Dependence of the sample estimates on the sample size

Grigorii Kozlov[1], *Mikhail* Pushkarev[1,*], and *Daniil* Belyaev[1]

[1]Saint-Petersburg State Institute of Technology, 190013, 26 Moskovsky prospect, St. Petersburg, Russia

Abstract. The paper provides data on the dependence of the sample indicators of the arithmetic mean, variance, asymmetry and excess of the length of the needles of European spruce (*Picea rubra*) and European larch (*Larix decidua*), the average length of a pair of needles of Scots pine (*Pinus sylvestris*) and Calabrian pine (*Pinus brutia*). The sizes of samples have been determined, which make it possible to obtain the values of the estimates of sample indicators that have stabilized around their general values. The data on the difference between the law of distribution of the length of the needles of coniferous plants from the normal one are confirmed. The possibility of using graphs of the dependence of sample indicators on sample sizes for the examination of scientific data is discussed.

1 Introduction

In everyday practice, most often researchers have to calculate the following sample indicators: arithmetic mean, variance, asymmetry and excess. Despite the routine and simplicity, the assessment of sample indicators is fraught with a number of pitfalls, some of which can be eliminated by the correct organization of data acquisition (for example, the "observer effect" [1]), and some require processing of already obtained data. An important point is the dependence of the sample value on the sample size [2]. Methods are well known that make it possible to obtain a reliable estimate of the interval [3] containing the desired value for small samples. However, in the case when obtaining samples of a large volume is not associated with technical difficulties or ethical obstacles (the use of animals [4]), primarily in ecological and botanical studies [5] or the processing of medical statistics [6], the accuracy of the results can be increased by studying the dependence of sample indicator on the sample size. In this work, an attempt is made to experimentally estimate the required sample size to assess the arithmetic mean, variance, asymmetry and excess coefficients for conifers.

* Corresponding author: malexpush@bk.ru

2 Materials and methods

Herbarium was collected in ecologically clean locations of Sergievka Park (St. Petersburg, Russia) in 2019 and Turkey resort (36 40' 58" N, 30 34' 11" E) in 2020. The length of the needles was measured with a ruler with an accuracy of 1 mm. The length of the needles was measured for European spruce (*Picea rubra*) and European larch (*Larix decidua*). The fallen needles were collected. For Scots pine (*Pinus sylvestris*) and Calabrian pine (*Pinus brutia*), the length of a pair of needles was measured and the average value was found. Herbarium collection and measurements were carried out by different researchers to ensure the "blind test" rule. Calculations of the arithmetic mean, variance, asymmetry and excess coefficients were carried out using Excel according to the following well-known formulas (1-4):

$$\overline{x} = \frac{\sum x_i}{n} \tag{1}$$

$$S_x{}^2 = \frac{\sum (x_i - \overline{x})^2}{(n-1)} \tag{2}$$

$$As = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{S_x}\right)^3 \tag{3}$$

$$Ec = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \overline{x}}{S_x}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{4}$$

where n-sample size; $S_x{}^2$ −variance;
S_x −-mean square deviation;
$\overline{x}$ −-arithmetic mean; x_i −variant.

Each sample estimate was calculated for a different sample size from 10 variants to sample size with a step of 1. Next, the graphs of the dependence of sample size on sample characteristic value were plotted, and the sample size sufficient for the calculated indicators to stabilize near their general value was estimated.

3 Results and discussion

The results are shown in Figures 1-16. All the graphs show a significant fluctuation in sample indicators with sample sizes up to 500. Thus, the assessment of the mean value, variance and asymmetry of the needles of European spruce stabilizes near its general value starting from the sample size of 1800-2000 variants, and the excess value - from 1200 variants. It should be noted that the asymmetry values exceed the critical value for p = 0.01 and a sample size of 2000, which allows rejecting the hypothesis of a normal distribution. The excess does not exceed the critical values. However, it stabilized at about 0.5 with a sample size of 1200 and does not tend to zero with a further increase in its size.

The mean value and asymmetry of the distribution of the length of the European larch needles stabilize around the general values with a sample size of 1000 variants, and the variance and excess - with 800 variants. The magnitude of the asymmetry exceeds the critical value for p = 0.01 and the sample size 1400, the excess does not exceed the critical value [7], but is stabilized around a nonzero value and does not tend to zero as the sample size increases.

The stabilization of the values of the sample estimates of the mean value, variance and asymmetry of the average value of a pair of Scots pine needles occurs when the sample size

is about 1400 variants, the excess value stabilizes at 800 variants. Neither excess nor asymmetry exceed critical values at a significance level of p = 0.01, but their values are stabilized around nonzero values.

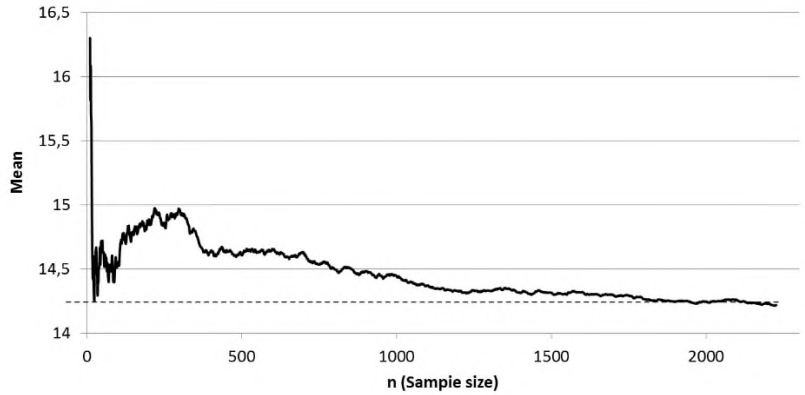

Fig. 1. Dependence of the arithmetic mean length of the needles of European spruce (*Picea rubra*) on the sample size (Sergievka park location).

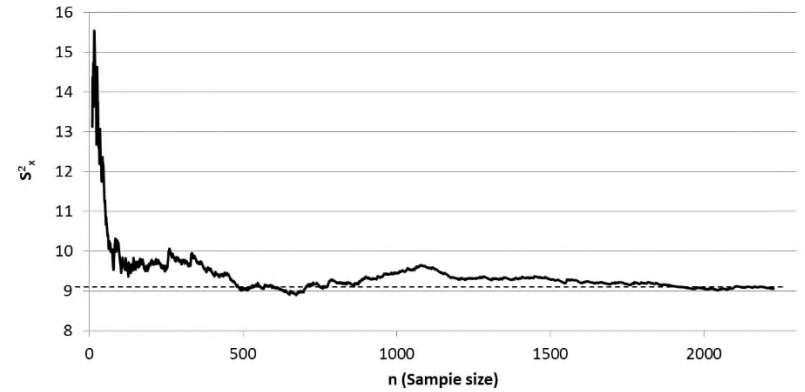

Fig. 2. Dependence of the value of the variance of the length of the needles of European spruce (*Picea rubra*) on the sample size (Sergievka park location).

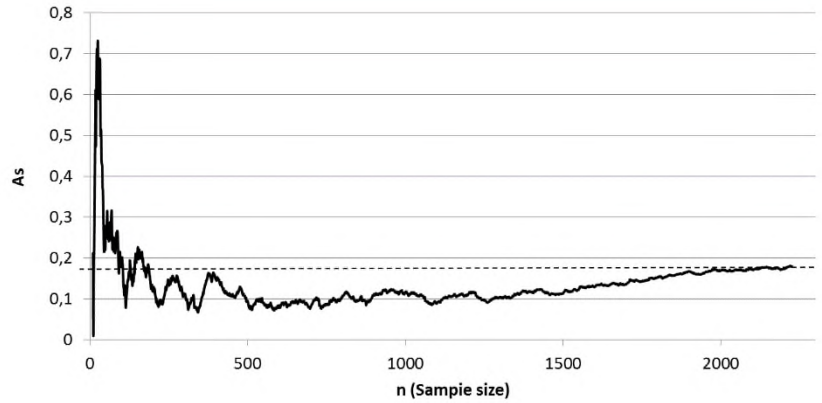

Fig. 3. Dependence of the value of the asymmetry of the distribution of the length of the needles of European spruce (*Picea rubra*) on the sample size (Sergievka park location).

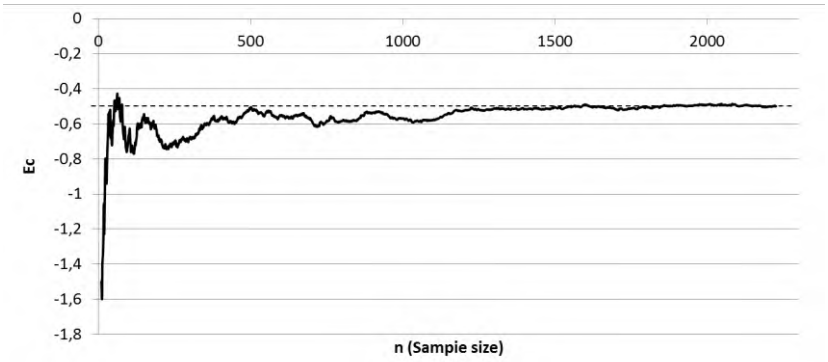

Fig. 4. Dependence of the excess value in the distribution of the length of the needles of European spruce (*Picea rubra*) on the sample size (Sergievka park location).

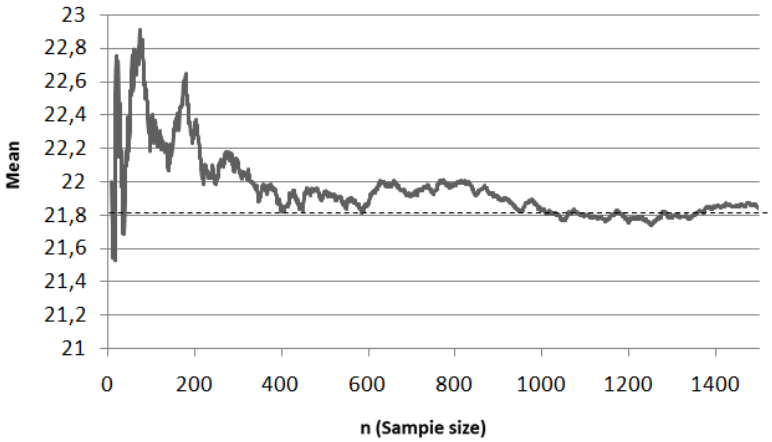

Fig. 5. Dependence of the arithmetic mean length of the needles of European larch (*Larix decidua*) on the sample size (Sergievka park location).

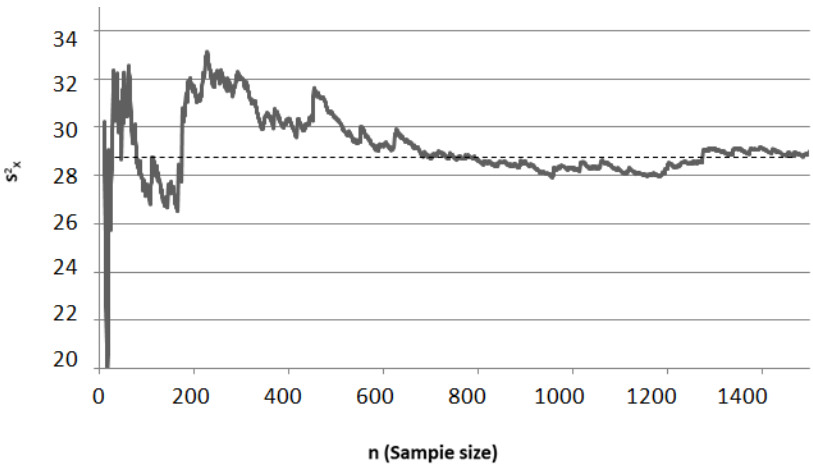

Fig. 6. Dependence of the variance of the length of the needles of European larch (*Larix decidua*) on the sample size (Sergievka park location).

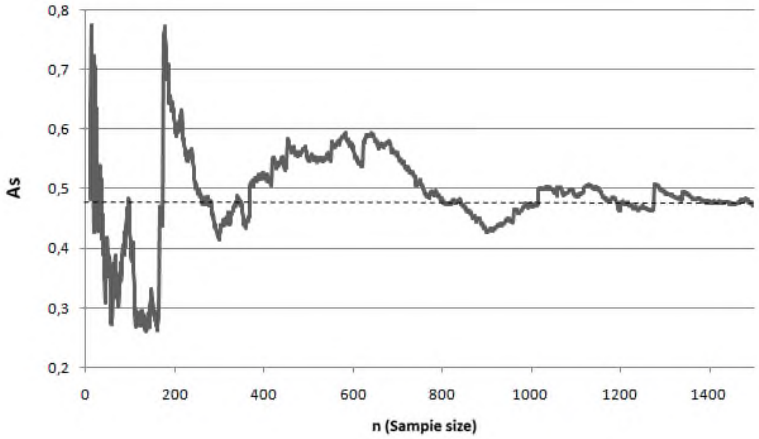

Fig. 7. Dependence of the value of the asymmetry of the distribution of the length of the needles of European larch (*Larix decidua*) on the sample size (Sergievka park location).

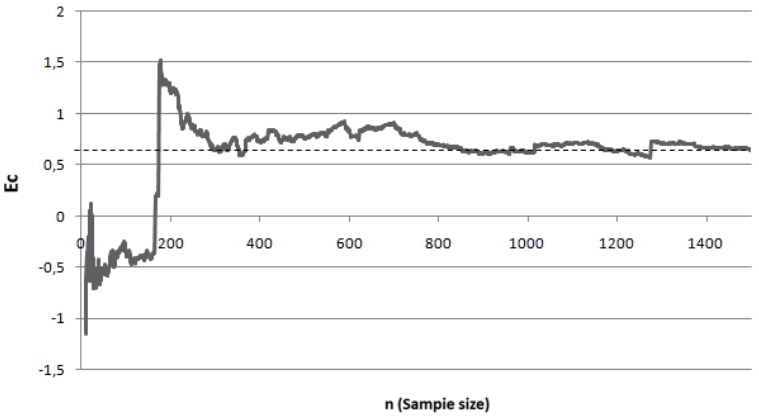

Fig. 8. Dependence of the excess value in the distribution of the length of the needles of European larch (*Larix decidua*) on the sample size (location Sergievka park).

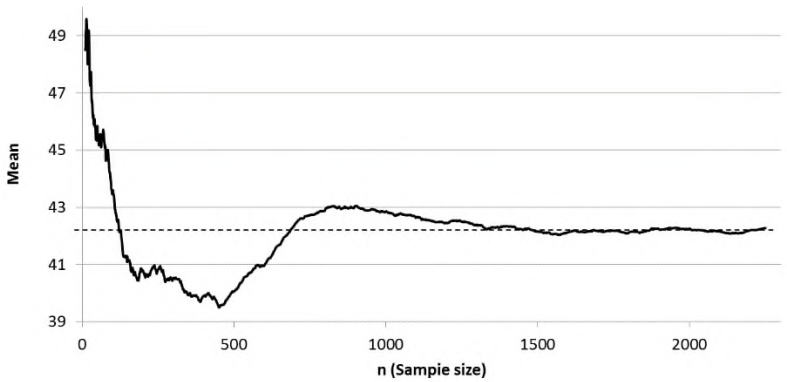

Fig. 9. Dependence of the arithmetic mean length of a pair of needles of Scots pine (*Pinus sylvestris*) on the sample size (Sergievka park location).

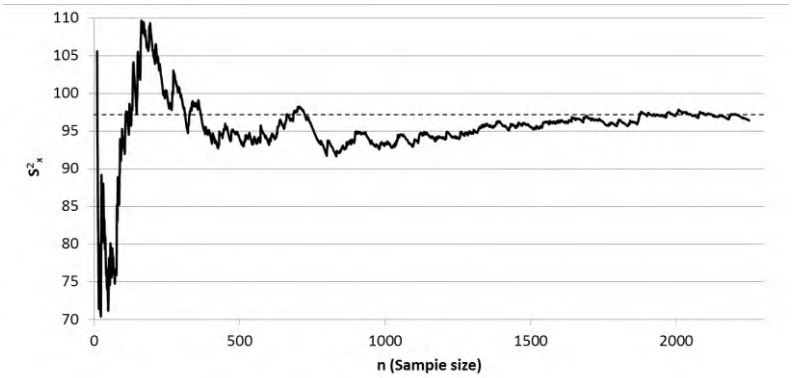

Fig. 10. Dependence of the value of the variance of the average length of a pair of needles of Scots pine (*Pinus sylvestris*) on the sample size (Sergievka park location).

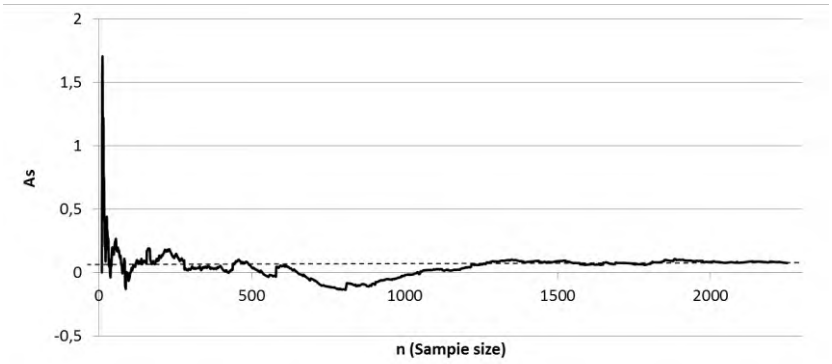

Fig. 11. Dependence of the value of the asymmetry of the distribution of the average length of a pair of needles of Scots pine (*Pinus sylvestris*) on the sample size (Sergievka park location).

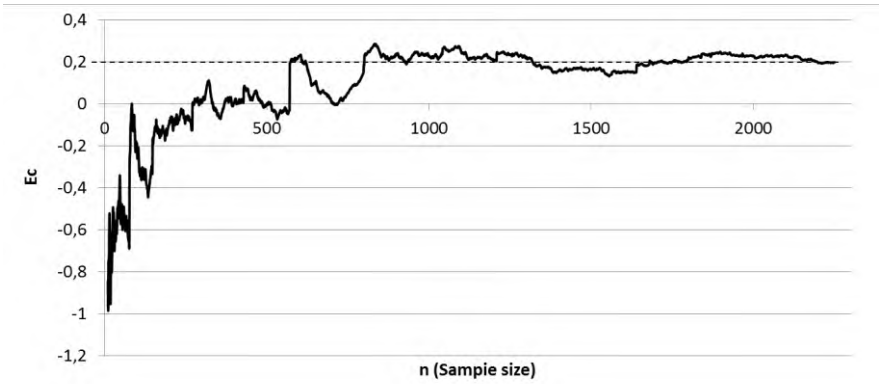

Fig. 12. Dependence of the excess value in the distribution of the average length of a pair of needles of Scots pine (*Pinus sylvestris*) on the sample size (Sergievka park location).

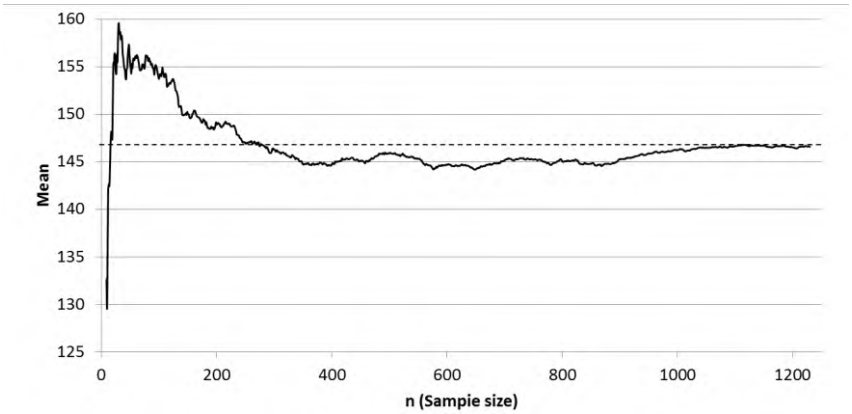

Fig. 13. Dependence of the value of the arithmetic mean length of a pair of needles of Calabrian pine (*Pinus brutia*) on the sample size (Turkey resort, 36 40' 58" N, 30 34' 11" E).

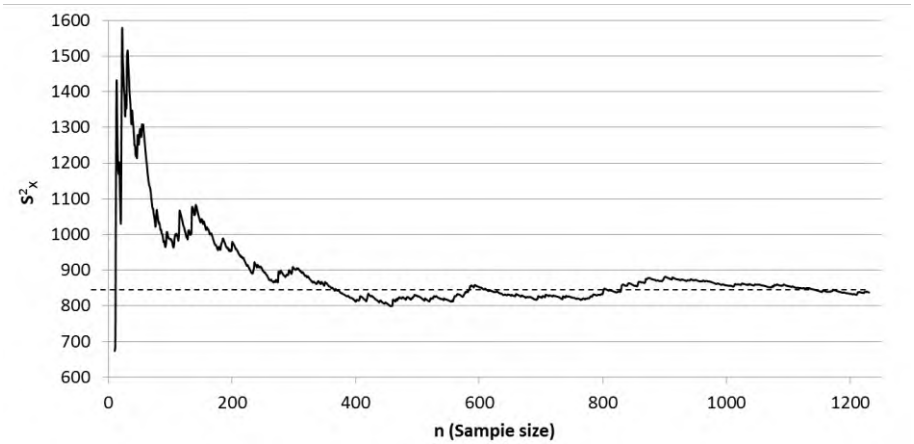

Fig. 14. Dependence of the value of the variance of the average length of a pair of needles of Calabrian pine (*Pinus brutia*) on the sample size (Turkey resort, 36 40' 58" N, 30 34' 11" E).

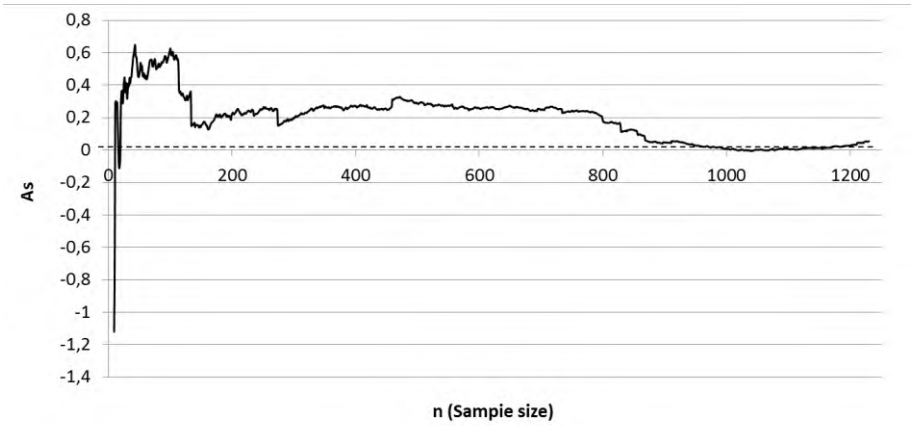

Fig. 15. Dependence of the value of the asymmetry of the distribution of the average length of a pair of needles of Calabrian pine (*Pinus brutia*) on the sample size (Turkey resort, 36 40' 58" N, 30 34' 11" E).

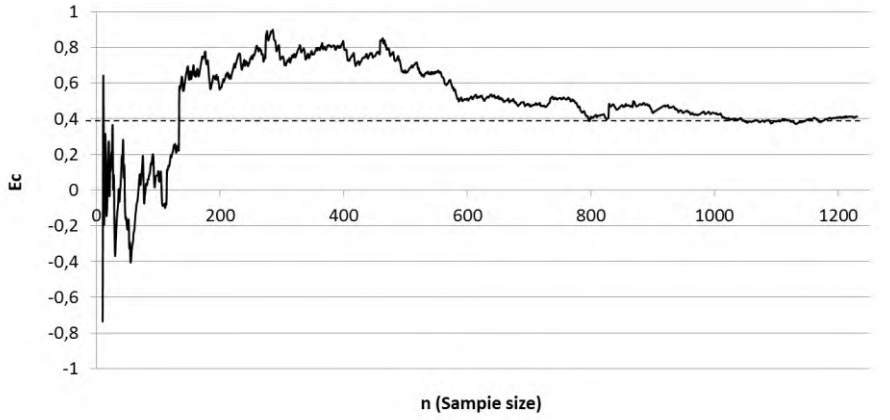

Fig. 16. Dependence of the excess value in the distribution of the average length of a pair of needles of Calabrian pine (*Pinus brutia*) on the sample size (Turkey resort, 36 40' 58" N, 30 34' 11" E).

The mean value and variance of the Calabrian pine stabilizes at a sample size of 1100, the asymmetry begins to fluctuate around zero starting from the 900 variants, and the excess stabilizes around 0.4 with a sample size of slightly more than 1000 variants. The values of asymmetry and excess do not exceed the critical ones, but the excess stabilizes around a nonzero value.

Thus, despite the fact that in a number of cases it is impossible to strictly reject the hypothesis of a normal distribution, the excess in all cases amounted to a nonzero value, which does not tend to zero with an increase in the sample size, while the asymmetry stabilized around zero only in the case of the Calabrian pine growing in the resort area. The normal distribution of the size assumes zero (tending to zero) values of asymmetry and excess [3].

In terms of processing experimental data, one should take into account the possibility of obtaining false positive or false negative results with small sample sizes. To avoid this, it is necessary to build a graph of the dependence of the value on the sample size.

3 Conclusions

1. Neglecting to substantiate the sample size can lead to false positive or false negative results.
2. In terms of the examination of experimental data, the possibility of deliberate adjustments to the data by changing the sample size should be taken into account. This method of adjusting the results is almost flawless and suitable for publications with open data, since checking the calculations will give the stated results. Revealing such moments should be done by plotting the dependence of the value on the sample size - if the graph goes steeply down or up and does not fluctuate around a certain value, the researcher took an extreme value precisely for the purpose of fine-adjusting the result.
3. The sample sizes for obtaining reliable estimates of the arithmetic mean, variance, indicators of asymmetry and excess for needles of spruce, Scots pine and Mediterranean pine are at least 800 variants. A sample size of more than 2000 variants is impractical.
4. The distributions of the length of the needles of Scots pine, European larch and average length of a pair of Scots pine have nonzero values of asymmetry and excess, the distribution of the average length of a pair of needles of Calabrian pine is characterized by zero asymmetry but nonzero excess.

This work was supported by the state mission of the Ministry of Science and Higher Education of the Russian Federation (785.00.X6019).

References

1. E. Rosenthal, Experimenter effects in behavioral research (New York: John Wiley, 1976) https://www.gwern.net/docs/statistics/bias/1976-rosenthal-experimenterexpectancyeffects.pdf
2. G. Kozlov, M. Pushkarev, A. Kozlov, E. Perepelitsa, *Bioindication for the search of microorganisms-destructors,* Advances in Intelligent Systems and Computing **1259** AISC, 676-684 (2021)
3. G.F. Lakin, Biometrics: Textbook for biology specialist of universities (4th ed., rev. and add. M.: Higher school, 1990)
4. T.K. Oleksyk, J.M. Novak, J.R. Purdue, S.P. Gashchak, M.H. Smith, *High levels of fluctuating asymmetry in populations of Apodemusflavicollis from the most contaminated areas in Chornobyl,* Journal of Environmental Radioactivity **73**, 1, 1-20 (2004)
5. E. Chudzinska, E.M. Pawlaczyk, K. Celinski, J. Diatta, *Response of Scots pine (Pinussylvestris L.) to stress induced by different types of pollutants – testing the fluctuating asymmetry,* Water and Environment Journal **28,** 533–539 (2014)
6. D. Scutt, G.A. Lancaster, J.T. Manning, *Breast asymmetry and predisposition to breast cancer,* BreastCancerRes **8**, R14 (2006) https://doi.org/10.1186/bcr1388
7. GOST R ISO 5479-2002. Statistical methods. Checking the deviation of the probability distribution from the normal distribution.