

# Исследование корреляции

# Корреляция

Между различными явлениями существуют сложные и многообразные связи. Их можно классифицировать.

В технике и естествознании часто говорят о функциональной зависимости. Например скорость выведения лекарственного вещества из организма.

Однако, многие явления происходят при воздействии многочисленных факторов, в этом случае, связь теряет свою строгую функциональность.

В результате, одна случайная переменная реагирует на изменения другой переменной изменением своего закона распределения.

**Корреляция** – это зависимость между двумя случайными величинами.

- Изучение статистических зависимостей основывается на исследовании таких связей между случайными переменными, при которых значение одной изменяется в зависимости от того, какие значения принимает другая.
- Так как понятие статистической зависимости относится к осредненным условиям, прогнозы не могут быть безошибочными. Применяя некоторые вероятностные методы, можно вычислить вероятность того, что ошибка прогноза не выйдет за определенные границы.
- В исследованиях между изучаемыми признаками чаще всего наблюдаются корреляционные взаимосвязи. (Связь роста с весом, прыжки в длину и бег на короткие дистанции).

# **Виды взаимосвязи**

# функциональная взаимосвязь

- Функциональной называется взаимосвязь, при которой каждому значению одного показателя соответствует строго определенное значение другого.

# Статистическая взаимосвязь

- Статистической взаимосвязью называется взаимосвязь, при которой одному значению первого показателя может соответствовать несколько значений второго показателя.

# Корреляционный анализ

- Корреляционный анализ состоит в определении степени связи между двумя случайными величинами ( $Y$  и  $X$ ).

# Основные задачи корреляционного анализа

- определение формы связи (линейная, нелинейная);
- определение направления связи (положительная связь или отрицательная);
- определение степени или тесноты взаимосвязи (слабая, средняя, сильная).



# Форма зависимости

# Форма зависимости

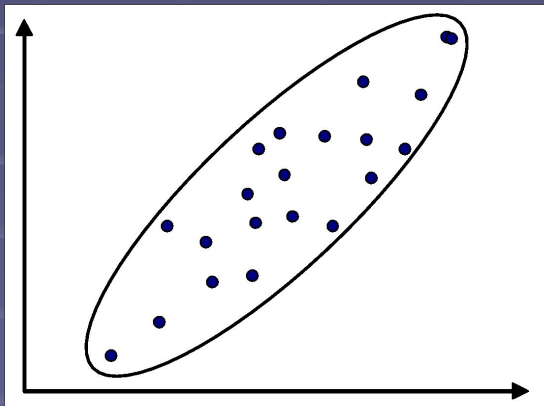


Рис 1. Линейная  
статистическая связь

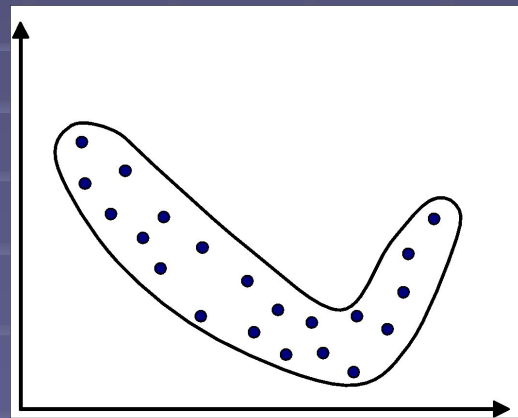


Рис 2. Нелинейная  
статистическая связь

# **Направленность взаимосвязи**

# Направленность взаимосвязи

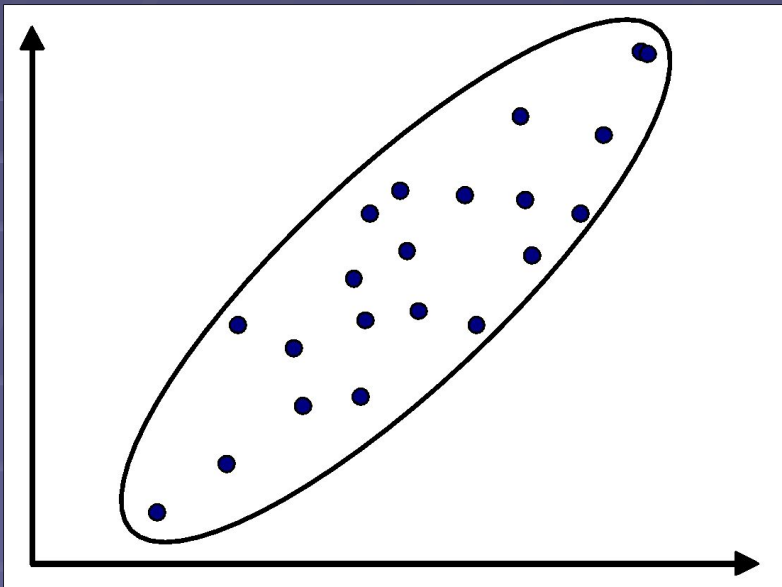


Рис 3. Положительная  
направленность

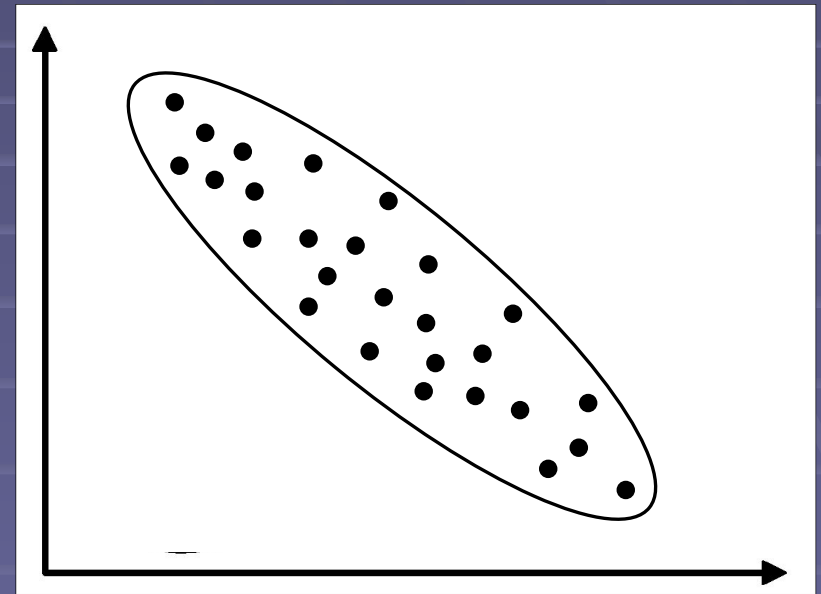


Рис 4. Отрицательная  
направленность

# **Теснота (сила) взаимосвязи**

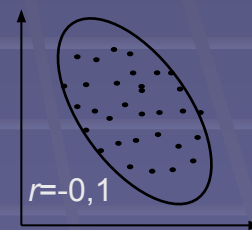
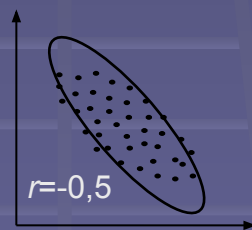
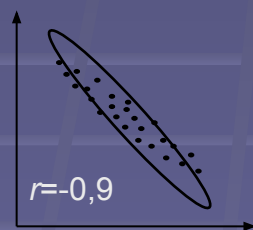
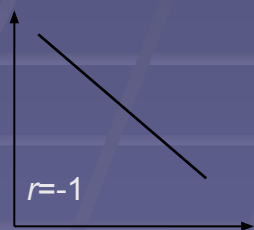
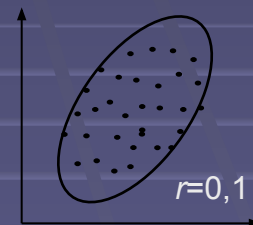
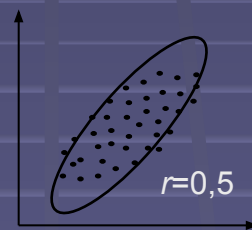
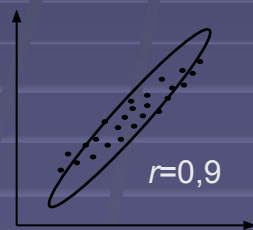
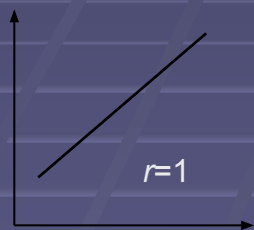
# Диапазон коэффициента корреляции

$$-1 \leq r \leq 1$$

# Построение корреляционного поля

- Пару случайных чисел  $x$  и  $y$ , представляющих собой результаты измерения спортивных результатов, можно изобразить **графически** в прямоугольной системе координат в виде совокупности точек с координатами  $x$ ,  $y$ . Множество этих точек образуют графическую зависимость, называемую **корреляционным полем** или **диаграммой рассеивания**.
- Визуальный анализ графика позволяет выявить как форму, так и направленность и силу взаимосвязи.
- Корреляционное поле необходимо обвести по краю и рассмотреть полученную фигуру, если обведенный ареал напоминает эллипс, то речь идет о **линейной зависимости**.
- Далее производится анализ графика, если эллипс узкий, то зависимость сильная. По графику можно увидеть положительную или отрицательную направленность.

# Корреляционные поля





# Критерии оценки силы взаимосвязи в корреляции

$|r| = 1$  (функциональная зависимость)

$0.7 \leq |r| \leq 0.99$  (сильная зависимость)

$0.5 \leq |r| \leq 0.69$  (средняя зависимость)

$0.2 \leq |r| \leq 0.49$  (слабая зависимость)

$0.09 \leq |r| \leq 0.19$  (очень слабая зависимость)

$r = 0$  (зависимости нет)

# Коэффициент детерминации

- Коэффициент детерминации ( $R^2$ ) - величина квадрата коэффициента корреляции.

$$D = r^2 \cdot 100\%$$

Величина  $R^2$  показывает долю (%) части варьирования одного из признаков, связанную с варьированием другого

**Коэффициент  
корреляции  
Браве-Пирсона**

# Вычисление коэффициента корреляции Браве-Пирсона

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{nS_x S_y}$$

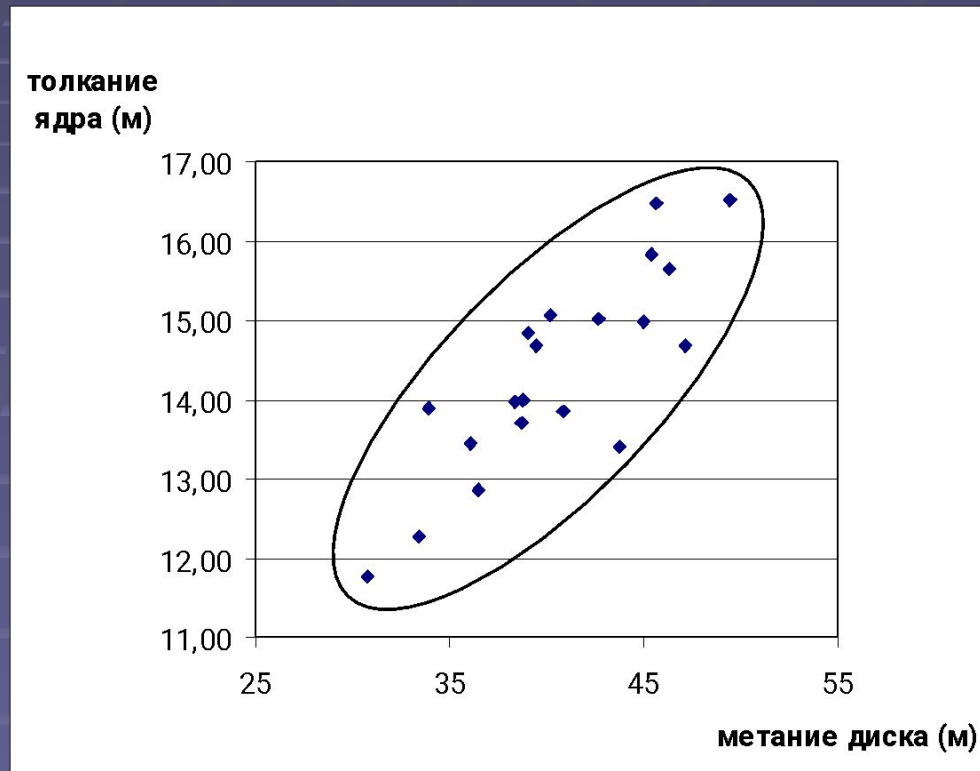
# Этапы проверки гипотезы

- 1. Задаются уровнем значимости  $\alpha=0,05$ .
- 2. Формулируют гипотезы  $H_0: r=0$   $H_1: r \neq 0$
- 3. Рассчитывают эмпирическое значение  $t$  критерия Стьюдента
- 4. Определяют критическое значение критерия  $t_{кр}$
- 5. Сравнивают эмпирическое значение критерия с критическим

# Пример исследования корреляции

- Результаты  
метания диска и  
толкания ядра

# Корреляционное поле



■Рис. 6. Корреляционное поле

1	2	3	4	5	6	7	8
$i$	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	40,9	13,84	0,318	-0,502	-0,159636	0,101124	0,252004
2	49,47	16,51	8,888	2,168	19,269184	78,996544	4,700224
3	45,44	15,83	4,858	1,488	7,228704	23,600164	2,214144
4	45,64	16,47	5,058	2,128	10,763424	25,583364	4,528384
5	43,76	13,40	3,178	-0,942	-2,993676	10,099684	0,887364
6	36,08	13,45	-4,502	-0,892	4,015784	20,268004	0,795664
7	33,92	13,88	-6,662	-0,462	3,077844	44,382244	0,213444
8	40,22	15,06	-0,362	0,718	-0,259916	0,131044	0,515524
9	39,47	14,68	-1,112	0,338	-0,375856	1,236544	0,114244
10	38,38	13,97	-2,202	-0,372	0,819144	4,848804	0,138384
11	38,68	13,70	-1,902	-0,642	1,221084	3,617604	0,412164
12	47,14	14,68	6,558	0,338	2,216604	43,007364	0,114244
13	36,47	12,85	-4,112	-1,492	6,135104	16,908544	2,226064
14	39,03	14,84	-1,552	0,498	-0,772896	2,408704	0,248004
15	46,3	15,65	5,718	1,308	7,479144	32,695524	1,710864
16	33,47	12,27	-7,112	-2,072	14,736064	50,580544	4,293184
17	44,97	14,97	4,388	0,628	2,755664	19,254544	0,394384
18	38,83	13,99	-1,752	-0,352	0,616704	3,069504	0,123904
19	42,68	15,03	2,098	0,688	1,443424	4,401604	0,473344
20	30,79	11,77	-9,792	-2,572	25,185024	95,883264	6,615184
Сумма	811,64	286,84	0	0	102,40092	481,0747	30,97072



# Вычисление суммы значений $x_i$ и $y_i$

$$\sum_{i=1}^{20} x_i = x_1 + x_2 + x_3 + \dots + x_{19} + x_{20} = 811,64$$

$$\sum_{i=1}^{20} y_i = y_1 + y_2 + y_3 + \dots + y_{19} + y_{20} = 286,84$$

# Определение средних значений признаков $x_i$ и $y_i$

$$\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{811,64}{20} = 40,58$$

$$\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = \frac{286,64}{20} = 14,34$$

# Соответствующие суммы

$$\sum_{i=1}^{20} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 102,4009$$

$$\sum_{i=1}^{20} (x_i - \bar{x})^2 = 481,0747$$

$$\sum_{i=1}^{20} (y_i - \bar{y})^2 = 30,9707$$

# Значение коэффициента корреляции Браве-Пирсона

$$r = \frac{102,4009}{\sqrt{481,0747 \cdot 30,9707}} = 0,839$$

- Коэффициент корреляции лежит в интервале  $0,7 \leq |r| \leq 0,99$ , поэтому можно сделать предположение о том, что между результатами, показанными спортсменами в метании диска, и результатами, показанными ими в толкании ядра, существует **линейная положительная сильная** статистическая взаимосвязь.

# Коэффициент детерминации

$$D = r^2 \cdot 100\% = 0,839 \cdot 0,839 \cdot 100\% = 70,4\%$$

- Таким образом, 70% взаимосвязи между двумя наборами данных объясняется их взаимовлиянием. Остальная часть вариации обусловлена воздействием других неучтенных причин.

# Вывод о статистической значимости коэффициента корреляции

- Между результатами, показанными спортсменами в метании диска, и результатами, показанными ими в толкании ядра, существует значимая положительная взаимосвязь.

# Коэффициенты вариации

$$V_x = \frac{\sigma_x}{\bar{x}} \cdot 100\% = \frac{5,03}{40,58} \cdot 100\% = 12,4\%$$

$$V_y = \frac{\sigma_y}{\bar{y}} \cdot 100\% = \frac{1,28}{14,3} \cdot 100\% = 8,9\%$$

- Поскольку коэффициент вариации у результатов в метании диска больше, чем у результатов в толкании ядра, то этот признак варьирует сильнее



# Алгоритм №1 вычисления коэффициента корреляции



1. Находим  $x$  и  $y$
2. Заполняем таблицу

3. Находим

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

4. Находим

$$r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot \sigma_x \cdot \sigma_y}$$



## 5. Проверка значимости выборочного коэффициента корреляции

**Вычислить наблюдаемое значение критерия**

$$T_{\text{набл}} = \frac{\bar{r}_{xy} \cdot \sqrt{n-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

**Сравнить числа  $|T_{\text{набл}}|$  и  $T_{\text{крит}}$  :**

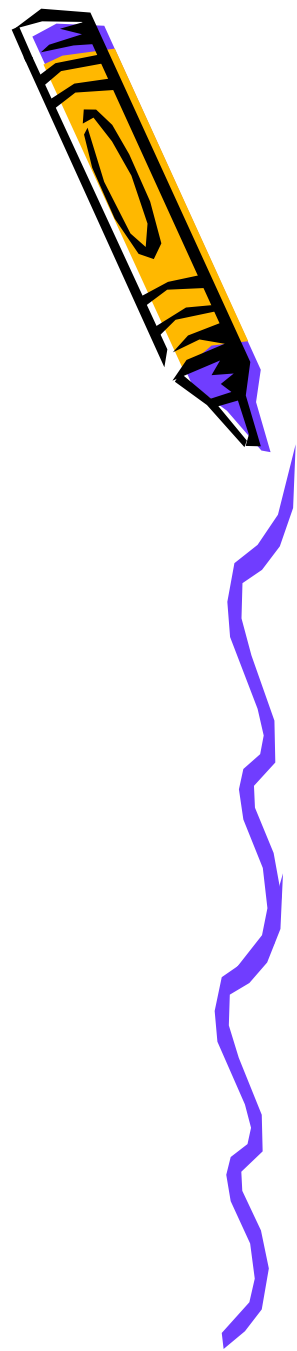
**если  $|T_{\text{набл}}| < T_{\text{крит}}$  , то принять гипотезу  $H_0$  ;**

**если  $|T_{\text{набл}}| > T_{\text{крит}}$  то гипотеза  $H_0$  отвергается**



## 6. Коэффициент детерминации

$$D = r^2 \cdot 100\%$$



# Вспомогательная таблица для расчета коэффициента корреляции

№	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1							
2							
3							
4							
5							
...							
$n$							
$\Sigma$							

# Алгоритм №2 вычисления коэффициента корреляции



1. Находим  $x$  и  $y$
2. Заполняем таблицу

3. Находим 
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad ; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

4. Находим 
$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2$$



5. Находим выборочный корреляционный момент:

$$\bar{\mu}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

6. Находим выборочный коэффициент корреляции:

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



7. Найти оценки параметров линейной регрессии по выборке.

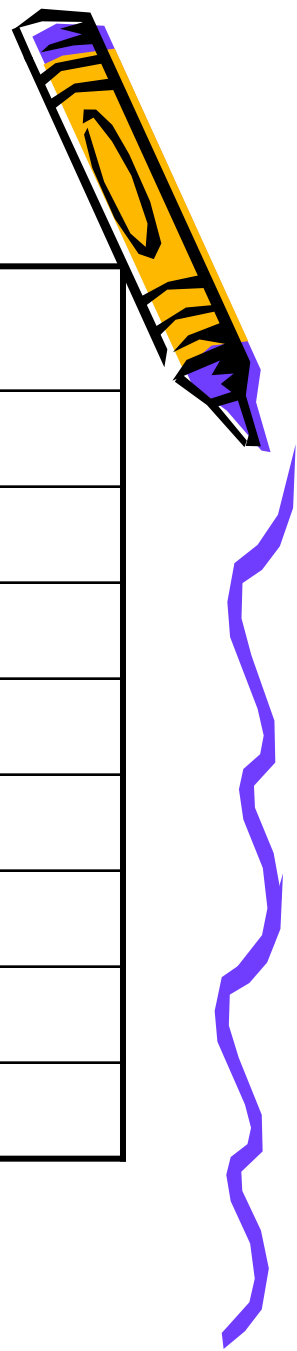
8. Изобразить заданные точки и прямую регрессии.

Уравнение искомой прямой

$$y = ax + b$$



# Вспомогательная таблица для расчета коэффициента корреляции



№	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1					
2					
3					
4					
5					
...					
n					
$\Sigma$					

