ПЕРЕВІРКА НАЯВНОСТІ ЗВ'ЯЗКУ МІЖ ЗМІННИМИ

Випадкові величини можуть бути пов’язані або функціональною залежністю, або статистичною, або бути незалежними.

Функціональний зв'язок

Функціональний зв'язок - кожному значенню змінної *X* поставлене в однозначну відповідність певне значення *Y*.

Статистичний зв'язок

Статистичний зв'язок – зміна однієї з величин приводить до зміни закону розподілу іншої.

Кореляційна залежність

Якщо статистична залежність проявляється в тому, що при зміні однієї з випадкових величин змінюється середнє значення іншої, то таку залежність називають ***кореляційною***

Кореляційний аналіз (КА)

Кореляційний аналіз застосовується, коли змінні вимірюються в шкалах відносин, інтервалів або порядку, тобто **мають числову природу** .

Кореляційний аналіз - статистичний метод, що дозволяє визначити, чи існує ***лінійна*** залежність між змінними і на скільки вона сильна.

Приклади

- **1. Менеджер** цікавиться, чи залежить обсяг продажів у цьому місяці від обсягу реклами в цьому ж періоді?
- **2. Викладач** прагне з'ясувати, чи існує залежність між кількістю годин, витрачених студентом на заняття, і результатами іспиту?
- **3. Лікар** досліджує, чи існує зв'язок між віком людини і його кров'яним тиском?
- **4. Соціолог** досліджує, який зв'язок між рівнем злочинності й рівнем безробіття в регіоні. Чи пов’язані дохід від професійної діяльності з тривалістю освіти

Коваріація

Характеристикою залежності між випадковими величинами *X* і *Y* служить коефіцієнт ***коваріації.***

$$\operatorname{cov}(x, y) = M[(X - MX)(Y - MY)]$$

Коваріація

Оцінкою коефіцієнта коваріації є ***вибірковий коефіцієнт коваріації:***

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\operatorname{cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}$$

Якщо при більших значеннях ***X*** більше ймовірні більші значення ***Y***, а при малих значеннях ***X*** більше ймовірні малі значення ***Y***, то в (1) додатні доданки домінують і ***cov(x,y)>0***. У цьому випадку говорять про ***прямий зв'язок***: із зростанням ***X*** випадкова величина ***Y*** має тенденцію до зростання.

Якщо ж більш ймовірні доданки $(x_i - \overline{x})(y_i - \overline{y})$ із співмножників різних знаків, то $\operatorname{cov}(x, y) < 0$, тобто буде мати місце ***зворотний зв'язок***, із зростанням X випадкова величина Y зменшується.

Якщо $\operatorname{cov}(x, y) \approx 0$, то додатні і від'ємні доданки «гасять» один одного, і зв'язок між X і Y не спостерігається.

Якщо X і Y ***незалежні***, то $\operatorname{cov}(x, y) = 0$.

Зворотного висновку зробити ***не можна!*** Випадкові величини можуть бути пов'язані функціональною залежністю, але коефіцієнт $\operatorname{cov}(x, y) = 0$.

Величина cov(x,y) – залежить від одиниць вимірювання, тому її незручно використовувати за показник зв'язку. У зв'язку з цим вводять коефіцієнт парної кореляції

Коефіцієнт парної кореляції використовують для вимірювання сили лінійних зв'язків різних пар ознак з їх множини.

Вибірковий коефіцієнт парної кореляції обчислюється за формулою

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

де

x_i, y_i – спостережувані значення;

$\bar{x}$, $\bar{y}$ – відповідні вибіркові середні значення для X та Y;

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

вибіркові середні квадратичні відхилення для X та Y; n – обсяг вибірки.

σ_x, σ_y – вибіркові середні квадратичні відхилення (**не виправлені!!**)

Функція в *EXCEL* ***СТАНДОТКЛОНП***(<діапазон>).

Функції в *EXCEL:*

Функції в *EXCEL:*

***КОВАР*(*масив1*; *масив2*)** – повертає коваріацію, тобто середнє добутків відхилень для кожної пари точок даних.

КОРРЕЛ*(*масив1*;*масив2*)** – повертає парний коефіцієнт кореляції. Можна також скористатися ***Сервис – Анализ данных – Корреляция

Властивості коефіцієнта кореляції

1 Коефіцієнт кореляції приймає значення з відрізку *[-1;1]* $-1 \le r_{xy} \le 1$.

2 Якщо $r_{xy} = \pm 1$, випадкові величини X і Y пов'язані лінійною залежністю і цей зв'язок є функціональним.

Властивості коефіцієнта кореляції

3 Якщо $r_{xy} > 0$, то між змінними існує прямий лінійний зв'язок, значення змінних збільшуються або зменшуються одночасно.

Якщо $r_{xy} < 0$, то між змінними зворотній лінійний зв'язок, і при збільшенні однієї змінної інша зменшується.

Якщо коефіцієнт вибіркової кореляції за модулем наближається до 1, це означає, що між випадковими величинами X і Y існує ***лінійний статистичний зв'язок.***

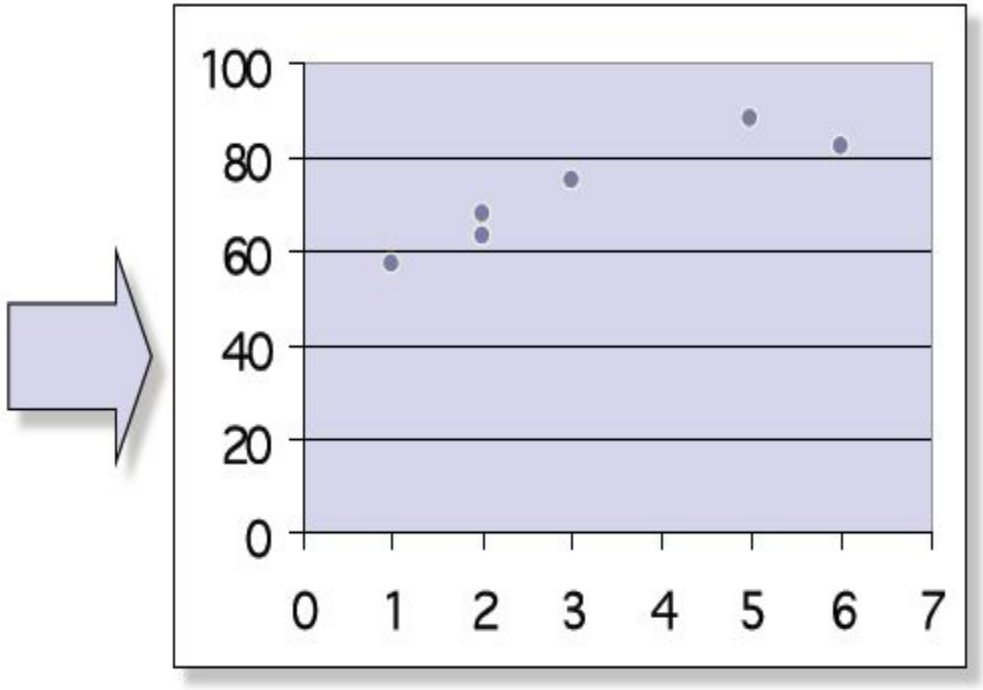
4 $r_{xy} = r_{yx}$.

Властивості коефіцієнта кореляції

Коефіцієнт кореляції показує тісноту тільки ***лінійного*** зв'язку, для більш складних залежностей (квадратичних, кубічних та ін.) коефіцієнт кореляції може показувати відсутність зв'язку. Для досліджень більш складних залежностей використовують регресійний аналіз

Рассматриваем две переменные: «продолжительность занятий» студентов перед экзаменом и «итоговая оценка» (из 100 балов). Пытаемся визуально определить связь. Правда ли, что **чем меньше времени занятий, тем выше оценка?**

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75



Визуально видно, что имеет место линейная зависимость, которая отрицательна. Это означает, что увеличение переменной *x* приводит к уменьшению второй переменной *y*.

Студент	Пропустил *x*	Оценка *y*
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

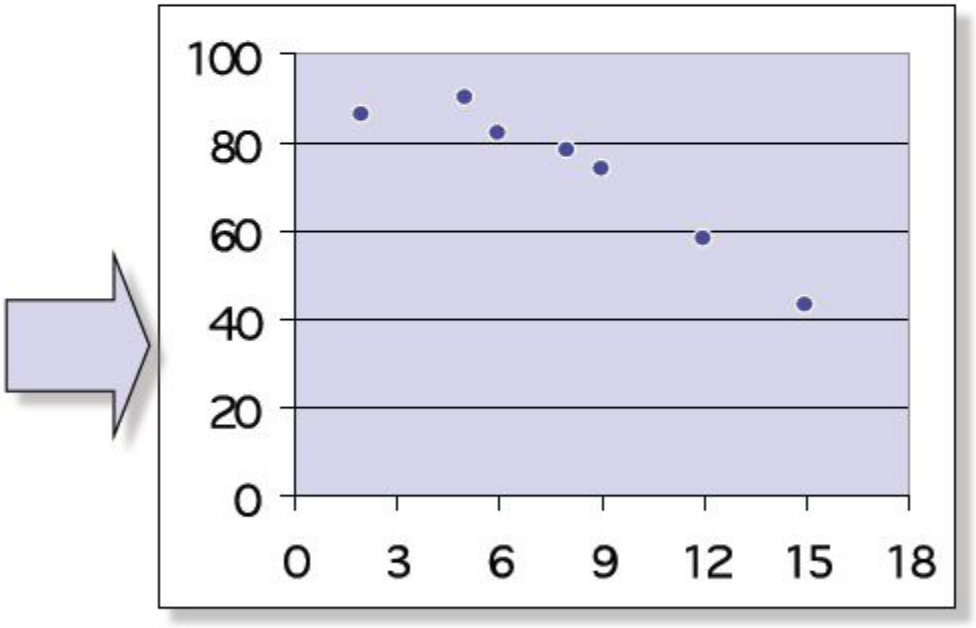


График показывает, что имеется зависимость, которая не является линейной. Возможно, эта зависимость квадратичная или какая-то иная.

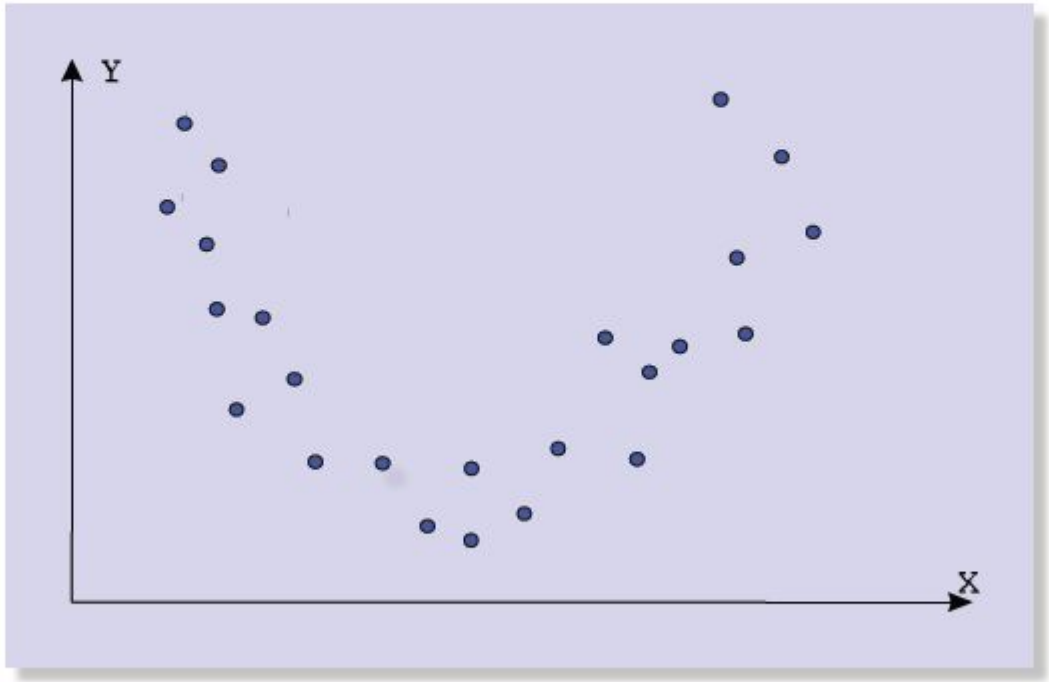
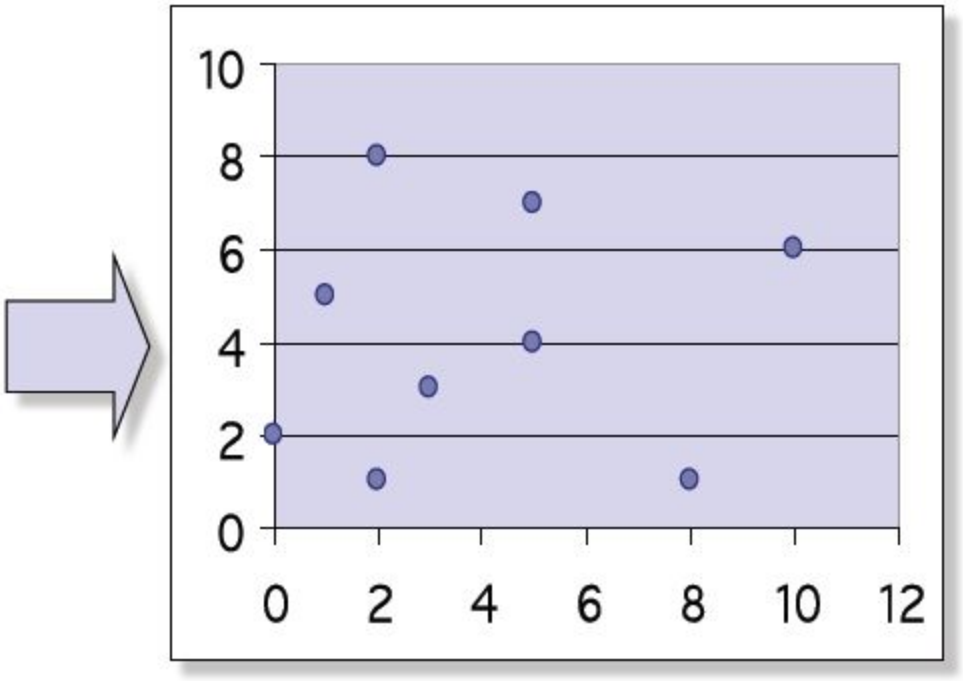


График сообщает нам об отсутствии зависимости времени на подготовку к экзамену и количества вопросов, заданных преподавателем на экзамене.

Студент	Часы *x*	Вопросы *y*
A	3	3
B	0	2
C	2	1
D	5	7
E	8	1
F	5	4
G	10	6
H	2	8
I	1	5

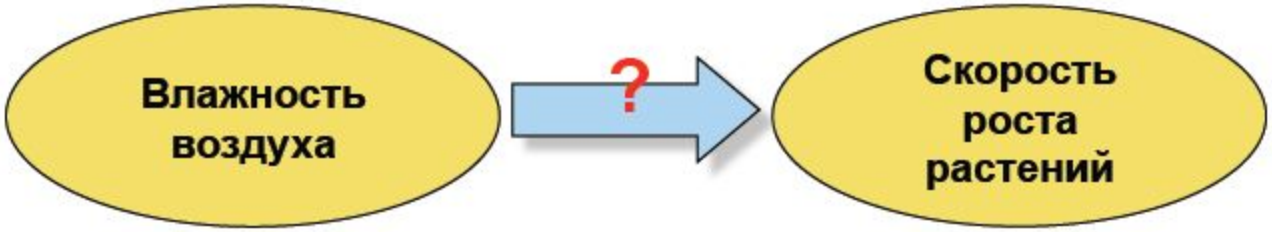


П'ять видів зв'язку між змінними

1. Прямий причинно-наслідковий зв'язок
2. Зворотній причинно-наслідковий зв'язок
3. Зв'язок викликаний третьою (прихованої) змінною.
4. Взаємозв'язок викликаний кількома прихованими змінними
5. Зв'язку немає, спостережувана залежність випадкова

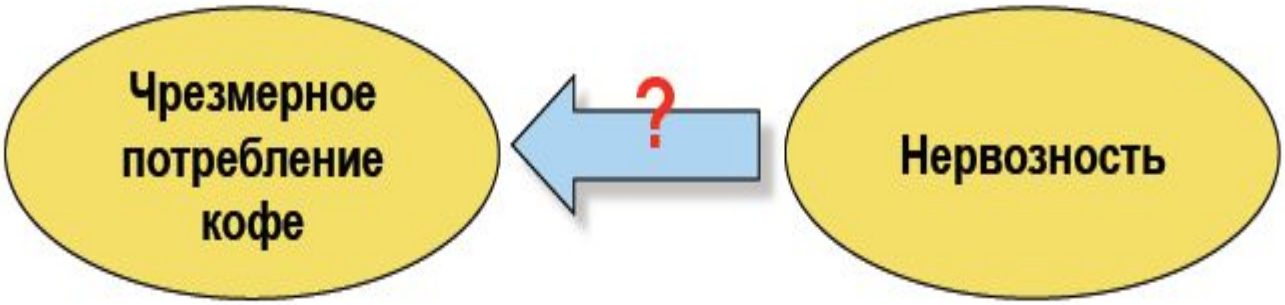
Прямая причинно-следственная связь между переменными (переменная x определяет значение переменной y).

Наличие воды ускоряет рост растений. Яд вызывает смерть.
Температура воздуха прямо влияет на скорость таяния льда.



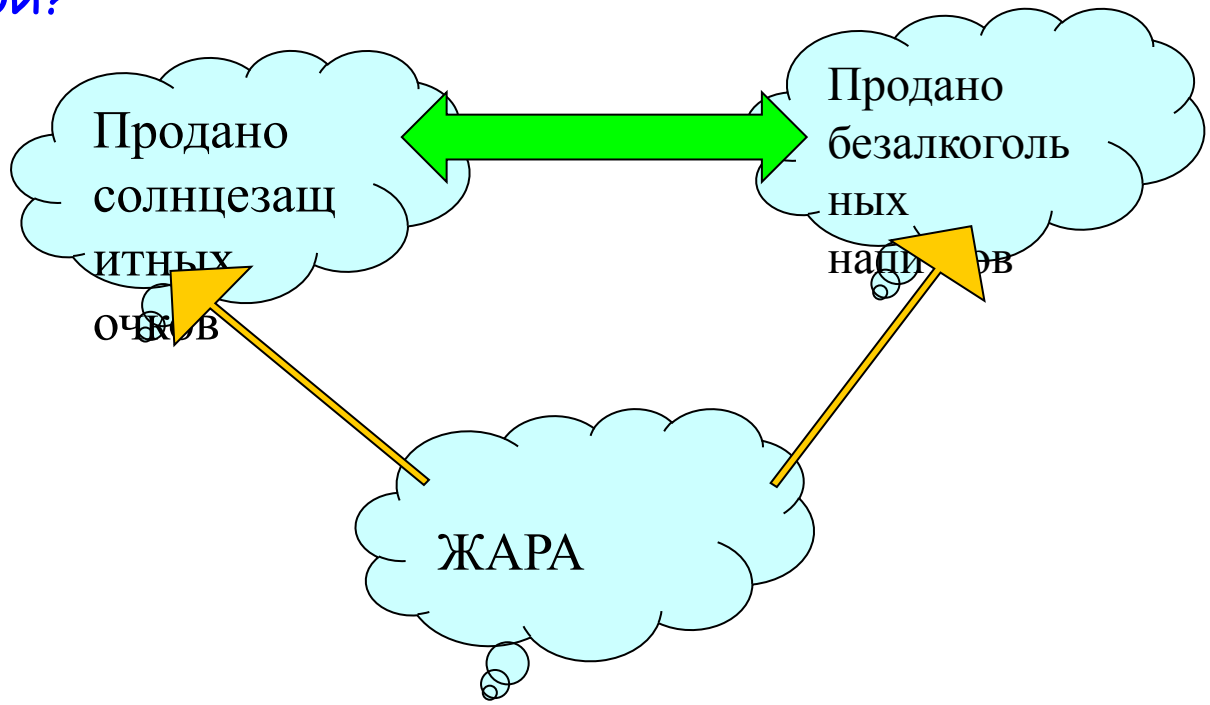
Обратная причинно-следственная связь между переменными (переменная у определяет значение переменной х).

Исследователь может думать, что чрезмерное потребление кофе вызывает нервозность. Но, может быть, очень нервный человек выпивает кофе, чтобы успокоить свои нервы?



Связь между переменными может быть вызвана третьей переменной

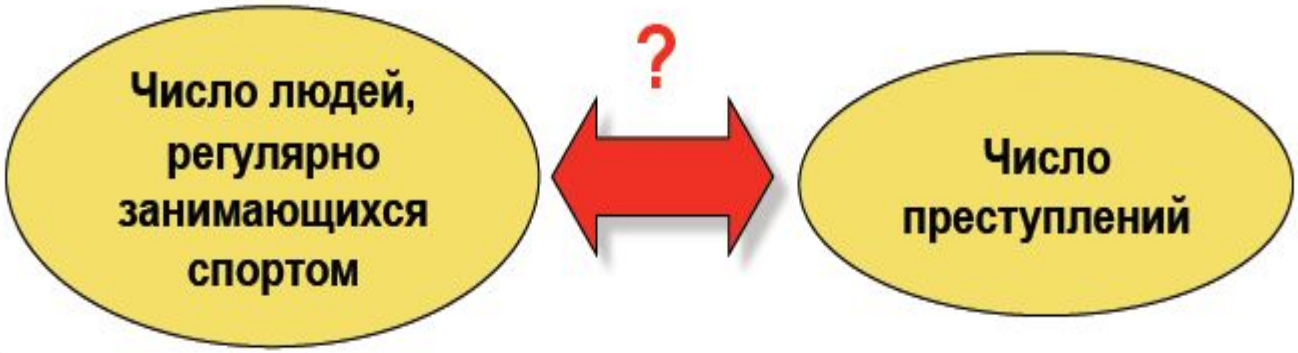
- Исследователь установил, сто существует некая зависимость между числом проданных солнцезащитных очков и числом выпитых безалкогольных напитков в летнее время. А может быть обе переменные связаны с жарой?



5. Зависимость случайна

Исследователь может найти значимую зависимость между увеличением количества людей, которые занимаются спортом и увеличением количества людей, которые совершают преступления.

Но здравый смысл говорит, что любая связь между этими двумя переменными должна быть случайной.



Значущість коефіцієнта кореляції

Щоб перевірити, чи значуще коефіцієнт кореляції відрізняється від 0, використовують критеріальне значення

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}},$$

яке є розподілом Стьюдента з k=N-2 ступенями вільності. При заданому рівні значущості α критичне значення $t_{кр}$ знаходять із рівняння $P(|t|>t_{кр}) = \alpha$.

Значущість коефіцієнта кореляціїіластивості

Якщо $|t| < t_{кр}(\alpha; N-2)$, то r_{xy} не значуще відрізняється від 0, і приймають гіпотезу про відсутність лінійного кореляційного зв'язку між змінними.

Якщо $|t| > t_{кр}(\alpha; N-2)$, то r_{xy} значуще відрізняється від 0, і приймають гіпотезу про наявність лінійного кореляційного зв'язку між змінними.

Значущість коефіцієнта кореляції

Залежно від значення r_{xy} розрізняють такі види зв'язку:

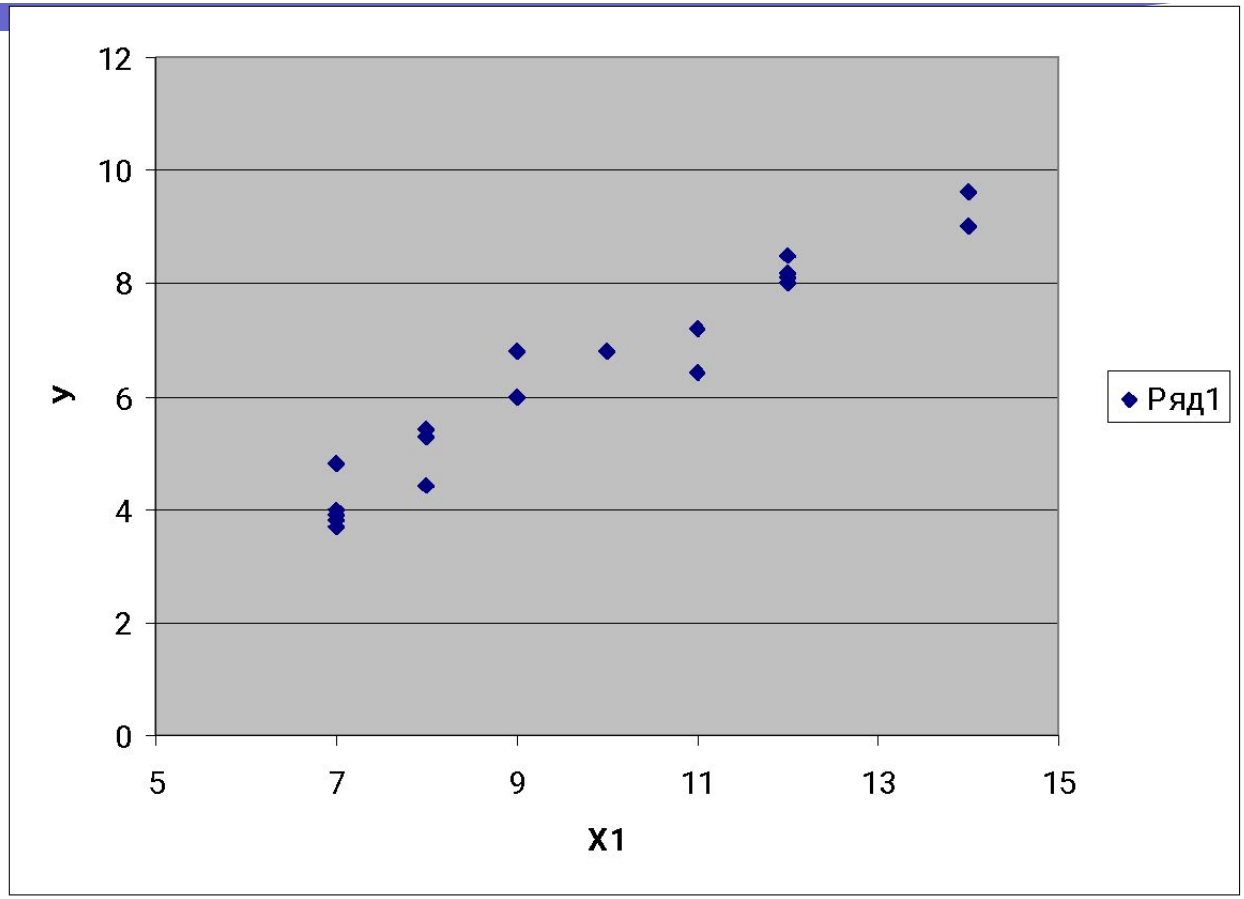
0 – 0,3 – слабкий зв'язок;

0,3 – 0,7 – середній зв'язок;

0,7 – 1 – сильний зв'язок

Приклад.

По 20 підприємствам регіону вивчається залежність виробітку продукції на одного працівника у (тис.грн.) від введення в дію нових основних фондів x1 (% від вартості фондів на кінець року)


12
10
8
6
4
2
0
y
5
7
9
11
13
15
X1
Ряд1

Номер предприятия	y	X_1	y-ysr	x-xsr	(x-xsr)(y-ysr)
1	7	3,9	-2,6	-2,29	5,954
2	7	3,9	-2,6	-2,29	5,954
3	7	3,7	-2,6	-2,49	6,474
4	7	4	-2,6	-2,19	5,694
5	7	3,8	-2,6	-2,39	6,214
6	7	4,8	-2,6	-1,39	3,614
7	8	5,4	-1,6	-0,79	1,264
8	8	4,4	-1,6	-1,79	2,864
9	8	5,3	-1,6	-0,89	1,424
10	10	6,8	0,4	0,61	0,244
11	9	6	-0,6	-0,19	0,114
12	11	6,4	1,4	0,21	0,294
13	9	6,8	-0,6	0,61	-0,366
14	11	7,2	1,4	1,01	1,414
15	12	8	2,4	1,81	4,344
16	12	8,2	2,4	2,01	4,824
17	12	8,1	2,4	1,91	4,584
18	12	8,5	2,4	2,31	5,544
19	14	9,6	4,4	3,41	15,004
20	14	9	4,4	2,81	12,364
среднее	9,6	6,19			4,391
σ=	2,396	1,890			

Приклад

,

$$r_{xy} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = 0{,}970$$

$$r_{xy} = КОРРЕЛ(<діапазон>) = 0{,}970.$$

Приклад

Перевіряємо значущість коефіцієнта кореляції з рівнем значущості $\alpha=0,05$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}},$$

$$t_p = \frac{0,970\sqrt{20-2}}{\sqrt{(1-(0,970)^2}} = 16,9;$$

$t_{кр}$= *СТЬЮДРАСПОБР*(0,05; 20-2)=2,1

$t_{кр} = 2,1;$

Приклад

$/t_p/ > t_{кр}$ – отже, коефіцієнт кореляції є статистично значущим.

Висновок: між виробітком продукції на одного робітника *y* та введенням в дію нових основних фондів *x1* існує прямий сильний кореляційний зв'язок.

.

Поняття про багатовимірний кореляційний аналіз

Нехай є багатовимірна нормальна сукупність із *m* ознаками $X_1, X_2, \ldots, X_m$... У цьому випадку взаємозалежність між ознаками можна описати за допомогою кореляційної матриці. Під кореляційною матрицею будемо розуміти матрицю, складену з парних коефіцієнтів кореляції. Оцінкою парного коефіцієнта кореляції є вибірковий парний коефіцієнт кореляції

Кореляційна матриця

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{n\sigma_{x_j}\sigma_{x_k}}, \quad (j,k = 1,2,..m).$$

Якщо знайдені вибіркові коефіцієнти кореляції, то можна одержати оцінену кореляційну матрицю

$$Q = \begin{bmatrix} q_{11} & q_{12} & \ldots . q_{1m} \\ \ldots & \ldots & \ldots \\ q_{m1} & q_{m2} & \ldots . q_{mm} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \ldots . & r_{1m} \\ \ldots & \ldots . & \ldots \\ r_{m1} & r_{m2} \ldots & 1 \end{bmatrix}. \quad (2)$$

Ця матриця симетрична.

Часткова кореляція

Часткові коефіцієнти кореляції характеризують тісноту зв'язку між результатом і відповідним фактором ***при усуненні впливу інших факторів***

Часткова кореляція

Формула вибіркового часткового коефіцієнта має вигляд

$$r_{jk,*} = -\frac{Q_{jk}}{\sqrt{Q_{jj}Q_{kk}}},$$

де Q_{jk}, Q_{jj}, Q_{kk} – алгебраїчні доповнення до відповідних елементів кореляційної матриці Q (2).

Часткова кореляція

Для випадку, коли на змінну Y діють 2 фактори X_1 і X_2, часткові коефіцієнти кореляції можна обчислити за формулами:

$$r_{Y1,2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1-r^2_{Y2}}\sqrt{1-r^2_{12}}},$$

$$r_{Y2,1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1-r^2_{Y1}}\sqrt{1-r^2_{21}}},$$

$$r_{12,Y} = \frac{r_{12} - r_{Y1}r_{Y2}}{\sqrt{1-r^2_{Y1}}\sqrt{1-r^2_{Y2}}}.$$

Часткова кореляція

Формулу вибіркового часткового коефіцієнта кореляції можна виразити через елементи матриці Z, оберненої до матриці Q.

$$r_{ij,*} = -\frac{Q_{ij}/|Q|}{\sqrt{Q_{ii}/|Q|}\sqrt{Q_{jj}/|Q|}} = -\frac{z_{ij}}{\sqrt{z_{ii}z_{jj}}},$$

де z_{ij}, z_{ii}, z_{jj} - елементи матриці $Z=Q^{-1}$.

Множинний коефіцієнт кореляції

Множинний коефіцієнт кореляції дозволяє оцінити зв'язок однієї з ознак з усіма іншими.

$$R_{j,*} = \sqrt{1 - \frac{|Q|}{Q_{jj}}} = \sqrt{1 - \frac{1}{z_{jj}}},$$

де $|Q|$ – визначник кореляційної матриці,

Q_{jj} – алгебраїчне доповнення до відповідного елемента кореляційної матриці,

z_{jj} – елементи матриці $Z = Q^{-1}$.

Квадрат коефіцієнта множинної кореляції називають ***множинним коефіцієнтом детермінації***. Коефіцієнти множинної кореляції й детермінації - величини додатні і набувають значення з відрізку [0;1]. Чим ближче їх значення до 1, тим тісніший зв'язок результативної ознаки з усім набором досліджуваних факторів

Приклад.

За даними річних звітів десяти (n=10) підприємств провести аналіз залежності собівартості товарної продукції *Y* (ум. од.) від обсягу валової продукції *X1* (млн. ум. од.) і продуктивності праці *X2* (тис. ум. од. на чол.)

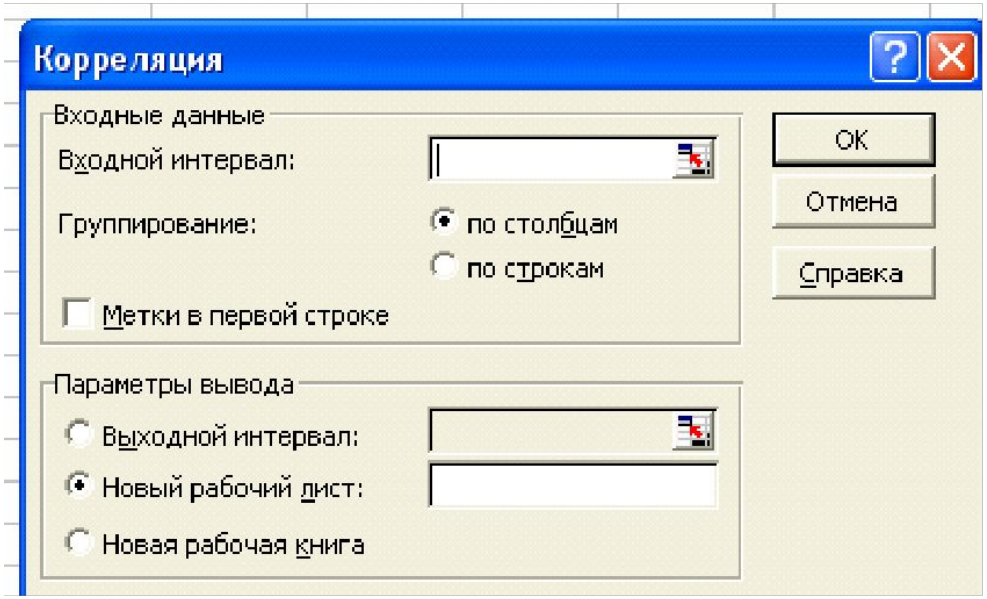
Приклад

– Собівартість товарної продукції, обсяг валової продукції і продуктивність праці

№	X_1, (млн. ум. од.)	X_2, (тис. ум. од. на чол.).	Y, (ум. од.)	№	X_1, (млн. ум. од.)	X_2, (тис. ум. од. на чол.).	Y, (ум. од.)
1	3	1,8	2,1	6	5	1,5	4,9
2	4	1,5	2,8	7	6	1,6	5,5
3	5	1,4	3,2	8	7	1,2	6,5
4	5	1,3	4,5	9	15	1,3	12,1
5	5	1,3	4,8	10	20	1,2	15

Приклад

1 Знаходимо матрицю парних коефіцієнтів кореляції за допомогою ***Сервис – Анализ данных – Корреляция.***



Приклад

Натисніть кнопку ***ОК***. Отримаємо таку таблицю:

	Столбец 1	*Столбец 2*	*Столбец 3*
Столбец 1	1		
Столбец 2	-0,565031	1	
Столбец 3	0,9872039	-0,6049951	1

Приклад

Будуємо кореляційну матрицю Q

1	-0,565031	0,9872039
-0,56503	1	-0,6049951
0,987204	-0,604995	1

Приклад

Розрахунок оцінок часткових коефіцієнтів кореляції.

Знаходимо обернену матрицю Z за допомогою функції EXCEL ***МОБР(<діапазон>).***

42,03319	-2,13634	-42,7878
-2,13634	1,685914	3,128971
-42,7878	3,128971	45,1333

Приклад

$$r_{ij,*} = -\frac{z_{ij}}{\sqrt{z_{ii} z_{jj}}}$$

Одержуємо такі значення:

$r_{12,y}=$	0,253779237
$r_{1y,2}=$	0,982370424
$r_{2y,1}=$	-0,35870321

Зв'язок обсягу валової продукції (*X1*) і собівартості товарної продукції (*Y*): *ry1*=0,987, *ry1,2*=0,982 сильний.

У даному прикладі *r12,y=0,25378*, а *r12=-0,565*, тобто чистий зв'язок між обсягом валової продукції (*X1*) і продуктивністю праці (*X2*) незначний

Приклад

Розрахунок оцінок множинних коефіцієнтів кореляції й детермінації. Оцінки множинних коефіцієнтів кореляції та детермінації розраховуються за формулами:

$$R_{y,12} = \sqrt{1 - \frac{1}{z_{33}}} = 0{,}989,$$

$$R^2_{y,12} = 1 - \frac{1}{z_{33}} = 0{,}978.$$

Приклад

Значення множинних коефіцієнтів кореляції й детермінації близькі до 1, що свідчить про наявність сильної лінійної залежності *Y* від *X1* і *X2*, тобто собівартість дійсно залежить від обсягу валової продукції і продуктивності праці.