ON THE DESIGN OF RANDOMIZED RESPONSE SAMPLING PLAN

Erdogan Gunel, West Virginia University

ABSTRACT

Consider a random response sampling plan where a respondent answers "yes" or "no" to a sensitive question Q: "Do you belong to group A?", or to the complementary question Q^c. The problem is to estimate the proportion of the sampled population who belong to group A. The information is elicited by asking each individual to choose question Q or Q^c by using a chance mechanism where the probability of choosing the question Q is p. We consider a Bayesian approach to choosing the value of p when n individuals are to be interviewed.

1. INTRODUCTION

We consider a problem in survey sampling where individuals are asked a sensitive question. If the individuals feel that answering could be used to their disadvantage they may choose not to respond. Randomized response sampling is an attempt to overcome such a nonresponse problem. For the mathematics and application of several randomized response sampling plans see Horvitz, Greenberg and Abernathy (1976), Greenberg, Kuebler, Abernathy and Horvitz (1971), Campbell and Joiner (1973). The original randomized response plan is due to Warner (1965). A Bayesian approach to Warner's randomized response model is considered by Winkler and Franklin (1979). In Warner's model a respondent answers "yes" or "no" to a sensitive question or to the complement of the question. For example, let group A be the population of women who had an abortion. Let question Q be "Do you belong to group A?" then the complementary question Q^c is "Do you belong to group A^c?" where A^c is the population of women who did not have an abortion. The information is elicited by asking each individual to choose question Q or Q^c by using a chance mechanism where the probability of choosing the question Q is p. This method assures the highest degree of confidentiality if $p = 1/2$ and as $p \to 0$ or $p \to 1$ the degree of confidentiality diminishes. The problem of interest is to estimate the proportion of the sampled population who belong to group A. We will denote this proportion by π.

In sampling surveys where randomization is not used the individuals interviewed are asked the question A and are given the opportunity to not respond if they wished to do so. It is assumed that if an individual chooses to respond then he/she does not falsify his/her answer.

We consider a Bayesian approach to choosing the value of p in the randomized response plan. The value of p is chosen by comparing the Bayes risks of the estimators of π under the randomized and voluntary response plans. The loss functions is taken to be the quadratic loss function

$$L[\pi, t(D)] = [\pi - t(D)]^2$$

where t(D) is the estimator of π based on data D. We take the prior distribution of π to be a Beta distribution and find the smallest value of $p(1/2<p<1)$, say p_0, such that the Bayes risk of the estimator of π under the voluntary response model is greater than the Bayes risk of the estimator of π under the randomized response plan. Then, it is more advantageous to use the randomized response plan when the p value is taken to be greater than p_0.

2. BAYES RISK

We first consider the voluntary response model. The question Q is asked to n individuals who are chosen at random. Suppose we have $n_{.2}$ nonrespondents and $n_{.1}$ respondents among whom n_{11} answered yes and n_{21} answered no. Clearly each individual belongs to one of the four mutually exclusive and exhaustive categories: (respond, belong to group A), (do not respond, belong to group A), (respond, do not belong to group A), (do not respond, do not belong to group A). Let θ_{11}, θ_{12}, θ_{21}, θ_{22} denote the probabilities of the four categories above respectively.

A mathematically tractable choice of the joint prior distribution for $\theta = (\theta_{11},\theta_{12},\theta_{21},\theta_{22})$ is the three-variate Dirichlet distribution with parameters a_{11}, a_{12}, a_{21}, a_{22}.

$$f(\theta_{11},\theta_{12},\theta_{21},\theta_{22}|a_{11},a_{12},a_{21},a_{22}) =$$

$$\frac{\Gamma(\sum_{ij} a_{ij})}{\prod_{ij} \Gamma(a_{ij})} \prod_{ij} \theta_{ij}^{a_{ij}-1}$$

if $\theta_{ij} > 0$, $\sum_{ij} \theta_{ij} = 1$ and zero elsewhere (Wilks, 1962). An important consequence of taking a Dirichlet prior distribution for θ is that

$$\pi = \theta_{11} + \theta_{12},\ Z_1 = \theta_{11}/(\theta_{11} + \theta_{12}),$$

$$Z_2 = \theta_{21}/(\theta_{21} + \theta_{22})$$

are independently distributed and

$$\pi \sim \text{Beta}\ (\alpha,\beta),\ Z_1 \sim \text{Beta}\ (a_{11},a_{12}),$$

$$Z_2 \sim \text{Beta}\ (a_{21},a_{22})$$

where $\alpha = a_{11} + a_{12}$, $\beta = a_{21} + a_{22}$. Moreover $\theta_{.1} = \theta_{11} + \theta_{21}$, $W_1 = \theta_{11}/(\theta_{11} + \theta_{21})$, $W_2 = \theta_{12}/(\theta_{12} + \theta_{22})$ are independently distributed and $\theta_{.1} \sim \text{Beta}\ (a_{.1}, a_{.2})$, $W_1 \sim \text{Beta}\ (a_{11},a_{21})$, $W_2 \sim \text{Beta}\ (a_{12},a_{22})$ where $a_{.1} = a_{11} + a_{21}$, $a_{.2} = a_{12} + a_{22}$. If $X \sim \text{Beta}\ (a,b)$ where $a > 0$, $b > 0$, the probability density function of X is

$$f_\beta(x|\ a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ x^{a-1}\ (1-x)^{b-1}$$

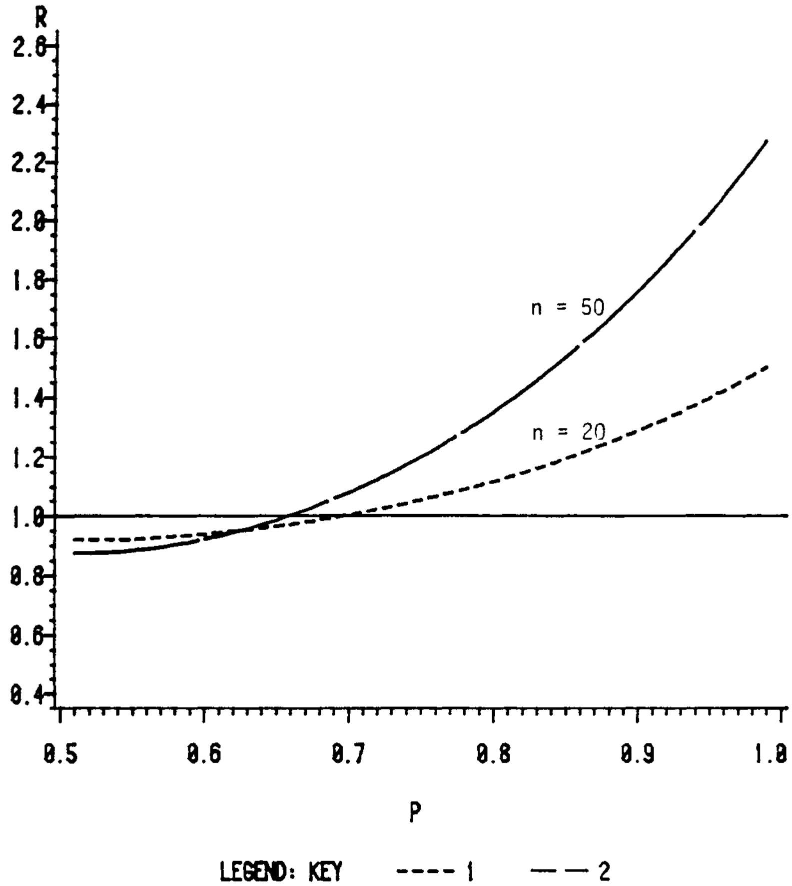
if $0<x<1$ and zero elsewhere.

Gunel (1985), obtained the Bayes risk ρ_V under the voluntary response sampling model

$$\rho_V = A[\alpha(\alpha+1) + (2\alpha+1)E(n_{1.}^*) + E(n_{1.}^{*\,2})$$

$$+E[n_{.2}(n_{.2}+a_{.2})]V(W_2)] - B[\alpha^2+2\alpha E(n_{1.}^*)+E(n_{1.}^{*2})]$$

where



1. n=20 2. n=50

$E(\theta_{.1}) = .2$ $V(\theta_{.1}) = .0051$

$E(\pi) = .333$ $V(\pi) = .0071$

$E(W_1) = .333$ $V(W_1) = .0317$

$E(W_2) = .333$ $V(W_2) = .0088$

$$n^*_{1.} = n_{11} + n_{.2}(a_{12}/a_{.2})$$

and

$$A = \frac{1}{(n+\alpha+\beta)(n+\alpha+\beta+1)} \qquad B = \frac{1}{(n+\alpha+\beta)^2}$$

$$E(n^*_{1.}) = n\, E(\theta_{1.})E(W_1)+n[1-E(\theta_{.1})]\, E(W_2)$$

$$E[n_{.2}(n_{.2}+a_{.2})] = n(n+\alpha+\beta)[V(\theta_{.1})+[1-E(\theta_{.1})]^2]$$

$$E(n_{1.}^{*2}) = C\, E(W_1)E(\theta_{.1}) + D\, E^2(W_1)E^2(\theta_{.1})$$
$$+ 2D\, E(W_1)\, E(\theta_{.1})[1-E(\theta_{.1})]\, E(W_2)$$
$$+ E^2(W_2)\{n^2 + C\, E(W_1)\, E(\theta_{.1}) + D\, E^2(W_1)\, E^2(\theta_{.1})$$
$$+ C\,[1-E(W_1)]\, E(\theta_{.1}) + D\,[1-E(W_1)]^2\, E^2(\theta_{.1})$$
$$- 2n^2\, E(\theta_{.1}) + 2D\, E(W_1)E(\theta_{.1})\,[1-E(W_1)][1-E(\theta_{.1})]\}$$

in which $C = \frac{n(n+\alpha+\beta)}{\alpha+\beta+1}$ $\qquad D = \frac{n(n-1)(\alpha+\beta)}{\alpha+\beta+1}$

$$E(\theta_{1.}) = a_{.1}/(\alpha+\beta), \quad V(\theta_{.1}) = \frac{a_{.1}a_{.2}}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$E(W_1) = a_{11}/a_{.1}$$

$$E(W_2) = a_{12}/a_{.2}, \quad V(W_2) = \frac{a_{12}a_{22}}{a_{.2}^2(a_{.2}+1)}$$

In the randomized response model, each individual answers the question Q or Q^c where the probability of choosing the question Q is p. Without loss of generality it can be assumed that $p > 1/2$. Suppose we have n individuals resulting in r "yes" answers. Gunel (1985), obtained the Bayes risk ρ_R by taking a Beta (α,β) prior distribution for π.

$$\rho_R = A[\alpha(\alpha+1) + (2\alpha+1)\, E(n_{1.}) + E(n^2_{1.})]$$
$$-B\,\{\alpha^2 + 2\alpha E(n_{1.}) + E_{r|n}[E^2(n_{1.}|n,r,p)]\}$$

where

$$E(n_{1.})\ n\ \alpha/(\alpha + \beta)$$

$$E(n^2_{1.}) = [n(n+\alpha+\beta)\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)]+n^2\alpha^2/(\alpha+\beta)^2$$

$$E_{r|n}[E^2_n(n_{1.}|r,n,p) = \sum_{r=0}^{n} \frac{\left[\sum_{n_{1.}=0}^{n} n_{1.} f_{\beta b}(n_{1.}|\alpha,\beta,n)f(r|n,n_{1.},p)\right]^2}{\sum_{n_{1.}=0}^{n} f_{\beta b}(n_{1.}|\alpha,\beta,n)f(r|n,n_{1.},p)}$$

$$f(r|n,n_{1.},p) = \sum_{j=\max(0,r-n+n_{1.})}^{\min(r,n_{1.})} \binom{n_{1.}}{j}\binom{n-n_{1.}}{r-j}$$
$$p^{n-n_{1.}-r+2j}\ (1-p)^{n_{1.}+r-2j}$$

and

$$f_{\beta b}(n_{1.}|\alpha,\beta,n) = \binom{n}{n_{1.}} \frac{B(n_{1.}+\alpha,\ n-n_{1.}+\beta)}{B(\alpha,\beta)}$$

in which $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

3. CHOOSING THE VALUE OF P

Since under both models $\pi \sim$ Beta (α,β) a priori, we can compare the Bayes risks (with respect to the beta prior) of the Bayes estimators of π under the voluntary and randomized response plans. We may examine the behavior of $R = \rho_V/\rho_R$ as a function of p under various prior opinions. It can be shown that R is an increasing function of p and for a given value of p, R is a decreasing function of the prior expectation of $\theta_{.1}$, the probability of responding under the voluntary response model.

Let us define p_0 as follows: $R>1$ if $p>p_0$. Then p_0 is the smallest value of p for which the randomized response plan is superior to the voluntary response plan. To find p_0, one has to plot R versus p. As an illustration consider the case where $a_{11} = 2$, $a_{12} = 8$, $a_{21} = 4$, $a_{22} = 16$, then $\alpha = 10$, $\beta = 20$ and we have these following prior expectations and variances: $E(\theta_{.1})=.2$, $E(\pi) = E(W_1) = E(W_2) = .333$, $V(\theta_{.1}) = .0051$, $V(\pi) = .0071$, $V(W_1) = .0317$, $V(W_2) = .0088$. We plot R versus p for n = 50 and n = 20. From the graph we see that when n = 50 we have $p_0 = .66$ and for n = 20 we have $p_0 = .7$.

REFERENCES

Campbell C., Joiner B., (1973). "How to Get the Answer Without Being Sure You've Asked the Question." The American Statistician, 27, 229-231.

Greenberg, B. G., Kuebler, R. T., Abernathy, J. R., Horvitz, D. G., (1971). "Application of Randomized Response Technique in Obtaining Quantitative Data." Journal of the American Statistical Association, 66, 243-250.

Gunel, E., (1985). "A Bayesian Comparison of Randomized and Voluntary Response Sampling Models." To appear in Communication in Statistics, Part A: Theory and Methods.

Horvitz, D. G., Greenberg, B. G., Abernathy, J. R., (1976). "Randomized Response. A Data Gathering Device for Sensitive Questions." International Statistical Review, 44, 181-196.

Warner, S. L., (1965). "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." Journal of the American Statistical Association, 60, 63-69.

Wilks, S. S., (1962). Mathematical Statistics. J. Wiley.

Winkler, R. and Franklin, A., (1979). "Warner's Randomized Response Model. A Bayesian Approach." Journal of the American Statistical Association, 74, 207-214.