

Правительство Российской Федерации
Государственное образовательное бюджетное учреждение
высшего профессионального образования

ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Высшая Школа Экономики
Нижегородский филиал

Кафедра математики
Кафедра информационных систем и технологий

Н.Н. Бобков, В.М. Дёмкин, О.М. Солычева

SPSS: Анализ данных в менеджменте
Описательные статистики

Методическая разработка по курсу
“Анализ данных в менеджменте”

Нижегород
2009

Составители: Н.Н.Бобков, В.М.Дёмкин, О.М.Солычева

УДК 004.67

ББК 32.973.26

Б 72

SPSS: Анализ данных в менеджменте. Описательные статистики. Методическая разработка по курсу “Анализ данных в менеджменте”/НФ ГУ-ВШЭ; Сост.: Н.Н.Бобков, В.М.Дёмкин, О.М.Солычева. Н.Новгород, 2009. 67 с.

Обсуждаются основные этапы анализа данных в среде SPSS (версия 13.0) на примере описательных статистик. Рассматриваются процедуры частотного анализа (Frequencies), описательных статистик (Descriptives), разведочного анализа (Explore), построения таблиц сопряженности (Crosstabs). Обсуждаются возможности SPSS для графического представления данных.

Изложение материала сопровождается теоретическими выкладками, а также иллюстрациями в виде скриншотов окон SPSS.

Издание предназначено для самостоятельного изучения основ статистического анализа данных в среде SPSS.

Работа обсуждена и одобрена на заседании секции “Математика и информатика” УМС НФ ГУ-ВШЭ 11.03.2009г., протокол №3.

© Н.Н.Бобков, В.М.Дёмкин, О.М.Солычева

© Государственный Университет–Высшая Школа Экономики
Нижегородский филиал, 2009

СОДЕРЖАНИЕ

Введение	4
I. Исследование данных: частотный анализ, статистические характеристики	4
II. Графическое представление данных в SPSS	24
1. Диаграмма Box plot («ящик с усами»)	24
2. Диаграмма Stem-and-leaf plot («ствол и лист»)	28
3. Диаграмма Парето	29
4. Примеры решения задач по визуализации данных	35
Реализация задачи с интервальной переменной	35
Реализация задачи с категориальной переменной	43
III. Таблицы сопряженности признаков	54
Библиографический список	66

Введение

В настоящее время в самых разнообразных областях бизнеса формирование имиджа и маркетинговой стратегии любой компании немислимо без применения тех или иных методов представления и пояснения коммерческой информации. Поэтому все большую значимость приобретает практика количественного анализа хозяйственной деятельности, который, в свою очередь, нельзя представить в отрыве от компьютерных технологий. Одним из наиболее распространенных программных продуктов, решающих широчайший спектр задач обработки и представления многомерной статистической информации, является пакет SPSS – Statistical Package for Social Science.

I. Исследование данных: частотный анализ, статистические характеристики

Вычисление статистических характеристик для некоторого набора данных позволяет в ряде случаев на основании нескольких величин (описательных статистик) получить представление о структуре анализируемых данных в целом.

К описательным статистикам традиционно относят характеристики центральной тенденции (выборочное среднее, мода, медиана и квартили (см. подробнее в параграфе Диаграмма Box plot («ящик с усами»)) и показатели изменчивости данных (дисперсия, стандартное отклонение, вариация, межквартильный размах и т.д.).

Для статистического исследования данных в SPSS наиболее часто используются следующие три пункта главного меню:

- ✓ Analyze (*Анализ*) → Descriptive Statistics (*Описательные статистики*) → Descriptives... (*Описательный...¹*);
- ✓ Analyze (*Анализ*) → Descriptive Statistics (*Описательные статистики*) → Frequencies... (*Частотный...*);
- ✓ Analyze (*Анализ*) → Descriptive Statistics (*Описательные статистики*) → Explore... (*Разведочный...*).

¹ Под «...» следует понимать текущий контекст, например, здесь он скрывает за собой слово «анализ»; необходимость этой оговорки вызвана стремлением следовать стилю SPSS.

Рассмотрим реализацию методов статистического анализа в SPSS на конкретном примере.

Ниже приведены данные о балансе (в рублях) телефонных счетов студентов группы, изучающих дисциплину «Анализ данных в менеджменте» с указанием пола студента²:

<i>Баланс</i>	<i>509</i>	<i>38</i>	<i>100</i>	<i>54</i>	<i>283</i>	<i>21</i>	<i>35</i>	<i>103</i>	<i>448</i>	<i>30</i>	<i>168</i>	<i>253</i>	<i>95</i>
<i>Пол</i>	<i>Ж</i>	<i>Ж</i>	<i>Ж</i>	<i>М</i>	<i>Ж</i>	<i>М</i>	<i>М</i>	<i>Ж</i>	<i>М</i>	<i>Ж</i>	<i>М</i>	<i>М</i>	<i>М</i>
<i>Баланс</i>	<i>0</i>	<i>62</i>	<i>328</i>	<i>805</i>	<i>33</i>	<i>158</i>	<i>350</i>	<i>150</i>	<i>36</i>	<i>354</i>	<i>104</i>	<i>224</i>	
<i>Пол</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>М</i>	<i>Ж</i>	<i>Ж</i>	<i>Ж</i>	

На основании этой информации необходимо провести частотный анализ данных о балансе, сначала для всех студентов группы, затем отдельно с разделением по половому признаку, и сделать необходимые выводы.

Создание файла данных. Окно редактора данных SPSS содержит две вкладки: Data View и Variable View. На второй вкладке определяются переменные, а на первой – хранятся сами значения данных для объявленных переменных. Эти данные называют также наблюдениями. Поскольку мы имеем дело с двумерным набором данных (баланс и пол студента), файл данных после операции сохранения будет содержать две переменные. На второй вкладке Variable View окна редактора данных каждая строка соответствует одной переменной.

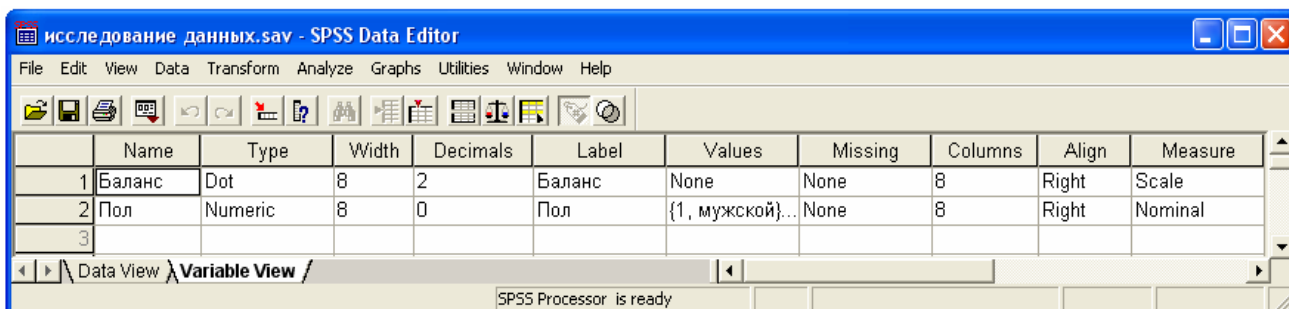
Определим переменные как **Баланс** и **Пол**. Для определения типа переменной можно воспользоваться следующей таблицей:

Numeric (<i>Числовой</i>)	К допустимым значениям относятся числа, перед которыми стоит знак плюс или минус и десятичный разделитель. Знак плюс перед числом, в отличие от минуса, не отображается. В текстовом поле Length (<i>Длина</i>) задается максимальное количество знаков, включая позицию для десятичного разделителя. В текстовом поле Decimals (<i>Десятичные разряды</i>) вводится количество отображаемых знаков дробной части числа.
Comma (<i>Запятая</i>)	К допустимым значениям относятся числа, перед которыми стоит знак плюс или минус, точка как десятичный разделитель и одна или несколько запятых в качестве разделителей

² Сбор данных проводился среди студентов НФ ГУ-ВШЭ путем анонимного письменного опроса.

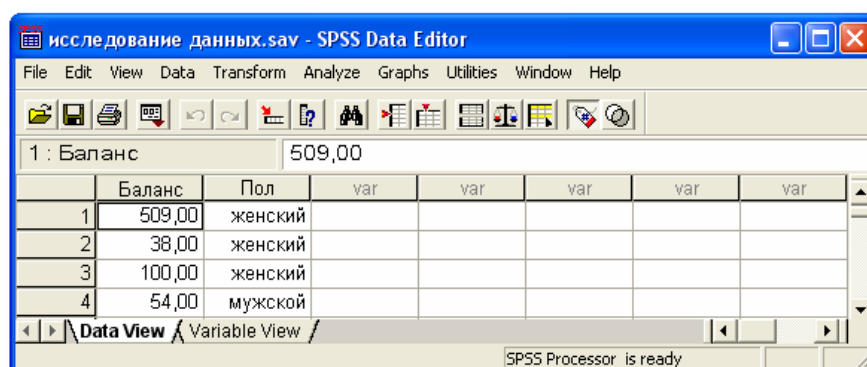
	<p>групп разрядов. Если запятые опускаются при вводе, они вставляются автоматически. Длина такой переменной равна максимальному количеству знаков, включая десятичный разделитель и запятые между группами разрядов.</p>
Dot (<i>Точка</i>)	<p>К допустимым значениям относятся числа, перед которыми стоит знак плюс или минус, запятая как десятичный разделитель и одна или несколько точек в качестве разделителей групп разрядов. Если точки опускаются при вводе, они вставляются автоматически.</p>
Scientific notation (<i>Научный формат</i>)	<p>При вводе данных разрешаются все допустимые числовые форматы, включая научный, о котором свидетельствует содержащаяся в числе буква E или D, а также знак плюс или минус.</p>
Date (<i>Дата</i>)	<p>Допустимые значения – дата и/или время.</p>
Dollar (<i>Доллар</i>)	<p>К допустимым значениям относятся: знак доллара, точка как десятичный разделитель и запятые как разделители групп разрядов. Если знак доллара или запятые опускаются при вводе, они вставляются автоматически.</p>
Special currency (<i>Специальная валюта</i>)	<p>Пользователь может задавать собственные форматы валюты. В поле Length в этом случае задается максимальное количество знаков, включая все знаки, заданные пользователем. Обозначение валюты при вводе не указывается; оно вставляется автоматически.</p>
String (<i>Строчный</i>)	<p>Строка символов. К допустимым значениям относятся: буквы, цифры и специальные символы. Различаются короткие и длинные строковые переменные. Короткие строковые переменные могут содержать не более восьми знаков. В большинстве процедур SPSS применение длинных строковых переменных ограничивается или вообще не допускается.</p>

Переменная **Баланс** имеет тип Dot и относится к метрическому (интервальному) типу переменных (в колонке Measure (*Шкала измерения*) указан тип Scale (*Интервальная*)), переменная **Пол** имеет тип Numeric и относится к категориальному типу переменных (в колонке Measure (*Шкала измерения*) указан тип Nominal (*Номинальная*)), в колонке Values (*Метки значений*) значению 1 приписан «мужской пол», значению 2 – «женский». Для решаемой задачи информация о переменных выглядит следующим образом:



*Рис. 1.1. Окно редактора данных для файла **Исследование данных.sav**, вкладка Variable View*

Далее, на вкладке Data View вводятся значения объявленных переменных. По завершении ввода наблюдений сохраним файл данных, воспользовавшись пунктами меню File (*Файл*) и Save as (*Сохранить как*). Каждый столбец таблицы, как видим, соответствует той переменной, имя которой указано в названии столбца. Порядок столбцов на вкладке Data View такой же, как и порядок строк на вкладке Variable View.



*Рис. 1.2. Окно редактора данных для файла **Исследование данных.sav**, вкладка Data View*

Частотный анализ. Меню Frequencies... При анализе большого количества информации ее необходимо прежде всего представить в более наглядном виде, поэтому первым этапом статистического анализа данных, как правило, является частотный анализ. Для того чтобы провести частотный анализ, выберем в

меню Analyze (*Анализ*) пункт Descriptive Statistics (*Описательные статистики*), а в выпавшем меню – пункт Frequencies... (*Частотный...*). Появится диалоговое окно Frequencies. Поскольку нас будут интересовать числовые характеристики, связанные с переменной **Баланс**, то из общего списка переменных отправляем **Баланс** в список анализируемых переменных – поле Variable(s) (*Переменные*), – кликнув на «стрелку»:

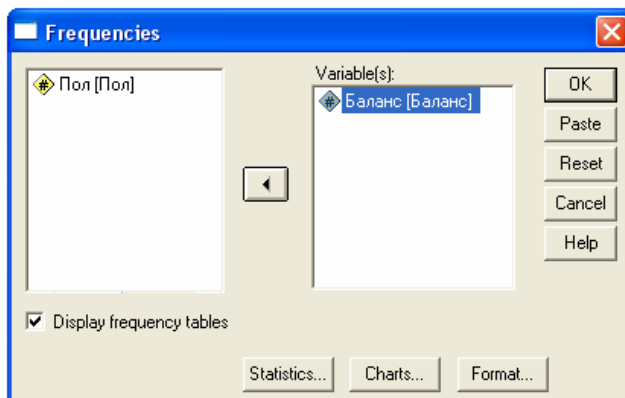


Рис. 1.3. Диалоговое окно Frequencies

Чтобы получить описательные статистики, кликнем в диалоговом окне Frequencies (*Частотный*) на кнопку Statistics... (*Статистики...*). Откроется диалоговое окно Frequencies: Statistics (*Частотный: Статистики*).

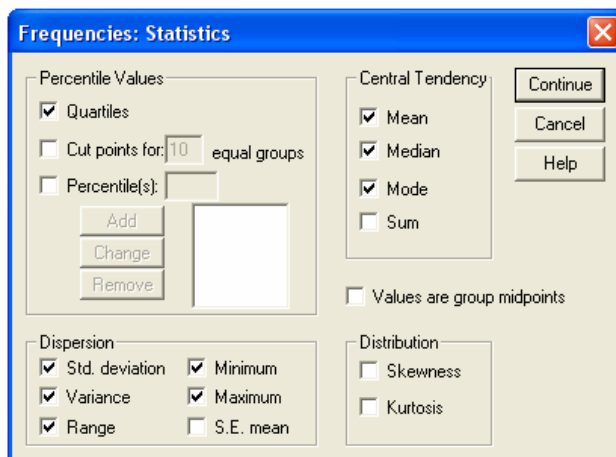


Рис. 1.4. Диалоговое окно Frequencies: Statistics

В блоке Percentile Values (*Значения процентов*) можно выбрать следующие варианты:

✓ **Quartiles** (*Квартили*). Будут показаны первый, второй и третий квартили. Первый квартиль (Q_1) – это точка на шкале измеренных значений, ниже (левее) которой располагаются 25% значений. Второй квартиль (Q_2) – это точка, ниже которой располагаются 50% измеренных значений. Второй квартиль также

называется медианой. Третий квартиль (Q_3) – это точка на шкале измеренных значений, ниже которой располагаются 75% значений. Если данные имеются только в форме порядкового отношения, то в качестве меры разброса используется межквартильная широта IQR . Она определяется как $IQR = Q_3 - Q_1$.

✓ **Cut points (Точки раздела)**. Будут вычислены значения процентилей, разделяющие выборку на группы наблюдений, которые имеют одинаковую ширину, то есть включают одно и то же количество измеренных значений. По умолчанию предлагается количество групп 10. Если задать, к примеру, 4, то будут показаны квартили, то есть квартили соответствуют процентилем 25, 50 и 75. Видно, что число показываемых процентилей на единицу меньше заданного числа групп.

✓ **Percentile(s) (Процентили)**. Здесь имеются в виду значения процентилей, определяемые пользователем. Введите значение перцентиля в пределах от 0 до 100 и кликнете на кнопку Add (*Добавить*). Повторите эти действия для всех желаемых значений процентилей. Значения в порядке возрастания будут показаны в списке. Например, если ввести значения 25, 50 и 75, то мы получим квартили. Можно задавать любые значения процентилей, например, 37 и 83. В первом случае (37) будет показано значение выбранной переменной, ниже которого лежат 37% наблюдений, а во втором случае (83) – значение, ниже которого располагаются 83% наблюдений.

В блоке Dispersion (*Разброс*) можно выбрать следующие меры разброса:

✓ **Std. deviation (Стандартное отклонение)**. Стандартное отклонение – это мера разброса измеренных величин; оно равно квадратному корню из дисперсии: $S = \sqrt{D(X)}$. В интервале шириной $2S$, который отложен по обе стороны от выборочного среднего значения \bar{x}_B , располагается примерно $2/3$ всех значений выборки, подчиняющейся нормальному распределению.

✓ **Variance (Дисперсия)**. Дисперсия – это квадрат стандартного отклонения и, следовательно, также является мерой разброса измеренных величин:

$$D(X) = \frac{\sum_{i=1}^n (x_i - \bar{x}_B)^2}{n-1}, \text{ где } n - \text{объем выборки, } x_i - \text{наблюдаемые значения, } \bar{x}_B -$$

выборочное среднее.

✓ **Range (Размах).** Размах – это разница между наибольшим (максимумом) и наименьшим (минимумом) значением: $\Delta = x_{\max} - x_{\min}$.

✓ **Minimum (Минимум).** Наименьшее значение в выборке x_{\min} .

✓ **Maximum (Максимум).** Наибольшее значение в выборке x_{\max} .

✓ **S.E. mean (Стандартная ошибка).** Это стандартная ошибка среднего значения: $S_E = \frac{S}{\sqrt{n}}$. В интервале шириной, равной $2S_E$, отложенному по обе сто-

роны от выборочного среднего значения \bar{x}_B , располагается среднее значение генеральной совокупности с вероятностью $2/3$.

В блоке Central Tendency (*Центральная тенденция*) имеются следующие характеристики:

✓ **Mean (Среднее).** Это выборочное среднее значение: $\bar{X}_B = \frac{1}{n} \left(\sum_{i=1}^n x_i \right)$.

✓ **Median (Медиана).** Медиана (*Me, m*) – это точка на шкале измеренных значений, выше и ниже которой лежит по половине всех измеренных значений. Например, если измеренные значения таковы: 37854639284, то сначала они располагаются в порядке возрастания: 23344567889. В данном случае медианой будет значение 5. Всего у нас 11 измеренных значений, следовательно, медианой является шестое значение. Выше него располагается 5 значений, и ниже – тоже 5. При нечетном количестве значений медиана всегда будет совпадать с одним из измеренных значений. При четном количестве медиана будет средним арифметическим двух соседних «центральных» значений. Например, если имеются следующие измеренные значения: 3445678899, то медиана в этом случае будет равна: $(6 + 7) / 2 = 6,5$.

✓ **Mode (Мода).** Мода (*Mo*) – это наиболее часто встречающееся значение в выборке. Если одна и та же наибольшая частота встречается у нескольких значений, то выбирается наименьшее из них.

- ✓ Sum (Сумма). Сумма всех значений.

В блоке Distribution (Распределение) можно выбрать следующие меры несимметричности распределения:

- ✓ Skewness (Коэффициент асимметрии). Коэффициент асимметрии – это мера отклонения распределения частоты от симметричного распределения, то есть такого, у которого на одинаковом удалении от среднего значения по обе стороны выборки данных располагается одинаковое количество значений. В зависимости от того, какая характеристика более информативна (а это, в свою очередь, связано с типом и шкалой измерения исследуемой переменной) – медиана или мода, – коэффициент асимметрии вычисляется по одной из формул:

$$Sk = \frac{3(\bar{x}_B - m)}{S} \quad \text{или} \quad Sk = \frac{\bar{x}_B - Mo}{S}.$$

Если наблюдения подчиняются нормальному распределению, то асимметрия равна нулю. Для проверки на нормальное распределение можно применять следующее правило: *если асимметрия значительно отличается от нуля, то гипотезу о том, что данные взяты из нормально распределенной генеральной совокупности, следует отвергнуть*. Если вершина асимметричного распределения сдвинута к меньшим значениям, то говорят о положительной асимметрии, в противном случае – об отрицательной.

- ✓ Kurtosis (Коэффициент вариации или эксцесс). Коэффициент вариации указывает, является ли распределение пологим (при большом – более

50% – значения коэффициента) или крутым: $K_x = \frac{S}{\bar{x}_B} \cdot 100\%$. Коэффициент

вариации равен нулю, если наблюдения подчиняются нормальному распределению. Поэтому для проверки на нормальное распределение можно применять еще одно правило: *если коэффициент вариации значительно отличается от нуля, то гипотезу о том, что данные взяты из нормально распределенной генеральной совокупности, следует отвергнуть*.

Как правило, для переменных, относящихся к метрической (интервальной) шкале и подчиняющихся нормальному распределению, в качестве основной ха-

рактические используют среднее значение, а в качестве меры разброса – стандартное отклонение или стандартную ошибку. Для порядковых или интервальных переменных, не подчиняющихся нормальному распределению, – соответственно медиану или первый и третий квартили. Для переменных, относящихся к номинальной шкале, нельзя дать других значимых характеристик кроме моды.

Для переменной **Баланс** вычислим следующие характеристики: квартили, стандартное отклонение, дисперсию, размах вариации, максимум, минимум, среднее, медиану и моду. Кликнем на кнопку Continue и вернемся в диалоговое окно Frequencies. После клика на кнопку Ok откроется окно вывода, которое сохраним как **Исследование данных.spo**.

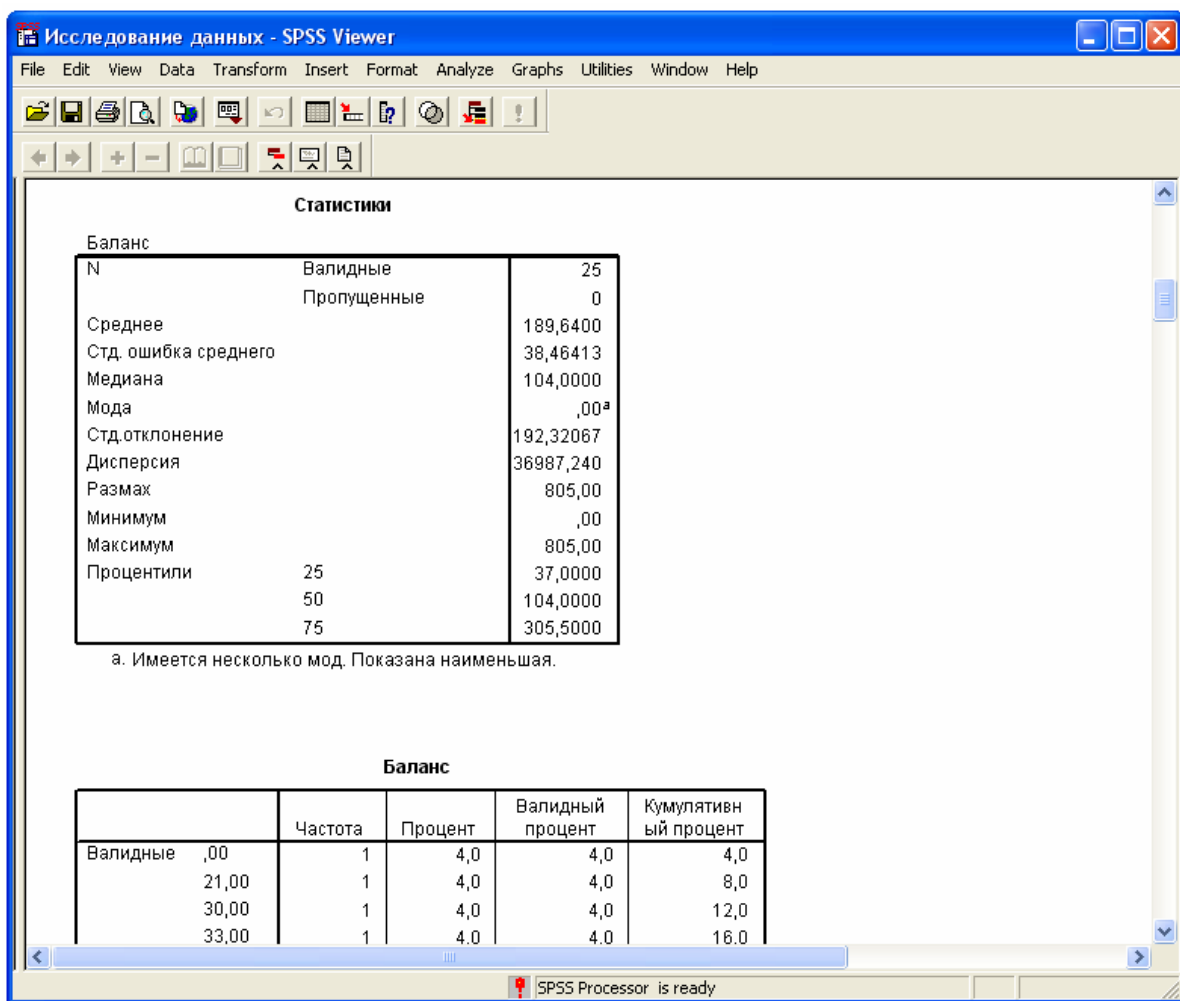


Рис. 1.5. Окно вывода для файла **Исследование данных.spo**

Первая таблица (Статистики) содержит значения выбранных числовых характеристик для переменной **Баланс**, вторая (Баланс) – значения абсолютных частот для этой переменной. Последняя таблица не несет значимой информации, поскольку все значения переменной **Баланс** различны, и соответствующие час-

тоты равны единице (такая таблица удобна, если исследуемая переменная относится к категориальному типу с небольшим числом категорий), а вот статистические характеристики, напротив, вычисляются в основном для переменных, относящихся к интервальной шкале.

Графическое представление данных в меню Frequencies... Результаты частотного распределения можно представить графически. Специальные возможности для этого собраны в группе процедур меню Graphs (*Графики*), но в наиболее простых случаях можно поступить, например, так: выбрать в меню Analyze (*Анализ*) пункт Descriptive Statistics (*Описательные статистики*), а в нем выбрать пункт Frequencies... (*Частотный...*) и в открывшемся окне Frequencies затем кликнуть на кнопку Charts... (*Диаграммы...*). Появится диалоговое окно Frequencies: Charts (*Частотный: Диаграммы*), в котором в блоке Chart Type (*Тип диаграммы*) по умолчанию включена радиокнопка None (*никакая*).

В блоке Chart Type можно выбрать один из следующих типов диаграмм:

- Bar charts (*Столбиковые диаграммы*);
- Pie charts (*Круговые (секторные) диаграммы*);
- Histograms (*Гистограммы*), при активации элемента управления With normal curve (*С нормальной кривой*) на гистограмме появится нормальная кривая, наиболее точно описывающая распределение данных.

В блоке Chart Value (*Значение диаграммы*) можно выбрать, что будут отражать диаграммы – частоты (Frequencies) или проценты (Percentages). Заметим, что для гистограммы эта опция недоступна, так как на гистограмме по умолчанию отражаются частоты.

Создадим, например, гистограмму частот для переменной **Баланс** вместе с кривой нормального распределения. Для этого выберем в меню Analyze (*Анализ*) пункт Descriptive Statistics (*Описательные статистики*), а в нем – пункт Frequencies... (*Частотный...*). В открывшемся окне Frequencies перенесем переменную **Баланс** в список анализируемых переменных. Далее, кликнув на кнопку Charts... (*Диаграммы...*), в открывшемся диалоговом окне Frequencies: Charts (*Частотный: Диаграммы*) выберем в блоке Chart Type (*Тип диаграммы*) пункт

Histograms (*Гистограммы*) и поставим галочку на элементе управления With normal curve. Подтвердим выбор кликом на кнопку Continue (*Продолжить*) и вернемся в окно Frequencies. Здесь снимем галочку с элемента управления Display frequency tables (*Отобразить частотные таблицы*), и после клика на кнопку ОК гистограмма будет показана в окне вывода.

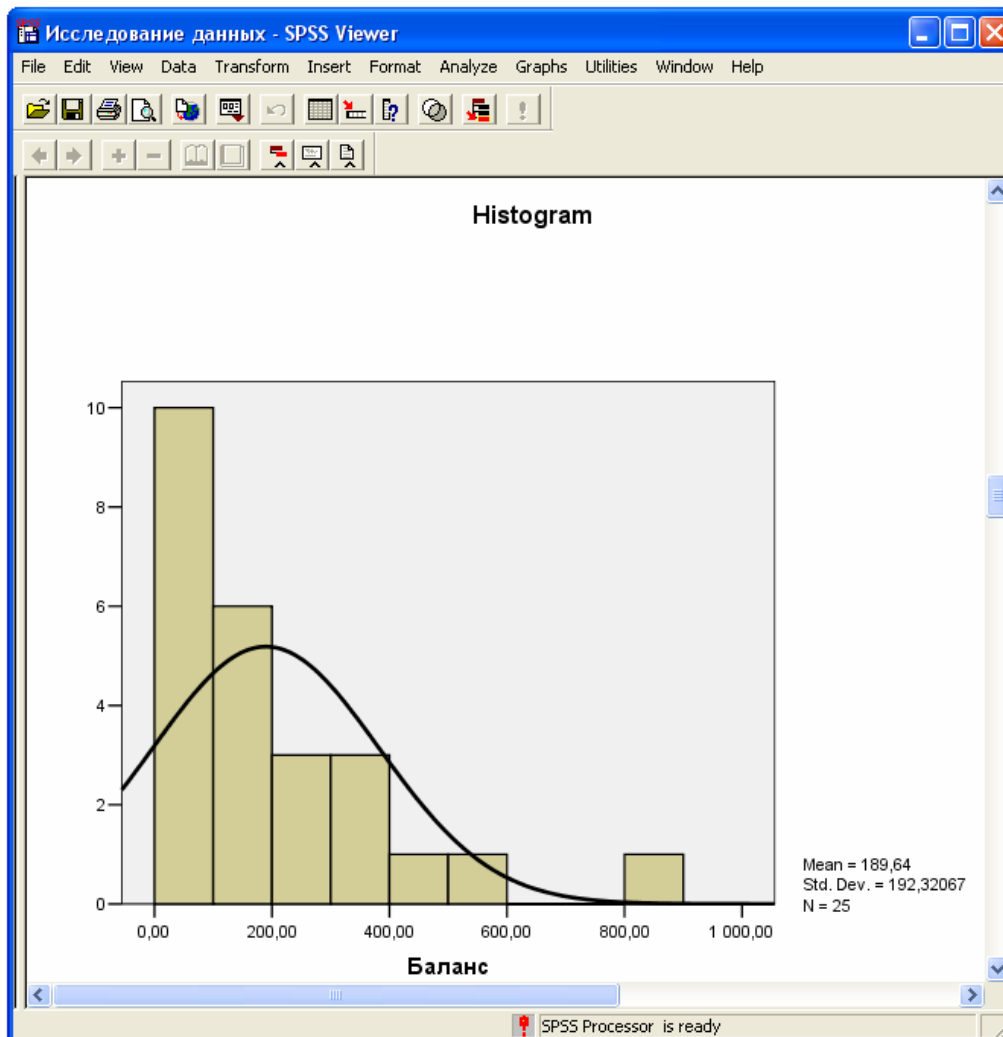


Рис. 1.6. Окно вывода для файла *Исследование данных.spo*

Вид гистограммы как графического объекта можно редактировать. Чтобы начать редактирование, нужно дважды кликнуть в области диаграммы. Диаграмма будет показана в редакторе диаграмм Chart Editor, в котором доступны различные опции (как в стандартном графическом редакторе). Среди прочих стоит отметить опцию Show Data Labels (*Показать метки данных*), позволяющую на каждом элементе диаграммы (в нашем случае – на столбцах гистограммы) показывать его значение (частоту). После клика на пиктограмму Show Data Labels откроется диалоговое окно Properties (*Свойства*). На вкладке Data Value Labels (*Значения меток данных*) в поле Displayed (*Отображаемые*) необходимо вы-

брать элемент Count (*Частота*), кликнуть на кнопку Apply (*Применить*) и затем на кнопку Close (*Заккрыть*). На каждом столбце появится надпись со значением частоты.

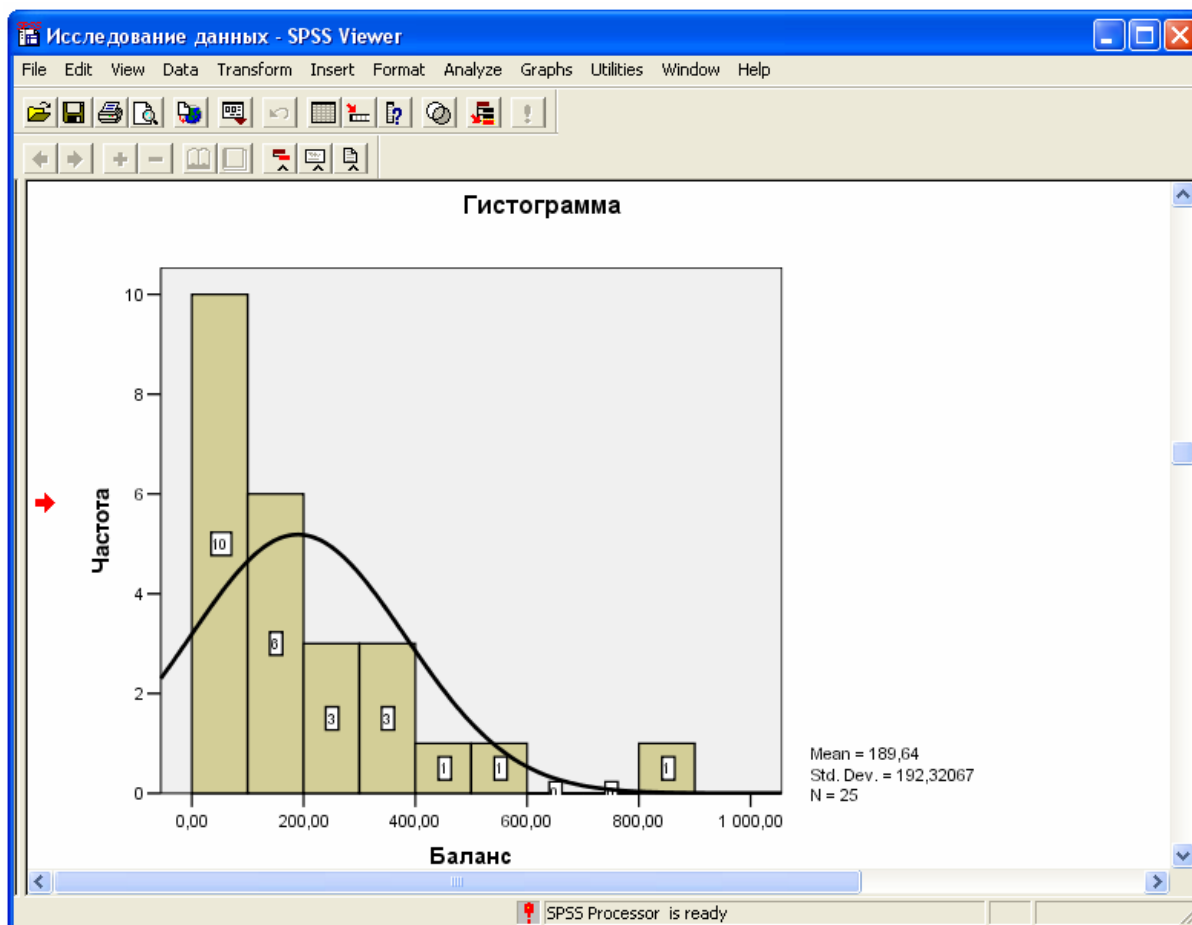


Рис. 1.7. Окно редактора диаграмм Chart Editor

Форматы частотных таблиц. Таблица частот допускает следующие форматы отображения, связанные с типом сортировки данных:

- ✓ Ascending values (*По возрастанию значений*): Данные сортируются по возрастанию значений. Это настройка по умолчанию.
- ✓ Descending values (*По убыванию значений*): Данные сортируются по убыванию значений.
- ✓ Ascending counts (*По возрастанию частот*): Данные сортируются по возрастанию частот.
- ✓ Descending counts (*По убыванию частот*): Данные сортируются по убыванию частот.

Кроме того, активация элемента управления Suppress tables with more than n categories (*Не выводить таблицы с более чем n категориями*) позволяет избегать вывода длинных частотных таблиц.

Описательные статистики. Меню Descriptives... Выберем в меню Analyze (*Анализ*) пункт Descriptive Statistics (*Описательные статистики*), а в нем – пункт Descriptives... (*Описательный...*). Откроется диалоговое окно Descriptives. Поместим переменную **Баланс** в список анализируемых переменных и кликнем на кнопку Options... (*Опции...*). Появится диалоговое окно Descriptives: Options (*Описательный: Опции*):

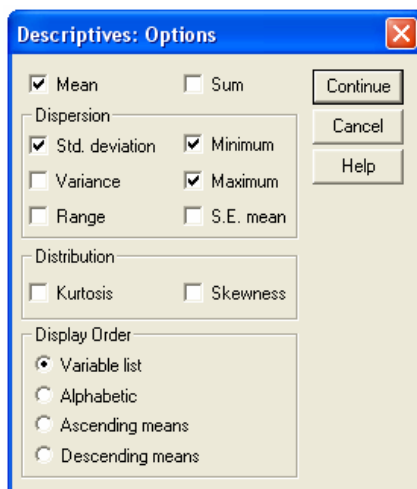


Рис. 1.8. Диалоговое окно Descriptives: Options

Поставим галочки для тех статистик, которые хотим получить в окне вывода. Кликнем на кнопку Continue, а затем на кнопку Ok. В окне вывода появится следующая таблица:

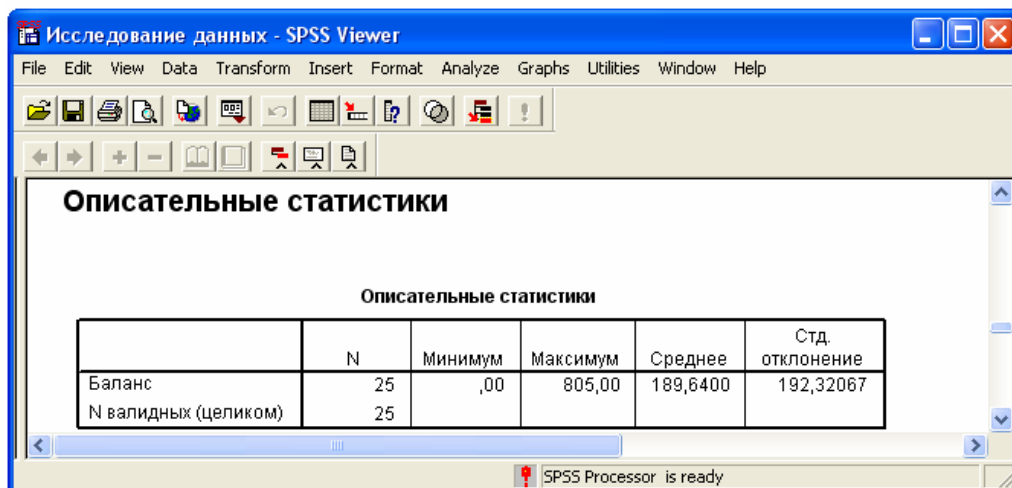


Рис. 1.9. Окно вывода для файла Исследование данных.spo

Как мог заметить читатель, процедуры, вызванные из меню Frequencies и Descriptives, вычисляют статистические характеристики, однако формат вывода и спектр возможностей меню Frequencies несколько шире. Наиболее же полную картину статистического анализа данных дает процедура Explore.

Исследование данных с помощью Explore... Процедура Explore позволяет провести наиболее полный статистический анализ, а также обеспечивает более широкие возможности графического представления данных. Выберем в меню Analyze (*Анализ*) пункт Descriptive Statistics (*Описательные статистики*), а в нем – пункт Explore... (*Разведочный...*). Откроется диалоговое окно Explore:

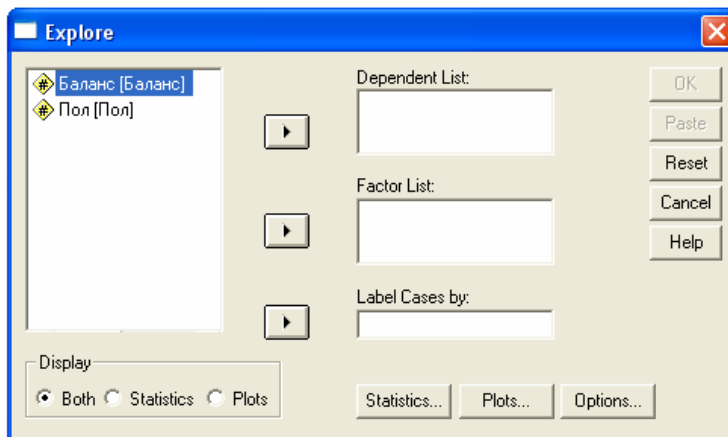


Рис. 1.10. Диалоговое окно Explore

Видим, что в этом диалоговом окне проводится различие между зависимыми и независимыми переменными (факторами). Это означает, что можно выполнять анализ как всего набора данных, так и отдельно по группам наблюдений. В этом случае анализируемой переменной будет зависимая переменная, а группирующей переменной – независимая (фактор). Если же такой отдельный анализ проводить не требуется, список факторов не используется. Продемонстрируем оба способа анализа на рассматриваемом примере.

Анализ без группирующей переменной. Перенесем переменную **Баланс** в список зависимых переменных (Dependent List), при этом поле Factor List (*Список факторов*) оставим незаполненным. Поначалу выясним, какие методы анализа выполняются по умолчанию, поэтому не будем вносить изменений в настройки. После клика на кнопку ОК в окне вывода будут созданы следующие объекты: таблица Case Processing Summary (*Сводка обработки наблюдений*) и таблица Descriptives (*Описательные статистики*) (см. рис. 1.11 ниже), диаграмма Stem-and-leaf plot («ствол и лист») (см. рис. 1.12 ниже) и диаграмма Box plot («ящик с усами») (см. рис. 1.13 ниже). Рассмотрим каждый из них.

Исследовать

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Всего	
	N	Процент	N	Процент	N	Процент
Баланс	25	100,0%	0	,0%	25	100,0%

Описательные

			Статистика	Стд. ошибка
Баланс	Среднее		189,6400	38,46413
	95% доверительный интервал для среднего	Нижняя граница	110,2539	
		Верхняя граница	269,0261	
	5% усеченное среднее		169,0444	
	Медиана		104,0000	
	Дисперсия		36987,240	
	Стд. отклонение		192,32067	
	Минимум		,00	
	Максимум		805,00	
	Размах		805,00	
	Межквартильный размах		268,50	
	Асимметрия		1,638	,464
	Эксцесс		3,079	,902

SPSS Processor is ready

Рис. 1.11. Окно вывода для файла *Исследование данных.spo*

Таблица Case Processing Summary (Сводка обработки наблюдений) содержит информацию об общем числе наблюдений, количестве валидных (используемых при анализе) и пропущенных значениях, выраженных также и в процентах.

Таблица Descriptives (Описательные) содержит основные статистические характеристики переменной *Баланс*. Большую их часть мы уже рассматривали ранее, но появились и новые характеристики (все они выводятся по умолчанию):

- ✓ 5% усеченное среднее – это среднее значение, вычисленное без учета 5% наименьших и 5% наибольших значений;
- ✓ 95% доверительный интервал для среднего;
- ✓ межквартильный размах (IQR).

Кроме статистических характеристик, вычисляемых по умолчанию, процедура Explore позволяет вычислить, например, M-оценки Губера, Тьюки, Эндрюса и Хампеля (*M-estimators*), найти выбросы (*Outliers*) и процентиля (*Percentiles*).

Основная идея М-оценок состоит в том, чтобы перед вычислением среднего значения присвоить отдельным наблюдениям разные веса. Так, в распространенных М-оценках применяются веса, уменьшающиеся с удалением от центра распределения. Кроме этого, в меню Explore есть обширный инструментарий, связанный с проверкой гипотез, например, о нормальном распределении данных и опирающийся на различные тесты: например, тесты Лиллифора (модификации теста Колмогорова-Смирнова) на нормальное распределение, а при объеме выборки менее 50 наблюдений, например, тест Шапиро-Уилкса. Но на этом мы подробно останавливаться не будем, предлагая читателю самостоятельно изучить данные возможности среды SPSS, поскольку наша цель – первичный статистический анализ.

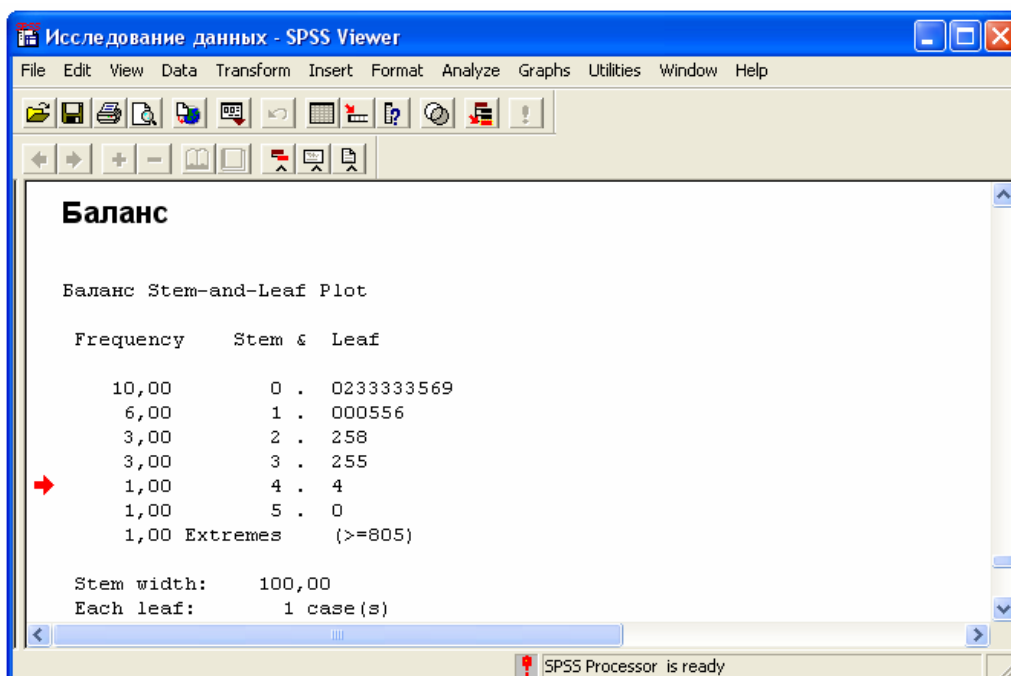


Рис. 1.12. Окно вывода для файла *Исследование данных.spo*

Диаграмма Stem-and-leaf plot («ствол и лист») представляет собой комбинацию гистограммы и табличного списка. Более подробно к рассмотрению вопроса о построении данной диаграммы мы так же вернемся несколько позже, а сейчас просто прокомментируем полученную диаграмму: ширина шага для «ствола» равна 100 руб., ширина шага для «листа» – 10 руб., оба шага выбираются программой автоматически; наибольшую частоту (10) имеют значения баланса от 0 до 99 руб. (ветви «0» на стволе соответствуют 10 «листьев»), в данных имеется один выброс (экстремальное значение).

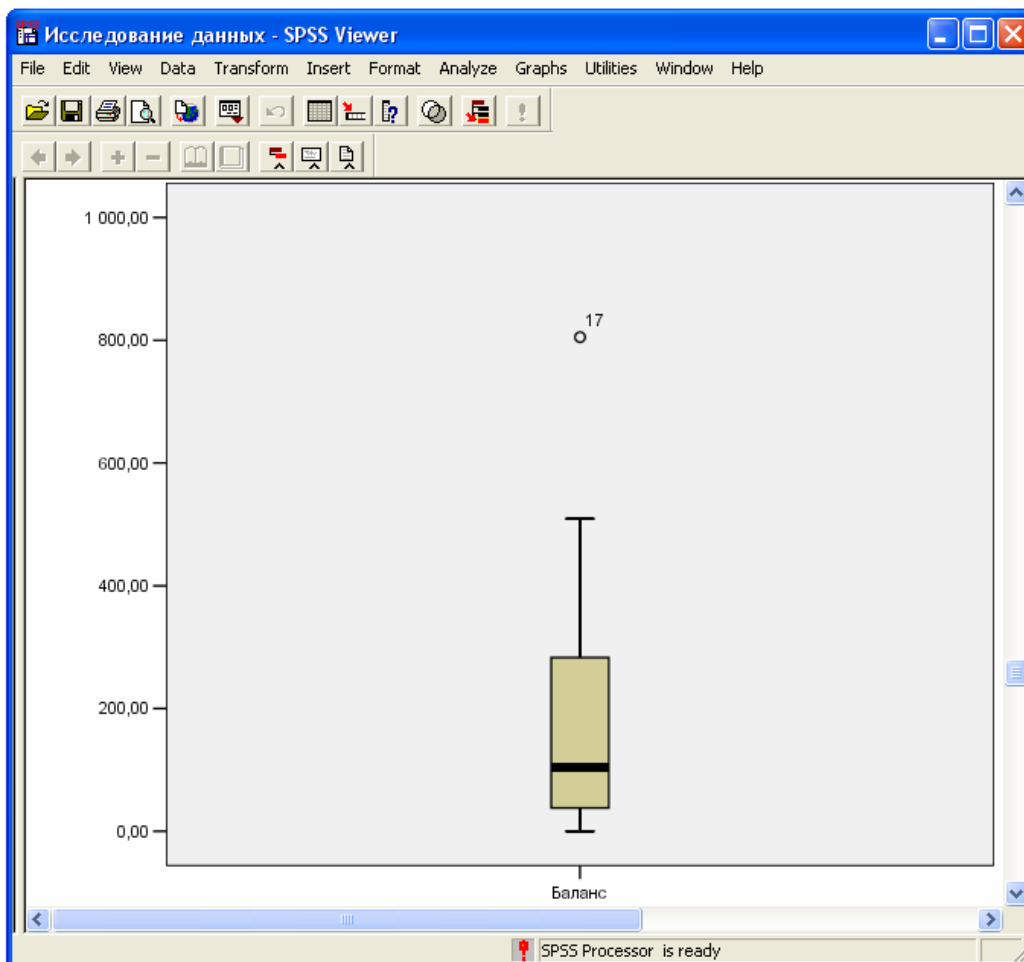


Рис. 1.13. Окно вывода для файла *Исследование данных.spo*

Диаграмма Box plot («ящик с усами») показывает распределение данных внутри выборки. По виду диаграммы можем сказать следующее: распределение данных несимметрично, медиана смещена ближе к первому квартилю (нижняя граница «ящика с усами»), имеется один умеренный выброс, соответствующий наблюдению №17, равный 805 руб. (умеренные выбросы обозначаются кружками, а экстремальные – звездочками). Более подробно этот тип диаграмм будет разобран несколько позже в главе «Графическое представление данных».

Анализ для групп наблюдений. Выполним анализ данных с учетом пола респондентов (исходная выборка теперь распадается на две части). В диалоговом окне Explore кнопкой Reset (*Сброс*) восстановим настройки по умолчанию, перенесем переменную **Баланс** в список зависимых переменных (Dependent List), а переменную **Пол** – в список факторов (Factor List). Затем кликнем на кнопку ОК. В результате будут вычислены описательные статистики (см. рис. 1.14 и 1.15 ниже), построены диаграмма «ствол и лист» (см. рис. 1.16 ниже) и диаграмма «ящик с усами» (см. рис. 1.17 ниже) отдельно по двум значениям фактора.

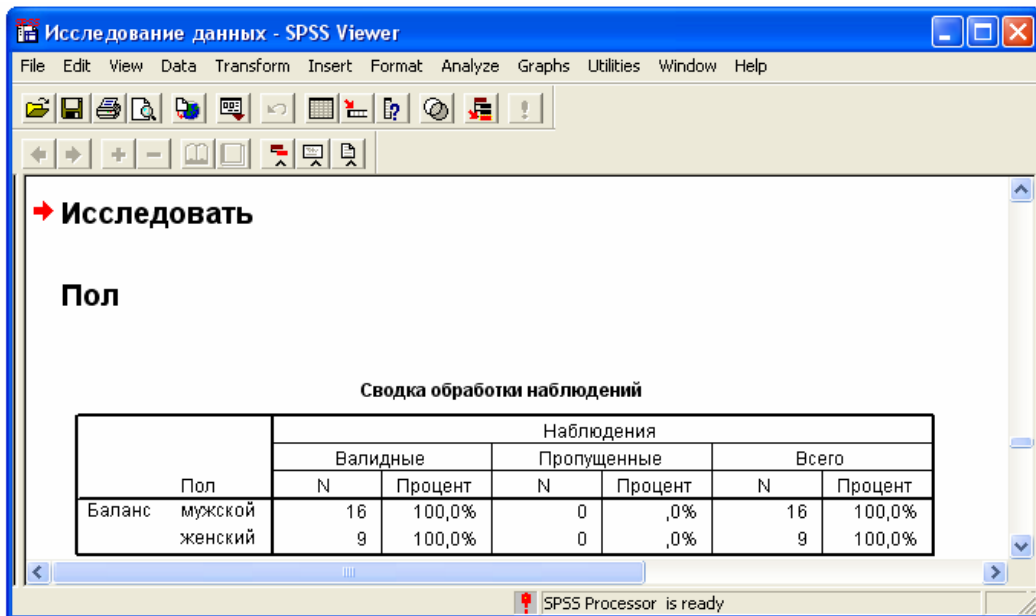


Рис. 1.14. Окно вывода для файла *Исследование данных.spo*

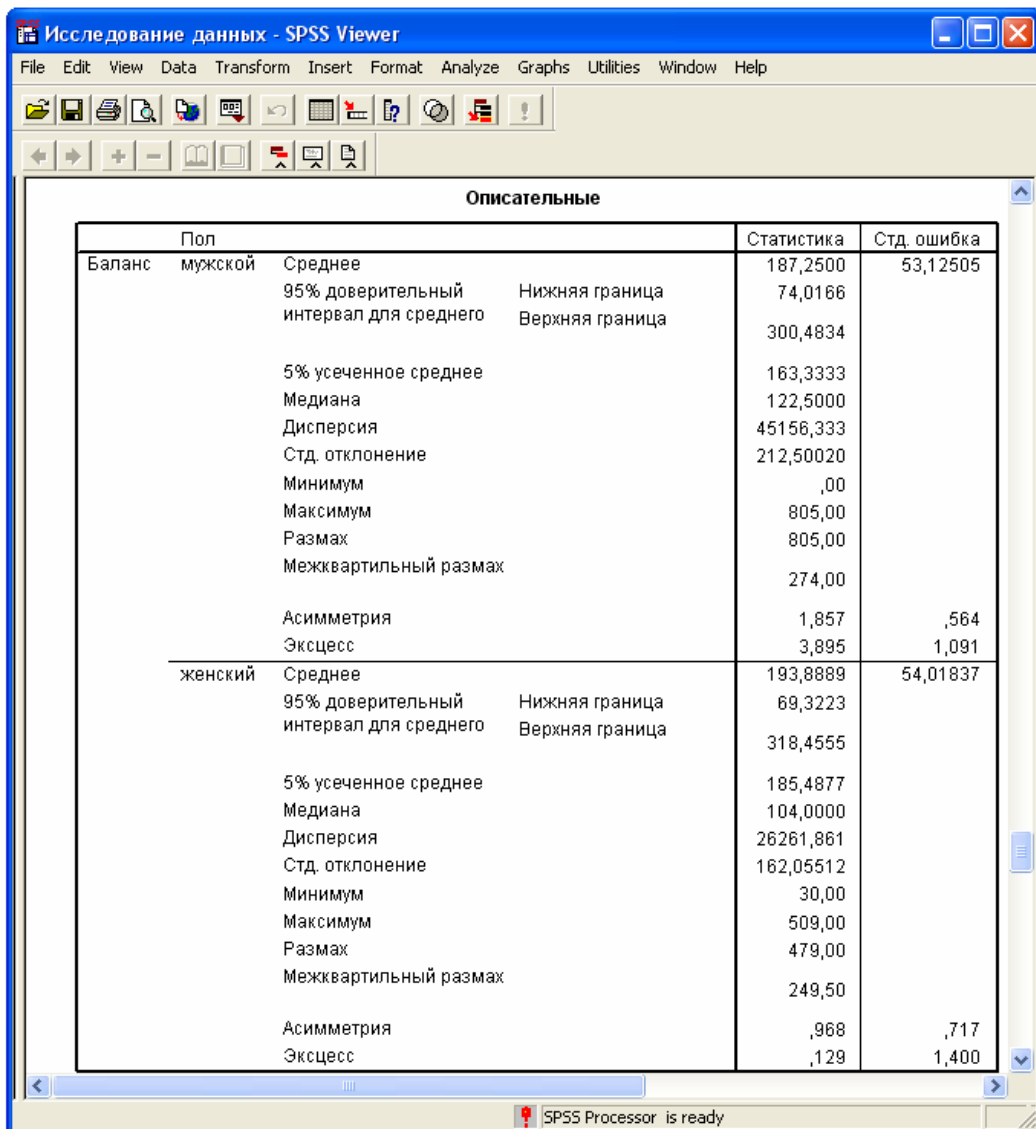


Рис. 1.15. Окно вывода для файла *Исследование данных.spo*

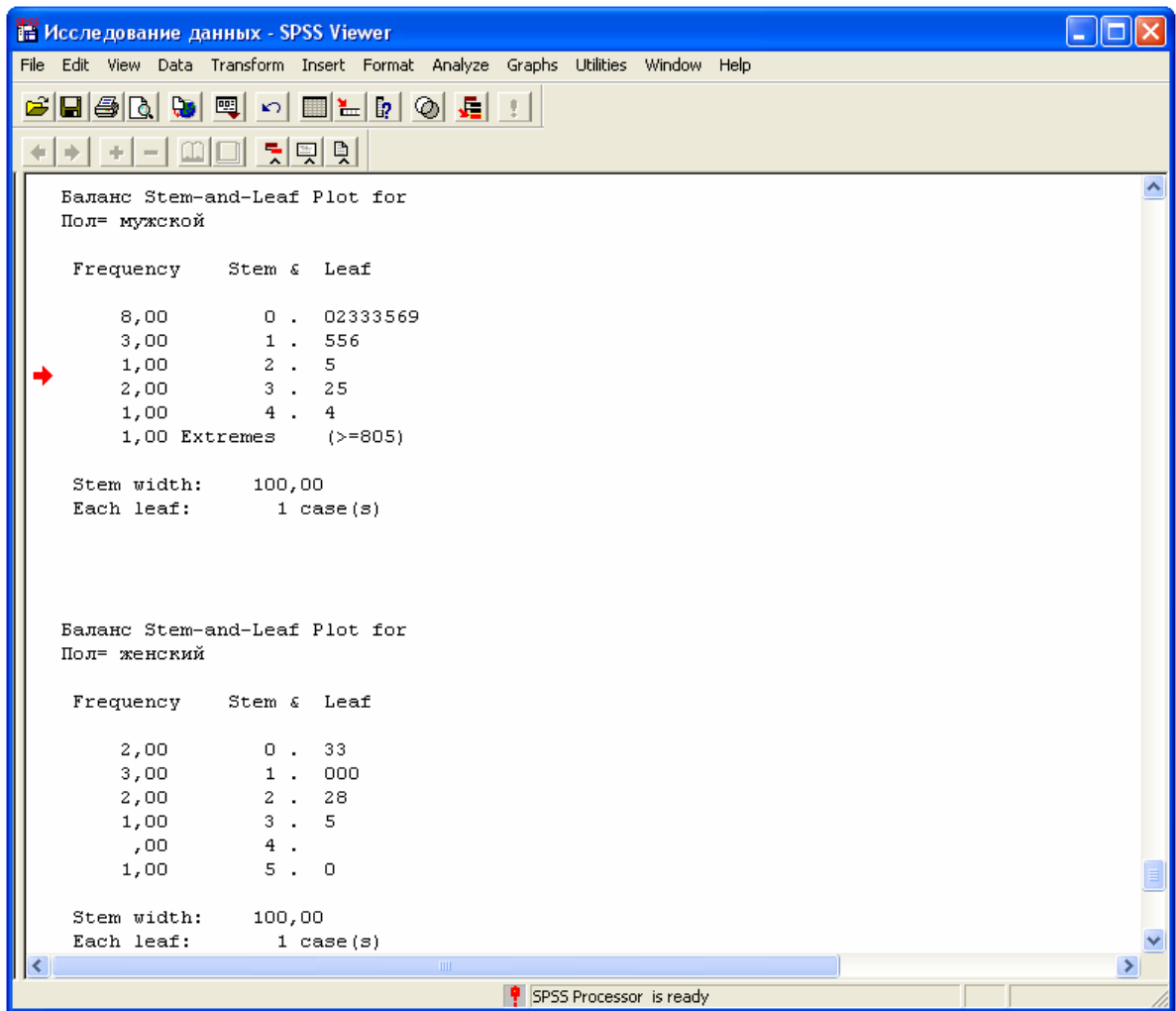


Рис. 1.16. Окно вывода для файла *Исследование данных.spo*

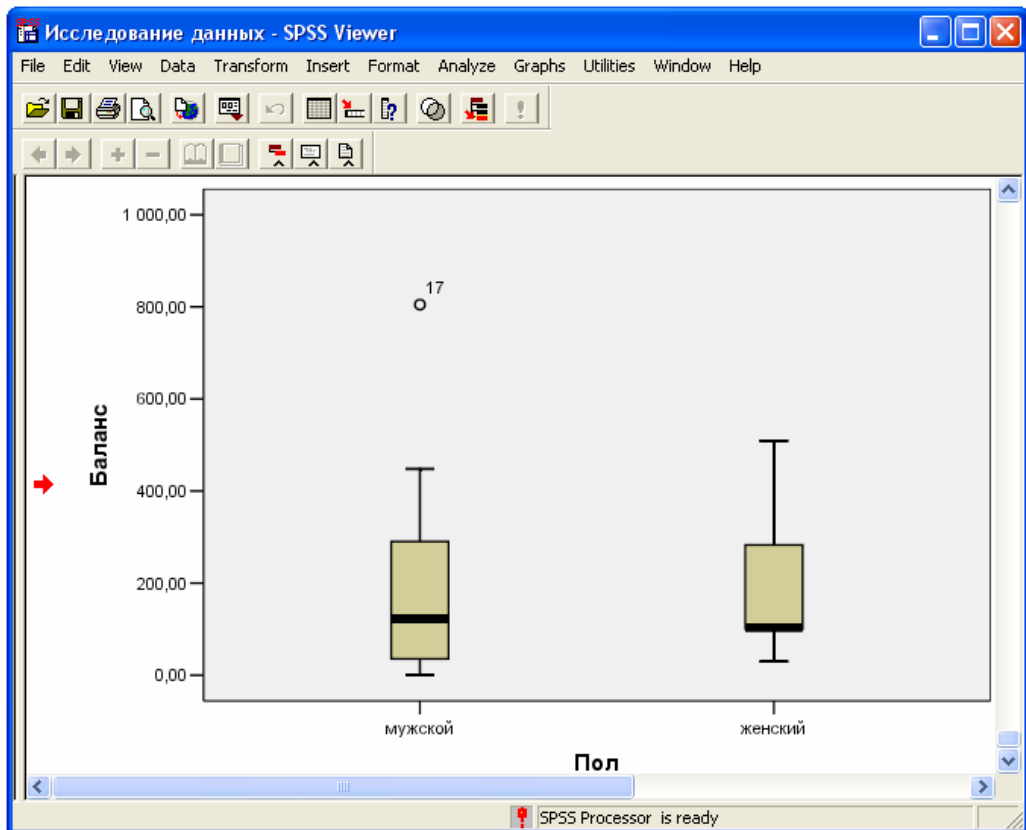


Рис. 1.17. Окно вывода для файла *Исследование данных.spo*

Диаграммы «ящик с усами» отображаются вместе для двух значений фактора, что позволяет быстро и наглядно сравнивать различные группы данных. Видим, что выброс со значением 805 руб. соответствует «мужской» части наблюдений, медианы расположены практически на одном уровне, при этом для значения пола «мужской» медиана расположена почти на середине «ящика», а для значения «женский» – почти совпадает с его нижней границей.

Мы рассмотрели основные инструменты статистического анализа, и теперь переходим к подробному изучению графических возможностей SPSS.

II. Графическое представление данных в SPSS

Одним из достоинств пакета SPSS является наличие большого количества разнообразных графиков, которые могут быть построены как при помощи процедур меню графиков, так и из разнообразных процедур меню статистик.

Начиная с версии SPSS 8.0, наряду с традиционными стандартными графиками существует возможность создавать и интерактивные графики. Стандартные графики строятся при помощи многочисленных процедур статистического меню или меню графиков, составные компоненты которых и соответственно их возможности несколько не изменились. Однако в меню графиков добавилась еще одна позиция – *Interactive (Интерактивные)*, которая открывает еще одно собственное меню, служащее для построения так называемых интерактивных графиков. Интерактивные графики дают довольно широкую палитру новых возможностей. Наряду с удобными глобальными возможностями менять отдельные стиливые элементы графиков и преобразовывать переменные, используемые для построения графика, при помощи интерактивных графиков становится также возможным одновременное построение нескольких графиков для отдельных категорий дополнительных переменных.

В рассмотренных примерах будут использованы как хорошо известные широкому кругу читателей графические объекты (такие как гистограмма абсолютных частот, столбиковая и секторная диаграммы, линейный график), так и более специализированные графики, с которыми, возможно, читатель не знаком. Поэтому, прежде всего, остановимся подробно на каждом из них.

1. Диаграмма **Box plot** («ящик с усами»)

Диаграмма типа «ящик с усами» (или коробчатая диаграмма) была предложена в 70-е годы XX века известным американским статистиком Дж. Тьюки, одним из создателей практического анализа данных. Для построения этой диаграммы используют так называемые 5 базовых показателей (медиану, квартили и наименьшее и наибольшее значения в выборке), несущих важную информацию как о диапазоне числовой оси, в котором оказались первичные данные, так и об их взаимном расположении внутри выборки. Рассмотрим эти понятия подробнее, хотя мы уже сталкивались с ними в первой главе.

Медианой (Me, m) выборки является значение (само оно может как принадлежать, так и не принадлежать этой выборке), которое разбивает ее на две равные части: половина наблюдений лежит ниже (не выше) медианы, и половина – выше (не ниже). Из сказанного вытекает, что для нахождения медианы выборочные значения нужно сначала упорядочить по возрастанию, получив так называемый вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Далее, если $n = 2k + 1$ (то есть n – нечетное число), то $m = x_{(k+1)}$, а если $n = 2k$ – четно, то $m = \frac{x_{(k)} + x_{(k+1)}}{2}$. Медиана является частным случаем понятия *квантиль*. Так, p -й квантиль, где $0 \leq p \leq 1$, есть такое значение x_p в выборке, ниже (не выше) которого в вариационном ряду лежит его p -я часть. Иными словами, p -я часть чисел выборки не превосходит числа x_p . Если указанная доля чисел выборки, не превосходящих x_p , выражена не в долях единицы, а в процентах, говорят о соответствующих процентилях. Наиболее часто употребляются квантили, связанные с долями $p = 0,25$ (25%), $p = 0,5$ (50%), $p = 0,75$ (75%). Их называют *квартилями*³: $x_{0,25} \equiv Q_1$ – 1-й (или нижний) квартиль, $x_{0,5} \equiv Q_2 \equiv m$ – 2-й (или центральный) квартиль, или медиана, $x_{0,75} \equiv Q_3$ – 3-й (или верхний) квартиль. В промежутке между нижним и верхним квартилями лежат 50% «центральных» значений выборки.

Одним из возможных методов вычисления квартилей является снабжение их порядковыми номерами, или рангами, на основании которых они затем разыскиваются по членам вариационного ряда. Для того чтобы найти квартили и медиану для выборки, необходимо сначала составить вариационный ряд, т.е. упорядочить выборку в порядке возрастания. Затем надо найти ранги (порядковые номера в вариационном ряду) искомым квартилей, например, так, как показано в **Таблице 1**.

Если информация представлена не самими наблюдениями, а интервальной таблицей частот (т.е. указаны только границы интервалов и частоты попадания в каждый из них), то в этом случае значения Q_1 , m , Q_3 находят интерполяцией по ломаной накопленных частот (именуемой иногда в статистике «стрелкой»).

³ От «кварта», лат., – четверть.

Квартиль	Ранг квартиля	Значение квартиля
Q_1	$i_1 = \frac{1 + [(n + 1) / 2]}{2}$	$Q_1 = \begin{cases} x_{i_1}, & \text{если } i_1 - \text{целое} \\ \frac{x_{[i_1]} + x_{[i_1]+1}}{2}, & \text{если } i_1 - \text{полуцелое} \end{cases}$
$Q_2 = m$	$i_2 = \frac{n + 1}{2}$	$m = \begin{cases} x_{i_2}, & \text{если } i_2 - \text{целое} \\ \frac{x_{[i_2]} + x_{[i_2]+1}}{2}, & \text{если } i_2 - \text{полуцелое} \end{cases}$
Q_3	$i_3 = (n + 1) - i_1$	$Q_3 = \begin{cases} x_{i_3}, & \text{если } i_3 - \text{целое} \\ \frac{x_{[i_3]} + x_{[i_3]+1}}{2}, & \text{если } i_3 - \text{полуцелое} \end{cases}$

Таблица 1. Значения рангов и квартилей

Итак, диаграмма «ящик с усами» состоит из прямоугольника («ящик»), занимающего пространство от первого до третьего квартиля (т.е. от 25-го до 75-го процентиля), линии внутри этого прямоугольника, соответствующей уровню (значению) медианы, и двух отрезков горизонтальной прямой («усы»), начинающихся справа и слева от «ящика» и заканчивающихся минимальным и максимальным значениями набора данных, которые **еще не являются выбросами**.

В терминологии Тьюки нетипичные значения выборки (выбросы) подразделяются на умеренные (moderate outliers) и экстремальные (extreme outliers). Первые представляют собой значения, отстоящие от Q_1, Q_3 не ближе полутора и не дальше трех межквартильных размахов $IQR = Q_3 - Q_1$, т.е. в области между $Q_1 - 3 \cdot IQR$ (внешнее нижнее ограждение «ящика с усами») и $Q_1 - 1,5 \cdot IQR$ (внутреннее нижнее ограждение), либо между $Q_3 + 1,5 \cdot IQR$ (внутреннее верхнее ограждение) и $Q_3 + 3 \cdot IQR$ (внешнее верхнее ограждение). Экстремальные выбросы лежат вне внешних ограждений, т.е. либо ниже $Q_1 - 3 \cdot IQR$, либо выше $Q_3 + 3 \cdot IQR$. В SPSS анализ выборки на наличие выбросов производится автоматически, и они соответствующим образом отображаются на окончательной диаграмме, а именно экстремальные выбросы маркируются звездочками, а умерен-

ные выбросы – кружками. Диаграмма в общем случае выглядит так (на ней в данном случае выбросов нет):

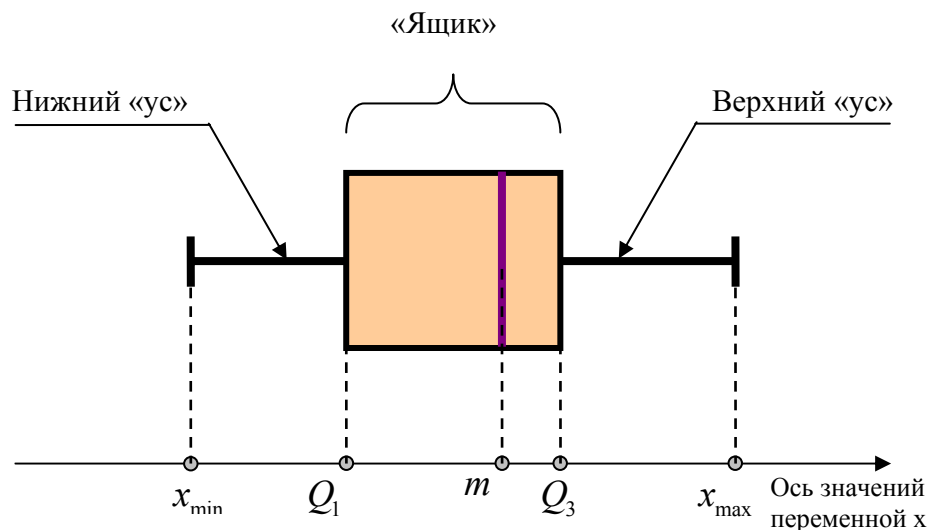


Рис. 2.1. Схема диаграммы «ящик с усами»

Являясь фактическим центральным значением выборки, медиана в этом смысле может считаться ее наиболее репрезентативным показателем «как целого». Легко заметить, что медиана, в отличие от выборочного среднего значительно менее чувствительна к выбросам. Действительно, если к набору данных добавить, скажем, одно значение, намного превосходящее все остальные, то вариационный ряд в конце пополнится этим добавленным значением без изменения порядка остальных членов, так что центральное значение в нем останется примерно «там же». Именно в связи с такой малой чувствительностью медианы к экстремальным значениям в ряде стран она, а не среднее значение, используется в качестве характеристики доходов в статистических отчетах.

Ясно, что выбросы в значительной степени затушевывают поведение остальных данных, которые «теряются» на фоне экстремальных значений, в результате чего наглядность графического представления результата снижается. Поэтому если обнаружено наличие выбросов (и оно не связано, например, с ошибками ввода данных), то общей рекомендацией является анализ выборки сначала с включенными, а затем с исключенными выбросами. Во всяком случае, появление выбросов заслуживает дополнительного внимания со стороны исследователя, поскольку может нести важную информацию о распределении изучаемой переменной.

В SPSS диаграмма «ящик с усами» может быть построена либо с помощью самостоятельной процедуры Boxplot меню Graphs (есть как стандартный, так и интерактивный графики), либо с помощью процедуры Explore меню Analyze.

2. Диаграмма Stem-and-leaf plot («ствол и лист»)

Диаграмма «ствол и лист» так же была предложена Дж. Тьюки, она представляет собой комбинацию гистограммы и табличного списка. Как на гистограмме, длина каждой строки соответствует количеству наблюдений, попадающих в определенный интервал. Но, сверх этого, на данной диаграмме выводится также численное значение для каждого наблюдения. Для этой цели численное значение разбивается на два компонента: «ствол» с ветвями, каждая из которых представляет собой первую цифру или группу цифр, и «лист» – последующие цифры. «Ствол» соответствует тем разрядам численного значения наблюдаемой переменной, которые не изменяются, а «листья» – разрядам, которые изменяются в пределах избранного интервала. Отметим, что шаг для «ствола» SPSS выбирает автоматически (он указан в предпоследней строке зоны диаграммы). В рассмотренном ранее примере «ствол» разбит разрядами сотен с шагом 100 руб. (первая ветвь «0» соответствует интервалу от 0 до 99 руб., вторая ветвь «1» – от 100 до 199 руб. и т.д.). В случае большого числа наблюдений, каждая ветвь «ствола» разбивается на две части: одна – для «листьев» с 0 по 4 и другая – для «листьев» с 5 по 9, так как иначе диаграмма потеряет наглядность ввиду большой частоты в каждой ветви. Первый столбец отражает частоты – количество наблюдений, попавших в каждый интервал, причем последняя строка в нем – это количество экстремальных значений (выбросов) в наборе данных (о том, что это такое и как они определяются уже говорилось ранее), второй столбец – это собственно диаграмма: первый вертикальный набор цифр – «ствол» с указанными на нем ветвями, а дальше вправо – «листья» на каждой ветви (наблюдаемые значения из указанного диапазона) Так, если взять ветвь, соответствующую «2», то мы попадаем в интервал от 200 до 299 руб. Видим, что частота попадания в этот интервал равна 3, и из этой ветви «растут» 3 «листа»: 2, 5, 8. Это значит, что там находятся не произвольные 3 значения из указанного интервала, а именно значения вида 22* руб., 25* руб. и 28* руб., т.е. «листьями» являются значения разряда десятков.

В пакете SPSS диаграмма «ствол и лист» может быть построена только с помощью процедуры Explore меню Analyze.

3. Диаграмма Парето

Диаграмма Парето представляет собой столбчатую диаграмму, столбцы которой соответствуют различным значениям некоторой категориальной переменной, высота каждого столбца отражает частоту появления соответствующего значения, и столбцы располагаются в **порядке убывания частот**. Кроме этого, диаграмма также включает в себя ломаную кумулятивного процента, позволяющую определить совокупную частоту, выраженную в процентах, двух, трех и т.д. наиболее часто встречающихся значений категориальной переменной.

Использование диаграммы Парето обоснованно и эффективно, например, в задачах, связанных с вопросами качества. Предположим, что мы изучаем группу некачественных компонентов чего-либо и классифицируем каждую единицу в соответствии с причиной дефекта. В этом случае диаграмма Парето показывает различные причины дефектов, упорядоченные от наиболее к наименее часто встречающимся (столбиковая диаграмма), и, кроме этого, процент дефектов, вызванный двумя, тремя, четырьмя и т.д. наиболее часто встречающимися причинами одновременно. Рассмотрим пример из практикума (Тема 1, задание 12).

В таблице приведены данные, полученные в литейной компании, производящей пластмассовые детали для компьютерных клавиатур, стиральных машин, автомобилей и телевизоров. В таблице указаны типы возможных дефектов компьютерных клавиатур и количество случаев их возникновения, зафиксированных на протяжении последних трех месяцев:

<i>Тип дефекта</i>	<i>Количество случаев</i>
<i>Черное пятно</i>	<i>413</i>
<i>Повреждение</i>	<i>1039</i>
<i>Впрыскивание</i>	<i>258</i>
<i>Отпечаток опоры</i>	<i>834</i>
<i>Царапина</i>	<i>442</i>
<i>Брызги</i>	<i>275</i>
<i>Серебряная полоска</i>	<i>413</i>
<i>Отпечаток формы</i>	<i>371</i>
<i>След пульверизатора</i>	<i>292</i>
<i>Деформация</i>	<i>1987</i>
<i>Всего</i>	<i>6324</i>

Построив диаграмму Парето для категориальной переменной «дефект», определить основные и второстепенные типы дефектов клавиатур и их про-

центный вклад в общее их количество, а также выяснить, какой процент брака соответствует трем наиболее распространенным типам дефектов.

Создание файла данных *Диаграмма_Парето_1_12.sav*. Файл данных организован следующим образом (см. рис. 2.2 и 2.3): имеем две переменные, первая из которых – *Тип_дефекта* – имеет тип Numeric, принимает значения от 1 до 10, и каждое из них имеет свое значение метки – название дефекта (колонка Values на вкладке Variable View редактора данных).

Вторая переменная – *Количество_дефектов* – имеет тип Numeric и отражает частоту, с которой встречается данный тип дефекта.

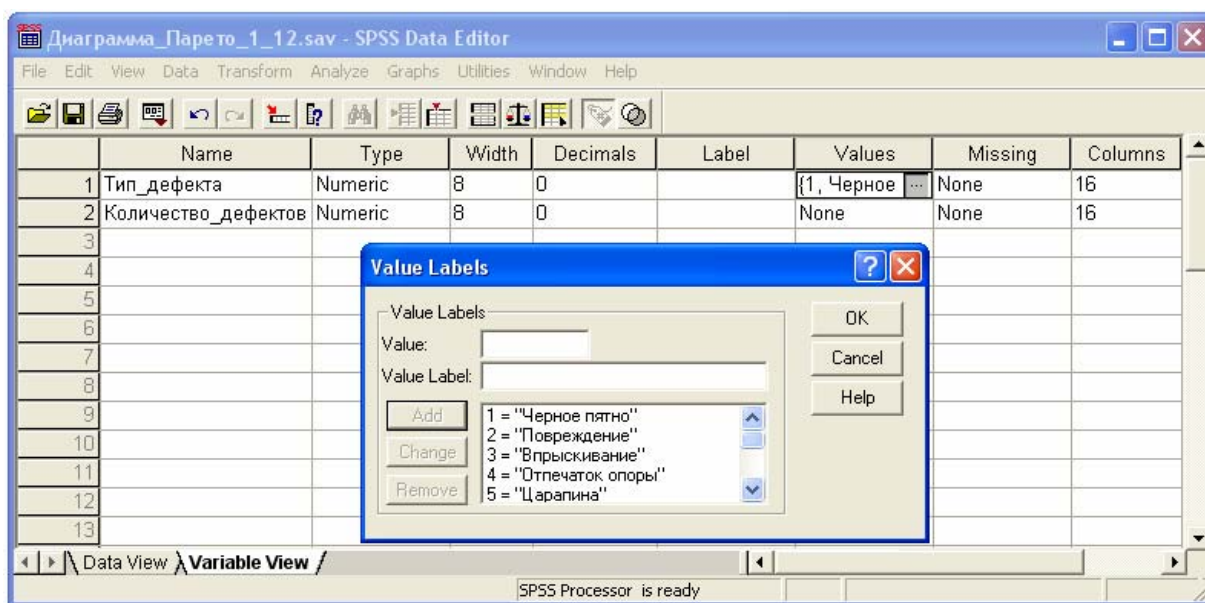


Рис. 2.2. Окно редактора данных для файла *Диаграмма_Парето_1_12.sav*, вкладка *Variable View*

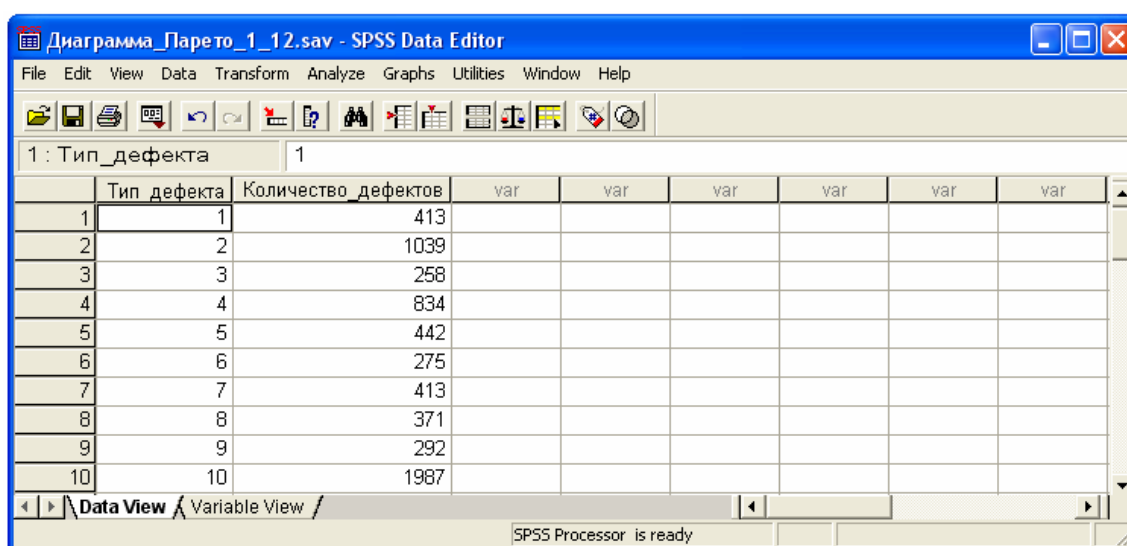


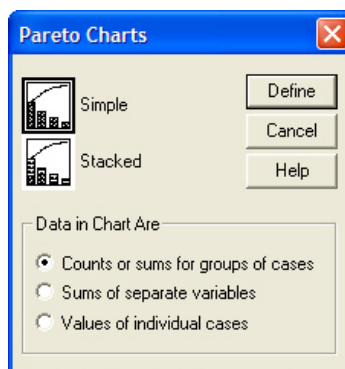
Рис. 2.3. Окно редактора данных для файла *Диаграмма_Парето_1_12.sav*, вкладка *Data View*

Для того чтобы в столбце переменной *Тип_дефекта* появились сами значения, необходимо в меню View поставить галочку на элементе управления Value Labels (*Значения меток*) либо кликнуть на соответствующую пиктограмму на панели управления, после чего в окне редактора данных увидим:

	Тип дефекта	Количество дефектов	var	var	var	var	var	var
1	Черное пятно	413						
2	Повреждение	1039						
3	Впрыскивание	258						
4	Отпечаток опоры	834						
5	Царапина	442						
6	Брызги	275						
7	Серебряная полоска	413						
8	Отпечаток формы	371						
9	След пульверизатора	292						
10	Деформация	1987						

*Рис. 2.4. Окно редактора данных для файла **Диаграмма_Парето_1_12.sav**, вкладка **Data View***

Построение диаграммы Парето. Выберем в меню Graphs (*Графики*) пункт Pareto... (*Парето...*). Откроется соответствующее диалоговое окно:



*Рис. 2.5. Диалоговое окно **Pareto Charts***

SPSS предоставляет возможность построить простую или состыкованную диаграмму Парето, причем и здесь существует три варианта представления данных. В диалоговом окне Pareto Charts (*Диаграммы Парето*) необходимо кликнуть на пиктограмму Simple (*Простая*) и оставить радиокнопку Counts or sums for groups of cases (*Частоты или суммы категорий для групп наблюдений*) в исходном состоянии. После клика на кнопку Define (*Определить*) откроется следующее диалоговое окно:

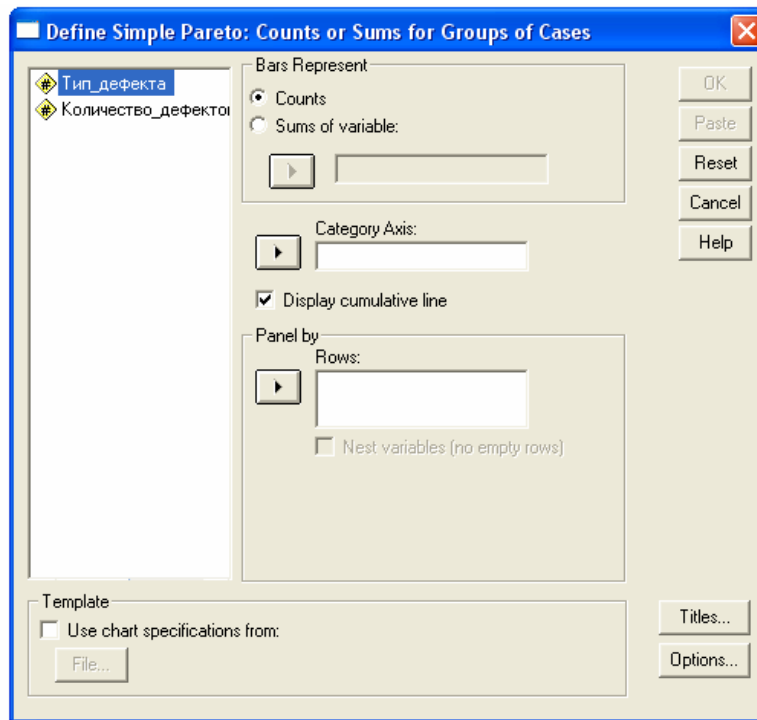


Рис. 2.6. Диалоговое окно Define Simple Pareto: Counts or Sums for Groups of Cases

Перенесем переменную **Тип_дефекта** в поле Category Axis: (*Ось категорий:*) диалогового окна. В блоке Bars Represent (*Столбцы отражают*) включим радиокнопку Sums of variable: (*Суммы переменной:*) и перенесем переменную **Количество_дефектов** в ставшее активным поле. Элемент управления Display cumulative line (*Отобразить кривую кумулятивного процента*) оставим в исходном состоянии. С помощью кнопки Titles... (*Заголовки...*) введем подходящий заголовок. После клика на кнопку ОК в окне вывода появится диаграмма Парето (см. рис. 2.7 ниже). Сохраним файл как **Диаграмма_Парето_1_12.spo**.

Как видим, из-за отображения кумулятивной (совокупной) кривой некоторые столбцы опустились довольно низко, так что различия в их величине стали незначительны. Тем не менее, каждый столбец снабжен значением соответствующей частоты, что делает диаграмму более наглядной. Однако иногда бывает удобнее запретить отображение кумулятивной кривой. Это приведет к изменению масштаба диаграммы и сделает ее более наглядной. Для того чтобы изменить масштаб высоты столбцов, следует дважды кликнуть в области диаграммы и в открывшемся окне редактора диаграмм Chart Editor двойным кликом выделить ось, отражающую частоты. В открывшемся диалоговом окне Properties (*Свойства*) (см. рис. 2.8 ниже) необходимо открыть вкладку Scale (*Шкала*).

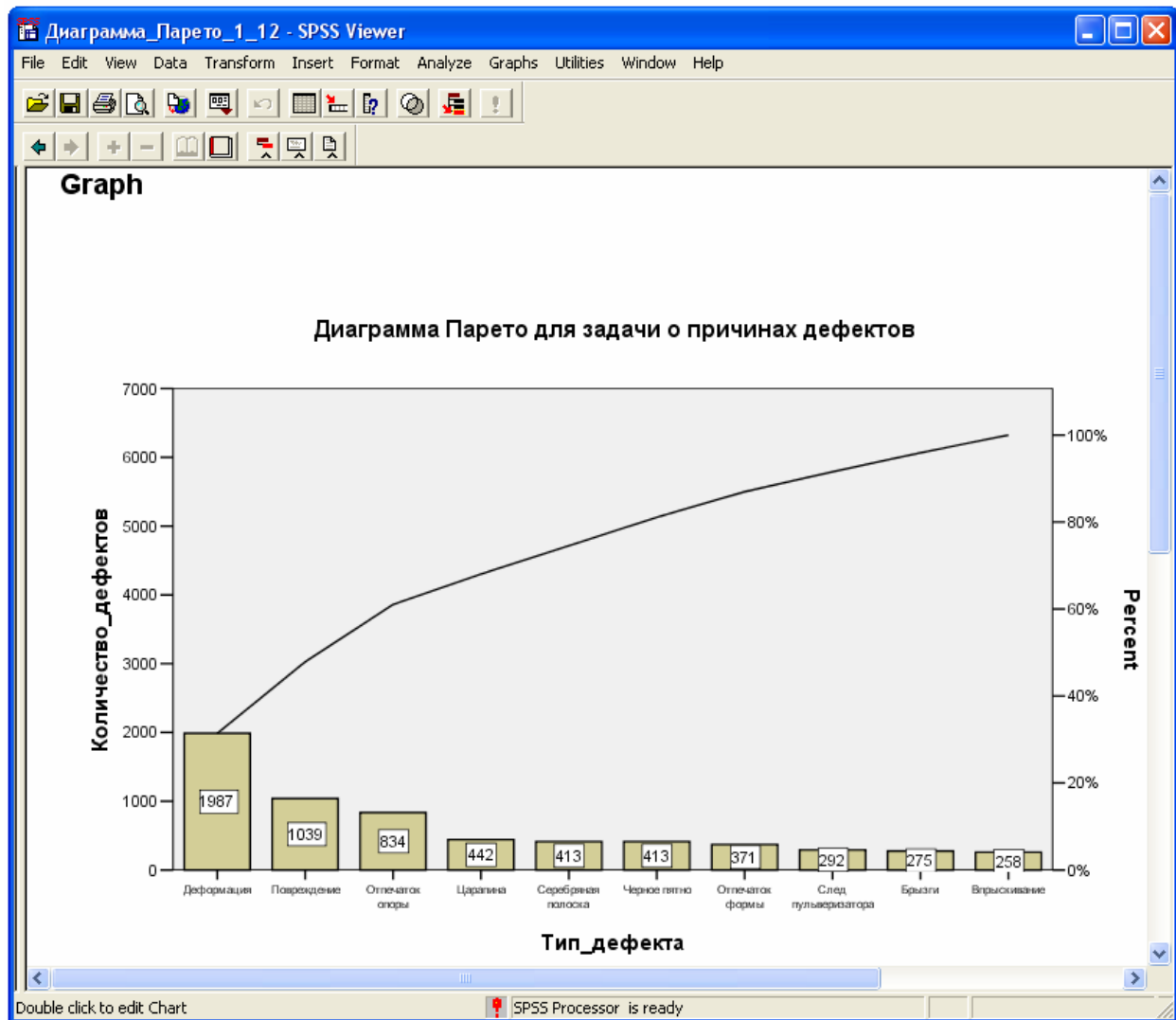


Рис. 2.7. Окно вывода для файла *Диаграмма Парето_1_12.spo*

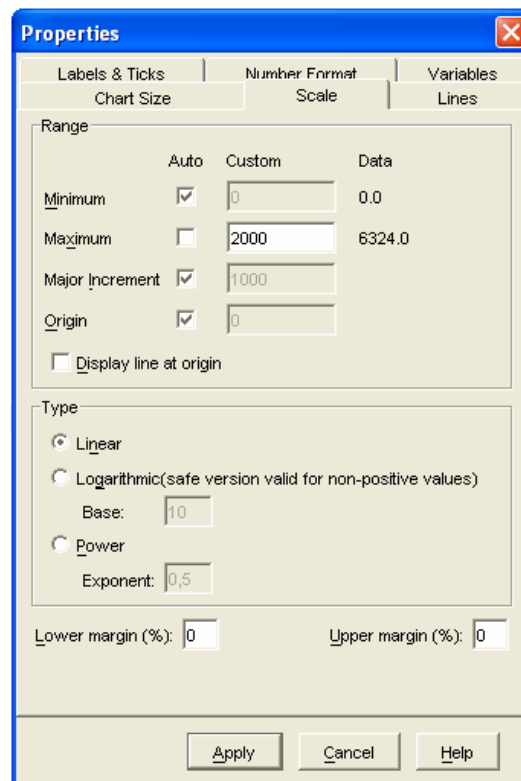


Рис. 2.8. Диалоговое окно *Properties*

На вкладке Scale (Шкала) (см. рис. 2.8) снимем галочку с элемента управления Auto (Автоматически) для поля Maximum, после чего введем в это поле значение наибольшей частоты (по умолчанию там стояла сумма всех частот). После этого диаграмма будет выглядеть так, как показано на рис. 2.9:

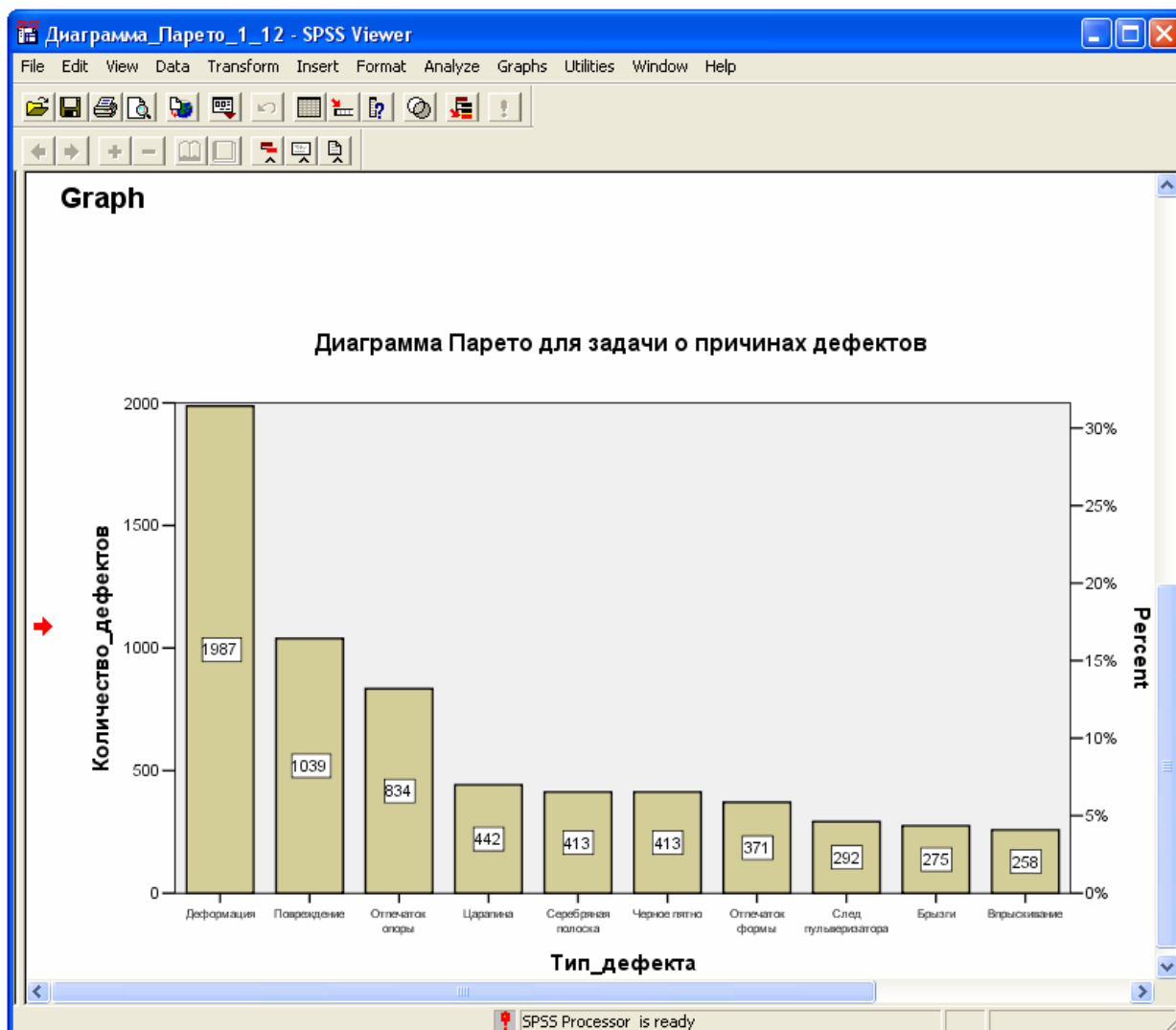


Рис. 2.9. Окно вывода для файла *Диаграмма Парето_1_12.spo*

Ответим теперь на вопросы задания. Нетрудно видеть (см. рис. 2.7), что первые три столбца диаграммы Парето существенно выше, чем все остальные, поэтому к основным причинам дефектов можно отнести деформацию, повреждение и отпечаток опоры. Остальные типы дефектов можно считать второстепенными. Процентный вклад основных типов дефектов таков: деформация составляет

$\frac{1987}{6324} \cdot 100\% = 31,42\%$ от общего числа дефектов, повреждение составляет

$\frac{1039}{6324} \cdot 100\% = 16,43\%$ и отпечаток опоры – $\frac{834}{6324} \cdot 100\% = 13,19\%$.

Правая вертикальная ось диаграммы Парето отражает кумулятивный (совокупный) процент дефектов, поэтому можем заключить, что деформация и повреждение (как два наиболее распространенных типа дефектов) вместе дают $31,42\% + 16,43\% = 47,85\%$ всех случаев брака; деформация, повреждение и отпечаток опоры (три наиболее распространенных типа дефектов) дают вместе $31,42\% + 16,43\% + 13,19\% = 61,04\%$ всех случаев брака, что составляет более половины всех случаев.

4. Примеры решения задач по визуализации данных

Реализация задачи с интервальной переменной

Пример из практикума (Тема 1, задание 1) демонстрирует возможности обработки в SPSS однородных данных (задача с интервальной переменной), когда есть ряд количественных наблюдений одного и того же признака.

Маркетинговой компании «Фриц энд Коль» заказали провести исследование распространения ряда журналов и газет на территории Великобритании. Нижеприведенные данные отражают количество читателей некой общенациональной газеты за период 50 дней (цифры приведены в 10 тыс. читателей).

121	102	132	142	139	114	136	142	156	145
135	140	148	117	125	134	120	137	107	134
110	150	94	135	144	111	145	128	133	146
137	127	146	154	136	105	138	153	143	124
123	145	114	130	125	149	128	133	118	136

1) Составьте таблицу частот на основании этих данных, 2) на основании таблицы частот постройте гистограмму, 3) изложите свою точку зрения на использование других видов графиков (например, линейных) для отображения такого рода данных.

Создание файла данных *Графический_отчет_1_1.sav*. В файле данных изначально присутствует одна переменная *Количество_читателей* типа Numeric, которая отражает количество читателей за каждый день наблюдения.

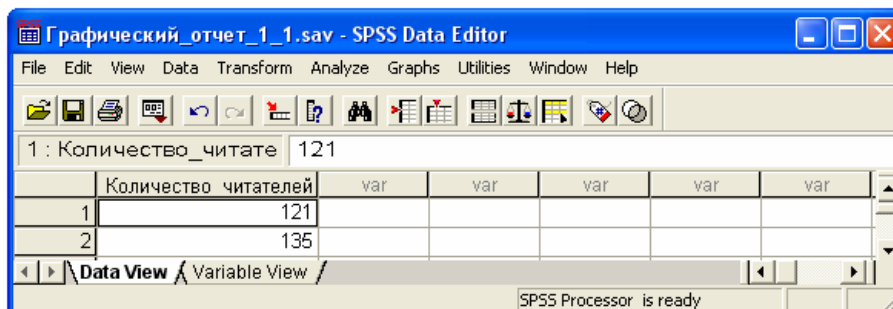


Рис. 2.10. Окно редактора данных для файла *Графический_отчет_1_1.sav*

Создание интервальной таблицы частот. Для данных такого типа наиболее удобным способом визуализации является гистограмма абсолютных частот, хотя иногда удобно использование и других типов графиков, например, линейного. Для того чтобы построить гистограмму частот, необходимо прежде всего сгруппировать данные, а именно построить интервальную таблицу частот. При построении гистограммы SPSS разбивает данные на интервалы автоматически и не выводит интервальную таблицу частот, а только соответствующую ей гистограмму. Однако сама таблица частот зачастую также бывает полезна. Для того чтобы ее получить, необходимо вызвать процедуру Visual Bander (*Визуальная категоризация*) меню Transform (*Преобразовать*). Эта процедура помогает подготовить к анализу количественные данные, например, доход или возраст или, как в нашем примере, наблюдаемое в течение 50 дней количество читателей. Она позволяет быстро в интерактивном режиме ознакомиться с имеющимися данными. В результате предварительного сканирования данных автоматически строится гистограмма, на которой можно задать границы интервалов. При помощи данной процедуры можно также быстро создать метки значений для заданных интервалов.

Процедура Visual Bander по переменной **Количество_читателей** создает значения для новой категориальной переменной, имя которой необходимо задать в поле Banded Variable (*Категориальная переменная*) (см. рис. 2.11 ниже). Как видим, в поле Current Variable (*Текущая переменная*) помещена переменная **Количество_читателей**, в поле Banded Variable – переменная **Шкала** (с меткой «Количество_читателей (Banded – категоризовано)» по умолчанию).

В столбце Value (*Значение*) вводятся значения границ интервалов. Так, в нашем примере количество читателей изменяется от 94 до 156, поэтому удобно разбить имеющийся диапазон на равные интервалы длиной 10 с первой правой границей 100. При этом получаем 7 интервалов (их правые границы вводятся в столбце Value, причем последнее значение HIGH стоит по умолчанию). После того как интервалы определены, необходимо кликнуть на кнопку Make Labels (*Создать метки*), после чего в столбце Label (*Метка*) появятся метки для интервалов – то, что мы увидим потом в интервальной таблице окна вывода.

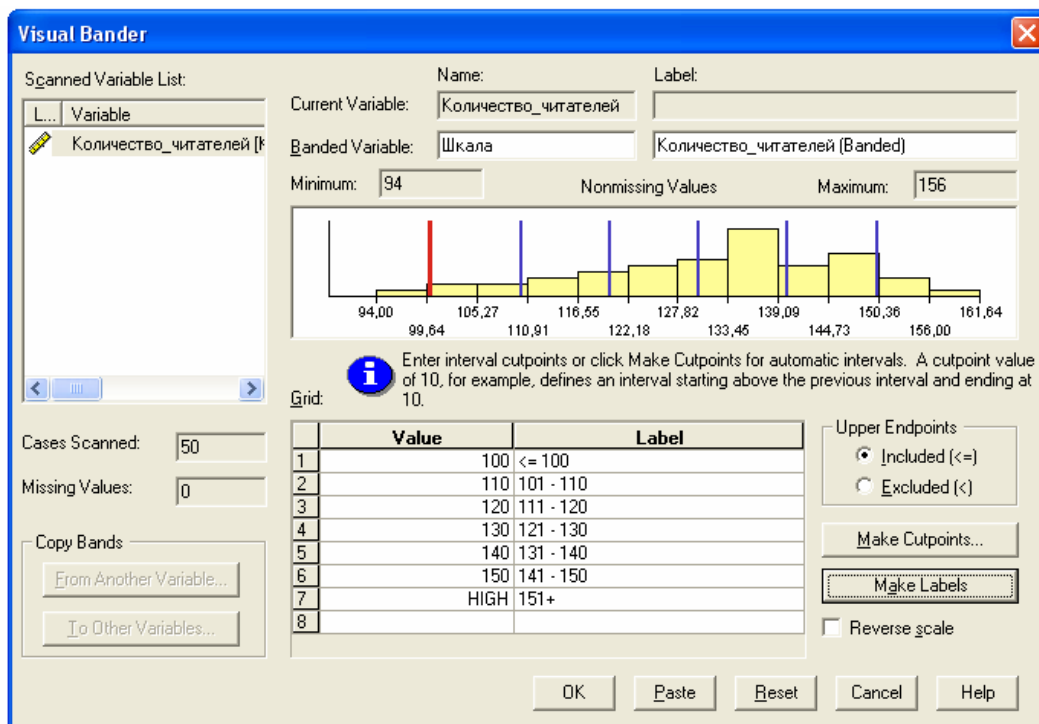


Рис. 2.11. Диалоговое окно процедуры Visual Bander

Интервалы можно задать и по-другому. Для этого нужно кликнуть на кнопку Make Cutpoints... (Создать точки разбиения...). В появившемся диалоговом окне нужно выбрать один из способов разбиения точками, например, такой, как показан на рис. 2.12:

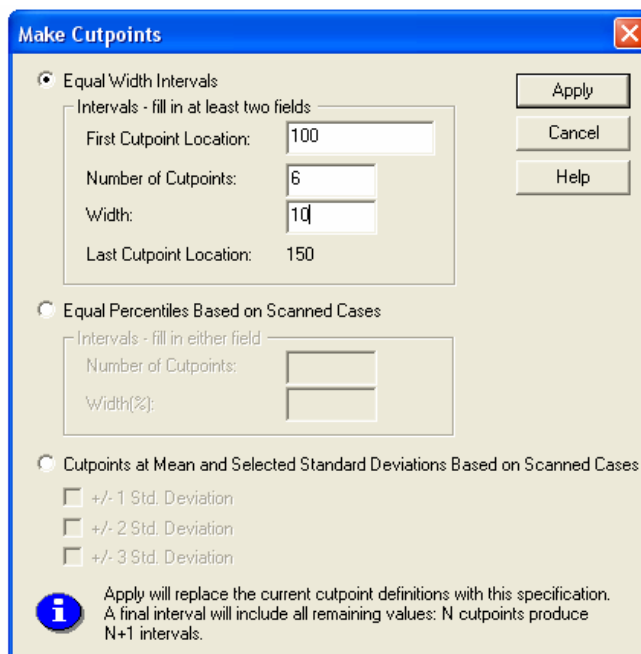


Рис. 2.12. Диалоговое окно Make Cutpoints

После клика на кнопку Ok в основном диалоговом окне процедуры Visual Bander завершается ее работа. При этом в окне редактора данных автоматически появляется вторая переменная **Шкала**. Теперь для этой переменной можно по-

строить таблицу частот уже известным нам способом (Analyze → Descriptive Statistics → Frequencies). Сохраним файл как *Графический_отчет_1_1.spo*. Отметим, что в окне вывода над таблицей частот будет отображаться не имя переменной *Шкала*, а соответствующая ей метка.

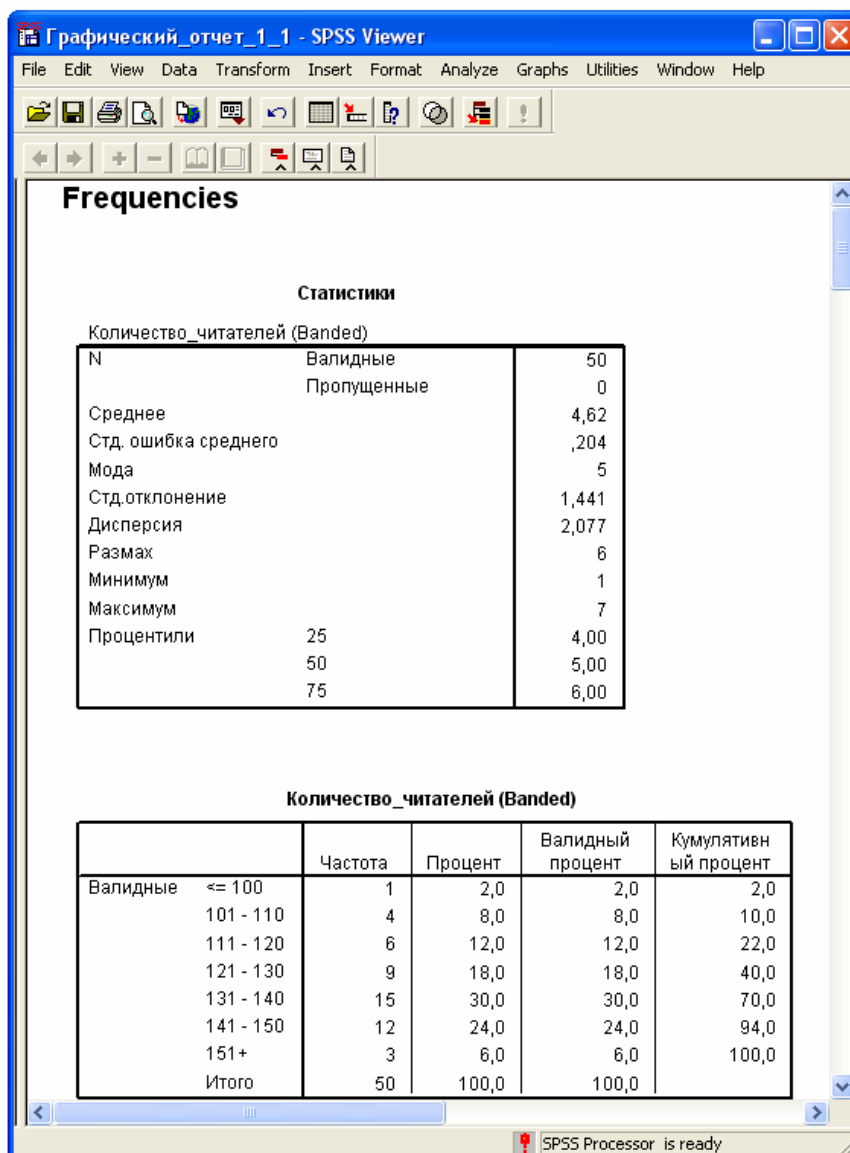


Рис. 2.13. Окно вывода для файла Графический_отчет_1_1.sav

Последующий этап построения гистограммы теперь может быть подкреплен соответствующей группировкой данных, а пока ограничимся группировкой данных по умолчанию.

Построение гистограммы абсолютных частот. Гистограмму можно построить двумя способами – при помощи стандартных или интерактивных графиков. Сначала построим гистограмму с помощью стандартного графика. Для этого выберем в меню Graphs (*Графики*) пункт Histogram...(Гистограмма...).

В открывшемся диалоговом окне Histogram (см. рис. 2.14) в поле Variable (Переменная) поместим переменную *Количество_читателей* и кликнем на кнопку Ok.

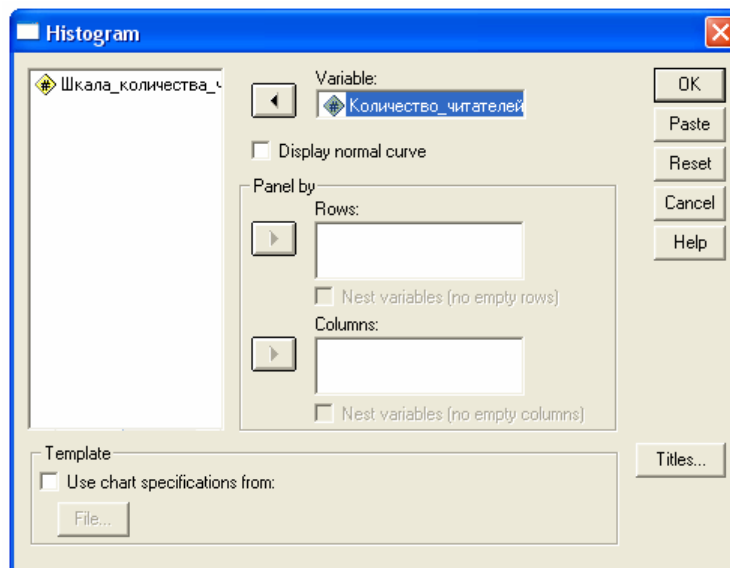


Рис. 2.14. Диалоговое окно Histogram

Отметим, что при создании гистограммы с помощью **стандартного** графика SPSS автоматически выбирает число интервалов группировки, и пользователь **не может его изменить**. В данном случае 14 интервалов шириной 5. При построении этой гистограммы, очевидно, была использована другая интервальная таблица частот, чем та, которую мы построили с помощью процедуры Visual Bander. Результат процедуры Histogram представлен на рис. 2.15 ниже: каждый столбец гистограммы снабжен абсолютным значением частоты попадания в данный диапазон (она отражает количество дней из 50, в которые наблюдалось количество читателей из этого диапазона). Напомним, что эти значения можно вывести, активировав элемент Show Data Labels (*Показывать метки данных*) в окне редактора диаграмм Chart Editor. Как видно, наибольшую частоту имеет столбец 10 (9 дней из 50 наблюдалось количество читателей в диапазоне от 1,35 млн. до 1,4 млн. читателей). Напротив, столбец, соответствующий диапазону от 950 тыс. до 1 млн. читателей, вообще не содержит данных.

Теперь построим гистограмму с использованием интерактивного графика. По сравнению со стандартной интерактивная гистограмма дает дополнительные возможности, в частности позволяет варьировать число интервалов, что удобно

при решении конкретных задач. Как увидим далее, внешний вид гистограммы будет несколько отличаться от предыдущего.

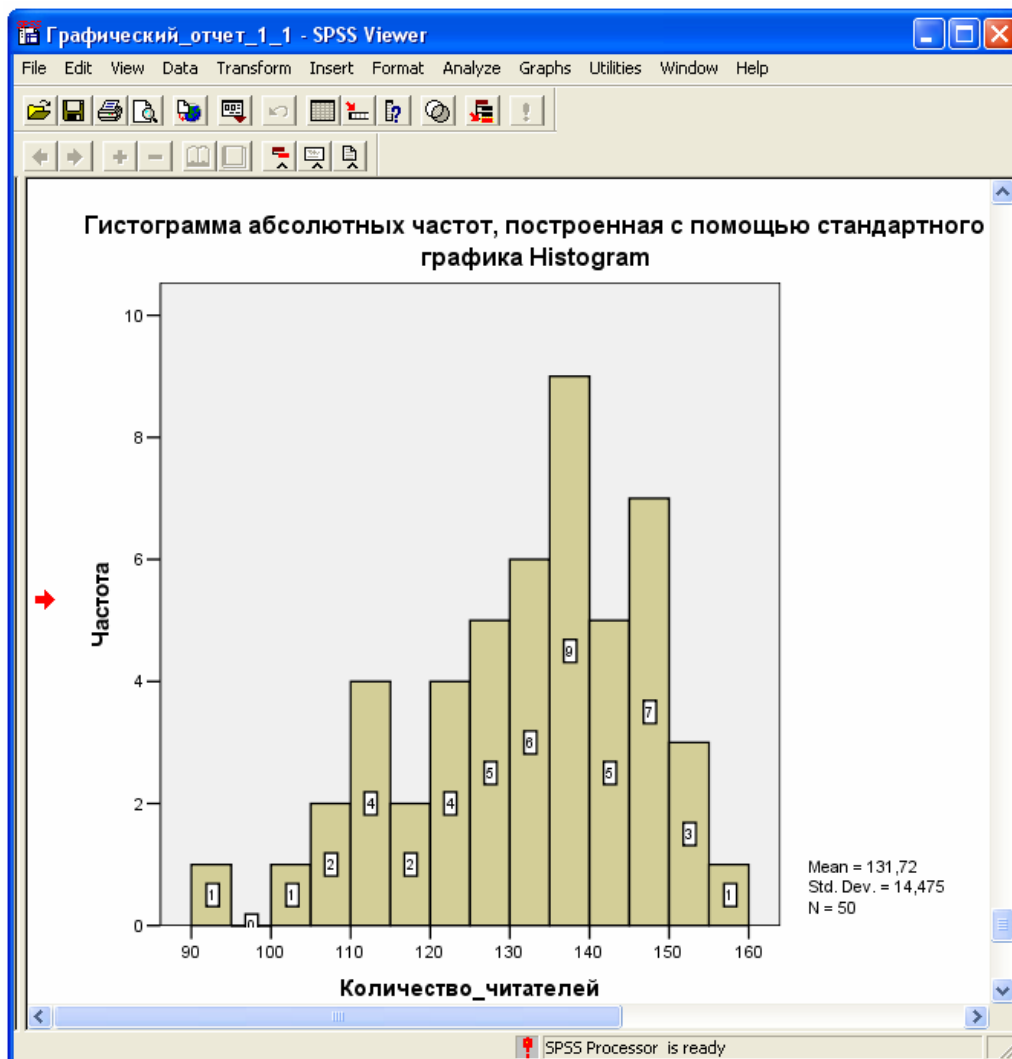


Рис. 2.15. Окно вывода для файла *Графический_отчет_1_1.sav*

Выберем в меню Graphs (*Графики*) пункт Interactive (*Интерактивный*) и в нем – Histogram... (*Гистограмма...*). В появившемся диалоговом окне Create Histogram (*Создать гистограмму*) (см. рис. 2.16 ниже) поместим переменную *Количество_читателей* в поле по оси абсцисс, а встроенную переменную *\$count* с меткой Count (*Частота*) оставим в поле по оси ординат.

На вкладке Histogram снимем галочку с элемента управления Set interval size automatically (*Установить размер интервала автоматически*), включим радиокнопку Number of intervals: (*Количество интервалов:*) и в ставшее активным поле ввода введем количество интервалов, равное, например, 7.

После клика на кнопку Ok появится окно вывода с соответствующим графиком, изображенным на рис. 2.17 ниже.

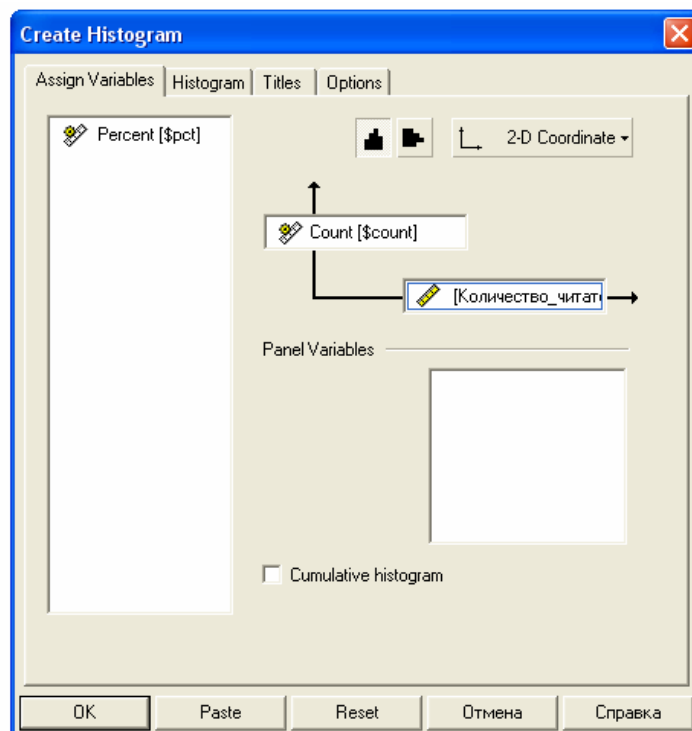


Рис. 2.16. Диалоговое окно *Create Histogram*

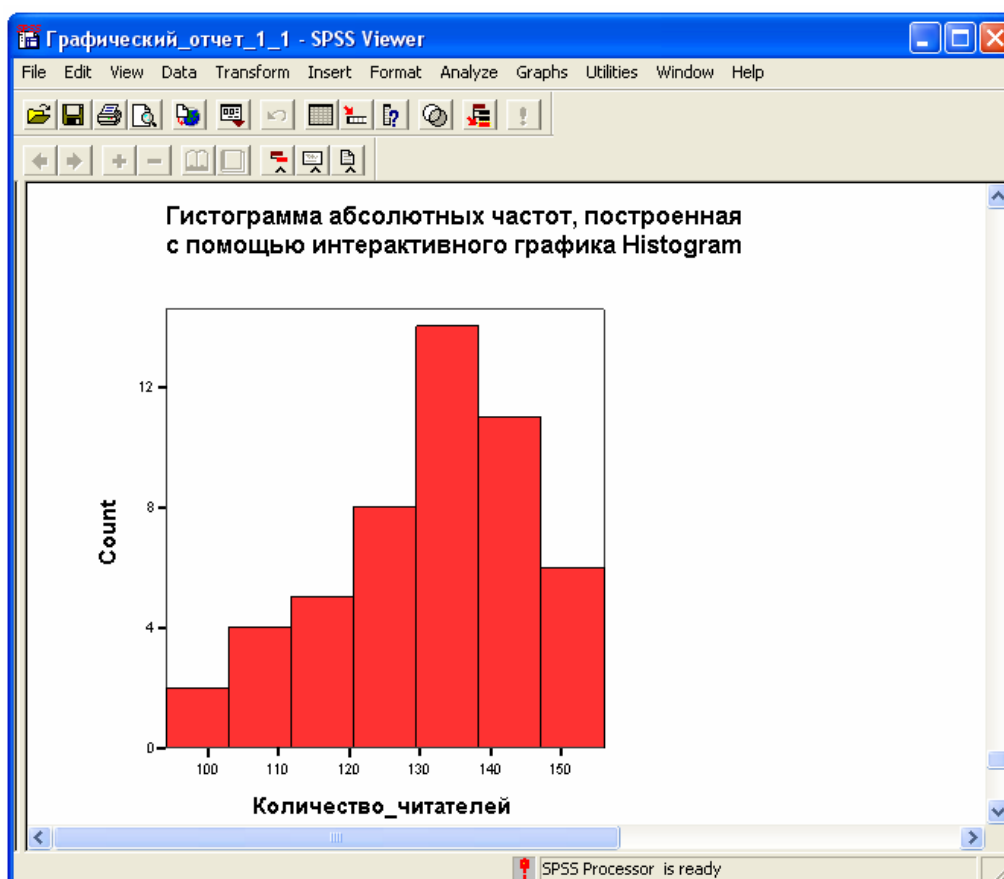


Рис. 2.17. Окно вывода для файла *Графический отчет_1_1.sav*

Построение линейных графиков. Для сравнения построим линейный график для того же набора данных. Для этого выберем в меню Graphs (*Графики*) пункт Line... (*Линейные...*). В открывшемся диалоговом окне Line Charts (*Линейные диаграммы*) укажем тип диаграммы Simple (*Простая*), остальные установки

оставим неизменными и нажмем на кнопку Define (*Определить*). В открывшемся диалоговом окне Define Simple Line: Summaries for Groups of Cases (см. далее рис. 2.18) поместим переменную **Количество_читателей** в поле Category Axis: (*Ось категорий:*), в блоке Line Represents (*Линия отражает*) оставим радиокнопку N of cases (*Количество наблюдений*) включенной и нажмем на кнопку Ok. Результат – окно вывода – представлен на рис. 2.19 ниже.

Аналогичным образом строится и интерактивный линейный график (Graphs → Interactive → Line). Линейный график в каждой точке оси абсцисс отражает число дней, в которые наблюдалось данное количество читателей.

С точки зрения репрезентативности (наглядности), гистограмма частот является наиболее удобным способом представления такого рода данных, когда мы имеем дело с большим количеством значений **одного** наблюдаемого признака. Линейный же график отражает частоту появления каждого конкретного значения наблюдаемого признака, и вполне реалистична ситуация, когда частота для каждого наблюдаемого значения окажется равной 1. В таком случае линейный график будет представлять собой горизонтальную прямую, что, по понятным причинам, не позволит сформировать представление о структуре данных. Поэтому использование в подобных задачах линейных графиков зачастую неэффективно.

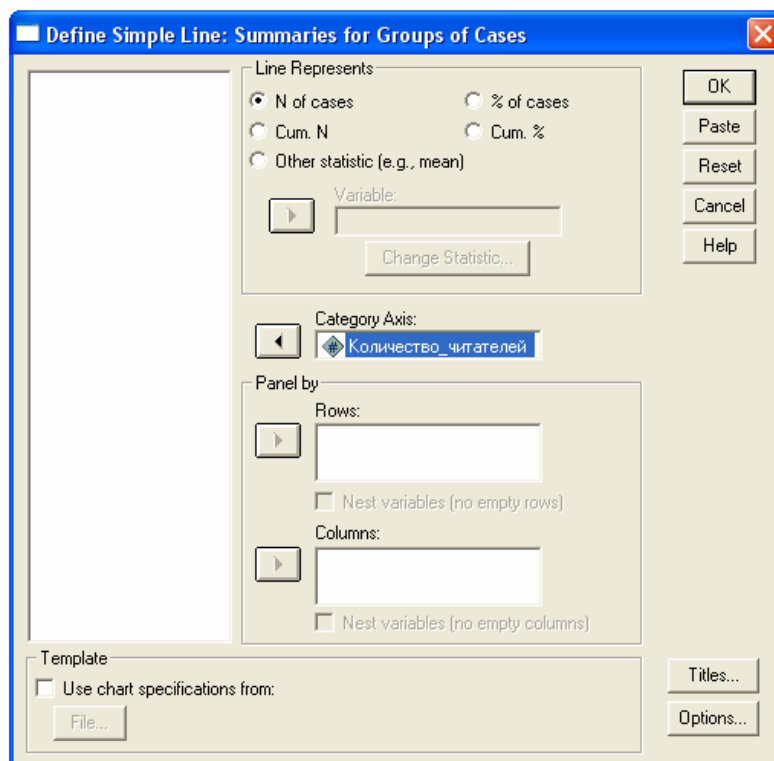


Рис. 2.18. Диалоговое окно Define Simple Line: Summaries for Groups of Cases

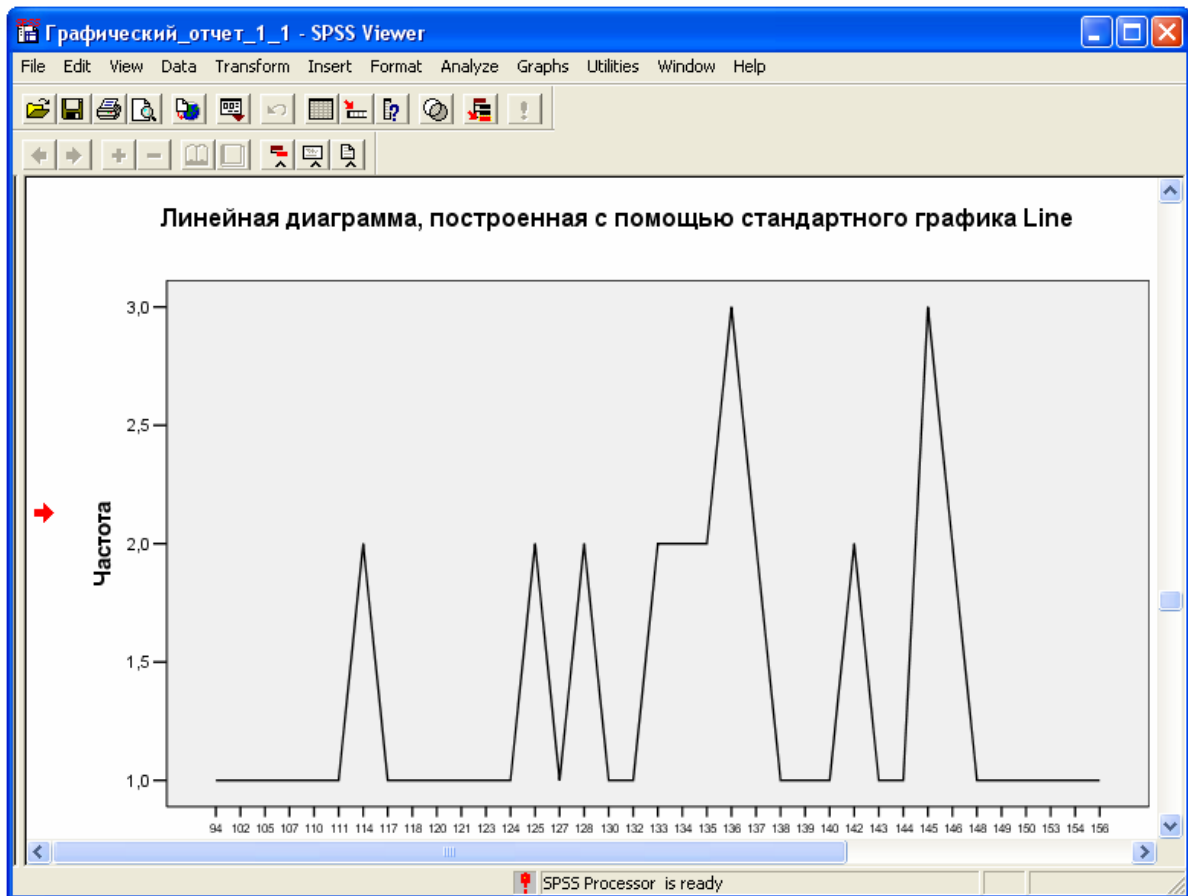


Рис. 2.19. Окно вывода для файла *Графический_отчет_1_1.sav*

Реализация задачи с категориальной переменной

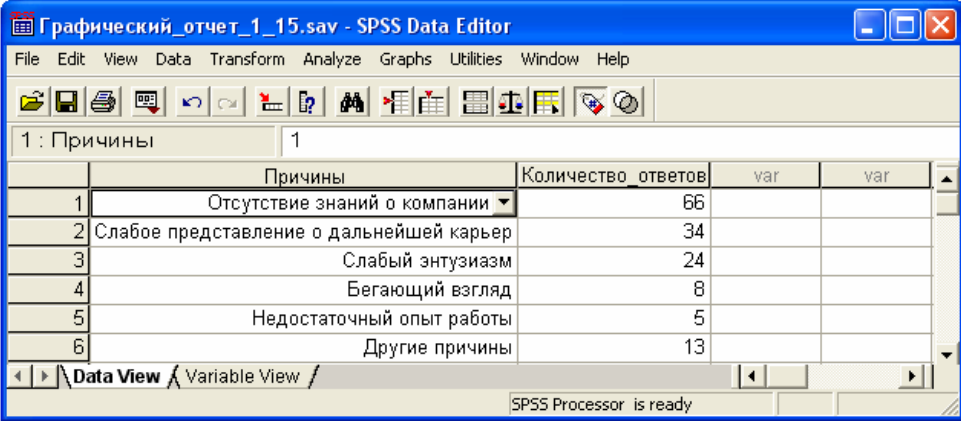
Проиллюстрируем теперь некоторые возможности графического анализа в SPSS для задачи с категориальной переменной на примере из практикума (Тема 1, задание 15).

В ходе опроса 150 менеджеров их попросили указать основные причины отказов в предоставлении работы ее соискателям в результате собеседования. Ответы респондентов приведены ниже (USA Today SNAPSHOTS, November, 19, 2001):

<i>Причина отказа</i>	<i>Количество ответов</i>
<i>Отсутствие знаний о компании</i>	<i>66</i>
<i>Слабое представление о дальнейшей карьере</i>	<i>34</i>
<i>Слабый энтузиазм</i>	<i>24</i>
<i>Бегающий взгляд</i>	<i>8</i>
<i>Недостаточный опыт работы</i>	<i>5</i>
<i>Другие причины</i>	<i>13</i>

Предоставьте иллюстрированный отчет анализа приведенных данных (в виде столбиковой, секторной диаграмм и диаграммы Парето). Сделайте соответствующие выводы. Если бы Вы были соискателем работы, каких ошибок при прохождении собеседования Вам следовало бы опасаться больше остальных?

Создание файла данных *Графический_отчет_1_15.sav*. Файл данных организован следующим образом: имеем две переменные, первая из которых – **Причины** – категориальная, имеет тип Numeric, принимает значения от 1 до 6, и каждое из них имеет свое значение метки – конкретную причину отказа. Вторая переменная – **Количество_ответов** – интервальная, имеет тип Numeric и отражает частоту, с которой встречается данный ответ респондента. Фрагмент окна файла данных приведен на рис. 2.20. Начнем создание графического отчета с построения столбиковой диаграммы. Это можно сделать двумя способами: с помощью стандартного или интерактивного графика.



	Причины	Количество_ответов	var	var
1	Отсутствие знаний о компании	66		
2	Слабое представление о дальнейшей карьере	34		
3	Слабый энтузиазм	24		
4	Бегаящий взгляд	8		
5	Недостаточный опыт работы	5		
6	Другие причины	13		

Рис. 2.20. Окно редактора данных для файла *Графический_отчет_1_15.sav*

Построение столбиковой диаграммы с помощью стандартного графика. Для этого выберем в меню **Graphs (Графики)** пункт **Bar... (Столбиковые...)** В открывшемся диалоговом окне **Bar Charts (Столбиковые диаграммы)** укажем тип диаграммы **Simple (Простая)**, а в блоке **Data in Chart Are (Данными для диаграммы являются)** оставим радиокнопку **Summaries for groups of cases (Обработка категорий для групп наблюдений)** включенной.

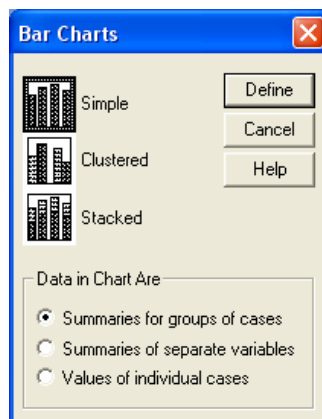


Рис. 2.21. Диалоговое окно **Bar Charts**

После клика на кнопку Define (*Определить*) откроется соответствующее диалоговое окно. В поле Category Axis: (*Ось категорий*) необходимо перенести категориальную переменную **Причины**, в блоке Bars Represent (*Столбцы отражают*) включить радиокнопку Other Statistics (e.g. mean) (*Другие статистики, например, среднее*) и в ставшее активным поле Variable (*Переменная*) перенести переменную **Количество_ответов**. После клика на кнопку Titles... (*Заголовки...*) необходимо ввести заголовок для диаграммы.

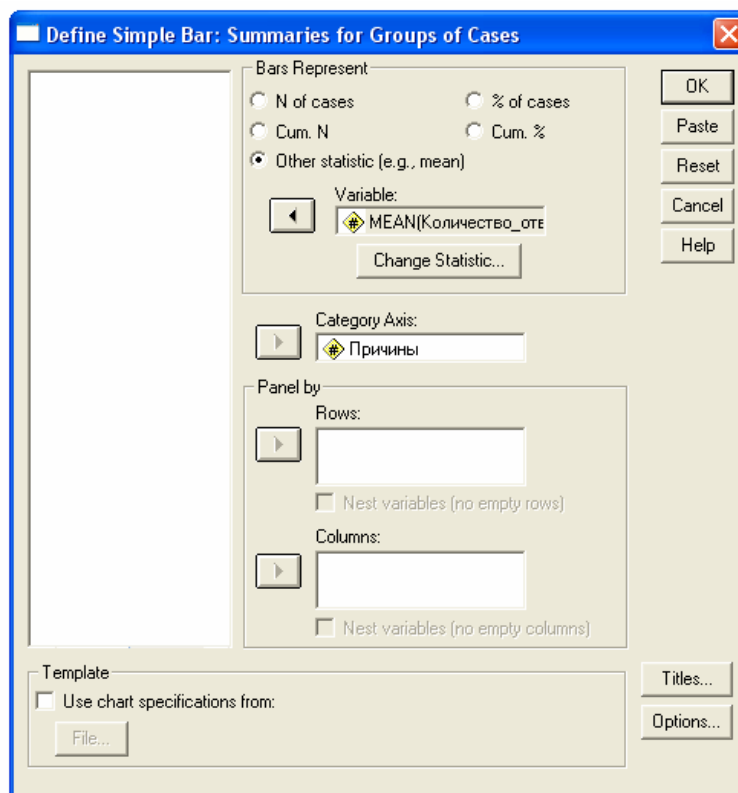


Рис. 2.22. Диалоговое окно Define Simple Bar: Summaries for Groups of Cases

После клика на кнопку Ok откроется окно вывода, которое сохраним как **Графический отчет_1_15.spo**. Ниже на рис. 2.23 показана диаграмма, в которой каждый столбец представлен двумя значениями данных – абсолютной частотой ответов и процентной долей от общего числа ответов. Для того чтобы вывести эти значения, необходимо дважды кликнуть на области диаграммы, после чего откроется окно редактора диаграмм Chart Editor. В нем необходимо выделить столбцы диаграммы и на панели инструментов кликнуть на пиктограмму Show Data Labels (*Показать метки данных*). В открывшемся диалоговом окне Properties (*Свойства*) на вкладке Data Value Labels (*Значения меток данных*) в поле Displayed (*Отображаемые*) переместить встроенную переменную Percent

(Процент), переменная **Количество_ответов** уже находится там по умолчанию. Выполнить необходимые настройки (изменить порядок переменных в поле Displayed, указать расположение метки в поле Label Position) и завершить начатое кликом на кнопку Apply (*Применить*). Затем закрыть окно Chart Editor.

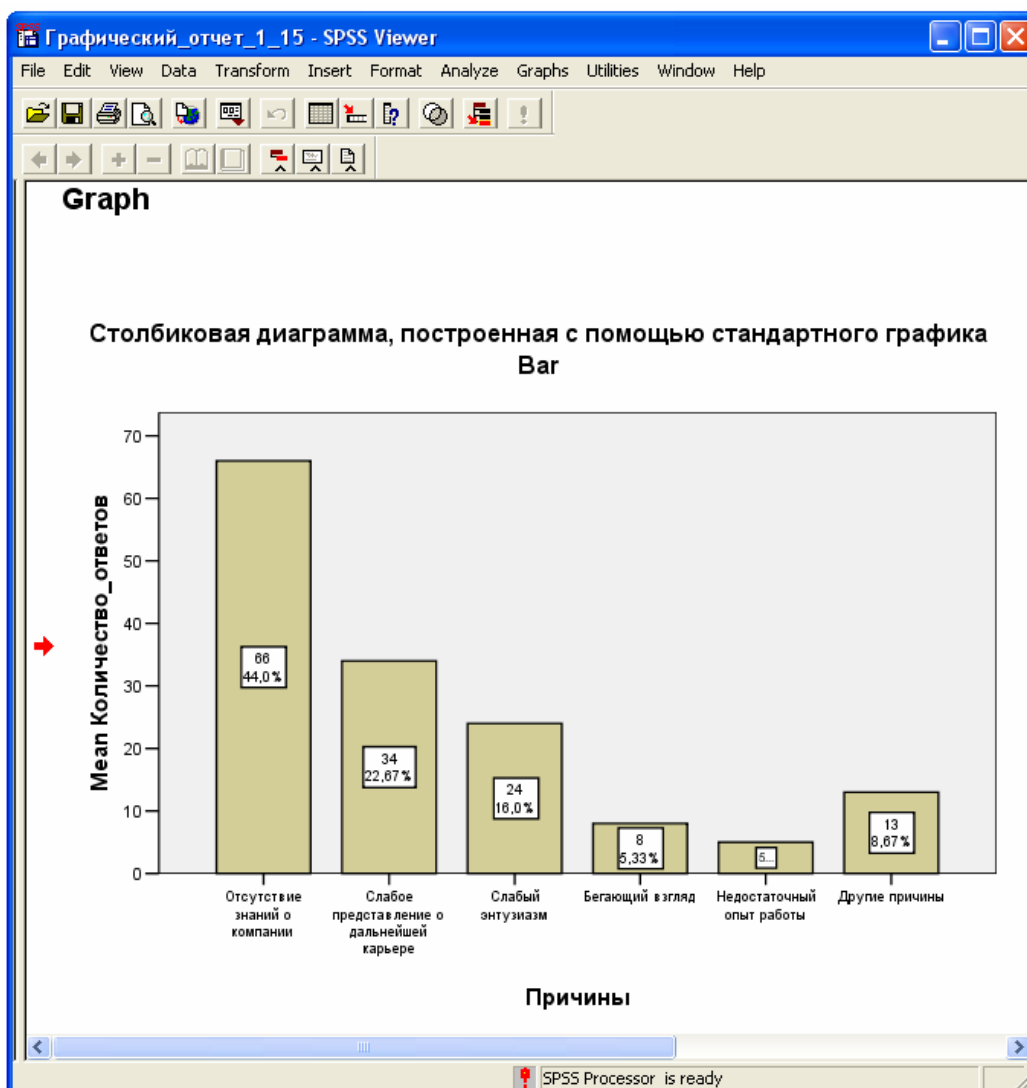


Рис. 2.23. Окно вывода для файла *Графический отчет_1_15.spo*

Построение столбиковой диаграммы с помощью интерактивного графика. Для этого выберем в меню Graphs (*Графики*) пункт Interactive (*Интерактивные*), а в нем – Bar... (*Столбиковые...*). В открывшемся диалоговом окне Create Bar Chart (*Создать столбиковую диаграмму*) на вкладке Assign Variables (*Назначение переменных*) в поля координатных осей переместим переменные **Причины** и **Количество_ответов**, а также поставим галочку на элементе управления Display Key (*Отобразить ключ*). Если выбор осей таков, как показано ниже на рис. 2.24, столбцы диаграммы будут расположены горизонтально.

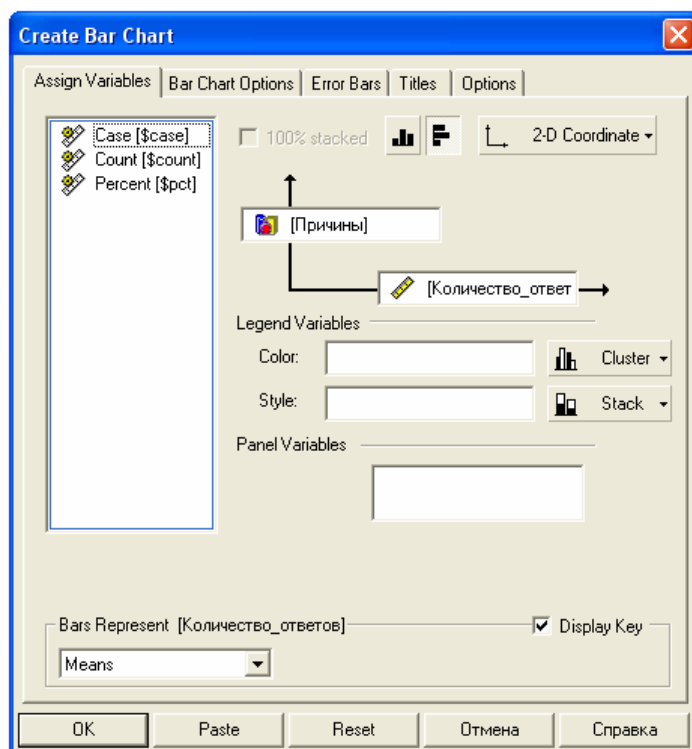
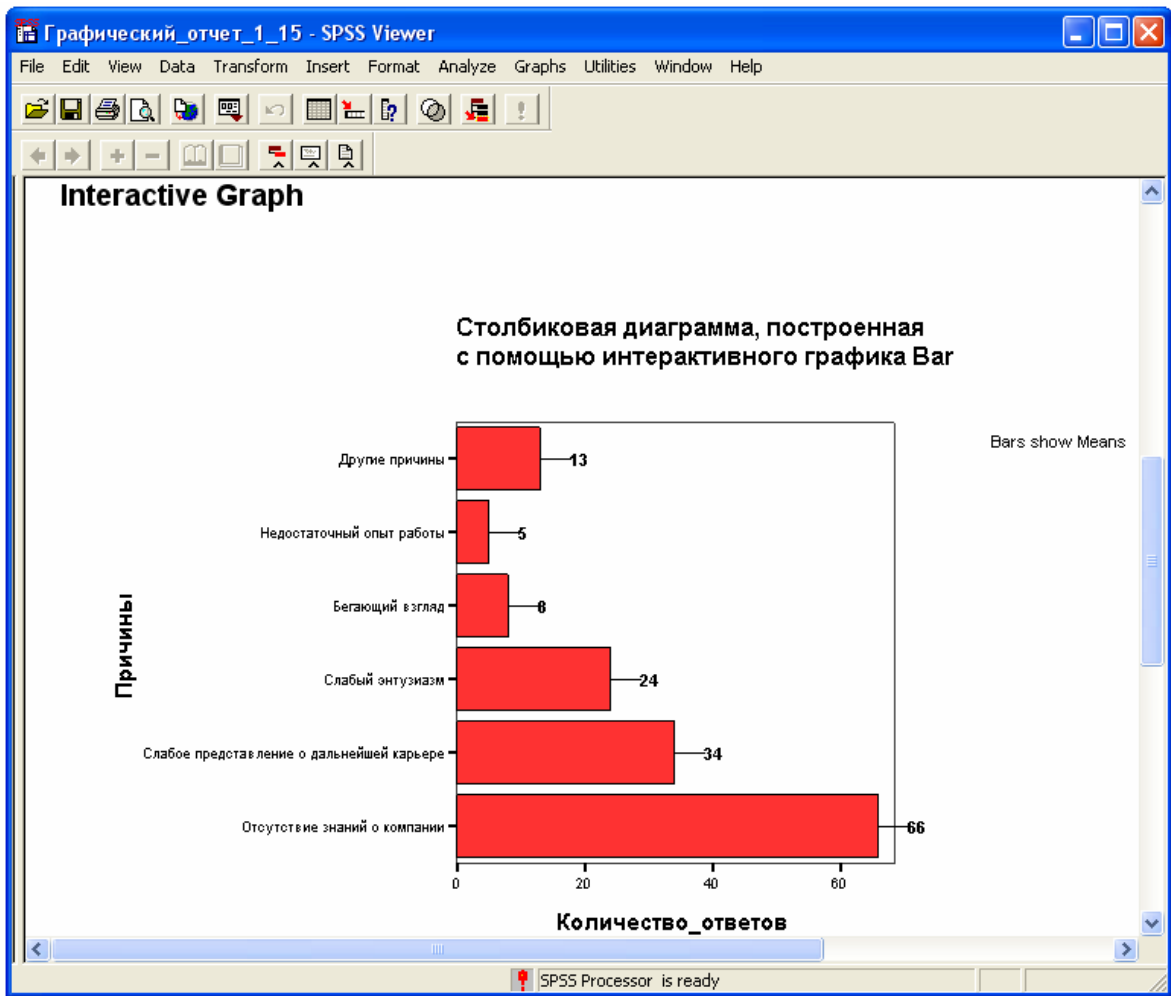


Рис. 2.24. Диалоговое окно *Create Bar Chart*

На вкладке *Titles* (*Заголовки*) введем заголовок диаграммы, а на вкладке *Bar Chart Options* (*Опции столбиковой диаграммы*) в поле *Bar Labels* (*Метки столбцов*) поставим галочку на элементе управления *Value* (*Значение*). После клика на кнопку *Ok* получим диаграмму, которую, например, можно отредактировать так, как показано ниже на рис. 2.25. Для этого необходимо дважды кликнуть на области диаграммы, после чего правым кликом на любой столбец вызвать контекстное меню, выбрать пункт *Select All Bars* (*Отобразить все столбцы*), затем выбрать в этом контекстном меню пункт *Label Connectors* (*Соединители меток*), а в выпавшем меню – пункт *Value* (*Значение*). После клика на пиктограмму *Display the Chart Manager* (*Отобразить менеджер диаграммы*) в диалоговом окне *Chart Manager* (*Менеджер диаграммы*) выбрать элемент *Bar* и кликнуть на кнопку *Edit* (*Редактировать*). На вкладке *Bar Options* диалогового окна *Bars* в поле *Bar Labels* выбрать значение *Outside End* (*Снаружи над*) для *Location:* (*Расположение:*) и кликнуть на кнопку *Ok*.

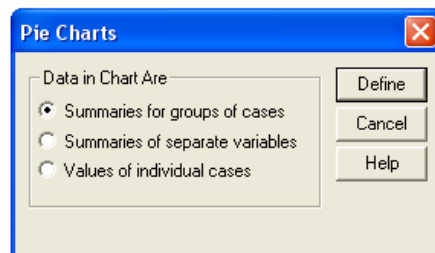
Как видим, умелое использование инструментальных средств редактирования интерактивных графиков позволяет добиваться лучшей иллюстративности диаграмм по сравнению со стандартными графиками.



*Рис. 2.25. Окно вывода для файла **Графический_отчет_1_15.spo***

Построение секторной диаграммы с помощью стандартного графика.

Выберем в меню Graphs (*Графики*) пункт Pie... (*Круговые...*). В открывшемся диалоговом окне Pie Charts (*Круговые диаграммы*) оставим радиокнопку Summaries for groups of cases (*Обработка категорий для групп наблюдений*) включенной. Затем кликнем на кнопку Define (*Определить*).



*Рис. 2.26. Диалоговое окно **Pie Charts***

В появившемся диалоговом окне Define Pie: Summaries for Groups of Cases перенесем в поле Define slices by: (*Создать сектора при помощи:*) переменную **Причины**. Кликнем на кнопку Options... (*Опции...*) и в открывшемся окне Options снимем галочку с элемента управления Display groups defined by missing

values (*Отобразить группы, образованные пропущенными значениями*). Затем после клика на кнопку Titles... (*Заголовки...*) введем заголовок.

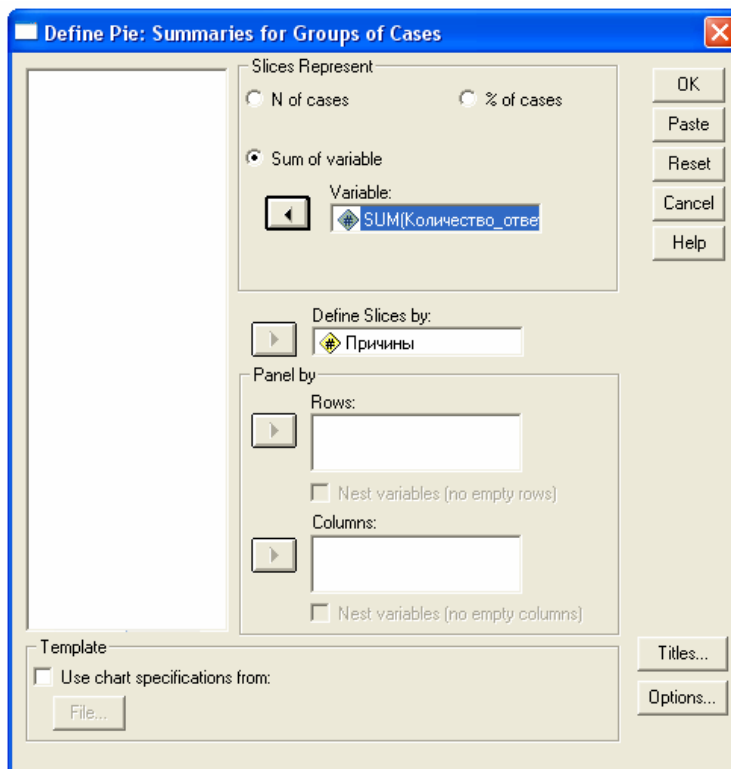


Рис. 2.27. Диалоговое окно *Define Pie: Summaries for Groups of Cases*

После клика на кнопку *Ок* в окне вывода появится диаграмма, которую, например, можно отредактировать так, как показано ниже на рис. 2.28. После двойного клика на область диаграммы откроется окно редактора диаграмм *Chart Editor*. В нем, аналогично рассмотренному ранее примеру редактирования столбиковой диаграммы, построенной с помощью стандартного графика, можно выполнить настройки для вывода абсолютных значений и/или процентов. Значение процента каждой из причин отказа наиболее наглядно отражает структуру данных, поскольку секторная диаграмма показывает распределение целого на доли. Справа от диаграммы приведена ее легенда – описание каждого сектора. Она выводится по умолчанию.

Для секторных диаграмм, построенных с помощью стандартного графика, как правило, отказываются от заливки секторов по умолчанию, если предполагается использовать черно-белый режим печати. В этом случае для заливки обращаются к набору образцов *Pattern* диалогового окна *Properties (Свойства)*, открывающегося в окне редактора диаграмм *Chart Editor*.

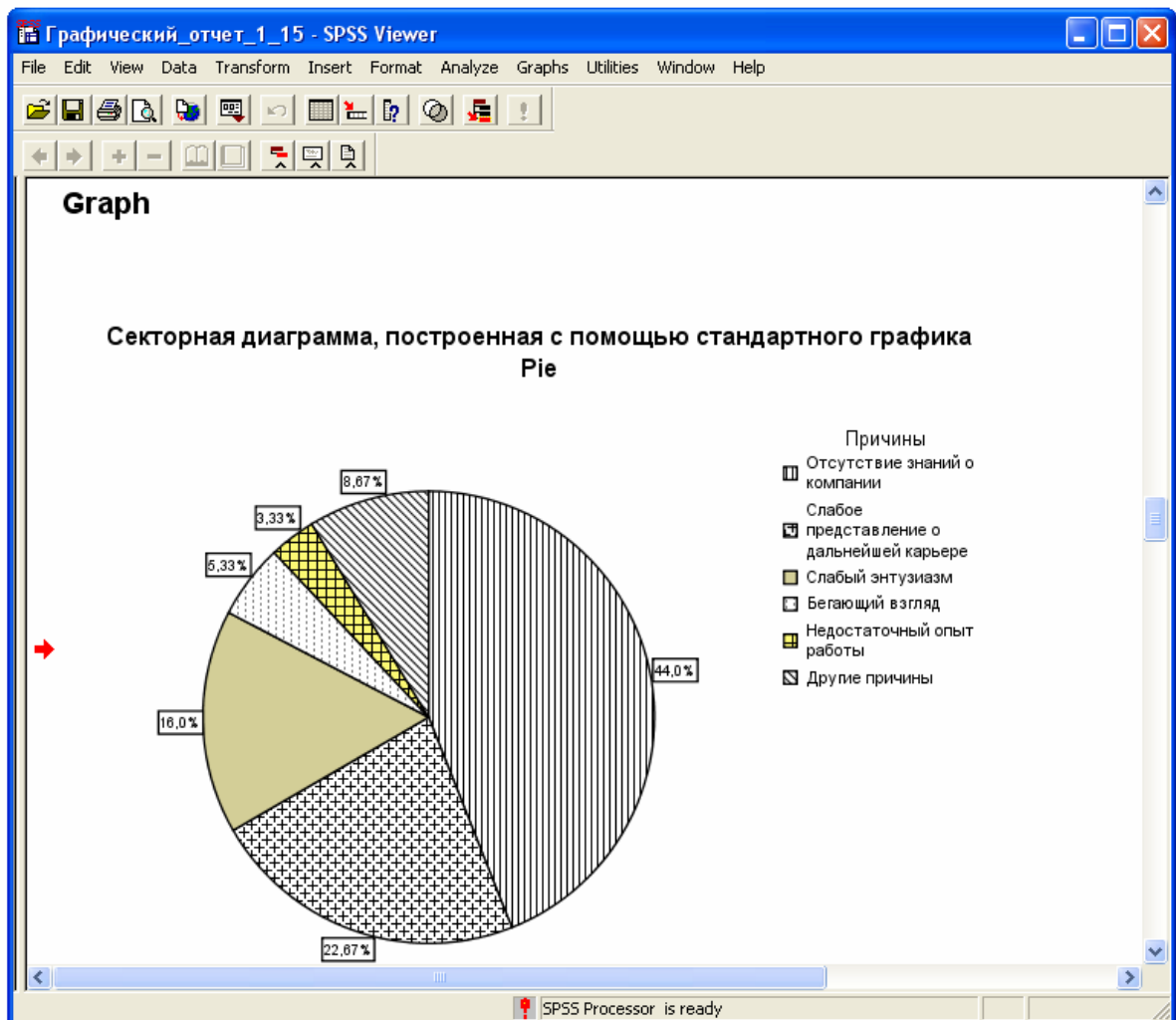


Рис. 2.28. Окно вывода для файла *Графический_отчет_1_15.spo*

Построение секторной диаграммы с помощью интерактивного графика. Выберем в меню Graphs (*Графики*) пункт Interactive (*Интерактивные*), в нем – Pie... (*Круговая...*), указав тип Simple... (*Простая...*).

В открывшемся диалоговом окне Create Simple Pie Chart (*Создать простую круговую диаграмму*) на вкладке Assign Variables (*Назначение переменных*) перенесем переменную **Причины** в поле Slice By (*Сектор при помощи*), а переменную **Количество_ответов** – в поле Slice Summary (*Сводка для сектора*) (см. рис. 2.29 ниже). Включим радиокнопку Style (*Стиль*) и поставим галочку на элементе управления Display Key (*Отобразить ключ*). Затем на вкладке Pies (*Круги*) в блоке Slice Labels (*Метки секторов*) необходимо поставить галочки для тех значений, которые должны отображаться на диаграмме. В нашем случае – это Percent (*Процент*). В поле Location: (*Расположение:*) оставим элемент управления All Outside (*Все снаружи*) в исходном состоянии. После клика на кнопку Ok появится окно вывода с построенной диаграммой (см. рис. 2.30 ниже).

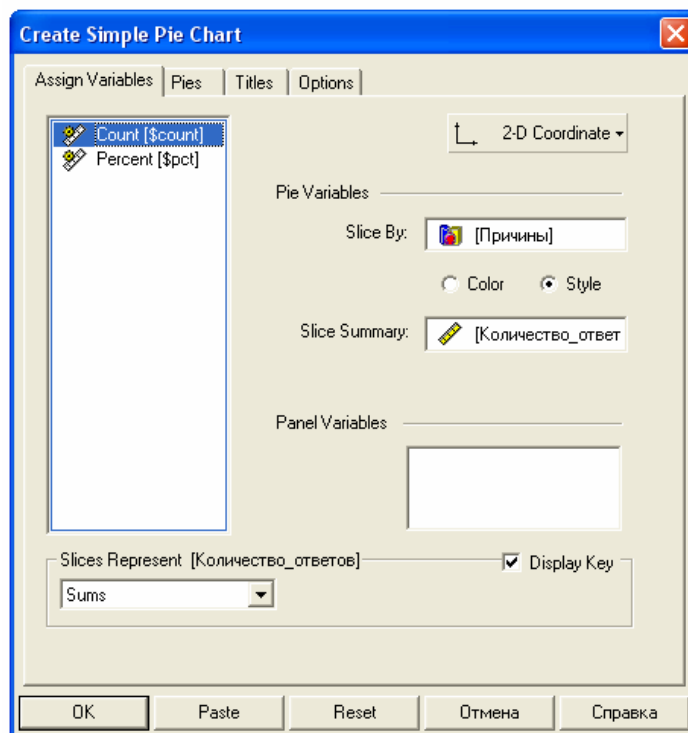


Рис. 2.29. Диалоговое окно *Create Simple Pie Chart*, вкладка *Assign Variables*

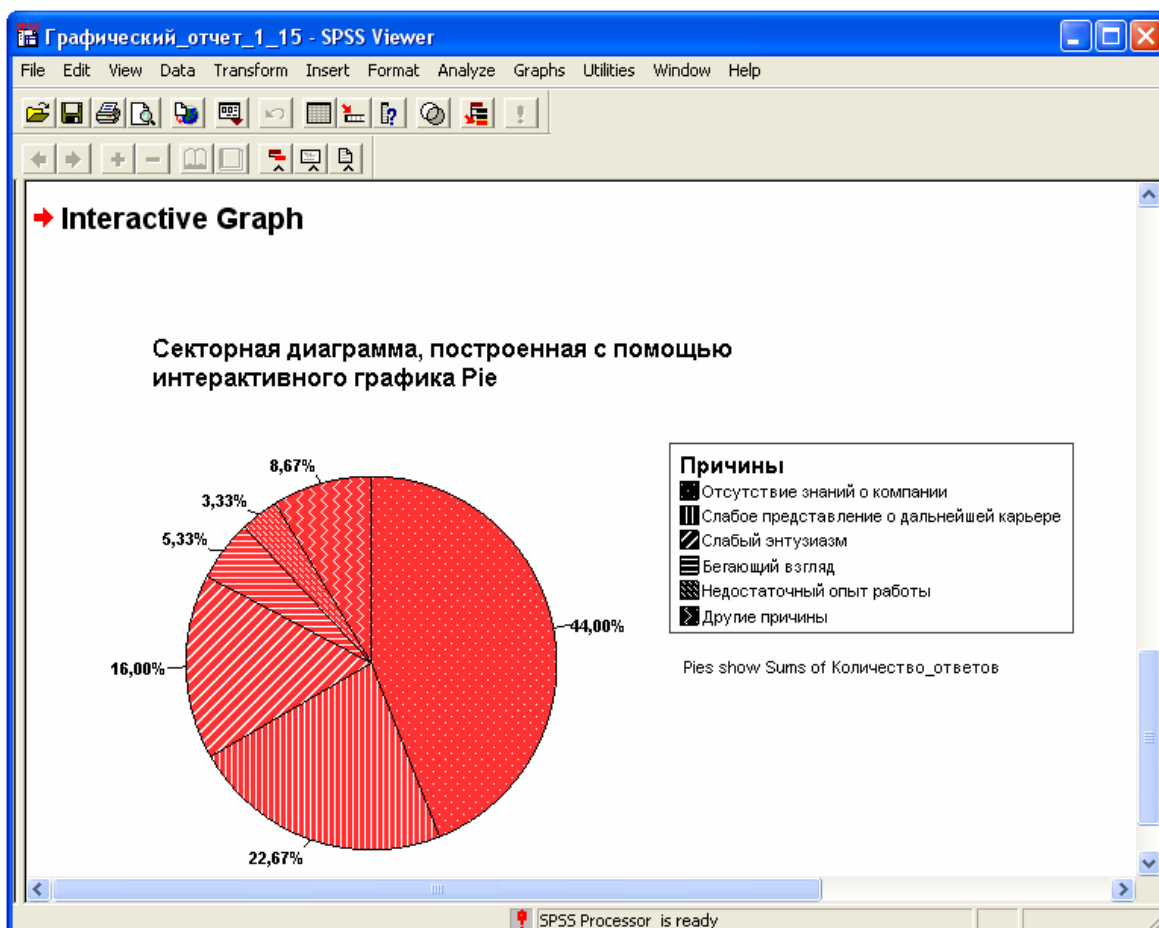


Рис. 2.30. Окно вывода для файла *Графический_отчет_1_15.spo*

Справа от диаграммы приведена ее легенда. Она, как и для стандартной секторной диаграммы, выводится по умолчанию.

Построение диаграммы Парето. В завершение графического отчета построим диаграмму Парето. Для этого в меню Graphs (*Графики*) необходимо выбрать пункт Pareto... (*Парето...*). Этапы построения диаграммы Парето были подробно разобраны ранее, поэтому ограничимся здесь приведением иллюстрации окна вывода (см. рис. 2.31) и общими комментариями.

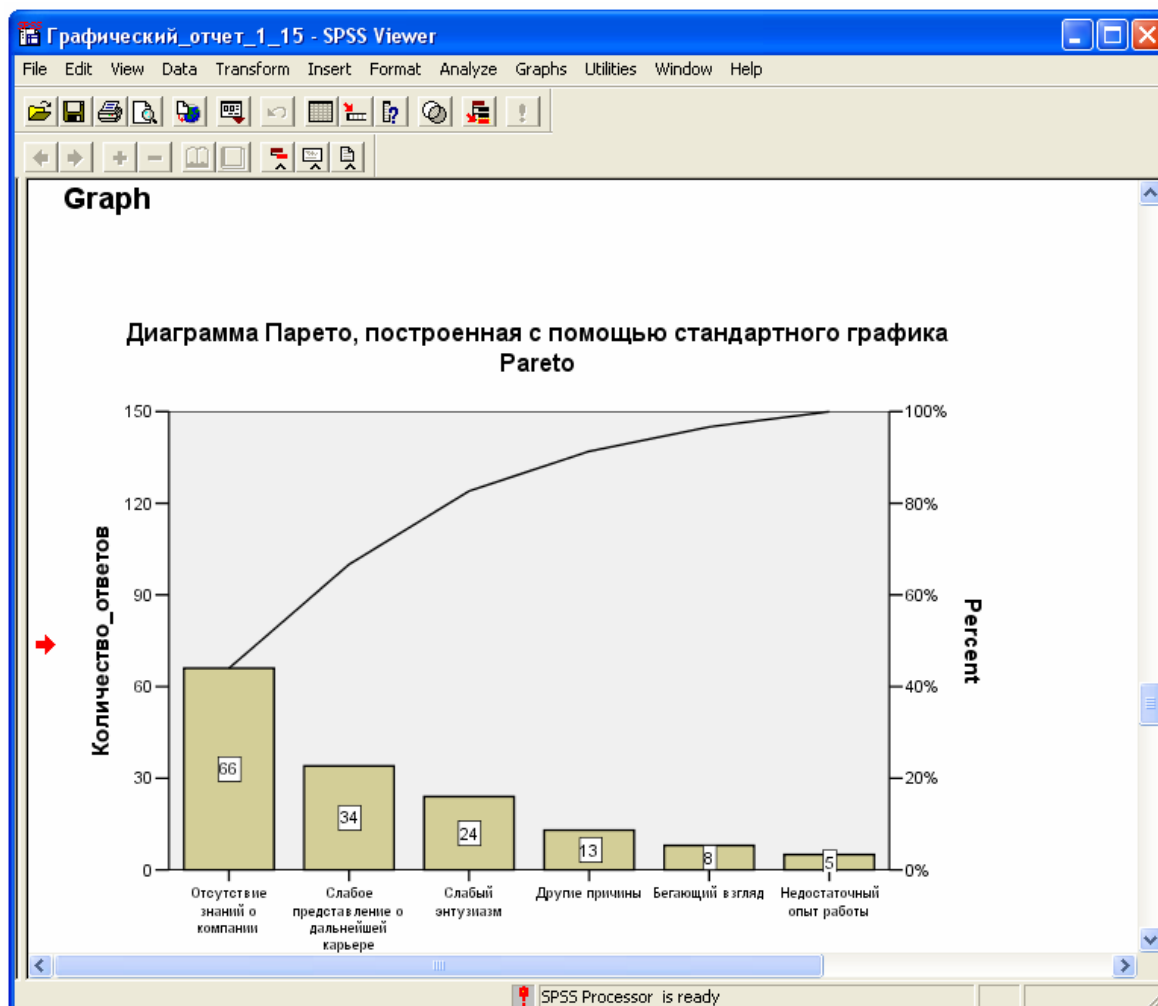


Рис. 2.31. Окно вывода для файла *Графический отчет_1_15.spo*

Анализируя построенные диаграммы, можно сделать следующие выводы: Наиболее часто встречающийся ответ, мотивирующий отказ в предоставлении работы, – это «Отсутствие знаний о компании». Из общего числа ответов таких было 66 или 44% (чуть меньше половины), на основании чего можно говорить о том, что компания отдаст предпочтение соискателям, которые не просто ищут работу и могут ее выполнять, а заинтересованы в получении вакансии именно в этой компании. Ответу «Недостаточный опыт работы», напротив, соответствует наименьшее число ответов – 5 из 150 или 3,33%, что говорит о том, что компания делает ставку скорее на энтузиазм соискателя, его желание работать и способ-

ность к самообразованию и получению опыта именно в стенах данной компании, чем на формальный трудовой стаж. Вероятно, у компании есть интерес к молодым специалистам, получившим образование недавно.

Диаграмма Парето показывает, что второй по численности ответ – это «Слабое представление о дальнейшей карьере» (таких ответов было 34 или 22,67% – почти четверть), что говорит о том, что компании интересны целеустремленные соискатели, имеющие свое направление развития карьеры, которые знают, чего хотят, и готовые приложить к этому достаточные усилия. Ответы «Недостаточное знание о компании» и «Слабое представление о дальнейшей карьере» (две наиболее распространенные причины отказа в предоставлении работы) вместе мотивируют 66,67% всех случаев отказа, а вместе с третьей по числу ответов причиной «Слабый энтузиазм» – 82,67%, что еще раз подтверждает ориентацию компании на энергичных, целеустремленных людей, заинтересованных в получении работы именно в данной компании. Поэтому соискателям, отправляющимся на собеседование, необходимо иметь полное представление о возможном работодателе (направление деятельности, масштабы, целевые группы на рынке и пр.), ясно видеть направление развития своей карьеры (иметь четкие планы, предполагаемые сроки их достижения) и всячески демонстрировать желание работать.

III. Таблицы сопряженности признаков

Таблица сопряженности является одним из распространенных инструментов многомерного анализа (методы анализа, такие как частотный анализ, вычисление статистических характеристик для отдельных переменных, называются одномерными). Она позволяет выяснить, существует ли взаимосвязь между двумя или более переменными, и наиболее удобна в случае, когда стоит вопрос о наличии взаимосвязи между категориальными переменными (переменными, относящимися к номинальной или порядковой шкалам) с не очень большим числом категорий. Продемонстрируем построение и анализ таблиц сопряженности на примере.

Ниже приведены данные социологического исследования, призванного выяснить, взаимосвязаны ли такие факторы, как семейное положение и жизненные приоритеты. Респондентам задавали два вопроса: «Ваше семейное положение?» и «Что Вы больше всего цените в жизни?», ответы на которые приведены в таблице:

<i>Жизненные ценности</i>	<i>Семейное положение</i>				
	<i>Женат</i>	<i>Разведен</i>	<i>Вдовец</i>	<i>Холост</i>	<i>Всего</i>
<i>Друзья</i>	<i>13</i>	<i>14</i>	<i>2</i>	<i>10</i>	<i>39</i>
<i>Интересная работа</i>	<i>18</i>	<i>4</i>	<i>2</i>	<i>10</i>	<i>34</i>
<i>Семья</i>	<i>30</i>	<i>5</i>	<i>5</i>	<i>5</i>	<i>45</i>
<i>Материальное благосостояние</i>	<i>37</i>	<i>4</i>	<i>3</i>	<i>5</i>	<i>49</i>
<i>Здоровье</i>	<i>19</i>	<i>8</i>	<i>7</i>	<i>5</i>	<i>39</i>
<i>Всего</i>	<i>117</i>	<i>35</i>	<i>19</i>	<i>35</i>	<i>206</i>

Необходимо построить таблицу сопряженности и на основании ее анализа сделать вывод о наличии либо отсутствии взаимосвязи между семейным положением и приоритетами человека.

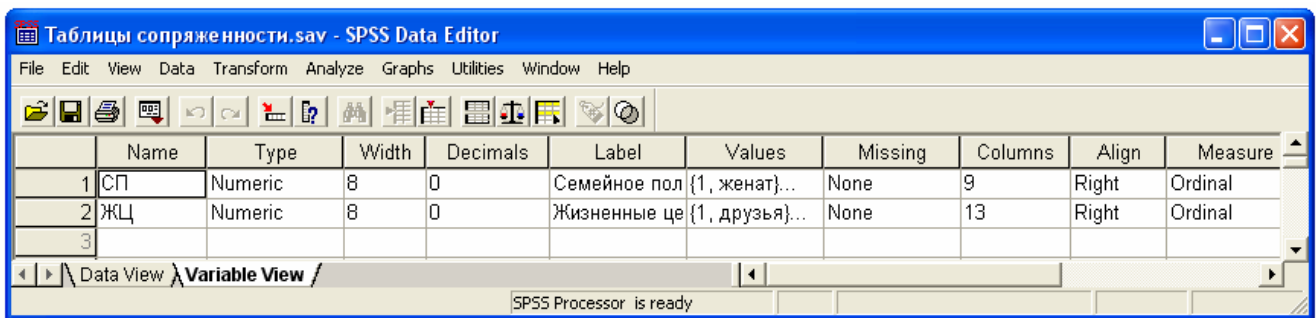
Создание файла данных *Таблицы сопряженности.sav*. Файл данных организован следующим образом: имеем две переменные – *СП* и *ЖЦ* типа Numeric со шкалой измерения Ordinal (*Порядковая*), относящейся к категориальному типу переменных. Метки переменных (Label) – названия, более подробно описывающие переменные – *Семейное положение* и *Жизненные ценности*

соответственно. Метки значений (Values) – это названия, подробно описывающие значения переменных. Для переменной **СП** имеем четыре различных значения:

- 1 – женат,
- 2 – разведен,
- 3 – вдовец,
- 4 – холост;

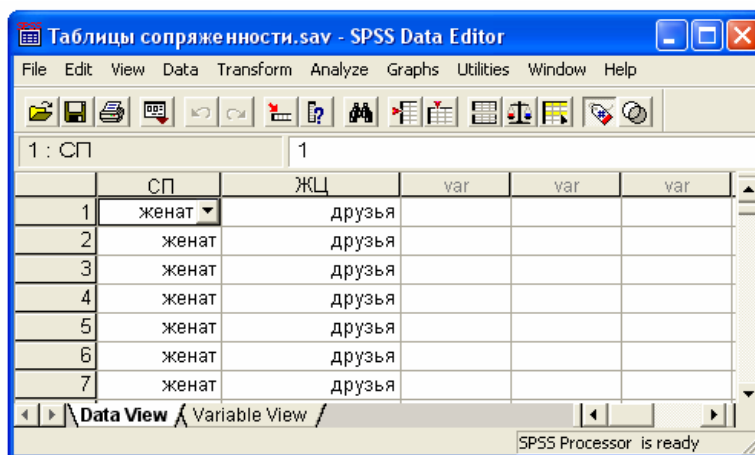
для переменной **ЖЦ** – пять различных значений:

- 1 – друзья,
- 2 – интересная работа,
- 3 – семья,
- 4 – материальное благосостояние,
- 5 – здоровье.



*Рис. 3.1. Окно редактора данных для файла **Таблицы сопряженности.sav**, вкладка **Variable View***

Файл данных содержит 206 (общее количество респондентов) записей, каждая из которых содержит по одному значению каждой переменной.



*Рис. 3.2. Окно редактора данных для файла **Таблицы сопряженности.sav**, вкладка **Data View***

Создание таблиц сопряженности. Выберем в меню Analyze (*Анализ*) пункт Descriptive Statistics (*Дескриптивные статистики*), а в нем – Crosstabs... (*Таблицы сопряженности...*). Откроется диалоговое окно Crosstabs:

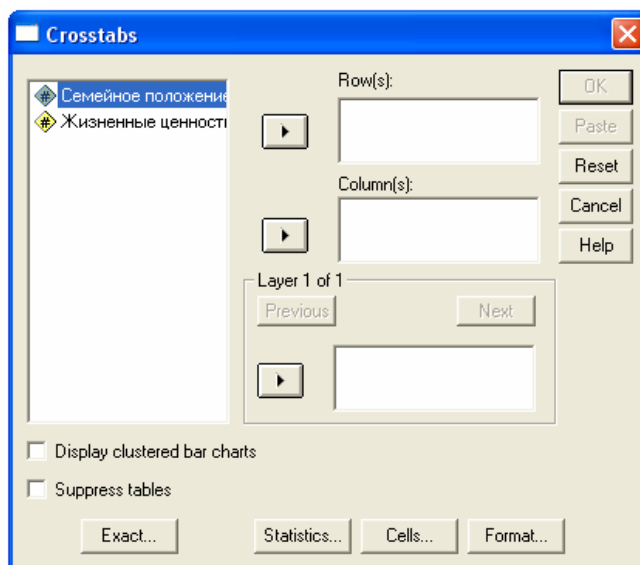


Рис. 3.3. Диалоговое окно Crosstabs

Список исходных переменных содержит переменные вновь созданного файла данных. Здесь можно выбрать переменные для строк и столбцов таблицы сопряженности. Построим таблицу сопряженности для имеющихся переменных **СП** (Семейное положение) и **ЖЦ** (Жизненные ценности). Для этого перенесем переменную **ЖЦ** в список строк (Rows), а переменную **СП** – в список столбцов (Columns). Поначалу для диалогового окна Crosstabs сохраним его настройки по умолчанию: элементы управления Display clustered bar charts (*Отобразить столбиковые кластеризованные диаграммы*) и Suppress tables (*Не выводить таблицы*) оставим в исходном состоянии. После клика на кнопку ОК в окне вывода будут построены две таблицы: Case Processing Summary (*Сводка обработки наблюдений*) и сама таблица сопряженности, в которой представлены частоты наблюдаемых ответов респондентов (см. рис. 3.4 ниже). Отметим, что таблица сопряженности по умолчанию строится в стандартном формате. Сохраним файл как *Таблицы сопряженности.spo*.

Переменная **СП** (Семейное положение) является «столбцовой» переменной, так как каждое ее значение (женат, разведен и т.д.) отображается в отдельном столбце. Переменная **ЖЦ** (Жизненные ценности) – это «строковая» переменная, так как каждое ее значение (друзья, интересная работа и т.д.) отобража-

ется в отдельной строке таблицы. Значение в каждой ячейке таблицы – количество наблюдений (частота). Так, например, видим, что 37 женатых (замужних) респондентов считают материальное благосостояние своей главной жизненной ценностью, а 7 респондентов-вдовцов – здоровье.

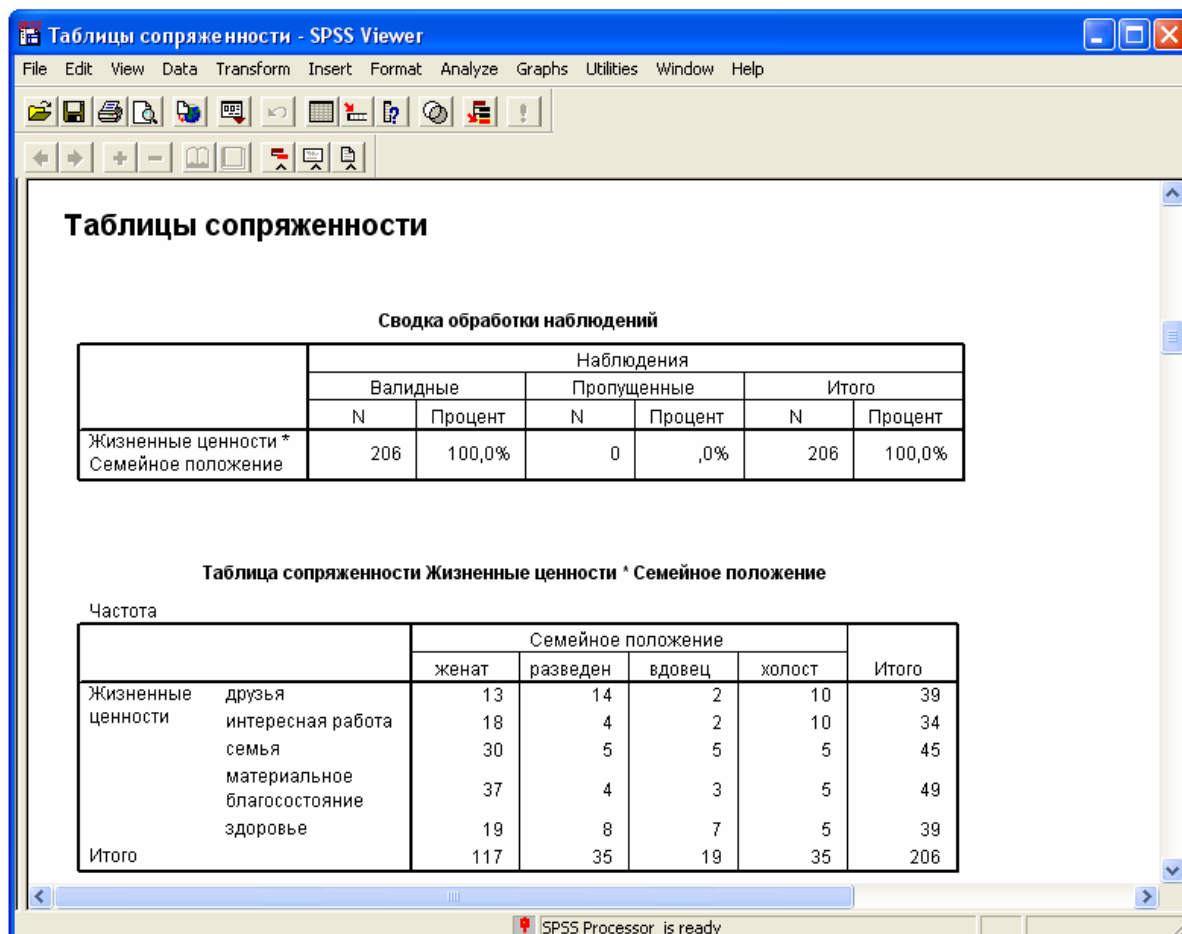


Рис. 3.4. Окно вывода для файла *Таблицы сопряженности.spo*

Если для таблицы сопряженности приняты параметры по умолчанию, в каждой ячейке отображаются только абсолютные частоты. Это и есть так называемый стандартный формат таблиц сопряженности. Метки переменных и значений в таблице соответствуют определениям переменных в файле данных SPSS. Числа в последней строке и в последнем столбце (Итого) показывают суммы значений соответственно по столбцам и по строкам. В данном примере суммы по столбцам указывают, например, что 45 (30+5+5+5) всех респондентов считают семью своей основной жизненной ценностью. Суммы по строкам показывают, например, что 35 опрошенных (14+4+5+4+8) являются разведенными людьми. При анализе принимались в расчет 206 допустимых наблюдений.

Полученные результаты можем интерпретировать следующим образом:

- из 206 опрошенных, которые учитывались при анализе, – 117 людей, состоящих в браке, 35 разведенных, 19 вдовых и 35 холостых людей;
- 18 респондентов, состоящих в браке, жизненной ценностью назвали интересную работу, тогда как среди вдовых – таких только двое;
- 5 холостых респондентов отметили своей жизненной ценностью здоровье, тогда как среди разведенных людей – таких 8 (при одинаковом числе тех и других).

Даже первое впечатление, которое возникает при анализе таблицы сопряженности, свидетельствует о том, что зависимость между переменными **СП** (Семейное положение) и **ЖЦ** (Жизненные ценности) существует. Холостые респонденты (вероятно, люди молодого возраста) склонны выбирать друзей и интересную работу, женатые – материальное положение и семью, вдовы – здоровье. Для более детального исследования зависимости потребуется ответить на следующие вопросы:

- ✓ Существует ли зависимость вообще?
- ✓ Что можно сказать об интенсивности этой зависимости?
- ✓ Что можно сказать о направлении и характере этой зависимости?

Более тщательно исследовать существование зависимости позволяет вычисление значений ожидаемых частот (в предыдущей таблице сопряженности приведены абсолютные (или наблюдаемые) частоты). Чтобы определить значения ожидаемых частот, выберем в меню *Analyze (Анализ)* пункт *Descriptive Statistics (Дескриптивные статистики)*, а в нем – пункт *Crosstabs... (Таблицы сопряженности...)*. Кликом на кнопку *Cells... (Ячейки...)* откроем диалоговое окно *Crosstabs: Cell Display (Таблицы сопряженности: Отображение ячеек)*. По умолчанию в ячейках таблицы сопряженности отображаются только наблюдаемые значения частот. В блоке *Counts (Частоты)* можно выбрать *Observed (Наблюдаемые)* или *Expected (Ожидаемые)*. Если поставить галочку на последнем элементе управления, будут отображаться ожидаемые частоты. Они вычисляются как произведения сумм соответствующей строки и столбца, деленные на общую сумму частот.

Итак, выберем *Expected* и получим следующую таблицу сопряженности:

Таблицы сопряженности

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
Жизненные ценности * Семейное положение	206	100,0%	0	,0%	206	100,0%

Таблица сопряженности Жизненные ценности * Семейное положение

			Семейное положение				Итого
			женат	разведен	вдовец	холост	
Жизненные ценности	друзья	Частота	13	14	2	10	39
		Ожидаемая частота	22,2	6,6	3,6	6,6	39,0
	интересная работа	Частота	18	4	2	10	34
		Ожидаемая частота	19,3	5,8	3,1	5,8	34,0
	семья	Частота	30	5	5	5	45
		Ожидаемая частота	25,6	7,6	4,2	7,6	45,0
	материальное благосостояние	Частота	37	4	3	5	49
		Ожидаемая частота	27,8	8,3	4,5	8,3	49,0
	здоровье	Частота	19	8	7	5	39
		Ожидаемая частота	22,2	6,6	3,6	6,6	39,0
Итого		Частота	117	35	19	35	206
		Ожидаемая частота	117,0	35,0	19,0	35,0	206,0

Рис. 3.5. Окно вывода для файла *Таблицы сопряженности.spo*

Теперь под наблюдаемыми частотами (Count) появились ожидаемые значения (Expected Count). Видим, что для одних позиций ожидаемые частоты оказались выше наблюдаемых, для других – наоборот, ниже. Так, если сравнить женатых и холостых респондентов, то видим, что для значений переменной **ЖЦ** (Жизненные ценности) «друзья», «интересная работа» и «здоровье» ожидаемые частоты выше наблюдаемых, а для значений «семья» и «материальное положение» – ниже наблюдаемых.

Таблицы сопряженности, которые мы рассмотрели выше, имеют тот недостаток, что в них приводятся только абсолютные значения. Чтобы узнать, насколько эти значения важны по отношению к общему количеству, надо определить их процентную долю. Для вычисления процентных значений в диалоговом окне Crosstabs... (*Таблицы сопряженности...*), не меняя прежних настроек, кликнем на кнопку Cells... (*Ячейки...*). Откроется диалоговое окно Crosstabs: Cell Display (*Таблицы сопряженности: Отображение ячеек*).

В блоке Percentages (*Проценты*) можно выбрать один или более из ниже-
следующих вариантов отображения:

- ✓ Row (*По строкам*). Вычисляются процентные значения по строкам: количество наблюдений в каждой ячейке, отнесенное к сумме по строке.
- ✓ Column (*По столбцам*). Вычисляются процентные значения по столбцам: количество наблюдений в каждой ячейке, отнесенное к сумме по столбцу.
- ✓ Total (*Полные*). Вычисляются полные процентные значения: количество наблюдений в каждой ячейке, отнесенное к общей сумме наблюдений.

Построим отдельно три различных таблицы сопряженности: полную (см. рис. 3.6), для сумм по строкам (см. рис. 3.7 ниже) и для сумм по столбцам (см. рис. 3.8 ниже).

Таблицы сопряженности

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
Жизненные ценности * Семейное положение	206	100,0%	0	,0%	206	100,0%

Таблица сопряженности Жизненные ценности * Семейное положение

			Семейное положение				Итого
			женат	разведен	вдовец	холост	
Жизненные ценности	друзья	Частота	13	14	2	10	39
		% по таблице (слою)	6,3%	6,8%	1,0%	4,9%	18,9%
	интересная работа	Частота	18	4	2	10	34
		% по таблице (слою)	8,7%	1,9%	1,0%	4,9%	16,5%
	семья	Частота	30	5	5	5	45
		% по таблице (слою)	14,6%	2,4%	2,4%	2,4%	21,8%
	материальное благосостояние	Частота	37	4	3	5	49
		% по таблице (слою)	18,0%	1,9%	1,5%	2,4%	23,8%
	здоровье	Частота	19	8	7	5	39
		% по таблице (слою)	9,2%	3,9%	3,4%	2,4%	18,9%
Итого		Частота	117	35	19	35	206
		% по таблице (слою)	56,8%	17,0%	9,2%	17,0%	100,0%

Рис. 3.6. Окно вывода для файла *Таблицы сопряженности.spo*

Анализируя эту общую таблицу, можем сделать следующие выводы:

- 14,6% всех респондентов – это люди, состоящие в браке и называющие своей основной жизненной ценностью семью;

- 4,9% – это холостяки, предпочитающие интересную работу;
- 9,2% всех респондентов являются вдовыми;
- 23,8% респондентов определяют материальное благополучие своей основной жизненной ценностью.

Рассмотрим теперь таблицу с суммой по строкам. В ней все респонденты делятся на пять групп (будучи отнесенными к каждой из жизненных ценностей), каждая из которых представляется как 100%:

The screenshot shows the SPSS Viewer window with the following tables:

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
Жизненные ценности * Семейное положение	206	100,0%	0	,0%	206	100,0%

Таблица сопряженности Жизненные ценности * Семейное положение

% по категории переменной Жизненные ценности

		Семейное положение				Итого
		женат	разведен	вдовец	холост	
Жизненные ценности	друзья	33,3%	35,9%	5,1%	25,6%	100,0%
	интересная работа	52,9%	11,8%	5,9%	29,4%	100,0%
	семья	66,7%	11,1%	11,1%	11,1%	100,0%
	материальное благополучие	75,5%	8,2%	6,1%	10,2%	100,0%
	здоровье	48,7%	20,5%	17,9%	12,8%	100,0%
Итого		56,8%	17,0%	9,2%	17,0%	100,0%

Рис. 3.7. Окно вывода для файла *Таблицы сопряженности.spo*

Анализируя эту таблицу, можем сделать следующие выводы:

- 35,9% отметивших своей главной жизненной ценностью друзей – разведенные люди и лишь 5,1% – вдовы;
- среди тех, кто заявил свою главную жизненную ценность – материальное благополучие – 75,5% людей, состоящих в браке, и лишь 8,2% разведенных людей.

Наиболее репрезентативной для данной задачи является таблица сопряженности для сумм по столбцам (см. рис. 3.8): каждое значение переменной *СП* (*Семейное положение*) делит респондентов на группы, и объем каждой такой

группы берется за 100%. Это позволяет определить процент женатых (разведенных, вдовых, холостых), выбирающих определенный жизненный приоритет.

The screenshot shows the SPSS Viewer window titled "Таблицы сопряженности - SPSS Viewer". It displays two tables. The first is a summary table for the variable "Жизненные ценности * Семейное положение". The second is a crosstabulation table showing the percentage distribution of life values across different marital statuses.

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
Жизненные ценности * Семейное положение	206	100,0%	0	,0%	206	100,0%

Таблица сопряженности Жизненные ценности * Семейное положение

% по категории переменной Семейное положение

		Семейное положение				Итого
		женат	разведен	вдовец	холост	
Жизненные ценности	друзья	11,1%	40,0%	10,5%	28,6%	18,9%
	интересная работа	15,4%	11,4%	10,5%	28,6%	16,5%
	семья	25,6%	14,3%	26,3%	14,3%	21,8%
	материальное благосостояние	31,6%	11,4%	15,8%	14,3%	23,8%
	здоровье	16,2%	22,9%	36,8%	14,3%	18,9%
Итого		100,0%	100,0%	100,0%	100,0%	100,0%

*Рис. 3.8. Окно вывода для файла **Таблицы сопряженности.spo***

Анализируя эту таблицу, можем сделать следующие выводы:

- приоритеты между различными жизненными ценностями всех респондентов распределились практически равномерно (процентные частоты в последнем столбце), но внутри каждой из групп распределение далеко от этого;
- 40% всех разведенных считают своей жизненной ценностью друзей, при этом среди женатых и вдовых этот процент в 4 раза ниже (11,1% и 10,5% соответственно);
- 28,6% холостых говорят о важности интересной работы, тогда как для женатых этот показатель почти в два раза ниже (15,4%);
- 31,6% всех респондентов, состоящих в браке, объявляют своей жизненной ценностью материальное благосостояние и только 11,1% – друзей;
- среди холостых респондентов 28,6% опрошенных «выбирают» друзей и только 14,3% – материальное благополучие;

- среди вдовых 36,8% опрошенных говорят о том, что их основная жизненная ценность – здоровье, при этом лишь 10,5% респондентов склоняются в пользу друзей и столько же в пользу интересной работы.

Как можно объяснить полученные результаты? Среди женатых респондентов 25,6% и 31,6% опрошенных говорят о приоритетах в пользу семьи и материального благополучия (это, вероятно, связано с материальной и моральной ответственностью людей за семью и детей) и только 11,1% – говорят в пользу друзей (остается мало времени на общение из-за большого количества забот и ответственности). Среди разведенных 40% респондентов говорят в пользу друзей и 22,9% – в пользу здоровья (люди свободные от семьи имеют больше возможностей тратить время только на себя); среди вдовых 36,8% респондентов «голосуют» за здоровье (скорее всего, возраст этой категории весьма почтительный и вопросы здоровья стоят острее, чем в молодости) и 26,3% – за семью (семья – это не только супруг и дети, но и внуки). Среди холостых респондентов (предположительно, это люди молодого возраста) 28,6% всех опрошенных выделяют приоритетом друзей и интересную работу (люди развиваются, нуждаются в общении, имеют возможность жить в удовольствие, будучи не обремененными семьей и болезнями), при этом материальное благополучие отмечают лишь 14,3% респондентов (опять же, присущая молодости жизнь «в удовольствие» – работать там, где нравится, а не там, где много платят). Но это всего лишь мнение авторов, и у читателя вполне может быть другое истолкование полученных результатов. В любом случае наличие взаимосвязи между переменными очевидно.

Все эти значения можно было бы вывести в одной таблице сопряженности, однако при этом пропадает наглядность и последовательность трактовки результатов.

Форматы таблиц сопряженности. В таблицах сопряженности можно изменить порядок сортировки переменных строк, кликнув в диалоговом окне Crosstabs на кнопку Format... (*Формат...*). Откроется диалоговое окно Crosstabs: Table Format (*Таблицы сопряженности: Формат таблицы*).

В блоке Row Order (*Порядок строки*) можно выбрать один из следующих вариантов сортировки значений:

- ✓ Ascending (*По возрастанию*): Значения переменных строк отображаются в порядке возрастания от наименьшего к наибольшему. Это настройка по умолчанию.
- ✓ Descending (*По убыванию*): Значения переменных строк отображаются в порядке убывания от наибольшего к наименьшему.

Графическое представление таблиц сопряженности. Данные, содержащиеся в таблицах сопряженности, можно представить в виде столбиковой диаграммы. Выберем в меню Graphs (*Графики*) пункт Bar... (*Столбиковые...*). Откроется диалоговое окно Bar Charts (*Столбиковые диаграммы*). Выберем тип диаграммы Clustered (*Кластеризованные*), оставим радиокнопку Summaries for groups of cases (*Обработка категорий для групп наблюдений*) в исходном состоянии и кликнем на кнопку Define (*Определить*). Откроется диалоговое окно Define Clustered Bar: Summaries for Groups of Cases (*Определить столбиковую диаграмму: Обработка категорий для групп наблюдений*). В блоке Bars Represent (*Столбцы отражают*) включим радиокнопку % of cases (*% наблюдений*), перенесем переменную **СП** в поле Category Axis: (*Ось категорий:*), а переменную **ЖЦ** – в поле Define Clusters by (*Определить группы при помощи*):

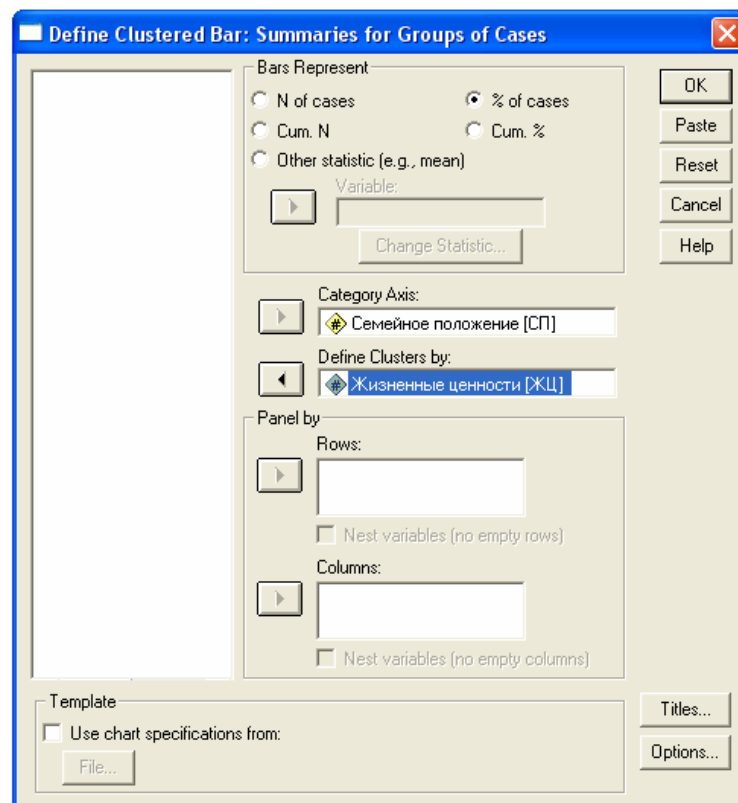


Рис. 3.9. Диалоговое окно Define Clustered Bar: Summaries for Groups of Cases

Затем кликнем на кнопку Options... (Опции...). Откроется диалоговое окно Options. Снимем в нем галочку с элемента управления Display groups defined by missing values (Отобразить группы, образованные пропущенными значениями). Кликнем на кнопку Continue, а затем – на кнопку ОК. В окне вывода появится график:

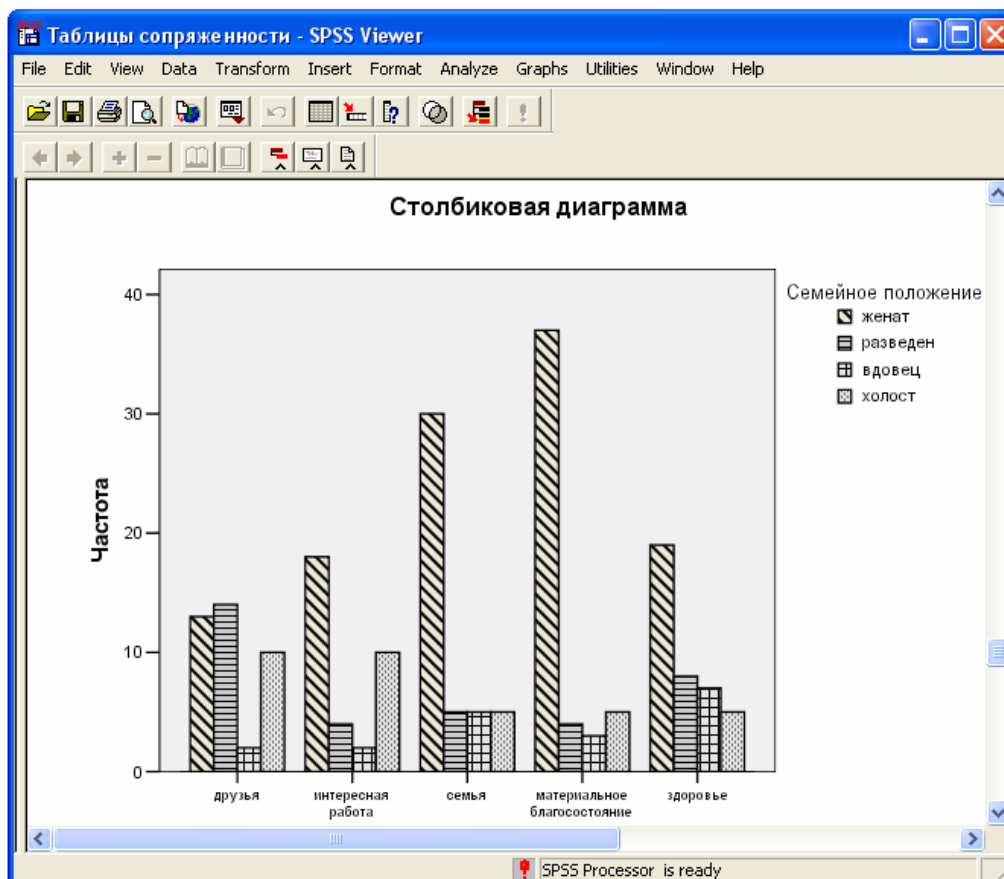


Рис. 3.10. Окно вывода для файла *Таблицы сопряженности.spo*

Можно не вызывать меню Graphs, а просто поставить галочку в диалоговом окне Crosstabs на элементе управления Display clustered bar charts (Отобразить столбиковые кластеризованные диаграммы), если выбрать в меню Analyze (Анализ) пункт Descriptive Statistics (Дескриптивные статистики), а в нем – Crosstabs... (Таблицы сопряженности...).

На диаграмме столбцы с одинаковой заливкой показывают долю респондентов определенного семейного положения в различных группах, определяемых приоритетной жизненной ценностью. График также отражает зависимость, полученную при анализе таблиц сопряженности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Бююль А., Цёфель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. : Пер. с нем. – СПб.: “ДиаСофтЮП”, 2002. – 608 с.
2. Дубнов П.Ю. Обработка статистической информации с помощью SPSS – М.: ООО “Издательство АСТ”: Издательство “НТ Пресс”, 2004. – 221 с.
3. Плис А.И., Сливина Н.А. Практикум по прикладной статистике в среде SPSS: Учеб. пособие. В 2-х ч. Ч. 1. Классические процедуры статистики – М.: Финансы и статистика, 2004. – 288 с.
4. Сигел Э. Практическая бизнес-статистика. : Пер. с англ. – М.: Издательский дом “Вильямс”, 2004. – 1056 с.

Бобков Николай Николаевич

Дёмкин Валерий Матвеевич

Солычева Ольга Михайловна

SPSS: Анализ данных в менеджменте

Описательные статистики

*Методическая разработка по курсу
“Анализ данных в менеджменте”*