# ЛАБОРАТОРНАЯ РАБОТА № 7
# ВЫЯВЛЕНИЕ СВЯЗИ МЕЖДУ ПРИЗНАКАМИ

Методы определения связи признаков заметно отличаются в зависимости от вида шкалы измерений этих признаков:

- для изучения связи качественных признаков, измеренных в номинальной шкале (например, признаков вида «да» или «нет») применяются таблицы сопряженности, статистика Фишера-Пирсона $\chi^2$, различные меры связи признаков (коэффициенты Юла, Крамера и др.) и логарифмические линейные модели;
- для признаков, измеренных в порядковой шкале — данных типа «лучше-хуже», тестовых баллов и т.д. — применяются ранжирование и коэффициенты корреляции Спирмена и Кендэлла;
- для данных, измеренных в количественных шкалах, применяются выборочные коэффициенты корреляции и модель простой линейной регрессии.

## КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ПОКАЗАТЕЛЕЙ ДЕЯТЕЛЬНОСТИ ПЕСЧАНЫХ КАРЬЕРОВ

**Задача.** Деятельность 8 карьеров характеризуется себестоимостью 1 т. песка ($x^{(1)}$), сменной добычей песка ($x^{(2)}$), и фондоотдачей ($x^{(3)}$). Значения показателей представлены в таблице

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x^{(1)}$ (тыс. руб) | 30 | 20 | 40 | 35 | 45 | 25 | 50 | 30 |
| $x^{(2)}$ (т.) | 20 | 30 | 50 | 70 | 80 | 20 | 90 | 25 |
| $x^{(3)}$ (%) | 20 | 25 | 20 | 15 | 10 | 30 | 10 | 20 |

Требуется в предположении нормальности распределения трехмерной случайной величины ($x^{(1)}$, $x^{(2)}$, $x^{(3)}$) построить корреляционную матрицу, найти частные коэффициенты корреляции.

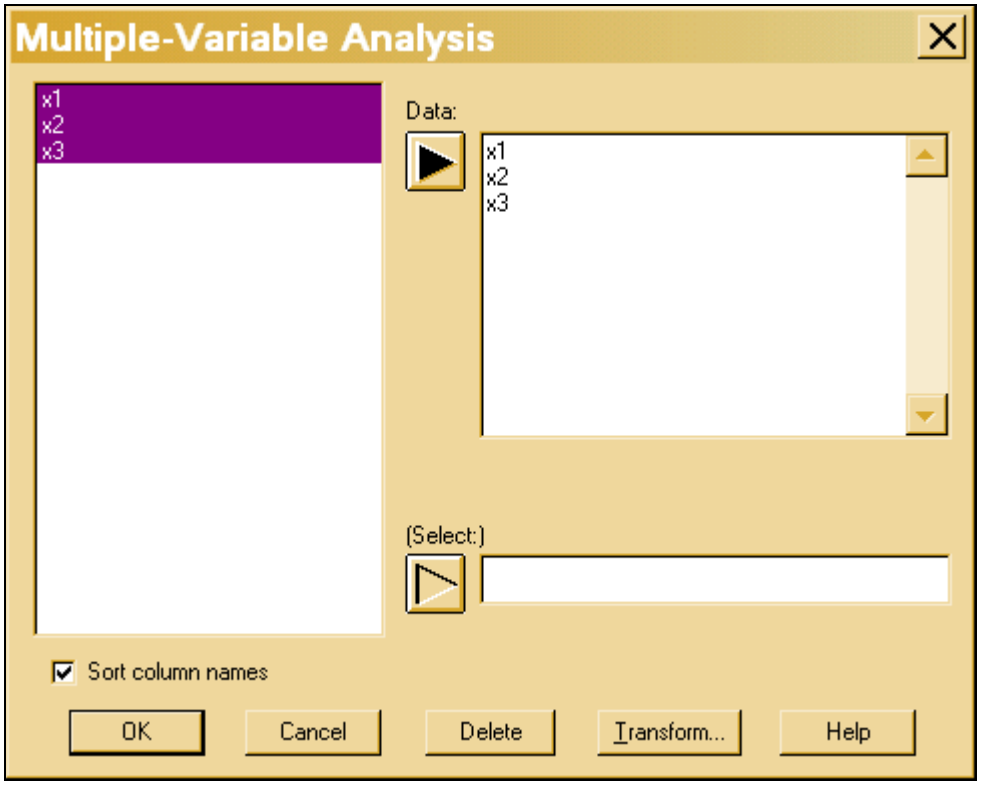
Здесь данные измерены в количественных шкалах, применим исследование с помощью коэффициента корреляции Пирсона.

Введите данные, должна получиться следующая таблица:

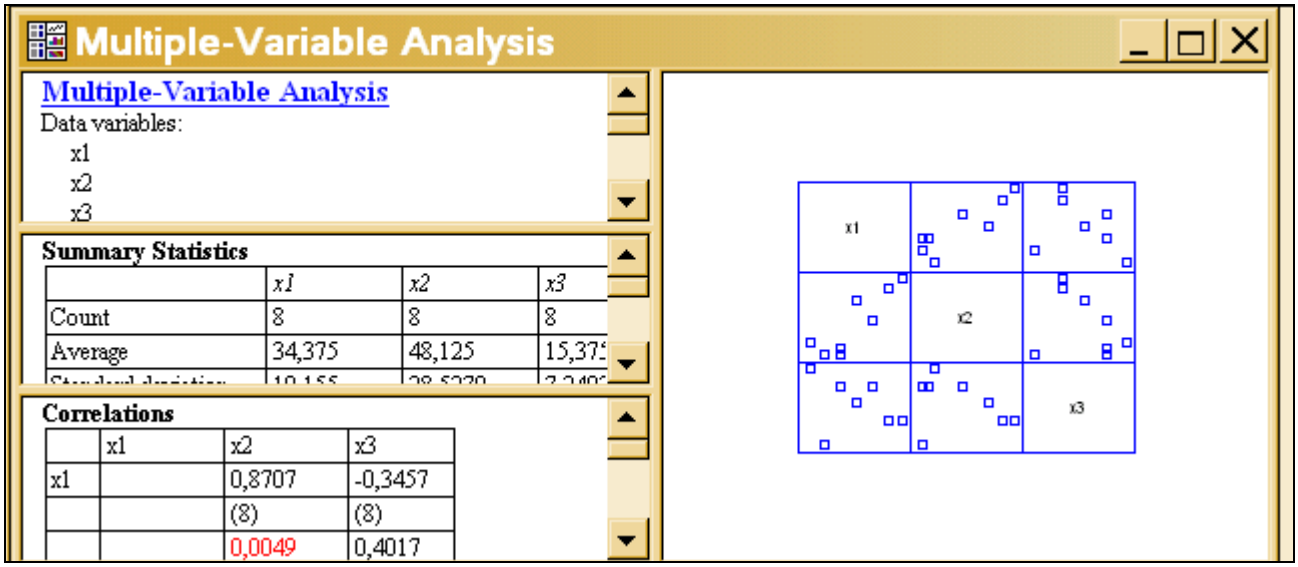
| | x1 | x2 | x3 | Col |
|---|---|---|---|---|
| 1 | 30 | 20 | 20 | |
| 2 | 20 | 30 | 25 | |
| 3 | 40 | 50 | 20 | |
| 4 | 35 | 70 | 15 | |
| 5 | 45 | 80 | 10 | |
| 6 | 25 | 20 | 3 | |
| 7 | 50 | 90 | 10 | |
| 8 | 30 | 25 | 20 | |
| 9 | | | | |

A B C D E F G H I J

В строке меню выберите ***Describe***, в раскрывшемся меню выберите ***Numeric Data***, затем ***Multiple-Variable Analysis***. Раскроется окно, в котором в поле ***Data*** необходимо перевести названия всех рассматриваемых столбцов,

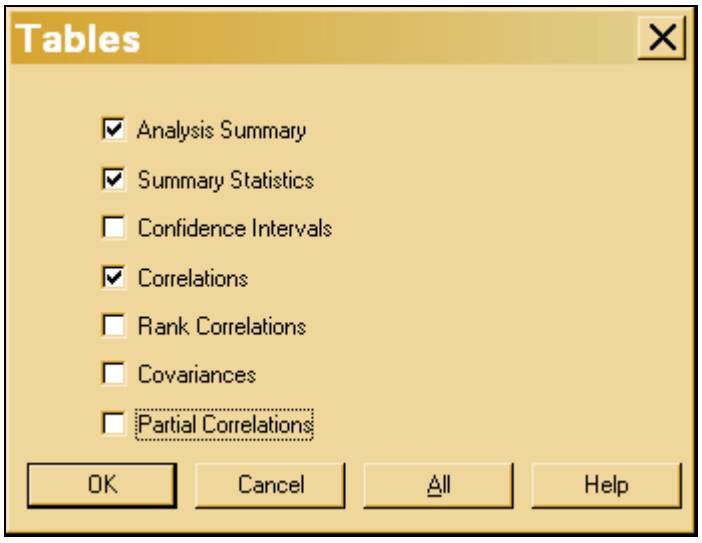



затем нажмите ОК. Раскроется окно ***Multiple-Variable Analysis***.




Обратите внимание на диаграмму рассеяния. Точки представляют из себя облачко, если оно вытянуто вдоль оси – значит, есть связь. В данном случае видно, что есть связь между x1 и x2. Связь между x1 и x3 тоже наблюдается, но наклон линии в другую сторону. Это – пример отрицательной связи.

Щелкните дважды по окну ***Correlations,*** если оно есть на экране. В противном случае нажмите кнопку ***Tables***, в диалоговом окне выберите ***Correlations***.




Раскройте окно ***Correlations,*** оно будет выглядеть следующим образом:

**Multiple-Variable Analysis**

**Correlations**

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | | 0,8707 | -0,8737 |
| | | (8) | (8) |
| | | 0,0049 | 0,0046 |
| x2 | 0,8707 | | -0,8789 |
| | (8) | | (8) |
| | 0,0049 | | 0,0040 |
| x3 | -0,8737 | -0,8789 | |
| | (8) | (8) | |
| | 0,0046 | 0,0040 | |

Correlation
(Sample Size)
P-Value

Эта таблица показывает коэффициенты корреляции Пирсона между каждой парой переменных. Здесь под коэффициентом корреляции в скобках стоит объем выборки, а ниже p-value данного коэффициента корреляции. Значение *p-value*, меньшее 0,05, означает статистическую значимость с 95 % доверительным интервалом. Эти *p-value* отображаются на экране красным цветом. Следующие пары переменных имеют *p-value* меньшее 0,05: $x_1$ и $x_2$, $x_1$ и $x_3$, $x_2$ и $x_3$.Коэффициент корреляции между: $x_1$ и $x_3$ отрицателен и статистически значим, следовательно, можно сделать вывод, что с ростом фондоотдачи себестоимость песка уменьшается.

Построим матрицу частных коэффициентов корреляции. Для этого нажмите кнопку ***Tables,*** выберите ***Partial Correlations.*** Перед вами раскроется окно:

**Partial Correlations**

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | | 0,4428 | -0,4622 |
| | | (8) | (8) |
| | | 0,3198 | 0,2964 |
| x2 | 0,4428 | | -0,4942 |
| | (8) | | (8) |
| | 0,3198 | | 0,2596 |
| x3 | -0,4622 | -0,4942 | |
| | (8) | (8) | |
| | 0,2964 | 0,2596 | |

Correlation
(Sample Size)
P-Value

Эта таблица показывает частные коэффициенты корреляции. Например, частный коэффициент $r_{13}$ характеризует степень тесноты линейной связи между $x_1$ и $x_3$ при исключенном влиянии фиксированной $x_2$. В скобках под коэффициентом частной корреляции стоит размер выборки, а ниже - *p-value*.

Сравним парный коэффициент корреляции между $x_1$ и $x_3$ и частный коэффициент корреляции $r_{13}$. Так как | -0,8737| > | -0,4622|, то можно утверждать, что $x_2$.усиливает тесноту связи между $x_1$ и $x_3$. Аналогично можно исследовать связь между остальными переменными.

Для того чтобы лучше понять, для чего нужен частный коэффициент корреляции, рассмотрим пример. Откройте файл E_H.sf. В нем находятся следующие данные:

- В столбце Year – год, с 1057 по 1966
- В столбце H – данные о рекордах по прыжкам в высоту с шестом по годам;
- В столбце E – данные о производстве электроэнергии в США.

Есть ли связь между производством электроэнергии и рекордам по прыжкам в высоту?

Вычислите самостоятельно коэффициент корреляции. Получится результат

**Correlations**

| | E | H | Year |
|---|---|---|---|
| E | | 0,9487 | 0,9905 |
| | | (10) | (10) |
| | | 0,0000 | 0,0000 |
| H | 0,9487 | | 0,9330 |
| | (10) | | (10) |
| | 0,0000 | | 0,0001 |
| Year | 0,9905 | 0,9330 | |
| | (10) | (10) | |
| | 0,0000 | 0,0001 | |

Correlation
(Sample Size)

Видим, что коэффициент корреляции между H и E равен 0,9487 (близок к единице) и статистически значим (p-value равно 0). Следовательно, имеется связь

4

между прыжками в высоту и производством электроэнергии. Результат достаточно странный. Найдем частный коэффициент корреляции между E и H при фиксированном параметре Year.

**Partial Correlations**

| | E | H | Year |
|---|---|---|---|
| E | | 0,4954 | 0,9263 |
| | | (10) | (10) |
| | | 0,1750 | 0,0003 |
| H | 0,4954 | | -0,1536 |
| | (10) | | (10) |
| | 0,1750 | | 0,6933 |
| Year | 0,9263 | -0,1536 | |
| | (10) | (10) | |
| | 0,0003 | 0,6933 | |

Correlation

Видим, что частный коэффициент равен 0,4954, он достаточно мал, p-value=0,1750 > 0,1 и это говорит о незначимости коэффициента. Следовательно, можно считать, что есть связь между ростом рекорда по прыжкам в зависимости от года и ростом производства электроэнергии **в зависимости от года** (т.е. производство электроэнергии растет с каждым годом и рекорды по прыжкам тоже растут из года в год). Если же убрать зависимость от времени, то связи между рекордом по прыжкам в высоту и производством электроэнергии нет.

## АНАЛИЗ СВЯЗИ ПО ТАБЛИЦАМ СОПРЯЖЕННОСТИ

**Задача.** По данным переписи населения 1939, 1959 и 1970 гг. получены следующие выборочные данные об образовании городских и сельских жителей.

| Образование | 1939 | | 1959 | | 1970 | |
|---|---|---|---|---|---|---|
| | город | село | город | село | город | село |
| Высшее или среднее | 107 | 31 | 376 | 210 | 800 | 358 |
| Начальное | 453 | 1094 | 621 | 879 | 552 | 703 |

1. Можно ли говорить о наличии связи между уровнем образования и местом проживания?

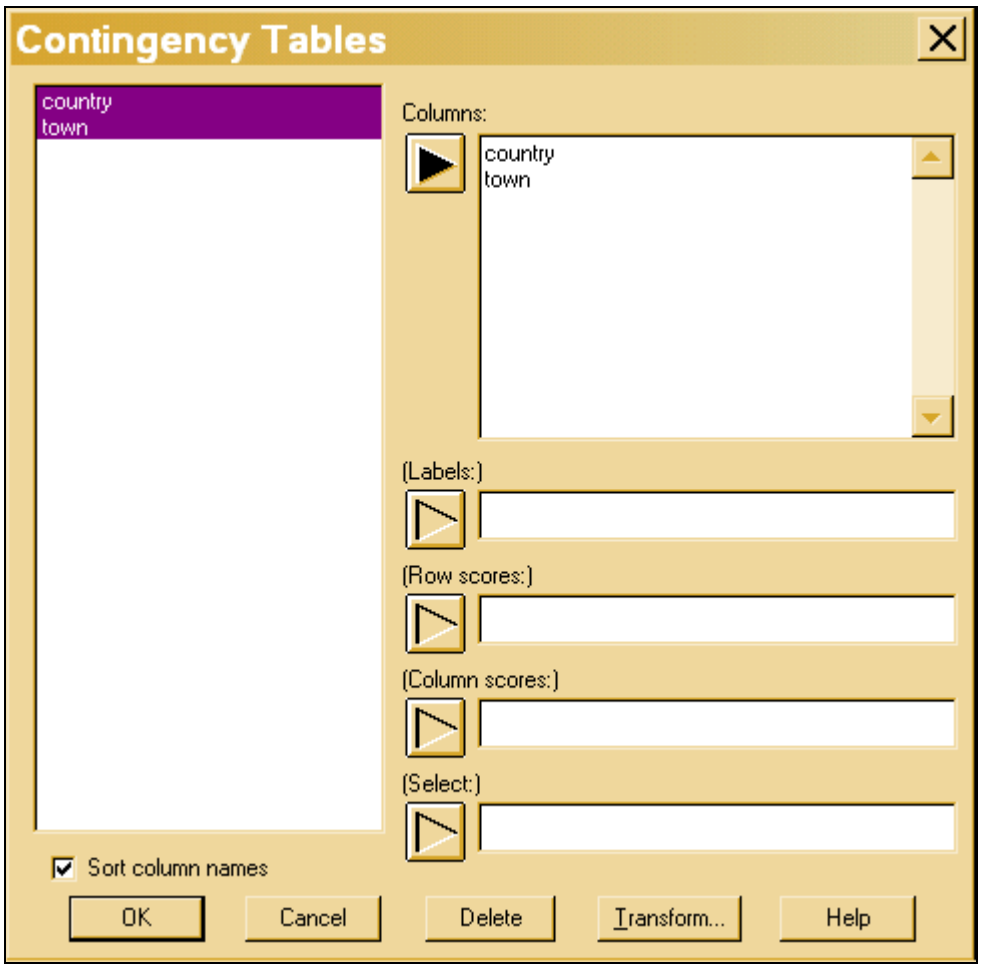
2. Если выборочные данные доказывают наличие такой связи, то можно ли проследить, как изменялась теснота этой связи в динамике (в 1939, 1959 и 1970 гг)?
3. По данным, представленным в следующей таблице, проанализируйте зависимость образовательного уровня мужчин и женщин от места их проживания в 1970 г.:

| Образование | Мужчины | | Женщины | |
|---|---|---|---|---|
| | город | село | город | село |
| Высшее или среднее | 388 | 188 | 412 | 170 |
| Начальное | 238 | 297 | 314 | 406 |

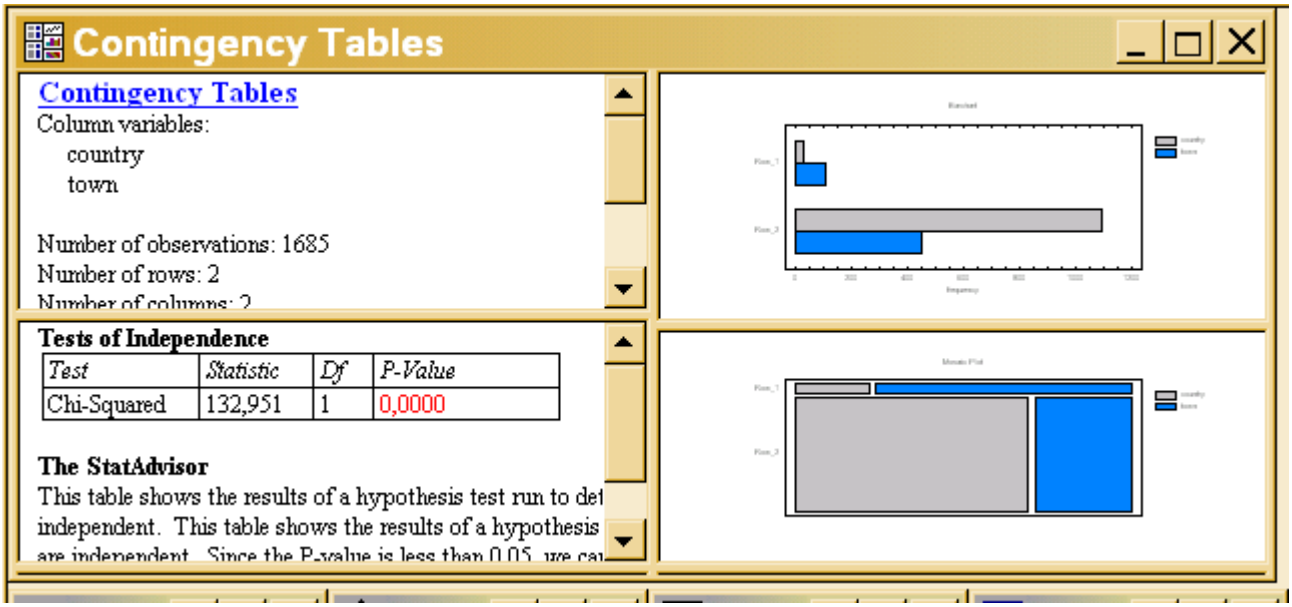
Введите данные в таблицу следующим образом:

| | town | country |
|---|---|---|
| | | |
| 1 | 107 | 31 |
| 2 | 453 | 1094 |
| 3 | | |

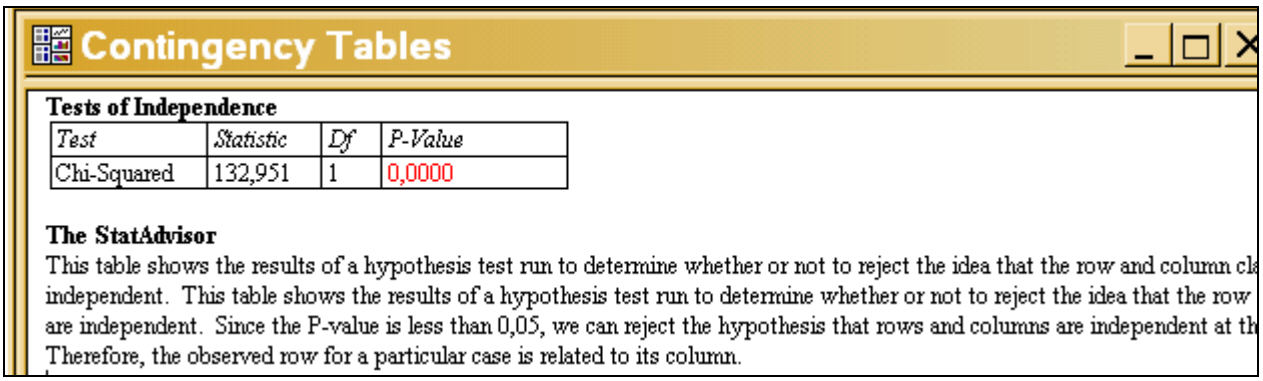
В строке меню выберите ***Describe***, в строке меню выберите ***Categorical Data***, затем ***Contingency Table***. Раскроется окно, в котором последовательно надо выбирать названия столбцов и нажимать на кнопку со стрелкой в поле ***Columns***.




Нажмите кнопку OK. Откроется окно ***Contingency Tables***.




Щелкните дважды по окну **Tests of Independence**, если оно есть на экране. В противном случае щелкните по кнопке ***Tables***, выберите **Tests of Independence.** ***Chi-Square Test***.


Contingency Tables

**Tests of Independence**

| *Test* | *Statistic* | *Df* | *P-Value* |
|---|---|---|---|
| Chi-Squared | 132,951 | 1 | 0,0000 |

**The StatAdvisor**

This table shows the results of a hypothesis test run to determine whether or not to reject the idea that the row and column cl
independent. This table shows the results of a hypothesis test run to determine whether or not to reject the idea that the row
are independent. Since the P-value is less than 0,05, we can reject the hypothesis that rows and columns are independent at th
Therefore, the observed row for a particular case is related to its column.

В ***StatAdvisor*** можно прочитать, что $\chi^2$ -тест представляет собой проверку гипотезы о том, можно или нет отклонить гипотезу о независимости столбцов и строк. Так как *p-value* меньше чем 0,01, то мы можем отклонить гипотезу о независимости с 99 % уровнем доверия.

Для анализа тесноты связи нажмите снова на кнопку ***Tabular Options,*** в раскрывшемся окне выберите ***Summary Statistics***. Раскроется следующее окно:

**Summary Statistics**

| | | *With Rows* | *With Columns* |
|---|---|---|---|
| *Statistic* | *Symmetric* | *Dependent* | *Dependent* |
| Lambda | 0,1089 | 0,0000 | 0,1357 |
| Uncertainty Coeff. | 0,0806 | 0,1308 | 0,0583 |
| Somer's D | -0,2443 | -0,1635 | -0,4825 |
| Eta | | 0,2809 | 0,2809 |

| *Statistic* | *Value* | *P-Value* | *Df* |
|---|---|---|---|
| Contingency Coeff. | 0,2704 | | |
| Cramer's V | 0,2809 | | |
| Conditional Gamma | -0,7858 | | |
| Pearson's R | -0,2809 | 0,0000 | 1683 |
| Kendall's Tau b | -0,2809 | 0,0000 | |
| Kendall's Tau c | -0,1451 | | |

Посмотрите значения **коэффициента сопряженности** (*Contingency Coeff.*) и **коэффициента Крамера** (*Cramer's V*). Эти коэффициенты всегда принимают значения от 0 до 1: они равны 0 в случае отсутствия связи между признаками и возрастают с увеличением тесноты связи. Они нужны для того чтобы сравнивать тесноту связи. Запишите в тетрадь коэффициент Крамера. Повторите анализ тесноты связи для остальных данных обеих таблиц. Сформулируйте выводы (растет ли теснота связи или уменьшается? Как можно это интерпретировать?). Покажите результаты преподавателю.

# КОЭФФИЦИЕНТЫ РАНГОВОЙ КОРРЕЛЯЦИИ

**Задача.** Директор фирмы выставил оценки своим сотрудникам по двадцатибалльной системе, учитывая два признака:

1. Степень соответствия образования занимаемой ими в данной фирме должности.
2. Качество выполнения ими служебных обязанностей.

Получились следующие результаты:

| Фамилия | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* | *I* | *L* |
|---|---|---|---|---|---|---|---|---|---|---|
| Образование | 5 | 8 | 18 | 9 | 10 | 10 | 14 | 16 | 19 | 20 |
| Качество | 8 | 10 | 15 | 8 | 12 | 13 | 18 | 17 | 18 | 20 |

Используя ранговые критерии, выясните, влияет ли на качество выполнения служебных обязанностей образование по специальности, соответствующей должности. Какова направленность этой связи (прямая или обратная)?

| | Name | Level | Quality |
|---|---|---|---|
| 1 | A | 5 | 8 |
| 2 | B | 8 | 10 |
| 3 | C | 18 | 15 |
| 4 | D | 9 | 8 |
| 5 | E | 10 | 12 |
| 6 | F | 10 | 13 |
| 7 | G | 14 | 18 |
| 8 | H | 16 | 17 |
| 9 | I | 19 | 18 |
| 10 | L | 20 | 20 |

Введите данные в соответствии с таблицей. Должна получиться следующая таблица. В строке меню выберите ***Describe***, в раскрывшемся меню выберите ***Numeric Data***, затем ***Multivariable Analysis***. Нажмите кнопку ***Tables,*** в раскрывшемся окне выберите ***Rank Correlations***. Раскроется окно анализа:

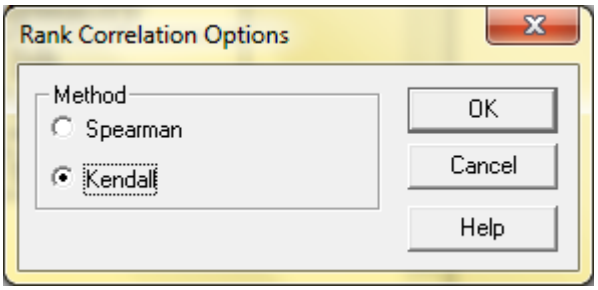
**Spearman Rank Correlations**

| | level | quality |
|---|---|---|
| level | | 0,9113 |
| | | (10) |
| | | 0,0063 |
| quality | 0,9113 | |
| | (10) | |
| | 0,0063 | |

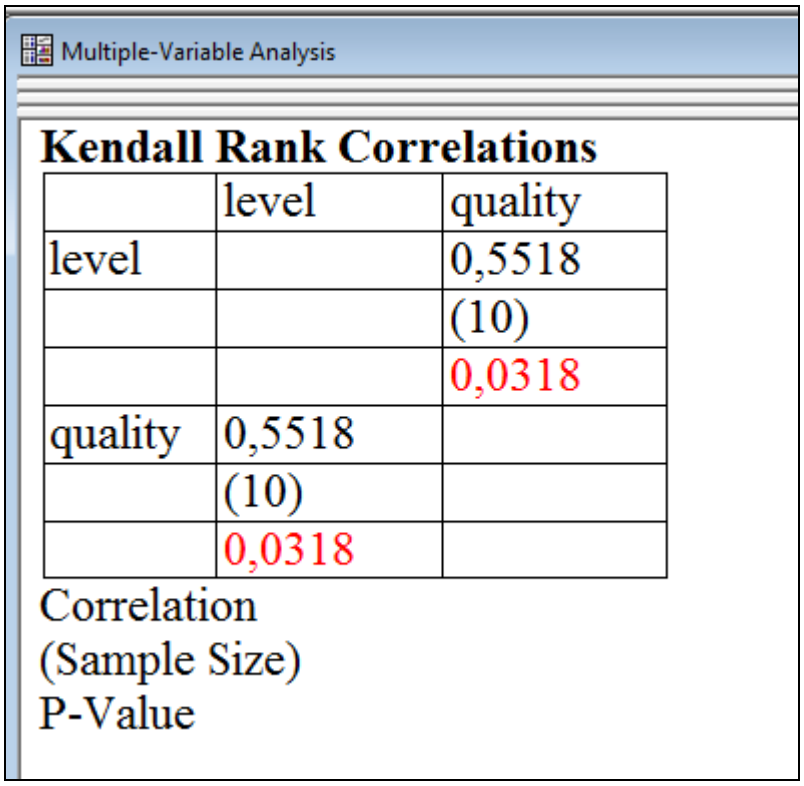
Correlation
(Sample Size)
P-Value

Эта таблица представляет нам матрицу ранговых коэффициентов корреляции Спирмена. Под коэффициентом в скобках стоит количество пар переменных, а ниже — значение *p-value*. Значение *p-value = 0,0063,* меньшее 0,05, означает

статистическую значимость коэффициента. (Здесь коэффициент статистически значим, так как 0,0063 < 0,05). Можно применить еще один метод – вычислить ранговый коэффициент Кендалла. Щелкните в окне правой кнопкой, выберите **Pane Options**, появится окно




Выберите в этом окне нужный вариант.




Известно, что коэффициент Кендалла всегда «более осторожный».

**ЗАДАНИЯ**

1. В таблице приведены результаты небольшого опроса о возможности в ближайшие 12 месяцев краха фондового рынка.

| | Акционеры | Не акционеры |
|---|---|---|
| Очень вероятно | 18 | 26 |
| Весьма вероятно | 41 | 65 |

| | | |
|---|---|---|
| Маловероятно | 52 | 68 |
| Невероятно | 19 | 31 |
| Не уверен | 8 | 13 |

Зависит ли ответ от того, является ли опрашиваемый акционером?

2. Влияет ли рост на быстроту бега?

| | Бегуны | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Рост (ранги) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Быстрота | 5 | 6 | 10 | 7 | 9 | 4 | 3 | 1 | 8 | 2 |

3. В таблице представлены темпы прироста (%) следующих макроэкономических показателей десяти развитых стран мира за 1992 г.: ВНП ($x^{(1)}$), промышленного производства ($x^{(2)}$), индекса цен ($x^{(3)}$) и доли безработных ($x^{(4)}$).

Хорошо бы эту задачу заменить!

| Страны | $x^{(1)}$ | $x^{(2)}$ | $x^{(3)}$ | $x^{(4)}$ |
|---|---|---|---|---|
| Япония | 3,5 | 4,3 | 2,1 | 2,3 |
| США | 3.1 | 4,6 | 3,9 | 6,3 |
| Германия | 2,2 | 2,0 | 3,4 | 5,1 |
| Франция | 2,7 | 3,1 | 2,9 | 9,7 |
| Италия | 2,7 | 3,0 | 5,6 | 11,1 |
| Великобритания | 1,6 | 1,4 | 4,0 | 9,5 |
| Канада | 3,1 | 3,4 | 3,0 | 10,0 |
| Австралия | 1,8 | 2,6 | 4,0 | 2,6 |
| Бельгия | 2,3 | 2,6 | 3,4 | 8,9 |
| Нидерланды | 2,3 | 2,4 | 3,5 | 6,4 |

Требуется:

а) найти оценку коэффициента корреляции между темпами прироста ВНП и промышленного производства, проверить его значимость;

б) оценить тесноту связи между $x^{(1)}$ и $x^{(3)}$, проверить значимость коэффициента корреляции;

в) влияет ли доля безработных на тесноту связи между промышленным производством и индексом цен?

4. По данным обследования получена информация о занятом населении по наличию второй работы и готовности к дополнительной занятости

**Таблица Данные о наличии работы и готовности к дополнительной занятости.**

| **Дополнительная занятость** | **Имеют работу** | | | |
|---|---|---|---|---|
| | Мужчины | | Женщины | |
| | **одну** | **две и более** | **одну** | **две и более** |
| Ищут | 212 | 29 | 145 | 20 |
| Не ищут | 2913 | 46 | 1915 | 45 |

Охарактеризуйте отдельно для мужчин и женщин связь поиска дополнительной занятости с наличием одной, двух и более видов работ. У кого связь теснее? Зависит ли от пола поиск работы среди имеющих одну? среди имеющих две и более?

## ВОПРОСЫ

1. С помощью какого критерия можно выявить связь между двумя количественными признаками?
2. Что характеризует выборочный коэффициент корреляции?
3. Что характеризует частный коэффициент корреляции?
4. Что Вы понимаете под порядковым признаком?
5. С помощью какого критерия можно выявить связь между двумя порядковыми признаками?
6. Для чего используются коэффициенты Спирмена и Кэнделла?
7. С помощью какого критерия можно выявить связь между двумя качественными признаками?
8. Что характеризует коэффициент Крамера?