Лекция 9

ВЗАИМОСВЯЗЬ ПРИЗНАКОВ.

Во многих практических задачах приходится исследовать объекты, обладающие несколькими признаками. Поэтому возникает вопрос, насколько эти признаки связаны между собой. Например, у каждого человека есть возраст, место рождения, уровень образования, увлечения и т.д. Вопрос состоит в том, можно ли эти признаки считать независимыми (в вероятностном смысле), либо между ними существует некоторая взаимосвязь. И, если она существует, насколько точно можно предсказать значение одного признака по значению другого. Таким образом, при изучении взаимосвязи признаков рассматриваются две задачи:

1. Проверки гипотезы об отсутствии статистической зависимости признаков
2. Если статистическая зависимость (связь признаков) существует, нахождение меры связи

(1) Понятие статистической независимости формулируется в теории вероятностей

Пусть A и B – некоторая пара признаков некоторого объекта. Предположим, что признак A может принимать значения $A_1, A_2, \ldots$, а признак B – значения $B_1, B_2, \ldots$

Определение. Признаки A и B называют независимыми, если (при случайном выборе объекта) для всех пар i и j оказываются независимыми события «признак A принимает значение A_i» и «признак B принимает значение B_j»

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

(2) Выявление связи между случайными признаками и оценка тесноты этой связи является задачей **корреляционного анализа**. Показателем тесноты связи является **коэффициент корреляции.**

Для качественных (номинальных и порядковых) и количественных (дискретных и непрерывных) признаков используются разные вида шкалы измерений (номинальная, порядковая или «шкалы») и соответственно по-разному решаются вопросы об их статистической зависимости:

- (I) Для изучения связи признаков, измеренных в номинальной шкале – таблицы сопряженности, критерий хи-квадрат
- (II) Для изучения связи признаков, измеренных в порядковой шкале – ранговая корреляция (коэффициенты корреляции Спирмена и Кендалла)

- (III) Для изучения связи количественных признаков («шкалы») – коэффициент корреляции Пирсона, модель простой линейной регрессии.

Рассмотрим сначала задачу проверки гипотезы об отсутствии статистической зависимости **номинальных** признаков.

Типичной ситуацией, в которой встречаются номинальные признак, является обработка социологических анкет. Многочисленный цифровой материал, имеющийся в распоряжении социолога, для удобного сопоставления и анализа оформляется в виде таблиц. По содержанию таблицы делятся на аналитические и неаналитические. В неаналитических таблицах помещаются необработанные статистические данные, необходимые лишь для информации и констатации. Аналитические таблицы являются результатом обработки, группировки и анализа цифровых показателей. Поэтому их еще называют **таблицами сопряженности**.

(I) Таблицы сопряженности (связь номинальных признаков)

Понятие статистической независимости формулируется в теории вероятностей

Пусть A и B – некоторая пара признаков некоторого объекта. Предположим, что признак A может принимать значения $A_1, A_2, \ldots$, а признак B – значения $B_1, B_2, \ldots$

Определение. Признаки A и B называют независимыми, если (при случайном выборе объекта) для всех пар i и j оказываются независимыми события «признак A принимает значение A_i» и «признак B принимает значение B_j»

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

В распоряжении исследователя имеется выборка объемом m из генеральной совокупности. По этой выборке можно определить частоты событий A_i и B_j по отдельности и в любых комбинациях

m_{ij} – частота события «признак A принимает значение A_i и одновременно «признак B принимает значение B_j»

$m_{(A)i} = \sum_j m_{ij}$ – «признак A принимает значение A_i» (при этом признак B принимает любое значение)

$m_{(B)j} = \sum_i m_{ij}$ – «признак В принимает значение B_j» (при этом признак А принимает любое значение)

$$\sum_i \sum_j m_{ij} = m$$

При бесконечном объеме выборки $m \to \infty$ относительные частоты неограниченно приближаются к вероятностям

$$\frac{m_{ij}}{m} \xrightarrow[m\to\infty]{} P(A_i \cap B_j)$$

$$\frac{m_{(A)i}}{m} \xrightarrow[m\to\infty]{} P(A_i)$$

$$\frac{m_{(B)j}}{m} \xrightarrow[m\to\infty]{} P(B_j)$$

Поэтому, если теоретическая модель и соответствующие вероятности известны, то можно найти ожидаемые частоты

$$m_{(ожс)ij} = m \cdot P(A_i \cap B_j)$$

В частности, если теоретически признаки А и В **статистически независимы**, то $P(A_i \cap B_j) = P(A_i)P(B_j)$, и

$$m_{(ожс)ij} = m \cdot P(A_i)P(B_j)$$

При ограниченном объеме выборки m наблюдаемые и ожидаемые частоты могут отличаться. В качестве меры расхождения используется статистика

$$\chi^2 = \sum_{i,j} \frac{(m_{ij} - m_{(ожс)ij})^2}{m_{(ожс)ij}}$$

Теорема (К.Пирсон, Р.Фишер). Если верна модель, по которой рассчитаны теоретические частоты $m_{(ожс)ij}$, то при неограниченном росте числа наблюдений распределение случайной величины $\chi^2 = \sum_{i,j} \frac{(m_{ij} - m_{(ожс)ij})^2}{m_{(ожс)ij}}$ стремится к распределению хи-квадрат. Число степеней свободы этого распределения определяется как разность между возможным числом событий

и числом связей, налагаемых моделью, и равно

$k = (i_{\max} - 1)(j_{\max} - 1)$.

Если расхождение наблюдаемых и ожидаемых частот не очень большое, величина χ^2 также не будет очень большой, и признаки можно считать независимыми (расхождения **не являются значимыми**). В том случае, если значение статистики велико χ^2 , расхождения между теоретическими и наблюдаемыми частотами **значимы.** А это повод говорить о наличии взаимосвязи между признаками и проводить более глубокое исследование – определять меру связи, искать функцию связи и т.д.

План проверки гипотезы о независимости признаков следующий:

1. **Выдвигаются нулевая и альтернативная гипотезы**

   H_0: признаки А и В генеральной совокупности статистически независимы

   H_1: между признаками А и В генеральной совокупности существует статистическая зависимость

2. **Строится таблица сопряженности признаков, вычисляются статистика критерия $\chi^2 = \sum_{i,j} \frac{(m_{ij} - m_{(ожс)ij})^2}{m_{(ожс)ij}}$ и число степеней свободы $k = (i_{\max} - 1)(j_{\max} - 1)$**
3. **Вычисляется соответствующее ей значение значимости $\alpha = 1 - \mathrm{CDF.CHISQ}(\chi^2, k)$**
4. **Полученное значение α сравнивается с заданным критическим (например, если задать 95% доверительный интервал, то это значение будет равно 0.05)**
5. **Принимается решение: если $\alpha > 0.05$ гипотеза «принимается», если $\alpha \le 0.05$ гипотеза «отклоняется»**

Примечание.

1) **как правило**, основные показатели (признаки, от которого логически могут зависеть другие признаки), должны быть строковыми переменными, а данные, характеризующие эти основные признаки – столбцовыми переменными.

2) Для корректного анализа объем выборки и число ожидаемых или наблюдаемых частот в каждой ячейке должно быть достаточно велико. На практике считается, что объем выборки должен быть **не менее 30**, а частота **не менее 5**. Ограничения в использовании этих коэффициентов соответствуют ограничениям критерия хи-квадрат Пирсона

3) Если таблица имеет размер 2x2 или минимальная частота менее 10, вводится поправочный коэффициент «поправка на непрерывность»

Пример. По результатам наблюдений получены следующие данные о цвете глаз 36 студентов

		цвет_глаз		
		карие	голубые	Всего
пол	"Ж"	7	8	15
	"М"	5	16	21
Всего		12	24	36

Используя критерий хи-квадрат, проверить гипотезу о независимости цвета глаз от пола студента.

Признак А – пол студента - может принимать два значения A_1 - женский и A_2 - мужской. Вероятности этих событий оценим по выборке

$$P(A_1)=\frac{15}{36},\quad P(A_2)=\frac{21}{36}$$

Признак В – цвет глаз студента - может принимать два значения B_1 - карие и B_2 -голубые. Вероятности этих событий также оценим по выборке

$$P(B_1)=\frac{11}{36},\quad P(B_2)=\frac{25}{36}$$

Тогда ожидаемые частоты $m_{(ож)ij}=m\cdot P(A_i)P(B_j)$ равны

$$m_{(ож)11}=m\cdot P(A_1)P(B_1)=36\cdot\frac{15}{36}\cdot\frac{12}{36}=5$$

$$m_{(ож)12}=m\cdot P(A_1)P(B_2)=36\cdot\frac{15}{36}\cdot\frac{24}{36}=10$$

$$m_{(ож)21}=m\cdot P(A_2)P(B_1)=36\cdot\frac{21}{36}\cdot\frac{12}{36}=7$$

$$m_{(ож)22}=m\cdot P(A_2)P(B_2)=36\cdot\frac{21}{36}\cdot\frac{24}{36}=14$$

Наблюдаемые частоты равны $m_{11}=7,\ m_{12}=8,\ m_{21}=5,\ m_{12}=16$

Вычислим значение статистики $\chi^2=\sum_{i,j}\frac{(m_{ij}-m_{(ож)ij})^2}{m_{(ож)ij}}$

$$\chi^2=\frac{(7-5)^2}{5}+\frac{(8-10)^2}{10}+\frac{(5-7)^2}{7}+\frac{(16-14)^2}{14}\approx 2,05714$$

Найдем число степеней свободы $k=(i_{\max}-1)(j_{\max}-1)=(2-1)(2-1)=1$

Найдем соответствующее значение

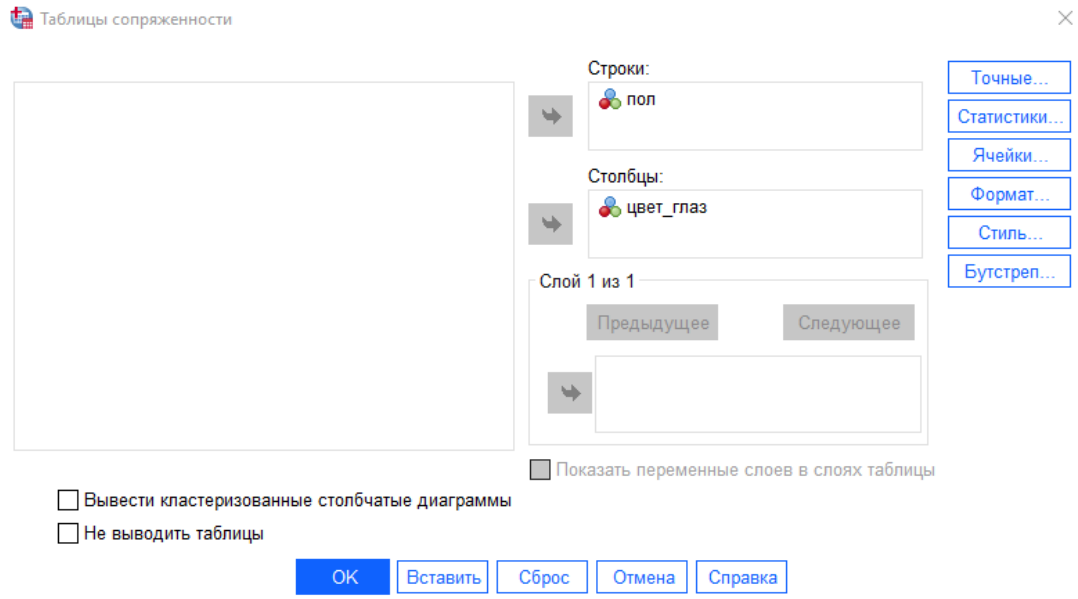
$$\alpha=1-\mathrm{CDF.CHISQ}(\chi^2,k)=1-\mathrm{CDF.CHISQ}(2.05714,1)\approx 0,15149$$

Поскольку 0,151 >0,05, гипотезу о независимости признаков пол - цвет глаз на уровне значимости 0, 05 принимаем

В SPSS для создания таблиц сопряженности и вычислении теста хи-квадрат используются опции

Анализ → ***Описательные статистики*** → ***Таблицы сопряженности***

Диалоговое окно выглядит следующим образом



Отметим в **Статистиках** «Хи-квадрат»

Отметим в **Ячейках** «Количества» наблюденные, ожидаемые

«Остатки» нестандартизированные

→ ***OK***

Таблицы сопряженности

Сводный отчет по наблюдениям

	Наблюдения					
	Допустимо		Пропущенные		Всего	
	N	Проценты	N	Проценты	N	Проценты
пол * цвет_глаз	36	100,0%	0	0,0%	36	100,0%

Таблица сопряженности пол * цвет_глаз

			цвет_глаз		
			карие	голубые	Всего
пол	"Ж"	Количество	7	8	15
		Ожидаемое количество	5,0	10,0	15,0
		Остаток	2,0	-2,0	
	"М"	Количество	5	16	21
		Ожидаемое количество	7,0	14,0	21,0
		Остаток	-2,0	2,0	
Всего		Количество	12	24	36
		Ожидаемое количество	12,0	24,0	36,0

Критерии хи-квадрат

	Значение	ст.св.	Асимптотическая значимость (2-сторонняя)	Точная знч. (2-сторонняя)	Точная знч. (1-сторонняя)
Хи-квадрат Пирсона	2,057[a]	1	,151		
Поправка на непрерывность[b]	1,157	1	,282		
Отношения правдоподобия	2,049	1	,152		
Точный критерий Фишера				,175	,141
Линейно-линейная связь	2,000	1	,157		
Количество допустимых наблюдений	36				

a. Для числа ячеек 0 (0,0%) предполагается значение, меньше 5. Минимальное предполагаемое число равно 5,00.

b. Вычисляется только для таблицы 2x2

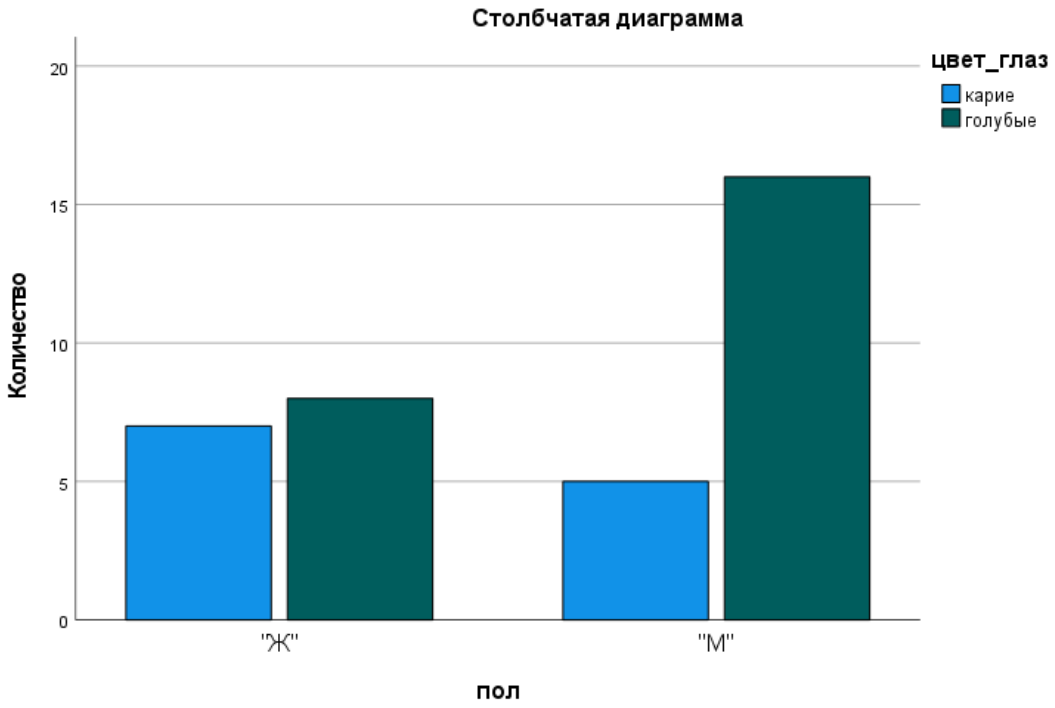
Чтобы увидеть в таблице не только абсолютные, но и относительные (процентные) значения, необходимо в диалоговом окне «Таблицы сопряженности: Вывод в ячейках» проставить

галочки напротив соответствующих окошек в разделе «Проценты». Здесь можно задать вывод процентов по строке, по столбцу и в целом по таблице.

Для наглядности можно построить столбчатые диаграммы. Применим опции ***Анализ*** → ***Описательные статистики*** → ***Таблицы сопряженности,*** отметим опции

Вывести кластеризованные столбчатые диаграммы и **Не выводить таблицы**. Нажав на ***ОК,***

получим столбчатые диаграммы, отражающие распределение



(II) Ранговая корреляция (связь порядковых признаков)

Представим себе, что теперь мы имеем дело с двумя признаками A и B, измерения которых проведены в порядковой шкале. Нас интересует, как влияет величина одного признака на степень выраженности другого. Если такого влияния нет, признаки естественно назвать независимыми. Как проверить гипотезу о независимости порядковых признаков (гипотезу H_0)? Первым решение этой задачи предложил психолог Чарльз Эдвард Спирмен в 1904 г. Он предложил метод, по которому количественно можно определить взаимосвязь двух порядковых признаков.

Каждому значению признака A ставится в соответствие его ранг r_i. Например, ранг «наименьшего» значения равен 1, следующего по порядку – 2. То есть, если выстроить все значения «в порядке возрастания», то ранг будет совпадать с порядковым номером этого

значения. Аналогично каждому значению признака B ставится в соответствие его ранг s_i. (Ранги можно присваивать и в порядке убывания!) Теперь каждому объекту можно поставить в соответствие не только пару значений признаков (A_i, B_i), но и пару значений их рангов (r_i, s_i). Если признаки A и B взаимосвязаны, то последовательности рангов $r_1, r_2, \ldots$ и $s_1, s_2, \ldots$ в какой-то мере взаимосвязаны.

Коэффициент ранговой корреляции Спирмена. Близость двух рядов чисел $r_1, r_2, \ldots$ и $s_1, s_2, \ldots$ отражает величина

$$S = \sum_{i=1}^{n} (r_i - s_i)^2$$

Она принимает значение $S = 0$ тогда и только тогда, когда последовательности полностью совпадают. Наибольшее возможное значение равно $S = \frac{1}{3}\left(n^3 - n\right)$ и соответствует случаю, когда последовательности рангов полностью противоположны. В том и другом случае взаимосвязь («прямая» либо «обратная») проявляется максимально. Коэффициент ранговой корреляции Спирмена равен

$$\rho = 1 - \frac{6S}{n^3 - n}$$

1. По абсолютной величине этот коэффициент ограничен единицей $|\rho| \le 1$, свои крайние значения $\rho = \pm 1$ он принимает в случаях полной предсказуемости одной ранговой последовательности по другой.
2. Минимальное абсолютное значение $\rho = 0$ соответствует отсутствию связи между признаками (при этом $S = 0$).
3. Чем больше величина ρ, тем более выражена связь между признаками.
4. Если $\rho > 0$, говорят о **прямой** корреляционной связи, если $\rho < 0$ - об **обратной** корреляционной связи

При использовании коэффициента ранговой корреляции условно оценивают тесноту связи между признаками, например, по шкале Чеддока:

Абсолютное значение ρ	Теснота (сила) корреляционной связи
менее 0.3	слабая
от 0.3 до 0.5	умеренная
от 0.5 до 0.7	заметная
от 0.7 до 0.9	высокая
более 0.9	весьма высокая

При проверке нулевой гипотезы об отсутствии взаимосвязи предполагается, что при большом объеме выборки *n* статистика $t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$ имеет распределение Стьюдента с $k = n-2$ степенями свободы/

Коэффициент ранговой корреляции применяется также для количественных признаков.

Коэффициент ранговой корреляции Кендалла. Другой коэффициент ранговой корреляции получил популярность после работ М. Кендэлла. Этот коэффициент в качестве меры сходства между двумя ранжировками использует **число инверсий** К - минимальное число перестановок соседних объектов, которые надо сделать, чтобы упорядочить объекты. Наименьшее возможное значение $K = 0$, наибольшее $K = \frac{n(n-1)}{2}$. Как и для *S*, эти значения получаются при полном совпадении и полной противоположности ранговых последовательностей.

Коэффициент ранговой корреляции Кендалла находится по формуле

$$\tau = 1 - \frac{4K}{n(n-1)}$$

1. По абсолютной величине этот коэффициент также ограничен единицей $|\tau| \le 1$, свои крайние значения $\tau = \pm 1$ он принимает в случаях полной предсказуемости одной ранговой последовательности по другой.
2. Минимальное абсолютное значение $\tau = 0$ соответствует отсутствию связи между признаками (при этом $K = 0$).
3. Чем больше величина τ, тем более выражена связь между признаками.
4. Если $\tau > 0$, говорят о **прямой** корреляционной связи, если $\tau < 0$ - об **обратной** корреляционной связи

В качестве примера использования ранговых коэффициентов Спирмена и Кендалла рассмотрим следующую задачу

Знания десяти студентов проверены по двум тестам: A и B. Оценки по стобалльной системе оказались следующими (в первой строке указано количество баллов по тесту A, а во второй — по тесту B):

95	90	86	84	75	70	62	60	57	50
92	93	83	80	55	60	45	72	62	70

Найти выборочный коэффициент ранговой корреляции Спирмена между оценками по двум тестам.

Р е ш е н и е. Присвоим ранги x_i оценкам по тесту A. Эти оценки расположены в убывающем порядке, поэтому их ранги x_i равны порядковым номерам:

ранги x_i	1	2	3	4	5	6	7	8	9	10
оценки по тесту A	95	90	86	84	75	70	62	60	57	50

Присвоим ранги y_i оценкам по тесту B, для чего сначала расположим эти оценки в убывающем порядке и пронумеруем их:

1	2	3	4	5	6	7	8	9	10	(**)
93	92	83	80	72	70	62	60	55	45	

Выпишем последовательности рангов x_i и y_i:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	1	3	4	9	8	10	5	7	6

Найдем искомый коэффициент ранговой корреляции Спирмена, учитывая, что $n=10$:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}, = 1 - \frac{6 \cdot 60}{10^3 - 10} = 0{,}64.$$

Найдем теперь для этой задачи коэффициент Кендалла. Условие то же:

95	90	86	84	75	70	62	60	57	50
92	93	83	80	55	60	45	72	62	70

Найти выборочный коэффициент ранговой корреляции Кендалла между оценками по двум тестам.

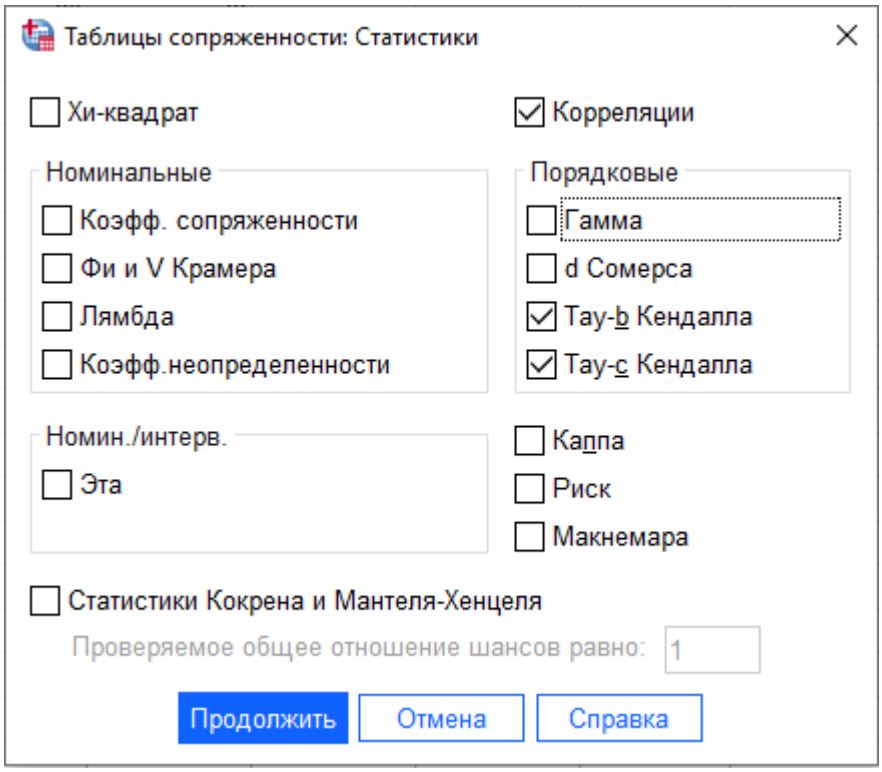
Выпишем последовательности рангов x_i и y_i:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	1	3	4	9	8	10	5	7	6

Число инверсий (нарушений порядка, когда большее число стоит слева от меньшего равно 1(для2)+4(для 9)+3(для 8)+3(для 10)+ 1 (для 7)=12)

$$\tau = 1 - \frac{4K}{n(n-1)} = 1 - \frac{4 \cdot 12}{10 \cdot 9} \approx 0{,}47, \quad 1{,}5 * \tau = 0.7 \approx 0.64$$

Коэффициенты ранговой корреляции вычисляются в SPSS с помощью последовательности функций ***Анализ*** → ***Описательные статистики*** → ***Таблицы сопряженности*** → ***Статистики***



Получаем таблицы

Таблицы сопряженности

Сводный отчет по наблюдениям

	Наблюдения					
	Допустимо		Пропущенные		Всего	
	N	Проценты	N	Проценты	N	Проценты
оценка1 * оценка2	10	100,0%	0	0,0%	10	100,0%

Симметричные меры

		Значение	Асимптотическая среднеквадратичная ошибка[a]	Приблизительная T[b]	Приблизительная значимость
Порядковый/порядковый	Тау-b Кендалла	,467	,228	2,044	,041
	Тау-c Кендалла	,467	,228	2,044	,041
	Корреляция Спирмена	,636	,260	2,333	,048[c]
Интервал/интервал	R Пирсона	,691	,129	2,704	,027[c]
Количество допустимых наблюдений		10			

a. Не предполагая нулевой гипотезы.

b. Использование асимптотической среднеквадратичной ошибки в предположении нулевой гипотезы.

c. Основано на нормальной аппроксимации.

Значение коэффициента Спирмена 0,634 свидетельствует о заметной прямой корреляционной связи между оценками. Небольшая значимость 0, 48 – о том, что гипотезу об отсутствии взаимосвязи следует отклонить. Аналогичные выводы следуют из значений коэффициентов Кенделла. Тау-b Кендалла - непараметрическая мера корреляции для порядковых или ранговых переменных, которая **учитывает возможные совпадения** значений (связи). Тау-с Кендалла. - **игнорирует возможные совпадения** значений (связи).