PROBABILITY QUOTA SAMPLING OF A DICHOTOMOUS VARIABLE

Myron J. Katzoff, U.S. Census Bureau

This paper presents some preliminary results on quota sampling. Let P denote the proportion of individuals that belong to a specific subset of a finite population for which a quota has been prescribed. New results include the exact expression for the variance of the UMV estimator of P and upper and lower bounds on the variance.

In the special case that only one subset of the population (and its complement) is under examination, there is an obvious relationship between quota sampling and sequential estimation of the proportion defective, P, in a finite population. The sequential estimation problem has been investigated by Knight (AMS, 1965, 1494-1503) and Olkin and David (Tech. Rept. 5, 1956, University of Chicago) using geometric methods. For the quota-type stopping rule, both papers give the unique unbiased minimum variance estimator for P and Knight gives an unbiased estimator for its variance.

Introduction

Quota sampling has been used extensively in opinion and market research surveys. Generally, the applications of quota sampling have involved the nonrandom selection of units at the last stage in a multistage design. For this reason, complex theoretical problems have been encountered in the development of formulas for estimating the sampling variability of quota sample estimates. The mathematical problems and advantages and disadvantages of quota sampling are discussed in Cochran [3], Stephan and McCarthy [8], and Sudman [9].

As discussed here, quota sampling will refer to a type of probability sampling which could be used for the purpose of the sequential estimation of the proportion of individuals, P, who belong to a specific subset, E, of a finite population. The paper examines the problems of obtaining an exact expression for the variance of the unique unbiased (and, therefore, MVU) estimator for P and upper and lower bounds on its variance when selections are made randomly one-at-a-time without replacement and sampling is continued until a quota of k members of E is satisfied. The stopping rule is the special case of one considered by Bershad and Perkins [1] at the Census Bureau.

The sequential estimation of P has been investigated by Knight [4] and Olkin and David [6] using geometric methods. Both papers give the unique unbiased estimator for P. Knight gives an unbiased estimator for the variance of the estimator of P. Olkin and David have demonstrated the uniqueness of the unbiased estimator for P by a completeness argument.

Preliminaries

If the above-described sampling plan is carried out, the sample size, n, will be a random variable which is distributed according to the negative hypergeometric distribution

$$p(n;k,N,P) = \begin{cases} \binom{n-1}{k-1}\binom{N-n}{NP-k}\Big/\binom{N}{NP} & \text{for } n=k,k+1,\dots,\ NQ+k \\ 0, & \text{otherwise} \end{cases}$$

where $Q=1-P$ and $3\leq k\leq NP$. (Wilks [11] has called $p(n;k,N,P)$ the probability mass function for a hypergeometric waiting-time distribution.) In the sequel, the operations involving $p(n;k,N,P)$ make use of the combinatorial identities and results on hypergeometric series which are summarized in this section.

Theorem 1: Let j be an integer. Then

$$\sum_{x=k}^{NQ+k}\binom{x-j-1}{k-j-1}\binom{N-x}{NP-k} = \binom{N-j}{NP-j}$$

proof: For $|t| < 1$, it follows from a generalization of the binomial theorem, after making suitable transformations of the variables of summation, that

$$\text{(i)}\quad (1-t)^{-(k-j)} = \sum_{x=k}^{\infty}\binom{x-j-1}{k-j-1}t^{x-k}$$

$$\text{(ii)}\quad (1-t)^{-(NP-k+1)} = \sum_{y=-\infty}^{NQ+k}\binom{N-y}{NP-k}t^{NQ-y+k}$$

and

$$\text{(iii)}\quad (1-t)^{-(NP-j+1)} = \sum_{z=0}^{\infty}\binom{NP-j+z}{NP-j}t^{z}$$

Here the symbol $\binom{n}{r}$ stands for the general binomial coefficient defined by Riordan [7, pp. 4,5]. The remaining details can be found in Wilks [11, p. 141].

Corollary 1.1: If j is any integer and r is any nonnegative integer

$$\sum_{x=k}^{NQ+k}\binom{x-j-r-1}{k-j-1}\binom{N-x}{NP-k} = \binom{N-j-r}{NP-j}\ .$$

The series

$$F(\alpha,\beta,\gamma;\delta,\varepsilon;x) = 1+\sum_{n=1}^{\infty}\frac{x^n}{n!}\prod_{k=0}^{n-1}\frac{(\alpha+k)(\beta+k)(\gamma+k)}{(\delta+k)(\varepsilon+k)}$$

is called a hypergeometric series. In the case that $\gamma = \varepsilon = 1$, the series may be denoted by $F(\alpha,\beta;\delta;x)$. The convergence of $F(\alpha,\beta;\delta;1)$ is examined in Whittaker and Watson [10], where it is shown, in particular, that if α,β, and δ are real numbers such that

$\delta - \alpha - \beta > 0$, then

$$F(\alpha,\beta;\delta;1) = \frac{\Gamma(\delta)\Gamma(\delta-\alpha-\beta)}{\Gamma(\delta-\alpha)\Gamma(\delta-\beta)}$$

In the situations dealt with below, $\alpha,\beta,\gamma,\delta$, and ε are all integers, with $\alpha=1$, δ and ε positive integers, and β and γ positive or negative integers. It should be noted that if either β or γ are negative, the hypergeometric series has a finite number of terms.

Estimators

By making direct use of the form of $p(n;k,N,P)$ and the standard definition of a maximum likelihood estimator for P, $\hat{P}_{ml}$, satisfies

$$\frac{k(N+1)-n}{n} \leq N\hat{P}_{ml} \leq \frac{k}{n}(N+1) .$$

On the other hand, it can be seen, by the application of theorem 1, that

$$\hat{P} = \frac{k-1}{n-1}$$

is the unbiased estimator for P. If one also observes, again, using theorem 1, that when $k \geq 3$,

$$E\left[\frac{(k-1)(k-2)}{(n-1)(n-2)}\right] = P\,\frac{(NP-1)}{N-1} ,$$

it follows that

$$E\left[\frac{(N-1)(k-1)(k-2)}{(n-1)(n-2)}\right] = NP^2 - E\left(\frac{k-1}{n-1}\right)$$

The last equation provides us with an unbiased estimator for P^2 which can be used to show that an unbiased estimator of the variance of $\hat{P}$ is

$$v(\hat{P}) = \frac{\hat{P}(1-\hat{P})(N-n+1)}{N(n-2)} .$$

This is Knight's estimator of the variance of $\hat{P}$.

Variance and Bounds on the Variance of $\hat{P}$

Theorem 2: Under the sampling plan described above, the variance of $\hat{P}$ is given by

$$\mathrm{Var}(P) = PF(1,k-1,-NQ;k,NP;1) - P^2$$

proof:

Observe that

$$\frac{k-1}{n-1} = \frac{\Gamma(k)\Gamma(n-1)}{\Gamma(k-1)\Gamma(n)} = F(1,-(n-k);k,1) ,$$

a hypergeometric series with n-k+1 terms. Thus, after some simplification one obtains

$$\frac{k-1}{n-1} = \sum_{r=0}^{n-k} (-1)^r \frac{(n-k)_r}{(k+r-1)_r} ,$$

where the notation $(H)_t$ is defined by

$$(H)_t = \frac{!H}{(H-t)!}$$

for H, t nonnegative integers with $H \geq t$. Therefor,

$$\binom{N}{NP} E\left[\left(\frac{k-1}{n-1}\right)^2\right] = \sum_{n=k}^{NQ+k} \sum_{r=o}^{n-k} \left(\frac{k-1}{n-1}\right)\binom{n-1}{k-1}\binom{N-n}{NP-k} (-1)^r \frac{(n-k)_r}{(k+r-1)_r}$$

$$= \sum_{r=o}^{NQ} \frac{(-1)^r}{(k+r-1)_r} \sum_{n=r+k}^{NQ+k} (n-k)_r \binom{n-2}{k-2}\binom{N-n}{NP-k}$$

upon reversing the order of summation.

But

$$(n-k)_r\binom{n-2}{k-2} = (k+r-2)_r\binom{n-2}{k+r-2}$$

so that

$$\binom{N}{NP} E\left[\left(\frac{k-1}{n-1}\right)^2\right] = \sum_{r=o}^{NQ} (-1)^r \frac{(k+r-2)_r}{(k+r-1)_r} \sum_{n=r+k}^{NQ+k} \binom{n-2}{k+r-2}\binom{N-n}{NP-k}$$

$$= \sum_{r=0}^{NQ} (-1)^r \frac{k-1}{k+r-1} \binom{N-1}{NP+r-1} ,$$

by corollary 1.1 with $j = -(r-1)$. This implies that

$$E\left[\left(\frac{k-1}{n-1}\right)^2\right] = P \sum_{r=o}^{NQ} (-1)^r \frac{k-1}{k+r-1} \frac{(NQ)_r}{(NP+r-1)_r}$$

$$= PR(1,k-1,-NQ;k,NP;1)$$

In theorem 3, an upper bound on the variance of $\hat{P}$ is obtained by an argument similar to one in Mikulski and Smith [5]. Lower bounds can be found with the aid of the approach taken by Chapman and Robbins in [2] for finding minimum variance bounds without requiring that regularity assumptions are satisfied. It should be noted that, if P remains constant as $N \to \infty$, the upper and lower bounds on $\mathrm{Var}(\hat{P})$ approach those obtained by Mikulski and Smith for the negative bionomial distribution.

Theorem 3: Under the above sampling plan,

$$\frac{Q(NP+2)(NP-k+1)}{N(N+1)k} \leq \mathrm{Var}(\hat{P}) \leq \frac{PQ(NP-k+1)}{(NP-1)+(k-2)(N-1)}$$

proof:

(i) upper bound:

$$E\left[\left(\frac{k-1}{n-1}\right)^2\right] = E\left[\frac{(k-1)^2}{(n-1)(n-2)}\right] - E\left[\frac{(k-1)^2}{(n-1)(n-2)} - \frac{(k-1)^2}{(n-1)^2}\right]$$

But

$$E\left[\frac{(k-1)^2}{(n-1)(n-2)}\right] = \frac{P(k-1)(NP-1)}{(k-2)(N-1)},$$

by theorem 1; and, letting $N' = N-2$, $n' = n-2$, $k' = k-2$, $P' = (NP-2)/N-2)$, and $Q' = 1-p'$,

$$E\left[\frac{(k-1)^2}{(n-1)(n-2)} - \frac{(k-1)^2}{(n-1)^2}\right] = \binom{N}{NP}^{-1} \frac{k-1}{k-2} \sum_{n=k}^{NQ+k} \frac{1}{n-1} \binom{n-3}{k-3}\binom{N-n}{NP-k}$$

$$= \frac{P(NP-1)(k-1)}{(N-1)(k-2)} \sum_{n'=k'}^{N'Q'+k'} \frac{1}{n'+1} \frac{\binom{n'-1}{k'-1}\binom{N'-n'}{N'P'-k'}}{\binom{N'}{N'P'}},$$

$$\geq \frac{P(NP-1)(k-1)}{(N-1)k-2)} \left[1 + \frac{(k-2)(N-1)}{(NP-1)}\right]^{-1}$$

by Jensen's Inequality because of the convexity of $g(x) = \frac{1}{1+x}$ for $x > o$. (Here, n' is distributed in accordance with $p(n'; k', N', P')$.)

Consequently, after some algebra, one finds that

$$E\left[\frac{k-1}{n-1}\right]^2 \leq \frac{P(k-1)(NP-1)}{(NP-1) + (k-2)(N-1)},$$

which implies that

$$\mathrm{Var}(\hat{P}) \leq \frac{PQ\ (NP-k+1)}{(NP-1) + (k-2)(N-1)}$$

(ii) lower bound:

Following the notation of Chapman and Robbins [2], for μ take the counting measure. It seems convenient to work with the following form of the probability mass function for n:

$$f(x,P) = \begin{cases} \dfrac{\binom{k+x-1}{k-1}\binom{N-k-x}{NP-k}}{\binom{N}{NP}}, & \text{for } x = 0,1,\ldots, NQ \\ 0, & \text{otherwise} \end{cases}$$

A new P must be an integral multiple of 1/N to belong to the parameter space and $S(P) = \{0,1,2,\ldots, NQ\}$ in the notation of [2]. Defining $h = t/N$, for $t=1,2,\ldots, NQ$,

$$S(P+h) = \{0,1,2,\ldots, NQ-t\} \subset S(P)$$

and

$$f(x,P+h) = \frac{\binom{k+x-1}{k-1}\binom{N-k-x}{NP+t-k}}{\binom{N}{NP+t}}, \text{ for } x \text{ in } S(P+h).$$

Thus

$$J = J(P,h) = \frac{1}{h^2}\left\{\left[\frac{f(x,P+h)}{f(x,P)}\right]^2 - 1\right\}$$

$$= \left(\frac{N}{t}\right)^2 \left\{\left[\frac{\binom{N-k-x}{NP+t-k}\binom{N}{NP}}{\binom{N-k-x}{NP-k}\binom{N}{NP+t}}\right]^2 - 1\right\}$$

Taking the expectation of J with respect to f(x,P), denoted by $E(J|P)$, one first establishes that

$$E([\ ]^2|P) = \frac{\binom{N}{NP}}{\binom{N}{NP+t}\binom{NP+t-k}{t}} \sum_{x=o}^{NQ} \frac{\binom{NQ-x}{t}\binom{k+x-1}{k-1}\binom{N-k-x}{NP+t-k}}{\binom{N}{NP+t}}$$

where the brackets refer to the bracketed term in J, from which it follows that if t=1, then

$$E(J|P) = \frac{N(N+1)\ k}{Q(NP-k+1)(NP+2)}$$

which implies that

$$\mathrm{Var}(\hat{P}) \geq \frac{Q(NP-k+1)(NP+2)}{N(N+1)\ k}$$

REFERENCES

[1] Dershad, M.A. and Perkins, W.M. "An Unbiased Step Rule for a Random Sample of Variance Size," Tech. Note 3, 1970, U.S. Department of Commerce, Bureau of the Census, pp. 1-4.

[2] Chapman, D.G. and Robbins, H. "Minimum Variance Estimation Without Regularity Conditions," Annals of Mathematical Statistics, vol. 22, 1951, pp. 581-586.

[3] Cochran, W.G. (1963). Sampling Techniques, second edition. John Wiley and Sons, New York.

[4] Knight, W. "A Method of Sequential Estimation Applicable to the Hypergeometric, Binomial, Poisson, and Exponential Distributions," Annals of Mathematical Statistics, vol. 36, 1965, pp. 1494-1503.

[5] Mikulski, P.W. and Smith, P.J. "A Variance Bound for Unbiased Estimation in Inverse Sampling," Biometrika, vol. 63, 1976, pp. 216-217.

[6] Olkin, I. (with Herbert David) "Sequential Estimation from a Finite Population," Tech. Report 5, 1956, University of Chicago.

[7] Riodan, J. (1958). An Introduction to Combinatorial Analysis. John Wiley and Sons, New York.

[8] Stephan, F.F. and McCarthy, P.J. (1958). Sampling Opinions. John Wiley and Sons, New York.

[9] Sudman, S. (1967). Reducing the Cost of Surveys. Aldine Publishing Co., Chicago.

[10] Whittaker, E.T. and Watson, G.N. (1927). A Course in Modern Analysis. Cambridge University Press, New York.

[11] Wilks, S.S. (1962). Mathematical Statistics. John Wiley and Sons, New York.