ВОПРОС 29:

ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика - наука, выявляющая закономерности повторяющихся случайных явлений на основе обработки статистических данных, полученных в результате наблюдений.

Основные задачи мат. статистики:

1. Разработка методов анализа наблюдаемых случайных данных (оценка неизвестной вероятности события, неизвестной функции распределения и ее параметров, оценка зависимостей от случайных величин и т.д., проверка статистических гипотез);

2. Синтез алгоритмов для решения задач выявления взаимосвязей, трендов, прогнозирования, поддержки принятия решений.

Определения.

Генеральная совокупность – все множество имеющихся наблюдений или объектов, относящихся к изучаемому явлению.

Выборка – набор наблюдений или объектов, случайно отобранных из генеральной совокупности.

Объем генеральной совокупности N и **объем выборки** n – число наблюдений

Виды выборки

Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;

Бесповторная – отобранный объект в генеральную совокупность не возвращается.

NB!

Выборка должна быть репрезентативной (представительной).

Пусть с.в. X принимает в выборке значение x_1 n_1 раз, x_2 – n_2 раз, …, $x_к$ – $n_к$ раз, причем $\sum_{i=1}^{k} n_k = n,$

n – объем выборки.

Тогда наблюдаемые значения случайной величины $x_1, x_2, \ldots, x_к$ называют **наблюдениями**, а n_1, n_2, …, $n_к$ – **частотами**.

О***тносительные частоты*** $w_i = \frac{n_i}{n}.$

Статистический ряд

X	x_1	x_2	…	x_k
n	n_1	n_2	…	n_k
w	w_1	w_2	…	w_k

Пример.

При проведении 20 бросков игральной кости число выпадений очков оказалось равным 2, 2, 5, 1, 2, 3, 2, 3, 3, 1, 5, 4, 4, 2, 1, 3, 2, 3, 6, 4.

Статистический ряд имеет вид:

x_i	1	2	3	4	5	6
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Определение.

Последовательность наблюдений, записанных в порядке возрастания или убывания

$$x(1),\ x(2),\ldots,\ x(к):$$

$$x(1) <= x(2) <= \ldots <= x(к)$$

или убывания

$$x(1),\ x(2),\ldots,\ x(к):\quad x(1) >= x(2) >= \ldots >= x(к)$$

называют **вариационным рядом**.

Определение.

Наблюдения, образующие вариационный ряд

$$x(1), x(2),\ldots, x(k),$$

называются

порядковыми статистиками,

а их номера в вариационном ряду – ***рангами***.

ВОПРОС 30:

Группированные данные

Статистический ряд для непрерывной с.в.

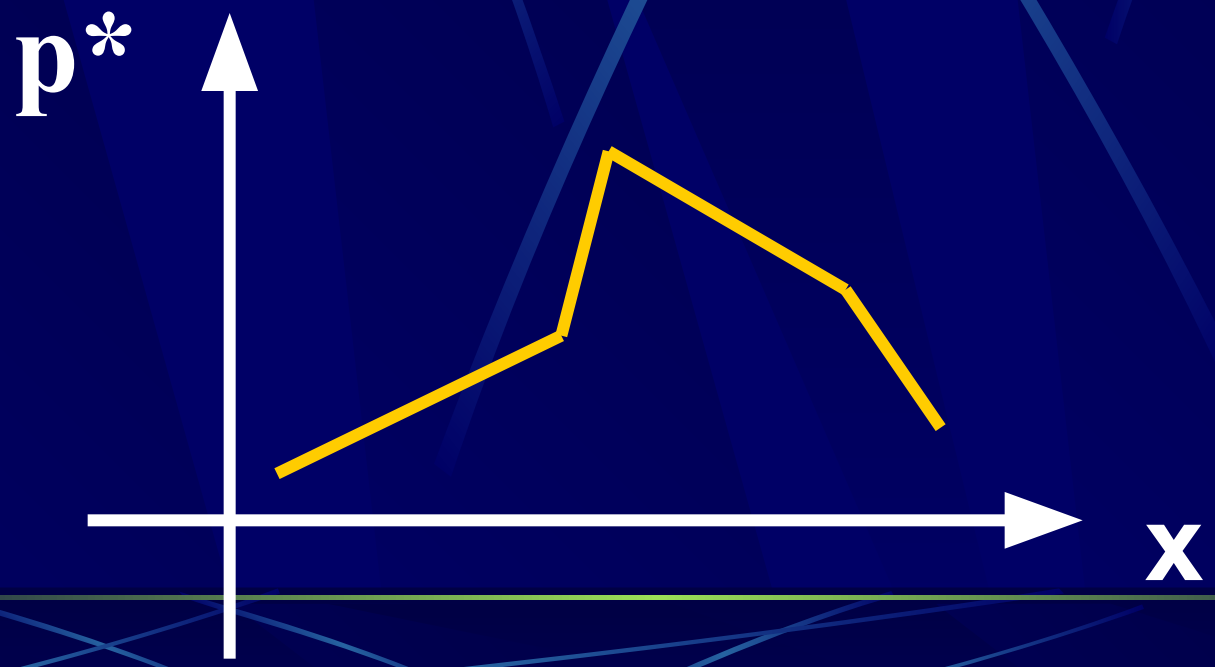
Номер интервала	Интервал	m_i	p_i^*
1	$[X_0, X_1)$	m_1	p_1^*
2	$[X_1, X_2)$	m_2	p_2^*
...	...	...	...
k	$[X_{k-1}, X_k]$	m_k	p_k^*

$$p_i^* = \frac{m_i}{n} \qquad \sum_{i=1}^{k} p_i^* = 1.$$

Полигон частот:

ломаная, отрезки которой соединяют точки с координатами

$$(x_1, p_1^*), (x_2, p_2^*), \ldots, (x_k, p_k^*)$$



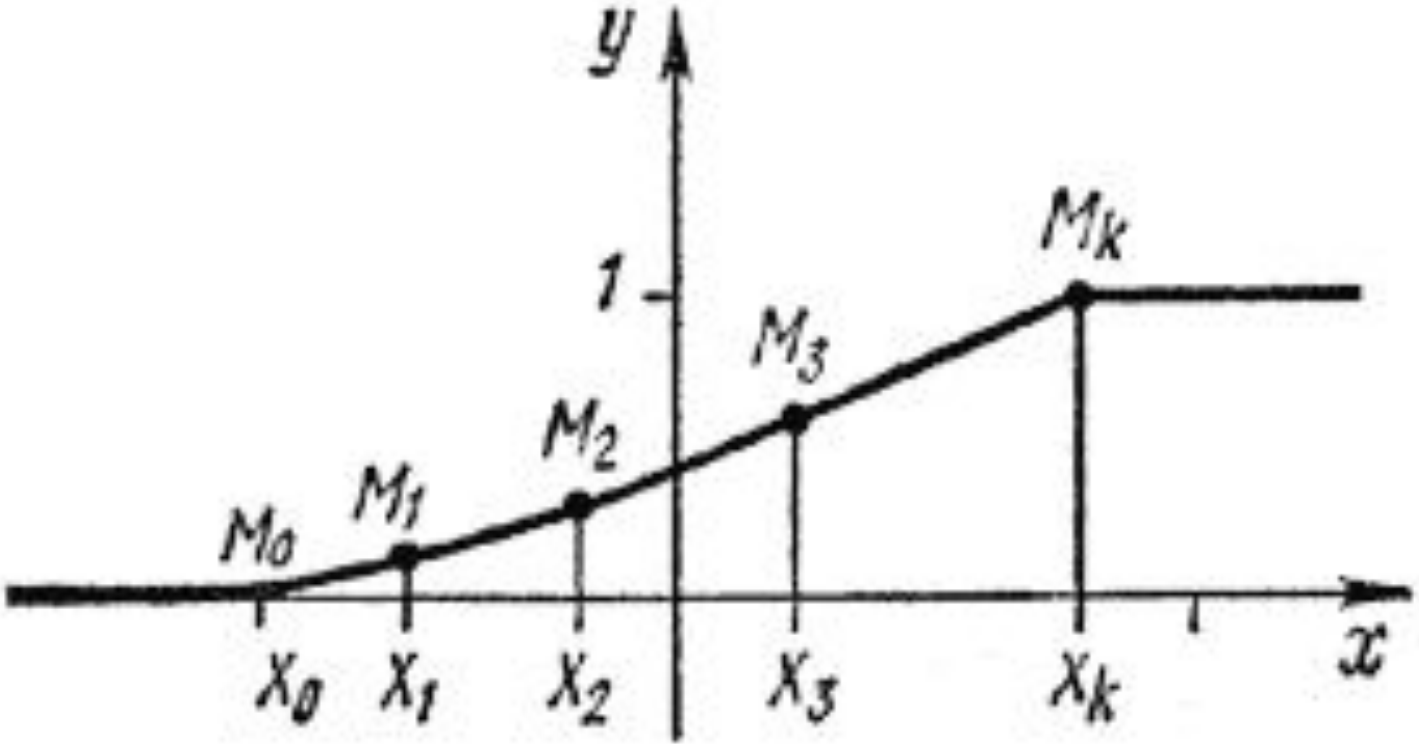
ВОПРОС 31:

Выборочная функция распределения и гистограмма

Определение.
Выборочной (эмпирической) **функцией распределения** называют функцию ***F*(x)***, определяющую для каждого значения *x* относительную частоту события ***X < x:***

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

$$\begin{cases} F^*(x_0) = P^*(\xi < x_0) = 0 \\ F^*(x_1) = P^*(\xi < x_1) = \dfrac{m_1}{n} = p_1^* \\ F^*(x_2) = P^*(\xi < x_2) = \dfrac{m_1 + m_2}{n} = p_1^* + p_2^* \\ F^*(x_k) = P^*(\xi < x_k) = \dfrac{m_1 + m_2 + \ldots + m_k}{n} = p_1^* + p_2^* + \ldots + p_k^* = 1 \end{cases}$$



$$F^*(x) \xrightarrow[n \to \infty]{} F(x).$$

Свойства F*(x)

(совпадают со свойствами $F(x)$):

1. $0 \le F^*(x) \le 1$.
2. $F^*(x)$ – неубывающая функция.
3. Если x_1 – наименьшее наблюдение, то $F^*(x) = 0$ при $x \le x_1$;

   если $x_к$ – наибольшее наблюдение, то $F^*(x) = 1$ при $x > x_к$

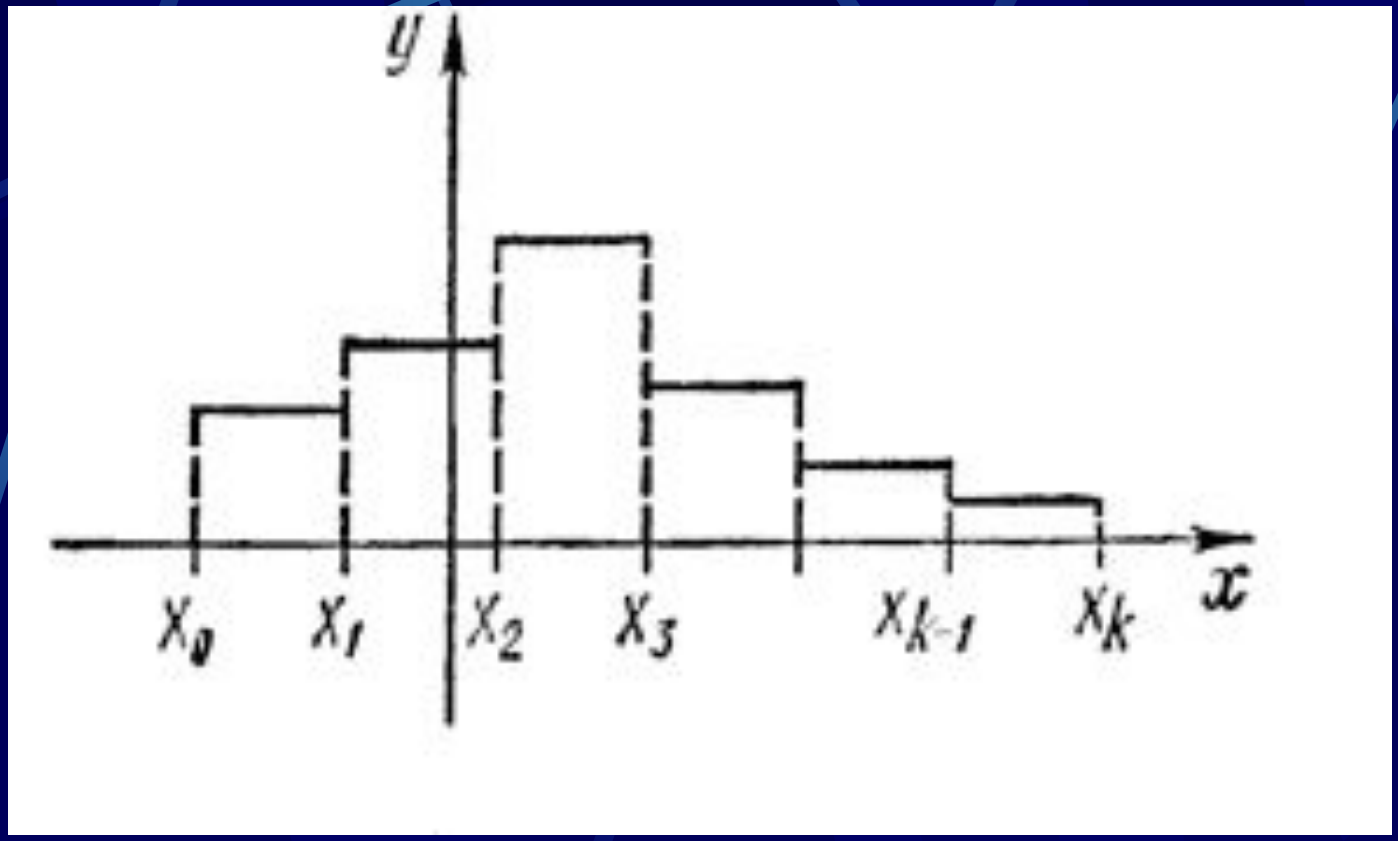
Эмпирическая плотность распределения

$$]y = \varphi^*(x):$$

которая в интервале (X_{i-1}, X_i] постоянна и равна

$$h_i = p_i^*$$

Гистограмма



$$\varphi^*(x) \underset{n\to\infty}{\to} \varphi(x).$$

ВОПРОС 32:

Оценки параметра положения:

выборочное среднее, оценки моды и медианы

ОПРЕДЕЛЕНИЯ.

Выборочное среднее:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{k} x_i}{n}$$

Мода:

$$Mod = x_{k^*}:$$

$$m_{k^*} = max(p_1, \ldots, p_n)$$

Медиана:

$$Med = \begin{cases} x_{k+1}, & n = 2k+1; \\ \dfrac{x_k + x_{k+1}}{2}, & n = 2k \end{cases}$$

ВОПРОС 33:

Оценки параметра масштаба:

оценки *дисперсии, начальных и центральных моментов*

ОПРЕДЕЛЕНИЯ.

Выборочной дисперсией

называется

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

$$\hat{\sigma}_n = \sqrt{\hat{D}_n}$$

$$\hat{D} = \overline{x^2} - (\bar{x})^2$$

Начальный эмпирический момент:

$$\hat{\alpha}_n^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Центральный эмпирический момент:

$$\hat{\mu}_n^k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^k$$

Пример 1.

Найти числовые характеристики выборки

x_i	2	5	7	8
m_i	3	8	7	2

$$\bar{x}_B = \frac{2 \cdot 3 + 5 \cdot 8 + 7 \cdot 7 + 8 \cdot 2}{20} = 5{,}55;$$

$$D_B = \frac{4 \cdot 3 + 25 \cdot 8 + 49 \cdot 7 + 64 \cdot 2}{20} - 5{,}55^2 =$$

$$= 3{,}3475;$$

$$\sigma_B = \sqrt{3{,}3475} = 1{,}83.$$

$$Mo = 5; \qquad Me = \frac{5 + 7}{2} = 6.$$

ВОПРОС 34:

Свойства оценок

Схема: ***k*** **выборок одного и того же объема** ***n*** **и вычислим для каждой из них оценку параметра Θ:**

$$\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_n.$$

Определение

Статистическая оценка $\hat{\theta}$ называется НЕСМЕЩЕННОЙ, если ее математическое ожидание равно оцениваемому параметру θ при любом объеме выборки:

$$E(\hat{\theta}) = \theta$$

Определение. **Оценка некоторого признака называется АСИМПТОТИЧЕСКИ НЕСМЕЩЕННОЙ, если для выборки $x_1, x_2, \ldots, x_n$**

$$\lim_{n \to \infty} \frac{x_1 + x_2 + \ldots + x_n}{n} = x$$

Определение. **Статистическая оценка называется ЭФФЕКТИВНОЙ, если она при заданном объеме выборки *n* имеет наименьшую возможную дисперсию**

$$D(\hat{\theta}) = min$$

Определение.

СОСТОЯТЕЛЬНОЙ называется статистическая оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру :

$$\theta_n^* \xrightarrow[n \to \infty]{P} \theta$$

Теорема.

Выборочное среднее представляет собой несмещенную оценку математического ожидания $E(X)$.

Доказать самостоятельно!

Выборочное дисперсия представляет собой смещенную оценку дисперсии:

$$E(\hat{D}_n) = \frac{n-1}{n} D$$

Исправленная выборочная дисперсия:

$$s^2 = \frac{n}{n-1}\hat{D}_n$$

Исправленное среднее квадратическое отклонение

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1}}$$

Определение. Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки $x_1, x_2, \ldots, x_n$

$$\lim_{n\to\infty} \frac{x_1 + x_2 + \ldots + x_n}{n} = x$$

СПОСОБЫ ПОСТРОЕНИЯ ОЦЕНОК

ВОПРОС 35:

Метод максимального правдоподобия

Модель.] X – дискретная с.в., которая в результате n испытаний приняла значения $x_1, x_2, \ldots, x_n$.

Предположим, что нам известен закон распределения этой величины, определяемый параметром Θ, но неизвестно численное значение этого параметра. Найдем его точечную оценку.

] $p(x_i, \Theta)$ – вероятность того, что в результате **i**-го испытания величина X примет значение x_i.

Определение. Назовем ***функцией правдоподобия*** **дискретной случайной величины X функцию аргумента Θ, определяемую по формуле:**

$$L(x_1, x_2, \ldots, x_n; \Theta) = = p(x_1, \Theta) p(x_2, \Theta) \ldots p(x_n, \Theta).$$

$$L(x_1, x_2, \ldots, x_n; \Theta) = f(x_1, \Theta) f(x_2, \Theta) \ldots f(x_n, \Theta).$$

Определение. Оценкой наибольшего правдоподобия называется оценка $\hat{\theta}_n$

при котором функция правдоподобия достигает максимума:

$$L(x, x, \ldots, x, \theta) = \max.$$

$\ln L$ – логарифмическая функция правдоподобия.

Алгоритм поиска ММП-оценки:

$$\frac{d \ln L}{d\Theta} = 0 \qquad \frac{d^2 \ln L}{d\Theta^2} < 0$$

ММП-оценки состоятельны (хотя могут быть смещенными), распределены асимптотически нормально и имеют наименьшую дисперсию по сравнению с другими асимптотически нормальными оценками.

ВОПРОС 36:

Метод моментов

] известный вид п.р. $f(x, \Theta_1, \Theta_2)$ определяется двумя неизвестными параметрами Θ_1 и Θ_2.

Требуется составить два уравнения, например $\Theta 1 = M1, \ \Theta 2 = m2.$

Отсюда

$$\begin{cases} E(X) = \overline{x}_n \\ D(X) = D_n \end{cases}$$

Решениями будут точечные оценки

$$\Theta_1 = \psi_1 (x_1, x_2, \ldots, x_n),$$
$$\Theta_2 = \psi_2 (x_1, x_2, \ldots, x_n).$$

ВОПРОС 37:

Метод наименьших квадратов

$$\sum_{i=1}^{n}(y_i-\varphi(x_i;a,b,c,\ldots))^2=\min.$$

$$\begin{cases}\sum_{i=1}^{n}(y_i-\varphi(x_i;a,b,c\ldots))\left(\frac{\partial\varphi}{\partial a}\right)_i=0\\ \sum_{i=1}^{n}(y_i-\varphi(x_i;a,b,c\ldots))\left(\frac{\partial\varphi}{\partial b}\right)_i=0\\ \sum_{i=1}^{n}(y_i-\varphi(x_i;a,b,c\ldots))\left(\frac{\partial\varphi}{\partial c}\right)_i=0\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\end{cases}$$

Пример: y=ax+b

$$\left(\frac{\partial \varphi}{\partial a}\right)_i = x_i; \qquad \left(\frac{\partial \varphi}{\partial b}\right)_i = 1.$$

$$\begin{cases} \sum_{i=1}^{n} (y_i - (ax_i + b))x_i = 0 \\ \sum_{i=1}^{n} (y_i - (ax_i + b)) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0 \end{cases}$$

$$a = \frac{(\hat{K}_{xy})_n}{(\hat{D}_x)_n}, \quad b = \overline{y}_B - \frac{(\hat{K}_{xy})_n}{(\hat{D}_x)_n}\overline{x}_n$$

ВОПРОС 38:

Байесовский подход к получению оценок

](*Y, X*) – случайный вектор,
для которого известна плотность $p(Y|x)$.

Для оценки некоторой заданной функции $\varphi(x)$ в качестве ее приближенного значения предлагается искать условное математическое ожидание $E(\varphi(x)|Y)$, вычисляемое по формуле

$$\psi(Y) = \frac{\int \varphi(x)\, p(Y \mid x)\, p(x)\, d\mu(x)}{q(Y)}$$

где $q(y) = \int p(y \mid x)\, p(x)\, d\mu(x)$

ДВУМЕРНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

ВОПРОС 39:

Двумерные случайные величины

Определение.

Двумерной случайной величиной **называют систему из двух случайных величин**

$$(\xi_1, \xi_2)$$

для которой определена вероятность совместного выполнения неравенств

$$P[(\xi_1 < x), (\xi_2 < y)]$$

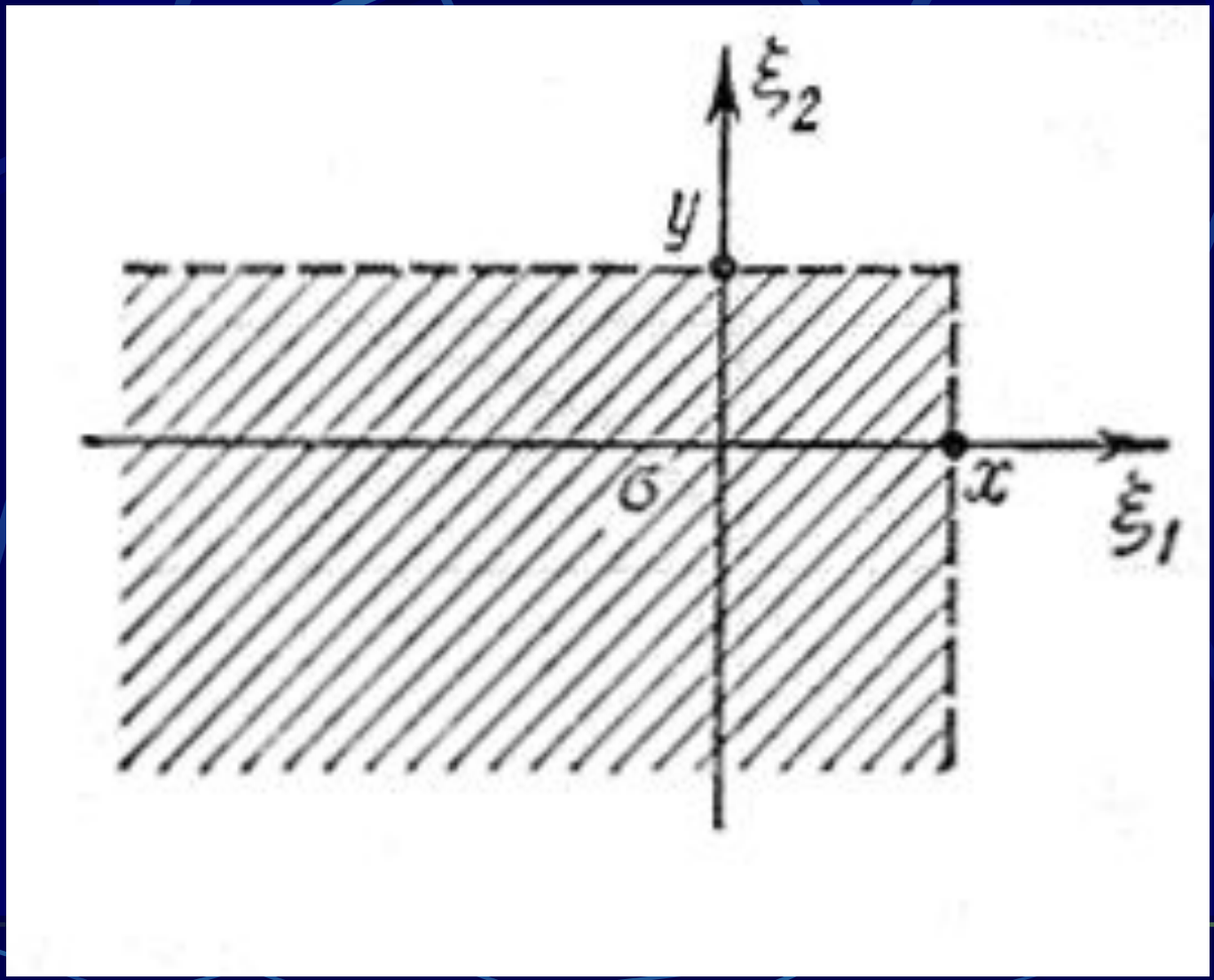
Определение.

Функция двух переменных

$$F(x,y) = P[(\xi_1 < x),(\xi_2 < y)]$$

определенная для любых *x* и *y*, называется *функцией распределения* системы двух случайных величин (ξ_1, ξ_2)

Пример



ДИСКРЕТНАЯ ДВУМЕРНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА

Определение.

Двумерная случайная величина

(ξ_1, ξ_2)

называется *дискретной*, если

ξ_1, ξ_2 **- дискретные величины.**

$$p_{ij} = P[(\xi_1 = x_i, \xi_2 = y_j)]$$

$$F(x, y) = \sum_i \sum_j p_{ij},$$

$$\forall i, j: \quad x_i < x, \quad y_j < y$$

/

Табличная форма задания двумерной случайной величины. Пример

ξ_1 / ξ_2	-1	0	1
0,1	$p_{11}=0,05$	$p_{12}=0,20$	$p_{13}=0,30$
0,2	$p_{21}=0,10$	$p_{22}=0,20$	$p_{23}=0,15$

$$\sum_{i=1}^{2}\sum_{j=1}^{3} p_{i,j} = 1.$$

Определение.

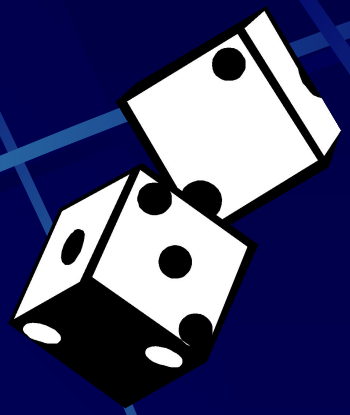
Две дискретные случайные величины называются независимыми, если

$$p_{ij} = P[(\xi_1 = x_i, \xi_2 = y_j)] =$$

$$= P(\xi_1 = x_i) \cdot P(\xi_2 = y_j)$$

$$для\ \forall (i, j)$$

Пример.



$$P[(\xi_1 = x_i, \xi_2 = y_j)] = \frac{1}{36}$$

$$P(\xi_1 = x_i) = \frac{1}{6}; \quad P(\xi_2 = y_j) = \frac{1}{6}.$$

$$P(\xi_1 = x_i) \cdot P(\xi_2 = y_j) = \frac{1}{36}.$$

НЕПРЕРЫВНАЯ ДВУМЕРНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА

Определение.

Двумерная случайная величина (ξ_1, ξ_2) **называется** ***непрерывной*****, если**

$$\exists\ \varphi(x, y) \geq 0 :$$

$$P\{(\xi_1, \xi_2) \in \sigma\} = \iint_{\sigma} \varphi(x, y)\,dx\,dy$$

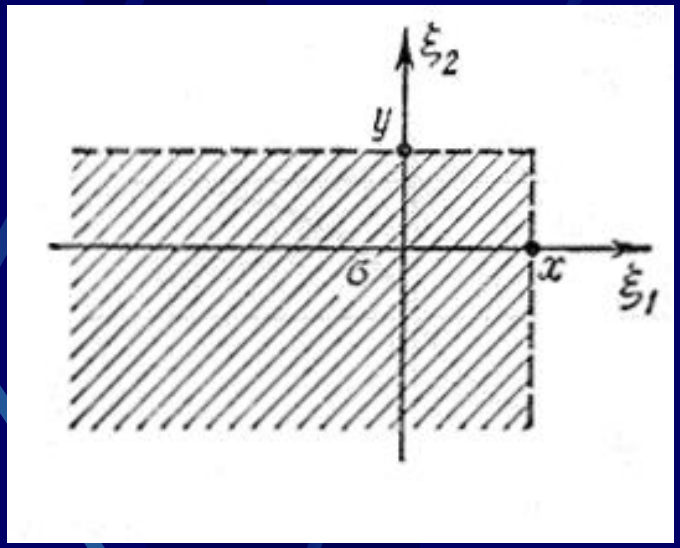
Определение.

Функция $\varphi(x, y)$:

$$P\{(\xi_1, \xi_2) \in \sigma\} = \iint_{\sigma} \varphi(x, y)\,dx\,dy$$

называется плотностью распределения вероятностей системы двух величин (ξ_1, ξ_2)

Пример:



$$F(x,y) =$$

$$= P((\xi_1 < x)\cdot(\xi_2 < y)] =$$

$$= \iint\limits_{\sigma} \varphi(x,y)dxdy =$$

$$= \int\limits_{-\infty}^{x} dx \int\limits_{-\infty}^{y} \varphi(x,y)dy$$

Определение.

Непрерывные случайные величины

$$(\xi_1, \xi_2)$$

называются независимыми, если

$$\varphi(x, y) = \varphi_1(x)\varphi_2(y)$$

Для независимых с.в.

$$F(x,y) = P[(\xi_1 < x)\cdot(\xi_2 < y)] =$$

$$= \int_{-\infty}^{x} dx \int_{-\infty}^{y} \varphi_1(x)\varphi_2(y)dy =$$

$$= \int_{-\infty}^{x} \varphi_1(x)dx \cdot \int_{-\infty}^{y} \varphi_2(y)dy = F_1(x)F_2(y)$$

$$F(x,y) \Rightarrow F_1(x),\ F_2(y)$$

$$F_1(x) = P(\xi_1 < x) =$$

$$= P[(\xi_1 < x),\ \xi_2 \in (-\infty, +\infty)] =$$

$$= \int_{-\infty}^{x} dx \int_{-\infty}^{\infty} \varphi(x,y)\,dy$$

Отсюда

$$\varphi_1(x) = F_1'(x) =$$

$$= \int_{-\infty}^{\infty} \varphi(x, y)\, dy$$

Аналогично:

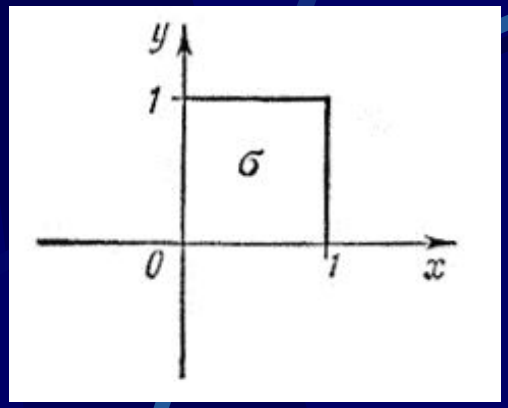
$$F_2(y) = P(\xi_2 < y) =$$

$$= \int_{-\infty}^{y} dy \int_{-\infty}^{\infty} \varphi(x, y) dx$$

$$\varphi_2(y) = F_2'(y) =$$

$$= \int_{-\infty}^{\infty} \varphi(x, y) dx$$

Пример



$$\varphi(x)=\frac{1}{\pi^2(1+x^2)(1+y^2)}$$

$$P((\xi_1,\xi_2)\in\sigma)=\iint_{\sigma}\varphi(x,y)dxdy$$

$$F(x,y)=P[(\xi_1<x)\cdot(\xi_2<y)]=\int_{-\infty}^{x}dx\int_{-\infty}^{y}\varphi(x,y)dxdy=$$

$$\varphi_1(x)=\int_{-\infty}^{\infty}\varphi(x,y)dy$$

Определение.

Двумерная случайная величина распределена нормально, если плотность распределения системы имеет вид:

$$\varphi(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-R^2}}e^{-\frac{1}{1-R^2}\left[\frac{(x-a_1)^2}{2\sigma_1^2}-R\frac{(x-a_1)(y-a_2)}{\sigma_1\sigma_2}+\frac{(y-a_2)^2}{2\sigma_2^2}\right]}$$

Для независимых с.в. (ξ_1, ξ_2)

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\left[\frac{(x-a_1)^2}{2\sigma_1^2} + \frac{(y-a_2)^2}{2\sigma_2^2}\right]}$$

ВОПРОС 40:

Числовые характеристики двумерных случайных величин

Определение.

Начальным моментом порядка (k, s) **двумерной случайной величины (*X, Y*) называется**

$$\alpha_{k,s} = E(X^k Y^s)$$

$$\alpha_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij},$$

$$\alpha_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^s f(x,y)dxdy.$$

Определение.

Центральным моментом порядка (k, s) **двумерной случайной величины (X, Y) называется**

$$\mu_{k,s} = E[(X - E_X)^k (Y - E_Y)^s]$$

$$\mu_{k,s} =$$

$$= \sum_i \sum_j (x_i - E(X))^k (y_j - E(Y))^s p_{ij},$$

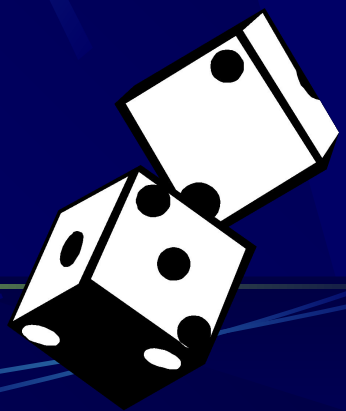
$$\mu_{k,s} =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))^k (y - E(Y))^s f(x,y) dx dy.$$

При этом *E(X) = α1,0, E(Y) = α0,1, D(X) = μ2,0, D(Y) = μ0,2.*

ВОПРОС 41:

Корреляционный момент и коэффициент корреляции



Определение.

Корреляционным моментом системы двух случайных величин называется второй смешанный центральный момент

$$K_{xy} = \mu_{1,1} = E((X - E_X)(Y - E_Y)).$$

$$K_{xy} =$$

$$= \sum_i \sum_j (x_i - E(X))(y_j - E(Y))\, p_{ij},$$

$$K_{xy} =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))\, f(x, y)\, dx\, dy.$$

Определение.

Безразмерной характеристикой коррелированности двух случайных величин является **коэффициент корреляции**

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$$

Для независимых *X* и *Y* f(x =f1(x)f2(y), тогда

$$K_{xy} =$$

$$= \int_{-\infty}^{\infty} (x - E(X)) f_1(x) dx \int_{-\infty}^{\infty} (y - E(Y)) f_2(y) dy = 0$$

Теорема.

$$| r_{xy} | \leq 1.$$

Доказательство

$$Z = \sigma_y X - \sigma_x Y$$

$$D_Z \equiv \alpha_2(Z) - (E_Z)^2 =$$

$$= \sum (\sigma_y x_i - \sigma_x x_i)^2 p_i -$$

$$-(\sum (\sigma_y x_i - \sigma_x y_i) p_i)^2 =$$

$$= 2\sigma_x^2 \sigma_y^2 - 2\sigma_x \sigma_y K_{xy} \geq 0.$$

$$\Rightarrow |K_{xy}| \leq \sigma_x \sigma_y. \Rightarrow \left| \frac{K_{xy}}{\sigma_x \sigma_y} \right| = |r_{xy}| \leq 1,$$

ВОПРОС 44:

Статистическое описание и вычисление характеристик двумерного случайного вектора

Двумерная выборка представляет собой набор значений случайного вектора:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Дескриптивный статистический анализ:

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$$

$$\bar{y}_x = \frac{\sum (y_i \mid x)}{n} \qquad \bar{x}_y = \frac{\sum (x_i \mid y)}{n},$$

$$\hat{\sigma}_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}, \quad \hat{\sigma}_y^2 = \frac{\sum (y_i - \bar{y})^2}{n},$$

$$K_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \bar{x})(y_j - \bar{y})$$

ВИДЫ ЗАВИСИМОСТЕЙ: :

- функциональная зависимость, если каждому возможному значению *X* соответствует одно значение *Y*;

- статистическая, при которой изменение одной величины приводит к изменению распределения другой.