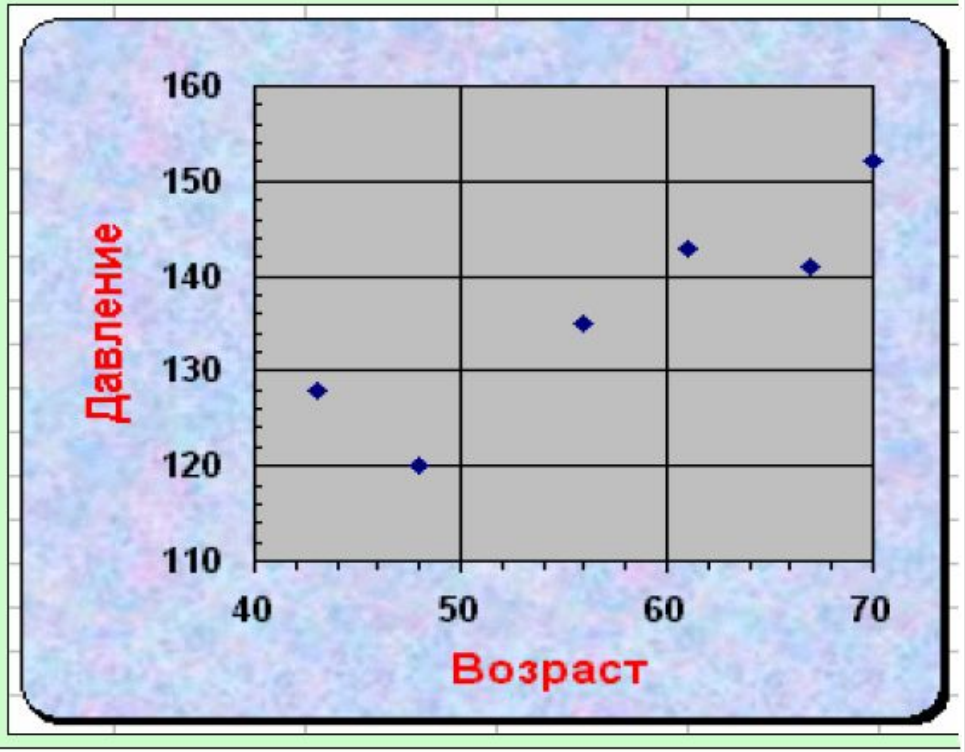
Тема 3.

Корреляционный анализ

Диаграмма рассеяния

Пример 1. Построить диаграмму рассеяния для результатов наблюдения за возрастом и артериальным давлением группы людей, приведенных в таблице.

№	Возраст, лет (x)	Давление, мм.рт.ст. (y)
1	43	128
2	48	120
3	56	135
4	61	143
5	67	141
6	70	152



Свойства коэффициента корреляции

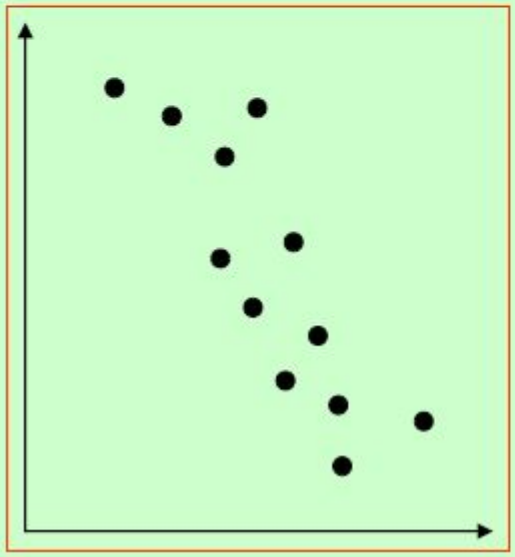
Основные свойства коэффициента корреляции:

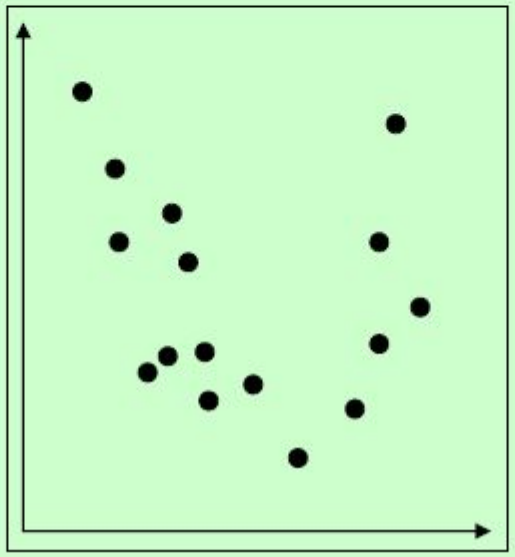
Сильная обратная связь

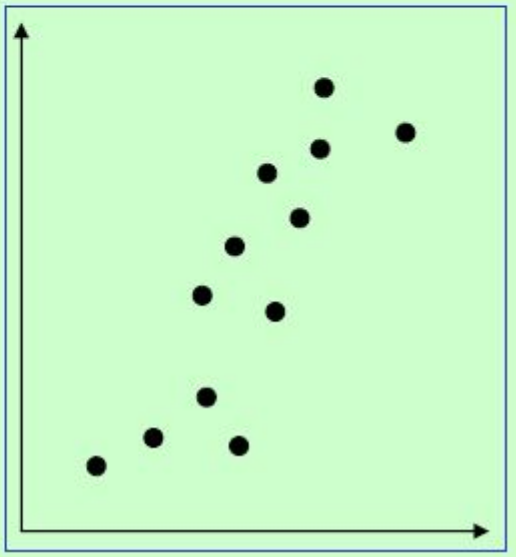
Нет линейной связи

Сильная прямая связь

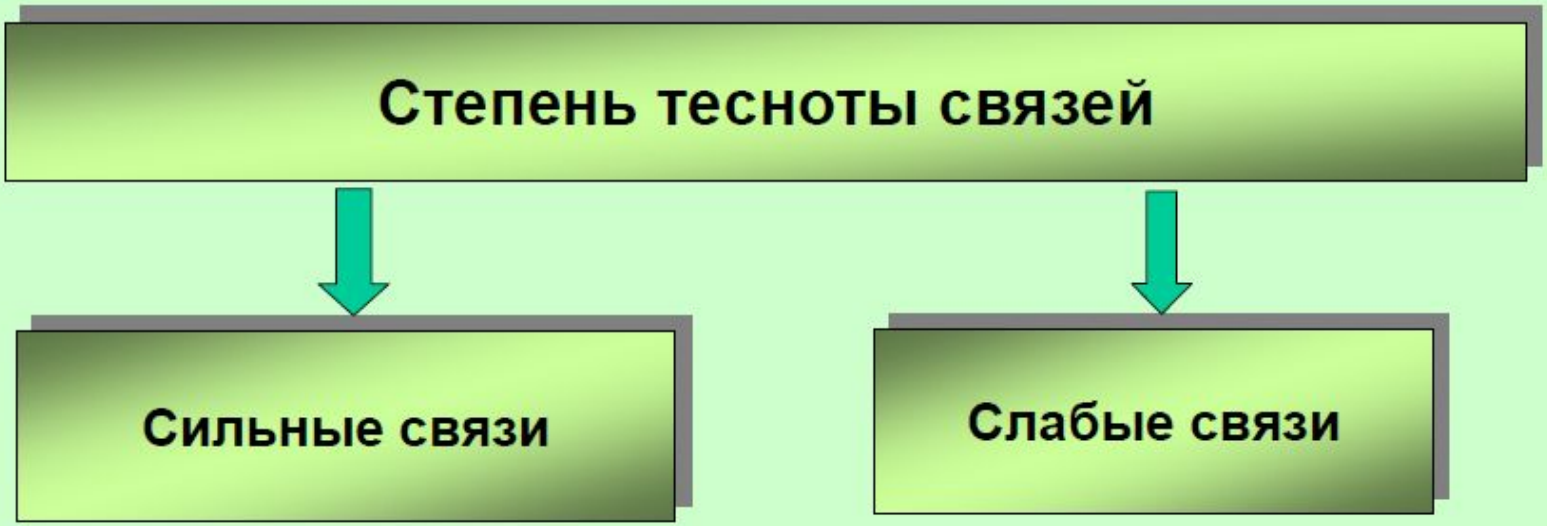
-1 0 +1







Шкала силы связи



шкала Чеддока

Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

средняя сильная

Корреляционная таблица

Двумерное распределение частот можно представить в виде таблицы:

X \ Y	y_1	y_2	…	y_l	m_{i*}
x_1	m_{11}	m_{12}	…	m_{1l}	m_{1*}
x_2	m_{21}	m_{22}	…	m_{2l}	m_{2*}
…	…	…	…	…	…
x_k	m_{k1}	m_{k2}	…	m_{kl}	m_{k*}
m_{*j}	m_{*1}	m_{*2}	…	m_{*l}	n

m_{ij} - частота наблюдаемой пары (x_i, y_j)

$$n = \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij} \qquad m_{i*} = \sum_{j=1}^{l} m_{ij} \qquad m_{*j} = \sum_{i=1}^{k} m_{ij}$$

Линейный коэффициент корреляции.

Определение. ***Линейным коэффициентом корреляции*** признаков ξ и η, вычисленным по выборке Z называется величина:

$$r_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_X}\right)\left(\frac{y_i - \bar{y}}{\sigma_Y}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\sigma_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$.

Вычисление коэффициента

(II способ вычисления коэфф.)

Парный коэффициент корреляции
если оба признака измерены
в количественных шкала:

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

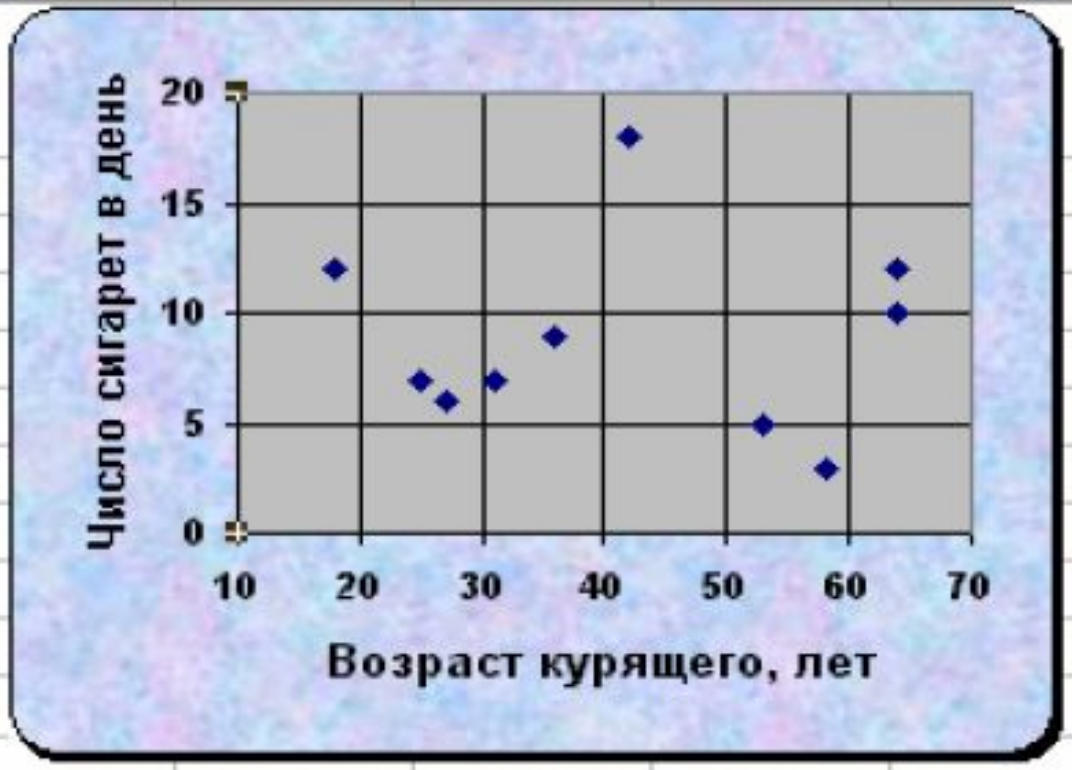
где σ_x, σ_y – среднее квадратическое отклонение признаков x, y соответственно,

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

– ковариация признаков x и y,
где m_{ij} – абсолютная частота наблюдаемой пары значений (x_i, y_j) в выборке объема n.

Пример коэффициента

	A	B
1	Возраст курящего, x	Число сигарет в день, y
2	27	6
3	64	10
4	36	9
5	42	18
6	31	7
7	18	12
8	53	5
9	64	12
10	58	3
11	25	7
12		
13	Коэффициент кор-реляции	-0,03507



Значимость корреляции

Корреляционный анализ
Проверка значимости коэффициента корреляции

$$t_{расч} = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}}$$

$t_р > t_{табл.}$ - существует зависимость между X и Y

$t_р <= t_{табл.}$ - величины X и Y независимы

Пример таблицы для определения $t_{табл.}$

Таблица значений критерия Стьюдента (t-критерия)

Критические значения коэффициента Стьюдента (t-критерия) для различной доверительной вероятности *p* и числа степеней свободы *f*:

f	*p*							
	0.80	0.90	0.95	0.98	0.99	0.995	0.998	0.999
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619
2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869

f=n-2 – число степеней свободы,

p – доверительная вероятность

Ранговая (порядковая) шкала

Теорема. Пусть признаки ξ и η измерены в ранговой шкале, а результаты наблюдений за этими признаками представлены выборками

$$P = \left\{ p_i, i = \overline{1,n} \right\} \quad \text{и} \quad Q = \left\{ q_i, i = \overline{1,n} \right\}$$

Тогда коэффициент корреляции между ξ и η равен:

$$\rho = 1 - 6 \frac{\sum_{i=1}^{n} (p_i - q_i)^2}{n(n^2 - 1)}$$

формула Спирмена

Номинальная (качественная) шкала

Коэффициент корреляции Пирсона
если оба признака измерены
в дихотомической шкале,
(т. е. принимают значения – 0 и 1):

$$\varphi = \frac{p(x,y) - p(x)\,p(y)}{\sqrt{p(x)(1-p(x))\,p(y)(1-p(y))}}$$

где $p(x)$ – доля выборочных значений признака x=1;

$p(y)$ – доля выборочных значений признака y=1;

$p(x, y)$ – доля вариант (x_i, y_i) с единичными значениями у обоих признаков.