Часть 2
ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Глава I
Выборочный метод

§ 1. Задачи математической статистики. Статистический материал

Пусть требуется определить функцию распределения $F(x)$ некоторой непрерывной случайной величины X. Кроме того, возможно, потребуется определение параметров распределения, таких как, например, математическое ожидание m, среднее квадратическое отклонение σ или дисперсия D. Предположим также, что никакими априорными сведениями об этой случайной величине мы не располагаем. Единственный выход в решении этой задачи состоит в проведении испытаний и получении реализаций этой случайной величины. Пусть в результате проведения n опытов получен ряд значений

$$x_1, x_2, ..., x_n. \tag{1}$$

Ряд (1) значений случайной величины x называется **выборкой из генеральной совокупности** с функцией распределения $F(x)$. Он служит основным статистическим материалом, несущим сведения о случайной величине. Выборка содержит приближенные сведения о генеральной совокупности. Процесс составления выборки называется **выбором**. Процесс выбора должен быть организован так, чтобы выборка хорошо представляла генеральную совокупность, или, как говорят, была **репрезентативной**. Непосредственные результаты измерений располагают в табличной форме (табл. 1) в том порядке, в котором они появились при проведении экспериментов. Такая таблица называется [11] **статистическим** рядом[1].

Содержание понятия генеральной совокупности зависит от конкретной задачи. Если есть ограниченное число объектов, например деталей, и требуется определить их характеристики по выбранному конечному

---

[1] В книге [16] этим термином обозначается последовательность элементов вариационного ряда с указанием частот n_i повторения результатов измерений.

Таблица 1. Пример статистического ряда

Номер опыта	1	2	...	n
Результаты измерения случайной величины $\boldsymbol{X}$	x_1	x_2	...	x_n

количеству объектов, то генеральной совокупностью в этом случае является перечень измеряемых характеристик всех имеющихся объектов. Если изучается случайная величина методом проведения ее испытаний и число испытаний может быть каким угодно, то объем генеральной совокупности можно считать бесконечным.

§ 2. Построение эмпирической функции распределения

Важным этапом статистической обработки выборки является построение эмпирической функции распределения случайной величины X. Для этого требуется сначала провести сортировку результатов измерения, приведенных в табл. 1.1, например по возрастанию значений случайной величины. Если статистический ряд окажется слишком большим, то, возможно, сортировку целесообразно провести с использованием компьютера. В качестве примера построения эмпирической функции распределения рассмотрим статистический ряд из 5 измерений водоизмещения корабля, представленный в табл. 1.

Таблица 1. Статистический ряд измерений водоизмещения корабля

Номер опыта	1	2	3	4	5
Водоизмещение $\boldsymbol{V_i}$, $м_3$	5250	5370	5390	5370	5180

Расположим результаты измерения в порядке возрастания водоизмещения. В результате получится **вариационный** ряд, представленный в табл. 2.

Определение. **Вариационным рядом** называется последовательность всех элементов выборки, расположенных в неубывающем порядке. Одинаковые результаты измерения повторяются. Одновременно изменяется нумерация опытов.

Таблица 2. Отсортированный статистический (вариационный) ряд измерений водоизмещения корабля

Номер опыта	1	2	3	4	5
Водоизмещение $\boldsymbol{V_i}$, $м_3$	5180	5250	5370	5370	5390

Экспериментальный график функции распределения вероятности водоизмещения корабля строится на основе определения самой функции $F(x)$: значение функции есть вероятность того, что в результате одного испытания случайная величина окажется меньше своего аргумента. Заменяя вероятность ее статистической оценкой, получаем, что статистическая функция распределения вычисляется по формуле

$$F^*(x) = \frac{\text{число измерений со значениями} < x}{\text{общее число измерений}}. \qquad (1)$$

В результате проведенной серии экспериментов величина водоизмещения ни разу не оказалась меньше значения 5180 м^3, поэтому ордината экспериментальной функции $F(x)$ будет нулем для значений аргумента, меньших 5180. На промежутке от 5180 до 5250 случайная величина только один раз из 5 приняла значение, меньшее 5250. Поэтому вероятность того, что в результате одного опыта водоизмещение примет значение в указанном промежутке, следует принять равной $\frac{1}{5}$ и т. д. В результате появится **статистическая функция распределения**, представленная на рис. 1.

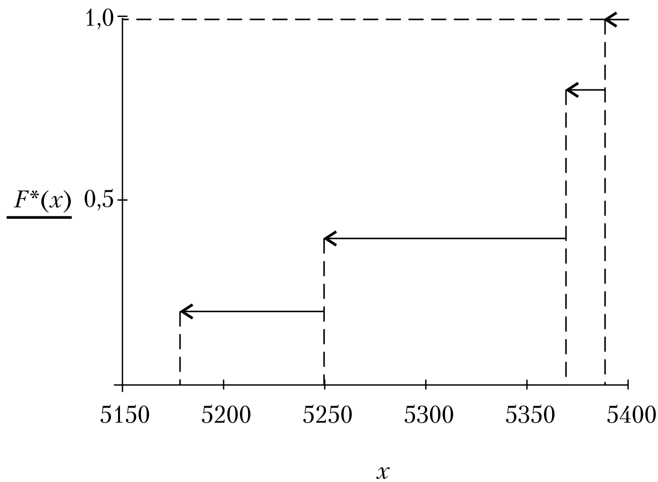

Рис. 1. Статистическая функция распределения

По закону больших чисел в форме Бернулли величина, определяемая формулой (1), сходится по вероятности к своему математическому ожиданию при неограниченном возрастании числа опытов.

Причем в качестве события А (см. ч. 1, п. III.5.4.3) следует считать попадание результата измерения в интервал $(-\infty, x)$.

§ 3. Построение гистограммы

При большом числе измерений использование статистического ряда вызывает затруднения. В этом случае результаты наблюдений объединяют в группы. Их группировка осуществляется предварительным разбиением области изменения случайной величины на промежутки: $\Delta_1 = [x_{\min},\ a_1),\ \Delta_2 = [a_1,\ a_2),\ ...,\ \Delta_k = [a_{k-1},\ x_{\max}]$. Их длины могут быть одинаковыми (h) или различными. Затем подсчитывается число измерений m_i, попавших в интервал i. Совокупность интервалов Δ_i и чисел m_i называется **группированным статистическим рядом** [16]. Для каждой группы сначала вычисляется отношение $\dfrac{m_i}{n}$, где n — общее число результатов измерений. Оно называется частотой[1]. Новый объект, состоящий из групп и частот, называется [11] **статистической совокупностью**. И наконец, для каждой группы вычисляется относительная частота $\dfrac{m_i}{nh}$, при этом получают еще один новый объект, состоящий из совокупности групп и относительных частот. Этот объект удобно представить графически. По оси абсцисс откладывается случайная величина, а по оси ординат — относительная частота, постоянная в промежутке своего определения. Такой график называется **гистограммой** (рис. 1). Площадь каждого прямоугольника, составляющего гистограмму, равна частоте попадания измерений случайной величины в заданный интервал, то есть является оценкой вероятности попадания измерения в заданный интервал. Это означает, что гистограмма представляет собой аналог функции плотности вероятностей. Число групп выбирают таким, чтобы в каждую из них попало достаточно много результатов измерений и число групп тем не менее было достаточно большим.

Кроме гистограммы для наглядного представления дискретной случайной величины (ДСВ) строят **полигон** распределения (рис. 2). Он похож на многоугольник распределения вероятностей ДСВ (см. ч. 1,

---

[1] В литературе существует разнобой в терминологии. В одних книгах [16] частотой называют m_i, отношение $\dfrac{m_i}{n}$ — относительной частотой, а $\dfrac{m_i}{nh}$ — приведенной частотой, в других [1] частотой называют отношение $\dfrac{m_i}{n}$.

гл. III, § 1, п. 1.1), полученный в результате эксперимента. Для построения полигона распределения дискретной случайной величины используют декартову систему координат, в которой по оси абсцисс откладывают середины групп значений СВ, а по оси ординат — относительные частоты ее появления в соответствующей группе, определенные из эксперимента. Полученные соседние точки соединяют отрезками прямых.

Пример. Последовательно измерены 100 амплитуд отклонений корабля θ_m на правый и левый борт в некоторых реальных условиях плавания [11]. Отклонения лежали между –20° и +20°. Результаты измерений приведены в табл. 1 при условии, что интервал для подсчета частот принят равным 5°.

Для построения гистограммы $G(x)$ (рис. 1) использованы первая и пятая, а при построении функции распределения F^* (рис. 3) — первая и шестая строки табл. 1.

Таблица 1. Статистическая совокупность для амплитуд отклонений корабля

Группы, град	–20 ÷ –15	–15 ÷ –10	–10 ÷ –5	–5 ÷ 0	0 ÷ 5	5 ÷ 10	10 ÷ 15	15 ÷ 20
Середины групп $\tilde{x}_i$, град	–17,5	–12,5	–7,5	–2,5	2,5	7,5	12,5	17,5
Число отклонений m_i в группе	2	8	17	24	26	13	6	4
Частота p^*	0,02	0,08	0,17	0,24	0,26	0,13	0,06	0,04
Относительная частота, $G \equiv \frac{p^*}{5}$; $\frac{1}{\text{град}}$	0,004	0,016	0,034	0,048	0,052	0,026	0,012	0,008
F^*	0,02	0,1	0,27	0,51	0,77	0,9	0,98	1

Рекомендации по выбору числа групп.

Очевидно, что для каждой случайной величины и для определенного объема выборки существует оптимальное число групп, включающих в себя диапазон изменения случайной величины, а также их размер и место расположения. Если взять много групп, то гистограмма будет иметь гребенчатую форму. Если взять мало групп, то гистограмма будет мало похожа на функцию плотности вероятностей.

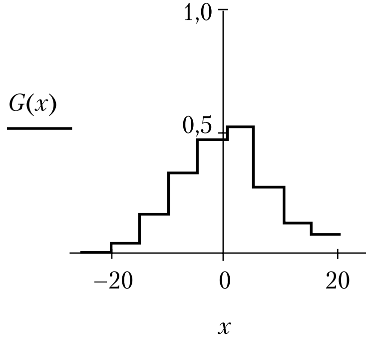

Рис. 1. Гистограмма амплитуд углов отклонений корабля

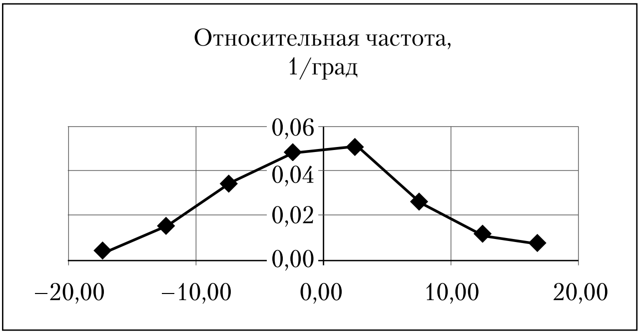

Рис. 2. Полигон амплитуд углов отклонений корабля

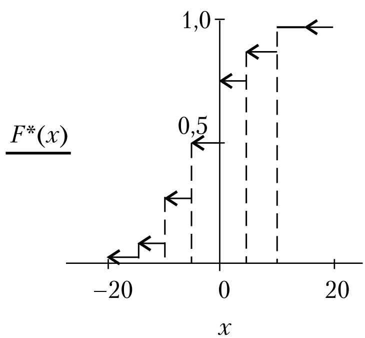

Рис. 3. Статистическая функция распределения амплитуд углов отклонений корабля

Возникает также следующий вопрос: каким должно быть число групп — четным или нечетным? Если выбирать число групп четным, то трудно будет уловить максимум функции плотности вероятности. Поэтому лучше выбирать это число нечетным. Рекомендации по выбору числа групп предлагаются в работе И. У. Алексеевой (журнал «Измерительная техника». № 5. 1975). Для законов, близких к равномерному, число k групп определяется формулой

$$k = \frac{4}{\varepsilon} \ln \frac{n}{10},$$

где $\varepsilon = \frac{\sigma^2}{\sqrt{\mu_4}}$ — контрэксцесс, σ — среднее квадратичное отклонение, μ_4 — четвертый центральный момент. Однако обычно упрощают эту формулу и используют следующее соотношение:

$$k = 5 \ln \frac{n}{10}. \qquad (1)$$

Для законов, имеющих ярко выраженный экстремум, число k групп увеличивается в 2 раза и оценивается выражением

$$k = 10 \ln \frac{n}{10}. \qquad (2)$$

Для определения числа k применяется также формула

$$k = \log_2 n + 1. \qquad (3)$$

Пример. Произведем расчет числа k по формулам (1–3) для $n = 100$:

$k \approx 11,5$ — по формуле (1);

$k \approx 23$ — по формуле (2);

$k \approx 7,6$ — по формуле (3).

Контрольные вопросы

1. Что такое генеральная совокупность и выборка из нее?
2. Какая выборка называется репрезентативной?
3. Что такое статистический, вариационный ряд?
4. Как составляется статистическая функция распределения?
5. Как составляется гистограмма?
6. Каково назначение гистограммы?
7. Как составляется полигон частот?