Lietuvos matematikos rinkinys. LMD darbai
52 tomas, 2011, 309–314

ISSN 0132-2818
www.mii.lt/LMR/

Approximations to distribution of median in stratified samples

Andrius Čiginas[1], Tomas Rudys[2]

[1] *Faculty of Mathematics and Informatics, Vilnius University*
Naugarduko 24, LT-03225 Vilnius
[2] *Faculty of Fundamental Sciences, Vilnius Gedimino Technical University*
Saulėtekio 11, LT-10223 Vilnius
E-mail: andrius.ciginas@mif.vu.lt; tomas.rudys@gmail.com

Abstract. We consider an Edgeworth type approximation to the distribution function of sample median in the case of stratified samples drawn without replacement. We give explicit expression of this approximation, and also its empirical version based on bootstrap. We compare their accuracy with that of the normal approximation by numerical examples.

Keywords: finite population, stratified sample without replacement, Hoeffding decomposition, Edgeworth expansion, bootstrap.

1 Introduction and results

Consider a population $\mathcal{X} = \{x_1, \dots, x_N\}$ of size N. We assume without loss of generality that $x_1 \leqslant \dots \leqslant x_N$. Let $\mathcal{X}$ be divided into $h \geqslant 1$ nonoverlapping strata $\mathcal{X} = \mathcal{X}_1 \cup \dots \cup \mathcal{X}_h$, where $\mathcal{X}_k = \{x_{k,1}, \dots, x_{k,N_k}\}$. Clearly, $N = N_1 + \dots + N_h$. Here, for convenience, we will also assume that $x_{k,1} \leqslant \dots \leqslant x_{k,N_k}$. Let $\mathbb{X}_k = \{X_{k,1}, \dots, X_{k,n_k}\}$ be the simple random sample of size $n_k < N_k$ drawn without replacement from the stratum $\mathcal{X}_k$. We assume that the samples $\mathbb{X}_1, \dots, \mathbb{X}_h$ are independent. Write $\mathbb{X} = \mathbb{X}_1 \cup \dots \cup \mathbb{X}_h$ and denote $n = n_1 + \dots + n_h$. Denote the distribution function of the stratum k and its empirical analogue by

$$F_{N,k}(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\{x_{k,i} \leqslant x\} \quad \text{and} \quad F_{n,k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}\{X_{k,i} \leqslant x\}$$

respectively. Here $\mathbb{I}\{\cdot\}$ is the indicator function. Then the distribution function of the population $\mathcal{X}$ and its estimator are

$$F_N(x) = \sum_{k=1}^{h} \frac{N_k}{N} F_{N,k}(x) \quad \text{and} \quad F_n(x) = \sum_{k=1}^{h} \frac{N_k}{N} F_{n,k}(x)$$

respectively. Consider the population median defined as follows $F_N^{-1}(0.5) = \inf\{x\colon F_N(x) \geqslant 0.5\}$. Define its estimator

$$X_{\text{med}} = F_n^{-1}(0.5) = \inf\big\{x\colon F_n(x) \geqslant 0.5\big\}.$$

Denote $\sigma^2 = \mathbf{Var} X_{\mathrm{med}}$. In the present paper we are interested in approximations to the distribution function $F_{\mathrm{med}}(x) = \mathbf{P}\{X_{\mathrm{med}} - \mathbf{E}X_{\mathrm{med}} \leqslant x\sigma\}$. The asymptotic normality of median X_{med} under stratified simple random sampling (STSRS) without replacement was considered in [4, 5]. Here we present an Edgeworth expansion for $F_{\mathrm{med}}(\cdot)$ and its empirical analogue. Our approach is based on Hoeffding's (orthogonal) decomposition $X_{\mathrm{med}} = \mathbf{E}X_{\mathrm{med}} + L + Q + R$ constructed in [1] for general symmetric statistics based on STSRS samples drawn without replacement. Here L and Q are called linear and quadratic parts of the decomposition, and R is a remainder term. In the case of U-statistics, where $R \equiv 0$, Edgeworth expansions were constructed and their second-order correctness was shown in [2]. Thus we expect that, if R is negligible, those Edgeworth expansions will also approximate $F_{\mathrm{med}}(\cdot)$ well. In particular, we propose to approximate $F_{\mathrm{med}}(\cdot)$ by

$$G(x) = \varPhi(x) - \frac{\alpha + 3\kappa}{6\sigma^3}\varPhi'(x)\bigl(x^2 - 1\bigr), \tag{1}$$

obtained in [2]. Here $\varPhi'(x)$ denotes the derivative of the standard normal distribution function $\varPhi(x)$, and

$$\alpha = \sum_{k=1}^{h}(1 - 2n_k/N_k)\tau_k^2\alpha_k \quad \text{and} \quad \kappa = \sum_{k=1}^{h}\tau_k^4\kappa_{kk} + 2\sum_{1\leqslant k<u\leqslant h}\tau_k^2\tau_u^2\kappa_{ku}, \tag{2}$$

with $\tau_k^2 = n_k(1 - n_k/N_k)$. Here the moments

$$\alpha_k = \frac{1}{N_k}\sum_{s=1}^{N_k} g_k^3(x_{k,s}), \qquad \kappa_{kk} = \frac{1}{\binom{N_k}{2}}\sum_{1\leqslant s<r\leqslant N_k}\psi_k(x_{k,s}, x_{k,r})g_k(x_{k,s})g_k(x_{k,r}),$$
$$\kappa_{ku} = \frac{1}{N_kN_u}\sum_{1\leqslant s\leqslant N_k,\, 1\leqslant r\leqslant N_u}\psi_{ku}(x_{k,s}, x_{u,r})g_k(x_{k,s})g_u(x_{u,r}),$$

established in [2], are based on the functions

$$g_k(x_{k,s}) = \frac{N_k - 1}{N_k - n_k}\sum_{i=1}^{N-1}(p_i(x_{k,s}) - p_i)\ \triangle_i, \tag{3}$$

$$\begin{aligned}\psi_k(x_{k,s}, x_{k,r}) =& \frac{N_k - 2}{N_k - n_k}\frac{N_k - 3}{N_k - n_k - 1}\sum_{i=1}^{N-1}\\ &\times\left(p_i(x_{k,s}, x_{k,r}) - \frac{N_k - 1}{N_k - 2}\bigl(p_i(x_{k,s}) + p_i(x_{k,r})\bigr) + \frac{N_k}{N_k - 2}p_i\right)\triangle_i,\end{aligned} \tag{4}$$

$$\begin{aligned}\psi_{ku}(x_{k,s}, x_{u,r}) =& \frac{N_k - 1}{N_k - n_k}\frac{N_u - 1}{N_u - n_u}\\ &\times\sum_{i=1}^{N-1}\bigl(p_i(x_{k,s}, x_{u,r}) - p_i(x_{k,s}) - p_i(x_{u,r}) + p_i\bigr)\ \triangle_i,\end{aligned} \tag{5}$$

where for $1 \leqslant i \leqslant N-1$ we write $\triangle_i = x_{i+1} - x_i$, and denote the probabilities

$$p_i = \mathbf{P}\{X_{\text{med}} > x_i\}, \qquad p_i(x_{k,s}) = \mathbf{P}\{X_{\text{med}} > x_i \mid X_{k,1} = x_{k,s}\},$$
$$p_i(x_{k,s}, x_{k,r}) = \mathbf{P}\{X_{\text{med}} > x_i \mid X_{k,1} = x_{k,s}, X_{k,2} = x_{k,r}\},$$
$$p_i(x_{k,s}, x_{u,r}) = \mathbf{P}\{X_{\text{med}} > x_i \mid X_{k,1} = x_{k,s}, X_{u,1} = x_{u,r}\}.$$

We give these probabilities in (6) and in Proposition 1 below. Note that expressions (3)–(5) are obtained directly from (11) in [1], using the definitions of expectation and conditional expectations, and applying summation by parts formula $\sum_{i=1}^{N}(p_{i-1} - p_i)x_i = -p_N x_N + p_0 x_1 + \sum_{i=1}^{N-1} p_i \triangle_i$ (in the case of expectation) and noting that, by definition, $p_N = 0$ and $p_0 = 1$, and so forth.

Let T be the set of h-tuples $(t_1, \dots, t_h) \in \{0, \dots, n_1\} \times \dots \times \{0, \dots, n_h\}$, which satisfy the condition $\sum_{j=1}^{h} w_j t_j < 0.5$. Here $w_j = N_j/(Nn_j)$. Denote $\mathcal{H}_{N,n,i}(j) = \binom{i}{j}\binom{N-i}{n-j}/\binom{N}{n}$ the probability that a hypergeometric random variable with parameters N, n and i attains the value j. Denote $d_{ij} := N_j F_{N,j}(x_i)$. In [4] is shown that for $0 \leqslant i \leqslant N$,

$$p_i = \sum_{T} \prod_{1 \leqslant j \leqslant h} \mathcal{H}_{N_j, n_j, d_{ij}}(t_j), \tag{6}$$

and then the variance of X_{med} in (1) is

$$\sigma^2 = \sum_{i=1}^{N}(p_{i-1} - p_i)x_i^2 - \left(\sum_{i=1}^{N}(p_{i-1} - p_i)x_i\right)^2. \tag{7}$$

Next we give explicit expressions of the conditional probabilities.

Proposition 1. *Let* $1 \leqslant i \leqslant N-1$.

(i) *For* $1 \leqslant k \leqslant h$ *and* $1 \leqslant s \leqslant N_k$ *we have*

$$p_i(x_{k,s}) = \sum_{T} \varphi_i(k,s) \prod_{1 \leqslant j \leqslant h,\, j \neq k} \mathcal{H}_{N_j, n_j, d_{ij}}(t_j),$$

where

$$\varphi_i(k,s) = \begin{cases} \mathcal{H}_{N_k-1, n_k-1, d_{ik}}(t_k) & \text{if } i \in \mathcal{I}_{21}, \\ \mathcal{H}_{N_k-1, n_k-1, d_{ik}-1}(t_k - 1) & \text{if } i \in \mathcal{I}_{22}, \end{cases}$$

with

$$\mathcal{I}_{21} = \{i\colon x_i < x_{k,s}\}, \qquad \mathcal{I}_{22} = \{i\colon x_i \geqslant x_{k,s}\}.$$

(ii) *For* $1 \leqslant k \leqslant h$ *and* $1 \leqslant s < r \leqslant N_k$ *we have*

$$p_i(x_{k,s}, x_{k,r}) = \sum_{T} \phi_i(k,s;k,r) \prod_{1 \leqslant j \leqslant h,\, j \neq k} \mathcal{H}_{N_j, n_j, d_{ij}}(t_j),$$

where

$$\phi_i(k,s;k,r) = \begin{cases} \mathcal{H}_{N_k-2, n_k-2, d_{ik}}(t_k) & \text{if } i \in \mathcal{I}_{31}, \\ \mathcal{H}_{N_k-2, n_k-2, d_{ik}-1}(t_k - 1) & \text{if } i \in \mathcal{I}_{32}, \\ \mathcal{H}_{N_k-2, n_k-2, d_{ik}-2}(t_k - 2) & \text{if } i \in \mathcal{I}_{33}, \end{cases}$$

with

$$\mathcal{I}_{31} = \{i\colon x_i < x_{k,s} \leqslant x_{k,r}\}, \qquad \mathcal{I}_{32} = \{i\colon x_{k,s} \leqslant x_i < x_{k,r}\},$$
$$\mathcal{I}_{33} = \{i\colon x_{k,s} \leqslant x_{k,r} \leqslant x_i\}.$$

(iii) *For* $1 \leqslant k < u \leqslant h$ *and* $1 \leqslant s \leqslant N_k$, $1 \leqslant r \leqslant N_u$ *we have*

$$p_i(x_{k,s}, x_{u,r}) = \sum_T \theta_i(k,s;u,r) \prod_{1 \leqslant j \leqslant h,\, j \neq k,u} \mathcal{H}_{N_j,n_j,d_{ij}}(t_j),$$

where

$$\theta_i(k,s;u,r) = \begin{cases} \mathcal{H}_{N_k-1,n_k-1,d_{ik}}(t_k)\mathcal{H}_{N_u-1,n_u-1,d_{iu}}(t_u) & \text{if } i \in \mathcal{I}_{41}, \\ \mathcal{H}_{N_k-1,n_k-1,d_{ik}-1}(t_k-1)\mathcal{H}_{N_u-1,n_u-1,d_{iu}}(t_u) & \text{if } i \in \mathcal{I}_{42}, \\ \mathcal{H}_{N_k-1,n_k-1,d_{ik}}(t_k)\mathcal{H}_{N_u-1,n_u-1,d_{iu}-1}(t_u-1) & \text{if } i \in \mathcal{I}_{43}, \\ \mathcal{H}_{N_k-1,n_k-1,d_{ik}-1}(t_k-1)\mathcal{H}_{N_u-1,n_u-1,d_{iu}-1}(t_u-1) & \text{if } i \in \mathcal{I}_{44}, \end{cases}$$

with

$$\mathcal{I}_{41} = \{i\colon x_i < x_{k,s},\ x_i < x_{u,r}\}, \qquad \mathcal{I}_{42} = \{i\colon x_i \geqslant x_{k,s},\ x_i < x_{u,r}\},$$
$$\mathcal{I}_{43} = \{i\colon x_i < x_{k,s},\ x_i \geqslant x_{u,r}\}, \qquad \mathcal{I}_{44} = \{i\colon x_i \geqslant x_{k,s},\ x_i \geqslant x_{u,r}\}.$$

Proof. Calculations of all conditional probabilities are based on the same arguments as the derivation of (6) in [4]. Here for every of cases (i)–(iii) we need to consider, under fixed conditions, a few different positions of x_i only. Note that the set T is the same for all probabilities, since we use the convention that $\binom{b}{a} = 0$ if $a < 0$; as well as the convention that $\binom{b}{a} = 0$ if $a > b$. □

Empirical approximation. The parameters $\alpha = \alpha(\mathcal{X})$, $\kappa = \kappa(\mathcal{X})$ and $\sigma^2 = \sigma^2(\mathcal{X})$ defining approximation (1) are usualy unknown characteristics of the population $\mathcal{X}$. Thus they should be estimated in practice. In [4], for the estimation of the parameter σ^2, convenient plug-in rule was proposed, where strata distribution functions were replaced by their corresponding empirical versions. However, it is not convenient for the estimation of α and κ. Another way is to replace the population parameters by their jackknife estimators, see [2]. But it is well known that in the case of sample median (or other empirical quantiles) jackknife estimators often fail.

Here we consider the finite population bootstrap of [3]. Let $\eta = \eta(\mathcal{X})$ be any characteristic of the population $\mathcal{X}$. For $1 \leqslant k \leqslant h$ write $N_k = m_k n_k + l_k$, where $0 \leqslant l_k < n_k$. Given the sample $\mathbb{X}_k$ drawn from the stratum $\mathcal{X}_k$ construct an empirical stratum $\mathcal{X}_k^*$ by combining m_k copies of $\mathbb{X}_k$ with a simple random sample without replacement $\mathbb{Y}_k = \{Y_{k,1}, \dots, Y_{k,l_k}\}$ of size l_k from $\mathbb{X}_k$. Then $\mathcal{X}^* = \mathcal{X}_1^* \cup \dots \cup \mathcal{X}_h^*$ is an empirical (bootstrap) population, and the bootstrap estimator of η is then defined as

$$\hat{\eta} = \mathbf{E}\bigl(\eta(\mathcal{X}^*) \bigm| \mathbb{X}\bigr). \tag{8}$$

Thus we have the bootstrap estimators $\hat{\alpha}$, $\hat{\kappa}$ and $\hat{\sigma}^2$ of α, κ and σ^2. However, it is difficult to obtain their explicit expresions. Therefore, here we apply Monte Carlo (MC) approximations for the parameters we are interested in. In particular, let

$\mathcal{X}^*_{(1)}, \ldots, \mathcal{X}^*_{(B)}$ be B empirical populations constructed independently as described above, i.e., we randomly and with replacement select B empirical populations from all possible $\prod_{k=1}^{h} \binom{n_k}{l_k}$. Then MC approximation to (8) is

$$\tilde{\eta} = \frac{1}{B} \sum_{b=1}^{B} \eta(\mathcal{X}^*_{(b)}). \tag{9}$$

Finally, replacing the true parameters α, κ and σ^2 in (1) by their estimates $\tilde{\alpha}$, $\tilde{\kappa}$ and $\tilde{\sigma}^2$ we obtain the empirical approximation $\tilde{G}(\cdot)$ to $F_{\mathrm{med}}(\cdot)$.

2 Simulation study and conclusions

We perform a numerical comparison of distribution of sample median $F_{\mathrm{med}}(\cdot)$ with its approximations $\varPhi(\cdot)$, $G(\cdot)$ and $\tilde{G}(\cdot)$. Here 'exact' distribution $F_{\mathrm{med}}(\cdot)$ is obtained by a MC simulations, i.e., by drawing independently 10^5 stratified samples from $\mathcal{X}$. In particular, in tables below we present q-quantiles $q = 0.01, 0.05, 0.10, 0.90, 0.95, 0.99$ of $F_{\mathrm{med}}(\cdot)$, $\varPhi(\cdot)$, $G(\cdot)$ and $\tilde{G}(\cdot)$. For empirical approximation $\tilde{G}(\cdot)$ we give two characteristics for each of the empirical q-quantiles: estimated values of its expectation $\mathbf{E}\tilde{G}^{-1}(q)$ and standard error $\mathbf{S}\tilde{G}^{-1}(q)$ based on 10^2 stratified samples drawn independently from $\mathcal{X}$. To estimate the parameters α, κ and σ^2 by (9) we take $B = 30$.

A population for two examples below consists of Lithuanian service enterprises with economic activity classified as 'combined facilities support activities'. For our purposes we take three completely sampled strata of sizes $N_1 = 25$, $N_2 = 7$ and $N_3 = 13$, and for our simulations we choose sample sizes $n_1 = 10$, $n_2 = 3$ and $n_3 = 5$. Tables 1 and 2 present simulation results for the populations $\mathcal{X}^{(1)} = \mathcal{X}_1^{(1)} \cup \mathcal{X}_2^{(1)} \cup \mathcal{X}_3^{(1)}$ and $\mathcal{X}^{(2)} = \mathcal{X}_1^{(2)} \cup \mathcal{X}_2^{(2)} \cup \mathcal{X}_3^{(2)}$, where elements of $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ are measurements of income and number of persons employed respectively. We use the first-quarter data of year 2011.

Table 1 shows that $G(\cdot)$ significantly improves $\varPhi(\cdot)$. However, it is not the case for its empirical version $\tilde{G}(\cdot)$, since for a large part of the samples this approximation to $F_{\mathrm{med}}(\cdot)$ is less accurate than $\varPhi(\cdot)$. Table 2 shows that $G(\cdot)$ evidently outperforms $\varPhi(\cdot)$. Here the variability $\mathbf{S}\tilde{G}^{-1}(\cdot)$ is comparatively small, thus $\tilde{G}(\cdot)$ is also more efficient than $\varPhi(\cdot)$.

We note that the proposed approximations may be very efficient in real surveys where we need to measure the accuracy of median in small domains of a population (for some collections of strata) and where populations are highly skewed. Our formulas with minor modifications are applicable for any quantile.

Table 1. The case of $\mathcal{X}^{(1)}$.

q	0.01	0.05	0.10	0.90	0.95	0.99
$F_{\mathrm{med}}^{-1}(q)$	−2.847	−2.013	−1.941	0.959	0.959	0.959
$\varPhi^{-1}(q)$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$G^{-1}(q)$	−2.816	−1.946	−1.413	1.212	1.458	1.805
$\mathbf{E}\tilde{G}^{-1}(q)$	−2.210	−1.635	−1.281	1.324	1.735	2.419
$\mathbf{S}\tilde{G}^{-1}(q)$	0.419	0.195	0.079	0.086	0.205	0.418

Table 2. The case of $\mathcal{X}^{(2)}$.

q	0.01	0.05	0.10	0.90	0.95	0.99
$F_{\mathrm{med}}^{-1}(q)$	−1.596	−1.260	−0.819	1.897	1.988	3.852
$\Phi^{-1}(q)$	−2.326	−1.645	−1.282	1.282	1.645	2.326
$G^{-1}(q)$	−1.836	−1.470	−1.215	1.398	1.918	2.787
$\mathbf{E}\tilde{G}^{-1}(q)$	−2.001	−1.529	−1.238	1.352	1.813	2.633
$\mathbf{S}\tilde{G}^{-1}(q)$	0.172	0.062	0.023	0.043	0.099	0.162

References

[1] M. Bloznelis. A note on the bias and consistency of the jackknife variance estimator in stratified samples. *Statistics*, **37**(6):489–504, 2003.

[2] M. Bloznelis. Second-order and resampling approximation of finite population U-statistics based on stratified samples. *Statistics*, **41**(4):321–332, 2007.

[3] J.G. Booth, R.W. Butler and P. Hall. Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.*, **89**:1282–1289, 1994.

[4] S. Gross. Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, pp. 181–184. American Statistical Association, 1980.

[5] J. Shao. L-statistics in complex survey problems. *Ann. Statist.*, **22**(2):946–967, 1994.

REZIUMĖ

Medianos skirstinio aproksimacijos sluoksninėms imtims

A. Čiginas, T. Rudys

Darbe tiriamos empirinės medianos pasiskirstymo funkcijos Edžvorto tipo aproksimacijos, kai imtis yra sluoksninė ir renkama be grąžinimo. Pateikiamas išreikštinis aproksimacijos pavidalas, bei jos saviranko̧s įvertinys. Skaitiniais pavyzdžiais parodoma, kad tokios aproksimacijos gali būti tikslesnės už normaliąją.

Raktiniai žodžiai: baigtinė populiacija, sluoksninis ėmimas be grąžinimo, Hoeffding'o skleidinys, Edžvorto skleidinys, saviranka.