MEDIAN ESTIMATION IN SAMPLE SURVEYS

Shulamith T. Gross, Baruch College, CUNY

ABSTRACT

In a recent paper Maritz and Jarrett (1978) proposed a small-sample estimate of the variance of sample medians from continuous population. In this paper their methods are adapted to median estimation in stratified sampling without replacement from finite populations. A weighted sample median for estimating the median of heavy-tailed or skewed populations is proposed. Its asymptotic normal distribution is derived, and optimal allocation is discussed. A systematic statistic based on the ordered observations in the complete sample, for the estimation of the variance of the weighted median is proposed. This variance estimate is shown to be a consistent estimate of the weighted median asymptotic variance. Initial simulation results suggest a trimmed version of this estimate to be an efficient and computationally feasible alternative. Some results are also obtained for stratified sampling of clusters. The weighted median, based on the pooled clusters in each stratum, is shown to be a consistent and asymptotically normal estimate of the population median. A "Bookstrap" estimate of the variance of this estimate is proposed.

1. INTRODUCTION

Traditionally the theory of sampling concerned itself mostly with estimation of means and totals. Sampling designs of increasing complexity have been developed, along with appropriate weighted sample means and their variance estimates. More recently, a different type of investigation associated with complex sampling designs, has been undertaken. Statistical methodology, originally developed for categorical or continuous data from simple random samples has been investigated and modified for use in complex surveys. The application of χ^2-tests to categorical data from stratified or cluster samples was recently considered by J.N.K. Rao and A.J. Scott (1979). Log-linear model techniques were applied by T.J. Tomberlin (1979) to classified data from complex designs.

In this preliminary report we present results pertaining to a similar investigation concerning median estimation. Sample medians have long been recognized as simple robust alternatives to sample means, for estimating location of heavy-tailed or markedly skewed populations from simple random samples. A large class of robust estimates of location, including the sample median, was investigated in the Princeton simulation study (D.F. Andres et. al. (1972)). Although the sample median did not emerge as "best" estimate in many nonstandard populations simulated in the study, its robustness in small samples for medium and large deviations from normality was clearly demonstrated. Its simplicity relative to other robust estimates, indicated its choice for investigation in designs other than simple random sampling. The extension of the median to stratified sampling leads naturally to weighted medians, similar in nature to Least Absolute Deviation estimates used in regression analysis (see e.g. S. Gross & W.L. Steiger (1979)). Their computation requires sorting the observations in ascending order and assigning each a weight according to the statum from which it originated. The weighted median is the ordered observation that splits the ordered sequence of weights into two subsequences of equal total weight. An exact definition of the weighted median is given in section 2, along with its asymptotic normal distribution. Exact conditions and proofs of statements regarding asymptotic distributions are omitted. Most asymptotic results presented in this paper follow from Hajek's (1961) theorems. When clustering is present, some further arguments, common in nonparametric theory, are required. The asymptotic variance obtained for simple stratified sampling is then used to discuss optimal allocation for estimating the total population median. The general rule that emerges states that strata for which the median is known to be close to the overall population median should be more heavily sampled. As we explain in section 2, the rule in fact requires larger sampling fractions for the more internally variable strata.

A small sample estimate of the variance of the weighted median is presented in section 3. Its derivation is analogous to that of Maritz & Jarrett's (1978) for median estimation in continuous populations. Except for small samples from a small number of strata the estimate requires automated computation. A FORTRAN program for the IBM-370 system is now available for computing the estimate. In its current form the estimate is a systematic statistic, i.e., a linear combination of the complete ordered sample. Computationsof the coefficients done so far suggest that a much simplified trimmed systematic statistic will perform equally well. Further experimentation is necessary before any precise recommendations can be made.

In section 4 results obtained in sections 2 and 3 for stratified sampling are extended to stratified sampling of clusters. Cluster sizes are allowed to vary within and between strata, and sampling fractions are not assumed constant, but no subsampling is carried out within clusters. A weighted sample median, with observations weighted according to the strata from which they originated, is shown to be a consistent, asymptotically normal estimate of the population median. Variance estimation is complicated by the clustering present in the design. Efron's (1979) "Bootstrap" estimation method is extended to cluster sampling without replacement from finite populations, and used to construct an estimate of the variance of the weighted median. The method calls for the construction of a "Bootstrap" population from a given sample, followed by direct computation of the variance of the estimate from the "Bootstrap" population. The theoretical justification of this procedure is not complete but the numerical results obtained in a simulated population and a real population indicate that it yields satisfactory results in practice. Concluding remarks are presented in section 5.

2. THE WEIGHTED MEDIAN IN STRATIFIED POPULATIONS

Consider the typical situation that customarily leads to stratified sampling. A population of size N is naturally divided into K strata of sizes N_1, N_2, ..., N_K. Assuming strata sizes are approximately known from previous surveys, it is desired to estimate the location of the complete population. If the population is known to be highly skewed or heavy-tailed, the population median rather than the population mean is sought. Independent samples of sizes n_1, n_2, ..., n_K are taken without replacement from the K strata. Each of the strata cumulative distribution functions (CDF's from here on) F_1, F_2, ..., F_K may be consistently estimated by the corresponding sample CDF among F_{n1}, F_{n2}, ..., F_{nK}. Here F_{ni} denotes the empirical cumulative ("less than or equal") for the j-th stratum. The population CDF

$$F = \sum_{i=1}^{K} p_i F_i \qquad \text{with } p_i = N_i/N$$

for i = 1, ..., K, is then consistently estimated by the empirical CDF

$$F_n = \sum_{i=1}^{K} p_i F_{ni} \; .$$

The 50-th percentile of F_n is therefore a consistent estimate of the population median, denoted by $\tilde{x}$, if the latter is unique. To compute the estimate X_{med} all K samples must be pooled and ordered and each assigned a weight

$$w_i = p_i \; / \; n_i \quad \text{for } i = 1, \ldots, K \tag{1}$$

The weights of the ascending sequence are then accumulated until .5 is first crossed. The first observation encountered after the crossing is X_{med} In small samples it is customary to take a convex combination, or simple average of that observation, $X_{(\ell)}$, and the one preceding it in the combined sample $X_{(\ell-1)}$. Thus X_{med} is defined as

$$X_{med} = .5(X_{(\ell)} + X_{(\ell-1)}). \tag{2}$$

As the sample sizes increase to the total strata sizes N_i, X_{med} becomes the actual population median. The estimate X_{med} is also consistent in the usual sense in probability when N, N_i, and each $n_i \to \infty$ as long as the population cumulative F_N (the N appended to indicate dependence on size) converge to a continuous CDF F that is strictly increasing at its median $\tilde{x}$. Under certain conditions it is also asymptotically normal with mean $\tilde{x}$ and variance given by

$$\sigma^2 = 1 \; / \; (Nf^2(\tilde{x})) \; (p_i \; / \; f_i) \; (1 - f_i) \; F_i(\tilde{x}) \; (1 - F_i(\tilde{x})) \tag{3}$$

where $f(\tilde{x})$ denotes the population density at its median $\tilde{x}$, $p_i = N_i/N$ and f_i are the sampling fractions n_i/N_i. This expression for the asymptotic variance contains the usual finite population corrections $1-f_i$ and the asymptotic density at the median which will be recognized from the uni-stratum case. In order to assess the dependence of this variance on the strata structure and sampling fractions, it is of some interest to determine optimal allocation for a fixed total sample size

$$n = \sum_{i=1}^{K} N_i f_i .$$

Optimal sampling fractions are given by

$$n_i/n = N_i(F_i(\tilde{x})(1-F_i(\tilde{x})))^{\frac{1}{2}} \; / \; \sum_{i=1}^{K} N_i \; (F_i(\tilde{x}) \; (1-F_i(\tilde{x})))^{\frac{1}{2}}. \tag{4}$$

Under optimal allocation, strata whose individual medians are situated away from the population median will have smaller allocation than those situated close to the population median $\tilde{x}$. This can be seen in (4) by noting that the farther the stratum median is from $\tilde{x}$, the smaller $F_i(\tilde{x})(1-F_i(\tilde{x}))$ and the smaller the optimal allocation for it becomes. We note here that asymptotic normality may be violated if the range of any stratum does not cover $\tilde{x}$. Thus, at least theoretically, all the fractions in (4) are positive, i.e., all strata must be sampled. Note also that the last expression reflects in fact the internal variability of the i-th stratum with respect to median estimation.

It is of course possible to derive the asymptotic variance for porportional sampling from formula (4), and then compare it to the corresponding variance for optimal and simple random sampling. We leave these simple derivations out for lack of space. Note that in proportional sampling, the weighted median becomes the usual sample median, as is the case in mean estimation.

3. VARIANCE ESTIMATION

In order to explain the rationale behind the small sample estimate of the variance of the weighted median X_{med} offered in this section, we begin by developing a formula for the population variance of X_{med}. A sample estimate of this variance will then be obtained by simply replacing all strata cumulatives by their corresponding sample CDF's. From our initial definition of X_{med} as the 50-th percentile of F_n, it is seen that

$$P(X_{med} > x_{(\ell)}) = P(\sum_{i=1}^{K} w_i t_i(\ell) < .5) \tag{5}$$

where $1 \leq \ell \leq N$ and $x_{(\ell)}$ denotes the ℓ-th ordered population value. The variable $t_i(\ell)$ denotes the number of values in the i-th stratum sample that precede or equal $x_{(\ell)}$. The random variable $t_i(\ell)$ for fixed ℓ is a Hypergeometric variable with parameters N_i, $N_iF_i(x_{(\ell)})$ and n_i. The variables $t_i(\ell)$ for i=1, ..., K are independent. Thus the probability in (5), now denoted by p_ℓ is given by

$$p_\ell = \sum_T \prod_{i=1}^{K} \binom{N_iF_i(x_{(\ell)})}{t_i} \binom{N_i(1-F_i(x_{(\ell)}))}{n_i-t_i} \Big/ \binom{N_i}{n_i} \tag{6}$$

for ℓ=1, ..., N and p_0=1. The set T of K-tupples $(t_1, \ldots, t_K)$ is composed of those K-tupples that satisfy the conditions

$$\sum_{i=1}^{K} w_i t_i < .5, \; \text{Max}(0, \; n_i+N_i(F_i(x_{(\ell)})-1)) \leq t_i \leq \text{Min}(n_i, \; N_iF_i(x_{(\ell)})$$

Finally, $P(X_{med} = x_{(\ell)}) = p_{\ell-1}-p_\ell$ and for any r, the r-th moment of X_{med} is given by

$$E(X^r_{med}) = \sum_{\ell=1}^{N} (p_{\ell-1}-p_\ell) \; x^r_{(\ell)} \tag{7}$$

By replacing all population CDF's by their sample estimates the estimate of the variance becomes

$$\hat{\mathrm{Var}}(X_{med}) = \sum_{\ell=1}^{n}(\hat{p}_{\ell-1}-\hat{p}_{\ell})\, X^2_{(\ell)} - \left(\sum_{\ell=1}^{n}(\hat{p}_{\ell-1}-\hat{p}_{\ell})\, X_{(\ell)}\right)^2 \qquad (8)$$

where $X_{(\ell)}$ denotes the ℓ-th ordered observation in the combined sample obtained by pooling together all the K strata samples. The estimates $\hat{p}_\ell$ are given by formula (6) with each $F_i(x_{(\ell)})$ replaced by the corresponding CDF F_{ni} evaluated at $X_{(\ell)}$. Computation of this variance estimate requires tables of Hypergeometric coefficients and a complete ordering of the stratified sample. Except in small samples with three or four strata, hand computation becomes prohibitively time-consuming, but automated computation is rather straightforward. Even in automated computation, much computing time can be saved by starting the computation of the p_ℓ array at $\ell = n/2$ and stopping as soon as the p_ℓ being computed becomes smaller than some small threshold value. Typically in our computations approximately half the array actually required computation. The resulting trimmed systematic estimate performed as well as the original untrimmed one. Under the conditions that assure asymptotic normality for X_{med} the estimate given in (8) is a consistent estimate of the asymptotic variance σ^2 given in (3) in the sense that $N(\mathrm{Var}(X_{med})-\sigma^2) \to 0$ in probability as n_i, N_i and N increase indefinitely.

4. STRATIFIED CLUSTER SAMPLING

In order to appreciate the complications introduced by clustering we first consider the case of simple random sampling of clusters. We assume the population is comprised of a known number M of clusters, each completely described by its CDF F_j and size N_j for j=1, ..., M. A sample of m clusters is taken without replacement, and completely inspected. The sample CDF's G_j and their sizes $N(G_j)$ are then available. The natural estimate of the population CDF F is seen to be

$$F_m = \sum_{j=1}^{m} N(G_j)G_j \Big/ \sum_{j=1}^{m} N(G_j) \qquad (9)$$

Note that since no subsampling is done, no special weighting is required. Hence the simple sample median is a consistent estimate of the population median if the latter is unique. Under conditions that impose restrictions on the variability in cluster sizes and cluster medians, the sample median may be shown to be asymptotically normal with the population median $\tilde{x}$ as mean, and variance given by σ_c^2 where

$$m\sigma_c^2 = (1-m/M)(M-1)^{-1} \sum_{j=1}^{M} N_j^2(F_j(\tilde{x})-F(\tilde{x}))^2\, (\bar{N}f(\tilde{x}))^{-2} \qquad (10)$$

where $f(\tilde{x}) > 0$ is the asymptotic density at $\tilde{x}$ which is assumed to exist when $M \to \infty$, and

$$\bar{N} = \sum_{j=1}^{M} N_j/M.$$

Comparison of (10) with the asymptotic variance $(1-n/N)/4f^2(\tilde{x})$ of the median in simple random sampling indicates the effectiveness of clustering when cluster median variation is small. Using the notation developed in section 2 for stratified sampling, in addition to that introduced in this section for clustering, we are now able to state the results pertaining to stratified cluster sampling with complete inspection of sample clusters. Two distinct cases will be considered.

Case I: When the total number of elementary units

$$N_i = \sum_{j=1}^{M_i} N_{ij}$$

in each stratum is known prior to sampling, a consistent estimate of the population cumulative is constructed by

$$F_n(x) = \sum_{i=1}^{K} (N_i/N)\, F_{mi}(x) \qquad (11)$$

where each stratum empirical cumulative F_{mi} is constructed as in (9). The 50-th percentile of the empirical cumulative given by (11) is then the weighted median described in section 2 obtained by pooling all clusters within a stratum and then pooling together all strata elementary units assigning each a weight w_i according to its stratum of origin. The weights w_i are random and computed by

$$w_i = N_i \,/\, (Nn_i) \qquad (12)$$

where $n_i = \sum_{j=1}^{m_i} N(G_{ij})$

denotes the total sample size in elementary units in the i-th stratum sample. Since the strata coefficients in (11)

$$p_i = N_i/N \qquad (13)$$

are not random, asymptotic normality of $F_n(x)$ and therefore of X_{med} is obtained directly from the single stratum case. Under conditions that impose restrictions on the variability in cluster sizes and indirectly on cluster medians in individual strata, the weighted median X_{med} is asymptotically normal with mean $\tilde{x}$ and variance given by

$$\sigma^2_{sc} = \sum_{j=1}^{K} p_i^2\, ((1-f_i)/m_i) \sum_{j=1}^{M_i} N_{ij}^2\, (F_{ij}(\tilde{x}) - F_i(\tilde{x}))^2 \,/\, (M_i-1)\, (\bar{N}_i\, f(\tilde{x}))^2 \qquad (14)$$

where $f_i = m_i/M_i$, the strata sampling fractions, $\bar{N}_i$ denotes the average cluster size in the i-th stratum, and $f(\tilde{x}) > 0$ is the density of the asymptotic distribution F which is assumed to exist at $\tilde{x}$.

Case II: In practical situations, the number of clusters M_i in the i-th stratum may be approximately known prior to sampling but the total stratum size N_i may not, and will therefore have to be estimated from the cluster sample by

$$\hat{N}_i = M_i n_i/m_i\ . \qquad (15)$$

The resulting estimate for the population cumulative will then be

$$F_m(x) = \sum_{i=1}^{K}(M_i/m_i) \sum_{j=1}^{m_i} N(G_{ij})\, G_{ij}(x) \,/ \sum_{i=1}^{K} (M_i/m_i)\, n_i\ . \qquad (16)$$

The weighted median X_{med} derived as the 50-th percentile of F_m is obtained as in case I with weights given by

$$\hat{w}_i = M_i/(m_i \sum_{i=1}^{K} (M_i n_i/m_i)). \tag{17}$$

Employing again Hajek's (1961) results and standard arguments necessary to deal with the random denominator in (16), the weighted median X_{med} may be shown to be asymptotically normal with mean $\tilde{x}$ and variance given by

$$\sigma^2_{sce} = \sum_{i=1}^{K} p_i^2((1-f_i)/m_i) \sum_{j=1}^{M_i} (N_{ij}(F_{ij}(\tilde{x})-.5) - \bar{N}_i(F_i(\tilde{x})-.5))^2/ ((M_i-1)(\bar{N}_i f(\tilde{x}))^2). \tag{18}$$

All the quantities in (18) are defined exactly as in the corresponding formula (14) for the asymptotic variance of case I.

Note that when K = 1, no weighting is done and the asymptotic variances given by (14) and (18) coincide since then $F_1(\tilde{x}) = .5$. In general it is expected that σ^2_{sc} will be smaller than σ^2_{sce} because additional estimation of strata sizes is involved in the latter case.

Under mild restrictions placed on the variability of cluster sizes in the K strata the empirical cumulatives given in (11) and (16) will converge to the asymptotic cumulative F in probability as m_i and $M_i \to \infty$. Thus if F has a unique median $\tilde{x}$, the weighted median based on the weights (12) or the estimated weights given by (17), will converge in probability to $\tilde{x}$.

Estimation of the variance of the weighted median in either one of the weighting schemes is complicated by the fact that no simple explicit formula for its population variance is available. Thus the Maritz-Jarrett method of replacing population cumulatives by empirical cumulatives is not directly applicable. Close inspection of the variance estimate derived using their method in the unclustered case in section 3 indicates a way of bypassing this difficulty. In fact the estimate given in (8) is the population variance of the weighted median when the population is the "Bootstrap" population constructed from the simple stratified sample as follows. Assuming that for each i, $k_i = N_i/n_i$ is an integer, a "Bootstrap" stratum is obtained by including k_i replications of each of the n_i sample observations in it. The "Bootstrap" estimation procedure for cluster stratified sampling may be described as follows:

(1) Reconstruct stratum i by assuming that it is made up of $M_i/m_i = k_i$ replications of the actual m_i clusters observed in that stratum.

(2) Take all possible stratified samples from the reconstructed population obtained by applying step (1) to all strata, and compute the corresponding weighted median for each of these samples. Note that any computer routine that calculates the weighted median for stratified populations can be used to compute the "Bootstrapped" medians in this step. Furthermore, the number of weighted medians that need to be computed is not as large as it may seem, due to the replication of clusters within strata.

(3) Compute the mean and variance, $\hat{Var}(X_{med})$, of the medians. The difference between the former and X_{med} will yield an estimate of the bias, whereas the latter will serve as an estimate of the variance of X_{med}.

This procedure is reminiscent of the BRR method, but differs from it in that it employs all possible reconstructed samples. We conjecture that under appropriate conditions $M(\hat{Var}(X_{med})-\sigma^2_{sce}) \to 0$ in probability as m_i and $M_i \to \infty$.

5. CONCLUDING REMARKS

In this paper median estimates in stratified (or) clustered designs were considered and estimates of their variances proposed. Examples of real and simulated data will be presented at the meeting, to display the computational procedures involved, and their efficacy in estimation of location of skewed populations as compared to the sample mean.

REFERENCES

Andrew, D.F. et. al. (1972) Robust estimates of location-surveys and advances. Princeton University Press.

Efron, B. (1979) "Bootstrap methods: Another look at the Jacknife". The Annals of Stat., Vol.7, no.1, 1-20.

Gross, S. and Steiger, W.L. (1979) "Lease absolute deviation estimates in autogression with infinite variance", Journal of Applied Probability, Vol. 16, 104-116.

Hajek, J. (1961) "Some extensions of the Wald-Wolfowitz-Noether theorem", AMS 32 506-523.

Maritz, J.S. and Jarrett, R.G. (1978) "A note on estimating the variance of the sample median", JASA Vol. 73, 194-196.

Rao, J.N.K. and Scott, A.J. (1979) "Chi-Square tests for analysis of categorical data from complex surveys". Paper presented at ASA 1979 meeting.

Tomberlin, T.J. (1979) "The analysis of contingency tables of data from complex samples". Paper presented at 1979 ASA annual meeting.

ACKNOWLEDGEMENTS

The author wishes to thank Martin R. Frankel for valuable discussions on the topic of this paper, and Ann Dean for programming assistance.