Ann. Inst. Statist. Math.
Vol. 47, No. 4, 637–644 (1995)

ON ESTIMATING DOMAIN TOTALS OVER A SUBPOPULATION*

KUN HE

Department of Mathematics, University of Kansas, Lawrence, KS 66045, U.S.A.

(Received June 15, 1994; revised March 3, 1995)

Abstract. A simple random sample is drawn over a finite population which is composed of several subpopulations. Each subpopulation consists several domains. The minimax estimator under squared error loss function for the domain totals over a subpopulation is derived, in which the number of sample units falling into the subpopulation is random.

Key words and phrases: Random sample size, subpopulation, domain totals.

1. Introduction

The purpose of this paper is to derive a minimax estimator of domain totals over a subpopulation (or stratum). Assume that a finite population π has k subpopulations (or strata) π_i with T_i units, $i = 1, \dots, k$. In each subpopulation π_i, there are l_i domains with $(U_{i1}, \dots, U_{il_i})$ as the domain totals. A simple random sample is drawn over the whole population π, and the purpose is to estimate $(U_{i1}, \dots, U_{il_i})$, $i = 1, \dots, k$ by using sample units falling into the subpopulation π_i, $i = 1, \dots, k$. Note that sample sizes n_i, $i = 1, \dots, k$ in each subpopulation are random in such situation.

The above problem arises frequently in practice. For example, if two subpopulations are formed by males and females, estimates of totals of unemployment for males and females separately may be wanted. The subpopulations may be formed by different age groups, and estimates of totals of voting for Republican, Democrat and Independent separately in each group may be wanted. If strata are formed by geographic locations, separate estimates of totals of unemployment for males and females over each location may be wanted. The basic formulas were given by Yates and Grundy (1953), and some further contributions were made by Durbin (1958) and Hartley (1959). There is one section in dealing with such problem in Cochran ((1977), p. 142).

Notice that the sample sizes n_i, $i = 1, \dots, k$ in each subpopulation are random which is common in finite sampling. For example, Haldane (1945) discusses the problem of inverse sampling with replacement, and Guenther (1969) discusses the

* Supported by the General Research Fund of the University of Kansas.

inverse sampling without replacement. Hill (1968, 1979) gives Bayes estimator both with replacement and without replacement. He (1993, 1995) discusses the estimation of the population mean and the stratum means vector with random sample sizes over a super-population.

The totals T_i, $i = 1, \ldots, k$ of the subpopulations are assumed to be known, and consequently the distributions of the n_i is assumed to be known. Thus n_i is an ancillary statistic.

The minimax estimator of domain totals $(U_{i1}, \ldots, U_{il_i})$, $i = 1, \ldots, k$ will be derived in the next section by using similar techniques as those in Trybula (1958), Hill (1968, 1979) and He (1990). A brief discussion will be given in the end.

2. The minimax estimator

For notational simplicity subscripts will be omitted and the results will be presented only for one subpopulation.

Assume that the subpopulation consisting of T units has been classified into l domains with the i-th domain containing U_i units ($i = 1, \ldots, l$). The number of sample units falling into the subpopulation is n, in which x_i units belonging to the i-th domain, $i = 1, \ldots, l$, are observed respectively. We are interested in the estimation of the domain totals $U = (U_1, \ldots, U_l)$, which is the parameters of a multivariate hypergeometric distribution.

Let $X = (X_1, \ldots, X_l)$ have the distribution

$$P(X_1 = x_1, \ldots, X_l = x_l) = \frac{\binom{U_1}{x_1} \cdots \binom{U_l}{x_l}}{\binom{T}{n}}, \tag{2.1}$$

where $\sum_1^l X_i = n$ and

$$U \in \Theta = \left\{ (\theta_1, \ldots, \theta_l) : \theta_i \geq 0(\text{integer}), i = 1, \ldots, l; \sum_1^l \theta_i = T \right\}. \tag{2.2}$$

Suppose that we want to find a minimax estimator of U under the squared error loss function

$$L(a, U) = \sum_1^l (a_i - U_i)^2, \tag{2.3}$$

where

$$a \in \mathcal{A} = \{(a_1, \ldots, a_l) : a_i \geq 0, i = 1, \ldots, l\}. \tag{2.4}$$

Let $\delta : \mathcal{R}^l \to \mathcal{A}$ be any nonrandomized estimator and, as usual, denote its risk by

$$R(\delta, U) = EL(\delta, U).$$

When n is fixed, the minimax estimator of U is found by Trybula (1958),

$$\delta_0(x,n)=\left(T\frac{x_1+\frac{1}{l}\sqrt{n\frac{T-n}{T-1}}}{n+\sqrt{n\frac{T-n}{T-1}}},\ldots,T\frac{x_l+\frac{1}{l}\sqrt{n\frac{T-n}{T-1}}}{n+\sqrt{n\frac{T-n}{T-1}}}\right). \tag{2.5}$$

Now suppose that n is the value of a random variable N, having a (known) distribution. Let

$$\Pr(N=n)=f(n),\qquad n=1,\ldots,T. \tag{2.6}$$

Thus N is an ancillary statistic.

The minimax estimator of U is given in the following theorem.

THEOREM 2.1. *If*

$$f(T-1)+f(T)<1, \tag{2.7}$$

then the estimator

$$\delta_1(x,n)=\left(\frac{(T+la_0)x_1+(T-n)a_0}{n+la_0},\ldots,\frac{(T+la_0)x_l+(T-n)a_0}{n+la_0}\right) \tag{2.8}$$

is minimax and admissible and

$$\begin{aligned}&\max_U R(\delta_1(X,N),U)\\&\quad=R(\delta_1(X,N),U)\\&\quad=E^N\left(\frac{1}{(N+la_0)^2}\left(\frac{(T+la_0)^2(T-N)N}{T-1}-(T-N)^2la_0^2\right)\right),\end{aligned} \tag{2.9}$$

where a_0 is the unique solution to

$$E^N\left(\frac{1}{(N+la)^2}\left(\frac{-(T+la)^2N(T-N)}{T^2(T-1)}+\left(\frac{(T-N)la}{T}\right)^2\right)\right)=0. \tag{2.10}$$

If (2.7) *is false, then the estimator $\delta_0(x,n)$ defined in* (2.5) *is still minimax and admissible and*

$$\max_U R(\delta_0(X,N),U)=R(\delta_0(X,N),U)=\left(1-\frac{1}{l}\right)f(T-1). \tag{2.11}$$

PROOF. The conjugate prior distribution of $(U_1,\ldots,U_l)$ is proportional to

$$\binom{u_1+a_1-1}{u_1}\cdots\binom{u_l+a_l-1}{u_l}\propto\frac{\Gamma(a_1+u_1)\cdots\Gamma(a_l+u_l)}{u_1!\cdots u_l!}.$$

If n is fixed, the Bayes estimator of U_i is

$$\delta_i(x,n) = E(U_i \mid X_1 = x_1, \ldots, X_l = x_l; N = n) = \frac{\sum_u u_i \prod_{j=1}^l \binom{u_j}{x_j} \dfrac{\Gamma(a_j + u_j)}{u_j!}}{\sum_u \prod_{j=1}^l \binom{u_j}{x_j} \dfrac{\Gamma(a_j + u_j)}{u_j!}},$$

where $\sum_u$ is over $\{u_1 \geq x_1, \ldots, u_l \geq x_l; u_1 + \cdots + u_l = T\}$. It is easily seen that $\delta_i(x,n)$ is equivalent to

$$\frac{\sum_u u_i \prod_{j=1}^l \dfrac{\Gamma(a_j + u_j)}{(u_j - x_j)!}}{\sum_u \prod_{j=1}^l \dfrac{\Gamma(a_j + u_j)}{(u_j - x_j)!}} = \frac{\sum_v [(a_i + x_i + v_i) - a_i] \prod_{j=1}^l \dfrac{\Gamma(a_j + x_j + v_j)}{v_j!}}{\sum_v \prod_{j=1}^l \dfrac{\Gamma(a_j + x_j + v_j)}{v_j!}}$$
$$= \frac{\sum_v \Gamma(a_i + x_i + v_i + 1)/v_i! \prod_{j \neq i}^l \dfrac{\Gamma(a_j + x_j + v_j)}{v_j!}}{\sum_v \prod_{j=1}^l \dfrac{\Gamma(a_j + x_j + v_j)}{v_j!}} - a_i,$$

where $\sum_v$ is over $\{v_1 \geq 0, \ldots, v_l \geq 0; v_1 + \cdots + v_l = T - n\}$. Note that

$$\begin{aligned}
&\sum_v \frac{(T-n)!\Gamma(b_1 + v_1) \cdots \Gamma(b_l + v_l)}{v_1! \cdots v_l! \Gamma(T - n + \sum_1^l b_j)} \\
&= \int \cdots \int_{p_1 \geq 0, \ldots, p_l \geq 0; p_1 + \cdots + p_l = 1} p_1^{b_1 - 1} \cdots p_l^{b_l - 1} \\
&\quad \cdot \sum_v \frac{(T-n)!}{v_1! \cdots v_l!} p_1^{v_1} \cdots p_l^{v_l} dp_1 \cdots dp_l \\
&= \int \cdots \int_{p_1 \geq 0, \ldots, p_l \geq 0; p_1 + \cdots + p_l = 1} p_1^{b_1 - 1} \cdots p_l^{b_l - 1} dp_1 \cdots dp_l \\
&= \frac{\Gamma(b_1) \cdots \Gamma(b_l)}{\Gamma(\sum_1^l b_j)}.
\end{aligned}$$

Applying this fact to the numerator of δ_i with $b_j = a_j + x_j$ for $j \neq i$ and $b_i = a_i + x_i + 1$, $i = 1, \ldots, l$; and to the denominator of δ_i with $b_j = a_j + x_j$ one gets

$$\delta_i(x,n) = \frac{(a_i + x_i)(T + \sum_1^l a_j)}{n + \sum_1^l a_j} - a_i = \frac{(T + \sum_1^l a_j)x_i + (T - n)a_i}{n + \sum_1^l a_j}.$$

This estimator is still Bayes if N is random.

The prior with $a_1 = \cdots = a_l = a$ will be used, and the Bayes estimator $\delta(x,n)$ then reduces to

$$\delta(x,n) = \left(\frac{(T + la)x_1 + (T - n)a}{n + la}, \ldots, \frac{(T + la)x_l + (T - n)a}{n + la} \right).$$

By the strict convexity of the loss function, this is the unique Bayes estimator for the given prior. Hence it is admissible (Lehmann (1983), p. 263). Furthermore, if one can find $a_0 > 0$ such that $R(\delta(X,N),U)$ is constant as a function of U, then $\delta(x,n) = \delta_1(x,n)$ is a minimax estimator (Lehmann (1983), p. 250).

Note that

$$\begin{aligned} R(\delta(X,N),U) &= E^{N,X} L(\delta(X,N),U) \\ &= E^N \left[E^X \sum_1^l \left(\frac{(T+la)X_i + (T-N)a}{N+la} - U_i \right)^2 \mid N \right] \\ &= E^N \left[E^X \frac{1}{(N+la)^2} \sum_1^l (AX_i + B - CU_i)^2 \mid N \right], \end{aligned}$$

where $A = T + la$, $B = (T-N)a$, $C = (N+la)$. Since $(X_1, \ldots, X_l)$ follows a hypergeometric, one knows

$$E(X_i \mid N) = N\frac{U_i}{T},$$

and

$$E(X_i^2 \mid N) = \frac{N(T-N)}{T^2(T-1)} U_i(T-U_i) + \left(N\frac{U_i}{T}\right)^2.$$

By using these facts and $\sum_1^l U_i = T$, after some manipulations, one gets

$$\begin{aligned} R(\delta(X,N),U) = E^N \Bigg[\frac{1}{(N+la)^2} \sum_1^l \Bigg(\Bigg(& \frac{-A^2N(T-N)}{T^2(T-1)} \\ & + \frac{A^2N^2}{T^2} + C^2 - \frac{2ACN}{T} \Bigg) U_i^2 \\ & + \left(\frac{A^2N(T-N)}{T(T-1)} + \frac{2ABN}{T} - 2BC \right) U_i + B^2 \Bigg) \Bigg] \\ = E^N(g_1(N,a)) \sum_1^l U_i^2 + E^N(g_2(N,a)), & \end{aligned}$$

where

$$g_1(N,a) = E^N \left(\frac{1}{(N+la)^2} \left(\frac{-(T+la)^2 N(T-N)}{T^2(T-1)} + \left(\frac{(T-N)la}{T} \right)^2 \right) \right),$$

and

$$g_2(N,a) = \frac{1}{(N+la)^2} \left(\frac{(T+la)^2(T-N)N}{T-1} - (T-N)^2 la^2 \right).$$

If (2.7) holds, one gets

$$\lim_{a\to 0} E^N[g_1(N,a)] = -E^N\left(\frac{(T-N)}{N(T-1)} \right) < 0,$$

and

$$\lim_{a\to\infty} E^N[g_1(N,a)] = E^N\left(\frac{-N(T-N)}{T^2(T-1)} + \frac{N^2}{T^2} + 1 - \frac{2N}{T}\right)$$
$$= E^N\left(\frac{(T-N)(T-N-1)}{T(T-1)}\right) > 0.$$

Thus, by the intermediate-value theorem, there exists an $a_0 > 0$ which satisfies equation (2.10). Therefore, δ_1 is minimax and admissible. The value a_0 must be unique. If there exist two distinct admissible estimators having the same constant risk, by the strict convexity of the loss function and Jensen inequality, the average of the two estimators dominates the two admissible estimators; this is a contradiction. Consequently, δ_0 is not a minimax estimator since δ_1 has constant risk

$$R(\delta_1(X,N),U) = E^N(g_2(N,a_0))$$

and is admissible.

If (2.7) is not valid, then

$$\delta_2(x,n) = \left(\frac{x_1+(T-n)}{l}, \dots, \frac{x_l+(T-n)}{l}\right)$$

is the unique Bayes estimator for the prior distribution of U defined by

$$\frac{T!}{u_1!\cdots u_l!} s_1^{u_1}\cdots s_l^{u_l}, \qquad 0 \le s_i \le 1,\ i=1,\dots,l,\ \ \sum_1^l u_i = T.$$

By the same arguments as above, this estimator is minimax and admissible, and

$$R(\delta_2(X,N),U) = E^N\left(\frac{1}{l^2}\left(\frac{l^2(T-N)N}{T-1} - (T-N)^2 l\right)\right).$$

Note that in this case, $f(T-1) + f(T) = 1$, then

$$R(\delta_2(X,N),U) = \left(1-\frac{1}{l}\right) f(T-1).$$

Notice that the risk of $\delta_0(x,n)$ in this case is

$$R(\delta_0(X,N),U) = E^{N,X}\left(\sum_1^l (\delta_{0,i} - U_i)^2\right)$$
$$= E^X(\delta_{0,T-1} - U_{T-1})^2 f(T-1) + E^X(\delta_{0,T} - U_T)^2 f(T).$$

It is easily seen that $E^X(\delta_{0,l}(X,T) - U_l)^2 \mid N = T) = 0$, and consequently

$$R(\delta_0(X,N),U) = R(\delta_2(X,N),U) = \left(1-\frac{1}{l}\right) f(T-1).$$

The proof is completed.

Remark 2.1. Note that the estimator $\delta_1(x,n)$ is identical with that of equation (11) of Hill ((1968), p. 683) in the case $\boldsymbol{a} = 1$ and $M = m$. The estimator $\delta_1(x,n)$ is more general in that it applies for other $\boldsymbol{a}$, but is less general in so far as the formula of Hill is appropriate when the number of subpopulations M is either known or unknown. The prior distribution used in the Theorem 2.1 and Hill when $\boldsymbol{a} = 1$ is the direct multivariate generalization, for sampling without replacement, of the classical Bayes-Laplace uniform prior distribution for a Bernoulli parameter.

Remark 2.2. Fisher (1935) first defines an ancillary statistic partly as a basis for conditioning. A widely held notion about ancillary statistics is that the distribution of the ancillary statistics should be irrelevant to the statistical inference. Brown (1990) shows that in multiple linear regression the admissibility of the ordinary estimator of the intercept depends on the distribution of the design matrix. Some other examples, including confidence interval estimation, the estimation of loss function, etc. are given in the discussions of Brown (1990).

In design-based sampling the probability sampling distribution is usually related to an ancillary statistic. But it may not be surprising to some readers that the minimax estimator for fixed sample size is not minimax estimator anymore for random sample size. The reason is that for an estimation problem which is a randomization among component problems, the minimax estimator is not a composite of the minimax estimators for each of the component problems. The above theorem shows that an estimator for a composite problem, which has constant risk and is separately minimax for each component problem, need not be minimax for the overall problem.

Acknowledgements

The author wishes to thank Professor Frank Martin for many helpful discussions which stimulated this problem, and two referees for their helpful comments.

References

Brown, L. D. (1990). An ancillarity paradox which appears in multiple linear regression (with discussion), *Ann. Statist.*, **18**, 471–538.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., Wiley, New York.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected, *Bull. Internat. Statist. Inst.*, **36**, 113–119.

Fisher, R. A. (1935). The logic of inductive inference (with discussion), *J. Roy. Statist. Soc. Ser. A*, **98**, 39–54.

Guenther, W. C. (1969). Modified sampling, binomial and hypergeometric cases, *Technometrics*, **11**, 639–647.

Haldane, J. (1945). On a method of estimating frequencies, *Biometrika*, **33**, 222–225.

Hartley, H. O. (1959). *Analytic Studies of Survey Data*, volume in honor of Corrado Gini, Istituto di Statistica, Rome.

He, K. (1990). An ancillarity paradox in the estimation of multinomial probabilities, *J. Amer. Statist. Assoc.*, **85**, 824–828.

He, K. (1993). The estimation of stratum means vector with random sample sizes, *J. Statist. Plann. Inference*, **37**, 43–50.

He, K. (1995). On estimating a linear combination of stratum means with random sample sizes, *J. Multivariate Anal.* (to appear).

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population, *J. Amer. Statist. Assoc.*, **63**, 677–691.

Hill, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species, *J. Amer. Statist. Assoc.*, **74**, 668–673.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Trybula, S. (1958). Some problems of simultaneous minimax estimation, *Ann. Math. Statist.*, **29**, 245–253.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size, *J. Roy. Statist. Soc. Ser. B*, **15**, 253–261.