УДК 53.08:51

Л.П.ВЛАСЬЕВСКИЙ
Иркутский государственный технический университет

ОПТИМИЗАЦИЯ КОЛИЧЕСТВА КЛАССОВ ПРИ ОЦЕНКЕ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ СВОЙСТВ ПРИРОДНЫХ ОБЪЕКТОВ

Сделан обзор подходов к определению числа интервалов при построении гистограмм. Показано противоречие между точностью и детальностью отображения кривых распределения вероятностей свойств природных объектов. Предложено разрешающее это противоречие новое аналитическое выражение оптимального числа классов.

The review of the approaches to definition of class intervals quantity at histograms construction is made. The contradiction between accuracy and detail of display of curve distribution of probabilities of natural objects properties is shown. Permitting this contradiction new analytical expression of optimum number of classes is offered.

Экспериментальные данные, получаемые в процессе определения свойств природных объектов, представлены, как правило, набором результатов измерений, обладающих стохастическими свойствами. Так, в области изучения месторождений полезных ископаемых типичными примерами являются значения содержания полезного компонента, мощности рудных тел, оценка угловых и линейных параметров трещиноватости горных пород. При статистической обработке большеобъемных выборок из совокупности таких случайных величин последние разносятся на ряд диапазонов (классов) для удобства визуализации, например, для построения графика распределения вероятностей в виде гистограммы, а также для вычисления таких параметров, как среднее значение, дисперсия и т.д. В связи с этим возникает задача определения необходимых интервалов группирования и, как следствие, общего их количества. Как указывают П.В.Новицкий и И.А.Зограф [1], выбор числа интервалов при построении гистограммы – это задача оптимальной фильтрации, а оптимальным числом интервалов K является такое, когда максимальное возможное сглаживание случайных флуктуаций данных сочетается с минимальным искажением от сглаживания самой кривой искомого распределения, т.е. на фоне случайных колебаний необходимо выделение основных, характерных черт распределения вероятностей [2].

Из рекомендованных формул по определению необходимого количества классовых интервалов наиболее популярна формула Старджеса, однако она не основывается на оптимизации процесса фильтрации графика распределения, так как выводится на основании свойства суммы некоторого количества биномиальных коэффициентов [3], в результате чего выражение для расчета числа интервалов K имеет вид

$$K = 3{,}32 \lg n + 1, \qquad (1)$$

где n – число обрабатываемых наблюдений.

Другой подход к выбору числа интервалов базируется на использовании критерия согласия χ^2, причем, согласно работе [1], оптимальное число интервалов с равной длиной существенно отличается от числа интервалов с равной вероятностью. В наиболее общем виде

$$K = 1{,}9 n^{0{,}4}. \qquad (2)$$

Нам представляется достаточно корректным рекомендовать следующие подходы для выражения оптимального количества классовых интервалов.

При разбиении группы статистических данных общим числом N на K интервалов по

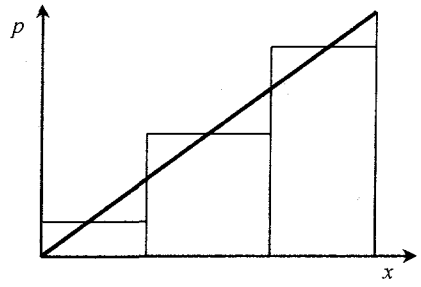


Рис.1. Аппроксимация треугольного распределения ступенчатой гистограммой

значениям исследуемой величины их число в отдельном классе может быть описано биномиальным распределением с вероятностью попадания отдельного значения в интервал $p = n/N$, где n – среднее число значений в классе, $n = N/K$. Тогда $p = 1/K$.

Дисперсия числа значений в интервале

$$D_n = N(p - p^2),$$

или дисперсия вероятности p в интервале

$$D_p = \frac{1}{N^2} N(p - p^2) = \frac{1}{N}\left(\frac{1}{K} - \frac{1}{K^2}\right).$$

Переходя к плотности вероятности $f = p/x$, найдем ее дисперсию:

$$D_f = \frac{D_p}{x^2} = \frac{K^2}{X^2}\frac{1}{N}\left(\frac{1}{K} - \frac{1}{K^2}\right) = \frac{1}{X^2}\frac{K-1}{N}. \quad (3)$$

Таким образом, чем большее число интервалов используется, тем выше дисперсия плотности вероятности попадания значения в интервал. С другой стороны, если распределение вероятностей отличается от равномерного и имеет, например, треугольную форму (рис.1), то при разнесении данных на интервалы ступенчатая гистограмма неточно отображает форму распределения вероятностей. В этом случае на одном из K интервалов вероятность изменяется в диапазоне $\Delta p = 2/K^2$ и имеет дисперсию

$$D_p = \frac{(\Delta p)^2}{12} = \frac{1}{3K^4},$$

а функция плотности вероятности, в свою очередь, характеризуется дисперсией

$$D_f = \frac{K^2}{X^2}\frac{1}{3K^4} = \frac{1}{X^2}\frac{1}{3K^2}. \quad (4)$$

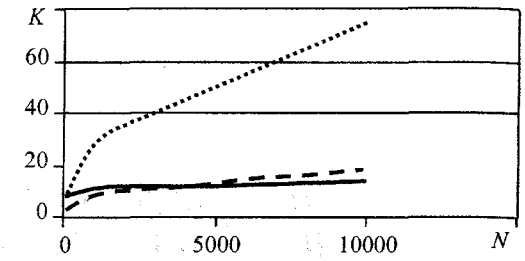


Рис.2. Определение по разным формулам зависимости числа классов K от числа исходных значений N

Компоновка выражений (3) и (4), за исключением константы X^2 в знаменателе, дает целевую функцию F нахождения оптимального количества интервалов при группировке. Эта величина определяется путем отыскания первой производной целевой функции и приравнивания ее нулю:

$$F = \frac{1}{3K^2} + \frac{K-1}{N};$$

$$F' = -\frac{2}{3K^3} + \frac{1}{N} = 0;$$

$$K = \sqrt[3]{\frac{2}{3}N}. \quad (5)$$

Выражение (5) и задает оптимальное количество классовых интервалов для группирования экспериментальных данных в процессе статистической обработки.

Графики, приведенные на рис.2, указывают на определенную близость предлагаемой формулы (5) и выражения Старджеса (1), тогда как формула (2) представляется дающей завышенный результат.

Использование выражения (5) поможет прийти к компромиссу между как излишним, так и недостаточным сглаживанием распределения, поскольку это выражение получено в результате решения задачи оптимальной фильтрации.

ЛИТЕРАТУРА

1. *Новицкий П.В.* Оценка погрешности результатов измерений / П.В.Новицкий, И.А.Зограф. Л.: Энергоатомиздат, 1991. 304 с.

2. *Хальд А.* Математическая статистика с техническими приложениями. М.: Изд-во иностранной литературы, 1956. 664 с.

3. *Sturgess H.A.* The choice of classic intervals // J. American statist.- Ass.- march 1926. 47 p.