MSC 62F10

МАТЕМАТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА СТАТИСТИЧЕСКИХ ДАННЫХ НЕГАУССОВСКОГО ТИПА

***М.М. Ошхунов, **З.М. Ошхунова, *М.А. Джанкулаева**

*Кабардино-Балкарский государственный университет им. Х.М. Бербекова,
ул. Чернышевского, 173, Нальчик, 360004, Россия, e-mail: muaed@inbox.ru;
**Кабардино-Балкарский филиал ОАО Ростелеком,
ул. Головко, 4, Нальчик, 360000, Россия, e-mail: zalina_oshhunova@mail.ru

Аннотация. В работе рассматриваются и обосновываются методы анализа статистических данных, распределенных «не по Гауссу», т.е. имеющих явно выраженную асимметрию и эксцесс. Даны алгоритмы, позволяющие модифицировать классические методы статистики, разработанные в основном для нормально распределенных случайных величин, в частности, метод доверительных интервалов, учитывающий указанные выше факторы. Получены простые формулы для вычисления доверительной точности с заданной надежностью по Лапласу и Стьюденту, если функция плотности распределения случайной величины отличается существенно от экспоненты. Приводятся практические примеры анализа информации, когда асимметрия и эксцесс существенны.

Ключевые слова: асимметрия, эксцесс, нормальное распределение, интервальная оценка.

Как известно [1], нормальный закон распределения случайных чисел встречается в природе наиболее часто. В теории вероятности и математической статистике справедлив закон больших чисел или центральная предельная теорема, суть которой такова: если на какой-то результат влияет бесчисленное количество независимых случайных факторов, отклоняющих его в ту или иную сторону, то суммарное их воздействие приводит к результату, близкому к нормальному распределению.

Нормальный закон распределения характеризуется плотностью распределения, зависящей от двух параметров a, σ:

$$f(x,a,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-a)^2}{2\sigma^2}}\,, \quad \sigma > 0\,. \tag{1}$$

Заметим, что при любых значениях a, σ справедливо условие

$$\int_{-\infty}^{\infty} f(x,a,\sigma)\,dx = 1.$$

Таким образом, нормальный закон распределения симметричен относительно точки $x = a$, имеет определённую «скорость роста» в окрестности этой же точки.

Для количественной характеристики отклонений от симметрии вводят число

$$A = \mu_3/\sigma^3\,,$$

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №12-07-00624 а, №13-07-01003 а.

где

$$\mu_3 = M\left(x - M\left(x\right)\right)^3 ,$$

которое называется коэффициентом асимметрии. Легко показать, что для нормального закона (1) этот коэффициент равен нулю.

Если закон роста плотности функции распределения не гауссовский, вводится новый коэффициент (эксцесс) по формуле

$$E = \frac{\mu_4}{\sigma^4} - 3,$$

$$\mu_4 = M\left(x - M\left(x\right)\right)^4 .$$

Очевидно, эксцесс также равен нулю для нормального закона (1).

Так как методы статистики разработаны, в основном, для нормально распределенных случайных величин, то выводы на их основе справедливы только для таких распределений. Возникает вопрос, как учесть отклонения от закона Гаусса, в частности асимметрию и эксцесс, например, в теории доверительных интервалов.

Попытаемся построить математическую модель, учитывающую отклонение скорости роста плотности распределения от классической, т.е. гауссовской. Как известно, классический нормальный закон имеет максимум в точке $x = a$ равный

$$f\left(x\right)_{x=a} = \frac{1}{\sqrt{2\pi}\sigma} .$$

Встречаются распределения, для которых это «пик» больше или меньше классического значения [2]. Для описания такой плотности введём функцию

$$f\left(x, a, \sigma, c\right) = \frac{\sqrt{\ln c}}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}\ln c}, \quad c > 1 .$$

Очевидно,

$$\int_{-\infty}^{\infty} f\left(x, a, \sigma, c\right) dx = 1$$

для любых параметров a, σ, c ($\sigma > 0$, $c > 1$).

Пик такой функции равен

$$\frac{\sqrt{\ln c}}{\sqrt{2\pi}\sigma} .$$

Он может быть больше или меньше, чем классическое значение для нормального закона в зависимости от c: если $c > e$, то $\sqrt{\ln c}\Big/\left(\sqrt{2\pi}\sigma\right) > 1\Big/\left(\sqrt{2\pi}\sigma\right)$; если $1 < c \leq e$ то $\sqrt{\ln c}\Big/\left(\sqrt{2\pi}\sigma\right) < 1\Big/\left(\sqrt{2\pi}\sigma\right)$.

Вычислим количественное значение эксцесса по формулам

$$E = \frac{\mu_4}{\sigma^4} - 3, \mu_4 = \int_{-\infty}^{\infty} (x-a)^4 f\left(x\right) dx .$$

Легко получить

$$E = \frac{3}{\ln^2 c} - 3 .$$

Очевидно, если $c > e$, то

$$E = 3\left(\frac{1}{\ln^2 c} - 1\right) < 0\,,$$

и, наоборот, если $1 < c \leq e$, то

$$E = 3\left(\frac{1}{\ln^2 c} - 1\right) > 0\,.$$

Таким образом, плотность распределения

$$f = \frac{\sqrt{\ln c}}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-a)^2}{2\sigma^2}\ln c\right\}, \quad c > 1$$

моделирует любое значение эксцесса, если оно симметрично относительно точки $x = a$.

Легко оценить доверительный интервал с заданной надежностью, когда эксцесс – ненулевой. Очевидно, аналог интегральной функции Лапласа имеет вид

$$\phi^*(x) = \frac{\sqrt{\ln c}}{\sqrt{2\pi}\sigma}\int_0^\infty \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\ln c\right\}dx = \phi\left(\frac{x-a}{\sigma}\sqrt{\ln c}\right),$$

где $\phi(x) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_0^x e^{-\frac{t^2}{2}}dt.$

Для классического нормального закона вероятность попадания случайной величины в интервал $\alpha < x < \beta$ равна

$$\phi\left(\frac{\beta-a}{\sigma}\right) - \phi\left(\frac{\alpha-a}{\sigma}\right).$$

При наличии эксцесса эта формула приобретает вид

$$P(\alpha < x < \beta) = \phi\left(\frac{\beta-a}{\sigma}\sqrt{\ln c}\right) - \phi\left(\frac{\alpha-a}{\sigma}\sqrt{\ln c}\right).$$

Аналогичным образом, вероятности заданного отклонения, очевидно, равны

$$\phi(|x-a| < \delta) = 2\phi\left(\frac{\delta}{\sigma}\right)$$

– по Лапласу;

$$\phi(|x-a| < \delta) = 2\phi\left(\frac{\delta\sqrt{\ln c}}{\sigma}\right)$$

– при $E \neq 0$.

В случае ненулевого эксцесса, когда доверительная точность с заданной надежностью может быть найдена по формулам

$$\delta = \frac{t\sigma}{\sqrt{\ln c}\cdot\sqrt{n}}\,, \tag{2}$$

$$a = \overline{x} \pm \delta.$$

Из формулы (2) следует, что, чем больше c, тем доверительная оценка выше. Если объем выборки небольшой, то, очевидно, параметр надежности t следует выбирать из таблицы из Стьюдента.

Теория доверительных интервалов для распределений с ненулевым эксцессом распространяется также на задачи о статистическом различии (или неразличии) средних двух выборок. Очевидно, также, что объем выборки, обеспечивающий необходимую статистическую точность с заданной надежностью, при наличии эксцесса равен

$$n = \frac{t^2\sigma^2}{\delta^2 \ln c} .$$

Приведенные формулы дают возможность такого учета. Из приведенного краткого анализа можно сделать вывод, что если полученное значение числа c отличается сильно от e, то эксцесс распределения экономической информации существенен и его учет является обязательным.

Литература

1. Васильев В.И., Класильников В.В., Млаксий С.И., Тягунова Т.Н. Статистический анализ многомерных объектов произвольной природы / М.: ИКАР, 2004. – 382 с.
2. Льюис К.Д. Методы прогнозирования экономических показателей / М.: Финансы и статистика, 1986.

MATHEMATICAL MODELS OF STATISTICAL DATA ANALYSIS OF NON-GAUSSIAN TYPE

***M.M. Oshkhunov, **Z.M. Oshkhunova, *M.A. Dzhankulaeva**

*Kabardino-Balkarian State University named H.M. Berbekov,
Chernishevkogo St., 173, Nalchik, 360004, Russia, e-mail: muaed@inbox.ru;
** Kabardino-Balkarian department of OJSC Rostelecom,
Golovko St., 4, Nalchik, 360000, Russia, e-mail: zalina_oshhunova@mail.ru

Abstract. Methods of statistical analysis of data distributed by non-gaussian way are under consideration. Asymmetry and excess are pronounced. Given algorithms, in particular, the method of confidence intervals, are modified to classic statistical methods developed primarily for normally distributed random variables. The simple formula is proposed to calculate the confidence of accuracy with a given Laplace and Student reliability if the probability density of random variable differs significantly from the exponents low. Practical examples are provided concerned the analysis of information when asymmetry and excess are significant.

Key words: asymmetry, excess, normal distribution, interval estimation.