Статистическими данными называются сведения о числе объектов, обладающих теми или иными признаками. Статистический метод – метод, опирающийся на рассмотрение статистических данных. Математическая сторона статистических методов, безразличная к природе изучаемых объектов составляет предмет математической статистики. Математическая сторона включает в себя:

1) методы статистического описания;
2) теорию оценок;
3) теорию проверок статистических гипотез;
4) дисперсионный анализ;
5) дискриминантный анализ;
6) регрессионный анализ;
7) теорию массового обслуживания;
8) теорию надежности;
9) теорию информации.

Связь математической статистики с теорией вероятностей осуществляется в силу закона больших чисел (теорема Чебышева, теорема Бернулли) и центральной предельной теоремы:

Теорема Чебышева. Если для последовательности попарно независимых случайных величин $\xi_i \left(i = 1,2.....\right)$ дисперсии $D\left(\xi_i\right)$ - равномерно ограничены, то

$$\bar{\xi}_n \underset{n\to\infty}{\overset{вер}{\to}} M\left(\bar{\xi}_n\right),$$

где $\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ (например: вероятности приближенно выражаются в виде относительных частот, математическое ожидание – в виде среднего арифметического).

Теорема Чебышева показывает, что при большом числе независимых случайных величин ξ_i величина $\bar{\xi}_n$ утрачивает характер случайной величины.

Теорема Бернулли. Если эксперимент проводится по схеме Бернулли, то

$$\lim_{n\to\infty} P(|W(A) - p| < \varepsilon) = 1.$$

Теорема Бернулли разъясняет, почему относительная частота $W(A)$ при достаточно большом числе испытаний обладает свойством статистической устойчивости и оправдывает статистическое определение вероятности.

Теорема Ляпунова. Если взаимнонезависимые случайные величины ξ_i имеют конечные абсолютные моменты третьего порядка

$\mu_i = M[|\xi_i - M(\xi_i)|^3], \quad (i = 1,2\ldots.)$
и если эти моменты удовлетворяют условию
$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mu_i}{\delta_n^3} = 0,$
то тогда величина $\xi_n = \xi_1 + \xi_2 + \cdots + \xi_n$ имеет асимптотически нормальное распределение с центром $\gamma_n = M(\xi_n)$ и средним квадратическим отклонением $\delta_n = \sqrt{D(\xi_n)}$.

Центральная предельная теорема объясняет тот факт, что нормальное распределение часто встречается на практике.

§1.1. Выборка, статистики

Математическая статистика изучает такие явления, которые обладают свойством статистической устойчивости.

Генеральная совокупность и выборка

Пусть X-наблюдаемая физическая величина. В статистике X рассматривают как случайную величину ξ, распределенную по неизвестному закону. Множество (конечное или бесконечное) **всех** возможных значений случайной величины ξ называют генеральной совокупностью. На практике мы имеем дело с конечным набором экспериментальных данных, полученных в результате проведения n измерений или наблюдений, составляющих лишь **часть** генеральной совокупности. Эту конечную часть генеральной совокупности экспериментальных данных x_1, x_2, x_3 ..... x_n называют выборкой объема n из генеральной совокупности ξ, а элементы выборки называют выборочными значениями. Выборки подразделяют на повторные и бесповторные. *Повторной* называют выборку, при которой отобранный объект перед отбором следующего возвращается в генеральную совокупность. *Бесповторной* называют выборку, при которой отобранный объект в генеральную совокупность не возвращается. Если обьем генеральной совокупности велик, а объем выборки составляет лишь малую ее часть, то различие между повторной и бесповторной выборками исчезает. Поэтому на практике обычно используются бесповторные выбоки. При этом выборка должна быть *репрезентативной*, которая является таковой, если ее осуществление происходит случайно (*каждый объект отобран случайно, если в силу закона больших чисел все объекты имеют*

одинаковую вероятность попасть в выборку). На практике прменяют различные способы отбора, которые подразделяют на два вида:

1. отбор, не требующий расчленения генеральной совокупности на

части (простой случайный бесповторный отбор, простой случайный повторный отбор);

2. отбор, при котором генеральная совокупность разбивается на

части (типический отбор, механический отбор, серийный отбор)

Статистическая устойчивость

Если существует функция $f(x_1, x_2, \ldots x_n)$, значение которой может быть предсказано с существенно лучшей точностью, чем результат отдельного измерения x_k физической величины X, то говорят, что изучаемое явление обладает свойством статистической устойчивости. Функция $f(x_1, x_2, \ldots x_n)$ называется *статистикой* (например: относительная частота, выборочное среднее, выборочная дисперсия).

§1.2. Классическая вероятностная модель

Пусть X – наблюдаемая физическая величина. Классическая вероятностная модель основывается на следующих предположениях:

1)величине X сопоставляется случайная величина ξ, где ξ - одномерная непрерывная случайная величина с подлежащей определению или оценке плотностью вероятности $\varphi_\xi(x)$;

2)любой выборке $(x_1, x_2, .., x_n)$ сопоставляются случайные величины $(\xi_1, \xi_2, ... \xi_n)$, где ξ_i - взаимно независимые случайные величины с одинаковой плотностью вероятности $\varphi_\xi(x)$. Множество $(\xi_1, \xi_2, ... \xi_n)$ представляет собой n-мерную случайную величину с плотностью распределения

$$L(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \varphi_\xi(x_i),$$

которая называется функцией правдоподобия.

§1.3. Полигон выборки и полигон частот, гистограмма и листовая диаграмма

Пусть

$$(x_1, x_2, .., x_n) \qquad (1)$$

выборка объема n. Выборочные значения, записанные в порядке их появления, называют первичным списком.

Пусть выборка (1) содержит k различных значений, x_{i_1} - минимальное выборочное значение в выборке (1) и пусть

$$x_{i_1} < x_{i_2} < x_{i_3} < .... < x_{i_{k-1}} < x_{i_k}$$

Каждое из различных выборочных значений x_{i_j} называют вариационными значениями. Последовательность вариационных значений, записанных в возрастающем порядке, называют вариационным рядом

$$x_{i_1}, x_{i_2}, x_{i_3}, ..., x_{i_{k-1}}, x_{i_k} \qquad (2)$$

Пусть x_{i_1} встречается m_1 раз, x_{i_2} - m_2 раза, ..., x_{i_k} - m_k раз и $\sum_{j=1}^{k} m_j = n$. Число m_j - называют частотой вариационного значения x_{i_j}. Отношение $W_j = \frac{m_j}{n}$ называют относительной частотой вариационного значения x_{i_j}. Последовательность пар (x_{i_j}, m_j) называют простым статистическим рядом. Обычно простой статистический ряд представляют в виде таблицы

x_{i_j}	x_{i_1}	x_{i_2}	.....	x_{i_k}
m_i	m_1	m_2	.....	m_k

Полигоном выборки называется ломаная линия с вершинами в точках (x_{i_j}, m_j).

Для сравнения двух признаков иногда удобно использовать ***«листовую диаграмму»***. ***«Листовая диаграмма»*** состоит из выдвинутого столбца («стебля») и строк («листьев»). Выборочное значение разлагается на стеблевую часть (например, десятки или сотни) и листовую часть (например, единицы или десятки и единицы).

Сначала строится стебель от самых малых значений и до самых больших. Затем заносится в строку листовая часть каждого выборочного значения на соответствующей высоте стебля. Данная диаграмма изготавливается из первичного списка, поэтому числа в листьях не упорядочены.

Эмпирическая функция распределения

Эмпирической функцией распределения, соответствующей выборке (1), называют функцию вида

$$F_n(x) = \frac{k_x}{n}, \tag{3}$$

где k_x - число элементов выборки, значения которых меньше произвольного x.

В соответствии с формулой (3) и вариационным рядом (2) эмпирическую функцию распределения можно представить в виде:

$$F_n(x) = \begin{cases} 0, \ x \le x_{i_1} \\ \frac{m_1}{n}, \ x_{i_1} < x \le x_{i_2} \\ \frac{m_1 + m_2}{n}, x_{i_2} < x \le x_{i_3} \\ \ldots\ldots\ldots \\ 1, x > x_{i_k} \end{cases} \tag{4}$$

Из выражения (4) видно, что $F_n(x)$ обладает известными из теории вероятностей свойствами: изменяется от 0 до 1, не убывает. При этом она кусочно-непрерывная, возрастает только в точках последовательности (2) и величина скачков равна

$$F_n\left(x_{i_{j+1}}\right) - F\left(x_{i_j}\right) = \frac{m_{j+1}}{n} \tag{5}$$

Пусть P_j - вероятность того, что случайная величина $\xi = x_{i_j}$. Тогда по теореме Бернулли

$$\lim_{n\to\infty} P\left(\left|\frac{m_j}{n} - P_j\right| < \varepsilon\right) = 1, \tag{6}$$

т.е. при достаточно большом объеме выборки n с вероятностью $P \approx 1$

$$P_j \approx \frac{m_j}{n}$$

(7)

Пусть $P(\xi < x)$ - вероятность того, что случайная величина принимает значения $(\xi < x)$. Так как по определению выборочная функция распределения равна относительной частоте $F_n(x) = W(\xi < x)$, а теоретическая функция распределения равна вероятности $F(x) = P(\xi < x)$, то тогда по теореме Бернулли

$$\lim_{n \to \infty} P(|F_n(x) - P| < \varepsilon) = 1, \qquad (8)$$

т.е. при $n \to \infty$ выборочная функция распределения $F_n(x)$ сходится по вероятности к теоретической функции распределения $F(x) = P(\xi < x)$.

Гистограмма

Пусть ξ - непрерывная случайная величина с неизвестной плотностью вероятности $f_\xi(x)$. Для приближенного представления $f_\xi(x)$ по выборке $x_i (i = 1,2...n)$ разобьем область значений выборки на интервалы $[x_i, x_{i+1}]$ длинной h. Пусть ν_i - число элементов выборки, попавших в i-ый интервал. Тогда по теореме Бернулли вероятность попадания в i-й интервал

$$P_i \approx \frac{\nu_i}{n}$$

С другой стороны, эту же вероятность можно выразить по известному из теории вероятностей свойству через интеграл от плотности вероятности $f_\xi(x)$ с учетом теоремы о среднем:

$$\frac{\nu_i}{n} \approx \int_{x_i}^{x_i+1} f(x)dx = f_\xi(x_i^*) \cdot h,$$

где x_i^* - некоторая точка в частичном интервале, в качестве которой обычно выбирается середина интервала. Отсюда получаем оценку плотности распределения:

$$f_n(x_i^*) = \frac{\nu_i}{n \cdot h} \qquad (9)$$

График функции (9) называется гистограммой. Если соединить точки $M(x_i^*, f(x_i^*))$ отрезками прямых, то получится график, который называется полигоном частот.

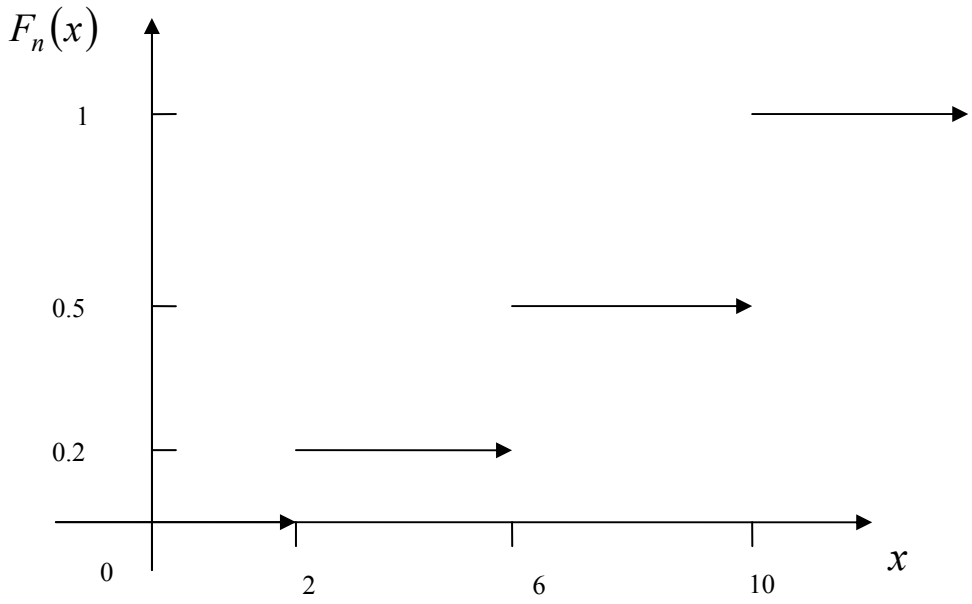
Пример. Дан простой статистический ряд

x_i	2	6	10
m_i	12	18	30

Используя статистический ряд, найдем выборочную функцию распределения

$$F_n(x)=\begin{cases}0, & x\le 2\\ \dfrac{12}{60}, & 2<x\le 6\\ \dfrac{30}{60}, & 6<x\le 10\\ 1, & x>10\end{cases}$$

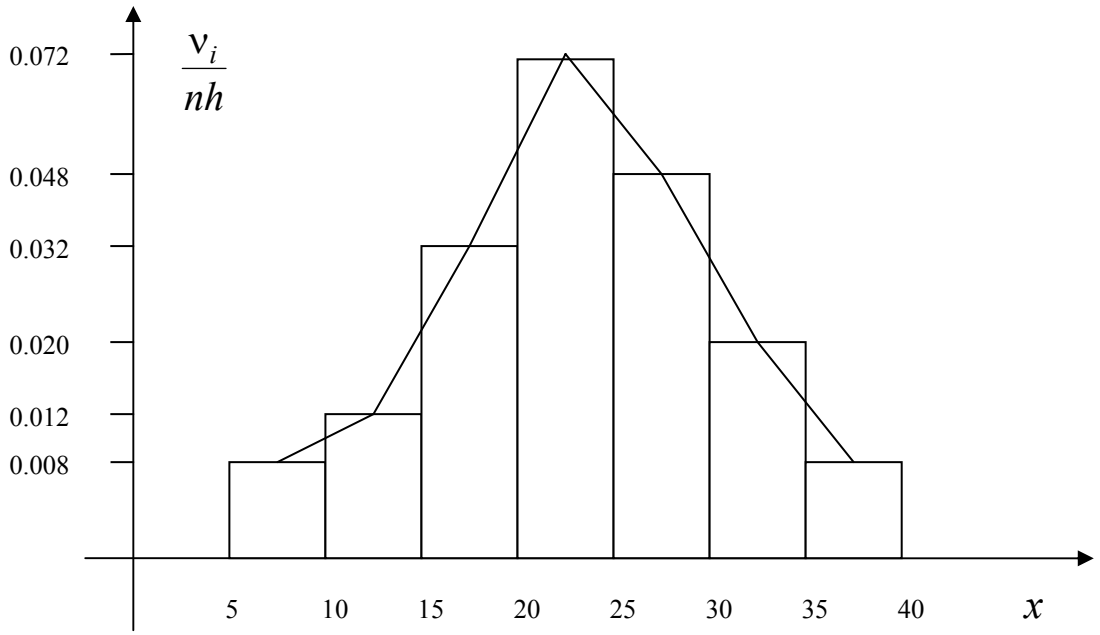
Построим график этой функции



Пример. Дан интервальный статистический ряд

Δx_i	5-10	10-15	15-20	20-25	25-30	30-35	35-40
ν_i	4	6	16	36	24	10	4
$\frac{\nu_i}{nh}$	0.008	0.012	0.032	0.072	0.048	0.020	0.008

Построим гистограмму и полигон частот



Пример. Первичный список уровня осадков в Воронеже состоит из 20 значений: 54, 51, 8, 50, 5, 53, 79, 71, 41, 21, 41, 23, 68, 88, 106, 34, 9, 41, 41, 36. Построим «Листовую диаграмму».

0	8 5 9
1	
2	1 3
3	4 6
4	1 1 1 1
5	4 1 0 3
6	8
7	9 1
8	8
9	
10	6

Здесь в столбце стоят десятки, единицы образуют листья. Стебель отделен от листьев вертикальной чертой. Между листьев оставляют небольшой пробел. Сравнение двух признаков осуществляют, располагая листовые диаграммы «спина к спине».

Для примера сравним сумму осадков в июле и августе в Воронеже с 1977 по 1996 год.

Июль					Август				
		9	5	8	0				
					1	5	4		
			3	1	2	2	5	0	
			6	4	3	9	3		
1	1	1	1		4	5	7	0	
3	0	1	4		5	6	7		
			8		6	9	5		
		1	9		7	7	2	7	1
			8		8				
					9				
			6		10				
					11				
					12	3			
					13				
					14	4			