**Анализ таблиц сопряженности признаков**

**Критерий независимости Хи-квадрат Пирсона**

К наиболее часто используемым инструментам изучения взаимосвязи двух переменных относятся методы анализа таблицы сопряженности. Анализ таблицы является весьма простым и наглядным, и вме­сте с тем эффективным инструментом изучения одновременно двух переменных. Таблица сопряжённости является наиболее универсальным средством изучения статистических связей, в ней могут быть представлены переменные, измеренные в любой шкале. Чтобы высказать предположение о зависимости признаков, следует анализировать не абсолютные частоты в таблице, а относительные, отнесенные либо к маргинальным частотам строк, либо к маргинальным частотам столбцов. Значительное расхождение в распределениях данных частот свидетельствует о наличии связи между признаками.

Основным критерием для проверки гипотезы о наличии связи между признаками на основе таблицы сопряженности является критерий Хи-квадрат Пирсона. Пусть таблица сопряженности двух признаков  и , содержащих, соответственно,  и  категорий имеет вид, представленный в таблице.

Таблица 2. Общий вид таблицы сопряженности двух признаков.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Относительные маргинальные частоты  и  оценивают, соответственно, одномерные распределения признаков  и . Если гипотеза «признаки  и независимы» верна, то оценкой вероятности события:  является произведение соответствующих маргинальных частот . Но тогда, величина  будет являться оценкой ожидаемой частоты для ячейки с индексами и  при объеме выборки .

Составим статистику:

,

которая характеризует сумму квадратов отклонений наблюдаемых значений  от ожидаемых по всем ячейкам. В соответствии с теоремой Пирсона, при истинности  данная статистика стремится к распределению  с  числом степеней свободы ( - это число неизвестных оцениваемых параметров двух распределений).

Возьмем в качестве критической точки  квантиль распределения Хи-квадрат уровня  с  числом степеней свободы. Тогда, если , то принимается гипотеза  о независимости двух признаков, в противном случае принимается гипотеза о наличии связи между двумя признаками.

Условие для корректного применения критерия Хи-квадрат остается прежним: желательно, чтобы ожидаемое значение  для каждой ячейки таблицы сопряженности было не меньше 5.

Заметим, что статистика критерия с точностью до обозначений совпадает со статистикой критерия Пирсона для проверки гипотезы однородности, то есть, по сути, это один и тот же критерий. Действительно, **любая задача однородности нескольких совокупностей может быть переформулирована в терминах независимости тех или иных признаков**. Допустим, мы проверяем гипотезу о равенстве (различии) средних доходов двух групп респондентов, которые различаются, скажем, по половому признаку. Данная гипотеза равенства (различия) средних доходов респондентов равносильна гипотезе о независимости (зависимости) средних доходов респондентов от пола респондента.

Помимо критерия Хи-квадрат Пирсона при анализе таблиц сопряженности используется также критерий Хи-квадрат, основанный на методе максимального правдоподобия. На практике статистика максимума правдоподобия Хи-квадрат очень близка по величине к рассмотренной статистике Пирсона Хи-квадрат.

Для [таблиц сопряженности 2x2](mk:@MSITStore:C:\PROGRA~1\StatSoft\STATIS~1\BasicStats.chm::/BasicStatistics/Overview/Crosstab/DescriptionofJointFrequencyDistributionsviaCrosstabulations2by2Tables.htm) используют критерий [Хи-квадрат](mk:@MSITStore:C:\PROGRA~1\StatSoft\STATIS~1\BasicStats.chm::/BasicStatistics/Overview/Crosstab/StatisticsinCrosstabulationsPearsonChisquare.htm) с поправкой Йетса на непрерывность, а также точный критерий Фишера. В случае зависимых выборок для таблиц сопряженности 2x2 используют критерий Mакнемара.

Для характеристики степени связи между признаками в таблице сопряженности наиболее часто используют коэффициент сопряженности Пирсона и, статистику V Крамера.

Коэффициент сопряженности Пирсона (коэффициент контингенции) вычисляется по формуле:

.

Значения коэффициента сопряженности лежат в пределах от 0 до 1. Чем ближе значения коэффициента сопряженности к 1, тем более зависимы признаки. Недостаток коэффициента сопряженности в том, что его максимальное значение зависит от размера таблицы. Этот коэффициент может достигать значения 1 только в случае, если число категорий не ограничено.

Более распространен коэффициент сопряженности Крамера V, вычисляемый по формуле:

,

где . Равно, как и значения коэффициента сопряженности Пирсона, значения коэффи­циента Крамера меняется от 0 до 1. Оба коэффициента принимают значение равное 0, при значении статистики , то есть при отсутствии связи между признаками. Однако, в отличие от коэффициента сопряженности Пирсона, который всегда меньше 1, коэффициент Крамера равен единице, в ситуации жестко детерминированной связи между перемен­ными.

Положив в формуле для статистики V Крамера , получим коэффициент сопряженности Фи, который используется при анализе таблиц сопряженности 2x2: .

**Анализ коэффициентов связи для количественных и порядковых данных**

В предыдущем разделе были рассмотрены коэффициенты связи между признаками, построенные на основе таблицы сопряженности. При этом не делалось никаких ограничений в отношении уровня измерения тех переменных, которые образуют таб­лицу. Не использовалась и информация о порядке следования градаций в переменных. Это является естественным для переменных, измеренных на номинальном уровне. Действительно, числовые категории, которые присваиваются градациям в таких переменных, имеют абсолютно условный смысл. Так, совершенно не имеет значения, присвоен ли в вопросе «Ваш пол» мужчинам код 0, 1 или 10. Главное, чтобы код, присвоенный мужчинам, отличался от кода, присвоенного женщинам. По этой причине то, что коэффициенты связи никак не реагируют на наш произвол в присвоении определенным градациям тех или иных числовых кодов, является вполне правильным для случая, когда исходные данные получены по номинальным шкалам. Однако, в случае анализа данных измеренных в порядковых или количественных шкалах, подобной характеристики связи не достаточно. Для такого рода переменных порядок расположения градаций уже существен, посколь­ку он фиксирует степень выраженности измеряемого свойства. Коэффициенты связи для порядковых и количественных данных, должны по возможности использовать всю содержащуюся в переменных информацию и отражать не только наличие связи, но и характер связи, ее направленность.

В качестве оценки степени связи между количественными признаками обычно используют значение выборочного коэффициента корреляции Пирсона:

,

где  - выборочная ковариация,

,  - выборочные дисперсии величин  и .

Выборочный коэффициент корреляции является оценкой коэффициента корреляции  генеральной совокупности, который характеризует степень линейной связи между признаками. Значения коэффициентакорреляции по модулю не превышают единицу: , причем значение коэффициента корреляции по модулю равно единице, только в случае линейной зависимости между  и . Если , то при возрастании одной переменной другая проявляет в среднем тенденцию также возрастать. Если , то при возрастании одной переменной другая проявляет в среднем тенденцию убывать. В первом случае говорят, что  и  положительно коррелированны, а во втором, соответственно, что  и отрицательно коррелированны. Если , то это означает только отсутствие линейной связи между значениями одной величины и средним значением другой величины, то есть, в общем случае, не говорит о отсутствии связи между  и .

Поскольку значения выборочного коэффициента корреляции , при условии , могут отличаться от нуля, необходима проверка значимости выборочного коэффициента корреляции. Для нормальной совокупности для проверки гипотезы  можно использовать Стьюдента, основанный на статистике:

.

Если , где  - квантиль распределения Стьюдента с числом степеней свободы  уровня , то гипотеза  отвергается и признается значимым отличие от нуля коэффициента корреляции . Заметим, что с точки зрения формулировки альтернативной гипотезы , следующие утверждения эквивалентны: «Коэффициент корреляции значимо отличается от нуля» и «Значение выборочного коэффициента корреляции значимо».

В качестве оценки степени связи между признаками, измеренными в порядковой шкале, используют коэффициенты ран­говой корреляции - Спирмена,  Кендэла,  Гудмена-Краскэла. Данные коэффициенты могут также применяться для характеристики степени связи между количественными признаками, если их распределение существенно отличается от нормального.

Коэффициент корреляции  Спирмена есть не что иное, как коэффициент корреляции Пирсона для случайных величин, представленных в ранговой шкале, и соответственно, характеризует силу линейной связи между рангами.

Пусть ,  - ранги элементов  в вариационных рядах  и  соответственно. Тогда:

,;

, .

Соответственно выборочный коэффициент корреляции Спирмена:

.

При наличии в рядах  и  совпадающих значений, им присваиваются ранги, равные среднеарифметическому соответствующих рангов в вариационном ряду и используют коэффициент корреляции Спирмена с поправками на равные ранги:

, , ,

где  - длина i-ой связи в ряду , - длина i-ой связи в ряду . Под связью понимается набор значений с одинаковыми рангами.

Проверку значимости коэффициента корреляции Спирмена осуществляют также, как и проверку значимости коэффициента корреляции Пирсона, используя статистику , закон распределения которой при условии истинности “корреляция между рангами равна нулю” приближенно описывается распределением Стьюдента, с числом степеней свободы . Можно также для проверки значимости  использовать то, что при условии истинности , закон распределения статистики  стремится к нормальному распределению с математическим ожиданием, равным нулю и дисперсией, равной . При небольших  для проверки значимости используют таблицы распределения статистики .

Коэффициент ранговой корреляции Кендалла вычисляется по формуле:

,

где  - число согласованных пар, то есть случаев, когда , при условии, что ;  - число несогласованных пар, то есть случаев, когда , при условии, что . Значения коэффициента корреляции Кендалла также, как и коэффициента корреляции Спирмена, лежат в интервале .

При условии истинности “корреляция между рангами равна нулю”, закон распределения статистики Кендалла стремится к нормальному распределению , что и используется для проверки гипотезы о значимости данного коэффициента. При наличии в рядах  и  совпадающих значений, им присваиваются ранги, равные среднеарифметическому соответствующих рангов в вариационном ряду, и используют коэффициент корреляции Кендалла с поправками на равные ранги:

, , , ,

где  длина i-ой связи в ряду ,  длина i-ой связи в ряду . Под связью понимается набор значений с одинаковыми рангами.

Если в данных имеется много совпадающих значений можно также использовать -статистику Гудмана-Краскала, которая вычисляется по формуле:

.

Эта мера имеет прямую вероятностную интерпретацию, поскольку она есть не что иное, как разность между относительными частотами правильного и неправильного порядка для двух совокупностей. При условии истинности “корреляция между рангами равна нулю”, закон распределения статистики  стремится к стандартному нормальному распределению, что и используется для проверки гипотезы о значимости данного коэффициента.

**Проверка гипотез о зависимости (независимости) признаков в пакете STATISTICA**

**Пример 1.** Результаты ответов 400 респондентов на вопросы анкеты «Томск 400» В\_13: «Как Вы оцениваете Ваше здоровье в сравнении со здоровьем Ваших сверстников» (варианты ответов: “Очень хорошее”, “Хорошее”, “Среднее”, “Плохое”, “Очень плохое”, “Затрудняюсь ответить”) и B\_33\_5: «Приходится ли Вам экономить на лечении и как часто» (варианты ответов: “Постоянно”, “Время от времени”, “Никогда”) оформлены в виде числовых выборок кодов ответов. Код ответа соответствует номеру ответа. Требуется проверить гипотезу: «респонденты, меньше экономящие на лечении, лучше оценивают свое здоровье».

В данном случае имеем классическую задачу оценки связи между признаками, представленными в порядковых шкалах (если исключить вариант ответа “Затрудняюсь ответить” для первого вопроса). Коды ответов 1-5 для первого вопроса соответствуют последовательным уровням ухудшения оцениваемого состояния здоровья. Коды ответов 1-3 для второго вопроса соответствуют последовательно убывающим уровням экономии на лечении.

Для выяснения наличия зависимости между признаками строим таблицу сопряженности признаков В\_13 и В\_33\_5 (см. пример 3 раздела «Выборочный метод»). Анализируя таблицу сопряженности (рис. 1) делаем предварительный вывод о наличии связи между признаками.

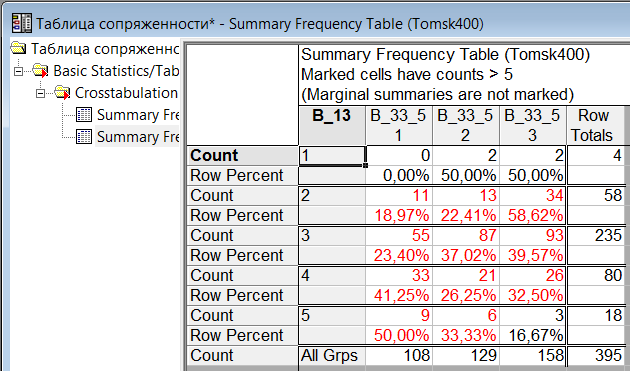


Рис. 1. Таблица сопряженности частот признаков В\_13 и В\_33\_5

Для корректного применения критерия Хи-квадрат необходимо, чтобы наблюдаемое число значений для каждой ячейки было не сильно мало (желательно, не менее 5). В нашем случае очень мало значений содержится в ячейках первой строки. Поэтому объединим ответы 1 и 2 на вопрос В\_13 в один, чтобы избавиться от ячеек с малыми частотами (то есть объединяем ответы “Очень хорошее” и “Хорошее” на вопрос B\_13 в один ответ). Для этого в таблице данных введем новую переменную (назовем ее B\_13\_1), заходим в свойства переменной и в поле «Long name» записываем формулу, как показано на рис. 2.

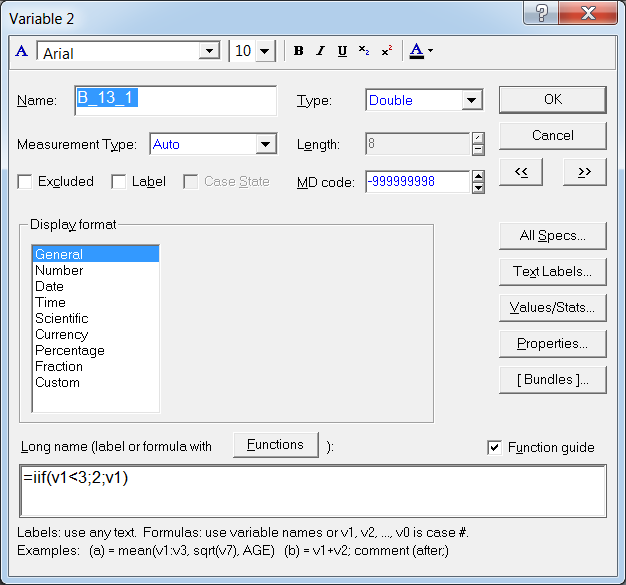


Рис. 2. Создание переменной B\_13\_1, объединяющей категории 1 и 2 переменной В\_13 (v1) в категорию 2

В результате категории 1 и 2 переменной B\_13 будут объединены в одну с кодом 2. Таблица сопряженности переменных B\_13\_1 и В\_33\_5 приведена на рис. 3.

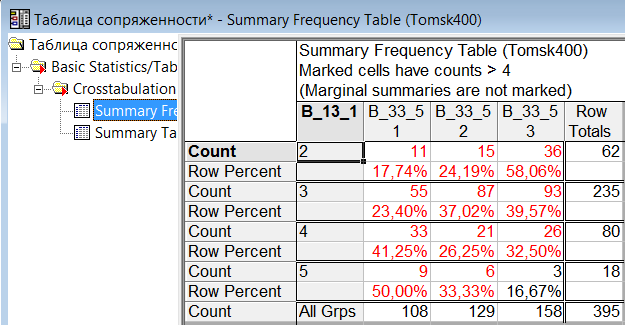


Рис. 3. Таблица сопряженности частот признаков В\_13\_1 и В\_33\_5

Для статистической проверки наличия связи между признаками в окне результатов кросстабуляции «Crosstabulation tables: Results» выбираем вкладку «Options» и отмечаем галочками параметры, как показано на рис. 4. Тем самым, мы заказываем тест Хи-квадрат (параметр «Pearson & M-L Chi-square»), коэффициенты сопряженности Пирсона и Крамера (параметр «Phi (2x2 tables) & Cramer's V & C», ранговый коэффициент корреляции Спирмена (параметр «Spearman rank order correlation»), ранговый коэффициент корреляции Кендалла (параметр «Kendal’s tau-b & tau-c»), коэффициент корреляции Гудмана-Краскала (параметр «Gamma»).

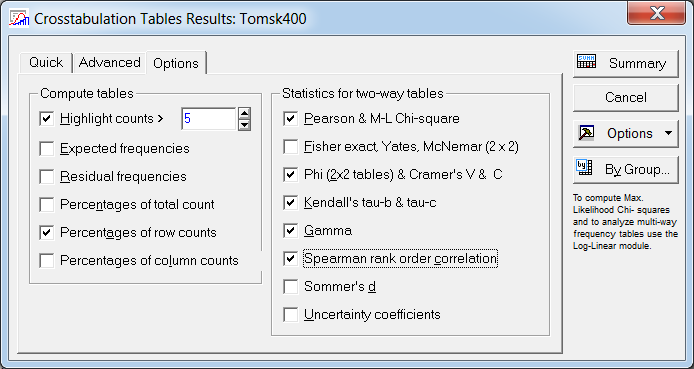


Рис. 4. Выбор величин, оценивающих наличие и степень связи между порядковыми признаками.

Переходим на вкладку «Advanced» и нажимаем кнопку «Detailed two-way tables» («подробные двухвходовые таблицы»). В результате, на странице «Statistics» в рабочей книге в разделе «Crosstabulation results dialog», получим значения всех заказанных статистик с указанными уровнями значимости (рис. 5).

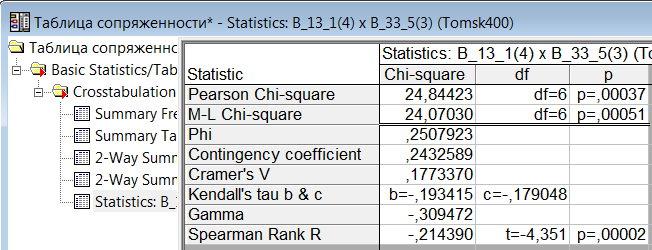


Рис. 5. Значения статистик, измеряющих связь между признаками.

В таблице сверху вниз перечислены значения статистик: Хи-квадрат Пирсона, Хи-квадрат максимального правдоподобия, Фи, коэффициента сопряженности Пирсона, V Крамера, тау-b и тау-с Кендалла, гамма Гудмана-Краскала, Спирмена.

Отметим, прежде всего, высоко значимое значение статистики Хи-квадрат: . Это означает, что нулевая гипотеза должна быть отвергнута и признано значимым наличие связи между рассматриваемыми признаками.

Поскольку данные измерены в порядковой шкале, можно попытаться установить характер этой связи. Заметим, что все использованные характеристики степени связи, которые учитывают направление связи, имеют отрицательное значение. Это означает, что мы должны признать наличие отрицательной корреляции между признаками. То есть возрастанию одного из признаков (чем хуже респондент оценивает здоровье) соответствует убывание другого признака (тем больше экономит на лечении). Значимость именно такого соответствия подтверждается значимостью коэффициента корреляции Спирмена (). Для остальных статистик в таблице уровни значимости не указаны, поэтому их можно использовать только, как информативные.

Таким образом, гипотеза: «респонденты, меньше экономящие на лечении, лучше оценивают свое здоровье» получила в ходе проверки значимое подтверждение.

**Пример 2.** Ежедневные доходности акций ПАО «Газпром» и ПАО «НК Роснефть» за период с 01.01.2015 по 01.09.2015 составляют соответственно выборки GAZP и ROSN. Требуется проверить гипотезу «Доходности акций ПАО «Газпром» и ПАО «НК Роснефть» в указанном периоде являются независимыми величинами».

Для проверки гипотезы «Доходности акций ПАО «Газпром» и ПАО «НК Роснефть» в указанном периоде являются независимыми величинами» используем критерий независимости Хи-квадрат. Для применения критерия нам надо построить таблицу сопряженности, где строки и столбцы будут соответствовать различным интервалам для каждой переменной. Поскольку пакет STATISTICA не позволяет строить таблицы сопряженности для интервальных данных, необходимо построить категориальные переменные, которые будут определять принадлежность выборочных данных по каждой переменной к тому или иному интервалу. Число ячеек двумерной таблицы можно ориентировочно определить по формуле Стерджеса. В нашем случае выборки имеют объем , соответственно, рекомендуемое число ячеек . Если ориентироваться на это значение, то наиболее разумно будет построить таблицу , то есть, для каждой переменной ввести три интервала группирования.

Чтобы оценить границы интервалов, предварительно построим статистические ряды для каждой выборки, содержащие по три интервала группирования данных. Полученные интервальные статистические ряды приведены на рис 6.

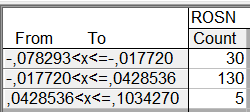
 

Рис. 6. Интервальные статистические ряды для двух выборок

Как видим, границы интервалов следует подкорректировать (уменьшить границы центральных интервалов), чтобы получить более равномерное распределение данных по интервалам. Выберем в качестве центрального интервала для каждой переменной интервал . Естественно, что для каждой переменной мы можем в данном случае задавать свои границы интервалов (вспомним, что при проверке гипотезы однородности по критерию Хи-квадрат, необходимо задавать одни и те же интервалы для каждой переменной).

Создаем новые переменные, значения которых (или код) будут указывать на принадлежность значений исходных переменных к тому или иному интервалу группирования. Для этого выбираем новые переменные (назовем их K\_GAZP и K\_ROSN), заходим в свойства переменной и в поле «Long name» записываем формулы, как показано на рис. 7. и рис. 8.

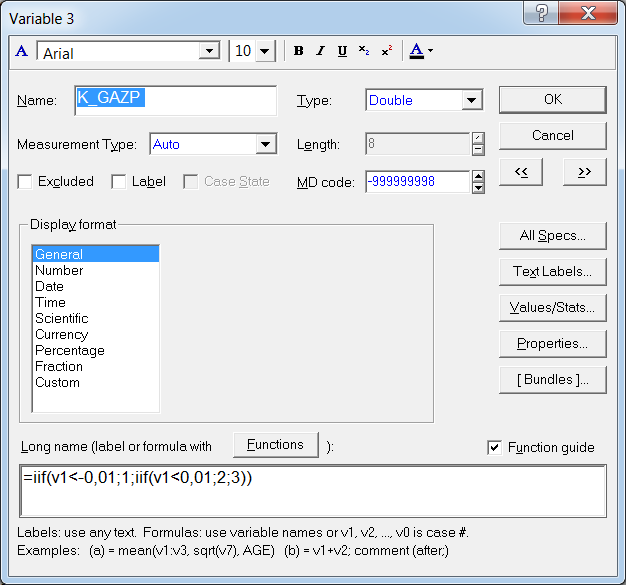


Рис. 7. Создание категориальной переменной K\_GAZP, указывающей на принадлежность переменной GAZP (v1) к соответствующему интервалу группирования

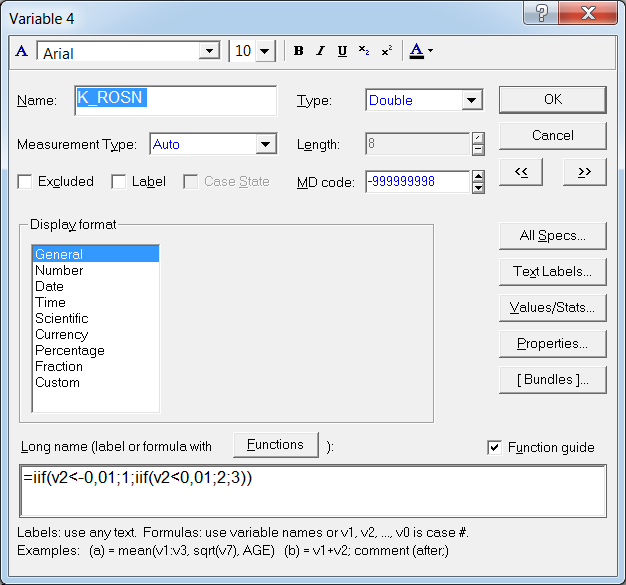


Рис. 8. Создание категориальной переменной K\_ROSN, указывающей на принадлежность переменной ROSN (v2) к соответствующему интервалу группирования

Строим таблицу сопряженности для признаков K\_GAZP и K\_ROSN – рис. 9. Как видим, различие распределений частот по строкам очень существенное, что позволяет сделать предварительный вывод о зависимости признаков. Затем, что в таблице есть ячейки с очень малыми частотами, что может привести к некорректным выводам при применении критерия Хи-квадрат. Однако, если посмотреть таблицу ожидаемых частот, построенную в предположении независимости признаков (рис. 10), то можно заметить, что она не содержит малых частот. Правильнее при применении критерия Хи-квадрат ориентироваться именно на таблицу ожидаемых частот, хотя на практике обычно смотрят на таблицу наблюдаемых частот, поскольку она является первичным статистическим материалам, а таблица ожидаемых частот уже строится на ее основе. Таким образом, применение критерия Хи-квадрат в данном случае можно считать корректным, а собственно значения статистики и уровень значимости можно получить в заголовке таблицы ожидаемых частот.

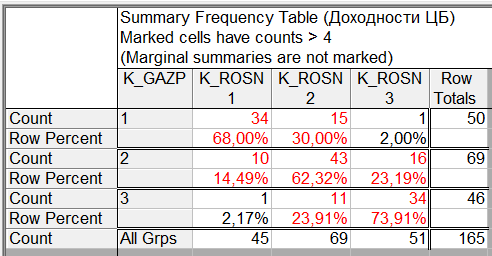


Рис. 9. Таблица сопряженности частот признаков K\_GAZP и K\_ROSN

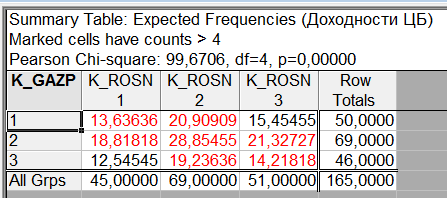


Рис. 10. Таблица ожидаемых частот признаков K\_GAZP и K\_ROSN при условии их независимости

Высоко значимое значение статистики Хи-квадрат () говорит о том, что нулевая гипотеза о независимости должна быть отвергнута и принята гипотеза о зависимости доходностей двух ценных бумаг.

Можно также проверить наличие связи между величинами GAZP и ROSN, не прибегая к построению таблиц сопряженности, а вычисляя значение коэффициента корреляции между этими величинами и анализируя его значимость. Для этого запускаем в головном меню модуль «Statistics», в стартовой панели выбираем пункт «Basic Statisics/Tables», и далее в появившемся меню (рис. 11) выбираем пункт «Correlation matrices» («Матрица корреляций»). В появившемся окне модуля «Product-Moment and Partial Correletions» (рис. 12) задаем переменные, выбрав «One variable list», и получаем матрицу корреляций, нажав на кнопку «Summary» (рис. 13). Помимо матрицы корреляций для всех величин в таблице приведены также значения выборочных средних и стандартных отклонений.

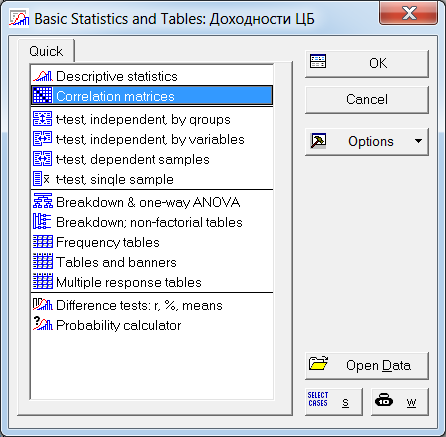


Рис. 11. Выбор пункта для построения матрицы корреляций в меню модуля «Basic Statistics and Tables»

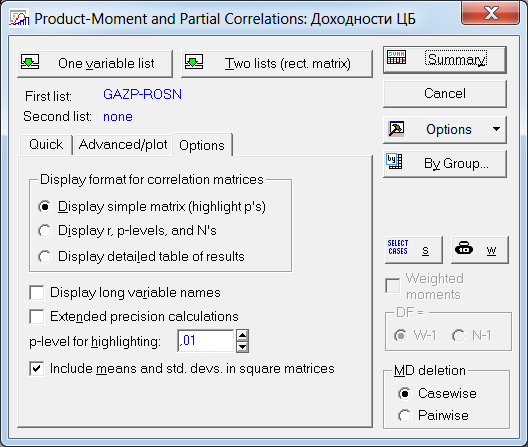


Рис. 12. Окно модуля «Product-Moment and Partial Correlations»

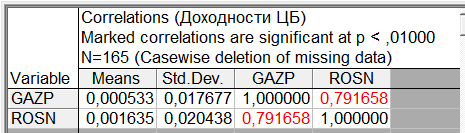


Рис. 13. Матрица корреляций величин GAZP и ROSN

Как видим, между величинами GAZP и ROSN существует сильная значимая положительная корреляция: . В данном случае программа выделяет красным значение коэффициентов корреляций, значимых на уровне 0,01. Данное значение можно задать на вкладке «Options» модуля «Product-Moment and Partial Correlations».

Таким образом, для данных величин, чтобы установить факт зависимости между ними, достаточно было вычислить коэффициент корреляции и оценить его значимость. Однако, следует не забывать, что коэффициент корреляции характеризует лишь степень линейной связи между величинами. Критерий же независимости Хи-квадрат является универсальным критерием и позволяет выявить факт зависимости между величинами произвольной природы.