Кафедра «Автоматизированные информационные системы»

Дисциплина
«Прикладные статистические методы анализа и прогнозирования социально-экономических и общественно-политических процессов»

Тема 3. «Выявление скрытых закономерностей между социально-экономическими и общественно-политическими показателями»

Занятие 3.4. Лекция: «Парный корреляционный анализ данных»

Учебные вопросы

1. **Корреляционная зависимость. Основные понятия и определения.**
2. **Показатели парной корреляционной зависимости количественных признаков.**
3. **Показатели парной корреляционной зависимости качественных признаков.**

Литература

1. Новиков Е.И. Статистические методы анализа данных в информационно-аналитической деятельности : пособие / Е.И. Новиков. – Орел : Академия ФСО России, 2013. – 190 с.
2. Гмурман В.Е. Теория вероятностей и математическая статистика. Учебное пособие для вузов. М. : Высшая школа, 2011.
3. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. – 1022 с.

1

КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ.

ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Корреляционная зависимость. Основные понятия и определения

Виды корреляционных зависимостей

1. **Функциональная.**
2. **Статистическая.**
3. **Корреляционная.**

Задачи корреляционного анализа

1. Оценивание силы или тесноты зависимости между переменными (верификация известных зависимостей).
2. Выявление неизвестных (скрытых) зависимостей между изучаемыми признаками.
3. Отбор факторов, оказывающих наиболее сильное влияние на результирующий признак.

Корреляционная зависимость. Основные понятия и определения

Относительно характера корреляции

- – положительная (прямая) корреляция
- – отрицательная (обратная) корреляция

Относительно формы связи

- – линейная корреляция
- – нелинейная корреляция

Относительно типа признака

- – корреляция между качественными признаками
- – корреляция между количественными признаками

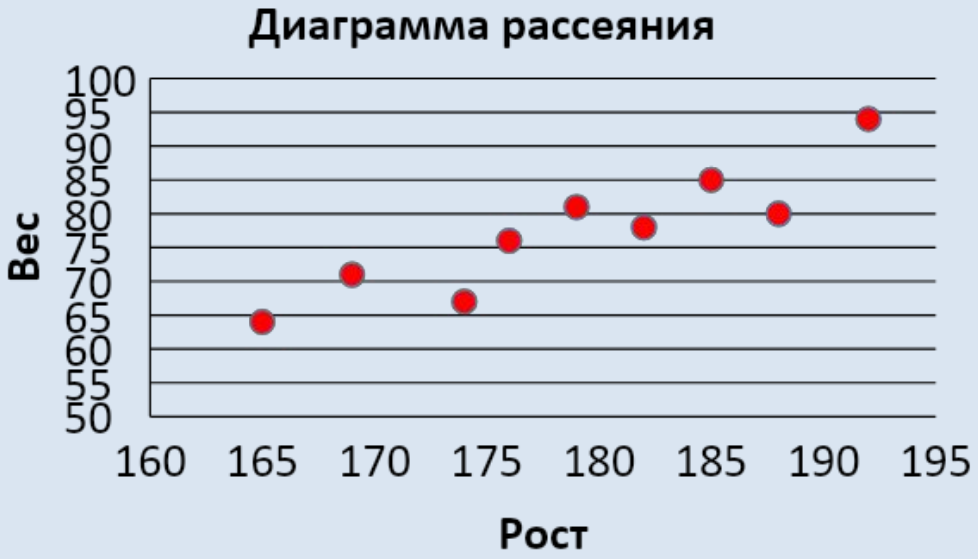
Относительно числа переменных

- – простая (парная) корреляция
- – множественная корреляция
- – частная корреляция

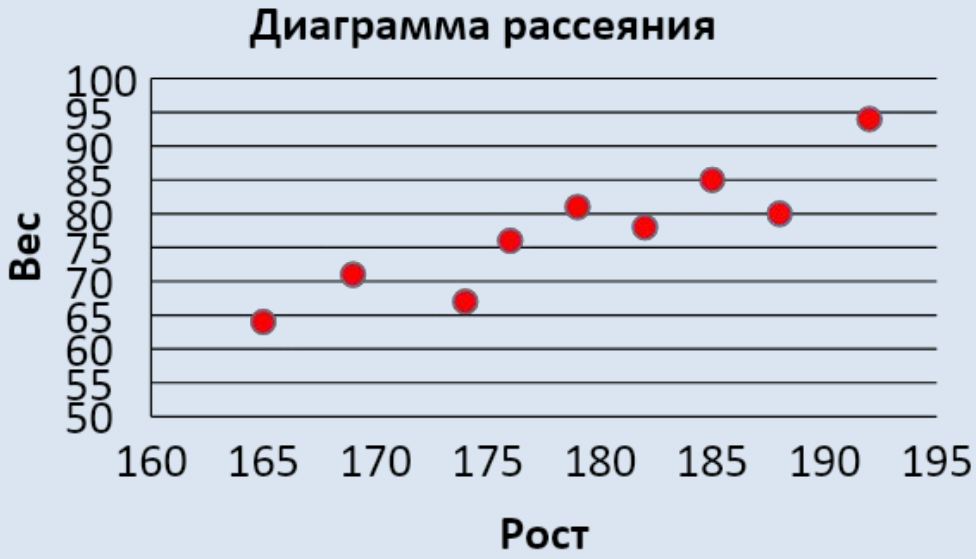
Корреляционная зависимость. Основные понятия и определения

Отрицательная корреляция

Положительная корреляция

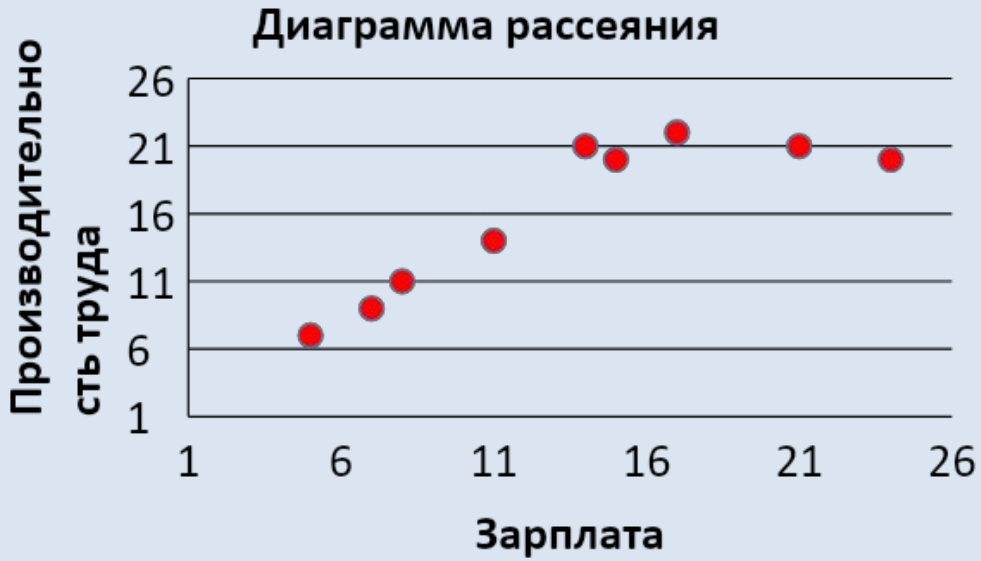


Корреляционная зависимость. Основные понятия и определения

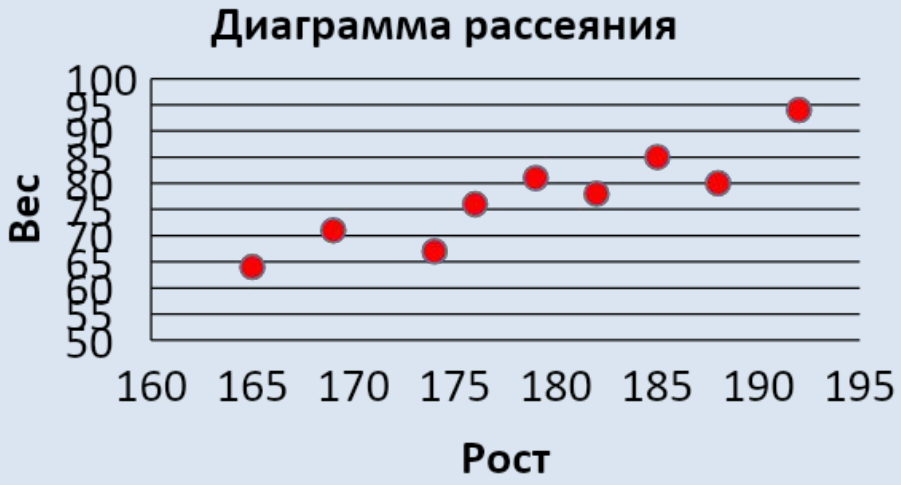


Линейная корреляция

Нелинейная корреляция

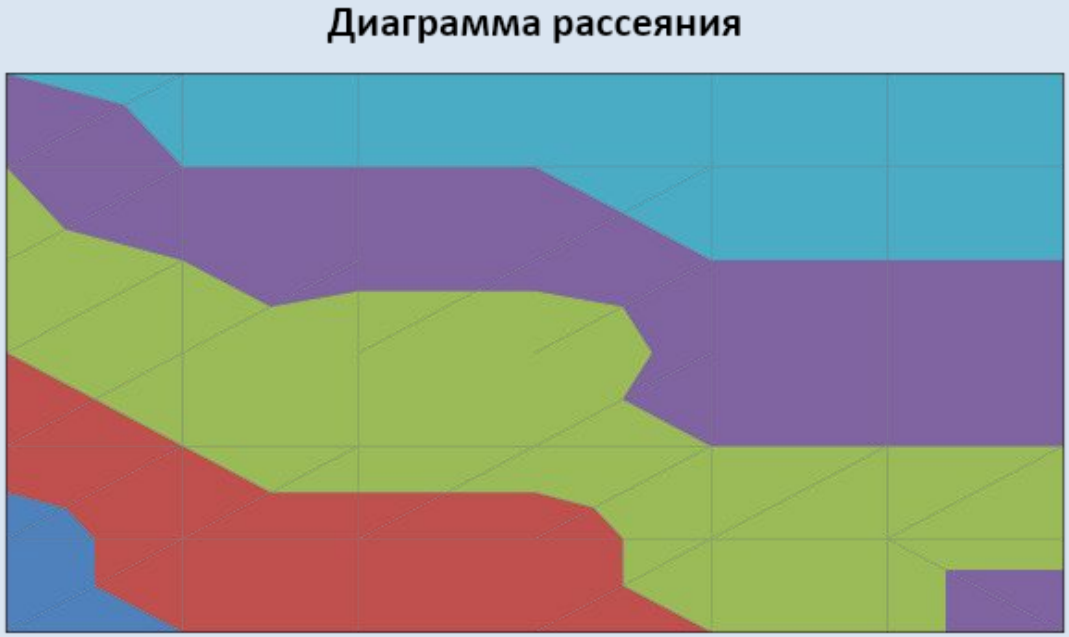


Корреляционная зависимость. Основные понятия и определения



Простая (парная) корреляция

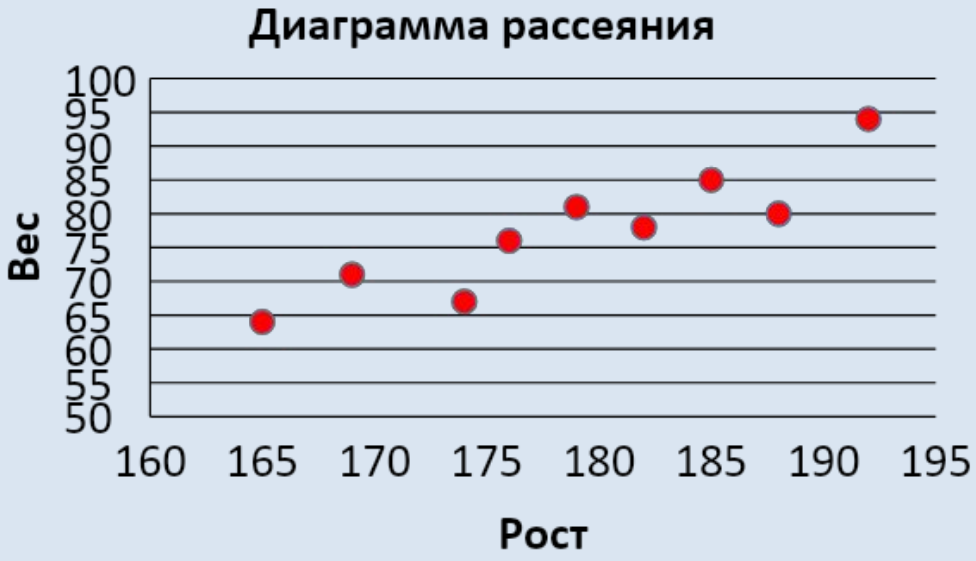
Множественная корреляция



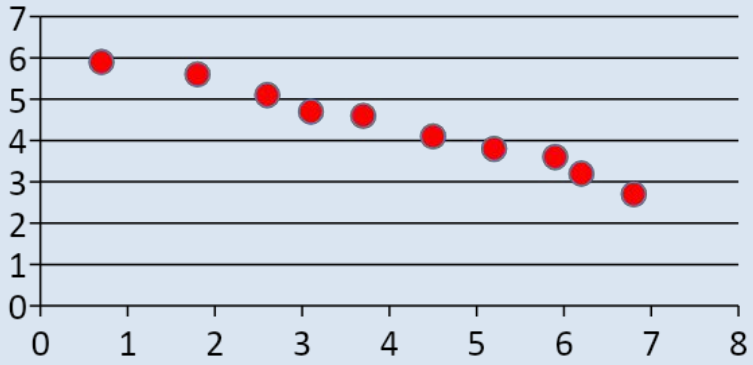
Корреляционная зависимость. Основные понятия и определения

Пример функциональной зависимости

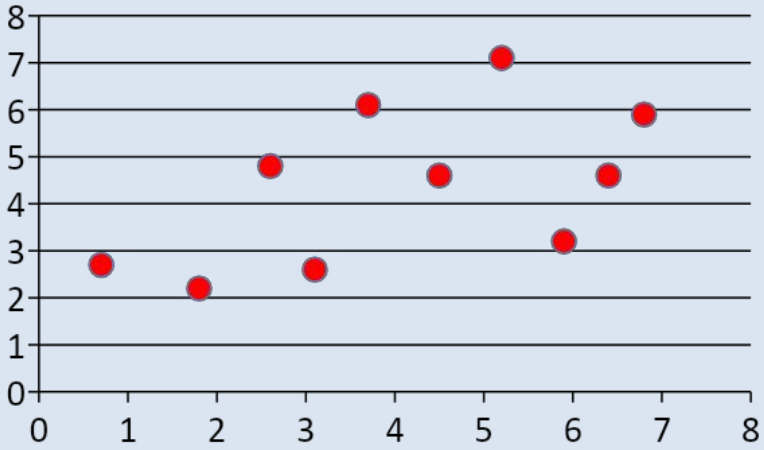
Пример корреляционной зависимости



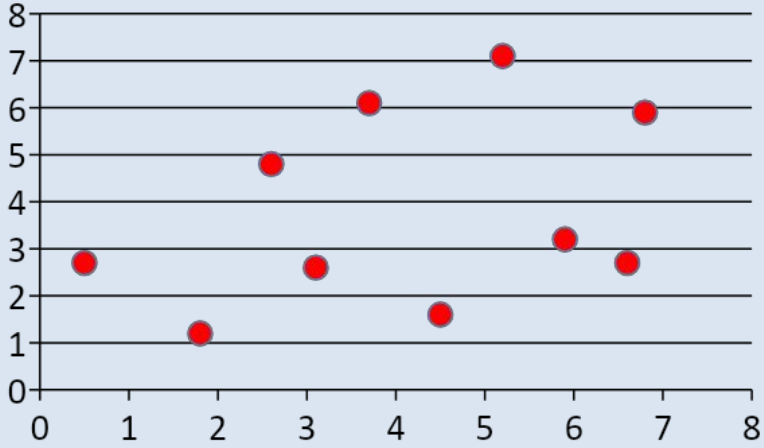
Корреляционная зависимость. Основные понятия и определения



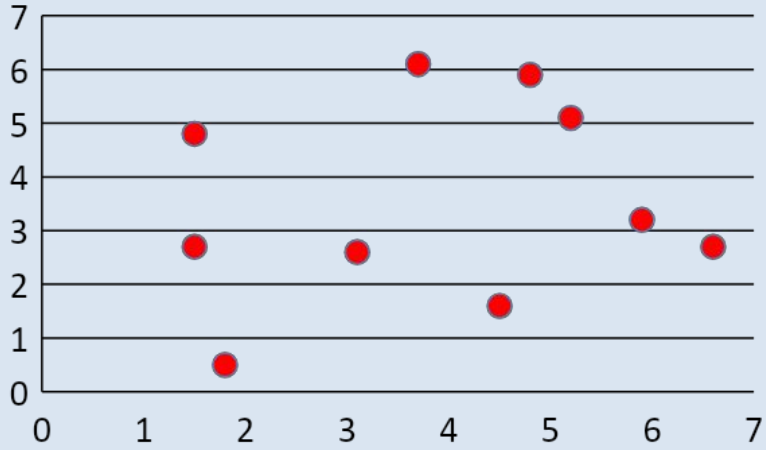
1



2



3



4

Корреляционная зависимость. Основные понятия и определения

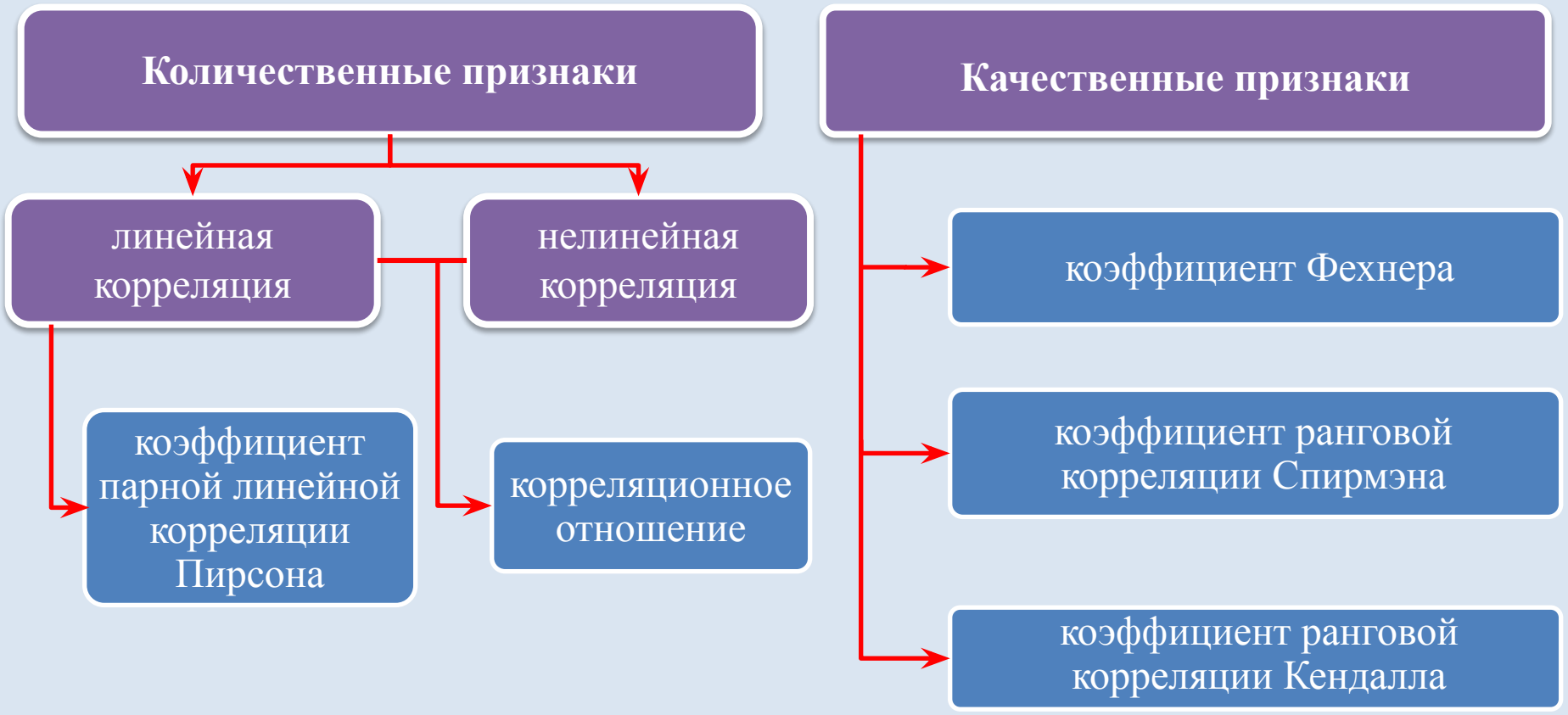
Число предприятий

Уровень безработицы

Производительность труда

Зарплата

Корреляционная зависимость. Основные понятия и определения



2

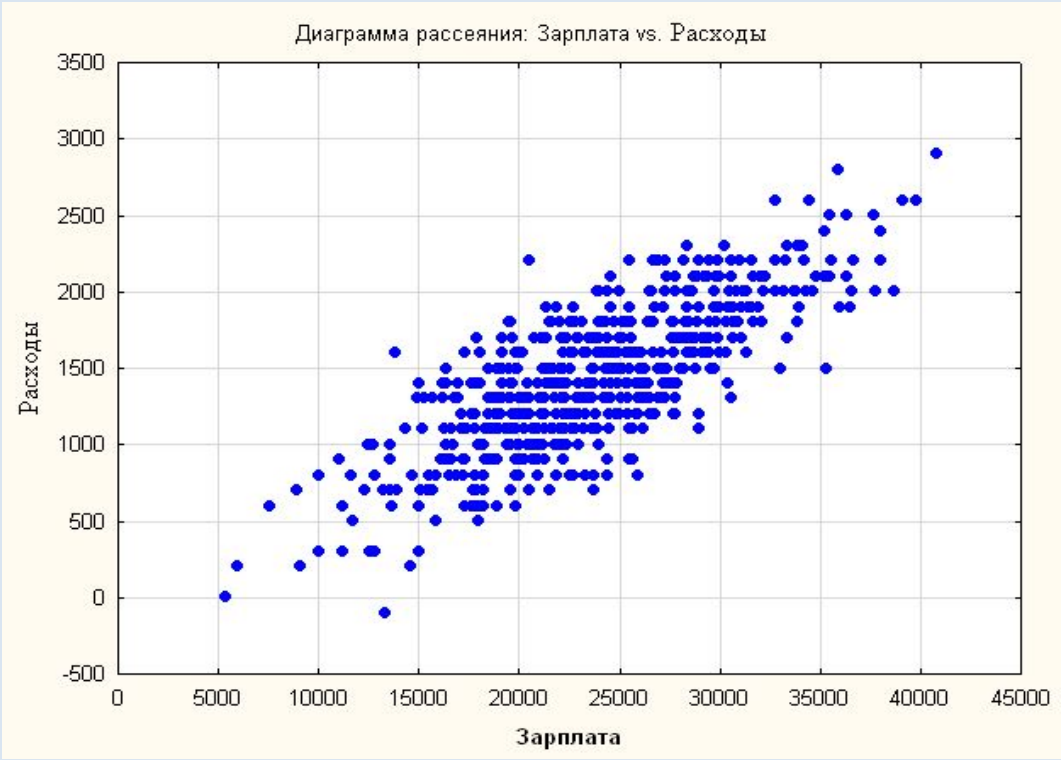
ПОКАЗАТЕЛИ ПАРНОЙ КОРРЕЛЯЦИОННОЙ ЗАВИСИМОСТИ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Показатели парной корреляционной зависимости количественных признаков

Пример 1. Оценить наличие и степень зависимости между зарплатой населения субъекта и расходами на лекарства.

Номер наблюдения	Расходы	Зарплата
1	1200	24900
2	1200	24900
3	1300	19500
4	1300	21100
5	1600	23400
6	1300	22900
7	1300	16900
8	2100	29800
9	1600	25700
10	2400	38000

Номер наблюдения	Расходы	Зарплата
590	2200	26700
591	900	17200
592	900	11000
593	1800	33900
594	2100	32300
595	2200	20500
596	1300	25900
597	1700	29100
598	1000	19900
599	1800	24100
600	1000	22100

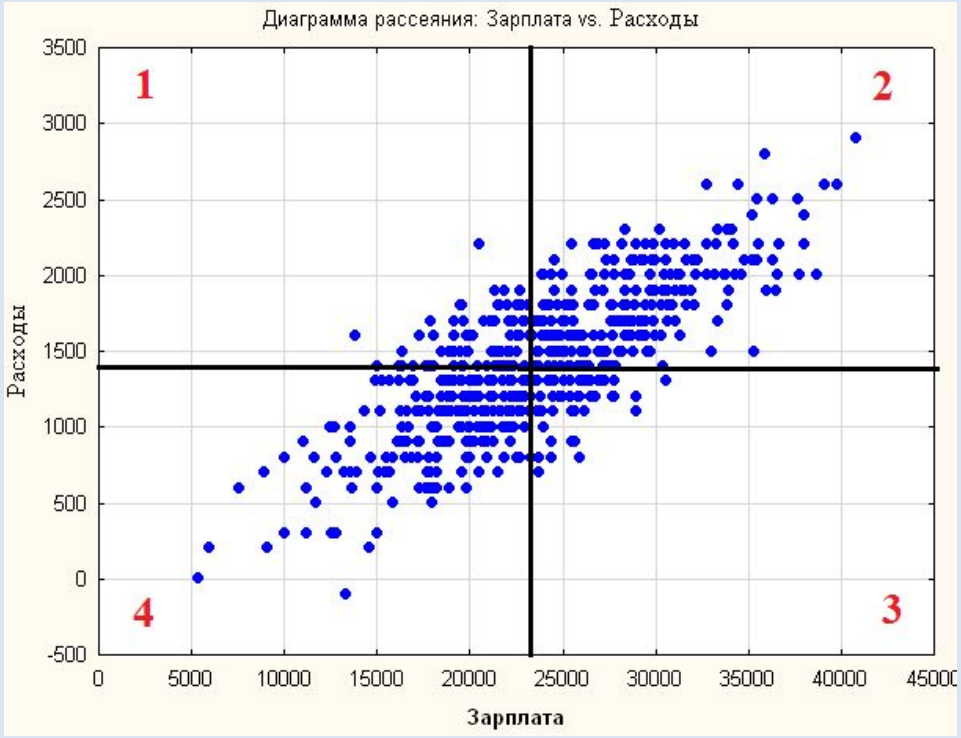


Показатели парной корреляционной зависимости количественных признаков

$$\rho = \frac{M[(x-M(x))\cdot(y-M(y))]}{\sigma_x \cdot \sigma_y} \quad (1)$$

$$r_в = \frac{\sum_{i=1}^{n}[(x_i-\bar{x}_в)\cdot(y_i-\bar{y}_в)]}{(n-1)\cdot s_x \cdot s_y} \quad (2)$$

$$r_в = \frac{\sum_{i=1}^{n}[(x_i-\bar{x}_в)\cdot(y_i-\bar{y}_в)]}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_в)^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3)$$



Показатели парной корреляционной зависимости количественных признаков

1. .

2. - линейная функциональная зависимость.

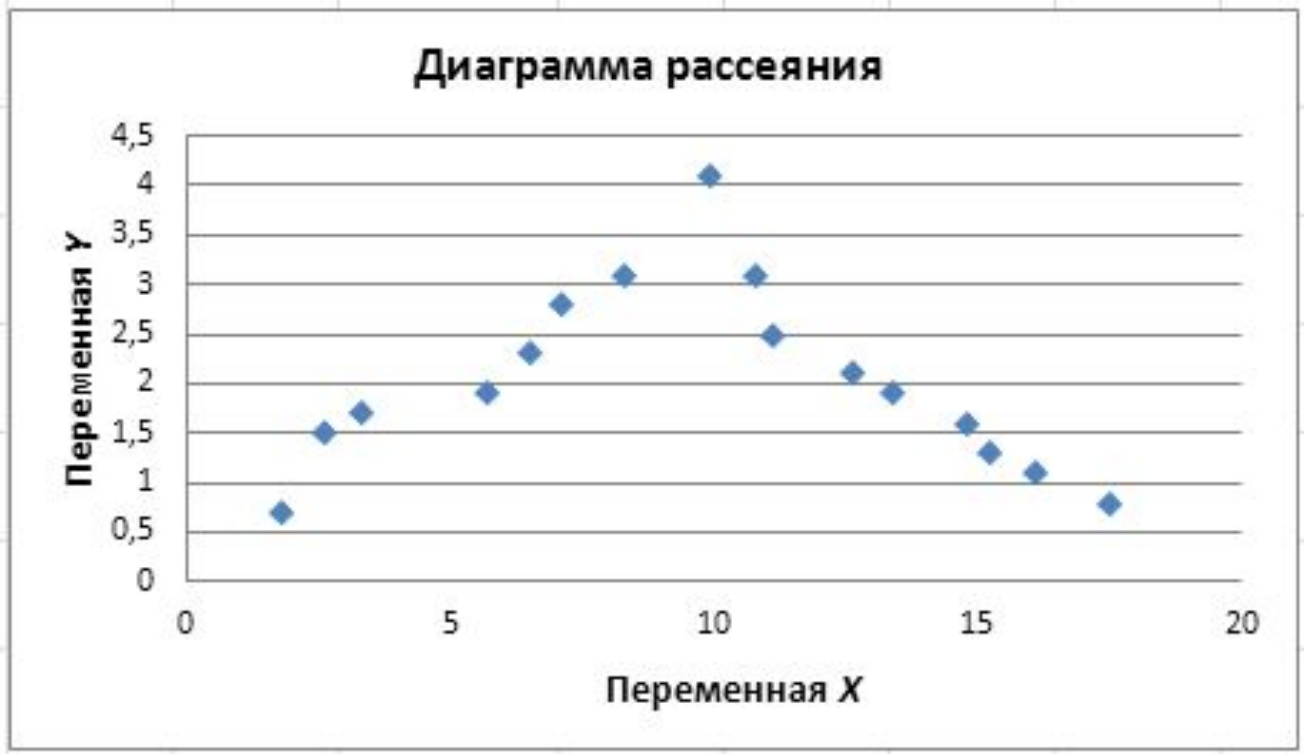
3. - линейная корреляционная связь отсутствует.

4. - корреляция между исследуемыми величинами.

5. .

6. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

Показатели парной корреляционной зависимости количественных признаков



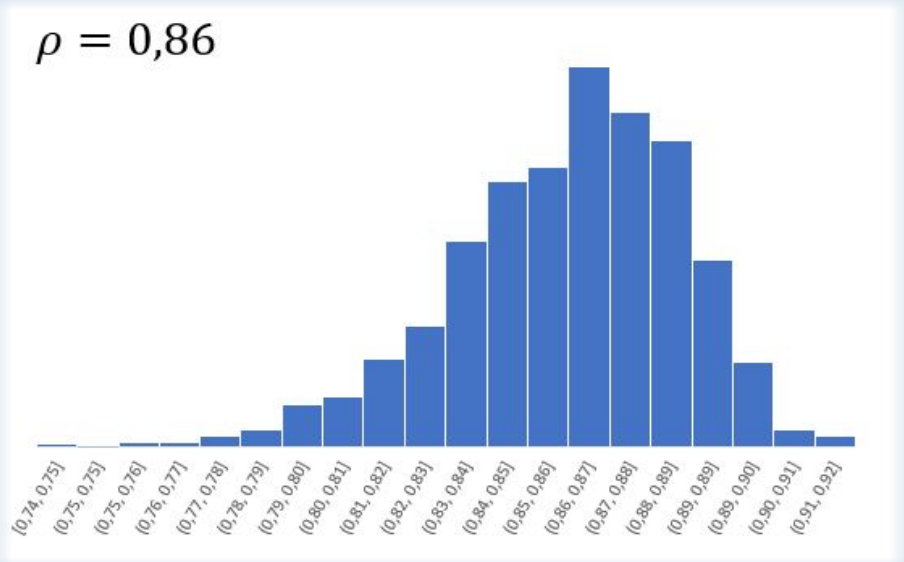
Коэффицент линейной корреляции	**-0,105**

Условия нормального распределения выборочного коэффициента корреляции

1. Значения случайной величины измерены в интервальной шкале или в шкале отношений.
2. В значениях случайных величин не должно быть значительных выбросов.
3. Объем выборочной совокупности должен быть достаточно большим.

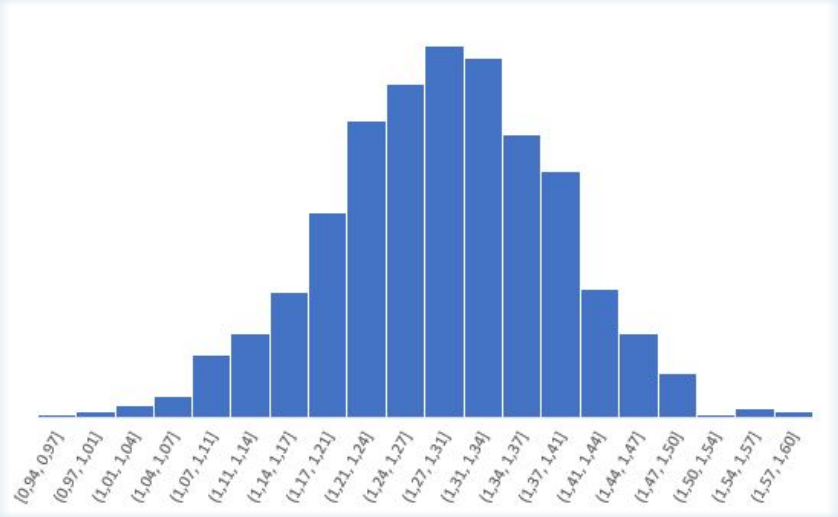
4. Совместное распределение исследуемых случайных величин является двумерным нормальным.

5. Абсолютное значение коэффициента корреляции не должно быть слишком близким к единице.



$$h = 0{,}5 \cdot ln\left[\frac{1+r}{1-r}\right]$$

$$M(h) = 0{,}5 \cdot ln\left[\frac{1+\rho}{1-\rho}\right] + \frac{\rho}{2\cdot(n-1)}, \quad D(h) = \frac{1}{\sqrt{n-3}}$$



Показатели парной корреляционной зависимости количественных признаков

$$P(h_{\text{н}} = h - \delta \le M(h) \le h_{\text{в}} = h + \delta) = \gamma.$$

$$\delta = z_{1-(\alpha/2)} \cdot D(h).$$

$$r_{\text{н}} = \frac{e^{2h_{\text{н}}}-1}{e^{2h_{\text{н}}}+1}; \quad r_{\text{в}} = \frac{e^{2h_{\text{в}}}-1}{e^{2h_{\text{в}}}+1}.$$

$$P(r_{\text{н}} \le \rho \le r_{\text{в}}) = \gamma.$$

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}_{\text{в}}) \cdot (y_i - \bar{y}_{\text{в}})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \approx 0{,}78$$

$$P\left(h - z_{1-(\alpha/2)} \cdot D(h) \le M(h) \le h + z_{1-(\alpha/2)} \cdot D(h)\right) = 0{,}95;$$

$$h = 0{,}5 \cdot ln\left[\frac{1+r}{1-r}\right] = 0{,}5 \cdot ln\left[\frac{1+0{,}78}{1-0{,}78}\right] \approx 1{,}05;$$

$$z_{1-(\alpha/2)} \approx 1{,}96;$$

$$D(h) = \frac{1}{\sqrt{n-3}} \approx 0{,}04;$$

$$P(0{,}97 \le M(h) \le 1{,}13) = 0{,}95$$

$$r_{\text{н}} = \frac{e^{2h_{\text{н}}}-1}{e^{2h_{\text{н}}}+1} \approx 0{,}75; \; r_{\text{в}} = \frac{e^{2h_{\text{в}}}-1}{e^{2h_{\text{в}}}+1} \approx 0{,}81;$$

$$P(0{,}75 \le \rho \le 0{,}81) = 0{,}95$$

Показатели парной корреляционной зависимости количественных признаков

$$H_0: \rho = 0; \quad H_1: \rho \neq 0$$

$$t_{\text{набл}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}, \quad v = n-2, \quad t_{\text{кр.п.}}(1-\alpha/2; v)$$

$$H_0: \rho = v_0; \quad H_1: \rho > v_0$$

$$t_{\text{набл}} = \frac{v-v_0}{D(v)} = \left[0{,}5 \cdot ln\left(\frac{1+r}{1-r}\right) - 0{,}5 \cdot ln\left(\frac{1+\rho}{1-\rho}\right)\right] \cdot \sqrt{n-3}$$

Показатели парной корреляционной зависимости количественных признаков

Пример 2. Оценить зависимость между потреблением молока и совокупным доходом семьи.

		Совокупный доход семьи, тыс. руб. (x)			
		интервалы (группы)			
		5-15	15-25	25-35	
Потребление молока, литров (y)		варианты			
интервалы	варианты	$x_1=10$	$x_2=20$	$x_3=30$	
10-20	$y_1=15$	4	28	6	38
20-30	$y_2=25$	6	–	6	12
		21	15	20	
		10	28	12	$n=50$

$$\eta_{yx} = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}} \quad (4)$$

$$\sigma_{\text{межгр}} = \sqrt{\frac{\sum_{j=1}^{m} n_{x_j} \cdot (\bar{y}_{x_j} - \bar{y})^2}{n}} \quad (5)$$

$$\sigma_{\text{общ}} = \sqrt{\frac{\sum_{i=1}^{l} n_{y_i} \cdot (y_i - \bar{y})^2}{n}} \quad (6)$$

$$\bar{y} = \frac{\sum_{i=1}^{l} n_{y_i} \cdot y_i}{n} = 17{,}4.$$

$$\sigma_{\text{общ}} = \sqrt{\frac{\sum_{i=1}^{l} n_{y_i} \cdot (y_i - \bar{y})^2}{n}} = 4{,}27.$$

$$\sigma_{\text{межгр}} = \sqrt{\frac{\sum_{j=1}^{m} n_{x_j} \cdot (\bar{y}_{x_j} - \bar{y})^2}{n}} = 2{,}73.$$

$$\eta_{yx} = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}} = \frac{2{,}73}{4{,}27} = 0{,}64.$$

Показатели корреляционной зависимости

1. .

2. - корреляционная зависимость

3. - функциональная зависимость

4. .

$$H_0: \eta_{yx} = 0; \quad H_1: \eta_{yx} > 0$$

$$F_{\text{набл}} = \frac{\eta_{yx}^2 \cdot (n-m)}{(1-\eta_{yx}^2) \cdot (m-1)}$$

$$v_1 = m-1, v_2 = n-m$$

$$F_{\text{кр.пр.}}(1-\alpha; v_1; v_2)$$

3

ПОКАЗАТЕЛИ ПАРНОЙ КОРРЕЛЯЦИОННОЙ ЗАВИСИМОСТИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

Коэффициент ранговой корреляции Спирмена

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_x - R_y)^2}{n \cdot (n^2 - 1)}.$$

$$r_s = \frac{\frac{n \cdot (n^2-1)}{6} - \sum_{i=1}^{n} (R_x - R_y)^2 - A - B}{\sqrt{\left(\frac{n \cdot (n^2-1)}{6} - 2A\right) \cdot \left(\frac{n \cdot (n^2-1)}{6} - 2B\right)}},$$

где c_x и c_y – количество связок в ранжированном списке переменной x и y соответственно, a_i и b_i – число одинаковых значений ранжированного ряда переменной x и y соответственно.

$$A = \frac{1}{12} \cdot \sum_{i=1}^{c_x} (a_i^3 - a_i), \quad B = \frac{1}{12} \cdot \sum_{i=1}^{c_y} (b_i^3 - b_i)$$

Показатели парной корреляционной зависимости качественных признаков

$$H_0: \rho_s = 0, \quad H_1: \rho_s \neq 0$$

$$t_{\text{набл}} = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

$$t_{\text{кр.п.}}(1-\alpha/2; v = n-2).$$

Пример 3. Необходимо исследовать зависимость между отношением населения субъекта к губернатору и Президенту.

Рейтинг	Субъект											
	1	2	3	4	5	6	7	8	9	10	11	12
Губернатор	4	8	12	5	1	6	7	10	2	9	11	3
Президент	6	5	10	7	3	4	9	8	1	11	12	2

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_x - R_y)^2}{n \cdot (n^2-1)} \approx 0{,}85$$

$$t_{\text{набл}} = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1-r_s^2}} \approx 5{,}07$$

$$t_{\text{кр.п.}}(1-\alpha/2; v = n-2) \approx 2{,}23$$

Коэффициент ранговой корреляции τ Кендалла

Пример 4. Необходимо исследовать зависимость между рейтингом предприятия бытовых услуг и их доходами.

Показатели	Предприятие							
	1	2	3	4	5	6	7	8
Доход	1	3	2	7	5	6	8	4
Рейтинг	3	2	1	4	7	8	6	5

$$\tau = \frac{2\cdot(P-Q)}{n\cdot(n-1)},$$

P – число совпадений, Q – число инверсий.

$$\tau = \frac{2\cdot(P-Q)}{n\cdot(n-1)} = \frac{2\cdot(21-7)}{8\cdot(8-1)} = 0{,}5$$

Показатели	Предприятие								Сумма
	1	3	2	8	5	6	7	8	
Доход	1	2	3	4	5	6	7	8	
Рейтинг	3	1	2	5	7	8	4	6	
Совпадения, P	5	6	5	3	1	0	1	0	
Инверсии, Q	2	0	0	1	2	2	0	0	

Показатели парной корреляционной зависимости качественных признаков

$$\tau = \frac{P-Q}{\sqrt{\frac{1}{2}\cdot n\cdot(n-1)-A}\cdot\sqrt{\frac{1}{2}\cdot n\cdot(n-1)-B}}, \qquad A = \frac{1}{2}\cdot\sum_{i=1}^{c_x} a_i\cdot(a_i-1), \quad B = \frac{1}{2}\cdot\sum_{i=1}^{c_y} b_i\cdot(b_i-1),$$

где c_x и c_y – количество связок в ранжированном списке переменной x и y соответственно, a_i и b_i – число одинаковых значений ранжированного ряда переменной x и y соответственно.

$$H_0\colon \tau_\Gamma = 0,\ H_1\colon \tau_\Gamma \neq 0$$

$$z_{\text{набл.}} = \frac{S^*}{\sqrt{\frac{n\cdot(n-1)\cdot(2\cdot n+5)}{18}}}$$

$S^* = S + 1$, если $S < 0$ и $S^* = S - 1$, если $S > 0$

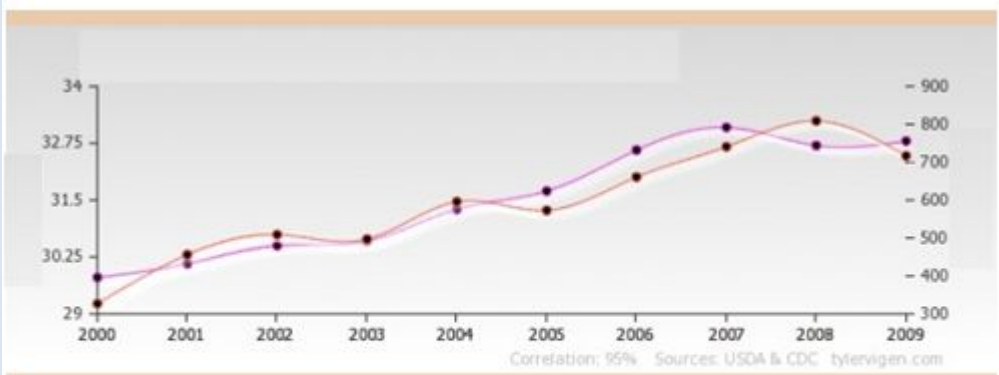
$$z_{\text{кр.п.}}(1-(\alpha/2)).$$

Пример косвенной корреляции

Количество пожарных

Степень разрушений

Примеры ложных корреляций



Потребление сыра
и количество людей, которые умерли, запутавшись в своих простынях.