Normal approximation for stratified samples

Mindaugas Bloznelis

Vilnius University, Lithuania
e-mail: mindaugas.bloznelis@maf.vu.lt

Abstract

Assume we want to estimate the Gini difference $\sum_{i,j}|x_i-x_j|$ of the finite population $\mathcal{X}=\{x_1,\dots,x_N\}$ and suppose that the information available is a stratified sample. The commonly used estimator has the form of a $U-$statistic. We evaluate the variance σ_U^2 of the estimator and study the normal approximation of its distribution. Furthermore, we study the bias and consistency of the jackknife estimator of variance S^2 and the normal approximation in the case where estimator is standardized by S.

The work is supported by Lithuanian State Science and Studies Foundation Grant T-04051.

1 Introduction

Consider the population $\mathcal{X}=\{x_1,\dots,x_N\}$ and assume that we want to estimate the population parameter $u=\sum_{1\le i<j\le N}t(x_i,x_j)$. Suppose that the population is divided in h non-intersecting strata $\mathcal{X}=\mathcal{X}_1\cup\ldots\cup\mathcal{X}_h$, where $\mathcal{X}_k=\{x_{k.1},\dots,x_{k.N_k}\}$ and $N_1+\ldots+N_k=N$. From every $\mathcal{X}_k$ we draw (without replacement) the simple random sample $\mathbb{X}_k=\{X_{k.1},\dots,X_{k.n_k}\}$, $k=1,\dots,h$, so that samples $\mathbb{X}_1,\dots,\mathbb{X}_h$ are independent. The sample $\mathbb{X}=(\mathbb{X}_1,\dots,\mathbb{X}_h)$ is called stratified sample without replacement (STSI sample for short). We assume that the function t is symmetric (i.e., $t(x,y)=t(y,x)$).The statistic

$$\hat{u}=\hat{u}(\mathbb{X})=\sum_{1\le j\le h}\sum_{\{x,y\}\subset\mathbb{X}_j}w_jt(x,y)+\sum_{1\le j<r\le h}\sum_{x\in\mathbb{X}_j}\sum_{y\in\mathbb{X}_r}w_{jr}t(x,y),\qquad(1)$$

where

$$w_j=w_j(\mathbb{X})=\binom{N_j}{2}\binom{n_j}{2}^{-1},\qquad w_{jr}=w_{jr}(\mathbb{X})=N_jN_r(n_jn_r)^{-1}\qquad(2)$$

is an unbiased estimator of the parameter u. In the particular case $t(x,y)=|x-y|$ the parameter u is the Gini difference and the statistic $\hat{u}(\mathbb{X})$ is its unbiased estimator.

Statistics of the form (1) are called $U-$statistics. This class of statistics contains many important estimators. For instance, the commonly used STSI estimator of the population total can be written in the form (1). Let us consider (real valued) measurements $g(x_1), \dots, g(x_N)$ of the population units $x_1, \dots, x_N$ and let $a = \sum_{i=1}^{N} g(x_i)$ be the population total. The STSI estimator

$$\hat{a} = \sum_{k=1}^{h} \frac{N_k}{n_k} \sum_{x \in \mathbb{X}_k} g(x) \tag{3}$$

can be written in the form (1) if we take $t(x_i, x_j) = (N-1)^{-1}(g(x_i) + g(x_j))$. Therefore, the theory given below applies to the estimator of the population total as well.

2 Some analysis of $U-$statistics

2.1 Decomposition

It is convenient to decompose the statistic $\hat{u}$ into the sum of the linear and quadratic part (L and Q below) which are *uncorrelated* (see Hoeffding (1948))

$$\hat{u} = u + L + Q, \qquad L = \sum_{1 \le r \le h} L_r, \qquad Q = \sum_{1 \le r \le s \le h} Q_{rs}, \tag{4}$$

$$L_r = \sum_{x \in \mathbb{X}_r} g_r(x), \qquad Q_{rr} = \sum_{\{x,y\} \subset \mathbb{X}_r} \psi_r(x, y), \qquad Q_{rs} = \sum_{x \in \mathbb{X}_r} \sum_{y \in \mathbb{X}_s} \psi_{rs}(x, y).$$

The functions (we call them kernels) g_r and ψ_r, ψ_{rs} are given in Appendix.
Remark. For estimator (3) we have $\hat{a} = a + L$, where

$$L = \sum_{r=1}^{h} L_r, \qquad L_r = \sum_{x \in \mathbb{X}_r} \frac{N_r}{n_r}(g(x) - a_r), \qquad a_r = \frac{1}{N_r} \sum_{x \in \mathcal{X}_r} g(x).$$

In this case the quadratic part is not present.
Statistics that have no quadratic part are called *linear* statistics.

2.2 Variance formula

The linear and quadratic part help to write the variance formula. We have

$$\sigma_U^2 = \sigma_L^2 + \sigma_Q^2, \tag{5}$$

where σ_U^2, σ_L^2 and σ_Q^2 denotes the variances of $\hat{u}$, L and Q. Let $\sigma^2(L_k)$, $\sigma^2(Q_{kk})$ and $\sigma^2(Q_{kr})$ denote the variance of L_k, Q_{kk} and Q_{kr}. One can show that

$$\sigma_L^2 = \sum_{1\le k\le h} \sigma^2(L_k), \qquad \sigma_Q^2 = \sum_{1\le k\le r\le h} \sigma^2(Q_{kr}), \tag{6}$$

$$\sigma^2(L_k) = \frac{N_k}{N_k-1}\tau_k^2\sigma_k^2, \qquad \sigma^2(Q_{kk}) = \frac{\binom{N_k-n_k}{2}\binom{n_k}{2}}{\binom{N_k-2}{2}}\sigma_{kk}^2,$$

$$\sigma^2(Q_{kr}) = \frac{N_k}{N_k-1}\frac{N_r}{N_r-1}\tau_k^2\tau_r^2\sigma_{kr}^2.$$

Here we denote

$$\sigma_k^2 = Eg_k^2(X_{k.1}), \quad \sigma_{kk}^2 = E\psi_k^2(X_{k.1}, X_{k.2}), \quad \sigma_{kr}^2 = E\psi_{kr}^2(X_{k.1}, X_{r.1}) \tag{7}$$

and

$$\tau_k^2 = N_k p_k q_k, \qquad p_k = n_k/N_k, \qquad q_k = (N_k - n)_k)/N_k.$$

3 Normal approximation

We shall assume that the linear part L has positive variance, $\sigma_L^2 > 0$. By the central limit theorem, for large $n = n_1 + \ldots + n_h$ and N, the distribution of L/σ_L can be approximated by the standard normal distribution,

$$P\{L \le x\sigma_L\} \approx \Phi(x), \qquad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

Let σ_U^2 denote the variance of $\hat{u}$. If the linear part dominates the statistic and we have $\sigma_U^2/\sigma_L^2 \approx 1$, the normal approximation applies to $\hat{u}/\sigma_U$ as well,

$$P\{\hat{u} - u \le x\sigma_U\} \approx \Phi(x). \tag{8}$$

The normal approximation is used for construction of confidence intervals.

3.1 Improvement by an Edgeworth expansion

An improvement over the normal approximation is provided by an Edgeworth expansion, see Cramér (1946). Edgeworth expansions include correcting terms, which capture the asymmetry and some other factors that cause the deviation of the distribution function from $\Phi(x)$. The one-term Edgeworth expansion includes only one correcting term

$$P\{L \le x\sigma_L\} \approx \Phi(x) - \frac{\alpha}{6\sigma_L^3}\Phi'(x)(x^2-1), \tag{9}$$

$$P\{\hat{u} - u \le x\sigma_U\} \approx \Phi(x) - \frac{\alpha + 3\kappa}{6\sigma_U^3}\Phi'(x)(x^2-1). \tag{10}$$

Here the parameter α reflects the asymetry of the linear statistic L and is determined by L. The parameter κ reflects the influence of the quadratic part. Both parameters are the population mean values of some functionals (”measurements”).

Since often the variances σ_L^2 and σ_U^2 and moments α and κ are not known the approximations above have little use for practice. One would like to have improved approximations like (9) and (10) that does *not* use unknown population parameters. One way is to replace the true parameters σ_L^2, σ_U^2 etc by their estimators. Another way is the bootstrap.

3.2 Jackknife variance estimator

The classical jackknife estimator S^2 of the variance σ_U^2 is defined as follows

$$\begin{aligned} S^2 &= S^2(\hat{u}) = \sum_{k=1}^{h} q_k \frac{n_k - 1}{n_k} v_k^2, \qquad (11) \\ v_k^2 &= \sum_{i=1}^{n_k} (\hat{u}(\mathbb{X}_{k|i}) - \overline{u}_k)^2, \qquad \overline{u}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \hat{u}(\mathbb{X}_{k|i}). \end{aligned}$$

Here $\mathbb{X}_{k|i}$ denotes the STSI sample obtained from $\mathbb{X}$ by removing the observation $X_{k.i}$. In order to get $\hat{u}(\mathbb{X}_{k|i})$ we replace $\mathbb{X}_k$ by $\mathbb{X}_k \setminus \{X_{k.i}\}$ in (1) and adjust the weights (2). We put $w_j(\mathbb{X}_{k|i}) = w_j(\mathbb{X})$ for $j \neq k$ and $w_k(\mathbb{X}_{k|i}) = w_k(\mathbb{X}) n_k/(n_k - 2)$. Similarly, $w_{jr}(\mathbb{X}_{k|i}) = w_{jr}(\mathbb{X})$ for $k \notin \{j, r\}$ and $w_{jr}(\mathbb{X}_{k|i}) = w_{jr}(\mathbb{X}) n_k/(n_k - 1)$ for $k \in \{j, r\}$. Note that $\hat{u}(\mathbb{X}_{k|i})$ is an unbiased estimator of u, i.e., $E\hat{u}(\mathbb{X}_{k|i}) = u$.

It is well known that $S^2(\hat{a})$ is an unbiased estimator of the variance of $\hat{a}$. More generally, $S^2(\hat{u})$ is an unbiased estimator of σ_U^2 if $\hat{u}$ is a linear statistic. For $U-$statistics which have the quadratic part the estimator $S^2(\hat{u})$ is biased upwards. The bias can be evaluated and expressed in terms of the quantities σ_{kk}^2 and σ_{kr}^2.

Lemma. *The following identity is true*

$$ES^2 = \sigma_U^2 + D. \qquad (12)$$

Here

$$\begin{aligned} D &= \sum_{1 \le r < k \le h} \sigma^2(Q_{rk}) + \sum_{1 \le k \le h} (c_k - 1)\sigma^2(Q_{kk}), \\ c_k &= 2 \frac{N_k - 1}{N_k} \frac{N_k - n_k}{N_k - n_k - 1} \frac{n_k - 1}{n_k - 2} \end{aligned}$$

Clearly, $c_k \ge 2$, for every $k = 1, \ldots, h$. Therefore, we have $D \ge 0$. Note that $D > 0$ whenever at least one of the quantities $\sigma_{kr}^2 > 0$, for $1 \le k \le r \le h$, i.e. whenever $\sigma_Q^2 > 0$ in (5).

The inequality $ES^2 \ge \sigma_U^2$ tells us that the jackknife variance estimator tends to be biased upwards. For statistics based on independent and identically distributed observations this fact was shown by Efron and Stein (Ann. Stat. 1981). For STSI samples similar result was shown by Bloznelis (Statistics 2003), but he considered a different version of the jackknife variance estimator.

3.3 Normal approximation of Studentized statistics

Often the variance σ_U^2 is unknown. By the law of large numbers we have $S^2 \approx \sigma_U^2$. Replacing σ_U by S in (8) we obtain

$$P\{\hat{u} - u \leq xS\} \approx \Phi(x), \tag{13}$$

Hence, the normal approximation applies to the so called Studentized statistic $(\hat{u} - u)S^{-1}$.
An improvement over the normal approximation (13) is provided by an Edgeworth expansion. Write the one-term Edgeworth expansion

$$P\{\hat{u} - u \leq xS\} \approx \Phi(x) + \frac{\alpha + \alpha' x^2 + 3\kappa(x^2+1)}{6\sigma_U^3}\Phi'(x). \tag{14}$$

Here α, α' and κ are the population mean values of some functionals (α, κ are the same as in (10)).

4 Resampling approximations

Replacing the quantities α, α' and κ by their estimators $\hat{\alpha}, \hat{\alpha'}$ and $\hat{\kappa}$ in (9), (10) and (14) we obtain empirical approximations which do not assume the knowledge of population parameters α, α' and κ. The simplest estimators are constructed using the jackknife type procedure as in Bloznelis (2003).
The bootstrap method provides approximations of the probabilities $P\{\hat{u} - u \leq x\sigma_U\}$ and $P\{\hat{u} - u \leq xS\}$ by corresponding empirical analogues which are obtained by simulating STSI samples from appropriately chosen empirical populations. Here it is important to mach the parameters of expansions (10), (14) with their empirical analogues, see Babu and Singh (1985), Rao and Wu (1988), Chen and Sitter (1993), Booth and Presnell (1994).

5 Appendix

5.1 Some more formulas for Hoeffding decomposition

Here we give formulas defining the kernels g_k, ψ_k and ψ_{kr} introduced in (4) above. Write $\tilde{t}_k(x,y) = t_k(x,y) - Et_k(X_{k.1}, X_{k.2})$ and $\tilde{t}_{kr}(x,y) = t_{kr}(x,y) - Et_{kr}(X_{k.1}, X_{r.1})$. We have

$$g_k(x) = (n_k - 1)t_k^*(x) + \sum_{1 \leq j \leq h,\, j \neq k} n_j t_{k|j}^*(x), \tag{15}$$

where

$$t^*_k(x) = \frac{N_k - 1}{N_k - 2} E(\tilde{t}_k(X_{k.1}, X_{k.2})|X_{k.1} = x),$$

$t^*_{k|r}(x) = E(\tilde{t}_{kr}(X_{k.1}, X_{r.1})|X_{k.1} = x)$, $t^*_{r|k}(x) = E(\tilde{t}_{kr}(X_{k.1}, X_{r.1})|X_{r.1} = x)$.

Furthermore, we have

$$\psi_k(x,y) = \tilde{t}_k(x,y) - t^*_k(x) - t^*_k(y), \quad \psi_{kr} = \tilde{t}_{kr}(x,y) - t^*_{k|r}(x) - t^*_{r|k}(y). \tag{16}$$

Note that for every k and r we have

$$Eg_k(X_{k.1}) = 0, \qquad E\psi_k(X_{k.1}, X_{k.2}) = 0, \qquad E\psi_{kr}(X_{k.1}, X_{r.1}) = 0. \tag{17}$$

Moreover, for every k and r we have

$$E(\psi_k(X_{k.i}, X_{k.j})|X_{k.j}) = 0, \qquad i \neq j, \tag{18}$$

$$E(\psi_{kr}(X_{k.i}, X_{r.j})|X_{r.j}) = 0, \qquad E(\psi_{kr}(X_{k.i}, X_{r.j})|X_{k.i}) = 0. \tag{19}$$

It follows from (18) and (19) that the parts L and Q are uncorrelated.

5.2 Proof of the Lemma

We can assume without of generality that $u = 0$.
We have $ES^2 = \sum_k \frac{n_k - 1}{n_k} q_k E v_k^2$. In order to prove (12) we show that

$$\frac{n_k - 1}{n_k} q_k E v_k^2 = \sigma^2(L_k) + \sum_{1 \le r \le h} \mathbb{I}_{\{r \neq k\}} \sigma^2(Q_{rk}) + c_k \sigma^2(Q_{kk}), \tag{20}$$

where c_k is given in (12) above.
Let us prove (20). Denote (for short) $\hat{u}_{k|i} = \hat{u}(\mathbb{X}_{k|i})$. It follows from the identity $v_k^2 = \sum_i \hat{u}_{k|i}^2 - n_k \overline{u}_k^2$, by symmetry, that

$$E v_k^2 = n_k E \hat{u}_{k|i}^2 - n_k E \overline{u}_k^2. \tag{21}$$

Let us evaluate the expectations $E\hat{u}_{k|i}^2$ and $E\overline{u}_k^2$.
5.2.1. Since $E\hat{u}_{k|i} = u = 0$, we have $E\hat{u}_{k|i}^2 = \sigma^2(\hat{u}_{k|i})$, where $\sigma^2(\hat{u}_{k|i})$ denotes the variance of $\hat{u}_{k|i}$. In order to evaluate $\sigma^2(\hat{u}_{k|i})$ we are going to apply (5). Let us write the decomposition (4) for $\hat{u}_{k|i}$. We have $\hat{u}_{k|i} = \tilde{L}_{(i)} + \tilde{Q}_{(i)}$, where $\tilde{L}_{(i)} = \sum_{r=1}^h \tilde{L}_r$ and $\tilde{Q}_{(i)} = \sum_{1 \le r \le s \le h} \tilde{Q}_{rs}$ denote the linear and quadratic part respectively. It is easy to see that

$$\tilde{L}_r = L_r, \qquad \tilde{Q}_{rs} = Q_{rs}, \qquad k \notin \{r, s\}. \tag{22}$$

Furthermore, denote $\mathbb{X}_k^i = \mathbb{X}_k \setminus \{X_{k.i}\}$. A calculation shows that

$$\begin{aligned}
\tilde{L}_k &= \sum_{x \in \mathbb{X}_k^i} \tilde{g}_k(x), \qquad \tilde{g}_k(x) = \frac{n_k}{n_k - 1} g_k(x), \qquad (23)\\
\tilde{Q}_{kk} &= \sum_{\{x,y\} \subset \mathbb{X}_k^i} \tilde{\psi}_k(x,y), \qquad \tilde{\psi}_k = \frac{n_k}{n_k - 2} \psi_k(x),\\
\tilde{Q}_{kr} &= \sum_{x \in \mathbb{X}_k^i} \sum_{y \in \mathbb{X}_r} \tilde{\psi}_{kr}(x,y), \qquad \tilde{\psi}_{kr} = \frac{n_k}{n_k - 1} \psi_{kr}(x), \quad r \neq k.
\end{aligned}$$

Let $\sigma^2(\tilde{L}_{(i)})$ and $\sigma^2(\tilde{Q}_{(i)})$ denote variances of $\tilde{L}_{(i)}$ and $\tilde{Q}_{(i)}$. (5) implies

$$E\hat{u}_{k|i}^2 = \sigma^2(\hat{u}_{k|i}) = \sigma^2(\tilde{L}_{(i)}) + \sigma^2(\tilde{Q}_{(i)}). \qquad (24)$$

Invoking formulas (6) and using (22) we obtain

$$\begin{aligned}
&\sigma^2(\tilde{L}_{(i)}) + \sigma^2(\tilde{Q}_{(i)}) = A + B, \quad A = \sigma^2(\tilde{L}_k) + \sum_{1 \le r \le h} \sigma^2(\tilde{Q}_{kr}), \qquad (25)\\
&B = \sum_{1 \le r \le h} \mathbb{I}_{\{r \neq k\}} \sigma^2(L_r) + \sum_{1 \le r \le s \le h} \mathbb{I}_{\{k \notin \{r,s\}\}} \sigma^2(Q_{rs}),
\end{aligned}$$

It follows from (23) that

$$\begin{aligned}
&\sigma^2(\tilde{L}_k) = \frac{N_k}{N_k - 1} \tilde{\tau}_k^2 \tilde{\sigma}_k^2, \qquad \sigma^2(\tilde{Q}_{kk}) = \frac{\binom{N_k - n_k + 1}{2}\binom{n_k - 1}{2}}{\binom{N_k - 2}{2}} \tilde{\sigma}_{kk}^2, \qquad (26)\\
&\sigma^2(\tilde{Q}_{kr}) = \frac{N_k}{N_k - 1} \frac{N_r}{N_r - 1} \tilde{\tau}_k^2 \tau_r^2 \tilde{\sigma}_{kr}^2.
\end{aligned}$$

Here we denote $\tilde{\tau}_k^2 = \frac{(n_k - 1)(N_k - n_k + 1)}{N_k}$ and $\tilde{\sigma}_k^2$, $\tilde{\sigma}_{kk}^2$, $\tilde{\sigma}_{kr}^2$ are defined by (7), but with g, ψ replaced by $\tilde{g}$, $\tilde{\psi}$. It follows from (23) that

$$\tilde{\sigma}_k^2 = \frac{n_k^2}{(n_k - 1)^2} \sigma_k^2, \quad \tilde{\sigma}_{kk}^2 = \frac{n_k^2}{(n_k - 2)^2} \sigma_{kk}^2, \quad \tilde{\sigma}_{kr}^2 = \frac{n_k^2}{(n_k - 1)^2} \sigma_{kr}^2. \qquad (27)$$

Substitution of (26) and (27) in (25) gives the explicit formula for A.

5.2.2. Let us show that

$$\begin{aligned}
E\overline{u}_k^2 &= E\hat{u}^2 = \sum_{1 \le r \le h} \sigma^2(L_r) + \sum_{1 \le r \le s \le h} \sigma^2(Q_{rs}), \qquad (28)\\
&= B + A', \qquad A' = \sigma^2(L_k) + \sum_{1 \le r \le h} \sigma^2(Q_{kr}).
\end{aligned}$$

The second and third identities follow from (5) and (6), and the fact that $E\hat{u} = u = 0$. In order to prove the first identity of (28) we show that the random variables $\overline{u}_k$ and $\hat{u}$ coincide. For this purpose we substitute the

decompositions $\hat{u}(\mathbb{X}_{k|i}) = \tilde{L}_{(i)} + \tilde{Q}_{(i)}$ in the formula $\overline{u}_k = n_k^{-1} \sum_i \hat{u}(\mathbb{X}_{k|i})$ and using (22), (23) verify the identity $\overline{u}_k = L + Q$. Here L and Q denote the linear and the quadratic part of the decomposition $\hat{u} = L + Q$.
5.2.3. Finally we obtain from (24), (25), (27) and (28) that

$$n_k(E\hat{u}_{k|i}^2 - E\hat{u}^2) = n_k(A - A') = d_k\sigma^2(L_k) + \sum_{1 \le r \le h} d_{kr}\sigma^2(Q_{kr}),$$

Here $d_k = \frac{n_k}{n_k-1}\frac{1}{q_k}$, $d_{kk} = 2\frac{N_k-1}{N_k-n_k-1}\frac{n_k}{n_k-2}$ and $d_{kr} = \frac{n_k}{n_k-1}\frac{1}{q_k}$, for $r \neq k$. This identity together with (21) shows (20).

References

Babu, J.G. and Singh, K. (1985) Edgeworth expansions for sampling without replacement from finite populations. *J. Multivar. Anal.* **17**, 261-278.

Bloznelis, M. (2003) An Edgeworth expansion for Studentized finite population statistics *Acta Applicand. Math.* **78**, 51-60.

Booth, J. and Presnell, B. (1994) Resampling Methods for Sample Surveys, *Technical report.*

Chen, J. and Sitter, R.R. (1993) Edgeworth expansion and teh bootstrap for stratified sampling without replacement from a finite population. *Canad. J. Statist.*, **21**, 347-357.

Cramér, H. (1946) *Mathematical methods of statistics.* (Princeton Mathematical series. 9) Princeton N. J.: Princeton University Press XVI, 575 p.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* 19, 293-325.

Rao, J.N.K. and Wu, C.F.J. (1988) Resampling inference with complex survey data. *J.Amer. Statist. Assoc.* **87**, 755-765.