СТАТЬИ

УДК 519.216/.224
DOI 10.17513/snt.39813

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ОБЪЕМА МАССИВА ДАННЫХ НА КЛЮЧЕВЫЕ ВЫБОРОЧНЫЕ ПАРАМЕТРЫ РАЗЛИЧНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

Акимов С.С., Трипкош В.А.

ФГБОУ ВО «Оренбургский государственный университет», Оренбург, e-mail: sergey_akimov_work@mail.ru

В статье рассматривается проблема постановки эксперимента в области идентификации закона распределения. Цель исследования: определить влияние размера выборки на ключевые параметры определенных законов распределения. Для получения необходимых массивов данных использовался генератор случайных чисел программы Mathcad 15. Литературные источники изучались на предмет рекомендаций количества значений массива для применимости методов, связанных с идентификацией закона распределения. Для определения точности при изменении количества данных сгенерированы массивы размером в 10000 данных. В качестве законов распределения рассмотрены наиболее распространенные. Из массивов были взяты подвыборки различного объема. Для полученных подвыборок определялись основные характеристики распределений, которые далее преобразовывались по специальной формуле. Определено, что совокупный размер отклонений конкретного параметра выборки от параметра генеральной совокупности заметно снижается в зависимости от размеров выборки. Для полученных рядов данных подбирались модели регрессии. Регрессионный анализ показал, что в большинстве случаев независимо от вида параметра и закона распределения наибольшей величиной достоверности аппроксимации обладает линейная функция. При этом коэффициенты линейной функции на интервале от 1000 до 10000 данных весьма незначительны. Таким образом, в работе описано влияние количества данных в выборке при определении ключевых параметров закона распределения. Проведен анализ литературных источников, выявлено, что чаще всего исследователи берут 10, 100 и 1000 значений для проведения эксперимента. Исследованы выборки различного объема на разных законах распределения, установлено, что с увеличением объема выборки величина отклонений существенно убывает.

Ключевые слова: объем выборки, размер отклонений, регрессия, закон распределения вероятности

STUDY OF THE INFLUENCE OF DATA VOLUME ON KEY SAMPLE PARAMETERS OF VARIOUS DISTRIBUTION LAWS

Akimov S.S., Tripkosh V.A.

Orenburg State University, Orenburg, e-mail: sergey_akimov_work@mail.ru

The article discusses the problem of setting up an experiment in the field of identifying the distribution law. Purpose of the study: to determine the influence of sample size on the key parameters of certain distribution laws. To obtain the necessary data arrays, a random number generator of the Mathcad 15 program was used. Literary sources were studied for recommendations on the number of array values for the applicability of methods related to the identification of the distribution law. To determine the accuracy when changing the amount of data, arrays of 10,000 data were generated. The most common distribution laws are considered. Subsamples of varying sizes were taken from the arrays. For the obtained subsamples, the main characteristics of the distributions were determined, which were then transformed using a special formula. It has been determined that the total size of deviations of a particular sample parameter from the general population parameter decreases noticeably depending on the sample size. Regression models were fitted for the obtained data series. Regression analysis showed that in most cases, regardless of the type of parameter and distribution law, the linear function has the greatest value of approximation reliability. In this case, the coefficients of the linear function in the interval from 1000 to 10000 data are very insignificant. Thus, the work describes the influence of the amount of data in the sample when determining the key parameters of the distribution law. An analysis of literary sources was carried out and it was revealed that most often researchers take 10, 100 and 1000 values to conduct an experiment. Samples of various sizes were studied using different distribution laws; it was found that with increasing sample size, the magnitude of deviations decreases significantly.

Keywords: sampling, the size of deviations, regression, the law of probability distribution

В настоящий момент проблеме обработки информации в экономике придается весьма большое значение, поскольку обработка представляет собой начальную стадию анализа. Обработка информации независимо от природы ее получения представляет собой целый комплекс специальных процедур с целью получения определенного результата [1].

Одной из особенностей информации является ее стохастическая природа [2]. Данное обстоятельство открывает широкие границы взаимодействия методов экономического анализа в совокупности с теорией вероятности и математической статистикой. Широко известно, что для полноценной характеристики вероятностных данных необ-

ходимо и достаточно знание закона, которому эти данные подчиняются [3].

Однако на практике достаточно распространена ситуация, когда исследователь подобным знанием не обладает, а использует лишь некий массив данных, ничего не зная о природе его распределения [4]. Данные ситуации проявляются достаточно часто в системах, обладающих свойством динамичности или же неопределенности [5, 6]. При этом необходимо отметить, что идентификация закона распределения – достаточно сложная задача, решение которой, учитывая некорректность даже самой постановки, является априори несостоятельным [7]. Потому на практике исследователь, как правило, принимает неизвестный ему массив как нормально распределенный, что в итоге может исказить получаемые результаты [8].

Во избежание подобных искажений имеется необходимость если не идентифицировать, то хотя бы сделать обоснованное предположения о характере распределения, основываясь на самих данных в исследуемом массиве [9]. Известно также, что ключевую роль в идентификации закона распределения играет количество данных, которым располагает исследователь для анализа [10]. Также определено, что количество данных в значительной степени оказывает влияние на результат любого эксперимента, в том числе и идентификацию закона распределения [11, 12]. Отсюда возникает вопрос степени влияния на итоговый результат [13].

Стоит отметить, что, несмотря на широкие исследования в данной области, на настоящий момент не существует строгих универсальных рекомендаций, касающихся необходимого и достаточного количества данных для получения достоверного результата исследования [14]. Существующие на сегодня рекомендации относительно объема распределения носят только достаточно общий характер или же касаются лишь определенных ограниченных критериев [15, 16].

Цель исследования: определить влияние размера выборки на ключевые параметры определенных законов распределения.

Задачи исследования:

– провести анализ литературных источников, посвященных восстановлению закона распределения, с упоминанием количеств данных в выборке;

– исследовать выборки различного объема, подчиняющиеся различным распределениям, на предмет отклонения основных параметров распределения выборки от параметров генеральной совокупности;

– определить наиболее подходящие функции регрессии для различных законов распределения.

Материалы и методы исследования

Исследования проведены на базе кафедры управления и информатики в технических системах Оренбургского государственного университета. Для получения необходимых массивов данных использовался генератор случайных чисел программы Mathcad 15. Часть данных была обработана посредством пакета прикладных программ MS Excel. Данный пакет использовался также для хранения исходных данных и полученных в ходе исследования результатов. Проводился анализ трудов как отечественных, так и зарубежных авторов. Предпочтение отдавалось как наиболее известным авторам, так и последним данным, посвященным анализу данного вопроса.

Литературные источники изучались на предмет рекомендаций количества значений массива для применимости того или иного метода, связанного с идентификацией закона распределения. В ряде работ [17, 18] упоминается применение критерия проверки нормальности распределения Шапиро–Уилка при количестве данных не менее 7. В других работах количество еще выше [19, 20]. В работе [21] указано, что количество значений в массиве при использовании процедуры определения закона распределения (в частности, при проверке нормальности) должно составлять не менее 7. В некоторых других работах [22, 23] также есть ссылка на то, что минимальный размер данных должен быть не ниже 7 исследований в выборке.

Таким образом, определен минимальный уровень количества значений. Работ, в которых бы рассматривалось менее 7 значений для идентификации, в процессе анализа не выявлено. В рамках проведенного литературного обзора рассмотрено более 20 литературных источников.

Однако литературный анализ не является показателем при постановке эксперимента [24]. Необходимо определить, насколько изменяется точность исследования при изменении количества данных. Для этих целей были сгенерированы массивы размером в 10000 данных для различных распределений. В качестве законов распределения рассмотрим наиболее распространенные из них: нормальный, экспоненциальный, равномерный, логнормальный, логистический, биноминальный, геометрический, гипергеометрический, распределения Рэлея, Коши, Пуассона.

Из массивов вновь при помощи генератора случайных чисел были взяты подвыборки различного объема. Объем подвыборок определялся исходя из общих рекомендаций, взятых из анализа литературных источников.

Для полученных подвыборок определялись основные характеристики распределений (среднее, стандартное отклонение и дисперсия для нормального, минимальное и максимальное значения, медиана для равномерного, интенсивность для экспоненциального распределения и т.д.), которые затем подвергались преобразованию для дальнейшей обработки по следующей формуле:

$$\Delta = |Pg - Ps|. \quad (1)$$

где Pg – параметр генеральной совокупности, Ps – параметр выборки.

Результаты исследования и их обсуждение

Объединяя данные, полученные из всех исследуемых литературных источников, необходимо составить общее представление о том количестве значений, которым оперируют исследователи в процессе проведения своих работ. Для этого составим рейтинг количества значений в исследуемых выборках, которые используются при постановке эксперимента и получении экспериментальных результатов (рис. 1).

Как показывают данные рисунка, наиболее популярными в исследованиях являются числа 10, 100 и 1000.

В результате определено, что совокупный размер отклонений конкретного параметра выборки от параметра генеральной совокупности заметно снижается в зависимости от размеров выборки, причем на данную выявленную тенденцию существенного влияния не оказывает ни вид закона распределения, ни конкретный параметр, а только размер совокупности изучаемых данных.

Для полученных рядов данных подбирались модели регрессии. Однако ни одна из распространенных функций (логарифмическая, экспоненциальная, линейная, степенная, полиномиальная) не давала значимой величины аппроксимации (максимальная – у логарифмической функции, равная 0,056).

Исходя из данного обстоятельства, исследуемые массивы были разбиты на участки. После перебора нескольких вариантов наиболее оптимальными оказались следующие интервалы: 7–100, 100–1000, 1000–10000. В качестве примера отобразим отклонение среднего значения нормального распределения (рис. 2–4).

Проведенный регрессионный анализ методом наименьших квадратов показал, что в абсолютном большинстве случаев независимо от вида параметра и закона распределения наибольшей величиной достоверности аппроксимации обладает линейная функция. Все остальные аппроксимирующие уравнения имеют весьма низкую точность и уровень достоверности. При этом коэффициенты линейной функции на интервале от 1000 до 10000 данных весьма незначительны, что говорит о слабом изменении параметра при исследовании свыше 1000 значений.

Рис. 1. Рейтинг количества значений, используемых в различных исследованиях

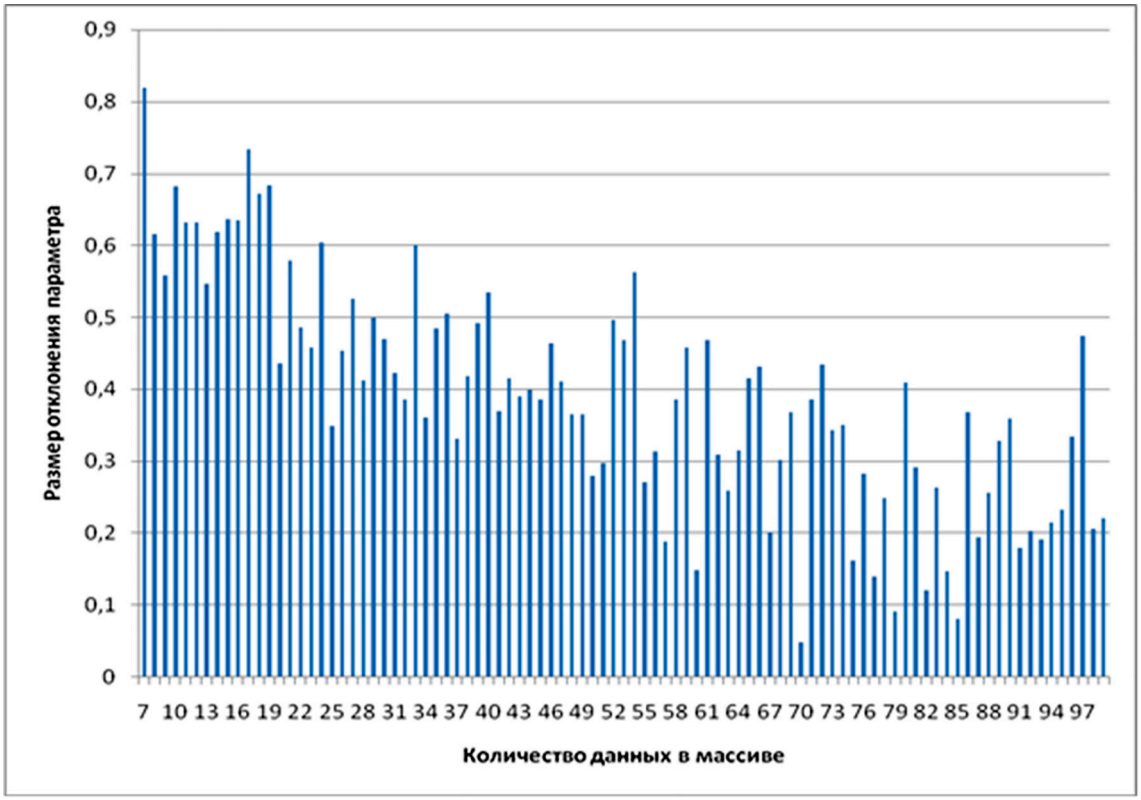

Рис. 2. Размер отклонений выборочного среднего от среднего значения генеральной совокупности на интервале 7–100

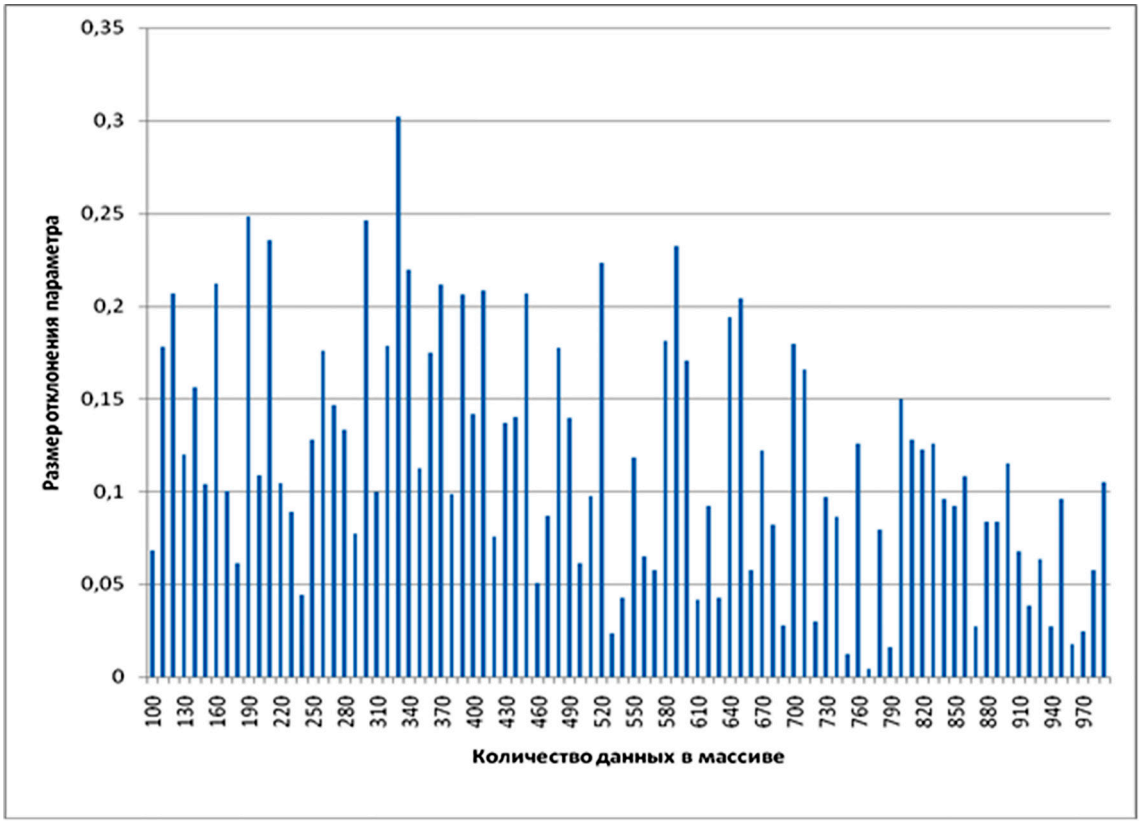

Рис. 3. Размер отклонений выборочного среднего от среднего значения генеральной совокупности на интервале 100–1000

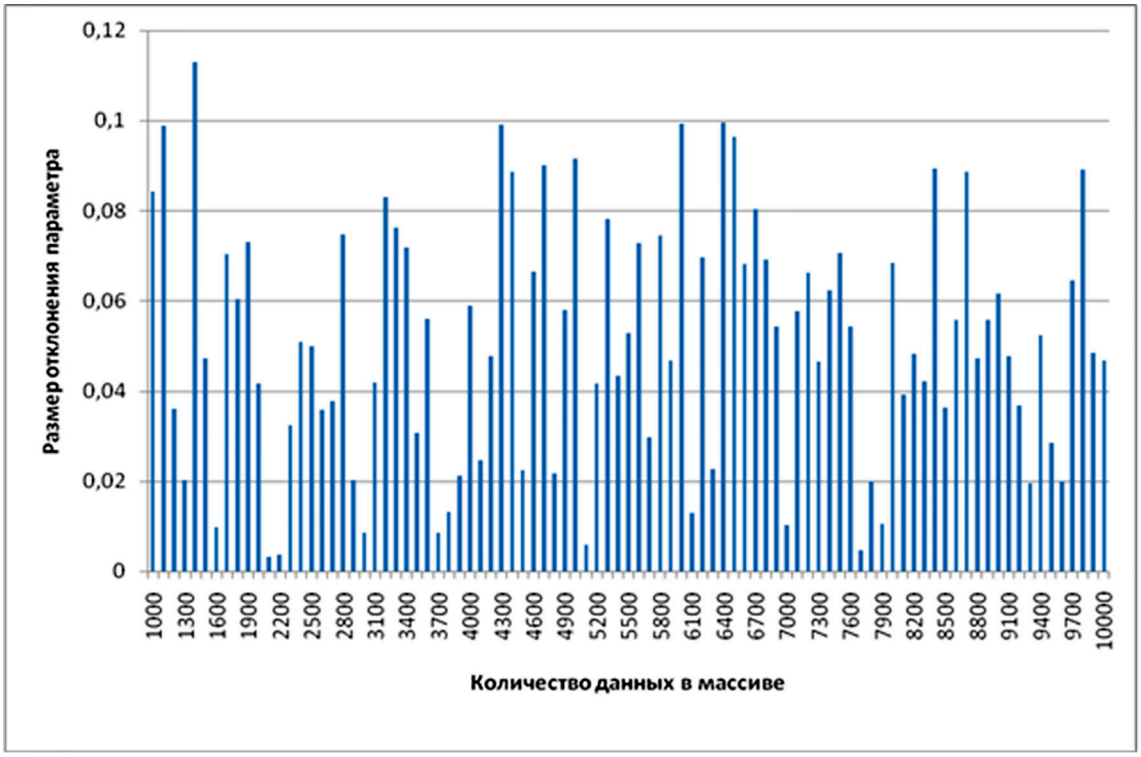

Рис. 4. Размер отклонений выборочного среднего от среднего значения генеральной совокупности на интервале 1000–10000

Заключение

В работе описано влияние количество данных в выборке при определении ключевых параметров закона распределения. Проведен анализ литературных источников, выявлено, что чаще всего исследователи берут 10, 100 и 1000 значений для проведения эксперимента.

Исследованы выборки различного объема на разных законах распределения, установлено, что с увеличением объема выборки величина отклонений достаточно существенно убывает.

Проведен регрессионный анализ, определено, что наиболее подходящим способом описания изменения данных в зависимости от количества является разбиение общего интервала на участки до 100 значений, 100–1000 значений, свыше 1000 значений, которые затем аппроксимируются линейной функцией, независимо от исследуемого параметра и закона распределения, которому подчиняется исследуемый массив.

Список литературы

1. Ji K., Chen T., Li B. et al. Study on Gas Diffusion Distribution Law of Inverted Oil Immersed Current Transformer // High Voltage Apparatus. 2021. Vol. 57, № 11. P. 164-170.

2. Nagaev S., Chebotarev V. On approximation of the tails of the binomial distribution with these of the poisson law // Mathematics. 2021. Vol. 9, №. 8. DOI: 10.3390/math9080845.

3. Орлов А.И. Статистика интервальных данных // Заводская лаборатория. Диагностика материалов. 2015. Т. 81, № 3. С. 61-69.

4. Шепель В.Н., Акимов С.С. Эвристическая процедура определения подходящего распределения вероятности // Компьютерная интеграция производства и ИПИ-технологии: V Всероссийская научно-практическая конференция с элементами научной школы-семинара молодых ученых и специалистов, посвященная 50-летию механического факультета Аэрокосмического института ОГУ. Оренбург: ОГУ, 2011. С. 137-139.

5. Umemoto D., Ito N. Power-law distribution in an urban traffic flow simulation // Journal of Computational Social Science. 2018. Vol. 1, № 2. P. 493-500.

6. Акимов С.С. Оптимизированный алгоритм определения закона распределения вероятности по выборке из генеральной совокупности // Известия Самарской государственной сельскохозяйственной академии. 2013. № 2. С. 52-56.

7. Vozhov S.S. Parametric and Nonparametric Identification of the Distribution Law from Interval Data // Measurement Techniques. 2018. Vol. 61, № 3. P. 216-222.

8. Акимов С.С. Использование коэффициентов асимметрии и эксцесса при гистограммном методе определения закона распределения вероятности // Известия Оренбургского государственного аграрного университета. 2014. № 1(45). С. 225-227.

9. Catillo M., Glozman L.Y. Distribution law of the Dirac eigenmodes in QCD // International Journal of Modern Physics A. 2018. Vol. 33, № 10. P. 1850054.

10. Шепель В.Н., Акимов С.С. Использование оценки Хилла для различения законов распределения вероятности //

Вестник Оренбургского государственного университета. 2014. № 1(162). С. 75-78.

11. Кареев И.А. Нижние границы для среднего объема выборки и эффективность последовательных процедур упорядочивания // Теория вероятностей и ее применения. 2013. Т. 58, № 3. С. 591-597.

12. Попов А.А. Некоторые проблемы применения статистического метода в ходе выборочного аудита // Экономические науки. 2009. № 1. С. 317-320.

13. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука. Главная редакция физико-математическом литературы, 1983. 416 с.

14. Соловьев И.А. Модифицированный закон нормального распределения // Математические методы в технике и технологиях – ММТТ. 2020. Т. 2. С. 3-8.

15. Филатов В.И., Борукаева А.О. Выборочный метод и параметры закона распределения выборки // Инновационное развитие. 2019. № 3(30). С. 33-36.

16. Шерматов Н., Одинаев Р.Н. Законы распределения и характеристики // Вестник Таджикского национального университета. 2019. № 2. С. 33-40.

17. Лаптева А.С. Нормальный закон распределения // Академия педагогических идей Новация. 2019. № 1. С. 192-194.

18. Слепов Н.А. Скорость сходимости распределений геометрических сумм к закону Лапласа // Теория вероятностей и ее применения. 2021. Т. 66, № 1. С. 149-174.

19. Warusawitharana M. Time-varying volatility and the power law distribution of stock returns // Journal of Empirical Finance. 2018. Vol. 49. P. 123-141.

20. Григорьев Ю.Д. Планы эксперимента для моделей регрессии типа сплайнов // Заводская лаборатория. Диагностика материалов. 2013. Т. 79, № 11. С. 60-66.

21. Акимов С.С. Расчет вероятности дискретности для массива данных // Научное обозрение. 2013. № 6. С. 78-83.

22. Balthrop A., Quan S. The power-law distribution of cumulative coal production // Physica A: Statistical Mechanics and its Applications. 2019. Vol. 530. P. 121573.

23. Шепель В.Н., Акимов С.С. Модернизация метода гистограмм для выявления принадлежности неизвестного массива данных определенному закону распределения вероятностей // Вестник Оренбургского государственного университета. 2014. № 9(170). С. 179-181.

24. Акимов С.С. Оценка Хилла как ключевая оценка для распознавания тяжело- и легкохвостовых законов распределения вероятности // Научное обозрение. 2014. № 10-2. С. 349-352.