Использование коэффициентов асимметрии и эксцесса при гистограммном методе определения закона распределения вероятности

***С.С. Акимов**, аспирант, Оренбургский ГУ*

В последнее время, в связи с совершенствованием компьютерных технологий, процесс обработки информации в экономических системах становится ускоренным и более доступным. При этом в ходе процесса обработки информации накапливается колоссальное количество эмпирических данных, имеющих, как правило, случайную природу. Для качественной их обработки необходима широкая теоретическая база обработки подобных данных.

Наиболее полной, исчерпывающей характеристикой случайной величины является закон распределения [1].

На сегодняшний момент существует целый ряд способов восстановления закона распределения вероятности по выборке из генеральной совокупности.

Необходимость в восстановлении закона распределения обусловлена особенностями всех современных статистических пакетов и программ: для проведения анализа данных закон распределения задаётся вручную. Незнание же закона

распределения, которому подчиняется выборка, приводит к тому, что исследователь берёт за основу нормальное распределение и далее анализирует совокупность, исходя из параметров нормального распределения [2].

Одним из наиболее известных и применяемых методов восстановления закона распределения служит гистограммный метод, подробно описанный в [3]. Однако данный метод, при всей его простоте использования, является весьма субъективным. Для снижения субъективности данного способа оценивания необходимо использовать математические методы.

При гистограммном методе оценки плотности распределения применяется разностная аппроксимация $P(y)$ в виде:

$$P(y) = F'(y) = \lim_{h\to 0} \frac{F(y+h)-F(y)}{h} \approx \frac{F(y+h)-F(y)}{h}, \quad (1)$$

а в качестве оценки функции $P(y)$ используется зависимость:

$$P_N(y) = \frac{F_N(y+h)-F_N(y)}{h} = \frac{1}{Nh}\sum_{i=1}^{N}\left[\theta(y+h-x_i)-\theta(y-x_i)\right] = \frac{\upsilon_y}{Nh}, \quad (2)$$

где υ_y – количество выборочных значений, попавших в интервал $(y; y+h]$ [1].

Гистограммный метод даёт исследователю графическое отображение экспериментальных данных, и по виду построенной гистограммы исследователь принимает гипотезу о виде закона распределения вероятности. Однако сам вид гистограммы зависит от ряда характеристик, из которых основными являются (для одновершинных распределений) сдвиг параметров относительно центра (асимметрия) и кривизна полученной гистограммы (эксцесс). Как известно, эти параметры рассчитываются как центральные моменты третьего и четвёртого порядка [4].

Гистограммный метод является далеко не единственным методом. Не менее распространены методы, предложенные Парзеном и Ронзенблатом, в которых используется сглаженная эмпирическая функция распределения в виде:

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} G\left(\frac{y-x_i}{h_N}\right), \quad (3)$$

а также введено понятие «ядерная функция».

Однако и данный метод сопряжён с рядом трудностей. Как известно, смещение и вариация оценки данной функции зависят от вида ядра $K(t)$ и значения параметра размытости h_N. И если для выделения среди числа функций $K(t)$ имеется достаточно подходящий критерий отбора, выраженный через информационный функционал:

$$J = \int \ln\left[K(\mathrm{t})\right]P(t)dt, \quad (4)$$

то задача оценивания оптимальной величины h_N является более сложной, нежели исходная задача восстановления плотности распределения [1]. Кроме того, существуют и другие проблемы, связанные с использованием этого метода, например проблема локальных сгущений или проблема «проклятия размерности» [5].

Целый ряд отечественных и зарубежных учёных описывают применение методов коэффициента асимметрии и эксцесса для проверки нормальности распределения [7, 8]. Однако ряд авторов также признают несостоятельность использования данного метода для проверки нормальности [6]. Основная причина несостоятельности заключается в том, что существует ряд распределений, имеющих коэффициент асимметрии и эксцесса, аналогичный нормальному закону распределения.

Таким образом, использование коэффициентов асимметрии и эксцесса не может дать однозначного ответа на вопрос о нормальности закона распределения в частности и виде закона распределения в целом. Однако этот метод весьма действенен, если использовать его как критерий для сортировки законов распределений.

Рассмотрим этот процесс более подробно. Для начала необходимо перечислить все наиболее часто встречающиеся распределения: распределение Коши, Фишера, Стьюдента, Пуассона, Вейбулла, Бернулли, Рэлея, нормальное, логнормальное, логистическое, равномерное непрерывное и дискретное, биноминальное, отрицательное биноминальное, геометрическое, гипергеометрическое, экспоненциальное, гамма, бета и хи-квадрат.

Прежде чем исследовать коэффициенты асимметрии и эксцесса, необходимо отметить главную сложность этого процесса: в ряде распределений с изменением параметра изменяются и моменты третьего и четвёртого порядков. Потому отнесём такие распределения в «зону неопределённости».

Итак, разобьём перечисленные выше законы распределения согласно коэффициенту асимметрии, представив данные в таблице 1.

Как видно по таблице, в зону неопределённости попадает гораздо больше законов распределений, чем во все другие. Отрицательный момент заключается в том, что при оценке коэффициента асимметрии «зону неопределённости» придётся учитывать как в симметричных, так и асимметричных распределениях, отсюда следует, что данный коэффициент лишь поможет отбросить те законы распределения, которые точно не являются симметричными или асимметричными.

Такая же ситуация и с коэффициентом эксцесса (табл. 2).

Замечание, написанное для коэффициента асимметрии, справедливо и для коэффициента эксцесса – расчёт коэффициента поможет лишь отбросить законы, не попадающие в «нормальную зону».

1. Сортировка законов распределения вероятностей согласно коэффициенту асимметрии

Симметричные	Асимметричные	Неопределённые
Биноминальное Нормальное Логистическое Стьюдента Равномерное непрерывное Равномерное дискретное	экспоненциальное Фишера геометрическое логнормальное	Коши Бета хи-квадрат гамма гипергеометрическое отр. биноминальное Пуассона Вейбулла Рэлея Бернулли

2. Сортировка законов распределения вероятностей согласно коэффициенту эксцесса

Нормальный эксцесс	Ненормальный эксцесс	Неопределённые
Биноминальное Гипергеометрическое Нормальное Стьюдента Равномерное непрерывное Равномерное дискретное	Коши экспоненциальное Фишера геометрическое логнормальное	бета хи-квадрат гамма логистическое отр. биноминальное Пуассона Вейбулла Бернулли Рэлея

По приведённым таблицам видно, что совмещение этих коэффициентов не даёт особого результата: в большинстве случаев симметричные распределения имеют нормальный эксцесс и наоборот. Поэтому наиболее информативными станут случаи, когда закон распределения симметричный, а эксцесс не соответствует нормальному или наоборот – в этих случаях отсеивается примерно половина приведённых законов распределения.

Кроме того, по приведённым таблицам видна несостоятельность использования коэффициентов асимметрии и эксцесса для проверки нормальности. Однако их использование для определения вида закона распределения позволяет отсеять от 4 до 10 законов, не попадающих под заданные условия.

Таким образом, из всего вышесказанного можно сделать следующие выводы.

1. Гистограммный метод является весьма простым, но очень субъективным методом.

2. Субъективность гистограммного метода можно снизить, используя различные математические методы, основанные на свойствах законов распределения.

3. Моменты третьего и четвёртого порядка (асимметрия и эксцесс) широко используются в качестве определения нормальности распределения, хотя и не являются состоятельными.

4. Использование коэффициентов асимметрии и эксцесса целесообразно для отсеивания ряда законов распределения с целью увеличения точности применения впоследствии гистограммного метода.

Литература

1. Сызранцев В.Н., Невелев Я.П., Голофаст С.Л. Адаптивные методы восстановления функции плотности распределения вероятности // Известия вузов. Машиностроение. 2006. № 12. С. 3–11.
2. Акимов С.С., Шепель В.Н. Эвристическая процедура определения подходящего распределения вероятности // Компьютерная интеграция производства и ИПИ-технологии: сб. матер. V Всеросс. науч.-практич. конф. Оренбург: Изд. ИП Осиночкин Я.В., 2011. С. 137–139.
3. Шепель В.Н. Алгоритм определения эмпирической функции плотности $f^{(n)}(x)$ по выборке из генеральной совокупности // Современные информационные технологии в науке и практике: матер. VIII всеросс. науч.-практич. конф. (с международным участием). Оренбург: ИПК ГОУ ОГУ, 2009. С. 224–226.
4. Акимов С.С. Применение коэффициентов асимметрии и эксцесса для определения закона распределения вероятностей // Новинанта за напреднали наука: матер. 9-й межд. науч.-практич. конф. 2013. Т. 53. Математика. София. «Бял ГРАД-БГ» ООД. С. 30–33.
5. Акимов С.С. Оптимизированный алгоритм определения закона распределения вероятности по выборке из генеральной совокупности // Известия Самарской государственной сельскохозяйственной академии. 2013. № 2. С. 52–56.
6. Орлов А.И. Типовые ошибки при вхождении в прикладную статистику. URL: //http://forum.orlovs.pp.ru/viewtopic.php?t=97.
7. D'Agostino, Ralph B.; Albert Belanger; Ralph B. D'Agostino, Jr (1990). «A suggestion for using powerful and informative tests of normality». The American Statistician 44 (4): 316–321.
8. Shenton L.R.; Bowman K.O. (1977). «A bivariate model for the distribution of b1 and b2». Journal of the American Statistical Association 72 (357): 206–211.