32.5. Закон больших чисел и предельные теоремы

Теория вероятностей изучает закономерности, свойственные массовым случайным явлениям. В том случае, если явление носит единичный немассовый характер, теория вероятностей способна обычно предсказать лишь результаты в очень широких пределах. Закономерности проявляются именно при большом числе случайных явлений, происходящих в одинаковых условиях. В этом случае характеристики случайных событий и случайных величин, наблюдаемых при испытании, становятся почти неслучайными. Например, частота появления события при большом числе испытаний устойчива, то же относится и к средним значениям случайных величин.

О Группа теорем, устанавливающих соответствие между теоретическими и экспериментальными характеристиками случайных величин и случайных событий при большом числе испытаний над ними, носит название ***предельных теорем*** теории вероятностей.

К ним относятся *закон больших чисел* (группа теорем, включающая, в частности, неравенство и теорему Чебышёва, теорему Бернулли) и центральная предельная теорема Ляпунова.

Неравенство Чебышёва. *Пусть случайная величина X имеет математическое ожидание $M(X)$ и дисперсию $D(X)$. Вероятность того, что отклонение случайной величины от ее математического ожидания будет по абсолютному значению не меньше любого положительного числа ε, ограничена сверху величиной $\frac{D(X)}{\varepsilon^2}$, т. е.* $P(|X - M(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$.

Геометрический смысл этого события заключается в том, что значение случайной величины X попадает в незаштрихованную область (рис. 32.4).

Следует помнить, что неравенство Чебышёва справедливо как для дискретных, так и для непрерывных случайных величин и полезно на практике лишь тогда, когда $\varepsilon > \sigma$; в противном случае оно дает тривиальную оценку. Например, если выбрать $\varepsilon = \frac{1}{2}\sigma(X)$, то $P\left(|X - M(X)| \geq \frac{1}{2}\sigma(X)\right) \leq \frac{D(X)}{\frac{1}{4}D(X)} = 4$, но и без того очевидно, что вероятность не может быть больше четырех; если же выбрать $\varepsilon = 10\sigma(X)$,

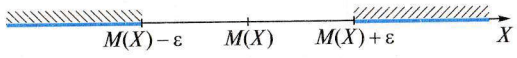

Рис. 32.4

то $P(|X - M(X)| \geq 10\sigma(X)) \leq \frac{D(X)}{100 \cdot D(X)} = 0{,}01$, а это уже достаточно хорошая оценка вероятности.

Неравенство Чебышёва можно применить и в другой форме: $P(|X - M(X)| < \varepsilon) > 1 - \frac{D(X)}{\varepsilon^2}$.

Т32.5 **Теорема Чебышёва.** Пусть имеется бесконечная последовательность $X_1, X_2, \ldots, X_n, \ldots$ случайных величин с одним и тем же математическим ожиданием a ($M(X_1) = M(X_2) = \ldots = M(X_n) = \ldots = a$) и дисперсиями, ограниченными одной и той же постоянной C ($D(X_1) < C$, $D(X_2) < C, \ldots, D(X_n) < C, \ldots$), тогда *каково бы ни было положительное число ε, вероятность события* $\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon$ *стремится к единице при* $n \to \infty$, *т. е.* $\lim\limits_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon\right) = 1$.

Т32.6 **Теорема Бернулли.** Пусть производится n независимых опытов, в каждом из которых с вероятностью p может наступить некоторое событие A. Рассмотрим случайную величину Y_n — число наступлений события A в n опытах. Тогда *каково бы ни было положительное число ε, вероятность события* $\left|\frac{Y_n}{n} - p\right| < \varepsilon$ *стремится к единице при* $n \to \infty$, *т. е.* $\lim\limits_{n\to\infty} P\left(\left|\frac{Y_n}{n} - p\right| < \varepsilon\right) = 1$ ($\frac{Y_n}{n}$ — частота наступления интересующего нас события).

Т32.7 **Центральная предельная теорема Ляпунова.** *При достаточно большом числе n независимых случайных величин $X_1, X_2, \ldots, X_n$, подчиненных каким угодно законам распределения (при соблюдении некоторых ограничений), их сумма будет иметь закон распределения, как угодно близкий к нормальному закону.*

Глава 33. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

33.1. Задачи математической статистики

В математической статистике рассматриваются две основные задачи.

Первая задача состоит в том, чтобы указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате поставленных экспериментов.

Вторая задача заключается в разработке методов анализа статистических данных в зависимости от целей исследования. К ним относятся:

1) оценка неизвестной вероятности события; оценка параметров распределения, вид которого не известен; оценка зависимости случайной величины от одной или нескольких случайных величин;

2) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого не известен.

Современная математическая статистика разрабатывает способы нахождения числа необходимых испытаний до начала исследования (планирование эксперимента) и решает многие другие задачи; ее определяют как науку о принятии решений в условиях неопределенности.

Итак, основная задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

В основе задач, решаемых методами математической статистики, лежит необходимость изучения совокупности однородных объектов, относительно некоторого качественного или количественного признака.

Например, при изучении партии однородных товаров, хранящихся на базе, качественным признаком могут служить сортовые, а количественным — весовые характеристики, изменяющиеся в процессе хранения.

33.2. Основные определения

О Совокупность всех объектов, подчиненных данному признаку, называют ***генеральной совокупностью***, а число таких объектов — *объемом генеральной совокупности*.

На практике, как правило, обследование всех объектов генеральной совокупности не производится в силу излишней трудоемкости такого процесса. Обычно из всей совокупности отбирают ограниченное число объектов, которые и подвергают изучению. Такую случайно отобранную совокупность называют *выборочной совокупностью*, или *выборкой*. Для того чтобы по данным выборки достаточно уверенно характеризовать всю генеральную совокупность, необходимо, чтобы отобранные элементы правильно ее представляли.

О Выборку, достаточно хорошо описывающую всю генеральную совокупность, называют ***репрезентативной (представительной)***.

Для получения репрезентативной выборки необходимо, чтобы все отобранные элементы имели одинаковую вероятность попасть в выборку. В случае большого объема N генеральной совокупности хорошие результаты в этом смысле дает использование таблицы «случайных чисел».

Для того чтобы отобрать, например, 20 объектов из пронумерованной генеральной совокупности, можно открыть любую страницу таблицы «случайных чисел» и выписать подряд 20 случайных чисел. В выборку включают те объекты, номера которых совпали с выписанными случайными числами (случайные числа, превышающие N, при выписывании пропускаются).

О Элементы $x_1, x_2, \ldots, x_n$, попавшие в выборку, называют ***вариантами***, а их количество n — *объемом выборки*. Отобранные элементы располагают обычно в порядке их возрастания: $x^{(1)} \le x^{(2)} \le \ldots \le x^{(n)}$. Такая последовательность вариант называется ***вариационным рядом***. Разность между максимальным и минимальным элементами выборки $z = x^{(n)} - x^{(1)}$ называют *размахом выборки*.

Среди n элементов выборки могут встречаться повторяющиеся. Если, например, элемент x_1 встречается n_1 раз, x_2 – n_2 раз, ..., x_k – n_k раз, то числа $n_1, n_2, \ldots, n_k$ называются *частотами вариант* $x_1, x_2, \ldots, x_k$. Очевидно, что сумма всех частот равна объему выборки, т. е. $n_1 + n_2 + \ldots + n_k = n$, или $\sum_{i=1}^{k} n_i = n$.

О Расположенная в порядке возрастания вариант последовательность пар чисел, составленная из вариант и их частот $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$, называется ***статистическим рядом***, или *статистическим распределением*.

При этом пользуются табличной записью:

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

Иногда вместо частоты варианты n_i требуется использование относительных частот $\omega_i = \frac{n_i}{n}$. Очевидно, что сумма частот всех элементов выборки равна единице:

$$\sum_{i=1}^{k} \omega_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{n_1}{n} + \frac{n_2}{n} + \cdots + \frac{n_k}{n} = \frac{n_1 + n_2 + \ldots + n_k}{n} = 1.$$

П 1. Записать вариационный ряд и статистическое распределение элементов выборки 5, 0, 3, 7, 0, 10, 5, 0, 5, 2, 10, 2, 0, 7, 2, 0, 4, 7, 7, 4 из числа рабочих дней в году, пропущенных по болезни работниками супермаркета. Определить размах выборки.

Решение. Объем выборки $n = 20$. Упорядочив элементы выборки по величине, получим вариационный ряд 0, 0, 0, 0, 0, 2, 2, 2, 3, 4, 4, 5, 5, 5, 7, 7, 7, 7, 10, 10 (размах выборки $z = 10 - 0 = 10$):

x_i	0	2	3	4	5	7	10
n_i	5	3	1	2	3	4	2

При большом объеме выборки для упрощения вычислений ее элементы объединяют в группы (разряды), представляя выборку в виде *группированного статистического ряда* (распределения). Для этого интервал, содержащий все элементы, разбивают на k непересекающихся интервалов. Вычисления значительно упрощаются, если эти интервалы имеют одинаковую длину $h = \frac{z}{k}$. После того как интервалы выбраны, определяются частоты — количество элементов n_i^* выборки, попавших в i-й интервал, а сами элементы, попавшие в этот интервал, считаются равными его середине x_i^* (элемент, совпадающий с верхней границей интервала, относится к последующему интервалу).

Число интервалов k группировки берется в зависимости от объема выборки. При этом следует помнить, что группировка выборки вносит погрешности в дальнейшие вычисления. Эти погрешности растут с уменьшением числа интервалов. В процессе составления группированного статистического распределения подсчитываются также *накопленные частоты* $\tilde{n}_i^*$ (накопленная частота i-го интервала равна сумме частот самого i-го и всех предыдущих интервалов, т. е. $\tilde{n}_3^* = n_1^* + n_2^* + n_3^*$), относительные частоты $\frac{n_i^*}{n}$ и накопленные частоты $\frac{\tilde{n}_i^*}{n} = \sum_{j=1}^{i} \frac{n_j^*}{n}$.

П 2. Известно время недельной загрузки электрических духовых шкафов 50 обследованных предприятий общественного питания в часах:

38 60 41 51 33 42 45 21 53 60 61 35 47 72 41 45 44 56 30 40
60 52 47 46 49 49 14 57 54 59 67 65 39 48 43 60 54 42 59 50
77 47 28 48 58 32 42 58 61 30

Найти размах выборки, число и длину интервалов, а также составить таблицу частот (записать группированное статистическое распределение). Первый интервал взять 14 — 23.

Решение. Будем проводить группировку по интервалам равной длины $h = 23 - 14 = 9$. Размах выборки $z = 77 - 14 = 63$. Тогда необходимое число интервалов $k = \frac{63}{9} = 7$. Результаты группировки сведены в таблицу (табл. 33.1).

Таблица 33.1

Номер интервала i	Границы интервала	Середина интервала x_i^*	Частота n_i^*	Относительная частота n_i^*/n	Накопленная относительная частота $\tilde{n}_i^*/n$
1	14 — 23	18,5	2	0,04	0,04
2	23 — 32	27,5	3	0,06	0,10
3	32 — 41	36,5	6	0,12	0,22
4	41 — 50	45,5	17	0,34	0,56
5	50 — 59	54,5	10	0,20	0,76
6	59 — 68	63,5	9	0,18	0,94
7	68 — 77	72,5	8	0,16	1,00

В целях наглядности строят различные графики статистического распределения. Они помогают лучше представить себе характер распределения элементов выборки, а иногда и сделать предварительные предположения о законе распределения генеральной совокупности.

Такими графиками являются полигон частот и гистограммы.

О ***Полигоном частот*** называют ломаную линию, вершинами которой являются точки (x_1, n_1), (x_2, n_2), ..., (x_k, n_k), определяемые элементами статистического ряда.

Для его построения на оси абсцисс откладываются варианты x_i, а по оси ординат — соответствующие им частоты n_i. Построенные точки (x_i, n_i) соединяются отрезками прямых и получается полигон частот (для группированной выборки полигон частот строится по точкам (x_i^*, n_i^*)).

Для группированной выборки обычно строится гистограмма частот.

О ***Гистограммой частот*** называют ступенчатую фигуру, составленную из прямоугольников, построенных на интервалах так, что площадь каждого прямоугольника численно равна частоте n_i^* варианты x_i^*, расположенной в середине i-го интервала.

Отсюда следует, что площадь гистограммы частот равна объему выборки n. При равных длинах интервалов разбиения h высоты прямоугольников равны $n_i = \frac{n_i^*}{h}$.

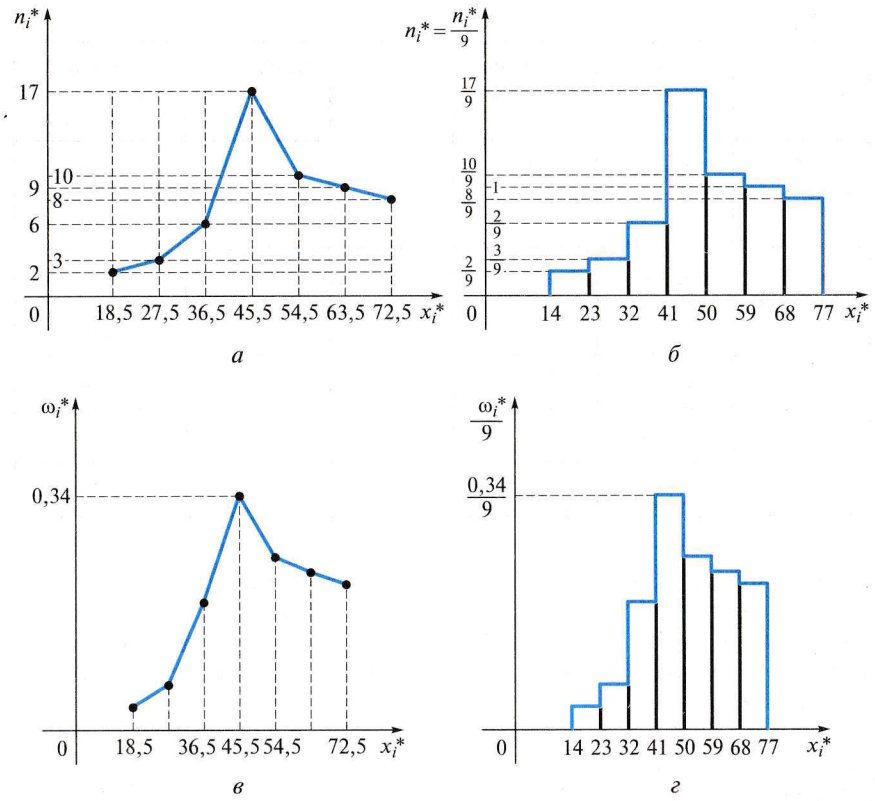

Рис. 33.1

Иногда вместо полигона и гистограммы частот строят *полигон* и *гистограмму относительных частот*, построение которых отличается от вышеописанных лишь тем, что по оси ординат откладывается не частота n_i, а относительная частота $\omega_i = \frac{n_i}{n}$ (высота прямоугольников гистограммы будет равна соответственно $\frac{\omega_i^*}{h}$, а площадь всей гистограммы относительных частот равна единице).

П Построить полигон и гистограмму частот и относительных частот по группированной выборке примера 2.

Решение. По результатам группировки (см. табл. 33.1) строим соответственно полигон частот (рис. 33.1, *а*), гистограмму частот (рис. 33.1, *б*), полигон и гистограмму относительных частот (рис. 33.1, *в*, *г*).

О ***Модой*** (обозначается Мо) называют варианту выборки, имеющую наибольшую частоту.

П По данным статистического распределения выборки найти мо ду (Мо).

x_i	1	3	5	7	9
n_i	1	5	6	5	3

Решение. Поскольку наибольшая частота выборки $n_3 = 6$ принад лежит элементу $x_3 = 5$, то Мо = 5.

О ***Медианой*** (обозначается Ме) называют варианту выборки, де лящую вариационный ряд на две части, равные по числу вариант.

При этом, если объем выборки — число нечетное, т. е. $n = 2k -$ то медианой будет являться средний элемент вариационного ряд Ме = $2k + 1$, если же n — четно, т. е. $n = 2k$, то медианой будет ве личина, равная полусумме двух вариант выборки, расположенны в середине вариационного ряда, т. е. $\text{Ме} = \frac{x_k + x_{k+1}}{2}$.

П По данным вариационных рядов определить медиану:
1) 2, 3, 6, 7, 9; 2) 2, 3, 5, 7; 3) 2, 2, 3, 5, 7, 7, 8, 9, 9, 11.
Решение.
1) Объем выборки $n = 5$ — нечетное число, следовательно, меди ной будет средний элемент, т. е. Ме = 6;
2) объем выборки $n = 4$ — четное число, следовательно, медиано будет полусумма двух средних значений: $\text{Ме} = \frac{x_2 + x_3}{2} = \frac{3+5}{2} = 4$;
3) объем выборки $n = 10$ — четное число $\Rightarrow$ $\text{Ме} = \frac{x_5 + x_6}{2} = \frac{7+7}{2} =$

? Контрольные вопросы

1. Дайте определение и укажите границы применения классическо формулы вероятности.
2. Приведите формулы вычисления основных комбинаторных соотн шений (размещения, перестановки, сочетания).
3. При каких условиях возможно применение формулы Бернулли?
4. Перечислите основные задачи математической статистики.

Задачи для самостоятельного решения

1. В опыте было получено 100 вариантов, составляющих выборочну совокупность, значения которых приведены в таблице (табл. 33.2).
Произведите статистическую обработку результатов опыта:
1) найдите размах выборки, число и длину интервалов; составьте табли частот (запишите группированное статистическое распределение). *Указан* длину интервала задать самостоятельно;