УДК 330.4: 519.2

Шепель В.Н.
Оренбургский государственный университет
E-mail: fit.cits@mail.ru

ПРОЦЕДУРА ПОСТРОЕНИЯ ВЫБОРОЧНОГО АНАЛОГА ФУНКЦИИ ПЛОТНОСТИ

Рассматривается проблема увеличения степени формализации процесса определения закона распределения по выборке из генеральной совокупности. Предлагается процедура определения эмпирической функции плотности $\hat{f}^{(n)}(x)$.

Ключевые слова: процедура, генеральная совокупность, выборка, эмпирическая функция плотности, проверка статистических гипотез.

В практике статистического анализа и моделирования точный вид закона распределения анализируемой генеральной совокупности, как правило, бывает неизвестен. Мы располагаем лишь выборкой из интересующей генеральной совокупности. Если это одномоментные наблюдения для одного признака, то матрица выборочных данных (в.д.) имеет вид

$$(\text{в.д.})=\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}. \qquad (1)$$

Строить свои выводы и принимать решения мы вынуждены на основании расчета ограниченного ряда выборочных характеристик. К основным выборочным (эмпирическим) характеристикам [1] относятся:

– эмпирическая функция распределения $\hat{F}^{(n)}(x)$;

– эмпирическая функция плотности $\hat{f}^{(n)}(x)$;

– эмпирическая относительная частота $\hat{p}_i^{(n)}$ появления i-го возможного значения x_i дискретной случайной величины;

– эмпирические начальные и центральные моменты анализируемой случайной величины $\hat{m}_k(n)$ и $\hat{m}_k^o(n)$;

– порядковые статистики $x_{(i)}(n)$ $(i=1,2,\dots,n)$.

Наиболее информативной для непрерывных случайных величин является выборочный аналог функции плотности (эмпирическая функция плотности) $\hat{f}^{(n)}(x)$. В статье предлагается алгоритм (процедура) определения $\hat{f}^{(n)}(x)$ пригодный для использования при работе в статистических пакетах или подготовке программных продуктов.

Процедура определения эмпирической функции плотности $\hat{f}^{(n)}(x)$:

1. Отмечаются наименьшее $x_{\min}(n)$ и наибольшее $x_{\max}(n)$ значения в выборке (1).

2. Диапазон $[\,x_{\min}(n)\,, x_{\max}(n)\,]$ разбивается на s равных интервалов группирования; при этом количество интервалов s должно быть в пределах 7–20. В выборе s можно пользоваться приближенной формулой $s \approx 1+3{,}321\lg(n)$.

3. Отмечаются крайние точки каждого из интервалов $c_0, c_1, c_2, \dots, c_s$ в порядке возрастания; для чего определяется длина интервала

$$\Delta_{k(x)} = (\,x_{\max}(n) - x_{\min}(n)\,)/[\,s\,],$$

затем $c_{k(x)} = c_{k(x)-1} + \Delta_{k(x)}$, а также их середины $x_1^0, x_2^0, \dots, x_s^0$.

4. Подсчитываются числа выборочных данных, попавших в каждый из интервалов: $\nu_1, \nu_2, \dots, \nu_s$ (очевидно, $\nu_1 + \nu_2 + \dots + \nu_s = n$); выборочные данные, попавшие на границы интервалов, либо равномерно распределяются по двум соседним интервалам, либо относятся только к какому-либо одному из них.

5. Для каждого интервала рассчитывается эмпирическая функция плотности

$$\hat{f}^{(n)}(x) = \frac{\nu_{k(x)}}{n\Delta_{k(x)}},$$

где $k(x)$ – порядковый номер группирования, который накрывает точку x; $\nu_{k(x)}$ – число наблюдений, попавших в этот интервал, $\Delta_{k(x)}$ – длина интервала.

6. Строится гистограмма, для чего на оси абсцисс откладываются крайние точки каждого из интервалов $c_0, c_1, c_2, \dots, c_s$, по оси ординат эмпирическая функция плотности $\hat{f}^{(n)}(x)$. Тог-

да k-му интервалу будет соответствовать прямоугольник, основанием которого является замкнутый слева интервал $[c_{k-1}, c_k)$, а высота

$$\hat{f}^{(n)}(x) = \frac{v_{k(x)}}{n\Delta_{k(x)}}.$$

7. По виду гистограммы принимается гипотеза о модели закона распределения анализируемой генеральной совокупности (например, нормальный, экспоненциальный, равномерный и т. д.).

Для облегчения выбора и улучшения его точности целесообразно использовать классификацию случайных величин и законов распределения. Наиболее используемыми в моделировании и практических приложениях [2] являются распределения непрерывных случайных величин: равномерное, экспоненциальное, Эрланга m-го порядка, нормальное, логнормальное, Вейбулла, Пирсона V типа, Пирсона VI типа, лог-логистическое, Безье, треугольное, гамма-распределение, бета-распределение, связанное распределение Джонсона, несвязанное распределение Джонсона.

8. Рассчитываются оценки неизвестных параметров гипотетического закона распределения $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ (например, для нормального закона распределения выборочное среднее $\bar{x}(n)$ и выборочную дисперсию $s^2(n)$).

Эмпирическая функция плотности $\hat{f}^{(n)}(x)$ – получена. Возникает необходимость в экспериментальной проверке гипотезы о виде закона распределения анализируемой генеральной совокупности, т. е. наша цель – проверить, не противоречит ли высказанная гипотеза H имеющимся выборочным данным.

Задача может быть решена качественно или количественно. Качественное решение задачи осуществляется на основе сравнения графиков эмпирической $\hat{f}^{(n)}(x)$ и модельной $\hat{f}(x)$ плотностей. Количественная оценка достоверности принятой гипотезы осуществляется с помощью того или иного статистического критерия.

Качественная проверка гипотезы о виде закона распределения анализируемой генеральной совокупности:

9. Для построения модельной кривой плотности используется гипотетический (принятый в соответствии с гипотезой) закон распределения, в который подставляются вместо неизвестных параметров значения соответствующих выборочных характеристик.

10. В результате сравнения вида гистограммы (полигона) и кривой плотности гипотетического закона распределения гипотеза отклоняется или не отклоняется.

Количественную (статистическую) проверку гипотезы о виде закона распределения анализируемой генеральной совокупности целесообразно проводить с помощью критерия согласия χ^2 Пирсона [3], для этого:

11. Подсчитываем значение критической статистики

$$\gamma^{(n)} = \sum_{j=1}^{s} \frac{(v_j - np_j(\hat{\hat{\theta}}))^2}{np_j(\hat{\hat{\theta}})};$$

где $p_j(\hat{\Theta}) = F_{мод}(c_j, \hat{\Theta}) - F_{мод}(c_{j-1}, \hat{\Theta})$ – результат модельного расчета вероятности попадания в j-й интервал группирования.

12. Задаемся уровнем значимости критерия α. По заданному уровню значимости критерия α из таблиц χ^2 распределения находим точки $\chi^2_{1-\alpha/2}(s-k-1)$ и $\chi^2_{\alpha/2}(s-k-1)$ с $(s-k-1)$ степенями свободы;

13. Если

$$\chi^2_{1-\alpha/2}(s-k-1) \le \gamma^{(n)} \le \chi^2_{\alpha/2}(s-k-1),$$

то гипотеза о виде закона распределения анализируемой генеральной совокупности не отклоняется, в противном случае гипотеза отклоняется.

Итак, в статье рассмотрена процедура определения эмпирической функции плотности $\hat{f}^{(n)}(x)$. Процедура пригодна для использования при работе в статистических пакетах или подготовке программных продуктов.

14.12.2011

Список литературы:
1. Сигел, Эндрю Практическая бизнес-статистика.: Пер. с англ. – М.: Издательский дом «Вильямс», 2002. – 1056 с.: ил.
2. Кельтон В., Лоу А. Имитационное моделирование. Классика CS. 3-е изд. – СПб.: Питер; Киев: Издательская группа BHV, 2004. – 847 с.: ил.
3. Паклин Н.Б., Орешков В.И. Бизнес- аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.: ил.

Сведения об авторе:
Шепель Вячеслав Николаевич, заведующий кафедрой управления и информатики в технических системах Оренбургского государственного университета, доктор экономических наук, профессор
460018, г. Оренбург, пр-т Победы, 13, ауд. 14336, тел. (3532) 372558, e-mail: fit.cits@mail.ru

UDC 330.4: 519.2
Shepel V.N.
Orenburg state university
E-mail: fit.cits@mail.ru

PROCEDURE FOR CONSTRUCTION OF SAMPLING ANALOG OF DENSITY FUNCTION

In paper we consider increasing the degree of formalization of the process of determining the distribution law for a sample of the population. The procedure of determining the empirical density function $f^{(n)}(x)$ is given.

Key words: procedure, general population, sample, empirical density function, testing statistical hypotheses.

Bibliography:

1. Siegel, Andrew Practical Business Statistics.: Trans. from English. – M.: Publishing house «Williams», 2002. – 1056 p.
2. Kelton W., Simulation A. Low. Simulation Modeling And Analysis. Classic CS. 3rd ed. – St. Petersburg.: Peter; Kiev: Publishing Group BHV, 2004. – 847 p.
3. Paklin N.B., Oreshkov V.I. Business Intelligence: From Data to Knowledge. – St. Petersburg.: Peter, 2009. – 624 p.