Понятие корреляционной зависимости

Многие задачи требуют установить и оценить зависимость между двумя или несколькими случайными величинами.

- Определение. Зависимость случайных величин называют *статистической*, если изменение одной величины влечет изменение распределения другой величины.
- Определение. Статистическая зависимость называется *корреляционной*, если при изменении одной величины изменяется среднее значение другой.

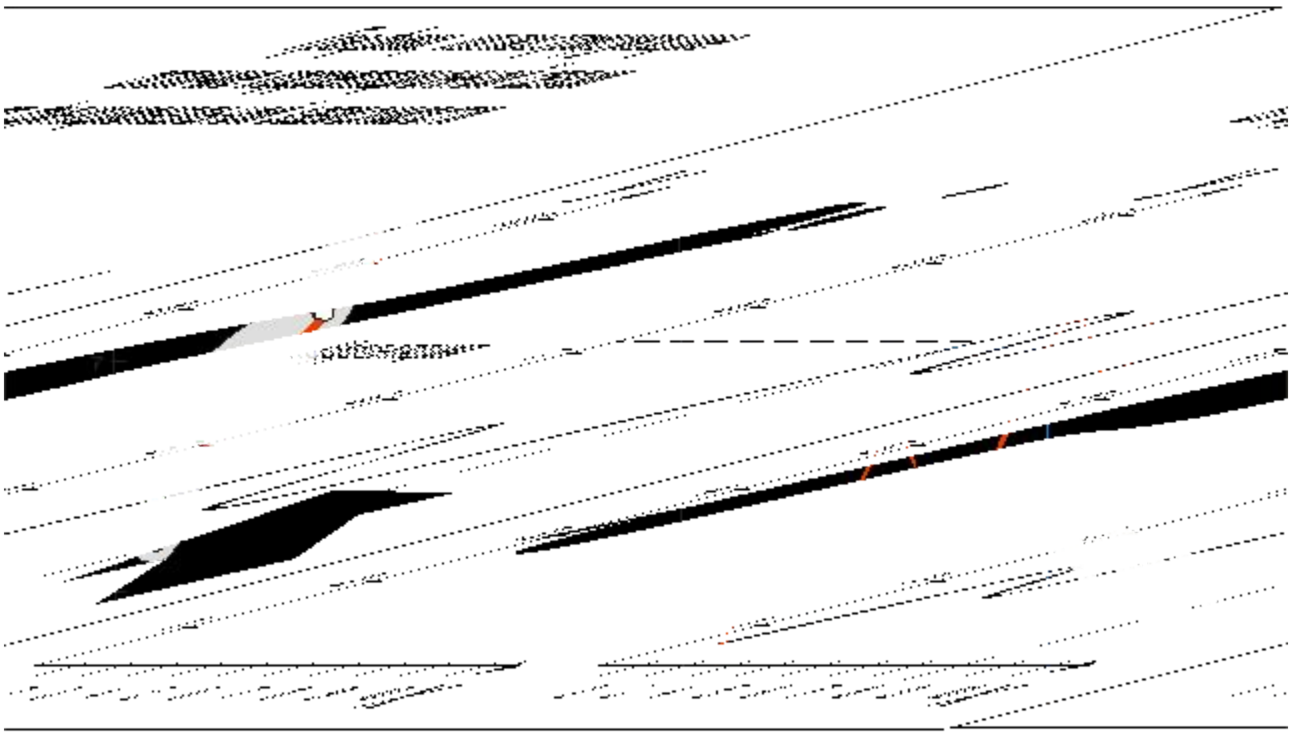
Если случайная величина представляет некоторый признак (например, статистические наблюдения некой экономической величины), то под **корреляцией** понимают – меру согласованности одного признака с другим, или с несколькими, либо взаимную согласованность группы признаков.

Ложная корреляция

- **Корреляционная зависимость** указывает на причинно-следственную связь изменений двух признаков. Однако, корреляционные методы не выявляют этой причинности, а лишь указывают на наличие некоторого соответствия. Признаки могут находиться не только во взаимной зависимости друг от друга, но и оба зависеть от какого-либо третьего воздействия, не включенного в область рассмотрения. Например, между двумя временными рядами (переменные, состоящие из наблюдений отстоящих на равные промежутки времени друг от друга) может быть сильная корреляционная зависимость, однако эта зависимость будет ложной, так как переменные сами зависят от времени.
- Таким образом, более корректно употреблять понятие корреляционная связь.

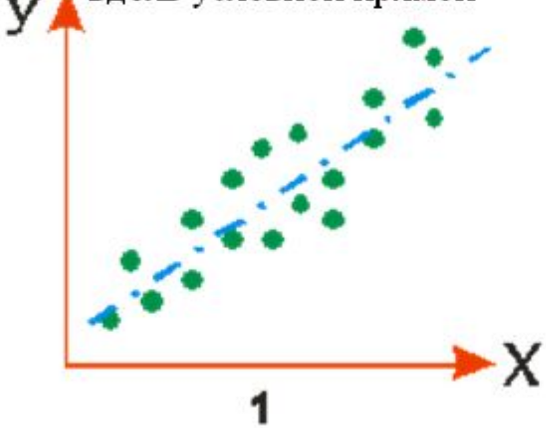
Отличие корреляционной от функциональной зависимости

Функциональная зависимость предполагает взаимно однозначное соответствие аргумента ***x*** и функции ***y=f(x)***, вероятностная же зависимость допускает некий условный диапазон, в который предположительно (с такой-то долей вероятности) попадает значение признака y_i при значении x_i признака **x**.

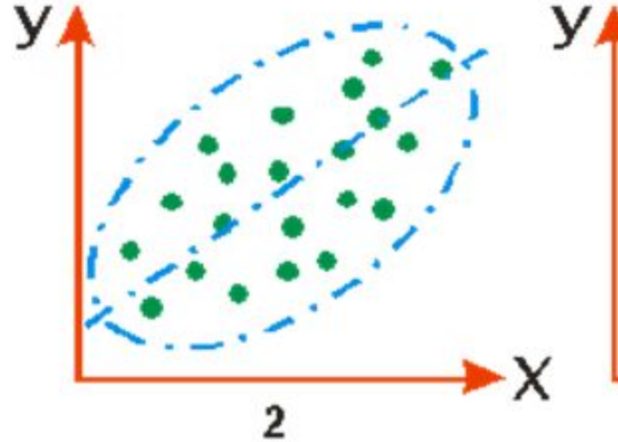


Примеры корреляционной зависимости

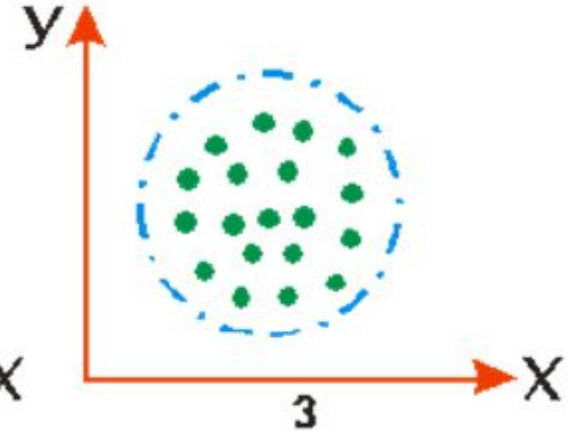
имеется значимая положительная корреляция: r>0,8
точки расположены примерно вдоль условной прямой



имеется некоторая корреляция, точки еще расположены вдоль прямой, но уже хаотично, вписываются в эллипс 0,5<r<0,6



корреляция отсутствует: точки расположены хаотично (вписываются в окружность) r=0



Имеется значимая положительная корреляция r=+1, точки расположены вдоль прямой
Иначе: функциальная зависимость

Диаграмма рассеяния показывает однозначное соответствие: точки расположены вдоль линии y=cos(x)
Однако r=0!
Имеется нелинейная взаимосвязь

Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона характеризует наличие линейной связи между признаками,

де x_i — значения, принимаемые в выборке X,

y_i — значения, принимаемые в выборке Y;

— средняя по X, — средняя по Y.

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

$$\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$$

Ведем обозначения: ковариация признаков X иY

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

Средние квадратичные отклонения $\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$ и $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

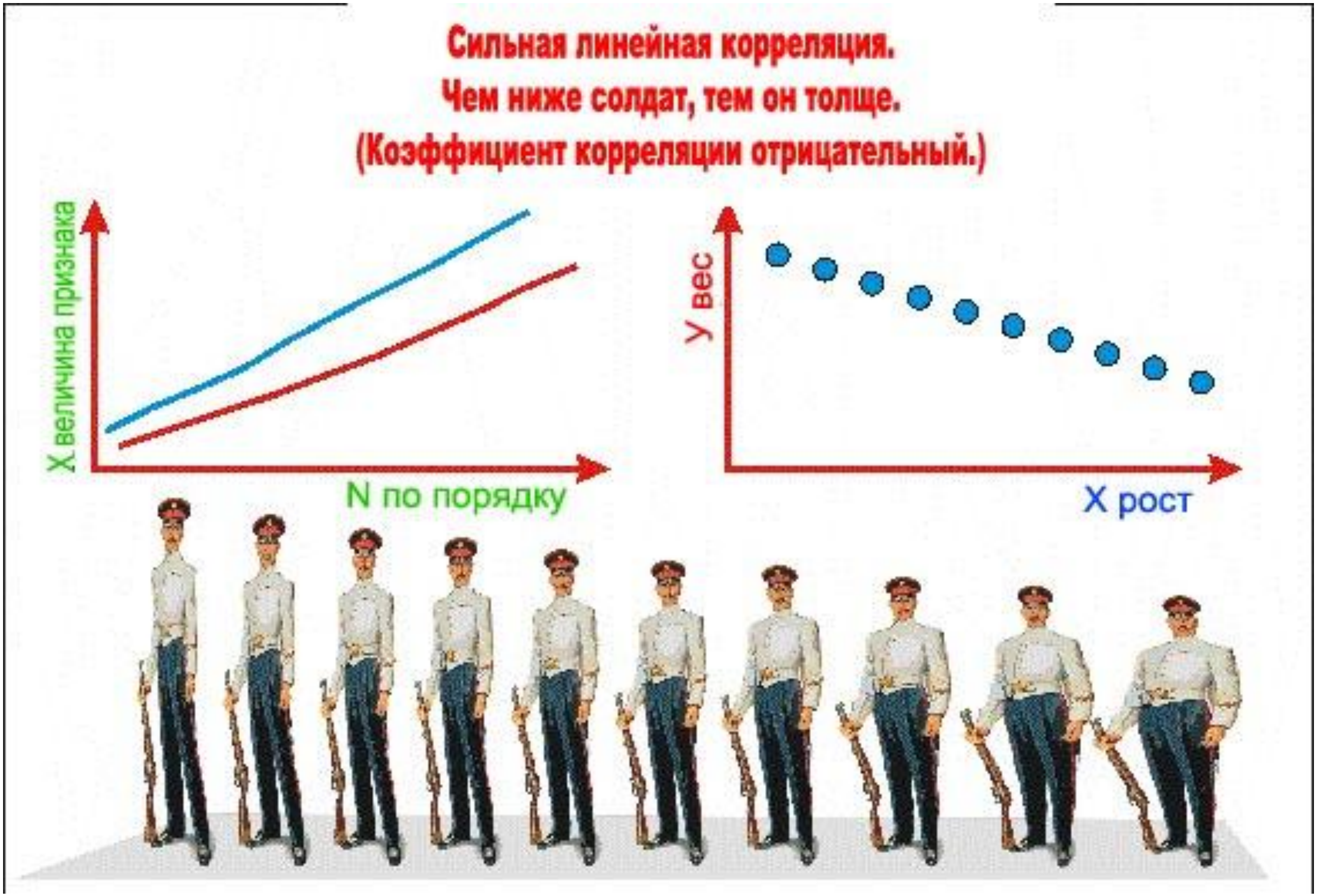
Тогда:

$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_y}$$

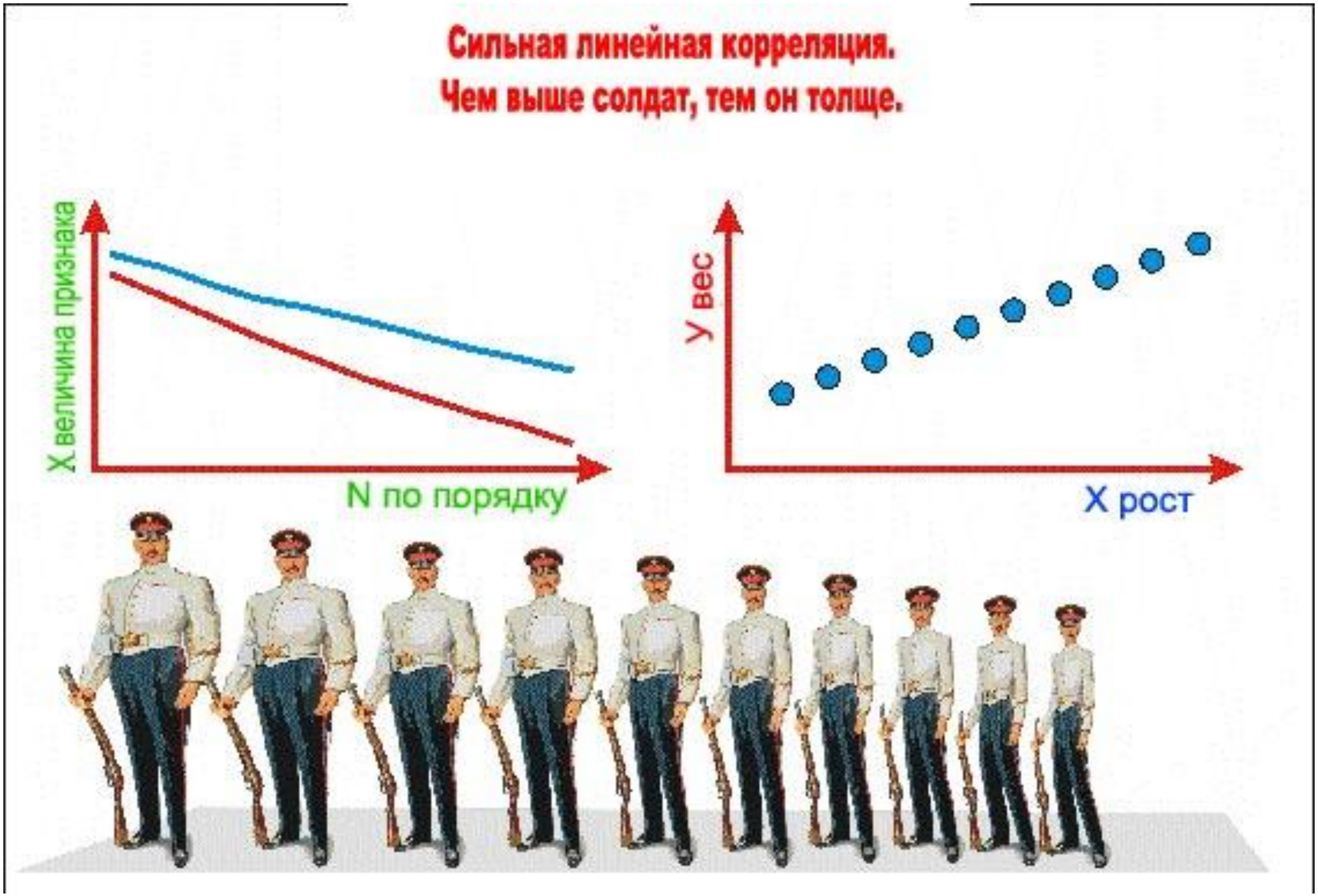
Значение коэффициента корреляции

- **сильная**, или **тесная** при коэффициенте корреляции $\boldsymbol{r}>0,70$;
- **средняя** при $0,50<\boldsymbol{r}<0,69$;
- **умеренная** при $0,30<\boldsymbol{r}<0,49$;
- **слабая** при $0,20<\boldsymbol{r}<0,29$;
- **очень слабая** при $\boldsymbol{r}<0,19$.
- Если коэффициент корреляции положительный, то связь между признаками прямая: увеличение одного признака приводит к увеличению другого
- Если коэффициент корреляции отрицательный, то связь между признаками обратная: увеличение одного признака приводит к уменьшению другого
- В случае, если $\boldsymbol{r}$=1, -1, то связь между признаками функциональная!


Слабая линейная корреляция.
Почти при одинаковом росте все солдаты
то худые, то толстые.
X величина признака
N по порядку
У вес
X рост



Сильная линейная корреляция.
Чем ниже солдат, тем он толще.
(Коэффициент корреляции отрицательный.)
X величина признака
N по порядку
У вес
X рост


Сильная линейная корреляция.
Чем выше солдат, тем он толще.
X величина признака
N по порядку
У вес
X рост

Проверка значимости коэффициента корреляции Пирсона

Нулевая и альтернативная гипотезы имеют вид:

H_0*: *коэффициент корреляции Пиросона* r *незначимый;

H_1*: *коэффициент корреляции Пирсона* r *значим.

- Рассчитывается t-статистика по формуле:

$$t_{расч.} = \frac{r}{\sqrt{1-r^2}}\sqrt{(n-2)}$$

- Определяется $t_{табл}$ по таблице Стьюдента со степенями свободы n-2 и уровнем значимости α
- *Если* $|t_{расч}| > t_{табл}$, *то* ***H_0*** отклоняют на заданном уровне значимости, и считаем, что коэффициент корреляции Пирсона значимый.

Непараметрические показатели корреляции

Определение. Под **качественным** подразумевается признак, который невозможно измерить точно, но он позволяет сравнить объекты между собой и расположить их в порядке убывания или возрастания качества.

Под ранжированием будем понимать упорядочивание объектов согласно убыванию качественного признака

Для оценки степени связи качественных признаков используют коэффициенты ранговой корреляции.

Коэффициент корреляции Спирмена — мера линейной связи между случайными величинами. Корреляция Спирмена является **ранговой**, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги.

Коэффициент корреляции Кендалла — мера линейной связи между случайными величинами

Схема нахождения коэффициента Корреляции Спирмена

1. Определить, какие два признака или две иерархии признаков будут участвовать в сопоставлении как переменные *X* и *Y*.
2. Проранжировать значения переменной *X*, присваивая ранг 1 наименьшему значению, и т.д. Занести ранги в первый столбец таблицы по порядку номеров испытуемых или признаков.
3. Проранжировать значения переменной *Y*, в соответствии с теми же правилами. Занести ранги во второй столбец таблицы по порядку номеров испытуемых или признаков.
4. Подсчитать разности *d* между рангами *X* и *Y* по каждой строке таблицы и занести в третий столбец таблицы.
5. Возвести каждую разность в квадрат: d^2. Эти значения занести в четвертый столбец таблицы.
6. Подсчитать сумму d^2.
7. При наличии одинаковых рангов рассчитать поправки:

$$T_a = \sum\left(a^3 - a\right)/12$$

$$T_b = \sum\left(b^3 - b\right)/12$$

где a - объем каждой группы одинаковых рангов в ранговом ряду *X*; b - объем каждой группы одинаковых рангов в ранговом ряду *Y*.

Схема нахождения коэффициента Корреляции Спирмена

8. Рассчитать **коэффициент ранговой корреляции** r_s по формуле:

при отсутствии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2}{N \cdot \left(N^2 - 1\right)}$$

при наличии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2 + T_a + T_b}{N \cdot \left(N^2 - 1\right)}$$

где sum(d^2) - сумма квадратов разностей между рангами;
T_a и T_b - поправки на одинаковые ранги;
N - количество наблюдений признаков, участвовавших в ранжировании.

Проверка значимости коэффициента ранговой корреляции Спирмена

Нулевая и альтернативная гипотезы имеют вид:

H_0: ***коэффициент ранговой корреляции Спирмена*** r_s ***незначимый***;

H_1: ***коэффициент ранговой корреляции Спирмена*** r_s ***значим.***

- Рассчитывается t-статистика по формуле:

$$t_{расч.} = \frac{r_s}{\sqrt{1 - r_s^2}}\sqrt{(n-2)}$$

- Определяется $t_{табл}$ по таблице Стьюдента со степенями свободы n-2 и уровнем значимости α
- *Если* $|t_{расч}| > t_{табл}$ *, то* $\boldsymbol{H_0}$ отклоняют на заданном уровне значимости, и считаем, что коэффициент ранговой корреляции Спирмена значимый.

Схема нахождения коэффициента корреляции Кендалла

1. В порядке возрастания признака ***X*** выстраивают сопряженные наблюдения пар (x_i , y_i) и записывают их в таблицу.

2. Для каждого значения y_i определяют его ранг s_i, записывается в таблицу.

3. На последовательности рангов s_1, s_2, …, s_N определяют количество *инверсий*, т.е. нарушений порядка следования. Например, при $N = 4$ и последовательности рангов {1, 3, 4, 2} имеем количество инверсий: 3 – количество инверсий для числа 1 (после числа 1 есть три значения, больше 1) и 1 – количество инверсий для числа 3 (после числа 3 есть одно значение, больше 3).

4. Формируют ряд значений в таблице из инверсий, если инверсий нет, то присваивают ячейке значение 0.

5. Рассчитывают сумму всех инверсий *K*: $K = \sum_{i=1}^{N} inv$

6. Определяют ***коэффициент ранговой корреляции по Кендаллу***:

$$\tau_K = 1 - \frac{4 \cdot K}{N * (N-1)}$$

Проверка значимости коэффициента ранговой корреляции Кендалла

Для проверки *значимости рангового коэффициента Кендалла*, то есть для проверки существенности корреляционной связи, выдвигают гипотезы:

H$_0$: коэффициент ранговой корреляции Кендалла τ_K незначимый (τ_K=0);

H$_1$: коэффициент ранговой корреляции Кендалла τ_K значим ($\tau_K \neq 0$);.

Рассчитывается Z-статистика по формуле: $z_{расч.} = \tau_K \sqrt{\dfrac{9N(N+1)}{2(2N+5)}}$

По таблице значений функции Лапласа определяем $z_{табл}$ из равенства для $\Phi(z_{табл}) = \dfrac{1-\alpha}{2}$ уровня значимости α.

<u>Примечание:</u> $z_{табл}$ можно определить также в модуле Вероятностный калькулятор, выбрав нормальное распределение Z, p=1–α , *mean*=0, *st.dev*=1, и отметив режим двусторонней проверки гипотезы.

Если $|z_{расч}| > z_{табл}$, следовательно, нулевую гипотезу о незначимости коэффициента Кендалла (τ_K=0), можно отклонить на заданном уровне значимости α.

Схема нахождения коэффициента конкордации

- Определение. Множественный коэффициент ранговой корреляции, позволяющий определить тесноту связи между несколькими ранжированными признаками, называется **коэффициентом конкордации.**

1. Определить, какие признаки будут участвовать в сопоставлении как переменные (X, Y, Z,…).
2. Проранжировать значения всех признаков, присваивая ранг 1 наименьшему значению, и т.д. Занести ранги в столбцы таблицы по порядку номеров признаков (R_x, R_y, R_z,…).
3. Сформировать в таблице столбец из суммы всех рангов ($R_s=R_x+R_y+R_z+\ldots$).
4. Сформировать в таблице столбец из квадратов сумм всех рангов, полученных в п.3. $R_s^{\,2}$
5. Определить по столбцу из сумм всех рангов (полученных в п.3) среднее значение, где n – число наблюдений.

$$\overline{R_S} = \frac{\sum_{i=1}^{n} R_{S_i}}{n}$$

Схема нахождения коэффициента конкордации

6. Определить отклонение суммы квадратов рангов от средне квадратов рангов. $S = \sum_{i=1}^{n} R_{S_i}^{2} - \left(\overline{R_S}\right)$

7. Вычислить **коэффициент конкордации**: $W = \dfrac{12 \cdot S}{m^2 \cdot (n^3 - n)}$

Где m- количество факторов (признаков сравнения),

n – число наблюдений.

Для проверки ***значимости коэффициента конкордации***, выдвигают гипотезы:

H_0: коэффициент конкордации W незначимый (W=0);

H_1: коэффициент конкордации W значим ($W \neq 0$);.

Рассчитывается χ^2-статистика по формуле: $\chi^2 = \dfrac{12 \cdot S}{m \cdot n(n-1)}$

По таблице значений χ^2-распределения определяем $\chi^2_{табл}$, для степени свободы $v=n$ и уровня значимости α.

Если $\chi^2 > \chi^2_{табл}$ следовательно, нулевую гипотезу о незначимости коэффициента конкордации (W=0), можно отклонить на заданном уровне значимости α.

Примечание, $\chi^2_{табл}$ можно определить из модуля **Вероятностный калькулятор** пакета Statistica.

Количественная оценка связи явлений различной природы: коэффициенты ассоциации и контингенции

Если качественные признаки состоят только из двух групп, то для определения тесноты связи двух качественных признаков применяют **коэффициенты ассоциации и контингенции.**

Схема нахождения коэффициентов

$\bar{x}$

1. Пусть I явление имеет две альтернативы ***a*** и ***b***, причем частоты их появления соответственно: n_a и $n_{b.}$

Пусть II явление имеет две альтернативы ***c*** и ***d***, причем частоты их появления соответственно: n_c и n_d

2. Составляется таблица:

I II	a	b	
c	n_{ac}	n_{bc}	n_c
d	n_{ad}	n_{bd}	n_d
	n_a	$n_{b.}$	

Схема нахождения коэффициентов *ассоциации и контингенции*

3. Причем $n_a = n_{ac} + n_{ad.}$ и $n_{b.} = n_{bc} + n_{bd}$
$n_c = n_{ac} + n_{bc.}$ и $n_{d.} = n_{ad} + n_{bd}$
4. Определяется **коэффициент ассоциации** как:

$$K_a = \frac{n_{ac} \cdot n_{bd} - n_{bc} \cdot n_{ad}}{n_{ac} \cdot n_{bd} + n_{bc} \cdot n_{ad}}$$

Определяется **коэффициент контингенции**:

$$K_k = \frac{n_{ac} \cdot n_{bd} - n_{bc} \cdot n_{ad}}{\sqrt{n_a \cdot n_b \cdot n_c \cdot n_d}}$$

5. Связь считается подтвержденной если $K_a > 0{,}5$, а $K_k > 0{,}3$.

Примечание. Коэффициент контингенции всегда меньше коэффициента ассоциации.

Коэффициенты взаимной сопряженности

Если качественные признаки состоят из более чем двух групп, то для определения тесноты связи качественных признаков применяют **коэффициенты сопряженности Пирсона и Чупрова.**

Схема нахождения коэффициентов сопряженности

1. Пусть I явление имеет альтернативы $\boldsymbol{a}_I$, $\boldsymbol{b}_I$, $\boldsymbol{c}_I$ и т.д., причем частоты их появления соответственно: n_{aI} , n_{bI}, n_{cI} …

Пусть II явление имеет альтернативы $\boldsymbol{a}_{II}$, $\boldsymbol{b}_{II}$, $\boldsymbol{c}_{II}$ и т.д, причем частоты их появления соответственно: n_{aII} , n_{bII}, n_{cII} …

2. Составляется таблица:

	a_I	b_I,	c_I		Итого
a_{II}	$n_{aI,aII}$	$n_{bI,aII}$	$n_{cI,aII}$	…	n_{aII}
b_{II}	$n_{aI,bII}$	$n_{bI,bII}$	$n_{cI,bII}$	…	n_{bII}
c_{II}	$n_{aI,cII}$	$n_{bI,cII}$	$n_{cI,cII}$	…	n_{cII}
	…	…	…	…	…
	n_{aI}	n_{bI}	n_{cI}	…	

Схема нахождения коэффициентов *взаимной сопряженности*

3. Причем $n_{aII} = n_{aI,aII} + n_{bI,aII} + n_{cI,aII}$; $n_{bII} = n_{aI,bII} + n_{bI,bII} + n_{cI,bII}$
$n_{cII} = n_{aI,cII} + n_{bI,cII} + n_{cI,cII}$
И $n_{aI} = n_{aI,aII} + n_{aI,bII} + n_{aI,cII}$; $n_{bI} = n_{bI,aII} + n_{bI,bII} + n_{bI,cII}$
$n_{cI} = n_{cI,aII} + n_{cI,bII} + n_{cI,cII}$
4. Определяется значение:

$$1+\varphi^2 = \frac{\frac{(n_{aI,aII})^2}{n_{aI}} + \frac{(n_{bI,aII})^2}{n_{bI}} + \frac{(n_{cI,aII})^2}{n_{cI}} + \ldots}{n_{aII}} + \frac{\frac{(n_{aI,bII})^2}{n_{aI}} + \frac{(n_{bI,bII})^2}{n_{bI}} + \frac{(n_{cI,bII})^2}{n_{cI}} + \ldots}{n_{bII}} +$$
$$+ \frac{\frac{(n_{aI,cII})^2}{n_{aI}} + \frac{(n_{bI,cII})^2}{n_{bI}} + \frac{(n_{cI,cII})^2}{n_{cI}} + \ldots}{n_{cII}} + \ldots$$

Схема нахождения коэффициентов *взаимной сопряженности*

5. Определяется **коэффициент взаимной сопряженности Пирсона**:

$$K_{П} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

6. Определяется **коэффициент взаимной сопряженности Чупрова**:

$$K_{Ч} = \sqrt{\frac{\varphi^2}{\sqrt{(K_1 - 1)(K_2 - 1)}}}$$

Где K_1 – число значений (групп) I-ого признака.

Где K_2 – число значений (групп) II-ого признака.

Чем ближе коэффициенты взаимной сопряженности **к единице**, тем теснее связь.

Биссериальный коэффициент корреляции

Связь между качественными альтернативами признака и количественными вариациями признака определяют на основе **биссериального коэффициента корреляции.**

Схема нахождения коэффициентов сопряженности

1.Пусть даны два качественных признака (категории или группы), для которых известны количественные характеристики. Количество наблюдений в I-ой группе – n_1, в II-ой группе – n_2. Общее количество наблюдений $n = n_1 + n_2$.

2. По каждому из признаков (группе) определяется среднее значение: $\overline{y_1}$ и $\overline{y_2}$

3. Определяются доли каждой группы в общем объеме:

Для I –ой группы: $\boldsymbol{p = n_1/n}$

Для II-ой группы: $\boldsymbol{q = n_2/n}$

4. Рассчитывается общее среднее значение для обеих групп (признаков)

$\overline{y_{общ}}$

Биссериальный коэффициент корреляции

5. Вычисляется среднее квадратичное отклонение фактических значений признака от среднего уровня:

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \overline{y_{общ}}\right)^2}{n}}$$

6. По таблице значений функции Лапласа определяем $z_{табл}$ из равенства для $\Phi(z_{табл}) = \frac{1-\alpha}{2}$ уровня значимости α.

7. Определяется **биссериальный коэффициент корреляции**:

$$r_б = \frac{\left|\overline{y_2} - \overline{y_1}\right|}{\sigma_y} \cdot \frac{pq}{z_{табл}}$$

Чем ближе значение коэффициента к единице, тем теснее связь между признаками.