DOI: 10.18287/2542-0461-2020-11-3-176-183
УДК 519.222
Научная статья / Scientific article

Дата: поступления статьи / Submitted: 10.05.2020
после рецензирования / Revised: 18.06.2020
принятия статьи / Accepted: 28.08.2020

В.М. Дуплякин
Самарский национальный исследовательский университет
имени академика С.П. Королева, г. Самара, Российская Федерация
E-mail: v.duplyakin@gmail.com. ORCID: http://orcid.org/0000-0001-6929-356X

Особенности идентификации нормального закона распределения

Аннотация: Статистический анализ выборочных данных является эффективным инструментом исследования трендов экономических процессов и их критических состояний. Широко используемый на практике инструментарий статистических исследований основывается на предположении нормального закона распределения рассматриваемых выборочных данных. В статье автор раскрывает, что применение популярного в таких задачах критерия согласия К. Пирсона для подтверждения нормальности распределений выборочных данных может приводить к ложным выводам, в случаях когда исходная генеральная совокупность распределена по нормальному закону, а критерий указывает на низкую вероятность реализации гипотезы нормальности. Предлагает численную процедуру исследования особенностей идентификации нормальности выборочных данных, использующую оригинальный инструмент в виде эталонных статистических рядов, которые соответствуют выборкам определенного объема при заданных статистических оценках математического ожидания и среднего квадратического отклонения. Автор представил методику численного моделирования и результаты исследования характеристик выборочных данных, влияющих на ошибки в идентификации принадлежности к генеральной совокупности, имеющей нормальное распределение. Проведенные численные эксперименты позволили получить статистические данные для исследования достоверности идентификации выборочных распределений. Автор привел рекомендации, позволяющие избежать ошибок идентификации нормального распределения выборочных данных.

Ключевые слова: нормальное распределение, выборочные данные, идентификация, эталонный статистический ряд, число интервалов, численный эксперимент, критерий Пирсона, число связей, формула Стерджесса, ошибки идентификации.
Цитирование. Дуплякин В.М. Особенности идентификации нормального закона распределения // Вестник Самарского университета. Экономика и управление. 2020. Т. 11, № 3. С. 176–183. DOI: http://doi.org/10.18287/2542-0461-2020-11-3-176-183.
Информация о конфликте интересов: автор заявляет об отсутствии конфликта интересов.

V.M. Duplyakin
Samara National Research University, Samara, Russian Federation
E-mail: v.duplyakin@gmail.com. ORCID: http://orcid.org/0000-0001-6929-356X

Nuances of identification for normal distribution

Abstract: Statistical analysis of sample data is an effective tool for researching trends in economic processes and their critical conditions. The techniques in statistical analysis that are widely used in practice are based on the assumption that the sample data being considered follows a normal distribution. In the article the author reveals that the application of the popular K. Pearson criterion of agreement in such problems to confirm normality distributions of sample data can lead to false conclusions, in cases where the original general population is distributed according to the normal law, and the criterion indicates a low probability of implementing the normality hypothesis. The author proposes a numerical procedure for studying the nuances of identifying the normality in sample data; it uses a novel technique that is based on reference statistical series which correspond to samples of a certain size with the given, fixed estimates of the expected value and standard deviation. The author presents a numerical modeling method and the results of studying the characteristics of sample data that affect the errors in the identification of the normality of the sampled populations. The performed numerical experiments allowed us to obtain statistical data for investigating the reliability of the identification of the sampled distributions. The author presented recommendations that can help to avoid errors in identifying normality.

Key words: normal distribution, sample data, identification, reference statistical series, number of intervals, numerical experiment, Pearson's test, number of links, Sturges' formula, identification error.
Citation. Duplyakin V.M. Nuances of identification for normal distribution. *Vestnik Samarskogo universiteta. Ekonomika i upravlenie = Vestnik of Samara University. Economics and Management*, 2020, vol. 11, no. 3, pp. 176–183. DOI: http://doi.org/10.18287/2542-0461-2020-11-3-176-183. (In Russ.)
Information on the conflict of interest: author declares no conflict of interest.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

© *Вячеслав Митрофанович Дуплякин* – доктор экономических наук, профессор кафедры экономики, Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Российская Федерация, г. Самара, Московское шоссе, 34.
© *Vyacheslav M. Duplyakin* – Doctor of Economics, professor of the Department of Economics, Samara National Research University, 34, Moskovskoye shosse, Samara, 443086, Russian Federation.

Введение

В практической экономической статистике зачастую без какого-либо обоснования принимается нормальный закон распределения исследуемых данных. Такой подход базируется на теоретическом фундаменте в виде Центральной Предельной Теоремы Теории Вероятностей, которую в наиболее строгом виде сформулировал и доказал в 1887 году методом моментов П.Л. Чебышев [1; 2]. Как известно, в этой теореме утверждается, что некий результат будет иметь нормальное распределение при соблюдении совокупности ограничений, выполняющихся на практике в подавляющем числе случаев, так, например, распределение рыночной стоимости производимой продукции, процент брака и многое другое. С другой стороны, большинство программно-инструментальных средств статистического анализа, таких как STATISTICA [3], SPSS [4], MATLAB [5], EXCEL [6], включающих, например, построение доверительных интервалов, разработаны для обработки выборок в предположении нормального закона распределения соответствующих им генеральных совокупностей. Если генеральная совокупность распределена по какому-то другому закону, то найти программное обеспечение для решения задач статистического анализа затруднительно. Более того, если необходимое программное обеспечение будет найдено, то, как правило, численные результаты его применения мало отличаются от того, которое разработано для случая нормального распределения генеральной совокупности. Поэтому разработки средств статистического анализа для распределений, отличающихся от нормального, практически не ведутся из-за малой востребованности в практической деятельности. Тем не менее вопрос остается открытым, поскольку во многих случаях исследуемые данные распределяются по закону равномерной плотности, по закону Вейбулла или как-то иначе, причем подмена таких законов нормальным распределением приводит к значительным искажениям численных оценок вероятностей критических событий, рисков и т. п. Добросовестный аналитик, желая обеспечить достоверность своих выводов, проводит исследование нормальности выборочных распределений, используя соответствующие статистические критерии, из которых здесь чаще других применяется критерий Пирсона [7–9].

Постановка задачи

Рассматривается ситуация, когда диагностика выборки из нормальной генеральной совокупности может приводить к ложным результатам диагностирования существенных различий с нормальным распределением, в то время как генеральная совокупность имеет именно нормальный закон распределения. Многие исследователи в этой связи в качестве причины ложной диагностики упоминают выбор числа интервалов предварительной обработки выборочных данных путем построения так называемого статистического ряда, однако каких-либо конкретных рекомендаций не приводится.

Ход исследования

С целью численного моделирования и последующего исследования особенностей идентификации выборок заданного объема из генеральной совокупности с нормальным распределением и известными параметрами предлагается трехэтапная численная процедура генерирования эталонных статистических рядов

$$G = G(N,\ k,\ m_x,\ s_x,\ dx) = \{(x_i,\ x_{i+1};\ n_i);\ i = 1,...,k\},$$

где N – объем выборки, k – число разрядов регистрации наблюдаемого признака, m_x и s_x – математическое ожидание и среднее квадратическое отклонение в соответствующей генеральной совокупности, $dx = x_{i+1} - x_i$ – ширина интервала регистрации данных, x_i, x_{i+1} – границы текущего интервала (разряда) статистического ряда.

На первом этапе генерирования эталонного статистического ряда вычисляются накопленные в отдельных разрядах частоты появления случайной величины, распределенной по нормальному закону:

$$n_i^* = ЦЕЛОЕ(p_i \cdot N),\ \text{где}\ p_i = F_{норм}(x_i, m_x, s_x) - F_{норм}(x_{i+1}, m_x, s_x);\quad i = 1,...,k, \tag{1}$$

где $F_{норм}$ – интегральная функция нормального распределения.

Несмотря на очевидную простоту данной операции, следует отметить, что использование выделения целой части произведений $(p_i \cdot N);\ i = 1,...,k$ приводит к искажению значения заданного объема выборки, т. е. получаем

$$N^* = \sum_{i=1}^{k} n_i^* \neq N. \tag{2}$$

Поэтому предлагается коррекция накопленных частот:

$$n_i = ЦЕЛОЕ(n_i^* \cdot N / N^*);\ i = 1,...,k. \tag{3}$$

На втором этапе коррекции эталонного статистического ряда производится коррекция, обеспечивающая равенство выборочного и задаваемого значений среднего квадратического отклонения. Для этого вводится поправочный множитель α, изменяющий границы разрядов:

$$x_i^* = \alpha \cdot x_i;\quad \alpha = s_x / s_{факт};\ i = 1,...,k+1, \tag{4}$$

где s_x и $s_{факт}$ – задаваемое и расчетное значения среднего квадратического отклонения моделируемого статистического ряда.

На третьем этапе коррекции сгенерированного эталонного статистического ряда обеспечивается равенство выборочного и задаваемого значений математического ожидания. Для этого вводится поправка границ разрядов Δ:

$$x_i^{**} = x_i^* + \Delta;\quad \Delta = m_x - m_{факт};\ i = 1,...,k+1, \tag{5}$$

где m_x и $m_{факт}$ задаваемое и расчетное значения математического ожидания моделируемого статистического ряда.

Выполнив коррекции, заменим границы всех разрядов эталонного ряда:

$$x_i = x_i^{**};\ i = 1,...,k+1. \tag{6}$$

В качестве примера обратимся к моделированию эталонного статистического ряда, соответствующему нормальному распределению с характеристиками N = 50000, m_x = 100, s_x = 5, k = 50. В этом случае корректирующие характеристики имеют следующие значения: коррекция среднего квадратического отклонения α = 0,997196, коррекция математического ожидания β = 0,280366, при этом Δ = 1,37114 , а сам смоделированный ряд приведен в таблице 1.

Таблица 1 – Пример эталонного статистического ряда (m_x = 100, s_x = 5, k = 50)
Table 1 – An example of a reference statistical series (mx = 100, sx = 5, k = 50)

x_i	79,43	80,80	82,18	83,55	84,92	86,29	87,66	89,03	90,40	91,77
n_i	2	6	15	38	87	184	362	661	1118	1755
x_i	93,14	94,52	95,89	97,26	98,63	100,00	101,37	102,74	104,11	105,48
n_i	2555	3451	4324	5025	5417	5417	5025	4324	3451	2555
x_i	106,86	108,23	109,60	110,97	112,34	113,71	115,08	116,45	117,82	119,20
n_i	1755	1118	661	362	184	87	38	15	6	2

Чтобы исследовать распределение выборочных данных в плане соответствия нормальному закону, воспользуемся наиболее распространенным решением, использующим критерий Пирсона [7; 8], в котором в качестве меры расхождения распределений предложена величина

$$U = N\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i}\ , \tag{7}$$

где k – число разрядов статистического ряда, используемого для предварительного анализа при вычислении статистических оценок математического ожидания m_x* и среднего квадратического отклонения s_x* , $N = \sum_{i=1}^{k} n_i$ – общее число наблюдений, n_i – число наблюдений, зафиксированное в i-м разряде, $p_i^* = \frac{n_i}{N}$ – частота появления в i-м разряде (статистическая оценка вероятности), p_i – вероятность появления события в данном разряде в соответствии с выбранным теоретическим законом распределения, в наших исследованиях это нормальный закон распределения.

К. Пирсон показал, что величина U имеет распределение, называемое распределением хи-квадрат χ^2, или распределением Пирсона, которое зависит от числа «степеней свободы»:

$$r = k - s, \tag{8}$$

где k – число разрядов; s – число связей, определяемое при использовании в качестве теоретического закона нормального распределения.

Вероятность β, равную значению доверительной вероятности приемлемости нормального закона распределения, найдем из решения следующего уравнения:

$$\chi^2(r, \beta) = U\ . \tag{9}$$

Рассмотрим результаты численного моделирования эталонных статистических рядов, соответствующих нормальному распределению с заданными характеристиками, с последующим определением вероятности соответствия нормальному закону распределения (рисунок 1).

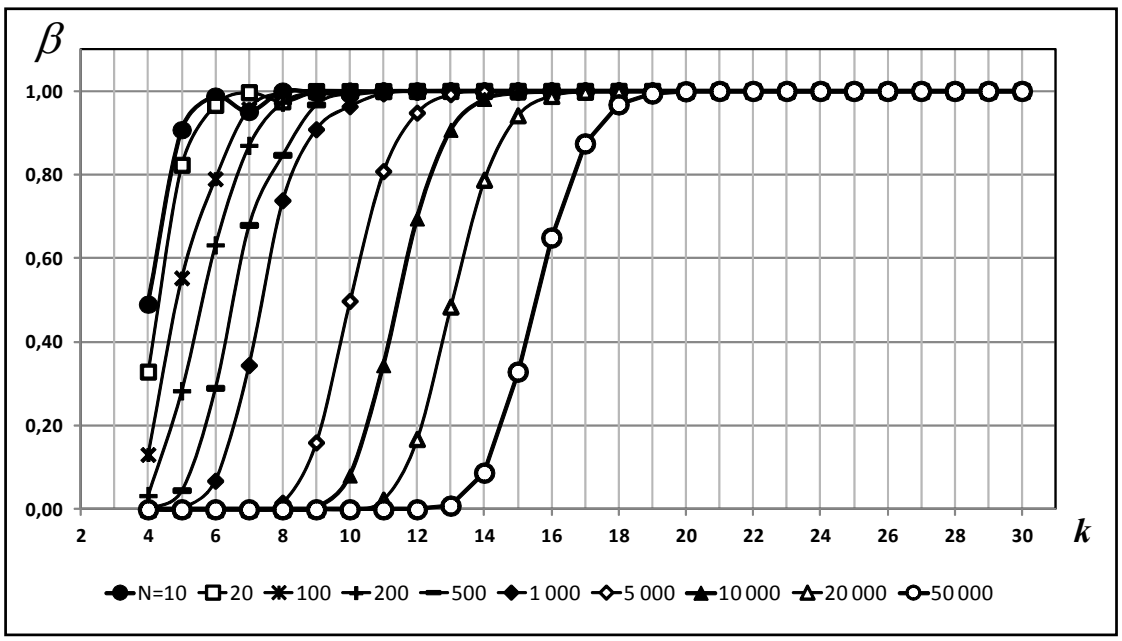

Рисунок 1 – Вероятность реализации гипотезы нормальности распределения – β (m_x = 100, s_x = 5, s = 3)

Figure 1 – Probability of realizing the hypothesis of normal distribution – β (m_x = 100, s_x = 5, s = 3)

Обработка полученных результатов моделирования в интервале объемов выборочных данных $20 < N < 50000$ позволила найти выражение для характеристик выборочных данных, удовлетворяющих условию правильной идентификации нормального распределения с вероятностью не меньше 0,95, в виде

$$k(0,95) = 3,84 \cdot N^{0,15}. \quad (10)$$

Расчетные значения необходимого числа разрядов $k(0,95)$ для различного объема выборок представлены в таблице 2.

Отметим широко используемый прием определения числа разрядов с помощью полуэмпирического соотношения, называемого формулой Стерджесса [10–12]:

$$k(\text{стр}) = 1 + \log_2 N. \quad (11)$$

Результаты оценки необходимого числа разрядов, вычисленные по формуле Стерджесса $k(\text{стр})$, представлены в таблице 2.

Рассмотрим значения вероятностей соответствия нормальному распределению $\beta(\text{стр})$, полученные при выборе числа разрядов, найденных по формуле Стерджесса. Как видно из таблицы 2, использование формулы Стерджесса приводит к значительным ошибкам в идентификации нормального распределения выборочных данных, если объемы выборок не превышают N = 20 или больше N = 20000.

Таблица 2 – К выбору числа интервалов регистрации выборочных данных (m_x = 100, s_x = 5, s = 3)
Table 2 – On the choice of the number of intervals for recording sample data (m_x = 100, s_x = 5, s = 3)

N	10	20	100	200	500	1000	5000	10000	20000	50000
K (0,95)	5	6	7	8	9	10	13	15	16	19
k (стр)	4	5	7	8	9	10	13	14	15	16
β(стр)	0,5	0,82	0,95	0,95	0,95	0,95	0,95	0,95	0,94	0,75

Продолжение численных экспериментов показало, что, вопреки прогнозам, изменение математического ожидания и среднего квадратического отклонения генерируемых эталонных статистических рядов в широком диапазоне характеристик выборочных данных никак не отражается на полученной зависимости расчетного значения вероятности гипотезы нормальности распределения от числа интервалов статистического ряда.

Следует отметить достаточно дискуссионный вопрос назначения числа степеней свободы r при использовании критерия Пирсона для проверки нормальности распределения. Например, в учебнике Е.С. Вентцель [9] рекомендуется принимать $r = k - s$ при s = 3, имея в виду, что число накладываемых связей s учитывает связь по математическим ожиданиям, по дисперсиям и связь в виде суммы накопленных связей, часто равной единице. Именно так назначено $r = k - 3$ при получении обсуждаемых результатов, т. е. принято s = 3. Однако, например, в упомянутых ранее программных продуктах во многих случаях назначается s = 2 или даже s = 1 , что, на наш взгляд, менее обоснованно.

Повторив численные эксперименты с последующей обработкой при минимальном числе связей s = 1, убеждаемся, что назначаемое число связей резко изменяет влияние объема выборки на идентификацию исходного нормального распределения генеральной совокупности, как это продемонстрировано на графиках рисунка 3 и в таблице 3, если сравнить их с графиками рисунка 1 и результатами в таблице 2.

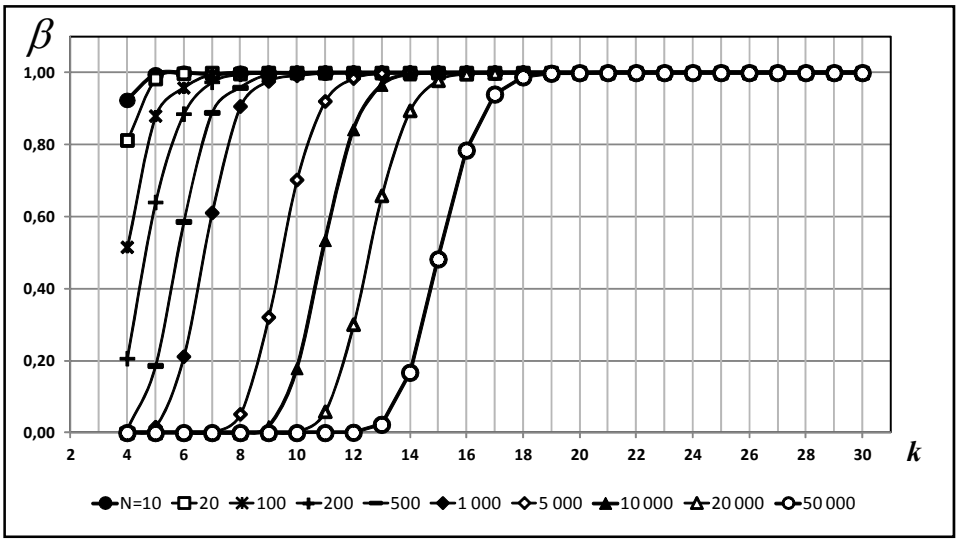

Рисунок 2 – Расчетные значения вероятности реализации гипотезы нормальности выборочного распределения ($m_x = 100$, $s_x = 1$, $s = 1$)

Figure 2 – Calculated values of the probability of realizing the hypothesis of normality of the sample distribution ($m_x = 100$, $s_x = 1$, $s = 1$)

Таблица 3 – К выбору числа интервалов регистрации выборочных данных ($m_x = 100$, $s_x = 5$, $s = 1$)
Table 3 – On the choice of the number of intervals for recording sample data ($m_x = 100$, $s_x = 5$, $s = 1$)

N	10	20	100	200	500	1000	5000	10000	20000	50000
$k(0,95)$	5	5	7	7	9	9	12	13	15	17
k (стр)	4	5	7	8	9	10	13	14	15	16
β(стр)	0,92	0,975	0,975	0,975	0,975	0,975	0,975	0,975	0,95	0,78

Число разрядов статистического ряда, обеспечивающего идентификацию нормального распределения с вероятностью не менее 0,95 с помощью критерия Пирсона при назначении числа связей $s = 1$, определяется следующей формулой:

$$k(0,95) = 3,71 \cdot N^{0,143}. \quad (12)$$

При выборе числа связей критерия Пирсона $s = 1$ можно заметить расширение интервала возможных значений числа разрядов, определяемого по формуле Стерджесса, а именно: недопустимые ошибки идентификации нормального распределения выборочных данных имеют место, только если объемы выборок больше $N = 20000$.

Необходимо отметить техническую особенность проведения рассматриваемых численных экспериментов. Сгенерировано 9000 эталонных статистических рядов с последующим определением оценок нормальности по критерию Пирсона, что представляет достаточно трудоемкую вычислительную задачу, для решения которой нами разработано программное обеспечение на языке Visual Basic for Application Excel.

Полученные результаты и выводы

1. В статье исследуются причины ложных выводов о распределении выборочных данных генеральной совокупности, которая, безусловно, подчиняется нормальному закону распределения.

2. Предложена трехэтапная процедура генерирования эталонных статистических рядов, необходимых для численного исследования особенностей идентификации нормальности выборочного распределения.

3. Проведена серия численных экспериментов, позволивших выявить условия неадекватного применения критерия согласия К. Пирсона при анализе нормальности выборочных распределений.

4. Показано, что надежность идентификации нормального распределения выборочных данных с использованием критерия Пирсона существенным образом зависит не только от объема выборочных данных, но и от выбранного числа степеней свободы при идентификации распределения с помощью критерия Пирсона и от числа разрядов статистического ряда.

5. Сформулированы рекомендации по выбору числа разрядов статистических рядов, обеспечивающие достоверную идентификацию распределения выборочных данных по нормальному закону. Определены ограничения использования известной формулы Стерджесса для назначения числа разрядов при предварительной обработке выборочных данных.

Библиографический список

1. Чебышев П.Л. Полное собрание сочинений. Т. III. Москва: Изд-во АН СССР, 1948. 404 с.

2. Прохоров А.В. Моментов метод // Математическая энциклопедия / гл. ред. И.М. Виноградов. Т. 3. Москва: Сов. энциклопедия, 1982. 1184 с.

3. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. СПб.: Издательский дом «Питер», 2003. 688 с.

4. Бююль А., Цефель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: пер. с нем. Санкт-Петербург: ДиаСофтЮП, 2005. 608 с.

5. Дьяконов В.П. MATLAB. Полный самоучитель. Москва: ДМК Пресс, 2012. 768 с.

6. Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. Статистические функции MSEXCEL в экономико-статистических расчетах. Москва: ЮНИТИ. 2003. 231 с.

7. Дуплякин В.М. Статистический анализ выборочных данных: учеб. пособие. Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2010. 110 с.

8. Митропольский А.К. Техника статистических вычислений. Изд. 2-е, перераб. и доп. Москва: Наука, 1971. 576 с.

9. Вентцель Е.С. Теория вероятностей: учебник. 12-е изд., стереотип. Москва: Изд-во Кнорус, 2018. 664 с.

10. Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия χ^2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. С. 61–67. URL: https://ami.nstu.ru/~headrd/seminar/publik_html/Z_lab_8.htm.

11. Эконометрика: учебник для магистров / И.И. Елисеева [и др.]; под ред. И.И. Елисеевой. Москва: Юрайт, 2014. 453 с.

12. Sturges H. The choice of a class-interval. J. Amer. Statist. Assoc., 1926, 21, pp. 65–66. DOI: http://doi.org/10.1080/01621459.1926.10502161.

References

1. Chebyshev P.L. Complete works. Vol. III, Moscow: Izd-vo AN SSSR, 1948, p. 404. (In Russ.)

2. Prokhorov A.V. Method of moments. In: *Vinogradov I.M. (Ed.) Mathematical encyclopedia*. Vol. 3, Moscow: Sov. entsiklopediia, 1982, 1184 p. (In Russ.)

3. Borovikov V. STATISTICA. The Art of Computer Data Analysis: For Professionals. 2nd edition. Saint Petersburg: Izdatel'skii dom «Piter», 2003, 688 p. (In Russ.)

4. Bühl A., Zöfel P. SPSS: The Art of Information Processing. Analysis of statistical data and restoration of hidden patterns. Translation from German. Saint Petersburg: DiaSoftIuP, 2005, 608 p. (In Russ.)

5. Dyakonov V.P. MATLAB. Complete tutorial. Moscow: DMK Press, 2012, 768 p. (In Russ.)

6. Kozlov A.Yu., Mkhitaryan V.S., Shishov V.F. Statistical functions of MSEXCEL in economic and statistical calculations. Moscow: IuNITI, 2003, 231 p. (In Russ.)

7. Duplyakin V.M. Statistical analysis of sample data: textbook. Samara: Izd-vo Samar. gos. aerokosm. un-ta, 2010, 110 p. (In Russ.)

8. Mitropol'skiy A.K. Technique of statistical computing. 2nd edition, revised and enlarged. Moscow: Nauka, 1971, 576 p. (In Russ.)

9. Ventsel E.S. Probability theory: textbook. 12th edition, stereotyped. Moscow: Knorus, 2018, 664 p. (In Russ.)

10. Lemeshko B.Yu., Chimitova E.V. On the choice of the number of intervals in the criteria of agreement χ^2. *Industrial Laboratory. Diagnostics of Materials*, 2003, vol. 69, pp. 61–67. Available at: https://ami.nstu.ru/~headrd/seminar/publik_html/Z_lab_8.htm. (In Russ.)

11. Eliseeva I.I. et al. Econometrics: textbook for masters. Moscow: Iurait, 2014, 453 p. (In Russ.)

12. Sturges H. The choice of a class-interval. *Journal of American Statistical Association*, 1926, 21, pp. 65–66. DOI: http://doi.org/10.1080/01621459.1926.10502161.