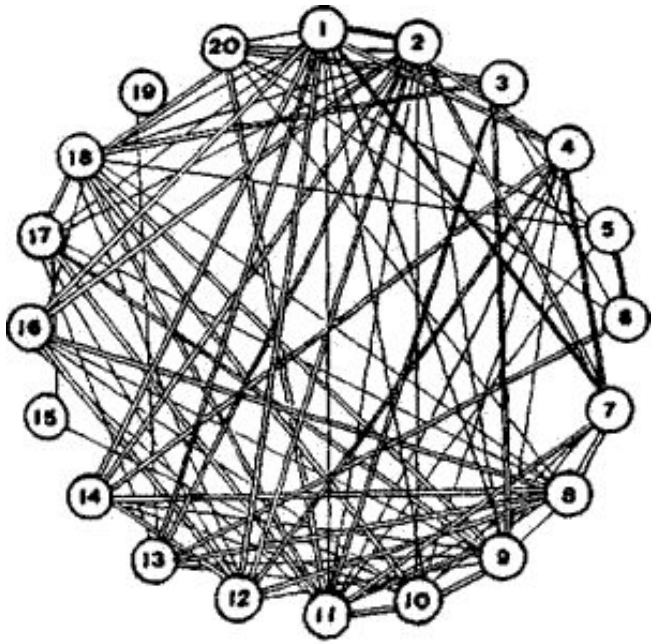
Корреляционный анализ

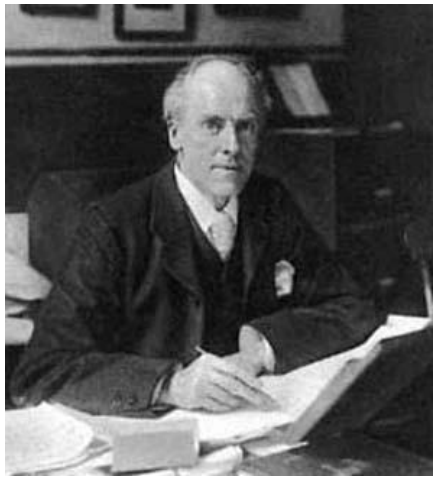
<u>План лекции</u>:

1. **Непараметрические методы измерения тесноты связи.**
2. **Параметрические методы.**



Непараметрические методы измерения тесноты связи

Карл Пирсон (27.03.1857-27.04.1936)



Джордж Юл (18.02.1871- 26.06.1951)

Методы измерения тесноты взаимосвязи

параметрические и *непараметрические*

Непараметрические методы применяются для измерения тесноты связи качественных и альтернативных признаков, а так же количественных признаков, распределение которых отличается от нормального.

Измерение связи альтернативных признаков

коэффициент ассоциации Д. Юла

$$K_a = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

коэффициент контингенции К. Пирсона.

$$K_к = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(d+c)(a+c)(b+d)}}$$

таблица сопряжённости (Таблица 1).

Таблица 1. Заболеваемость сотрудников фирмы

Число сотрудников	не заболевших	заболевших	*Итого*
сделавших прививку	$a = 86$	$b = 14$	$a + b = 100$
не сделавших прививку	$c = 22$	$d = 28$	$c + d = 50$
Итого	$a + c = 108$	$b + d = 42$	$a + b + c + d = 150$

$|K_a| \geq 0{,}5$

$|K_к| \geq 0{,}3$

Таблица 1. Заболеваемость сотрудников фирмы

Число сотрудников	не заболевших	заболевших	*Итого*
сделавших прививку	$a = 86$	$b = 14$	$a + b = 100$
не сделавших прививку	$c = 22$	$d = 28$	$c + d = 50$
Итого	$a + c = 108$	$b + d = 42$	$a + b + c + d = 150$

$$K_a = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

$$K_к = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(d+c)(a+c)(b+d)}}$$

$$K_a = \frac{86 \cdot 28 - 14 \cdot 22}{86 \cdot 28 + 14 \cdot 22} = 0{,}773$$

$$K_к = \frac{86 \cdot 28 - 14 \cdot 22}{\sqrt{100 \cdot 50 \cdot 108 \cdot 42}} = 0{,}441$$

коэффициенты взаимной сопряжённости признаков К. Пирсона или *А.А. Чупрова*:

(1874-1926)

$$K_{П} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

$$K_{Ч} = \sqrt{\frac{\varphi^2}{\sqrt{(K_1-1)(K_2-1)}}}$$

K_1 и K_2 – число групп первого и второго признака, соответственно

$$\varphi^2 = \sum \frac{n_{xy}^2}{n_x n_y} - 1$$

Таблица 2. Корреляционная таблица

$Y \setminus X$	10	20	30	40	50	60	n_y
2			4	4	2	10	20
4		2	15	10	24	1	52
6		10	42	40	11		103
8	3	25	12	16	3		59
10	9	3	3	1			16
n_x	12	40	76	71	40	11	250

$$\varphi^2 = \left(\frac{10^2}{11 \cdot 20} + \frac{2^2}{40 \cdot 20} + \frac{4^2}{71 \cdot 20} + \frac{4^2}{76 \cdot 20} + \frac{1^2}{11 \cdot 52} + \ldots + \frac{9^2}{12 \cdot 16} \right) - 1 = 1{,}163$$

$$K_{П} = \sqrt{\frac{1{,}163}{1 + 1{,}163}} = 0{,}733$$

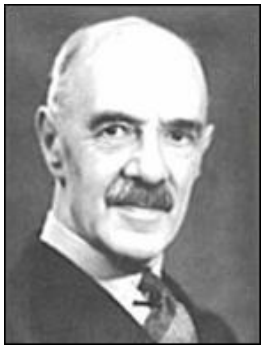
$$K_{Ч} = \sqrt{\frac{1{,}163}{\sqrt{(5-1)(6-1)}}} = 0{,}510$$

$$K_{П} > K_{Ч}$$

Методы оценки силы взаимодействия

Ранжирование (от английского *rank – ранг*, *класс*, *звание*)) – это упорядочение объектов в порядке убывания (возрастания) степени проявления в них изучаемого свойства. *Ранг* равен порядковому месту значений признака в упорядоченном таким образом ряду.

коэффициенты корреляции рангов Спирмена и *Кендалла*



Чарльз Эдвард Спирмен

(10.09.1863 – 17.09.1945)

Моррис Джордж Кендалл

(06.09.1907 – 29.03.1983)

Коэффициент Спирмена

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i – разность рангов двух показателей,

n – число наблюдаемых пар значений.

ρ может принимать значения от -1 до +1.

x	1	2	3	4	5	6	7	8	9	
y	4	3	8	1	9	7	5	2	6	
d	-3	-1	-5	3	-4	-1	2	6	3	Σ
d^2	9	1	25	9	16	1	4	36	9	**110**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 110}{9(9^2-1)} = 0{,}083$$

Коэффициент корреляции рангов Кендалла

P Q

$S = P + Q$:

$$P_{\max} = (n-1) + (n-2) + \ldots + 3 + 2 + 1 = \frac{n(n-1)}{2}$$

$$\tau = \frac{S}{n(n-1)/2} = \frac{2S}{n(n-1)}$$

М. Кендалл и Б. Смит: *коэффициент конкордации* (множественный коэффициент ранговой корреляции)

$$W = \frac{12S}{m^2(n^3 - n)}$$

где S – сумма квадратов отклонений суммы m рангов от их средней величины; m – число ранжируемых признаков; n – число ранжируемых единиц (число наблюдений).

Коэффициент конкордации W принимает значения от 0 до 1.

Параметрические методы

коэффициент корреляции

$$r = \frac{\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)}{n} = \frac{\overline{(x-\bar{x})(y-\bar{y})}}{\sigma_x\sigma_y} = \frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}$$

ковариация

$$\mathrm{cov} = \overline{(x-\bar{x})(y-\bar{y})} = \overline{xy}-\bar{x}\bar{y}$$

$$r = \rho_{yx} \frac{\sigma_y}{\sigma_x} = \rho_{xy} \frac{\sigma_x}{\sigma_y}$$

$r = 0$ между признаками отсутствует линейная зависимость, при $r = \pm 1$ –зависимость между ними функциональная

r^2 - *коэффициент детерминации*

$$\overline{y^2} = \frac{1}{250}(11 \cdot 45^2 + 40 \cdot 55^2 + 71 \cdot 65^2 + 76 \cdot 75^2 +$$

$$+ 40 \cdot 85^2 + 12 \cdot 95^2) = 5072{,}2$$

$$\sigma_y^2 = \overline{y^2} - \bar{y}^2 = 5072{,}2 - 70{,}2^2 = 102{,}0; \quad \sigma_y = \sqrt{102{,}0} = 10{,}10$$

$$\sigma_x = \sqrt{144{,}16} = 12{,}01$$

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{12355{,}6 - 174{,}96 \cdot 70{,}2}{12{,}01 \cdot 10{,}10} = 0{,}605$$

Шкала Чеддока

Таблица 3. Шкала Чеддока

Коэффициент корреляции $\|r\|$	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99	1,0
Характеристика связи	Слабая	Умеренная	Заметная	Тесная	Очень тесная	Функциональная

корреляционное отношение (*коэффициент корреляции по Пирсону*)

эмпирическое

теоретическое

$\bar{y}_x$

$$\eta_{эмп} = \sqrt{\frac{\delta_y^2}{\sigma_y^2}}$$

$$\eta_{теор} = \sqrt{\frac{\delta_y^2}{\sigma_y^2}}$$

σ_y^2 – общая дисперсия,

$$\delta_y^2 = \frac{\sum(\bar{y}_x - \bar{y})^2}{n}$$

δ_y^2 – межгрупповая дисперсия.