МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Математическая статистика — прикладная математическая дисциплина, примыкающая к теории вероятностей.

Ее задача — получить обоснованные выводы о параметрах, виде распределения и других свойствах случайной величины по конечной совокупности наблюдений над ней (наборе значений, полученных в n испытаниях - ***выборке***).

В частности определение точечных и интервальных оценок параметров распределения, проверка гипотез о виде распределения. Проводится также корреляционный и регрессионный анализ совокупности двух и более случайных величин и т.д.

Выборки и их характеристики.

Выборочной совокупностью (*выборкой*) называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Число объектов (наблюдений) в совокупности называется ее *объемом.*

Пусть x_1 , x_2 ,..., x_n – выборка объема n. По этой выборке можно оценить основные числовые характеристики генеральной совокупности. Различные элементы выборки x_i называются вариантами. Ряд вариант, расположенных в порядке **возрастания** их значений называется **вариационным рядом**. Им пользуются, в основном, при малых n .

Если n велико, то ряд преобразуют в группировки по отдельным значениям признака x (для дискретной случайной величины) или по интервалам изменения признака (интервальная группировка), для чего разбивают интервал изменения признака x (x_{max}, x_{min}) , на K равных интервалов. Подсчитывают частоты m_i - число значений x_k , попавших в i-й интервал.

В случае непрерывной случайной величины для определения количества интервалов рекомендуется правило $K = \sqrt{n}$, где $5 \le K \le 20$.

Результат группировки представляют рядом интервалов и рядом соответствующих частот. Под частотой m_i интервала понимают число членов выборки, варианты которых x_i лежат в i – м интервале.

Пример 2. Дана выборка объема n = 30 .

20.3; 15,4; 17,2; 19,2; 23,1; 18,1; 21,9; 15,3; 16,8; 13,2; 20,4; 16,5;
19,7; 20,5; 14,3; 20,1; 16,8; 14,7; 20,8; 19,5; 15,4; 19,3; 17,8; 16,2; 15,7;
22,8; 21,9; 12,5; 10,1; 21,1.

Сделать интервальную группировку этой выборки.

Решение: Диапазон изменения признака x (размах выборки) $R = x_{\max} - x_{\min} = 23{,}1 - 10{,}1 = 13$. Количество интервалов K можно взять равным 5 ($K \approx \sqrt{30}$), ширина интервала $d = \frac{R}{K} = \frac{13}{5} = 2{,}6$. Таблица интервальной группировки будет выглядеть следующим образом:

Интервалы	10,1–12,7	12,7–15,3	15,3–17,9	17,9–20,5	20,5–23,1
m_i	2	4	9	9	6
h_i	0,067	0,133	0,3	0,3	0,2
середины интервалов u_i	11,4	14,0	16,6	19,2	21,8

m_i - частота, т. е. число значений x_i, попавших в i-й интервал.

Расчет характеристик выборки (приближенных значений M(X) и D(X)).

Итак, пусть $x_1, x_2, ..., x_n$ − выборка объема n из генеральной совокупности, имеющей функцию распределения $F(x)$. Числовые характеристики выборки называются ***выборочными*** (эмпирическими) числовыми характеристиками. Введем основные числовые характеристики:

− ***среднее арифметическое (среднее)***

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,;$$

Выше приведена формула вычисления по **несгруппированным** данным, т. е. без интервального представления данных.

При **интервальной группировке:**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{K} m_i u_i$$

- вычисление по сгруппированным данным.

n — полное число исходных значений x_i , u_i — середины интервалов.

− ***выборочная (эмпирическая) дисперсия***

- вычисление по несгруппированным данным.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\cdot\left(\sum_{i=1}^{n} x_i^2 - n\left(\overline{x}\right)^2\right)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{K} m_i (u_i - \overline{x})^2$$

- вычисление по сгруппированным данным.

– ***стандартное (среднее квадратическое) отклонение*** является корнем из выборочной дисперсии:

$$S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{\frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n\left(\overline{x}\right)^2 \right)}. \qquad (1.7)$$

Для сгруппированных данных, соответственно, имеем формулы:

$$S = \sqrt{\frac{1}{n-1} \cdot \left(\sum_{i=1}^{K} m_i x_i^2 - n\left(\overline{x}\right)^2 \right)} \qquad (1.8)$$

– ***выборочные начальные и центральные моменты порядка*** ℓ ($\ell = 1, 2, \ldots$), соответственно, определяются формулами

$$\alpha_\ell = \frac{1}{n} \sum_{i=1}^{n} x_i^\ell, \qquad \beta_\ell = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^\ell \qquad (1.11)$$

Если данные сгруппированы, то получаем следующие формулы:

$$\alpha_\ell = \frac{1}{n} \sum_{i=1}^{K} m_i u_i^\ell, \qquad \beta_\ell = \frac{1}{n} \sum_{i=1}^{K} m_i (u_i - \overline{x})^\ell$$

При интервальной группировке:

Гистограмма.

Гистограмма строится следующим образом. На оси X откладываются интервалы, и на каждом из них строится прямоугольник, площадь которого равна относительной частоте, соответствующей этому интервалу, т.е. высота прямоугольника h_i равна

$h_i = m_i/(n*d)$, так что полная площадь соответствующей ступенчатой фигуры равна 1. Здесь m_i/ - число значений x в i-м интервале, n — объем выборки, d — ширина каждого интервала.

Отметим, что гистограмма является эмпирическим аналогом плотности распределения. При $n \to \infty$ и $d \to 0$ гистограмма переходит в график точной плотности распределения $f(x)$.

Полигон частот - ломаная линия, которая получается, если из середины каждого интервала u_i восстановить перпендикуляр высотой h_i и соединить вершины этих перпендикуляров.

Пример 1. Построение гистограммы

Таблица 1

N	Значение	N	Значение	N	Значение	N	Значение
1	133,5	14	141,5	27	144,0	40	137,5
2	142,0	15	139,0	28	142,5	41	141,5
3	145,5	16	140,5	29	139,0	42	141,0
4	144,5	17	139,0	30	137,0	43	142,5
5	134,5	18	143,5	31	136,0	44	143,5
6	138,5	19	139,5	32	137,0	45	141,0
7	144,0	20	140,5	33	138,5	46	147,0
8	141,0	21	140,0	34	139,0	47	139,5
9	141,5	22	138,5	35	139,5	48	136,5
10	139,5	23	135,0	36	140,5	49	142,0
11	140,0	24	139,5	37	139,5	50	140,0
12	145,0	25	139,0	38	140,0		
13	141,5	26	138,0	39	140,5		

$n = 50$ x_{min} □ =133, 5 , x_{max}□ =147, 0 , R =□ 13, 5

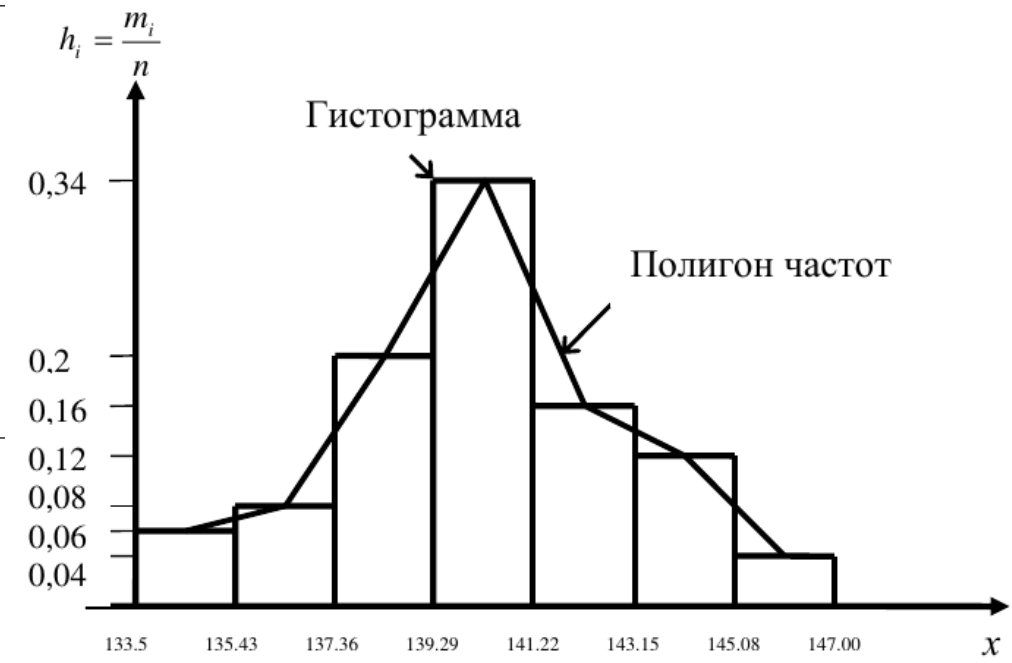
Количество интервалов: $K \approx \sqrt{n},\ K = 7$

,

Ширина интервала: $d = \dfrac{R}{K} = 1,93.$

Группировка:

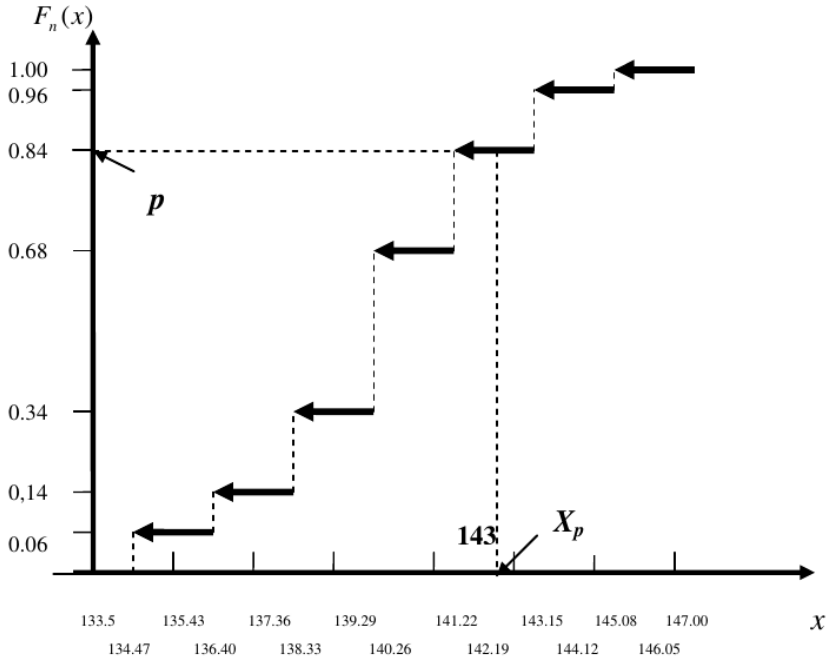
Номер интервала	Границы	m_i	$h_i = m_i / n$	$\sum_{j=1}^{i} h_j$	u_i
1					
2					
3					
4					
5					
6					
7					



Эмпирическая функция распределения

Эмпирическую функцию распределения F_n (x) получают построением ступенчатой кривой относительных накопленных частот (в случае несгруппированных данных F_n (x) равно количеству x_i, меньших x, деленному на n); F_n (x) имеет скачки в точках, соответствующих серединам интервалов u_i .

По гистограмме и полигону частот судят о виде плотности распределения исследуемой непрерывной случайной величины или о распределении вероятностей дискретной случайной величины. Эмпирическая функция распределения дает представление о функции распределения и используется в основном *в статистической проверке гипотез*.



Эмпирическая функция распределения

Кроме того, эмпирическая функция распределения используется также для определения эмпирических (выборочных) квантилей. Строгим квантилем порядка p $(0 < p < 1)$ называется величина x_p, определяемая из соотношения $P\{X < x_p\} = F(x_p) = p$. Выборочный квантиль x_p определяется из соотношения $F_n(x_p) = p$.

Задавая p, можно по графику эмпирической функции распределения оценить, например, значение x p, которое исследуемая величина не превзойдет с вероятностью p, либо, задавая x, по тому же графику оценить соответствующую вероятность p.