Лекция 2

Корреляционный анализ

Корреляционный анализ

Корреляция — статистическая взаимосвязь двух или более случайных величин. При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Математической мерой корреляции двух случайных величин служит коэффициент корреляции.

Корреляционный анализ

Корреляция может указывать на некоторую взаимосвязь между переменными, которую можно с успехом применять на практике.

Например, электростанция может генерировать меньше электроэнергии в погожий день из-за корреляции между спросом на электроэнергию и погодой.

Корреляционный анализ

В этом примере существует причинно-следственная связь, поскольку экстремальные погодные условия заставляют людей использовать больше электроэнергии для обогрева или охлаждения.

Однако в целом наличия корреляции недостаточно, чтобы сделать вывод о наличии причинно-следственной связи.

Корреляционный анализ

В этом примере существует причинно-следственная связь, поскольку экстремальные погодные условия заставляют людей использовать больше электроэнергии для обогрева или охлаждения.

Однако в целом наличия корреляции недостаточно, чтобы сделать вывод о наличии причинно-следственной связи.

Ковариация — в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин.

Пусть X, Y — две случайные величины, определённые на одном и том же вероятностном пространстве. Тогда их ковариация определяется следующим образом:

$$\sigma_{XY} = cov(X, Y) = \mathbf{M}[(X - \mathbf{M}X)(Y - \mathbf{M}Y)],$$

где $\mathbf{M}$ — математическое ожидание.

Корреляционный анализ

При анализе выборок используют выборочный коэффициент ковариации, определяемый как :

$$\bar{s}_{XY} = cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}),$$

где N — объем выборок, $\bar{X}, \bar{Y}$ - выборочные средние.

Используя свойство линейности математических ожиданий, можно записать

$$\sigma_{XY} = cov(X, Y) = \mathbf{M}[XY] - \mathbf{M}X \cdot \mathbf{M}Y,$$

Аналогично и для выборочного коэффициента:

$$\bar{s}_{XY} = cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

Свойства ковариации:

- Если X, Y - независимые случайные величины, то

$$cov(X, Y) = 0.$$

Обратное утверждение неверно: из отсутствия ковариации не следует независимость.

- Ковариация случайной величины с собой равна дисперсии:

$$cov(X, X) = \mathbf{D}X.$$

- Ковариация симметрична:

$$cov(X, Y) = cov(Y, X).$$

Свойства ковариации:

- Если a, b - константы, то

$$cov(X, a) = 0.$$
$$cov(aX, bY) = ab \cdot cov(X, Y).$$
$$cov(X + a, Y + b) = cov(X, Y).$$

- Неравенство Коши-Буняковского:

$$cov^2(X, Y) \leq \mathbf{D}X \cdot \mathbf{D}Y.$$

Корреляционный анализ

Если ковариация положительна, то с ростом значений одной случайной величины, значения второй имеют тенденцию возрастать, а если знак отрицательный — то убывать.

Однако только по абсолютному значению ковариации нельзя судить о том, насколько сильно величины взаимосвязаны, так как масштаб ковариации зависит от их дисперсий.

Корреляционный анализ

Ковариация имеет размерность, равную произведению размерности случайных величин, то есть величина ковариации зависит от единиц измерения независимых величин.

Данная особенность ковариации затрудняет её использование в целях корреляционного анализа.

Корреляционный анализ

Значение ковариации можно нормировать, поделив её на произведение среднеквадратических отклонений (квадратных корней из дисперсий) случайных величин.

Полученная величина называется коэффициентом корреляции Пирсона $\mathbf{r}_{XY}$, который всегда находится в интервале от −1 до 1:

$$\mathbf{r}_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}.$$

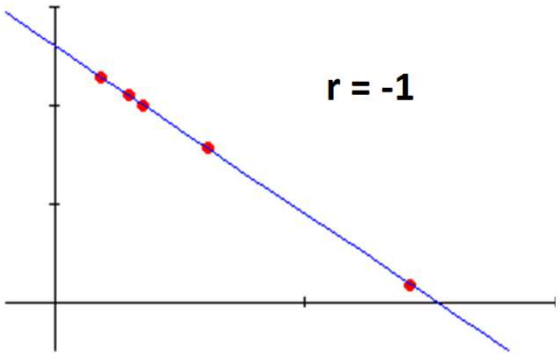
Эквивалентные формулы для выборочного коэффициента корреляции Пирсона:

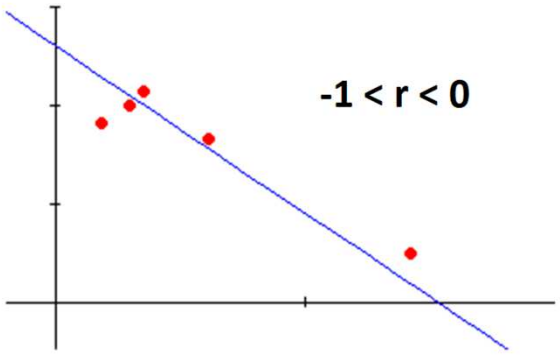
$$\mathbf{r}_{XY} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{(N-1)s_X s_Y} =$$

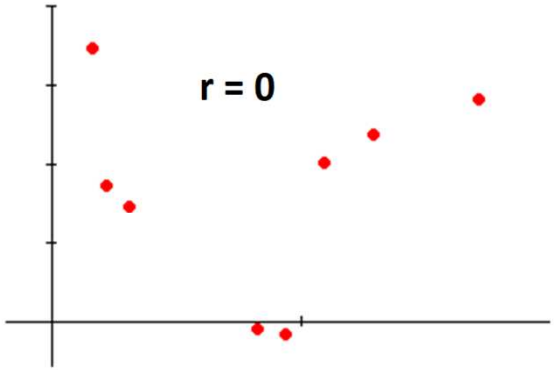
$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right).$$

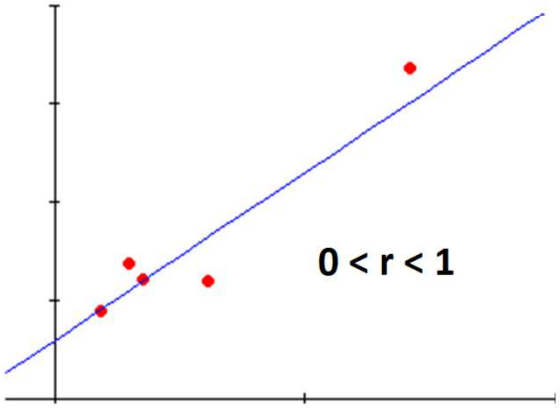
где $s_Z = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2}$ - выборочное среднее квадратическое отклонение.

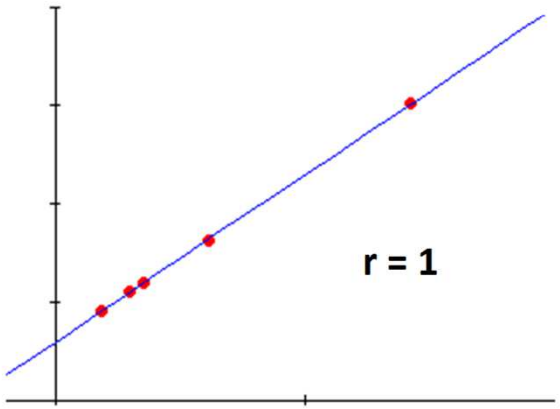
Корреляционный анализ



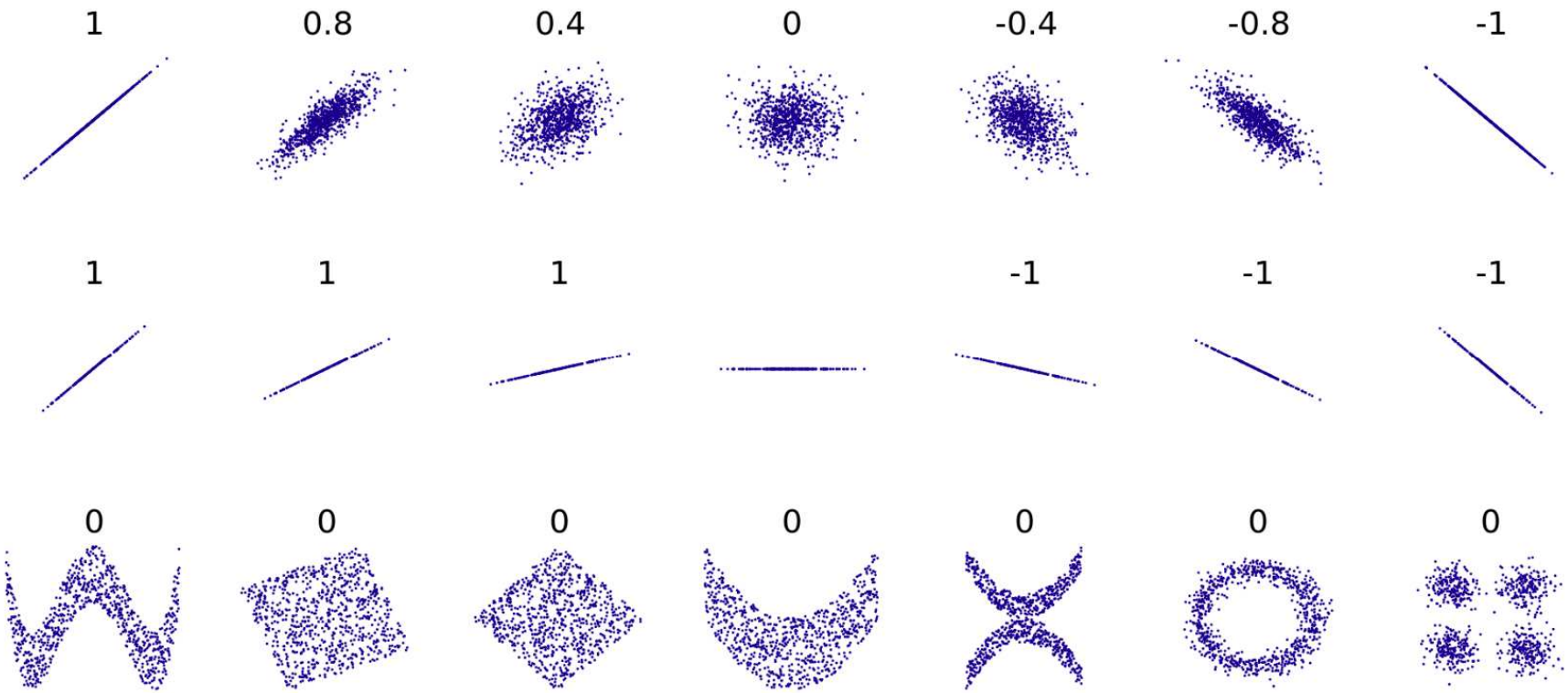








Корреляционный анализ



Пример 1:

$X = \{-2, -0.5, 0, 1, 1.5, 3.5, 4, 4.5\}$,
$Y = \{-3, 0, 0.5, 0, 3, 2.5, 3, 4\}$.

Найти значения ковариации и коэффициента корреляции Пирсона.

Пример 1:

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y});$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}.$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5	✖	1,25		4,28125	6,1563
		=				
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563

1,875 → 4,28125

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	Y	Y-Ycp	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	-3	-4,25	14,875	6
	-0,5	-2	0	-1,25	2,5	0
	0	-1,5	0,5	-0,75	1,125	0
	1	-0,5	0	-1,25	0,625	0
	1,5	0	3	1,75	0	4,5
	3,5	2	2,5	1,25	2,5	8,75
	4	2,5	3	1,75	4,375	12
	4,5	3	4	2,75	8,25	18
Σ	12		10		34,25	49,25
Σ/N	1,5		1,25		4,28125	6,1563
		1,875			4,28125	

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

Пример 1:

$$\mathbf{r}_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{N}(Y_i - \bar{Y})^2}};$$

$$\mathbf{r}_{XY} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{(N-1)s_X s_Y}.$$

$$s_Z = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2} \qquad \sigma_Z = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z})^2}$$

Корреляционный анализ

Пример 1:

	X	X-Xcp	(X-Xcp)^2	Y	Y-Ycp	(Y-Ycp)^2	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	12,25	-3	-4,25	18,063	14,875	6
	-0,5	-2	4	0	-1,25	1,5625	2,5	0
	0	-1,5	2,25	0,5	-0,75	0,5625	1,125	0
	1	-0,5	0,25	0	-1,25	1,5625	0,625	0
	1,5	0	0	3	1,75	3,0625	0	4,5
	3,5	2	4	2,5	1,25	1,5625	2,5	8,75
	4	2,5	6,25	3	1,75	3,0625	4,375	12
	4,5	3	9	4	2,75	7,5625	8,25	18
Σ	12		38	10		37	34,25	49,25
Σ/N	1,5			1,25			4,28125	6,1563

XcpYcp = 1,875 **cov(X,Y) = 4,28125**

σX = 2,1794 **σY = 2,1506**

sX = 2,3299 **sY = 2,2991**

rXY = 0,91341453

rXY = 0,91341453

Корреляционный анализ

Пример 1:

	X	X-Xcp	(X-Xcp)^2	Y	Y-Ycp	(Y-Ycp)^2	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	12,25	-3	-4,25	18,063	14,875	6
	-0,5	-2	4	0	-1,25	1,5625	2,5	0
	0	-1,5	2,25	0,5	-0,75	0,5625	1,125	0
	1	-0,5	0,25	0	-1,25	1,5625	0,625	0
	1,5	0	0	3	1,75	3,0625	0	4,5
	3,5	2	4	2,5	1,25	1,5625	2,5	8,75
	4	2,5	6,25	3	1,75	3,0625	4,375	12
	4,5	3	9	4	2,75	7,5625	8,25	18
Σ	12		38	10		37	34,25	49,25
Σ/N	1,5			1,25			4,28125	6,1563

XcpYcp = **1,875**

cov(X,Y) = **4,28125**

σX = 2,1794
sX = 2,3299

σY = 2,1506
sY = 2,2991

rXY = 0,91341453

rXY = 0,91341453

Корреляционный анализ

Пример 1:

	X	X-Xcp	(X-Xcp)^2	Y	Y-Ycp	(Y-Ycp)^2	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	12,25	-3	-4,25	18,063	14,875	6
	-0,5	-2	4	0	-1,25	1,5625	2,5	0
	0	-1,5	2,25	0,5	-0,75	0,5625	1,125	0
	1	-0,5	0,25	0	-1,25	1,5625	0,625	0
	1,5	0	0	3	1,75	3,0625	0	4,5
	3,5	2	4	2,5	1,25	1,5625	2,5	8,75
	4	2,5	6,25	3	1,75	3,0625	4,375	12
	4,5	3	9	4	2,75	7,5625	8,25	18
Σ	12		38	10		37	34,25	49,25
Σ/N	1,5			1,25			4,28125	6,1563

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

XcpYcp = **1,875**

cov(X,Y) = **4,28125**

σX = **2,1794**

σY = **2,1506**

sX = **2,3299**

sY = **2,2991**

rXY = 0,91341453

rXY = 0,91341453

Корреляционный анализ

Пример 1:

	X	X-Xcp	(X-Xcp)^2	Y	Y-Ycp	(Y-Ycp)^2	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	12,25	-3	-4,25	18,063	14,875	6
	-0,5	-2	4	0	-1,25	1,5625	2,5	0
	0	-1,5	2,25	0,5	-0,75	0,5625	1,125	0
	1	-0,5	0,25	0	-1,25	1,5625	0,625	0
	1,5	0	0	3	1,75	3,0625	0	4,5
	3,5	2	4	2,5	1,25	1,5625	2,5	8,75
	4	2,5	6,25	3	1,75	3,0625	4,375	12
	4,5	3	9	4	2,75	7,5625	8,25	18
Σ	12		38	10		37	34,25	49,25
Σ/N	1,5			1,25			4,28125	6,1563

$$\mathbf{r}_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\mathbf{r}_{XY} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{(N-1)s_X s_Y}$$

XcpYcp = 1,875 **cov(X,Y) = 4,28125**

σX = 2,1794 **σY = 2,1506**

sX = 2,3299 **sY = 2,2991**

rXY = 0,91341453

rXY = 0,91341453

Корреляционный анализ

Пример 1:

	X	X-Xcp	(X-Xcp)^2	Y	Y-Ycp	(Y-Ycp)^2	(X-Xcp)(Y-Ycp)	XY
	-2	-3,5	12,25	-3	-4,25	18,063	14,875	6
	-0,5	-2	4	0	-1,25	1,5625	2,5	0
	0	-1,5	2,25	0,5	-0,75	0,5625	1,125	0
	1	-0,5	0,25	0	-1,25	1,5625	0,625	0
	1,5	0	0	3	1,75	3,0625	0	4,5
	3,5	2	4	2,5	1,25	1,5625	2,5	8,75
	4	2,5	6,25	3	1,75	3,0625	4,375	12
	4,5	3	9	4	2,75	7,5625	8,25	18
Σ	12		38	10		37	34,25	49,25
Σ/N	1,5			1,25			4,28125	6,1563

$$\mathbf{r}_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\mathbf{r}_{XY} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{(N-1)s_X s_Y}$$

XcpYcp = 1,875

cov(X,Y) = **4,28125**

σX = **2,1794**

σY = **2,1506**

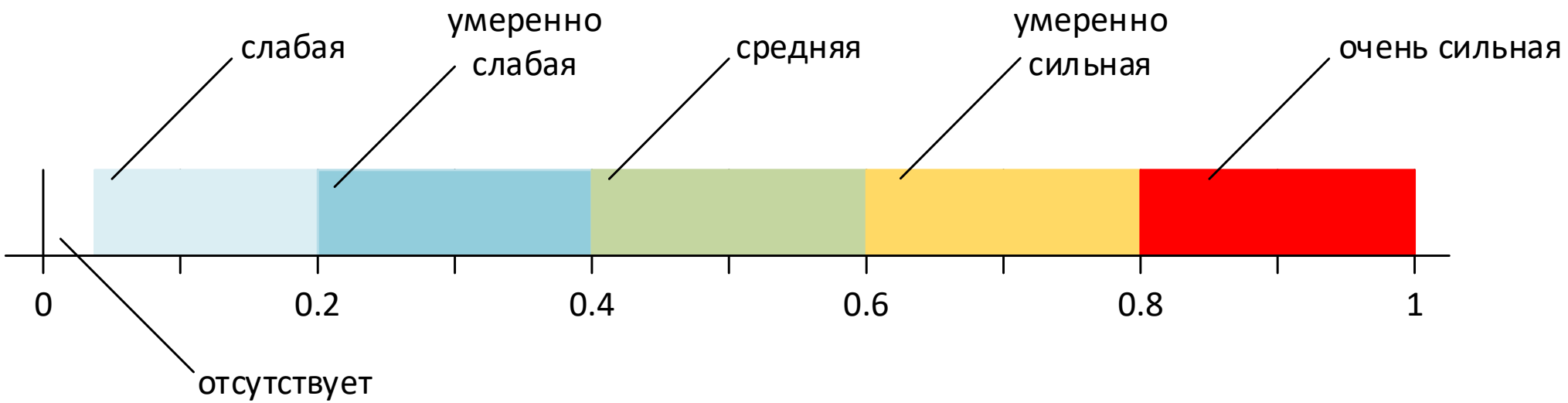
sX = **2,3299**

sY = **2,2991**

rXY = 0,91341453

rXY = 0,91341453

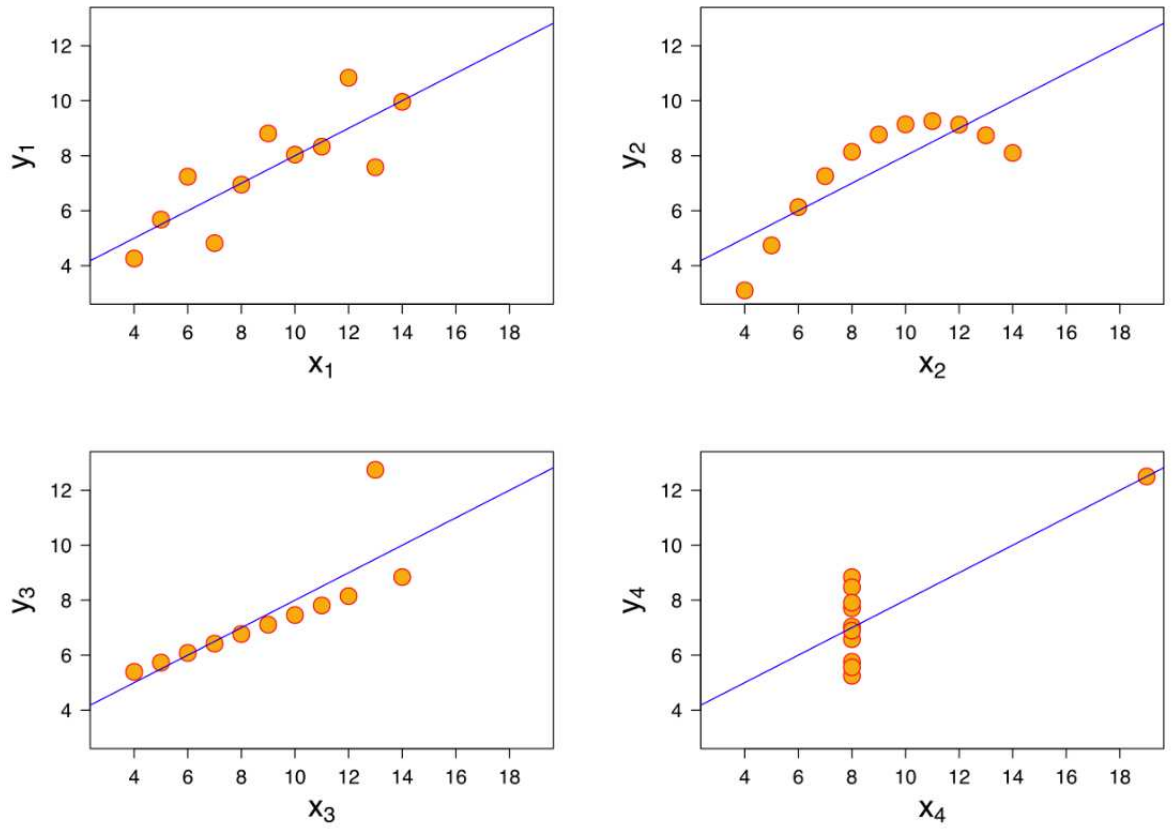
Корреляционный анализ



Корреляционный анализ

Квартет Энскомба

$$\mathbf{r}_{XY} = 0{,}816$$



Корреляционный анализ

Корреляция не подразумевает причинно-следственную связь!

Невозможно обоснованно вывести причинно-следственную связь между двумя переменными исключительно на основе корреляции между ними.

Корреляция не подразумевает причинно-следственную связь!

Идея о том, что «корреляция подразумевает причинно-следственную связь», является примером логической ошибки, в которой два события, происходящие вместе, считаются установившими причинно-следственную связь.

Cum hoc ergo propter hoc – лат. «вместе с этим, значит, из-за этого».

Корреляционный анализ

Для любых двух коррелированных событий, A и B, их возможные отношения включают:

- A влечёт B (прямая причинность);
- B влечёт A (обратная причинность);
- C влечёт одновременно A и B (общий фактор);
- A влечёт B, и B влечёт A (двунаправленная или циклическая причинность);
- Нет связи между A и B; корреляция - совпадение.

Корреляционный анализ

Обратная причинность

Чем быстрее вращаются лопасти мельницы, тем сильнее ветер. Следовательно, мельницы вызывают ветер.

Дети, которые много смотрят телевизор, являются более жестокими. Очевидно, телевидение делает детей жестокими.
(*неясно, что является причиной, а что - следствием*)

Общий фактор

С увеличением продаж мороженого растет число утонувших. Следовательно, употребление мороженого приводит к утоплению.

Корреляционный анализ

Для проверки гипотезы о независимости (некоррелированности) факторов X и Y ($H_0: r = 0$) используется статистика

$$t = \frac{\hat{\mathbf{r}}_{XY}\sqrt{N-2}}{\sqrt{1-\hat{\mathbf{r}}_{XY}{}^2}},$$

которая имеет t-распределение Стьюдента с числом степеней свободы $(N-2)$.

Корреляционный анализ

Факторы X и Y считаются линейно зависимыми, когда величина выборочного коэффициента корреляции $\hat{\mathbf{r}}_{XY}$ отлична от нуля на уровне значимости α, то есть если выполняется неравенство

$$\hat{\mathbf{r}}_{XY}{}^2 > \left[1 + \frac{N-2}{\left(t_{N-2}^{1-\alpha/2}\right)^2}\right]^{-1},$$

где $t_{N-2}^{1-\alpha/2}$ - квантиль на уровне $1 - \alpha/2$ t-распределения Стьюдента с $(N-2)$ степенями свободы.

Чем с меньшим уровнем значимости отвергается гипотеза о некоррелированности, тем меньше вероятность того, что это сделано ошибочно.

Таким образом, уровень значимости статистической гипотезы – это вероятность отвергнуть эту гипотезу при условии, что на самом деле она верна.

Корреляционный анализ

$$\hat{\mathbf{r}}_{XY}{}^2 > \left[1 + \frac{N-2}{\left(t_{N-2}^{1-\alpha/2}\right)^2}\right]^{-1}$$

Если решить неравенство относительно α, то можно найти граничный уровень значимости α_0, для которого гипотеза о некоррелированности всё ещё отвергается при том, что при чуть меньшем уровне значимости она уже была бы принята при заданных выборочном коэффициенте корреляции $\hat{\mathbf{r}}_{XY}$ и объёме выборки N:

Корреляционный анализ

$$\alpha < \alpha_0 = 2 - 2F_{t(N-2)}\left(\frac{|\hat{\mathbf{r}}_{XY}|\sqrt{N-2}}{\sqrt{1-\hat{\mathbf{r}}_{XY}{}^2}}\right),$$

где $F_{t(N-2)}(\cdot)$ - интегральная функция t-распределения Стьюдента с $(N-2)$ степенями свободы.

Значение граничного уровня значимости α_0 в западной литературе носит название *p*-value.

Ранговые коэффициенты корреляции

Иногда возникает потребность в статистическом анализе нечисловых факторов. Например, требуется проверить наличие статистической зависимости между результатами двух кругов чемпионата по футболу.

Так может быть и в случае, когда признаками являются некоторые натуральные порядковые номера или другие целые числа, абсолютные значения которых несут только информацию о порядке объектов.

Корреляционный анализ

В этом случае целесообразно вычислять корреляцию не между самими признаками, а между порядковыми номерами объектов, упорядоченных по этим признакам.

В этом случае вместо коэффициента корреляции Пирсона используются коэффициенты ранговой корреляции Кендалла и Спирмена.

Коэффициент ранговой корреляции Спирмена

Коэффициент ранговой корреляции Спирмена ρ является частным случаем коэффициента корреляции Пирсона, если в качестве значений факторов X, Y использовать значения их рангов x^i, y^i, то есть индексов в упорядоченных последовательностях их значений.

Корреляционный анализ

Коэффициент ранговой корреляции Спирмена ρ можно рассчитать по формуле

$$\rho = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^{N} \left(x^i - y^i\right)^2,$$

где x^i – ранг фактора X_i, y^i – ранг фактора Y_i, i – номер объекта (наблюдения), N – объем выборок (число наблюдений).

Однако эту формулу можно применять только в случае, когда ранги всех элементов в обоих выборках различны.

В ином случае необходимо использовать формулу для коэффициента корреляции Пирсона, заменяя значения элементов выборки на их ранги:

$$\rho = \frac{cov(\boldsymbol{rg}_X, \boldsymbol{rg}_Y)}{\sigma_{\boldsymbol{rg}_X}\sigma_{\boldsymbol{rg}_Y}},$$

где $\boldsymbol{rg}$ – вектор рангов.

Пример 2:

Вычислить коэффициент корреляции Спирмена между IQ человека и временем, проведенным за просмотром телевизора.

IQ	Время, ч в неделю	IQ	Время, ч в неделю
106	7	103	29
100	27	97	20
86	2	113	12
101	50	112	6
99	28	110	17

Пример 2:

IQ, X_i	Время, Y_i	Ранг, x^i	Ранг, y^i	$d_i = x^i - y^i$	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Корреляционный анализ

Пример 2:

IQ, X_i	Время, Y_i	Ранг, x^i	Ранг, y^i	$d_i = x^i - y^i$	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\rho = 1 - \frac{6}{N(N^2-1)} \sum_{i=1}^{N} \left(x^i - y^i\right)^2$$

$$\rho = 1 - \frac{6 \cdot 194}{10(100-1)} \approx -0.1758$$

Корреляционный анализ

Для проверки гипотезы о независимости (некоррелированности) факторов X и Y ($H_0: \rho = 0$), как и для коэффициента корреляции Пирсона, используется статистика

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}},$$

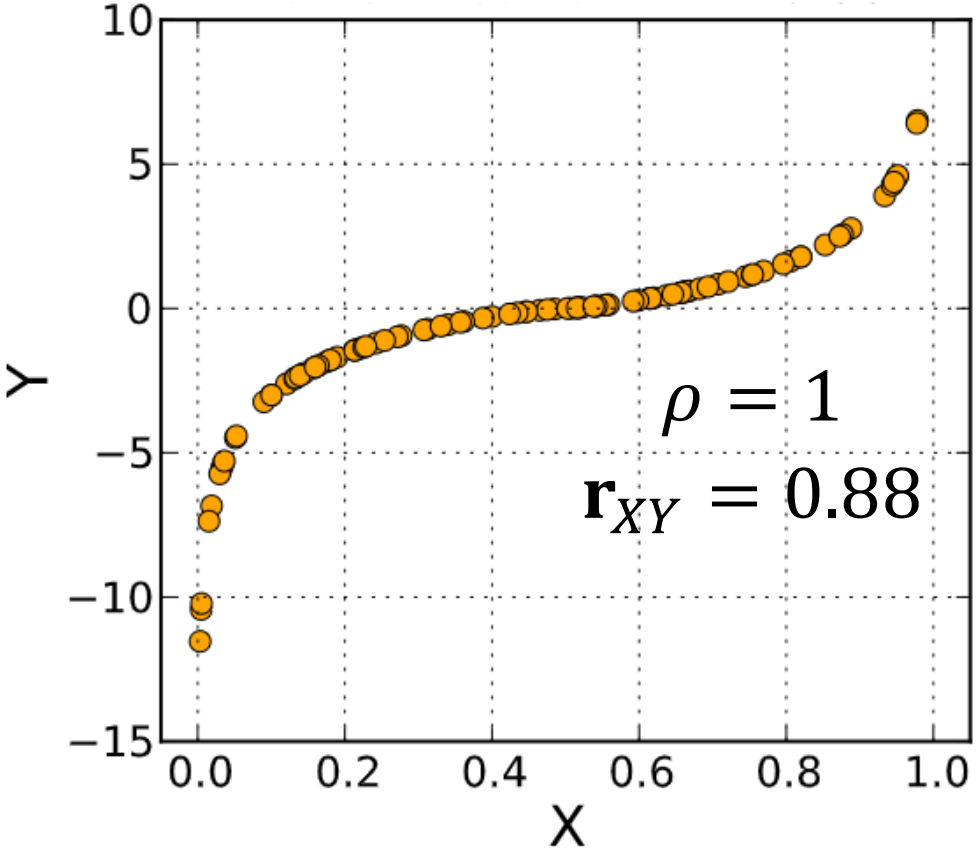
которая имеет t-распределение Стьюдента с числом степеней свободы $(N-2)$.

Граничный (предельный) уровень значимости:

$$\alpha_0 = 2 - 2F_{t(N-2)}\left(\frac{|\rho|\sqrt{N-2}}{\sqrt{1-\rho^2}}\right),$$

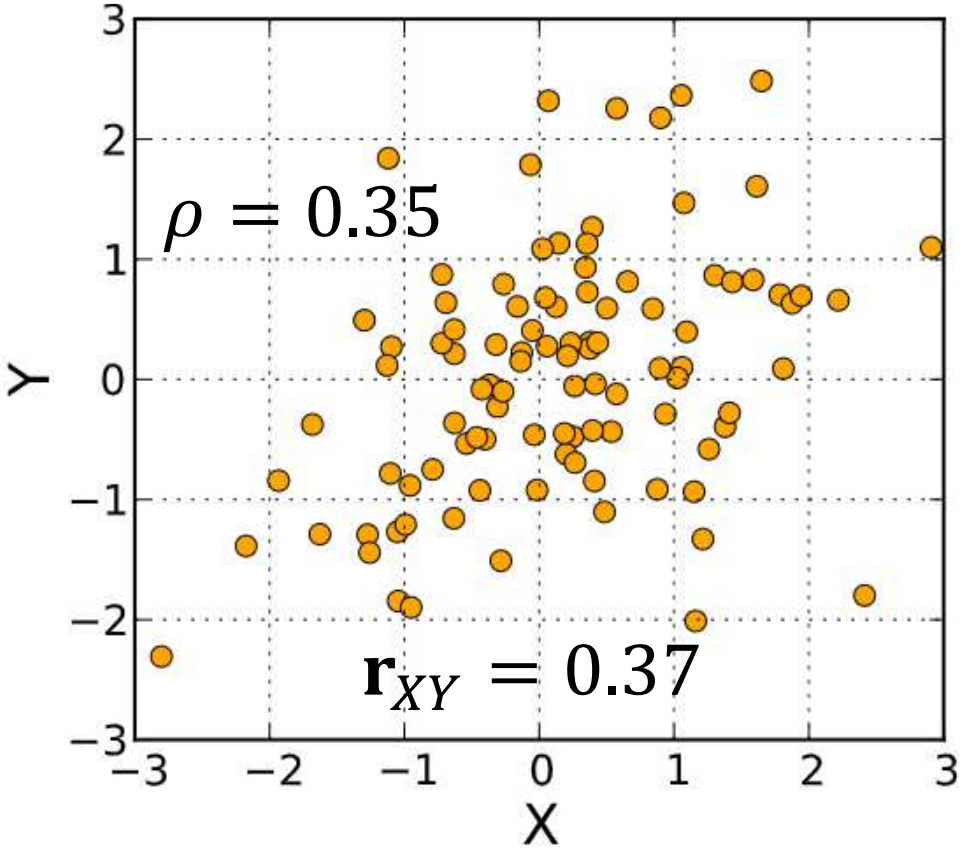
где $F_{t(N-2)}(\cdot)$ - интегральная функция t-распределения Стьюдента с $(N-2)$ степенями свободы.

Корреляционный анализ



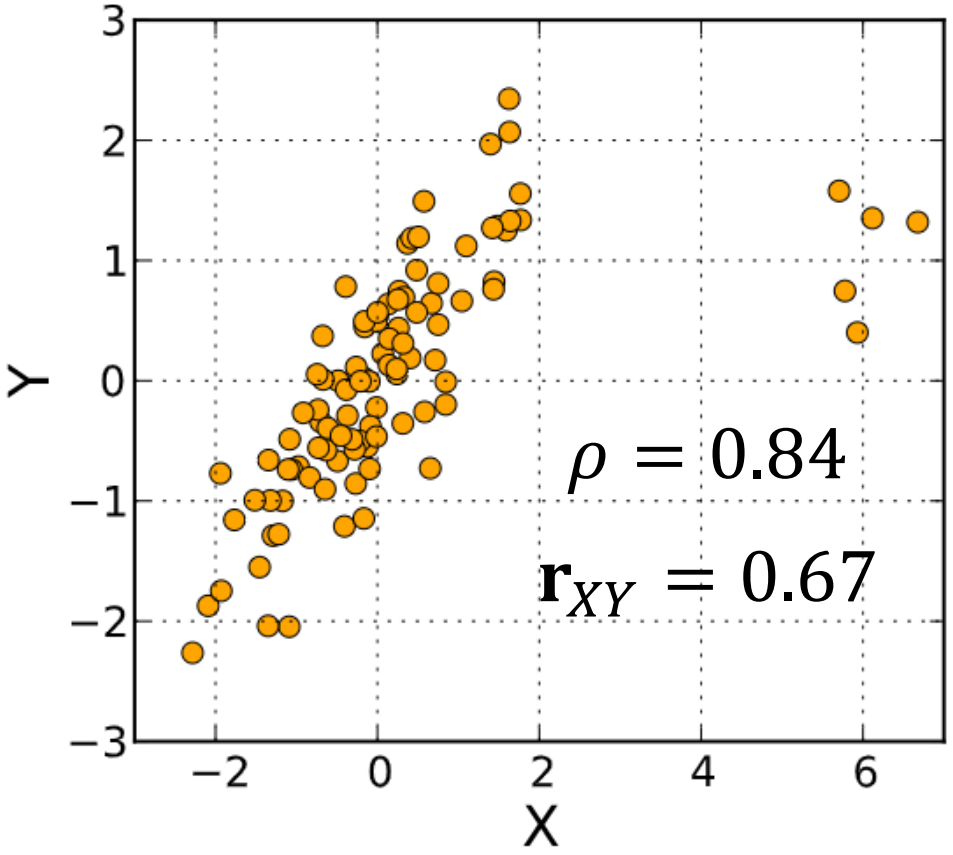
Связь между факторами – нелинейная, однако факторы монотонно связаны.

Корреляционный анализ



Данные распределены более равномерно, без значительных выбросов.

Корреляционный анализ



Корреляция Спирмена менее чувствительна к значительным выбросам, поскольку имеет дело не с самими значениями, а с их рангами.

Коэффициент ранговой корреляции Кендалла

Коэффициент ранговой корреляции Спирмена τ вычисляют на основе ранжирования признаков. Если рассмотреть все возможные пары объектов (i, j), $i, j = 1, 2, \ldots, N$, то про каждую пару можно сказать, одинаково ли или нет упорядочены признаки в этой паре.

Обозначим через P – количество пар объектов, у которых признаки x и y упорядочены одинаково, а через Q – количество пар объектов, у которых признаки x и y упорядочены по-разному.

Тогда

$$P - Q = \sum_{i<j} \mathbf{sgn}\left(x^i - x^j\right)\mathbf{sgn}\left(y^i - y^j\right),$$

где x^i - ранг признака x, y^i - ранг признака y, i – номер объекта,

$$\mathbf{sgn}(u) = \begin{cases} 1, & u > 0; \\ 0, & u = 0; \\ -1, & u < 0. \end{cases}$$

Коэффициент ранговой корреляции Кендалла вычисляют с использованием показателя $(P - Q)$:

$$\tau = \frac{2(P - Q)}{N(N - 1)}.$$

При «ручном» расчёте параметров P и Q можно использовать следующий алгоритм:

Корреляционный анализ

1. Значения фактора X выставляют в порядке возрастания и присваивают ранги.
2. Ранжируют значения показателя Y.
3. Рассчитывают: P – суммарное число наблюдений, следующих за текущими наблюдениями, с большим значением рангов Y; Q – суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y (равные ранги не учитываются).
4. Рассчитывают коэффициент корреляции:

$$\tau = \frac{2(P - Q)}{N(N - 1)}.$$

Коэффициент τ часто используется при проверке статистической гипотезы о независимости факторов X и Y.

При отсутствии зависимости факторов X и Y коэффициент ранговой корреляции τ Кендалла при $N \geq 10$ имеет распределение, близкое к нормальному с параметрами

$$\mathbf{M}\tau = 0; \ \mathbf{D}\tau = \frac{2(2N+5)}{9N(N-1)},$$

то есть статистика $u = \tau \Big/ \sqrt{\frac{2(2N+5)}{9N(N-1)}}$ имеет приблизительно стандартное нормальное распределение.

Пример 3:

Вычислить выборочные коэффициенты корреляции Кендалла τ для двух пар факторов: Мастер-Студент 1 и Мастер-Студент 2.

Мастер	Студент 1	Студент 2
1	2	12
2	1	2
3	4	3
4	3	4
5	6	5
6	5	6
7	8	7
8	7	8
9	10	9
10	9	10
11	12	11
12	11	1

Корреляционный анализ

Пример 3:

Мастер	Студент 1	P	Q
1	2	10	1
2	1	10	0
3	4	8	1
4	3	8	0
5	6	6	1
6	5	6	0
7	8	4	1
8	7	4	0
9	10	2	1
10	9	2	0
11	12	0	1
12	11	0	0

$$\tau = \frac{2 \cdot (60 - 6)}{12 \cdot 11} = \frac{54}{66} \approx 0{,}818$$

Корреляционный анализ

Пример 3:

Мастер	Студент 2	P	Q
1	12	0	11
2	2	9	1
3	3	8	1
4	4	7	1
5	5	6	1
6	6	5	1
7	7	4	1
8	8	3	1
9	9	2	1
10	10	1	1
11	11	0	1
12	1	0	0

$$\tau = \frac{2 \cdot (45 - 21)}{12 \cdot 11} = \frac{24}{66} \approx 0{,}364$$

Пример 4:

Вычислить выборочные коэффициенты корреляции Пирсона $\mathbf{r}$, Спирмена ρ и Кендалла τ двух *бинарных факторов*: X – пациенту проведена или не проведена вакцинация от гриппа, Y – пациент заболел или не заболел гриппом. Протокол прививок и наблюдений задан следующей таблицей (число объектов $N = 9$):

№ пациента		1	2	3	4	5	6	7	8	9
X	Прививка	1	1	1	0	1	0	0	1	0
Y	Заболевание	0	1	1	1	1	0	0	1	0

Пример 4:

	№ пациента	1	2	3	4	5	6	7	8	9
X	Прививка	1	1	1	0	1	0	0	1	0
Y	Заболевание	0	1	1	1	1	0	0	1	0

$$\mathbf{r} = \frac{\sum_{i=1}^{N} X_i Y_i - N \cdot \bar{X}\bar{Y}}{(N-1)s_X s_Y};\ \ s_Z = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2};$$

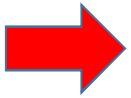
$$\sum X_i Y_i = 4;\ \ \bar{X} = \frac{5}{9};\ \ \bar{Y} = \frac{5}{9};\ \ s_X = s_Y = \frac{180}{81};$$

$$\mathbf{r} = \frac{11}{20} = 0{,}55.$$

Пример 4:

Для вычисления ранговой корреляции Спирмена проведём ранжирование факторов X и Y:

	№ пациента	1	2	3	4	5	6	7	8	9
X	Прививка	1	1	1	0	1	0	0	1	0
Y	Заболевание	0	1	1	1	1	0	0	1	0



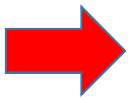
	№ пациента	1	2	3	4	5	6	7	8	9
X	Прививка	5	6	7	1	8	2	3	9	4
Y	Заболевание	1	5	6	7	8	2	3	9	4

$$\rho = 1 - \frac{6}{N(N^2-1)}\sum_{i=1}^{N}\left(x^i - y^i\right)^2 = 1 - \frac{6}{9 \cdot 80}(4^2 + 1^2 + 1^2 + 6^2 + 0 + 0 + 0 + 0 + 0) = \frac{11}{20} = 0{,}55$$

Пример 4:

Для вычисления ранговой корреляции Кендалла расположим один их факторов (например, фактор X) в порядке возрастания рангов и запишем соответствующие ранги фактора Y:

	№ пациента	1	2	3	4	5	6	7	8	9
X	Прививка	1	1	1	0	1	0	0	1	0
Y	Заболевание	0	1	1	1	1	0	0	1	0



	№ пациента	4	6	7	9	1	2	3	5	8
X	Прививка	1	2	3	4	5	6	7	8	9
Y	Заболевание	7	2	3	4	1	5	6	8	9

$$P = 2 + 6 + 5 + 4 + 4 + 3 + 2 + 1 + 0 = 27$$

$$Q = 6 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 = 9$$

$$\tau = \frac{2(P - Q)}{N(N - 1)} = \frac{2 \cdot 18}{9 \cdot 8} = 0{,}5$$