非標本誤差管理技法

1993. 8

統計企劃局
調査管理課

목 차

非標本誤差管理技法

美國商務省 센서스局
(BUREAU OF THE CENSUS)
金 鍾 益 博士

目　　次

A Glimpse of 1990 U.S. Census Evaluation Studies for the Purpose of Undercount Adjustment

Jongik Kim

In the middle of 1980's, the U.S. Deparment of Commerce, in particluar, the Bureau of the Census, was sued by many large cities such as New York City and Chicago, to name a few, to force them to adjust the population counts which are to be derived from the 1990 decennial census. The evaluation of the previous censuses indicated that there were undercounts for some racial groups, even if there were sometimes overcounts in some parts of the country. The undercount of the population is a serious problem, since the apportionment, i.e., allocation of congressional seats among the states, and the allocation of funds among different governmental units depend on the census population counts. The Federal judge ordered the Commerce Deparment to investigate the undercount issues. Thus the Bureau did an extensive study of the undercount and census adjustment issues. Subsequently, the Bureau conducted the 1990 decennial census on April 1, 1990. They also designed and conducted the Post Enumeration Survey (PES) to estimate the undercount rate of the 1990 census. Since PES itself is a survey and subject to errors, a followup survey called PES Evaluation Survey was designed and conducted. Note that the PES Evaluation Survey is to estimate the bias or systematic error of the PES estimate.

In a general terminology, census is an interview and PES is a reinterview. Concentrating on PES and PES Evaluation, the former is an interview and the latter is a reinterview.

In the following, reinterview, nonsampling errors in surveys and some examples of census/PES evaluation studies will be dealt with.

II. Reinterview

To measure the nonsampling errors in a survey, reinterview is usually conducted. Usually reinterview is conducted on a subset of the original survey including the respondents and questions. The reinterviews are conducted mainly for 3 purposes: i). check on the interviewer's performance and deter falsification; ii). estimate the bias and iii). estimate the variance. The bias here does not mean the sampling bias, but systematic nonsampling error such as caused by the interviewers framing questions favoring a specific response.

Reinterview is conducted differently whether the primary goal of the reinterview is the estimation of bias or variance. If the estimation of the bias is the primary reason for the reinterview,

best possible interviewers with a lot of interviewing experience need to be hired to get the response as close as to the truth. Thus usually supervisory interviewers (or field representatives) are hired for the reinterview. Probing questions are sometimes added to the original questions to get to the truth and if the response to the reinterview differs from that from the original interview for the same item, the interviewer is sent back to get more information from the same respondent for the reconcilliation of the difference. Or if it were a Computer Assisted Interview such as Computer Assisted Telephone Interview (CATI) or Computer Assisted Personal Interview (CAPI), the difference is resolved on the spot. That is, when the interviewer keys in the response to the reinterview, if is different from the original response, the computer pops up the original response so that the interviewer can tell that there is a difference. Thus the interviewer asks further questions and resolve the difference during the reinterview.

Again the variance here does not mean the sampling variance, rather it means the response variance. In order to estimate the response variance, reinterview situation should be created which is as close to the initial interview. Thus, regular interviewers, not the supervisory interviewers are used for the reinterview and the same questions without probing questions are asked.

The reinterview to estimate bias can be used to improve the estimator. The resulting estimator is called the Dual System Estimator (DSE). For the DSE, see the following table.

Table 1. Situation for Dual System Estimator

		Reinterview		
		in	out	sum
Initial Interview	in	X_{11}	X_{12}	$X_{1.}$
	out	X_{21}	X_{22}	$X_{2.}$
	sum	$X_{.1}$	$X_{.2}$	$X_{..} = N$

where $X_{ab} = \sum_{i=1} X_{abi}$, and X_{abi} is an indicator signifing whether or not the i^{th} respondent is in cell (a,b) or not for a,b = 1,2, .

In terms of probability for the respondent i, the table can be expressed as follows;

Table 2. Table 1 reexpressed in Probability

		Reinterview: in	Reinterview: out	sum
Initial Interview	in	P_{11i}	P_{12i}	$P_{1.i}$
	out	P_{21i}	P_{22i}	$P_{2.i}$
	sum	$P_{.1i}$	$P_{.2i}$	1

If the event of being included in the reinterview is independent of the event of being in the initial interview (which is called causal independence), then

$$P_{11i}P_{22i} = P_{12i}P_{21i}, \qquad \text{for } i=,2,, \quad , \quad N.$$

If the causal indenpendence is assumed in Table 1, then

$$X_{22} = \frac{X_{12}X_{21}}{X_{11}}.$$

The above is the formula which can be used to estimate those who were missed in both interviews.

From Table 1, we can see that

$$\frac{X_{11}}{X_{1.}} = \frac{X_{.1}}{X_{..}}$$

Thus

$$X_{..} = N = \frac{X_{.1}X_{1.}}{X_{11}}$$

Note in the above N is the DSE. In the census, $X_{.1}$ and $X_{1.}$ are a little complicated, which will be shown later.

It should be noted that for the census evaluation studies, we used the reinterview approach for estimating the bias. The experienced interviewers for the Bureau of the Census were hired to do the PES interview (or reinterview).

III. Adjusting the 1990 U.S. Census

The Bureau had two tools to measure the coverage (undercount rate) of the 1990 census: Demographic Analysis (DA) and PES. DA was handled by the demographers and was not considered as a good approach as the PES. Thus I will concentrate on the PES.

PES is composed of two samples: E- and P- sample. Based on the 1980 undercount rate, the nation was stratified and from each stratum a predetermined number of blocks were selected. Note that the sampling units are blocks even if ultimate sampling units (USU's) are persons in the housing units. Around 9,000 blocks amounting to 150,000 housing units were selected in the PES.

From the same blocks, 2 enumerations were taken; census on April 1, 1990 and PES on July 15, 1990. The terminology of "PES" has double meanings. It most often means the E- and P- samples, but sometimes it basically means only P-sample. The PES conducted on July 15 is essentially P-sample. Thus,

E-sample is composed of households enumerated during the census from the selected blocks;

P-sample is composed of households enumerated during the PES from the same blocks as before;

Thus, in terms of geography, the two samples are the same. Only difference is that E-sample is composed of the initial interviews and P-sample the reinterviews.

Table 1 can be rephrased as follows;

Table 3. PES

		P-Sample		
		in	out	sum
E-Sample	in	$X_{11}=M$	$X_{12}=C-M$	$X_{1.}=C$
	out	$X_{21}=N_p-M$	X_{22}	$X_{2.}$
	sum	$X_{.1}=N_p$	$X_{.2}$	$X_{..}=N$

In the above table, $X_{11}=M$ is the total number of matches between interview and reinterview. To determine whether a person in E-sample is a match to a perspon in P-sample, we used "Matcher," whcih is a computerized version of Felligi-Sunter algorithm. Based on 5 to 6 key variables, computer determines whether a case is "mathch," "non-match" or "unresolved." The unresolved cases come to the clerks' attention for resolution. Then they investigate the cases further. Sometimes we send out the interviewers to the field or call the respondents from the processing offices for more information.

Roughly speaking, $X_{21}=N_p-M$ is omissions, i.e., they should have been counted in the census, but not enumerated. C is the census count and $X_{12}=C-M$ is the number of erroneous enumerations, which are mainly duplications and falsifications.

Now the DSE can be expressed more precisely as follows:

$$\hat{N} = \frac{N_p C^*}{M} = C^* \frac{N_p}{M}$$

$$= C^* \frac{M + X_{21}}{M} .$$

C^* is modified C excluding the erroneous enumerations and substituted persons.

So far we have talked about the total persons in the sample blocks. In order to extend this to the whole population, we weight each sample person up, and thus if weighted C^*, N_p and M are used, we have the the estimated national population total.

Let $W\hat{N}$, WN_p, WC^*, WC and WM are the weighted counts of $\hat{N}$, N_p, C^*, C and M, respectively. Then the adjustment factor is

$$\frac{W\hat{N}}{WC} = \frac{WC^*/WC}{WM/WN_p} .$$

Let the above be denoted by R. Then the undercount rate is

$$1 - \frac{1}{R} .$$

IV. Nonsampling Errors Affecting the Adjustment

Eight potential sources of error affect coverage measurements produced by PES. They are

i). errors committed in matching each person in P-sample to the original census enumeration, including

false matches
false nonmatches

ii). errors in the reporting in the P-sample interview of census day address, including false reports by movers that they did not move

iii). fabrications in the P-sample interview, where the interviewer completes a questionnaire with fictitious people or ficticious characteristics instead of conducting a proper interview

iv). errors in the measurement of enumeration status of the original census enumeration, including

false erroneous enumeration
false correct enumeration

v). "correlation bias" due to heterogeneous census capture pobabilities whithin a poststratum

vi). errors introduced by the statistical treatment of any missing data

vii). uncertainty in balancing the estimates of gross undercount and gross overcount created by misspecification or inconsistent application of a common reference (or search) area in the P- and E-sample.

viii). random error due to sampling or to various forms of random nonsampling error

The first seven components of error are known to bias the DSE of population size. The eighth component causes its variance.

In the following we will discuss the estimation of correlation bias, capture probability and evaluation of synthetic assumption.

V. Estimation of Correlation Bias Using Conditional Logistic Regression

1. BACKGROUND

The dual system estimation used for the U.S. Bureau of the Census 1990 Post Enumeration Survey (PES) estimates is based on three independence assumptions: causality, homogeneity, and autonomy. Basically these assumptions say, respectively, that inclusion in the PES sample and the census are independent, that everyone has the same probability of inclusion, and that everyone acts on their own as to whether they are included in the PES sample population or the census. The violation of any of these three assumptions may cause the estimate of the proportion of the population enumerated in the census, and thereby the estimates of the population, to be biased. Such a bias is known as a correlation bias. The focus of this paper is on evaluating whether the homogeneity assumption holds.

2. STATISTICAL METHODOLOGY

To discuss the estimation of correlation bias, we need to define the dual system estimator (DSE). The present application of the dual system estimator involves two incomplete lists of the population. The census enumerations of the population not living in institutions or homeless comprise the first list. The second is an implicit list of those persons covered by the sampling frame for the P sample of the PES, which we will call the P-sample population; this list would be obtained if the P sample were conducted for the entire U.S. (instead of a sample) with no measurement errors or missing data.

Whether the i-th individual in the population of size N is in the census or not and in the P sample or not are assumed to be random events with probabilities as shown in Table 2.1. The true population size in each category is also shown in Table 2.1, and $N_{++} = N$ is the total population size. Even if we could observe the N_{jk}'s in the first row and first column, the N_{jk}'s in parentheses would not be observed directly but would have to be estimated. The estimator, $\hat{N} = N_{1+}N_{+1}/N_{11}$, is called the *DSE*. The DSE is accurate only to the extent that N_{11}/N_{+1} is an accurate estimate of the proportion of the population enumerated in the census. Accuracy depends on certain independence assumptions being satisfied (Wolter 1986):

Table 2.1. Probabilities of Inclusion and Population Sizes in a Cell

Inclusion Probability | True Population Size

		Original Enumeration		
		In	Out	Total
P-sample Pop.	In	$P_{11}\|N_{11}$	$P_{12}\|N_{12}$	$P_{1+}\|N_{1+}$
	Out	$P_{21}\|N_{21}$	$P_{22}\|(N_{22})$	$P_{1+}\|(N_{2+})$
	Total	$P_{+1}\|N_{+1}$	$P_{+2}\|(N_{+2})$	$P_{++}\|(N_{++})$

Causal Independence. The event of being included in the census is independent of the event of being included in the P-sample population. That is, the cross-product ratio $\theta_i = P_{11}P_{22}/P_{12}P_{21}$ is equal to 1 for each person $i = 1$, for $i = 1, \ldots, N$.

Autonomous Independence. The two lists, census and the P-sample populations, are formed in N mutually independent trials.

Heterogeneous Independence. The covariance between P_{1+} and P_{+1} is 0, with covariance defined as $N^{-1}\Sigma(P_{1+} - \bar{p}_{1+})(P_{+1} - \bar{p}_{+1})$, with $\bar{p}_{1+} = N^{-1}\Sigma P_{1+}$ and $\bar{p}_{+1} = N^{-1}\Sigma P_{+1}$. A sufficient condition for heterogeneous independence is *homogeneity*, i.e., that $P_{1+} = P_{2+}$ or $P_{+1} = P_{+2}$ for $i = 1, \ldots, N$.

Sekar and Deming (1949) suggested forming poststrata, groupings of the population by demographics (e.g., age, race, sex) and geography, so that the homogeneity assumption holds within each poststratum.

The Census Bureau poststratifies the persons in the PES according to demographic and geographic variables (Alberti et al. 1988). An estimate of the population size in each poststratum is calculated and then the estimates are summed to give an estimate for the total population.

Poststratification reduces but does not eliminate the effect of failure of the heterogenous independence assumption. Having independent field operations avoids failure of the causality assumption. Failure of autonomy tends to increase variance but has only a negligible effect on the bias; see Cowan and Malec (1986) and Wolter (1986).

Let $\theta = N_{11}N_{22}/(N_{12}N_{21})$ be the overall cross-product ratio and let $\tau = \theta - 1$. We will refer to τ

as the *correlation bias factor* that reflects failure of the independence assumptions. If the independence assumptions hold then $1 = \theta = \theta_i$ for $i = 1,\ldots,N$. The correlation bias may be expressed as follows:

$$\bar{N} - N = -\tau N_{12} N_{21} / N_{11} + O_p(\bar{N}^{1/2})$$

with the O_p term the random component of correlation bias that is negligible in this application (Wolter 1986).

Our goal is to estimate the correlation bias factor. A conditional logistic estimation procedure (Alho, 1990) is used to estimate the probabilities of inclusion in the census and the P sample, P_{i+} and P_{i+1}. This method allows analysis of dual system data using individual level covariate information. No grouping of the data is required as in the method of Sekar and Deming, and no completely independent source of information, such as demographic analysis, is needed. Having estimated the inclusion probabilities, we can estimate the correlation and obtain an estimate of τ, the *correlation bias factor*.

2.1 Calculation of Inclusion Probabilities

For ease of notation, let $P_{i1} = P_{i+1}$ be the probability of the i-th individual being included in the census, and let $P_{i2} = P_{i+}$ be probability of the i-th individual being included in the P-Sample population.

Conditional logistic regression requires assuming that we have vectors X_{i1} and X_{i2} of "explanatory" variables giving the characteristics (e.g. age, sex, tenure) of individuals correlated with inclusion. The inclusion probabilities, P_{i1} and P_{i2}, can be modeled as follows:

$$\log\left(\frac{P_{i1}}{(1 - P_{i1})}\right) = X_{i1}^T a_1$$

$$\log\left(\frac{P_{i2}}{(1 - P_{i2})}\right) = X_{i2}^T a_2$$

where a_1 and a_2 are vectors of parameters, which are estimated. Newton's method is used to estimate the parameters, a_1 and a_2, iteratively.

2.2 Estimation

A conditional logistic regression model produces an E-sample inclusion probability and a P-sample inclusion probability for each person. These probabilities may be used to calculate a correlation bias factor using Spencer's estimator.

2.2.2 Spencer's Method of Calculating τ Using Only Resolved Cases

Spencer has developed an estimator of τ using the covariance of P_{i1} and P_{i2} (1991). When only resolved cases are used, the estimator has the following form.

$$\tau = \frac{Cov(P_{i1}, P_{i2})}{(\bar{P}_1 - \bar{P}_2)(\bar{P}_2 - \bar{P}_2)}$$

where

$$Cov(P_{i1}, P_{i2}) = \bar{P}_2 - \bar{P}_1\bar{P}_2,$$

$$\bar{P}_2 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i1} P_{i2}}{\phi_i}}{\sum_{i=1}^{n} \frac{W_i}{\phi_i}}, \quad \bar{P}_1 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i1}}{\phi_i}}{\sum_{i=1}^{n} \frac{W_i}{\phi_i}},$$

$$\bar{P}_2 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i2}}{\phi_i}}{\sum_{i=1}^{n} \frac{W_i}{\phi_i}}$$

W_i = stratum weight for the i-th individual.
$\phi_i = P_{i1} + P_{i2} - P_{i1} * P_{i2}$ for $i = 1,\ldots,n$.
n = number of resolved cases.

2.2.3 Spencer's Method of Calculating τ Using Resolved and Unresolved Cases

The difference between the estimator using unresolved E-sample cases and the estimator using unresolved P-sample cases is in the calculation of $\bar{P}_1$, $\bar{P}_2$ and $\bar{P}_r$. Using unresolved E-sample cases, they are calculated as follows:

$$\bar{P}_2 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i1} P_{i2}}{\phi_i} + \sum_{i=1}^{n_2} W_i \gamma_{i1} P_{i2}}{\hat{N}_2},$$

$$\bar{P}_1 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i1}}{\phi_i} + \sum_{i=1}^{n_2} W_i \gamma_{i1}}{\hat{N}_2},$$

$$\bar{P}_2 = \frac{\sum_{i=1}^{n} \frac{W_i P_{i2}}{\phi_i} + \sum_{i=1}^{n_2} \frac{W_i \gamma_{i1} P_{i2}}{P_{i1}}}{\hat{N}_2},$$

where $\hat{N}_2 = \sum_{i=1}^{n} \frac{W_i}{\phi_i} + \sum_{i=1}^{n_2} \frac{W_i \gamma_{i1}}{P_{i1}}$

U_E = number of unresolved E-sample cases.
γ_u = the probability of being correctly enumerated, which is estimated for each unresolved E-sample case.

Using P-sample unresolved cases, $\bar{P}_1$, $\bar{P}_2$, and $\bar{P}_3$ are calculated as follows:

$$\bar{P}_3 = \frac{\sum_{i=1}^{n} \frac{W_i P_{1i} P_{2i}}{\phi_i} + \sum_{i=1}^{U_P} W_i P_{1i}}{\hat{N}_P},$$

$$\bar{P}_1 = \frac{\sum_{i=1}^{n} \frac{W_i P_{1i}}{\phi_i} + \sum_{i=1}^{U_P} \frac{W_i P_{1i}}{P_{2i}}}{\hat{N}_P},$$

$$\bar{P}_2 = \frac{\sum_{i=1}^{n} \frac{W_i P_{2i}}{\phi} + \sum_{i=1}^{U_P} W_i}{\hat{N}_P},$$

$$\text{where } \hat{N}_P = \sum_{i=1}^{n} \frac{W_i}{\phi_i} + \sum_{i=1}^{U_P} \frac{W_i}{P_{2i}}$$

U_P = number of unresolved P-sample cases.

3. IMPLEMENTATION

3.1 Adjustment of P-sample Capture Probabilities for Migration

Applying the conditional logistic modeling in the PES setting is complicated because of migration. Some people move in to PES sample blocks, and others move out. Only those individuals who were present in the PES sample area at census time should be included in the estimation of correlation bias. Therefore the inclusion probability for each person in the P-sample is multiplied by an estimate of the probability that the person was in the PES sample area at census time. The interpretation of the E-sample probabilities is in no way confounded by migration.

3.2 Accounting for Data Errors

The application of the conditional logistic model also is complicated by errors during the PES data collection. These data errors create issues of which cases to include in model fitting and which to include in estimation. One issue of this is the cases which remain unresolved at the end of the matching operation. These cases are excluded from the model fitting but are included in the two of three estimators in Section 2.2. The underlying assumption is that the unresolved individuals have the same capture characteristics as those individuals that were resolved, given equal covariates.

Another issue is caused by geocoding error. Some P-sample people match to census people in the search area of the PES block. The search area is a ring of blocks surrounding the PES block. Such a match is allowed to compensate for minor geocoding errors in the census or PES. Such cases are included in the model fitting as a match. This formulation has the effect of adding the census people who match P-sample people to the E-sample.

4. DATA ANALYSIS

4.1 Model Building

Models were built for the following four minority evaluation poststrata which are aggregates of PES poststrata: Northeast central cities, South central cities, West central cities, and Midwest central cities. Models were built for E- and P-sample separately in each evaluation poststratum. Table 4.1 displays the regression coefficients for the E-sample and P-sample models.

Sex, race, tenure, place type 0, relationship, marital status, and census division (CD) are indicator variables, i.e. they receive a value of 0 or 1. The values for the indicator variables are assigned as follows: sex is '1' if female, race is '1' if black, tenure is '1' if renter, place type 0 is '1' if living in central cities in primary metropolitan statistical areas (the most densely populated areas), relationship is '1' if not related to the person who completed the questionnaire, marital status is '1' if married, and census division is '1' if living in the particular CD.

The 'rate' variables, such as renter rate, are block level variables. The variables age, household size, and the block level variables are standardized to give them a level of magnitude equivalent to the indicator variables.

Variables are selected based on significance tests, multicollinearity, and assumed sociological importance. There is a group of "core variables" which are common to each of the eight models. Not all the "core variables" have an impact on each of the four models, however, each of these variables does have an impact for some of the evaluation poststrata and was included for each poststrata in order to make comparisons and to simplify the model building process. The remaining variables are place type, place type x race, census division,

Table 4.1 Regression Coefficients (and Standard Errors) for the Minority, Central City Evaluation Poststrata

	E Sample				P Sample			
	Northeast	South	Midwest	West	Northeast	South	Midwest	West
Intercept	1.863 (0.119)	2.757 (0.116)	2.422 (0.109)	2.853 (0.115)	1.190 (0.109)	2.387 (0.096)	2.291 (0.100)	2.208 (0.126)
Age	-0.054 (0.064)	0.326 (0.112)	-0.112 (0.089)	0.143 (0.069)	-0.032 (0.058)	0.353 (0.108)	0.014 (0.010)	0.105 (0.056)
(Age)2	0.088 (0.029)	0.764 (0.110)	0.132 (0.036)	0.059 (0.040)	0.144 (0.028)	0.618 (0.081)	0.110 (0.032)	0.141 (0.035)
(Age)3	-0.021 (0.018)	-0.357 (0.125)	-0.012 (0.024)	-0.022 (0.023)	-0.058 (0.015)	-0.889 (0.149)	-0.072 (0.017)	-0.055 (0.020)
Sex	0.199 (0.043)	0.281 (0.041)	0.256 (0.049)	0.111 (0.058)	0.212 (0.040)	0.187 (0.034)	0.262 (0.044)	0.140 (0.049)
Race (Black)	-0.166 (0.129)	-0.315 (0.144)	-0.187 (0.071)	-0.589 (0.155)	0.351 (0.118)	-0.464 (0.119)	-0.288 (0.068)	-0.143 (0.137)
Hispanic				0.077 (0.128)				0.006 (0.101)
Tenure	-0.657 (0.108)	-0.603 (0.086)	-0.715 (0.063)	-0.755 (0.089)	-0.446 (0.101)	-0.773 (0.074)	-0.578 (0.056)	-0.357 (0.071)
HH Size		-0.163 (0.032)	-0.155 (0.037)	-0.338 (0.042)		-0.146 (0.030)	-0.165 (0.031)	-0.308 (0.039)
Renter Rate	0.103 (0.037)	-0.041 (0.034)	0.136 (0.042)	-0.080 (0.052)	-0.299 (0.033)	-0.424 (0.028)	0.128 (0.038)	-0.380 (0.045)
Black Rate	-0.012 (0.037)	0.026 (0.057)		-0.337 (0.066)	-0.283 (0.036)	0.041 (0.046)		0.000 (0.054)
Hispanic Rate				-0.169 (0.070)				-0.008 (0.054)
Multiunit Rate	-0.124 (0.034)	0.023 (0.029)	-0.260 (0.039)	0.067 (0.045)	0.155 (0.030)	0.009 (0.022)	-0.003 (0.035)	0.108 (0.037)
Vacancy Rate	-0.047 (0.021)	-0.196 (0.019)	-0.109 (0.024)	-0.018 (0.031)	0.017 (0.021)	-0.108 (0.016)	-0.156 (0.022)	-0.066 (0.025)
Place Type 0	0.045 (0.087)	-0.340 (0.087)	-0.275 (0.071)		0.017 (0.081)	-0.532 (0.069)	-0.401 (0.066)	
Relationship	-1.020 (0.090)	-0.962 (0.085)	-0.797 (0.099)	-1.141 (0.096)	-0.878 (0.081)	-1.054 (0.071)	-0.884 (0.089)	-1.180 (0.082)
Marital Status	0.180 (0.610)				0.326 (0.057)			
Tenure*Race	-0.035 (0.117)	0.167 (0.094)		0.752 (0.127)	0.089 (0.109)	0.487 (0.081)		0.037 (0.120)
Tenure*HH Size		0.194 (0.038)	0.240 (0.048)	0.163 (0.053)		0.149 (0.035)	-0.010 (0.041)	0.164 (0.048)
Age*Race	0.139 (0.054)	0.202 (0.115)	0.270 (0.078)	0.157 (0.065)	0.171 (0.049)	0.263 (0.088)	0.128 (0.072)	0.163 (0.060)
Race*Pl. Type 0	0.065 (0.100)	0.021 (0.115)			-0.285 (0.092)	0.256 (0.093)		
Age*HH Size	0.043 (0.024)	0.313 (0.055)	0.075 (0.030)	0.044 (0.098)	-0.053 (0.023)	0.004 (0.041)	0.006 (0.024)	-0.355 (0.084)
Sex*Age	0.094 (0.045)	0.440 (0.096)	0.119 (0.054)	-0.027 (0.062)	0.020 (0.042)	0.204 (0.076)	0.132 (0.045)	-0.025 (0.054)
South Atlantic Census Division		-0.366 (0.065)				-0.117 (0.052)		
East South Central Census Division		-0.287 (0.072)				-0.001 (0.059)		
East North Central Census Division			0.114 (0.097)				0.099 (0.089)	
Pacific Census Division				-0.344 (0.031)				-0.031 (0.027)

and Hispanic origin. Place type and census division are geographic variables which vary according to evaluation poststrata. The Hispanic indicator and Hispanic rate, a block level variable, are used in the West poststratum because this poststratum includes a substantial number of Asians.

4.2 Analysis of Odds Ratios

If the sex variable is coded as 0 for male and 1 for female, then the odds of being captured, or enumerated, for females is defined as P(1)/[1-P(1)], where P(1) is the capture probability for females. Similarly, the odds of being captured for males is defined as P(0)/[1-P(0)], where P(0) is the capture probability of male. The odds ratio, denoted by Ψ, is defined as the ratio of odds for females to the ratio of odds for males. Thus

$$\Psi = \frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]}.$$

The odds ratios for both the E and P sample for the Northeast and Midwest minority/central city evaluation poststrata are given in tables 4.2.1 and 4.2.2. The odds ratios for the South and West are similar to those for the Midwest.

Among the five effects considered, three effects, renter among black females in place type 0, non-relative, and black renter among females in place type 0, have an odds ratio consistently less than one for both the E and P sample. This implies that each of the three groups has a lower inclusion probability than its respective counterpart. The odds ratio for female among black renters in place type 0 is greater than one for both the E and P samples in both poststrata. Except for the P-sample for the Northeast, the odds ratio for black among female renters in place type 0 is less than one. Non-relatives and black renters have the consistently lowest odds ratios.

4.3 Inclusion Probabilities

Table 4.3 shows the average and range of inclusion probabilities for the four minority, central city poststrata. The South minority/central city poststratum had the highest average inclusion probability for both the E-sample and P-sample, .920 and .813 respectively, and the Northeast had the lowest for both the E and P sample, at .816 and .751 respectively. For each of the four poststrata, the average E-sample inclusion probability is higher than the average P-sample inclusion probability.

Table 4.2.1 Estimated Odds Ratios for Northeast, Minority, Central City Evaluation Poststratum

Effect	Among	Odds Ratio E	Odds Ratio P
Female	Black renter place type 0	1.220	1.236
Black	Female renter place type 0	.847	1.421
Renter	Black female place type 0	.519	.640
Non-relative	All	.503	.416
Black Renter	Female place type 0	.398	.748

Table 4.2.2 Estimated Odds Ratios for Midwest, Minority, Central City Evaluation Poststratum

Effect	Among	Odds Ratio E	Odds Ratio P
Female	Black renter place type 0	1.292	1.298
Black	Female renter place type 0	.829	.750
Renter	Black female place type 0	.489	.561
Non-relative	All	.451	.413
Black Renter	Female place type 0	.406	.421

Table 4.3 Average and Range of Inclusion Probabilities for Minority, Central City Evaluation Poststrata

	E-sample			P-sample		
	Ave	Max	Min	Ave.	Max	Min
NE	.816	.944	.496	.751	.948	.350
SO	.920	.998	.507	.813	.979	.429
MW	.864	.991	.373	.794	.963	.368
WE	.879	.981	.308	.785	.974	.249

4.4 Correlation Bias

The correlation bias factors are estimated using the capture probabilities for E- and P-sample based on Spencer's method for these four evaluation poststrata. Table 4.4 shows the correlation bias factors, their conditional standard errors, the undercount rates calculated using the usual DSEs, and undercount rates calculated using DSEs adjusted for correlation bias. The inclusion of unresolved cases has little impact on the estimates probably because the number of unresolved cases is small relative to the number of resolved cases. Thus, estimates which include unresolved cases are given for the Midwest only. MW(ue) includes unresolved E-sample cases, and MW(up) includes unresolved P-sample cases. T-test values are 2.022 for MW and MW(ue) and 2.454 for MW and MW(up). The correlation bias factor estimate for the South using unresolved P-sample cases was the only other estimate using unresolved cases which differed significantly, at the .05 level, from the corresponding estimate using only resolved cases.

Table 4.4 Correlation Bias Factors and the Effect of Correlation Bias on Undercount Rates for Minority, Central City Evaluation Poststrata

	Corr. Bias Factor	Std.* Error	Undct Rate	Adj. Undct Rate
NE	0.14	0.02	6.83%	7.31%
SO	0.34	0.04	5.68%	6.13%
WE	0.42	0.03	6.14%	7.27%
MW	0.25	0.03	3.97%	4.12%
MW(ue)	0.26	0.03	3.97%	4.12%
MW(up)	0.27	0.03	3.97%	4.12%

* Standard errors are conditional on the models.

We conclude by noting that the undercount estimates based on conditional logistic regression in table 4.4 are all higher than the ones based on the usual DSEs. This suggests that there has been some residual heterogeneity in the inclusion probabilities that the logistic regression has revealed. The E and P samples appear to have had a higher positive correlation than the one expected based on the usual stratified analysis.

5. REFERENCES

Alberti,N., Diffendal, G., Hogan, H., Isaki, C., Monsour, N., Passel, J., Robinson, G., Schenker, N., Thompson, J., Wolter, K., and Woltman, H. (1988) "Preliminary Poststratification Schemes for the 1990 Census Coverage by Measurement Programs," unpublished manuscript August 3, 1988. Bureau of the Census, Washington, D.C.

Alho, J. M. (1990) "Logistic Regression in Capture-Recapture Models," Biometrics, 46, 623-635.

Cowan C.D. and Male, D.J. (1986) "Capture-Recapture Models When Both Sources Have Clustered Observation," Journal of the American Statistical Association, 81, 347-353.

Ericksen, E.P. and Kadane, J.B. (1985) "Estimating the Population in a Census Year: 1980 and Beyond," Journal of the American Statistical Association, 80, 98-108, 129-131.

Sekar, C. C. and Deming, W. E. (1949) "On a Method of Estimating Birth and Death Rates and the Extent of Registration," Journal of the American Statistical Association, 44, 101-115.

Spencer, B.D. (1991) Derivation of *r*, March 4, 1991, unpublished manuscript.

Wolter, K. M. (1986) "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-346.

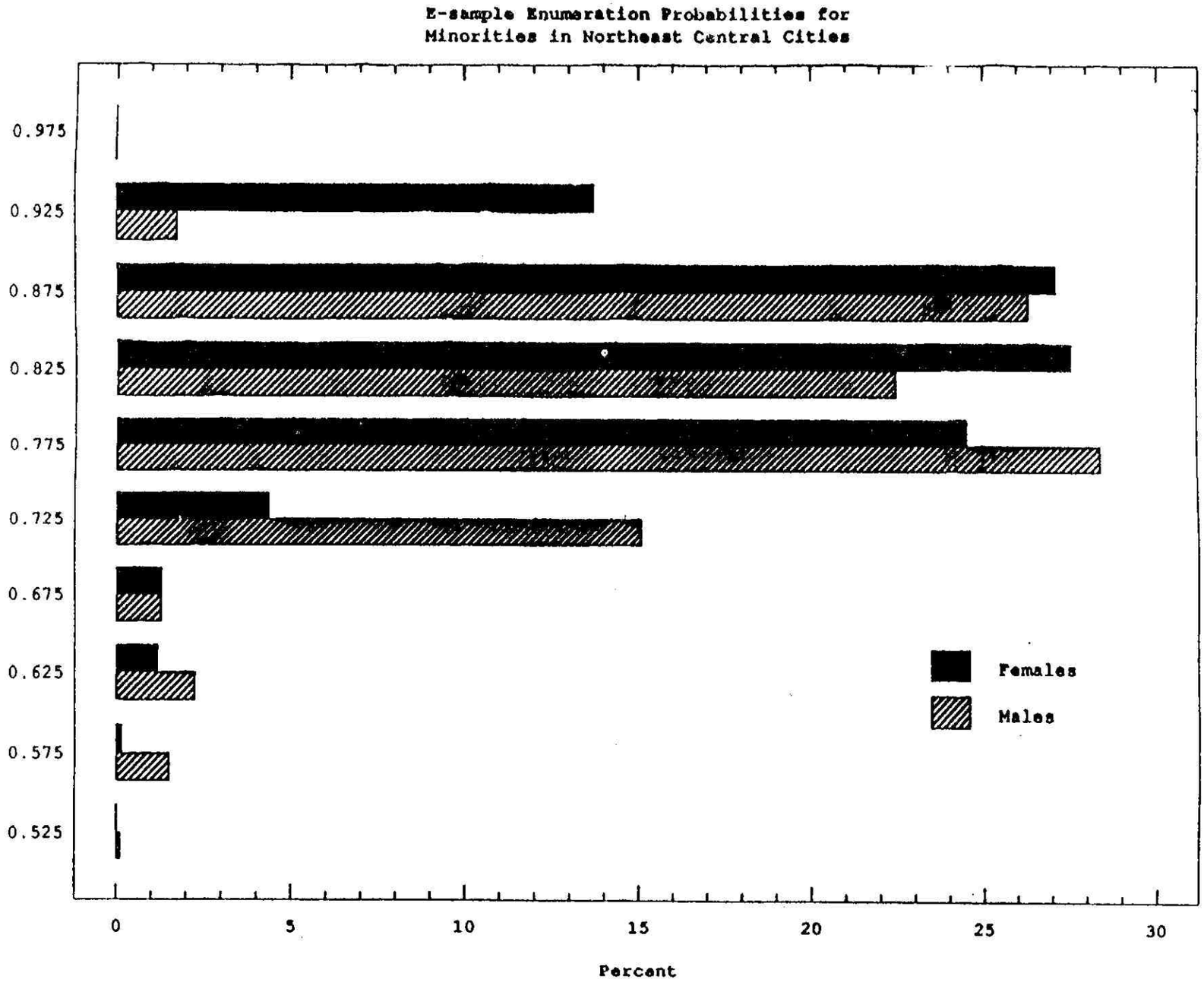
Wolter, K. M. (1986) "A Combined Coverage Error Model for Individuals and Housing Units," SRD Research Report Number Census/SRD/RR-86/27, Statistical Research Division Report Series, U.S. Bureau of the Census, Washington, D.C.

Acknowledgements

The authors wish to thank Ven Sathyamoorthy, David Carr, and Robert Fay for their contributions.

Juha Alho wishes to thank the Census Bureau for supporting this research through a joint statistical agreement with the NORC.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.


E-sample Enumeration Probabilities for
Minorities in Northeast Central Cities
0.975
0.925
0.875
0.825
0.775
0.725
0.675
0.625
0.575
0.525
0
5
10
15
20
25
30
Percent
Females
Males

VI. EVALUATION OF THE SYNTHETIC ASSUMPTION IN THE 1990 POST ENUMERATION SURVEY

Abstract

The Census Bureau considered adjusting the 1990 decennial census counts by a "synthetic adjustment" procedure which is valid under the assumption that undercount rates are homogeneous within poststrata defined by geography and demographic characteristics. This paper reports on evaluations of this assumption with an emphasis on detecting interstate heterogeneity within poststratum (PS).

Five surrogate variables (allocation rate, mail return rate, multiunit structure rate, mail universe rate and substitution rate) believed to be associated with undercount rate were analyzed. Significant interstate heterogeneity was found within most poststratum groups (poststrata collapsed over age and sex) by logistic regression.

A second analysis focused on data from the Post Enumeration Survey (PES). Estimated adjustment factors were calculated for state by poststratum group (PSG) domains (state parts). In every division, PSG explained more of the variance of these factors than did state, supporting the decision to use PSG as the adjustment cell.

Linear models were applied to a linearized version of the influence of each block part on the overall undercount rate to test the significance of state effects within PS. Although in some divisions there were several PSG's with significant state effects on adjustment factor, in only one division was the additive effect of state significant ($\alpha = .05$) in the state × PSG model when unweighted data was used. However, when weighted data was used, no division showed a significant state effect ($\alpha = .05$). Thus, there was not significant evidence that in the aggregate the poststratification was biased against certain states, although in 19 out of 99 PSG's the state effects were significant. The majority of these PSG's corresponded to non-urban areas, suggesting room for improvement in poststratification in those areas.

Key Words: Undercount Rate, Heterogeneity, Poststratum, Linearized Statistics

1. Introduction

The Post Enumeration Survey (PES) of the 1990 Decennial Census was designed to produce coverage estimates for 1392 poststrata. The nation was first divided into 116 domains, called poststratum groups (PSG's) according to geography, race/Spanish origin and tenure (owner vs renter). With only 4 exceptions, all PSG's are defined within a census division, one

of nine contiguous geographic areas each composed of several states. Each PSG was further divided into 12 age by sex groups, the poststrata. For example, roughly all Black renters in New York city constitute a PSG and all females age 0-9 of this PSG make a poststratum (PS). Further details on the PES are in Hogan (1992,1993).

Small area undercount rates were calculated by synthetic estimation; the same adjustment factor was applied to persons from a given PS in all areas. This procedure is accurate under the "synthetic assumption" of homogeneity of undercount rate within a PS. This paper reports research conducted as a part of the PES evaluation project on whether or not this objective of poststratification was achieved. It uses two data sets, one drawn from the Census and the other a combination of PES and Census data. The census data are a stratified cluster sample extract of 1990 Census data whose sample design is the same as that of the PES; the census extract has 204,394 blocks while the PES has 12,144 blocks.

In the analysis of the Census data, we selected variables which were considered highly correlated with the undercount rate to act as surrogates for undercount (Isaki, 1988). The selected surrogates are the allocation (item nonresponse) rate, mail return rate, multiunit structure rate, mail universe rate (fraction of units receiving mail questionnaire) and substitution (unit nonresponse) rate.

Under the homogeneity assumption, the rates are the same within a PS regardless of state. Thus, this assumption can be tested by comparing the surrogate variables or undercount rate from state to state within a PS; this test focuses attention on the question of whether synthetic estimation is "unfair" to certain state. The unit of the analysis is the intersection of a census block and a PS or PSG, called a block part (BP). A census block is a small area bounded by visible features such as streets, streams etc and/or by political boundaries. In fact, most of our analyses are performed on PSG's, since the age-sex breakdown of the PSG did not vary much from state to state. Hence, the analysis focuses on whether BP's differ between states within PSG.

A two-way ANOVA is fitted to undercount rates for the state parts, intersections of a state and PSG. This helps to compare the state effect with the PSG effect in undercount rates.

2. Analysis of Surrogate Variables

Surrogate variables are analyzed by logistic regression. Two forms of logistic regression model were used. For the within-PSG analysis, the model for PSG i is

$$\log [P_{ij} /(1-P_{ij})]= A + C_j$$

and for the within-division analysis,

$$\log [P_{ij} /(1-P_{ij})] = A + B_i + C_j$$

where P_{ij} is the rate for a surrogate variable in the i^{th} PSG and j^{th} state, A is the intercept, B_i is the i^{th} PSG effect and C_j is the j^{th} state effect. The models used only the 99 PSG's astride two or more states. Models were built for surrogate variables in the 99 PSG's and in each of nine divisions. SAS PROC CATMOD estimated the parameters by maximum likelihood and provided Wald statistics for testing the significance of state effects.

The data were collected with a cluster sample rather than a simple random sample so the test statistics must be divided by a design effect. We estimate a design effect,

$$\hat{D}_{ij} = \frac{\sum_{k=1}^{k=k_{ij}} n_{ijk}(\hat{p}_{ijk} - \hat{p}_{ij})^2}{K_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})}$$

where $\hat{p}_{ijk}$ is the rate for the i^{th} PSG, j^{th} state and k^{th} BP; n_{ijk} is the size of the BP; K is the number of BP's in the i^{th} PSG in the j^{th} state and $\hat{p}_{ij}$ is the rate for the i^{th} PSG and j^{th} state. The fraction is the ratio of the observed between-block variance to that expected under binomial sampling.

For division models, which span multiple PSG's, design effects were multiplied by the judgmentally chosen factor, 1.25, to account for the fact that the sampling unit is a block, which sometimes includes parts of more than one PSG. Without the factor, the design effect accounts for clustering only within a PSG.

A design effect was calculated for each surrogate variable and PSG. It is small (around 2) in most PSG's for the allocation and substitution rate. The effect is slightly higher for mail return rate, but it tends to be large (as much as 20) for multiunit structure and mail universe rate, since these factors are usually fairly uniform within a block.

Table 1 summarizes the design-corrected tests for state effects within PSG. Nationally, for each surrogate variable the state effect is significant for at least 84% of the PSG's. (The total number of PSG's varies because when a PSG falls entirely within one state or when only one state has non-zero observations for a particular variable, the corresponding model cannot be fit.). The results are summarized at the division level. (Divisions 1 through 9 are New England, Mid-Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain and Pacific Division.)

Table 2 shows the magnitude of state effects, expressed as χ^2 values of test statistics adjusted for design effect, for three

rates having relatively high correlation with the undercount rate.

In Table 2, the χ^2 values have from 1 to 8 degrees of freedom.

3. Analysis of Undercount Rate

The results described above for surrogate variables were obtained early in the census process, but they have limited relevance to homogeneity of undercount itself. After PES data were processed, direct analysis of the distribution of undercount became possible.

The data set for these analyses merged two data sets for the 12,144 PES sample blocks, one for the E-sample (Census followup) and the other for the P-sample (PES). Correct enumerations and E-sample total counts are on the E-sample file and match and P-sample total counts are on the P-sample file.

3.1 Variance Explained by State and PSG

For each division, a two-way ANOVA was fitted to undercount rates for state parts. Table 3 shows the ratio of the sum of squares due to PSG's to that due to states within a division. The ratio is always greater than one and in Division 9 it is 40.28, showing much larger effects for PSG than for state. The mean square for group also exceeds the mean square for state in each

division except Division 2. This supports the decision to use the PS rather than the state as the cell for undercount estimation and adjustment.

3.2 Tests for State Effects on Undercount Rates

Assuming the substitution rate (fraction of units imputed for nonresponse) is negligible, the adjustment factor ($\hat{R}$) for a domain is

$$\hat{R}=\frac{WCE/WE}{WM/WP},$$

and the undercount rate is

$$1-1/\hat{R},$$

where WE and WP are the estimated population sizes weighted up from the E and P-sample, respectively. WCE is the weighted number of correct enumerations and WM is the weighted number of matches in the PES.

The statistic for the influence (see Appendix) of the i^{th} BP on the adjustment factor or undercount rate is

$$I_i=\hat{R}\left(\frac{WCE_i}{WCE}+\frac{WP_i}{WP}-\frac{WE_i}{WE}-\frac{WM_i}{WM}\right),$$

where WCE_i, WP_i, WE_i and WM_i are contributions from the i^{th} BP to totals above.

A linear model was fitted to BP influence statistics to test for state effects. Under the null hypothesis, all the state

parts in a PSG have the same undercount rate and the expected mean of the influence statistics for each state is 0 within each PSG. The influence statistics can be analyzed with one way ANOVA within a single PSG or two way ANOVA for all PSG's within a division.

Table 4 summarizes the tests for state effects on linearized statistics within each PSG. The tests reveal significant heterogeneity between states in 19 out of 99 groups at the 5% significance level. The magnitude of the estimated state effect ranges from a few percent up to 20%, but the standard errors of these estimates are very large.

Table 5 summarizes the results of these analysis by place type. Place types 0, 1, 2 and 3 are large central cities in PMSA, place types 4, 5 and 6 are non-central cities in PMSA with large central cities and place types 7, 9 and 9 are other areas. The significant results are concentrated in PSG's for small areas (place types 7, 8 and 9). 10 out of 19 such groups show significant interstate heterogeneity at the 5% level. This suggests that the poststratification can be improved in those areas.

A linear model for linearized undercount at the division level showed no significant state effects when both PSG and state effects were included in the model.

Table 6 shows the F statistics and p-value for state effect for state x PSG models, once weighted by the size of domain and once without weights.

The additive effect of state was significant in only one division (p=.01) in the unweighted state × PSG model; when data were weighted by size of domain, the smallest p-value for the state effect was .18. In either case, the most significant effect was observed in Division 2, in which New Jersey appeared to have higher undercount rate, controlling for PSG, than New York, possibly because the most undercounted area in New York (New York City) had its own poststrata. Elsewhere, because the state effects in different PSG's varied in magnitude and sometimes in sign, and because only within a minority of PSG's in any division were there significant state effects, there was not significant evidence that in the aggregate the poststratification was biased against certain states.

4. Discussion

This paper evaluates the homogeneity, or synthetic, assumption for the 1990 Post Enumeration Survey.

The evaluation used 1990 Census data and 1990 PES data. Surrogate variables from the 1990 Census were tested for significant heterogeneity among states within PSG. At the PSG level, state effect was significant ($\alpha = .05$) for 84%-95% of its PSG's for the various surrogate variables.

ANOVA on linearized undercount based on the PES data at the PSG level showed significant ($\alpha = .05$) state effects for 19 out of 99 PSG's. The significant results were concentrated in the

poststratification in the relatively unurbanized areas was not as successful as in the more urbanized areas.

How can we explain the different findings of the two studies? The two data sets had very different sample sizes, i.e., the Census data had 204,394 blocks but the PES data had 12,144 blocks. Furthermore, the correlation between the undercount rate and the surrogate variables are low as shown in Table 7. It is therefore not surprising that small differences between states on surrogate variables would be statistically significant although corresponding differences would not be demonstrable with respect to undercount rates.

When this research was first embarked upon, the PES data were unavailable and were not expected to become available for analysis before the scheduled completion date.

Given the modest correlation between undercount rates and surrogate variables, we prefer to give greater weight to the analysis of the PES data.

We conclude from these data that there are no demonstrable differences in average undercount rate between states within each division, after adjusting for PSG effects. While there is weak evidence for a difference between New Jersey and New York within the Mid-Atlantic division, this result must be downweighted in the context of the number of divisions (nine) for which the test was performed. We conclude that if adjustment of population counts had been carried out based on the 1990 PES, no state would have been able to show that the poststratification was manifestly

counts had been carried out based on the 1990 PES, no state would have been able to show that the poststratification was manifestly unfair in that it underadjusted that state relative to what direct state estimates showed that it deserved.

Our question does not address another form of heterogeneity within the poststrata that has been mentioned by some critics as a source of error in the proposed undercount adjustment. Wachter and Freedman (1993) and Hengartner and Speed (1993) have suggested that undercount rates vary within PSG, considering units smaller than the state part (such as local political divisions or even individual blocks). They argue that this is an additional source of error in synthetic estimates for small areas that is considered in the Census Bureau's error models. The research of Fay and Thompson (1993) suggests that this type of heterogeneity probably does not have a systematic effect on comparisons of the accuracy of the adjusted and unadjusted counts.

This paper reports the general results of research undertaken by the authors. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Harvard University.

The second author was with the Bureau of the Census when the research was in progress.

Appendix : Testing for Interstate Differences Using Linearized Statistics

A two-way ANOVA for adjustment factors in state parts yields an intuitively meaningful summary of the relative contributions of state and PSG effects to the variation in adjustment factors. Because the sampling unit of the PES is the block cluster rather than the state part, these models do not yield valid statistical tests of the significance of the state effects.

Consider a statistic whose sample estimate for a state or state part is a weighted mean of the sample estimates in each component block or BP. Significance of the state effects for this statistic within a PSG could be evaluated by one-way ANOVA with the included block parts as units (corresponding to PSUs), or aggregated across PSGs by two-way ANOVA for state and PSG effects.

The sample adjustment factor estimate (WCE/WE)/(WM/WP) is a nonlinear function of sample counts. In small primary sampling units (PSUs) such as block parts this nonlinearity may be very noticeable, especially when the number of matches in a PSU is very small or zero so that the sample estimate of the adjustment factor is large or infinite. In this situation, if PSU sample estimates are treated as data, the additive assumptions of ANOVA are violated. Useful tests may be recovered, however, by using a linearized version of the statistic of interest.

Suppose that we are interested in a parameter Z = f(X) where X is a vector of population proportions in certain cells. Let $\bar{x}$, x_i represent the corresponding sample proportions in the entire sample and in PSU i respectively, so $\bar{x} = \sum N_i x_i / \sum N_i$ is a size-weighted average of block cell proportions. Let $f_1(X)$ be the gradient of f at X. Then by Taylor linearization $f(\bar{x}) - f(X) \approx f_1(X)'(\bar{x}-X) = \sum N_i f_1(X)' x_i / \sum N_i - f_1(X)'X$, i.e. we may treat the problem as one of inference regarding the quantities (pseudo-observations) $z_i = f_1(X)' x_i$. Because the approximate (linearized) influence of PSU i on the estimate $f(\bar{x})$, that is, the difference between the estimate with and without PSU i included, is $N_i f_1(X)'(x_i - \bar{x})$, we may describe this as a method based on influence statistics (Hampel et al. 1986) or the infinitesimal jackknife (Efron 1982, Chapter 6).

To derive a sensible variance model, suppose that we may regard PSU i as sample (not necessarily SRS) from a superpopulation with cell proportions X_i. A simple model is then, for some covariance matrices U_i and V_i,

superpopulation model:

$$E(X_i) = X, \quad Var(X_i) = V_i$$

and

sampling model:

$$E(x_i | X_i) = X_i, \quad Var(x_i | X_i) = U_i.$$

The sampling covariance U_i will typically be proportional to N_i^{-1}. A plausible and mathematically convenient specification for V_i is $V_i \propto N_i^{-1}$ (i.e. smaller PSUs more variable than larger ones), so Var $z_i = \sigma^2/N_i$ for some constant σ^2. The corresponding linear model weight for PSU i is N_i so the model-based estimate of the mean agrees with the design-based estimate obtained by aggregating the cell counts if N_i is a weighted size measure.

In the case of the adjustment factor $\hat{R}$ = (WCE/WE)/(WM/WP), the pseudo-observations are of the form $z_i = f_i(X)'(x_i - \bar{x})$ =

$$\hat{R}\left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM}\right),$$

where WCE_i, WP_i, WE_i and WM_i are similar to the above for the i^{th} BP. We approximate the appropriate weight of a block part by N_i = $(WE_i + WP_i)/2$.

If the variance specifications of the model are inaccurate so there is some heteroskedasticity, or if the distribution is very long-tailed, then there will be a long-tailed distribution of residuals, making the tests at least slightly liberal. Some care must be taken to note the presence of outliers signalling this heteroskedasticity, for example, outlying blocks due to large-scale geocoding errors.

The assumption of approximately independent observations in ANOVA may be violated in two ways. First, the PSUs are not selected by SRS but rather by a geographical stratification somewhat finer than reflected in the poststratification scheme.

To the extent that this geographical stratification reduces the sampling variance of the state effect estimates, inferences under the independence model will be somewhat conservative. Second, there will be correlations between adjustment factors for different block parts from the same block (in multi-PSG models). These will tend to make inferences assuming independence somewhat liberal. On the balance, we regard the tests performed here as useful.

References

1. Efron, B. (1982), "The Jackknife, the Bootstrap and Other Resampling Plans," SIAM, Philadelphia, Pa.

2. Fay, R.E. and Thompson, J.H. (1993), "The 1990 Post Enumeration Survey Statistical Lessons, in Hindsight," Proceedings, Bureau of the Census Annual Research Conference (in press).

3. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), "Robust Statistics: The Approach Based on Influence Functions," John Wiley & Sons, New York.

4. Hengartner, N. and Speed, T.P. (1993), "Assessing Between-Block Heterogeneity within the Poststrata of the 1990 Post

Enumeration Survey," Journal of the American Statistical Association (in press)

5. Hogan H. (1992), "The 1990 Post Enumeration Survey : An Overview," American Statistician, Vol.46, pp. 261-269.

6. Hogan H. (1993), "The 1990 Post Enumeration Survey : Operations and Results," Journal of the American Statistical Association," in press.

7. Isaki, C.T., Schultz, L.K., Diffendal, G.J. and Huang, E.T. (1988), "On Estimating Census Undercount in Small Areas," Journal of Official Statistics, Vol.4, No.2, pp. 95-112.

8. Wachter, K.W. and Freedman, D.A. (1993), "Measuring Local Homogeneity 1990 Census Data," unpublished report.

Table 1. Number of PSG's with Significant (α=.05) State Effect (Logistic Regression)

Div.	No. Grp	Alloc	Mail Ret	Mult Str	Mail Unv	Sub
1	5	5	5	5	1(1)	3(4)
2	12	11	11	12	7(10)	12
3	16	15	16	16	3(3)	12(12)
4	8	8	8	7	5(6)	5(8)
5	10	10	9	10	4(4)	7(8)
6	15	15	13	15	5(7)	15
7	9	8	9	9	4(4)	8(8)
8	7	7	7	7	2(3)	6(6)
9	17	15	14	14	5(5)	6(12)
Sum	99	94	92	95	36(43)	74(84)

The numbers in ( ) are the number of PSG's for which test statistics are available when they are less than the number of groups.

Table 2. Magnitude of State Effects with respect to Test Statistics

	Allocation Rate	Mail Return Rate	Substitution Rate
Minimum	4.3	0.28	5.46
25 %-ile	27.5	102.83	49.80
50 %-ile	68.9	254.49	97.35
75 %-ile	140.3	644.05	260.88
Maximum	945.2	8779.88	1815.12

Table 3. Variance of Undercount Rate Explained by State and PSG

Div.	No. of Groups	No. of States	SS(Group) / SS(State)	MS(Group) / MS(State)
1	5	6	4.51	5.64
2	12	3	4.88	.89
3	16	9	12.69	6.77
4	8	4	8.73	3.74
5	10	4	8.17	2.72
6	15	5	7.67	2.19
7	9	7	2.78	2.09
8	7	8	1.31	1.53
9	17	5	40.28	10.07

* States include D.C.

Table 4. Analysis of Linearized Undercount at the PSG Level

Division	Number of PSG	Number of PSG with P<.05
1	5	0
2	12	3
3	16	4
4	8	5
5	10	2
6	15	1
7	9	0
8	7	1
9	17	3
sum	99	19

Table 5. Summary of Analysis of Linearized Undercount by Place Type

Place Type	Number of PSG	Number of PSG with P<.05
0	11	3
1	23	1
2	12	1
3	8	1
4	0	0
5	6	2
6	6	1
7	11	3
8	11	4
9	10	3

Table 6. State Effects by Division - Weighted and Unweighted Data

		Unweighted Models		Weighted Models	
Division	D.F.	F	p	F	p
1	5	.57	.72	.40	.85
2	2	4.64	.01	1.72	.18
3	8	.43	.91	.65	.74
4	3	.64	.59	.66	.58
5	3	.66	.58	1.37	.25
6	4	.60	.66	.24	.92
7	6	.39	.88	.22	.97
8	7	.62	.74	.76	.62
9	4	.77	.54	.48	.75

Table 7. Correlation Coefficients between the Surrogate Variable and Undercount Rate by PSG

Variable	Correlation
Allocation Rate	.44
Mail Return Rate	-.57
Multiunit Str Rate	.39
Mail Universe Rate	.08
Substitution Rate	.47

Reference

BIOMETRICS 46, 623–635
September 1990

Logistic Regression in Capture–Recapture Models

Juha M. Alho

Institute for Environmental Studies and Department of Statistics,
University of Illinois, 1101 W. Peabody Drive, Urbana, Illinois 61801, U.S.A.

SUMMARY

The effect of population heterogeneity in capture–recapture, or dual registration, models is discussed. An estimator of the unknown population size based on a logistic regression model is introduced. The model allows different capture probabilities across individuals and across capture times. The probabilities are estimated from the observed data using conditional maximum likelihood. The resulting population estimator is shown to be consistent and asymptotically normal. A variance estimator under population heterogeneity is derived. The finite-sample properties of the estimators are studied via simulation. An application to Finnish occupational disease registration data is presented.

1. Introduction

We consider the problem of estimating the size of a closed population based on a capture and a single recapture (e.g., Seber, 1982, Chap. 3; Seber, 1986, p. 273). In demography this is known as *dual-system estimation* (e.g., Ericksen and Kadane, 1985, pp. 102–103). Several authors have studied the problems caused by heterogeneity in capture probabilities (Seber, 1982, pp. 85–88). Burnham and Overton (1978, 1979) and Otis et al. (1978) postulated a model of unobservable heterogeneity in which the individual capture probabilities are a random sample from an unknown distribution. A jackknife estimator based on several recaptures is used to estimate the population size. This work has been extended by, e.g., Pollock and Otto (1983) and Chao (1987, 1988), who study the bias, variance, and robustness of alternative estimators. Rodrigues, Bolfarine, and Leite (1988) propose a Bayesian analysis using both noninformative and informative priors. Closer to our contribution is Pollock, Hines, and Nichols (1984), who introduced a logistic regression technique to account for observable population heterogeneity in the capture probabilities. In other words, the characteristics of the captured individuals are used to explain their probabilities of capture. To avoid problems connected with the unobservable part of the likelihood (due to those members of the population that are not captured at all), they categorized the independent variables to carry out the estimation (Pollock et al., 1984, p. 332). We circumvent these problems by conditioning (cf. Sanathanan, 1972; Bishop, Fienberg, and Holland, 1975, Chap. 6; Seber, 1982, pp. 489–490). This allows us to use independent variables without any grouping. Huggins (1989) has independently suggested a similar approach to the problem.

In Section 2 we generalize the classical estimator of population size to cover, for instance, the case in which all capture probabilities are different between individuals and captures. In Section 3 we develop the conditional maximum likelihood estimation procedures and establish sufficient conditions for the strong consistency and asymptotic normality of the proposed estimator. Part of the material is somewhat technical and can be skipped by a

Key words: Asymptotic theory; Capture–recapture models; Logistic regression; Simulation.

reader interested in applications. In Section 4 we generalize the variance formula of Sekar and Deming (1949) to cover our model. The finite-sample properties of our estimators are investigated via simulation in Section 5. We show that the analysis based on logistic regression corrects for the bias caused by observable population heterogeneity. However, in very small populations the estimator becomes unstable. Finally, in Section 6 we present an application to Finnish occupational disease data.

The notation used in capture–recapture/dual registration literature is not standard (cf. Cormack, 1968, p. 457), and the existing loose conventions (e.g., in Seber, 1982) differ from those used in many other parts of statistics. An attempt is made here to formulate the results in a language familiar in the capture–recapture literature.

2. Population Heterogeneity

2.1 *Classical Model*

Suppose we sample twice from a closed population of unknown size N. Let n_1 be the number of individuals captured the first time, n_2 the number captured the second time, and m the number captured twice. Let p_1 be the probability of capture on the first occasion, p_2 the probability of capture on the second occasion, and p_{12} the probability of being captured twice. We assume that the captures are *independent*, so that $p_{12} = p_1 p_2$.

Define $u_1 = n_1 - m$, $u_2 = n_2 - m$, and $M = n_1 + n_2 - m$. Assuming that different individuals are registered independently of each other, we have a *multinomial* model (e.g., Seber, 1986, p. 274)

$$(u_1, u_2, m, N - M) \sim \text{Mult}(N; p_1(1 - p_2); (1 - p_1)p_2; p_1 p_2; 1 - \phi),$$

where $\phi = p_1 + p_2 - p_1 p_2$. The classical model is often phrased conditionally on the values of n_1 and n_2. This leads to a *hypergeometric* distribution for m, with the well-known maximum likelihood estimators

$$\hat{p}_1 = \frac{m}{n_2}, \quad \hat{p}_2 = \frac{m}{n_1}, \quad \hat{N} = \frac{n_1 n_2}{m},$$

(e.g., Feller, 1968, pp. 45–46). Sekar and Deming (1949, pp. 114–115) derived an estimator for the asymptotic variance of $\hat{N}$ in the hypergeometric setting:

$$V_1 = \frac{n_1 n_2 u_1 u_2}{m^3}.$$

We shall show that $\hat{N}$ is maximum likelihood, and derive V_1 under the multinomial model as a special case in Examples 3.1 and 4.2.

We shall now extend the model to allow for variation in the individual probabilities of registration, by treating each individual as a separate stratum. Define indicator variables u_{1i}, u_{2i}, and m_i, for $i = 1, \ldots, N$,

$$u_{ji} = \begin{cases} 1, & \text{if individual } i \text{ is captured on occasion } j \text{ only, } j = 1, 2; \\ 0, & \text{otherwise;} \end{cases}$$

$$m_i = \begin{cases} 1, & \text{if individual } i \text{ is captured twice;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $n_{ji} = u_{ji} + m_i$ $(j = 1, 2)$, $M_i = u_{1i} + u_{2i} + m_i$, and define for each individual the probabilities of being registered as $p_{ji} = \text{E}[n_{ji}]$ $(j = 1, 2)$, and $p_{12i} = \text{E}[m_i]$. Assume that these probabilities are strictly between 0 and 1. We shall complete the definition of the

model allowing for population heterogeneity, by assuming that the registers operate *independently* on the individual level, or $p_{12i} = p_{1i}p_{2i}$, and that the multinomial vectors

$$(u_{1i},\ u_{2i},\ m_i,\ 1 - M_i) \sim \text{Mult}(1;\ p_{1i}(1 - p_{2i});\ (1 - p_{1i})p_{2i};\ p_{1i}p_{2i};\ 1 - \phi_i),$$

where $\phi_i = p_{1i} + p_{2i} + p_{1i}p_{2i}$, are *independent* for $i = 1, \ldots, N$. As we shall see below, this model allows for a population-level correlation between the captures.

It is well known both empirically (Seber, 1982, p. 565) and theoretically (e.g, Burnham and Overton, 1979, Table 4, pp. 931–932) that the classical estimator may be severely biased under population heterogeneity. The expected bias can be expressed in terms of the sample covariance of the pairs $(p_{1i},\ p_{2i})$, $i = 1, \ldots, N$: $\hat{N}$ gives asymptotically an underestimate if the covariance is positive, and an overestimate if it is negative. Under zero covariance the estimator is consistent (cf. Sekar and Deming, 1949, pp. 105–106; Seber, 1982, p. 86).

2.2 *Estimation Under Observable Heterogeneity*

Suppose for the moment that we know the probabilities of being captured at least once, ϕ_i, and consider the estimator

$$\tilde{N} = \sum_{i=1}^{N} \frac{M_i}{\phi_i} = \sum_{M_i=1} \frac{1}{\phi_i}.$$

The summation on the right means summation over those indices i for which $M_i = 1$. $\tilde{N}$ is obviously an unbiased estimator of N. One can also show (using sufficiency and completeness) that it has minimum variance among such estimators. However, for our purposes it is more important to note that $\tilde{N}$ can be calculated when ϕ_i *is known only for those individuals that have been captured at least once.* This means that if we can estimate ϕ_i's from the data concerning the *captured* individuals, then we can replace $\tilde{N}$ by its estimator, and get an estimator of N (cf. Sanathanan, 1972, p. 144). This is precisely what we shall do in Section 3.

However, some regularity conditions must be imposed on the true underlying ϕ_i's to guarantee the consistency of $\tilde{N}$ in large samples. Despite the fact that $\tilde{N}$ is always unbiased, its variance may explode unless the ϕ_i's are bounded in some way. We shall now prove a technical result that gives a sufficient condition for consistency. The result will also be used in the proof of Proposition 3.2.

Proposition 2.1 Suppose there is a constant $q > 0$ such that $q \leq \phi_i$ for all $i = 1, 2, \ldots$. Then $\tilde{N}$ is strongly consistent for N,

$$\frac{\tilde{N}}{N} \to 1 \quad \text{a.s. as } N \to \infty.$$

If in addition $\phi_i \leq 1 - q$ for all $i = 1, 2, \ldots$, then $\tilde{N}$ has an asymptotic normal distribution.

Proof Since $\text{var}(M_i/\phi_i) = (1 - \phi_i)/\phi_i \leq (1 - q)/q$, the strong law of large numbers (e.g., Chung, 1974, Theorem 5.1.2, p. 103) implies that $\tilde{N}/N \to 1$ almost surely. As a sum of independent variables with variances bounded from above and away from zero, $\tilde{N}$ is asymptotically normal.

This gives a sufficient condition for the strong consistency of $\tilde{N}$. It does exclude certain practical situations from consideration, such as the case in which part of the population is effectively uncatchable, i.e., it has positive, but very small capture probabilities. [Seber (1982, p. 72) relates an example due to Ayre of an ant population for which such conditions

hold: Only a fraction of ants go foraging, for others remain in the ant hill and so are uncatchable.] Then $\hat{N}$ is still unbiased, but it may not be consistent. Furthermore, by considering, e.g., $\phi_i = i^{-\delta}$ $(0 < \delta;\ i = 1, 2, \ldots)$, we have that $\text{var}(\hat{N}/N) = O(N^{\delta-1})$. For $\delta \geq 1$ the variance does not vanish as $N \to \infty$, so we do not get even weak consistency. For $0 < \delta < 1$, the terms $\text{var}(M_i/\phi_i)i^{-2} = i^{\delta-2} - i^{-2}$ $(i = 1, 2, \ldots)$ form a convergent series, so the strong consistency follows from Kolmogorov's strong law of large numbers (Chung, 1974, Corollary, p. 125). These examples show that the existence of a bound $c > 0$ is not necessary for Proposition 2.1 to hold, but some bound for the frequency of extreme values of ϕ_i is essential.

In practice the ϕ_i's would not be known. Instead, we shall suppose that there is a parameter vector θ such that $p_{1i} = p_{1i}(\theta)$ and $p_{2i} = p_{2i}(\theta)$. We shall write $\hat{p}_{1i}$, $\hat{p}_{2i}$, $\hat{\phi}_i$, and $\hat{N}$, for p_{1i}, p_{2i}, ϕ_i, and $\hat{N}$, when θ has been estimated by a maximum likelihood estimator $\hat{\theta}$. Under population homogeneity

$$\hat{N} = \sum_{M_i=1} \frac{1}{\hat{\phi}_i} = \sum_{M_i=1} \frac{1}{\phi_i(\hat{\theta})}$$

will agree with the classical estimator so that the same symbol can be used for the estimator we have introduced.

3. Logistic Analysis

We shall now derive a conditional maximum likelihood estimator of θ under logistic regression. The iterative formulas for the solution of the likelihood equations are given below, before Example 3.1, so a reader interested in applications may want to proceed there. Assume we have vectors $\mathbf{X}_{1i} = (X_{1i1}, \ldots, X_{1ik})^\intercal$ and $\mathbf{X}_{2i} = (X_{2i1}, \ldots, X_{2ih})^\intercal$ of "explanatory" variables giving the characteristics of individual i relevant to capture, with $X_{1i1} = X_{2i1} = 1$, for $i = 1, \ldots, N$. We model the p_{ji}'s by letting $\log(p_{ji}/(1 - p_{ji})) = \mathbf{X}_{ji}^\intercal \mathbf{a}_j$ $(j = 1, 2)$, where $\mathbf{a}_1 = (a_{11}, \ldots, a_{1k})^\intercal$ and $\mathbf{a}_2 = (a_{21}, \ldots, a_{2h})^\intercal$ are vectors of parameters.

Let $\mathbf{M} = (M_1, \ldots, M_N)^\intercal$, and define $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{m}$ correspondingly. The *conditional likelihood* of $\theta = (\mathbf{a}_1^\intercal, \mathbf{a}_2^\intercal)^\intercal$, given $\mathbf{M}$, can (with some algebra) be shown to be

$$L(\theta \mid \mathbf{u}_1, \mathbf{u}_2, \mathbf{m}; \mathbf{M}) = \prod_{M_i=1} \Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1),$$

where

$$\Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1) = \frac{\exp(u_{1i}\mathbf{X}_{1i}^\intercal \mathbf{a}_1 + u_{2i}\mathbf{X}_{2i}^\intercal \mathbf{a}_2 + m_i(\mathbf{X}_{1i}^\intercal \mathbf{a}_1 + \mathbf{X}_{2i}^\intercal \mathbf{a}_2))}{K_i(\theta)},$$

with

$$K_i(\theta) = \exp(\mathbf{X}_{1i}^\intercal \mathbf{a}_1) + \exp(\mathbf{X}_{2i}^\intercal \mathbf{a}_2) + \exp(\mathbf{X}_{1i}^\intercal \mathbf{a}_1 + \mathbf{X}_{2i}^\intercal \mathbf{a}_2).$$

Consequently, we can write

$$L(\theta \mid \mathbf{u}_1, \mathbf{u}_2, \mathbf{m}; \mathbf{M}) = \exp(\mathbf{T}_1^\intercal \mathbf{a}_1 + \mathbf{T}_2^\intercal \mathbf{a}_2) \prod_{M_i=1} K_i(\theta)^{-1},$$

where

$$\mathbf{T}_j = \sum_{M_i=1} n_{ji} \mathbf{X}_{ji}, \quad j = 1, 2.$$

This shows that the conditional distribution of $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{m}$, given $\mathbf{M}$, belongs to the exponential family of degree $k + h$, and the vectors $\mathbf{T}_1$ and $\mathbf{T}_2$ form the minimal sufficient

statistic for θ (Andersen, 1980, p. 28). This is also a generalized linear model with the natural link function (Nelder and Wedderburn, 1972, p. 372; Fahrmeir and Kaufmann, 1985, p. 345). It follows (cf. Andersen, 1980, Theorem 3.1, p. 56, for the i.i.d. case) that the likelihood equations are

$$\mathrm{E}[\mathbf{T}_j \mid \mathbf{M}] = \mathbf{t}_j, \quad j = 1, 2,$$

where $\mathbf{t}_j$ is the observed value of $\mathbf{T}_j$ ($j = 1, 2$). Let us write

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix},$$

where $\mathbf{X}_j^{\mathsf{T}} = (X_{j1}, \ldots, X_{jN})$, $j = 1, 2$. Write also $\mathbf{Y} = (n_{11}, \ldots, n_{1N}, n_{21}, \ldots, n_{2N})^{\mathsf{T}}$, and note that $\mathbf{T} = \mathbf{X}^{\mathsf{T}}\mathbf{Y}$. The likelihood equations can now be written as

$$\mathbf{t} - \mathbf{X}^{\mathsf{T}}\mathrm{E}[\mathbf{Y} \mid \mathbf{M}] = \mathbf{0},$$

where for $i = 1, \ldots, N$ we have $\mathrm{E}[Y_i \mid M_i = 1] = \Pr(1, 0, 0 \mid M_i = 1) + \Pr(0, 0, 1 \mid M_i = 1)$, with $\Pr(u_{1i}, u_{2i}, m_i \mid M_i = 1)$ as above; for $i = N + 1, \ldots, 2N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = \Pr(0, 1, 0 \mid M_i = 1) + \Pr(0, 0, 1 \mid M_i = 1)$. Note that for all i, we have $\mathrm{E}[Y_i \mid M_i = 0] = 0$, so the $\mathbf{X}_{ji}$'s belonging to unobserved individuals do not enter into the likelihood equations even through they are formally included in the formulas.

To solve the equations we need $\mathrm{cov}(\mathbf{T} \mid \mathbf{M}) = \mathbf{X}^{\mathsf{T}}\mathrm{cov}(\mathbf{Y} \mid \mathbf{M})\mathbf{X}$ (cf. Andersen, 1980, proof of Lemma 3.3, p. 59, and Theorem 3.4, p. 65, for the i.i.d. case). Components of $\mathbf{Y}$ relating to different individuals are independent by assumption. However, conditionally on M_i, Y_i is dependent on Y_{N+i}, $i = 1, \ldots, N$. It follows that $\mathrm{cov}(\mathbf{Y} \mid \mathbf{M}) \equiv \mathbf{W}$ is of the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_3 \\ \mathbf{W}_4 & \mathbf{W}_2 \end{bmatrix},$$

where $\mathbf{W}_j$'s are diagonal $N \times N$ matrices. To define their elements, denote $\tilde{\mathbf{W}}_j = (\tilde{w}_{ik}^j)$ ($j = 1, \ldots, 4$; $i, k = 1, \ldots, N$), where

$$\tilde{w}_{ii}^j = \mathrm{var}(n_{ji} \mid M_i = 1) = \frac{p_{ji}}{\phi_i} - \frac{p_{ji}^2}{\phi_i^2}, \quad j = 1, 2,$$

$$\tilde{w}_{ii}^3 = \tilde{w}_{ii}^4 = \mathrm{cov}(n_{1i}, n_{2i} \mid M_i = 1) = \frac{p_{1i}p_{2i}}{\phi_i} - \frac{p_{1i}p_{2i}}{\phi_i^2}.$$

Now define the elements of $\mathbf{W}_j = (w_{ik}^j)$ by taking $w_{ik}^j = M_i M_k \tilde{w}_{ik}^j$.

With these notations Newton's method yields the recursion

$$\theta_{s+1} = \theta_s + (\mathbf{X}^{\mathsf{T}}\mathbf{W}_s\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{Y} - \mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}]), \quad s = 0, 1, \ldots,$$

where $s = 0$ corresponds to an initial value we use to start the iteration, and $\mathbf{W}_s$ and $\mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}]$ contain the estimates of $\mathbf{W}$ and $\mathrm{E}[\mathbf{Y} \mid \mathbf{M}]$ based on θ_s. Using the "working variate"

$$\mathbf{g}_s = \mathbf{X}\theta_s + \mathbf{W}_s^{-1}(\mathbf{Y} - \mathrm{E}_s[\mathbf{Y} \mid \mathbf{M}])$$

renders the recursion into the familiar regression form

$$\theta_{s+1} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}_s\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}_s\mathbf{g}_s.$$

This can be implemented by running regressions with any statistical package that allows weighting of observations in regression or that has matrix algebra.

An estimator for the covariance matrix of $\hat{\theta}$ is

$$\widehat{\text{cov}}(\hat{\theta} \mid \mathbf{M}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1},$$

where $\hat{\mathbf{W}}$ is the matrix $\mathbf{W}_s$ corresponding to $\theta_s = \hat{\theta}$.

Example 3.1 Assume $k = h = 1$, i.e., there is no population heterogeneity. Then $t_1 = n_1$ and $t_2 = n_2$. For $i = 1, \ldots, N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = p_1/\phi$, and for $i = N + 1, \ldots, 2N$, we have $\mathrm{E}[Y_i \mid M_i = 1] = p_2/\phi$. Conditionally on M, the maximum likelihood estimators of these are n_1/M and n_2/M because they satisfy the likelihood equations. Solving for p_1 and p_2 gives us $\hat{p}_1 = m/n_2$ and $\hat{p}_2 = m/n_1$ as the maximum likelihood estimators. The estimator for $1/\phi$ is $(n_1/M)(n_2/m)$, and consequently $\hat{N} = n_1 n_2/m$, or the classical estimator.

Example 3.2 Suppose $k = 2$ and $h = 1$, i.e., there is no heterogeneity at second capture time. Two of the three likelihood equations are

$$n_1 = \sum_{M_i = 1} \frac{p_{1i}}{\phi_i}, \quad n_2 = \sum_{M_i = 1} \frac{p_2}{\phi_i} = p_2 \hat{N},$$

when we write $p_{2i} = p_2$ for all i. Since $p_{1i} = (\phi_i - p_2)/(1 - p_2)$, the first equation becomes $n_1 = (M - p_2 \hat{N})/(1 - p_2)$. Solving these equations gives $\hat{p}_2 = m/n_1$ and $\hat{N} = n_1 n_2/m$. Consequently, the classical estimator $\hat{p}_2$ is conditionally maximum likelihood even when there is heterogeneity of the probabilities of n_{1i}'s. Similarly, our proposed estimator for N agrees with the classical one.

The existence, consistency, and asymptotic normality of the maximum likelihood estimator follow from the theory of generalized linear models under suitable conditions. One set of sufficient conditions is given below. The proofs can be found in the Appendix.

Proposition 3.1 Assume that the elements of $\mathbf{X}$ are bounded, and that $\mathbf{X}^T\mathbf{X}/N$ converges to a positive-definite matrix, as $N \to \infty$. Then $\hat{\theta} \to \theta$ a.s. and $\hat{\theta}$ has an asymptotic normal distribution that does not depend on $\mathbf{M}$ a.s.

Proposition 3.2 Under the conditions of Proposition 3.1, $\hat{N}$ is strongly consistent for N,

$$\frac{\hat{N}}{N} \to 1 \quad \text{a.s. as } N \to \infty,$$

and it has an asymptotic normal distribution.

An advantage of the logistic analysis over the stratified analyses proposed by Sekar and Deming (1949, pp. 106–107) is that we can use continuous explanatory variables (such as age) in our models. Second, we have the standard theory of exponential families at our disposal for inference. Even in the situation of Example 3.2 the logistic analysis may be helpful in understanding the underlying capture mechanisms. The logistic analysis provides also a simple means of estimating the distribution of explanatory variables in the population of interest, such as age distribution in a human population, or, say, the distribution of "size" (weight, length, . . .) in a fish population.

4. Variance Estimation

We shall first derive an estimator of the conditional asymptotic variance of $\hat{N}$, given $\mathbf{M}$. This estimator does not account for the variability in $\mathbf{M}$ itself. Then we present an approximation to the unconditional variance. The resulting estimator can be thought of as

a generalization of V_1 introduced by Sekar and Deming (1949). We saw in Section 3 that the conditional maximum likelihood estimators combined with the estimator $\tilde{N}$ of Section 2 result in estimators of N that are consistent and asymptotically normal. The unconditional variance estimator derived below will allow us to present *unconditional* confidence intervals for N under population heterogeneity even though a conditional likelihood was used in the estimation of $\boldsymbol{\theta}$.

Let us make the dependency of $\hat{N}$ on $\boldsymbol{\theta}$ explicit by writing $\hat{N} = \tilde{N}(\hat{\boldsymbol{\theta}}) = \tilde{N}_1(\hat{\boldsymbol{\theta}}) + \cdots + \tilde{N}_N(\hat{\boldsymbol{\theta}})$, where $\tilde{N}_i(\hat{\boldsymbol{\theta}}) = M_i/\phi_i(\hat{\boldsymbol{\theta}})$. We shall calculate the conditional asymptotic variance of $\tilde{N}(\hat{\boldsymbol{\theta}})$, given $\mathbf{M}$. For $i = 1, \ldots, N$, define the vectors

$$\mathbf{V}_i(\boldsymbol{\theta}) = \left(\frac{\partial \tilde{N}_i}{\partial \theta_1}(\boldsymbol{\theta}), \ldots, \frac{\partial \tilde{N}_i}{\partial \theta_{k+h}}(\boldsymbol{\theta})\right)^{\mathsf{T}},$$

and let $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}_1(\boldsymbol{\theta}) + \cdots + \mathbf{V}_N(\boldsymbol{\theta})$. The first-degree Taylor approximation gives the asymptotic variance of $\tilde{N}(\hat{\boldsymbol{\theta}})$, given $\mathbf{M}$, as

$$\operatorname{var}(\tilde{N}(\hat{\boldsymbol{\theta}}) \mid \mathbf{M}) = \sum_{i,j=1}^{N} \mathbf{V}_i(\boldsymbol{\theta})^{\mathsf{T}} \operatorname{cov}(\hat{\boldsymbol{\theta}} \mid \mathbf{M}) \mathbf{V}_j(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta})^{\mathsf{T}} \operatorname{cov}(\hat{\boldsymbol{\theta}} \mid \mathbf{M}) \mathbf{V}(\boldsymbol{\theta}).$$

A straightforward calculation shows that, for $i = 1, \ldots, N$, and $j = 1, \ldots, k$,

$$\frac{\partial \tilde{N}_i}{\partial \theta_j}(\boldsymbol{\theta}) = -X_{1ij}\psi_i(\boldsymbol{\theta}),$$

where $\psi_i(\boldsymbol{\theta}) = M_i \exp(\mathbf{X}_{1i}^{\mathsf{T}}\mathbf{a}_1)(1 + \exp(\mathbf{X}_{2i}^{\mathsf{T}}\mathbf{a}_2))/K_i(\boldsymbol{\theta})^2$. Similarly, for $j = k + 1, \ldots, k + h$, we have

$$\frac{\partial \tilde{N}_i}{\partial \theta_j}(\boldsymbol{\theta}) = -X_{2i,j-k}\psi_{N+i}(\boldsymbol{\theta}),$$

where $\psi_{N+i}(\boldsymbol{\theta}) = M_i \exp(\mathbf{X}_{2i}^{\mathsf{T}}\mathbf{a}_2)(1 + \exp(\mathbf{X}_{1i}^{\mathsf{T}}\mathbf{a}_1))/K_i(\boldsymbol{\theta})^2$. Let $\boldsymbol{\psi}(\boldsymbol{\theta}) = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_{2N}(\boldsymbol{\theta}))^{\mathsf{T}}$. Then we can write

$$\mathbf{V}(\boldsymbol{\theta}) = -\mathbf{X}^{\mathsf{T}}\boldsymbol{\psi}(\boldsymbol{\theta}),$$

and our formula for the estimator V_2 of the conditional asymptotic variance $\operatorname{var}(\tilde{N}(\hat{\boldsymbol{\theta}}) \mid \mathbf{M})$ becomes

$$V_2 = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}),$$

where $\hat{\mathbf{W}}$ is as in Section 3.

To show that conditioning on $\mathbf{M}$ reduces the variability of $\tilde{N}$, let us consider the case of no population heterogeneity.

Example 4.1 We saw in Example 3.1 that under population homogeneity, our point estimators coincide with the classical ones. In this case $\mathbf{X}_1 = \mathbf{X}_2 =$ vector of N ones. A direct substitution into the formula for V_2 yields after some algebra that

$$V_2 = V_1 \frac{u_1 + u_2}{M},$$

or $V_2 < V_1$.

We shall now derive an estimator for the unconditional asymptotic variance of $\hat{N}$. Conditioning on $\mathbf{M}$, we have

$$\operatorname{var}(\hat{N}) = \mathrm{E}[\operatorname{var}(\hat{N} \mid \mathbf{M})] + \operatorname{var}(\mathrm{E}[\hat{N} \mid \mathbf{M}]).$$

We estimate $E[\text{var}(\hat{N} \mid \mathbf{M})]$ by V_2. To estimate the latter term, note that

$$E[\hat{N} \mid \mathbf{M}] = \sum_{i=1}^{N} M_i E\left[\frac{1}{\hat{\phi}_i} \middle| \mathbf{M}\right].$$

We shall first approximate $E[1/\hat{\phi}_i \mid \mathbf{M}]$ by its limit value $1/\phi_i$, so

$$\text{var}(E[\hat{N} \mid \mathbf{M}]) \approx \sum_{i=1}^{N} \frac{\phi_i(1 - \phi_i)}{\phi_i^2}.$$

Estimating the first ϕ_i by M_i, and the remaining ϕ_i's by $\hat{\phi}_i$, gives us an estimator V_3, which can be written as

$$V_3 = \sum_{M_i=1} \frac{1 - \hat{\phi}_i}{\hat{\phi}_i^2}.$$

Combining the results, we get the unconditional estimator V_0 of $\text{var}(\hat{N})$ as

$$V_0 = V_2 + V_3.$$

Example 4.2 A direct calculation shows that under population homogeneity,

$$V_3 = M\left(1 - \frac{mM}{n_1 n_2}\right)\left(\frac{n_1 n_2}{mM}\right)^2 = V_1 \frac{m}{M}.$$

Together with Example 4.1, this shows that $V_0 = V_2 + V_3 = V_1$, so that our unconditional estimator agrees with the classical one under population homogeneity.

5. Finite-Sample Properties

We conducted a simulation study to see how the logistic analysis compares with the classical one in terms of bias, variance, and accuracy of confidence intervals, in small samples. First, we assumed that the probability mechanism generating population heterogeneity was correctly specified. Then the case of missing covariates was considered.

Three unknown population sizes, $N = 100$, $N = 300$, and $N = 1{,}000$, were considered. One covariate $X \sim N(0, 1)$ was used to generate the observations under two models.

Under Model I the n_{1i}'s were generated by taking $\text{logit}(p_{1i}) = .5 + .8X_i$, and the n_{2i}'s by taking $\text{logit}(p_{2i}) = 1.5 + .4X_i$, $i = 1, \ldots, N$. This implies that we have $E[n_1] = .608N$, $E[n_2] = .810N$, $E[m] = .503N$, and $E[M] = .915N$. In other words, about 92% of the unknown population are expected to be registered by at least one of the registers. The distributions of the p_{1i}'s and p_{2i}'s are slightly skewed to the left with standard deviations .18 and .07, respectively.

Under Model II we took $\text{logit}(p_{1i}) = -.5 + .8X_i$ and $\text{logit}(p_{2i}) = -1.0 + .4X_i$. Then we have $E[n_1] = .392N$, $E[n_2] = .276N$, $E[m] = .121$, and $E[M] = .547N$. This time the distributions of p_{1i}'s and p_{2i}'s are slightly skewed to the right with standard deviations .17 and .08, respectively.

Under both models the pairs (p_{1i}, p_{2i}) are essentially perfectly correlated. Based on the bias results referred to in Section 2.1, we expect the classical estimator to be about 2.5% downward biased under Model I and about 11.2% downward biased under Model II.

The simulations were carried out as follows. (1) N independent observations from $N(0, 1)$ were generated. (2) n_{1i} and n_{2i} were generated independently under Model I for $i = 1, \ldots, N$. (3) The classical estimator for N and its variance were calculated. (4) Newton's method was used to estimate the logistic regression models, as outlined in Section 3, and an unconditional variance estimator was calculated using formulas of Section 4.

Steps (1)–(4) were repeated 600 times for $N = 100$, $N = 300$, and $N = 1{,}000$. After that the same was done with Model II. The calculations were carried out using MINITAB on a personal computer.

Table 1 presents results from these simulations. Under both Models I and II the classical estimator $\hat{N}$ is downward biased but slightly less so than the first-order asymptotics would imply. Its standard deviation is adequately estimated by $\sqrt{V_1}$, and plots (not shown here) indicate that its distribution is very close to normal. Due to the bias, the purported "95% confidence intervals" for N do not come close to reaching the nominal level of coverage.

In contrast, the "95% confidence intervals" based on the *logistic estimator*, denoted by $\hat{N}'$, are very nearly adequate for both models and all three values of N. However, the way this is accomplished needs a closer look.

The logistic analysis corrects for the bias of $\hat{N}$. We pay for this in the increased variance of the estimator. Under Model I, the classical estimator $\hat{N}$ and the proposed logistic estimator $\hat{N}'$ have approximately the same mean squared errors (MSE) for $N = 300$. For $N = 100$ the MSE of $\hat{N}'$ is larger, and for $N = 1{,}000$ it is smaller than that for $\hat{N}$. For Model II the break-even point of MSEs is further between $N = 300$ and $N = 1{,}000$. A comparison of the mean and the median of $\hat{N}'$ for $N = 100$ under Model II indicates that its distribution is highly skewed to the right. As N increases, the skewness decreases.

The results of Table 1 do not give a direct indication of how well the *method* based on logistic regression might perform in a real situation, in which one would test whether the coefficients a_{12} and a_{22} vanish. If either one would be deemed not to be significantly different from 0, then, in view of Example 3.2, a classical analysis could be performed. Any reasonable testing strategies should give results that fall between those obtained using exclusively $\hat{N}$ or $\hat{N}'$, both in terms of bias and the coverage of confidence intervals.

Additional simulations (600 repetitions) were carried out with Model II and $N = 1{,}000$ to study the effect of model misspecification. The situation was modified by replacing X by $(X' + X'')/\sqrt{2}$, where X' and X'' are independent $N(0, 1)$ variables, and by assuming that only X' was observable. From the point of view of classical estimation this situation

Table 1
Small-sample properties of the classical population size estimator $\hat{N}$ and the estimator applying logistic regression $\hat{N}'$, based on 600 simulations of Models I and II for $N = 100, 300, 1{,}000$.

		Model I		Model II	
	N	$\hat{N}$	$\hat{N}'$	$\hat{N}$	$\hat{N}'$
Average of estimates divided by N	100	.981	1.011	.931	1.434
	300	.980	1.004	.898	1.044
	1,000	.980	1.002	.894	1.010
Median of estimates divided by N	100	.982	1.002	.896	1.021
	300	.980	1.003	.887	1.006
	1,000	.981	1.001	.891	.999
Standard deviation of estimator divided by N	100	.037	.063	.207	2.723
	300	.022	.030	.096	.191
	1.000	.012	.016	.053	.088
Average of estimated standard deviations of estimator divided by N	100	.035	.058	.185	1.082
	300	.020	.031	.095	.183
	1,000	.011	.016	.051	.087
Coverage probability of "95% confidence intervals"	100	.845	.933	.802	.933
	300	.798	.953	.703	.945
	1.000	.582	.975	.432	.955

is identical to Model II. However, the proposed logistic regression procedure performed worse than under Model II, since only 50% of the population heterogeneity was observable: average/1,000 = .948; median/1,000 = .943; standard deviation/1,000 = .071; average of estimated standard deviations/1,000 = .067; coverage of "95% confidence intervals" = .802.

A further effect of model misspecification is that the parameter estimates become biased. Under the original Model II we had $a_{12} = .8$ and $a_{22} = .4$. When the model was correctly specified, the average of $\hat{a}_{12}$'s was .809, and the average of $\hat{a}_{22}$'s was .399. Under the modified model we had $a_{12} = .8/\sqrt{2} = .566$, and $a_{22} = .4/\sqrt{2} = .283$. In this case the average of $\hat{a}_{12}$'s was .524, and the average of $\hat{a}_{22}$'s was .265. Their empirical standard errors were .005 and .004, respectively. This bias appears analogous to the effect of "errors in explanatory variables" in ordinary regression.

So far, all our simulations have been concerned with the case of positive correlation between the captures. Since a negative correlation is also a possibility, we generated additional data in the set-up where $N = 1{,}000$ and Model II has been modified to have $\text{logit}(p_{2i}) = 1.0 - .4X_i$. In other words, the sign of the coefficient of X has been reversed. Based on Section 2.1, the classical estimator was expected to be about 14% *upward* biased. In 600 simulations this turned out to be the case. The logistic estimator was again nearly unbiased in terms of both mean and median. The coverage probabilities of the "95% confidence intervals" were .652 for the classical method and .948 for the proposed method. Interestingly, in this case $\sqrt{\text{var}(\hat{N})} = 86.7 > \sqrt{\text{var}(\hat{N}')} = 74.7$. The estimators of both variances were slightly downward biased. In this case both the bias *and* the variance of the classical estimator are larger than those of the logistic estimator.

6. An Application to Occupational Disease Registration

We illustrate the logistic analysis by an application to occupational disease data from Finland in 1981. The Finnish Register of Occupational Diseases was founded in 1964. At that time it was agreed that accident insurance companies would report all new cases of occupational disease to the Register, irrespective of the compensation decision.

From 1975 every physician has been required to report all new cases of occupational disease directly to a government agency, which reports the cases to the Register. Despite the legal obligation, physicians neglect to report a large number of cases, presumably because such a report causes paperwork, but does not directly benefit the patient. Unfortunately, all cases are not reported via the other channel either.

The Finnish Register of Occupational Diseases can, thus, be thought of as a dual registration system. However, the probabilities of registration are not constant over different types of cases, nor are the two information channels independent (on a population level). Alho (paper presented at 11th Nordic Conference on Mathematical Statistics, Uppsala, 1986) has shown previously that diagnosis has a strong impact on the probabilities of registration. Using data from 1981, we found $M = 5{,}231$. Stratifying the data into four groups of diagnoses, (1) noise-induced hearing loss, (2) diseases of the musculo-skeletal system caused by repetitive or monotonous work, (3) skin diseases, and (4) other diseases, yielded $\hat{N} = 8{,}258$, as opposed to $\hat{N} = 7{,}232$ obtained without stratification.

Using the logistic techniques, we can check whether these findings hold when we control for the possible effect of age. Age at which a disease is diagnosed may be correlated with the severity of the case, making cases of older workers more likely to be reported through either channel. On the other hand, one may argue that older workers are less likely to have their diseases diagnosed as occupational ones due to fear of losing one's job at an advanced age. A logistic analysis with age as an explanatory variable might, thus, reveal additional population heterogeneity within the groups of diagnoses. This turned out to be the case for

noise-induced hearing loss and the "other" category, but in both cases the heterogeneity pertained to only one channel, so it did not influence $\hat{N}$.

As an example, let us look at noise-induced hearing loss, which had $M = 1{,}854$ in 1981. Let p_1 = probability that a case of noise-induced hearing loss is reported to the Register from insurance companies, and p_2 = the corresponding probability for the other channel. Let X = age. We estimated the model $\text{logit}(p_1) = a_{11} + a_{12}X$, and $\text{logit}(p_2) = a_{21} + a_{22}X$, but found $\hat{a}_{22}$ not to be significant at 5% level. Taking $a_{22} = 0$ gave $\hat{a}_{11} = -1.543$, $\hat{a}_{12} = .0438$ (estimated standard error = .0020), and $\hat{a}_{21} = .0409$. The range of p_{1i}'s was from .359 (corresponding to $X = 22$) to .881 (for $X = 81$), with a roughly normal distribution with mean .663 and standard deviation .0870. It is likely that the cases of noise-induced hearing loss that are diagnosed at a late age are more severe, due to a longer exposure time, than the ones diagnosed at an early age. Consequently, the likelihood of positive compensation decision probably increases with age.

In this case the classical analysis and the logistic analysis give $\tilde{N} = 2{,}218$, with $\sqrt{V_1} = 33.3$.

7. Discussion

We have introduced a conditional logistic estimation procedure that allows us to analyze capture–recapture data using individual-level covariate information. It is worth noting that the distributions of the covariates typically are *not* the same in the observed and unobserved segments of the population (cf. Cormack, 1989, p. 412). Our work can be viewed as an extension of the classical procedures and a model introduced by Pollock et al. (1984). Both asymptotic and finite-sample properties of the proposed estimator have been studied. This indicates that the model may be of wide use when the required covariate information exists. The model can be generalized to a multiple-recapture situation. It may be possible to formulate the problem in terms of a nonparametric logistic regression model. Perhaps a Bayesian approach could be used to reduce the instability of the estimator in small samples, if suitable prior information exists.

ACKNOWLEDGEMENTS

The author would like to thank D. Simpson for comments on an early version of the paper. The helpful suggestions of an associate editor and the referees are also gratefully acknowledged.

RÉSUMÉ

L'effet d'une hétérogénéité de la population dans les modèles de capture–recapture on de double enregistrement est discuté. Un estimateur de l'effectif inconnu de la population basé sur un modèle de regression logistique est obtenu. Le modèle permet des probabilités de capture différentes entre individus et entre dates de capture. Les probabilités sont estimées à partir des données observées en utilisant le maximum de vraisemblance conditionnel. On montre que l'estimateur résultant de l'effectif de la population est convergent et asymptotiquement normal. Un estimateur de la variance sous l'hypothèse d'hétérogénéité de la population est obtenu. Les propriétés des estimateurs pour des tailles d'échantillon finies sont étudiées par simulation. Une application à des données d'enregistrement de maladies professionnelles en Finlande est présentée.

REFERENCES

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, Massachusetts: MIT Press.

Burnham, P. K. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**, 625–633.
Burnham, P. K. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**, 927–936.
Chao, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43**, 783–791.
Chao, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management* **52**, 295–300.
Chung, K. L. (1974). *A Course in Probability Theory*, 2nd edition. New York: Academic Press.
Cormack, R. M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology Annual Review* **6**, 455–506.
Cormack, R. M. (1989). Log-linear models for capture–recapture. *Biometrics* **45**, 395–413.
Ericksen, E. P. and Kadane, J. B. (1985). Estimating the population in a census year. *Journal of the American Statistical Association* **80**, 98–109.
Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* **13**, 342–368.
Feller, W. (1968). *An Introduction to Probability Theory and Its Applications. Vol.* I, 3rd edition. New York: Wiley.
Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.
Neider, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* No. 62.
Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). The use of auxiliary variables in capture–recapture and removal experiments. *Biometrics* **40**, 329–340.
Pollock, K. H. and Otto, M. C. (1983). Robust estimation of population size in closed animal populations from capture–recapture experiments. *Biometrics* **39**, 1035–1049.
Rodrigues, J., Bolfarine, H., and Leite, J. G. (1988). A Bayesian analysis in closed animal populations from capture–recapture experiments with trap response. *Communications in Statistics—Simulation and Computation* **17**, 407–430.
Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics* **43**, 142–152.
Seber, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd edition. New York: Griffin.
Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42**, 267–292.
Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**, 101–115.

Received November 1987; *revised December* 1988 *and August* 1989; *accepted September* 1989.

APPENDIX

A.1 Proof of Proposition 3.1

The boundedness of the elements of $\mathbf{X}$ implies that $M \to \infty$ a.s. as $N \to \infty$. Moreover, under our hypothesis the conditions of Fahrmeir and Kaufmann's (1985, p. 355) Corollary 1 are easily satisfied. It follows that $\hat{\boldsymbol{\theta}}$ is strongly consistent for $\boldsymbol{\theta}$ and has an asymptotic normal distribution, conditionally on $\mathbf{M}$. The unconditional strong consistency follows from $\Pr(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > c \text{ i.o.}) = \mathrm{E}[\Pr(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > c \text{ i.o.} \mid \mathbf{M})] = 0$, where "i.o." means infinitely often in N (cf. Chung, 1974, Theorem 4.2.2, p. 73). Consider the normality now.

The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is given by $(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}/N)^{-1}$. We shall show that $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}/N$ converges a.s. to a matrix (say) $\boldsymbol{\Sigma}$ that does not depend on $\mathbf{M}$. Partition $\boldsymbol{\Sigma}$ into four submatrices corresponding to those of $\mathbf{W}$. For instance, take the upper left-hand corner to be the limit of $\mathbf{X}^{\mathrm{T}}\mathbf{W}_1\mathbf{X}/N$. Note that the (i, j) element of this can be written as

$$\frac{1}{N}\sum_{k=1}^{N} M_k \bar{w}_{kk} X_{1ki} X_{1kj}.$$

By the boundedness of the summands this converges to its asymptotic expectation a.s. Hence, it does not depend on the particular realization $\mathbf{M}$ a.s. For the other elements we reason analogously. It follows that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$ asymptotically, conditionally on $\mathbf{M}$.

We can uncondition by noting that the normality of $\sqrt{N}(\hat{\theta} - \theta)$ is equivalent to the condition that for every vector $\nu \neq 0$, $\sqrt{N}\nu^T(\hat{\theta} - \theta) \sim N(0, \nu^T \Sigma^{-1} \nu)$ asymptotically. This, in turn, is the same as $E[f(\sqrt{N}\nu^T(\hat{\theta} - \theta))| M] \to E[f(\nu^T \xi)]$ as $N \to \infty$ for $\xi \sim N(0, \Sigma^{-1})$ and any continuous function f that vanishes outside a compact set (Chung, 1974, Theorem 4.4.1, p. 87). Using the dominated convergence theorem, we get that

$$E[f(\sqrt{N}\nu^T(\hat{\theta} - \theta))] \to E[f(\nu^T \xi)] \quad \text{as } N \to \infty.$$

This proves the unconditional asymptotic normality.

A.2 Proof of Proposition 3.2

The boundedness of the elements of X implies that ϕ_i's are bounded away from both 0 and 1. Hence, by Proposition 2.1, $(\hat{N} - N)/N \to 0$ a.s. Another consequence is that there is a neighborhood U of the true value θ such that the first derivatives S_i, and the second derivatives H_i, of $1/\phi_i$, are bounded in it. Write the Taylor series expansion

$$\frac{\hat{N} - \tilde{N}}{\sqrt{N}} = R_1 + R_2,$$

where

$$R_1 = \left(\frac{1}{N} \sum_{i=1}^{N} M_i S_i(\theta)^T\right)\sqrt{N}(\hat{\theta} - \theta),$$

and

$$R_2 = \frac{1}{2}\sqrt{N}(\hat{\theta} - \theta)^T\left(\frac{1}{N^{3/2}} \sum_{i=1}^{N} M_i H_i(\eta)\right)\sqrt{N}(\hat{\theta} - \theta),$$

where η is between $\hat{\theta}$ and θ. The boundedness of H_i's ensures that $R_2 \to 0$ a.s. as $N \to \infty$. The sum in R_1 converges to its asymptotic expectation a.s. It follows from Proposition 3.1 that $R_1/\sqrt{N} \to 0$ a.s., so that $\hat{N}$ is strongly consistent. Second, R_1 has an asymptotic normal distribution that does not depend on M a.s. It follows that $(\hat{N} - N)/\sqrt{N}$, which is a function of M, and $(\hat{N} - \tilde{N})/\sqrt{N}$ are asymptotically independent. As a sum of two asymptotically independent normal variables $(\hat{N} - N)/\sqrt{N}$ is asymptotically normal.

確率化應答技法
(RANDOMIZED RESPONSE TECHNIQUE)

류 제 복(청주대)
홍 기 학(동신대)
이 기 성(동국대)

目　　次

확률화응답기법
(Randomized Response Technique)

1. 서론

사회 여러분야의 조사에서 최근 연구의 관심은 응답을 회피하거나 고의적인 거짓응답으로 인한 비표본오차를 줄이는 데 있다. 이러한 오차는 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 증가하게 된다. 이에 Warner (1965)는 확률장치를 이용하여 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대해 정보를 이끌어낼 수 있는 확률화응답기법(randomized response technique:RRT)을 처음으로 제시하였다. 응답자의 신분이나 비밀을 보장함으로써 응답자로부터 민감한 질문에 보다 더 정확한 정보를 얻을 수 있는 확률화응답기법은 많은 학자들에 의해 연구, 발전되어 왔다. 특히, Abul-Ela et al.(1967)등은 이지모집단(dichotomous population)에 대한 Warner의 관련질문기법을 다지모집단(polychotomous population)의 경우로 확장하였고, Greenberg et al.(1969)등은 무관질문기법의 이론적 체계를 완성하였으며, Moors(1971)와 Folsom et al.(1973) 등은 이를 개선, 보완하였다. 또한 Drane(1976)은 강요질문기법을 제시하였으며, Liu 와 Chow(1976)는 Warner 기법의 확장으로써 반복시행기법을 제시하였고, 아울러 새로운 양적확률화응답기법(quantitative randomized response technique)도 제시 하고 있다. 그리고 Chaudhuri 와 Mukerjee(1988)는 확률화응답기법에 대한 이론을 정리하여 체계화시켰으며, Mangat 와 Singh(1990)는 2개의 확률장치를 이용하는 2단계 확률화응답기법을 제시하고 있다.

본 논문에서는 내용을 모두 서론을 포함해서 5장으로 구성하여 제 2장에서 Warner류의 관련질문기법과 제3장에서 Greenberg et al.류의 무관질문기법을 비교적 자세히 살펴보고 제4장에서는 확률화응답기법의 여러가지 활용 방안들에 대하여 논하였으며, 끝으로 제5장에서는 확률화응답기법의 실제 적용 사례들을 몇가지 살펴보았다.

2. 관련질문기법(The Related Technique)

사회여러분야의 조사에서 개인 신상에 관계되는 조사에 응답자들이 남들에게 숨기고 싶어하는 사항들이 있게 된다. 대부분의 자료는 질문을 함으로써 어렵지 않게 얻을 수 있으나 사람들이 응답을 회피하거나 정직하게 응답하지 않는 일부분의 민감한 질문이 있게 된다. 예를 들면, 개인소득, 재산상태, 불로소득, 탈세여부, 전과경험, 알콜중독, 환각제사용, 낙태경험, 동성연애등과 같은 통상적으로 사회에서 인정하지 않는 사항들이다.

이러한 민감한 질문에 공개적이거나 직접(open or direct)질문을 하게 되면 신뢰할 수 있는 자료를 얻기가 힘들게 된다. 이는 무응답이나 거짓응답 또는 응답회피로부터 생기는 편의(bias)때문에 응답자들로부터 정확한 자료를 얻을 수 없기 때문에 표본자료를 가지고 모집단에 대한 올바르고 공정한 판단(결론)을 할 수 없게 된다. 따라서 신상에 관한 문제(민감한 질문)에 보다 신뢰할 수 있는 정보(자료)를 얻기 위해서는 직접질문보다는 간접적인 대체 질문방식이 필요하게 된다. 이에 Warner(1965)는 확률화응답기법(randomized response technique : RRT)을 제시하였다. 이 기법은 조사과정에 확률장치를 사용해서 응답자의 신분이 노출되지 않게 함으로써 사생활이나 비밀을 보장하여주어 응답자들이 진실되게 응답하게 유도하여 신뢰할 수 있는 정보를 얻을 수 있는 방법이다.

2.1 Warner기법

모집단의 구성원이 두 그룹(민감한 그룹 A와 민감하지 않은 그룹 A^c)중 하나에 속해 있는 이지모집단(dichotomous population)을 가정한다.

표본조사에 의하여 그룹 A에 속해 있는 모집단비율 π ($0 < \pi < 1$)를 추정하기 위하여 단순임의복원추출(SRSWR)에 의하여 n명의 표본을 뽑는다. 이들에게 직접질문을 하게 되면 응답자들이 응답을 회피하거나 거짓응답을 하게되므로 추정량의 효율이 떨어지게 된다.

이러한 점을 해결하기 위하여 Warner(1965)는 그림 2.1과 같은 돌림 판을 확률장치로 사용하여 응답자 자신들이 돌림 판을 돌려서 지침이 가르키는 곳에 해당

하는 질문에 대한 답을 단순히 "예" 또는 "아니오"로만 응답하게 하였다. 이러한 확률장치에 의해서 선택된 질문은 조사자는 알 수가 없고 단지 응답자만이 알 수 있게 함으로써 응답자의 신분을 보장해줄 수 있게 된다.

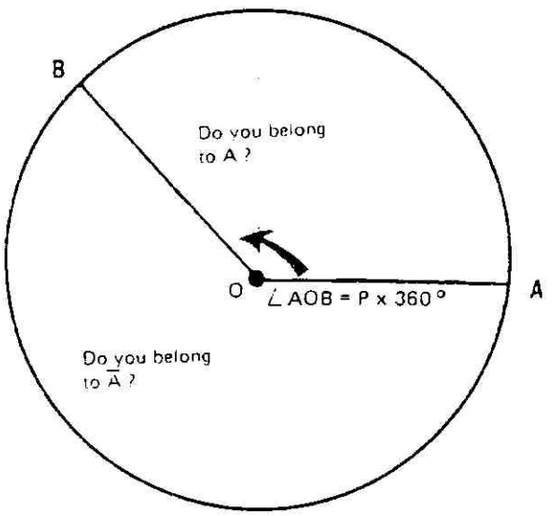


그림 2.1 확률장치

여기서 사용한 확률장치 대신에 주사위나 카드 등을 확률장치로 사용할 수 있다. 한편 질문이 선택될 확률 p는 조사자가 확률장치에서 사전에 조정할 수 있다.

확률장치를 사용하는 확률화응답기법에 의해서 얻은 자료로부터 모집단 비율 π를 추정한다. 이 때 응답자들의 응답은 진실되게 응답했다는 가정에서, 응답자들이 "예"라고 응답할 확률은 다음과 같다.

$$\begin{aligned} \lambda &= \pi p + (1-\pi)(1-p) \\ &= (1-p) + (2p-1)\pi \end{aligned} \tag{2.1}$$

n명의 표본중에서 "예"라고 응답한 사람의 수를 n_1이라 하면 n_1은 $b(n,\lambda)$를 하므로 우도함수(likelihood function)는

$$L = [\pi p + (1-\pi)(1-p)]^{n_1}[(1-\pi)p + \pi(1-p)]^{n-n_1}$$

가 된다. 이로부터 π의 추정량과 추정량의 분산을 구하면 다음과 같다.

$$\hat{\pi}_w = \frac{p - 1}{2p - 1} + \frac{n_1}{(2p - 1)n}, \quad p \neq \frac{1}{2} \tag{2.2}$$

$$\begin{aligned} Var(\hat{\pi}_w) &= \frac{\lambda(1 - \lambda)}{n(2p - 1)^2} \\ &= \frac{\pi(1-\pi)}{n} + \frac{1}{n}\left[\frac{1}{16(p-0.5)^2} - \frac{1}{4}\right] \end{aligned} \tag{2.3}$$

이 때 $\hat{\pi}_w$는 π의 불편추정량(unbiased estimator : UE)이 된다.

한편 p가 1이거나 0인 경우는 직접질문이 되겠고 이 때의 추정량과 추정량의 분산은 각각 $\hat{\pi}_d = \frac{n_1}{n}$ 과 $Var(\hat{\pi}_d) = \frac{\pi(1 - \pi)}{n}$ 가 된다.

따라서, 확률화응답기법에 의해서 얻어진 추정량의 분산식(2.3)은 직접질문으로부터의 분산에 확률장치를 사용함으로써 생기는 분산의 합으로 되어 있다. 식(2.3)은 미지모수 λ를 포함하고 있으므로 분산의 추정량를 구하려면

$$E(n_1) = n\lambda, \quad E(n_1^2) = Var(n_1) + (E(n_1))^2 = n\lambda + n(n-1)\lambda^2$$

이므로

$$\begin{aligned} E\left[\frac{\hat{\lambda}(1 - \hat{\lambda})}{n-1}\right] &= E\left[\frac{nn_1 - n_1^2}{n^2(n-1)}\right] \\ &= \frac{\lambda(1-\lambda)}{n} \end{aligned}$$

이고 $Var(\hat{\pi}_w)$의 불편추정량은 다음과 같게 된다.

$$\hat{Var}(\hat{\pi}_w) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(2p-1)^2}$$
$$= \frac{\hat{\pi}_w(1-\hat{\pi}_w)}{(n-1)} + \frac{1}{(n-1)}\left[\frac{1}{16(p-0.5)^2} + \frac{1}{4}\right] \quad (2.4)$$

$Var(\hat{\pi}_w)$는 p의 함수로서 p가 0이나 1에 가까와지면 추정량의 효율이 상당히 좋아지는 반면에 응답자의 신분보호가 충분치 못하게 되고, p가 1/2에 가까와지면 응답자들의 신분은 충분히 보장할 수는 있으나 분산이 급격히 커지게 된다. p = 1/2 이면 우도함수가 π에 의존하지 않게 되므로 π를 추정할 수 없게 된다. 그러므로 p의 값에 따라 응답자와 조사자간의 협동의 정도가 다르게 되므로 RRT에서는 신분보호와 효율의 문제가 상존하게 된다. 한편 주어진 정도(precision)하에서 표본크기의 문제는 p가 0 또는 1에 가까와지면 p가 1/2에 가까운 경우보다 표본크기가 작아진다. 즉, p가 1/2에 가까와지면 조사로부터 적은 정보를 얻게 되므로 표본수가 커져야만 주어진 정도를 유지할 수 있다. 예를 들면, π = 0.5 이고 표준편차가 0.05일 때 p = 0.75이면 n = 400이 필요하게 되고 p = 1이면 n = 100으로 충분하게 된다. 물론 이는 응답자들이 모두 진실되게 응답한 것으로 가정한 경우이다.

2.2 Warner의 확률화응답기법의 효과

Warner가 제시한 확률화응답기법은 응답자들의 신분을 보장해 줄 수 있는 반면에 직접질문은 이를 보장할 수 없다. 모집단 비율 π에 대한 추정량은

$$\hat{\pi} = \sum_{i=1}^{n} \frac{Y_i}{n} \quad (2.5)$$

이 된다. 이때 응답자들이 그룹 A 또는 그룹 A^c에 속한다고 응답하면 Y_i = 1 또는 0이 된다. 그러나 직접질문을 할 경우, 그룹 A와 그룹 A^c에 속해 있는 사람

들이 진실되게 응답할 확률을 각각 T_a와 T_b로 가정하면

$$E(\hat{\pi}) = \pi T_a + (1-\pi)(1-T_b)$$

가 되므로 $\hat{\pi}$의 편의와 분산은 다음과 같게 된다.

$$B(\hat{\pi}) = \pi(T_a + T_b - 2) + (1 - T_b) \tag{2.6}$$

$$Var(\hat{\pi}) = \frac{\{\pi T_a + (1-\pi)(1-T_b)\}\{(1-\pi T_a) - (1-\pi)(1-T_b)\}}{n} \tag{2.7}$$

Warner는 직접질문에 대한 확률화응답기법의 효율을 평균제곱오차(mean square error : MSE)로 비교하였는데 T_a와 T_b가 작을수록, p가 증가할수록 확률화응답기법의 효율이 증가하고 또한 표본수 n이 증가할수록 확률화응답기법이 선호됨을 보여주고 있다.

Moors(1985)는 식(2.1)로부터

$$1-p \le \lambda \le p\ ,\ p > \frac{1}{2} \text{ 또는 } p \le \lambda \le 1-p\ ,\ p < \frac{1}{2}$$

가 되어 λ의 가용 구간이 [0,1]이 되지 않으므로 Warner의 추정량이 불편추정량이 아님을 주장하였다. 그리고, Raghavarao(1978)은 Warner(1965)나 Singh (1976)의 추정량이 $n_1 / n \notin [p^c, p]$ 이면 좋은 추정량이 되지 않으므로 (특히 Warner의 경우는 [0,1]을 벗어남) 다음과 같은 축소된 형태(shrinkage-type)의 추정량을 제시하였다.

$$\hat{\pi}_r = \frac{\{\exp(\frac{25\,\hat{\pi}_w}{6} - \frac{25}{12})\}}{\{1 + \exp(\frac{25\,\hat{\pi}_w}{6} - \frac{25}{12})\}} \tag{2.8}$$

한편 Flinger et al.(1977)과 Devore(1977)은 Warner의 추정량이 모비율 π에 대한 최우추정량(maximum likelihood estimator : MLE)이 아님을 입증하였다.

이는 모수에 대한 MLE가 모수공간내의 원소이어야하나 $\hat{\pi}_w$가 모수공간 [0,1]에 반드시 속하지는 않는다. 따라서 다음과 같이 수정된 MLE를 제시하였다.

$$\tilde{\pi}_w = \begin{cases} 1 & \hat{\lambda} \geq p > 1/2 \\ 0 & \hat{\lambda} \leq (1-p) < 1/2 \\ \hat{\pi}_w & 1-p < \hat{\lambda} < p \,,\ p > 1/2 \\ \hat{\pi}_w & p < \hat{\lambda} < 1-p \,,\ p > 1/2 \\ 1 & \hat{\lambda} \leq p < 1/2 \\ 0 & \hat{\lambda} \geq (1-p) > 1/2 \end{cases} \tag{2.9}$$

여기서 $\hat{\lambda} = \frac{n_1}{n}$도 λ의 MLE가 아니다. 실제 MLE $\hat{\lambda}$ 도 위와같이 절단(truncation)에 의하여 얻는다. 또한 이러한 MLE는 UE가 아니기 때문에 오차의 측도로서 다음과 같은 π에 대한 MSE를 사용해야한다.

$$| \tilde{\pi}_w - \pi | < | \hat{\pi}_w - \pi |$$

이므로

$$MSE(\tilde{\pi}_w) < MSE(\hat{\pi}_w) = Var(\hat{\pi}_w)$$

가 된다.

Devore(1977)는 Warner의 RRT를 약간 변형시켜서 π의 MLE인 UE를 얻는 과정을 다음과 같이 제시하였다.

n 명의 표본(짝수로 가정)을 뽑아 n개 카드로 구성된 묶음으로부터 비복원으로 각각 하나의 카드를 뽑도록 한 다음 그 중 반수 $(n/2)$에게 무조건 "예"라고 응답하게 하고 나머지 반에게는 특성 A를 갖고 있는 지의 여부를 진실되게 "예" 또는 "아니오"라고 응답하게 한다. 이 때 X를 "예"라고 응답한 사람수라 하면 $(X - \frac{n}{2})$의 응답은 서로 독립적이 되고 $Y = (X - \frac{n}{2})$는

$b(\frac{n}{2}, \pi)$분포를 하므로

$$\hat{\pi} = \frac{2Y}{n} = \frac{2(X - \frac{n}{2})}{n} = \frac{2X}{n} - 1 \qquad (2.10)$$

이 된다. [0, 1]구간의 값을 갖는 $\hat{\pi}$는 π에 대한 UE이고 MLE가 되며 분산과 이의 불편추정량은 다음과 같이 된다.

$$Var(\hat{\pi}) = \frac{2}{n}\pi(1-\pi)$$
$$Var(\hat{\pi}) = \frac{2}{n-2}\hat{\pi}(1-\hat{\pi}) \qquad (2.11)$$

Warner는 표본을 추출하고 질문을 선택할 때 SRSWR을 사용하였다. 그러나 일반적으로 조사에서는 단순임의비복원추출(SRSWOR)이 요구된다. Kim과 Flueck(1978a)는 Warner 형태를 변형시킨 다음과 같은 4가지 형태의 표본추출과정을 사용하였다.

I. 응답자와 설문들을 복원(WR)으로 추출

II. 응답자는 비복원(WOR)으로 설문들은 복원으로 추출

III. 응답자는 복원으로 설문들은 비복원으로 추출

IV. 응답자와 설문들을 비복원으로 추출

여기서, 유한모집단 크기는 N이고 설문은 M이다. M개의 설문중에서 민감한 속성을 갖는 설문은 B개로 가정한다.($B / M = p$)

X_i 는 i번째 응답자가 민감한 그룹에 속하면 1, 아니면 0의 값을 갖고 Y_i 는 i번째 응답자가 민감한 속성을 갖는 설문을 선택하면 1, 아니면 0의 값을 갖으며 Z_i 는 i번째 응답자가 "예"라고 응답하면 1, "아니오"라고 응답하면 0의 값을 갖는 확률변수이고 $X = \sum X_i$, $Y = \sum Y_i$, $Z = \sum Z_i$ 라 정의하면

$$Z_i = X_iY_i + (1-X_i)(1-Y_i) = 2X_iY_i - X_i - Y_i + 1$$

$$Var(Z) = E(\sum Z_i^2) + E(\sum_{i \neq j} Z_iZ_j) - \{E(\sum Z_i)\}^2$$

이 된다.

확률변수 X_i와 Y_i는 추출방법이 복원인가 비복원인가에 따라 이항분포와 초기하분포를 하게 되므로 위 네가지 방법에 따른 $\hat{\pi}$의 분산은 $p \neq \frac{1}{2}$에서 각각 다음과 같이 된다.

$$V_1(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}$$

$$V_2(\hat{\pi}) = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right) + \frac{p(1-p)}{n(2p-1)^2}$$

$$V_3(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{4\pi(1-\pi)p(1-p)}{n(2p-1)^2}\left(1-\frac{M-n}{M-1}\right) + \frac{p(1-p)}{n(2p-1)^2}\left(\frac{M-n}{M-1}\right)$$

$$V_4(\hat{\pi}) = \frac{\pi(1-\pi)}{n(2p-1)^2}\left[(4p^2-4p)\left\{\frac{N(M-n)-n(M-1)}{(N-1)(M-1)}\right\} + \frac{N-n}{N-1}\right] + \frac{p(1-p)}{n(2p-1)^2}\left(\frac{M-n}{M-1}\right)$$

위 식에서 $N \to \infty$ 이면 $V_4(\hat{\pi}) \to V_3(\hat{\pi})$가 되고 마찬가지로 M이 증가하면 $V_4(\hat{\pi}) \to V_2(\hat{\pi})$에 접근해 간다. 또한 위의 4가지 방법에 대한 비교는

1. $V_1(\hat{\pi}) > V_i(\hat{\pi})$, $i = 2, 3, 4$
2. $V_2(\hat{\pi}) \geq V_4(\hat{\pi})$, $4\pi(1-\pi) \leq \frac{N-1}{N}$
3. $V_3(\hat{\pi}) \geq V_4(\hat{\pi})$, $4p(1-p) \leq \frac{M-1}{M}$

이 되므로 방법 III과 방법 IV의 분산이 방법 I과 방법 II의 분산보다 상당히 작게 된다. 그러나 π가 1/2에 가까와지면 분산의 감소도 작아진다. 그리고 Greenberg et al.(1969)의 무관질문기법에 비복원추출방법을 적용한 결과도 Warner의 경우와 유사하게 방법 IV(WOR인 경우)의 분산이 가장 작게 됨을 보여 주었다.

한편 Abul-Ela et al.(1967)은 SRSWR로뽑힌 여러개의 표본을 이용하여 이지모집단(dichotomous population)만을 다룬 Warner기법을 다지모집단(polychotomous population)의 경우로 확장시켰고, Kim과 Fleuck(1978b) Bourke(1978,1984)등은 SRSWR로뽑힌 하나의 표본에서 응답자로 하여금 다지응답(polychotimous response)을 허용함으로써 다지모집단의 민감한 모비율 추정으로 확장 시켰다.

3. 무관질문기법(The Unrelated-Question Technique)

이 장에서도 앞의 2장에서와 같이 민감한 그룹 A의 모집단비율을 추정하는 문제를 생각한다.

확률화응답기법에서는 응답자가 선택한 질문을 응답자 자신 이외에는 알 수가 없게 되므로 응답자의 신분이 보장된다. 그러나 Warner기법에서의 두 질문중 하나는 민감한 질문이고 다른 하나는 민감한 질문과 배반되는 즉, 부의 관계를 갖는 질문이므로 결과적으로 두 질문 모두가 민감한 질문과 관련되게 돼서 질문에 응답할 때 응답자들로부터 반발을 받을 수 있게 된다. 그리고 응답자들의 신분보호를 위하여 두 질문 모두가 반드시 민감한 질문과 관련을 갖을 필요가 없고 오히려 민감한 질문과 부의 관계를 갖고 있는 질문대신 민감한 질문과 전혀 관련이 없는(민감한 그룹과 전혀 관련이 없는 그룹 Y) 질문을 사용하게 되면 그 질문에 "예" 또는 "아니오"라고 응답하는 것이 응답자들의 신분에 영향을 주지 않게 되므로 응답자들의 신분을 더 보장해줄 수 있게 된다. 예를 들면, 청소년들의 흡연실태를 조사하기 위하여 다음과 같은 설문으로 구성된 확률화응답기법을 생각할 수 있다.

설문1 : 당신은 흡연을 습관적으로 합니까?

설문2 : 당신은 야구를 좋아합니까?

위에서 설문2는 민감한 설문1과는 전혀 무관한 설문으로 이러한 확률화응답기법을 무관질문기법이라한다.

3.1 그룹 Y의 모집단비율(π_y)을 알 때

민감한 그룹을 A라하고 이와 전혀 무관한 그룹을 Y라 하자. 모집단으로부터 SRSWR로 n명의 표본을 뽑아 확률장치를 사용하여 설문1(선택확률 p) 이나 설문2(선택확률 $1-p$)에 대하여 "예" 또는 "아니오"라 응답한다. 여기서도 2장에서와 같이 동일한 확률장치를 사용할 수 있다. 무관질문기법에서 "예"라고 응답할 확률은 다음과 같다.

$$\lambda = p\pi + (1-p)\pi_y \tag{3.1}$$

만약 n명의 표본 중에서 "예"라고 응답한 사람의 수를 n_1이라 하면 $\hat{\lambda} = \frac{n_1}{n}$ 이므로 π의 불편추정량과 이의 분산은 다음과 같다.

$$\hat{\pi}_{u_1} = \frac{\hat{\lambda} - (1-p)\pi_y}{p} \tag{3.2}$$

$$Var(\hat{\pi}_{u_1}) = \frac{\lambda(1-\lambda)}{(np)^2} \tag{3.3}$$

그리고 분산에 대한 불편추정량은 다음과 같게 된다.

$$\hat{Var}(\hat{\pi}_{u_1}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)p^2} \tag{3.4}$$

여기서, π_y를 알고 있다는 가정은 일반적으로 타당하지 않다. Horvitz et al.(1976)은 Richard Morton이 제시한 다음과 같은 확률장치를 이용하여 두개의 표본을 사용하지 않고 π_y의 값을 항상 알 수 있게 하였다.

설문1 : 민감한 설문

설문2 : "예" 라고 응답

설문3 : "아니오" 라고 응답

이 때 각 설문이 선택될 확률은 p_i (i = 1, 2, 3)이고 이들의 합은 1이 된다.

예를 들면, Abernathy et al.(1970)은 붉은 콩, 흰 콩 그리고 푸른 콩이 들어 있는 상자에서 응답자가 임의로 하나의 콩을 뽑아 뽑힌 콩의 색깔에 따라 응답한다. 즉 붉은 콩 이면 설문1에 응답하고 흰 콩이면 설문2 그리고 푸른 콩이면 설문3에 응답한다. 이때 각 설문이 선택될 확률을 p_i (i = 1, 2, 3)라하면 응답자들이 "예"라고 응답할 확률은 다음과 같다.

$$\lambda = p_1\pi + p_2 \tag{3.5}$$

이는 식(3.1)에서 $p = p_1$이고 $\pi_y = \frac{p_2}{p_2 + p_3}$인 경우와 같게 된다.

한편, Boruch(1972)는 확률장치의 결과에 따라 응답자들이 진실되게 응답하지 않을 경우 이를 오염된 모형(contamination design)이라 했고, 이에 앞서 Warner(1971)는 민감한 그룹에 속하지 않은 응답자들로부터 효율의 감소를 줄이기 위하여 Boruch의 기법을 확장한 것을 제시하였다. 이러한 손실은 거짓으로 "예"라고 응답하는 비율을 낮게 유지함으로써 줄여질 수 있다. 예를들면, 응답자들이 확률 p_j , $j = 1, 2, 3$ 를 갖고 아래의 세개의 응답군 중에 하나를 선택하여 응답하도록 한다.

1. (항상 거짓응답)
   - 집단1에 속한 사람이 집단2에 속해 있다고 응답하고
   - 집단2에 속한 사람이 집단1에 속해 있다고 응답한다.
2. (집단1에 있는 경우만 거짓응답)
   - 집단1에 속한 사람이 집단2에 속해 있다고 응답하고
   - 집단2에 속한 사람이 집단2에 속해 있다고 응답한다.
3. (항상 진실하게 응답)
   - 집단1에 속한 사람이 집단1에 속해 있다고 응답하고
   - 집단2에 속한 사람이 집단2에 속해 있다고 응답한다.

여기서, 집단1은 민감한 집단이고 집단2는 민감하지 않은 집단을 나타내며 첫번째 응답군이 선택될 확률을 낮게 하면(예를들어 0.1) 거짓으로 "예"라고 응답할 확률이 작게 될 수 있다.

3.2 그룹 Y의 모집단비율(π_y)을 모를 때

구하고자 하는 미지 모수가 π 하나일 때는 단지 하나의 표본이면 충분하나 두개의 모수 π와 π_y 를 모두 모를 때는 최소한 두개의 표본이 필요하게 된다. 따라서 모집단으로부터 단순임의복원(SRSWR)으로 크기가 n_1과 n_2인 두개의 독립 표본을 추출한다. 두개의 표본을 사용해야 되므로 i, ($i = 1, 2$)번째 표본에

서 민감한 질문이 선택될 확률이 p_i가 되는 두 조의 확률장치가 필요하게 된다. 응답자들이 진실되게 응답한다는 가정하에서 i번째 표본에서 "예"라고 응답할 확률은

$$\begin{aligned}\lambda_i &= p_i\pi + (1-p_i)\pi_y \\ &= p_i(\pi-\pi_y) + \pi_y\end{aligned} \tag{3.6}$$

가 되며 n_{i1}을 i번째 표본에서 "예"라고 응답한 사람들의 수라하면 $\hat{\lambda}_i = \frac{n_{i1}}{n_i}$

가 된다. 또한 $E(\hat{\lambda}_i) = \lambda_i = p_i\pi + (1-p_i)\pi_y$ 로부터

π의 불편추정량은

$$\hat{\pi}_{u_2} = \frac{\hat{\lambda}_1(1-p_2) - \hat{\lambda}_2(1-p_1)}{p_1 - p_2}, \quad p_1 \neq p_2 \tag{3.7}$$

이 된다. 그리고, $Var(\hat{\lambda}_i) = \lambda_i(1-\lambda_i)/n_i$ 이고 $\hat{\lambda}_1$와 $\hat{\lambda}_2$가 독립이므로

$$Var(\hat{\pi}_{u_2}) = \frac{\left[\frac{(1-p_2)^2\lambda_1(1-\lambda_1)}{n_1} + \frac{(1-p_1)^2\lambda_2(1-\lambda_2)}{n_2}\right]}{(p_1 - p_2)^2} \tag{3.8}$$

가 되고 분산의 불편추정량은 다음과 같다.

$$Var(\hat{\pi}_{u_2}) = \frac{\left[\frac{(1-p_2)^2\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1 - 1} + \frac{(1-p_1)^2\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2 - 1}\right]}{(p_1 - p_2)^2}$$

무관질문기법과 Warner기법과의 비교를 위하여 $n_1 = n_2 = \bar{n}$ 로 하면 Warner기법에서의 표본크기는 $n = 2\bar{n}$가 된다.

3.3 무관질문기법에서 모수들의 최적선택

무관질문기법의 가장 바람직한 사용을 위해서는 개인의 사생활을 보호하고 효율의 유지를 유념해야 한다. 이를 위해서는 모수들 p_i, π_y, 그리고 n_1과 n_2를

적절히 잘 선택하여야 한다. Greenberg et al.(1969)는 이러한 무관질문기법에 대한 전반적인 이론적 체계를 수립하였고 Moors(1971)와 Lanke(1975) 등도 무관기법에서 모수 추정문제를 다루었다.

무관질문에서 추정량의 분산은 p_1과 p_2및 무관질문의 선택(π_y) 그리고 표본 n을 어떻게 n_1 과 n_2 할당하는 가에 따라 달라진다.

Greenberg et al.(1969)는 각각 다음과 같은 모수의 선택을 제시하였다.

첫째, 적절한 p_i를 선택하기 위하여 Warner기법에서 p의 선택기준과 같이 p_1을 0.5로부터 멀리 떨어진 값 (0.2 ± 0.1 또는 0.8 ± 0.1)으로 고정하고 p_2의 선택은 분산을 최소로 하는 값으로 결정한다. 그러나 (3.8)식을 p_2로 미분하여 0으로 놓고 풀면 유일한 p_2의 값을 얻을 수 없다. 따라서 분산을 줄이기 위한 방법으로 p_2를 가급적 p_1으로부터 멀리 떨어진 값으로 선택하는 것이 바람직하다. 식(3.7)에서 p_2가 p_1에 가까우면 π의 추정치가 1보다 크게 되므로 응답자의 협조를 해치지 않는 범위에서 p_2와 p_1의 차를 크게하는 것이 바람직하다. 이러한 요구를 충족시키는 간단한 방법은 $p_1 + p_2 = 1$로 둔다. 이는 두 표본에 같은 영향을 주는 이점을 갖게 된다.

둘째, 실제상황에서 표본크기는 $n(=n_1+n_2)$으로 고정한다. 이 조건에서 분산식(3.8)을 최소로 하는 최적의 n_1과 n_2를 구하기 위하여 다음과 같은 Cauchy-Schwarz의 부등식을 이용한다.

즉, $(\sum a_i^2)(\sum b_i^2) \geq (\sum a_i b_i)^2$ 에서 $\frac{b_i}{a_i}$ = 상수 일때 등호가 성립하므로,

$$\left[\frac{(1-p_2)^2\lambda_1(1-\lambda_1)}{n_1} + \frac{(1-p_1)^2\lambda_2(1-\lambda_2)}{n_2}\right](n_1+n_2) \geq \left\{(1-p_2)\sqrt{\lambda_1(1-\lambda_1)} + (1-p_1)\sqrt{\lambda_2(1-\lambda_2)}\right\}^2 \quad (3.9)$$

이고 최적할당은

$$\frac{n_1}{n_2} = \sqrt{\frac{(1-p_2)^2\lambda_1(1-\lambda_1)}{(1-p_1)^2\lambda_2(1-\lambda_2)}} \tag{3.10}$$

이 된다. 이러한 최적할당값을 식(3.8)에 대입하면 다음을 얻는다.

$$Var(\hat{\pi}_{u2}) = \left\{ \frac{(1-p_2)\sqrt{\lambda_1(1-\lambda_1)} + (1-p_1)\sqrt{\lambda_2(1-\lambda_2)}}{\sqrt{n}(p_1 - p_2)} \right\}^2 \tag{3.11}$$

셋째, 무관질문의 모집단비율 π_y는 π값에 관계없이 π_y를 너무 0에 가까운 값을 취하는 것은 바람직스럽지 못하다. 왜냐하면 그러한 선택은 무관질문기법을 사용하는 전반적인 목적에 위배가 될 수 있기 때문이다. 식(3.11)는 미지의 값 λ_i을 포함하고 있고 $\sqrt{\lambda_i(1-\lambda_i)}$는 $\lambda_i = 0.5$ 에서 최대값을 갖고 0.5에서 대칭이며 오목(concave)하므로 λ_i를 가급적 0.5로부터 멀리 떨어져 있는 값을 선택하는 것이 바람직하다. 식(3.6)에서 π_y는 0.5를 기준으로 π와 같은 방향의 값을 갖고 $\max|\pi_y - 0.5|$일 때 분산이 작게 된다. 따라서 대체로 바람직한 π_y는 0.1 (또는 0.9)주변의 값이면 바람직하다.

한편 Moors(1971)는 Greenberg et al.(1969)에서 제시한 $p_1 + p_2 = 1$인 조건에서 p_1과 p_2의 선택이 최선이 아니고 $p_2 = 0$ 가 바람직하다는 것을 제시하였다. $p_1 \neq p_2$이기 때문에 π는 추정가능하고 $p_1 > p_2$라는 가정은 두 독립표본의 역할이 바뀔 수 있기 때문에 전혀 제약이 되지 않는다. 따라서 식(3.11)를 p_1에 관하여 미분해서 단순화시키면

$$sgn\left[\frac{\partial Var(\hat{\pi}_{u2})}{\partial p_1}\right] = \frac{\left(\frac{1}{2} - \lambda_1\right)(\pi - \pi_y)}{\sqrt{\lambda_1(1-\lambda_1)}} - \frac{\sqrt{\lambda_1(1-\lambda_1)} + \sqrt{\lambda_2(1-\lambda_2)}}{p_1 - p_2} \tag{3.12}$$

이 되므로 이 식에 $(p_1 - p_2)\sqrt{\lambda_i(1-\lambda_i)}$을 곱하고 식(3.6)를 이용하면 식(3.12)는

$$sgn[\frac{\partial Var(\hat{\pi}_{u2})}{\partial p_1}] = -\frac{1}{2}\lambda_1(1-\lambda_2) - \frac{1}{2}\lambda_2(1-\lambda_1) - \sqrt{\lambda_1\lambda_2(1-\lambda_1)(1-\lambda_2)} \leq 0 \quad (3.13)$$

가 되므로 p_1이 가급적 클수록 분산이 최소가 된다. 마찬가지로 식(3.11)를 p_2에 관하여 미분하면

$$sgn[\frac{\partial Var(\hat{\pi}_{u2})}{\partial p_2}] = \frac{(\frac{1}{2} - \lambda_2)(\pi - \pi_y)}{\sqrt{\lambda_2(1-\lambda_2)}} + \frac{\sqrt{\lambda_1(1-\lambda_1)} + \sqrt{\lambda_2(1-\lambda_2)}}{p_1 - p_2} \geq 0 \quad (3.14)$$

이 되므로 p_2가 가급적 작을 때 즉, p_2=0일 때가 분산이 최소가 된다. p_2 = 0이면 하나의 표본에 민감하지 않은 질문만 하게 되므로 이를 π_y를 추정하는 데만 사용할 수 있고 또한 추출계획(sampling plan)도 매우 쉽게 된다.

Greenberg et al.의 $p_1 + p_2 = 1$에 비해 Moors의 $p_2 = 0$인 경우 다음과 같은 장점이 있다.

1. $p_2 = 0$에서 얻은 추정량의 분산이 더 작다.
2. 확률장치가 전체 표본 가운데 일부를 설명하는 데 만 필요하므로 면접 비용(interview cost)을 줄일 수 있다.
3. 특성 Y가 조사자(연구자)에 의해 선택될 수 있기 때문에 개인 면접 (personal interview)보다 훨씬 적은 비용으로 π_y를 추정할 수 있다.

Lanke(1975)는 무관질문기법에서 A는 민감한 성질이고 A^c는 민감하지 않는 성질이며 $\pi < 0.5$ 인 가정에서 π_y의 선택문제를 다루었다.

Greenberg et al.와 Moors는 π_y의 선택을 0.5를 기준으로 π와 같은 방향에서 응답자들의 협조를 해치지 않는 범위에서 가급적 작은 값의 사용을 제시하고 있

으나 응답자들의 입장에서는 단지 "예"라는 응답이 당혹스럽게 되고 빈도가 많아지면 민감하지 않은 질문에 대하여도 "예"라는 응답을 회피하게 된다. π_y를 0.1로 택한 경우보다 π_y를 0.9로 택한 경우가 오히려 응답자들에 대한 의심을 덜어줄 수 있게 된다. 그래서 Lanke는 π_y를 1로 선택하는 것이 바람직하다고 주장하였다.

Moors의 조건 $p_2 = 0$인 경우 분산식(3.11)는 다음과 같이 π_y의 함수가 된다.

$$Var(\hat{\pi}_{u_2}) = \frac{\{\sqrt{\lambda_1(1-\lambda_1)} + (1-p_1)\sqrt{\pi_y(1-\pi_y)}\}^2}{np_1^2} \qquad (3.15)$$

여기서, $\pi_y \in (\pi_-, \pi_+)$, $\pi_- < 0.5$이고, $\pi_+ > 0.5$이며 $\pi_+ + \pi_- > 1$이다.

식(3.15)은 오목(concave)이므로 π_-이나 π_+에서 최소값을 갖게되고 $\pi_+ + \pi_- = 1$ 인 경우는 π_-가 바람직하나 $\pi_+ + \pi_- > 1$ 가정하 에서는 π_+에서 분산이 최소가 된다. 따라서 위의 조건에서는 π_y를 1로 선택하는 것이 바람직하게 된다.

3.4 Warner기법과 무관질문기법과의 비교

Warner기법과 무관질문기법하에서 모집단비율 π에 대한 추정량들의 효율을 비교하기 위하여 Dowling과 Shachtman(1975)는 π_y를 알 때와 π_y를 모를 때로 나누어 비교하였다. 먼저 π_y를 알 때는 하나의 표본만 필요하게 되므로 식(2.3)과 식(3.3)으로부터 추정량들의 분산은 다음과 같게 된다.

$$\begin{aligned} Var(\hat{\pi}_w) &= \frac{\lambda(1-\lambda)}{n(2p-1)^2} \\ &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \end{aligned} \qquad (3.16)$$

$$Var(\hat{\pi}_{u_1}) = \frac{\lambda(1-\lambda)}{(np)^2} = \frac{\pi(1-\pi)}{n} + \frac{(1-p)}{np^2} K_p(\pi_y, \pi) \quad (3.17)$$

여기서,

$$K_p(\pi_y, \pi) = p(1-2\pi_y)\pi + \pi_y[1-(1-p)\pi_y] \quad (3.18)$$

정리 3.1 π_y를 알고 $p \in (p_0, 1)$이면 모든 $(\pi_y, \pi) \in [0,1]^2$에서 다음이 성립한다.

$$Var(\hat{\pi}_{u1}) < Var(\hat{\pi}_w)$$

단, p_0 는 $[0, \frac{1}{2}]$에서 $1/(1+p^2) = 4p(1-p)$의 유일한 해가된다.

증명 : 식(3.16)와 식(3.17)으로부터 $Var(\hat{\pi}_{u_1}) < Var(\hat{\pi}_w)$이 성립할 필요충분 조건이 다음과 같으므로 식(3.19)가 모든 $(\pi_y, \pi) \in [0,1]^2$에서 성립함을 보여주면 된다.

$$K_p(\pi_y, \pi) < p^3 / (2p-1)^2 \quad (3.19)$$

식(3.18)은 집합 $[0,1]^2$ 에서 연속이므로 최대값이 존재하고, 이는 우선 π에 관하여 최대화하고 그 다음 π_y에 관하여 최대화 함으로써 얻어질 수 있다. 더욱이

$$K_p(\pi_y, \pi) = K_p(1-\pi_y, 1-\pi)$$

가 되므로 $(\pi_y, \pi) \in [0, \frac{1}{2}] \times [0, 1]$로 제한할 수 있다. $\pi_y \in [0, \frac{1}{2}]$로 고정하면

$$\max_{\pi \in [0,1]} K_p(\pi_y, \pi) = K_p(\pi_y, 1) = -(1-p)\pi_y^2 + \pi_y(1-2p) + p$$

가 되므로 이는 π_y에 관하여 오목이차함수가 되므로 다음값에서 최대가 된다.

$$\pi_y^* = (1-2p) / [2(1-p)]$$

$p \neq \frac{1}{2}$을 가정하고 $p \in (0, \frac{1}{2})$이면 $\pi_y^* \in (0, \frac{1}{2})$가 되고 $p \in (\frac{1}{2}, 1)$이면 $\pi_y^* < 0$가 되므로

$$\max_{(\pi_y, \pi) \in [0,1]^2} K_p(\pi_y, \pi)$$

$$= \max_{\pi_y \in [0, 1/2]} K_p(\pi_y, 1)$$

$$= \begin{cases} K_p(\pi_y^*, 1) = \frac{1}{4(1-p)}, & p \in (0, \frac{1}{2}) \\ K_p(0,1) = p, & p \in (\frac{1}{2}, 1) \end{cases} \quad (3.20)$$

식(3.20)을 식(3.19)에 대입하면 모든 $(\pi_y, \pi) \in [0,1]^2$에서 $Var(\hat{\pi}_{u_1}) < Var(\hat{\pi}_w)$이 성립할 조건은 다음중 하나만 성립하면 된다.

① $p \in (0, \frac{1}{2})$이고 $\frac{1}{4(1-p)} < \frac{p^3}{(2p-1)^2}$

② $p \in (\frac{1}{2}, 1)$이고 $p < \frac{p^3}{(2p-1)^2}$

여기서, ②의 부등식은 모든 $p \in (1/2, 1)$에서 성립하고 ①의 부등식은 $1/(1+p^2) < 4p(1-p)$와 같게 되므로 $p \in (p_0, 1/2)$에서 성립한다.

따라서 $p \in (p_0, 1)$에서 $Var(\hat{\pi}_{u_1}) < Var(\hat{\pi}_w)$이 성립한다. ■

한편 Dowling과 Shachtman(1975)는 π_y를 모를 때 무관질문기법에서 $p_1 = p$, $p_2 = 0$를 주장한 Moors(1971)의 최적기법을 사용하였다. 따라서 식(3.15)은 다음과 같이 된다.

$$Var(\hat{\pi}_{u_2}) = \frac{\{\sqrt{\lambda_1(1-\lambda_1)} + (1-p)\sqrt{\pi_y(1-\pi_y)}\}^2}{np^2} \qquad (3.21)$$

단, $\lambda_1 = p\pi + (1-p)\pi_y$가 된다.

정리 3.2 $p \in (p_0, 1)$이면 모든 $(\pi_y, \pi) \in [0,1]^2$에서 다음이 성립한다. 여기서, $p_0 = \frac{3-\sqrt{5}}{2}$

$$Var(\hat{\pi}_{u_2}) < Var(\hat{\pi}_w)$$

증명 : 식(3.21)에서

$$L_p(\pi_{y,\pi}) = \{\sqrt{\lambda_1(1-\lambda_1)} + (1-p)\sqrt{\pi_y(1-\pi_y)}\}$$

로 놓으면 $L_p(\pi_{y,\pi})$는

$$\pi_y = (1-p\pi) / (2-p)$$

일 때 최대가 되며 이때 분산은 아래와 같이 된다.

$$\max_{0 \le \pi_y \le 1} Var(\hat{\pi}_{u_2}) = \frac{\pi(1-\pi)}{n} + \frac{1-p}{np^2} \qquad (3.22)$$

그러므로 식(3.22)와 식(3.16)로부터 $Var(\hat{\pi}_{u_2}) < Var(\hat{\pi}_w)$이 성립할 필요충분조건은

$$\frac{1-p}{p^2} < \frac{p(1-p)}{(2p-1)^2}$$

이거나

$$p^3 - 4p^2 + 4p - 1 > 0 \qquad (3.23)$$

이 된다. 식(3.23)의 3차식의 해는 $p_0 = 1$, $\frac{3-\sqrt{5}}{2}$ 와 $\frac{3+\sqrt{5}}{2} > 1$이 되므로 식(3.23)은 $p \in (p_0, 1)$ 조건하에서 $p \in (0,1)$에서성립한다. ■

π_y를 알고 있는 경우와 모르고 있는 경우의 비교를 위하여 $Var(\hat{\pi}_{u_1})$과 $Var(\hat{\pi}_{u_2})$를 각각 다음과 같이 나타낼 수 있다.

$$Var(\hat{\pi}_{u_1}) = \frac{\pi(1-\pi)}{n} + \frac{(1-p)}{np^2} \times [p(1-2\pi_y)\pi + \pi_y \{1-(1-p)\pi_y\}] \quad (3.24)$$

$$Var(\hat{\pi}_{u_2}) = \frac{\pi(1-\pi)}{n} + \frac{(1-p)}{np^2} \times [\{2-p-2(1-p)\pi_y\}\pi_y + p\pi(1-2\pi_y) + 2\sqrt{\pi_y(1-\pi_y)\lambda(1-\lambda)}] \quad (3.25)$$

식(3.24)과 식(3.25)로부터 $Var(\hat{\pi}_{u_1}) \leq Var(\hat{\pi}_{u_2})$ 일 필요충분조건은

$$1-(1-p)\pi_y \leq 2-p-2(1-p)\pi_y$$

또는

$$(1-p)(1-\pi_y) \geq 0$$

이 된다.

한편 Folsom et al.(1973)은 민감한 성질과 전혀 관계없는 두 개의 무관질문 Y_1 과 Y_2 를 민감한 질문과 함께 사용하는 기법을 제시하고, 두개의 무관한 성질 Y_1, Y_2와 민감한 성질 A가 상호독립이고 $n_1 = n_2 = \frac{n}{2}$ 이면, 하나의 무관질문기법에서 π_y가 미지인 경우 Moors 추정량보다 효율적이고,한편, π_y가 기지(known)인 경우에 하나의 무관질문을 사용한 기법에서 π_y가 기지(known)인 경우보다 효율적이지 못함을 보였다.

4. 확률화응답기법의 활용 (Application of Randomized Response Technique)

조사시 발생하는 두가지의 주된 고질병인 응답률의 감소와 응답편의(response bias)의 증가는 민감한 질문에 대한 직접질문과 항상 연관되어 있다. 습관적인 도박, 약물이나 환각제의 남용, 알콜중독, 탈세, 인공조산, 음주운전, 전과경력 그리고 동성연애 등은 일반적으로 사람들이 누설하기 원치 않는 성질들이 된다. 사회적으로 중대한 이러한 문제들에 관하여 조사자들은 표본으로 뽑힌 대상들에 민감하고 자극적이며 죄가 될만한 위와같은 문제들에 대해 직접질문하는 것을 꺼려하게 된다. 또한 이러한 범주형변수 이외의 양적변수들에 대하여도 여러 사람들의 관심을 갖고 있다. 예를들면, 대 사업가 또는 세계적인 영화나 스포츠의 슈퍼스타들의 부의 축적, 은밀하고 불공정한 수단을 통해 벌어들인 돈이나 불법적이거나 비도덕적인 목적으로 돈을 사용한 것, 탈세, 의무의 회피 등등. 자연히 관련된 사람들은 이러한 문제들에 대하여 비밀을 지키려고 하고 진실을 말하는 것을 거려하게 된다.

확률화응답기법(RRT)들은 무응답, 고의적인 거짓진술 그리고 거짓말의 수준을 줄일 수 있는 유용한 방법들 중의 하나가 된다. 그래서 추정의 편의를 줄이고 검정의 효율을 높이고자 한다. 우선은 이러한 RRT들이 응답자들의 사생활을 보호해 준다고 믿게하고 더 중요한 것은 후에 이를 믿게 되서 자발적으로 정확한 응답을 하게하기 위해서 한번 계획된 RRT는 효율적이고 확실한 방법으로 활용될 필요가 있다. 어떤 RRT가 주어 졌을 때 이의 효율을 평가하는 것이 가장 어렵고 한번 사용된 후 이의 신뢰성을 추측하고 측정하는 것도 아주 위험하다. 물론 이론적으로 RRT의 가능성을 측정하는 것은 어렵지 않으나 실제로 그것이 얼마나 잘 활용(진행)되는 가를 평가하는 것은 쉽지 않다. 그럼에도 불구하고 RRT들이 이 책에서 검토한 많은 문헌(논문)에 다양하게 제시되고 놀라울 정도로 빠르게 출현하여 활용되고 있다.

문헌조사에 의하면 이들 기법들중 상당수가 성공적인 결과를 제시해 주고 있는 반면에 일부는 그렇지 못하였다. 그러나 여전히 이러한 경험들로부터 미래에는 RRT들이 널리 사용하리라고 예견하는 것이 어리석지는 않을 것이다. 가치있는 응용이나 연구에 관한 문헌(논문)들로부터 RRT들의 효율을 실제로 재 평가 할

수 있을 것이다.

Warne(1965)가 그의 논문에서 RRT를 실제 적용한 후 많은 사람들에 의해서 특정문제에 대하여 다양한 RRT의 활용을 제시하고 여러 사례를 통하여 그들의 효율성을 입증하였다. 또한 대체 방법들간의 비교연구도 하였다.

Emrich(1983)은 RRT에 관한 다음과 같은 단점을 지적하였다.

첫째, RRT는 조사자들에게 특별한 기술을 요구하고 RRT를 사용하는 방법을 조사 대상자들에게 설명하는 데 많은 시간과 비용이들고 어려움이 따르기 때문에 이들을 도구로 사용하기가 곤란하게 된다.

둘째, RRT들이 응답자들의 거짓응답을 허용하는 오염된 응답(contaminated responses)을 할수 있게 사전에 장치를 마련하고 있다는 것을 알게 되면 응답자들이 진실된 응답을 하지 않게될 수 있다.

이러한 단점에도 불구하고 기존 RRT의 지속적인 혁신, 새로운 기법의 개발과 시종일관 균형(조화)있는 활용등 문헌에서의 빠른 성장을 고려한다면 우리들은 이론과 실제에서 RRT의 전반적인 관심과 흥미가 증가한다고 결론지울 수 있다.

RRT가 때로는 응답비율을 높여주지 못한 다는 주장은 응답자들이 RRT에 대한 이해가 부족하거나 조사 결과의 비밀보장에 대한 믿음이 결핍되어 생긴 것으로 볼 수 있다.

5. 확률화응답기법의 적용사례

민감한 질문에 대한 조사시 발생하는 두 가지의 주된 고질병은 응답률의 감소와 응답편의(response bias)의 증가인데 이는 직접질문과 연관되어 있다.

습관적인 도박, 약물이나 환각제의 남용, 알콜중독, 탈세, 인공조산, 음주운전, 전과경력 그리고 동성연애 등은 일반적으로 사람들이 누설하기 원치 않는 사항들이 된다. 사회적으로 중대한 사항들에 관하여 조사자들은 표본으로 뽑힌 대상들에게 민감하고 자극적이며 죄가 될만한 위와같은 문제들에 대해 직접질문하는 것을 꺼려하게 된다. 또한 범주형변수 이외의 양적변수들에 대하여도 사람들은 관심을 갖게 된다. 예를들면, 대 사업가 또는 영화나 스포츠의 슈퍼스타들의 부의 축적, 은밀하고 불공정한 수단을 통해 벌어들인 돈, 불법적이거나 비도덕적인 목적으로 돈을 사용한 것, 탈세, 의무의 회피 등등. 자연히 이러한 문제들에 연관된 사람들은 비밀을 지키고 진실을 말하는 것을 꺼려하게 된다.

확률화응답기법(RRT)들은 무응답, 고의적인 거짓진술 그리고 거짓말의 수준을 줄이고 추정의 편의를 작게해서 검정의 효율을 높일 수 있는 유용한 방법들 중의 하나가 된다. 우선은 이러한 RRT들이 응답자들의 사생활을 보호해 준다고 믿게하고 더 중요한 것은 후에 이를 믿게 되서 자발적으로 정확한 응답을 하게하기 위해서 한번 계획된 RRT는 효율적이고 확실한 방법으로 활용될 필요가 있다. 어떤 RRT가 주어 졌을 때 이의 효율을 평가하는 것이 가장 어렵고 한번 사용된 후 이의 신뢰성을 추측하고 측정하는 것도 아주 위험하다. 물론 이론적으로 RRT의 가능성을 측정하는 것은 어렵지 않으나 실제로 그것이 얼마나 잘 활용(진행)되는가를 평가하는 것은 쉽지 않다. 그럼에도 불구하고 RRT들이 이 책에서 검토한 많은 문헌(논문)에 다양하게 제시되고 놀라울 정도로 빠르게 출현하여 활용되고 있다.

문헌조사에 의하면 이들 기법들중 상당수가 성공적인 결과를 제시해 주고 있는 반면에 일부는 그렇지 못하였다. 그러나 가치있는 응용이나 연구에 관한 문헌(논문)들로부터 RRT들의 효율을 실제로 재 평가 할 수 있으므로 미래에는 RRT들이 널리 사용하리라고 예견하는 것이 어리석지는 않을 것이다.

Warner(1965)가 그의 논문에서 RRT를 실제 적용한 후 많은 사람들에 의해서 특정문제에 대하여 다양한 RRT의 활용을 제시하고 여러 사례를 통하여 그들의 효율성을 입증하였다. 또한 대체 방법들간의 비교연구도 하였다.

Emrich(1983)은 다음과 같은 RRT에 관한 단점을 지적하였다.

첫째, RRT는 조사자들에게 특별한 기술을 요구하고 RRT를 사용하는 방법을 조사 대상자들에게 설명하는 데 많은 시간과 비용이들고 어려움이 따르기 때문에 이들을 도구로 사용하기가 곤란하게 된다.

둘째, RRT가 거짓응답을 허용하는 오염된 응답(contaminated responses)을 할 수 있게 사전에 장치를 마련하고 있다는 것을 응답자들이 알게 되면 진실된 응답을 하지 않게 될 수 있다.

이러한 단점에도 불구하고 기존 RRT의 지속적인 혁신, 새로운 기법의 개발과 균형(조화)있는 활용등 문헌에서의 빠른 발전을 고려한다면 이론과 실제에서 RRT의 전반적인 관심과 흥미가 증가한다고 결론지을 수 있다.

RRT가 때로는 응답비율을 높여주지 못한 다는 주장은 응답자들이 RRT에 대한 이해가 부족하거나 조사 결과의 비밀보장에 대한 믿음이 결핍되어 생긴 것으로 볼 수 있다.

< 사례 1 >

Reinmuth, J. E. and Geurts, M. D.(1975). "The collection of sensitive information using a two-stage, randomized response technique".

목적 : Honolulu에 있는 대규모 소매상가에서 들치기(shoplifter)들의 피해도를 조사하고자 함. 즉 들치기배들 중에서 들치기의 빈도를 추정하고자 한다.

방법 : 양적이고 질적인 확률화응답기법으로 부터 얻은 결과를 혼합한 2단계 확률화응답기법인 비추정량을 사용.

Honolulu에 있는 쇼핑센타의 모든 고객중 들치기배들의 비율, $\hat{\pi}_s$, 과 고객당 평균 들치기 빈도, $\hat{\mu}_s$, 는 각각 질적확률응답기법과 양적확률응답기법에 의하여 추정된다. 이들로부터 들치기배들의 평균 들치기빈도의 추정은 비추정에 의하여 얻을 수 있다. 즉, $\hat{\mu}_s / \hat{\pi}_s = \hat{\theta}$.

비추정량의 편의는 점근적으로

$$B(\hat{\theta}) = E(\hat{\theta} - \theta) \doteq (\theta / \pi_s^2) V(\hat{\pi}_s)$$

이고 이의 추정량은 다음과 같다.

$$\hat{B}(\hat{\theta}) = (\hat{\theta} / \hat{\pi}_s^2) V(\hat{\pi}_s)$$

따라서 수정된 θ의 추정량과 이의 근사 분산은 각각 다음과 같다.

$$\theta' = \hat{\theta} - \hat{B}(\hat{\theta}) = \hat{\theta}\left\{1 - \frac{V(\hat{\pi}_s)}{\hat{\pi}_s^2}\right\}$$

$$\hat{V}(\hat{\theta}) = \frac{1}{\hat{\pi}_s^2}[V(\hat{\mu}_s) + \theta'^2 V(\hat{\pi}_s)]$$

가정 : $\hat{\pi}_s$ 와 $\hat{\mu}_s$ 는 독립이고 배반인 표본으로부터 얻는다. 왜냐하면 첫째, 조사대상자들이 하나 이상의 민감한 질문에 응답하는 것을 꺼려하고 더욱이 들치기의 경험이 없다고 응답한 사람들에게 들치기의 횟수를 물어보는 것은 어리석은 질문이 된다. 따라서 각기 다른 표본으로부터의 추정이 보다 더 실뢰할 수 있는 결과를 기대할 수 있다. 둘째, 두 추정치가 통계적으로 독립이면 이론적으로 비추정량, $\hat{\theta}$, 의 분포성질을 단순화 시킬 수 있다.

조사대상 : Honolulu에 있는 Ala Moana쇼핑센터의 고객중 342명을 임의로 추출하여 잘 훈련된 두 명의 조사원이 5일간 계속하여 조사하였다.

조사방법 : 75개의 검은 공과 25개의 흰 공이 들어 있는 상자를 확률장치로 사용하였고 사전에 응답자들에게 확률장치의 사용에 대하여 상세히 설명을 하여주었다. 조사는 2단계로 실시 하였는데 1단계로 고객중 들치기배들의 비율을 추정하기 위하여 임의로 선택한 184명을 $n_1 = 138$, $n_2 = 46$ 인 두개의 부표본으로 분할하여 다음과 같은 무관질문기법을 사용하였다.

(질문1) "지난 1년 동안 Ala Moana에 있는 쇼핑센터의 소매상들로 부터 들치기 한 적이 있는 가?"

(질문2) "지난 1주일 동안 Ala Moana에서 쇼핑을 하였는 가?"

첫번째 부표본의 응답자들은 확률장치에서 검은 공이 나오면 질문1에 응답하고 두번째 부표본의 응답자들은 흰 공이 나오면 질문1에 응답한다. 또한 들치기의 정도를 추정하기 위하여 중복되지 않는 158명의 표본을 뽑아 $n_1 = 126$, $n_2 = 42$ 인 두개의 부표본으로 분할하여 다음 질문을 한다. 방법은 들치기배들의 비율을 추정한 경우와 동일하게 한다.

(질문1) "지난 1년 동안 Ala Moana의 소매상들로 부터 몇 번이나 들치기를 하였는 가?"

(질문2) "지난 한 달 동안 Ala Moana에서 몇 번이나 쇼핑을 하였는 가?"

결과 : 확률화응답기법을 사용하여 조사한 결과는 표5-1에 요약되어 있다. 표

5-1로부터 모든 고객의 약 20%가 Ala Moana에서 들치기 한 것으로 나타났고 들치기 횟수는 고객 한 사람당 평균 1.7회이고 들치기배들 중에서는 평균 7.9회나 되었다. θ에 대한 95% 신뢰구간은 0.25에서 15.57로 상당히 넓은 데 이는 들치기를 한번도 하지 않은 사람이 상당히 많고 들치기 횟수가 넓게 퍼져있어 분산이 상대적으로 크기 때문이다. 이 조사에서 총 응답자 가운데 단 3명 만이 응답을 거절하였는데 이는 확률화응답기법이 민감한 사항에 대한 조사에 무응답 편의를 줄여줄 수 있는 것으로 볼 수 있다.

표5-1. 확률화응답기법을 사용한 조사결과

	질적조사	양적조사	비추정량
평 균	$\hat{\pi}_s$ = .19565	$\hat{\mu}_s$ = 1.7142	θ' = 7.9117
분 산	$V(\hat{\pi}_s)$ = .00369	$V(\hat{\mu}_s)$ = .3315	$\hat{V}(\hat{\theta})$ = 14.6941
95%근사	.07565	0.5642	0.2517
신뢰구간	~ .31565	~ 2.8642	~ 15.5719

확률화응답기법이 시장조사에 대한 연구에 유용하게 사용되기 위해서는 조사의 전 과정들이 조사의 결과를 대외적으로 입증할 수 있게 계획되어져야 한다. 일반적인 표본조사에서 예전에는 적절치 못하다고 생각되는 사항에 관하여 확률화응답기법을 이용하면 신뢰성있는 정보를 얻을 수 있게 된다. 게다가 2단계 확률화응답기법은 민감한 사항에 관하여 추가적인 정보를 제공해 준다. 본 논문에서 다룬 2단계기법은 다음과 같은 문제를 연구하는 데 유용하게 활용될 수 잇다.

1. 약물 사용자들에 대한 사용의 빈도를 추정.
2. 신경안정제를 사용하는 사람들의 평균 구입 횟수를 추정.
3. 범법자들 중에서 사무직 계급의 범죄빈도를 추정.

< 사례 2 >

Charles W. Lamb, Jr. and Donald E. Stem, Jr.(1978). "An empirical validation of the randomized response technique".

RRT를 평가하기 위하여 다음과 같은 두가지 점을 생각 할 수 있다.

1. RRT가 모수와 유의적으로 차이가 나지 않는 추정치를 제공하는 가?

2. RRT가 기존 방법에 의한 것보다 더 정확한 추정치를 제공해 주는 가?

RRT에 의한 추정치가 모수와 통계적으로 유의한 차가 없다는 것을 검정하기 위해서는 모수를 알아야 한다. RRT의 타당성에 대한 대부분의 연구가 어려움을 겪는 이유는 모수에 근사한 정보를 얻지 못하기 때문이다. 또한 RRT가 전통적인 조사방법보다 더 정확한 추정치를 제공하는지 어떤지를 검정하기 위해서는 모수를 알고 있는 상황에서 RRT와 종전 방법에의한 추정치들을 비교하여야 한다.

만약 RRT를 사용함으로써 줄어든 측정오차가 RRT를 사용하는 데 따른 비용과 자체 방법상의 비효율성을 상쇄시키지 못한다면 RRT의 활용에 한계가 된다. 따라서 RRT의 타당성에 대한 연구가 지속적으로 이루어지고 이러한 연구가 RRT 활용의 한계를 극복하는 계기가 된다.

목적 : RRT의 실증적 연구의 결과를 제시하기 위하여 민감한 질문으로 대학생들이 재학중 F학점 받은 과목 수를 양적, 질적 그리고 2단계(Reinmuth and Guerts,1975)확률응답기법과 직접질문을 사용하여 얻은 추정치들을 실제 대상 학생들의 F학점 수를 갖고 비교하였다.

방법 : 330명의 상급학년을 대상으로 F학점 수를 조사하였는 데 이중 18명이 조사를 거부하여 총 312명(응답비율 96%)을 표본으로 사용하였다. 조사에 참여한 응답자들에게 참여의 댓가로 상금을 걸었다. 각 응답자들은 색깔로 구분된 상금추천 카드에 이름, 전화번호, 학번등을 기입하도록 하였다. 각 카드의 색깔은 조사에 사용된 5가지의 처리방법을 나타낸다. 카드를 작성한 후 이를 특별히 설계된 상자에 넣으며 색깔에 따른 면접표를 받는다. 이때 상자는 원래의 카드 투입순위가 유지되도록 설계되어져 있다. 색깔로 분류된 조사표와 상금추천 카드의 조합이 응답자의 고유번호가 된다. 색깔로 구분된 상금추천카드에 기록된 학생들의 자료를 이용하여 학적과로부터 각 응답자들의 실제 F학점과목 수를 알 수 있다. 5개의 부표본 또는 5개의 처리방법중 하나는 종래의 직접질문을 사용한 통제 집단이고 이들에게 "F학점과목이 있다면 몇과목이나 되는가?" 라는 질문의 응답을 조사표에 기입한다. 나머지 4개의 부표본을 질적이고 양적인 확률화응답기법으로 나눈다. 즉 두개의 확률화응답기법은 각각 두개의 독립이고 배반인 처리집단으로 나눈다.응답자들은 붉은 돌과 흰 돌이 3:1 또는 1:3으로 구성된 두개의 상자중 하나의 상자로부터 뽑은 돌의 색깔에 따라 다음 질문에 응답한다.

가. 질적질문

붉은 돌 : 재학중 F학점 과목이 있는 가?

흰 돌 : 지난 1년간 취소한 과목이 있는 가?

나. 양적질문

붉은 돌 : 재학중 F학점 받은 과목이 얼마나 되는 가?

흰 돌 : 이번 학년에 취소한 과목이 얼마나 되는 가?

표본크기와 추출확률에 따른 처리의 내용이 표5-2에 나타나 있다.

표5-2. 5가지 처리방법의 비교

처리	기법	민감한 질문의 선택확률(p)	표본수(n)	추출확률(p)
1	직접질문	n.a	63	.20
2a	질 적	.75	96	.30
b	질 적	.25	25	.10
3a	양 적	.75	97	.30
b	양 적	.25	31	.10

결과 : 실제값과 5가지 처리방법에 따른 추정치들 간의 차를 검정하기 위하여 student의 t-검정을 사용하였고 직접질문과 RRT를 사용한 조사결과가 표7-3에 요약되어 있다.

직접질문으로부터 얻은 추정치와 실제 값과의 차가 유의적이지 않으면 질문 내용이 민감한 내용으로 볼수 없다. 또한 RRT를 통하여 얻은 추정치와 실제 값과의 차가 유의적이지 않으면 RRT에 의하여 얻은 추정치가 정확하다고 볼 수 있다.

표5-3. 직접질문과 RRT를 사용한 조사결과

구 분	추정치	95% 신뢰구간	실제값	차	t	α
F학점 받은 학생 비율	.288	0.176-0.400	0.302	0.014	0.278	0.799
평균 F학점 수	.460	0.236-0.684	0.714	0.254	2.227	0.030
F학점 받은 학생중 평균 F학점 수	1.602	1.181-2.023	2.368	0.766	3.565	<.001
질적(처리1, $\hat{\pi}$)	.363	0.189-0.537	0.347	-0.016	0.1745	0.865
양적(처리2, $\hat{\mu}$)	.536	0.316-0.810	0.641	0.078	0.619	0.535
수정된 비추정치($\hat{\theta}$)	1.460	0.482-2.438	1.97	0.510	1.023	0.308

표5-3으로부터 직접질문으로 얻은 F학점 받은 학생 비율의 추정치(0.288)가 실제비율(0.302)보다 작으나 그 차는 통계적으로 유의하지 않다. 반면에 직접질문으로 추정한 평균 F학점 수와 실제 평균 F학점수와의 차는 통계적으로 유의하다고 볼 수 있다. 마찬가지로 직접질문에서 F학점 받은 과목이 있다고 응답한 사람중 평균 F학점수와 실제 F학점 받은 과목이 있는 사람중 평균 F학점수와의 차이는 통계적으로 유의하다(α<0.001). 이러한 결과로부터 응답자들이 F학점을 받았다고 응답하는 것을 꺼리지 않는다고 결론지을 수 있다.즉, 과목을 낙제한 것이 민감한 질문이라고 볼 수 없지만 몇 과목이나 낙제를 했는가 하는 질문은 민감한 질문으로 볼 수 있다. 한편 RRT를 사용하여 얻은 추정치들과 실제 값들과의 차이는 유의하지 않으므로 RRT에 의한 추정치들이 모수에 대한 보다 정확한 추정치로 볼 수 있다. 결론적으로 일부의 질문이 민감한 질문이 아니더라도 직접질문보다 RRT를 사용하여 얻은 추정치가 정확하다고 말할 수 있다. 그러나 음주운전이나 쇼핑에서 들치기 등에 대한 질문은 언제나 민감한 질문이 되므로 이러한 경우는 RRT의 사용이 보다 효율적이라 할 수 있다.

< 사례 3 >

Shimizu, I. M. and Bonhman,G. S. (1978). "Randomized response technique in a national survey".

목적 : 1973년 미국의 가족수 증가에 대한 전국조사(National Survey of Family Growth : NSFG)에서 지난 1년간 낙태(유산)한 경험이 있는 부인의 수를 추정하기 위함.

방법 : 전체 표본을 반으로 나눈 두개의 표본에 두 개의 무관질문을 이용한 Folsom et al(1973)의 확률응답기법을 사용하였다. 확률장치로는 질문이 선택될 확률이 1/2인 동전을 사용하였다.

조사시기 : 실사는 1973년 7월부터 1974년 2월사이에 시카고대학의 국민여론쎈타(Natiomal Opinion Research Center)에 의해 수행되었다.

조사대상 : 미국 가정에 거주하는 15세부터 44세까지 적어도 한번 이상 결혼한 경험이 있거나 자식이 있는 부인을 대상으로 다단계 확률표본을 기초로 하였다.

조사방법 : 조사는 보건통계를 위한 국가기관(National Center for Health Statistics)에서 주기적으로 실시된다. 본 연구는 1973년에 실시된 첫번째 조사에서 얻어진 자료를 사용하였다. 모든 표본가구는 표본으로 추출됨에 따라 번호가 메겨

지고 번호의 끝자가 홀수이면 첫번째 표본에(응답자수:4,926명)할당 되고 짝수이면 두번째 표본(응답자수:4,871)에 할당 되어 각각 다음과 같은 질문을 받게된다.

<표본1>

직접질문 : 작년 이맘때 당신은 지금과 다른 군이나 주에 살았습니까?

(This time last year, Did you live in a different county or state than this one ?)

RRT : 그림5-1과 같은 카드를 받고 동전의 위치에 따라 단지 "예" 또는 "아니오"로 응답한다.

그림5-1. 표본1에 사용된 카드

표면 : 지난 12개월 내에 낙태를 했다.
이면 : 당신의 어머님이 4월에 태어나셨습니까?

<표본2>

직접질문 : 당신의 어머님은 몇년 몇월에 태어나셨습니까?

RRT : 그림5-2과 같은 카드를 받고 동전의 위치에 따라 단지 "예" 또는 "아니오"로 응답한다.

그림5-2. 표본2에 사용된 카드

표면 : 지난 12개월 내에 낙태를 했다.
이면 : 작년 이맘때 당신은 지금과 다른 군이나 주에 살았습니까?

표본1과 표본2에서 얻은 추정량들의 가중평균치를 최종 추정량으로 사용한다.

즉,

$$\hat{\pi} = w\hat{\pi}(1) + (1-w)\hat{\pi}(2)$$

단, $\hat{\pi}(i)$, $i=1,2$ 는 i번째 표본에서 얻은 낙태한 모집단 비율에 대한 불편추정량이다.

결과 : 조사의 결과가 표5-4에 있다. 표에서 괄호안의 숫자는 추정치들의 표준오차가 된다.

표5-4. 1973년 미국의 15세-44세까지 부인들의 낙태에 대한 조사결과

구 분	총계	미혼모	결혼 경험이 있음	배우자 유무	
				현재 결혼중	과부 이혼녀 별거중
총 부인 수(천명)	31,018 (395)	771 (59)	30,247 (390)	26,646 (364)	3,601 (128)
12개월내에 낙태한 부인 수(천명)	930* (248)	77* (29)	847* (842)	693* (213)	194* (61)
$\hat{\pi}$ (최종추정량)	3.0%* (0.8)	10.0%* (3.7)	2.8%* (0.8)	2.6%* (0.8)	5.4%* (1.7)
$\hat{\pi}(1)$(표본1)	5.3% (1.1)	9.0%* (4.8)	5.3% (1.1)	5.1% (1.1)	6.4%* (2.3)
$\hat{\pi}(2)$(표본2)	0.6%* (1.1)	11.4%* (5.6)	0.3%* (1.1)	-0.2%* (1.1)	4.0%* (2.7)

*는 상대표준오차(= 표준오차/추정치)가 25%를 넘는 것을 나타낸다.

표5-4로부터 낙태에 대한 전체 추정비율은 3.0%이고 표준오차는 0.8%이다. 표본1과 표본2로부터 낙태에 대한 추정비율은 각각 5.3%와 0.6%로 두 추정치의 차는 차에 대한 표준오차의 3배나 된다. 이러한 큰 차이는 우연으로 보기 어렵다. 따라서 측정오차에 대한 검토가 요구된다. 차이가 큰 이유로는 표본2에 대한 질문중 county를 country로 잘못 읽음으로써 생길 수 있고 표본1에서는 응답자들이 어머님의 출생 월을 모름으로 영향을 받을 수 있다. 한편 직접질문한 경우 8.2%가 어머님의 출생월을 모르고 있었다.

RRT에 의한 추정치는 약 3%가 지난 1년간 낙태한 적이 있다고 응답한 반면에 1970년 직접질문에 의한 전국출산력연구(National Fertility Study)조사에서는 15세-44세 결혼한 여성중 1년내 낙태한 적이 있다고 응답한 비율은 0.3%에 불과했다. 낙태자료는 주로 질병관리센터(Center for Diseas Control : CDC)와 Alan

Guttmacher단체(AGI)에 있는데 1973년 이들 기관의 조사자료는 조사대상 모집단이 다르기 대문에 직접비교는 곤란하나 이들 자료를 조정해 줌으로써 비교는 가능하리라 본다. CDC는 24개 주에서 현재 결혼중에 있는 부인의 합법적인 낙태비율을 만약 모든 주로 환산하면 1년내 낙태한 사람이 총 167,000명으로 볼 수 있고 AGI는 203,000으로 볼 수 있으나 본연구인 NSFG는 15세-44세의 최근 결혼중에 있는 사람중 693,000명으로 추정되므로 CDC보고서 보다는 4.1배, AGI보고서 보다는 3.4배가 많은 것으로 평가된다. 결론적으로 낙태에 관한 민감한 조사에 RRT를 사용하여 얻은 추정치가 종전의 직접질문에 의한 추정치보다 높은 값을 얻게 되므로 NSFG에서 RRT의 사용은 가치가 있다고 볼 수 있다.

< 참 고 문 헌 >

1. Abul-Ela et al.(1967). A multiproportions randomized response model, J. Amer. Statist. Assoc. 62, 990-1008.

2. Bourke,P.D.(1982). Randomized response multivariate designs for categorical data, Commun. Statist.- theory and method -, 11(25), 2889-2901.

3. Chaudhuri, A. and Mukerjee, R. (1988). Randomized Response : theory and technique. Marcel Dekker, Inc.

4. Charles W. Lamb, Jr. and Donald E. Stem, Jr.(1978). An empirical validation of the randomized response technique. *J. Marketing Research* 15, 616-621.

5. Devore, J. L. (1977). A note on the RR techniques. *Comm. Statist.- Theory Methods* 6, 1525-1529.

6. Dowering,T.A. and Shachtman, R.H.(1975). On the relative efficiency of randomized response design, *J. Amer. Statist. Assoc.* 70, 84-87.

7. Drane,W.(1976). On the theory of randomized response to two sensitive questions,Commun.Statist.- theory and method -, A5(6), 565-574.

8. Emirch, L.(1983). RR Techniques. In : Incomplete Data in Sample Surveys, vol 2, ed, W.G. Madow, I. Olkin, Academic Press New Yorks 73-80.

9. Flinger et al.(1977). A comparison of two RR survey methods with consideration for the level of respondent protection. *Comm. Statist.- Theory Methods* 6, 1511.1524.

10. Folsom et al. (1973). The two questions randomized response model for human surveys. *J. Amer. Statist. Assoc.* 68, 158-163.

11. Greenberg et al.(1971). Applications of the randomized response technique in obtaining quantitative data , J. Amer. Statist. Assoc. 66, 243-250.

12. Greenberg et al.(1969). The unrelated question randomized response model : theoretical framework. J. Amer. Statist. Assoc. June , 521-539.

13. Iris, M.Shimizu and Gordon Scott Bonham (1978). Randomized Response Technique in a National Survey. *J. Amer. Statist. Assoc.* 73, 35-39

14. Kim, Jong-Ik, and Flueck, John A. (1978a). Modifications of the randomized response technique for sampling without replacement. *proc.* ASA. *Sec. Surv. Res. Methods*, 346-350.

15. Kim, Jong-Ik, and Flueck, John A. (1978b). An additive randomized response technique. *proc.* ASA. *Sec. Surv. Res. Methods*, 351.355.

16. Lanke, J.(1976). On the choice of unrelated question in Simmons' version of randomized response, *J. Amer. Statist. Assoc.* 70, 80-83.

17. Liu,P.T. and Chow,L.P.(1976).The efficiency of the multiple trial randomized response technique, Biometrics, 32, 607-618.

18. Mangat,N.S. and Singh, R.(1990). An alternative ranomized response procedure, Biometrika, 77, 439-442.

19. Moors,J.J.A. (1971). Optimization of the unrelated question randomized response model, J. Amer. Statist. Assoc., 66, 627-629.

20. Raghavarao, D. (1978). On an estimation problem in Warner's RR technique. *Biometrics* 34, 87-90.

21. Reinmuth, J. E. and Geurts, M. D.(1975). The collection of sensitive information using a two-stage, randomized response technique. *J. Marketing Research* 12, 402-407.

22. Warner, S. L. (1965), Randomized response : A survey technique for elliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 63-69.

23. Weissman, Arlenen(1981). Randomized response vs. directed question-ing: Two methods for asking sensitive questions over the telephone. *Paper presented at the American Educational Research Assoc. meeting, Los An-gles, Calif.*

韓國輿論調査의 實態와 調査方法

한국갤럽조사연구소장

朴　武　益

目　　次

한국여론조사 실태와 조사방법

1.여론조사의 현황

일반적으로 여론조사라 하면 정치, 사회여론조사를 연상하는 경우가 많으나 여기서는 기업이 행하는 시장조사 혹은 마아케팅조사를 포함하여 다루어 보고자 한다.

왜냐하면 우리나라 조사의 역사는 1970년대부터 기업의 마아케팅조사로 출발되었다고 할 수 있고 현재에 행해지는 각종 조사의 빈도면에서 마아케팅조사가 주류를 이루고 있기 때문이다.

흔히 한국갤럽조사 연구소를 사회정치여론조사기관으로 알고 있으나 실제로는 이 분야의 비중은 10%에 불과하고 기업이 의뢰해 오는 광고효과조사, 신제품조사, 수요예측 등 Marketing조사 비중이 90%를 점하고 있다.

여하간 여기서는 둘을 통합하여 다루고자 한다.

현재 우리나라에서 조사기능을 수행하는 기관을 형태별로 분류하면 다음과 같이 요약될 수 있다.

1)독립된 조사전문기관

인원이 30명이 넘는 조사기관으로는 한국갤럽, A,C,Nielsen, 한국리서치, 리스 PR,코리아리서치센터 등을 꼽을 수 있으며, 그 수는 10개 전후이다. 이들 기관은 대개 마아케팅조사를 주로 하고 있으며 사회여론조사도 일부 수행하고 있다.

이들 조사기관의 매출은 50억원 전후로 추정된다.

2)광고대행사 내부의 조사부

광고의 양적성장과 더불어 광고대행사의 규모도 커졌으며 인쇄매체의 수수료를 인정받고 있는 광고대행사에는 모두 조사부를 갖고 있다.

현재 순수조사부 인원이 10명이 넘는 대행사로는 제일기획, 오리콤의 2개 회사이며, 15개의 광고대행사의 조사매출은 20억원 전후로 추정된다.

3)대학내의 연구소

현재 종합대학의 사회학과, 신방과 등을 중심으로 각종 연구소가 설립되어 있으며 정부기관 및 언론사로부터 의뢰받아 사회정치여론조사를 행하고 있다. 아마 우리나라의 전체 사회여론조사용역의 90%이상이 이들 연구소가 맡고 있는것으로 보여지며 조사용역비 총규모는 추정하기 어렵다.

4)기업내의 조사부

옛부터 대기업의 조직에는 조사부가 있어 왔다. 그러나 이들의 역할은 2차 자료의 수집, 정리나 경쟁사 정보의 입수에 주력해 왔다고 할 수 있다.

그러나 1980년대에 들어서면서 소비자정보나 유통정보의 수집에 관심을 두기 시작했고, 소비자정보수집을 위해 표본조사를 행하는 곳도 생겨나게 되었다. 우리나라 100대 기업의 조사비는 30억원 전후가 아닐까 한다.

이상에서 일별해 본 바 대로 작년까지 우리나라의 총 조사비는 대학내 연구소를 제외하면 약 100억원으로 추정되며 이는 미국이나 일본에 비하면 아주 보잘 것 없는 시장크기라 할 수 있다.

예컨데 작년 일본에서는 대기업이 쓰는 조사비는 광고비의 1.2%수준인데, 우리나라는 0.2%에도 미달하고 있다. 또한 일본에는 전문조사기관의 수가 200여개이고 그 규모면에서도 우리와 비교할 바가 아니다.

우리의 조사기관의 수가 적은 원인은 주로 조사의 수요가 적은 것에서 비롯한다고 볼 수 있으며 그 원인으로는 다음과 같이 요약할 수 있겠다.

첫째, 기업쪽에서 합리적 의사결정의 바탕으로써 조사를 이용하지 않은 데 있다.

70년대 이전에 우리나라 기업은 상품을 만들면 팔리는 호시절을 경험하였다. 이른바 Seller Market이었다 할 수 있다.

그러나 80년대에 들어서면서 경쟁이 치열해지고 만드는 것보다 판매가 더욱 중요한 시대로 접어들게 되었다. 그러나 최고경영층이나 경영풍토가 자료나 조사를 바탕으로 한 의사결정이 소홀이 되고 있다.

예컨데 작년 국내 연간광고비가 1조원으로 그 규모면에서 크게 성장했음에도 100대 광고주 가운데 그 광고가 효과있는지 없는지를 알아보는 효과측정조사에 광고비의 1%이상 쓴 기업은 한둘에 지나지 않는다.

둘째, 정부기관이나 지방단체들이 정책입안에 있어서 국민의 소리를 듣기 위한 노력이 부족했던 점이다.

그동안 우리나라에 지방자치제가 실시되지 않았기 때문에 정책입안자들이 자기 주관대로 적당히 결정하는 경우가 흔했으며, 국민의 소리를 듣기위해 체계적인 여론조사를 활용하는 면이 아주 부족했다고 할 수 있다.

이의 예로 막대한 예산을 집행한 서울특별시가 작년도에 시민여론을 알기 위해 얼마의 예산을 사용했는지를 확인해 봄으로 명확해질 것이다.

2. 조사방법의 문제점

전문 여론조사기관의 수가 적고 영세함에도 불구하고 우리 언론기관에서는 거의 매일이라 할 정도로 여론조사의 결과로 보도 혹은 인용하는 것을 볼 수 있다. 신문의 보도가 당위론이나 이래야 된다는 사설식 위주의 편집에서 탈피하여 국민의 생각과 여론을 보여주는 Precision Journalism의 지향이라는 면에의 반가운 일이라 할 수 있다.

그러나 여기에서 제기되는 큰 문제는 동일한 질문 문항으로 조사하였음에도 불구하고 한두달사이에 조사가 이상하게도 서로 크게 틀리고 있는 점이다. 또한 동일 시점에서 조사된것이라도 서로 상반된 결과가 발표되는 경우도 있어 독자를 어리둥절하게 만들고 있는 점이다.

예를 하나 들면 최근에 신문에서 비슷한 시점에서 행해진 조사결과를 발표한 것에서 한 신문에서는 민정당 선호도가 21.1%였고 또한 한 신문에서는 15.6%였다. 이들 두 가지 조사는 모두 우리나라 전국을 모집단으로 했으며 무작위표본추출에 가구방문 1:1개별 면접으로 표본오차는 2.8%라고 밝히고 있다.

그렇다면 많은 독자들은 어느 한 조사 혹은 둘 모두 틀렸다고 할 지 모르겠다.

독자의 혼란을 더욱 가중시키고 있는 이런 현상은 우리 신문에 보도 인용되고 있는 조사 결과 중 95%가 전국을 모집단으로 한 조사가 아니고 일부지역 혹은 일부계층만을 대상으로 한 조사라는 점이다. 지역과 계층이 다르면 조사결과가 서로 틀릴 수 있는 것 또한 어쩌면 당연하다고 하겠다.

그럼에도 불구하고 조사결과를 보도할 때 모집단이나 지역, 계층 등 조사방법을 명기하지 않거나 지나가는 식으로 다루면서 굵직한 제목을 뽑는 경우가 흔하다. 또한 조사의 개요를 자세히 명기하였더라도 읽는 독자는 그것을 헤아려 읽으려 하지 않고 서로 틀리는 조사결과만 보고 「조사는 역시 조사다」라고 틀릴 수 있는 것으로 불신감을 깊게 하고 있다.

최근에 신문에 도시영세민 가운데 70몇%가 1년에 한두번밖에 쇠고기를 못 먹는다는 조사결과가 크게 발표되고 신문사설에까지 등장한 바 있다. 신문은 조사대상이 된 도시영세민의 정의가 무엇인지도 밝히지도 않았고 또한 도시영세민이 서울시민의 몇 %라는 수치는 전혀 밝히지 않았다. 이럴 경우, 일반 독자를 정확하다는 여론조사의 수치(%)의 힘을 빌어 무엇을 증명하려 하는 선동에 휩싸이게 하는 것과 무엇이 다른지 알 수 없게 된다.

여론조사가 사실의 파악이라는 면에서 유행되는 것은 바람직하다고 하겠으나 수준 낮은 조사나 지켜야 될 조사의 룰을 제대로 잘 지키지 않은 조사가 많이 돌아다니는 것은 결코 바람직하다고 할 수 없다.

필자는 우리 주위에서 흔히 행해지는여론조사의 실시방법에 관해 다음과 같은 문제점을 지적하고 싶다.

1)母集團의 규정 문제이다.

조사의 목적과 용도에 따라 조사지역이나 대상자의 성, 나이 등 모집단(Population)의 규정을 달리하게 되었다.

그런데 앞서 지적했듯이 우리 신문에 발표되는 조사결과중에는 거의 대부분이 일부 지역이나 계층에 한정한 조사이다. 경우에 따라 조사가 그럴수도 있으나 전국을 모집단으로 해야함이 필요한 조사인 경우에도 일부지역을 대상으로 해놓고 마치 「한국인은」식으로 확대해석하거나 독자로 하여금 오해하도록 하는 경우도 너무 흔하다.

작년 대통령 선거를 앞두고 어떤 신문에서는 일부지역(5대도시)만 대상으로 선거 여론조사를 한 바 있는데 이들 결과는 5대도시 성향만을 반영하는 것임에도 분명히 밝혔음에도 독자로 하여금 전국결과로 오해하기 쉽게 한다. 이 점에서 선거 여론조사는 조사지역을 전국으로 함이 바람직했다고 보여진다. 특히, 기업에서 마아케팅조사를 함에 그 목적이 전국수요의 크기와 특성을 아는 것에 있었다면 모집단이 전국이여야 함이 필수적이다.

그럼에도 일부지역만을 조사해 놓고 수요크기를 추정할 때 전국적인 것으로 가중치를 주는 경우가 흔한데 이는 아주 위험한 일이 아닐 수 없다.

2)標本抽出이 문제이다.

표본추출(Sampling)에는 무작위추출, 할당추출, 유의추출, 임의추출 등 여러가지가 있다. 흔히 표본추출방식의 결정에는 조사목적과 조사비의 한도가 고려요인이 된다.

그러나 양적인 자료(Quantitative Data)가 필요할 때나 前과 後의 변화를 추적하는 추적조사(Tracking Study)에서는 임의추출방식은 여기서 논할 가치가 전혀 없겠으나 할당추출방식도 문제가 있음을 지적하고 싶다.

현재 학계나 조사기관에서 割當抽出法을 흔히 사용하고 있으며, 이 때문에 독자에게 큰 오해를 일으키는 일이 많으므로 이 점을 밝혀둔다.

할당추출법의 원리를 간단히 설명하면 시도읍면의 가구수에 따라 면접해야 할 수를 미리 결정하고 면접대상자를 性, 나이, 소득, 직업, 교육수준별로 모집단과 유사하게 배분하는 방식이다. 이 방법의 문제점은 첫째, 나이, 교육, 소득수준별로 미리 할당하는 것은 어렵지 않다고 하더라도 전국적으로 뽑힌 각 조사시점에서 미리 결정된 각 계층을 골고루 면접하기가 어려운 점이다. 둘째, 할당의 기준이 性, 나이, 교육수준의 세가지 기준으로만 할 경우 직업, 종교 등과 같은 요인은 제외된다. 1960년 미국대통령 선거때 할당추출법에 의해 조사한 기관은 큰 오차를 보였는데 그 원인은 카톨릭신자가 「케네디」에게 투표한 사실을 놓쳐버렸기 때문이다.

세째, 이 방법의 결점은 실사중에 크게 나타난다. 면접원이 면접대상을 할당표에 따라 마음대로 선정하기 때문에 면접원은 가장 면접하기 쉬운 사람을 면접하게 된다. 가기 싫은 지저분한 판자촌에 갈 필요도 없고, 사무실이나 길거리 아무데서나 손쉬운 상대를 골라 면접하면 그만인것이다. 교통이 나쁜 시골길을 걸어가서 면접해야하고, 면접해야 할 대상자가 집에 없는 경우 재방문해야 하는 방식과는 큰 차이를 보이는 것이다.

네째, 할당추출법으로 진행되었을 경우에는 완성된 질문지의 20~30%는 버리게 된다. 이 경우 어떤 질문지는 버리고 어떤 질문지는 유효하게 하느냐 하는 기준에 대한 불신이 문제인 것이다. 이 방법에 의한 조사결과가 「ㅇㅇ신문의 구독률이 가장 높다」고 발표될 경우 다른측에서 「無效로 해버린 20%의 질문지는 어디에 있는가? 버려진 20%는 다른 신문구독가구가 아닌가?」하고 웃어 넘기더라도 할말이 없게 된다. 그러나 구독률의 경우는 별문제가 아닌 것으로 한다 하더라도 그것이 정치가나 정책입안자에게 국민의 진정한 여론을 보여주는 기초자료를 제공하는 것이라면 문제의 심각성이 더해진다.

한편, 표본추출문제와 상관하여 「무작위 표본추출」하였다고 해 놓고 무작위표본추출법을 제대로 지키지 않고 조사한 경우를 들 수 있다.

올해 어느 신문사가 무작위 표본추출로 진행되었다고 하는 전국 조사의 특성을 살펴보면 중졸이하가 34%이고 고졸이상이 66%를 점하고 있다.

현재 우리나라에서 만 20세 이상의 성인남녀의 교육수준은 고졸이상이 약 44%이다. 이에서 본다면 그 조사는 애초부터 무작위표본추출법을 사용하지 않았거나 혹은 그 추출법을 지키지 않고 편의적으로 진행되었다고 밖에 볼 수 없다.

나이나 교육수준에 따라 정치적인 견해를 보이고 있는 우리 현실에서 조사의 신뢰성과 타당성 면에서 큰 문제를 제기하고 있다고 할 수 있다.

3)실사(Fieldwork)의 문제이다.

일반인들이 조사에 대해 갖고있는 의문 중에는 다음 두 가지가 있다.

「과연 면접원이 정직하게 면접했고 중립적인 태도를 지켜서 조사를 진행했는가?」

「과연 응답자가 솔직하게 대답했는가?」

조사는 처음부터 끝까지 오차가 생길 수 있다. 정확한 조사는 각 단계마다 생길 수 있는 오차를 어떻게 줄이느냐는 싸움이랄 수 있다.

필자가 보기로는 흔히 행해지는 우리 조사에서 큰 문제는 실사과정의 엄격성이 잘 지켜지지 않는 데 있다고 본다. 앞에서 필자는 우리나라에 조사전문기관의 수가 적고 영세하다고 지적한 바 있다. 조사용역이 정기적으로 이어지지 않기 때문에 전국실사조사망(Network)를 갖추거나 유지하기 어렵다고 보여지며 흔히 조사를 행하는 대학의 연구소도 예외는 아니다.조사가 있을 때 마다 실사지도원이 출장가거나 대학의 선후배나 사제지간의 관계로 실사를 위임하는 경우도 흔한 일이다.

면접원의 선발, 교육,실사체크 등 모두 전문성이 요구되는 일이며 실제 면접 여부를 체크하는 검증 또한 필수적이다. 외국에서는 실사만 전문적으로 담당하는 조사기관이 있다. 앞으로 우리나라에서 이런 조사기관이 생겨난다면, 대학연구소나 소형조사기관도 공동으로 수준높은 실사를 활용할 수 있게 될 것이다.

4.질문지작성의 문제이다.

묻는방식(wording)은 조사의 운명이라고 할 수 있다. 「아 다르고 어 다르다」란 옛 우리속담이 딱 맞는 세계가 바로 조사라고 할 수있다. 오랜 경험과 훈련이 필요하다. 여기서 필자는 흔히 주위에서 행해지는 질문지작성의 문제점을 다음과 같이 지적하고 싶다.

1)외국의 조사척도를 마구 사용하는 경우이다.

현재 우리나라의 조사역사가 짧고 또한 대학에서도 조사의 척도(Scale)에 관한 연구도 아주 빈약하다. 따라서 질문지를 작성하는 연구자들이 아무 검증받지않은 외국의 척도를 적당히 사용하고 있다.

그러나 필자의 경험으로는 이는 아주 위험한 경우가 많았다. 조사척도의 결정에는 조사지역,응답자. 면접방식 등이 고려되어야 한다고 본다.

예컨데, 전국을 무작위로 한 조사에서는 교육수준이 낮은 계층이 많기 때문에 소위 SD(Semantic Differene)방식 7점 척도 이상은 피하는 것이 좋다고 본다. 가급적 단순한 척도가 바람직하다고 하겠다.

2)조사인 위주의 질문작성 태도가 문제이다.

외국의 기업으로부터 의뢰받은 마아케팅조사를 할 때 마다 느끼는 점인데 조사인 위주의 질문(필자는 이를 Researcher Oriented Questionnaire라고 부른다)을 꼭 그대로 강요하는 점이다. 다국적 기업들이 여러 나라를 동일한 질문으로 조사하여 서로 비교한다는 점에서 불가피한 경우도 있겠으나 조사를 끝내고 자료를 분석해 보면 조사 전체의 질이 형편없이 파괴되어 있음을 경험하고 놀라워 한다.

요즈음 사회과학이나 마아케팅 전문학계지를 보면 MDS 등 고급통계조사기법을 사용한 분석을 많이 볼 수있다. 그러나 필자는 전국조사에서는 그러한 예쁜 그래프를 얻을 수 있으되 조사의 전체 質이 떨어질 가능성이 높다고 본다.

예컨데, MDS분석을 위해서는 많은질문의 개수를 사용하여야 하며 응답자도 마치 원숭이 놀림감이 되는 것 같은 느낌을 받기 때문에 건성으로 대답하는 경향이 있다.

3)사전조사가 기본이다.

질문의 순서,질문지의 길이, 사용한 단어 등 모두 중요하다. 신문에 나는 어려운 문어체, 개념투의 언어의 사용이나 응답시간이 30분 이내 불가능한 긴 질문지 등을 흔히 본다. 지난번 대통령 선거때 행해진 필자가 입수한 여러 선거여론조사의 질문지를 보면 너무 현학적이란 느낌이 드는 것이 많았다. 그리고 핵심적인 질문(예:당신은 어느 후보에게 투표하실 예정입니까?)을 60여개 의 질문을 하고 난 다음 맨 끝에 위치하게 한 것도 있었는데 이는 오차를 크게 한다고 본다.

각 후보들의 측면별 이미지 측정에 관한 문항에 20여개의 질문지를 사용하고, 과거의 나쁜 정치문제를 묻는 문항에 20개 등을 공부시킨 다음 핵심질문을 물을 경우 학습효과로 인해 진짜 투표성향을 읽을 수 없게 된다고 본다. 따라서 연구원들은 질문지를 확정하기 전에 사전조사(Pre-Test)를 행하여 여러가지 생길 수있는 응답자 Bias를 사전에 제거해야 할 것이다.

5.분석방법에서의 문제이다.

지난 대통령선거에서 언론기관 및 여러 단체들에서 선거여론조사를 한 바 있으며 후보지지도가 서로 다른 경우가 많았다.

한국갤럽조사 연구소는 투표 끝난 시점인 6시 1분에 대통령선거 예측을 한 바 있으며, 적은 오차로 적중한 바 있다. 그 원인의 하나로 「무응답」의 투표성향을 분류해 낸 분석방법을 꼽을 수 있을 것 같다. 참고로 이를 설명하고자 한다.

한국갤럽의 선거예측에서는 "만일 내일이 대통령선거날이라면 당신은 누구에게 투표하시겠습니까?"라는 질문에 「무응답/말할 수 없다」는 응답자의 지시성향을 판별분석(Discriminant Analysis)을 통해 분류하였다. 판별분석이란 여러가지 변인 즉, 性,나이, 교육수준,직업,소득수준,지역 등을 고려하여 특정후보의 지지성향을 추출해내고,「무응답 집단」이 어느 후보를 지지할 것인지를 예측하는 다변량 분석기법(Multivatiate Analysis)으로, 미국 Gallup을 비롯하여 구미제국의 선거예측조사에서 쓰이는 방식이다.

그동안 한국갤럽의 여러번에 걸친 전국조사에 의하면 다음과 같은 문제점이 제기되었다. 즉,누구에게 투표할 것인지를 물었을 때 노태우 후보의 지지자들은 「무응답/모르겠다」는 비중이 높았다.

야당 후보자를 많이 지지한 젊은 계층은 분명히 이야기하는데 비해 나이 많은 계층이나 저소득계층은 분명히 밝히지 않았기 때문이다. 구체적인 예를 하나 들면, 한국갤럽의 마지막 선거조사에서 보면 노태우 27.8%, 김영삼 26.0%, 김대중 25.1%, 김종필 7.9%, 기타 0.2%, 무응답 13.1%였다.

이를 보면 3후보가 백중지세임이 분명하고 표본오차 2.0% 범위안에 있기 때문에 어느 후보가 당선될 것인가 예측이 어렵다.

따라서 한국갤럽은 「무응답」의 성향을 분석하기 위해 판별분석을 행하여 마지막 선거예측을 행하였으며, 그 결과는 노태우 35.3%, 김영삼 28.4%, 김대중 27.5%, 김종필 8.3%, 기타 0.5%로 당선자를 기준으로 할 때 1.3%의 적은 오차로 적중할 수 있었다. 선거여론 조사의 분석에서 「무응답」은 단순히 빼버린다면 사실에 접근하기에 어렵다고 본다.

맺는 말

흔히들 우리나라의 조사수준이 구미제국에 비해 낮다고들 한다. 외국에 비해 조사전문기관이 수도 적고, 조사분야를 연구하는 분도 적어 그럴지 모른다는 생각도 든다.

그러나 필자는 이런 논의보다 앞으로 우리의 조사수준이 높아지기 위해서는 다음과같은 우리 풍토가 개선되는 데 있다고 보여진다.

첫째, 조사를 의뢰하는 쪽의 인식전환이다. 기업의 경영자나 정부기관,공공단체 등에서 합리적 의사결정이나 정책수립에 객관적인 자료가 필수적임을 알고 이를 잘 활용하려는 인식이 필요하다. 필요성은 인정하더라도 숫자(%)가 나오면 조사라는 안이한 생각, 예산의 제약 때문에 아무렇게나 조사를 요청하는 풍토가 하루빨리 사라져야 할 것이다.

둘째, 조사하는 쪽의 각성이다.

조사도 비지니스이고 용역을 맡아야 조직을 유지.발전시킬 수도 있는 일이다. 그러나 경쟁입찰에서 형편없이 낮은 금액으로 써내거나, 조사를 의뢰하는 쪽에서 수준낮은 조사를 의뢰하거나 예산의 제약때문에 조사의 질을 유지 할 수 없다고 느껴질 때는 거절 할 수 있어야 한다고 본다.

세째, 조사를 보도하는 쪽의 전문성이 요구된다.

오늘의 신문을 보면 가히 조사의 홍수라는 느낌까지 든다. 그러나 표본의 설계조사 접근 방법 등을 검토하지 않고 마구 큰 타이틀을 뽑아 다루고 있다. 이는 독자로 하여금 조사에 대한 회의나 불신을 가속화시키는 오보라고 할 수 있다. 먼저 수준 높은 조사인가, 아닌가를 검토할 줄 알아야 하고, 제대로 지켜야 하는 조사방법으로 진행되지 않은 수준낮은 조사는 보도하지 않을때 신문의 권위를 떨어뜨리지 않게 되고 일반인들도 조사를 신뢰하게 될 것이다.