DOI: 10.18287/2542-0461-2021-12-4-165-172

НАУЧНАЯ СТАТЬЯ

УДК 519.222

Дата поступления: 13.10.2021
рецензирования: 20.11.2021
принятия: 26.11.2021

Вопросы идентификации распределения выборочных данных при ограничении нижней границы рассеивания наблюдаемых значений

В.М. Дуплякин
Самарский национальный исследовательский университет имени С.П. Королева,
г. Самара, Российская Федерация
E-mail: v.duplyakin@gmail.com. ORCID: http://orcid.org/0000-0002-7433-3188

Аннотация: Статистический анализ выборочных данных является широко распространенным инструментом исследований в различных отраслях научных знаний и в их приложениях, в том числе в исследовании экономических процессов и критических состояний, но в то же время вызывает ряд вопросов в связи с выбором теоретического закона распределения в генеральной совокупности, включающей исследуемую выборку данных. Последнее требуется для достоверного прогнозирования рисков и надежности, поскольку в этих задачах требуется прогнозировать достаточно малые или, наоборот, близкие к единице вероятности. Для исследования вопросов идентификации выборочных данных численным путем разработано программное обеспечение, включающее генерирование псевдослучайных выборок, подчиняющихся распределению Вейбулла с заданной нижней границей рассеивания, с последующим определением принадлежности как к исходному распределению, так и к нормальному распределению. Проведен численный эксперимент с широким интервалом варьирования параметров рассматриваемых распределений и с использованием критерия согласия Пирсона для идентификации распределения выборочных данных. Анализ результатов численного моделирования при широком диапазоне варьирования объема выборочных данных и их параметров показал высокую вероятность ложной идентификации нормального распределения выборочных данных, в то время как на самом деле они соответствуют распределению Вейбулла с фиксированной нижней границей рассеивания.

Ключевые слова: моделирование выборочных данных; нормальное распределение; распределение Вейбулла; нижняя граница рассеивания; согласование выборочных данных; критерий согласия Пирсона; численный эксперимент; объем выборки; параметры распределения Вейбулла; ложная идентификация нормального распределения.

Цитирование. Дуплякин В.М. Вопросы идентификации распределения выборочных данных при ограничении нижней границы рассеивания наблюдаемых значений // Вестник Самарского университета. Экономика и управление. 2021. Т. 12, № 4. С. 165–172. DOI: http://doi.org/10.18287/2542-0461-2021-12-4-165-172.

Информация о конфликте интересов: автор заявляет об отсутствии конфликта интересов.

Вячеслав Митрофанович Дуплякин – доктор технических наук, профессор кафедры экономики, Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Российская Федерация, г. Самара, Московское шоссе, 34.

SCIENTIFIC ARTICLE

Submitted: 13.10.2021
Revised: 20.11.2021
Accepted: 26.11.2021

Issues of identification of the distribution of sampled data with the limitation of the lower boundary of scattering of the observed values

V.M. Duplyakin
Samara National Research University, Samara, Russian Federation
E-mail: v.duplyakin@gmail.com. ORCID: http://orcid.org/0000-0002-7433-3188

Abstract: Statistical analysis of empirical data is a commonly used approach for research in various fields of science and in applications, including studies of economic processes and critical conditions, but at the same time, there are numerous questions regarding the selection of theoretical distribution laws in general populations that include the sample data being studied. This selection is required for reliable forecasting of risks and reliability because these tasks require the prediction of rather small probabilities or, conversely, the probabilities that approach 1.0. For studying issues with numerical identification of empirical data, a software tool has been developed; it includes drawing of pseudo-random samples from Weibull distribution with a given lower threshold of dispersion, followed by the determination of whether the samples belong to the original distribution or to the normal distribution. A numerical experiment has been carried out with a wide range of variation in the parameters of the considered distributions and using the Pearson's goodness-of-fit test for identification of the sample data's distribution. An analysis of the results of the numerical modeling, which incorporated significant variation of the volume of the samples and their parameters, showed the high probability of false identification of the normal distribution of the sample data, whereas, in fact, the samples were drawn from Weibull distribution with a fixed lower threshold of dispersion.

Key words: modeling of sample data; normal distribution; Weibull distribution; lower scattering bound; agreement of sample data; Pearson's goodness-of-fit test; numerical experiment; sample size; Weibull distribution parameters; false identification of the normal distribution.

Citation. Duplyakin V.M. Issues of identification of the distribution of sample data when limiting the lower boundary of the dispersion of the observed values. *Vestnik Samarskogo universiteta. Economika i upravlenie = Vestnik of Samara University. Economics and management,* vol. 12, no. 4. pp. 165–172. DOI: http://doi.org/10.18287/2542-0461-2021-12-4-165-172. (In Russ.)

Information on the conflict of interest: author declares no conflict of interest.

Vyacheslav M. Duplyakin – Doctor of Technical Sciences, professor of the Department of Economics, Samara National Research University, 34, Moskovskoye shosse, Samara, 443086, Russian Federation.

Введение

Актуальность рассматриваемой темы обусловлена развитием технической базы информационных технологий, стимулирующих регулярный рост числа прикладных исследований, в которых авторы, пользуясь возможностями современных специализированных программных средств, ограничиваются большей частью формальной проверкой гипотезы нормальности распределения. Получив высокие значения вероятности соответствия нормальному закону в генеральной совокупности данных, исследователи зачастую не задумываются о том, что на самом деле с еще большей достоверностью генеральная совокупность исследуемых данных может подчиняться другому закону распределения, например, трехпараметрическому распределению Вейбулла с заданной нижней границей рассеивания.

Остановимся на причинах ограничивающих исследователей в выборе гипотез о законе распределения генеральной совокупности, и как следствие безоговорочно отдающих предпочтение нормальному распределению. Во-первых, известная Центральная Предельная Теорема Теории Вероятностей [1], формулировка которой создает впечатление об аксиоматическом превосходстве гипотезы нормального распределения экспериментальных данных. Во-вторых, высокие значения доверительных вероятностей, получаемые с использованием известных критериев согласия [2]. Кроме того, дополнительную уверенность в универсальности нормального распределения создает известное свойство этого закона, в соответствии с которым сумма нормально распределенных случайных величин также подчиняется нормальному закону [3].

С другой стороны, в диссонансе с предыдущим выступают соображения, вызываемые известными свойствами нормального распределения, такими как симметрия и бесконечный интервал возможных значений, которые обычно не наблюдаются в реальных выборочных данных. Возникающие на этой почве недоумения обычно рассеиваются апелляцией к тезису об ограниченном объеме любой выборки по сравнению с неограниченным объемом ее генеральной совокупности.

Существуют в какой-то мере технические проблемы идентификации распределения генеральной совокупности по имеющимся выборочным данным. У этой проблемы два аспекта, во-первых, выбор инструмента идентификации среди множества критериев согласия. Здесь можно воспользоваться рекомендациями, которые представлены в работах [4; 5].

Многочисленные практические руководства и рекомендации [4–6], а с другой стороны, наличие разнообразных критериев согласия и доступных программных средств для их применения. Такая ситуация чисто психологически не стимулирует исследователей заниматься их детальным изучением и освоением различных критериев согласия, выбрав по старинке один из популярных критериев проверки нормальности распределения и здесь на первом месте традиционно выступает критерий Пирсона [7].

Если реальные выборочные данные при использовании одного из критериев согласия показывают высокую вероятность соответствия нормальному распределению, то надо понимать, что это в некотором смысле интегральная оценка. В то время как при оценке рисков или надежности в прикладных задачах используются "хвосты" распределений, соответствующие вероятностям меньше 0,05 (например, вероятности убытков и разрушений) или превышающие 0,95 (например, вероятности выполнения плановых заданий и достижения нормативных значений). Но именно здесь особенно важно не ошибиться с выбором теоретического распределения, которому подчиняются выборочные данные.

Постановка задачи

На основе результатов применения разработанной численной процедуры статистического имитационного моделирования рассматриваются условия, при которых выборка из генеральной совокупности с явным заданием нижней границы рассеивания описываемая распределением Вейбулла уверенно идентифицируется критерием Пирсона как выборка из нормального распределения. Ложная идентификация нормального распределения при его последующем применении приводит к существенным ошибкам в прогнозировании рисков надежности функционирования систем различной природы и назначения.

Ход исследования

Обратимся в качестве возможной альтернативы нормальному распределению к распределению Вейбулла с заданной нижней границей рассеивания, которое в математическом плане является моделью распределения экстремальных значений, построенной на основе так называемой "теории слабого звена" [8].

Можно привести многочисленные примеры распределений выборочных данных, в которых нижняя граница рассеивания имеет физические ограничения. Например, в механике сплошных сред: предел прочности материалов при различных видах деформирования. В энергетике – мощность источников энергии различной природы (тепловые электростанции, электростанции с возобновляемыми источниками энергии). В экономике: себестоимость продукции машиностроения в условиях ее гарантированного спроса, доходность высоколиквидных финансовых обязательств. Распределение Вейбулла применяется для описания ресурса объектов машиностроения, для характеристик внешних воздействий, таких как сила ветра, интенсивность дождя, в биологии – время прорастания семян, в промышленности – продолжительность простоев и во многих других задачах.

Интегральная функция распределения Вейбулла задается выражением [8]

$$F(x) = P(X < x) = 1 - \exp\left[-\left(\frac{x-u}{\Theta}\right)^{\alpha}\right], \qquad (1)$$

где u – нижняя граница рассеивания наблюдаемых значений, α – параметр формы, Θ – масштабный фактор.

Математическое ожидание и среднее квадратическое отклонение распределения Вейбулла находятся по формулам

$$m_x = \Theta \cdot \Gamma\left(1+\frac{1}{\alpha}\right) + u, \qquad s_x = \Theta\sqrt{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)}. \qquad (2, 3)$$

В формулах 2 и 3 используется известная неаналитическая гамма-функция Даниэля Бернулли [10]

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt. \qquad (4)$$

Численные значений гамма-функции можно получить из таблиц [9] или воспользовавшись каким-либо программным обеспечением, например, встроенной функцией ГАММА(x) в Excel [11].

Рассмотрим иллюстрацию распределения Вейбулла в сопоставлении с соответствующими данными нормального распределения, изображенную на рисунках 1 и 2. Здесь представлены функции распределения Вейбулла с фиксированными параметрами $u = 64{,}00$ и $\theta = 40{,}00$ при двух вариантах параметра формы $\alpha = 0{,}80$ и $\alpha = 6{,}00$ и согласованные функции нормального распределения, т. е. имеющие такие же математическое отклонение и среднее квадратическое отклонение.

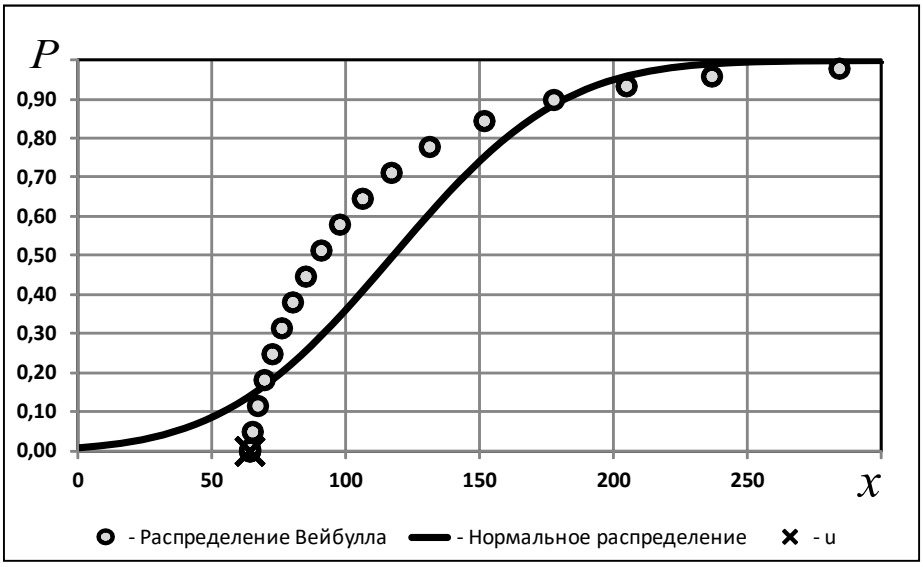

Рисунок 1 – Согласование распределений: $\alpha = 0{,}80$, $u = 64{,}00$, $\Theta = 40{,}00$, $m_x = 117{,}32$, $s_x = 49{,}74$

Figure 1 – Alignment of distributions: $\alpha = 0{,}80$, $u = 64{,}00$, $\Theta = 40{,}00$, $m_x = 117{,}32$, $s_x = 49{,}74$

Рисунок 2 – Согласование распределений: $\alpha = 6{,}00$, $u = 64{,}00$, $\Theta = 40{,}00$, $m_x = 101{,}10$, $s_x = 7{,}22$

Figure 2 – Alignment of distributions: $\alpha = 6{,}00$, $u = 64{,}00$, $\Theta = 40{,}00$, $m_x = 101{,}10$, $s_x = 7{,}22$

Как видно из рисунков 1 и 2, параметр α кардинально влияет на вид функции распределения Вейбулла и в любом случае наблюдается видимое отклонение от согласованной функции нормального распределения, не говоря о том, что функция распределения Вейбулла в отличие от нормального распределения ограничена снизу значением параметра u .

Рассмотрим имитационное моделирование выборки из генеральной совокупности, подчиняющейся распределению Вейбулла с заданными параметрами, описываемого выражением (1). Составив обратную функцию распределения Вейбулла и, подставляя в нее в качестве аргументов случайные числа $u(i)$ распределенные по закону равномерной плотности, получим формулу для генерирования случайной выборки из распределения Вейбулла в виде

$$x(i) = u + \Theta\left[-\ln(1 - P(i)\right]^{1/\alpha};\ \ i = 1,..., N. \tag{5}$$

Зная значения параметров, можно определить математическое ожидание и среднее квадратическое отклонение генеральной совокупности распределения Вейбулла по формулам (2, 3).

Рассмотрим характеристики численного эксперимента: Параметры распределения Вейбулла выбраны следующим образом u=64,00 и Θ =40,00. Параметр формы распределения α принимает фиксированные значения, указанные в таблице 1. Объем генерируемых выборок выбирался в соответствии с таблицей 2.

Таблица 1 – Значения параметра α
Table 1 – Values of the parameter α

i	1	2	3	4	5	6	7	8
α(i)	0,20	0,40	0,60	0,80	1,00	2,00	4,00	6,00

Таблица 2 – Объем выборочных данных
Table 2 – The amount of sample data

i	1	2	3	4	5	6	7	8	9	10	11
NV(i)	10	20	40	100	200	1000	2000	5000	10000	20000	50000

Число выборок в очередной серии численного эксперимента NN=20. Всего выполнено 1760 реализаций выборок из распределения Вейбулла. В таблице 3 приведены численные значения теоретических значений математического ожидания моделируемого распределения Вейбулла.

Таблица 3 – Характеристики сгенерированных данных
Table 3 – Characteristics of the generated data

α	0,20	0,40	0,60	0,80	1,00	2,00	4,00	6,00
m_x	4864,00	196,93	124,20	117,32	104,00	99,45	100,26	101,10
s_x	76046,3	417,527	105,765	49,7427	40,00	18,53	10,172	7,2157

Рассмотрим типовой фрагмент полученных данных, представленный в таблице 4. Эти данные соответствуют размеру выборки N=200, при числе разрядов для группирования k=20 и числе выборок в серии NN=20. Параметр формы распределения имеет значение α=0,20. Из приведенных данных видно, что исходное распределение Вейбулла идентифицируется по критерию Пирсона с высокими вероятностями P(B-B) равными 1,00000 и только в одном случае эта вероятность снижается до 0,99989, оставаясь весьма высокой, что и должно быть. Вероятности идентификации нормального распределения P(H-B) тех же данных к удивлению, составляют 0,98230 – 0,99944 и только в некоторых случаях опускаются до, казалось бы, очевидных значений 0,00000 – 0,00480.

Для обозрения всех полученных результатов составлена таблица 5, в которой приведена группировка по вероятностям идентификации соответствующего распределения в интервалах вероятностей (P(i); P(i+1)) при i=1,…,10. Здесь числа наблюдений m(i), n(i) в i-м интервале соответственно для

распределения Вейбулла и нормального распределения, P*(В-В)=m(i)/N, P*(Н-В)=n(i)/N – накопленные частоты.

Таблица 4 – Фрагмент смоделированных реализаций
Table 4 – Fragment of simulated implementations

i	P(В-В)	P(Н-В)	*i*	P(В-В)	P(Н-В)
92	1,00000	0,99944	96	1,00000	0,99772
93	1,00000	0,00000	97	0,99989	0,99940
94	1,00000	0,98230	98	1,00000	0,99943
95	1,00000	0,00000	99	1,00000	0,00480
96	1,00000	0,99772	100	1,00000	0,99945

Таблица 5 – Сводные данные численного эксперимента
Table 5 – Summary data of the numerical experiment

i	P(i)	P(i+1)	m(i)	n(i)	P*(В-В)	P*(Н-В)
1	0,90	1,00	1746	900	0,99205	0,51136
2	0,80	0,90	3	26	0,00170	0,01477
3	0,70	0,80	2	9	0,00114	0,00511
4	0,60	0,70	2	17	0,00114	0,00966
5	0,50	0,60	2	7	0,00114	0,00398
6	0,40	0,50	0	11	0,00000	0,00625
7	0,30	0,40	0	9	0,00000	0,00511
8	0,20	0,30	2	6	0,00114	0,00341
9	0,10	0,20	0	5	0,00000	0,00284
10	0,00	0,10	3	770	0,00170	0,43750

Из таблицы 5 видно, что более половины данных регистрируются как нормальное распределение с вероятностью, превышающей 0,90 в условиях, когда исходное распределение представляет собой распределение Вейбулла с заданной нижней границей рассеивания.

Полученные результаты и выводы

1. Сравнительный анализ результатов выполненных численных экспериментов по диагностике выборочного распределения с использованием критерия Пирсона наглядно показывает, что исходные данные, подчиняющиеся по своей природе распределению Вейбулла с явной нижней границей рассеивания, безусловно, идентифицируются с высокой доверительной вероятностью независимо от технических параметров процедуры идентификации в широком диапазоне изменения параметров исходного распределения Вейбулла. при этом в половине случаев те же данные могут идентифицироваться как нормальное распределение с весьма высокой доверительной вероятностью.

2. Наблюдаемая особенность идентификации может приводить к ошибочным выводам о нормальности распределения выборочных данных, если формально подходить к использованию критерия

Пирсона, ориентируясь только на подтверждаемые высокие значения доверительных вероятностей. Рассмотрение таких характеристик формы функции распределения, как эксцесс и асимметрия (скошенность) в ряде случаев дает надежные основания для отказа от гипотезы нормального распределения генеральной совокупности в пользу выбора распределения Вейбулла в качестве закона распределения рассматриваемых статистических данных. Кроме того, выбор в пользу распределения Вейбулла во многих случаях можно сделать априорно на основе физических соображений.

3. Высокие доверительные вероятности соответствия нормальному распределению являются недостаточным основанием для определения закона распределения рассматриваемых данных.

Библиографический список

1. Прохоров А.В. Моментов метод // Математическая энциклопедия / гл. ред. И.М. Виноградов. Москва: Сов. энциклопедия, 1982. Т. 3. 1184 с. URL: https://www.nehudlit.ru/books/matematicheskaya-entsiklopediya-tom-3.html.

2. Зыков С.В., Незнанов А.А., Максименкова О.В. Критерии отклонения распределения случайных величин от нормального в математическом обеспечении программных систем поддержки измерений в образовании // Программные системы: теория и приложения. 2018. 9:4(39), С. 199–218. DOI: http://doi.org/10.25209/2079-3316-2018-9-4-199-218.

3. Дёмин С.Е., Дёмина Е.Л. Теория вероятностей. Ч. 3. Системы и функции случайных величин. Случайные процессы: учеб.-метод. пособие / Нижнетагил. технол. ин-т (фил.). Нижний Тагил: НТИ (ф) УрФУ, 2017. 295 с. URL: https://elar.urfu.ru/bitstream/10995/54458/1/978-5-9544-0081-6_2017.pdf.

4. Александровская Л.Н., Кириллин А.В. Рекомендации по применению ряда критериев проверки отклонения распределения вероятностей от нормального закона в практике инженерного статистического анализа // Известия Самарского научного центра РАН, серия «Авиационная и ракетно-космическая техника». 2017. Т. 19, № 1, С. 82–90. URL: http://www.ssc.smr.ru/media/journals/izvestia/2017/2017_1_82_90.pdf; https://www.elibrary.ru/item.asp?id=29409494.

5. Лемешко Б.Ю. Критерии проверки отклонения распределения от нормального закона. Руководство по применению. Москва: НИЦ ИНФРА-М, 2015. 160 с. URL: https://www.elibrary.ru/item.asp?id=23743254.

6. Дуплякин В.М. Особенности идентификации нормального закона распределения // Вестник Самарского университета. Экономика и управление. 2020. Том 11, № 3, С. 176–183. DOI: http://doi.org/10.18287/2542-0461-2020-11-3-176-183.

7. Вентцель Е.С. Теория вероятностей. 11-е изд. стер. Москва: КНОРУС, 2010. 664 с.

8. Вейбулл В. Усталостные испытания и анализ их результатов. Москва: Машиностроение, 1964. 276 с.

9. Митропольский А.К. Техника статистических вычислений. Москва: Главная редакция физико-математической литературы изд-ва «Наука», 1971. 570 с.

10. Янке Е., Эмде Ф., Леш Ф. Специальные функции. Формулы, графики, таблицы / пер. с нем. 2 изд. Москва: Наука, 1968, 344 с. URL: https://bookree.org/reader?file=446837&pg=1.

11. URL: https://msoffice-prowork.com/ref/excel/excelfunc/statistical/gamma.

References

1. Prokhorov A.V. Moment method. In: *Vinogradov I.M. (Ed.) Mathematical encyclopedia*. Moscow: Sov. entsiklopediia, 1982, vol. 3, 1184 p. Available at: https://www.nehudlit.ru/books/matematicheskaya-entsiklopediya-tom-3.html. (In Russ.)

2. Zykov S.V., Neznanov A.A., Maksimenkova O.V. Tests for normality as mathematical support for educational management software. *Program Systems: Theory and Applications*, 2018, vol. 9, issue 4, pp. 199–218. DOI: http://doi.org/10.25209/2079-3316-2018-9-4-199-218. (In Russ.)

3. Demin S.E., Demina E.L. Probability theory. Part 3. Systems and functions of random variables. Random processes: study guide. Nizhny Tagil: NTI (f) UrFU, 2017, 295 p. Available at: https://elar.urfu.ru/bitstream/10995/54458/1/978-5-9544-0081-6_2017.pdf. (In Russ.)

4. Aleksandrovskaya L.N., Kirillin A.V. Recommendations for the use some of tests for the probability distribution of deviation from the normal distribution law in practice of the statistical engineering analysis. *Izvestia of Samara Scientific Center of the Russian Academy of Sciences*, 2017, vol. 19, no. 1, p. 82–90. Available

at: http://www.ssc.smr.ru/media/journals/izvestia/2017/2017_1_82_90.pdf; https://www.elibrary.ru/item.asp?id=29409494. (In Russ.)

5. Lemeshko B.Yu. Tests for checking the deviation from normal distribution law. Guide on the application. Moscow: Research Center INFRA-M, 2015, 160 p. Available at: https://www.elibrary.ru/item.asp?id=23743254. (In Russ.)

6. 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.)

7. Wentzel E.S. Probability theory. 11 edition, stereotyped. Moscow: KNORUS, 2010, 664 p. (In Russ.)

8. Weibull V. Fatigue tests and analysis of their results. Moscow: Mashinostroenie, 1964, 276 p. (In Russ.)

9. Mitropolsky A.K. Technique of statistical calculations. Moscow: Glavnaia redaktsiia fiziko-matematicheskoi literatury izd-va»Nauka», 1971, 570 p. Available at: https://bookree.org/reader?file=448678. (In Russ.)

10. Janke E., Emde F., Lösch F. Special functions. Formulas, graphs, tables, translated from German, 2nd edition. Moscow: Nauka, 1968, 344 p. Available at: https://bookree.org/reader?file=446837&pg=1. (In Russ.)

11. Available at: https://msoffice-prowork.com/ref/excel/excelfunc/statistical/gamma. (In Russ.)