Раздел 2. Анализ распределений в социологическом/маркетинговом исследовании.

1. Одномерное распределение.

2. Двумерное распределение.

1. Одномерное распределение.

1. Понятие частотной таблицы или вариационного ряда.
2. Показатели центра распределения.
3. Показатели вариации.
4. Показатели формы распределения.

Первым этапом статистического анализа данных, как правило, является частотный анализ данных.

Распределение частот является самым удобным способом представления различных значений переменной. Статистики, связанные с распределением частот можно разделить на три группы:

- показатели центра распределения;
- показатели вариации;
- показатели формы распределения.

Показатели центра распределения. Показатели центра распределения характеризуют положение центра распределения, вокруг которого концентрируются данные. Это - среднее арифметическое, мода и медиана.

Наиболее часто используемым показателем является среднее арифметическое, или, как его еще называют, выборочное среднее. Эта величина получается делением суммы всех имеющихся значений переменной на число значений:

$$\bar{x} = \sum_{i=1}^{n} x_i \ ,$$

где x_i - полученные значения переменной x, n - число наблюдений или размер выборочной совокупности.

Следующим показателем центром распределения является мода. Мода – это значение переменной, которое чаще всего встречается в выборочном распределении.

Еще одним показателем центра распределения является медиана. Медианой называется значение переменной в середине ряда данных, расположенных в порядке возрастания или убывания. Если число данных четное, то медиана равна полусумме двух срединных значений. Следует отметить, что медиана является так называемым 50-й процентилем.

Можно дать несколько рекомендации по использованию показателей центра распределения. Если переменная измеряется по номинальной шкале, то лучше использовать моду. Если переменная измеряется по порядковой шкале, то лучше использовать медиану Для интервальной или относительной шкалы лучше использовать среднее арифметическое. Этот показатель учитывает всю доступную информацию, но, в тоже время, он очень чувствителен к выбросам значений, которые, как правило, бывают либо экстремально большими, либо экстремально малыми. При наличии таких выбросов значений нужно дополнительно использовать медиану.

Показатели вариации. В отличие от показателей центра распределения, показатели вариации показывают не разброс данных вокруг центра распределения, а просто меру разброса данных. Показатели вариации включают размах вариации, межквартильный размах, дисперсию, стандартное отклонение и коэффициент вариации.

Проще всего рассчитать размах вариации. Он равен разности между наибольшим и наименьшим значениями переменной в вариационном ряду и отражает разброс данных. Таким образом, его можно рассчитать следующим образом:

$$R = x_{\max} - x_{\min},$$

где $x_{\max}$ - максимальное значение переменной в вариационном ряду, $x_{\min}$ - минимальное значение вариационном ряду.

Следующий показатель вариации межквартильный размах – это разность между 75-м и 25-м процентилем.

Важными показателями являются дисперсия и среднеквадратическое отклонение. Дисперсия является средним из квадратов отклонений переменной от её средней величины. Нужно помнить, что если значения данных сгруппированы вокруг среднего, то дисперсия невелика. Если же значения данных разбросаны, то дисперсия является большой. Дисперсия вычисляется следующим образом:

$$\sigma^2 = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n-1},$$

где x_i - полученные значения переменной x, $\bar{x}$ - среднее арифметическое (выборочное среднее), n - число наблюдений или размер выборочной совокупности.

Следует заметить, что деление происходит не на n, а на $n-1$, поскольку генеральное среднее неизвестно, а вместо него используется выборочное среднее. Это делает выборку менее изменчивой, чем фактически. Таким образом, мы корректируем более слабую изменчивость значений переменной, наблюдаемую в выборке.

Среднеквадратическое (стандартное) отклонение равно корню квадратному из значения дисперсии:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n-1}}.$$

Последним показателем, на котором нужно остановиться, является коэффициент вариации. Он рассчитывается как отношение стандартного отклонения к среднему арифметическому, выраженное в процентах:

$$CV = \frac{\sigma}{\bar{x}},$$

где σ - стандартное отклонение, $\bar{x}$ - среднее арифметическое (выборочное среднее).

Показатели формы распределения. К показателям формы распределения относятся показатели ассиметрии и эксцесса. Первый из них

показывает, насколько симметричную форму имеет распределение, а второй показатель рассчитывается для симметричных распределений и показывает, является ли распределение островершинным или, напротив, плосковершинным.

Построение частотных таблиц. Рассмотрим процесс построения частотных таблиц на следующем примере. Мы опросили сотрудников небольшой фирмы о размере их заработной платы (Таб. 4.1).

Таблица 4.1

Данные о заработной плате сотрудников

№ сотрудника	Заработная плата, руб.
1	5000
2	4500
3	5000
4	6500
5	10000
6	4500
7	6000
8	3000
9	7000
10	7500
11	8000
12	5000
13	4500
14	7000
15	7500

16	5000
17	5000
18	4500
19	5000
20	5500

В соответствии с рекомендациями, данными ранее, создаем в SPSS файл данных.

Для того, чтобы приступить к анализу мы должны выбрать в меню следующие команды:

Analyze (Анализ)

Descriptive Statistics (Описательная статистика)

Frequencies (Частоты).

В результате выполнения этих команд появляется Диалоговое окно **Frequencies** (Частоты) (рис. 4.1)**.**

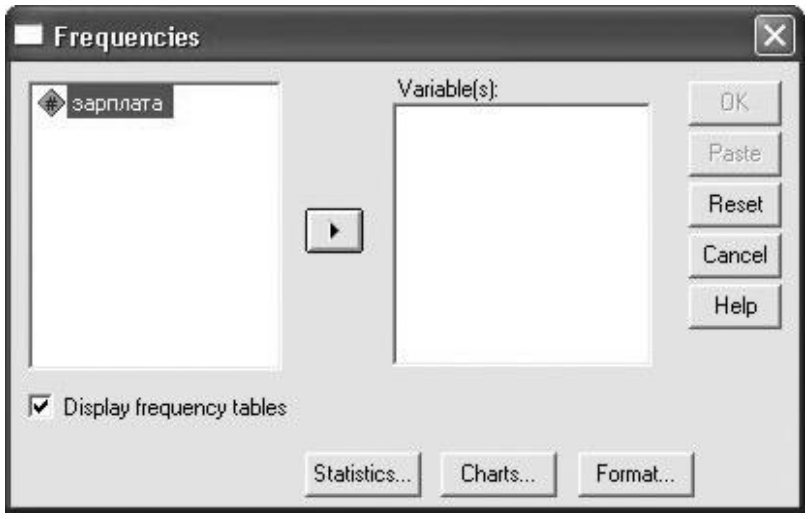

Рис. 4.1. Диалоговое окно Frequencies (Частоты)

С помощью кнопки с треугольником переменная «зарплата» может быть перемещена в список выходных переменных.

Для того, чтобы вычислить необходимые статистические характеристики нужно щелкнуть на кнопке Statistics… (Статистика). В результате выполнения этой операции откроется диалоговое окно **Frequencies**: **Statistics** (Частоты: Статистика) (рис. 4.2)

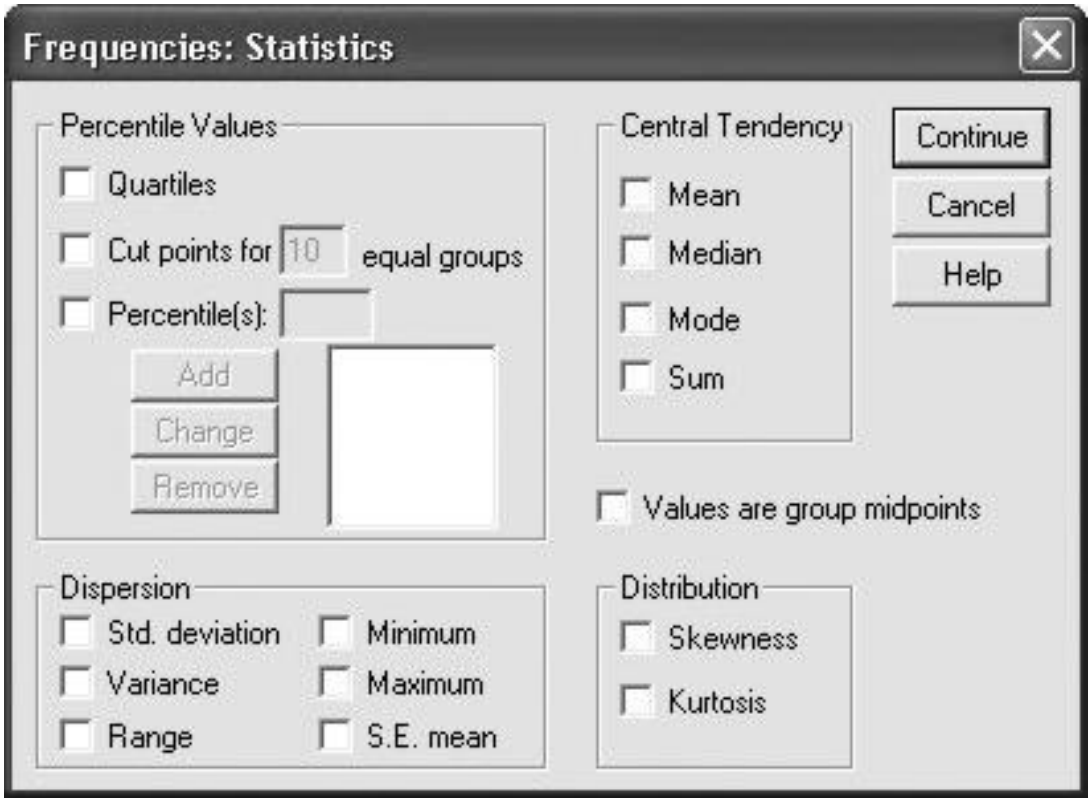

Рис. 4.2. диалоговое окно Frequencies: Statistics (Частоты: Статистика)

Данное диалоговое окно разделено на четыре группы.

Первая группа Percentile Values (Значения перцентилей) позволяет выбрать три варианта:

- **Quartiles** (Квартили). Позволяет вычислить первый, второй и третий квартили. Первый квартиль – точка на шкале измеренных значений, ниже которой располагаются 25% измеренных значений. Второй квартиль – точка на шкале измеренных значений, ниже которой располагаются 50% измеренных значений, а третий квартиль - точка на

шкале измеренных значений, ниже которой располагаются 75% измеренных значений.

- **Cut points** (Точки раздела). Позволяет вычислить такие значения процентилей, которые разделяют выборку на группы случаев имеющих одинаковую ширину.
- **Percentile(s)** (Процентили). Позволяет вычислить любые значения процентилей, которые потребуются и будут заданы пользователем.

Вторая группа Dispersion (Разброс) позволяет выбрать шесть показателей вариации:

- **Std. deviation** (Стандартное отклонение) Позволяет вычислить стандартное отклонение.
- **Variance** (Дисперсия). Позволяет вычислить дисперсию.
- **Range** (Размах). Позволяет вычислить размах.
- **Minimum** (Минимум). Позволяет определить минимум
- **Maximum** (Максимум). Позволяет определить максимум
- **S.E. mean** (Стандартная ошибка). Позволяет вычислить стандартую ошибку – стандартное отклонение деленное на квадратный корень от объема выборки.

Третья группа Central Tendency (Средние) позволяет выбрать четыре показателя центра распределения:

- **Mean** (Среднее значение). Позволяет вычислить среднее арифметическое.
- **Median** (Медиана). Позволяет вычислить медиану.
- **Mode** (Мода). Позволяет вычислить моду
- **Sum** (Сумма). Позволяет вычислить сумму всех значений.

Последняя, четвертая группа Distribution (Распределение) позволяет определить меры несимметричности распределения:

- **Skewness** (Коэффициент асимметрии). Позволяет вычислить коэффициент ассиметрии.
- **Kurtosis** (Коэффициент вариации или эксцесс). Позволяет вычислить эксцесс, на основании которого можно определить, является ли распределение пологим или крутым.

После того, как заданы все необходимые параметры, необходимо подтвердить их выбор, щелкнув по кнопке Continue (Продолжить). Это действие приведет к возвращению в диалоговое окно **Frequencies** (Частоты). В этом окне необходимо щелкнуть по кнопке Ok. В результате этого откроется окно просмотра, в котором будут представлены построенные таблицы (рис. 4.3).

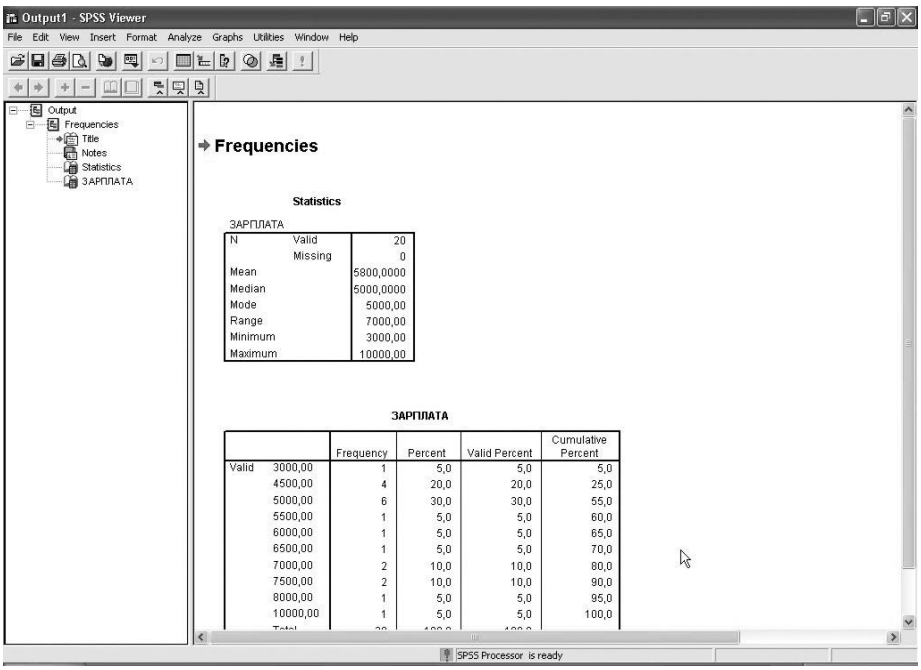

Рис. 4.3. Окно просмотра с построенными таблицами

Остановимся на этих таблицах. В первой таблице (таб. 2.2) приведены значения выбранных статистических характеристик (в данном примере были выбраны не все характеристики):

Таблица 2.2.

Таблица значений выбранных статистических характеристик

Statistics

ЗАРПЛАТА

	Valid	20
	Missing	0

Mean	5800,0000
Median	5000,0000
Mode	5000,00
Range	7000,00
Minimum	3000,00
Maximum	10000,00

На основании этой таблицы можно сделать вывод, что средняя зарплата сотрудников фирмы составляет 5800 рублей. Минимальная зарплата сотрудников равна 3000 рублей, а максимальная – 10000 рублей. Разница между минимальной и максимальной зарплатой – 7000 рублей. Наибольшее число сотрудников получает заработную плату в размере 5000 рублей, причем выше этой суммы лежат доходы половины сотрудников фирмы.

Следующая таблица, которую мы получили, позволяет определить сколько сотрудников фирмы получают определенную зарплату (таб. 2.3).

Таблица 2.3.

Частотная таблица (сведения о заработной плате)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3000,00	1	5	5,0	5,0

		,0		
4500,00	4	20,0	20,0	25,0
5000,00	6	30,0	30,0	55,0
5500,00	1	5,0	5,0	60,0
6000,00	1	5,0	5,0	65,0
6500,00	1	5,0	5,0	70,0
7000,00	2	10,0	10,0	80,0
7500,00	2	10,0	10,0	90,0
8000,00	1	5,0	5,0	95,0
10000,00	1	5,0	5,0	100,0
Total	20	100,0	100,0	

В первом столбике таблицы, имеющим название Frequency (Частоты), представлена частота определенных событий. Например, мы можем сказать, что 6 сотрудников фирмы получают зарплату 5000 рублей. Второй столбик, носящий название Percent (Проценты), позволяет делать вывод о процентном соотношении между теми или иными характеристиками респондентов. Так, например, из таблицы видно, что зарплату в 5000 рублей получают 30% сотрудников фирмы. Два остальных столба можно оставить без рассмотрения.

Вопросы домашнего задания:

1. Каковы особенности построения частотных таблиц?
2. Как получить показатели центра распределения?
3. Как получить показатели вариации?
4. Как получить показатели формы распределения?
5. Что характеризуют мода, медиана и среднее арифметическое?

2. Двумерное распределение.

1. Понятие таблиц сопряженности.
2. Построение таблиц сопряженности признаков.
3. Статистики таблиц сопряженности.

Одной из важнейших задач социологического исследования является определение связи конкретной переменной с другими переменными. Для этого существует специальный статистический метод, который называется построением таблиц сопряженности признаков. Его иногда еще называют кросс-табуляцией. Этот метод позволяет одновременно охарактеризовать две и больше переменных, а его суть заключается в создании таблиц сопряженности признаков, отражающих совместное распределение двух или более переменных с ограниченным числом категорий или определенными значениями.

Для оценки статистической значимости и тесноты связи переменных, содержащихся в таблице сопряженности, используется ряд статистических параметров.

Самым важным таким параметром является критерий χ^2, который позволяет измерить статистическую значимость наблюдаемой связи. Он помогает определить наличие или отсутствие систематической связи между двумя переменными.

Остановимся подробнее на его расчете. Прежде всего, делается предположение, что между двумя переменными не существует никакой связи. Проверка этого предположения выполняется вычислением частот распределения признаков анализируемых переменных в ячейках таблицы, которые можно было бы ожидать, если бы зависимости между переменными

не существовало, и при данных итоговых числах в каждой строке и столбце таблице. Ожидаемая частота для каждой ячейки таблицы вычисляется с помощью следующей формулы:

$$f_e = \frac{n_r n_c}{n},$$

где n_r - итоговое число в строке, n_c - итоговое число в столбце, n - полный размер выборки.

Затем ожидаемые частоты сравниваются фактическими наблюдаемыми частотами распределения признаков, соответствующим ячейкам таблицы. Очевидно, что чем больше это разница, тем выше значение статистики.

Значение критерия χ^2 вычисляется следующим образом:

$$\chi^2 = \sum_{\text{все ячейки}} \frac{(f_0 - f_e)^2}{f_e}.$$

Важной характеристикой критерия χ^2 является число степеней свободы df, которое вычисляется следующим образом:

$$df = (r-1)\times(c-1),$$

где r - число сток, c - число столбцов.

Предположение, что между двумя переменными не существует никакой связи, отклоняется, если полученное значение χ^2 больше, чем критическое значение χ^2 распределения с соответствующим числом степеней свободы. Для того, чтобы определить критическое значение χ^2 необходимо обратиться к специальной таблице – таблице значений χ^2.

Для выяснения тесноты связи вычисляется коэффициент корреляции φ, коэффициент сопряженности признаков, V -коэффициент Крамера.

Коэффициент корреляции φ используется очень редко – в случае, если таблица сопряженности состоит из двух столбцов и двух строк. Для его нахождения используют формулу:

$$\varphi = \sqrt{\frac{\chi^2}{n}},$$

где n - размер выборки.

Коэффициент корреляции φ принимает значение 0, если связь между переменными отсутствует и значение 1, если связь является сильной.

Коэффициент сопряженности признаков C используется для оценки тесноты связи в таблицах любого размера.

Для его вычисления используют формулу:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

где n - размер выборки.

Как и коэффициент корреляции φ, коэффициент сопряженности признаков C принимает значение 0, если связь между переменными отсутствует и стремиться к значению 1, но никогда его не достигает, если связь является сильной.

Еще одной статистикой, которую можно вычислить для измерения тесноты связи в таблицах любого размера, является V - коэффициент Крамера.

Он получается корректировкой коэффициента корреляции φ или по числу рядов, или по числу столбцов таблицы, причем выбирается меньшее значение.

V -коэффициент Крамера рассчитывается по формуле:

$$V = \sqrt{\frac{\chi^2 / n}{\min(r-1),(c-1)}},$$

где n - размер выборки, r - число сток, c - число столбцов.

V -коэффициент Крамера принимает значение 0, если связь между переменными отсутствует и значение 1, если связь является сильной.

Рассмотрим построение таблиц сопряженности признаков на примере, который мы использовали, когда изучали частотные распределения. Теперь мы введем дополнительную переменную – пол сотрудников (таб. 5.1).

Таблица 5.1

Данные о заработной плате и поле сотрудников

№ сотрудника	Пол сотрудника	Заработная плата, руб.
1	женский	5000
2	женский	4500
3	мужской	5000
4	женский	6500
5	мужской	10000
6	женский	4500
7	мужской	6000
8	женский	3000
9	мужской	7000
10	женский	7500
11	мужской	8000
12	женский	5000
13	женский	4500
14	мужской	7000
15	мужской	7500

16	женский	5000
17	женский	5000
18	женский	4500
19	мужской	5000
20	мужской	5500

В соответствии с рекомендациями, данными ранее, создаем в SPSS файл данных, где мужской пол будем кодировать значением «1», а женский «2»

Для того чтобы приступить к анализу мы должны выбрать в меню следующие команды:

Analyze (Анализ)

Descriptive Statistics (Описательная статистика)

Crosstabs… (Таблицы сопряженности).

В результате выполнения этого действия откроется диалоговое окно **Crosstabs** (Перекрестные таблицы) (рис. 5.1).

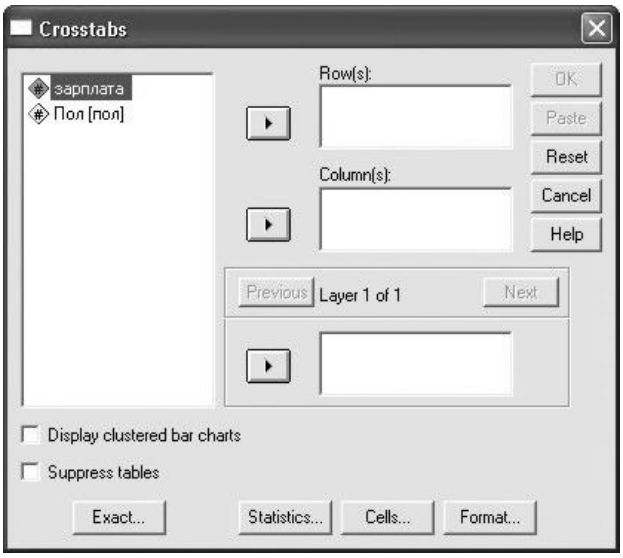

Рис. 5.1. Диалоговое окно Crosstabs (Перекрестные таблицы)

Исходные переменные с помощью кнопки с треугольником можно перенести либо в окно **Row(s)** (Строка(ки)), либо в окно **Column(s)** (Столбец(цы)). В нашем случае, мы связанную с полом переменную переносим в окно **Column(s),** а переменную связанную с зарплатой**,** в окно **Row(s).**

Нажатием кнопки Cells (Ячейки), можно указать вид представляемых данных. В открывшемся диалоговом окне **Crosstabs: Cells Display** (Таблицы сопряженности: Отображение ячеек) в группе Percentages (Проценты) необходимо поставить флажок Column.

Нажатием кнопки Statistics… (Статистики), вызывается диалоговое окно Crosstabs: Statistics (Таблицы сопряженности: Статистика), где можно задать необходимые статистики (рис.5.2).

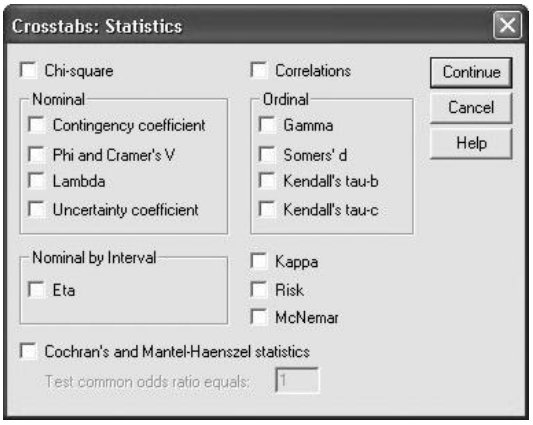

Рис. 5.2. Диалоговое окно Crosstabs: Statistics

(Таблицы сопряженности: Статистика)

Поставив в этом диалоговом окне флажок Chi-square (критерий χ^2), подтвердим наш выбор нажатием кнопки Continue и возвратимся в диалоговое окно **Crosstabs** (Перекрестные таблицы). Нажатие в этом окне кнопки Ok приведет к открытию окна просмотра, где будут построены необходимые нам таблицы.

Первая из трех полученных таблиц называется **Case Processing Summary** (Обработанные случаи), соответственно содержит информацию об обработанных случаях и в подробном рассмотрении не нуждается.

Две оставшиеся таблицы существенно важнее.

Вторая таблица (таб. 5.2.) – собственно таблица сопряженности.

Таблица 5.2

Данные о заработной плате и поле сотрудников

ЗАРПЛАТА * Пол Crosstabulation

			Пол		Total
			мужской	женский	
ЗАРПЛАТА	3000,00	Count		1	1
		% within Пол		9,1%	5,0%
	4500,00	Count		4	4
		% within Пол		36,4%	20,0%
	5000,00	Count	2	4	6
		% within Пол	22,2%	36,4%	30,0%
	5500,00	Count	1		1
		% within Пол	11,1%		5,0%
	6000,00	Count	1		1
		% within Пол	11,1%		5,0%
	6500,00	Count		1	1
		% within Пол		9,1%	5,0%
	7000,00	Count	2		2
		% within Пол	22,2%		10,0%
	7500,00	Count	1	1	2

		% within Пол	11,1%	9,1%	10,0%
	8000,00	Count	1		1
		% within Пол	11,1%		5,0%
	10000,00	Count	1		1
		% within Пол	11,1%		5,0%
Total		Count	9	11	20
		% within Пол	100,0%	100,0%	100,0%

Благодаря этой таблице мы можем проследить, как зависит распределение зарплаты от пола сотрудников. Очевидно, что в нашем случае такая зависимость есть, так как данные показывают, что мужчины получают зарплату выше, чем женщины.

Действительно ли это так можно было бы судить на основании третьей таблицы, которая носит название Chi-Square Tests (Тесты хи-квадрат), в которой рассчитывается необходимая нам характеристика (таб 5.2).

Таблица 5.2

Тесты хи-квадрат

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12,593(a)	9	,182
Likelihood Ratio	17,115	9	,047
Linear-by-Linear Association	5,921	1	,015
N of Valid Cases	20		

a 20 cells (100,0%) have expected count less than 5. The minimum expected count is ,45.

Однако, хотя характеристика и рассчитана, мы видим внизу предупреждение, что 20 ячеек (100%) имеют частоту менее 5. Тест же корректен при соблюдении двух условий:

- ожидаемые частоты меньше 5 должны встречаться не более чем в 20% полей таблицы;
- суммы по строкам и столбцам должны быть больше нуля.

Таким образом, в нашем случае тест не является корректным и для того, чтобы сделать выводы о наличии зависимости между полом и зарплатой необходимо увеличение выборочной совокупности.

Вопросы домашнего задания:

1. Как строятся перекрестные таблицы?
2. Как устроен модуль Custom Tables?
3. В чем различие предлагаемых типов перекрестных таблиц модуля Custom Tables?
4. Какие статистики можно рассчитать?
5. Какие заключения можно сделать на основе этих расчетов?
6. Как строятся трехмерные перекрестные таблицы?
7. Как провести анализ множественных ответов?