ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ -приёмы и способы научного сбора, анализа и обработки данных для научных и практических целей.

Статистическое наблюдение

Статистическая сводка, группировка

Анализ статистических материала

Шкалы измерений (классификация Стивенсона)

Шкала наименований

Шкала порядка

Шкала отношений

Точность измерений

Абсолютная погрешность измерения *(цена деления прибора)*

$\Delta A = | A - A_0 |$, где A – показание измерительного прибора,

A_0 - истинное значение измеряемой величины.

Относительная погрешность измерения

$$\delta A = \frac{\Delta A}{|A|} \cdot 100\%$$

Табличное и графическое представление экспериментальных данных

Выборка – ряд результатов измерений

Объём выборки – количество результатов измерений

Первичная обработка выборки:

а) данные ранжируются;

б) данные разбиваются на интервалы;

в) составляется вариационный ряд;

г) строятся полигон и гистограмма распределения

Результаты сдачи ЕГЭ по математике (в баллах), группы школьников 11 класса (20 человек)

24, 42, 37, 27, 47, 32, 50, 34, 49, 37,

26, 40, 26, 45, 45, 52, 36, 50, 34, 32

а) **ранжирование** – расстановка результатов измерений в порядке возрастания или убывания

24, 26, 26, 27, 32, 32, 34, 34, 36, 37,

37, 40, 42, 45, 45, 47, 49, 50, 50, 52

б) разбиение на интервалы

Рекомендуемое число интервалов для выборок разного объёма

Объём выборки	10-30	30-50	50-100	100-200	300-400
Число интервалов (k)	4	5-6	7	8	9

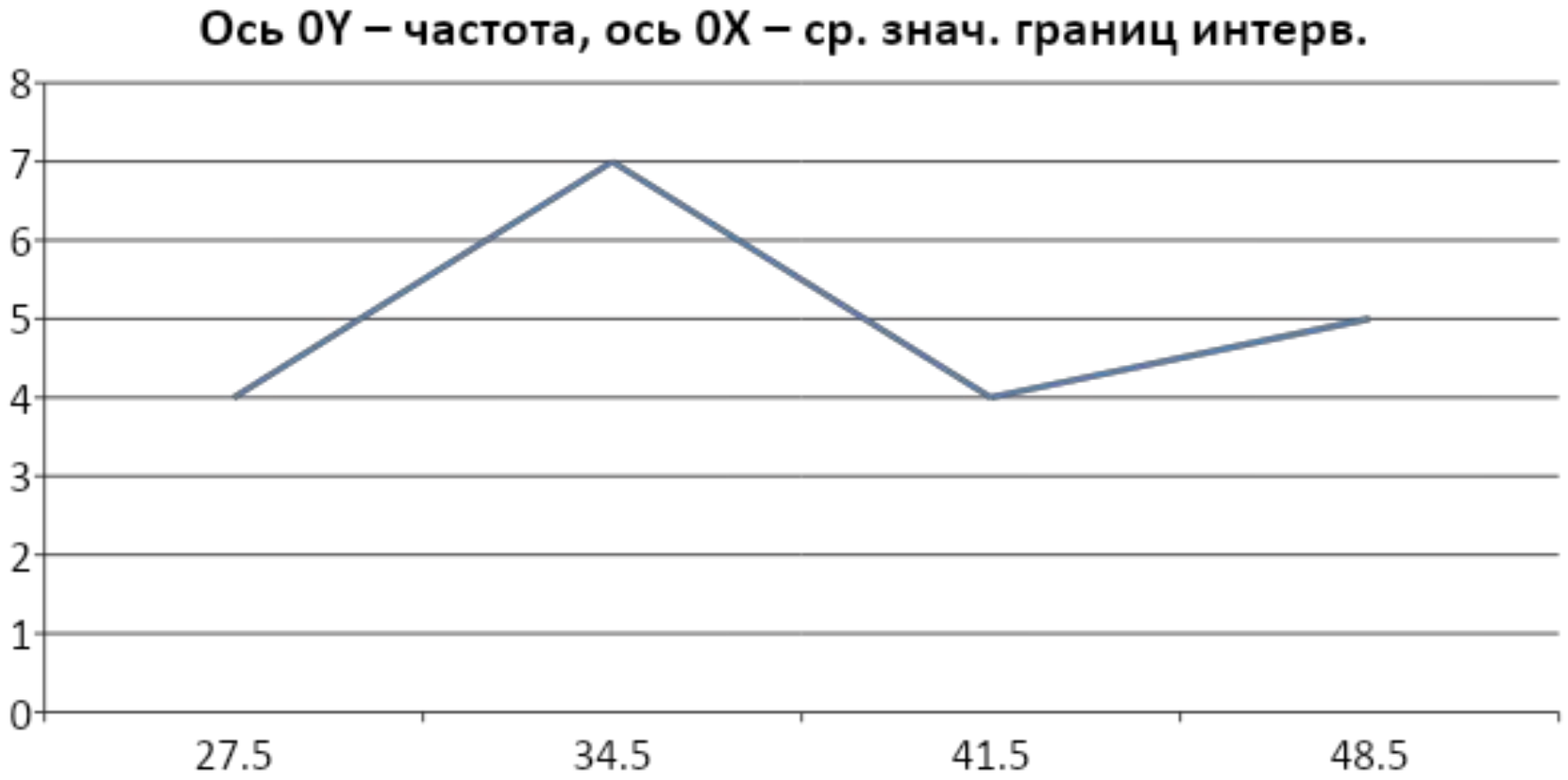
Шаг интервала $h = \frac{X_{max} - X_{min}}{k}$ (Шаг округляется в большую сторону)

h= (52-24):4=7

Вариационный ряд

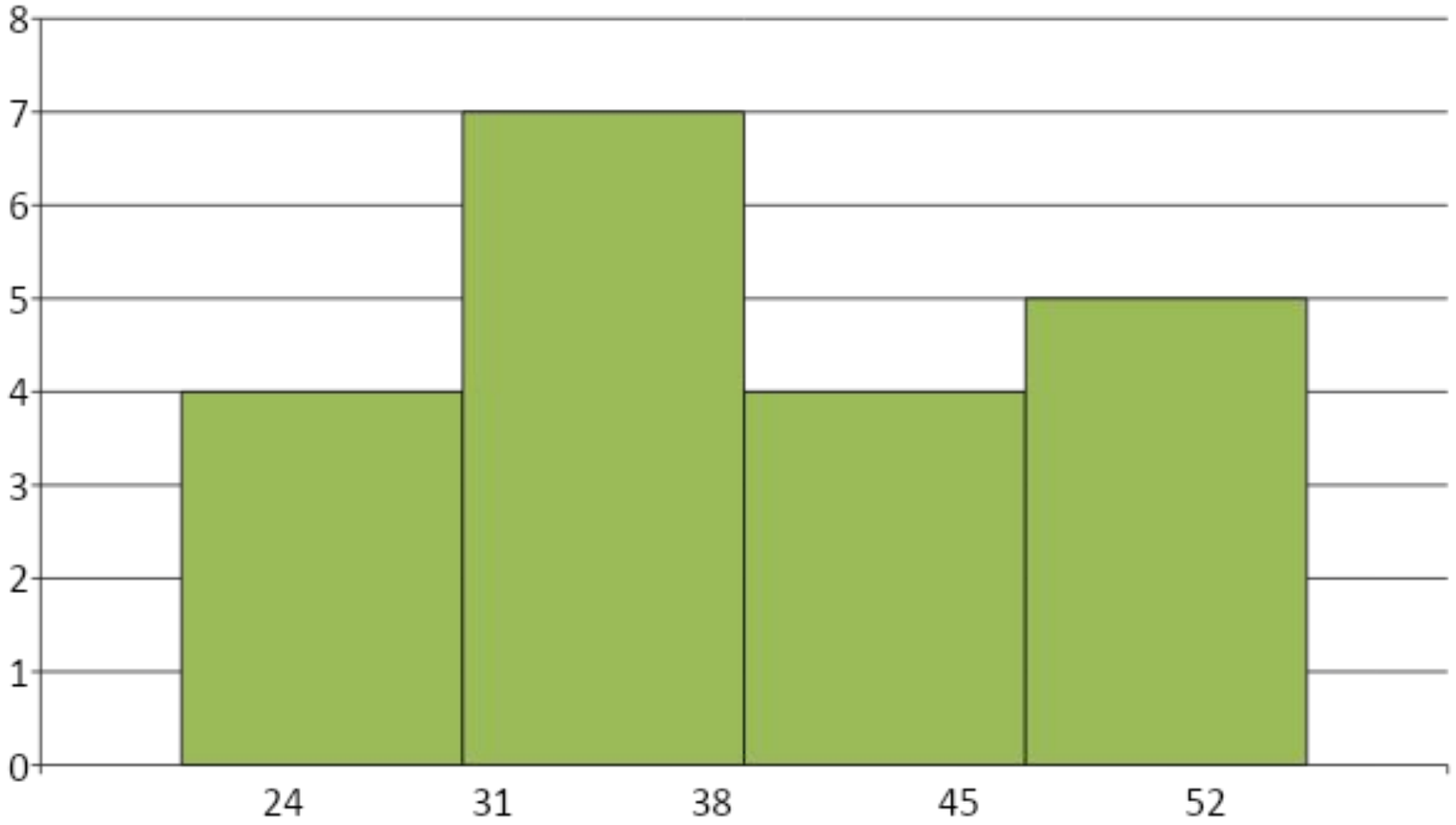
N	Границы интервалов	Серединные значения	Частота	Накопленная частота
1	24, 31	27,5	4	4
2	31, 38	34,5	7	11
3	38, 45	41,5	4	15
4	45, 52	48,5	5	20

Полигон распределения



Гистограмма распределения

ось 0Y – частота, ось 0X – гр. интервалов



Характеристики положения

1. **Среднее арифметическое значение** $\overline{X}$ для ряда измерений $X_1, X_2, \ldots, X_n$ вычисляется по формуле:

$$\overline{X} = \frac{n_1 \cdot X_1 + n_2 \cdot X_2 + \ldots + n_k \cdot X_k}{n_1 + n_2 + \ldots + n_k},$$

где $n_1, n_2, \ldots, n_k$ - частоты соответствующих интервалов,
$X_1, X_2, \ldots$ Хк - среднее арифметическое значение интервалов.

Характеристики положения

2. **Мода** (Мо) – значение в множестве наблюдений, которое встречается наиболее часто (наиболее представительное значение выборки).

Примеры:

2, 3, 4, 5, 5, 5 Мо = 5

5, 5, 1, 1, 2, 2 Мо - не определяется

1, 2, 2, 2, 3, 3, 3, 4 Мо = 2,5

10,11,11,11,13,14,15,15,15,20 Мо = 11, M_o = 15

Модальный интервал – интервал группировки с наибольшей частотой

Характеристики положения

Для вычисления моды используется следующая формула:

$$Mo = X_{\text{нижняя гр.мод.инт.}} + h \cdot \frac{K-L}{(K-L)+(K-P)}$$, где

K - частота модального интервала,

L - частота интервала, ему предшествующего,

P - частота интервала, следующего за модальным.

Характеристики положения

Медиана (Me) – значение в множестве наблюдений, когда одна половина значений экспериментальных данных меньше её, а вторая половина – больше.

Примеры:

2, 2, 3, 4, 5, 5, 6 Me=4

2, 2, 3, 4, 5, 5 Me=(3+4)/2=3,5

Медианный интервал – интервал, в котором накопленная частота впервые больше N/2 (или накопленная частотность больше 0,5).

$$Me = X_{\text{н.гр.мед.инт.}} + h \cdot \frac{N/2 - (\text{накопл.част.инт.,предш.медианному})}{\text{частота медианного интервала}}$$

Характеристики рассеяния (меры изменчивости)

Колеблемость признака характеризуется величинами: размах варьирования, дисперсия, среднее квадратическое отклонение и коэффициент вариации.

Размах варирования $R = X_{max} - X_{min}$

Дисперсия – средний квадрат отклонения значений признака от среднего арифметического

$$б^2 = \frac{(X_1-X)^2 + (X_2-X)^2 + \ldots + (X_n-X)^2}{n}$$

(Если n < 30, то в знаменателе дроби – (n-1)).

Характеристики рассеяния (меры изменчивости)

Если данные сгруппированы в интервальный вариационный ряд, то $\boldsymbol{б^2}$ вычисляется по формуле:

$$б^2 = \frac{n_1 (X_1-\overline{X})^2+n_2(X_2-\overline{X})^2+. . .+n_k(X_k-\overline{X})^2}{n}$$

где к - число интервалов, n_i - частота i-го интервала,

Xi-серединное значение i-го интервала.

(Если n < 30, то в знаменателе дроби – (n-1)).

Характеристики рассеяния (меры изменчивости)

Стандартное отклонение (среднее квадратическое отклонение) – корень квадратный из дисперсии

$$б = \sqrt{б^2}$$

Коэффициент вариации - отношение среднего квадратического отклонения к среднему арифметическому (выражается в процентах):

$$V = \frac{б}{X} \cdot 100\ \%$$

Если V≤10% , то выборка считается однородной.

Нормальный закон распределения результатов эксперимента

Теоретическое распределение большинства результатов измерений описывается формулой нормального распределения, которая впервые была найдена английским математиком де Муавром в 1733 г.:

$$U = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{\frac{-(X-\ddot{X})^2}{2\sigma^2}}$$

, где π=3,14 e=2,718

Нормальный закон распределения результатов эксперимента

- Кривая нормального распределения – графическая интерпретация формулы нормального распределения, колоколообразная кривая, симметричная относительно центра группировки ($\bar{X}$).

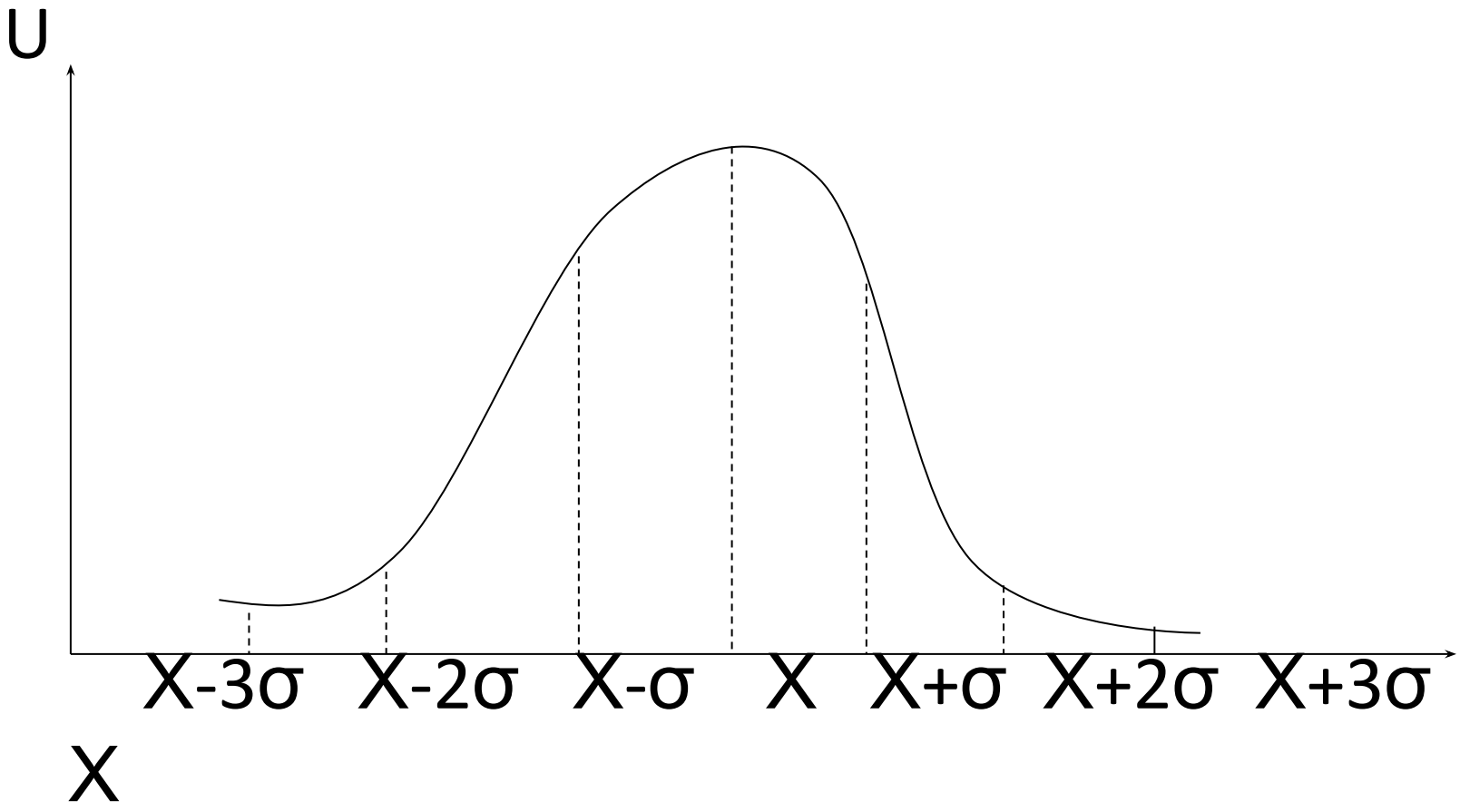
Для оценки варирования результатов измерений используют соотношения:

$X \pm \sigma$ - включает 68,27% всех результатов;

$X \pm 2\sigma$ - включает 95,45% всех результатов;

$X \pm 3\sigma$ - включает 99,73% всех результатов

Кривая нормального распределения



Корреляционный анализ

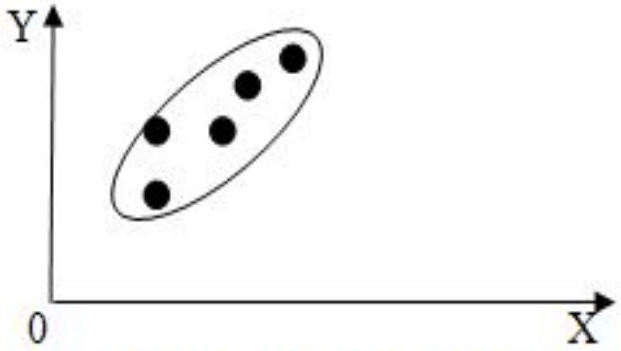
- **Функциональная зависимость** – зависимость, при которой каждому значению одного показателя строго соответствует определённое значение другого.

- **Статистическая взаимосвязь** – зависимость, при которой одному значению показателя соответствует несколько значений другого.

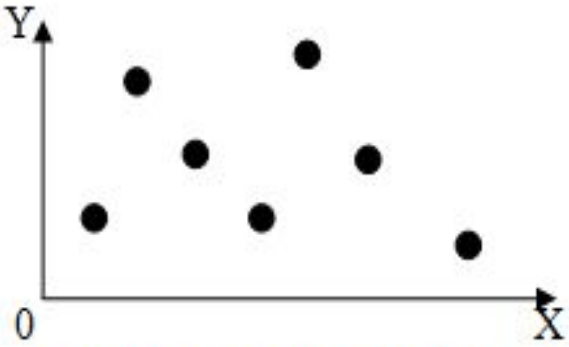
- **Корреляционный анализ** – определение связи между двумя случайными величинами и направления этой связи.

- **Корреляционное поле** (диаграмма рассеивания) - графическое представление измерений в системе координат, когда каждая пара результатов X_i , Y_i будет отображаться точкой (X_i ,Y_i).

Корреляционное поле (диаграмма рассеивания)



Линейная форма взаимосвязи данных



Нелинейная форма взаимосвязи данных

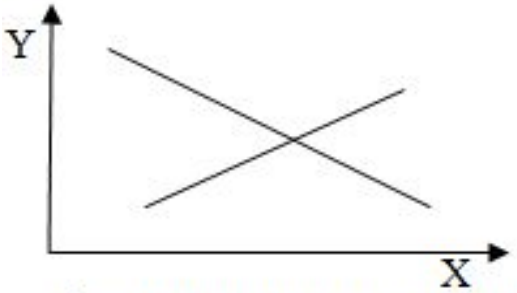
Оценка тесноты взаимосвязи

r – коэффициент корреляции *(абсолютная величина тесноты взаимосвязи).* $0 \leq r \leq 1$

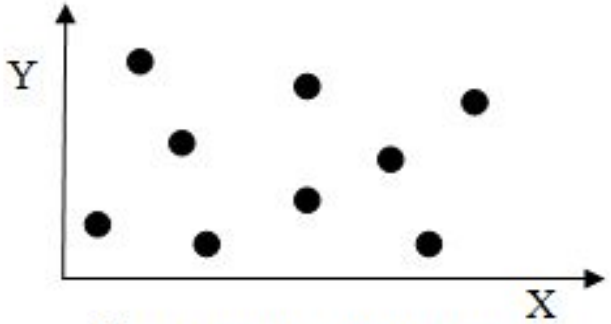
- Если r = 1, то взаимосвязь функциональная.
- Если $0,7 \leq r \leq 0,99$, взаимосвязь сильная статистическая.
- Если $0,5 \leq r \leq 0,69$, то взаимосвязь средняя статистическая.
- Если $0,2 \leq r \leq 0,49$, то слабая статистической связи.
- Если $0,09 \leq r \leq 0,19$, то очень слабая статистическая связь.
- Если r=0, то корреляции нет.

Если знак значения r «+», то связь положительная, если знак «-», то связь отрицательная.

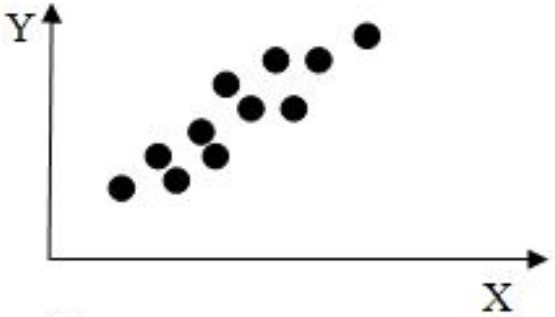
Корреляционное поле
(некоторых значений коэффициента корреляции)



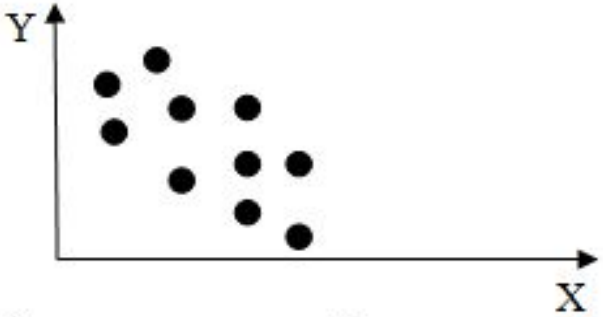
Функциональная зависимость



Отсутствие корреляции



Положительная сильная статистическая связь



Отрицательная слабая статистическая связь

Парный коэффициент корреляции Браве-Пирсона

$$r = \frac{\sum_{i=1}^{n}(X_i - X_{ср.}) \cdot (Y_i - Y_{ср.})}{n \cdot \sigma_x \cdot \sigma_y}$$

где n – число испытуемых,

Xср. и Yср. – средние арифметические значения выборок,

σ_x и σ_y – средние квадратические отклонения.

Парный коэффициент корреляции Браве-Пирсона

Пример: экспериментальные данные представляют собой результаты в запоминании основных формул характеристики рассеяния (меры изменчивости) (в мин.), показанные группой студентов (20 человек).

15,5	16,3	16,7	17,1	15,7	16,4	16,8	17,4	15,8	16,4
16,2	16,5	17,0	17,6	16,2	16,6	17,0	17,9	17,4	16,8

Эти же студенты участвовали в соревнованиях по прыжкам в высоту. Данные этих испытаний представлены в таблице:

100	95	100	105	105	110	95	100	80	85
90	105	90	100	90	95	100	110	105	95

Данные для расчёта: n=20, Xср.=16,7; Yср.=97,7; $\sigma_x = 0{,}637$ и $\sigma_y = 8{,}03$

$$r = 0{,}391$$

Коэффициент корреляции Спирмена

$$R = 1 - \frac{6}{n \cdot (n^2 - 1)} \cdot \sum_{i=1}^{n} d_i$$

где $d_i = dx_i - dy_i$ – разность рангов данной пары показателей x_i и y_i,

n – объём выборки.

Пример: Учащимся 9 класса предложили по 5-бальной школе оценить такое своё качество как справедливость. Затем это качество у каждого оценил классный руководитель. С помощью коэффициента корреляции выяснить вид взаимосвязи между этими данными.

№	1	2	3	4	5	6	7	8	9	10	11	12	13
Уч-ся (x_i)	1	2	2	2	3	3	4	4	4	4	5	5	5
Кл.рук. (y_i)	3	3	4	4	2	1	3	3	5	5	4	3	5
dx_i	1	3	3	3	5,5	5,5	8,5	8,5	8,5	8,5	12	12	12
dy_i	5	5	9	9	2	1	5	5	12	12	9	5	12
$d_i=dx_i-dy_i$	-4	-2	-6	-6	3,5	4,5	3,5	3,5	-3,5	-3,5	3	7	0

$r = 0{,}364$

Вывод: между самооценкой учеников и оценками классного руководителя слабая положительная связь.

Тетрахорический коэффициент корреляции Пирсона r_A

Пример: Группа испытуемых решала две математические задачи: одну по теории вероятностей, другую по математической статистике. Задачу по математической статистике верно решили испытуемые, записаны в списке под номерами: 1, 2, 4, 5, 7, 8, 9, 10. А задачу по математической статистике: 2, 3, 4, 9, 10. Определить степень эквивалентности этих значений.

Тетрахорический коэффициент корреляции Пирсона r_A

№ испытуемого	1	2	3	4	5	6	7	8	9	10
Результата решения задачи 1	1	1	0	1	1	0	1	1	1	1
Результат решения задачи 2	0	1	1	1	0	0	0	0	1	1

«1-1» - обозначим количество таких ситуаций A;
«0-1» - обозначим количество таких ситуаций B;
«1-0» - обозначим количество таких ситуаций C;
«0-0» - обозначим количество таких ситуаций D.

Тетрахорический коэффициент корреляции Пирсона r_A

$$r_A = \frac{|A \cdot D - B \cdot C| - 0{,}5 \cdot n}{\sqrt{(A+B) \cdot (C+D) \cdot (A+C) \cdot (B+D)}}$$

В нашем примере A=4, B=1, C=4, D=1.

$$r_A = \frac{|4 \cdot 1 - 4 \cdot 1| - 0{,}5 \cdot 10}{\sqrt{(4+1) \cdot (4+1) \cdot (4+4) \cdot (1+1)}} = \frac{-5}{20\sqrt{2}} = -0{,}18$$

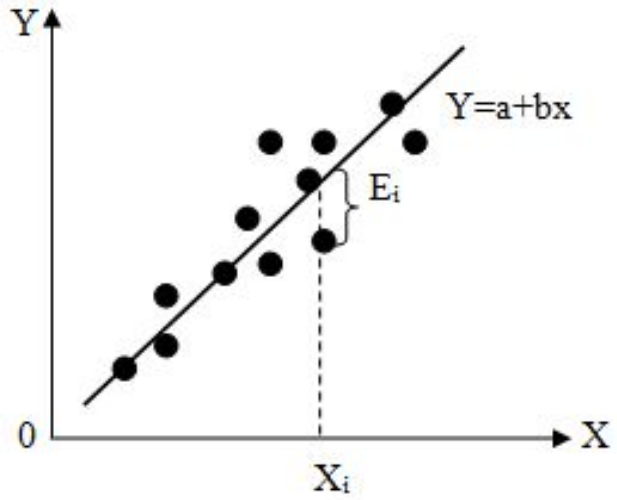
Регрессивный анализ

- это установление зависимости между случайной величиной Y и значением одной или нескольких переменных величин. Зависимость описывается ***уравнением регрессии***.

$$Y = a \cdot X + b$$

a - ? b - ?

Регрессивный анализ



$$E_i = y_i - Y_i = y_i - (a+bx_i) \quad (*),$$

где y_i – фактическое, а Y_i – расчётное значение зависимой переменной.

Регрессивный анализ

$$\Sigma_{i=1}^{n} \mathrm{Ei}^2 \longrightarrow \min.$$

Используя необходимое условие существования минимума функции (МНК), находят для функции $Q = \sum_{i=1}^{n}(\mathrm{yi} - \mathrm{a} - \mathrm{b} \cdot \mathrm{xi})^2$ частные производные по a и b, и приравнивают их к нулю. Учитывая то, что величины a и b являются постоянными, приходят к системе уравнений. Решая её, получают значения коэффициентов регрессии:

$$\begin{cases} \mathrm{b} = \dfrac{n \cdot \sum_{i=1}^{n}(Xi \cdot Yi) - (\sum_{i-1}^{n} Xi) \cdot (\sum_{i=1}^{n} Yi)}{n \cdot (\sum_{i-1}^{n} Xi^2) - (\sum_{i=n}^{n} Xi)^2} \\ \mathrm{a} = \overline{\mathrm{Y}} - \mathrm{b} \cdot \overline{\mathrm{X}}, \end{cases}$$

где n – число данных в каждой из выборок, $\overline{\mathrm{X}}$ и $\overline{\mathrm{Y}}$ – выборочные средние арифметические.

Пример. В таблице приведены результаты, показанные группой школьников (10 человек) в беге на дистанции 30м (x_i,) и 100м (y_i) в секундах.

№	1	2	3	4	5	6	7	8	9	10
Бег на 30м	4,6	4,6	4,7	4,8	4,8	4,8	4,9	4,9	4,9	5,0
Бег на 100м	12,4	12,7	13,0	13,3	13,1	13,1	13,2	13,5	13,6	13,7

Регрессивный анализ

$$\sum_{i=1}^{10} Xi = 48,\ \sum_{i=1}^{10} Yi = 132,\ \sum_{i=1}^{10} Xi^2 = 230{,}56,$$

$$\sum_{i=1}^{10} Xi \cdot Yi = 634{,}08,\ \overline{X} = 4{,}8\ ,\ \overline{Y} = 13{,}2.$$

Подставим найденные значения в формулы, получим

$b = (10 \cdot 634{,}08 - 48 \cdot 132) : (10 \cdot 230{,}56 - 48^2) = 3$

$a = 13{,}2 - 3 \cdot 4{,}8 = -1{,}2$

Таким образом, $Y = -1{,}2 + 3 \cdot X$