Janusz L. Wywiał
Uniwersytet Ekonomiczny w Katowicach

ON LIMIT DISTRIBUTION OF HORVITZ-THOMPSON STATISTIC UNDER POISSON SAMPLING DESIGN

Introduction

Let U_N be a fixed population of the size N, so, $N = 2, 3, ...$ The elements of the population are identified. So, the population can be represented by the set: $U_N = \{1, ..., N\}$. The observation of a variable under study is denoted by $y_{k,N}$, $k = 1, ..., N$, $N = 2, 3, ...$ So, the vector $\mathbf{y}_N = [y_{1,N}\ y_{2,N} \ldots y_{N,N}]$ is attached to the set U_N. Particularly, when we assume that $U_N \subset U_{N+1}$ then the observations of a variable in the population can be represented more simply by the vector: $\mathbf{y}_{N+1} = [\mathbf{y}_N y_{N+1}]$ where $\mathbf{y}_N = [y_1, y_2 \ldots y_N]$. A more particular case is as follows. Let $(y_i, w_i N_t)$ means that the value y_i is $w_i N_t$ times replicated in the population, where $\Sigma_{i=1}^{k} w_i = 1$ and $0 < w_1 < 1$ for $i = 1, 2, ..., k$, $t = 1, 2, ...$, and the vector $\mathbf{y}_N = [y_k\ y_k, ..., y_k]$ is fixed. The size N_t of the population U_{N_t} is determined in such a way that $w_i N_t$ is an integer for all $i = 1, ..., k$. So,

$$\mathbf{y}_{N_t} = [(y_1, w_1 N_t)(y_2, w_2 N_t) \ldots (y_k, w_k N_t)].$$

We assume that the all elements of the population can be selected for the sample with different probabilities. A k-th population element, $k \in U_N$, be selected for the sample with the inclusion probability $0 < \pi_{k,N} < 1$, $k = 1, ..., N$. More precisely, let $S_N = [S_{1,N} \ldots S_{N,N}]$ be the vector of independent binary random variables and

$$P(S_{k,N}=1)=\pi_{k,N}=1-P(S_{k,N}=0) \tag{1}$$

So, $s_N=[s_{1,N}\ldots s_{N,N}]$ is the realization of the random sample S.

The probability distribution of the random sample S is known as Poisson sampling design [see, e.g. Tille 2006]:

$$P(S_N=s_N)=\prod_{k=1}^{N}\pi_{k,N}^{s_{k,N}}(1-\pi_{k,N})^{1-s_{k,N}}.$$

The population total $\tilde{y}=\sum_{k\in U_N}y_{k,N}$ can be estimated on the basis of the Horvitz-Thompson statistic [1952]:

$$y_{HTS_N}=\sum_{k\in U_N}\frac{y_{k,N}S_{k,N}}{\pi_{k,N}}.$$

It is well known that $E(y_{HTS_N})=\tilde{y}$ if $\pi_{k,N}>0$ for all $k=1,...,N$. Because of $P(S_{k,N}=1,S_{h,N}=1)=\pi_{k,N}\pi_{h,N}$ for all $k=1,...,N$, $h=1,...,N$ and $k\neq h$ the variance of the statistic y_{HTS_N} is:

$$V\left(y_{HTS_N}\right)=\sum_{k\in U_N}\frac{y_{k,N}^2(1-\pi_{k,N})}{\pi_{k,N}}. \tag{2}$$

Its unbiased estimator is:

$$V\left(y_{HTS_N}\right)=\sum_{k\in U_N}\frac{y_{k,N}^2(1-\pi_{k,N})S_{k,N}}{\pi_{k,N}^2}. \tag{3}$$

Let

$$b_{3,N}=\frac{1}{N}\sum_{k\in U_N}\left|y_{k,N}\right|^3 \text{ and } v_{2,N}=\frac{1}{N}\sum_{k\in U_N}y_{k,N}^2 \quad v_{4,N}=\frac{1}{N}\sum_{k\in U_N}y_{k,N}^4.$$

The original version and the proof of the Lapunov's [2001] theorem, which is slightly less general, can be found in the monograph by Fisz [1963]. On the basis of the books by Billingsley [2009] or Jakubowski and Sztencel [2004], the following more general version of the theorem is presented.

Theorem 1. Let $Z_{k,N}$, $k = 1, ..., N$, $N = 1, 2, ...$ be a sequence of independent random variables and for some $\delta > 0$

$$\beta_N = \frac{B_N^{\delta}}{C_N^{2+\delta}} \to 0 \text{ if } N \to \infty \tag{4}$$

where:

$$B_N^{\delta} = \sum_{k=1}^{N} E\left|Z_{k,N} - E(Z_{k,N})\right|^{2+\delta}, \; C_N = \sqrt{\sum_{k=1}^{N} V(Z_{k,N})}. \tag{5}$$

Under these Lapunov's conditions, the random variable:

$$Z_N = \frac{\sum_{k=1}^{N}\left(Z_{k,N} - E(Z_{k,N})\right)}{C_N}$$

converges in distribution to the normal standard distribution if $N \to \infty$.

Hájek [1964] considered a limit distribution for the following statistic

$$\overline{H}_S = y_{HTS_N} - \frac{r}{N}\sum_{k=1}^{N}\left(\pi_k - S_k\right)$$

where:

$$r = \frac{\sum_{k=1}^{N} y_{k,N}\left(1-\pi_k\right)}{\sum_{k=1}^{N} \pi_k\left(1-\pi_k\right)}.$$

He proved that the probability distribution of the statistic $\overline{H}_S$ tends to the normal distribution because it fulfils the well known Lindeberg condition.

In the next section, the limit theorem for the estimator y_{HTS_N} will be considered.

1. Limit theorem

Firstly, let us formulate the following statistics and the theorem.

$$T_N = \frac{y_{HTS_N} - \widetilde{y}}{\sqrt{V(y_{HTS_N})}},$$

$$\hat{T}_N = \frac{y_{HTS_N} - \widetilde{y}}{\sqrt{V_{S_N}(y_{HTS_N})}}. \tag{6}$$

We say that $\pi_{k,N} = O(N^{-\alpha})$ if for all $0 \le \alpha < 1$ there exists such a_1 and a_0 that $0 \le a_1 \le a_0 < 1$ and

$$0 < a_1 N^{-\alpha} \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\} \le a_0 N^{-\alpha} < 1 \tag{7}$$

or

$$0 < a_1 \le \frac{\max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\}}{N^{\alpha}} \le a_0 < 1$$

Particularly, if $\alpha = 0$,

$$0 < a_1 \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\} \le a_0 < 1$$

Moreover, $\pi_{k,N}^{-1} = O(N^{\alpha})$ because for all $0 \le \alpha < 1$ there exists such $1 < c_1 \le \frac{1}{a_0} < \frac{1}{a_1} \le c_0$ that

$$1 < c_1 N^{\alpha} \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\frac{1}{\pi_{k,N}}\right\}\right\} \le c_0 N^{\alpha} \tag{8}$$

$\pi_{k,N}^{-1} - 1 = O(N^{\alpha})$ because for all $0 \le \alpha < 1$ there exists such $0 < d_1 \le c_1 - N^{-\alpha} < c_0 - N^{-\alpha} \le d_0$ that

$$0 < d_1 N^{\alpha} \le \max_{N=1,2,...} \left\{ \max_{k=1,...,N} \left\{ \frac{1}{\pi_{k,N}} \right\} \right\} \le d_0 N^{\alpha} \tag{9}$$

$O(N^{\alpha})O(N^{\gamma}) = O(N^{\alpha+\gamma})$ because for all $\alpha \ge 0$ and $\gamma \ge 0$ from the inequalities $0 < d_1 \le N^{\alpha} \le d_0$ and $0 < g_1 \le N^{\gamma} \le g_0$ results the following one

$$0 < e_1 \le d_1 g_1 \le N^{\alpha+\gamma} \le d_0 g_0 \le e_0 \tag{10}$$

Finally, $O(N^{\alpha})O(N^{-\gamma}) = O(N^{\alpha-\gamma})$ because for all $\alpha \ge 0$ and $\gamma \ge 0$ from the inequalities $0 < d_1 \le N^{\alpha} \le d_0$ and $0 < g_1 \le N^{\gamma} \le g_0$ results the following one

$$0 < l_1 \le \frac{d_1}{g_0} \le N^{\alpha-\gamma} \le \frac{d_0}{g_1} \le l_0 \tag{11}$$

Theorem 2. Let $\{\pi_{k,N}\} = O(N^{-\alpha})$ for all $0 \le \alpha < 1$ and $0 < v_0 \le v_{2,N} \le v_2 < \infty$ and $b_{3,N} \le b_3 < \infty$ for $N = 1, 2, ...$ When $N \to \infty$ then $T_N \xrightarrow[d]{} T \sim N(0,1)$. When additionally $v_{4,N} \le v_4 < \infty$ then $\hat{T}_N \xrightarrow[d]{} T \sim N(0,1)$.

Proof: On the basis of the theorem 1, it is sufficient to assume that $\delta = 1$. The Horvitz-Thompson statistic can be rewritten in the following way.

$$y_{HTS_N} = \sum_{k=1}^{N} Z_{k,N}$$

where:

$$Z_{k,N} = \frac{y_{k,N} S_{k,N}}{\pi_{k,N}} \quad k = 1, ..., N$$

On the basis of the expression (1), we have:

$$P(Z_{k,N} = z_{k,N}) = \begin{cases} \pi_{k,N} & \text{if } z_{k,N} = \dfrac{y_{k,N}}{\pi_{k,N}}, \\ 1 - \pi_{k,N} & \text{if } z_{k,N} = 0 \end{cases} \tag{12}$$

and

$$E(Z_{k,N}) = y_{k,N},\ E(Z_{k,N}^2) = \frac{y_{k,N}^2}{\pi_{k,N}},\ V(Z_{k,N}) = y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right),\ k = 1, ..., N$$

$$E\left|Z_{k,N} - E(Z_{k,N})\right|^3 = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^3} E\left|S_{k,N} - \pi_{k,N}\right|^3 = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^3}\left(\left(1-\pi_{k,N}\right)^3 \pi_{k,N} +\right.$$

$$\left. + \pi_{k,N}^3\left(1-\pi_{k,N}\right)\right) = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right)\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right) \le \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right).$$

This and the expression (5) for γ=1 lead to the following.

$$B_N = \sum_{k=1}^{N} \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right)\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right)$$

,\

$$C_N = \sqrt{\sum_{k=1}^{N} \frac{y_{k,N}^2\left(1-\pi_{k,N}\right)}{\pi_{k,N}}} = \sqrt{V(y_{HTS_N})}$$

,

On the basis of the expressions (7)-(11) we have:

$$\beta_N = \frac{B_N}{C_N^3} = \frac{\sum_{k=1}^{N} \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}(1-\pi_{k,N})\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right)\right)^{\frac{3}{2}}} \le$$

$$\le \frac{\sum_{k=1}^{N}\left|y_{k,N}\right|^3 \frac{1}{\pi_{k,N}}\left(\frac{1}{\pi_{k,N}} - 1\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right)\right)^{\frac{3}{2}}} = \frac{\sum_{k=1}^{N}\left|y_{k,N}\right|^3 O\left(N^{\alpha}\right)\left(O\left(N^{\alpha}\right) - 1\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(O\left(N^{\alpha}\right) - 1\right)\right)^{\frac{3}{2}}} =$$

$$= \frac{O\left(N^{2\alpha}\right)\sum_{k=1}^{N}\left|y_{k,N}\right|^3}{\left(O\left(N^{\alpha}\right)\sum_{k=1}^{N} y_{k,N}^2\right)^{\frac{3}{2}}} = \frac{O\left(N^{2\alpha+1}\right)b_{3,N}}{O\left(N^{3(\alpha+1)/2}\right)v_{2,N}^{3/2}} \le \frac{O\left(N^{2\alpha+1}\right)b_3}{O\left(N^{3(\alpha+1)/2}\right)v_0^{3/2}} = O\left(N^{(\alpha-1)/2}\right).$$

It is easy to show that $\beta \to 0$, when $N \to 0$ to and $0 \le \alpha \le \alpha_1 < 1$. This and the theorem 1 lead to the conclusion that $T_N \to T \sim N(0,1)$.

In order to prove the second part of the theorem, we firstly show that $R_N = V_{S_N}(y_{HTS_N}) / V(y_{HTS_N})$ converge in probability to 1. The expression (3) leads to the variance of the sample variance of Horvitz-Thompson statistic:

$$V(V_{S_N}(y_{HTS_N})) = V\left(\sum_{k=1}^{N} \frac{y_k^2(1-\pi_{k,N})S_{k,N}}{\pi_k^2(N)}\right) =$$

$$= \sum_{k=1} \frac{y_k^4(1-\pi_{k,N})^2}{\pi_k^4(N)} V(S_{k,N}) = \sum_{k\in U} y_k^4 \left(\frac{1}{\pi_{k,N}} - 1\right)^3 = \sum_{k\in U} y_k^4 \left(O(N^\alpha) - 1\right)^3 =$$

$$= O\left(N^{(3\alpha)}\right) \sum_{k\in U} y_k^4 = O(N^{3\alpha+1}) v_{4,N}.$$

Hence, on the basis of the expression (2) we have

$$V(R_N) = \frac{V(V_{S_N}(y_{HTS_N}))}{V^2(y_{HTS_N})} = \frac{\sum_{k\in U_N} y_k^4 \left(\frac{1}{\pi_{k,N}} - 1\right)^3}{\left(\sum_{k\in U_N} y_k^2 \left(\frac{1}{\pi_{k,N}} - 1\right)\right)^2} = \frac{\sum_{k\in U} y_k^4 \left(O(N^\alpha) - 1\right)^3}{\left(\sum_{k\in U_N} y_k^2 \left(O(N^\alpha) - 1\right)\right)^2} =$$

$$= \frac{O(N^{3\alpha+1}) v_{4,N}}{O(N^{2(\alpha+1)}) v_{2,N}^2} \le \frac{O(N^{3\alpha+1}) v_4}{O(N^{2(\alpha+1)}) v_0^2} = O(N^{\alpha-1})$$

Hence, $V(R_N) = O(N^{\alpha-1}) \to 0$ when $N \to \infty$ and $0 \le \alpha < 1$. So, this and the well known Tchebyshev's inequality lead to the conclusion that that $R_N = V_{S_N}(y_{HTS_N}) / V(y_{HTS_N})$ converges in probability to 1 (in short: $R_N \xrightarrow[p]{} 1$ if $v_0 > 0$, $v_4 < \infty$ and $N \to \infty$. Let us note that

$$\hat{U}_N = \frac{U_N}{R_N}.$$

Hence, when $N \to \infty$ then $T_N \xrightarrow[d]{} T \sim N(0,1)$ and $R_N \xrightarrow[p]{} 1$. So, this and the well known Slucky's lemma, see e.g. Van der Vaart [2007], let us conclude that $\hat{T}_N \xrightarrow[d]{} T \sim N(0,1)$. So, the proof of Theorem 2 has been completed.

2. Applications

The Poisson sampling design is frequently used to model non-response. In this case, $\pi_{k,N}$ is the probability that a k-th population element will respond. The Poisson sampling design can be treated as a model of the Internet research. In this case, $\pi_{k,N}$ is the probability that a k-th Internet user will respond. Moreover, the Poisson sampling design can be considered in an audit sampling. Let us note that, in the cases mentioned the probabilities $\pi_{k,N}$, $k = 1, ..., N$, $N = 2, ...$ are usually defined as follows.

$$\pi_{k,N} = \mathrm{n}\frac{\mathrm{x_{k,N}}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} x_{i,N}}$$

where n is the expected sample size and x_k is a value of a positive auxiliary variable x observed in all the population. Let us assume that $0 < \alpha \le x_{k,N} \le b < \infty$ and $n = wN$ for all $k = 1, ..., N$, $N = 2, ...$ where $0 < w < \frac{a}{b} \le 1$. So, in this case, the first assumption of the Theorem 2 is fulfilled, because

$$0 < a_0 = w\frac{a}{b} = \frac{na}{bN} \le \pi_{k,N} \le \frac{nb}{aN} = w\frac{b}{a} = a_1 < 1$$

Theorem 2 lets us construct the confidence interval for the mean value estimated by means of the Poisson-Horvitz-Thompson strategy. Let γ be the confidence level and let u_γ be such a quantile that $\phi(u_\gamma) = \frac{1+\gamma}{2}$ where $\phi(u)$ is the distribution function of the standard normal variable. When N is sufficient-

ly large, the confidence interval for the population mean is determined by the expression:

$$P\left(y_{HTS_N} - u_\gamma \sqrt{V_{S_N}(y_{HTS_N})} < \tilde{y} < y_{HTS_N} + u_\gamma \sqrt{V_{S_N}(y_{HTS_N})}\right) = \gamma .$$

It is possible to test the hypothesis on the population mean. The hypothesis $H_0 : \tilde{y} = \tilde{y}_0$ can be tested on the basis of the statistic defined by the expression (6) when N is sufficiently large.

Finally, let us note that if the Lapunov's condition is fulfilled, the Lindeberg's condition is fulfilled [see, e.g. Billingsley 2009], too. Hence, if the assumptions of the above theorem 2 are fulfilled, the assumptions of Hájek's theorem are fulfilled, too. Moreover, it seems that in our case the assumptions of the theorem 2 are verified more simply than the Lindeberg's ones.

Acknowledgements

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

Literature

Billingsley P. (2009): *Prawdopodobieństwo i miara* (*Probability and Measure)*. Wydawnictwo Naukowe PWN, Warszawa.

Fisz M. (1963): *Probability Theory and Mathematical Statistics*. Wiley and Sons, New York.

Hájek J. (1964): *Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population*. "The Annals of Mathematical Statistics", No. 35, 4.

Horvitz D.G., Thompson D.J. (1952): *A Generalization of the Sampling without Replacement from Finite Universe*. "Journal of the American Statistical Association", No. 47.

Jakubowski J., Sztencel R. (2004): *Wstęp do teorii prawdopodobieństwa* (*Introduction to Probability Theory*). SCRIPT, Warszawa.

Lapunov A.M. (1901): *Nouvell forme du theorem sur la limite de probabilite.* „Mem. Acad. Sci. St. Pétersburg", No. 12.

Tillé Y. (2006): *Sampling Algorithms*. Springer, New York.

Van der Vaart A.W. (2007): *Asymptotic Statistic*. Cambridge University Press, Cambridge, New York, Melbourne, Madrit, Cape Town, Singapore, Sao Paulo.

O ROZKŁADZIE GRANICZNYM STATYSTYKI HORVITZA-THOMPSONA DLA PRÓBY DOBIERANEJ ZGODNIE Z PLANEM LOSOWANIA POISSONA

Streszczenie

W pracy na podstawie znanego twierdzenia centralnego Lapunowa jest wyprowadzany rozkład graniczny prawdopodobieństwa znanej statystyki Horvitza-Thompsona (HT). Okazało się, że jeśli określane przez plan losowania Poissona prawdopodobieństwa wylosowania do próby poszczególnych elementów populacji spełniają pewne założenia oraz rozmiar populacji rośnie nieograniczenie, to rozkład standardowej postaci statystyki HT zmierza do rozkładu normalnego standardowego. Taki sam wynik otrzymano przy dodatkowym założeniu narzuconym na prawdopodobieństwa wylosowania elementów populacji do próby, gdy w standardowej postaci statystyki HT jej odchylenie standardowe zastąpimy przez pierwiastek z nieobciążonego estymatora tej wariancji.

Rezultaty pracy znajdują zastosowania np. w pewnych typach badań ankietowych, a w szczególności internetowych, wykorzystujących wnioskowanie statystyczne, czyli estymację przedziałową lub testowanie hipotez statystycznych.