*STATISTICS IN TRANSITION new series, Autumn 2014*
*Vol. 15, No. 4, pp. 525–534*

# ON CONDITIONAL SIMPLE RANDOM SAMPLE

**Janusz L. Wywiał** [1]

## ABSTRACT

Estimation of the population average in a finite and fixed population on the basis of the conditional simple random sampling design dependent on order statistics of the auxiliary variable is studied. The sampling scheme implementing the sampling design is proposed. The inclusion probabilities are derived. The well known Horvitz-Thompson statistic under the conditional simple random sampling designs is considered as the estimator of population mean. Moreover, it was shown that the Horvitz-Thompson estimator under some particular cases of the conditional simple random sampling design is more accurate than the ordinary mean from the simple random sample.

**Key words:** simple random sample, conditional sampling design, sampling scheme, inclusion probabilities, auxiliary variable, order statistics.

## 1. Introduction

The sampling designs dependent on an auxiliary variables are constructed in order to improve accuracy of population parameters estimation. Rao (1985) considered problems of conditional statistical inference in survey sampling. Applications of auxiliary information to construction of the conditional versions of sampling designs were discussed in the literature, for instance by Tillé (1998, 2006). This paper was inspired by Royall and Cumberland (1981) proposition of conditional simple sampling design.

Let $U$ be a fixed population of size $N$. The observation of a variable under study and an auxiliary variable are identifiable and denoted by $y_i$ and $x_i, i = 1, \ldots, N$, respectively. We assume that $x_i \leq x_{i+1}, i = 1, \ldots, N-1$. Our main purpose is to estimate the population average: $\bar{y} = \frac{1}{N}\sum_{i\in U} y_i$.

Let us consider the sample space $\mathbf{S}$ of the samples $s$ of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$ where $P(s) > 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$. As it is well known the simple sampling design is defined as follows:

$$P_0(s) = \binom{N}{n}^{-1} \qquad \text{for all} \quad s \in \mathbf{S}. \tag{1}$$

[1]Department of Statistics, University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland. E-mail:janusz.wywial@ue.katowice.pl.

Royall and Cumberland (1981) considered drawing the simple random sample $s$ until the inequality $|\bar{x}_s - \bar{x}| \leq c$ where $\bar{x} = \frac{1}{N}\sum_{i\in U} x_i$, $c > 0$, is fulfilled. This sampling scheme can be called the conditional simple random sampling. Of course conditions can be stated by means of other inequalities, see e.g. Wywiał (2003) using computer simulation analysis because the inclusion probabilities of the conditional simple random sampling design are not known. Derivation of those probabilities is one of our purposes. In the considered case the condition is defined on the basis of the properties of the order statistics of the auxiliary variable. In the next section, the conditional simple random sampling design is defined and the inclusion probabilities are derived. The sampling scheme described in the third section. In the fourth section the Horvitz-Thompson estimator is considered. Next we can find some general conclusions about the properties of the considered estimation strategy. The proof of the theorems is in the Appendix.

Let $s = \{s_1, i, s_2\}$ where $s_1 = \{i_1, ..., i_{r-1}\}$, $s_2 = \{i_{r+1}, ..., i_n\}$, $i_j < i$ for $j = 1, ..., r$, $i_r = i$ and $i_j > i$ for $j = r+1, ..., n$. Hence, $x_i$ is one of the possible observations of the order statistic $X_{(r)}$ of the rank $r$ $(r = 1, ..., n)$ from the sample $s$. Let $\mathbf{S}(r,i) = \{s : X_{(r)} = x_i\}$ be the set of all samples whose $r$ -th order statistic of the auxiliary variable is equal to $x_i$ where $r \leq i \leq N - n + r$. Hence, $\bigcup_{i=r}^{N-n+r} \mathbf{S}(r,i) = \mathbf{S}$. The size of the set $\mathbf{S}(r,i)$ is denoted by $g(r,i) = Card(\mathbf{S}(r,i))$ and

$$g(r,i) = \binom{i-1}{r-1}\binom{N-i}{n-r}, \qquad \sum_{i=r}^{N-n+r} g(r,i) = \binom{N}{n}.$$

The probability that the $r$ -th order statistic from simple random sample of an auxiliary variable takes value $x_i$ is as follows (see Wilks (1962), pp. 243-244 or Guenther (1975) or Hogg and Craig (1970)):

$$P\left(X_{(r)} = x_i\right) = \frac{g(r,i)}{\binom{N}{n}}, \quad i = r, ..., N - n + r.$$

$$E\left(X_{(r)}\right) = \sum_{i=r}^{N-n+r} x_i P\left(X_{(r)} = x_i\right) = \frac{1}{\binom{N}{n}} \sum_{i=r}^{N-n+r} x_i g(r,i).$$

The sample quantile of order $\alpha \in (0;1)$ is defined as $Q_{s,\alpha} = X_{(r)}$. The rank $r$ can be determined as follows: $r = [n\alpha] + 1$ where $[.]$ is the integer part of the value $n\alpha$. Hence, $r = 1, 2, ..., n$ and $X_{(r)} = Q_{s,\alpha}$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$. So, it will be more convenient to consider the order statistics than the quantiles.

The conditional (truncated) version of the order statistic distribution is as follows:

$$P\left(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w\right) = \frac{P\left(X_{(r)} = x_i\right)}{P\left(x_u \leq X_{(r)} \leq x_w\right)} = \frac{g(r,i)}{z(r,u,w)} = \\ = P\left(X_{(r)} = x_i | r, u, w\right)$$

where

$$P\left(x_u \leq X_{(r)} \leq x_w\right) = \frac{z(r,u,w)}{\binom{N}{n}},$$

$$z(r,u,w) = \sum_{t=u}^{w} g(r,t). \tag{2}$$

## 2. Sampling design

On the basis of the previous section we have obtained:

$$P_0\left(s \in \mathbf{S}(r,i)\right) = \sum_{s \in \mathbf{S}(r,i)} P_0(s) = \frac{z(r,u,w)}{\binom{N}{n}} = P\left(x_u \leq X_{(r)} \leq x_w\right).$$

Hence,

$$P_0\left(s|s \in \mathbf{S}(r,i)\right) = \frac{P_0(s)}{P_0\left(s \in \mathbf{S}(r,i)\right)} = \frac{1}{z(r,u,w)} =$$
$$= \frac{P_0(s)}{P\left(X_{(r)} = x_i|r,u,w\right)} = P_0\left(s|x_u \leq X_{(r)} \leq x_w\right) = P_0\left(s|r,u,w\right). \tag{3}$$

***Definition 2.1***. *The sampling design expressed by the equations (3) and (2) will be called the conditional simple random sampling design.*

So, the introduced sampling design provides such the simple random samples where $r$-the order $X_{(r)}$ takes value from the interval $[x_u;\ x_u]$ where $u \leq r \leq w$.

The inclusion probability of the first and second orders are defined by the following equation: $\pi_k = \sum_{\{s:k \in s\}} P(s)$ and $\pi_{k,t} = \sum_{\{s:k \in s, t \in s\, k \neq t\}} P(s)$, respectively where $k,t = 1,...,N$. Let us assume that if $x \leq 0$, $\delta(x) = 0$ else $\delta(x) = 1$. Let us note that $\delta(x)\delta(x-1) = \delta(x-1)$.

In the Appendix the following theorem is proved on the basis of Wywial's (2008) results.

***Theorem 2.1.*** *The inclusion probabilities of the first order for the conditional simple random sampling design $P_0(s|r,u,w)$ are as follows: if $k < u$,*

$$\pi_k^{(r)}(u,w) = \frac{\delta(r-1)\delta(w-1)\delta(u-1)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-2}{r-2}\binom{N-i}{n-r},$$

*If* $u \le k \le w$,

$$\pi_k^{(r)}(u,w) = \\ = \frac{1}{z_r(u,w)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}+\right. \\ \left.+\binom{k-1}{r-1}\binom{N-k}{n-r}+\delta(r-1)\delta(w-k)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i}{n-r}\right),$$

*if* $k > w$,

$$\pi_k^{(r)}(u,w) = \frac{\delta(n-r)\delta(N-w)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1},$$

*The inclusion probabilities of the second order for the conditional simple random sampling design* $P_0(s|r,u,w)$ *are as follows:*
*If* $k < u$, $t < u$ *and* $t \neq k$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(r-2)\delta(w-2)\delta(u-2)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}.$$

*If* $k > w$, $t > w$ *and* $t \neq k$,

$$\pi_{k,t}^{(r)}(u,w) = \\ = \frac{\delta(n-r-1)\delta(N-w-1)\delta(N-u-1)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2}.$$

*If* $k < u$ *and* $t > w$ *or* $t < u$ *and* $k > w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(r-1)\delta(n-r)\delta(u-1)\delta(N-w)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}.$$

*If* $k < u$ *and* $u \le t \le w$ *or* $t < u$ *and* $u \le k \le w$,

$$\pi_{k,t}^{(r)}(u,w) = \\ = \frac{\delta(r-1)}{z_r(u,w)}\left(\delta(n-r)\delta(t-u)\delta(t-2)\sum_{i=u}^{t-1}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}+\right. \\ +\delta(t-1)\binom{t-2}{r-2}\binom{N-t}{n-r}+ \\ \left.+\delta(r-2)\delta(w-t)\delta(w-2)\delta(t-1)\sum_{i=t+1}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}\right).$$

*If* $u \le k \le w$ *and* $t > w$ *or* $u \le t \le w$ *and* $k > w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(n-r)}{z_r(u,w)}\Bigg(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1)\cdot$$
$$\cdot\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2} + \delta(N-k)\binom{k-1}{r-1}\binom{N-k-1}{n-r-1}+$$
$$+\delta(r-1)\delta(w-k)\delta(N-w)\delta(w-1)\delta(N-k-1)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}\Bigg).$$

*If* $u \le k < t \le w$ *or* $u \le t < k \le w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(w-u)}{z_r(u,w)}\Bigg(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1)\cdot$$
$$\cdot\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2} + \delta(n-r)\delta(N-k)\binom{k-1}{r-1}\binom{N-k-1}{n-r-1}+$$
$$+\delta(r-1)\delta(n-r)\delta(t-k-1)\delta(t-2)\delta(N-k-1)\sum_{i=k+1}^{t-1}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}+$$
$$+\delta(r-1)\delta(t-1)\binom{t-2}{r-2}\binom{N-t}{n-r}+$$
$$+\delta(r-2)\delta(w-t)\delta(w-2)\delta(t-1)\sum_{i=t+1}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}\Bigg).$$

***Example 2.1***. *Let us assume that* $N = 11$, $n = 5$, $r = 3$.
*When* $u = 4$ *and* $w = 8$, *then* $\pi_k = 0.431$ *and* $\pi_t = 0.483$ *for* $k = 1, 2, 3, 9, 10, 11$ *and* $t = 4, 5, 6, 7, 8$.
*When* $u = 5$ *and* $w = 7$, *then* $\pi_k = 0.411$ *and* $\pi_t = 0.571$ *for* $k = 1, 2, 3, 4, 8, 9, 10, 11$ *and* $t = 5, 6, 7$.
*When* $u = 6$ *and* $w = 6$, *then* $\pi_6 = 1$ *and* $\pi_t = 0.4$ *for* $t \neq 6$.
*Finally, when* $u = 3$ *and* $w = 9$, *then* $\pi_k = \frac{5}{11} = 0.45(45)$ *for* $k = 1, ..., 11$. *In this case the conditional simple random sampling design reduces to the simple random sample drawn without replacement.*

*Hence, when the parameters* $u$ *and* $w$ *are closer and closer to each other then the probability of selecting to the sample the central population elements increases.*

## 3. Sampling scheme

The sampling scheme implementing the conditional simple random sampling design $P_0(s|r,u,w)$, where $r \le u \le w \le N-n+r$ is as follows. Firstly, population

elements are ordered according to increasing values of the auxiliary variable. Next, the $i$-th element of the population where $i = u, u+1, ..., w$, is drawn with the probability

$$P(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w) = \frac{P(X_{(r)} = x_i)}{P(x_u \leq X_{(r)} \leq x_w)} = \frac{g(r,i)}{\sum_{j=u}^{w} g(r,j)} \tag{4}$$

where $r = [n\alpha] + 1$.

Finally, two simple samples $s_1(i)$ and $s_2(i)$ are drawn without replacement from the subpopulations $U_1 = \{1, ..., i-1\}$ and $U_2 = \{i+1, i+2, ..., N\}$, respectively. The sample $s_1(i)$ is of the size $r-1$ and the sample $s_2(i)$ is of the size $n-r$. The sampling designs of these samples are independent and

$$P_0(s_1(i)) = \frac{1}{\binom{i-1}{r-1}}, \quad P_0(s_2(i)) = \frac{1}{\binom{N-i}{n-r}} \tag{5}$$

Hence, the selected sample is: $s = \{s_1(i), i, s_2)(i)\}$ and its probability is:

$$P(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w) P_0(s_1(i)) P_0(s_2(i)) = P_0(s|r,u,w)$$

where $r = u, u+1, ..., w$.

## 4. The Horvitz-Thompson estimator

The well-known Horvitz-Thompson (1952) estimator is given by:

$$\bar{y}_{HT,s} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} \tag{6}$$

The estimation strategy $(\bar{y}_{HT,s}, P(s))$ is unbiased for the population mean $\bar{y}$ if $\pi_k > 0$ for $k = 1, ..., N$, where $\pi_k$ is the inclusion probability of the sampling design $P(s)$. The variance of the strategy is:

$$V_0\left(\bar{y}_{HT,s}, P(s)\right) = \frac{1}{N^2} \left( \sum_{k \in U} \sum_{l \in U} \Delta_{k,l} \frac{y_k y_l}{\pi_k \pi_l} \right), \qquad \Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l \tag{7}$$

Particularly, under the simple random sampling design $P_0(s)$ the strategy $(t_{HT,s}, P(s))$ reduces to simple random sample mean denoted by $(\bar{y}_s, P_0(s))$, where

$$\bar{y}_s = \frac{1}{n} \sum_{k \in s} y_k. \tag{8}$$

It is an unbiased estimator of the population mean and its variance is given by:

$$V_0(\bar{y}_s) = \frac{N-n}{Nn} v_*(y), \qquad v_*(y) = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y})^2.$$

***Example 4.1***. *Let us assume that in the population of the size* $N = 11$ *the following values* $(x, y)$ *of the two dimensional variable is observed* $\{(1,2), (2,6), (3,10), (4,14), (5,15), (6,16), (7,17), (8,18), (9,22), (10,26), (11,30)\}$. *Let the sample*

*of size $n = 5$ be selected from that population according to the conditional sampling design. The variance of the simple sample mean is: $V(\bar{y}_s) = V(\bar{y}_{HTs}, P_0(s)) = V(\bar{y}_{HTs}, P_0(s|3,3,9)) = 7.353$. The variances of the Horvitz-Thompson estimator under the conditional design of simple sample are: $V(\bar{y}_{HTs}, P_0(s|3,4,8)) = 5.954$, $V(\bar{y}_{HTs}, P_0(s|3,5,7)) = 4.918$, $V(\bar{y}_{HTs}, P_0(s|3,6,6)) = 3.694$. The inclusion probabilities of the conditional simple random sample are shown in the Example 2.1. Hence, the accuracy of estimation of the population mean on the basis of the Horvitz-Thompson statistic under the considered variants of the conditional simple random sample is better than the accuracy of the mean from the unconditional simple random sample.*

## 5. Conclusions

The sampling design belonging to the class of the sampling designs dependent on the sample parameters of an auxiliary variable was proposed. It is the conditional version of the simple random sampling design explained by Definition 2.1 and denoted by $P_0(s|x_u \leq X_{(r)} \leq x_w)$. Let $M_s$ be the sample median of the auxiliary variable. So, when we assume that the distribution of an auxiliary variable is symmetric then $\bar{x} = M$, where $M$ is the population median of the auxiliary variable. When we assume that the distribution of the sample median is approximation of the distribution of the sample mean $\bar{x}_s$ then the simple random sample design $P_0(s|x_u \leq M_s \leq x_w)$ can be treated as approximation of the simple random sampling design $P_0(s|x_u \leq \bar{x}_s \leq x_w)$, defined by Royall and Cumberland (1981). Our consideration can be generalized to the case when the distribution of the auxiliary variable is not necessary symmetric. It is possible to find such rank $r$ of the order statistic $|E(X_{(r)}) - \bar{x}| = minim$. So, when we assume that the distribution of the sample mean $\bar{x}_s$ is sufficiently approximated by the distribution of the order statistic $X_{(r)}$ then the sampling design $P_0(s|x_u \leq \bar{x}_s \leq x_w)$ can be approximated by the sampling design $P_0(s|x_u \leq X_{(r)} \leq x_w)$.

We can expect that the sampling design can be useful in the case of censored observations of the auxiliary variable as well as when the outliers exist. The precision of the Horvitz-Thompson estimator depends on the parameters $u$ and $w$ through probabilities of the inclusion of the first and second order.

The derived properties of the sampling designs lead to the conclusion that without an additional extensive analysis it is not possible to determine precisely how the sampling strategies depend on the parameters of the conditional simple random sampling design as well as on the joint distribution of the variable under study and the auxiliary variable. This problem will be considered on the basis of computer simulation analysis in another papers. Moreover, such analysis makes it possible to compare the accuracy of the proposed estimation strategies with accuracy of the strategies typically used in statistical research.

**Acknowledgement**

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

## REFERENCES

GUENTER W. (1975). The inverse hypergeometric - a useful model. Statistica Neerlandica, Vol. 29, pp. 129–144.

HOGG, R. V., CRAIG, A. T., (1970). Introduction to Mathematical Statistics, 3rd edition. MacMillian, New York.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of the sampling without replacement from finite universe. Journal of the American Statistical Association, Vol. 47, pp. 663–685.

ROYALL, R. M., Cumberland W. G., (1981). An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association, Vol. 76.

TILLÉ, Y., (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. International Statistical Review, 66, 303–322.

TILLÉ, Y., (2006). Sampling Algorithms. Springer.

WILKS, S. S., (1962). Mathematical Statistics. John Wiley and Sons, Inc. New York, London.

WYWIAŁ, J. L., (2003). On conditional sampling strategies. Statistical Papers, Vol. 44, 3, pp. 397–419.

WYWIAŁ, J. L., (2008). Sampling design proportional to order statistic of auxiliary variable. Statistical Papers, Vol. 49, No. 2/April, pp. 277–289.

## APPENDIX

The theorems 4.1 formulated in the previous sections is proved here.
Let $\mathbf{S}(U(1,...,i-1),s_1(i))$ and $\mathbf{S}(U(i+1,...,N),s_2(i))$ be the sample spaces of the samples $s_1(i)$ and $s_2(i)$ and $s = s_1(i) \cup \{i\} \cup s_2(i)$, defined in Section 1. Hence,

$$\mathbf{S}(r,i) = \mathbf{S}(U(1,...,i-1),s_1(i)) \times \{i\} \times \mathbf{S}(U(i+1,...,N),s_2(i))$$

and

$$\mathbf{S}(r;u,w) = \mathbf{S}(r,u) \times \mathbf{S}(r,u+1) \times ... \times \mathbf{S}(r,i) \times ... \times \mathbf{S}(r,w)$$

where $\mathbf{S}(r,i)$ was defined in Section 1.

Wywiał (2008) proposed the following conditional sampling design:

***Definition 6.1***. *The conditional sampling design proportional to the values $x_i$, $i = u,...,w \leq N-n+r$, $u \geq r$, of the order statistic $X_{(r)}$ is as follows:*

$$P_r(s|u,w) = \frac{x_i}{\sum_{j=u}^{w} x_j g(r,j)}$$

*where* $\quad i \in s \in \boldsymbol{S}(r,i), \quad r \leq u \leq i \leq w \leq N-n+r.$

Moreover, Wywiał (2008) proved the theorem:

***Theorem 6.1.*** *The inclusion probabilities of the first order for the conditional simple random sampling design $P_r(s|u,w)$ are as follows:*
*If $k < u$,*

$$\pi_k^{(r)}(u,w) = \frac{\delta(r-1)\delta(w-1)\delta(u-1)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-2}{r-2}\binom{N-i}{n-r} x_i,$$

*If $u \leq k \leq w$,*

$$\pi_k^{(r)}(u,w) =$$

$$= \frac{1}{z_r(u,w)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}x_i + \right.$$

$$\left. + \binom{k-1}{r-1}\binom{N-k}{n-r}x_k + \delta(r-1)\delta(w-k)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i}{n-r}x_i\right),$$

*if $k > w$,*

$$\pi_k^{(r)}(u,w) = \frac{\delta(n-r)\delta(N-w)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} x_i,$$

When we replace $x_i$ by 1 for all $i = 1, ..., N$, then the above definition 6.1 and the expression (3) lead to the Definition 2.1. The same operation and the above Theorem 6.1 lead straightforward to the derivation of the first order inclusion probabilities of the conditional simple random sampling design $P_0(s|r, u, v)$, given by expressions (2) and (3). The inclusion probabilities of the second order presented by Theorem 6.1. can be straightforward derived in the same way but on the basis of the appropriate theorem proven by Wywiał (2008).