МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ПРЕДМЕТ И ЗАДАЧИ

Теория вероятностей изучает математические модели массовых случайных явлений – вероятностные модели. При этом сама математическая модель является заданной. Если изучается некоторое случайное событие A, то известно $P(A)$, если речь идет о случайной величине X, то известен закон распределения вероятностей в какой-либо форме.

В практических задачах исследования случайных явлений, а в реальности таковыми являются практически все явления и процессы, их характеристики, как правило, неизвестны, но имеются некоторые экспериментальные (опытные) данные – результаты наблюдений, измерений, испытаний. **Требуется на основании этих данных построить подходящую теоретико-вероятностную модель изучаемого явления. Это и является основной задачей математической статистики.**

Определение. **Математическая статистика** – раздел математики, изучающий методы сбора, систематизации, обработки и интерпретации результатов наблюдений (опытов, измерений) с целью построения теоретико-вероятностной модели изучаемого случайного явления.

Математическая статистика опирается на теорию вероятностей. Обе математические дисциплины изучают массовые случайные явления. Связующим звеном между ними являются предельные теоремы вероятности, устанавливающих связь между теоретическими и экспериментальными характеристиками случайных величин при большом числе испытаний над ними.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ПРЕДМЕТ И ЗАДАЧИ

Если теория вероятностей предоставляет исследователю набор математических моделей, предназначенных для описания закономерностей в поведении реальных явлений или систем, **то средства математической статистики** позволяют подбирать среди множества возможных теоретико-вероятностных моделей ту, которая в определенном смысле наилучшим образом соответствует экспериментальным данным, характеризующим реальное поведение конкретной исследуемой системы.

Предметом математической статистики являются результаты наблюдений случайных событий, явлений и процессов. Эти результаты называются статистическими данными.

Первая задача математической статистики состоит в упорядочении и систематизации полученных данных, представлении их в удобном для анализа виде.

Вторая задача – оценить хотя бы приблизительно интересующие нас характеристики наблюдаемой случайной величины. Например, дать оценку неизвестной вероятности события, оценку закона распределения или числовых характеристик исследуемой случайной величины.

Третья задача – проверка статистических гипотез. Например, выдвигается гипотеза о нормальном распределении наблюдаемой случайной величины. С использованием выбранного критерия необходимо принять решение о принятии или отвержении этой гипотезы.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ПРЕДМЕТ И ЗАДАЧИ

Одной из важнейших задач математической статистики является разработка методов, позволяющих по результатам обработки статистических данных делать выводы о законе распределения и характеристиках наблюдаемой случайной величины.

Методы математической статистики широко применяются в различных отраслях естествознания и техники: в теоретической физике, радиотехнике, общей теории связи, теории автоматического управления и т.д. Результаты исследования статистических данных используются для принятия решений в условиях неопределенности.

Для обработки статистических данных созданы специальные программные пакеты, которые выполняют трудоемкую работу по расчету различных статистик, построению таблиц и графиков. Функции статистической обработки данных имеются в пакетах Mathcad и Matlab. Для обработки статистических данных широко используется табличный процессор Excel.

Таким образом, математическая статистика занимается как статистическим описанием результатов опытов или наблюдений, так и построением и проверкой подходящих математических моделей изучаемых случайных явлений, содержащих понятие вероятности. Такие модели называются теоретико-вероятностными или статистическими.

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Определение. **Совокупность всех подлежащих изучению объектов или всех возможных мыслимых результатов наблюдений (измерений, испытаний)**, производимых в неизменных условиях над одним объектом, **называется генеральной совокупностью**.

Введенное понятие генеральной совокупности, как совокупности всех мыслимых результатов наблюдений случайной величины, будем считать синонимом понятия случайной величины и в дальнейшем не различать эти понятия.

Сплошное обследование генеральной совокупности, как правило, экономически нецелесообразно, а во многих случаях невозможно. В статистических исследованиях используется **выборочный метод**, в основе которого лежит понятие выборки.

Определение. **Выборкой или выборочной совокупностью называются** результаты ограниченного числа наблюдений (измерений), проведенных над исследуемой случайной величиной – генеральной совокупностью. Число наблюдений, образующих выборку, называется **объемом выборки.**

Метод статистического исследования, состоящий в том, что на основе изучения выборочной совокупности делается заключение о свойствах всей генеральной совокупности, называется выборочным методом.

ПОНЯТИЕ ВЫБОРКИ

К выборкам в математической статистике предъявляется **требование репрезентативности (представительности)**. Это означает:

– каждое значение из генеральной совокупности должно иметь одинаковую вероятность попасть в выборку (полностью случайный отбор), причем значения в выборке должны быть взаимно независимы;

– выборка должна иметь достаточно большой объем ($n > 40–50$).

В статистике **интерпретация выборки и ее отдельных элементов** допускает в зависимости от контекста два различных варианта.

При первом (практическом) варианте интерпретации под выборкой понимаются фактически наблюденные в данном конкретном эксперименте значения исследуемой случайной величины, т.е. конкретные числа.

В соответствии со вторым (гипотетическим) вариантом интерпретации **под выборкой понимается последовательность** $X_1, X_2, ..., X_n$ **независимых одинаково распределенных случайных величин, распределение каждой из которых совпадает с распределением генеральной совокупности**.

Значения выборки, полученные в результате наблюдений (измерений, испытаний), называют реализацией выборки и обозначают строчными буквами $x_1, x_2, ..., x_n$. Конкретное численное значение элемента выборки x_i, полученной в результате данного наблюдения (измерения), есть одно из возможных значений случайной величины X_i.

Результаты n наблюдений $x_1, x_2, ..., x_n$ генеральной совокупности (случайной величины X), записанные в порядке их получения, обычно труднообозримы и неудобны для дальнейшего анализа. **Задачей статистического описания выборки** является получение такого ее представления, которое позволяет выявить характерные особенности совокупности исходных данных. Первый шаг к осмыслению данных – это их упорядочение, расположение в порядке возрастания значений.

Определение. **Вариационным рядом выборки $x_1, x_2, ..., x_n$ называется способ ее записи**, при котором элементы упорядочиваются по величине, т.е. записываются в виде неубывающей последовательности $x^{(1)}, x^{(2)}, ..., x^{(n)}$, где $x^{(1)} \le x^{(2)} \le ... \le x^{(n)}$.

Отдельные элементы вариационного ряда называются **порядковыми статистиками.** Примерами порядковых статистик является наименьшее значение выборки, равное первому элементу вариационного ряда $x^{(1)}$; наибольшее значение выборки, равное $x^{(n)}$; выборочная медиана Me^*, равная значению элемента вариационного ряда с номером $(n+1)/2$, если n нечетно, и среднему арифметическому значений вариационного ряда с номерами $n/2$ и $(n/2+1)$.

Разность между наименьшим и наибольшим значениями выборки $\Delta = x^{(n)} - x^{(1)}$ называется **размахом выборки**.

Пусть выборка (вариационный ряд выборки) содержит k **различных** чисел $x_1, x_2, ..., x_k$, $(k \le n)$. Значения x_1, x_2, ..., x_k называются **вариантами** выборки.

Пусть варианта x_1 встречается в выборке n_1 раз, варианта x_2 – n_2 раз, …, варианта x_k – n_k. При этом $n_1 + n_2 + ... + n_k = n$.

Число n_i называется **частотой** варианты x_i. Очевидно, что $\sum_{i=1}^{k} n_i = n$.

Относительной частотой или **частостью** варианты x_i называется отношение числа n_i к объему выборки: $W_i = n_i / n$. Очевидно, что $\sum_{i=1}^{k} W_i = 1$.

Определение. **Статистическим рядом или статистическим распределением выборки** называется перечень её вариант и соответствующих им частот (или частостей).

Статистический ряд или статистическое распределение записывается в виде таблицы из трех строк, первая строка которой содержит варианты x_i, а вторая – их частоты n_i, третья – частости W_i.

Пример 1. В процессе испытаний цифровой линии передачи данных проведена серия измерений числа ошибок при приеме сообщений. Результаты измерений представлены рядом чисел – выборкой объема n=6:

5, 3, 6, 3, 5, 4.

Представим эту выборку в виде вариационного ряда, определим порядковые статистики и построим статистический ряд распределения.

1) Вариационный ряд: 3, 3, 4, 5, 5, 6.

Порядковые статистики: наименьшее значение равно 3; наибольшее значение равно 6; размах выборки $\Delta x = 6 - 3 = 3$; так как n четное, то значение выборочной медианы равно полусумме значений элементов вариационного ряда с номерами $n/2$ и $(n/2+1)$: $Me^* = (5-4)/2 = 4{,}5$.

2) Статистический ряд. Вариантами являются числа $x_1 = 3$, $x_2 = 4$, $x_3 = 5$, $x_4 = 6$. Подсчитав для каждой из вариант частоты и частости, получаем следующие статистические ряды распределения числа ошибок при приеме сообщений:

x_i	3	4	5	6
n_i	2	1	2	1
W_i	1/3	1/6	1/3	1/6

Контроль: $\sum_{i=1}^{k} n_i = 6 = n$; $\sum_{i=1}^{k} W_i = 1$.

ИНТЕРВАЛЬНЫЙ СТАТИСТИЧЕСКИЙ РЯД

В случае, когда число вариант наблюдаемой дискретной СВ велико или исследуемая СВ является непрерывной, то результаты наблюдений (измерений) представляют в виде **интервального (группированного) статистического ряда**.

Для этого диапазон значений выборки (размах выборки) разбивается на M непересекающихся, обычно равных по длине интервалов (разрядов). Рекомендуемое число интервалов, на которые разбивается диапазон значений выборки объема n, определяется по формуле Стерджеса:

$$M \approx 1 + \operatorname{int}(3{,}32 \cdot \lg n) = 1 + \log_2 n, \ (\text{здесь } \operatorname{int}(a) - \text{целая часть числа } a).$$

При этом длина интервала h равна отношению размаха выборки числу интервалов $h = \Delta / M$, начало первого интервала $a_1 = x_{(1)} - h/2$, конец первого или начало второго интервала $a_2 = a_1 + h$, конец второго или начало третьего $a_3 = a_2 + h$ и т.д.

После того как частичные интервалы выбраны, определяют частоты – число n_i элементов выборки, попавших в i-тый интервал. При этом элемент выборки, значение которого совпадает с верхней границей интервала, относится к следующему интервалу. По полученным для каждого интервала значениям частот n_i рассчитывают частости $W_i = n_i / n$.

Интервальный статистический ряд представляет собой таблицу, в верхнюю строку которой заносятся значения интервалов (начало – конец), во второй строке указываются значения середин интервалов, в третью и четвертую строки записываются соответствующие интервалам значения частот и частостей.

ИНТЕРВАЛЬНЫЙ СТАТИСТИЧЕСКИЙ РЯД

Пример 2. При испытаниях источника питания 5 В проведено 100 измерений выходного напряжения. Результаты измерений представлены в таблице 1.

Таблица 1 – Результаты измерений напряжения источника питания

4,9	4,9	4,7	5,6	4,8	4,8	4,8	4,9	5,0	5,0	4,7	4,7	4,7	5,0	4,8	5,2	4,9	4,7	5,1	5,0
5,1	4,9	5,3	5,0	4,7	4,6	4,7	5,0	5,3	4,7	5,3	5,0	5,0	4,9	5,0	5,3	5,1	4,7	5,1	5,1
5,0	5,0	4,9	5,0	4,6	5,2	4,9	5,2	5,0	5,0	5,1	5,2	4,8	5,3	5,3	5,2	4,6	5,3	5,3	4,8
4,6	5,1	5,0	4,9	4,8	5,1	4,7	4,9	5,0	5,1	4,9	5,1	5,3	4,7	4,9	5,2	5,0	5,0	5,3	5,1
4,5	5,2	5,2	4,9	5,1	5,1	4,9	5,3	5,0	4,9	5,2	5,6	5,0	5,0	5,1	5,0	4,9	5,0	4,8	5,0

Требуется построить интервальный статистический ряд.

1) Строим вариационный ряд (приведен в таблице 2).

Таблица 2 – Вариационный ряд

4,5	4,6	4,6	4,6	4,6	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,8	4,8	4,8	4,8
4,8	4,8	4,8	4,8	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9	4,9
5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0
5,0	5,0	5,0	5,0	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,1	5,2	5,2
5,2	5,2	5,2	5,2	5,2	5,2	5,2	5,3	5,3	5,3	5,3	5,3	5,3	5,3	5,3	5,3	5,3	5,3	5,5	5,5

По вариационному ряду определяем размах выборки: $\Delta = x_{(100)} - x_{(1)} = 5,5 - 4,5 = 1,0$.

2) Определяем по формуле Стерджеса число интервалов для объема выборки n=100: $M \approx 1 + \mathrm{int}(3,32 \cdot \lg 100) = 7$. Принимаем M=5. Тогда длина интервала $h = \Delta / M = 0,2$.

3) Определяем границы интервалов. За начало 1-ого интервала примем $a_1 = x_{(1)} = 4,5$, тогда конец 1-ого интервала и начало 2-ого будет равно $a_2 = x_{(1)} + h = 4,7$. Аналогичным образом определяем границы других интервалов.

ИНТЕРВАЛЬНЫЙ СТАТИСТИЧЕСКИЙ РЯД

4) Подсчитываем количества элементов выборки, попавших в каждый из интервалов, – частоты n_i для каждого из пяти интервалов вычисляем частости (относительные частоты) $W_i = n_i / n$.

5) По полученным значениям строим интервальный статистический ряд:

Интервалы напряжения, В	4,5 - 4,7	4,7 - 4,9	4,9 - 5,1	5,1 - 5,3	5,3 - 5,5
Середины интервалов, В	4,6	4,8	5,0	5,2	5,4
Частота	5	19	40	24	12
Частость	0,05	0,19	0,40	0,24	0,12

ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

При построении вероятностных характеристик случайной величины X используется всё множество её значений. Такие характеристики называют теоретическими. Характеристики, построенные на основании выборочных данных, называют эмпирическими или выборочными.

Пусть имеем выборку $x_1,\ x_2,\ \ldots,\ x_n$.

Определение. **Эмпирической функцией распределения** или функцией распределения выборки называется функция

$$F_n^*(x) = \frac{n_x}{n}$$

где n_x – число элементов выборки, значения которых меньше x, n – объём выборки,

Отличие теоретической функции распределения от эмпирической заключается в том, что теоретическая определяет вероятность события $\{X < x\}$, а эмпирическая функция распределения определяет относительную частоту (частость) этого же события, являющуюся оценкой его вероятности при проведении n экспериментов.

Эмпирическая функция распределения определяется по значениям частот или относительных частот соотношением

$$F_n^*(x) = \frac{1}{n}\sum_{x_i<x} n_i = \sum_{x_i<x} W_i\ ,$$

где суммируются частоты тех элементов выборки, для которых выполняется неравенство $x_i < x$.

ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами теоретической функции распределения $F(x)$, а именно:

1) $0 \le F_n^*(x) \le 1$; 2) $F_n^*(x)$ – неубывающая функция; 3) если $x_{\min}$ есть наименьшее значение в выборке, то $F_n^*(x) = 0$ при $x \le x_{\min}$, и если $x_{\max}$ – наибольшее значение, то $F_n^*(x) = 1$ при $x > x_{\max}$.

Пример 3. Построить эмпирическую функцию распределения, используя результаты примера 1.

В примере 1 получен следующий статистический ряд относительных частот:

x_i	3	4	5	6
W_i	1/3	1/6	1/3	1/6

Используя формулу для эмпирической функции распределения $F_n^*(x) = \sum_{x_i < x} W_i$, имеем

$$F_n^*(x) = \begin{cases} 0, & \text{при } x \le 3, \\ 1/3, & \text{при } 3 < x \le 4, \\ 1/2, & \text{при } 4 < x \le 5, \\ 5/6 & \text{при } 5 < x \le 6, \\ 1 & \text{при } x > 6. \end{cases}$$

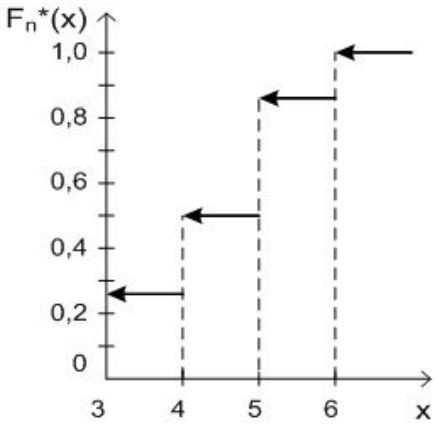

Рисунок – График функции распределения

ГРАФИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Для наглядного графического представления статистического распределения дискретной случайной величины используют полигон частот или относительных частот.

Полигоном частот называют ломаную, отрезки которой соединяют точки (x_1, n_1), (x_2, n_2), …, (x_k, n_k), где x_i – варианты выборки и n_i – соответствующие им частоты.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки (x_1, W_1), (x_2, W_2), …, (x_k, W_k), где x_i – варианты выборки и W_i – соответствующие им относительные частоты.

Значения вариант откладывают по оси абсцисс, значения частот – по оси ординат. На рисунках 1 и 2 показаны полигоны частот и относительных частот статистического распределения ошибок из примера 1.

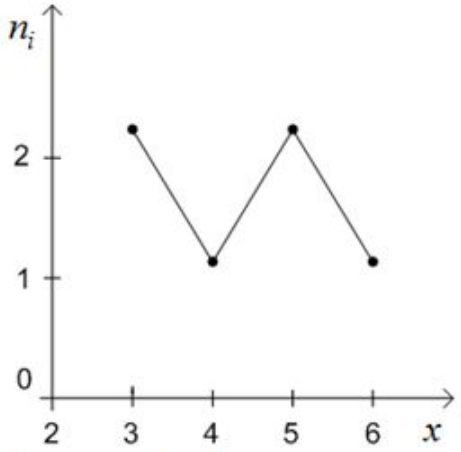

Рисунок 1 – Полигон частот

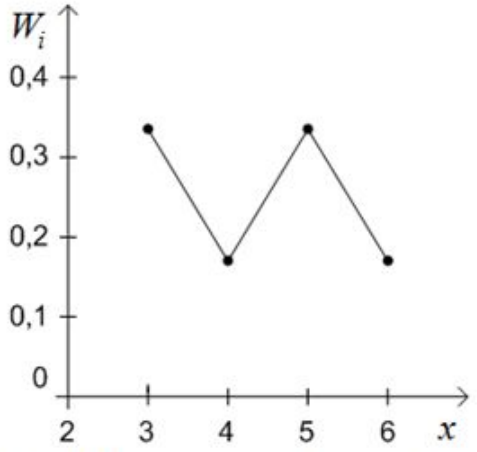

Рисунок 2 – Полигон относительных частот

ГРАФИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

В качестве графического представления выборки из непрерывной случайной величины используется гистограмма, которая строится по интервальному статистическому ряду.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высоты равны отношению n_i/h – **плотности частоты**. Площадь гистограммы частот равна сумме всех частот, т.е. объему выборки.

Гистограммой относительных частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высоты равны отношению W_i/h – плотности относительной частоты (**плотности частости**). Площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.

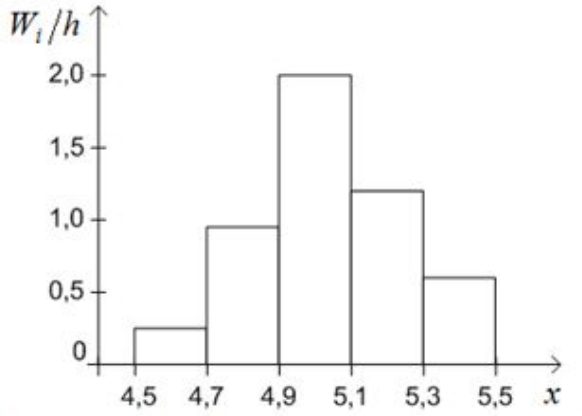

Рисунок 3 – Гистограмма частостей

На рисунке 3 приведена гистограмма относительных частот выборки из примера 2. При построении гистограммы использован интервальный статистический ряд распределения выборки.

Гистограмма относительных частот (частостей) является статистическим аналогом плотности распределения случайной величины.

ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Для выборки можно определить ряд числовых характеристик, аналогичным тем, что определялись в теории вероятностей для случайных величин.

Пусть статистическое распределение выборки объема n из дискретной случайной величины имеет вид

x_i	x_1	x_2	x_3	...	x_k
n_i	n_1	n_2	n_3	...	n_k

Здесь x_i - варианты выборки, n_i - соответствующие им частоты.

Выборочным средним $\overline{x}$ называется среднее арифметическое всех элементов выборки:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \ . \qquad (*)$$

Выборочное среднее можно записать и так:

$$\overline{x} = \sum_{i=1}^{k} x_i W_i \ ,$$

где W_i – частость. Для обозначения выборочного среднего используются и такие символы, как $m_x^*, M^*(X)$.

В случае представления выборки интервальным статистическим рядом в формуле (*) и формулах, приводимых ниже, в качестве x_i берут середины интервалов, а в качестве n_i – соответствующие этим интервалам частоты.

Выборочной дисперсией называется среднее арифметическое квадратов отклонений вариант выборки от выборочного среднего:

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \text{ или, что то же самое } D_{\text{в}} = \sum_{i=1}^{k}(x_i - \bar{x})^2 W_i .$$

Легко показать, что выборочная дисперсия может быть вычислена также по формуле

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\bar{x})^2 .$$

Выборочное среднее квадратическое отклонение выборки определяется формулой $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$.

При решении практических задач используется так называемая **исправленная выборочная дисперсия**, обозначаемая как S^2 вычисляемая по формуле

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i .$$

Исправленная выборочная дисперсия связана с выборочной дисперсией следующим выражением: $S^2 = \frac{n}{n-1} D_{\text{в}}$. Величина $S = \sqrt{S^2}$ называется исправленным средним квадратическим отклонением.