A C T A  U N I V E R S I T A T I S  L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

*Dorota Pekasiewicz*[*]

# SEQUENTIAL METHOD FOR ESTIMATING THE SAMPLE SIZE REQUIRED FOR TESTING HYPOTHESES ON THE POPULATION MEAN

**Abstract.** In the case of statistical inference on parameters from population of unknown distribution, we use limit theorems. The minimum sample size which allows to make statistical inference, depends on the type of the population distribution.

In the paper we propose an application of sequential method in estimating sample size needed for the verification of hypotheses on expected value of population. Apart from theoretical considerations, the results of simulations for different populations with different values of asymmetry coefficient and dispersion, are presented.

**Key works:** statistical test, sequential method, sample size

## I. INTRODUCTION

The application of limit theorems, in the verification of statistical hypothesis, is connected with using random samples of large sizes. It is impossible to determine exactly the minimum sample size needed, in the verification procedure, with the test's statistic following asymptotic normal distribution. It depends on population distribution, thus on the rate of the asymptotic convergence of estimator's distribution to the normal distribution.

Statistical significance test for expected value and the application of this test for populations with distributions different from the normal distribution is presented. A sequential method of determining sample size for which this test can be applied is proposed. This sample size is determined, so that, the sample arithmetic mean would be actually asymptotically normally distributed.

When applying the sequential method of determining the sample size, a special criterion should be defined. Meeting the criterion for the current sample size allows to finish the increasing the sample size and to apply the testing procedure. This procedure causes that the test statistic, as well as the sample size, are random variables.

[*] Ph.D., Chair of Statistical Methods, University of Łódź.

## II. TEST FOR EXPECTED VALUE OF POPULATION WITH SEQUENTIAL PROCEDURE OF ESTIMATING SAMPLE SIZE

Let us assume that $X$ is a random variable with unknown continuous distribution. Let $\mu$ be the expected value of $X$.

We consider the null hypothesis about the value of parameter $\mu$:

$$H_0: \quad \mu = \mu_0$$

against the alternative:

$$H_1: \quad \mu \neq \mu_0,$$

where $\mu_0$ is a fixed constant.

The above hypotheses are verified by the significance test whose statistic is the following:

$$U = \frac{\bar{x} - \mu_0}{s}\sqrt{n}, \tag{1}$$

where $\bar{x}$ is arithmetic mean and $s$ is standard deviation calculated on the basis of the sequence of values $x_1,...,x_n$ of the simple sample $X_1,...,X_n$.

First, we assume that we have $n$ observations $x_1,...,x_n$. For fixed $\varepsilon$ we calculate:

$$\left|\overline{X}_{n-1} - \overline{X}_n\right| \leq \varepsilon, \tag{2}$$

where $\overline{X}_{n-1}, \overline{X}_n$ are arithmetic means, respectively for $n$-1 and $n$ elements (cf. Lalu, Krishnan, 1978).

The connection between sample size, variance of population $\sigma^2$ and fixed $\varepsilon$ results from the Czebyszew inequality:

$$P\left(\left|Y - EY\right| > k\sigma_Y\right) < \frac{1}{k^2}. \tag{3}$$

For random variable $Y = \overline{X}_{n-1} - \overline{X}_n$ we have:

$$EY = EY\left|n = 0, \right. \tag{4}$$

$$D^2Y|n = D^2\left(\overline{X}_{n-1} - \overline{X}_n\right) = D^2\left(\frac{1}{n-1}\sum_{i=1}^{n-1} X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right) =$$

$$= D^2\left(\sum_{i=1}^{n-1}\left(\frac{1}{n-1} - \frac{1}{n}\right)X_i - \frac{1}{n}X_n\right) =$$

$$= D^2\left(\frac{1}{n}\left(\overline{X}_{n-1} - X_n\right)\right) = \frac{1}{n^2}\left(D^2\left(\overline{X}_{n-1}\right) + D^2\left(X_n\right)\right) = \tag{5}$$

$$= \frac{1}{n^2}\left[\frac{1}{(n-1)^2}\sum_{i=1}^{n-1} D^2\left(X_i\right) + D^2\left(X_n\right)\right] =$$

$$= \frac{1}{n^2}\cdot\left[\frac{1}{(n-1)^2}(n-1)\sigma^2 + \sigma^2\right] = \frac{1}{n(n-1)}\sigma^2.$$

For $k = \frac{\varepsilon}{\sigma_Y}$ we obtain:

$$P\left(\left|\overline{X}_{n-1} - \overline{X}_n\right| > \varepsilon\right) < \frac{D^2Y|n}{\varepsilon^2},$$

$$P\left(\left|\overline{X}_{n-1} - \overline{X}_n\right| \le \varepsilon\right) \ge 1 - \frac{\sigma^2}{\varepsilon^2 n(n-1)}. \tag{6}$$

If the inequality (2) is true, for the sample $x_1,...,x_n$, then we regard this sample as sufficient and we verify hypothesis for expected value. In other case, we enlarge the sample and repeat this procedure until condition (2) is met. In this way, we obtain the sample which allows us to apply the test considered.

## III SIMULATION ANALYSIS OF PROPERTIES OF THE TEST FOR EXPECTED VALUE WITH SEQUENTIAL METHOD OF ESTIMATING SAMPLE SIZE

The analyse the significance test for expected value we generate populations using simulation method. Distributions of these populations are different, but all of them are characterized by expected value equal to 4.

Next, we define the following hypothesis $H_0 : \mu = 4,$ against $H_1 : \mu \neq 4$ and verify these hypotheses 10 000 times for every population for a fixed level of significance $\alpha = 0.05$.

We consider the following distributions:

- exponential distribution: $f(x)=\begin{cases}\frac{1}{4}\exp\left(-\frac{x}{4}\right) & \text{for} \quad x>0 \\ 0 & \text{for} \quad x\le 0\end{cases}$,
- $\chi^2$ distribution with $k = 4$ degrees of freedom:

$$f(x)=\begin{cases}\frac{1}{4}x\exp\left(-\frac{x}{2}\right) & \text{for} \quad x>0 \\ 0 & \text{for} \quad x\le 0\end{cases},$$

- gamma distribution: $f(x)=\begin{cases}\frac{1}{\lambda^p\Gamma(p)}x^{p-1}\exp\left(-\frac{x}{\lambda}\right) & \text{for} \quad x>0 \\ 0 & \text{for} \quad x\le 0\end{cases}$,

for $\lambda=1,\ p=4$ and $\lambda=10,\ p=0{,}4$,

- Pareto distribution: $f(x)=\begin{cases}\frac{\alpha\beta^{\alpha}}{(\beta+x)^{\alpha+1}} & \text{for} \quad x\ge 0 \\ 0 & \text{for} \quad x<0\end{cases}$,

for $\alpha=2,\ \beta=2$ and $\alpha=4,\ \beta=3$,

- normal left-truncated distribution (cf. Pekasiewicz, (2007)):

$$f(x)=\begin{cases}\frac{\exp\left(-(x-\mu)^2/2\sigma^2\right)}{\sqrt{2\pi}\sigma(1-\Phi((c-\mu)/\sigma))} & \text{for} \quad x\ge c \\ 0 & \text{for} \quad x<c\end{cases},$$

where $\Phi$ is cumulative distribution function of normal standardized distribution, for $\mu=3;\ \sigma=2;\ c=2$ and $\mu=3{,}5;\ \sigma=2{,}5;\ c=2{,}2$,

- uniform distribution on the interval [0,8]:

$$f(x)=\begin{cases}\frac{1}{8} & \text{for} \quad 0\le x\le 8 \\ 0 & \text{for} \quad x<0 \quad \text{or} \quad x>8\end{cases}.$$

Table 1 presents numbers of wrong rejections of the true hypothesis $H_0$, when we apply tests with large sample sizes. The sample whose sizes are greater than 30 elements are regarded as large sample (cf. Hellwig, 1998). Then in simulation study the following sizes are used: 30, 50, 70. The values of wrong

rejections can be used to estimate the probability of type I error and compared with a fixed level of significance $\alpha = 0{,}05$. In these cases the number of wrong decisions should not be greater than 500.

Table 1. The number of rejections of the true hypothesis $H_0 : \mu = 4$ for different populations for significance test and $\alpha = 0.05$

| Nº | Type of distribution | Sample size | | |
|---|---|---|---|---|
| | | 30 | 50 | 70 |
| 1 | Exponential distribution with parameter $\lambda = 4$ | 1013 | 804 | 729 |
| 2 | $\chi^2$ Distribution with $k$ =4 degrees of freedom | 898 | 732 | 682 |
| 3 | Gamma distribution with parameters $\lambda = 1$ and $p = 4$ | 791 | 680 | 646 |
| 4 | Gamma distribution with parameters $\lambda = 10$ and $p = 0{,}4$ | 1289 | 1007 | 836 |
| 5 | Pareto distribution with parameters $\alpha = 2$ and $\beta = 2$ | 2595 | 2235 | 2029 |
| 6 | Pareto distribution with parameters $\alpha = 4$ and $\beta = 3$ | 1517 | 1335 | 1205 |
| 7 | Normal distribution N(3; 2) left-truncated by $c = 2$ | 698 | 605 | 558 |
| 8 | Normal distribution N(3,5; 1,5) left-truncated by $c$ =2,2 | 671 | 613 | 575 |
| 9 | Uniform distribution on interval [0,8] | 601 | 558 | 541 |

Source: Own calculations.

For the same class of populations, the number of wrong decisions depends on the value of the standard deviation of distribution and its asymmetry. For example, for the Pareto distribution with parameters $\alpha = 2$, $\beta = 2$ standard deviation is equal to 6,41. For the Pareto distribution with parameters $\beta = 3$, $\alpha = 4$ standard deviation is equal to 1,409. The number of wrong decisions is greater in the case of the Pareto distribution with greater measure of dispersion. We obtain similar results for another class of population distribution. The fixed sample sizes are too small to verify hypothesis about expected value of population for a fixed level of significance. It is necessary to fix the sample size for every population considering its distribution parameters. In these cases we can apply the sequential method of estimating sample size.

We start from $n = 30$ elements and apply the sequential method of estimating sample size for a fixed $\varepsilon$. In simulation analysis, the values of $\varepsilon$ are fixed in dependence on the value of $\mu_0$ in the null hypothesis. In the results presented the $\varepsilon$ values are the following: 1%, 0,5%, 0,25% and 0,125% of the $\mu_0$. The numbers of wrong rejections of the null hypothesis for fixed values of $\varepsilon$, are presented in Table 2.

Table 2. The number of rejections of the true hypothesis $H_0 : \mu = 4$ for different populations for significance test with sequential method of estimating sample size and $\alpha = 0.05$

| Nº | Type of distribution | Value of $\varepsilon$ | | | |
|---|---|---|---|---|---|
| | | 0,04 | 0,02 | 0,01 | 0,005 |
| 1 | Exponential distribution with parameter $\lambda = 4$ | 807 | 713 | 674 | 617 |
| 2 | $\chi^2$ Distribution with $k$ =4 degrees of freedom | 760 | 637 | 575 | 543 |
| 3 | Gamma distribution with parameters $\lambda = 1$ and $p = 4$ | 682 | 636 | 584 | 592 |
| 4 | Gamma distribution with parameters $\lambda = 10$ and $p = 0{,}4$ | 1112 | 882 | 742 | 571 |
| 5 | Pareto distribution with parameters $\alpha = 2$ and $\beta = 2$ | 2212 | 2167 | 2175 | 592 |
| 6 | Pareto distribution with parameters $\alpha = 4$ and $\beta = 3$ | 1217 | 1219 | 1114 | 538 |
| 7 | Normal distribution N(3; 2) Left-truncated by $c = 2$ | 620 | 597 | 558 | 1613 |
| 8 | Normal distribution N(3,5; 1,5) Left-truncated by $c$ =2,2 | 634 | 611 | 535 | 924 |
| 9 | Uniform distribution on interval [0,8] | 595 | 553 | 521 | 509 |

Source: Own calculations.

Application of significance test with sequential method of increasing the sample size causes that the number of wrong decisions becomes smaller. Both, the number of wrong rejections of the true null hypothesis and sample size depend on the choice of the value of $\varepsilon$.

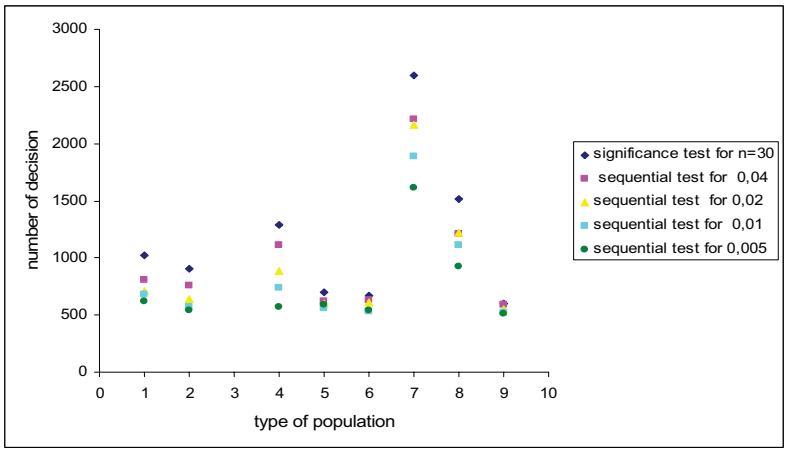


Figure 1. The number of wrong rejections of hypothesis $H_0$ for analyzed populations
Source: Own calculations

In the significance test with sequential method of estimating sample size, the sample size is a random variable. Chosen parameters of its distribution: the means, standard deviations and medians are presented in Table 3 for the considered cases. In Table 4 we present the empirical distribution of sample size

for population distribution $\chi^2$ with the number of the degrees of freedom equal to 4 and for $\varepsilon = 0{,}01$. Figure 2 presents the histogram of this distribution. The distribution of sample size is highly skewed.

Table 3. Parameters of distribution of sample size for test for expected value of populations for chosen values of $\varepsilon$

| Type of distribution | $\varepsilon$ | Parameters of distribution of sample size | | |
|---|---|---|---|---|
| | | $\overline{n}$ | $s$ | $Me(n)$ |
| Exponential distribution with parameter $\lambda = 4$ | 0,04 | 75,78 | 62,13 | 62,5 |
| | 0,02 | 149,73 | 136,46 | 127 |
| | 0,01 | 281,62 | 262,05 | 244 |
| | 0,005 | 581,68 | 553,78 | 506 |
| $\chi^2$ Distribution with $k$ =4 degrees of freedom | 0,04 | 59,54 | 41,33 | 46 |
| | 0,02 | 111,36 | 90,58 | 94 |
| | 0,01 | 214,31 | 180,27 | 183 |
| | 0,005 | 431,13 | 368,74 | 375 |
| Gamma distribution with parameters $\lambda = 1$ and $p$ = 4 | 0,04 | 46,58 | 26,37 | 34 |
| | 0,02 | 81,28 | 58,99 | 66 |
| | 0,01 | 158,64 | 122,55 | 137,5 |
| | 0,005 | 318,43 | 259,81 | 273 |
| Gamma distribution with parameters $\lambda = 10$ and $p$ = 0,4 | 0,04 | 105,52 | 119,54 | 81 |
| | 0,02 | 207,86 | 233,61 | 169 |
| | 0,01 | 416,10 | 486,02 | 341 |
| | 0,005 | 829,72 | 926,97 | 699 |
| Pareto distribution with parameters $\alpha = 2$ and $\beta = 2$ | 0,04 | 54,4 | 131,05 | 33 |
| | 0,02 | 96,60 | 201,93 | 63 |
| | 0,01 | 193,06 | 528,17 | 133 |
| | 0,005 | 378,47 | 838,46 | 277 |
| Pareto distribution with parameters $\alpha = 4$ and $\beta = 3$ | 0,04 | 34,95 | 24,37 | 30 |
| | 0,02 | 48,10 | 48,19 | 32 |
| | 0,01 | 85,82 | 106,89 | 65 |
| | 0,005 | 169,56 | 227,74 | 133 |
| Normal distribution N(3; 2) Left-truncated by $c$ = 2 | 0,04 | 37,41 | 14,23 | 30 |
| | 0,02 | 61,06 | 36,55 | 51 |
| | 0,01 | 114,21 | 78,32 | 102 |
| | 0,005 | 225,84 | 166,05 | 205 |
| Normal distribution N(3,5; 1,5) Left-truncated by $c$ =2,2 | 0,04 | 34,40 | 9,75 | 30 |
| | 0,02 | 52,69 | 28,44 | 42 |
| | 0,01 | 96,46 | 63,95 | 84 |
| | 0,005 | 186,31 | 133,15 | 165 |
| Uniform distribution on interval [0,8] | 0,04 | 55,17 | 23,94 | 51 |
| | 0,02 | 102,12 | 55,02 | 100 |
| | 0,01 | 198,99 | 115,64 | 198 |
| | 0,005 | 394,03 | 233,83 | 392 |

Source: Own calculations

Table 4. Distribution of sample size for distribution population $\chi^2$ with 4 degrees of freedom for $\varepsilon = 0.01$

| Ends of intervals of sample sizes | | Number of cases |
|---|---|---|
| 30 | 130 | 3508 |
| 130 | 230 | 2795 |
| 230 | 330 | 2015 |
| 330 | 430 | 892 |
| 430 | 530 | 246 |
| 530 | 630 | 203 |
| 630 | 730 | 112 |
| 730 | 830 | 83 |
| 830 | 930 | 47 |
| 930 | 1030 | 31 |
| 1030 | 1130 | 29 |
| 1130 | 1230 | 15 |
| 1230 | 1330 | 9 |
| 1330 | 1430 | 8 |
| 1430 | 1530 | 3 |
| 1530 | 1630 | 4 |

Source: Own calculations.

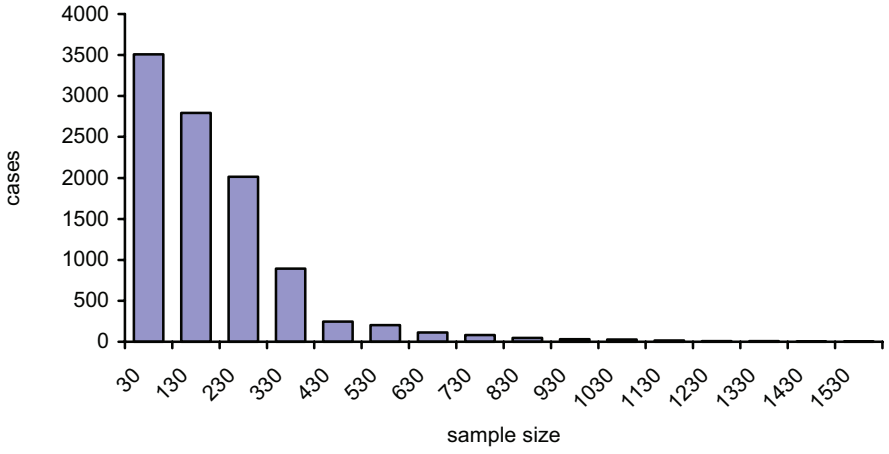


Figure 2. Histogram of the distribution of sample size for the $\chi^2$ population with 4 degrees of freedom, $\varepsilon = 0.01$

Source: Own calculations.

We can compare the considered tests using their power functions. Figure 3 presents the power functions of two significance tests with fixed sample sizes $n$=30 and $n$=50 and the power of tests with sequential estimation at the sample size for the chosen distribution and $\varepsilon = 0.01$.

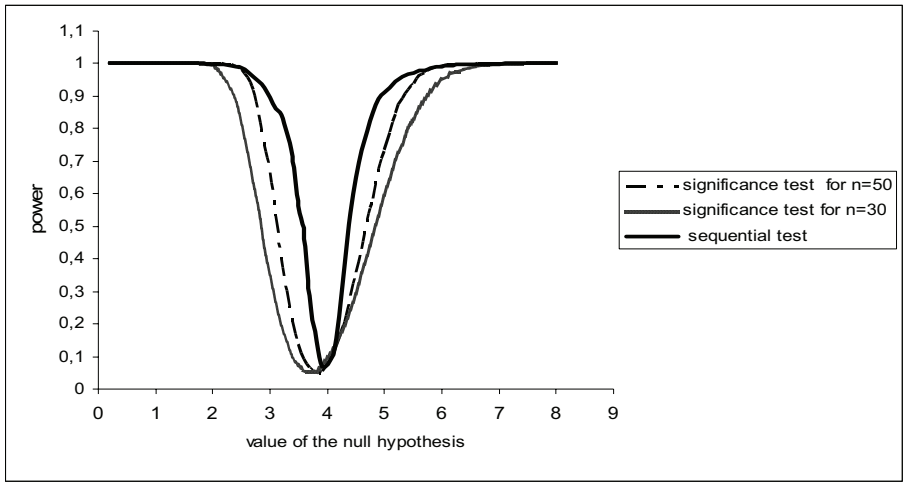


Figure 3. Power functions of tests for $\chi^2$ distribution with 4 degrees of freedom

Source: Own calculations

## IV. FINAL REMARKS

Sequential method of estimating sample size causes that the sample size is random variable. The sample size depends on the population distribution and the values drawn. The application of this method for the verification of hypotheses on the expected value of population allows to make a decision for a fixed level of significance. Moreover, the test with the sequential method of estimating sample size has power greater than the significance tests with fixed sample sizes $n$=30, 50 and 70. It is only important to fix properly the value of $\varepsilon$ in the sequential critertion.

The presented significance test for the expected value of population constitutes an example. The sequential method for estimating sample size can be applied in other methods of statistical inference making use of asymptotic properties of the sample arithmetic mean.

## REFERENCES

Hellwig Z. (1998), *Elementy rachunku prawdopodobieństwa i statystyki matematycznej* Wydawnictwo Naukowe PWN, Warszawa.

Lalu N. M., Krishnan P. ( ), *A Sequential Procedure for Estimating the Sample Size Needed for Normal Approximation in Finite Population Sampling*, Proceedings of the Survey Research Methods Section, 621–624, American Statistical Association.

Pekasiewicz D. (2007), *Testy sekwencyjne dla parametrów rozkładu normalnego lewostronnie uciętego*, [w:] Statystyka w praktyce społeczno-gospodarczej red. W Ostasiewicz, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, s. 230–245

*Dorota Pekasiewicz*

## SEKWENCYJNA METODA OKREŚLANIA LICZEBNOŚCI PRÓBY NIEZBĘDNEJ DO WERYFIKACJI HIPOTEZ O WARTOŚCI OCZEKIWANEJ POPULACJI

W przypadku wnioskowania statystycznego o parametrach populacji o nieznanym rozkładzie korzystamy z twierdzeń granicznych. Liczebność próby, pozwalająca przeprowadzić wnioskowanie statystyczne zależy od typu rozkładu populacji.

W pracy proponowane jest zastosowanie sekwencyjnej metody wyznaczania liczby elementów próby niezbędnej do weryfikacji hipotez o wartości oczekiwanej populacji. Oprócz rozważań teoretycznych, przedstawione są wyniki badań symulacyjnych dla populacji charakteryzujących się między innymi różną wielkością współczynnika asymetrii i różnym zróżnicowaniem.