Глава 3. Элементы математической статистики

Введение

Математическая статистика – это раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов.

Основной целью математической статистики является разработка научно обоснованных выводов о массовых случайных явлениях и процессах на основе статистических данных.

Основные задачи математической статистики:

1. Изучение большого числа объектов по сравнительно малому числу выбранных случайно объектов (*выборочный метод*).
2. Нахождение приближенных значений параметров закона распределения (*статистическая оценка параметров распределения*).
3. Проверка гипотез о законе распределения (*статистическая оценка гипотез статистического распределения*).
4. Установление формы и силы связи между случайными величинами (*корреляционный и регрессионный анализ*).

§1. Генеральная и выборочная совокупности. Статистическое распределение выборки

Пусть из большого числа совокупности каких-то объектов случайно отбирается небольшое число объектов для того, чтобы охарактеризовать данную совокупность объектов по некоторому признаку (количественному или качественному). Такой метод обследования совокупности объектов называется *выборочным методом*.

Определение. *Генеральной совокупностью* называется вся обширная совокупность изучаемых объектов.

Число объектов генеральной совокупности называется ее объемом, обозначается N.

Определение. *Выборочной совокупностью* или *выборкой* называется совокупность случайно отобранных объектов.

Число объектов выборки называется ее объемом, обозначается n.

Определение. Выборка называется *репрезентативной* (представительной), если она наилучшим образом отражает генеральную совокупность.

Пусть из генеральной совокупности объема N с количественным признаком X извлечена выборка объема n.

Определение. Наблюдаемые числовые значения признака X называются *вариантами*.

Обозначаются варианты $x_1, x_2, \ldots, x_k$ или x_i, $i = \overline{1,k}$.

Определение. Число, показывающее сколько раз варианта x_i встречается в наблюдении, называется *частотой* и обозначается n_i, а отношение $\frac{n_i}{n}$ – относительной частотой варианты и обозначается w_i. При этом $\sum_{i=1}^{k} n_i = n$ и $\sum_{i=1}^{k} w_i = 1$.

Определение. Последовательность вариант, записанных в возрастающем порядке, называется *вариационным рядом*.

Вариационный ряд может быть *дискретным,* если исследуемый признак X является дискретной случайной величиной и *интервальным*, если – непрерывной.

Определение. Совокупность вариационного ряда с соответствующими частотами (относительными частотами) называется *статистическим распределением частот (относительных частот)* выборки.

Статистическое распределение может быть представлено тремя способами: *табличным, графическим и аналитическим*.

Табличный способ. Статистическое распределение выборки имеет вид таблицы, первой строкой которой является вариационный ряд, а второй соответствующие частоты (относительные частоты).

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

$\sum_{i=1}^{k} n_i = n$

При большом объеме выборки (или в случае непрерывной случайной величины) статистическое распределение получают в виде перечня интервалов и соответствующих частот или относительных частот (будет подробно рассмотрено позднее в §4). При этом подсчитывают сумму частот вариант, попавших в каждый интервал. Значения, попавшие на границу интервала, относят либо к левому, либо к правому, либо делят (пополам при четном числе, при нечетном на единицу больше к одной из границ).

§2. Графические характеристики выборки

Графический способ. Для наглядного изображения дискретного распределения используется *полигон*, а интервального – *гистограмма*.

Определение. *Полигоном* частот (или относительных частот) называется ломаная линия, соединяющая точки M_i (x_i,n_i) (или N_i (x_i, w_i)), построенные в прямоугольной системе координат.

Пример. Построить полигон относительных частот

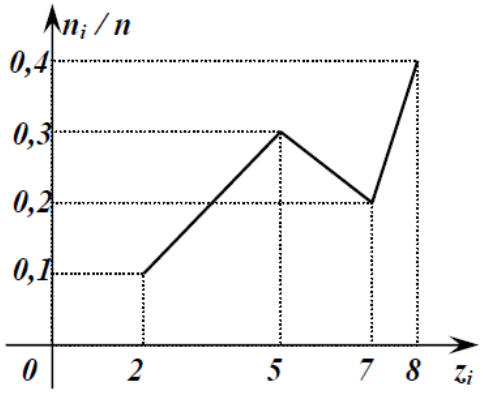
x_i	2	5	7	8
n_i	1	3	2	4

Решение. Объём выборки ***n = 1 + 3 + 2 + 4 = 10***.

Добавим в таблицу относительные частоты

x_i	2	5	7	8
n_i	1	3	2	4
w_i	0,1	0,3	0,2	0,4

Полигон относительных частот



Здесь z_i означает x_i

Определение. *Гистограммой* частот (или относительных частот) называется плоская ступенчатая фигура, состоящая из прямоугольников, основаниями которых являются частичные интервалы (x'_i, x'_{i+1}) длины h, а высоты вычисляются по формуле $\frac{n_i}{h}$ (или $\frac{w_i}{h}$), где h – длина интервала, n_i - сумма частот

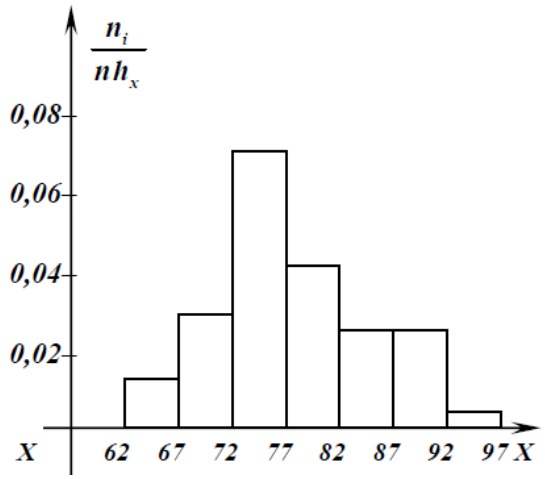
вариант, попавших в данный интервал. Величина $\frac{n_i}{h}$ (или $\frac{w_i}{h}$), называется плотностью частоты (или относительной частоты).

Замечание . В гистограмме площади прямоугольников равны частотам соответствующих интервалов $S_i = h\frac{n_i}{h} = n_i$, а сумма площадей $S = \sum_{i=1}^{k} S_i = n$ – объему выборки.

Пример. Построить гистограмму относительных частот

Интервалы	(62,67)	(67,72)	(72,77)	(77,82)	(82,87)	(87,92)	(92,97)
n_i	3	7	17	10	6	6	1
w_i	0,06	0,14	0,34	0,20	0,12	0,12	0,02
$\frac{w_i}{h}$	0,012	0,028	0,068	0,040	0,024	0,024	0,004

Объем выборки n=50.



Аналитический способ. Пусть дано статистическое распределение частот выборки в виде таблицы:

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

Рассмотрим варианту x_i и совокупность всех вариант меньших, чем x_i. Эта совокупность есть сумма относительных частот $\frac{x_i}{n}$. С изменением x_i будет изменяться и дробь, следовательно, данная дробь является функцией от x_i .

Определение. *Эмпирической функцией* распределения выборки называется функция $F^*(x)$, которая каждому числу x, ставит в соответствие относительную частоту $\frac{n_x}{n}$ события $X<x$: $F^*(x) = \frac{n_x}{n}$, где n_x – число вариант, меньших x; n – объем выборки.

Замечание. Эмпирическая функция $F^*(x)$ служит аналогом функции распределения случайной величины $F(x)$, поэтому имеет такие же свойства.

Пример. Использует статистическое распределение из примера для полигона

x_i	2	5	7	8

n_i	1	3	2	4

Построим эмпирическую функцию распределения $F^*(x)$.

$F^*(x)=\frac{n_x}{n}$, n=10

$-\infty<x\le 2$ $\quad F^*(x)=\frac{0}{10}=0$

$2<x\le 5$ $\quad F^*(x)=\frac{1}{10}=0{,}1$

$5<x\le 7$ $\quad F^*(x)=\frac{1+3}{10}=0{,}4$

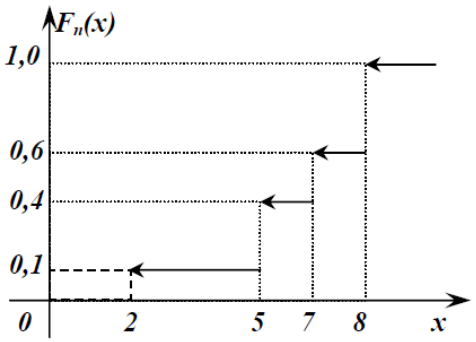
$7<x\le 8$ $\quad F^*(x)=\frac{1+3+2}{10}=0{,}6$

$8<x\le\infty$ $\quad F^*(x)=\frac{1+3+2+4}{10}=1$

Итого

$$F_n(x)=\begin{cases}0, & x\le 2\\ 0{,}1, & 2<x\le 5\\ 0{,}4, & 5<x\le 7\\ 0{,}6, & 7<x\le 8\\ 1, & x>8\end{cases}.$$

Здесь $F_n(x)=F(x)$



§3. Некоторые простейшие характеристики вариационного ряда

Определение. *Размах вариации* – разность между максимальным и минимальным значением признака: $R=x_{max}-x_{min}$

Определение. Модой M_0 называется величина признака (варианта), которая чаще всего встречается в данной совокупности. В вариационном ряду это будет варианта, имеющая наибольшую частоту.

Определение. Медиана m_e – это варианта, расположенная в середине упорядоченного вариационного ряда. Медиана делит ряд пополам, по обе стороны от нее находится одинаковое количество единиц совокупности.

Пример.

x_i	0	1	2	3	4	5
n_i	10	30	75	35	20	15

n=185

Мода M_0=2 - варианта с наибольшей частотой.

Размах R=5-0 =5

Чтобы найти медиану в дискретном вариационном ряде с нечетным объемом, нужно n разделить пополам и к полученному результату добавить ½:

185/2 + ½ = 93, т.е. 93-я варианта, которая делит упорядоченный ряд пополам. Сумма частот 1-й и 2-й вариант равна 40. Ясно, что здесь 93 варианты нет. Если прибавить к 40 частоту 3-й варианты, то получим сумму, равную 40 + 75 = 115. Следовательно, 93-я варианта соответствует третьему значению варьирующего признака, и медиана m_e =2.

Мода и медиана в данном примере совпали. Если бы у нас была четная сумма частот (например, 184), то, применяя указанную выше формулу, получим номер медианной варианты, 184/2 + ½ =92,5. Поскольку варианты с дробным номером не существует, полученный результат указывает, что медиана находится посередине между 92 и 93 вариантами.

§4. Группированный статистический ряд

При большом объёме выборки рекомендуется производить группировку данных, представляя результаты наблюдений в виде *группированного статистического ряда*. Для этого промежуток $[z_1, z_k]$, содержащий все элементы выборки, разбивается на r частичных непересекающихся интервалов $[a_{i-1}, a_i)$, $i = 1, 2, ..., r$. Вычисления упрощаются, если частичные интервалы имеют одинаковую длину (*шаг разбиения*) $h = \frac{z_k - z_1}{r} = \frac{R}{r}$, тогда $a_0 = z_1$, $a_i = a_{i-1} + h$, $(i = 1, 2, ..., r)$. В дальнейшем рассматривается именно этот случай.

Рекомендуемое число r интервалов зависит от объёма выборки n и может быть определено по формуле

$r = \sqrt{n}$ в простейшем случае. Есть и другие формулы для нахождения количества интервалов.

Замечание. Для того чтобы шаг разбиения был более удобным, допускается его небольшое увеличение. Например, пусть $z_1 = 3{,}1$, $R = 15{,}1$, $r = 8$. Тогда $h = \frac{R}{r} = \frac{15{,}1}{8} = 1{,}8875$. Если взять более удобное значение $h = 2$, то расширение промежутка разбиения составит $h \cdot r - R = 2 \cdot 8 - 15{,}1 = 0{,}9$.

При определении границ интервалов $[a_{i-1}, a_i)$ рекомендуется сдвигать левую границу a_0 первого интервала влево примерно на половину расширения, например, в точку $z_1 - 0{,}4 = 3{,}1 - 0{,}4 = 2{,}7$. Тогда правая граница a_8 последнего, восьмого, интервала окажется в точке $z_1 + R + 0{,}5 = 3{,}1 + 15{,}1 + 0{,}5 = 18{,}7$.

В данном примере получаются следующие интервалы:

$[2{,}7; 4{,}7)$, $[4{,}7; 6{,}7)$, $[6{,}7; 8{,}7)$, $[8{,}7; 10{,}7)$, $[10{,}7; 12{,}7)$, $[12{,}7; 14{,}7)$,

$[14,7;\ 16,7)$, $[16,7;\ 18,7)$.

После того как частные интервалы $[a_{i-1},\ a_i)$ выбраны, определяют частоты – количество n_i^* элементов выборки, попавших в i-й интервал (элемент выборки, совпадающий с правой границей интервала, относится к последующему интервалу). Очевидно, что $\sum_{i=1}^{r} n_i^* = n$.

Замечание. Результаты группировки удобно записывать в виде таблицы. В итоге получаем статистическое распределение в виде последовательности интервалов и соответствующих частот. Для построения полигона и эмпирической функции распределения используют середины интервалов.