УДК 621.9 DOI: 10.15222/TKEA2017.3.49

К. т. н. А. Н. ТЫНЫНЫКА

Украина, Одесский национальный политехнический университет
E-mail: nikal1091@gmail.com

ПРИМЕНЕНИЕ ЭНТРОПИЙНОГО КОЭФФИЦИЕНТА ДЛЯ ОПТИМИЗАЦИИ ЧИСЛА ИНТЕРВАЛОВ ПРИ ИНТЕРВАЛЬНЫХ ОЦЕНКАХ

Показано, что в качестве критерия выбора числа интервалов группирования опытных данных при интервальных оценках может использоваться энтропийный коэффициент. В соответствии с описанной процедурой быстрого определения числа интервалов на массиве данных исследована точность имеющихся в литературе и предложенных новых формул. Проведен анализ в сравнении с ранее опубликованными результатами применения для этих целей критерия согласия Пирсона. Сделаны расчеты с целью сравнения эффективности применения одних и тех же формул при распределении выборочных данных по нормальному закону и по закону Рэлея.

Ключевые слова: энтропийный коэффициент, число интервалов группирования, интервальные оценки, распределение Рэлея, нормальное распределение.

В процесс управления качеством продукции промышленного предприятия входит сбор экспериментальных данных с последующей их обработкой. В процедуру первичной обработки результатов эксперимента чаще всего включают сопоставление гипотез о законе распределения данных, описывающем с наименьшей погрешностью случайную величину по наблюдаемой выборке. В этом случае выборка представляется в виде гистограммы, состоящей из k столбцов, построенных на интервалах протяженностью d. Идентификации формы распределения результатов (или погрешностей) измерений требует также ряд задач, эффективность решения которых различна для различных распределений (например, использование метода наименьших квадратов или вычисление оценок энтропии). Идентификация распределений нужна еще и потому, что рассеяние всех оценок (среднего квадратического отклонения, эксцесса, контрэксцесса и др.) опять-таки зависит от формы закона распределения, как показано в исследованиях, на которые принято ссылаться как на классические [1–4].

От объема выборки зависит успешность идентификации формы распределения экспериментальных данных, и если он мал, особенности распределения оказываются замаскированными случайностью самой выборки. В такой ситуации важно наилучшим образом распределить выборочные данные по интервалам, когда для дальнейшего анализа и расчетов интервальный ряд необходим. И тут сразу же возникает вопрос о назначении числа интервалов k, потому что от этого зависит успешность идентификации. А. Хальд в книге [1] пространно убеждает, что существует оптимальное число интервалов группирования, когда ступенчатая огибающая гистограммы, построенной на этих интервалах, наиболее близка к плавной кривой распределения генеральной совокупности. Одним из практических признаков приближения к оптимуму может служить исчезновение в гистограмме провалов, и тогда близким к оптимуму считается наибольшее k, при котором гистограмма еще сохраняет плавный характер.

Очевидно, что вид гистограммы зависит от построения интервалов принадлежности случайной величины. Однако даже в случае равномерного разбиения до сих пор нет удовлетворительного способа такого построения. Разбиение, которое можно было бы считать правильным, приводит к тому, что ошибка аппроксимации кусочно-постоянной функцией предположительно непрерывной плотности распределения (гистограммой) будет минимальной. Затруднения тут вызваны тем, что оцениваемая плотность неизвестна, поэтому число интервалов сильно сказывается на виде распределения частот конечной выборки. При фиксированной ее длине укрупнение интервалов разбиения ведет не только к уточнению эмпирической вероятности попадания в них, но и к неизбежной потере информации (как в общем широком смысле, так и в смыс-

ле кривой распределения плотности вероятности), поэтому при дальнейшем необоснованном укрупнении слишком сильно сглаживается изучаемое распределение.

Однажды возникнув, задача оптимального разбиения размаха под гистограмму не исчезает из поля зрения специалистов, и пока не появится единственное устоявшееся мнение относительно ее решения, задача будет оставаться актуальной. Решения время от времени предлагаются — либо эмпирические (откровенно или завуалированно), либо авторы сильно упрощают задачу, считая априори известным закон распределения вероятностей. Иногда рекомендации имеют произвольно-директивный характер типа «число интервалов не должно выходить за пределы 6…20», при этом игнорируется даже то очевидное обстоятельство, что названный диапазон слишком широк и делать выбор нужно в нем.

Первым, вероятно, был Старджес, который еще в 1926 году в [2] рассмотрел идеализированную гистограмму из k интервалов, где i-е значение было равно биномиальному коэффициенту: $k_i = C_{k-1}^{i-1}$. Если считать это правомерным, то дальше на основании формальных преобразований можно записать для суммы коэффициентов (групповых частот) следующее:

$$\sum_{1}^{k} C_{k-1}^{i-1} = \sum_{1}^{k} k_i = 2^{k-1} = n,$$

откуда

$$k = 1 + 3{,}3\lg n, \qquad (1)$$

где n — объем выборки.

В таком виде формула попала практически во все учебники по статистике. При этом она статистически не обоснована, но ее, пожалуй, все же можно считать полуэмпирической, а не полностью подобранной.

В 1942 г. Манн и Вальд ушли от логарифмической зависимости и дали оценку оптимального числа интервалов в виде степенной функции [3]

$$k = 4\cdot[0{,}75(n-1)^2]^{1/5}. \qquad (2)$$

В 1950 г. Н. В. Смирнов показал, что отклонение гистограммы от неизвестного графика плотности убывает с увеличением выборки по закону $1/n^{1/3}$ [4].

В [5] Д. Скотт для оценки длины интервала гистограммы минимизировал среднеквадратическую ошибку и получил для случая дифференцируемой плотности асимптотическую оценку оптимальной длины интервала

$$d = \sqrt[3]{\frac{6}{n\int\limits_{+\infty}^{-\infty}\left[f'(x)\right]^2 dx}},$$

т. е. здесь число интервалов пропорционально $n^{1/3}$.

Для нормального распределения

$$d \approx \frac{3{,}5\sigma}{\sqrt[3]{n}},$$

где σ — среднеквадратическое отклонение.

Следует отметить, что часто эту формулу применяют для первоначальной оценки длины интервалов при неизвестном законе распределения.

В случае простого линейного распределения $f(x) = 2x$ оптимальная длина интервала

$$d = \sqrt[3]{\frac{3}{2n}},$$

и если функция $f(x)$ задана на отрезке [0; 1], получим

$$k = \sqrt[3]{\frac{2n}{3}}. \qquad (3)$$

В [6, с. 51] была приведена оценка числа интервалов оптимального разбиения для аппроксимации дважды дифференцируемой плотности $f(x)$:

$$k = \sqrt[5]{\frac{n\left|f''(x)\right|_{\max}}{4\left|f(x)\right|_{\max}}}. \qquad (4)$$

Формулы (3) и (4) нельзя применять к равномерному распределению и к плоским трапециевидным распределениям.

Еще одна формула без каких-либо пояснений и доказательств приводится в [7, с. 178; 8, с. 81] (в [7] — со ссылкой на автореферат кандидатской диссертации И. У. Алексеевой, 1975 г.):

$$k = \frac{1}{6}\left(1+\chi\right)\sqrt[5]{n^2}, \qquad (5)$$

где χ — эксцесс распределения.

В [7] собрано несколько эвристических формул нахождения числа интервалов в зависимости от выборки, предложенных разными авторами. Но поскольку при значительных объемах выборок разброс значений интервалов, задаваемых различными формулами, довольно велик, в [9] была поставлена задача выяснить, какая из имеющихся формул наилучшая. Предполагалось, что генеральная совокупность экспериментальных данных, из которой взята исследуемая выборка, имеет гладкую кривую распределения, чтобы можно было считать, что появляющиеся при группировании провалы и всплески на отдельных интервалах порождаются случайностью попадания измеренных значений в малую

выборку. На заводе из 500 заготовок, к которым при изготовлении предъявлялось требование по массе $m = 17^{06}_{-04}$ кг, было отобрано 80 и измерена их масса. Число интервалов определялось по шести формулам, строились равноинтервальные ряды, на их основе — гистограммы и делалось заключение о наиболее точной формуле. В качестве критерия использовался критерий согласия Пирсона.

Критерий Пирсона, как известно, требует разбиения выборки на интервалы — именно в них производится оценка отличия между принятой моделью и сравниваемой выборкой. Однако применение этого критерия в случае интервалов постоянной длины, используемых обычно для построения гистограмм, неэффективно. Поэтому в работах по эффективности критерия Пирсона рассматриваются интервалы не с равной длиной, а с равной вероятностью в соответствии с принимаемой моделью. При этом, однако, число интервалов равной длины и число интервалов равной вероятности различаются в разы (за исключением равновероятного распределения), что позволяет сомневаться в достоверности полученных в [9] результатов.

В настоящей работе с использованием энтропийного коэффициента в качестве критерия близости исследована правомочность применения имеющихся в литературе и предложенных автором формул, предназначенных для определения оптимального числа интервалов группирования экспериментальных данных, а также их эффективность при распределении плотности вероятности, отличном от нормального.

Исследуемые формулы и методика проведения их оценки

Оценим одиннадцать выражений — шесть, которые были рассмотрены в [9], и пять, предложенных в настоящей работе. Кроме приведенной выше формулы Старджеса (1), это следующие:

— Брукса и Каррузера

$$k = 5\lg n; \quad (6)$$

— И. Хайнхольда и К. Гаеде

$$k = n^{1/2}; \quad (7)$$

— З. Таушанова

$$k = 4\lg n; \quad (8)$$

— Е. Тоневой

$$k = 5\lg(0{,}1n); \quad (9)$$

— К. Уильямса

$$k = 2[0{,}85(n-1)]^{0{,}4} - 1; \quad (10)$$

— $k = 5\lg n - 1$; (11)

— $k = 10(\lg n - 1)$ (12)

— $k = 2n^{1/3}$; (13)

— $k = 6\lg(0{,}1n) + 6$; (14)

— $k = 2(2n)^{1/3}$. (15)

Для выбора наиболее точной формулы воспользуемся энтропийным коэффициентом $k_э$, предложенным в качестве числовой характеристики формы распределения в [10]. По гистограмме он вычисляется как

$$k_э = \frac{dn}{2\sigma}10^{\beta}, \quad (16)$$

$$\beta = -\frac{1}{n}\sum_{j=1}^{k}\left(n_j \lg n_j\right),$$

где n_j — число наблюдений в j-м интервале, $j = 0, \ldots, k$.

Процедура, по которой будем проводить оценку точности формул, следующая.

1) Из исходной, большой, выборки путем удаления всех нечетных членов сформируем меньшую выборку.

2) Найдем значения числа интервалов по меньшей выборке в соответствии с приведенными формулами для расчета k.

3) Определим диапазон значений k, в котором будут проводиться расчеты энтропийного коэффициента $k_э$.

4) Вычислим энтропийный коэффициент по большой выборке для максимального значения k и примем его за эталонный ($k_{ээ}$).

5) Вычислим энтропийный коэффициент по меньшей выборке для всех, кроме максимального, значений k ($k_{эk}$).

6) Путем сравнения полученных данных установим, при каком числе интервалов значение $k_{эk}$ будет наиболее близким к эталонному, а затем, какие формулы дали такое же число интервалов — именно они будут считаться самыми точными.

Проверка формул на массиве данных, распределенных по закону Гаусса

Для корректного сравнения результатов наших исследований с полученными в [9] будем рассматривать экспериментальную выборку с такими же, как и в [9], параметрами — объем $n = 80$, размах $R = 0{,}98$ кг, среднеквадратическое отклонение $\sigma = 0{,}238$ кг. Тогда, в соответствии с описанной выше процедурой, сформируем меньшую выборку объемом 40 и рассчитаем по исследуемым формулам значения числа интервалов k.

Как видно из **табл. 1**, значения k лежат в диапазоне 3—10. Поскольку задавать $k = 3$ нерационально, для дальнейших расчетов будем рассматривать значения k в диапазоне от 4 до 12.

При разбиении исходной выборки на 12 равных частей ($k = 12$) получаем по формуле (16) эталонный энтропийный коэффициент $k_{ээ} = 1{,}89$.

ISSN 2225-5818

Таблица 1

Результаты расчета k по малой выборке для двух видов распределения

Номер формулы	Число интервалов k при распределении	
	Гаусса	Рэлея
(1)	6	7
(6)	8	8
(7)	6	7
(8)	6	7
(9)	3	4
(10)	7	8
(11)	7	8
(12)	6	7
(13)	7	7
(14)	10	10
(15)	9	9

Результаты расчета $k_{эk}$ при числе интервалов от 4 до 11 приведены в **табл. 2**. Сравнение этих значений с эталонным показывает, что наилучшая гистограмма будет построена на семи интервалах. При этом, как видно из табл. 1, значение $k = 7$ дают формула Уильямса (10) и предложенные в данной работе формулы (11) и (13). Одинаково ухудшается результат при разбиении на 6 и 8 интервалов, при этом $k = 6$ получается по формулам (1), (7), (8), (12), а $k = 8$ — по формуле (6). Еще хуже будет гистограмма на 10 интервалах (формула (14)) и на 9 (формула (15)).

Таким образом, при использовании энтропийного коэффициента формулы, рассмотренные в [9], по убыванию точности располагаются следующим образом: (10), (1), (6), (7), (8), (9). Последовательность же, полученная в [9], другая, а именно: (7), (1), (10), (8), (6), (9).

Проверка формул на массиве данных, распределенных по закону Рэлея

Проверим, изменится ли полученный рейтинг формул для распределения, отличного от нормального. Для оценки выберем закон Рэлея $f(x) = x \cdot \exp(-x^2/(2\sigma^2))/\sigma^2$.

Чтобы получить исследуемую выборку объемом 100, с помощью программы Mathworks Matlab были сгенерированы 100 чисел, распределенных по этому закону в диапазоне от 0 до 100 (с точностью пяти знаков после запятой), и округлены до целых для внесения помехи. Размах полученного вариационного ряда составил 98. Соответственно, объем меньшей выборки составляет 50. Рассчитанные при $n = 50$ значения числа интервалов k также приведены в табл. 1.

В данном случае разобъем большую выборку на 14 равных интервалов и получим $k_{ээ} = 2{,}23$.

Определим значения $k_э$ по малой выборке для числа интервалов 4 — 10. Из результатов, представленных в табл. 2, видно, что самый близкий к $k_{ээ}$ энтропийный коэффициент — $k_{э9}$, а не $k_{э7}$, как в предыдущем случае. Причиной сдвига оптимального числа интервалов в большую сторону, очевидно, является увеличение на 25% объема выборки, а также изменение закона распределения выборочных данных.

Как можно видеть из табл. 2, по убыванию точности формулы теперь расположились в такой последовательности: (15), (6), (10), (11), (1), (7), (8), (12), (13), (14), (9).

Таким образом, формулы, которые были лучшими при исследовании нормальной выборки, в

Таблица 2

Результаты расчета по малой выборке для двух видов распределения при различных заданных значениях k

k	Распределение Гаусса		Распределение Рэлея	
	$k_{эk}$	Формула*	$k_{эk}$	Формула*
4	$k_{э4} = 1{,}51$		$k_{э4} = 1{,}66$	(9)
5	$k_{э5} = 1{,}81$		$k_{э5} = 1{,}72$	
6	$k_{э6} = 1{,}82$	(1); (7); (8); (12)	$k_{э6} = 1{,}73$	
7	$k_{э7} = 1{,}85$	(10); (11); (13)	$k_{э7} = 1{,}74$	(1); (7); (8); (12); (13)
8	$k_{э8} = 1{,}82$	(6)	$k_{э8} = 1{,}75$	(6); (10); (11)
9	$k_{э9} = 1{,}77$	(15)	$k_{э9} = 1{,}95$	(15)
10	$k_{э10} = 1{,}8$	(14)	$k_{э10} = 1{,}71$	(14)
11	$k_{э11} = 1{,}78$			

*Номера формул, при расчете по которым по меньшей выборке получается число интервалов, указанное в графе «k»

случае распределения выборочных данных по закону Рэлея показали хорошие, но все же не лучшие результаты, что подтверждает зависимость оптимального числа интервалов от вида закона распределения экспериментальных данных.

Заключение

В результате проведенных расчетов были обнаружены различия в точности рассмотренных формул при определении числа интервалов для построения наилучшей гистограммы при интервальных оценках по критериям Пирсона и при использовании энтропийного коэффициента $k_э$. Во втором случае лучшей из шести исследованных в [9] была определена формула Брукса и Каррузера (6) — и для распределения Гаусса, и для распределения Рэлея, тогда как в [9] она была предпоследней в рейтинге.

Таким образом, вычисления, проведенные в диапазоне распространенных на производстве малых выборок, подтвердили эффективность использования энтропийного коэффициента вместо критерия Пирсона для выбора числа интервалов в случае построения равноинтервальной гистограммы по экспериментальным данным.

Далее интересным представляется проверить возможность использования других характеристик распределения случайных величин, например эксцесса и контрэксцесса, и сравнить точность и удобство расчетов с применением энтропийного коэффициента.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Хальд А. Математическая статистика с техническими приложениями. — Москва: Изд-во иностр. лит., 1956.

2. Sturges H. A. The choice of a class interval // JASA. — 1926. — V. 21. — С. 65—66.

3. Mann H. B., Wald A. On the choice of the number of intervals in the application of the chi-square test // Ann. Math. Statist. — 1942. — Vol. 13. — С. 478—479.

4. Смирнов Н. В. О построении доверительной области для плотности распределения случайной величины // Доклады АН СССР. — 1950. — Т. 74, № 2. — С. 189—192.

5. Scott D. W. On optimal and data-based histograms // Biometrika. — 1979. — Vol. 66. — С. 605—610.

6. Лившиц М. Е., Иванов-Муромский К. А., Заславский С. Я. и др. Численные методы анализа случайных процессов. — Москва: Наука, 1976.

7. Новицкий П. В., Зограф И. А. Оценка погрешностей результатов измерений. — Ленинград: Энергоатомиздат, 1991.

8. Булашев С. В. Статистика для трейдеров. — Москва: Компания Спутник+, 2003.

9. Калмыков В. В., Антонюк Ф. И., Зенкин Н. В. Определение оптимального количества классов группирования экспериментальных данных при интервальных оценках // Южносибирский научный вестник. — 2014. — №3. — С. 56—58.

10. Новицкий П. В. Понятие энтропийного значения погрешности // Измерительная техника. — 1966. — № 7. — С. 11—14.

Дата поступления рукописи в редакцию 03.05 2017 г.

О. М. ТИНИНИКА

Україна, Одеський національний політехнічний університет
E-mail: nikal1091@gmail.com

ЗАСТОСУВАННЯ ЕНТРОПІЙНОГО КОЕФІЦІЄНТА ДЛЯ ОПТИМІЗАЦІЇ ЧИСЛА ІНТЕРВАЛІВ ПРИ ІНТЕРВАЛЬНИХ ОЦІНКАХ

Показано, що як критерій вибору числа інтервалів групування досліджених даних при інтервальних оцінках можна використовувати ентропійний коефіцієнт. Відповідно до описаної процедури швидкого визначення числа інтервалів на масиві даних досліджено точність наявних в літературі і запропонованих нових формул. Проведено аналіз в порівнянні з раніше опублікованими результатами застосування для цих цілей критерію згоди Пірсона. Зроблено розрахунки з метою порівняння ефективності застосування одних і тих самих формул при розподілі вибіркових даних за нормальним законом і за законом Релея.

Ключові слова: ентропійний коефіцієнт, число інтервалів групування, інтервальні оцінки, розподіл Релея.

DOI: 10.15222/TKEA2017.3.49
UDC 621.9

A. N. TYNYNYKA

Ukraine, Odessa National Polytechnic University
E-mail: nikal1091@gmail.com

APPLICATION OF THE ENTROPIC COEFFICIENT FOR INTERVAL NUMBER OPTIMIZATION DURING INTERVAL ASSESSMENT

In solving many statistical problems, the most precise choice of the distribution law of a random variable is required, the sample of which the authors observe. This choice requires the construction of an interval series.

Therefore, the problem arises of assigning an optimal number of intervals, and this study proposes a number of formulas for solving it. Which of these formulas solves the problem more accurately?

In [9], this question is investigated using the Pearson criterion. This article describes the procedure and on its basis gives formulas available in literature and proposed new formulas using the entropy coefficient. A comparison is made with the previously published results of applying Pearson's concord criterion for these purposes. Differences in the estimates of the accuracy of the formulas are found. The proposed new formulas for calculating the number of intervals showed the best results.

Calculations have been made to compare the work of the same formulas for the distribution of sample data according to the normal law and the Rayleigh law.

Keywords: entropy coefficient, grouping intervals number, interval estimates, Rayleigh distribution.

REFERENCES

1. Hald A. *Matematicheskaya statistika s tekhnicheskimi prilozheniyami* [Mathematical statistics with technical applications]. Moskow, Izd-vo inostr. lit, 1956.

2. Sturges H. A. The choice of a class interval. *JASA*, 1926, vol. 21, pp. 65-66.

3. Mann H. B., Wald A. On the choice of the number of intervals in the application of the chi-square test. *Ann. Math. Statist*, 1942, vol. 13, pp. 478-479.

4. Smirnov N. V. [On the construction of a confidence domain for the distribution density of a random variable]. *Doklady Akademii Nauk SSSR*, 1950, vol. 74, no 2, pp. 189-192 (Rus)

5. Scott D. W. On optimal and data-based histograms. *Biometrika*, 1979, vol. 66, pp. 605-610.

6. Livshits M. E., Ivanov-Muromsky K. A., Zaslavsky S. Ya., Voitinsky E. Ya., Lerner V. F, Romm B. I. *Chislovye metody analiza sluchainyh protsesov* [Numerical methods of analysis of random processes]. Moscow, Nauka, 1976. (Rus)

7. Novitsky P. V., Zograf I. A. *Otcenka pogreshnostei resultatov ismerenii* [Estimation of errors in measurement results]. Leningrad, Energoatomizdat, 1991. (Rus)

8. Bulashev S. V. *Statistika dlya treiderov* [Statistics for traders]. Moscow, Sputnik company+, 2003. (Rus)

9. Kalmykov V. V., Antonyuk F. I., Zenkin N. V. [Determination of the optimal number of classes of grouping of experimental data for interval estimates]. *Yuzhnosibirskii nauchnyi vestneyk,* 2014, no 3, pp. 56-58. (Rus)

10. Novitsky, P. V. [The concept of the entropy value of error]. *Izmeritel`naya tekhnika,* 1966, no 7, pp. 11-14. (Rus)

РЕЦЕНЗЕНТЫ НОМЕРА