6. Application to Health Survey Data

6.1 Dual Frame NHIS/RDD Data

The U.S. National Health Interview Survey (NHIS) is a national level face-to-face survey carried out in all 50 states of United States. For some applications, sample sizes were considered insufficient to evaluate state level estimates.

The purpose of the Dual Frame NHIS/RDD Methodology and Field Test was to evaluate the feasibility of supplementing NHIS face-to-face interviews with Random-Digit-Dialing (RDD) telephone interviews. This study was conducted in two states, here labeled States A and B. These states were selected for the study due to their relatively large NHIS sample sizes (Biemer, 1997). In NHIS data, the initial interview was conducted face-to-face and the reinterview was conducted by telephone. For the RDD data, both interviews were conducted over the telephone.

From the questionnaire used for NHIS and RDD, I selected question $G1$, "Are any firearms now kept in or around your home?", with possible responses "yes" or "no". The hypothesis in which I are interested is $H_0: P(G1 = Yes \mid State\ A) - P(G1 = Yes \mid State\ B) = 0$. I combined NHIS and RDD data; and for purposes of this analysis I considered the second interviews to give the true responses.

6.2 Effect of Heterogeneous Misclassification Probabilities

To examine whether there are any auxiliary variables associated with probability of saying "yes" on question $G1$ on the second interview, I estimated the coefficients for the logistic regression model in (4.1). Some potentially important explanatory variables are a person's state of residence, gender, age and second interview modes; specific explanatory indicator variables are reported in Table 6.1. Exploratory analysis led to the final model coefficient estimates reported in Table 6.2. Based on Table 6.2, I constructed eight groups of respondents based on the combination of binary classification by gender, mode and Age40. For each group, estimates $\hat{A}_{ic}$ are obtained for both States A and B, $i=1,2$, $c=1,\cdots,8$, for the combined data. For the NHIS and RDD data, there are only four groups within each state, that is, $c=1,\cdots,4$. I considered the estimates as known to evaluate powers. The estimator of variance of $\hat{e}_{i\cdot}$, $\hat{V}(\hat{e}_{i\cdot})$, was obtained by the linearization method (StataCorp, 1997, Reference P-Z, p. 418). Table 6.3 shows point estimates of $\hat{\pi}_i$ and their standard errors, $\sqrt{\hat{V}(\hat{\pi}_i)}$, for the NHIS and combined data, respectively.

Table 6.1 Explanatory indicator variables for the logistic regression model.

Variable Name	Group Indicated
(Baseline Gender)	(Female)
Male	Male
(Baseline Interview Mode)	(NHIS)
RDD	RDD
Fire2	Second Interview of $G1 = Yes$
Fire2_RDD	Interaction between Fire2 and RDD
(Baseline Age)	age $\in$ [18, 39] years old
Age40	age $\geq$ 40 years old
Age40_Fire2	Interaction between Age40 and Fire2

Table 6.2 Logistic regression coefficient point estimates, standard errors, approximate 95% confidence interval and p-values for $H_0: \beta_k = 0$

Predictor	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	p-value	$(\hat{\beta}_{kL}, \hat{\beta}_{kU})$
Constant	-4.0272	0.3062	0.000	(-4.6278, -3.4266)
Male	0.4805	0.1812	0.008	(0.1251, 0.8359)
RDD	-0.5973	0.2796	0.033	(-1.1457, -0.0488)
Fire2	5.8095	0.3328	0.000	(5.1568, 6.4622)
Fire2_RDD	1.4697	0.3843	0.000	(0.7159, 2.2235)
Age40	1.5805	0.3137	0.000	(0.9654, 2.1957)
Age40_Fire2	-1.6857	0.3676	0.000	(-2.4066, -0.9648)

Table 6.3 Estimates of cell proportions and their standard errors under heterogeneous misclassification probabilities.

Data	Point Estimate	State A	State B
NHIS	$\hat{\pi}_i$	0.5060	0.2713
	$se(\hat{\pi}_i)$	0.0467	0.0217
Combined	$\hat{\pi}_i$	0.4679	0.2709
	$se(\hat{\pi}_i)$	0.0267	0.0163

Table 6.4 Estimates of cell proportions and their standard errors under homogeneous misclassification probabilities.

Data	Point Estimate	State A	State B
NHIS	$\hat{\pi}^*_i$	0.4837	0.2857
	$se(\hat{\pi}^*_i)$	0.0283	0.0197
Combined	$\hat{\pi}^*_i$	0.4573	0.2779
	$se(\hat{\pi}^*_i)$	0.0186	0.0153

For homogeneous misclassification probabilities, each $a_{i,jk}$ in matrix A_i is estimated by

$$\hat{a}_{i,jk} = \hat{M}_{ij}^{-1} \sum_{t \in s_{ij}} w_t I_{tk}$$

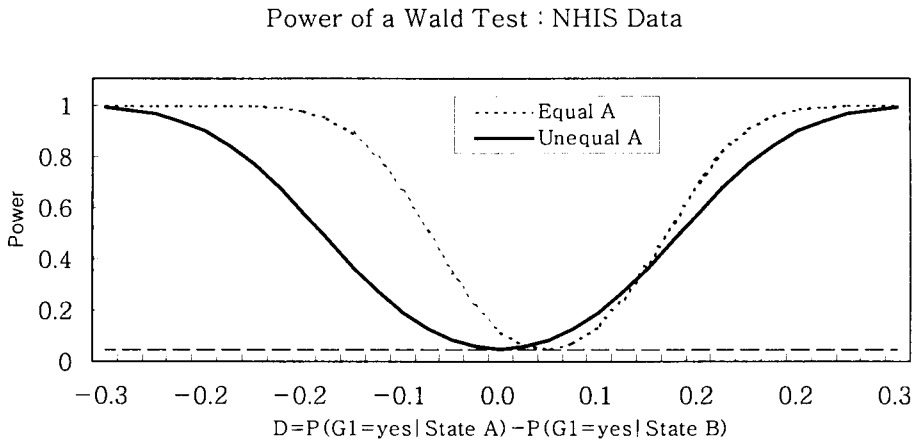

Figure 6.1 Power of a Wald test statistic with one degree of freedom for the NHIS data allowing for possible unequal misclassification probabilities.

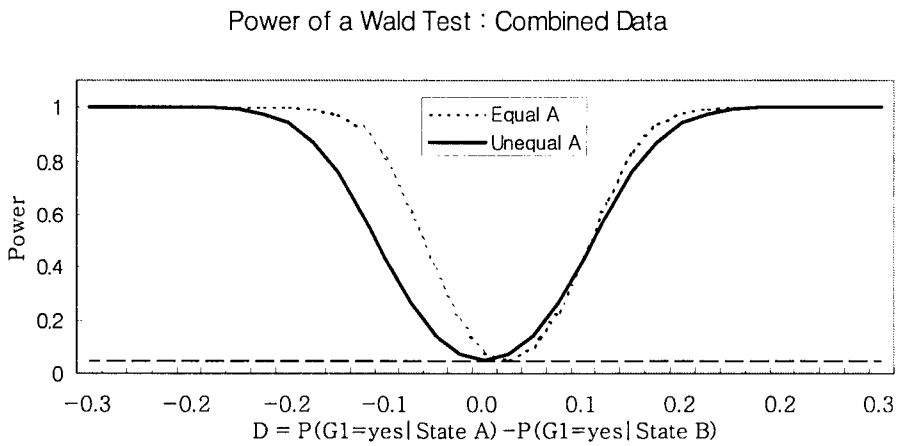

Figure 6.2 Power of a Wald test statistic with one degree of freedom for the Combined data allowing for possible unequal misclassification probabilities.

where $\hat{M}_{ij} = \sum_{t \in s_{ij}} w_t$ and s_{ij} is the set of sample units in belonging to category j in the second phase within population i; I_{tk} equals one if a person gives answer k in the first phase and zero otherwise. I consider these estimates $\hat{A}_i$ as known. The variance of $\hat{p}_i$ is estimated by the linearization method. Table 6.4 reports point estimates of $\hat{\pi}^*_i$ and their

standard errors, $\sqrt{\hat{V}(\hat{\pi}^*_i)}$, for the NHIS and combined data, respectively. The bias of $\hat{\pi}^*_1 - \hat{\pi}^*_2$ is

$$B = b_1 - b_2 = \{E(\hat{\pi}^*_1) - E(\hat{\pi}^*_2)\} - (\pi_1 - \pi_2)$$

and is estimated by

$$\hat{B} = (\hat{\pi}^*_1 - \hat{\pi}^*_2) - (\hat{\pi}_1 - \hat{\pi}_2)$$

since $\hat{\pi}_i$ is an unbiased estimator of π_i. These estimated biases are used to evaluate the power of a test based on an incorrect assumption of homogeneous misclassification probabilities when it is not true. Figure 6.1 and Figure 6.2 show powers from tests adjusted with homogeneous (dotted line) and with heterogeneous (solid line) misclassification probabilities. Figure 6.1 shows powers from the NHIS data and Figure 6.2 shows powers from the combined NHIS and RDD data. Both plots display a similar pattern.

In both graph, the test based on assuming $A_{ic} = A_i$ appears to have a positive bias, and the type I error rate is inflated accordingly. On the other hand, the inflation of variance due to accounting for heterogeneity of misclassification probabilities is nontrivial relative to the biasedness caused by incorrectly assuming their equality. The loss of power due to accounting for heterogeneous misclassification probabilities appears to be more severe for the NHIS data.

For the RDD data the difference between the two power curves is relatively small when it is compared to the NHIS and combined data, even though there is some positive biasedness exhibited when homogeneity is assumed.

7. Conclusions

I discussed chi-squared tests for homogeneity based on complex sample survey data subject to misclassification errors. I considered estimation based on either homogeneous or heterogeneous misclassification matrices. In addition, I evaluated power curved under assumption that the misclassification probabilities might be heterogeneous. Wald tests are used for power evaluation. I modeled misclassification mechanisms with logistic regression.

The proposed methods were applied to data from the Dual Frame National Health Interview Survey (NHIS)/ Random-Digit-Dialing (RDD) Methodology and Field Test Project conducted by Research Triangle Institute (RTI) in U.S..

The resulting power curves showed that the inflation of variance relative to the biasedness caused by incorrectly assuming their equality. Therefore, the loss of power of that arises in accounting for the heterogeneity of misclassification probabilities is of serious concern.

In this discussion, I assumed the coefficients of logistic regression and the heterogeneous misclassification matrices were known. However, in practical cases, these matrices may be estimated with nontrivial error. One could extend this ideas and method for these estimation errors.

References

[1] Biemer, P. P. (1997). Dual frame NHIS/RDD methodology and field test. Analysis report, Research Triangle Institute, Research Triangle Park, NC.

[2] Hochberg, Y. and Tenenbein, A. (1983). On triple sampling schemes for estimating from binomial data with misclassification errors. *Communications in Statistics - Series A, Theory and Methods*, **12**, 1523-1533.

[3] Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* **9**, 1010-1019.

[4] 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.

[5] 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*, 637-663. New York: John Wiley & Sons.

[6] Scott, A. J. and Rao, J. N. K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In D. Krewski, R. Platek, and J. N. K. Rao (eds.), *Current Topics in Survey Sampling*, 247-265. New York: Academic Press.

[7] Selén, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association* **81**, 75-81.

[8] Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics* **27**, 203-254.

[9] StataCorp (ed.). (1997). *Stata User's Guide, Release 5*. College Station, TX: Stata Press.

[10] Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics* **14**, 187-202.

[2002년 8월 접수, 2002년 11월 채택]