ЛЕКЦИЯ 9

Теория вероятностей и математическая статистика

Введение в математическую статистику

Основные понятия

Математическая статистика – это раздел математики который занимается разработкой методов сбора, описания и анализа экспериментальных результатов наблюдений, массовых случайных явлений.

Фундаментальными понятиями математической статистики являются *генеральная совокупность и выборка*.

Математическая статистика базируется на понятиях и методах теории вероятностей, но решает в каком –то смысле обратные задачи. Как и всякая математическая теория, она развивается в рамках некоторых моделей, описывающих определенный круг явлений.

Основные понятия

В МС предполагается, что вероятность *P* в модели наблюдаемого случайного явления не известна полностью. Известно только, что *P* из некоторого заданного класса вероятностей P. Способы задания класса вероятностей P могут быть различными.

Если задан класс допустимых распределений P, то говорят, что задана *статистическая модель.*

Т.о., статистическая модель описывает такие ситуации, когда в вероятностной модели изучаемого эксперимента имеется неопределенность в задании вероятности *P.*

Основные понятия

Задача математической статистики уменьшить неопределенность модели, используя информацию полученную из наблюдаемых исходов эксперимента.

Итак, о математической статистике имеет смысл вспоминать, если

- имеется случайный эксперимент, свойства которого частично или полностью неизвестны,
- мы умеем воспроизводить этот эксперимент в одних и тех же условиях некоторое (а лучше — какое угодно) число раз.

Основные понятия

Исходным материалом всякого статистического исследования является ***совокупность результатов наблюдений***.

В большинстве случаев исходные статистические данные $X = (X_1,...,X_n)$ – результат наблюдения некоторой конечной совокупности случайных величин, характеризующий исход изучаемого эксперимента.

Предполагается, что эксперимент состоит в проведении n испытаний и результат i –го эксперимента описывается случайной величиной X_i , i =1,..., n.

Основные понятия

- Совокупность наблюдаемых случайных величин $X = (X_1, ..., X_n)$ называется *выборкой*, сами величины X_i , i =1,..., n, – *элементами выборки*, а их число n – ее *объемом*.
- Реализации выборки *X* будем обозначать строчными буквами $x = (x_1,..., x_n)$.

Основные понятия

- Пусть $X = \{x\}$ – множество всех возможных значений выборки X, которое называется ***выборочным пространством***.
- ***Статистической моделью*** <F> называется класс распределений, допустимых для выборки.

Основные понятия

- Обычно рассматривают ситуации, когда компоненты выборки независимы и распределены так же, как некоторая случайная величина ξ с функцией распределения $F_\xi(x)$.
- Множество возможных значений ξ с распределением $F = F_\xi(x)$ называется ***генеральной совокупностью***, из которой производят случайную выборку.

Важно!

- **Таким образом, мы рассматриваем генеральную совокупность как случайную величину ξ, а выборку – как *n* – мерную случайную величину $(\xi_1, \ldots, \xi_n)$, компоненты которой независимы и одинаково распределены (так же, как ξ).**
- Такие выборки называются ***простыми***.

Порядковые статистики

Упорядочим выборку $x = (x_1, ..., x_n)$ (реализацию) по возрастанию, получим последовательность $x^* = (x_1^*, x_2^*, ..., x_n^{*)}$, где $x_1^* \le x_2^* ... \le x_n^*$.

<u>Пример.</u> $x = (2, 1, 4, 2, 3)$. $x^* = (1, 2, 2, 3, 4)$.

Если теперь через X_k^* обозначить случайную величину, которая для каждой реализации принимает значение x_k^* , k =1, …, n, (k-е по величине), то X_k^* называется *k - ой порядковой статистикой выборки.*

Порядковые статистики

- Очевидно, что порядковые статистики удовлетворяют неравенствам

$$X_1^* \le X_2^* \le \ldots \le X_n^*$$

- X_1^* и X_n^* называются ***экстремальными значениями выборки***.
- $X_1^* = X_{min}$, $X_n^* = X_{max}$.
- Последовательность $X_1^*, X_2^*, \ldots, X_n^*$ называют ***вариационным рядом***.

Способы представления выборки

- ***Вариационным рядом выборки*** называется способ ее записи, при котором элементы упорядочиваются по величине, т. е. записываются в виде упорядоченной последовательности.
- Разность между максимальным и минимальным элементами выборки называется ***размахом выборки***.

Способы представления выборки

Статистическим рядом называется последовательность пар (x_j, n_j).

Здесь x_j – значения, а n_j – ***частота элемента выборки***

X_i		X_1	X_2	…	X_{k-1}	X_k
n_i		n_1	n_2	…	n_{k-1}	n_k

Группированный статистический ряд

Интер –валы		$X_1 - X_2$	…	$X_{k-1} - X_k$	$X_k - X_{k+1}$
n_i		n_1	…	n_{k-1}	n_k

Эмпирическая функция распределения

Пусть $X=(X_1, ..., X_n)$ – выборка из генеральной совокупности наблюдаемой случайной величины.

Эмпирической функцией распределения называется случайная функция от $F_n(x)$, вычисляемая по формуле

$$F_n(x) = \frac{\nu_n}{n},$$

где ν_n – число элементов выборки X, значения которых меньше x.

Пример

- Выборка: X = {1, 2, 2, 3}

$$F_n(x) = \begin{cases} 0, & x \le 1 \\ 1/4, & 1 < x \le 2 \\ 3/4, & 2 < x \le 3 \\ 1, & x > 3 \end{cases}$$

Важно!

- Эмпирическая функция распределения выборки совпадает с функцией распределения дискретной случайной величины X, заданной рядом распределения:

X	x_1^*	x_2^*	…	x_n^*
P	1/n	1/n	…	1/n

Почему это важно:

- Это означает, что выборку можно рассматривать как дискретную случайную величину, и применять к ней то, что мы уже знаем о случайных величинах.

X	X_1^*	X_2^*	…	X_n^*
P	1/n	1/n	…	1/n

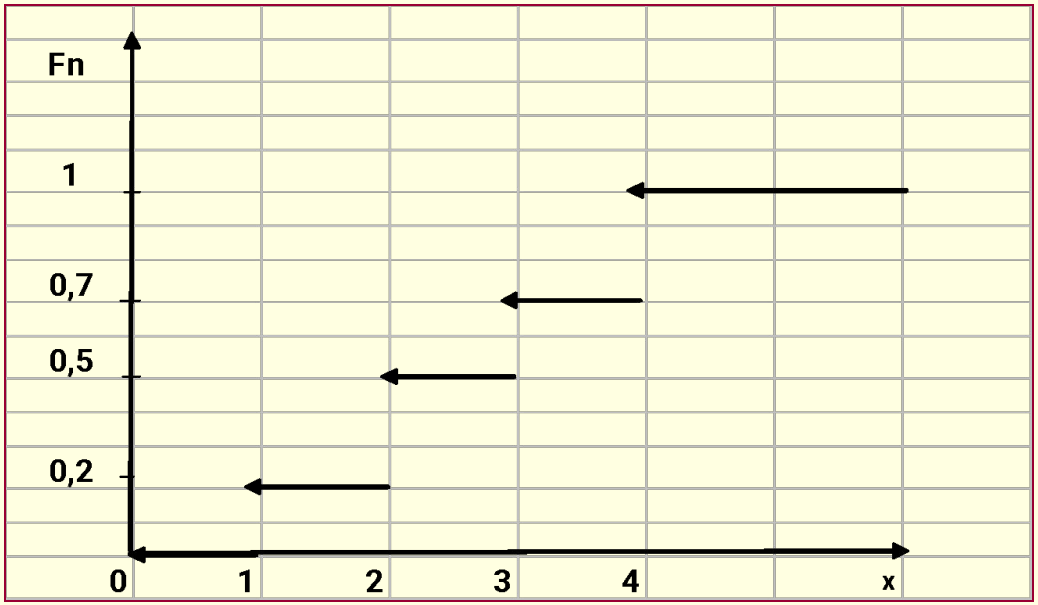
X	X_1^*	X_2^*	…	X_K^*
P	n_1/n	n_2/n	…	n_k/n

Еще один пример

X	1	2	3	4
n_i	2	3	2	3

$$F_n(x) = \begin{cases} 0, & x \le 1 \\ 0.2, & 1 < x \le 2 \\ 0.5, & 2 < x \le 3 \\ 0.7, & 3 < x \le 4 \\ 1 & x > 4 \end{cases}$$

График



Общая запись эмпирической функции распределения

$$F_n(x)=\begin{cases} 0 \ \ при \ \ x \le x_1^*, \\ \dfrac{k}{n} \ \ при \ \ x_k^* \le x \le x_{k+1}^*, \ \ k=1,2,\ldots,n-1 \\ 1 \ \ при \ \ x > x_n^*, \end{cases}$$

Замечание

- По эмпирической функции распределения легко построить другие способы представления выборки, например, статистический или вариационный ряд.

Пример

$$F_n(x) = \begin{cases} 0, & x \le 5 \\ 3/20, & 5 < x \le 7 \\ 5/20, & 7 < x \le 8 \\ 8/20, & 8 < x \le 12 \\ 13/20, & 12 < x \le 15 \\ 14/20, & 15 < x \le 20 \\ 17/20, & 20 < x \le 23 \\ 1, & x > 23 \end{cases}$$

Пример

- Этой эмпирической функции распределения $F_n(x)$ соответствует выборка, заданная статистическим рядом:

X	**5**	**7**	**8**	**12**	**15**	**20**	**23**
$\mathbf{n_i}$	**3**	**2**	**3**	**5**	**1**	**3**	**3**

Пример

- **Задача.** Дана $F_n(x)$ из предыдущего примера. Сколько в выборке значений:

а) равных 15,

б) не больших 11?

- **Решение.**

а) 1 значение равно 15,

б) 8 значений не больше 11.

Свойства эмпирической функции распределения

Эмпирическая функция распределения – сжатая характеристика выборки. Для каждой реализации $x = (x_1,\dots ,x_n)$ функция однозначно определена и обладает всеми свойствами функции распределения:

- изменяется от 0 до 1;
- не убывает;
- непрерывна слева;
- $F_n(x)=0$ при $x < x^*$ и $F_n(x) =1$ при $x > x^*$,
- она кусочно –постоянна и возрастает только в точках последовательности.

Свойства эмпирической функции распределения

Пусть $F_n(x)$ – эмпирическая функция распределения, построенная по выборке X из распределения ξ, и $F_\xi(x)$ – соответствующая теоретическая функция.

Тогда:

$$1)\ M[F_n(x)] = F_\xi(x)$$

$$2)\ F_n(x) \xrightarrow{p} F_\xi(x)$$

Теорема 1

Пусть $F_n(x)$ – эмпирическая функция распределения, построенная по выборке X из распределения ξ, и $F_\xi(x)$ – соответствующая теоретическая функция распределения. Тогда для любого $-\infty < x < +\infty$ и любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(|F_n(x) - F(x)| < \varepsilon) = 1.$$

Теорема 2 (теорема Колмогорова)

Если функция $F(x)$ непрерывна, то при любом фиксированном t > 0

$$\lim_{n\to\infty} P\left(\left|\sqrt{n}D_n\right| \le t\right) = K(t) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2t^2}$$

где

$$D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right|$$

функция Колмогорова (хорошее приближение при ≥ 20).

Теорема Колмогорова

Теорема справедлива для любой непрерывной функции и позволяет найти границы, в которых с заданной вероятностью $0<\alpha<1$ находится теоретическая функция $F(x)$. Если задана вероятность α, то при больших n с вероятностью, близкой к α $F(x)$ удовлетворяет неравенству

$$\left|F_n(x)-F(x)\right| \le \frac{t_\alpha}{\sqrt{n}}$$

где величина t_α вычисляется как корень уравнения $K(t)=\alpha$

Группировка выборки

Частота элемента выборки

При большом объеме выборки ее элементы объединяют в группы, представляя результаты опытов в виде ***группированного статистического ряда***.

Для этого интервал, содержащий все элементы выборки, разбивается на k непересекающихся интервалов. Вычисления значительно упрощаются, если эти интервалы имеют одинаковую длину h. Результаты сводятся в таблицу, называемую ***таблицей частот группированной выборки.***

Группированный статистический ряд

- **Вспомним вид этого ряда. Чтобы его построить, надо найти число интервалов k и ширину интервала h.**

Интер –валы		$X_1 - X_2$	…	$X_{k-2} - X_{k-1}$	$X_k - X_{k+1}$
n_i		n_1	…	n_{k-1}	n_k

Группировка выборки

- Разность между максимальным и минимальным элементами выборки называется ***размахом выборки R***.
- ***Число интервалов k*** находится из условия

$$2^{k-1} \approx n,$$

где n – объем выборки.

- ***Длину интервала h*** находят по формуле

$$h = R/k.$$

Все интервалы имеют одинаковую длину.

Пример. Неупорядоченная выборка

38	68	77	61	67
60	52	47	35	65
41	47	28	47	39
51	46	48	72	48
33	49	58	41	43
42	49	32	45	60
45	14	42	44	54
21	57	58	55	42
53	54	61	30	59
60	59	30	40	50

Упорядоченная выборка

14	40	46	52	60
21	41	47	53	60
28	41	47	54	60
30	42	47	54	61
30	42	48	55	61
32	42	48	57	65
33	43	49	58	67
35	44	49	58	68
38	45	50	59	72
39	45	51	59	**77**

Нахождение числа интервалов k и длины интервала h

$$n = 50$$

$$R = 77 - 14 = 63$$

$$2^{k-1} \approx 50$$

$$2^6 = 64 \Rightarrow k - 1 = 6$$

$$k = 7$$

$$h = \frac{R}{l} = \frac{63}{7} = 9$$

Таблица частот группированной выборки

k	x_i	x_{i+1}	n_i	$\overline{x}_i$	n_i^*
1	14	23	2	18.50	2
2	23	32	3	27.50	5
3	32	41	6	36.50	11
4	41	50	17	45.50	28
5	50	59	10	54.50	38
6	59	68	9	63.50	47
7	68	77	3	72.50	50

Группированная выборка

Интер –валы	[14 –23)	[23 –32)	[32 –41)	[41 –50)	[50 –59)	[59 –68)	[68 –77)
n_i	2	3	6	17	10	9	3

Графические характеристики выборки

- Если на каждом интервале построить прямоугольник с высотой n_i/h, получим ***гистограмму***.
- Кривая, соединяющая середины верхних оснований гистограммы, называется ***полигоном*** (***частот***). Полигон — непрерывная функция (ломаная).

Замечание

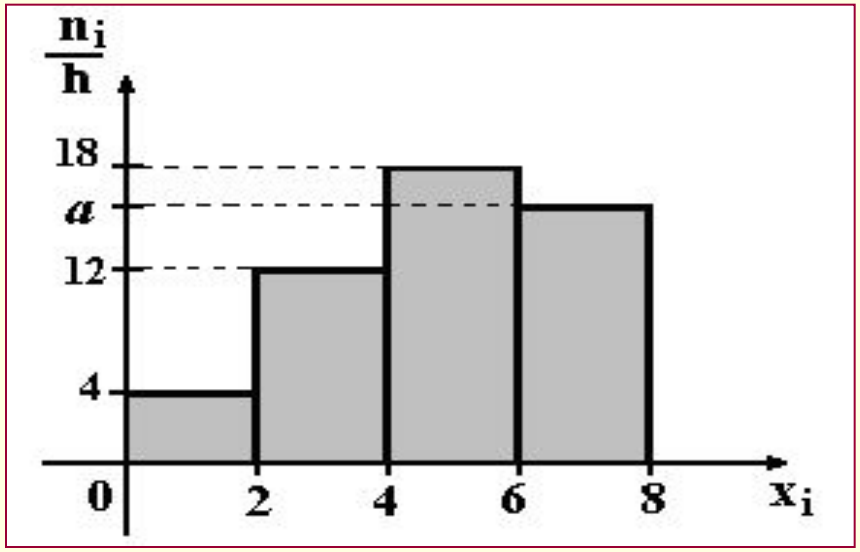
Если по оси ординат откладываются высоты n_i/h, то площадь ступенчатой фигуры под графиком гистограммы равна объему выборки n. В этом случае мы имеем *гистограмму частот.*

Если по оси ординат откладываются высоты n_i/nh, то получаем *гистограмму относительных частот.* Площадь соответствующей ступенчатой фигуры для нее равна единице.

Задача

- По выборке объема n = 100 построена гистограмма частот. Чему равно значение a?
- Решение. Площадь S = n = 100.

S = 2(4 + 12 + a + 18) = 2(34 + a) = 100, отсюда a = 16.



Смысл гистограммы и полигона

При увеличении объема выборки и уменьшении интервала группировки гистограмма относительных частот является статистическим аналогом плотности распределения генеральной совокупности.

Т.о., они дают представление о графике плотности.

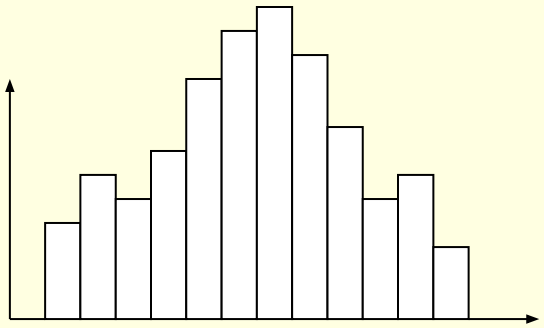

гистограмма
20
15
10
5
0
частота
14-23
23-32
32-41
41-50
50-59
59-68
68-77
значение фактора


полигон частот
20
15
10
5
0
частота
18.5
27.5
36.5
45.5
54.5
63.5
72.5
значение фактора

Замечание

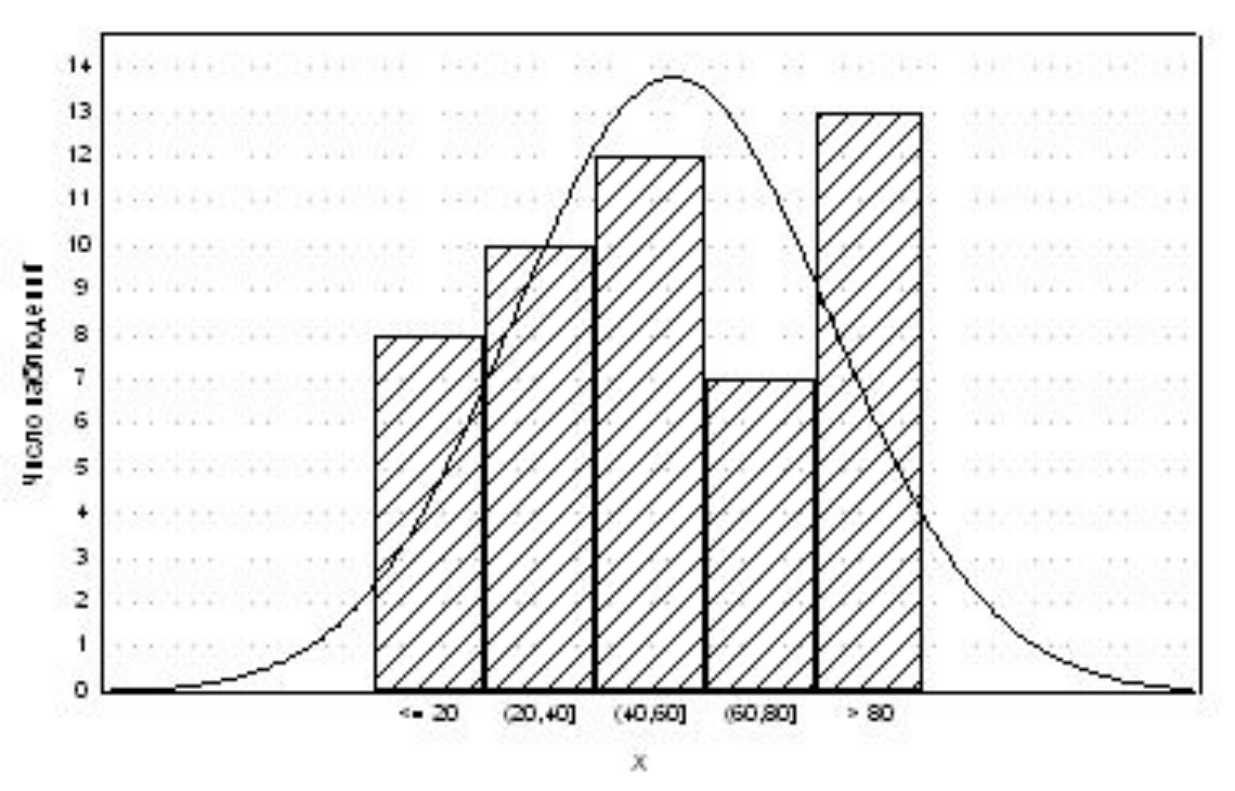
- Для лучшего приближения плотности столбики гистограммы рекомендуется строить без пробелов.

- **Гистограмма**



- **Полигон**

Гистограмма и плотность



Кумулята

- **Кумулята** относительных частот – это ломаная, соединяющая точки с координатами $(x_i, n_i^*/n)$. **Кумулята** частот соединяет точки с координатами (x_i, n_i^*).
- Напомним, что n_i^* – это накопленная сумма частот, $n_i^* = n_1 + n_2 + \ldots + n_i$
- Кумулята дает представление о графике функции распределения.


Кумулята частот
Накопленная чичтота
60
50
40
30
20
10
0
14
23
32
41
50
59
68
значения фактора