Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

5(965)

ISSN 1898-6447
Zesz. Nauk. UEK, 2017; 5(965): 17–29
DOI: 10.15678/ZNUEK.2017.0965.0502

Janusz L. Wywiał

On the Evaluation of Sample Size Required for a Good Approximation by the Normal Curve for Some Statistics*

Abstract

Testing hypotheses or evaluation confidence intervals requires knowledge of some statistics' distributions. It is convenient if the probability distribution of the statistic converges to normal distribution when the sample size is sufficiently large. This paper examines the problem of how to evaluate sample size in order to determine that a statistic's distribution does not depart from normal distribution by more than an assumed amount. Two procedures are proposed to evaluate the necessary sample size. The first is based on Berry-Esseen inequality while the second is based on simulation procedure. In order to evaluate the necessary sample size, the distribution of the sample mean is generated by replicating samples of a fixed size. Next, the normal distribution of the evaluated sample means is tested. The size of the generated samples is gradually increased until the hypothesis on the normality of the sample mean distribution is not rejected. This procedure is applied in the cases of statistics other than sample mean.

Keywords: sample size, central theorem, sampling scheme, computer simulation, chi-square test of goodness of fit.
JEL Classification: C12, C15.

Janusz L. Wywiał, University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, 1 Maja 50, 40-287 Katowice, e-mail: janusz.wywial@ue.katowice.pl

* This paper presents the results of a research project conducted with the financial support of the National Science Centre in Poland (grant number: DEC-2012/07/B/HS4/03073).

1. Introduction

Statistical inference procedures such as testing hypotheses or evaluating confidence intervals depend on the distribution properties of test statistics or estimators which can be evaluated on the basis of complex samples. Usually, a statistic's exact distribution is not known except as some function of a normal simple random sample. In this situation, it is convenient if the probability distribution of the statistic converges to normal distribution when the sample size is sufficiently large. This leads to the problem of how to evaluate sample size so that the departure of the distribution of the statistic from normal distribution is not larger than the level assumed. This problem is frequently taken into account in statistics literature, but usually where simple random sample is involved. This paper discusses the problem under complex samples drawn from fixed and finite populations.

In some situations, it is possible to observe all values of an auxiliary variable in an entire population. Moreover, let us assume that the value of the correlation coefficient between the auxiliary variable and the variable under consideration here is close to one. In this case, we can expect that the degree of convergence to normal distribution of, e.g., the sample average of the auxiliary variable and the distribution of the sample mean of the variable under consideration will be similar. This allows us to assess the size of the sample, providing a sufficient degree of convergence of the sample mean distribution to normal distribution.

Two procedures are proposed to evaluate the necessary sample size. The first is based on the Berry-Esseen equality, while the second is based on a simulation procedure. The sample mean's distribution is generated by replicating samples of fixed size. Next, the normal distribution of the evaluated sample means is then tested. The size of the generated samples is gradually increased until the hypothesis on the normality of the sample mean distribution is not rejected. The normality of generated values of the sample mean is tested by means of the chi-square test of goodness of fit. The hypothesis on normal distribution is verified under the assumed significance level as well as the power of the test. The outlined procedure is used to assess the necessary sample size of statistics other than sample mean. Complex sampling schemes are also taken into account.

The properties of central limit theorems allow us to evaluate sample sizes in such a way that the probability distribution of, e.g., the standardised sample mean does not differ from standard normal distribution by more than an assumed level. The distribution of the simple random sample frequency is approximated by means of several methods reviewed e.g. by G. A. F. Seber (2013) and T. P. Ryan (2013). In the case of continuous or integer random variables, a bootstrap version of the statistics can be analysed. In this case, the statistical distribution can be

approximated by means of the well-known F. Y. Edgeworth (1907) expansion, which has been detailed by P. Hall (1992). In the case of sampling from a fixed population, the central theorems have been considered e.g. by Y. G. Berger (1998), W. A. Fuller (2009) and J. Hájek (1964, 1981).

Using appropriately prepared computer simulation experiments, it is possible to determine what sample size is necessary to assure sufficient convergence of the distribution of a statistic to the appropriate asymptotic distribution. This problem has been considered e.g. by M. R. Chernick and C. Y. Liu (2002) and T. P. Ryan (2013) in the context of sample frequency. F. Greselin and M. Zenga (2006) considered the simulation analysis for determining sample size, which assures sufficient convergence of Gini's statistic to normality. Some similar ideas are developed below.

Let $(x; y)$ be highly correlated variables observed in a population U of size N. These variables' values are denoted by $(x_i; y_i)$, $i = 1, \ldots, N$. We assume that the values of the auxiliary variable x are observed in the whole population U but the values of the variable under study y are observed only in a sample s of size $n < N$ drawn from U. The random sample will be denoted by S and its observation by s treated as the set consisting of the population elements. The sample is drawn from the population according to sampling design denoted by $P(s) > 0$ for all $s \in \mathbf{S}$ and $\sum_{s \in \mathbf{S}} P(s) = 1$, where $\mathbf{S}$ is the sample space, see e.g. C. M. Cassel *et al*. (1977) or Y. Tillé (2006).

2. Numerical Approximation of Sample Size

Let $z_{x,S}$ and $z_{y,S}$ be statistics evaluated based on data observed in sample S. Because values of x are observed in the whole population U, it is possible to observe values $z_{x,s}$ of $z_{x,S}$ in all samples which can be drawn from population U. In practice, values of variable y are observed only in one sample s. Let us assume that the two-dimensional normal distribution (with the marginal distributions equal to standard normal distribution and the correlation coefficient close to one) is the limit distribution of statistics $(z_{x,S}, z_{y,S})$. This convergence can be proved using H. Cramér's (1946) results. Hence, we can expect that when we evaluate sample size n_o, which assures the sufficient convergence of statistic $z_{x,S}$ to standard normal distribution, then the same sample size is also sufficient for the convergence of $z_{y,S}$ to standard normal distribution.

Let us assume that $x_i = y_i + d_i$, $i = 1, \ldots, N$. x_i can be treated as a measure of y_i contaminated by error d_i. The following notation will be useful:

$$\bar{x}=\frac{1}{N}\sum_{i\in U}x_i,\quad c_{x,r}=\frac{1}{N}\sum_{i\in U}(x_i-\bar{x})^r,\quad v_x=c_{x,2},\quad \tau_{x,r}=\frac{1}{Nv_x^{r/2}}\sum_{i\in U}\left|x_i-\bar{x}\right|^r,$$

$$\eta_{x,r}=\frac{c_{x,r}}{v_x^{r/2}},\quad \bar{x}_S=\frac{1}{n}\sum_{i\in S}x_i,\quad v_{x,S}=\frac{1}{n-1}\sum_{i\in S}(x_i-\bar{x})^2,\ r=2,3,\ldots$$

Definitions of the parameters $\bar{y},\eta_{y,r},\eta_{d,r},\tau_{y,r},\tau_{d,r},\lambda_{y,r},\lambda_{d,r},\bar{y}_S,v_y,v_d,v_{y,S},v_{d,S}$ are analogous with the above ones. Under the assumption that variables y and d are independent, the squared correlation coefficient between x and y is equal to:

$$\kappa=\frac{v_y}{v_y+v_d}=\frac{v_y}{v_x}=1-\frac{v_d}{v_x},\quad 0<\kappa\le 1.$$

The following standardised sample means will be considered:

$$z_{y,S}=\frac{\bar{y}_S-\bar{y}}{\sqrt{v_{y,S}}}\sqrt{n},\quad z_{x,S}=\frac{\bar{x}_S-\bar{x}}{\sqrt{v_{x,S}}}\sqrt{n}. \tag{1}$$

When S is the simple random sample drawn with replacement from U, the Berry--Esseen inequality, following M. Krzyśko (2000), becomes:

$$\sup_s\left|F_{y,s}(z)-\Phi(z)\right|\le\zeta\frac{\tau_{y,3}}{\sqrt{n}},$$

where $F_{x,s}(z)$ is the sample distribution of $z_{x,S}$, $\Phi(z)$ is the distribution of standard normal random variable, and $\pi^{-1/2}\le\zeta<0{,}8$. J. L. Wywiał (2016) showed that:

$$\sup_s\left|F_{y,s}(z)-\Phi(z)\right|\le 0.8\frac{\tau_{y,3}}{\sqrt{n}}\le 0.8\sqrt{\frac{\eta_{y,4}}{n}}.$$

Under the assumption that variables x and d are independent, after appropriate algebraic computations it can be shown that:

$$v_y^2\eta_{y,4}=c_{y,4}=c_{x,4}+6v_xv_d+c_{d,4}=c_{x,4}+6v_x^2v_d^2+v_x^2(1-\kappa)^2\eta_{d,4}=$$
$$=\frac{1}{\kappa^2}\left(\eta_{x,4}+6(1-\kappa)+\eta_{x,4}\frac{v_d^2}{v_x^2}\right).$$

Finally:

$$\eta_{y,4}=\kappa^{-2}\left(\eta_{x,4}+6(1+\kappa)+\eta_{d,4}(1-\kappa)^2\right)=f(\kappa),$$

$\eta_{y,4}$ is a strictly decreasing function of κ because we can show that $f'(\kappa)<0$ for $\kappa\in(0;1]$. Hence, inequality:

$$\sup_{s\in\mathbf{S}}\left|F(z_{y,s})-\Phi(z_{y,s})\right|\le 0.8\sqrt{\frac{f(\kappa)}{n}}.$$

Let us assess the necessary sample size when the approximate values of the parameters $\eta_{x,4},\eta_{d,4}$, and κ are known. $f(0)$ takes an infinitely large value and $f(1)=\eta_{x,4}$. When we assume that $0.8\sqrt{\frac{f(\kappa)}{n}}\le\Delta_0$, where Δ_0 is an admissible

difference between the sample distribution of the statistic and the standard normal distribution, the necessary sample size yields the following expression:

$$n \geq n_o = \left\lceil \frac{0.64\left(\eta_{x,4} + 6(1+\kappa) + \eta_{d,4}(1-\kappa)^2\right)}{\Delta_0^2 \kappa^2} \right\rceil .$$

Hence, necessary sample size n_o is a decreasing function of coefficient κ.

Example 1. For instance, when $\eta_{x,4} = 4, \eta_{d,4} = 4, \kappa = 0.99$ and $\Delta = 0.01$, then $n > n_o = 26\,676$. If $\eta_{x,4} = 4, \eta_{d,4} = 4, \kappa = 0.9$ and $\Delta = 0.01$ then $n > n_o = 42\,539$.

3. Simulation Evaluation of Sample Size

3.1. Chi-square Test of Goodness of Fit

Sample size n_o will be evaluated on the basis of the following simulation experiment. Under assumed sample size, the normality of $z_{x,S}$ is tested on the basis of its simulated values. In order to do that, a series of samples $(s_j, j = 1, \ldots, r)$ each of size n are drawn independently from population U according to assumed sampling design. Let $\mathbf{z}_{x,s}^{(n)} = \left(z_{x,s_1}^{(n)}, \ldots, z_{x,s_r}^{(n)}\right)$ be the sequence of the statistics evaluated based on the sequence of samples. Next the normal distribution of $z_{x,S}$ can be tested on the basis of data $\mathbf{z}_{x,s}^{(n)}$. When the hypothesis on normality is not rejected, we can expect that the distribution of statistic $z_{x,S}$ for the sample size $n_o = n$ is sufficiently close to standard normal distribution. If the hypothesis on normality is rejected, a new series of samples is drawn, but each of them is of size $n + d$, where $d \geq 1$. Using these samples, the sequence $\mathbf{z}_{x,s}^{(n+d)}$ is evaluated, allowing us to again test the normality of $z_{x,S}$ but for larger sample size $n + d$. The procedure is repeated until the hypothesis on the normality is not rejected. In order to verify the hypothesis that $z_{x,S}$ has standard normal distribution, several test statistics can be used, e.g. Kolmogorov or Shapiro-Wilk statistics. However, the powers of these tests cannot be easily controlled. That is why we use the chi-square test of goodness of fit.

Usually, $z_{x,S}$ is used to construct confidence intervals or test statistics on the expected value of variable x. In this case only the quantiles of high or small degrees of $z_{x,S}$ have to be close to the appropriate quantiles of standard normal distribution. J. L. Wywiał (2016) proposed the following procedure for evaluating the necessary sample size. When $Z \sim N(0;1)$, we expect that z, where $P(Z < z_\lambda) = \lambda$ is close to $z_{n,\lambda}$ where $P(z_{x;n} < z_{n,\lambda}) = \lambda$. Usually, λ = 0.01; 0.05; 0.1; 0.9; 0.95; 0.99. Let $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$, where $P(Z < z_{k,\lambda}) = \lambda_k$, $k = 1, \ldots, K$. More formally, the following hypothesis must be tested:

$$H_0: \boldsymbol{\lambda} = \boldsymbol{\lambda}_0, \quad H_1: \boldsymbol{\lambda} = \boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_0.$$

Let $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_{K+1}]$, where $\omega_k = \lambda_k - \lambda_{k-1}$, $k = 2, \ldots, K$, $\omega_{1k} = \lambda_1, \omega_{K+1} = 1 - \lambda_K$. The above hypotheses can be rewritten as follows:

$$H_0: \boldsymbol{\omega} = \boldsymbol{\omega}_0, \quad H_1: \boldsymbol{\omega} = \boldsymbol{\omega}_1 \neq \boldsymbol{\omega}_0.$$

We can verify this using chi-square test of goodness of fit under fixed significance level α and power α. Several variants of the test have been considered – e.g. by F. C. Drost *et al*. (1989) and T. J. Santer and D. E. Duffy (1989). The test statistic is as follows:

$$Q_{n,r} = r \sum_{k=1}^{K+1} \frac{(W_k - \omega_k)^2}{\omega_k}, \tag{2}$$

where:

$$W_k = \frac{1}{r} \sum_{j=1}^{r} I_k\left(z_{x,S_j}^{(n)}\right), \tag{3}$$

if $z_{\lambda_{k-1}} < z_{x,s_j}^{(n)} \leq z_{\lambda_k}$, then $I_k\left(z_{x,s_j}^{(n)}\right) = 1$ otherwise $I_k\left(z_{x,s_j}^{(n)}\right) = 0$, $k = 1, \ldots, K+1$, $z_{\lambda_0} = -\infty$, $z_{\lambda_{K+1}} = \infty$. Under a sufficiently large number r, the statistic $Q_{n,r}$ has chi-square distribution with K degrees of freedom (denoted by $\chi_K^2(\kappa)$) and the following non-centrality parameter:

$$r\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1) = \sum_{k=1}^{K+1} \frac{(\omega_{0,k} - \omega_{1,k})^2}{\omega_{0,k}}. \tag{4}$$

The quantity $\delta(.,.)$ can be treated as distance between distributions specified by the hypotheses H_0 and H_1. Particularly, we will consider the following vector of probabilities:

$$\boldsymbol{\omega}_0 = [0.01 \;\; 0.04 \;\; 0.05 \;\; 0.8 \;\; 0.05 \;\; 0.04 \;\; 0.01],$$

$$\boldsymbol{\omega}_1^{(1)} = [0.012 \;\; 0.048 \;\; 0.06 \;\; 0.76 \;\; 0.06 \;\; 0.048 \;\; 0.012],$$

$$\boldsymbol{\omega}_1^{(2)} = [0.011 \;\; 0.044 \;\; 0.055 \;\; 0.78 \;\; 0.055 \;\; 0.044 \;\; 0.011].$$

Let us note that we do not consider, e.g., the alternative: $\boldsymbol{\omega}_1^{(*1)} = [0.008 \;\; 0.036 \;\; 0.04 \;\; 0.832 \;\; 0.04 \;\; 0.036 \;\; 0.008]$ because the chi-square test does not select the difference between $\boldsymbol{\omega}_1^{(1)}$ alternatives and $\boldsymbol{\omega}_1^{(*1)}$. In this case the non-centrality coefficient takes the same value. Expression (4) allows us to calculate that $\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1^{(1)}) = 0.01$, $\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1^{(2)}) = 0.0025$.

When hypothesis H_0 is true, the test statistic $Q_{n,r}$ has the central chi-square distribution χ_K^2 with $K = 6$ degrees of freedom, provided that r is large. W. G. Cochran (1952) wrote that convergence to asymptotic distribution is sufficiently accurate when $r_0 = 5/\omega_0$ where $\omega_0 = \min_{k=1,\ldots,K+1}\{\omega_k\}$. Hence, in our

case, for $\omega_0 = 0.01$, $r_0 = 500$. The algorithm for evaluating r is as follows. Firstly, based on χ^2_K distribution, the critical value q_α of the test is determined under an assumed significance level α. Next, the power of the test is calculated for $r \geq r_0$ according to $\beta_r = P\left(\chi^2_K\left(r\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1)\right) \geq q_\alpha \,\middle|\, H_1\right)$ and $\alpha = P\left(\chi^2_K \geq q_\alpha \,\middle|\, H_0\right)$. The r_0 is treated as the start number of sample replication. If for fixed $r \geq r_0$, β_r is not less than the assumed level β, then size r is sufficiently large and it will be denoted by $r_\#$. Otherwise, the power is calculated for $r + 10$ and so on.

Example 2. Consider these hypotheses:

$$H_0: \boldsymbol{\omega} = \boldsymbol{\omega}_0, \quad H_1: \boldsymbol{\omega} = \boldsymbol{\omega}_1^{(2)}. \tag{5}$$

The significance level is $\alpha = 0.05$ and power $\beta = 0.95$, $\delta\left(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1^{(2)}\right) = 0.0025$. The above algorithm leads to the necessary number of the sample replication being $r_\# = 8350$. The next variants for calculating $r_\#$ are presented in the first three columns of Table 1 – also see (Wywiał 2016).

3.2. Evaluation of Sample Size in Order to Assure the Normal Distribution of Some Statistics

Let us consider the determination of sample size n of a simple random sample in order to assure convergence of the standardised sample mean distribution to standard normal distribution when the sample is drawn from non-normal distribution. Let us suppose that a sample of size n is drawn with replacement from a population of size N, where values of variable x are observed. Next, the statistic $z_{x,S}$ is evaluated in the case when simple random sample is drawn with replacement. For sampling without replacement, the test statistic is as follows:

$$z_{1x,S} = \frac{\bar{x}_S - \bar{x}}{\sqrt{(N-n)v_{x,S}}}\sqrt{Nn}\,.$$

Our purpose is to evaluate the sample sizes so that $z_{x,S}$ and $z_{1x,S}$ converge sufficiently well to standard normal distribution. In order to do this, the sample sizes are replicated r-times. Values of the statistics are calculated on the basis of the replicated samples. Next, the value of the chi-square test statistic is calculated by means of expression (2). Bear in mind that under assumed significance level α and the number of sample replications $r_\#$, the chi-square test has power β. If the test rejects hypothesis H_0, then an increase of d has to be added to sample size n and the described algorithm has to be repeated for $n + d$. When the test does not reject the hypothesis, we state that $n = n_{\alpha,\beta}$.

Table 1. The Necessary Sample Sizes for Testing Normal Distributions of Statistics $z_{x,S}$ and $z_{1x,S}$ under Assumed Significance Levels and Powers of the chi-square Test

α	β	$r_{\#}$	p	$\underline{n}$	$\underline{n}_1$
0.1	0.9	5870	1	920	930
			2	520	510
			4	300	320
0.05	0.95	8350	1	1060	1040
			2	590	560
			4	320	340
0.01	0.99	14 010	1	1440	1380
			2	780	790
			4	450	430
0.005	0.995	16 420	1	1530	1550
			2	860	780
			4	470	410

Source: the author's own calculations.

Example 3. Let us consider a population of $N = 100{,}000$ values generated according to gamma probability distribution with shape parameter p and a scale parameter of one. Using the algorithm, the necessary sample sizes are evaluated. A computer simulation implements the above algorithm under the hypothesis given by (5) and several combinations of the significance levels and powers. The obtained results lead to Table 1. The algorithm for evaluating necessary sample sizes is replicated 10-times, which lets us compute the mean sample sizes denoted by $\underline{n}$ and $\underline{n}_1$ in the case of statistics $z_{x,S}$ and $z_{1x,S}$, respectively. In Table 1 we take into account only such α and β that $\alpha + \beta = 1$. Note, however, that this assumption is unnecessary.

Table 1 shows that the assessed mean sample sizes for both sampling without replacement and sampling with replacement are comparable. In general, when the significance level decreases and the power increases, the necessary sample size increases.

3.3. Evaluation of Complex Random Sample Size in Order to Assure the Normality of Some Statistics

Using the results of the previous sub-section, we can evaluate the necessary sample size for two complex sampling designs. The first is the well-known D. B. Lahiri (1951), H. Midzuno (1952) and A. R. Sen (1953) sampling design, which is defined by the following probability function:

$$P_2(s)=\binom{N}{n}^{-1}\frac{\bar{x}_s}{\bar{x}},\quad s\in\mathbf{S},$$

where **S** is sample space. The sampling design is defined for a positive valued variable observed in the whole population. The inclusion probabilities of the first and second order are as follows:

$$\pi_k=\frac{N-n}{(N-1)N}\frac{x_k-\bar{x}}{\bar{x}}+\frac{n}{N},\quad \pi_{k,l}=\frac{(n-1)n}{(N-1)N}\frac{n-1}{N-2}\left(\pi_k+\pi_l-\frac{2n}{N}\right),\quad k\neq l=1,\ldots,N.$$

The sampling scheme implementing the sampling design is as follows. Let $p_k = x_k/x_U$, where $k = 1, \ldots, N$ and $x_U = N\bar{x}$. The first element is drawn from the population into the sample with probability p_k, $k = 1, \ldots, N$. The next $n - 1$ elements are drawn without replacement from the remaining $N - 1$ elements of the population as a simple sample of size $n - 1$.

The population mean $\bar{y}$ is estimated by means of the following Horvitz-Thompson (1952) estimator:

$$\bar{y}_{HT,S}=\frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}.$$

This is an unbiased estimator of $\bar{y}$, when $\pi_k > 0$ for all $k = 1, \ldots, N$. The unbiased estimator of variance proposed by A. R. Sen (1953), F. Yates and P. M. Grundy (1953) is as follows:

$$V_{1S}(\bar{y}_{HT,S})=-\frac{1}{2N}\sum_{k\in S}\sum_{l\in S,k\neq l}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)\frac{\Delta_{k,l}}{\pi_{k,l}},$$

where $\Delta_{k,l}=\pi_{k,l}-\pi_k\pi_l$ for $k \neq l$ and $\Delta_{k,k}=\pi_k(1-\pi_k)$. This estimator is useful only when $\pi_{k,l} > 0$ for all $k, l = 1, \ldots, N$ and $k \neq l$.

Next, the sampling design provides samples s of fixed size n drawn with replacement from population U with the above defined probabilities p_k, $k, \ldots, N$. This is a particular case of the multinomial sampling design (Tillé 2006, pp. 70–73). In this case the parameter $\bar{y}$ is estimated by means of the following Hansen-Hurvitz (1943) estimator:

$$\bar{y}_{HH,S}=\frac{1}{n}\sum_{k\in S}\frac{y_k}{p_k}.$$

The unbiased estimator of its variance is:

$$V_{2S}(\bar{y}_{HH,S})=\frac{1}{n(n-1)}\sum_{k\in S}\left(\frac{y_k}{p_k}-\bar{y}_{HH,S}\right)^2.$$

Let us consider the following statistic:

$$t_{HT,S}=\frac{\bar{y}_{HT,S}-\bar{y}}{\sqrt{V_{1S}(\bar{y}_{HT,S})}},\qquad t_{HH,S}=\frac{\bar{y}_{HH,S}-\bar{y}}{\sqrt{V_{2S}(\bar{y}_{HH,S})}}.\tag{6}$$

As we did in section 3.2, here we evaluate the necessary sample size in order to assure sufficient convergence of the distributions of statistics $t_{HT,S}$ and $t_{HH,S}$ to standard normal distribution.

Example 4. The simulation analysis is based on three appropriately generated sets of values $(y_k, x_{i,k})$, $k = 1, \ldots, N$ of a two-dimensional random variable denoted by (Y, X_i), where $X_i = Y + D_i$ and $i = 1, 2, 3$. Variable Y has gamma distribution with a shape parameter equal to 4 and a scale parameter of 1. Moreover, $D_1 \sim N(0; 0.425)$, $D_2 \sim N(0; 0.294)$ and $D_3 \sim N(0; 0.125)$. The correlation coefficients are: $\rho(Y, X_1) = 0.9$, $\rho(Y, X_2) = 0.95$ and $\rho(Y, X_3) = 0.99$. For instance, values of X_i can be treated as observations of Y contaminated by errors which are values of D_i, $i = 1, 2, 3$. According to the above sampling design, samples are replicated r-times. On the basis of such samples, the values of statistic $t_{HT,S}$ and $t_{HH,S}$ are calculated. As in Example 2, the hypothesis given by (5) is tested by means of the chi-square statistic. If the hypothesis is rejected, then sample size n is increased to $n + 10$ and the procedure is repeated. The algorithm is replicated until the hypothesis is not rejected.

Table 2. The Necessary Sample Sizes for Ensuring Normal Distributions of the Statistics $t_{HT,S}$ and $t_{HH,S}$ under Assumed Significance Levels and Powers of the Test

α	β	r	ρ	$\underline{n}$	
				$\bar{y}_{HT,S}$	$\bar{y}_{HH,S}$
0.1	0.9	5870	0.9	360	50
			0.95	440	50
			0.99	450	60
0.05	0.95	8350	0.9	530	40
			0.95	620	60
			0.99	760	60
0.01	0.99	14 010	0.9	570	60
			0.95	660	60
			0.99	1330	60

Source: the author's own calculations.

A sample size obtained in such a way is treated as sufficient for normal distribution of the statistic being considered under the assumed significance level as well as the power of the chi-square test statistic. The algorithm for evaluating the necessary sample sizes is replicated 10-times, allowing us to compute $\underline{n}$. The results of the simulation experiments can be found in Table 2.

Based on Table 2, we can say that the necessary sample size, in order to assure sufficient convergence of statistic $t_{HT,S}$ to normality under sampling design $P_2(s)$, is at least seven times larger than it is for $t_{HH,S}$ under a sample drawn with replacement with probabilities proportional to the auxiliary variable values. Under the considered variants of significance levels and powers of the test, the necessary sample size in order to ensure the normal distribution of $t_{HH,S}$ distribution oscillates around 60. For $t_{HT,S}$ distribution, the necessary sample size increases when the significance level decreases and the power increases.

4. Conclusions

Both of the methods considered for evaluating necessary sample sizes in order to ensure statistics are normally distributed were based on the assumption that an auxiliary variable is known from a whole population. The method requires the assumption that standard normal distribution is the asymptotic distribution of the statistics under analysis. The first method of determining the sample size is based on the Berry-Esseen inequality. The particular case of dependence between variables x and y considered in Example 2 lets us conclude that the necessary sample size decreases when the correlation coefficient between these variables increases.

In the case of the next method, the necessary sample size is evaluated by means of appropriate formulation of hypotheses on the normality coefficient. The tested and alternative hypotheses (see expression (5)) are constructed in such a way that the tails of the standard normal distribution are especially taken into account. The proposed simulation algorithm based on testing appropriate statistical hypothesis leads to the following conclusion. Under the evaluated sample size n_o, the hypothesis H_0 on normal distribution of considered statistic is not rejected. This decision is wrong (type II error) with a probability of $\nu = 1 - \beta$. Moreover, in the previous steps of the algorithm, when the sample size was shorter than the optimal level n_o, the alternative hypothesis H_1 (determining the non-admissible distribution) was accepted. This was the wrong decision (type I error) with probability α. Hence, in consequence, the probabilities α and ν measure the risk of assessing the necessary sample size on the level n_o. When under the fixed distance measure $\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1)$ (see expression (4)) between the distributions specified by the hypotheses H_0 and H_1, we decrease the level of α or the level of ν, then size n_o increases. This usually causes the costs of data observations to rise. Hence, compromise levels for δ, α, ν and n_o need to be found. This procedure can be applied to more complicated statistics or more complex sampling schemes than those considered in this paper. Moreover, it is possible to generalise the obtained results on distributions of bootstrap type statistics.

Bibliography

Berger Y. G. (1998), *Rate of Convergence to Normal Distribution for Horvitz-Thompson Estimator*, "Journal of Statistical Planning and Inference", vol. 67, https://doi.org/10.1016/s0378-3758(97)00107-9.

Cassel C. M., Särndal C. E., Wretman J. H. (1977), *Foundation of Inference in Survey Sampling*, John Wiley & Sons, New York–London–Sydney–Toronto.

Chernick M. R., Liu C. Y. (2002), *The Saw-toothed Behavior of the Power versus Sample and Software Solutions: Single Binomial Proportion Using Exact Methods*, "The American Statistician", vol. 56, https://doi.org/10.1198/000313002317572835.

Cochran W. G. (1952), *The chi-squared Test of Goodness of Fit,* "Annals of Mathematical Statistics", vol. 23, https://doi.org/10.1214/aoms/1177729380.

Cramér H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

Drost F. C., Kallenberg W. C. M., Moore D. S., Oosterhoff J. (1989), *Power Approximations to Multinomial Tests of Fit*, "Journal of the American Statistical Association", vol. 84, https://doi.org/10.2307/2289856.

Edgeworth F. Y. (1907), *On the Representation of a Statistical Frequency by a Series*, "Journal of the Royal Statistical Society", vol. A 70.

Fuller W. A. (2009), *Sampling Statistics*, John Wiley & Sons, Hoboken, New Jersey.

Greselin F., Zenga M. (2006), *Convergence of the Sample Mean Difference to the Normal Distribution: Simulation Results*, "Statistica & Applicazioni", vol. 4, no 1.

Hájek J. (1964), *Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population*, "Annals of Mathematical Statistics", vol. 35, https://doi.org/10.1214/aoms/1177700375.

Hájek J. (1981), *Sampling from a Finite Population*, ed. V. Dupač, Marcel Dekker, Inc., New York–Basel.

Hall P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Hansen M. H., Hurvitz W. N. (1943), *On the Theory of Sampling from Finite Population*, "Annals of Mathematical Statistics", vol. 14, https://doi.org/10.1214/aoms/1177731356.

Horvitz D. G., Thompson D. J. (1952), *A Generalization of Sampling without Replacement from a Finite Universe*, "Journal of the American Statistical Association", vol. 47, https://doi.org/10.1080/01621459.1952.10483446.

Krzyśko M. (2000), *Statystyka matematyczna*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań.

Lahiri D. B. (1951), *A Method of Sample Selection Providing Unbiased Ratio Estimator*, "Bulletin of the International Statistical Institute", vol. 33.

Midzuno H. (1952), *On the Sampling System with Probability Proportional to Sum of Size*, "Annals of the Institute of Statistical Mathematics", vol. 3, https://doi.org/10.1007/bf02949779.

Ryan T. P. (2013), *Sample Size Determination and Power*, John Wiley & Sons, Hoboken, New Jersey.

Santer T. J., Duffy D. E. (1989), *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York.

Seber G. A. F. (2013), *Statistical Models for Proportions and Probabilities*, Springer Briefs in Statistics, Heidelberg–New York–Dordrecht–London.

Sen A. R. (1953), *On the Estimate of the Variance in Sampling with Varying Probabilities*, "Journal of the Indian Society of Agricultural Statistics", vol. 5.
Tillé Y. (2006), *Sampling Algorithms*, Springer, New York.
Wywiał J. L. (2016), *Contributions to Testing Statistical Hypotheses in Auditing*, Wydawnictwo Naukowe PWN, Warszawa.
Yates F., Grundy P. M. (1953), *Selection without Replacement from Within Strata with Probability Proportional to Size*, "Journal of the Royal Statistical Society", Series B, vol. 15.

Symulacyjne wyznaczanie niezbędnego rozmiaru próby zapewniającego wystarczającą zbieżność rozkładu pewnych statystyk do rozkładu normalnego

(Streszczenie)

Podczas testowania hipotez lub wyznaczania przedziałów ufności rozkłady pewnych statystyk zwykle nie są znane. Wygodne jest, gdy rozkłady takich statystyk można przybliżać rozkładem normalnym. Celem pracy jest wyznaczenie takiej liczebności próby, przy której rozkład statystyki jest dostatecznie dobrze aproksymowany rozkładem normalnym. Zaproponowano dwie procedury postępowania. Jedna z nich daje aproksymację liczebności próby na podstawie nierówności Berry-Esseena. Druga metoda polega na generowaniu serii prób o ustalonej liczebności, na podstawie których wyznacza się wartości statystyki. Opierając się na tych wartościach, testuje się normalność rozkładu statystyki. W razie odrzucenia hipotezy o normalności zwiększa się rozmiar generowanych prób. Procedurę tę powtarza się aż do ustalenia liczebności próby, przy której hipoteza o normalności nie jest odrzucona.

Słowa kluczowe: rozmiar próby, twierdzenia centralne, schemat losowania, symulacja komputerowa, test chi-kwadrat zgodności.