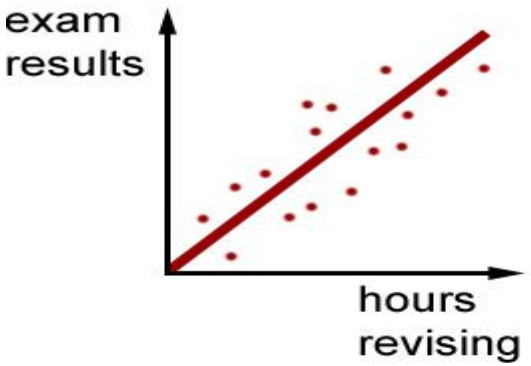
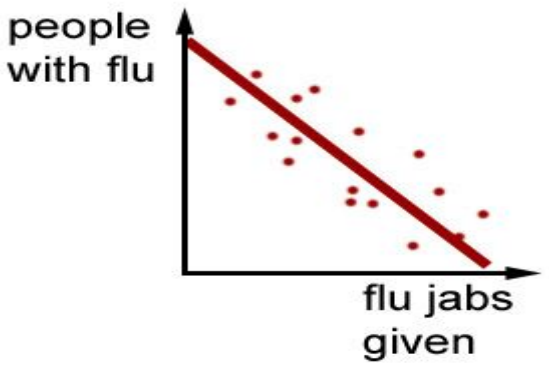


POSITIVE CORRELATION

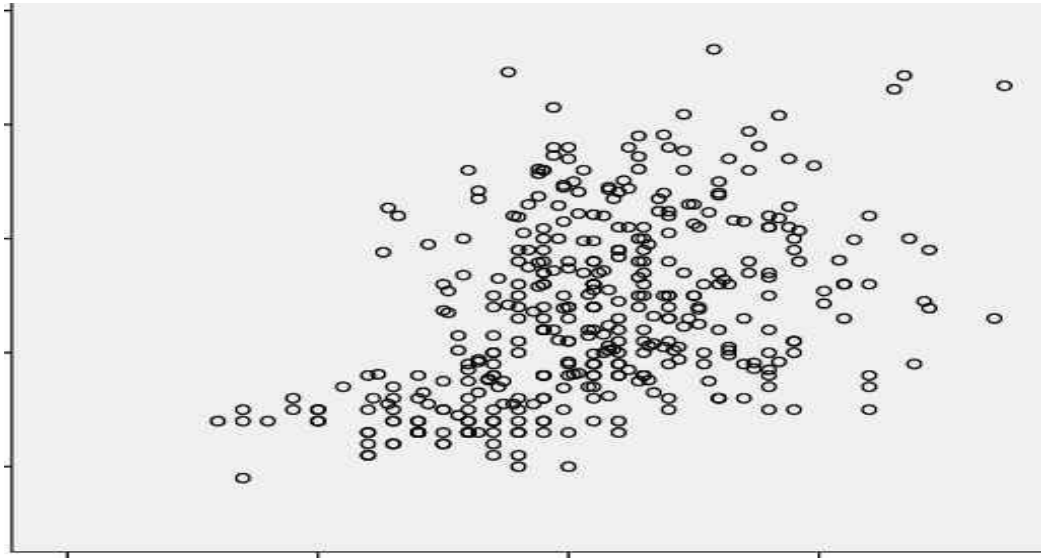
- people who do more revision get higher exam results.
- revising increases success.



NEGATIVE CORRELATION

- when more jabs are given the number of people with flu falls.
- flu jabs prevent flu.

Основы корреляционного анализа



Многомерный корреляционный анализ

При исследование реальных экономических явлений приходится сталкиваться с анализом многомерной генеральной совокупности в которой каждый объект характеризуется набором признаков

$$X_1, X_2, ..., X_n$$

- Исследователь располагает случайной выборкой

$$x^{(1)}, x^{(2)}, ..., x^{(k)}$$

- Необходимо сделать вывод о генеральной совокупности (многомерной случайной величине)

$$\xi = (\xi^{(1)}, \xi^{(2)}, ..., \xi^{(\kappa)})^T$$

Многомерный корреляционный анализ

Закон распределения не известен

Обычно ограничиваются оцениваем по выборке

- вектора математических ожиданий

$$a = (a_1, a_2, ..., a_к)$$

- ковариационной матрицы Σ

По существу вся специфика многомерной случайности сосредоточена в ковариационной матрице Σ.

Многомерный корреляционный анализ

Ковариационная матрица Σ позволяет строить и анализировать

- характеристики вариации
- характеристики статистической взаимосвязи

(коррелированности) компонент многомерного признака.

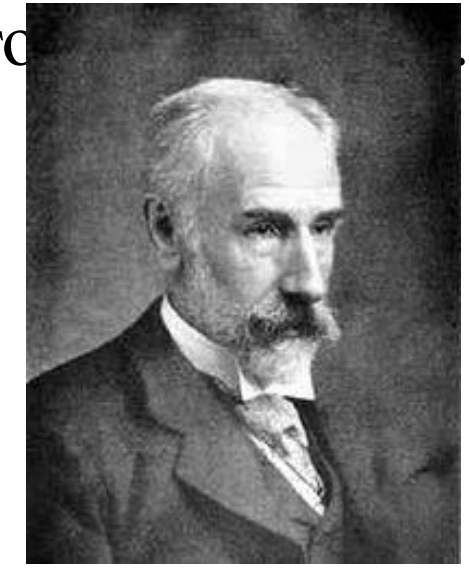
Ковариация

Для устранения недостатка ковариации был введён **линейный коэффициент корреляции**

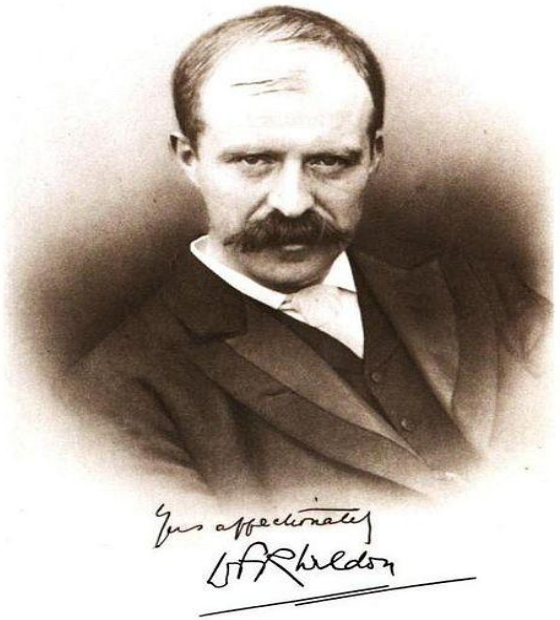
(или **коэффициент корреляции Пирсона**),

который разработали Карл Пирсон который разработали Карл Пирсон, Фрэнсис Эджуорт который разработали Карл Пирсон, Фрэнсис Эджуорт и Рафаэль Уэл х го .

1857-1936



1845-1926



1860-1906

Основатели корреляционного анализа

Карл (Чарлз) Пирсон
(Karl (Charles) Pearson)
(1857- 1936)
английский математик, статистик, биолог и философ;
основатель математической статистики

Correlation* – *взаимосвязь, взаимозависимость

Pearson product moment correlation
correlation coefficient r
(парный коэффициент корреляции Пирсона, парный коэффициент корреляции)

Ковариация

Коэффициент корреляции рассчитывается по формуле:

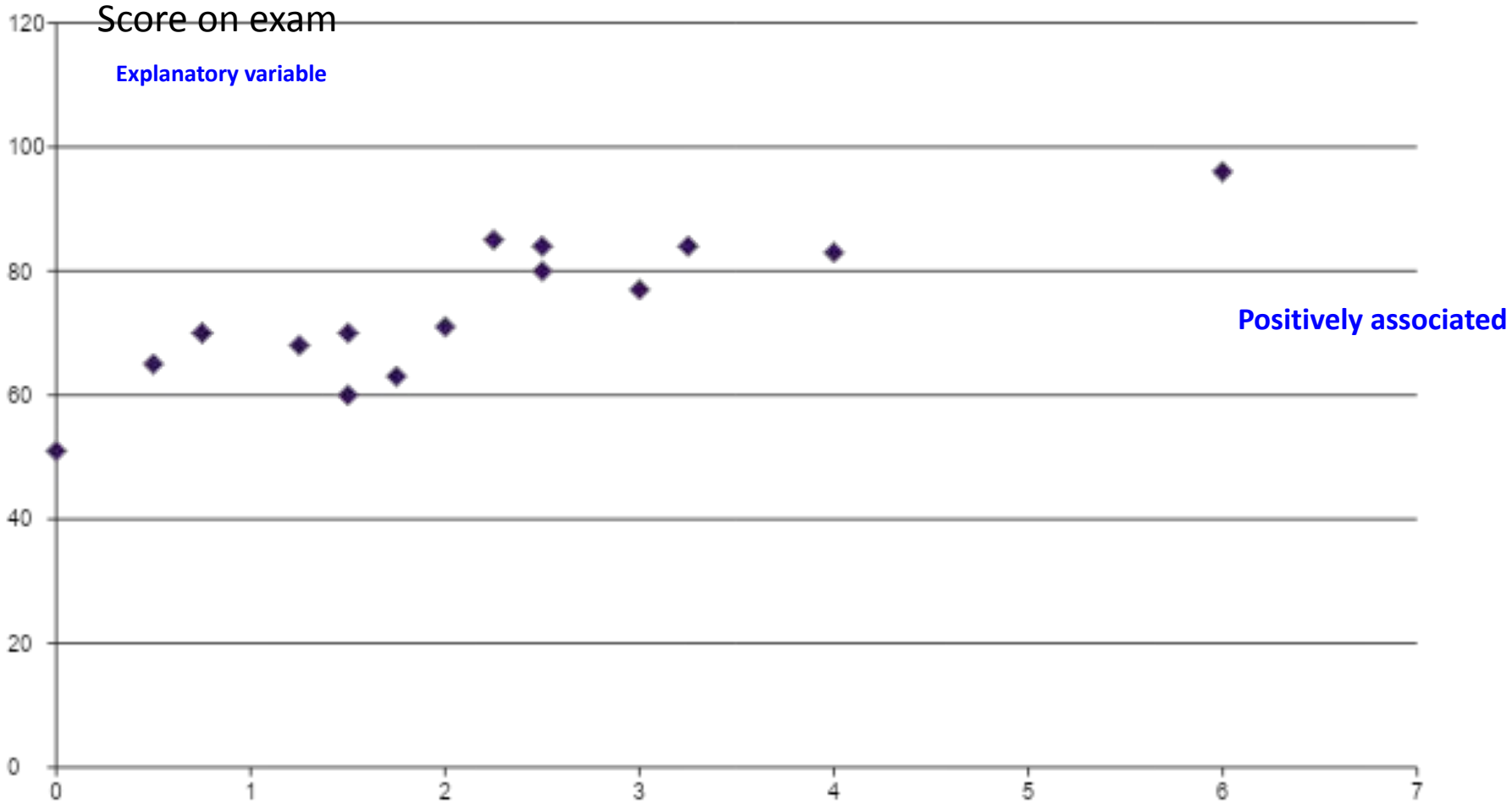
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Исследование зависимости между 2 переменными

Пример: Преподаватель попросил студентов (n=15) записать, сколько часов они потратили на подготовку к промежуточному экзамену. Результаты приведены в табл.

Student	Hours studied	Score on exam
A	0,5	65
B	2,5	80
C	3,0	77
D	1,5	60
E	1,25	68
F	0,75	70
G	4,0	83
H	2,25	85
I	1,5	70
J	6,0	96
K	3,25	84
L	2,5	84
M	0,0	51
N	1,75	63
O	2,0	71

Диаграмма рассеяния (scatterplot)



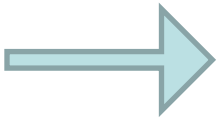
«Существует ли зависимость между доходом семьи и ее расходами на питание?»

- «Связан ли уровень безработицы в стране с ВВП?»
- «Оказывают ли влияние научные исследования на инновационную активность?»
- ……..

Корреляционный анализ – один из методов статистического анализа взаимозависимости нескольких признаков на основе выборочных данных.

Характеристики статистической связи, рассматриваемые в корреляционном анализе используются в качестве «входной» информации при решении следующих задач эконометрики и МСМ:

- Определение вида зависимости между переменными (РА);
- Снижение размерности анализируемого признакового пространства (ФА, МГК);
- Классификации объектов и признаков (КА).



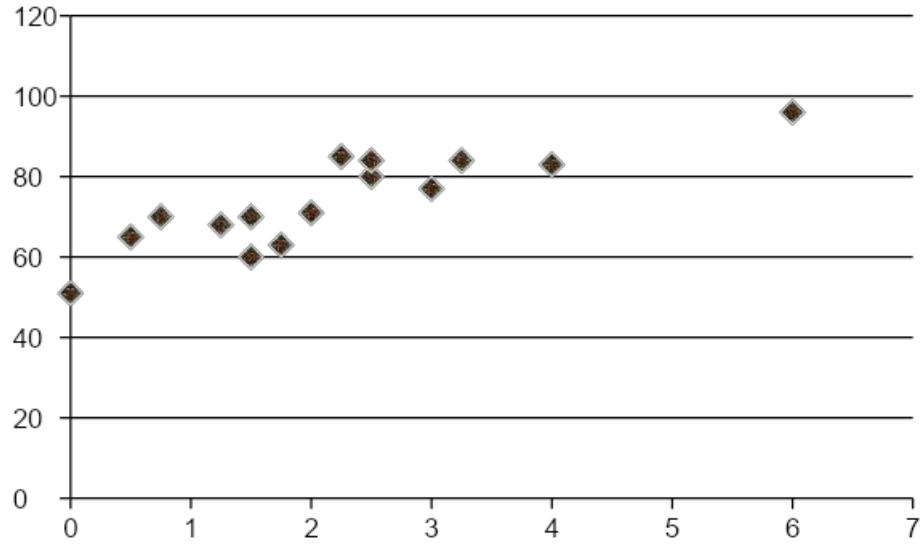
с корреляционного анализа начинаются практически все многомерные статистические исследования.

Корреляционный анализ

Основные понятия

Коэффициент корреляции –

- ✔измеритель силы линейной взаимосвязи между двумя переменными,
- ✔направления линейной взаимосвязи (прямая или обратная)



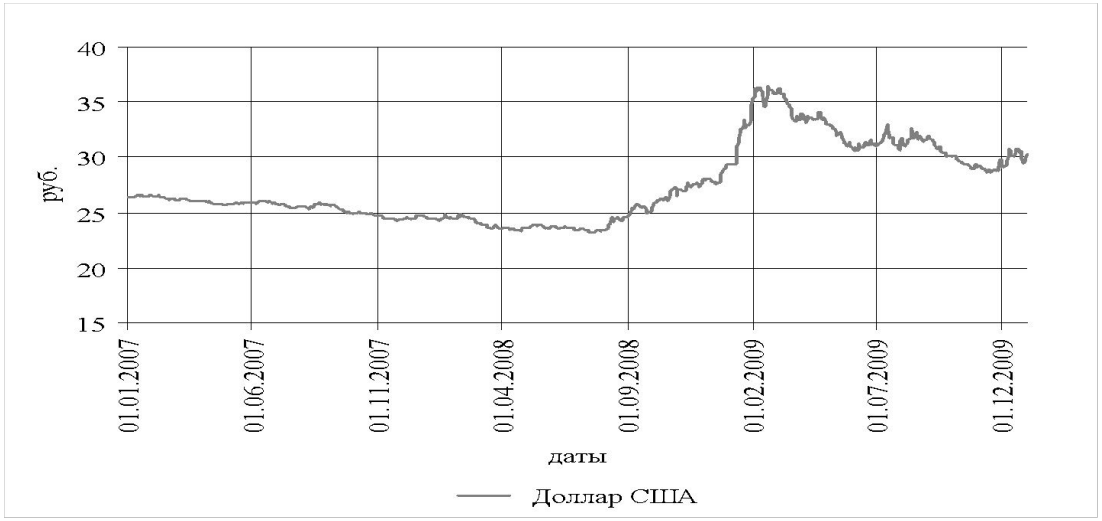
Корреляционный анализ

Основные понятия

Случайные величины X и Y могут быть либо зависимыми, либо независимыми

Типы зависимостей случайных величин

Функциональной зависимостью переменной ***Y*** от переменной ***X*** ***называют зависимость вида*** $Y = f(X)$ ***, где каждому допустимому значению X ставится в соответствие по определенному правилу единственно возможное значение переменной Y.***



На формирование значений СВ X и Y оказывают влияние различные факторы. Под воздействием этих факторов и формируются конкретные значения X и Y.

Типы зависимостей случайных величин

Пример:

1. Допустим, что на **X** и **Y** влияют одни и те же факторы, например, **Z1, Z2, Z3**, тогда X и Y находятся в полном соответствии с друг другом и связаны ……

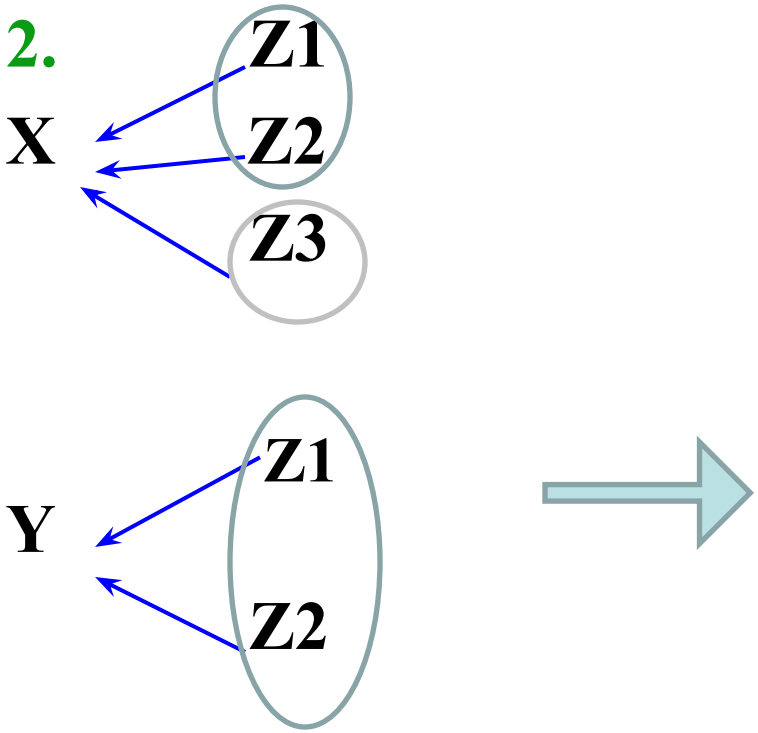
Типы зависимостей случайных величин

Пример:

1. Допустим, что на **X** и **Y** влияют одни и те же факторы, например, **Z1, Z2, Z3**, тогда X и Y находятся в полном соответствии с друг другом и связаны ***функционально***.

Типы зависимостей случайных величин

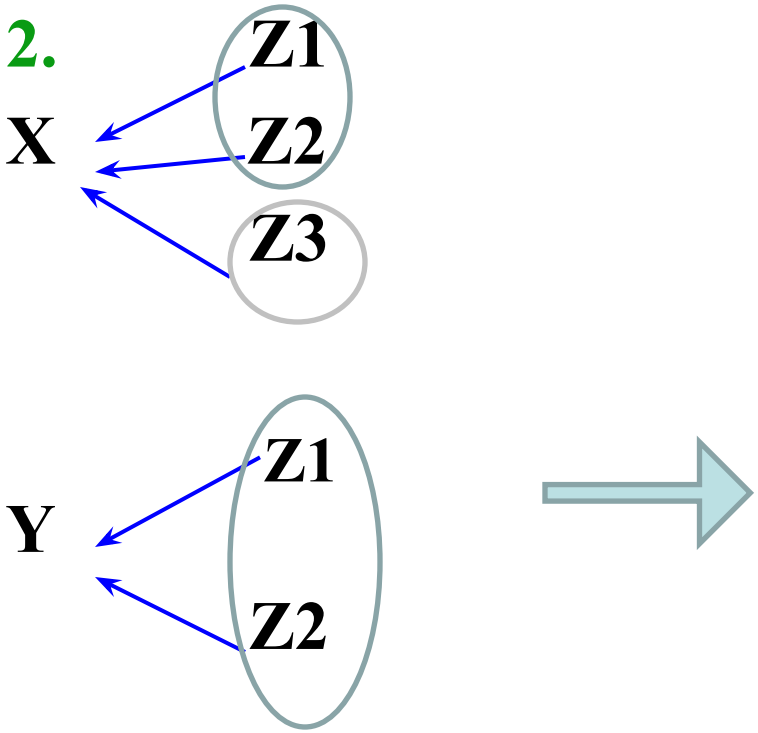
2.



величины X и Y являются *случайными*, но так как имеются общие факторы Z1 и Z2, оказывающие влияние и на X и на Y,
значения X и Y обязательно будут *взаимосвязаны*

Типы зависимостей случайных величин

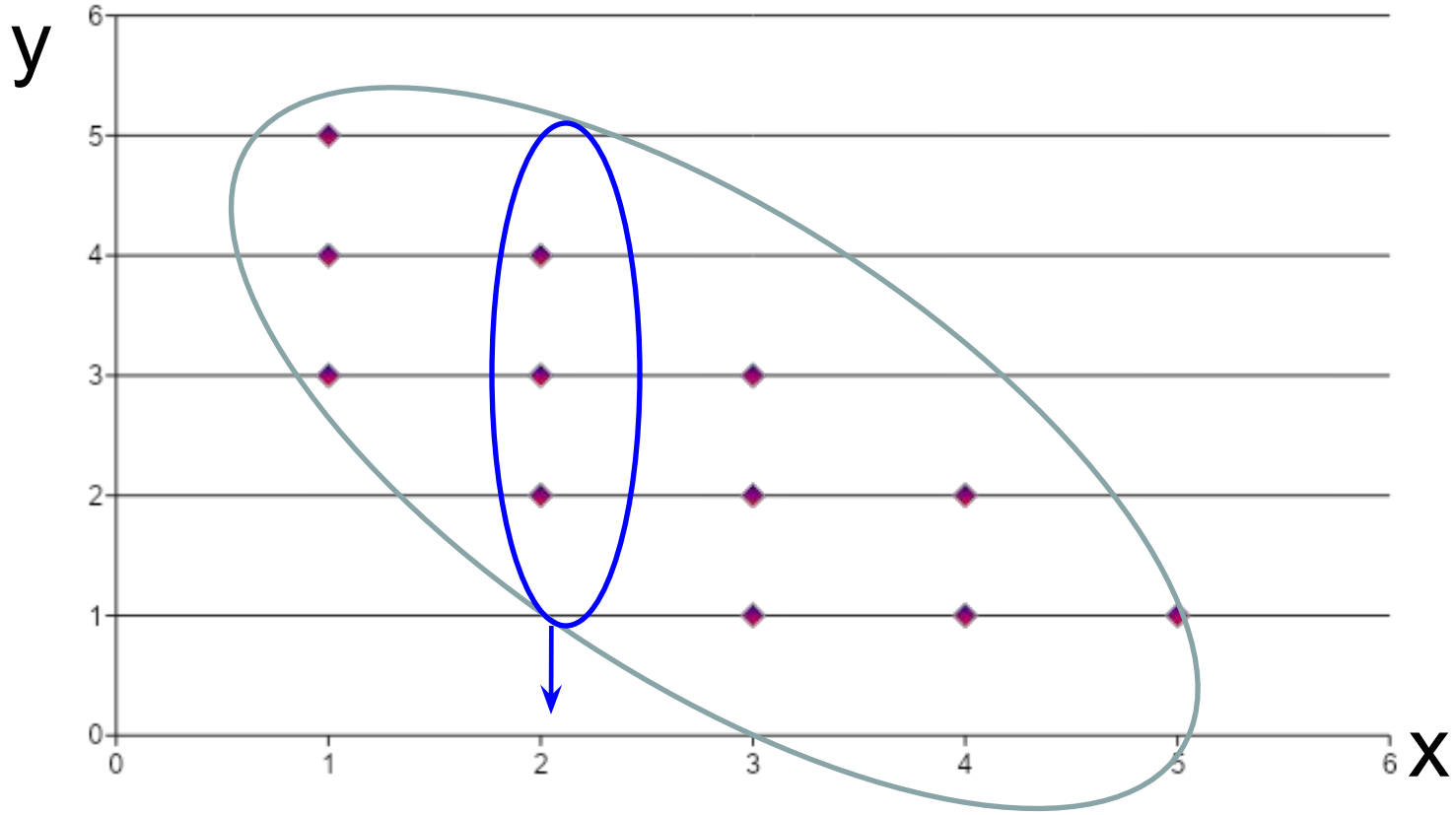
2.



величины X и Y являются *случайными*, но так как имеются общие факторы Z1 и Z2, оказывающие влияние и на X и на Y,
значения X и Y обязательно будут *взаимосвязаны*

- Связь уже не функциональная
- Носит вероятностный, случайный характер и меняется от испытания к испытанию.
- Такая зависимость называется ***стохастической.*** Каждому ***значению*** X может соответствовать не одно значение Y, а целое множество значений.

Типы зависимостей случайных величин

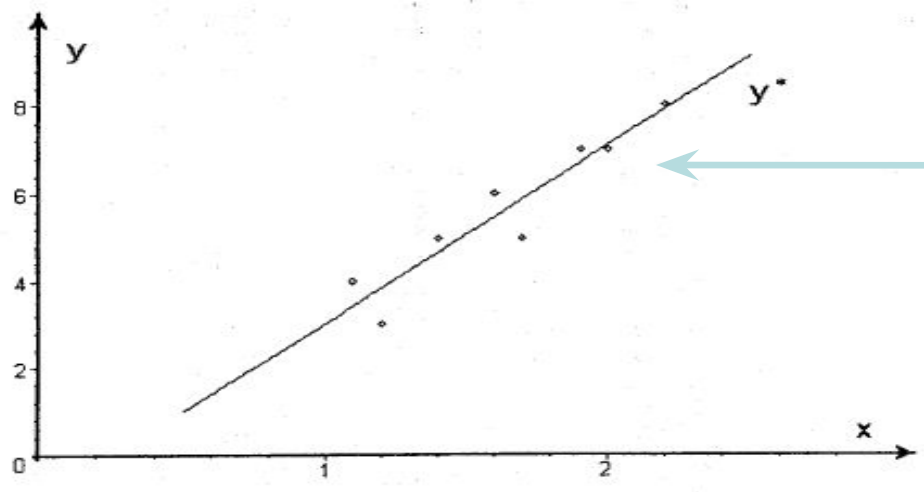


Типы зависимостей случайных величин

Среди множества значений Y можно найти среднее значение $M(Y/X=x)$, которое для каждого значения **x** свое. Множество этих значений на графике образуют линию

$$\hat{y} = M(Y/X=x) = M(Y/X)$$

вид которой может быть самым разнообразным (прямая, парабола, экспонента и т.д.) и определяется СВ X и Y.



Линия регрессии Y на X

Типы зависимостей случайных величин

Если изменение одной из СВ приводит к изменению среднего значения другой СВ, то такую зависимость называют ***корреляционной.***

Примеры:

- ***Урожайность зерновых культур (влажность, освещенность..);***
- ***зависимость массы тела от роста;***
- ***Зависимость заболеваемости от воздействия внешних факторов;***
- ***уровень жизни и процент смертности и т.д.***

Исследование зависимости между 2 переменными (bivariate date)

Вопросы исследования:

- Существует ли линейная взаимосвязь между переменными?
- Как по изменению одной переменной можно предсказать изменение другой переменной?

Линейный коэффициент корреляции

Двумерная корреляционная модель

Исходной для анализа является матрица

$$X = \begin{pmatrix} x_{11} x_{12} \\ \cdots\cdots \\ x_{i1} x_{i2} \\ \cdots\cdots \\ x_{n1} x_{n2} \end{pmatrix}$$

- матрица **«объект–свойство»**

размерности (n x 2),

i -я строка характеризует i-е наблюдение (объект) по двум показателям (j=1, 2).

Корреляционный анализ

Двумерная корреляционная модель

Двумерная корреляционная модель определяется 5 параметрами:

$$(X,Y) \in N\ (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho_{xy})$$

ρ – генеральный **парный коэффициент корреляции**, характеризующий тесноту связи между переменными X и Y.

Коэффициенты корреляции

Парный коэффициент корреляции ρ_{12}

характеризует тесноту **линейной** взаимосвязи между двумя переменными (x_1 и x_2) *на фоне действия всех остальных переменных, входящих в модель*.

ρ_{12} ***изменяется в пределах от -1 до +1.***

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y}\right)$$

В нашем примере r=0,81. Это индикатор сильной положительной взаимосвязи между временем, потраченным на изучение материала и экзаменационной оценкой.

Корреляционный анализ

Точечные оценки параметров двумерной корреляционной модели

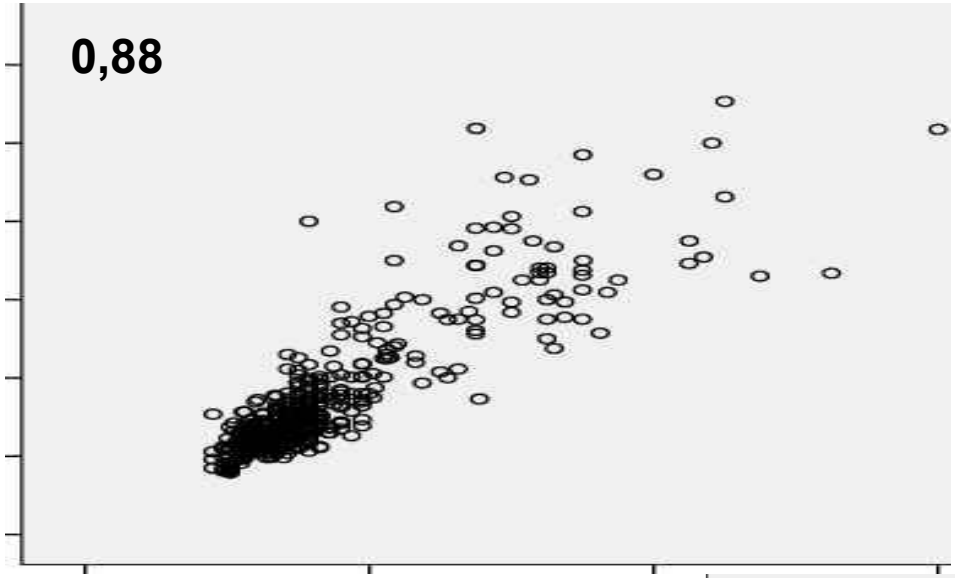
Генераль-ные характер.	Их оценки (выборочные характеристики)	
	n мало (данные не сгруппированы)	n велико (данные сгруппированы)
μ_x	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot m_{ix}$
μ_y	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j \cdot m_{jy}$
M(xy)	$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$	$\overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot y_j \cdot m_{ij}$
σ_x^2, σ_y^2	$S_x^2 = \overline{x^2} - (\bar{x})^2$	$S_y^2 = \overline{y^2} - (\bar{y})^2$
ρ Выборочный коэффициент корреляции	$r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x \cdot S_y}$	$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y}\right)$

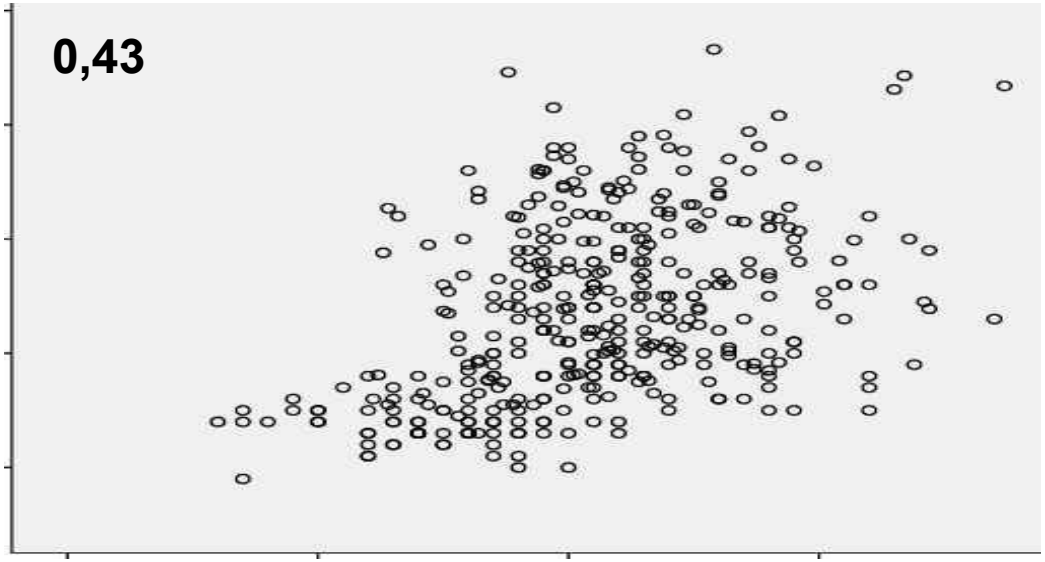
Диаграмма рассеяния

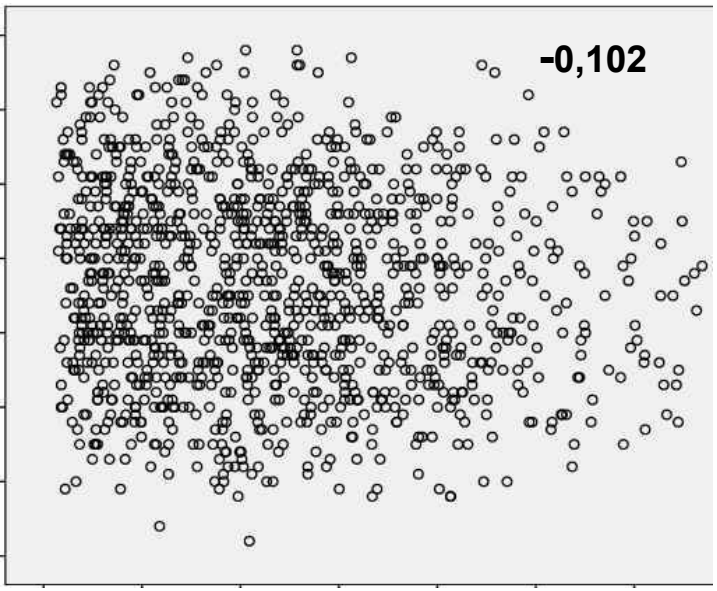
На практике изучение зависимости между двумя СВ необходимо начинать с построения поля корреляции (диаграммы рассеяния), с помощью которого можно

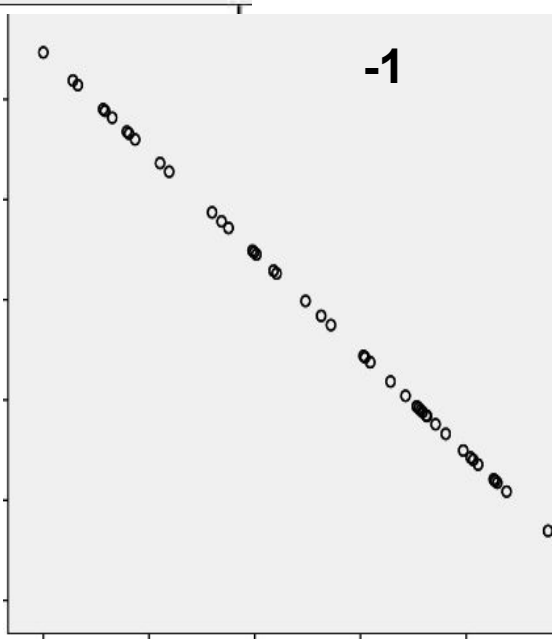
- установить наличие корреляционной зависимости,
- силу взаимосвязи,
- выявить аномальные наблюдения.

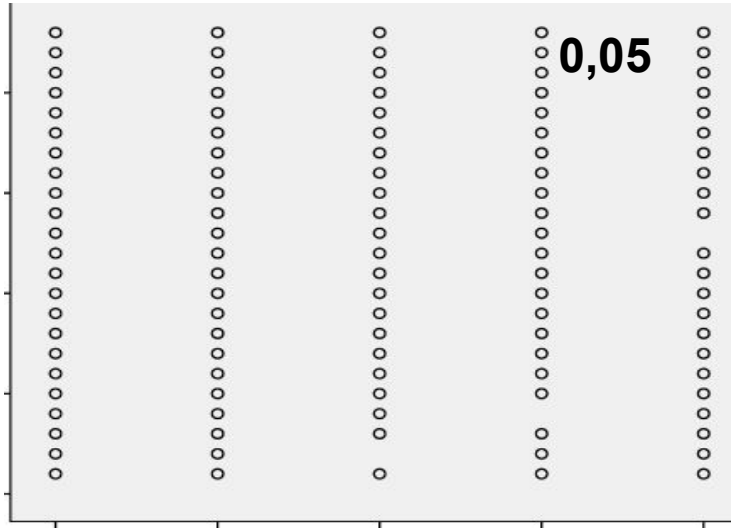
Диаграммы рассеивания

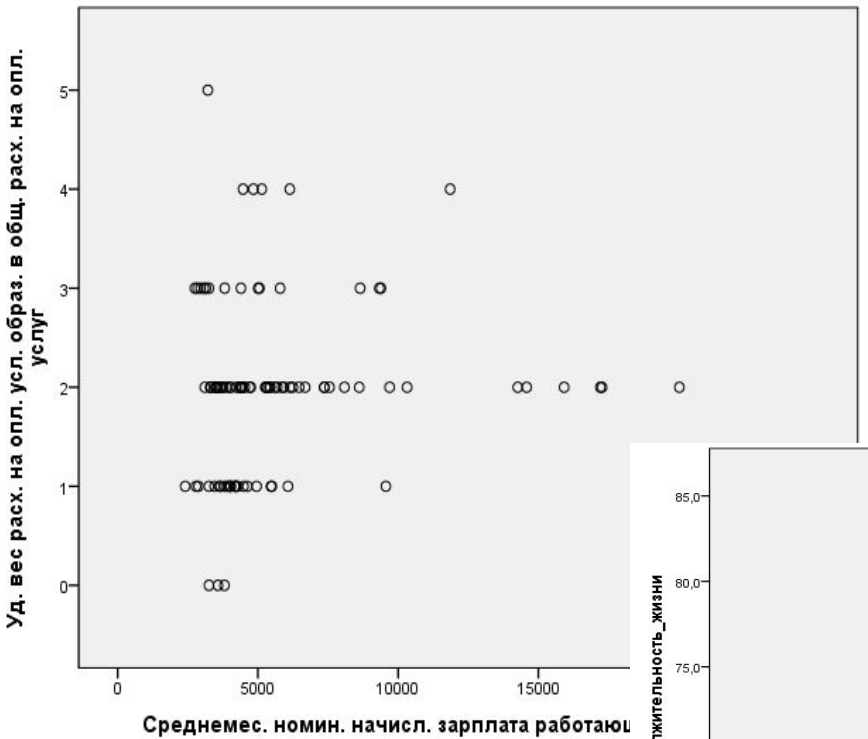


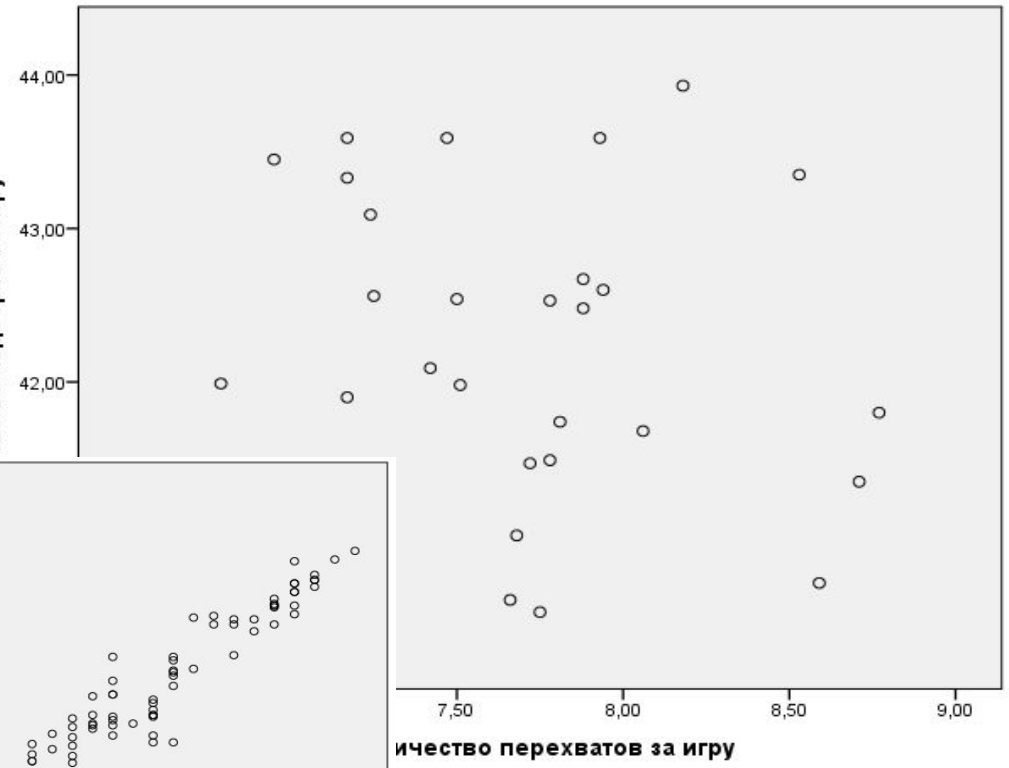


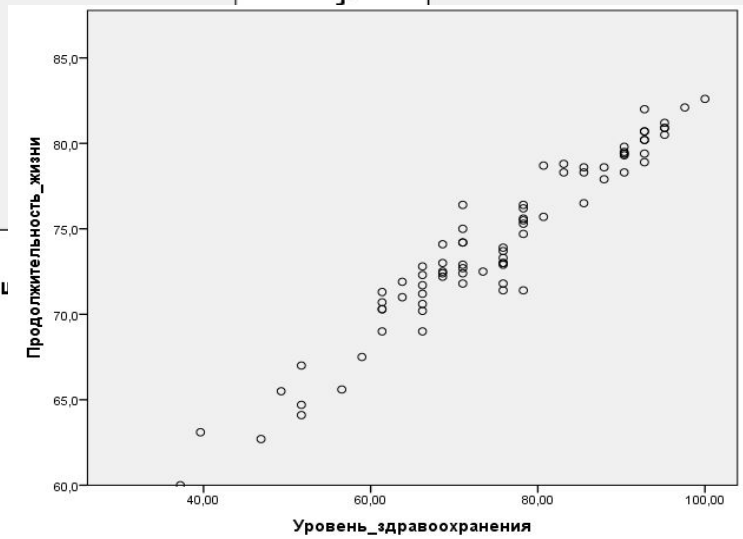






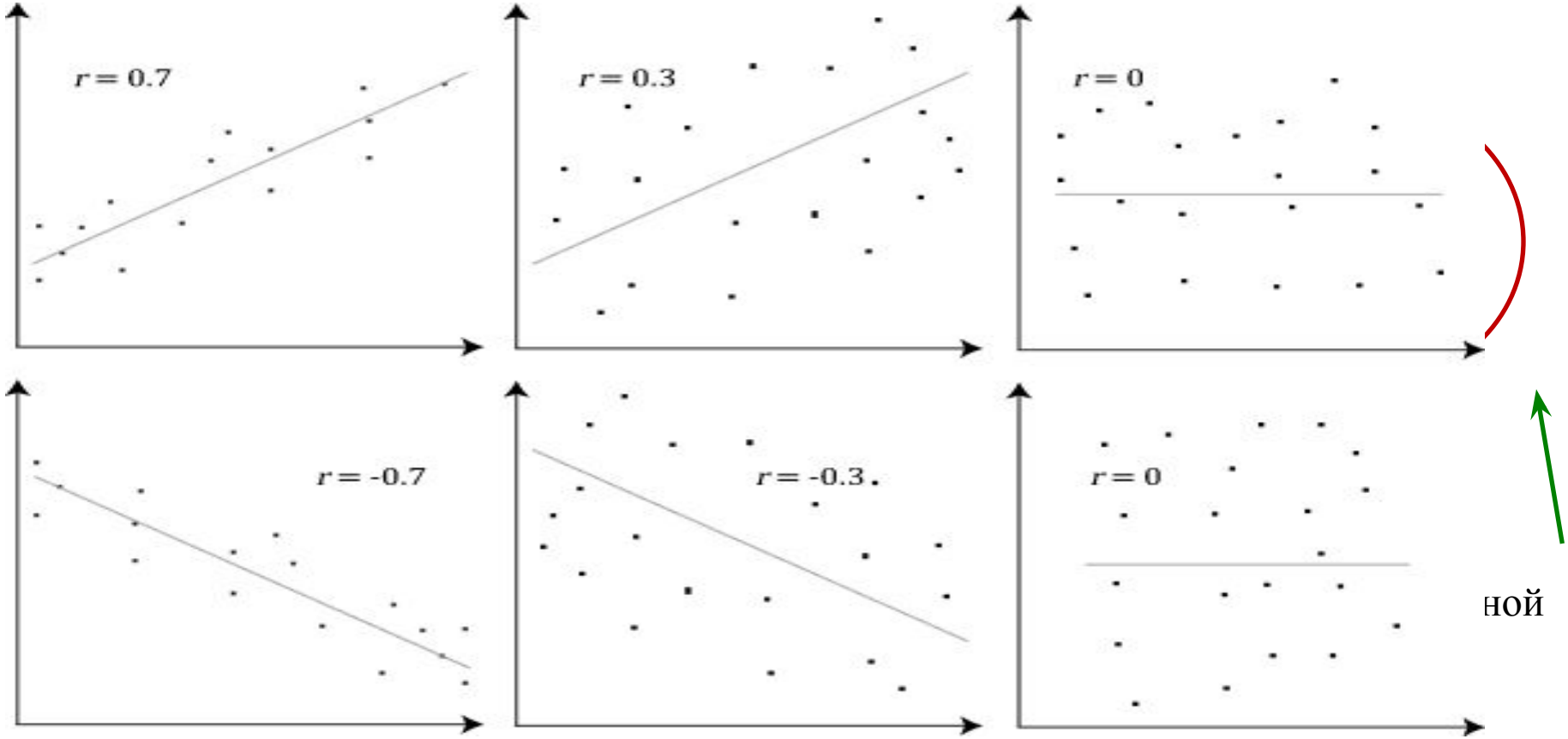

Уд. вес расх. на опл. усл. образ. в общ. расх. на опл. услуг
5
4
3
2
1
0
0
5000
10000
15000
Среднемес. номин. начисл. зарплата работаю


чество подборов за игру
44,00
43,00
42,00
7,50
8,00
8,50
9,00
ичество перехватов за игру


Продолжительность_жизни
85,0
80,0
75,0
70,0
65,0
60,0
40,00
60,00
80,00
100,00
Уровень_здравоохранения

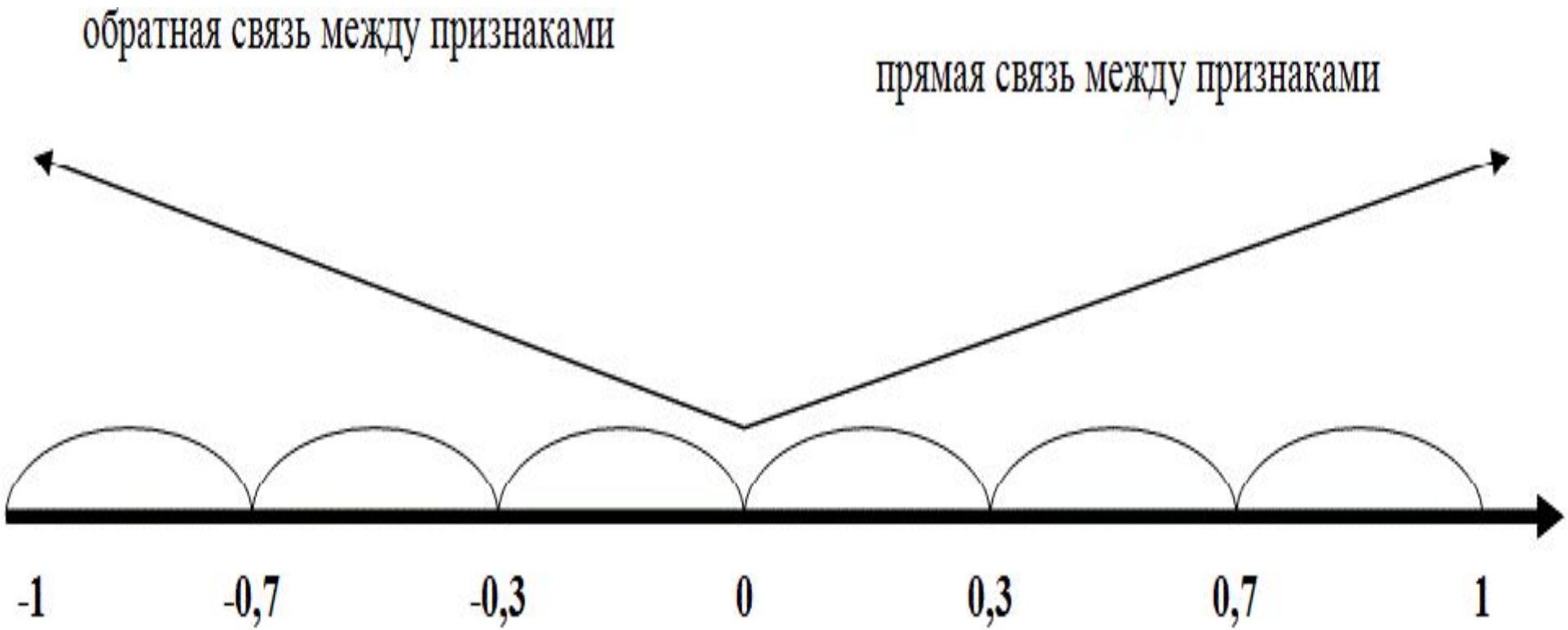
Свойства коэффициента корреляции:

Если точки не выстраиваются по прямой линии, а образуют «облако», коэффициент корреляции по абсолютной величине становится меньше единицы и по мере округления этого облака приближается к нулю.



Свойства коэффициента корреляции

1. $-1 \leq \rho \leq 1$



Свойства коэффициента корреляции

2. Если случайные величины x_j и x_l статистически независимы, то $\rho_{jl} = 0$, а в случае нормального распределения из некоррелированности x_j и x_l, когда $\rho_{jl} = 0$, следует их независимость.

(это не означает отсутствие любой зависимости между переменными, just not a linear one!)

Свойства коэффициента корреляции

2. Из условия $|\rho_{jl}| = 1$ следует наличие функциональной линейной связи между x_j и x_l и, наоборот, если x_j и x_l связаны линейной функциональной зависимостью, то

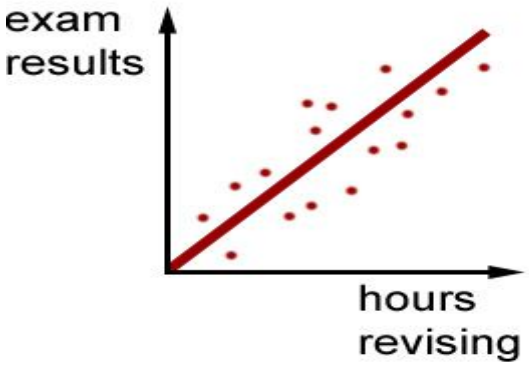
$$|\rho_{jl}| = 1$$

Чем ближе ρ к ± 1, тем теснее связь между X и Y.

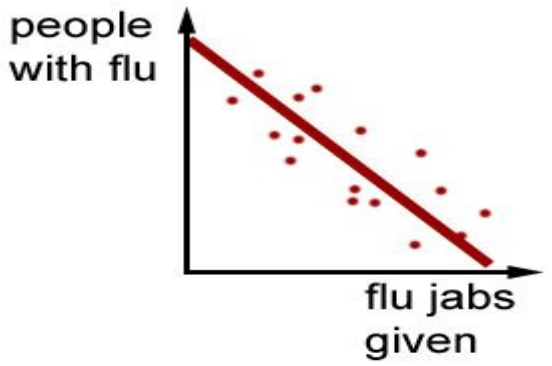
Свойства коэффициента корреляции:

3. $\rho > 0$ - свидетельствует о прямой зависимости между переменными (при увеличении значений одной переменной значения другой переменной также увеличиваются).

   $\rho < 0$ свидетельствует об обратной зависимости между переменными (при увеличении значений одной переменной значения другой переменной уменьшаются).



POSITIVE CORRELATION
- people who do more revision get higher exam results.
- revising increases success.

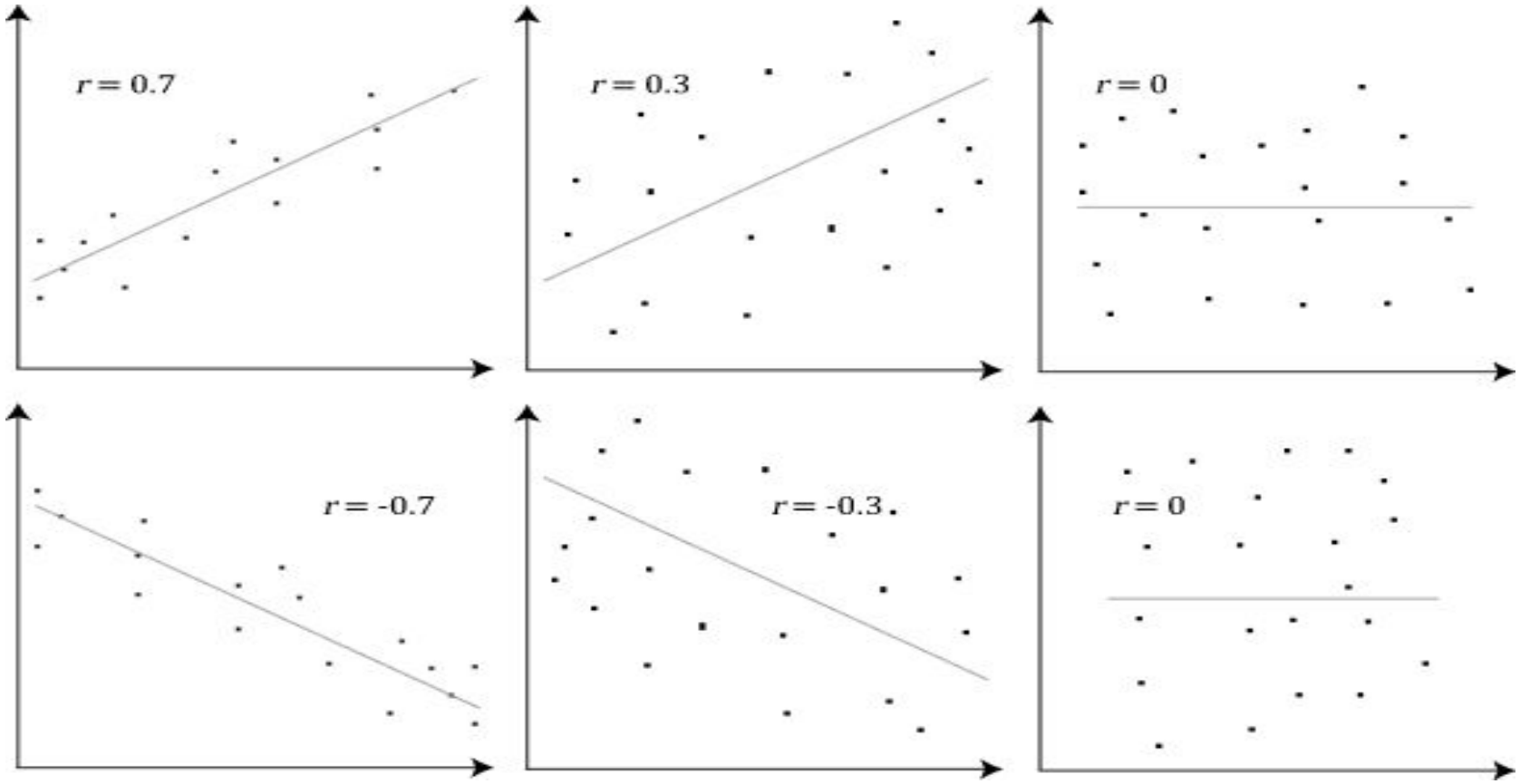


NEGATIVE CORRELATION
- when more jabs are given the number of people with flu falls.
- flu jabs prevent flu.

Свойства коэффициента корреляции:

3. $\rho > 0$ - свидетельствует о прямой зависимости между переменными

   $\rho < 0$ свидетельствует об обратной зависимости между переменными.



Свойства коэффициента корреляции

5. Сила корреляционной связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. Существуют различные рекомендации по интерпретации силы корреляционной взаимосвязи.

Значение коэффициента корреляции	**STRENGHT OF LINEAR RELATIONSHIP**
$0,7 < \|r\| \leq 1$	Сильная взаимосвязь, близкая к функциональной (strong)
$0,3 < \|r\| \leq 0,7$	Взаимосвязь средней силы (moderate)
$0,0 < \|r\| \leq 0,3$	Слабая взаимосвязь (weak)

Свойства коэффициента корреляции

Пример

Значение коэффициента корреляции (Value of r)	Сила линейной взаимосвязи (STRENGHT OF LINEAR RELATIONSHIP)
$0{,}8 \le r \le 1$ $-0{,}8 \le r \le -1$	Сильная взаимосвязь, близкая к функциональной (strong)
$0{,}6 \le r \le 0{,}8$ $-0{,}6 \le r \le -0{,}8$	Взаимосвязь средней силы (moderate)
$0{,}40 < \lvert r \rvert \le 0{,}6$	Умеренная
$0{,}20 < \lvert r \rvert \le 0{,}4$	Слабая взаимосвязь (weak)
$0 \le \lvert r \rvert \le 0{,}2$	очень слабая взаимосвязь

Свойства коэффициента корреляции

5. Неважно, какую переменную мы назовем x, а какую y.

   Коэффициент корреляции зависит только от выборочных данных, а не от названия переменных.

6. Парный коэффициент корреляции является симметричной характеристикой, т.е. $\rho_{jl} = \rho_{lj}$, что непосредственно следует из определения.

Свойства коэффициента корреляции

7. Коэффициент корреляции не имеет размерности и, следовательно, его можно сопоставлять для разных выборок. (В нашем примере часы или минуты, затраченные на подготовку к экзамену, не изменят величину r).

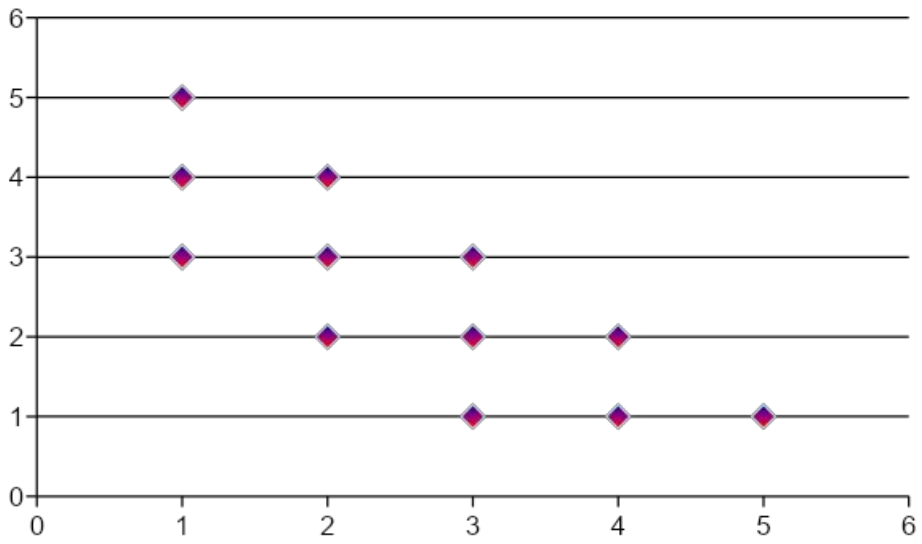
Свойства коэффициента корреляции

8. 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.

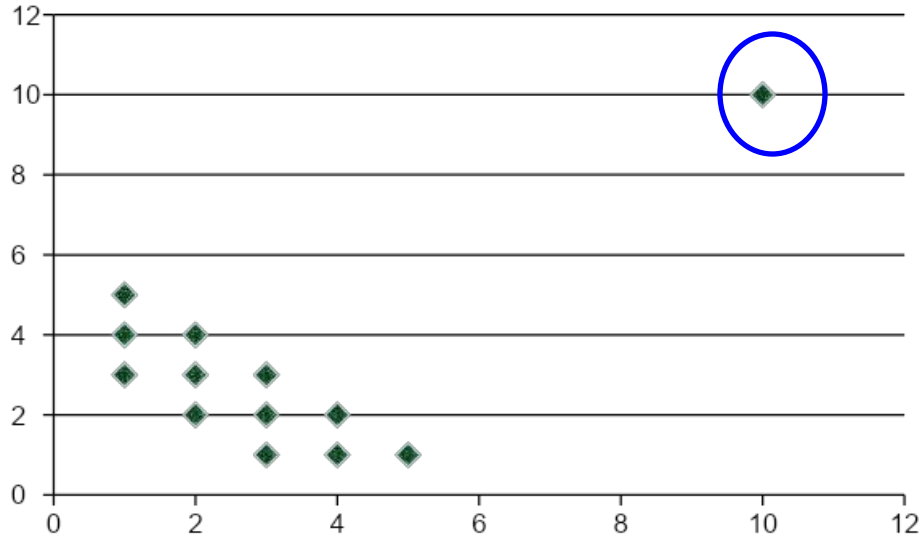
Свойства коэффициента корреляции:

9. Коэффициент корреляции очень чувствителен к выбросам (аномальным наблюдениям). Единичное extreme значение может иметь мощное воздействие на r и привести к неправильным выводам (?) .

Пример



Обратная связь
r=-0,80

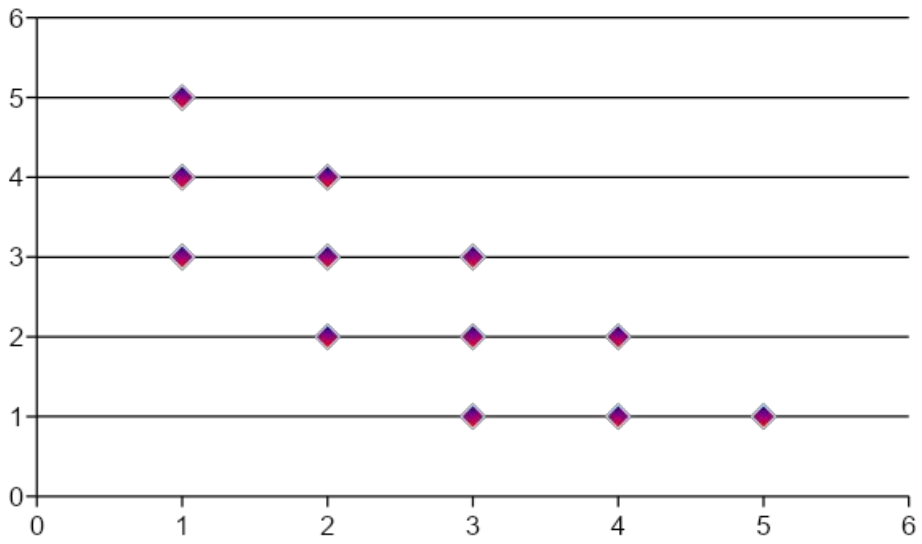


Прямая связь
r=0,51

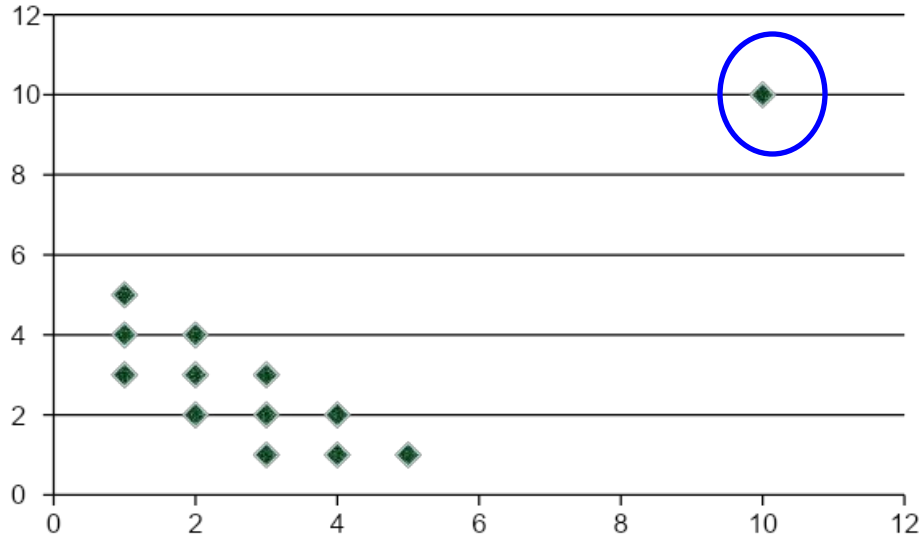
Свойства коэффициента корреляции:

9. Коэффициент корреляции очень чувствителен к выбросам (аномальным наблюдениям). Единичное extreme значение может иметь мощное воздействие на r и привести к неправильным выводам (так как базируется на среднем) .

Пример



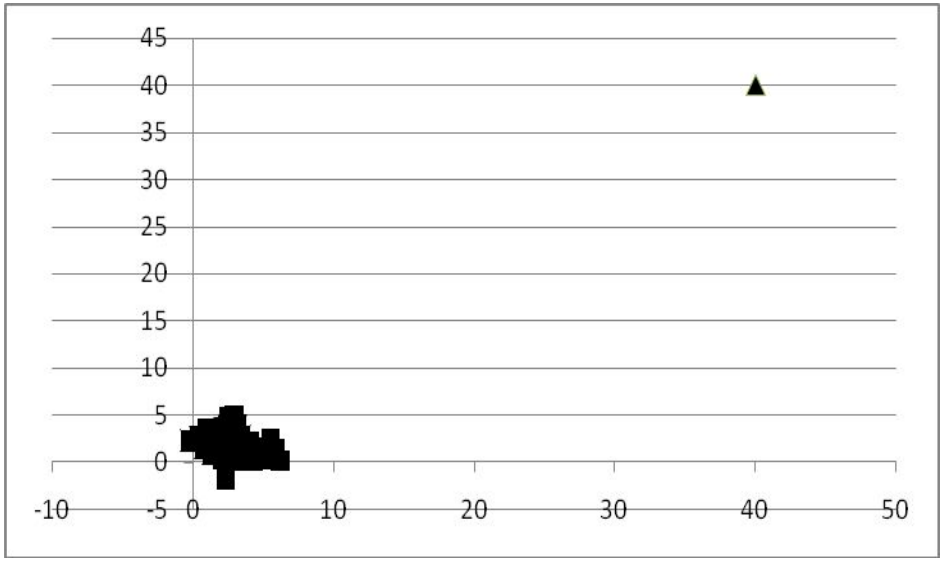
Обратная связь
r=-0,80

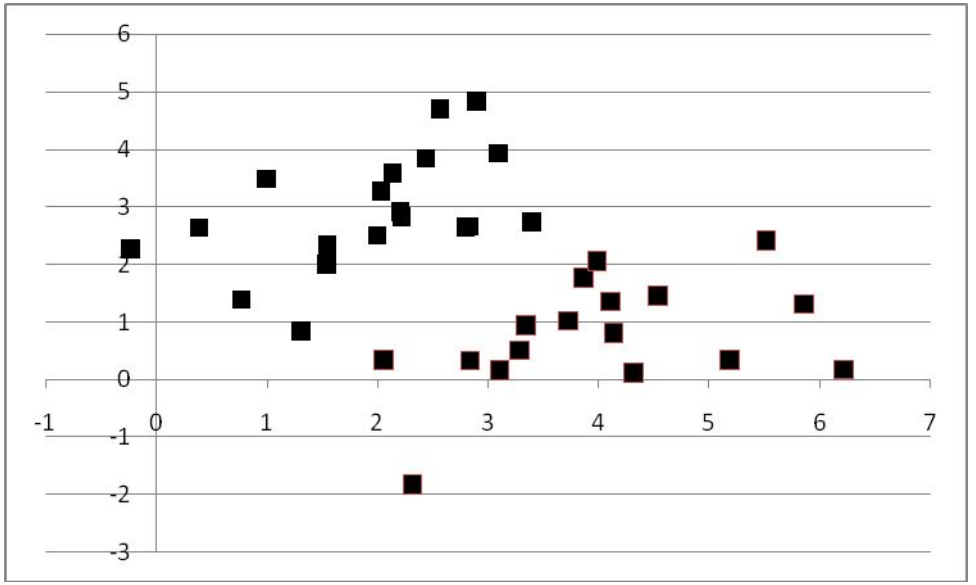


Прямая связь
r=0,51

Свойства коэффициента корреляции:

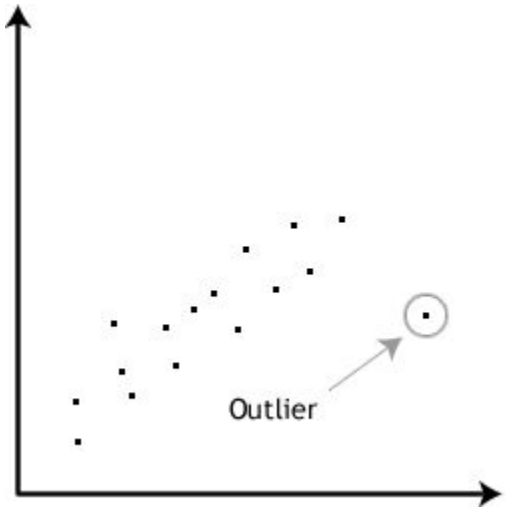
Наблюдения до и после удаления выброса





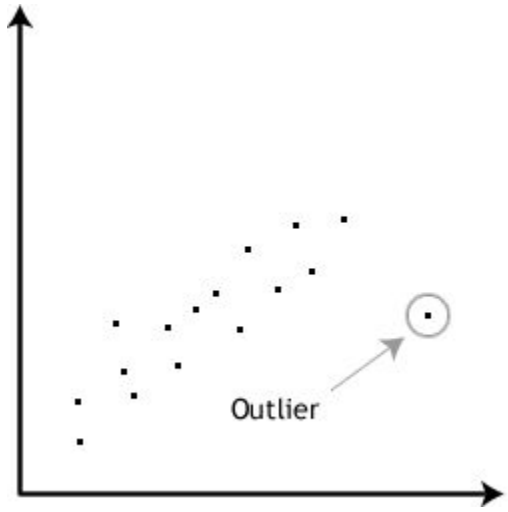
Свойства коэффициента корреляции:

9. if you cannot justify removing the data point(s), you can run a non-parametric test such as Spearman's rank-order correlation or Kendall's Tau Correlation instead, which are much less sensitive to outliers. This might be your best approach if you cannot justify removing the outlier. The diagram below indicates what a potential outlier might look



Свойства коэффициента корреляции:

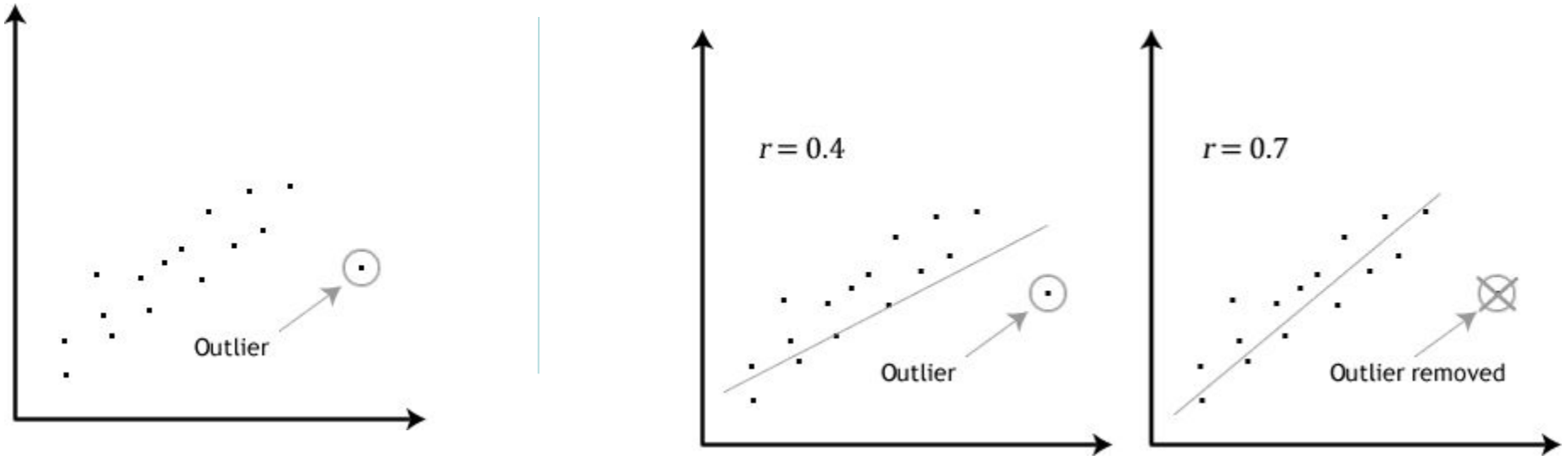
9. if you cannot justify removing the data point(s), you can run a non-parametric test such as Spearman's rank-order correlation or Kendall's Tau Correlation instead, which are much less sensitive to outliers. This might be your best approach if you cannot justify removing the outlier. The diagram below indicates what a potential outlier might look



Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions regarding your data. This point is most easily illustrated by studying scatterplots of a linear relationship with an outlier included and after its removal, with respect to both the line of best fit and the correlation coefficient. This is illustrated in the diagram below:

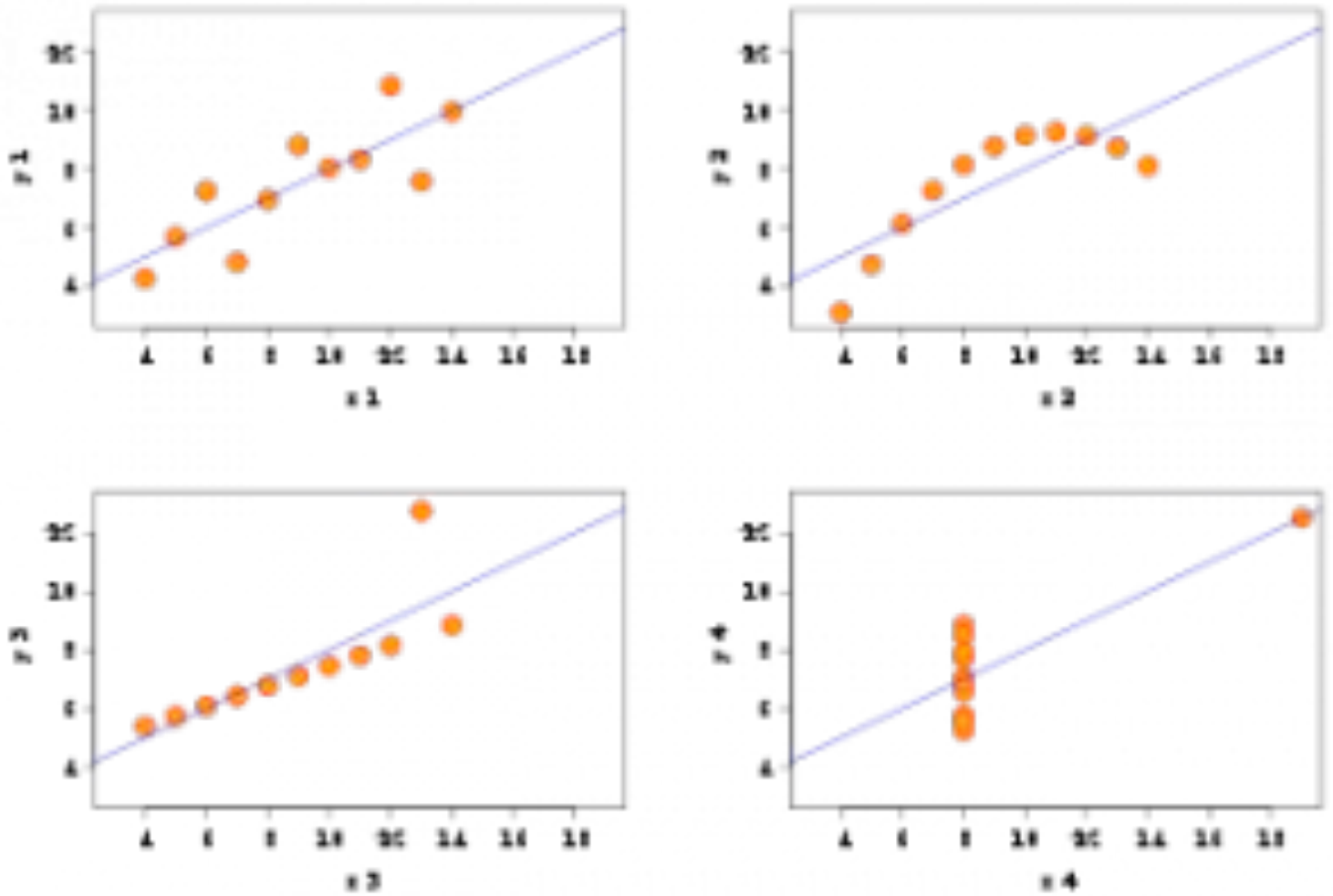
Свойства коэффициента корреляции:

9. if you cannot justify removing the data point(s), you can run a non-parametric test such as Spearman's rank-order correlation or Kendall's Tau Correlation instead, which are much less sensitive to outliers. This might be your best approach if you cannot justify removing the outlier. The diagram below indicates what a potential outlier might look



Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions regarding your data. This point is most easily illustrated by studying scatterplots of a linear relationship with an outlier included and after its removal, with respect to both the line of best fit and the correlation coefficient. This is illustrated in the diagram below:

Свойства коэффициента корреляции:

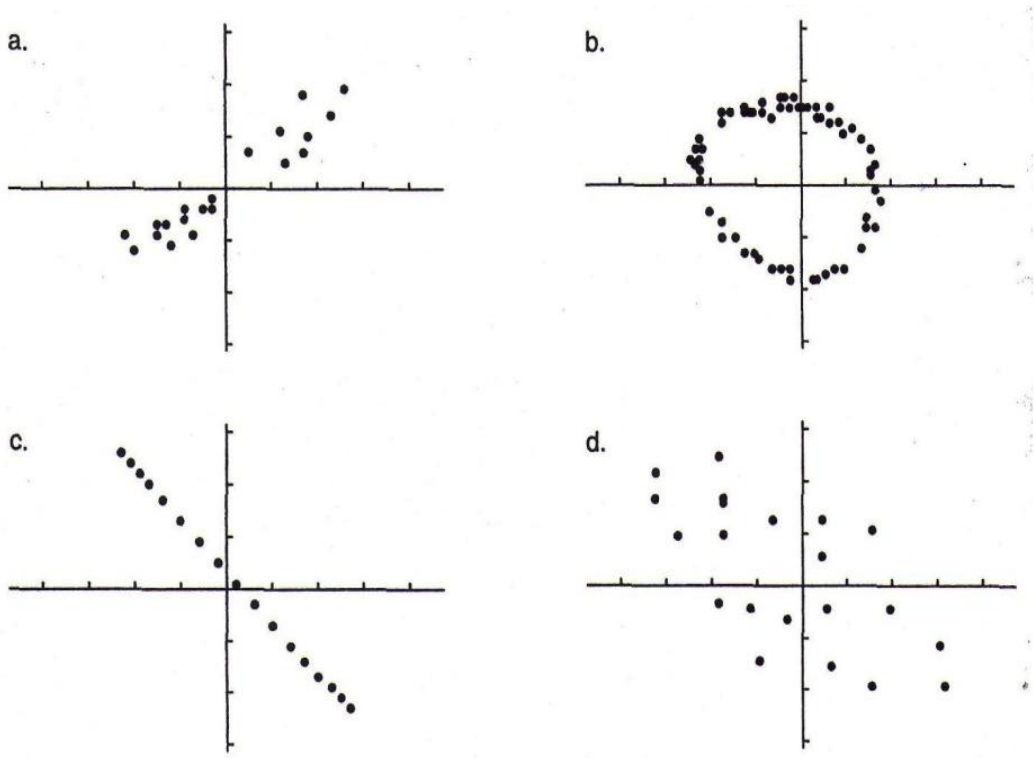


Четыре различных набора данных, коэффициент корреляции на которых равен 0.81

- Неустойчивость к выбросам.

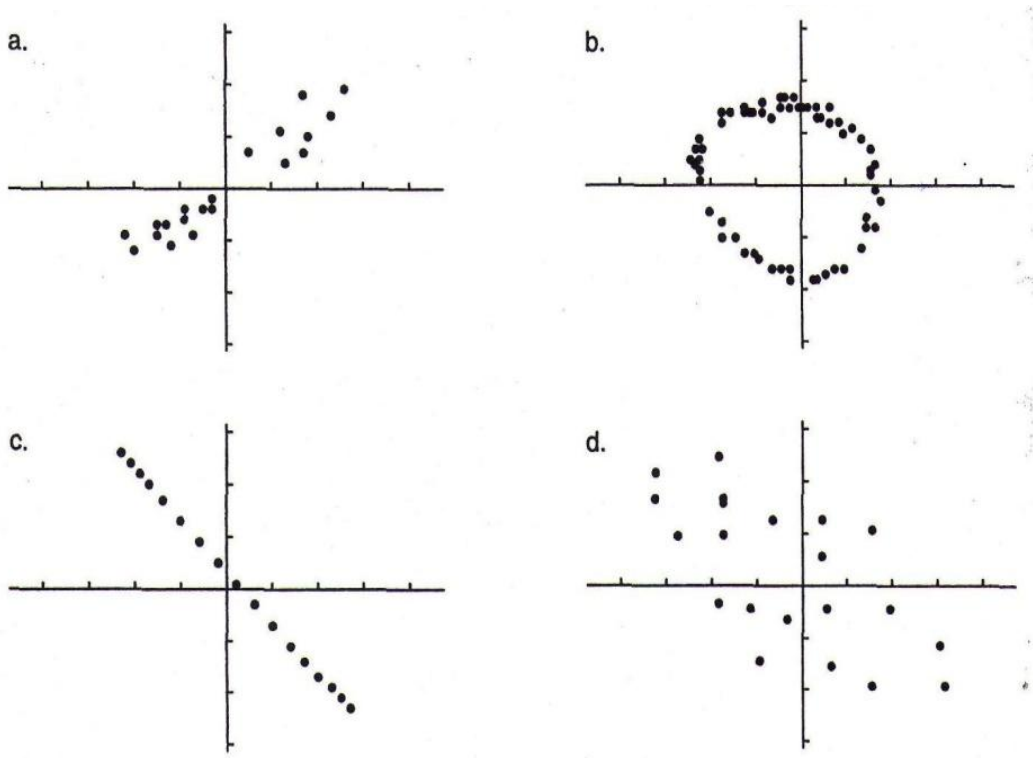
Пример

Оцените значение коэффициента корреляции *r* для каждого из представленных ниже графиков:



Пример

Оцените значение коэффициента корреляции *r* для каждого из представленных ниже графиков:



Ответ

а) 0,8;
б) 0;
с) -1;
d) -0,5

Проверка значимости коэффициента корреляции

Значимость **парных коэффициентов корреляции** проверяется с помощью t-критерия Стьюдента.

$$H_0: \quad \rho = 0$$

$$H_1: \quad \rho \neq 0$$ (двухсторонняя критическая область)

1. Расчет наблюдаемого значения статистики по формуле:

$$t_{набл} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

где **r** - оценка парного коэффициент корреляции.

Проверка значимости коэффициента корреляции

2. Нахождение критического значения статистики по таблицам распределения

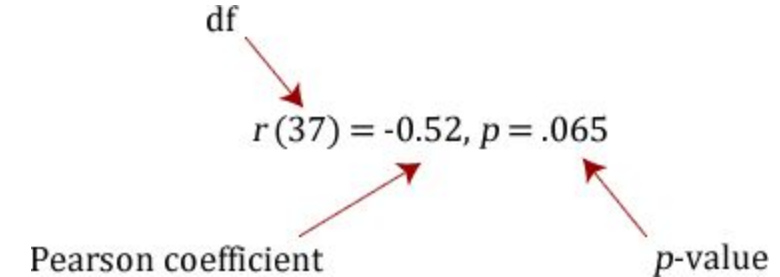
$t_{кр}$ определяется по таблице распределения Стьюдента для заданного уровня значимости α и $\nu = n - 2$

Уровень значимости	надежность
0,05	**95%**
0,01	**99 %**

3. Вывод по гипотезе

проверяемый коэффициент корреляции считается значимым, т. е. гипотеза H_0: ρ=0 отвергается с вероятностью ошибки α,

если $|t_{набл}| > t_{кр}$



Критические области для распределения Стьюдента (t-распределения)

n	Вероятность $a = St(t) = P(\|T\| > t_{табл})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,043	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,327	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,583

$$t_{кр}(\alpha=0,05;\ \nu=6) = 2,447$$

http://www.e-biblio.ru/book/bib/10_statistika/tv_i_ms/book/docs/

Корреляционный анализ

II способ. С использованием критерия Фишера-Иейтса

1. За $r_н$ принимается выборочное значение коэффициента корреляции **r**
2. $r_{кр}$ (α, **v=n-2**) находится по таб. Фишера-Иейтса (таб.8)
3. Вывод по гипотезе Рассчитанное значение ***r*** сравнивается с $r_{кр}$:

Если $|r| > r_{кр}$ => гипотеза H_0 отвергается =>

ρ – значим (с вероятностью ошибки α)

v	Двусторонние границы				v	Двусторонние границы			
	0,05	0,02	0,01	0,001		0,05	0,02	0,01	0,001
1	0,997	1,000	1,000	1,000	16	0,468	0,543	0,590	0,708
2	0,950	0,980	0,990	0,999	17	0,456	0,529	0,575	0,693
3	0,878	0,934	0,959	0,991	18	0,444	0,516	0,561	0,679
4	0,811	0,882	0,917	0,974	19	0,433	0,503	0,549	0,665
5	0,754	0,833	0,875	0,951	20	0,423	0,492	0,537	0,652
6	0,707	0,789	0,834	0,925	25	0,381	0,445	0,487	0,597
7	0,666	0,750	0,798	0,898	30	0,349	0,409	0,449	0,554
8	0,632	0,715	0,765	0,872	35	0,325	0,381	0,418	0,519
9	0,602	0,685	0,735	0,847	40	0,304	0,358	0,393	0,490
10	0,576	0,658	0,708	0,823	45	0,288	0,338	0,372	0,465

Пример: Преподаватель попросил студентов (n=15) записать, сколько часов они потратили на подготовку к промежуточному экзамену. Результаты приведены в табл.

Student	Hours studied	Score on exam
A	0,5	65
B	2,5	80
C	3,0	77
D	1,5	60
E	1,25	68
F	0,75	70
G	4,0	83
H	2,25	85
I	1,5	70
J	6,0	96
K	3,25	84
L	2,5	84
M	0,0	51
N	1,75	63
O	2,0	71

Пример: Преподаватель попросил студентов (n=15) записать, сколько часов они потратили на подготовку к промежуточному экзамену. Результаты приведены в табл.

Student	Hours studied	Score on exam
A	0,5	65
B	2,5	80
C	3,0	77
D	1,5	60
E	1,25	68
F	0,75	70
G	4,0	83
H	2,25	85
I	1,5	70
J	6,0	96
K	3,25	84
L	2,5	84
M	0,0	51
N	1,75	63
O	2,0	71

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$r = 0{,}887$$

Проверка независимости (значимости) признаков

$$H_0: \ \rho_{xy} = 0$$

Используем критерий Стьюдента для проверки гипотезы

1. $$t_{набл} = \frac{0{,}887}{\sqrt{1 - 0{,}887^2}} \cdot \sqrt{10 - 2} = 5{,}4327$$

2. $$t_{kp}(\alpha = 0{,}05; \nu = 10 - 2 = 8) = 3{,}833$$

3. Вывод

$$\left|t_{набл}\right| = 5{,}4327 > t_{kp} = 3{,}833$$

Коэффициент детерминации в двумерной модели

Квадрат парного коэффициент корреляции $\rho^2_{1,2}$ называется **коэффициентом детерминации**.

$\rho^2_{1,2}$ **характеризует долю дисперсии одной переменной (результативной), обусловленную влиянием другой переменной.**

Соответственно (1- $\rho^2_{1,2}$) показывает долю **остаточной дисперсии** случайной величины X_1, обусловленную влиянием не включённых в корреляционную модель факторов.

Плотность вероятности выборочного коэффициента корреляции имеет сложный вид, поэтому используют специально подобранные функции от выборочного коэффициента корреляции, которые подчиняются хорошо изученным законам, например нормальному или Стьюдента.

При нахождении доверительного интервала для коэффициента корреляции ρ чаще используют преобразование Фишера:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

Эта статистика уже при n>10 распределена приблизительно нормально, с параметрами $M(z) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$.

Интервальные оценки параметров связи

I. Для значимых параметров связи (коэффициентов корреляции) с надежностью γ определяют интервальные оценки.

Алгоритм

1. **Нахождение интервальной оценки для вспомогательной статистики Z с помощью Z-преобразования Фишера**

$$Z' - t_\gamma \sqrt{\frac{1}{n-3}} \le Z \le Z' + t_\gamma \sqrt{\frac{1}{n-3}}$$

δ δ

t_γ вычисляют по таблице интегральной функции Лапласа (табл. 1) из условия $\Phi(t_\gamma)=\gamma$

- Значение Z' (Z_r)определяют по таблице Z - преобразования (табл. 6) по найденному значению r.
- **!** Функция Z_r нечетная:
  $Z'(-r) = -Z'(r)$ нечетная

$$r \xrightarrow{\text{таб.6}} z_r$$

$$\gamma = \Phi(t) \xrightarrow{\text{таб.1}} t_\gamma \rightarrow \Delta z = \frac{t_\gamma}{\sqrt{n-3}}$$

Нормальный закон распределения (значения функции Лапласа)

Целые и десятичные доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412

http://www.e-biblio.ru/book/bib/10_statistika/tv_i_ms/book/docs/

Таблица Z-преобразования Фишера

$$Z = \frac{1}{2}\{\ln(1 + r) - \ln(1 - r)\}$$

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0101	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
1	0,1003	0,1104	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3767	0,3884	0,4001	0,4118
4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
5	0,5493	0,5627	0,5764	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
6	0,6932	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
7	0,8673	0,8872	0,9077	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
9	1,4722	1,5275	1,5890	1,6584	1,7381	1,8318	1,9459	2,0923	2,2976	2,6467
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

http://www.e-biblio.ru/book/bib/10_statistika/tv_i_ms/book/docs/

Интервальные оценки параметров связи

2. Обратный переход от *Z* к r

осуществляют также по таблице Z – преобразования.

3. Получение интервальной оценки для ρ с надежностью γ :

$$r_{\min} \le \rho \le r_{\max}$$

Таким образом, с вероятностью γ гарантируется, что генеральный коэффициент корреляции ρ будет находиться в интервале от r_{min} до r_{max}.

С помощью доверительного интервала можно проверить ***значимость коэффициента корреляции ρ:***

если ноль попадает в доверительный интервал, то коэффициент корреляции незначимый.

Трёхмерная корреляционная модель

Пусть признаки X, Y, Z образуют трехмерную нормально распределенную генеральную совокупность, которая определяется девятью параметрами:

$$(X,Y,Z) \leftrightarrow N(\mu_x, \mu_y, \mu_z, \sigma_x, \sigma_y, \sigma_z, \rho_{xy}, \rho_{yz}, \rho_{xz})$$

Трёхмерная корреляционная модель

Пусть признаки X, Y, Z образуют трехмерную нормально распределенную генеральную совокупность, которая определяется девятью параметрами:

$$(X,Y,Z) \leftrightarrow N(\mu_x, \mu_y, \mu_z, \sigma_x, \sigma_y, \sigma_z, \rho_{xy}, \rho_{yz}, \rho_{xz})$$

! Одномерные распределения X, Y, Z

и двумерные [(X, Y), (X,Z), (Y, Z)] распределения компонент,

а так же условные распределения при фиксированных одной [(X,Y)/Z; (X,Z)/Y; (Y,Z)/X]

и двух переменных [X/(Y,Z); Y/(X,Z); z/(X,Y)]

являются нормальными. Поэтому поверхности и линии регрессии являются плоскостями и прямыми соответственно.

Трёхмерная корреляционная модель

Для изучения разнообразия связей между тремя случайными величинами рассчитывают

- парные,
- частные
- множественные

коэффициенты корреляции (детерминации)

Трёхмерная (многомерная) корреляционная модель

Исходной для анализа является матрица:

$$X=\begin{pmatrix} x_{11}x_{12}x_{13} \\ \cdots\cdots\cdots \\ x_{i1}x_{i2}x_{i3} \\ \cdots\cdots\cdots \\ x_{n1}x_{n2}x_{n3} \end{pmatrix}$$

размерности (**n x 3**),

$$\begin{pmatrix} x_{11}\cdots x_{1j}\cdots x_{1k} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{i1}\cdots x_{ij}\cdots x_{ik} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n1}\cdots x_{nj}\cdots x_{nk} \end{pmatrix}$$

размерности (**n x k**)

***i*-я** строка которой характеризует *i*-е наблюдение (объект) по всем показателям (j=1, 2, 3,…,к).

Трёхмерная (многомерная) корреляционная модель

Парный коэффициент корреляции, например, ρ_{xy} характеризует тесноту связи между переменными X и Y на фоне действия пепеменной Z (на фоне действия всех остальных переменных, включенных в модель).

Матрица парных коэффициентов корреляции

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & r_{12} & . & . & r_{1\kappa} \\ r_{21} & 1 & . & . & r_{2\kappa} \\ r_{\kappa 1} & r_{\kappa 2} & . & . & 1 \end{pmatrix}$$

Матрица **R** является симметричной и положительно определенной, на главной диагонали стоят единицы.

Трёхмерная корреляционная модель

Частный коэффициент корреляции, например, $\rho_{xy/z}$ характеризует тесноту связи между переменными X и Y при фиксированном значении переменной Z (независимо от её влияния).

Если парный коэффициент корреляции больше частного , т.е.

$\rho_{xy} > \rho_{xy/z}$, то переменная Z усиливает связь между переменными X и Y.

Если $\rho_{xy} < \rho_{xy/z}$, то переменная Z ослабляет связь между переменными X и Y.

Трёхмерная корреляционная модель

Частный коэффициент корреляции обладает **всеми свойствами** парного коэффициента корреляции , т.к. он является коэффициентом корреляции двумерного условного распределения.

Сравнение частных коэффициентов корреляции позволяет ранжировать факторы по тесноте их связи с результатом (у).

$$\mathbf{R}_{\text{частн}} = \begin{pmatrix} 1 & r_{12/3,\ldots k} & \cdot & \cdot & r_{1\kappa/2,3,\ldots,k-1} \\ r_{21/3,\ldots,k} & 1 & \cdot & \cdot & r_{2\kappa/1,3,\ldots k-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{k1/2,3,\ldots,k-1} & r_{k2/1,3,\ldots,k-1} & \cdot & \cdot & 1 \end{pmatrix}$$

Трёхмерная корреляционная модель

Частный коэффициент корреляции

например,

$$\rho_{xy/z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{(1-\rho_{xz}^2)\cdot(1-\rho_{yz}^2)}}$$

$$-1 \le \rho_{xy/z} \le 1$$

Точечная оценка частного коэффициента корреляции:

$$r_{12/3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1-r_{13}^2)\cdot(1-r_{23}^2)}} = -\frac{A_{12}}{\sqrt{A_{11} \cdot A_{22}}}$$

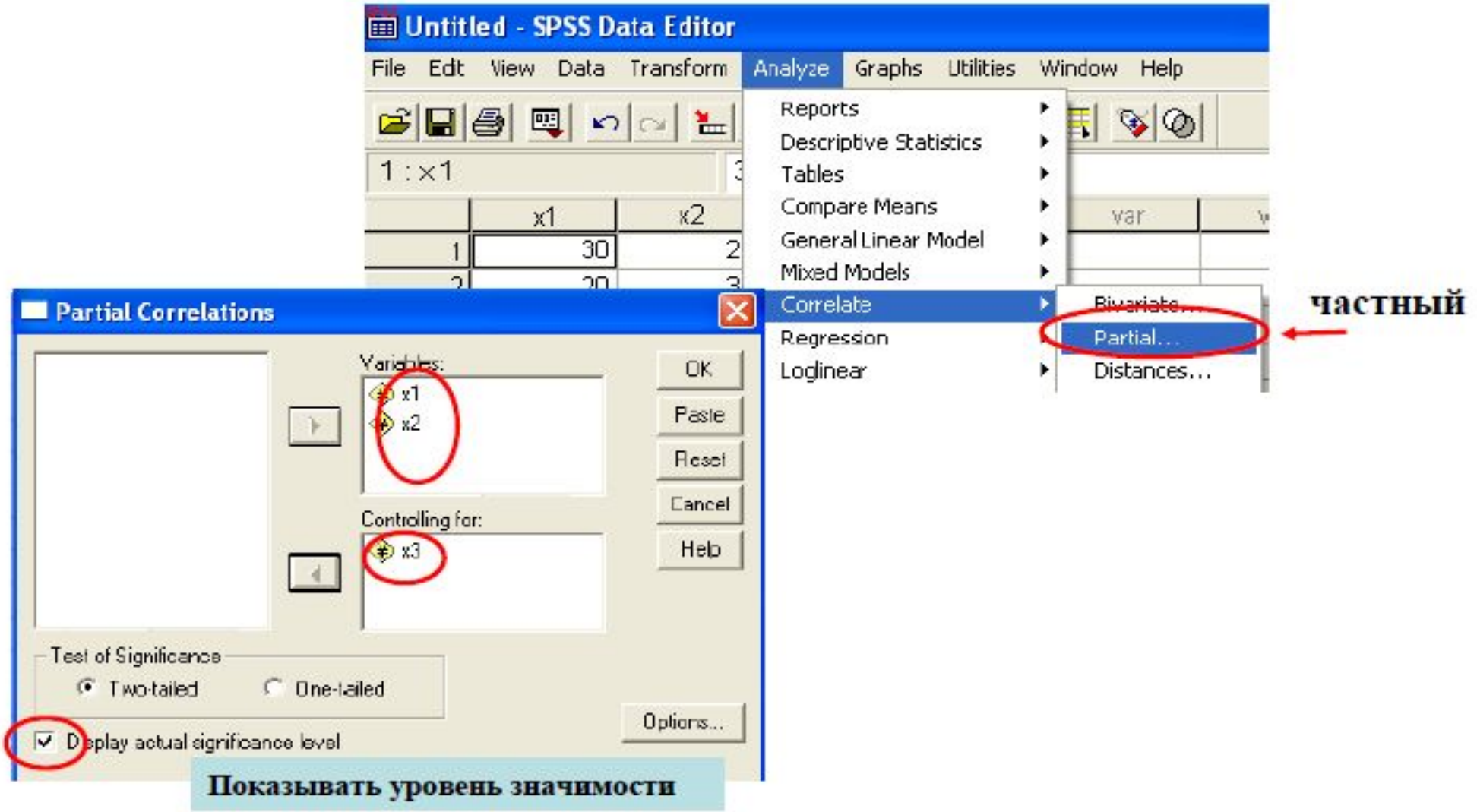
где A_{ij} - алгебраическое дополнение элемента r_{ij} корреляционной матрицы R.

$A_{ij} = (-1)^{i+j} \times M_{ij}$, где M_{ij} - минор, определитель матрицы, получаемой из матрицы R путем

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

i — строка, *j* — столбец

Частный коэффициент корреляции в SPSS



Частный коэффициент корреляции в SPSS

Результат:

Correlations

Control Variables			x1	x2
x3	x1	Correlation	1.000	.443
		Significance (2-tailed)	.	.320
		df	0	5
	x2	Correlation	.443	1.000
		Significance (2-tailed)	.320	.
		df	5	0

коэффициент

Значимость коэффициента

Трёхмерная корреляционная модель

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

Проверка значимости парного и частного КК

I способ. t – критерий Стьюдента (таб.2)

2. Рассчитывается наблюдаемое значение статистики $t_н$:

$$t_H = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-l-2}$$

3. Находится критическое значение статистики $t_{кр}$:

$t_{кр}$ (α, **v=** n-l-2)

4. Вывод по гипотезе

II способ. Критерий Фишера-Иейтса (таб.8) с учетом порядка КК

Трёхмерная корреляционная модель

Интервальная оценка для значимого парного и частного коэффициента корреляции

Аналогично построению ИО для парного коэффициента корреляции в двумерной модели.

Отличие

$$\Delta Z = \frac{t_\gamma}{\sqrt{n-l-3}} = \frac{t_\gamma}{\sqrt{n-4}}$$

Трёхмерная корреляционная модель

Множественный коэффициент корреляции

Множественный коэффициент корреляции в трёхмерной модели служит показателем тесноты линейной связи между одной переменной и *двумерным* массивом двух других переменных.

Например, $\rho_{y/xz}$ (ρ_y) служит показателем тесноты линейной связи между переменной У и двумерной величиной (X,Z).

Множественный коэффициент корреляции в многомерной модели служит показателем тесноты линейной связи между одной переменной и массивом других переменных.

Трёхмерная корреляционная модель

Множественный коэффициент корреляции

Точечная оценка множественного коэффициента корреляции:

$$r_{1/2,3} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 \cdot r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}} = \sqrt{1 - \frac{|R|}{A_{11}}}$$

где |R| - определитель матрицы парных коэффициентов корреляции,

A_{ij} - алгебраическое дополнение элемента r_{ij} корреляционной матрицы R.

$A_{ij} = (-1)^{i+j} \times M_{ij}$, где M_{ij} - минор, определитель матрицы, получаемой из матрицы R путем вычеркивания *i*-й строки и *j*-го столбца.

Коэффициент детерминации

Квадрат множественного коэффициент корреляции $\rho^2_{1/2,3}$ называется **множественным коэффициентом детерминации**.

Он характеризует долю дисперсии одной переменной (результативной), обусловленной влиянием всех остальных переменных (аргументов), включенных в модель.

Многомерная корреляционная модель

Множественный коэффициент детерминации в общем случае многомерной корреляционной модели, например, $\rho^2_{1/2,3,\ldots к}$ показывает долю дисперсии случайной величины X_1, обусловленную влиянием остальных переменных X_2, X_3, … $X_к$, включённых в корреляционную модель.

Соответственно $(1- \rho^2_{1/2,3,\ldots к})$ показывает долю **остаточной дисперсии** случайной величины X_1, обусловленную влиянием других, не включённых в корреляционную модель факторов.

Множественный коэффициент корреляции и его свойства

1. Множественный коэффициент корреляции изменяется в интервале

$$0 \le \rho_y \le 1$$

Множественный коэффициент корреляции и его свойства

1. Множественный коэффициент корреляции изменяется в интервале

$$0 \le \rho_y \le 1$$

2. Минимальное значение $\rho_y = 0$ соответствует случаю полного отсутствия корреляционной связи между y и остальными переменными.

⇨ усредненная дисперсия «регрессионных остатков» в точности равна общей вариации результирующего показателя.

Если в трехмерной модели $\rho_y = 0$,
то одномерная случайная величина У и
двумерная случайная величина (X, Z)
являются *независимыми* (в силу нормальности распределения).

Множественный коэффициент корреляции и его свойства

3. Максимальное значение $\rho_y = 1$ соответствует случаю полного отсутствия варьирования «регрессионных остатков»,что означает наличие функциональной связи между величиной y и остальными переменными.

В этом случае мы имеем возможность точно восстановить условные значения $y(X)=\{y/\xi=X\}$ по значениям факторных (предикторных) переменных X.

Свойства множественного коэффициента корреляции

4. Множественный коэффициент корреляции превышает любой парный или частный коэффициент корреляции, характеризующий статистическую связь результирующего показателя.

Свойства множественного коэффициента корреляции

5. Присоединение любой новой предсказывающей переменной не может уменьшить величины R (независимо от порядка присоединения).

$$R_{y/x_1} \le R_{y/x_1,x_2} \le R_{y/x_1,x_2,x_3} \le \ldots \le \mathrm{R}_{y/\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_к}$$

Коэффициент детерминации

Наибольшему множественному коэффициенту детерминации соответствуют большие частные коэффициенты корреляции.

Например, если

$$R_x^2 > R_z^2, \quad R_x^2 > R_y^2$$

$$\rho_{xz/y} > \rho_{zy/x}$$

$$\rho_{xy/z} > \rho_{zy/x}$$

Трёхмерная корреляционная модель

Множественный коэффициент детерминации

Проверка значимости множественного коэффициента (и корреляции (детерминации), например,

H_0: $\rho_{1/2,3}$ =0, осуществляется с помощью F-критерия.

1. Вычисляется

$$F_{набл} = \frac{\frac{1}{2} \cdot r^2_{1/2,3}}{\frac{1}{n-3} \cdot (1 - r^2_{1/2,3})}$$

- для трехмерного случая

$$F_{набл} = \frac{\frac{1}{\kappa - 1} r^2_{1/2,\ldots,\kappa}}{\frac{1}{n - \kappa} \left(1 - r^2_{1/2,\ldots,\kappa}\right)}$$

- для многомерного случая

Трёхмерная корреляционная модель

Множественный коэффициент детерминации

2. По таблице F-распределения Фишера-Снедекора (таб.4) определяют $F_{кр}$:

$F_{кр}(\alpha; \nu_1=2; \nu_2=n-3)$ | $F_{кр}(\alpha; \nu_1= \quad ; \nu_2=$

$$F_{набл} = \frac{\frac{1}{\kappa-1} r^2_{1/2,\ldots,\kappa}}{\frac{1}{n-\kappa}\left(1-r^2_{1/2,\ldots,\kappa}\right)}$$

3. Если $F_н > F_{кр}$, то гипотеза H_0 отвергается с вероятностью ошибки α и множественный коэффициент корреляции (и соответствующий коэффициент детерминации) считается статистически *значимым*.

Распределение Фишера-Снедекора
(F-распределение)

n_1 / n_2	1	2	3	4	5	6	8	12	24	∞	t
1	161,4 4052 406523	199,5 4999 500016	215,7 5403 536700	224,6 5625 562527	230,2 5764 576449	234,0 5859 585953	238,9 5981 598149	243,9 6106 610598	249,0 6234 623432	253,3 6366 636535	12,71 63,66 636,2
2	18,51 98,49 998,46	19,00 99,01 999,00	19,16 00,17 999,20	19,25 99,25 999,20	19,30 99,30 999,20	19,33 99,33 999,20	19,37 99,36 999,40	19,41 99,42 999,60	19,45 99,46 999,40	19,50 99,50 999,40	4,30 9,92 31,00
3	10,13 34,12 67,47	9,55 30,81 148,51	9,28 29,46 141,10	9,12 28,71 137,10	9,01 28,24 134,60	8,94 27,91 132,90	8,84 27,49 130,60	8,74 27,05 128,30	8,64 26,60 125,90	8,53 26,12 123,50	3,18 5,84 12,94
4	7,71 21,20 74,13	6,94 18,00 61,24	6,59 16,69 56,18	6,39 15,98 53,43	6,26 15,52 51,71	6,16 15,21 50,52	6,04 14,80 49,00	5,91 14,37 47,41	**α=0.05** **α=0.01** **α=0.001**	,63 3,46 4,05	2,78 4,60 8,61

$F_{кр}(\ \alpha=0.05;\ v_1=12;\ v_2=4\ )=5,91$

http://www.e-biblio.ru/book/bib/10_statistika/tv_i_ms/book/docs/

Корреляционный анализ

Коэффициент корреляции	Что характеризует?
парный тесноту линейной зависимости между двумя переменными на фоне действия всех остальных показателей	тесноту линейной зависимости между двумя переменными на фоне действия всех остальных показателей $-1 \le \rho_{jl} \le 1$
частный тесноту линейной зависимости между двумя переменными при исключении влияния всех остальных показателей, входящих в модель	тесноту линейной зависимости между двумя переменными при исключении влияния всех остальных показателей, входящих в модель $-1 \le \rho_{jl/1,2,..,k} \le 1$
множественный тесноту линейной связи между одной переменной (результативной) и остальными показателями	тесноту линейной связи между одной переменной (результативной) и остальными показателями $0 \le \rho_{j} \le 1$

Число наблюдений достаточно велико

Если число наблюдений достаточно велико и особенно если наблюдения <u>объединяются поинтервально</u>, т.е. все значения, попавшие в интервал, округляются до значения середины интервала

(например, рост измеряется с точность до целых сантиметров, а вес – с точностью до целых килограммов),

то каждая из наблюдаемых пар значений может встретится несколько раз.

строят таблицы с учетом частот встречаемости.

Такую табл. по сгруппированным данным называют <u>корреляционной</u>.

Пример соотношения роста (X) и массы тела (У)

Y/X	x_1	x_2	…	x_j	…	x_k	m_y
y_1	m_{11}	m_{21}	…	m_{i1}	…	m_{kl}	m_{*1}
y_2	m_{12}	m_{22}	..	m_{i2}	…	m_{k2}	m_{*2}
…	…	…	…	…	…	…	…
y_i	m_{1j}	m_{2j}	…	m_{ij}	…	m_{kj}	m_{*j}
…			…		…		…
y_l	m_{11}	m_{2l}	…	m_{il}	…	m_{kl}	m_{*l}
m_y	m_{1*}	m_{2*}		m_{i*}		m_{k*}	**n**

В первой строке в возрастающем порядке расположены варианты x_i, а в первом столбце – варианты y_j. На пересечении строк и столбцов находится частота $\mathbf{m_{ij}}$, обозначающая число точек выборки, значения признаков у которых равны (xi,yj).

Корреляционная таблица

Некоторые mij=0.

В последней строке (столбце) показаны суммы соответствующих частот для значений X и Y.

$$m_{1*} = m_{11} + m_{12} + \ldots + m_{1l}$$

$$m_{*1} = m_{11} + m_{21} + \ldots + m_{k1}$$

Сумма всех возможных m_{ij} равна m и сумме частот по строкам и столбцам

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

Корреляционная таблица

Каждому числу x_i соответствует целый набор значений $y_1, y_2, \ldots, y_l$ с конкретными частотами $m_{i1}, m_{i2}, \ldots, m_{il}$

Среднее этих значений обозначается $\bar{y}_x$

(условное среднее значение y при условии, что $X=x_i$)

И находится по формуле:

$$\bar{y}_x = \bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j \cdot m_{*j} \qquad \bar{x}_y = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot m_{i*}$$

Условные средние значения У

x	x_1	x2	…	x_k
$\bar{y}_x$	$\bar{y}_{x1}$	$\bar{y}_{x2}$	…	$\bar{y}_{xk}$
m_x	m_{x1}	m_{x2}	…	m_{xk}

Пример: Соотношения роста (X) и массы тела (У)

Y /X	170	172	174	176	178	180	182	my
65	8	4	-	2	-	-	-	14
70	15	19	11	5	-	1	-	51
75	7	10	16	11	3	-	-	47
80	2	8	12	3	1	1	2	29
85	-	3	2	-	5	4	5	19
mx	32	44	41	21	9	6	7	160

Решение

Выборочный коэффициент корреляции в случае сгруппированных данных по корреляционной таблице вычисляется следующим образом:

$$r = \frac{\sum m_{xy}(x - \bar{x})(y - \bar{y})}{\sqrt{\sum m_x (x - \bar{x})^2 \sum m_y (y - \bar{y})^2}},$$

или

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}, \qquad \overline{xy} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} x_i y_j m_{ij}$$

$$s_x^2 = \overline{x^2} - (\bar{x})^2, \qquad s_y^2 = \overline{y^2} - (\bar{y})^2$$

Решение

- Суммирование распространяется в знаменателе на все возможные x или y,
- в числителе - на все возможные пары (x,y).

Упростим выражение в числителе

$$r = \frac{\sum m_{xy} xy - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum m_x (x - \bar{x})^2 \sum m_y (y - \bar{y})^2}} =$$

$$= \frac{2075700 - 160 * 173,7 \cdot 74,6}{\sqrt{1566,9 \cdot 5277,5}} = 0,547,$$

$$\bar{x} = 173,7, \qquad \bar{y} = 74,6$$

Корреляционный анализ

Точечные оценки параметров двумерной корреляционной модели

Генераль-ные характер.	Их оценки (выборочные характеристики)	
	n мало (данные не сгруппированы)	**n** велико (данные сгруппированы)
μ_x	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{l} x_i \cdot m_{ix}$
μ_y	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j \cdot m_{jy}$
$M(x^2)$	$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$	$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{l} x_i^2 \cdot m_{ix}$
$M(y^2)$	$\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$	$\overline{y^2} = \frac{1}{n}\sum_{j=1}^{l} y_j^2 \cdot m_{jy}$
$M(xy)$	$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$	$\overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} x_i \cdot y_j \cdot m_{ij}$

Проверка независимости (значимости) признаков

Значимость парных коэффициентов корреляции можно проверить 2 способами: **1. С помощью t-критерия Стьюдента**.

Нулевая гипотеза $H_0: \rho_{xy} = 0$

Альтернативная гипотеза $H_1: \rho_{xy} \neq 0$

1. Вычисление наблюдаемого значения критерия $\mathbf{t_н}$:

$$\mathbf{t_н = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}}$$

где r – выборочная оценка парного коэффициента корреляции;

2. Нахождение критического значения $\mathbf{t_{кр}(\alpha, \nu=n-2)}$ по таб. 2

3. Вывод по гипотезе Рассчитанное значение $\mathbf{t_н}$ сравнивается с $\mathbf{t_{кр}}$: Если $\mathbf{|t_н| > t_{кр}}$ => гипотеза H_0 отвергается => ρ - значим

Корреляционный анализ

Iспособ. С использованием критерия Фишера-Иейтса

1. За $r_н$ принимается выборочное значение коэффициента корреляции **r**
2. $r_{кр}$ (α, **v=n-2**) находится по таб. Фишера-Иейтса (таб.8)
3. Вывод по гипотезе Рассчитанное значение ***r*** сравнивается с $r_{кр}$:

Если $|r| > r_{кр}$ => гипотеза H_0 отвергается =>

ρ – значим значим (с вероятностью ошибки α)

v	Двусторонние границы				v	Двусторонние границы			
	0,05	0,02	0,01	0,001		0,05	0,02	0,01	0,001
1	0,997	1,000	1,000	1,000	16	0,468	0,543	0,590	0,708
2	0,950	0,980	0,990	0,999	17	0,456	0,529	0,575	0,693
3	0,878	0,934	0,959	0,991	18	0,444	0,516	0,561	0,679
4	0,811	0,882	0,917	0,974	19	0,433	0,503	0,549	0,665
5	0,754	0,833	0,875	0,951	20	0,423	0,492	0,537	0,652
6	0,707	0,789	0,834	0,925	25	0,381	0,445	0,487	0,597
7	0,666	0,750	0,798	0,898	30	0,349	0,409	0,449	0,554
8	0,632	0,715	0,765	0,872	35	0,325	0,381	0,418	0,519
9	0,602	0,685	0,735	0,847	40	0,304	0,358	0,393	0,490
10	0,576	0,658	0,708	0,823	45	0,288	0,338	0,372	0,465

Интервальные оценки параметров связи

Для значимых параметров связи (парных и частных коэффициентов корреляции находят интервальные оценки с надежностью γ .

1. **Нахождение интервальной оценки для вспомогательной статистики Z с помощью Z-преобразования Фишера**

$$Z' - t_\gamma \sqrt{\frac{1}{n-l-3}} \le Z \le Z' + t_\gamma \sqrt{\frac{1}{n-l-3}}$$

t_γ вычисляют по таблице интегральной функции Лапласа (табл. 1) из условия $\Phi(t_\gamma)=\gamma$

- Значение Z' (Z_r)определяют по таблице Z - преобразования (табл. 6) по найденному значению r.
- Функция Z_r нечетная:
  $Z'(-r) = -Z'(r)$ нечетная

$$r \xrightarrow{\text{таб.6}} z_r$$

$$\gamma = \Phi(t) \xrightarrow{\text{таб.1}} t_\gamma \rightarrow \Delta z = \frac{t_\gamma}{\sqrt{n-3}}$$

Интервальные оценки параметров связи

2. Обратный переход от *Z* к r

осуществляют также по таблице *Z* – преобразования.

$$z_{min} = z_r - \Delta z \xrightarrow{\text{таб.6}} \rho_{min}$$

$$z_{max} = z_r + \Delta z \xrightarrow{\text{таб.6}} \rho_{max}$$

3. Получение интервальной оценки для r с надежностью γ :

$$r_{min} \le \rho \le r_{max}$$

Таким образом, с вероятностью γ гарантируется, что генеральный коэффициент корреляции ρ будет находиться в интервале от r_{min} до r_{max}.

С помощью доверительного интервала можно проверить ***значимость ρ: если ноль попадает в доверительный интервал, то коэффициент корреляции не значимый.***

Корреляционный анализ

Генеральная совокупность	Выборочная совокупность
μ - математическое ожидание	$\overline{x}$ - выборочное среднее
σ^2 - дисперсия	s^2 - выборочная дисперсия
Σ - среднее квадратическое отклонение	s - выборочное ср. квадр. отклонение
P - вероятность	$\frac{m}{n}$ - частость
ρ - коэффициент корреляции	r - выборочный коэффициент корреляции
β - коэффициент регрессии	b - выборочный коэффициент регрессии

Коэффициент детерминации

Квадрат парного коэффициента корреляции (для двумерного случая) называется **множественным коэффициентом детерминации** .

Он характеризует долю дисперсии одной переменной (результативной), обусловленной влиянием всех остальных переменных (аргументов), входящих в модель.

Матрица парных коэффициентов корреляции (многомерный случай)

$$R = \begin{pmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & r_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & . & . & 1 \end{pmatrix}$$

Матрица R является симметричной и положительно определенной, на главной диагонали стоят единицы.

Корреляционный анализ

В двумерном корреляционном анализе обычно строят

- корреляционную таблицу,
- поле корреляции,
- рассчитывают точечные оценки параметров корреляционной модели,
- проверяют значимость параметров связи
- для значимых параметров строят интервальные оценки.

Имея оценки параметров модели $\hat{x},\ \hat{y},\ s_x,\ s_y,\ r$ можно рассчитать оценки уравнений регрессии.

Корреляционный анализ

При небольших объемах выборки часто используют более предпочтительные оценки коэффициентов корреляции и детерминации, чем выборочные коэффициенты:

- **более предпочтительная оценка коэффициента корреляции** –

$$\tilde{r}^2 = r(1 + \frac{1 - r^2}{2 \cdot (n - 4)}),$$

- **более предпочтительная оценка коэффициента детерминации**

$$\tilde{r}^2 = \frac{(n-1) \cdot r^2 - 1}{n - 2}),$$

Корреляционный анализ

Уравнения линий регрессии

Если наблюдаемые значения У и Х представляют собой выборку из двумерного нормального распределения, то формально можно рассматривать два уравнения регрессии:

$$\tilde{Y} = MY/x = \mu_y + \beta_{yx} \cdot (x - \mu_x)$$ -прямая регрессии Y на X

$$\tilde{X} = MX/y = \mu_x + \beta_{xy} \cdot (y - \mu_y)$$ -прямая регрессии X на Y

Корреляционный анализ

$$\beta_{yx} = \rho \cdot \frac{\sigma_y}{\sigma_x}$$

$$\beta_{xy} = \rho \cdot \frac{\sigma_x}{\sigma_y}$$

βyx - генеральный коэффициент регрессии Y на X. Показывает на сколько единиц в среднем изменяется переменная Y при увеличении переменной X на единицу своего измерения

βxy - генеральный коэффициент регрессии X на Y. Показывает на сколько единиц в среднем изменяется переменная X при увеличении переменной Y на единицу своего измерения

Корреляционный анализ

II. Интервальные оценки генеральных коэффициентов корреляции и регрессии

Построение с надёжностью γ доверительных интервалов для генеральных коэффициентов регрессии

Y по X $\beta_{yx\,min} \le \beta_{yx} \le \beta_{yx\,max}$

$$\beta_{yx} \in \left[b_{yx} \pm t_{\alpha} \cdot \frac{S_y \sqrt{1-r^2}}{S_x \sqrt{n-2}} \right]$$

и X по Y $\beta_{xy\,min} \le \beta_{xy} \le \beta_{xy\,max}$

$$\beta_{xy} \in \left[b_{xy} \pm t_{\alpha} \cdot \frac{S_x \sqrt{1-r^2}}{S_y \sqrt{n-2}} \right]$$

t_α определяется по таб.2 (распределение Стьюдента) для уровня значимости $\alpha=1-\gamma$ и числа степеней свободы $v=n-2$

При $n \to \infty$ ($n>30$) t определяется по таб.1 для $\gamma=\Phi(t)$

Двумерная корреляционная модель

Остаточная дисперсия

Выборочная дисперсия переменной Y может быть представлена:

$$S_y^2 = S_y^2 \cdot r^2 + S_y^2 \cdot (1 - r^2)$$

S^2_r — выборочная дисперсия регрессии Y по X, объясняемая вариацией переменной X

$S^2_{y/x}$ — остаточная дисперсия, объясняемая неучтёнными в модели факторами

Остаточная (условная) дисперсия:

$S^2_{y/x} = S^2_y \cdot (1 - r^2)$ – регрессии Y по X

Корреляционный анализ

Точечные оценки параметров двумерной корреляционной модели

Генеральные характеристики	Их оценки (выборочные характеристики)			
σ_x^2, σ_y^2	$S_x^2 = \overline{x^2} - (\overline{x})^2$	$S_y^2 = \overline{y^2} - (\overline{y})^2$		
ρ	**Выборочный коэффициент корреляции**	$r = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y}$	**Выборочные коэффициенты регрессии**	
β_{yx}, β_{xy}	$b_{yx} = r \cdot \dfrac{S_y}{S_x}$	$b_{xy} = r \cdot \dfrac{S_x}{S_y}$	$b_{xy} \cdot b_{yx} = r^2$	$\dfrac{b_{yx}}{b_{xy}} = \dfrac{S_y^2}{S_x^2}$
$\tilde{Y}$=MY/x	$y = \overline{y}/x = \overline{y} + b_{yx} \cdot (x - \overline{x})$		**Оценки уравнений регрессии**	
$\tilde{X}$=MX/y	$x = \overline{x}/y = \overline{x} + b_{xy} \cdot (y - \overline{y})$			

Трёхмерная корреляционная модель

условные дисперсии

$$R^2_{x/yz} = \frac{\sigma^2_x - \sigma^2_{x/yz}}{\sigma^2_x}$$

$$r^2_{x/yz} = \frac{s^2_x - s^2_{x/yz}}{s^2_x} = 1 - \frac{s^2_{x/yz}}{s^2_x}$$

Трёхмерная корреляционная модель

Множественный коэффициент детерминации

Проверка значимости множественного коэффициента (и корреляции (детерминации), например,

H_0: $\rho_{1/2,3}$ =0, осуществляется с помощью F-критерия.

1. Вычисляется

$$F_{набл} = \frac{\frac{1}{2} \cdot r^2_{1/2,3}}{\frac{1}{n-3} \cdot (1 - r^2_{1/2,3})}$$

- для трехмерного случая

$$F_{набл} = \frac{\frac{1}{\kappa - 1} r^2_{1/2,\ldots,\kappa}}{\frac{1}{n - \kappa} \left(1 - r^2_{1/2,\ldots,\kappa}\right)}$$

- для многомерного случая

Трёхмерная корреляционная модель

Множественный коэффициент детерминации

2. По таблице F-распределения Фишера-Снедекора (таб.4) определяют $F_{кр}$:

   $F_{кр}(\alpha; v_1=2; v_2=n-3)$ – для трехмерной модели

   $F_{кр}(\alpha; v_1=к-1; v_2=n-к)$ – для многомерной модели

3. Если $F_н > F_{кр}$, то гипотеза H_0 отвергается с вероятностью ошибки α и коэффициент детерминации (и соответствующий множественный коэффициент корреляции) считается *значимым*.