УДК 621.391, 621.372

Филатов Владимир Иванович, к.т.н., доцент
Борукаева Александра Олеговна, студент
Бердиков Павел Геннадьевич, студент
Московский государственный технический университет имени Н. Э. Баумана (Россия, г. Москва)

ВЫБОРОЧНЫЙ МЕТОД И ПАРАМЕТРЫ ЗАКОНА РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

В работе рассмотрено, что на значения наблюдаемого признака оказывают влияние большое число различных факторов. Одни из этих факторов оказывают на наблюдаемый признак существенное влияние и поэтому считаются основными факторами. Таких факторов (основных) на практике учитывается небольшое число, что существенно упрощает математическую модель изучаемой совокупности объектов. Влияние же других факторов относительно мало, и они считаются второстепенными факторами. Таких факторов (второстепенных) на практике учитывается любое, сколь угодно большое число. Однако их учет несущественно усложняет математическую модель изучаемой совокупности объектов.

Ключевые слова: признак, фактор, функция распределения, метод, распределение, выборка.

МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ

Пусть в результате наблюдений требуется изучить признак X, в качестве которого может рассматриваться случайное событие, величина, вектор или процесс. Если бы на изучаемый признак оказывали влияние только основные факторы испытаний, то после некоторого числа наблюдений n можно было бы установить с любой наперед заданной ошибкой связь между X и основными факторами (в том числе и истинное значение самого X).Этот вывод основывается на том, что в процессе испытаний основные факторы строго контролируются (измеряются) и значение X оказывается неслучайным.

Поскольку на признак X кроме основных факторов оказывают влияние и второстепенные факторы, которые при испытаниях не контролируются (неизвестны их конкретные значения), то значение признака оказывается случайным (непредсказуемым). В этом случае в наиболее общем виде признак X может быть описан законом распределения $F(x)$ зависящим как от основных, так и от второстепенных факторов. Этот закон распределения называют законом распределения генеральной совокупности. Его математическое ожидание $M[X]$ и дисперсия$D[X]$называются, соответственно, генеральным средним и генеральной дисперсией. Аналогично называются и другие числовые характеристики закона распределения $F(x)$. Очевидно, что этот закон распределения и его числовые характеристики нам неизвестны и должны быть получены в процессе испытаний.

Испытания организуются в соответствии с выборочным методом.

Суть выборочного метода заключается в том, что из генеральной совокупности случайно выбирается в результате наблюдения признака егозначения $x_1, x_2, ..., x_i, ..., x_n$, образующие выборочную совокупность. Наблюдавшиеся различные значения x_i, $i=\overline{1,n}$ признака X называются вариантами. Выборочная совокупность подвергается детальномуисследованию, по результатам которого требуется описать всю генеральную совокупность или какие-нибудь ее свойства, характеристики.

Здесь уместно сделать одно предупреждение. На первый взгляд выборочный метод мало чем отличается от обычного метода малых проб. Однако это не так. При методе малых проб мы заведомо знаем, что интересующий нас признак распределен по генеральной совокупности равномерно и поэтому любая малая проба является точной мини-копией генеральной совокупности. При выборочном же методе исследуемый признак его значения распределен по генеральной совокупности неравномерно, причем даже характер этой неравномерности неизвестен. Поэтому далеко не всякая выборка хорошо отражает структуру всей генеральной совокупности (является репрезентативной).

Не имея в общем случае никаких сведений о генеральной совокупности, мы, делая выборку, можем полагаться только на случай — все прочие способы отбора будут необъективными, носящими следы влияния посторонних факторов.

Сопоставление закона распределения генеральной совокупности с законом распределения выборочной совокупности основано на теореме Гливенко: с вероятностью I при $n \to \infty$ максимальная разность между функциями распределения $F^*(x)$ и $F(x)$ стремится к нулю, где $F^*(x)$ —закон распределения выборочной совокупности;$F(x)$ —закон распределения генеральной совокупности.

Очевидно, что справедливы аналогичные утверждения относительно числовых характеристик (математического ожидания, дисперсии и др.) генеральной и выборочной совокупностей.

В математической статистике по результатам исследования выборочной совокупности решается два различных класса задач:

1. оценивание признаков генеральной совокупности;

2. проверка правдоподобия статистических гипотез относительно признаков генеральной совокупности.

Если элементы выборки $x_1, x_2, ..., x_i, ..., x_n$ расположить в порядке их наблюдения в серии испытаний, то они образуют простую статистическую совокупность (простой статистический ряд). Обилие и

неупорядоченность информации в простом статистическом ряду затрудняет его использование при дальнейшем анализе выборки. Для дальнейшего анализа подвергнем ранжированию простой статистический ряд, расположив наблюдавшиеся элементы выборки в возрастающем порядке $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i, \ldots, \tilde{x}_n$ (где $\tilde{x}_1 \le \tilde{x}_2 \le \ldots \le \tilde{x}_i \le \ldots \le \tilde{x}_n$). Величины $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ называютпорядковыми статистиками выборки. Затем в ранжированном простом статистическом ряду сгруппируем его члены по одинаковым значениям(вариантам) $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_j, \ldots, \hat{x}_m$, где m — число различных вариант в простом статистическом ряду. Очевидно, что $m \le n$. Полученный ряд значения признака называется вариационным рядом. Вариационный ряд и является той первичной информацией о выборке, на основании изучения которой строится закон распределения выборки.

Вариационный ряд записывается в виде таблицы

Таблица 1 - Вариационный ряд

X	$\hat{x}_1$	$\hat{x}_2$	…	$\hat{x}_j$	…	$\hat{x}_m$
n_j	n_1	n_2	…	n_j	…	n_m

где $\hat{x}_j$ — варианты статистического ряда;

n_j— частота наблюдения этих вариант.

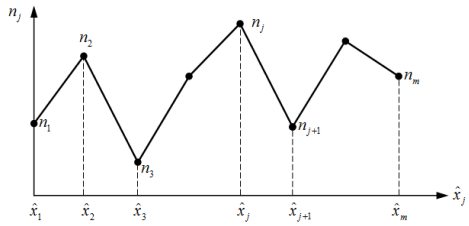

Рис.1. Зависимость частоты наблюдения от варианты статического ряда.

Очевидно, что

$$n_1 + n_2 + \ldots + n_j + \ldots + n_m = \sum_{j=1}^{m} n_j = n$$

Этот ряд можно представить графически в виде полигона частот.

Вариационный ряд можно задать и в другой форме, если частоты n_j в нем заменить относительными частотами $P_j^* = \frac{n_j}{n}$. При этом $\sum_{j=1}^{m} P_j^* = 1$.

Если число вариант велико, то более удобна форма задания вариационного ряда в виде гистограммы. Для этого все варианты разобьем на «разряды» (по нескольку вариант в разряде) с общим их числом Sи подсчитаем число наблюдавшихся значений признаков n_rв каждомr-м разряде (где $\sum_{r=1}^{S} n_r = n$). Тогда гистограммой частот будем называть ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат «длины разрядов» (h_r, а высоты равны частотам n_r, деленным на «длину разряда»n_r / h_r.

Для определения оптимальной «длины разряда» hможно использовать формулу Стэрджеса

$$h = \left(\hat{x}_{\max} - \hat{x}_{\min}\right) / \left(1 + 3{,}322 \lg n\right),$$

а за начало первого интервала целесообразно принять значение

$$x_{\min} - \frac{h}{2}.$$

Введем понятие накопленной (кумулятивной) частоты, равной числу элементов выборки, имеющих значение признака X меньше данного числа. Обозначим накопленную частоту через $W^*(x)$, тогда

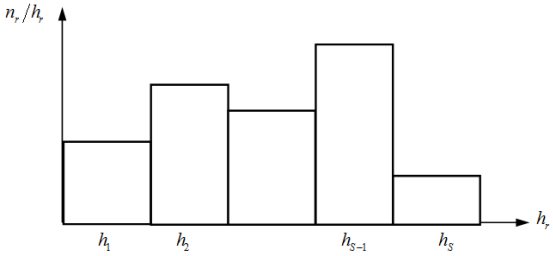

Рис.2. Гистограмма частот

$$W^*(x)=\begin{cases}0 & \text{при } x\le \hat{x}_1;\\ n_1 & \text{при } \hat{x}_1<x\le \hat{x}_2;\\ n_1+n_2 & \text{при } \hat{x}_2<x\le \hat{x}_3;\\ n_1+n_2+n_3 & \text{при } \hat{x}_3<x\le \hat{x}_4;\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ n_1+n_2+\ldots+n_j+\ldots+n_m=n & \text{при } \hat{x}_m<x.\end{cases}$$

Если в этом выражении от частот n_j перейти к относительным частотам $P_j^*=n_j/n$, то получим накопленную относительную частоту $F(x)$, которая будет называться кумулятой.

$$F^*(x)=\begin{cases}0 & \text{при } x\le \hat{x}_1;\\ P_1^* & \text{при } \hat{x}_1<x\le \hat{x}_2;\\ P_1^*+P_2^* & \text{при } \hat{x}_2<x\le \hat{x}_3;\\ P_1^*+P_2^*+P_2^* & \text{при } \hat{x}_3<x\le \hat{x}_4;\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ P_1^*+P_2^*+\ldots+P_m^*=1 & \text{при } \hat{x}_m<x.\end{cases}$$

Кумулята и является функцией (законом) распределения выборки, по которому производится изучение признака генеральной совокупности.

ПАРАМЕТРЫ ЗАКОНА РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

Важнейшая задача математической статистики, решение которой позволило бы в принципе решить и все остальные задачи, — это нахождение функции распределения $F(x)$ наблюдаемого признака. К счастью, при обработке наблюдений редко приходится прибегать к построениювыборочной функции распределения $F^*(x)$. Даже простейший анализ условий испытаний позволяет с достаточной степенью уверенности определить тип неизвестной функции распределения $F(x)$. В подобном случае окончательное уточнение неизвестной функции распределения сводится к определению только числовых параметров распределения. Эти параметры определяются по выборке, разумеется, приближенно. Однако нужная точность здесь достигается при гораздо меньших объемах выборки, чем при непосредственном использовании теоремы Гливенко.

Для определения параметров распределения выборки построим рассуждения следующим образом. Вместо того, чтобы говорить, что наблюдается наступление случайной переменной (признака) X в n испытаниях, которые и образуют случайную выборку, говорят: случайная выборка образуется проявлением ряда случайных переменных $X_1^*,\ X_2^*,\ldots,X_n^*$,каждая из которых сама подчиняется одному и тому же закону распределения $F^*(x)$. Это замечание следует из условия однородности испытаний в серии. Действительно, если все наблюдения X_k^* в выборке подчиняются одному и тому же закону распределения $F^*(x_1)=F^*(x_2)=\ldots=F^*(x_k)=\ldots=F^*(x)$, то это и подтверждает условие однородности испытаний. В свою очередь, условие однородности испытаний позволяет утверждать в первом приближении, что отдельные значения признака X в выборке равновероятны т. е. $P^*(x_1)=P^*(x_2)=\ldots=P^*(x_n)=\dfrac{1}{n}$. Именно это предположение и положено в основу определения параметров закона распределения выборки.

На основании сделанного предположения о равновероятности значений признака определим значение обычного выборочного момента порядка k по формуле математического ожидания при равномерном законе распределения

$$M_k^*[X-c]=\sum_{i=1}^{n}(x_i-c)^k\cdot P^*(x_i)=\frac{\sum_{i=1}^{n}(x_i-c)^k}{n}=\overline{(X-c)^k}$$

т. е. совпадает со средним арифметическим величины $(X-c)^k$.

Очевидно, что при c = 0 мы получаем начальные моменты $L_k^*(x)$ распределения выборки, а при $c = M[X] = m_x$ получаем центральные моменты $\mu_k^*(x)$ распределения. Практическая сложность применениязтой формулы состоит в необходимости знать до испытаний (априори) математическое ожидание $m_x = M[X]$ признака Xв генеральной совокупности.

Справедливость замены статистических средних (моментов) выборки ее арифметическими средними основана на законе больших чисел. Одной из наиболее известных форм закона больших чисел является неравенство Чебышева: Если X есть случайная переменная с каким угодно законом распределения и имеет конечное математическое ожиданиеm_x и дисперсию D_x, то для произвольно взятого положительного числа ε справедливо неравенство

$$P\{|X - m_x| < \varepsilon\} \geq 1 - \frac{D_x}{\varepsilon^2}.$$

Если принять $\varepsilon = k\sqrt{D_x}$ (где k — произвольное, заранее заданноеположительное число), то неравенство Чебышева примет вид

$$P\left\{|X - m_x| < k\sqrt{D_x}\right\} \geq 1 - \frac{1}{k^2} \text{ или}$$

$$P\left\{|X - m_x| < k\sqrt{D_x}\right\} < \frac{1}{k^2}.$$

Существует также более общая лемма Маркова: Для любой положительной случайной величины X вероятность того, что она примет значение, не превосходящее некоторого положительного числа ε, равна

$$P\{|X| \leq \varepsilon\} > 1 - \frac{m_x}{\varepsilon},$$

которая может быть также положена в основу доказательства неравенства Чебышева.

Таким образом, мы показали, что определение параметров распределения выборки может быть сведено к определению выборочных арифметических средних и на основании неравенства Чебышева ошибка от такой «подмены» уменьшается с ростом объема выборки n.

Для практического вычисления параметров (моментов) распределения выборки по значениям вариант $\hat{x}_j\left(j = \overline{1,m}\right)$ используются метод произведений и метод сумм.

Список литературы

1. Гмурман В.Е. Теория вероятностей и математическая статистика. Учебное пособие для вузов. Изд. 7, стер. – М.: Высш. шк., 2004.

2. Гмурман В.Е Руководство к решению задач по теории вероятностей и математической статистике. Учебное пособие для вузов. – М.: Высш. шк., 2004.

3. Серебряный М.С Руководство к решению задач по теории вероятностей и математической статистике. Учебное пособие.-Новогорск, 2001.

4. Окунев Ю. Б. Цифровая передача информации фазомодулированными сигналами. – М.: Радио и связь, 1991.–296 с.

5. Пенин П. И. Системы передачи цифровой информации. – М.: Сов. Радио, 1976.–364 с.

Filatov Vladimir Ivanovich, candidate of technical science, assistant professor
Borukaeva Alexandra Olegovna, student
Berdikov Pavel Gennadievich, student
Bauman Moscow State Technical University (Russia, Moscow)

SAMPLING METHODS AND PARAMETERS OF THE LAW DISTRIBUTION OF THE SAMPLE

The paper will be considered that the observed feature values for influencing a number of different factors. Some of these factors have to see signs of a significant impact and are considered the main factors. Such factors (basic) in practice is taken into account a small number, which greatly simplifies the mathematical model of the studied population of objects. The impact of the other factors is relatively small, and they are regarded as secondary factors. Such factors (secondary), in practice, take into account any arbitrarily large number. However, their accounting is immaterial complicates the mathematical model of the studied population of objects.

Keywords: symptom, factor, distribution function, method, the distribution, the sample.