Описательные статистики
Корреляционная матрица

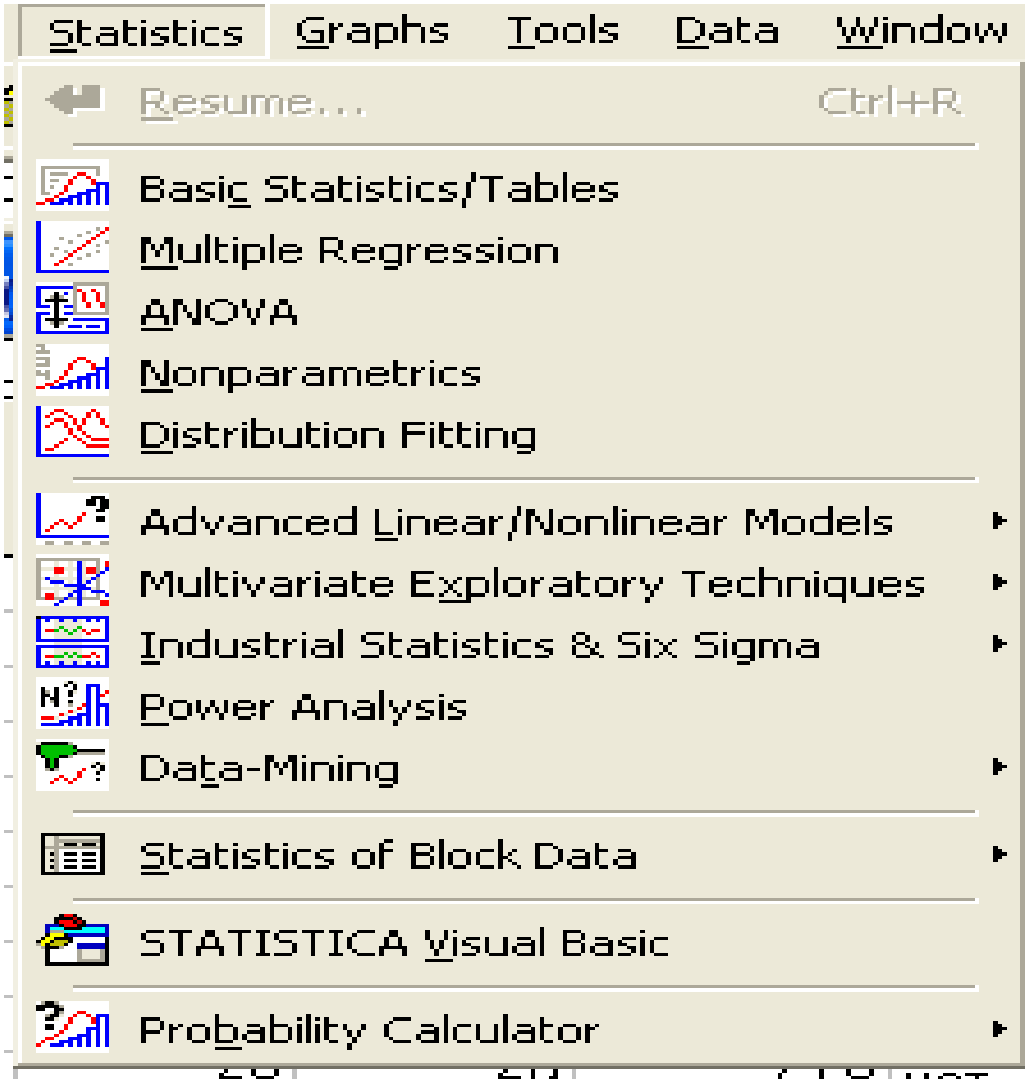
Для выбора переменной, описательные статистики которой нас интересуют, надо нажать кнопку **Variables** и в открывшемся окне щелкнуть на имени переменной (переменных) (рис.3).

Для просмотра результатов надо нажать кнопку **Summary. Descriptive statistics**. Откроется таблица с основными статистиками. Если нас интересуют другие статистики, необходимо указать их на вкладке **Advanced,** установив флажки напротив соответствующих статистик.

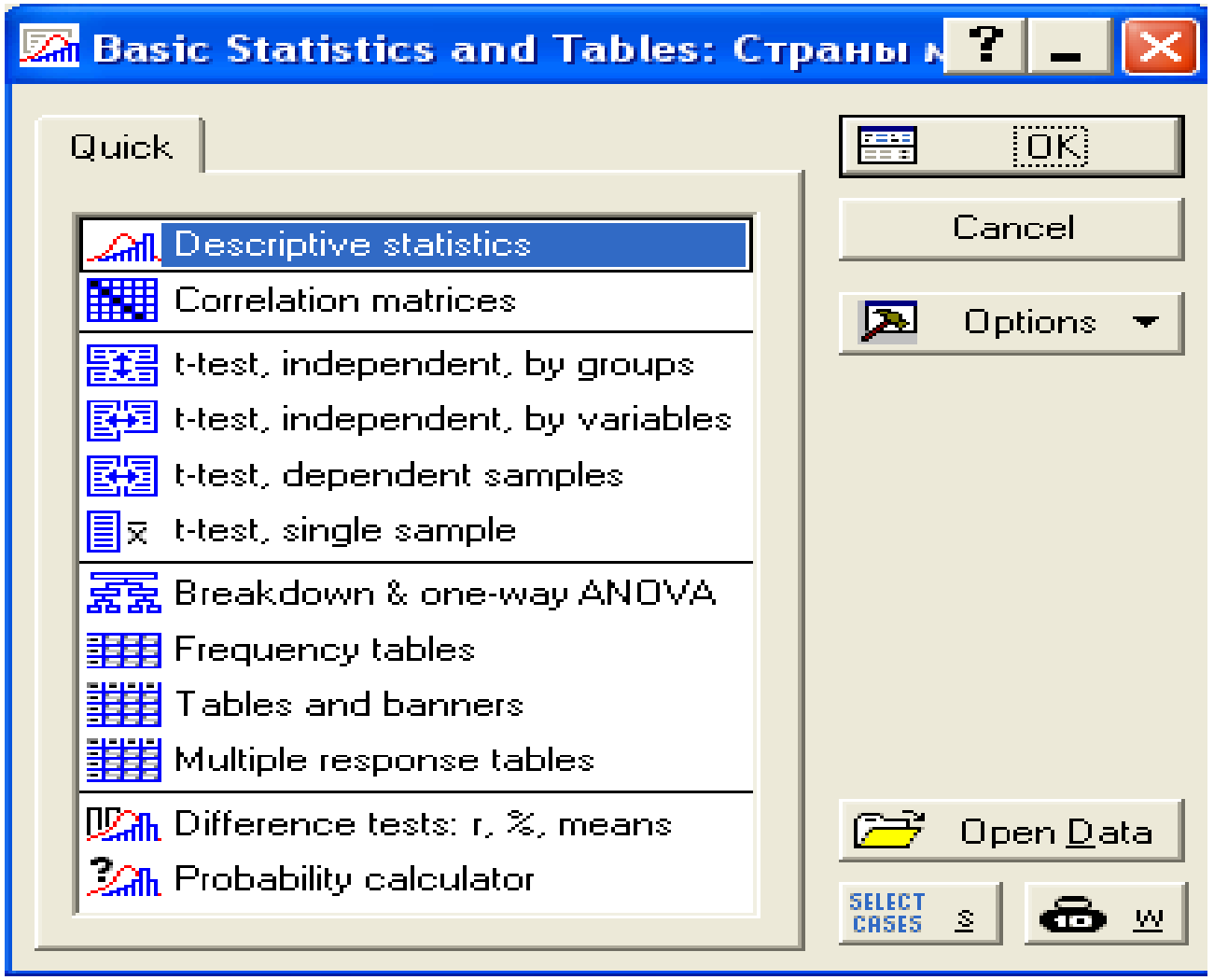
Посчитаем описательные статистики для файла Страны мира, рассмотренном на предыдущем занятии.

	Крупнейшие страны мира по численности населения								
	1 1990г	2 1995г	3 2000г	4 Город 1995г	5 Нефть	6 Газ	7 Пром.	8 С/х	9 Услуги
Китай	1120	1121	1275	30,3	есть	есть	48	21	31
Индия	830	935	1010	26,8	нет	нет	30	29	41
США	250	263	250	76,2	есть	есть	26	2	72
Бразилия	150	162	170	78,2	нет	нет	37	14	49
Россия	289	149	146	73	есть	есть	38	7	55
Япония	124	125	126	77,6	нет	нет	38	2	60
ФРГ	80	82	82	86,5	нет	нет	38	2	60
Индонезия	180	198	215	35,4	есть	есть	42	17	41
Великобритания	57	57	69	89,5	нет	нет	32	2	66
Франция	56	58	59	72,8	нет	нет	27	2	71

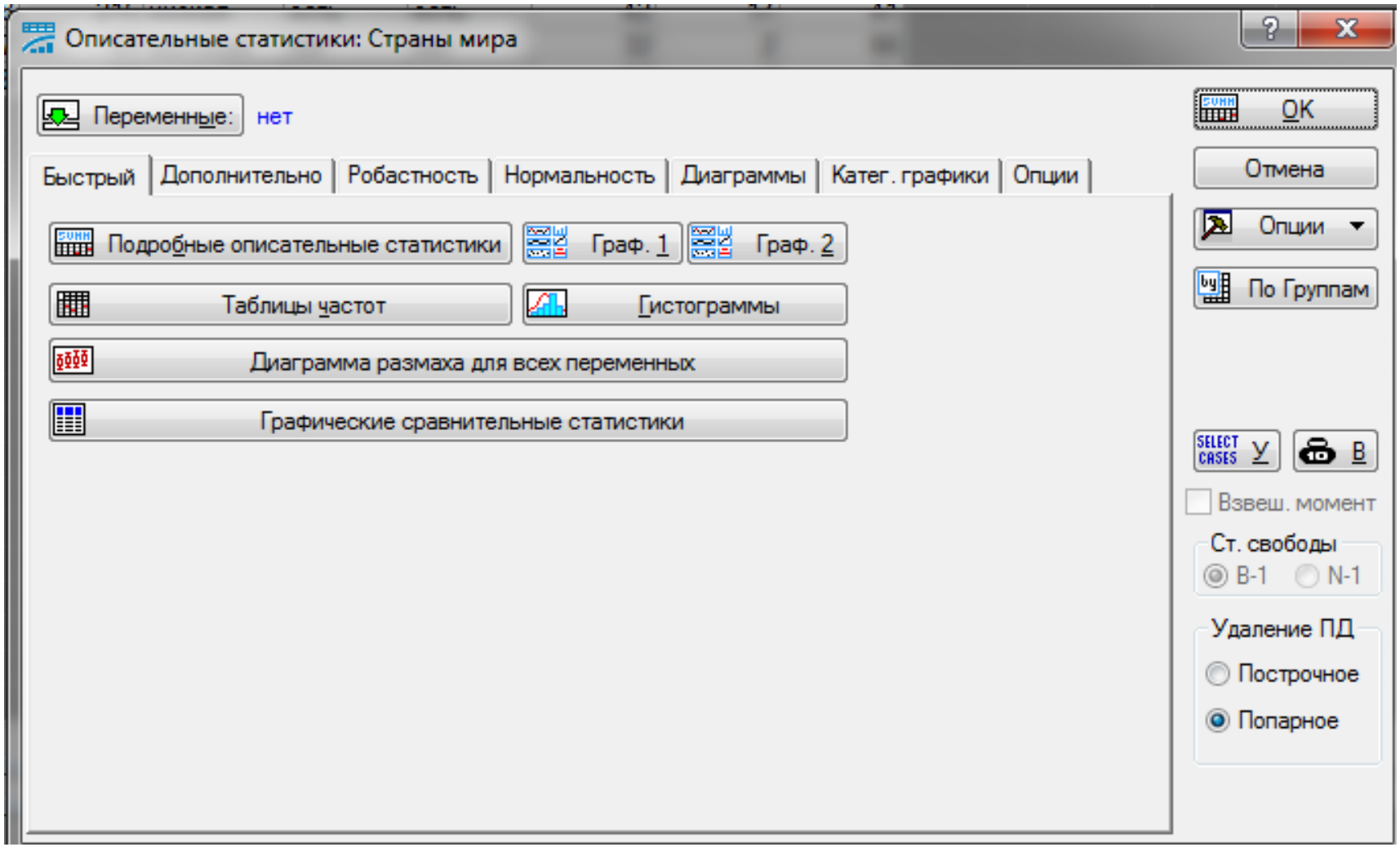
Для запуска программы в верхнем меню **Statistics** надо выбрать команду **Basic Statistic Tables** (основные статистики/таблицы).



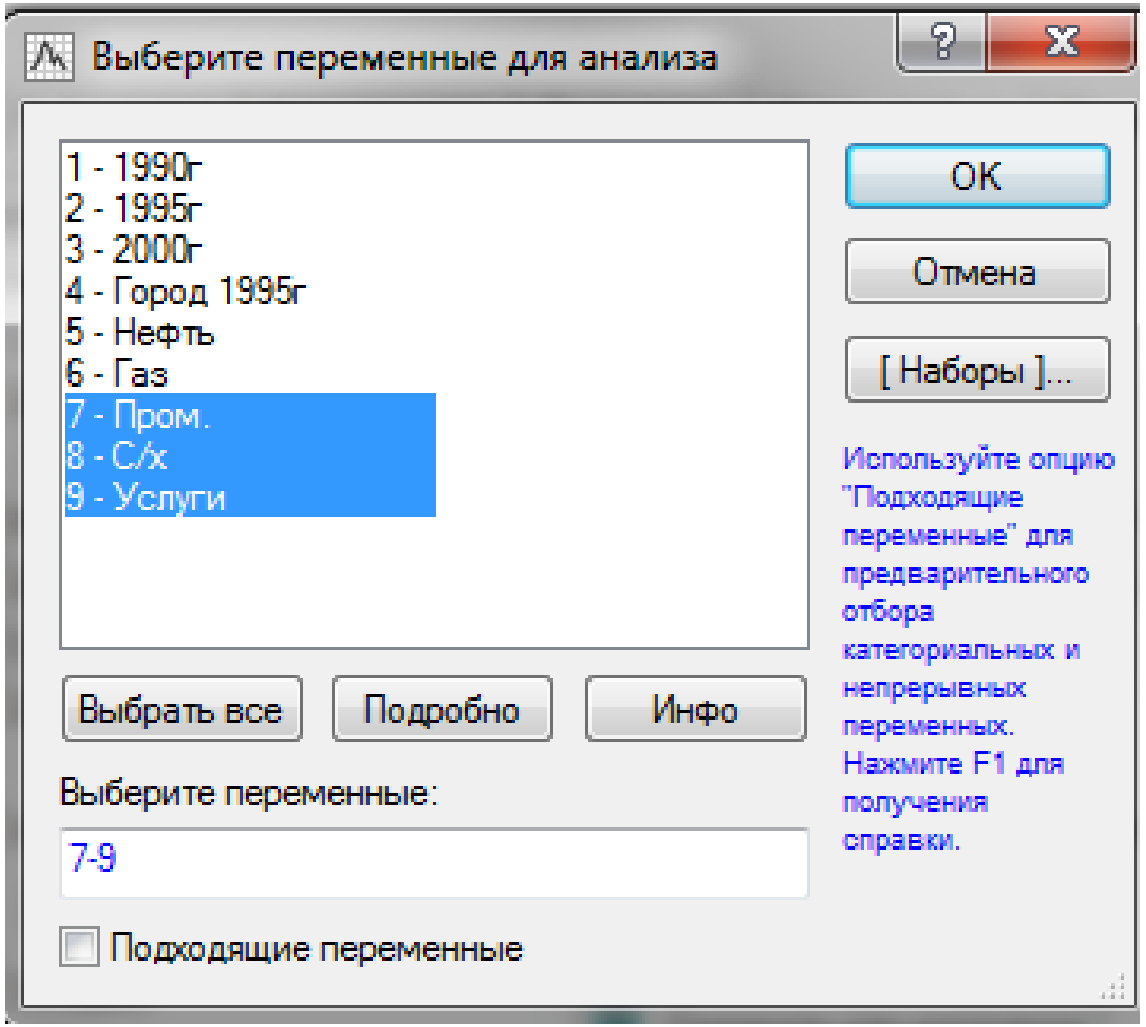
В появившемся меню надо выбрать команду **Descriptive statistics** (описательные статистики)

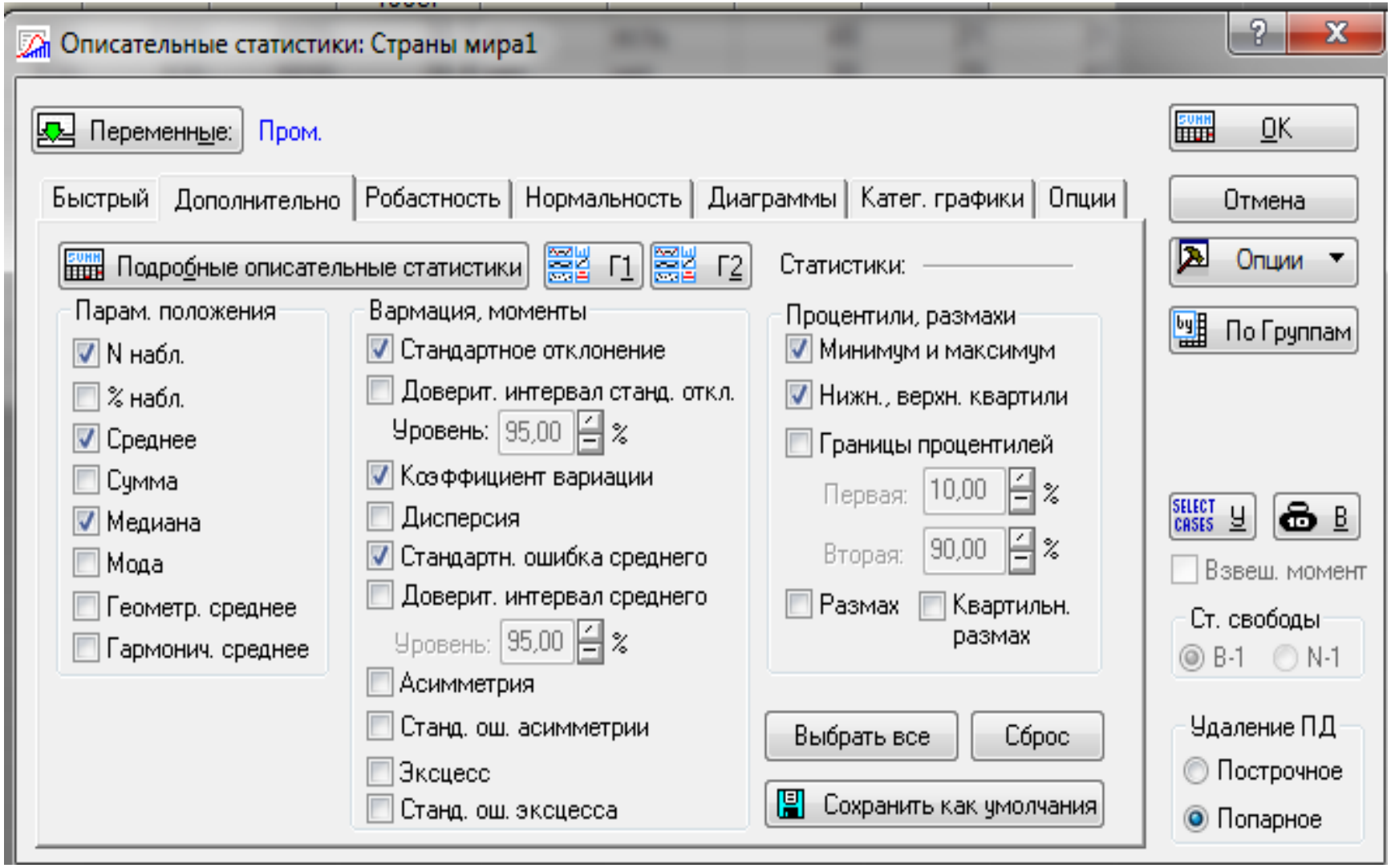


Откроется окно на вкладке Быстрый, воспользуемся кнопкой Переменные,



Укажем 3 количественные переменные и перейдем на вкладку дополнительно , где приведены основные статистики, характеризующие случайные величины




Описательные статистики: Страны мира1
Переменные:
Пром.
OK
Отмена
Опции
По Группам
Быстрый
Дополнительно
Робастность
Нормальность
Диаграммы
Катег. графики
Опции
Подробные описательные статистики
Г1
Г2
Статистики:
Парам. положения
N набл.
% набл.
Среднее
Сумма
Медиана
Мода
Геометр. среднее
Гармонич. среднее
Вармация, моменты
Стандартное отклонение
Доверит. интервал станд. откл.
Уровень: 95,00 %
Коэффициент вариации
Дисперсия
Стандартн. ошибка среднего
Доверит. интервал среднего
Уровень: 95,00 %
Асимметрия
Станд. ош. асимметрии
Эксцесс
Станд. ош. эксцесса
Процентили, размахи
Минимум и максимум
Нижн., верхн. квартили
Границы процентилей
Первая: 10,00 %
Вторая: 90,00 %
Размах
Квартильн. размах
Выбрать все
Сброс
Сохранить как умолчания
SELECT CASES
Взвеш. момент
Ст. свободы
B-1
N-1
Удаление ПД
Построчное
Попарное

Рассмотрим более подробно дополнительные статистики, предусмотренные в этом модуле. Статистики, используемые в данном модуле, в основном очень просты. Применение тех или иных статистик определяется использованием шкал, в которых произведено измерение признаков исследуемых объектов.

Mean (среднее арифметическое) – показывает центральное положение (центр) переменной и рассматривается совместно с доверительным интервалом. Доверительный интервал представляет интервал значений вокруг оценки, где с данным уровнем доверия находится «истинное» (неизвестное) среднее генеральной совокупности. Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p = 0,95 равны соответственно 19 и 27, то можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее совокупности.

Если установить больший уровень доверия, то интервал станет шире, поэтому возрастает вероятность, с которой он «накрывает» неизвестное среднее, и наоборот.

Квантиль, соответствующая вероятности *p*, это значение переменной, ниже которой находится *p-я* часть (доля) выборки. Квантили, соответствующие вероятностям 0,25 и 0,75, называются соответственно *Lower & upper quartiles* (*нижней* и *верхней квартилью*; кварта – четверть).

Альтернативной оценкой среднего являются *median* (медиана) и *mode* (мода).

Медиана – это квантиль, соответствующая вероятности 0,5, т.е. это значение, которое разбивает выборку на две равные части по количеству элементов. Одна половина наблюдений лежит ниже медианы, вторая половина – выше. Если число наблюдений в выборке четно, то медиана вычисляется как среднее двух средних значений. Нижняя квартиль, медиана, верхняя квартиль делят выборку на 4 равные части. Как правило, используется для оценки среднего, если переменная измерена в порядковой шкале.

Мода – это значение переменной, соответствующее наибольшей частоте появления переменной в выборке. Как правило, используется для оценки среднего, если переменная измерена в номинальной или порядковой шкале.

Std.dev. (стандартное отклонение) – это корень квадратный из суммы квадратов отклонений значений переменной от среднего значения, деленное на *n*-1.

Std.err.of mean (стандартная ошибка среднего) – это стандартное отклонение, деленное на корень квадратный из объема выборки.

Varience (коэффициент вариации) – это отношение стандартного отклонения к среднему.

Minimum (минимум) или *Maximum* (максимум) – это соответственно минимальное или максимальное значение выборки.

Range (размах) – это разность между максимальным и минимальным значениями выборки.

Quartiles range (квартильный размах) равен разности значений верхней и нижней квартилей, т.е. это интервал, содержащий медиану, в который попадает 50% выборки.

Skewness (асимметрия) – это мера симметричности распределения. Если распределение симметрично, то асимметрия равна нулю, если асимметрия существенно отличается от 0, то распределение несимметрично. Нормальное и равномерное распределения абсолютно симметричны. Асимметрия распределения с длинным правым хвостом положительна. Если распределение имеет длинный левый хвост, то его асимметрия отрицательна.

Kurtosis (эксцесс) – мера остроты пика распределения. Если распределение нормальное, то эксцесс равен 0. Если эксцесс положителен, то пик заострен, если отрицателен, то пик закруглен.

Нажмем на кнопку Подробные описательные статистики на вкладке Дополнительно, или Быстро.

Переменная	Описательные статистики (Страны мира)						
	N набл.	Среднее	Доверит. -95,000%	Доверит. 95,000%	Медиана	Мода	Частота моды
Пром.	10	35,60000	30,68765	40,51235	37,50000	38,00000	3
С/х	10	9,80000	2,74407	16,85593	4,50000	2,000000	5
Услуги	10	54,60000	44,71530	64,48470	57,50000	Множест.	2

Переменная	Описательные статистики (Страны мира)					
	Минимум	Максим.	Нижняя Квартиль	Верхняя Квартиль	Дисперсия	Ст.откл.
Пром.	26,00000	48,00000	30,00000	38,00000	47,1556	6,86699
С/х	2,00000	29,00000	2,00000	17,00000	97,2889	9,86351
Услуги	31,00000	72,00000	41,00000	66,00000	190,9333	13,81786

Переменная	Описательные статистики (Страны мира)	
	Коэф.Вар.	Станд. ошибки
Пром.	19,2893	2,171533
С/х	100,6481	3,119117
Услуги	25,3074	4,369592

Вычисление медианы для Пром.:

26, 27, **30**, 32, **37, 38**, 38, **38**, 42, 48

(37+38)/2=37,5

Корреляция

Между переменными (случайными величинами) может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. Но между переменными может существовать и связь другого рода, проявляющаяся в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Такую связь называют стохастической. Она появляется в том случае, когда имеются общие случайные факторы, влияющие на обе переменные.

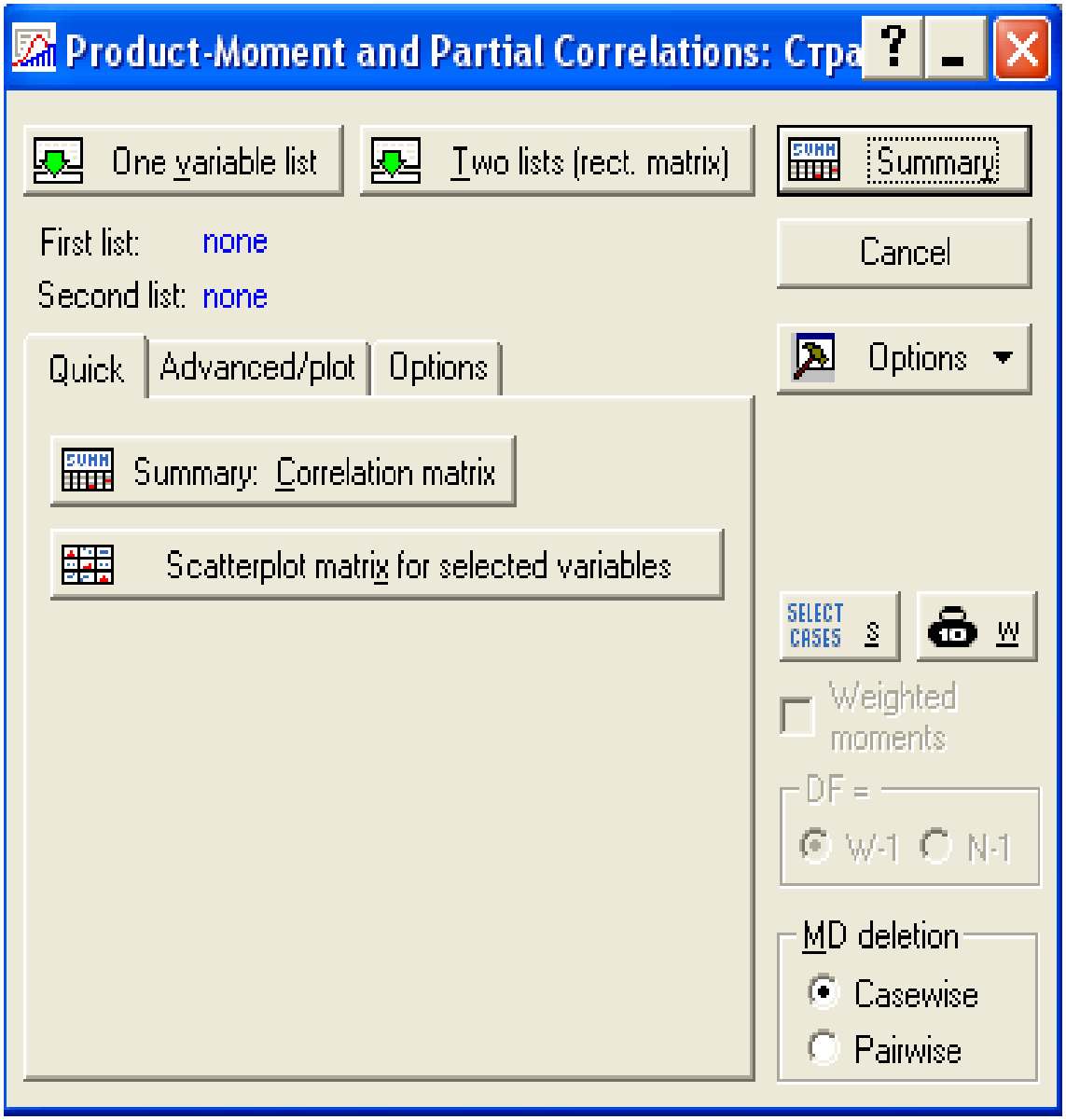
В качестве меры зависимости между переменными используется коэффициент корреляции (r), который изменяется в пределах от –1 до +1. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения другой убывают.

• Если переменные независимы, то коэффициент корреляции равен 0 (обратное утверждение верно только для переменных, имеющих нормальное распределение).

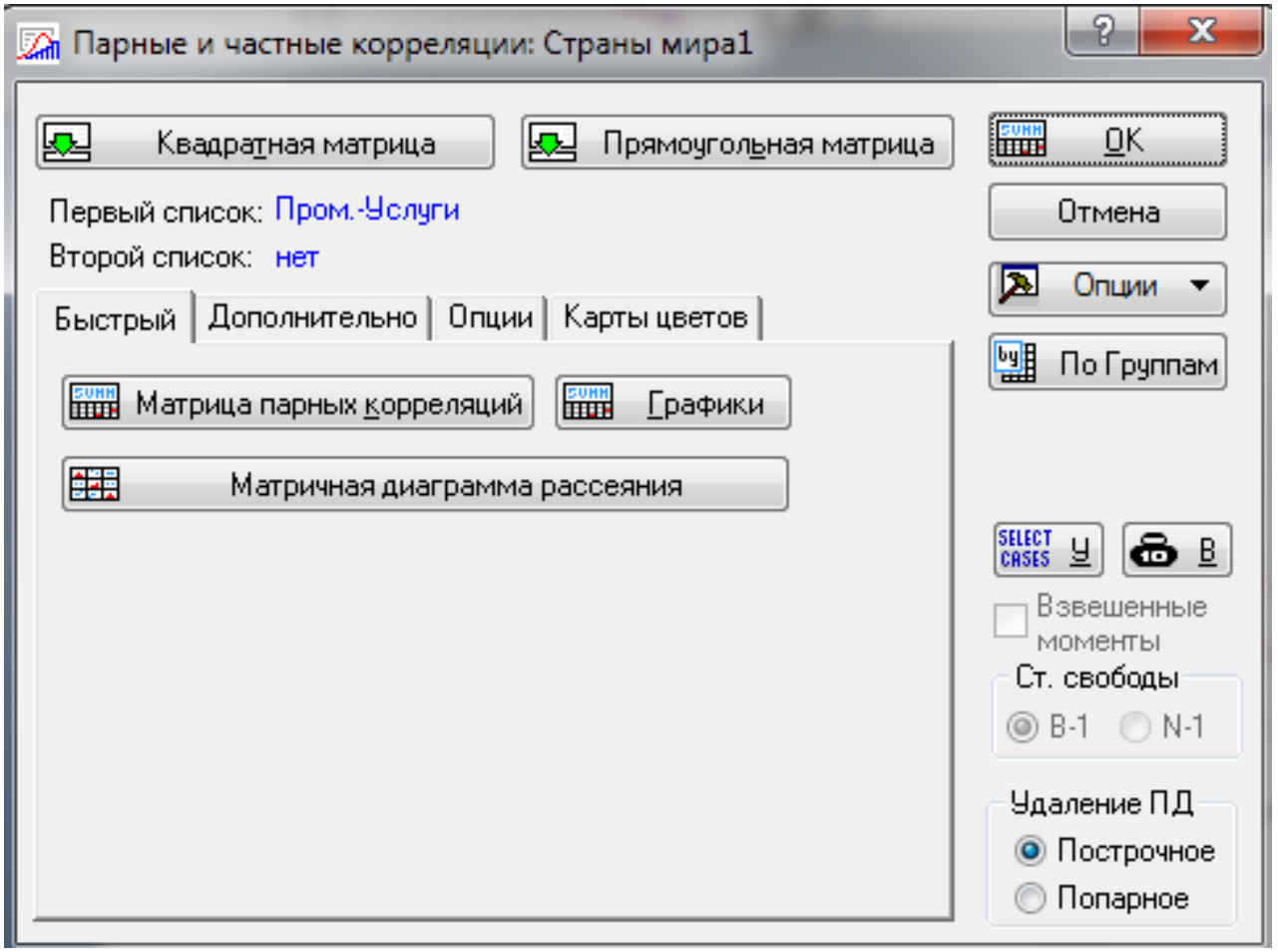
• Но если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Чем ближе значение r к 1, тем зависимость сильнее. Коэффициент корреляции достигает своих предельных значений +1 или -1, тогда и только тогда, когда зависимость между переменными линейная. . В модуле **Descriptive statistics** вычисляется коэффициент корреляции Пирсона, в предположении, что переменные измерены, как минимум, в интервальной шкале. Некоторые другие коэффициенты корреляции (например, корреляция Спирмена или тау Кендала) могут быть вычислены для более слабых шкал.

• Принято считать, что при $|r| \leq 0{,}25$ – корреляция слабая, $0{,}25 < |r| \leq 0{,}75$ – умеренная, при $|r| \geq 0{,}75$ – сильная [12]. Сильная корреляция означает, что связь между переменными может быть близкой к линейной, но может быть явно нелинейной.

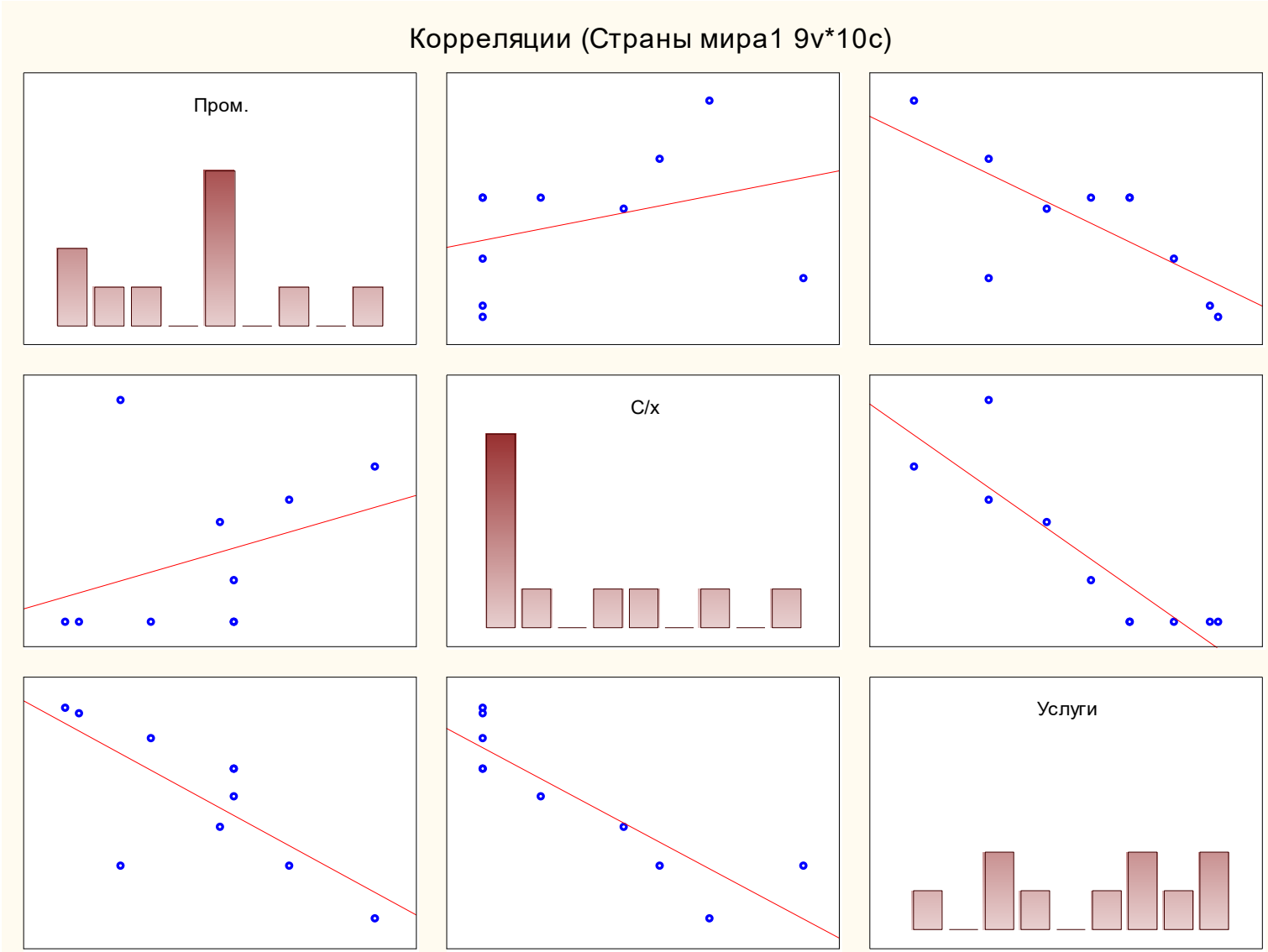
- Для построения корреляционной матрицы в верхнем меню **Statistics** надо выбрать команду **Basic Statistic Tables,** откроется меню команды (рис.2). После выбора команды **Correlation Matrices** откроется рабочее окно модуля. Имена переменных можно задать одним списком (кнопка **One variables list**) или двумя списками (кнопка **Two lists**).
- В первом случае будет построена квадратная корреляционная матрица, строки и столбцы которой представлены списком переменных. Элементы матрицы – коэффициенты корреляции между переменными, расположенными на пересечении строки и столбца.
- Во втором случае будет построена прямоугольная матрица, строки и столбцы которой представлены соответственно первым и вторым списком .


Product-Moment and Partial Correlations: Стра
One variable list
Two lists (rect. matrix)
Summary
First list: none
Second list: none
Cancel
Quick
Advanced/plot
Options
Options
Summary: Correlation matrix
Scatterplot matrix for selected variables
SELECT CASES
S
W
Weighted moments
DF =
W-1
N-1
MD deletion
Casewise
Pairwise

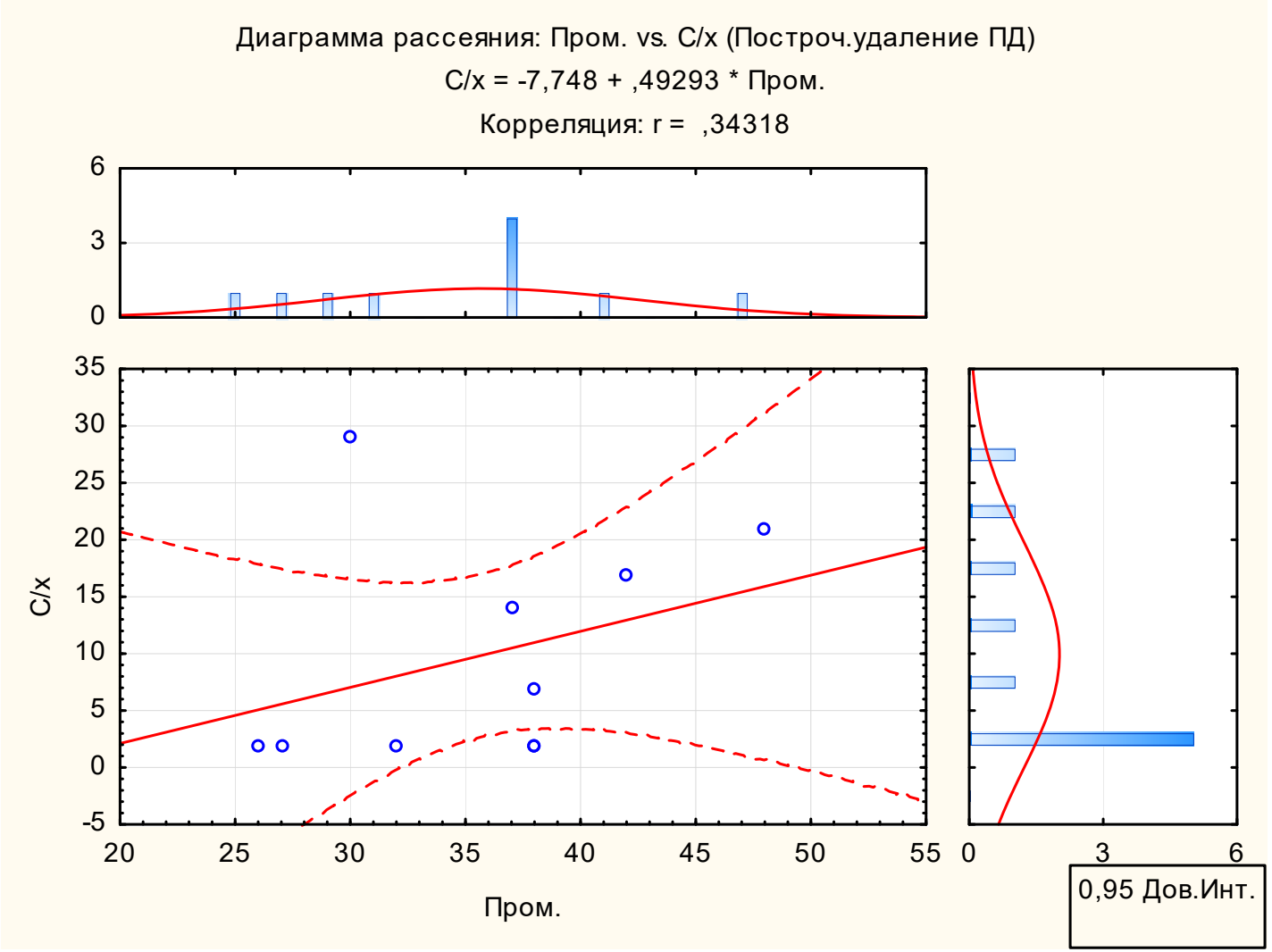
Variable	Correlations (Страны мир Marked correlations are si N=10 (Casewise deletion		
	Пром.	С/х	Услуги
Пром.	1,00	0,34	-0,74
С/х	0,34	1,00	-0,88
Услуги	-0,74	-0,88	1,00



Если нажать на кнопку Матричная диаграмма рассеяния, то появится график на котором будут изображены парные диаграммы всех со всеми и гистограммы


Корреляции (Страны мира1 9v*10c)
Пром.
С/х
Услуги

Если нажать на кнопку Графики, то появится 3 диаграммы рассеяний с доверительными интервалами на 3 отдельных графиках



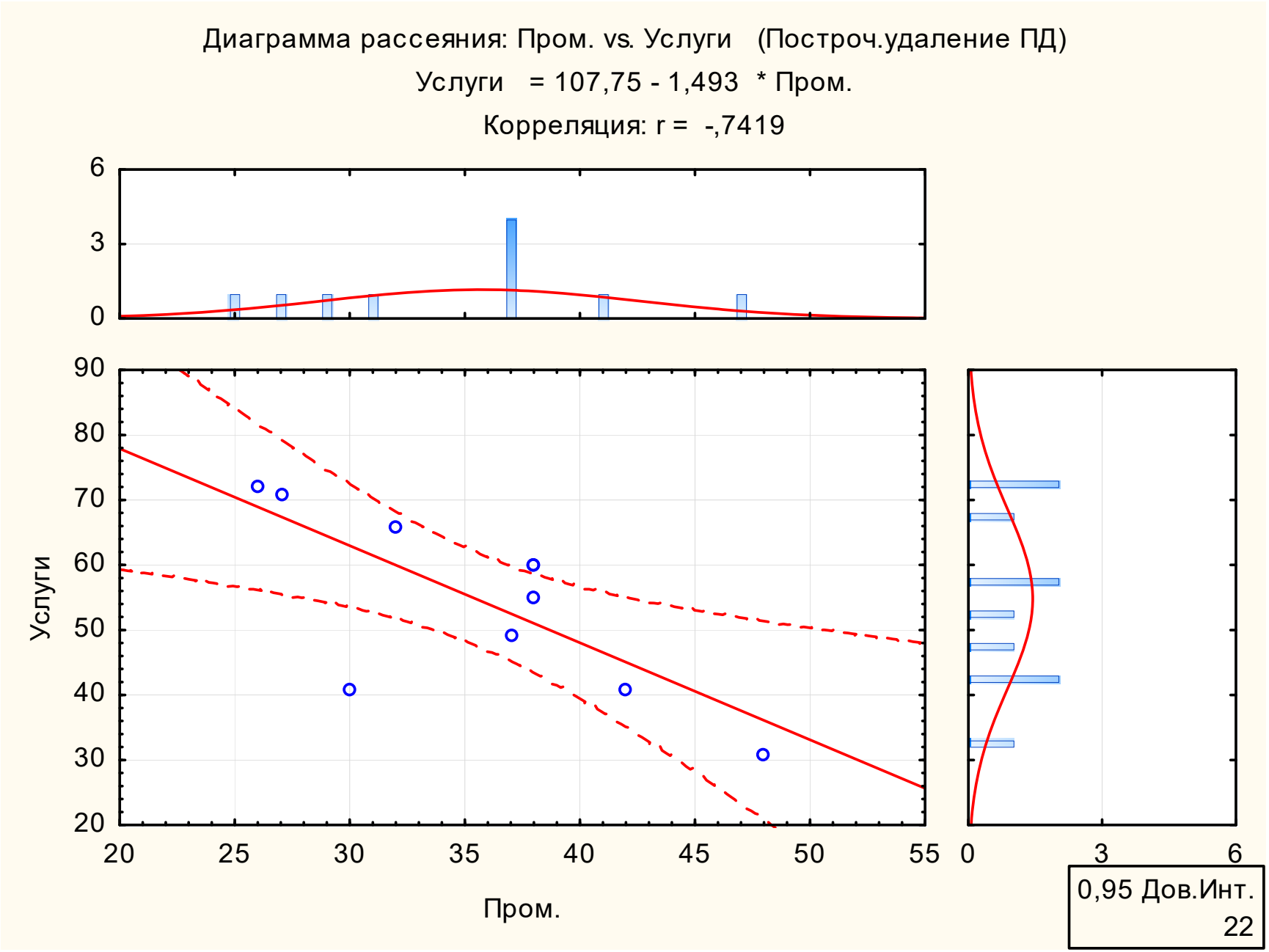


Диаграмма рассеяния: Пром. vs. Услуги (Построч.удаление ПД)
Услуги = 107,75 - 1,493 * Пром.
Корреляция: r = -,7419
6
3
0
90
80
70
60
50
40
30
20
Услуги
20
25
30
35
40
45
50
55
Пром.
0
3
6
0,95 Дов.Инт.
22

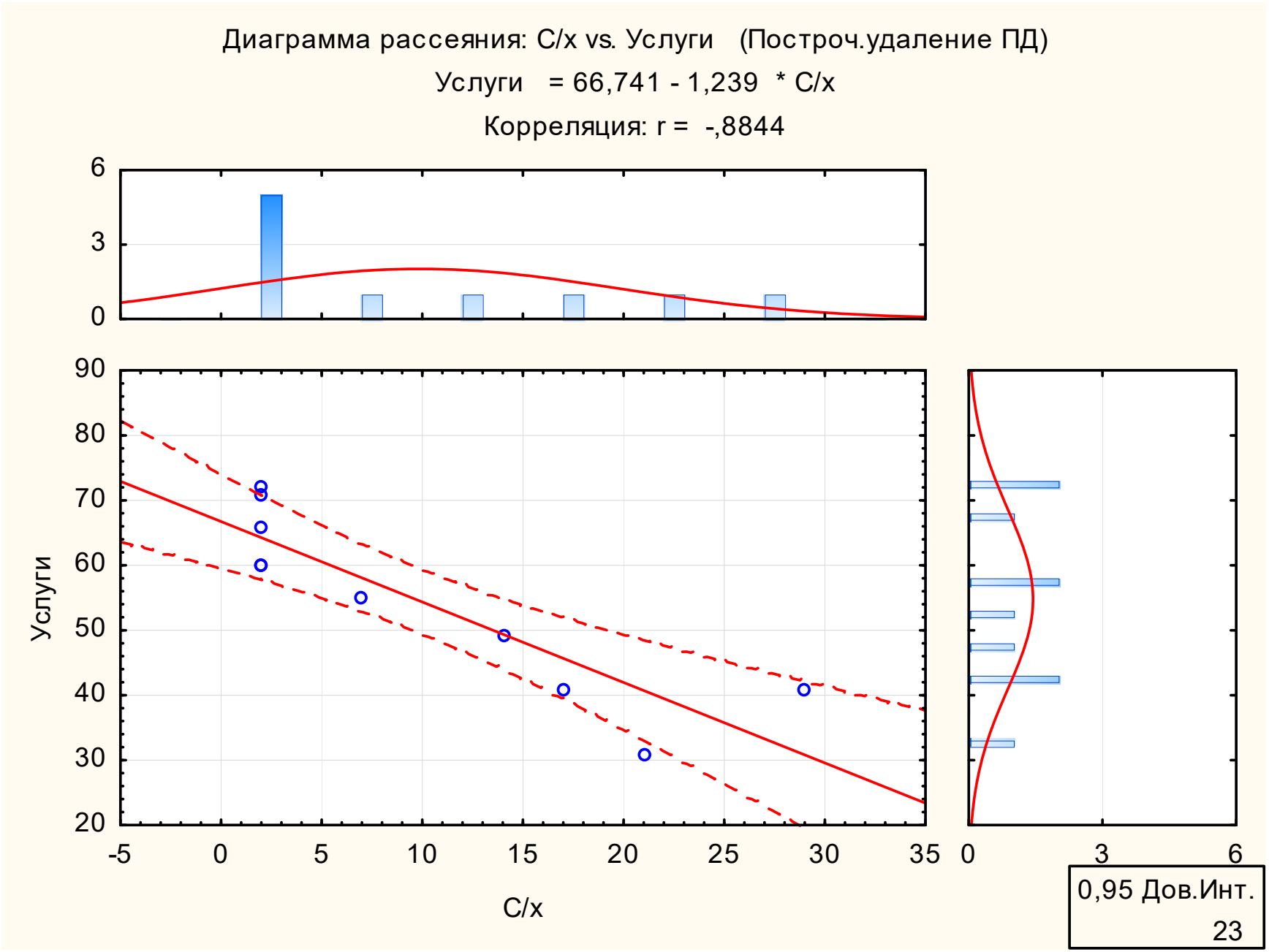
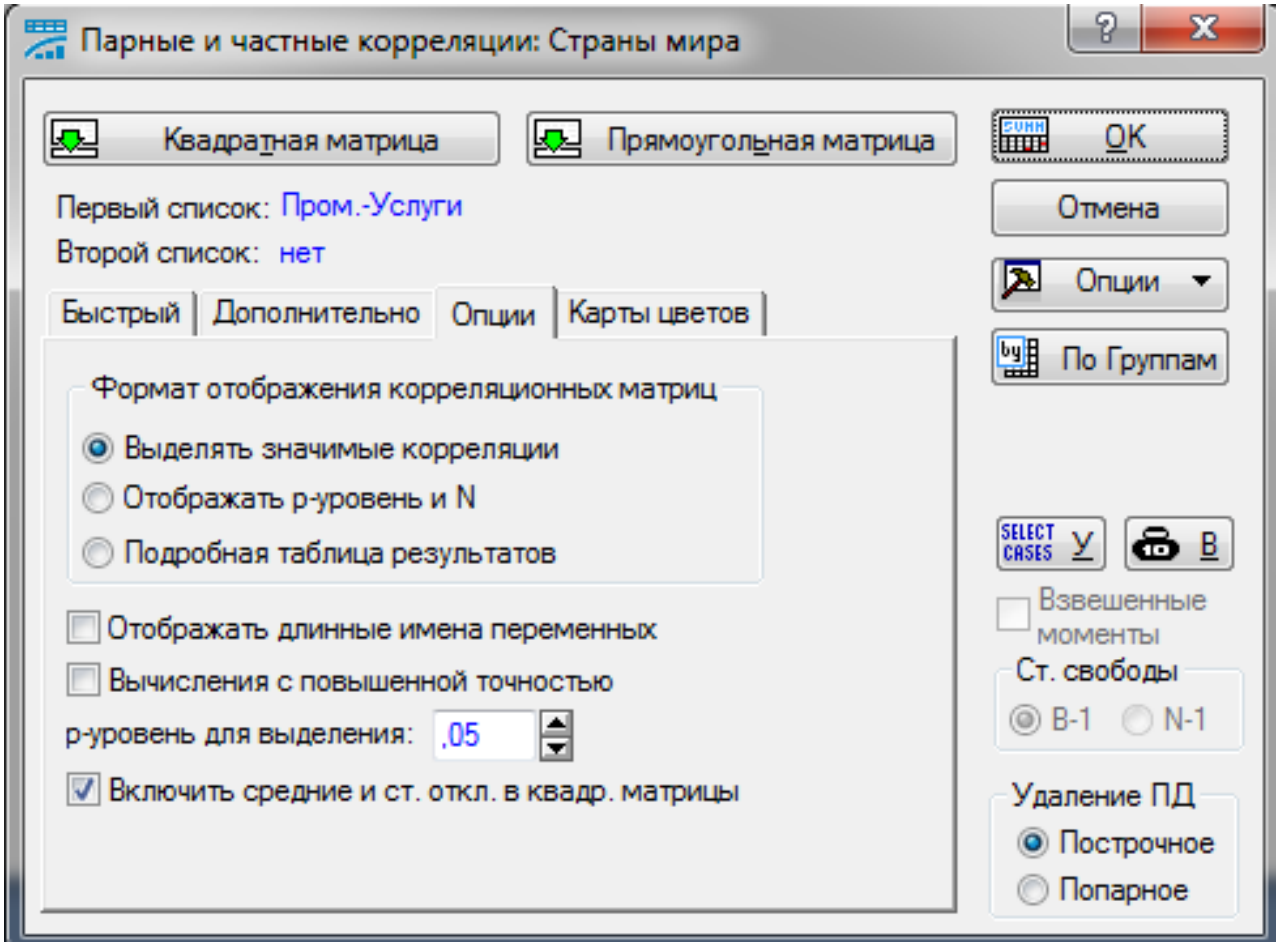


Диаграмма рассеяния: C/x vs. Услуги (Построч.удаление ПД)
Услуги = 66,741 - 1,239 * C/x
Корреляция: r = -,8844
6
3
0
90
80
70
60
50
40
30
20
Услуги
-5
0
5
10
15
20
25
30
35
C/x
0
3
6
0,95 Дов.Инт.
23

Если перейти на вкладку Опции, то можно в таблицу отобразить уровни значимости коэффициентов корреляции, построить более подробную таблицу результатов со средними значениями, стандартными отклонениями и т.д., также можно произвести вычисления с повышенной точностью



Переменная	Корреляции (Страны мира) Отмеченные корреляции значимы на уровне p <,05000 N=10 (Построчное удаление ПД)		
	Пром.	С/х	Услуги
Пром.	1,0000	0,3432	-0,7419
	p= ---	p=0,332	p=0,014
С/х	0,3432	1,0000	-0,8844
	p=,332	p= ---	p=,001
Услуги	-0,7419	-0,8844	1,0000
	P=0,014	p=0,001	p= ---

Если перейти на вкладку Карты цветов, то можно построить таблицу корреляций в абсолютных величинах, или построить карту цветов, на которой разными цветами будут изображены отрицательные, положительные, нулевые корреляции, Разными оттенками цветов будут изображены силы корреляционных связей