ПРЕДИСЛОВИЕ

Широкий спектр исследований в различных областях науки – в социологии, экономике, медицине, биологии, криминалистике и др. основан на использовании методов математической статистики и компьютерных программ, объединенных единым понятием «анализ данных». Применение анализа данных в каждой области имеет соответствующие особенности, связанные со структурой информации, содержанием задач и интерпретацией результатов. Данное учебно-методическое пособие содержит методику применения анализа данных в области социологии.

При подготовке учебно-методического пособия использованы общеизвестные, но не всегда доступные российскому читателю учебные пособия по статистическому анализу, такие как курс эконометрического анализа Грина [1], настольная книга по статистической методологии под редакцией Армингера, Клогга, Собела (G. Arminger, C. Clogg, M. E. Sobel) [2], объемистый учебник по прикладному статистическому анализу С. А. Айвазяна и В. С. Мхиторяна [3], учебник Ю. Н. Толстовой [4].

В пособии рассматриваются преимущественно методы, представленные пакетом программ по обработке и статистическому анализу социологических данных: *Statistical Package for Social Science* (*SPSS*). Он содержит все основные разделы анализа данных и во многих зарубежных и отечественных университетах является базовым при подготовке студентов гуманитарных факультетов.

Наше пособие включает лишь ключевые моменты практического анализа данных с использованием SPSS. Официальный дилер SPSS в России (http://www.spss.ru) предоставляет три учебника по применению пакета: «Руководство пользователя SPSS. Книга 1» [5], «Руководство пользователя SPSS» [6] и «Руководство по применению SPSS» [7]. При подготовке материалов использовались также: путеводитель по синтаксису SPSS [8], документация SPSS по регрессионному анализу [9], точным статистическим тестам [10], документация по кластерному анализу и многомерному шкалированию [11], другие материалы по SPSS. Учебники содержат достаточно полное описание методики применения пакета по многим разделам, поэтому для получения дополнительной информации мы отправляем читателя к этим руководствам. Однако ориентированы они преимущественно на работу с пакетом в режиме диалога. В нашем учебно-методическом посо-

бии баланс от диалогового режима смещен в сторону использования языка программирования заданий для SPSS, поскольку серьезная работа с данными требует определенных навыков в этой области.

Практически ежегодно выпускается новая версия SPSS, постоянно меняется дизайн, появляются новые программы и возможности работы с пакетом. Хотя данное учебно-методическое пособие ориентировано на 9-ю версию SPSS, его целью является донести до читателя основные принципы работы с SPSS и описать основные команды управления, которые остаются практически неизменными в течение уже 20 лет. При этом авторы старались не упустить из вида и новые возможности.

В пособие также включена отечественная разработка – метод анализа связи между неальтернативными вопросами [12]. Мы попытались доступным языком раскрыть сложную тему анализа множественных сравнений оценок значимости связи по таблицам для неальтернативных вопросов.

Большинство известных статистических пакетов реализует такие же методы, что и SPSS, и предполагает аналогичную структуру данных, поэтому освоение SPSS позволяет приобрести необходимые навыки для компьютерного анализа данных вообще.

Замечание по оформлению таблиц . Практически все они получены непосредственно пакетом SPSS и оформлены как машинные выдачи.

Глава 1. ИНФОРМАЦИЯ, ОБРАБАТЫВАЕМАЯ СТАТИСТИЧЕСКИМ ПАКЕТОМ

1.1. Анкетные данные

В большинстве социологических исследований анализируется анкетная информация. Условно эти данные можно представить в виде матрицы, строкам которой соответствуют объекты (анкеты), а столбцам – признаки (отдельные вопросы и подвопросы анкеты). Синонимом термина «признак» является термин «переменная», в дальнейшем мы будем употреблять их равноправно.

В современных статистических пакетах такую информацию принято представлять в виде таблицы. Обычно обрабатывается один файл данных, представленных в виде матрицы, которая на экране напоминает лист таблицы «Excel».

При кодировании информации для заполнения матрицы необходимо пользоваться определенными правилами в соответствии со структурой обрабатываемой анкеты.

Пример

Анкета обследования жалоб и проблем населения (шутка)

1. Пол
    1. Мужской
    2. Женский
2. Возраст ……
3. Проблемы (укажите 3 основные проблемы):
    1. Учеба
    2. Свободное время
    3. Любовь
    4. Музыка
4. Жалобы:
    1. Служба
    2. Здоровье
    3. Зарплата
    4. Жена
    5. Собака соседа

Соответствующая анкете матрица данных изображена на рис. 1.1. Пол закодирован в соответствии с содержимым анкеты кодами: 1 – мужчины, 2 – женщины; возраст введен непосредственно отдельным столбцом; проблемы закодированы в трех переменных, в которых указаны коды обведенных при опросе подсказок. Для каждой жалобы отведена своя переменная.

N анкеты	1. Пол	2. Возраст	3. Проблемы			4. Жалобы				
						1. Служба	2. Здоровье	3. Зарплата	4. Жена	5. Собака соседа
1	1	20	1	4	.	1	0	0	0	1
2	1	25	2	3	4	1	0	1	0	1
3	2	34	1	2	4	1	0	0	0	1
4	1	18	1	2	.	0	0	0	0	1
.	.	.	.	.	.	.	.	.	.	.

Рис. 1.1. Структура матрицы данных обследования жалоб и проблем населения

Итого 11 переменных закодированы в 11 столбцах. Приведенная матрица содержит информацию по 4 анкетам.

В нашем пособии работа пакета иллюстрируется на данных реального опроса населения восточных регионов России за 1991 г. о передаче островов Японии (анкета «Курильские острова», текст которой приведен в приложении 1, а файл с анкетными данными называется OCT. sav). В некоторых случаях использованы фактические данные «Российского мониторинга экономического положения и здоровья населения» (RLMS, [13]).

1.2. Типы переменных

Пакет допускает числовую или символьную кодировку информации.

1.2.1. Типы кодирования переменных

В статистическом пакете SPSS ***предусмотрено* 8 *типов кодирования*** переменных. Подробнее о них можно узнать в книге [5]. Мы остановимся лишь на строчных (`string`) и числовых (`numeric`) переменных. Строчные переменные используются достаточно редко, в основном для введения ответов на открытые вопросы или фамилий респондентов. Например, строчная переменная `dj56.1.1` 8-й волны RLMS содержит именно такие ответы на вопрос «В чем состояла эта Ваша работа?».

Но обычно при внесении в компьютер информации для статистической обработки ответы на вопросы анкеты кодируются числами. Хотя с формальной точки зрения практически любая обрабатывающая программа может использовать цифры независимо от того, кодируется ли профессия, возраст или сведения о цвете глаз, различные методы анализа данных ориентированы на данные различающихся типов. Для получения интерпретируемых результатов исследователь должен различать тип обрабатываемых соответствующим методом переменных.

Данные, закодированные числами, различаются в соответствии с типами шкал измерения переменных.

1.2.2. Тип шкалы измерения переменной

Формируя данные, исследователь ставит в соответствие значениям переменной, имеющей содержательный смысл («пол», «профессия»), числовые значения («мужской» = 1, «женский» = 0 или «учитель» = 1). Используемые числовые коды для представления значений переменных называются шкалой измерения переменной. В приведенном примере это 0 и 1. В зависимости от свойств переменной выделяют неколичественные шкалы (номинальную, ординальную (ранговую)) и количественные (интервальную и шкалу отношений).

1.2.3. Неколичественные шкалы

Номинальная шкала является самым «низким» уровнем измерения. Примером таких шкал являются числовые коды для переменных «пол», «профессия». В этом случае абсолютно не важен порядок используемых числовых кодов. Принципиальное значение имеет только равенство или неравенство значений переменной.

Ординальная, или ***ранговая***. Часто значения переменной выражают степень проявления какого-либо свойства и могут быть упорядочены. Например, работа «интересна», «безразлична» или «не интересна». В этом случае шкала называется ранговой или ординальной.

1.2.4. Количественные шкалы

Количественные шкалы всегда несут информацию о порядке данных.

Интервальная шкала предполагает, что можно определить не только порядок значений, но и расстояние между значениями. Эта шкала, однако, такова, что не имеет смысла рассматривать, во сколько раз одно значение больше другого. Пример: шкала измерения температуры по Цельсию.

Шкала отношений в дополнение к определению порядка значений позволяет измерять пропорции значений. Например, мы можем смело заявить, что зарплата в 1 000 $ вдвое выше зарплаты в 500 $. Шкалу отношений имеют переменные, несущие количественную информацию (доход, возраст, количество лет проживания в данной местности и т. д.). Для нас не очень важно различие интервальной шкалы и шкалы отношений. Техника анализа переменных, измеренных в количественных шкалах (интервальной и шкале отношений) обычно одинакова.

В соответствии с типом шкалы переменные относят к номинальным, ординальным (ранговым) и количественным типам переменных.

К особому типу номинальных переменных относятся переменные, имеющие два ответа: «да» и «нет» (например, «Имеете ли Вы телевизор?»). Эти переменные называют ***дихотомическими***. Их удобно кодировать цифрами 1 («да») и 0 («нет»). Они представляют простейший вид номинальных переменных, закодированных числами (0 или 1) и поэтому могут использоваться в количественном анализе.

Приведенная классификация шкал включает не все типы возможных отношений между значениями переменной. Например, переменная «время суток» при исследовании бюджета времени имеет «кольцевую» структуру, поскольку 0 часов эквивалентно 24 часам.

В некоторых переменных часть значений упорядочена, а часть нет. К таким переменным формально не может быть применена ни одна из шкал указанных видов.

Например, ответ на вопрос о доходах личного подсобного хозяйства может представлять денежную сумму, быть ответом «не имею подсобного хозяйства» или ответом «не знаю». Здесь значения переменной только частично являются количественными и упорядоченными. При кодировании неколичественных значений рекомендуется использовать коды специального вида, которые в принципе не могут встретиться в данных. Например, в RLMS в вопросе о весе респондента ответы «затрудняюсь ответить», «отказ от ответа» и «нет ответа» кодируются кодами 997, 998 и 999 соответственно. Для анализа таких переменных часто переходят к переменным с количественной шкалой, отбросив объекты с кодами специального вида.

Для этого можно использовать специальные команды SPSS (см. ниже команду `MISSING VALUES`). Например, объявить эти числовые значения кодами неопределенности, чтобы по ошибке не получить средний вес респондента больше 300 килограммов.

1.2.5. Неальтернативные признаки

Еще более сложны данные по так называемым неальтернативным (многозначным) вопросам. Часто встречаются вопросы типа: «Какие варианты ответов, предлагаемых анкетой, Вам кажутся разумными?» В анкете на такой вопрос предлагается несколько ответов. В этих случаях признаки принято называть неальтернативными или многозначными. Неальтернативный признак можно кодировать одним из двух способов:

1. Для каждой подсказки заводится переменная, которая соответствует столбцу матрицы и заполняется нулем, если подсказка в анкете не обведена, и единицей, если обведена (рис. 1.1). В этом случае количество столбцов матрицы, содержащих ответы по данному многозначному вопросу, равно количеству подсказок в анкете. Так, для 5 ответов на четвертый вопрос анкеты примера 1.1 отводится 5 столбцов матрицы данных, заполненных нулями и единицами. Нередко вместо кодов 0 и 1 используются другие коды, тогда в программах получения таблиц по неальтернативным вопросам нужно специально указывать код, соответствующей ответу «да». Например, вопрос может быть задан следующим образом:

Согласны ли вы с тем, что:

А. Нужна новая конституция?

1. Нет 2. Да 3. Не знаю

Б. Нужно переизбрать Думу?

1. Нет 2. Да 3. Не знаю

В. Нужен новый президент

1. Нет 2. Да 3. Не знаю

В этом случае положительный ответ определяется кодом 2 и отрицательный остальными кодами. В соответствии со сказанным выше, код 2 воспринимается как 1, остальные коды как 0. В ряде программ SPSS для обозначения дихотомического представления данных используется текст `Dichotomous counted value.`

2. Второй способ кодирования неальтернативных переменных носит название кодирования ***списком.*** Список представляет порядковые номера обведенных респондентом подсказок в тексте анкеты. Кодирование списком использовано при формировании 3-й группы столбцов матрицы из примера 1.1, рис. 1.1. В этом случае количество столбцов матрицы, отведенных для ответов на вопрос, зависит от числа возможных ответов и может быть значительно меньше, чем количество подсказок в вопросе. Например, для третьего вопроса анкеты с 4 подсказками достаточно отвести три столбца матрицы данных, т. к. никто не обвел все 5 подсказок. Представление данных списком делается с единственной целью – экономией памяти машины, но вызывает затруднения при обработке. Очевидно, что перед работой с этими переменными необходимо сообщить пакету, что данные закодированы списком. Для задания списков таких переменных в командах меню **Multiple response, General tables** и соответствующих командах синтаксиса SPSS используется ключевое слово `categories.`

1.3. Имена и метки переменных

Каждый столбец при организации матрицы данных должен иметь наименование. При этом предусмотрена возможность задания переменным двух имен. Кроме коротких имен – кодов, используемых в командах, можно завести содержательные имена – метки, удобные для выдачи результатов расчетов. В примере 1.1 можем обозначить признаки следующим образом:

```
v1, v2, v3s1, v3s2, v3s3, v4d1, v4d2, v4d3, v4d4, v4d5
```

или:

```
sex, age, problem1, …, problem3, compl1, …, compl5.
```

Меткой переменной может быть непосредственная формулировка вопроса или переработанный текст вопроса, например, «Назовите, пожалуйста, ваш пол» или «Пол».

Метки значений – это задание текстовой расшифровки кодов значений переменных (для пола: 1 – «мужской», 2 – «женский»).

Использование меток переменных и значений необязательно, но оно значительно облегчает расшифровку результатов счета и экономит время при формировании окончательных отчетов в текстовом виде.

1.4. Коды неопределенных значений

Неопределенные значения переменных возникают в случаях, когда респондент пропустил вопрос, или использованы особые кодировки для ответов, которыми следует пренебречь, или рассчитываемая переменная принимает неопределенный характер. Часто возникает необходимость исключить из рассмотрения переменные, коды которых соответствуют неопределенным значениям. В пакете предусмотрена такая возможность, если эти коды задать заранее.

Ниже мы увидим, каким образом информация о метках и неопределенных значениях заносится в данные.

Глава 2. ОБЩЕЕ ОПИСАНИЕ СТАТИСТИЧЕСКОГО ПАКЕТА ДЛЯ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ И ПОДГОТОВКА ДАННЫХ

2.1. Структура пакета

При применении пакета целесообразно различать команды определения данных, преобразования данных, команды выбора объектов, т. е. сервисную часть пакета, и команды по реализации методов статистической обработки информации. К последним относятся:

– суммарные статистики для отдельных переменных;

– частоты, суммарные статистики и графики для произвольного числа переменных;

– построение *n*-мерных таблиц сопряженности и получение мер связи;

– средние, стандартные отклонения и суммы по группам;

– дисперсионный анализ и множественные сравнения;

– корреляционный анализ;

– дискриминантный анализ;

– однофакторный дисперсионный анализ;

– общая линейная модель дисперсионного анализа (GLM);

– факторный анализ;

– кластерный анализ;

– иерархический кластерный анализ;
– иерархический лог-линейный анализ;
– многомерный дисперсионный анализ;
– непараметрические тесты;
– множественная регрессия;
– методы неметрического шкалирования и др.

В пакете достаточно развито графическое представление результатов. Он позволяет получать разнообразные графики – столбиковые и круговые, ящичковые диаграммы, поля рассеяния и гистограммы и др.

2.2. Схема организации данных, окна SPSS

В пакете предусмотрена целая система входных (файлов данных) и выходных файлов (создаваемых пакетом в процессе его работы).

К входным данным в системе SPSS относятся:

1. Исходные данные. Они могут быть представлены как в виде ASCII-файла, электронной таблицы, в виде баз данных, а также в виде собственного системного SPSS-файла данных.

Системные данные SPSS включают оболочку файла, где хранятся краткие и расширенные имена переменных, метки значений, а также информация о кодах неопределенных значений. Начиная с 8-й версии SPSS хранит также информацию о неальтернативных переменных файла.

Имена системных файлов исходных данных в SPSS имеют расширение **.sav**, например, **D:\city.sav**. Непосредственный ввод данных и просмотр таких файлов в SPSS осуществляется через окно редактирования данных с названием **SPSS for Windows Data Editor**.

2. Данные, полученные из диалогов. Команды, запущенные из меню, вызывают диалоговые окна, которые позволяют в процессе работы назначить параметры и переменные для программ обработки данных.

3. Файлы синтаксиса, содержащие задание на специализированном языке пакета.

Имена файлов с программами на языке пакета имеют расширение **.sps**, например, **D:\workl.sps.** По умолчанию они будут иметь имена **syntax1.sps** или **syntax2.sps** и т. д. При необходимости эти файлы можно сохранять для дальнейшей работы.

Для работы с программами на языке SPSS в SPSS предусмотрено окно синтаксиса (**Syntax**).

К выходным данным относятся:

– Файлы результатов, содержащие таблицы, текстовые результаты, графики расчетов, имеющие имена с расширением **.spo**. По умолчанию файлам результатов даются имена **output1.spo, output2.spo** … . Для просмотра этих файлов используется окно навигатора вывода (**Output**). Часть окна

навигатора вывода отведена для дерева выдачи, что облегчает просмотр результатов расчетов.

– Все файлы, которые в дальнейшем могут представлять собой также входную информацию. К ним можно отнести файлы синтаксиса, результатов и эмпирических данных.

– Преобразованные данные входного файла (с расширением **.sav**) и файл синтаксиса (**.sps**) также могут стать выходными данными.

Следует заметить, что кроме указанных окон в пакете могут открываться и другие окна, связанные с просмотром и редактированием графиков, просмотром и редактированием таблиц, написанием программ на языке более низкого уровня (*Scripts*), чем язык синтаксиса. Язык скриптов в данном учебно-методическом пособии мы не будем рассматривать.

Поскольку содержимое всех файлов можно просматривать и редактировать, выделение входных и выходных данных условно и определяется скорее основным их назначением.

2.3. Управление работой пакета

При управлении работой пакета через меню соблюдаются стандарты системы Windows. Каждое окно имеет свое меню. Многие команды меню доступны из различных окон.

2.3.1. Основные команды меню SPSS верхнего уровня

File. Обеспечивает доступ к файлам трех типов: эмпирическим данным, выходным файлам результатов анализа и программам. С файлами каждого типа связываются соответствующие им окна. Если текущее окно содержит эмпирические данные, то команда **File** обслуживает сохранение и замену этих данных. Если окно содержит файл синтаксиса (**Syntax**) или выдачи результатов счета (**Output**), то обеспечивается обработка файла синтаксиса или выдачи. Таким образом, операции по сохранению или редактированию осуществляются в текущий момент для активизированного (верхнего) окна. На панель экрана внизу обычно выведены типы файлов, и указателем мыши можно активизировать любой из них. Либо, задавая вложенный размер окнам, можно активизировать нужное окно, нажав указатель мыши на его поле. Окно с исходными данными является обязательным. Окно вывода результатов появляется после расчетов, либо вводится пользователем. Окно, содержащее тексты выполняемых команд, необязательно и используется только по желанию пользователя.

Edit обеспечивает редактирование командных файлов, выходных файлов и файлов данных статистических наблюдений и др.

Data обеспечивает операции над данными – сортировку, слияние различных файлов данных, агрегирование, организацию подвыборки из данных. Эта команда имеется только в меню окна редактора данных.

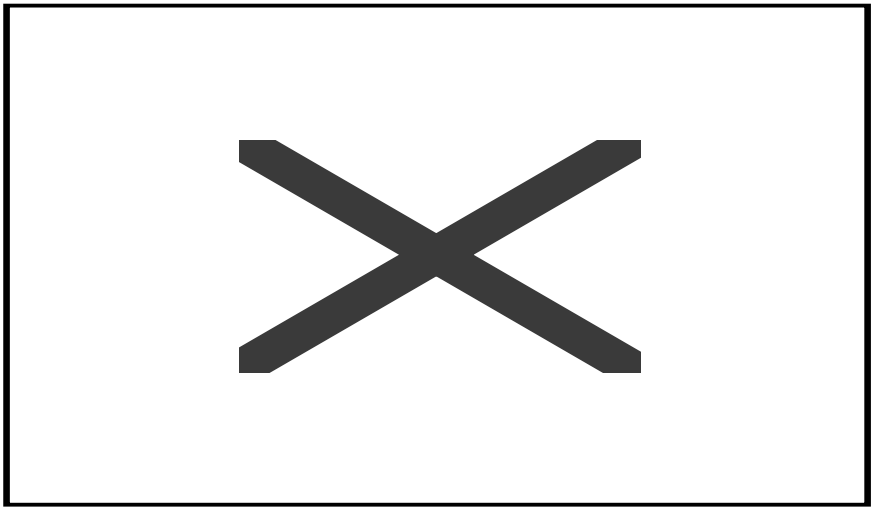


Рис. 2.1. Пример диалогового окна. Вычисление суммы переменных

Transform обеспечивает преобразование данных. Эта команда также имеется только в меню окна редактора данных.

Statistics обеспечивает доступ и реализацию статистических методов анализа данных; в 9-й и 10-й версиях SPSS ее заменяет команда **Analyze.**

Graphs позволяет получить графическое представление данных.

Utilities содержит обслуживающие программы.

Window обеспечивает переключение окон.

Help содержит справочную информацию.

При работе с графиками и мобильными таблицами (**Pivot tables**) появляются меню специального назначения.

Приведенные команды представляют далеко не полное описание меню, а лишь наиболее используемую его часть. Как принято в современном интерфейсе программ, под меню в верхней части окна в обычном режиме работы находится строка с панелью инструментов – кнопок, с которыми связаны различные действия пакета. При движении курсора по этим кнопкам ***на статусной строке*** внизу во внешней части экрана высвечиваются ***сведения о назначении кнопки***.

2.3.2. Статусная строка

Статусная строка показывает текущее состояние данных и процесса счета, например:

Transformations pending – задержка преобразований (например, если за преобразованиями не следует команда `EXECUTE` или статистическая процедура).

Weight on – данные взвешены;
Split on – данные для проведения расчетов разбиты на группы;
Filter on – включена временная выборка данных и др.

2.3.3. Ввод данных с экрана

При загрузке пакета появляется таблица, похожая на электронные таблицы. Данные можно вводить непосредственно с экрана. По умолчанию переменные (столбцы матрицы) будут иметь имена `var0001, var0002` и т. д. Для изменения имен переменных, назначения их типов и расширенных названий (меток) можно щелкнуть мышкой дважды на существующих названиях столбцов. При этом открывается окно диалога для описания переменной. Можно также применить команду `RENAME VARIABLES`, синтаксис которой мы не приводим из-за ее достаточно редкого использования.

Ниже будут приведены команды `VARIABLE LABELS, VALUE LABELS, MISSING VALUES`, осуществляющие основные функции этого диалога.

2.4. Режим диалога и командный режим

Самый простой, но достаточно медленный способ работы в пакете – использование диалоговых окон для формирования команд. Окна появляются на экране при вызове названия команды из меню. Диалоговые окна имеют многоуровневую структуру, соответствующую системе вложенных подпрограмм, реализующих данную команду. Последовательно вызываемые, они позволяют задать весь набор параметров, необходимых для осуществления задуманного статистического исследования или преобразования данных.

Диалоговый способ удобен тем, что в окне всегда присутствует подсказка о параметрах процедуры преобразования или анализа данных. Параметры вводятся в жестко закрепленные поля, поэтому ошибки в нем практически невозможны.

Важно то, что при диалоговом задании команды и ее параметров пакет программно формирует текст выполняемой команды и при желании его можно запомнить в командном файле. Для этого необходимо выполнять сформированную команду, используя в диалоговом окне не «кнопку» **Ok** – непосредственное исполнение команды, а кнопку **Paste** – дописать команду в файл **Syntax** (рис. 2.1). В результате команда будет записана в конце командного файла. В пакете предусмотрена возможность выполнения всех команд, записанных в командный файл синтаксиса и автономное выполнение отдельной команды или подмножества команд. Для выполнения нужных команд необходимо выделить их текст в окне синтаксиса и запустить их на выполнение с помощью специальной кнопки  . Таким образом, диалоговый режим позволяет составлять последовательность команд и це-

лые законченные программы, не зная языка программирования, предусмотренного в пакете.

Использование в анализе исключительно диалоговых окон удобно только для новичка. Для эффективной работы в пакете необходимо знать и понимать язык программирования SPSS. Написание программ на языке пакета предпочтительнее при достаточно большом объеме преобразований данных. Исследователь должен иметь перед глазами программу выполненных действий для уверенности в правильности результата. Кроме того, появляется возможность копирования и редактирования текста программы. Программы позволяют в любой момент повторить расчеты, упрощают контроль и поиск ошибок преобразования данных. Они легко модифицируются для решения других задач.

Впрочем, важно оптимальное сочетание диалоговых окон и языка.

2.4.1. Командный режим работы с пакетом. Основные правила написания команд на языке пакета

– Команды, имена переменных, ключевые слова могут вводиться большими или маленькими буквами.

– Список последовательно расположенных в активном файле переменных можно задавать в тексте команды, пользуясь сокращением: <первая переменная TO последняя переменная>.

– Ключевые слова могут усекаться до первых трех символов.

– В метках переменных и значений учитывается регистр буквы.

– Команды могут начинаться с любой позиции и должны кончаться символом конца команды – точкой.

– Продолжение команды начинается с любой позиции строки.

– Подкоманды разделяются слэшами (/).

– Имена файлов заключаются в апострофы или кавычки.

– Символ «*» в начале строки означает, что на данной строке расположен комментарий, комментарий также должен заканчиваться точкой.

Редактирование программ осуществляется по правилам, схожим с правилами, принятыми в распространенных редакторах системы **Windows**.

Напомним, что, набрав программу в окне **Syntax**, ее можно запустить полностью или частично (выделив блок), нажав кнопку  либо воспользовавшись пунктом меню **Run**. При выделении команд для запуска необходимо внимательно следить, чтобы выделение начиналось с первого символа команды и заканчивалось точкой – признаком конца команды.

Среди инструментов в окне редактирования файла **Syntax** имеется ***кнопка***  ***для вызова подсказки*** – схемы подкоманд команды. Подсказку можно получить, установив курсор на команде и щелкнув левой клавишей мыши указанную кнопку.

Среди команд SPSS условно можно выделить три основных типа команд: описания данных, преобразования данных и команды, выполняющие статистические процедуры.

Команды описания данных позволяют задать имя переменной и метки для ее значений, неопределенные значения, типы переменных, формат выдачи и др.

Команды преобразования данных предназначены для вычисления новых переменных и модификации имеющихся. Выполнение этих команд не вызывает непосредственного преобразования данных до тех пор, пока не будет выполнена следующая команда с участием новой переменной, либо после запуска команды **Execute**. Поэтому, если сразу после выполнения команды преобразования данных просмотреть файл данных, то в нем не будет никаких изменений. Например, чтобы обнаружить новую переменную, записанную в конец файла данных, необходимо выполнить **Execute**, либо какие-то расчеты с участием этой переменной. Такая организация необходима для уменьшения числа обращений к данным на магнитном носителе.

Статистические процедуры предназначены для получения статистик, оценки параметров моделей, получения графиков и др.

Деление это условно. Например, статистические программы также могут вычислять новые переменные, а команды агрегирования данных, как мы увидим ниже, вычисляют статистики для групп объектов. Кроме того, имеются команды управления данными, манипуляции файлами и другие команды, не вписывающиеся в эти три группы команд.

2.4.2. Порядок выполнения команд

При выполнении команд необходимо, чтобы данные для них были определены. Например, если заранее не вычислена переменная `x`, нельзя запустить команды

```
COMPUTE y = x+1.
DESCRIPTIVE var = y.
```

Команда `COMPUTE` не может вычислить переменную `y`, так как отсутствует переменная `x`, а команда `DESCRIPTIVE` не будет выполнена, так как будет отсутствовать `y`.

2.4.3. Команды вызова `GET` и сохранения данных `SAVE`

Хотя для вызова файла данных удобнее непосредственно использовать меню, команда полезна при многократном использовании данных, или использовании части данных.

Примеры

```
GET FILE = 'D:\mydir\city' /KEEP = x1 to x10, x15.
```

```
GET FILE = 'D:\mydir\city' /DROP = Z1, z5, z10.
```

Ключевое слово KEEP в первом примере говорит о том, что будут использованы лишь переменные начиная с x1 до x10 и x15.

Ключевое слово подкоманды DROP во втором примере исключает из анализа Z1, z5, z10.

Сохранение данных производится командой SAVE

```
SAVE FILE = 'D:\mydir\city' /KEEP = x1 to x10,
x15 /compressed.
```

Подкоманда /COMPRESSED необходима для сжатия информации. Подкоманды KEEP и DROP применяются для сохранения и отбрасывания части переменных.

2.4.4. Основные команды описания данных

Команда VARIABLE LABELS назначает ***переменным*** метки (расширенные текстовые наименования), которые используются при оформлении листингов.

```
VARIABLE LABELS V8 'Пол'
                V9 'Возраст'.
```

Здесь V8 – переменная, а «Пол» – метка.

Синтаксис: за именем переменной указывается в апострофах ее текстовое наименование – метка. Вы должны помечать каждую переменную отдельно. Максимальная длина метки 255 символов.

Команда VALUE LABELS назначает ***значениям*** переменных расширенные текстовые наименования – метки, которые используются при оформлении листингов

```
VALUE LABELS     V1   1 "расчет на свои силы"
                      2 "пределы"
                      3 "помощь"/
                 V8   1 "МУЖЧИНА"
                      2 "ЖЕНЩИНА"/
               x1 TO x10 1 "да" 2 "нет" 3 "не знаю".
```

Синтаксис: за именем переменной (например, V8) или списком переменных (например, x1 to x10) и кодом значения (например, 2) в апострофах следует метка («ЖЕНЩИНА» или «НЕТ» соответственно). Максимальная длина метки для значений переменных не больше 60 символов. Назначения меток для значений разных переменных должны разделяться слэшами, в качестве образца используйте приведенный пример.

Команда ADD VALUE LABELS делает то же, что и команда VALUE LABELS, но если VALUE LABELS при повторном запуске замещает все ранее назначенные метки указанных в ней переменных, команда ADD VALUE LABELS назначает метки только указанным кодам.

Команда MISSING VALUES. На практике приходится обрабатывать информацию с пропущенными данными. При кодировании неопределенных данных (таких как ответы «не знаю», отказ от ответа) необходимо выбрать символы или цифры – коды отсутствующих значений и сообщить пакету, что они соответствует пропущенным данным. Это делается командой MISSING VALUES, которая сохраняет в справочной информации файла данных объявленные пользователем коды для неопределенных значений переменной или списка переменных. В дальнейшем в статистических процедурах и при преобразовании данных эти коды обрабатываются специальным образом. Для переменной возможно назначение до трех неопределенных кодов либо интервала кодов и одного (но не более) отдельного кода.

Примеры

```
  MISSING VALUES X Y Z(-1)/ R(9, 99, 999)/ S1 TO
S20(999 thru 100000)/ SEX (9).
  MISSING VALUES v2 (Lowest thru –1, 99)/ v10
(-1, 900 THRU Highest).
```

В указанном выше примере –1 назначается кодом неопределенного значения для X, Y и Z; 9, 99, 999 – для R; от 999 до 100 000 – коды неопределенности переменных от S1 до S20; 9 – для SEX; от минимального кода до –1 и 99 – для v2; –1 и коды от 900 до максимального – для v10.

Ключевое слово thru определяет интервал кодов; Lowest, Highest – минимальный и максимальный коды соответственно. Возможны сокращения этих ключевых слов до 2 букв (th, lo, hi).

В команде указывается список переменных (разделять символом «/» необязательно), у которых может встретиться неопределенное значение и за которым в круглых скобках указан объявленный код. Объекты с такими значениями переменных при выполнении многих пакетных процедур просто исключаются из рассмотрения.

Неопределенные значения, описанные командой MISSING VALUES, называются пользовательскими неопределенными значениями. Однако и в процессе счета могут возникнуть ситуации, когда невозможно осуществить преобразование данных: деление на 0; корень из отрицательного числа; в вычисления попал код отсутствующего значения; при чтении данных нет совпадения типа (число, символ) данных и т. д. Пакет таким неопределенным значениям присваивает специальный системный код, который в дан-

ных изображается точкой. ***Системный код неопределенности*** в процедурах и командах обозначается ключевым словом SYSMIS.

Объявление пользовательских неопределенных значений можно отменить командой MISSING VALUES с пустыми скобками:

```
MISSING VALUES X Y Z() R()/ S1 TO S20()/ SEX().
```

2.5. Основные команды преобразования данных

Для преобразования данных в меню окна редактора данных имеется пункт **Transformations.** Тексты команд можно получать, пользуясь этим пунктом.

Преобразования в анализе данных – одна из самых трудоемких частей работы. Специалист, освоивший технику преобразования данных, имеет существенный шанс для получения содержательных результатов. На практике в большинстве случаев можно обойтись следующими командами:

COMPUTE – арифметические операции над переменными;

IF – условные арифметические операции над переменными;

RECODE – перекодирование переменных;

COUNT – подсчет числа заданных кодов в списке переменных.

2.5.1. Команды COMPUTE и IF

Команда COMPUTE вычисляет новую переменную или заменяет существующую.

Например, для приведенной в Приложении 1 анкеты требуется рассчитать, сколько лет респондент проживал за Уралом (см. анкету, Приложение 1).

```
COMPUTE  Y = V15 + V16 + V17.
```

В матрице данных создается новая переменная Y.

В команде указывается имя создаваемой переменной, за которым после обязательного знака « = » следует арифметическое выражение. Создаваемая переменная может быть функцией от других переменных.

После выполнения команды в матрицу данных в активный файл будет дописан столбец с новым именем. Если какой-либо член арифметического выражения не определен, то результатом будет системный код отсутствующего значения (SYSMIS). Например, если в команде COMPUTE Y = X-5/Z. значение переменной X не определено в соответствии с командой MISSING VALUES или имеет системный код неопределенности или, если Z = 0, то переменной Y присваивается системный код неопределенности SYSMIS.

Команда IF при выполнении указанного в ней условия создает новые переменные или заменяет существующие переменные арифметическими выражениями.

```
IF (R > D OR (R >= E AND B > 0)) STATUS = 1.
IF (STATE = 'IL') COST = COST + 0.07 * COST.
```

В ней указывается логическое выражение, за которым следует арифметическое присвоение. Логическое выражение должно быть заключено в круглые скобки. Логическое выражение в команде IF может быть ложно не только в результате выводов с позиций формальной математической логики, но в случае, если в выражении встретилось неопределенное значение. Для оператора присваивания в случае неопределенных значений переменных действуют те же правила, что и в команде COMPUTE.

В качестве логического выражения может быть и обычная числовая переменная или числовая константа. Считается, что она принимает значение «истина», если она равна 1, в противном случае ее значение – «ложь».

Область действия IF – один оператор присваивания, приведенный в тексте команды.

Пусть, например, требуется вычислить переменную D, характеризующую отклонение веса (W) от нормального (для мужчин (код значения переменной P «пол» равен 1) нормальный вес должен быть равен величине роста минус 100, для женщин (p = 2) – величине роста минус 105).

```
IF (P = 1) d = W - (R - 100).
IF (P = 2) d = W - (R - 105).
```

В результате выполнения этих команд появляется переменная D, которая вычисляется в зависимости от значений переменной P.

В диалоговом окне команд содержится подробный список функций и операторов. Чтобы читатель имел представление о возможностях команд IF и COMPUTE, ниже мы представим их основные типы.

2.5.1.1. Основные функции и операторы команд COMPUTE и IF

Арифметические операторы +, -, *, / в этих командах употребляются обычным порядком, две звездочки ** означают возведение в степень.

Результатом логической операции будет 1, если логическое выражение истинно, и 0, если выражение ложно (логическое выражение (v9 > 30) равно 1, если v9 > 30, и равно 0, если v9 ≤ 30).

Допустимы операторы сравнения <, <=, <, <=, ~=, где последний оператор означает «не равно» и логические операторы ~ – отрицание (not), & – логическое «и» (and) и логическое «или» | (or).

При вычислении логического выражения, если порядок выполнения не задан скобками, сначала выполняются арифметические операции, затем сравнения, затем логические операции. Приоритетность выполнения операций естественна – так она обычно определяется в математике и языках

программирования. Но следует заметить, что операции сравнения находятся на одном уровне. В частности значение выражения (5 > 3 > 2), будет равно 0 («ложь»), так как в соответствии с порядком выполнения операций в этом выражении (5 > 3 > 2) = ((5 > 3) > 2) = (1 > 2) = 0!

Наряду с арифметическими операторами в арифметических выражениях могут использоваться логические выражения, что позволяет достаточно компактно осуществлять преобразования данных:

```
COMPUTE  x = (v9 > 30) + v10 > x + z.
```

Эта хитроумная команда превращает вначале выражение (v9 > 30) в 0 или 1 в зависимости от его истинности, затем производит вычисления левой ((v9 > 30) + v10) и правой (x + z) частей неравенства и в зависимости от результата сравнения присваивает переменной x значение 0 или 1.

Кроме того, имеется возможность использовать:

Арифметические функции, такие как: ABS – абсолютное значение, RND – округление, TRUNC – целая часть, EXP – экспонента, LN – натуральный логарифм, и др. Например,

```
COMPUTE LNv9 = LN(V9).
```

Переменной LNv9 присваиваются логарифмы значений переменной V9.

Статистические функции: SUM – сумма, MEAN – среднее, SD – стандартное отклонение, VARIANCE – дисперсия**,** MIN – минимум и MAX – максимум. Например, команда

```
COMPUTE S =  MEAN (d1 to d10).
```

вычисляет переменную, равную среднему валидных (т. е. определенных) значений переменных d1, …, d10.

Функции распределения, например:

CDF.CHISQ(q, a) – распределения хи-квадрат, CDF.EXP(q, a) – экспоненциального распределения, CDF.T(q, a) – Стьюдента, и др. (q – аргумент функции распределения, a – параметр соответствующего распределения). Команда

```
COMPUTE Y = CDF.T(X, 10).
```

Эти функции могут быть использованы для проверки предположения о виде распределения переменной. Например, если мы для расчета переменной используем функцию распределения Стьюдента с 10 степенями свободы и построим значения от переменной X, которая распределена по Стьюденту с 10 степенями свободы, то получим переменную Y, равномерно распределенную на отрезке (0, 1).

Таким образом, если есть подозрение, что X имеет именно такое распределение, то можно проверить это предположение, построив переменную Y и проверив ее на равномерность распределения на отрезке (0, 1).

То же самое можно предпринять для проверки других видов распределений.

Обратные функции распределения, например:

IDF.CHISQ(p, a) – обратная функция распределения (по сути дела, квантиль) хи-квадрат, IDF.F(p, a, b) – квантиль распределения Фишера, IDF.T(p, a) – квантиль распределения Стьюдента, и др. (p – вероятность, a и b – параметры соответствующего распределения). Например,

```
COMPUTE  Z = IDF.CHISQ(X, 10).
```

вычисляет квантиль порядка X распределения *хи-квадрат* с 10 степенями свободы. Такие функции полезны для вычисления значимости статистик для подмножеств исследуемого множества, например значимости отклонения среднего возраста по городам региона, в котором произведен сбор данных.

Датчики случайных чисел, например:

RV.LNORMAL(a, b) – датчик лог-нормального распределения**.**

RV.NORMAL(a, b) – датчик нормального распределения,

RV.UNIFORM(a, b) – датчик равномерного распределения (a, b – параметры соответствующего распределения).

Функция, дающая значения переменной на предыдущем объекте LAG. Пример использования (см. рис. 1.1, данные «Проблем и жалоб»):

```
COMPUTE age1 = LAG(age).
COMPUTE age2 = LAG(age, 3).
EXECUTE.
```

Указанное преобразование осуществляет сдвиг информации, показанный в табл. 2.1. В скобках второй параметр задает длину лагового сдвига.

Таблица 2.1

Сдвиг, произведенный функцией LAG (данные «Проблем и жалоб»)

N Анкеты	Пол (Sex)	Возраст (Age)	Возраст (Age1)	Возраст (Age2)
1	1	20		
2	1	25	20	
3	2	34	25	
4	1	18	34	20
.	.	.		

Функция полезна для анализа временных рядов, при анализе анкетных данных – для поиска повторов объектов и других вспомогательных операций.

Логические функции:

RANGE(v, a1, b1, a2, b2, …) принимают значение 1, если значение V попало хотя бы в один из интервалов [a1, b1], [a2, b2], и 0 – в противном случае.

ANY(v, a1, a2, …) принимают значение 1, если значение V совпало хотя бы с одним из значений a1, a2, … и 0 – в противном случае.

Кроме того, в пакете имеются ***строчные функции, функции обработки данных типа даты и времени***.

2.5.1.2. Работа с неопределенными значениями

Вообще говоря, если в арифметическом выражении встретится переменная с неопределенным значением, результат не будет определен, однако значения выражения 0***«неопределенное значение»** (нуль, умноженный на неопределенное значение) и 0/ **«неопределенное значение»** (нуль, деленный на неопределенное значение) приравниваются к нулю.

2.5.1.3. Функции для неопределенных значений

VALUE – функция игнорирования назначения пользовательского неопределенного значения;

MISSING – логическая функция для обнаружения пользовательского или системного отсутствующего значения; ее значение – истина (единица), если значение аргумента не определено, ложь (нуль) – в противном случае;

SYSMIS – то же, но только для системных неопределенных значений;

NMISS – подсчитывает число неопределенных значений в списке аргументов;

NVALID – число определенных значений в списке аргументов.

2.5.1.4. Работа с пользовательскими неопределенными значениями

В матрице данных по вопросу о Курильских островах переменные V15, V16, v17 означают время проживания в Западной Сибири, Восточной Сибири и на Дальнем Востоке. Допустим, для удобства проведения текущих расчетов нулевые коды этих переменных объявлены неопределенными:

```
MISSING VALUES V15, V16, v17 (0).
```

Тогда вычисление времени проживания за Уралом вычисляется командой

```
COMPUTE Y = V15 + V16 + v17.
```

приведет в большинстве случаев к неопределенным значениям Y.

В этом случае функция VALUE позволяет работать с пользовательскими неопределенными значениями без отмены объявления о неопределенности кодов, как с определенными:

```
COMPUTE Y = VAL(V15) + VAL(V16)+VAL(V17).
```

2.5.1.5. Работа с функциями MISSING и SYSMIS.

В RLMS [13] (Российском мониторинге экономики и здоровья), волна 2, имеется переменная BO2a – ответ на вопрос «Сколько времени в течение последних 7 дней Вы потратили на работу … ?», причем коды 997, 998, 999 соответствуют ответам «ЗАТРУДНЯЮСЬ ОТВЕТИТЬ», «ОТКАЗ ОТ ОТВЕТА», «НЕТ ОТВЕТА». Имеет смысл эти коды объявить пользовательскими неопределенными, а системные неопределенные коды перекодировать в 0. Делается это следующими командами:

```
MISSING VALUES BO2a (997, 998, 999).
If (SYSMIS(BO2a)) BO2a = 0.
EXECUTE.
```

Аналогичным путем в других обстоятельствах можно употребить и функцию MISSING.

2.5.2. Команда RECODE

Назначение команды: перекодирование значений переменной в задаваемые. Формат команды:

```
RECODE V9 (0 THRU 25 = 1)(26 THRU 45 = 2)(ELSE = 3).
```

или

```
RECODE V9 (0 THRU 25 = 1)(26 THRU 45 = 2)(ELSE = 3)
INTO W9.
```

В первом случае будут заменены новыми кодами исходные значения переменной V9, и ее первоначальное содержимое будет потеряно на все время сеанса работы с пакетом. Во втором случае эта переменная сохранится, так как результат перекодирования заносится в новую переменную W9.

В команде указывается переменная или список переменных со спецификациями в круглых скобках. Перекодируемые переменные в списке разделяются слэшами (/). По этой команде значения перечисленных переменных в указанных в скобках пределах будут заменены числами, следующими за знаком равенства.

Ключевое слово INTO указывает, в какую переменную (список переменных) переслать результат перекодирования, при этом соответствие ме-

жду исходным списком переменных и переменными результата устанавливаются естественным образом.

Команда RECODE перекодирует данные исключительно в соответствии со списками старых и новых значений и не изменит переменную назначения, если в перекодируемой переменной не нашлось значений для перекодирования.

Список переменных можно задать через ключевое слово TO, но всегда следует указывать переменные в том порядке, в каком они следуют слева направо в матрице данных.

Ключевые слова для задания входных значений переменных в команде RECODE:

LOWEST или LO – наименьшее значение переменной;

THRU или THR – значения переменной из указанного диапазона;

HIGHEST или HI – наибольшее значение переменной;

MISSING – отсутствующее значение, определяемое пользователем;

SYSMIS – отсутствующее значение, определяемое системой;

ELSE – все неспецифицированные значения (не включаемые в SYSMIS).

В новой переменной W9, если ее специально предварительно не заполнить информацией, для всех объектов до выполнения команды находятся системные коды неопределенности. Тогда результатом перекодирования будет заданный код или системный код неопределенности SYSMIS**.** Однако, если вместе с ключевым словом ELSE употребить слово COPY, то значения переменной V9, не включенные в списки перекодирования, будут скопированы в новую переменную.

```
RECODE educat (1 = 2)(2 = 1)(ELSE = COPY) INTO edu-
cat1.
```

Без (ELSE = COPY) в переменную educat1 будут внесены лишь перекодированные значения.

Среди списка значений для переменной, имеющей неопределенные значения, могут стоять слова MISSING и SYSMIS.

```
RECODE K9 TO K12 (0 THRU 25 = 1)(MISSING = 10)
(SYSMIS = 5).
```

Команда RECODE позволяет также интервалировать, группировать значения (рис. 2.1).

```
RECODE V11 V13 (8, 9, 2, 4, 7 = 1) (ELSE = 2).
```

```
   V11 V12  V13              V11  V12  V13
```

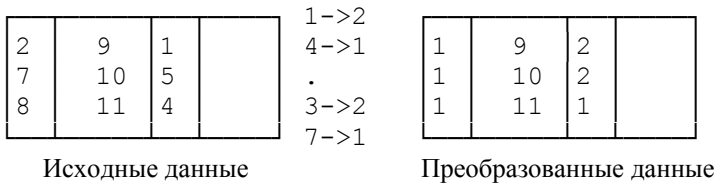


Рис. 2.1. Перекодирование данных

Что происходит при этом с матрицей данных? Как видно из приведенной выше схемы, происходит замена значений в соответствии с приведенными в команде списками значений.

Рассмотрим примеры перекодирования кодов неопределенности. При ответах на вопросы анкеты «Курильские острова» (Приложение 1) кто-то не ответил на первый вопрос, кто-то сказал «Затрудняюсь». В первом случае переменная принимает значение кода неопределенности, во втором равняется 4. Объединим этих респондентов. Это можно осуществить командой

```
RECODE V1 (SYSMIS = 4).
```

и таким образом перекодировать системный код неопределенности в код 4. Можно провести обратную операцию:

```
RECODE V1 (4 = SYSMIS).
```

Этой командой код 4 перекодируется в системный код неопределенности. Но при обработке данных по этому признаку объекты, для которых значение V1 было когда-то равно 4, будут исключены из статистической обработки.

Тот же эффект можно получить, воспользовавшись командой

```
MISSING VALUES V1(4).
```

При этом таблица данных не изменится; но во внутренней для SPSS информации сохранятся сведения о том, что указанный в данной команде код является пользовательским кодом неопределенности для `V1`.

В SPSS запрещено писать `MISSING` справа от знака равенства, т. е. команда

```
RECODE V1(4 = MISSING).
```

недопустима!

Для выполнения команды `RECODE` с созданием новой переменной используется ключевое слово `INTO`:

```
RECODE V11 (8, 9, 2, 4, 7 = 1) INTO W11.
```

При таком использовании команды в большинстве случаев необходимо перечислять все принимаемые исходной переменной значения, поскольку неуказанные значения переходят в системные неопределенные значения в переменной W11.

2.5.3. Команда COUNT

Команда COUNT подсчитывает для каждого объекта (для строки матрицы) число появлений указанных в ней кодов в заданном списке переменных и размещает результат в новую переменную или заменяет содержимое существующей.

В команде указывается имя переменной, куда будет заноситься результат подсчета, затем, после обязательного знака «=», приводится список переменных, для которых нужно вести подсчет, и далее в круглых скобках приводится список значений переменных, число которых следует пересчитать. Значения строковых переменных должны быть заключены в апострофы. Ключевое слово SYSMIS используется для подсчета системных отсутствующих значений; MISSING позволяет подсчитать все отсутствующие значения – и пользовательские, и системные. Команда допускает также ключевые слова LOWEST, HIGHEST и THRU. В отличие от команды RECODE команда подсчета значений в переменных при их отсутствии присваивает 0 в результирующую переменную.

Пример. Пусть нам необходимо вычислить число разумных вариантов решения проблемы островов (неальтернативный вопрос 7 анкеты о Курильских островах), а затем подсчитать число ответов на все неальтернативные вопросы анкеты.

```
  COUNT nofvari = v7s1 to v7s7 (1 thru 11)/
  nofans = v3s1 to v3s8 (1 thru 8) v5s1 to v6s8
(1 thru 8).
```

Пример. По результатам сессии (объекты – студенты, переменные – результаты экзаменов по математике (M), микроэкономике (E), и социологии (S)) необходимо создать переменную M45, в которой будет число пятерок и четверок, встречающихся в перечисленных переменных. У троечников и двоечников M45 примет значение 0. Значения новой переменной M45 будут изменяться от 0 до 3. Тройка будет присвоена, если студент получал только 4 и 5 по всем 3 дисциплинам.

```
COUNT M45 = I M E S (4,5).
```

2.5.4. Условное выполнение команд

Команды DO IF, ELSE IF, ELSE и ENDIF используются для преобразования переменных на подмножестве объектов, выбираемых по условию сразу несколькими командами. Между DO IF и ENDIF может быть написана целая программа. После ENDIF отбор по условию не действует.

Пусть, например, в файле «Курильские острова» требуется проинтервалировать возраст (v9), т. е. создать переменную, значениями которой будут номера соответствующего возрастного интервала. При построении интервалов должна учитываться разница в пенсионном возрасте для мужчин и женщин (табл. 2.2). Таким образом, при построении интервалов используется, также, переменная «пол» (v8).

Таблица 2.2

Интервалы для мужчин и женщин

Интервалы возраста	1	2	3	4	5
Мужчины	до 18	До 33	до 45	До 60	> 60 лет
Женщины	до 18	До 33	до 45	До 55	> 55 лет

```
DO IF (v8 = 1).
 RECODE v9 (LO THRU 18 = 1)(18 THRU 33 = 2)(33 THRU
  45 = 3)(45 THRU 60 = 4)(60 THRU hi = 5) INTO w9.
ELSE IF (v8 = 2).
 RECODE v9 (LO THRU 18 = 1)(18 THRU 33 = 2)(33 THRU
  45 = 3)(45 THRU 55 = 4)(55 THRU HI = 5) INTO w9.
END IF.
```

Здесь для мужчин в переменной w9 получаются одни интервалы возраста, для женщин – другие. Если бы не было неопределенных значений у переменной v8, можно было бы вместо «ELSE IF (v8 = 2).» использовать просто «ELSE.».

Заметим, что команды RECODE и COUNT непосредственно не могут выполняться на подмножествах объектов, но с помощью команд DO IF и END IF можно организовать для необходимой подвыборки объектов их выполнение.

Напомним, что команды, запущенные без команды EXECUTE, накапливаются в памяти, но не выполняются (**Transformations pending** в статусной строке). Так, команды IF, COMPUTE, COUNT, RECODE преобразуют данные не сразу после их запуска, а только после запуска команды EXECUTE**.** Поэтому в случае ошибки в командах, написанных между DO IF и END IF, успевает выполниться и попасть в память только команда DO IF. После исправления ошибки и повторного выполнения программы за-

пущенных команд DO IF оказывается больше, чем END IF, и появляется сообщение о новой ошибке. Для того чтобы справиться с этой ситуацией, после исправления ошибки, перед повторным запуском программы, следует выполнить отдельно команду

CLEAR TRANSFORMATIONS., которая очистит память от невыполненных команд.

2.5.5. Команда RANK

Анализируя доходы населения, мы можем работать непосредственно с доходами, вычисляя средние, корреляции и др. Можем изучать иерархию семей или индивидуумов по этой переменной. Для этого нужно перейти к порядковым номерам объектов, упорядоченным по доходам. Такие порядковые номера называются рангами. Например, иерархию семей можно изучать, определив для каждой семьи долю (процент) семей, которые беднее ее. Наконец, можно разбить семьи по уровню доходов на равные 5 частей (квинтили) или на 10 частей (децили). Ранги, процентили, *n*-тили суть преобразованные в соответствии с ранжированием объектов переменные.

Команда RANK весьма полезна, когда нужно перейти от исходных значений любых количественных переменных к их рангам, процентилям, децилям и квинтилям и др., а можно перекодировать переменную в соответствии с нормальным распределением.

Пусть нам необходимо получить переменные «ранг по доходам», «процентили по доходам» и «квинтильные группы по доходам». («Курильские» данные). Команда RANK создаст нам нужные переменные:

```
  RANK VARIABLES = v14 (A) /RANK into rangv14/NTILES
(5)into v14_5 /PERCENT percv14/PRINT = YES /TIES = MEAN.
  VARIABLE LABELS rangv14 "ранг по доходам"/
  v14_5 "квинтильные группы по доходам"/
  percv14 "процентили по доходам".
```

Подробнее о команде RANK см. в [1. С. 115].

2.5.6. Отбор подмножеств наблюдений

Для выбора в матрице данных в диалоговом режиме подмножества наблюдений необходимо использовать в главном меню **Data** окно **Select Cases.**

После выполнения этих команд появляется окно диалога, в котором пользователь задает условия отбора данных. Невыбранные объекты будут исключены из сеанса работы или временно отфильтрованы. Имеется возможность организовать случайную выборку данных заданного объема, например, выбрать 10 % случайных объектов из множества данных. Вся ра-

бота пакета будет осуществляться для отобранных объектов, пока действие **Select Cases** не буде аннулировано**.**

Если необходимость во временной выборке отпала, нужно снова обратиться к этому же пункту меню и указать, что необходимы все объекты (**All Cases**).

Если мы хотим, чтобы пакет сохранил наши действия в диалоговом режиме в виде соответствующих команд в файле синтаксиса, необходимо запустить их на выполнение с использованием диалогового окна **Paste**. Это приведет к появлению в конце текста файла синтаксиса целой серии следующих команд:

```
USE ALL.
COMPUTE filter_$ = (v8  =  1).
VARIABLE LABEL filter_$ 'v8  = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

Как видно из сгенерированного SPSS текста, в случае использования условия для временной подвыборки объектов программа выборки создает переменную фильтра (**filter_$**) и использует команду `FILTER BY filter_$.`

Таким образом, можно для временной выборки объектов самим написать программу, создающую переменную фильтра, и выполнить. Например, для выбора мужчин в нашем учебном массиве можно воспользоваться командой

```
FILTER BY V8.
```

Это возможно, так как для мужчин в переменной V8 указан код 1. А если хотим выбрать женщин таким же способом, то нужно заменить кодировку для женщин на 1, а для мужчин на любое другое число.

Для отмены фильтра необходимо запустить команду `FILTER OFF`**.**

Следует внимательно следить в процессе работы за состоянием ваших фильтров и вовремя отменять фильтрацию. В статусной строке, если включен фильтр, горит сообщение **Filter on**.

Если необходимо отдельно сохранить данные только для отобранных объектов, нужно в команде `SAVE` использовать подкоманду `/UNSELECTED DELETE`:

```
SAVE FILE = 'D:\mydir\city' /KEEP = x1 to x10, x15
/UNSELECTED DELETE/COMPRESSED.
```

В подкоманде KEEP указываются сохраняемые переменные (столбцы), а подкоманда UNSELECTED DELETE позволяет сохранять только отфильтрованные объекты (строки матрицы).

Если необходимо исключить некоторые объекты из матрицы данных на все время данного сеанса работы с пакетом, диалог позволяет выполнить последовательность команд такого типа:

```
USE ALL.
SELECT IF(v8 = 1).
EXECUTE .
```

Можно обойтись и одной командой SELECT IF (v8 = 1). Строки, в которых переменная V8 = 0, становятся недоступными.

Обратим еще раз внимание на то, что в результате применения команды SELECT IF невыбранные объекты для данного сеанса работы теряются полностью. Если потеря данных имела смысл только для этого сеанса, то либо не следует сохранять исходную информацию после окончания исследований, либо сохранить ее под другим именем.

2.5.7. Команда SPLIT FILE

Нередко возникает необходимость получить однотипные таблицы для различных значений некоторой переменной (переменных) и даже сравнивать их. С этой целью предусмотрена команда SPLIT FILE. Ее удобнее запускать из меню редактора данных, нежели из программы. Команда SPLIT FILE требует предварительной сортировки данных по переменным разбиения. В ней указываются переменные разбиения выборки, а также цель расщепления – получение независимых выдач для различных групп объектов (ключевое слово SEPARATE) или сравнение данных по группам (LAYERED). В последнем случае для большинства статистических программ выдачи по группам объединяются в единую таблицу.

Например, расщепление наших учебных данных выборки по полу с целью сравнения описательных статистик, получаемых для групп, можно сделать программой:

```
SORT CASES BY v8.
SPLIT FILE LAYERED BY v8.
DESCRIPTIVES VARIABLES = v9 v14.
```

Команда DESCRIPTIVES получает описательные статистики переменных. В табл. 2.3 (здесь и далее для большинства таблиц использованы машинные выдачи) благодаря команде SPLIT результаты работы команды DESCRIPTIVES для разных групп по полу объединены в одну таблицу.

Таблица 2.3

Описательные статистики для групп, полученные при расщеплении данных для сравнения

V8 Пол		N	Minimum	Maximum	Mean	Std. Deviation
1 муж.	V9 Возраст	354	16,0	76,0	39,6	13,0
	V14 Ср. мес. душевой доход	341	21,0	1254,0	237,9	168,2
	Valid N (listwise)	335				
2 жен.	V9 Возраст	344	16,0	74,0	39,5	12,2
	V14 Ср. мес. душевой доход	324	50,0	1500,0	219,8	132,8
	Valid N (listwise)	317				

При получении результатов для отдельных групп программой

```
SORT CASES BY v8.
SPLIT FILE SEPARATE BY v8.
DESCRIPTIVES VARIABLES = v9 v14.
```

будут получены две отдельные таблицы.

2.5.8. Взвешивание выборки `WEIGHT`

Социологи достаточно часто работают с некорректными статистическими данными. К примеру, необходимо изучить социальные характеристики людей, занятых в правовых органах. Но известно, что в органах юстиции занято всего 2 % трудоспособного населения, и, если будет отобрано 500 человек, то среди них может оказаться только 10 занятых в органах юстиции. В этом случае данных будет недостаточно для формирования выводов. Поэтому социологи осознанно опрашивают большее число занятых в правовых органах, например 50 из 500. Иногда они рассчитывают целую половозрастную, отраслевую и т. д. таблицу, по которой решают, сколько человек в каждой социальной группе опросить. Это деформирует выборку; ее характеристики не соответствуют параметрам генеральной совокупности, т. е. она становится нерепрезентативной.

Чтобы уменьшить влияние деформированности выборки на результаты статистического анализа, применяют взвешивание объектов: группы, которые были искусственно уменьшены, выбираются с весовым коэффициентом, превышающим единицу. Обычно суммарный вес всех объектов равен числу объектов в рассматриваемом файле.

Пусть, например, опрошено 300 человек, из них 100 мужчин, 200 женщин. Однако из накопленного опыта известно, что в генеральной совокупности 50 % мужчин, 50 % женщин. Поэтому целесообразно для

всех статистических расчетов учитывать мужчину с весом 1,5, а женщину – с весом 0,75, тогда с учетом весов их воздействие на результаты расчетов по выборке будет выравнено. Суммарный вес равен 1,5 × 100 + 0,75 × 200 = 300.

Пример. Пусть переменная `SEX` содержит сведения по полу респондентов (1 – мужской, 2 – женский). Соответствующие веса будут назначены командами

```
RECODE SEX (1 = 1.5)(2 = 0.75) into wsex.
WEIGHT BY wsex
EXECUTE.
```

Вообще, если известно распределение объектов k групп в генеральной совокупности $p_1, \ldots, p_k$; получено частотное распределение $n_1, \ldots, n_k$, то i-й группе должен быть приписан вес $w_i = p_{i*N}/n_i$, где $N = \sum_i n_i$.

Назначить веса можно через меню редактора данных (**Data>Weight Cases**).

Замечание. Взвешивание – это не физическое повторение наблюдения. Если значение веса отрицательное или неопределенное (предварительно определенное как `SYSMIS`), то оно обрабатывается статистическими процедурами как вес, равный нулю.

Пример. Приемы использования команд описания и преобразования данных рассмотрим на примере обработки анкеты «Курильские острова».

Задача. На основании ответов на вопросы анкеты получить переменную, отражающую степень противостояния СССР и Японии.

Решением этой задачи, по мнению исследователя, может быть новая переменная, в зависимости от ответов респондентов имеющая значения 1, 2, 3, обозначающие:

1. Япония противостоит Союзу и Союз – Японии, т. е. противостояние взаимно.

2. Одна из сторон (Япония или Союз) против контактов.

3. Стороны не противостоят по отношению друг к другу.

За основу конструирования такой переменной возьмем ответы на третий вопрос анкеты «III. Как Вы считаете, что мешает подписать мирный договор между СССР и Японией? » с подсказками:

1 – нет настоятельной необходимости, отношения и без того нормальные.

2 – традиционное недоверие друг к другу в результате войн в прошлом.

3 – слабая экономическая заинтересованность Японии.

4 – разные политические симпатии СССР и Японии.

5 – нежелание Японии признать послевоенные границы с СССР.

6 – нежелание СССР рассматривать вопрос о спорных островах.

7 – другое (что именно).

8 – не знаю, затрудняюсь сказать.

Под ответы на вопрос III в матрице данных отведено восемь столбцов, наименованных V3S1 – V3S8; для заполнения ответов используется кодирование в виде списка. Анализируя ответы, строим переменную TP, соответствующую трем типам, определенным в задаче. Для этого построим вспомогательные переменные T1 и T2, являющиеся индикаторами того, что Япония противостоит СССР и СССР противостоит Японии соответственно.

Построить такие переменные можно, воспользовавшись командами

```
COUNT T1 = V3S1 to V3S7 (2,5)/
      T2 = V3S1 to V3S7 (2,6).
```

В результате выполнения команды переменной T1 присваивается либо 1 (когда в анкете была обведена одна из двух подсказок: 2 или 5); либо 2 (когда обведены обе подсказки) и 0, если респондент не обвел ни подсказку 2, ни подсказку 5. По аналогии заполнена значениями – количеством обведенных соответствующих подсказок – переменная T2.

```
COMPUTE OPPOS = 3.
IF (T1 > 0 | T2>0) OPPOS = 2.
IF (T1 > 0 & T2>0) OPPOS = 1.
EXECUTE.
VARIABLE LABELS OPPOS 'Степень противостояния СССР
и Японии'
T1 'Противостояние Японии' T2. 'Противостояние
СССР'.
VALUE LABELS OPPOS 1 'Взаимное' 2 'Одна из сторон'
3 'Нет противостояния'.
```

Здесь первая команда IF «затирает» значение 3 кодом 2, а вторая команда IF «затирает» код 3 кодом 1. Есть и другой путь решения этой задачи:

```
COUNT T1 = V3S1 to V3S7 (2,5) /
T2 = V3S1 to V3S7 (2,6).
RECODE T1 T2 (2 = 1).
COMPUTE OPPOS = 3 - (T1 + T2).
```

А можно и так:

```
COUNT T1 = V3S1 to V3S7 (2,5) /
T2 = V3S1 to V3S7 (2,6).
COMPUTE OPPOS = 3 - ((T1 > 0)+(T2 > 0)).
```

Таким образом, OPPOS= 1 для первого типа респондентов, OPPOS= 2 для второго, OPPOS = 3 – для третьего. Построенная переменная позволяет проводить в дальнейшем многосторонний анализ выделенных типов населения, например, возрастной структуры, социального положения, образования и т. д.

2.6. Операции с файлами

2.6.1. Агрегирование данных (команда AGGREGATE)

Нередко на основе собранных данных необходимо получить статистические сведения об укрупненных объектах. Для этого на базе исходной матрицы создается и обрабатывается новая матрица агрегированных данных.

Пример. На рис. 2.2 приведены анкетные данные обследования рабочих нескольких заводов. Объекты – информация о рабочих. В данных в виде переменных содержатся номер завода и номер цеха, в котором трудится респондент. На основе собранной информации вычисляется новый массив данных, в которых объектами являются цеха, признаками – статистические сведения по цехам, например, доля мужчин в цехе (в %), средний возраст и т. д. Соотношение двух массивов информации приведено на рис. 2.2.

Новую матрицу агрегированных данных, организованную по тому же принципу «объект – признак», что и исходная матрица, можно получить с помощью команды AGGREGATE.

```
AGGREGATE /OUTFILE = 'ZECH.SPS'/BREAK ZAVOD ZECH
       /PERCM = PLT(POL,2) /SRWOZR = MEAN(WOZR).
```

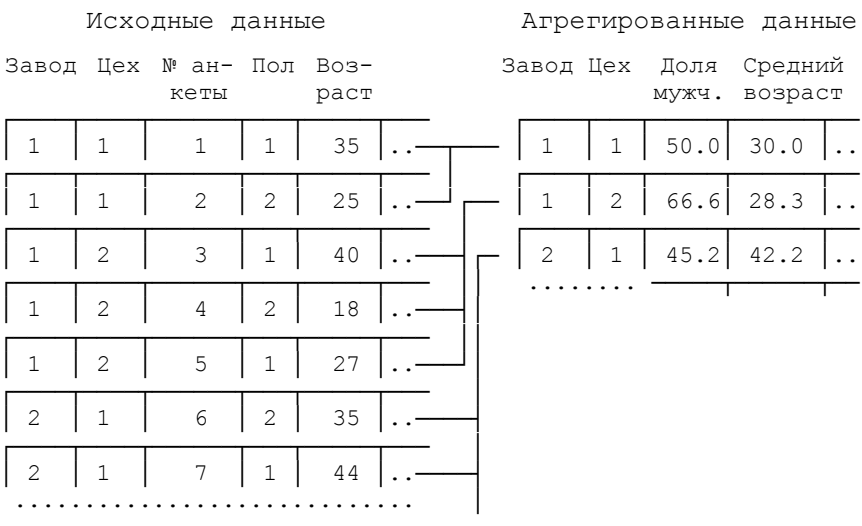


Рис. 2.2. Агрегирование данных

В подкоманде /OUTFILE указывается имя выходного файла; в подкоманде /BREAK назначаются переменные «разрыва» файла данных, которые определяют агрегируемые группы объектов. Далее задаются разделенные слэшами «/» имена новых переменных и функций (статистики), с использованием которых агрегируются исходные переменные, например:

```
Z9 "средний возраст" =  MEAN(V9)/PM = PLT(V8,2).
```

Перед именем функции агрегирования знак равенства «=» ***обязателен***. В списке допускается указание нескольких переменных для одной функции, в списках переменных можно использовать ключевое слово TO (Z9 Z14 = MEAN(V9 V14)/d1 to d6 = pgt(d1 to d6,0)). Число переменных в аргументе функции должно совпадать с числом новых переменных.

2.6.1.1. Функции агрегирования

В приведенном ниже списке функций идентификатор VARS означает список переменных или переменную.

N(VARS) – число объектов, для которых VARS определены;

N – без указания переменных – число объектов в агрегируемой группе;

MIN(VARS) – минимум;

MAX(VARS) – максимум;

SD(VARS) – стандартное отклонение;

PGT(VARS, значение) – процент объектов, у которых переменная имеет значение большее, чем указанное в команде;

PLT(VARS, значение) – процент объектов, у которых переменная имеет значение меньшее, чем указанное в команде;

PIN(VARS, значение1, значение2) – процент объектов, которые находятся в интервале [значение1, значение2];

POUT(VARS, значение1, значение2) – процент объектов, которые находятся вне интервала [значение1, значение2];

FGT(VARS, значение) – доля объектов, у которых переменная имеет значение большее, чем указанное в команде;

FLT(VARS, значение) – доля объектов, у которых переменная имеет значение меньшее, чем указанное в команде;

FIN(VARS, значение1, значение2) – доля объектов, которые находятся в интервале [значение1, значение2];

FOUT(VARS, значение1, значение2) – доля объектов, которые находятся вне интервала [значение1, значение2];

FIRST(VARS) – первое значение переменной;

LAST(VARS) – последнее значение переменной.

2.6.1.2. Пример агрегирования файла

Задача. Получить на базе исходного файла данных агрегированный по городам файл данных (переменная G является переменной разрыва в файле **oct.sps**). Файл должен содержать переменные:

NG – число опрошенных в городе;

W1 – доля рассчитывающих на свои силы;

W2 – доля отрицательно относящихся к свободным зонам;

W3D1 TO W3D6 – доли по подсказкам на вопрос III о причинах неподписания договора;

W4 – доля считающих, что острова нужно отдать;

W8 – доля женщин; W9 – средний возраст;

W10 – доля лиц с высшим образованием;

WR – регион.

Все переменные, кроме W3D1 TO W3D6, могут быть непосредственно получены с использованием функций агрегирования; для формирования переменных W3D1 TO W3D6 придется специально подготовиться, пользуясь командой COUNT.

```
GET FILE "D:\oct.sav".
COUNT d1 = v3s1 to v3s8(1)/ d2 = v3s1 to v3s8(2)
/ d3 = v3s1 to v3s8(3)/ d4 = v3s1 to v3s8(4)
/ d5 = v3s1 to v3s8(5) / d6 = v3s1 to v3s8(6).
AGGREGATE /OUTFILE = "D: aggr.sps"/BREAK g/NG "число
опрошенных в городе" = N/
```

```
  W1 'рассч на св силы' = PIN(v1,1,1)/
  W2 ' % отриц. относящ' = PIN(v2,3,4)/w3d1 to w3d6 =
PGT(d1 to d6,0)/
  W4 'мнен: острова отдать' = PIN(v4,1,1)/
  W8 'доля мужчин' = PIN(v8,2,2)/
  W9 'средний возраст' = MEAN(v9)/
  W10 'доля с высшим образованием' = PIN(v10,1,1)/
  WR = FIRST(r).
```

В новом файле будут созданы переменные W1, W2, W3D1, W3D2, W3D3, W3D4, W3D5, W3D6, W4, W8, W9, W10, WR. Так как после выполнения агрегирования остается активным исходный файл, то, чтобы начать работу с вновь созданным файлом, необходимо вызвать его командой GET.

По данным нового файла можно, например, командой MEANS вычислить средние значения переменных по регионам:

```
MEAN W3D1 TO W3D6 BY R.
```

или рассчитать корреляции долей, рассчитанных для городов:

```
CORR W1 W2 WITH W3D1 TO W3D6/OPTIONS 5.
```

и т. д. Напомним, что объектами агрегированного файла данных являются города, и нужно серьезно подумать над интерпретацией получаемых статистик. В частности, среднее значение переменной W9 будет не средним возрастом, а средним средних возрастов по городам.

2.6.2. Объединение файлов (MERGE FILES)

В пакете имеется возможность объединения данных различных файлов. Это предпочтительно делать с помощью меню **Data/ Merge**.

Рассмотрим, какие виды объединения файлов возможны (рис. 2.4).

Во-первых, это дополнение массива данных новыми строками – ***объектами*** (функция ADD). На практике такая операция необходима, если

– происходит многоэтапное исследование по одной и той же анкете, опрос в нескольких регионах и т. п.;

– исследователю повезло – удалось получить информацию другого обследования (не панельного, то есть опрос проводился по другой анкете и других людей). Информация частично совместима по переменным с имеющимися данными, и необходимо составить общий массив данных.

Во-вторых, дополнение данных ***новыми переменными*** (функция MATCH). Такое пополнение массива данных обычно необходимо, если

– не удается сразу закодировать все данные; на подмножестве данных нужно произвести срочные расчеты, другую часть необходимо еще подготовить к вводу;

– необходимо соединить данные панельных обследований;
– дополнение данными из агрегированного файла (функция TABLE).

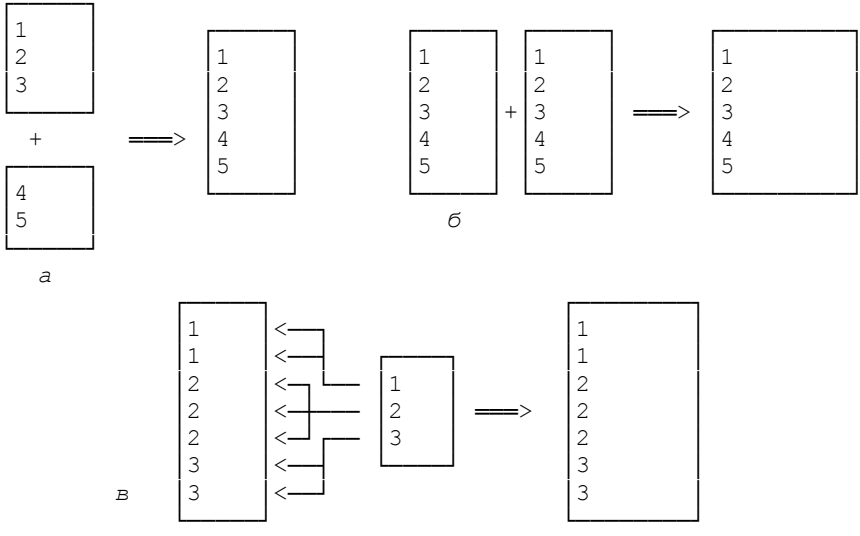


Рис. 2.3. Объединение данных: *а* – по объектам (**ADD**); *б* – по признакам (**MATCH**); *в* – внесение данных для групп объектов (**TABLE**)

Пусть, например, получены точные сведения о промышленности города, детской смертности, загрязнении атмосферы и т. д. Эти данные необходимо внести в каждую анкету жителей городов. Их можно закодировать, но экономичнее и быстрее сделать файл агрегированных данных и процедурой приписать дополнительно к объектам-анкетам в исходный файл (см. рис. 2.4).

Подробно объединение файлов описано в учебнике [1].

В качестве примера проведем присоединение данных агрегированного файла **Aggr.sav** (см. пример из предыдущего раздела) к анкетным данным курильского обследования, находящимся в файле **oct.sav** :

```
GET FILE "D:\oct.sav".
SORT CASES BY g (A) .
MATCH FILES /FILE = * /TABLE = 'D: Aggr.sav' /BY g.
EXECUTE.
```

Сортировка файлов данных по ключевой переменной здесь обязательна; если данные не отсортированы, есть риск их потерять.

После объединения, в файле **D:\oct.sps** появятся переменные d1, d2, d3, d4, d5 и d6, а также w1, w2, w4, w8, w9, w10 и wr. Это объеди-

нение позволяет изучать, как связано «общественное мнение» с индивидуальными характеристиками респондентов.

Заметим, что «ручное» написание команды в данном случае требует особой внимательности, так как диагностирование ошибок в этой команде не на высоком уровне.

Глава 3. ПРОЦЕДУРЫ ПОЛУЧЕНИЯ ОПИСАТЕЛЬНЫХ СТАТИСТИК И ТАБЛИЦ СОПРЯЖЕННОСТИ

Разнообразные режимы работы процедур статистического анализа и расчета таблиц распределений реализуются большим числом команд. При этом требуется задать множество параметров, что делает использование подсказок для таких процедур в режиме синтаксиса утомительным. Формирование текста этих команд намного удобнее в диалоговых окнах и практически не требует знания их синтаксиса. Но готовый текст команд рекомендуется запоминать в файле **Syntax**. Режим запуска статистических процедур из программного файла значительно экономит время, особенно когда приходится многократно повторять расчет, корректируя лишь параметры. Для первичного анализа данных достаточно процедур, реализуемых следующими командами раздела меню: **Analyze** (или **Statistics** в 6 – 8 версиях SPSS), содержащихся в пунктах:

- **Descriptive Statistics** – команды:
  Frequencies – распределения;
  Descriptives – одномерные описательные статистики;
  Explore (Examine) – одномерные описательные статистики в группах объектов;
  Crosstabs – таблицы сопряженности;
- **Compare Means** – команда:
  Means – средние;
- **Custom Tables** – команда:
  Multiple Response, General Table – таблицы для неальтернативных признаков.

Следует помнить, что команда меню **Explore** в языке программирования SPSS имеет имя `Examine`.

3.1. Команды получения распределений и описательных статистик

3.1.1. `FREQUENCIES` – получение одномерных распределений переменных

Процедура `FREQUENCIES` позволяет получить только самые основные статистические характеристики случайной переменной: перечень значений и частотное распределение, т. е. сколько раз переменная принимала каждое

из этих значений. Частотное распределение выдается в числовом виде, в виде процентов и в зависимости от желания пользователя представляется в виде таблицы и/или графика. По умолчанию выдается таблица.

Пример

```
FREQUENCIES VAR V1 V8 / HISTOGRAM /STATISTICS = MEANS.
```

Синтаксис: указываются через пробел переменные для табулирования. Допустимы числовые и строковые переменные. Параметры процедуры необязательны и задаются ключевыми словами, разделенными косыми чертами «/». В параметрах могут быть подпараметры.

На рис. 3.1 и в табл. 3.1 дан пример полученного процедурой `FREQUENCIES` частотного распределения респондентов анкеты «Курильские острова» и его столбиковой диаграммы по результатам их ответов на вопрос о точке зрения на иностранную помощь.

Наиболее распространенным (433 ответа) было мнение, что островам нужна ограниченная иностранная помощь. Из текста таблицы и подписей гистограммы видно, насколько удобно в практической работе использовать `VAR LAB` и `VAL LAB` – команды присвоения признакам текстовых имен. В колонке «Percent» проценты даны относительно всего объема выборки с учетом неопределенных кодов. В колонке «Valid Percent» приведены проценты в выборке без неопределенных кодов. В колонке «Cum Percent» – суммарный процент с нарастающим итогом, рассчитанный без учета объектов с неопределенными значениями.

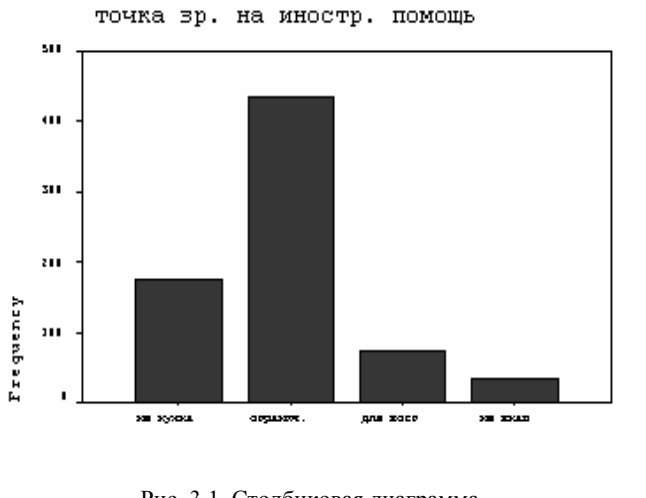


Рис. 3.1. Столбиковая диаграмма

Таблица 3.1

Таблица распределения числа респондентов курильского обследования по значениям переменной V1 «Точка зрения на иностранную помощь»

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 не нужна	177	24,5	24,7	24,7
	2 огранич.	433	60,1	60,5	85,2
	3 нужна	73	10,1	10,2	95,4
	4 не знаю	33	4,6	4,6	100
	Total	716	99,3	100	
Missing	0	5	0,7		
Total		721	100		

Пример

```
MISSING VALUES V1(0).
FREQUENCIES V1 /BARCHART.
```

В выборке 5 респондентов из 721 не ответили на первый вопрос и были закодированы при наборе данных «0». В данном примере мы указываем пакету, что нулевой код следует воспринимать как неопределенные пользовательские значения.

В процедуре `FREQUENCIES` полезно использовать следующие необязательные параметры:

`/BARCHART` – столбиковая диаграмма;

`/PIECHART` – круговая диаграмма;

`/HISTOGRAM` – гистограмма;

`/NTILES` – *n*-тили (квартили, квинтили, децили и др.);

`/PERCENTILES` – процентили;

`/STATISTICS` – все статистики, реализованные в команде**.**

3.1.1.1. Подкоманды `/BARCHART`, `/PIECHART` и `/HISTOGRAM` – диаграммы распределения

Столбиковая и круговая диаграммы обычно используются для неколичественных переменных.

Гистограмма необходима для графического представления количественных данных. Для ее построения SPSS подбирает интервалы группирования значений переменной и представляет графически частоты или доли числа объектов, попавших в соответствующие интервалы. К сожалению, принцип определения числа интервалов в имеющейся у нас документации SPSS не описан. В синтаксисе команды можно задать интервал значений, для которых будет выдаваться гистограмма.

На рис. 3.2 представлен график, полученный командой, в которой задан интервал:

```
FREQUENCIES VARIABLES = V9/ HISTOGRAM min(30), max(50).
```

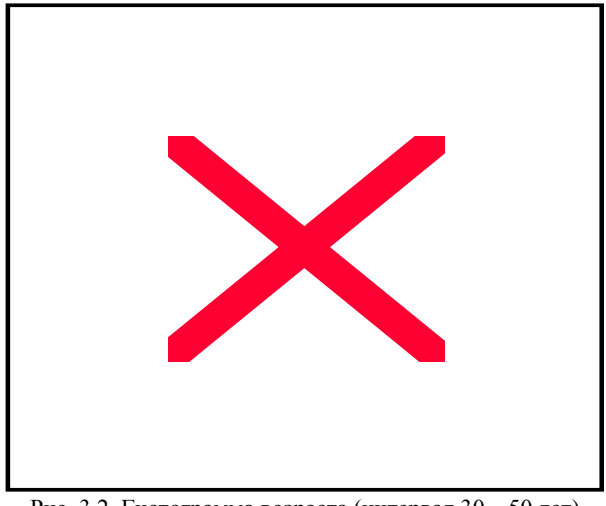


Рис. 3.2. Гистограмма возраста (интервал 30 – 50 лет)

Соотношение высоты столбиков отражает соотношение количества респондентов, имеющих возраст из соответствующего двухлетнего интервала. Например, из гистограммы видим, что более всего в выборке было 36 – 38-летних. Или: с увеличением возраста после 44 лет численность опрашиваемых сокращалась почти в равных пропорциях для трех последующих интервалов. Можно отметить также активное включение в опрос лиц в возрасте 50 – 52 года.

3.1.1.2. Подкоманды /NTILES, /PERCENTILES – *n*-тили, процентили

Подкоманда NTILES задает печать *n*-тилей – значений переменной, делящих распределение на заданное число групп с равным числом объектов. Следующая команда выдает квинтили (деление на 5 частей) переменной, содержащей данные по доходу:

```
FREQUENCIES /VARIABLES = V14 /NTILES = 5.
```

Подкоманда PERCENTILES печатает процентили (процентиль – это квантиль, рассчитанная по доле, указанной в процентах). Процентили являются значениями переменной, отделяющими указанную в процентах долю совокупности объектов. Пример: найдем значения дохода, отделяющие 10 % выборки, 50 % (медиану) и 90 %:

```
FREQUENCIES /VARIABLES = V14 /PERCENTILES 10 50 90.
```

Процентили удобно использовать, если нам нужно разбить упорядоченные значения переменной на интервалы, которые содержали бы задаваемое нами количество объектов (анкет).

3.1.1.3. Подкоманда /STATISTICS – описательные статистики

Подкоманда позволяет получить одномерные описательные статистики.

```
FREQUENCIES V1 V2 V4 /STATISTICS DEFAULT.
```

Ключевые слова:
MEAN – среднее;
SEMEAN – стандартная ошибка среднего;
MEDIAN– медиана (процентиль с 50 %)
MODE – мода (наиболее частое значение)
STDDEV – стандартное отклонение;
VARIANCE – дисперсия;
KURTOSIS – эксцесс (пикообразность);
SEKURT – стандартная ошибка эксцесса;
SKEWNESS – коэффициент асимметрии (скошенность);
SESKEW – стандартная ошибка коэффициента асимметрии;
RANGE – разброс = (MAX - MIN);
MINIMUM – минимум;
MAXIMUM – максимум;
SUM – сумма всех значений переменной;
ALL – все статистики;
DEFAULTS – статистики MEAN, STDDEV, MIN, MAX.

Статистика MEAN вычисляется по известной формуле $\bar{x} = (x_1 + x_2 + ... + x_n)/n$, VARIANCE – несмещенная оценки дисперсии – по формуле $S_x{}^2 = \sum(x_i - \bar{x})^2/(n-1)$, SEMEAN – стандартная ошибка среднего – по формуле $S_{\bar{x}} = \sqrt{S_x{}^2/n}$.

Стандартную ошибку можно использовать для оценки доверительного интервала матожидания (в случае нормального распределения генеральной совокупности границы $(1 - \alpha) \times 100$ % доверительного интервала имеют вид $\mu_{1,2} = \bar{x} \pm t_{1-\alpha/2}(n-1)S_{\bar{x}}$). Напомним, что доверительным интервалом параметра называется интервал со случайными границами, накрывающий значение параметра с заданной (доверительной) вероятностью. В частности, приближенными оценками границ 95 %-го двустороннего доверительного интервала для матожидания являются значения $\bar{x} \pm 1{,}96 S_{\bar{x}}$ (истинное значение матожидания с вероятностью 0,95 находится в этих пределах).

Примерно в пределах $\bar{x} \pm S_{\bar{x}}$ должно находиться около 68 % наблюдений совокупности.

На практике постоянно возникает вопрос, нормально ли распределение переменной, так как многие статистические методы разработаны в предположении нормальности. Исследуемые распределения обычно отличаются от нормального закона, а в этом случае оценки некоторых параметров будут смещены. Например, будет некорректно вычислена наблюдаемая значимость оценки. Исследователю важно понять, опасно ли смещение выборочного распределения от нормального. Приближенно и быстро оценить масштабы отклонения распределения от нормального можно, используя скошенность и пико$\left(\sum (x_i - \bar{x})^3 / S^3 \right) / n$ образность.

Скошенность SKEWNESS определяется расчетом третьего момента по формуле $\left(\sum (x_i - \bar{x})^3 / S^3 \right) / n$ – коэффициент асимметрии.

Если полученная величина меньше нуля, то распределение растянуто влево, если больше нуля – то вправо. Чем больше отличие от нуля, тем значительнее отклонения распределения от нормального.

Пикообразность KURTOSIS определяется значением четвертого момента: $\left(\sum (x_i - \bar{x})^4 / S^4 \right) / n - 3$ – эксцесс.

Нулевое значение **Kurtosis** означает, что пикообразность распределения совпадает с пикообразностью нормального распределения. Чем больше четвертый момент, тем больше пикообразность распределения и, следовательно, отличие от нормального. В этом случае существенность отклонений статистик от теоретических можно проверить, используя стандартные ошибки этих статистик (**Std. Error of Skewness и Std. Error of Kurtosis**). В основе лежит факт, что отношение статистики к ее стандартной ошибке имеет распределение, близкое к нормальному). Например, если это отношение превышает 1,96, то мы должны отклонить гипотезу о равенстве **Kurtosis** нулю в генеральной совокупности и, следовательно, о нормальном распределении переменной.

Полезность этих двух статистик не ограничивается проверкой нормальности распределения. Приобретя некоторый опыт, можно использовать эти статистики для качественного анализа распределения. Например, при исследовании доходов можно использовать **Kurtosis** как измеритель степени неравенства доходов населения. Чем больше пикообразность, тем однороднее доходы.

Перечисленные описательные статистики команды FREQUENCIES играют в анализе данных особую роль. Они позволяют провести первый этап

статистических исследований выборки. Ниже приведен пример описательных статистик, полученных для переменной «Среднемесячный душевой доход в семье», построенной по ответам на 14-й вопрос анкеты «Курильские острова».

```
  FREQUENCIES VARIABLES = V14 /NTILES = 4
/PERCENTILES = 10 90
/STATISTICS = STDDEV VARIANCE RANGE MINIMUM MAXIMUM
SEMEAN MEAN MEDIAN MODE SUM SKEWNESS SESKEW KURTOSIS
SEKURT.
```

Команда вычисляет также *n*-тили и процентили.

Таблица 3.2

Статистики переменной V14 – «Душевой доход», выданные командой FREQUENCIES

N	Valid	673
	Missing	48
Mean		229,11
Std. Error of Mean		5,83
Median		200
Mode		200
Std. Deviation		151,342
Variance		22904,531
Skewness		3,035
Std. Error of Skewness		0,094
Kurtosis		15,080
Std. Error of Kurtosis		0,188
Range		1479
Minimum		21
Maximum		1500
Sum		154190
Percentiles	10	100
	25	140
	50	200
	75	280
	90	400

Анализируя полученные данные (табл. 3.2), видим, что доход в семьях меняется в диапазоне от 21 до 1 500 р. (разброс равен 1 479). При этом средний доход составил около 229,11 р. Приближенными границами пятипроцентного доверительного интервала для матожидания будут значения

229,11 ± 1,96 × 5,83, где 1,96 – критическое значение нормального распределения для ***p*** = 0,05/2 = 0,025. Скошенность ***skewness*** = 3,035 и пикообразность ***kurtosis*** = 15,08 значительно больше нуля. Их стандартные ошибки (0,094 и 0,188 соответственно) свидетельствуют о статистической значимости такого отличия. Действительно, отношение коэффициентов к ошибкам достаточно велико и попадает в критическую область, что позволяет отклонить гипотезу о равенстве полученных статистик нулю.

Результатом задания процентилей и *n*-тилей являются выданные в таблице процентили (у 10 % опрошенных респондентов доход меньше 100 р., у 90 % – меньше 400; имеются также процентили, ограничивающие уровни дохода для 25, 50, 75 % респондентов).

3.1.2. `DESCRIPTIVES` – описательные статистики

Если команда `FREQUENCIES` получает описательные статистики «попутно», то `DESCRIPTIVES` специально для этого предназначена. Ею удобнее пользоваться для анализа количественных переменных.

```
  DESCRIPTIVES VAR = V9 V14 /STATISTICS = MEAN MIN MAX
/ SAVE.
```

Синтаксис: указывается список переменных, список необходимых статистик, подкоманда сохранения в файле полученных стандартизованных переменных (`/save`).

Список вычисляемых статистик (10) здесь значительно меньше, чем в команде `Frequencies` (16):

```
  MEAN MIN SKEWNESS STDDEV SEMEAN MAX KURTOSIS
VARIANCE SUM RANGE.
```

Стандартизованные переменные. Командой `DESCRIPTIVES` необходимо пользоваться для получения нормированных переменных. Потребность в них может появиться, например, для проведения кластерного или регрессионного анализа. Иногда это связано с необходимостью сопоставления разномасштабной информации.

Пример. Мы имеем данные по заработной плате за два последних года. На основании этих данных необходимо определить, в каком социальном слое находятся респонденты. Но это затруднительно сделать, поскольку за 2 года ***существенно*** изменился масштаб цен. Для сравнения преобразуем к стандартному виду данные по каждому году, что позволит нам провести сравнительный анализ для определенных социальных слоев:

$Z = \left(X - \overline{X}\right)/S$, где S – стандартное отклонение переменной X; $\left(S_z{}^2 = 1; \overline{Z} = 0\right)$.

Стандартизованные переменные можно получить, указав в скобках за переменной имя новой, стандартизованной, переменной:

```
DESCRIPTIVES VAR V14(Z14) V9(Z9).
```

Если используется подкоманда SAVE, то сообщать имена нет необходимости. Стандартизованные переменные запишутся в конец файла данных под именами, которые будут автоматически образованы добавлением буквы Z слева к имени исходной переменной.

Например,

```
DESCRIPTIVES VAR = V9 V14/SAVE.
```

Новым переменным пакет присвоит имена ZV9 и ZV14.

Напомним, что более разнообразные преобразования переменных можно получить командой RANK. С помощью этой команды можно ранжировать значения переменной, перекодировать переменную с целью получения нормального распределения, получать процентили и др.

3.1.3. EXPLORE – исследование распределений и сравнение групп объектов

Команда меню **Explore** на языке программирования имеет имя EXAMINE. Она реализует удобный инструмент исследования распределения данных в подвыборках объектов и рассчитывает статистики для проверки нормальности распределения и однородности дисперсий в группах. Мы не будем подробно описывать эту процедуру, поскольку она хорошо описана в книге [7. С. 43 – 71].

Команда отличается развитыми графическими возможностями. В ней предусмотрены гистограммы, диаграммы типа «ствол с листьями», ящичковые диаграммы, графики сравнения эмпирического распределения с нормальным. Для описательного анализа удобны ящичковые диаграммы. Для примера рассмотрим диаграмму распределения по возрасту в группах по семейному положению, полученную командой

```
EXAMINE  VARIABLES = V9 BY V11
/PLOT BOXPLOT HISTOGRAM NPPLOT SPREADLEVEL(1)
/COMPARE GROUP /STATISTICS DESCRIPTIVES
/CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

Нижние и верхние границы «ящичков» показывают 25 % и 75 % процентили распределений, черта посередине – медиана, «усы» показывают максимальные и минимальные значения в группах, если они не отстоят от верхнего (нижнего) края ящичка более чем на 1,5 его длины. Иначе они показывают эту границу, а вышедшие за эти пределы значения отмечаются отдельными точками или кружками (рис. 3.3).

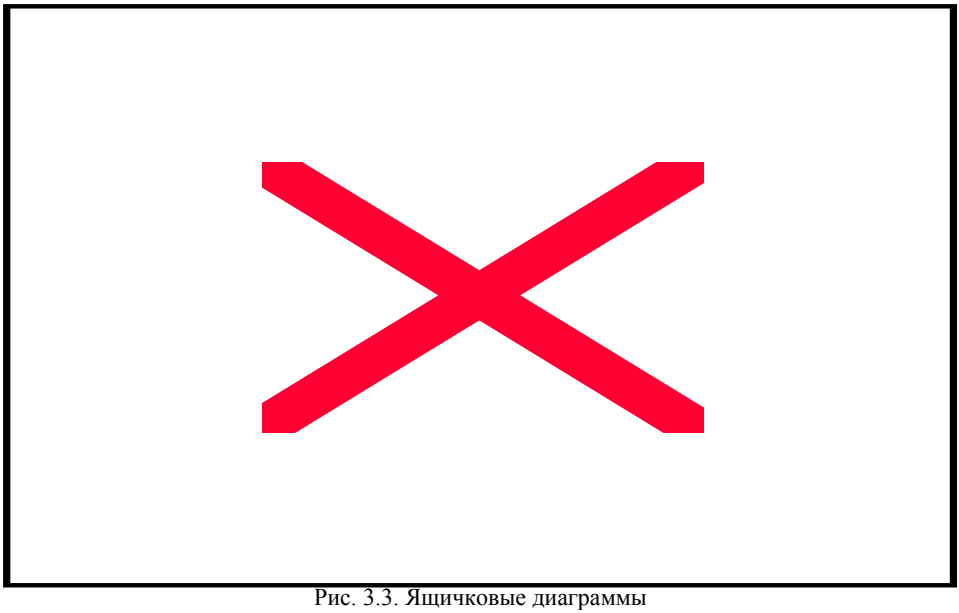


Рис. 3.3. Ящичковые диаграммы

На диаграмме видим, что для всех групп медиана находится не в центре возрастных интервалов. Особенно велик перекос для не состоявших в браке, т. е. среди этой группы опрошенные были в возрасте 20 лет. Выборка ограничена трудоспособным возрастом 16 лет, что отчасти объясняет перекошенность в последней группе. Людям более старшего возраста свойственно заводить семью. Для женатых медиана находилось в возрасте около 40 лет. Медиана возраста разведенных приходится примерно на 44 года и вдовцов в возрасте 52 лет. Т. е. «центральный» возраст разведенных примерно на 4 года больше «центрального» возраста разведенных.

3.2 Анализ связи между неколичественными переменными

3.2.1. CROSSTABS – таблицы сопряженности

CROSSTABS позволяет получать таблицы сопряженности многомерных распределений и связей двух и более переменных. Рекомендуется использовать CROSSTABS для переменных с небольшим числом значений (обычно для неколичественных переменных), так как каждая комбинация значений соответствует новой клетке в таблице.

```
CROSSTABS /TABLES = v1 v2 BY v10 BY pol.
```

Таблицы сопряженности для пары переменных X и Y содержат частоты N_{ij}, с которыми встретилось сочетание i-го значения X и j-го значения Y. Кроме того, в таблице обязательно присутствуют маргинальные частоты

$N_{i.}$, равные сумме чисел N_{ij} по i-строке; $N_{.j}$ – сумме по столбцу (частоты i-го значения X и j-го значения Y, подсчитанные независимо) и N – общее число объектов.

Основные подкоманды CROSSTABS:

/TABLES – задание таблиц;

/CELLS – статистики клеток таблицы;

/STATISTICS – статистики взаимосвязи переменных;

/METHOD – метод проверки значимости связи переменных;

/BARCHART – столбиковая диаграмма.

3.2.1.1. Подкоманда /TABLES – задание таблиц

Параметр TABLES может быть опущен:

```
CROSSTABS v1 TO v5 BY v10.
```

Строки таблицы сопряженности соответствуют значениям переменной, указанной в тексте команды перед ключевым словом «BY»; столбцы матрицы соответствуют значениям переменной, расположенной после «BY».

Пример. Совместное распределение по региону (*R*), точке зрения на иностранную помощь (v1) и полу (V8):

```
CROSSTABS TABLES R BY v1 BY v8/cells = COUNT ROW.
```

В результате выполнения этой команды рассчитывается табл. 3.3. Перед ключевым словом BY указываются переменные, по которым вычисляется двувходовая таблица (переменная, значения которой идентифицируют строки), после ключевого слова BY указываются переменные, идентифицирующие столбцы. За следующими BY идут переменные условий, определяющие подвыборки, на которых рассчитываются таблицы. Хотя в современной версии пакета эти таблицы объединяются в одну таблицу, их статистический анализ производится по отдельности. Ключевым словом BY могут разделяться и списки переменных. В этом случае процедурой получаются таблицы по всем парам таблиц из первого и второго списка. Например,

```
CROSSTABS V8 V11 V12 BY V4 V1.
```

Эта команда выведет таблицу сопряженности: V8 с V4, V8 с V1, V11 с V4, V11 с V1 и т. д., то есть сочетания по всем переменным, перечисленным в команде. Всего будет выдано на печать 6 таблиц. Если более двух списков переменных разделены ключевыми словами BY, то переменные, стоящие за вторым, третьим и т. д. BY, задают условия получения таблиц. Таблицы формируются на подвыборках, соответствующих сочетаниям значений этих переменных.

Таблица 3.3

Распределение переменной «Точка зрения на иностранную помощь» в разрезе региона и пола респондентов

V8 Пол	R регион	Содержимое ячеек	V1 точка зрения на иностр. помощь				
			Не нужна	Огранич.	Нужна	Не знаю	Total
Муж.	Дальн./В.	Count	25	91	22	7	145
		%	17,2	62,8	15,2	4,8	100
	Вост./Сиб.	Count	25	56	13	1	95
		%	26,3	58,9	13,7	1,1	100
	Зап./Сиб.	Count	38	65	13	3	119
		%	31,9	54,6	10,9	2,5	100
	Total	Count	88	212	48	11	359
		%	24,5	59,1	13,4	3,1	100
Жен.	Дальн./В.	Count	26	87	9	6	128
		%	20,3	68,0	7,0	4,7	100
	Вост./Сиб.	Count	23	54	6	7	90
		%	25,6	60,0	6,7	7,8	100
	Зап./Сиб.	Count	40	75	9	7	131
		%	30,5	57,3	6,9	5,3	100
	Total	Count	89	216	24	20	349
		%	25,5	61,9	6,9	5,7	100

Употребление `BY` в команде `CROSSTABS` возможно до 10 раз, но и этого достаточно, чтобы занять все ресурсы компьютера.

Если мы хотим получить в одной команде `CROSSTABS` несколько независимых таблиц, то следует отделять списки переменных символом /:

```
CROSSTABS V8 V11 BY V4 V1/ V12 BY V1 /CELLS row.
```

Таблица, заполненная одними частотами N_{ij}, обычно не имеет смысла, так как она не проясняет должным образом взаимосвязи переменных. Для исследования взаимосвязи необходимы статистики оценки взаимосвязи самих переменных и статистики оценки связи их значений.

3.2.1.2. Подкоманда /`CELLS`

Параметр `CELLS` задает вывод некоторых статистик (см. ниже параметры подкоманды `CELLS`) для клеток таблицы сопряженности. *Cells* переводится как «клетка». Если этот параметр не указан, то в клетках таблицы выводятся только абсолютные частоты.

Пример задания статистик клеток:

```
CROSSTABS V1 BY V4 /CELLS = COUNT ROW COLUMN.
```

Параметры подкоманды /CELLS:

COUNT – абсолютное число объектов (N_{ij});

ROW – проценты по строке;

COLUMN – проценты по столбцу;

TOTAL – проценты по отношению ко всей выборке;

EXPECTED – частоты ($E_{ij} = N_{i.} \times N_{.j}/N$), ожидаемые в случае независимости переменных (N – общая сумма частот в таблице);

RESID – изменение ($N_{ij} - E_{ij}$) частоты *Nij* по сравнению с ожидаемым *Eij* в условиях независимости переменных;

SRESID – стандартизованное изменение частоты по сравнению с ожидаемым $(N_{ij}–E_{ij})/ E_{ij}^{1/2}$. Напомним, что статистика хи-квадрат, вычисляемая для проверки гипотезы независимости рассматриваемых переменных, является суммой квадратов этих величин.

ASRESID – стандартизованные изменения частоты $Z_{ij} = (N_{ij}–E_{ij})/\sigma_{ij}$, где σ_{ij} вычисляется исходя из гипергеометрического распределения N_{ij} (см. ниже п. 3.3.1). Статистика Z_{ij} имеет асимптотически нормальное распределение $N(0,1)$;

ALL – вывод для клетки всех статистик.

Табл. 3.4 получена в результате преобразования данных и применения процедуры CROSSTABS с параметром CELLS:

```
RECODE v4 (1,2 = 1)(3 = 2)(4 = 3) into W4.
VAR LAB W4 "Возможность удовлетворить территориаль-
ные требования Японии".
VAL LAB W4 1 "отдать" 2 "не надо" "не знаю".
CROSSTABS /TABLES = v1 BY W4 /CELLS = COUNT ROW col.
```

Верхний процент в клетке означает отношение числа объектов, попавших в эту клетку, к итоговой сумме числа объектов по строке. Нижний процент соответствует отношению значения клетки к итоговой сумме по столбцу. По величине процентов, приведенных в клетках, можно сравнивать группы респондентов по распределению как по значениям «вертикальной» переменной, так и по «горизонтальной».

В частности, анализируя первую строку матрицы (она соответствует ответам тех респондентов, которые считают, что иностранная помощь не нужна), видим, что основная часть – 81,7 % этой группы респондентов против передачи островов Японии. При этом их доля среди тех, кто против передачи островов, составляет всего 27,2 %; а основная часть (62,0 %) про-

тивников передачи островов допускает возможность получения ограниченной иностранной помощи. В последнем столбце таблицы расположены итоги по каждой строке, которые совпадают с распределением по переменной v1. Так как до выполнения команды CROSSTABS были объявлены неопределенные значения v1 и v4, таблица рассчитывалась без их учета, поэтому объем выборки, учтенный в таблице, составил 712 анкет из 721 имеющихся. Аналогичные данные для столбцов приведены в строке TOTAL.

Таблица 3.4

Связь ответов на вопросы «Точки зрения на иностранную помощь» и «Возможности удовлетворения территориальных требований Японии» (частоты и проценты)

V1 точка зрения на иностранную помощь		V4 Возможность удовлетворить территориальные требования Японии			Total
		1 отдать	2 не надо	3 не знаю	
Не нужна	Count	21	143	11	175
	% row	12,0	**81,7**	6,3	100,0
	% col	19,6	27,2	13,9	24,6
Огранич.	Count	57	326	48	431
	% row	13,2	75,6	11,1	100,0
	% col	53,3	**62,0**	60,8	60,5
Нужна	Count	27	32	14	73
	% row	37,0	43,8	19,2	100,0
	% col	25,2	6,1	17,7	10,3
Не знаю	Count	2	25	6	33
	% row	6,1	75,8	18,2	100,0
	% col	1,9	4,8	7,6	4,6
Total	Count	107	526	79	712
	% row	15,0	73,9	11,1	100,0
	% col	100,0	100,0	100,0	100,0

Проценты в CROSSTABS позволяют изучать взаимосвязь переменных, а не только структуру таблицы. В частности, сравнивая строки, можно сделать заключение, что более склонны отдать острова те, кто считает, что нужна помощь восточным регионам (37 %), чем те, кто считает, что помощи не нужно. Можно взять в качестве точки отсчета распределение в целом по совокупности (всего 15 % в среднем по массиву готовы отдать все или часть островов).

3.2.1.3. Статистики смещения частот

Реализованные в параметре CELLS статистики позволяют провести более сложный анализ связи переменных. Например, в табл. 3.4 можно уви-

деть, что среди полагающих ненужной иностранную помощь 12 % готовы отдать острова Японии. Среди считающих, что помощь нужна, их 37 %. В то же время в целом по совокупности лишь 15 % готовы передать острова. Существенны ли полученные отличия долей подмножеств соответственно на 3 % и 22 % от доли в целом по совокупности? Может ли в следующем обследовании связь оказаться противоположной? Основой для исследования смещения выборки от истинного распределения служат теоретические значения, ожидаемые в случае независимости выборки. Подпараметр `EXPECTED` параметра `CELLS` позволяет вывести в клетках абсолютные значения частот (N_{ij}) и ожидаемые в предположении независимости переменных (теоретические) частоты (E_{ij}). Отклонение ($N_{ij} - E_{ij}$) наблюдаемой частоты от ожидаемой – более удобная величина для анализа, она достаточно наглядна, но остается неясным, насколько это отклонение статистически значимо.

Более полезна статистика $Z_{ij} = (N_{ij} - E_{ij})/\sigma_{ij}$ – стандартизованное смещение частоты; Z_{ij} выдается в клетке при указании подпараметра `ASRESID` (*Adjusted residuals*). Иными словами, Z_{ij} представляет собой отклонение наблюдаемой частоты от ожидаемой, измеренное в числе стандартных отклонений. При этом стандартное отклонение σ_{ij} вычисляется исходя из предположения, что N_{ij} – случайная величина, имеющая гипергеометрическое распределение:

$$\sigma_{ij}^{\,2} = \left\{ N_{i.} * N_{.j} \left(N - N_{i.}\right)\left(N - N_{.j}\right) \right\Big/ \left(N^2 \times (N-1) \right).$$

Если переменные независимы, то при больших N случайная величина Z_{ij} имеет нормальное распределение с параметрами (0,1). Для нее практически невероятно принять значение, большее трех стандартных отклонений, т. к. вероятность такого значения составляет менее 0,0027 (правило «трех сигм»). Поэтому, если мы получаем значение Z_{ij}, превышающее 3, то можем считать, что i-е значение и j-е значение X и Y связаны. На практике, когда анализируется единственная клетка таблицы, выставляются более слабые требования. Существенными считаются уже те односторонние отклонения, которые превышают лишь $1{,}65\sigma_{ij}$ – вероятность их получения составляет 5 %. Таким образом, начиная с отклонения 1,65 (Z_{ij} имеет $\sigma_{ij} = 1$) и большего, можно высказывать гипотезу о существовании связи между значениями. (См. таблицу нормального распределения в любом статистическом справочнике).

В практических расчетах принято считать теоретическое распределение Z_{ij} близким к нормальному, если $\sigma_{ij}^{\,2} > 9$. Хотя последнее ограничение дос-

таточно жестко, так как можно показать, что для его выполнения в выборке должно быть по крайней мере 144 наблюдения.

К сожалению, получив данные расчетов, указывающие на зависимость ($Z_{ij} > 1{,}96$ в случае 5 %-го двустороннего критерия) значений, мы не вправе быть уверенными, что эта зависимость существует.

На практике мы рассчитываем показатели значимости для множества клеток. Чем их больше, тем выше вероятность ***случайно*** получить хотя бы одно значение, превышающее указанный порог. Из теории следует, что если клетки независимы, то при критическом значении статистики Z_{ij}, равном 1,96 (5 %-й уровень значимости), мы в среднем найдем 5 «значимых» из 100 клеток таблицы. А хотя бы одну статистику, превзошедшую критическое значение ($|Z_{ij}| > 1{,}96$), в условиях независимости клеток мы можем получить с вероятностью $(1 - 0{,}95^{100}) = 0{,}9941$! Таким образом, если мы получили значимые связи, то это дает нам лишь повод для высказывания гипотезы об их наличии и требует содержательной дополнительной проверки. Поэтому сложившаяся практика руководствоваться отклонением 1,96 оберегает нас только от грубейших ошибок. В то же время, если мы не получили значимых связей, то можем делать вывод либо об их отсутствии, либо о недостаточном количестве данных для их обнаружения.

Величина `SRESID` – стандартизованное изменение частоты по сравнению с ожидаемым $(N_{ij}–E_{ij})/ E_{ij}^{1/2}$ – связана с распределением Пуассона. Напомним, что распределение Пуассона – это распределение числа успехов для редко случающихся событий при большом числе испытаний. Если попадание наблюдения в клетку (i,j) считать этим редким событием, то ожидаемое значение E_{ij} можно считать оценкой параметра распределения Пуассона (λ). Дисперсия распределения Пуассона совпадает с его математическим ожиданием, поэтому $(N_{ij}–E_{ij})/ E_{ij}^{1/2}$ является отклонением, вычисленным в числе стандартных отклонений. При больших ожидаемых частотах E_{ij} так же, как, `ASRESID` – распределение Пуассона, асимптотически нормально, что позволяет нам решать вопрос о независимости ответов, проверив попадание наблюдаемого значения `SRESID` в критическую область.

Пример. (См. табл. 3.5.) Определим зависимость между отношением к получению иностранной помощи и «Возможностью удовлетворить территориальные требований Японии»:

```
  CROSSTABS /TABLES = v1 BY W4/CELLS = COUNT EXPECTED
RESID ASRESID.
```

Так как в `CELLS` указан параметр `COUNT, EXPECTED, RESID` и `ASRESID`, то в клетках выведены реальные и ожидаемые значения, а также

абсолютная разность расчетной частоты от ожидаемой. В нижней строке клеток выведена эта же разность, но в числе стандартных отклонений.

Таблица 3.5

Связь ответов на вопросы «Точки зрения на иностранную помощь» и «Возможностью удовлетворить территориальные требования Японии» (статистики смещений частот)

V1 точка зрения на иностр. помощь		W4 Возможн. удовлетворить территориальные требования Японии			Total
		Отдать	Не надо	Не знаю	
Не нужна	Count	21,0	143,0	11,0	175
	Expected Count	26,3	129,3	19,4	175
	Residual	–5,3	13,7	–8,4	
	Adjusted Residual	–1,3	2,7	–2,3	
Огранич.	Count	57,0	326,0	48,0	431
	Expected Count	64,8	318,4	47,8	431
	Residual	–7,8	7,6	0,2	
	Adjusted Residual	–1,7	1,3	0,0	
Нужна	Count	27,0	32,0	14,0	73
	Expected Count	11,0	53,9	8,1	73
	Residual	16,0	–21,9	5,9	
	Adjusted Residual	5,5	–6,2	2,3	
Не знаю	Count	2,0	25,0	6,0	33
	Expected Count	5,0	24,4	3,7	33
	Residual	–3,0	0,6	2,3	
	Adjusted Residual	–1,5	0,3	1,3	

В табл. 3.5 получен ответ на поставленный в начале раздела вопрос: смещение частоты в клетке «Отдать острова» – «Нужна помощь» (*residual* = 16) оказалось существенным, так как $Z = 5{,}5 >> 1{,}96$! В то же время смещение частоты на 5,3 в клетке «помощь не нужна – отдать» – не значимо ($Z = 1{,}3 < 1{,}96$, и гипотеза независимости значений принимается).

В статистической взаимосвязи значений переменных можно еще раз убедиться, рассмотрев табл. 3.6 с процентными распределениями (в среднем по совокупности 15 % считают, что острова можно отдать, в то время как в этой группе таковых 37 %!). В то же время, судя по статистикам, хотя и видна отрицательная связь значений «нужна ограниченная помощь» – «отдать острова», она все же не достаточно значима. Конечного потребителя полученных результатов чаще интересует не значение *Z*-статистик, а величина смещения процентов.

Надеемся, что нам удалось показать, что эти статистики наиболее интересны для интерпретации. К сожалению, в SPSS расчет Z_{ij} реализован без учета размеров выборки, что необходимо иметь в виду, так как для малых выборок эти вероятностные рассуждения оказываются неточными.

3.2.1.4. Подкоманда /STATISTICS – исследование связи неколичественных переменных

В предыдущем разделе изучалась связь отдельных значений переменных. Для получения ответа на вопрос о связи самих переменных используется подкоманда STATISTICS команды CROSSTABS**.** Пользователю необходимо указать статистику или параметр, выбранный для исследования связи переменных. Вот некоторые из этих статистик:

CHISQ – позволяет оценить связь с помощью критерия хи-квадрат; кроме значения коэффициента хи-квадрат при задании этого ключевого слова выдается отношение правдоподобия (*Likelihood Ratio*), а также статистика для проверки линейной связи. Последняя статистика редко используется и поэтому не рассматривается в нашем учебно-методическом пособии.

PHI – коэффициент фи-Пирсона; вместе с этим коэффициентом выдаются:

V – коэффициент Крамера;

CC – коэффициент контингенции;

BTAU – тау-B Кендалла для ранговых переменных;

CTAU – тау-C Стюарта для ранговых переменных;

ALL – все статистики (около десятка), включая вышеперечисленные.

Как можно охарактеризовать в целом связь неколичественных переменных? Для характеристики их связи наиболее часто используется критерий хи-квадрат (*CHISQ*), основанный на вычислении статистики:

$$CHISQ = \sum_{i,j} \left(N_{ij} - E_{ij}\right)^2 \Big/ E_{ij} \ .$$

Эта величина показывает расстояние эмпирически полученной (расчитанной нами по результатам обследования на основании выборки) таблицы сопряженности от ожидаемой теоретически. В ее основе лежит расстояние между значениями N_{ij} выборочной таблицы и E_{ij} – ожидаемыми в условиях независимости переменных. Само по себе значение статистики ни о чем не говорит. Важно знать ***вероятность*** получения расстояния *CHISQ*, большего, чем оно может быть для случайной выборки в условиях независимости переменных. Напомним, что такая вероятность называется ***наблюдаемой значимостью*** и обозначается словом *Significance* (возможны сокращения *Sig., P*-значения).

Пакет выдает выборочное значение *CHISQ* и его значимость. Традиционно считается, что значение *Significance*, меньшее 0,05, свидетельствует о взаимосвязи переменных, т. к. значение статистики попадает в критическую область и гипотезу о независимости переменных следует отвергнуть.

CHISQ в условиях независимости и при достаточном числе наблюдений имеет распределение, близкое к распределению хи-квадрат с $(r-1)(c-1)$ степенями свободы, где r – число строк в таблице, c – число столбцов ($CHISQ_{\text{теор.}} \approx \chi^2((r-1)(c-1))$). Существует эмпирическое правило, по которому считается, что *CHISQ* достаточно точно аппроксимируется теоретическим распределением $\chi^2((r-1)(c-1))$, если не более 20 % клеток имеют ожидаемые частоты $E_{ij} < 5$ и нет $E_{ij} < 1$. В выдаче всегда присутствует информация о числе клеток, где это соотношение не выполняется. Рекомендуется использовать в CROSSTABS критерий хи-квадрат для переменных **с небольшим числом значений,** что достигается перекодировкой переменных.

Вместе с критерием хи-квадрат выдается также логарифм отношения правдоподобия *LI*:

$$LI = 2\sum_{i,j} N_{ij} \ln\left(\frac{N_{ij}}{N_{i.}N_{i.}/N}\right).$$

Этот показатель также имеет асимптотическое хи-квадрат – распределение, но более устойчивое к объему выборки. Поэтому при оценке связи пары признаков мы рекомендуем пользоваться отношением правдоподобия.

Таблица 3.6

Тесты хи-квадрат

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,517	3	.015
Likelihood Ratio	10,708	3	.013
Linear-by-Linear Association	0,156	1	.693
N of Valid Cases	708		

a 0 cells (.0 %) have expected count less than 5. The minimum expected count is 22,25.

Наблюдаемая значимость (*Significance*) – это вероятность случайно получить большее значение, чем выборочное. Таким образом, для *CHISQ* наблюдаемая значимость (SIG) равна $P\{CHISQ > CHISQ_{\text{выбороч.}}\}$, и, аналогично, для отношения правдоподобия *LI* наблюдаемая значимость (SIG) равна $P\{LI > LI_{\text{выбороч.}}\}$.

Пример задания для исследования связи ответа на вопрос о необходимости иностранной помощи (v1) и полом (v8):

```
  CROSSTABS v8 BY v1 /CELLS COUNT ROW COL ASRESID
/STATISTICS = CHISQ.
```

В приведенном примере наблюдаемая значимость *CHISQ* составила около 1,5 % (см. Asymp. Sig. (2-sided)), значимость *LI* примерно 1,3 %. С такой незначительной вероятностью в условиях независимости можно случайно получить большие значения соответствующих статистик. Поэтому в соответствии с 5 %-м уровнем значимости переменные v8 и v1 следует считать связанными (1,5 % < 5 %). Таким образом, можно сделать вывод, что мужчины и женщины имеют разные мнения в вопросе об иностранной помощи.

Текст под таблицей «a 0 cells (.0 %) have expected count less than 5. The minimum expected count is 22,25» свидетельствует, что все ожидаемые частоты больше 5, их минимум равен 22,25. Это свидетельствует о корректности использования критерия.

В расчетах нами было получено для клетки «мужчины» – «помощь нужна», значение *Z*-статистики, равное 2,9, что больше 1,65, и, следовательно, ответы зависимы. Кроме того, из таблицы следует, что о необходимости помощи говорят вдвое больше мужчин, чем женщин. Мы не будем приводить здесь эту таблицу, покажем лишь столбиковую диаграмму на рис. 3.4, полученную командой

```
CROSSTABS v8 BY v4 / CELLS COUNT ROW COL ASRESID
/BARCHART.
```

На диаграмме ясно видно, что среди респондентов, сказавших, что помощь нужна, столбик, соответсвующий количеству мужчин, существенно больше столбика, соответствующего количеству женщин.

3.2.1.5. Измерение силы связи между номинальными переменными

В условиях, когда связь значима и величина значимости (*Significance*) близка к нулю, появляется необходимость оценить силу этой связи и выявить наиболее связанные переменные. Непосредственное использование коэффициента хи-квадрат неудобно – он зависит от числа объектов, из-за чего одинаковые по пропорциям распределений таблицы на выборках разного объема будут оценены по-разному.

Коэффициент Пирсона $PHI = \sqrt{CHISQ/N}$ лишен этого недостатка, но диапазон его изменения зависит от размерности таблиц:

$0 \le PHI \le \sqrt{\min(r-1, c-1)}$, где r – число строк, $\boldsymbol{c}$ – число столбцов.

Более устойчив к размерности выборки коэффициент контингенции:

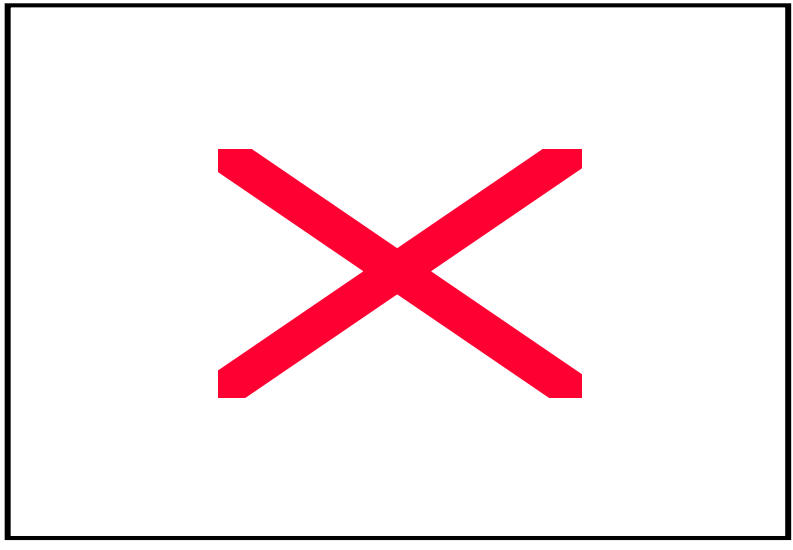


Рис. 3.4. Столбиковая диаграмма, полученная CROSSTABS

$CC = \sqrt{CHISQ/(CHISQ+N)}$, $0 \le CC < 1$;

еще лучше в этом отношении коэффициент Крамера

$V = \sqrt{CHISQ / N(k-1)}$, где $\kappa = \min [r, c]$, $0 \le V \le 1$.

Все эти коэффициенты можно использовать для оценки силы связи и, сравнивая их по величине, делать вывод о более тесной или менее тесной связи. Но эти коэффициенты не носят точного характера, поэтому их использование – дело вкуса каждого исследователя.

Заметим, что коэффициенты анализа связи переменных хи-квадрат (*CHISQ*), Фи (*PHI*) и обычный коэффициент корреляции изобретены Пирсоном.

3.2.1.6. Коэффициенты связи между ранговыми переменными

Коэффициенты *BTAU* (Кендалла) и *CTAU* (Стюарта) служат для оценки взаимосвязи ранговых переменных.

Напомним, что ранговыми переменными называются переменные, в которых можно установить порядок между значениями. Например, ответы на вопрос, требующий ответа «плохо», «средне» или «хорошо». Количественные переменные, такие как возраст, доход также можно использовать в качестве ранговых.

Рассмотрим пары всех объектов (строк матрицы данных). Для пары объектов (*i, j*) рассматривается, одинаково ли упорядочиваются объекты и по переменной *X* и по переменной *Y*. Если $X_i < X_j$ и $Y_i < Y_j$ или $X_i > X_j$ и $Y_i > Y_j$, то упорядочения одинаковы, если $X_i < X_j$ и $Y_i > Y_j$ или $X_i > X_j$ и $Y_i < Y_j$ – упорядочения не одинаковы. Число одинаковых упорядочений для всех пар объектов по *X, Y* обозначим *P*; число разных – *Q*. Кендалл предложил рассматривать величину $BTAU = (P - Q) / T$, где T – нормирующий знаменатель, такой, чтобы величина *BTAU* изменялась от –1 до 1. $BTAU = -1$ означает, что получена полная отрицательная связь *X* и *Y*, $BTAU = 1$ – полная положительная связь.

Коэффициент *CTAU* несколько отличается нормирующим знаменателем. С точки зрения использования отличие их в том, что *BTAU* предпочтительнее использовать для квадратных таблиц сопряженности, то есть когда $r = c$. Например, с помощью этих коэффициентов можно проверить гипотезу независимости переменных «степень противостояния СССР и Японии» и «степень альтруизма» против гипотезы их зависимости: одинаковой или противоположной упорядоченности, предварительно построив эти переменные на основе данных по нашей учебной анкете.

Рассчитаем коэффициенты *BTAU* и *CTAU* для наших переменных `V1` «Точка зрения на иностранную помощь» и `V4` «Возможность удовлетворить территориальные требований Японии». Следует заметить, что код значения «не знаю» этих переменных максимален – 4 (см. анкету в Приложении). Это нарушит порядок градаций и неясно, каким образом повлияет на результаты. Поэтому самым простым выходом будет пожертвовать данными и провести расчеты, объявив этот код кодом неопределенности:

```
MISSING VALUES v1 v4(4).
CROSSTABS /TABLES = v4 BY v1
/STATISTIC = CHISQ BTAU CTAU CMH(1)
/CELLS = COUNT ROW COL.
```

Таблица 3.7

Коэффициенты для ранговых переменных

	Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Kendall's tau-b	–0,158	0,043	–3,571	0,000
Kendall's tau-c	–0,094	0,026	–3,571	0,000

N of Valid Cases 606

Поскольку *BTAU* = – 0,158 меньше нуля и значима, можно с уверенностью утверждать, что превалирует обратная связь между рангами: чем меньше желание отдать острова, тем больше преобладает мнение, что помощь необходима. То же самое дает *CTAU*.

3.2.1.7. Точные методы оценки значимости

Что же делать, когда количество наблюдений не позволяет воспользоваться аппроксимацией распределения статистики *CHISQ* распределением хи-квадрат (больше 20 % клеток имеют $E_{ij} < 5$)? В действительности аппроксимация необходима лишь для того, чтобы можно было вычислить наблюдаемую значимость, т. е. вероятность $P\{CHISQ. > CHISQ_{\text{выбороч.}}\}$. То же касается и значимости коэффициентов *CTAU*, *BTAU*. Современная вычислительная техника позволяет во многих случаях обойтись без использования аппроксимации, вычислить вероятности за счет имитации сбора данных в условиях их независимости (метод Монте-Карло) или воспользовавшись непосредственным вычислением вероятности.

Во многих процедурах SPSS, в том числе и в `CROSSTABS`, реализованы ***метод Монте-Карло*** и ***метод прямого вычисления*** вероятностей.

В ***методе Монте-Карло*** проводятся компьютерные эксперименты, в которых многократно случайно перемешиваются данные. В каждом эксперименте вычисляется значение статистики значимости и сравнивается с ее выборочной величиной. Доля случаев, когда статистика превысила выборочное значение, является оценкой уровня значимости $P\{CHISQ. > CHISQ_{\text{выбороч.}}\}$. Поскольку здесь оценка вычисляется на основе случайных экспериментов, в дополнение к оценке уровня значимости выдается ее доверительный интервал. Число экспериментов и доверительная вероятность задается заранее.

В ***методе прямого вычисления*** рассматривается обобщение гипергеометрического распределения для таблицы сопряженности. Процедура весьма трудоемка и имеет смысл для небольших данных. Заранее задается

время счета, и, если программа не успела справиться с вычислениями, выдается результат, полученный на основе аппроксимаций.

Метод Монте-Карло практически всегда позволяет получить оценку значимости за реальное время, но с определенной точностью. Метод прямого вычисления определяет вероятность точно, но расчеты требуют слишком много времени.

В диалоговом окне CROSSTABS (как, впрочем, и в окнах для других непараметрических процедур) указанные методы включаются с помощью кнопки **Exact**.

Пример. Решается вопрос, как связаны «Точка зрения на иностранную помощь» и «Возможность удовлетворить территориальные требований Японии» на выборке, ограниченной жителями Дальнего Востока (276 наблюдений). Для решения используется

```
  CROSSTABS /TABLES = v4 BY v1 /STATISTIC = CHISQ
/CELLS = COUNT Row Col /METHOD = MC CIN(99)
SAMPLES(10000).
```

Параметры последней подкоманды «/METHOD = MC CIN(99) SAMPLES(10000)» говорят о том, что значимость оценивается методом Монте-Карло (МС), будет получен 99 %-й доверительный интервал для оценки наболюдаемой значимости (CIN(99)) с использованием 10 000 экспериментов (SAMPLES(10000)).

В результате получаем табл. 3.8, в которой размещены значимости всех исследуемых статистик. Исследуемые в эксперименте статистики включают дополнительно обобщение точного теста Фишера (Fisher's Exact Test). Статистика для этого теста имеет вид $FI = -2 \log (\gamma P)$, где γ – константа, зависящая от итоговых частот таблицы, а P – вероятность получить наблюдаемую таблицу в условиях независимости переменных. Статистика FI также имеет асимптотическое распределение хи-квадрат (в условиях гипотезы независимости). Следует заметить, что значимость, вычисленная на основе аппроксимации, выглядит значительно оптимистичнее с точки зрения обнаружения связи, чем при прямых вычислениях, да это и не мудрено – доля клеток, в которых ожидаемая частота меньше 5, равна 56,3 %, а минимальная ожидаемая частота равна 0,47.

Опыт показывает, что точный тест на основе прямого вычисления вероятности требует больших затрат времени. Для нашей задачи оказалось недостаточно 25 мин. на персональном компьютере с процессором 200 mhz.

Таблица 3.8

Хи-квадрат тесты, оценка значимости методом Монте-Карло

	Value	Df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)		
				Sig.	99 % Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	21,6	9	0,010	0,0155	0,012	0,019
Likelihood Ratio	18,9	9	0,026	0,0327	0,028	0,037
Fisher's Exact Test	19,1			0,0103	0,008	0,013
Linear-by-Linear Association	0,3	1	0,611	0,6492	0,637	0,661
N of Valid Cases	276					

a 9 cells (56,3 %) have expected count less than 5. The minimum expected count is .47.

3.3. Сложные табличные отчеты. Таблицы для неальтернативных вопросов

Получить сложные многоуровневые таблицы, содержащие описательные статистики по числовым переменным, можно, используя раздел меню **Custom Tables**. Этот раздел соответствует в языке программирования команде `TABLES`. Синтаксис этой команды весьма сложен, и при «ручном» наборе команды `TABLES` можно легко ошибиться. Поэтому здесь мы не будем даже пытаться знакомить читателя с ее текстовым заданием и рекомедуем при написании использовать преимущества диалога.

Хотя раздел меню состоит из четырех команд: **Basic Tables, General Tables, Multiple Responcse Tables и Tables of Frequencies,** мы не будем описывать все тонкости работы с этими командами, покажем лишь принципиально новые возможности по сравнению с `CROSSTABS`.

Ячейки таблицы, получаемой с помощью **Basic Tables**, соответствуют комбинациям значений переменных. В этих ячейках могут располагаться частоты, всевозможные проценты, средние по количественным переменным. Например, можно вычислить средние возраст и доход при различных сочетаниях пола, семейного положения и образования. Всего в диалоговом окне может быть задано около 30 статистик. Но нет ни одной статистики, по которой можно было бы проверить значимость связи переменных и значимость различия средних в группах. Недоступны для обработки и неальтернативные вопросы.

Команда **Tables of Frequencies** по сути объединяет в одну таблицу множество одномерных распределений одних переменных в группах по комбинациям значений других переменных и выдает только самые простые статистики – частоты и проценты.

Мы предлагаем читателю самостоятельно разобраться с простыми командами **Basic Tables** и **Tables of Frequencies**, но подробно рассмотрим команду **General Tables**, имеющую принципиальное значение для анализа неальтернативных вопросов.

3.3.1. Работа с командой General Tables

Итак, команда **General Tables** отличается тем, что с ее помощью можно обрабатывать ответы на неальтернативные вопросы и комбинации этих ответов. В клетках таблиц для неальтернативных и обычных вопросов можно также получать средние значения количественных переменных.

Для получения таблицы с использованием неальтернативных вопросов необходимо через диалоговое окно **General Tables** (см. рис. 3.5) выйти в ***окно задания списков переменных для неальтернативных вопросов*** (см. кнопку **Mult Response Sets**, рис. 3.6) и задать списки этих переменных. Словами **Dichotomies Counted Value** обозначается дихотомическое кодирование этих вопросов, словом **Categories** – кодирование в виде списка подсказок.

При вычислении процентов в таблицах для неальтернативных вопросов рассматриваются две возможности использовать в качестве знаменателя сумму ответов или число наблюдений (анкет). Причем в последнем случае берутся не все объекты, а только анкеты ответивших на соответствующий вопрос.

После задания групп переменных в списке **Mult Response** в диалоговом

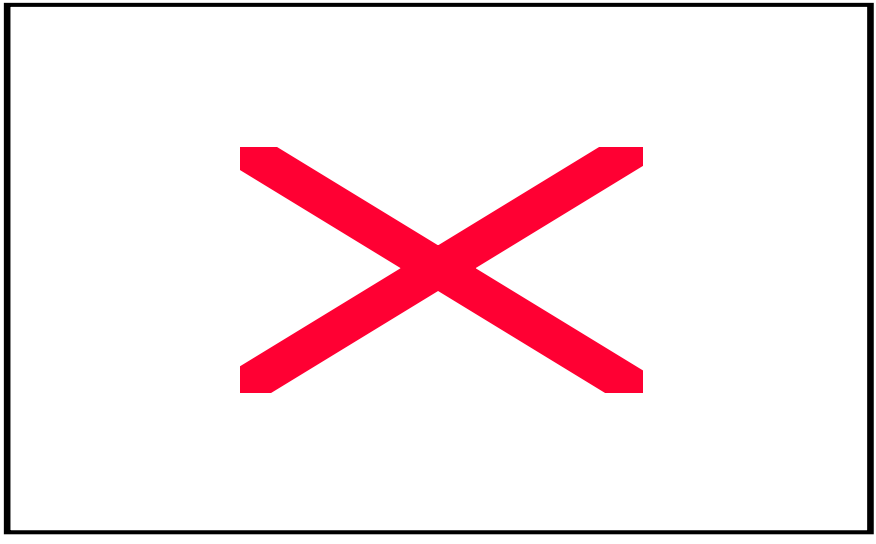


Рис. 3.5. Задание средних в General Tables

окне **General Table** появятся их имена, начинающиеся со знака доллара. Эти имена могут использоваться для задания строк, столбцов, слоев таблицы.

В SPSS начиная с 8-й версии информация о неальтернативных вопросах сохраняется в файле данных. Поэтому, если группы переменных были уже сформированы в прошлых сеансах работы с SPSS, соответствующие имена можно использовать непосредственно; при вызове **General Tables** появятся в упомянутом списке **Mult Response** без дополнительных манипуляций.

Для того чтобы в таблице были статистики количественной переменной, нужно эту переменную разместить в окно **Layers** и отметить, что она суммируема (**Is summarized** в сведенниях о выбранной переменной в основном диалоговом окне **General Tables**). По умолчанию средние выводятся в целом формате, что часто неудобно, поэтому обычно нужно его исправить, используя кнопку **Format**.

Итоговые строки и столбцы назначаются специально (кнопка **Totals**). При вычислении частотных таблиц следует позаботиться о задании процентов в числе статистик. Не забудьте, что частотные таблицы без задания процентов в большинстве случаев бессмысленны.

Следует обратить внимание, что в **General tables** *итоговые строки и столбцы таблицы формируются по сумме ответов*. Поэтому итоговые средние подсчитываются некорректно.

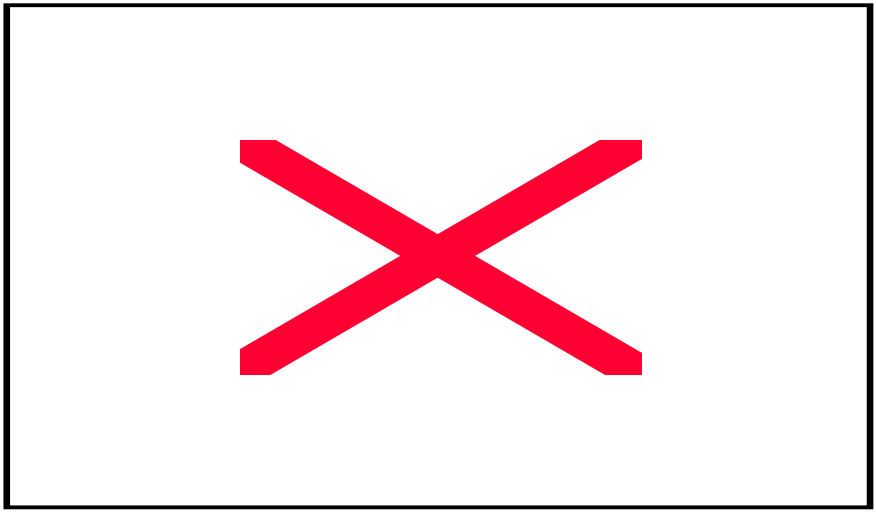


Рис. 3.6. Задание списков переменных

Пример. Синтаксис команд для выполнения расчета среднего возраста в группах по ответам на вопрос III «Что мешает заключить договор?» для мужчин и женщин имеет следующий вид:

```
   * General Tables.
TABLES /OBSERVATION = v9 /MRGROUP $v3 v3s1 to v3s8
   /GBASE = CASES /FTOTAL =  $t000001 "Total"
$t000003 "Total"
   /TABLE = $v3 + $t000001 BY v8 > (STATISTICS) +
$t000003 BY v9
   /STATISTICS MEAN(v9(COMMA7.1)) VALIDN(v9(COMMA5.0)).
```

Результат представлен на табл. 3.9. Самая «старая» группа – те, кто считает, что мешает взаимное недоверие, для респондентов как мужского пола, так и женского. К сожалению, насколько это отличие статистически значимо, выяснить по полученной таблице невозможно, так как отсутствуют соответствующие статистики (такие как *t* Стьюдента).

Таблица 3.9

Средний возраст в группах по ответам на вопрос III «Что мешает заключить договор?» для мужчин и женщин

$V3	Пол				Total	
	1 муж.		2 жен.		Mean	Valid N
	Mean	Valid N	Mean	Valid N		
1 нет необх.	38,0	38	40,5	22	38,9	60
2 недоверие	45,4	41	44,0	45	44,7	86
3 незаинт. Яп.	37,4	32	36,5	56	36,8	88
4 разн. полит.	39,8	41	36,5	30	38,4	71
5 непризн. гр.	39,8	163	40,8	151	40,2	314
6 нежел. СССР	38,2	82	39,3	61	38,7	143
7 другое	38,6	5	44,3	3	40,8	8
8 не знаю	35,0	24	36,5	53	36,0	77
Total	39,4	426	39,5	421	39,4	847

Обратите внимание, что общая сумма здесь составила 847 ответов, что на 135 больше, чем объектов в выборке. Это произошло из-за того, что один респондент может дать несколько ответов.

Команда **Multiple Response Tables**, по сути, несколько облегченный вариант **General Tables**, предназначенный для счета частотных таблиц.

3.3.2. Типичные примеры использования Multiple Response Tables

Пример* 1. *Подготовка дихотомически закодированного неальтернативного признака.

В анкете имеются вопросы «Сколько лет проживали

14. В Западной Сибири?

15. В Восточной Сибири?

16. На Дальнем Востоке? »

Рассмотрим, как можно получить в одной таблице распределение по неальтернативному признаку «Места проживания», полученному по ответам на эти вопросы. Элементарные дихотомические переменные, соответствующие данному признаку, можно построить с помощью следующих команд:

```
COMPUTE m1 = V14.
COMPUTE m2 = V15.
COMPUTE m3 = V16.
RECODE m1 m2 m3 (1 THR HI = 1).
VAR LAB m1 "Зап Сиб" m2 "Вост Сиб" m3 "Дальн Вост".
* General Tables.
TABLES
/MRGROUP $v3 'Мешает договору' v3s1 to v3s8
/MDGROUP $region m1 m2 m3 ( 1 )
/GBASE = RESPONSES
 /FTOTAL =  $t000005 "Total" $t000006 "Total"
 /TABLE = $region + $t000005 BY $v3 + $t000006
 /STATISTICS  count( $v3( F5.0 ))
 rpct( $v3( PCT5.1 ) 'Row Response %':$region )
 rpct( $v3( PCT5.1 ) 'Col Response %':$v3 ).
```

Пример* 2. *Объединение подсказок в неальтернативном признаке, закодированном в виде списка. Объединение подсказок можно сделать за счет приведения этих переменных в дихотомическую форму.

Задача: объединить в 7-м вопросе ответы: «продажа островов» и «продажа с компенсацией» и исследовать его связь с регионом проживания респондента (переменная R). Для этого следует выполнить программу:

```
COUNT D1 = V7S1 TO V7S7 (1)/
D2 = V7S1 TO V7S7 (2,3)/
D3 = V7S1 TO V7S7(4 TO 10).
RECODE D1 TO D3(1 THR 10 = 1).
*метки переменных.
  VAR LAB D1 'Жесткий вариант'
          D2 'Совместное использование'
          D3 'мягкий вариант'.
```

```
  TABLES MDGROUPS D "Степень жесткости позиции"
D1 D2 D3(1)
  /TABLES D+T BY R+T/ STAT COUNT(D) CPCT(D:D) CPCT(D:R).
```

3.4. Множественные сравнения в таблицах для неальтернативных вопросов. Программа Typology Tables

Как уже было отмечено, в сложных табличных отчетах SPSS отсутствуют статистики значимости. Это касается также таблиц для неальтернативных вопросов. Этот пробел восполнила программа **Typology Tables**, разработанная в Институте экономики и ОПП СО РАН, г. Новосибирск (исследование финансировалось грантом РФФИ № 00-06-80221).

В программе рассматриваются двумерные таблицы частотных распределений и таблицы средних по количественным переменным в группах по сочетаниям ответов на неальтернативные вопросы. Исследуется значимость отклонений частот от ожидаемых в условиях независимости ответов на два вопроса и отклонений эмпирических средних от теоретических средних в итоговых ячейках. Эта программа может быть вставлена пунктом командой меню в SPSS версий 8, 9, 10.

3.4.1. *Z*-статистика значимости отклонения частот

В качестве статистики значимости используется асимптотически нормально ($\sim N(0,1)$) распределенная статистика $Z = (N_{11} - E_{11}) / \sigma$. Мы уже рассматривали эту статистику под названием `ASRESID` (**Adjusted residuals**) в `CROSSTABS`. Для малых выборок эта статистика корректируется на основе прямого вычисления вероятностей так, чтобы для нее выполнялись соотношения нормального распределения.

3.4.2. *Z* -статистика отклонения средних

При анализе средних в таблицах для неальтернативных признаков каждая ячейка рассматривается по отдельности, при этом среднее в группе, соответствующей ячейке, сравнивается со средними по объектам, не содержащимся в группе.

Обозначим A совокупность объектов, соответствующую i-му ответу вертикального и j-му ответу горизонтального вопроса, B – ее дополнение. Число объектов в группе A равно $N_A = N_{ij}$. Группа объектов B может иметь разное содержание в зависимости от того, с чем мы хотим сравнить среднее в этой группе: 1) со средним по всей совокупности, тогда B – дополнение A до всей совокупности и содержит $N_B = N - N_{ij}$ объектов; 2) с итоговым средним по строке, тогда B – дополнение A до i-й группы по вертикальному вопросу, а $N_B = N_{i.} - N_{ij}$; 3) с итоговым средним по столбцу, тогда B – дополнение A до j-й группы по горизонтальному вопросу, а $N_B = N_{.j} - N_{ij}$.

Для проверки значимости различия средних в группах A и B в предположении теоретического нормального распределения, при несовпадении

дисперсии в группах используется статистика $Z = \frac{\left(\overline{X}_A - \overline{X}_B\right)}{S\sqrt{1/N_A + 1/N_B}}$, где S – оценка дисперсии X на исследуемой совокупности. Автором данного методического пособия показано, что в случае, когда случайные величины $\overline{X}_A$ и $\overline{X}_B$ получаются за счет перемешивания выборки по переменной X, знаменатель выражения для Z является точным значением дисперсии разности $\overline{X}_A$ и $\overline{X}_B$. Распределение Z асимптотически нормально.

3.4.3. Как выяснить надежность результата?

В соответствии с общепринятым использованием 5 %-го уровня значимости мы можем заявить, что величина стандартизованного смещения Z, превышающая 1,96, свидетельствует о существенности связи (вероятность в условиях независимости получить большее смещение равна 5 %, см. выделенные клетки со значимыми смещениями в табл. 2). Однако это утверждение о значимости верно только для отдельно взятой клетки таблицы, как мы ранее показали, вероятность того, что в этой таблице из 100 независимых клеток имеется хотя бы одна «значимая» статистика, равна 99,41. Это результат множественных сравнений статистик.

Чтобы снизить вероятность принятия случайных отклонений за закономерные, нужно использовать более жесткий критерий, хотя, конечно, и обычное применение Z-статистик позволяет избежать очевидных ошибок.

К сожалению, таблицу с Z-статистиками, подобную табл. 2, обычными средствами статистических пакетов получить сложно, поскольку в них нет средств анализа значимости по неальтернативным вопросам.

3.4.4. Критические значения *Z*-статистики при множественных сравнениях

Для выяснения значимости вычисляется *критическое значение максимальной по модулю Z-статистики таблицы* ($\max |Z_{ij}|$), и значимыми считаем Z_{ij}, превышающие это значение. Как обычно, критическое значение выбирается так, чтобы вероятность случайно его превзойти была равна заданному значению (обычно 5 %).

3.4.5. Статистические эксперименты

Для выяснения критического значения $\max |Z_{ij}|$ многократно (заданное число раз) имитируется ситуация независимости ответов, соответствующих строкам и столбцам. В ходе имитации в клетках таблицы получаются значения Z-статистик. Такая имитация осуществляется за счет случайного перемешивания данных, которое можно представить так, будто мы рассыпали листочки с разными вопросами анкеты и собираем их вместе в случайном порядке.

По эмпирической функции распределения получаются критические значения для максимума *Z*-статистики.

Эксперименты позволяют также оценить в каждой клетке *наблюдаемую множественную значимость Z-статистики – вероятность на всей таблице случайно получить большее значение Z-статистики.*

3.4.6. Работа с программой Typology Tables

Коротко статистический анализ таблиц при помощи Typology Tables можно представить последовательностью следующих естественных действий.

– Задание групповых переменных

– Выбор переменных для строк, столбцов, если необходимо – переменных для вычисления средних и условий (слоев).

– Выбор таблицы сопряженности или средних (на основе числа валидных («немиссинговых») объектов внутри таблицы.

– Статистический эксперимент.

– Выдача результатов. Программа может выводить результат в текстовый файл, формат, применяемый в Интернет (HTML) и в виде файла EXCEL.

Каждое из этих действий в программе обеспечено своей экранной формой; переход от одной формы к другой происходит естественным путем (запуском очередных расчетов) или с помощью специальных кнопок-переключателей.

3.4.7. Примеры использования программы Typology Tables

В информации RLMS содержатся сведения о покупках 3 700 семей, сделанных в течение 1 недели (молочных продуктов, спиртного и табачных изделий, сладостей и др.), о размерах жилья и имеющихся в жилье удобствах, о наличии в семье дорогостоящих предметов и недвижимости.

3.4.7.1. Частотная таблица. Наличие крупной собственности и покупки спиртного и табака.

Связаны ли ответы о покупках спиртного и табака с наличием автомобиля, дачи и других предметов длительного пользования? Этот вопрос мы проанализируем с помощью **Typology Tables**. Табл. 3.10, полученная по совокупности городских семей (подвыборка из RLMS 2604 семей), показывает такую связь. Строки таблицы соответствуют ответам по вопросу о благосостоянии, столбцы – ответам по вопросу о пристрастиях к напиткам и курению. Отличие таблицы для неальтернативных признаков от обычной таблицы частот заключается в том, что группы объектов (семей), соответствующие разным ответам, могут пересекаться.

Явно видно, что в семьях, владеющих крупной собственностью, употребляют больше алкоголя и табака (может быть, сказывается наличие в них

большего числа мужчин?). Однако насколько надежен этот вывод? Особенно для группы владельцев грузового автомобиля – уж слишком мала эта группа для надежных выводов.

Таблица 3.10

Покупка алкоголя и табачных изделий и наличие крупной собственности (фрагмент таблицы сопряженности, частоты и % по строкам)

Крупная собственность	Алкоголь и табачные изделия				Итого
	Пиво	Табачные изделия	Водка	Вина	
Легковой автомобиль	169 26,2%	313 48,5%	142 22.00%	60 9,3%	646 100%
Наличие второй квартиры	46 27,2%	93 55.00%	40 23,7%	16 9,5%	169 100%
Грузовой автомобиль	10 38,5%	12 46,2%	7 26,9%	2 7,7%	26 100%
Итого	462 17,7%	1076 41,3%	415 15,9%	175 6,7%	2604 100%

Z-статистики в табл. 3.11 показывают значимость связей некоторых ответов. Однако множественные сравнения не позволяют полностью доверять этим результатам.

Таблица 3.11

***Z*-статистики и значимость (%) связи покупки алкоголя и табачных изделий и наличие крупной собственности (фрагмент таблицы, *Z*-статистики)**

Крупная собственность	Алкоголь и табачные изделия			
	Пиво	Табачные изделия	Водка	Вина
Легковой автомобиль	6,46 0,00%	4,24 0,00%	4,84 0,00%	3,01 0,26%
Другая квартира	3,33 0,09%	3,74 0,02%	2,84 0,45%	1,47 14,16%
Грузовой автомобиль	2,49 1,28%	0,5 61,71%	1,44 14,99%	0,28 77,95%

В табл. 3.12 отмечены значимые с точки зрения множественных сравнений *Z*-статистики. При этом оценка 5 %-го критического значения *Z* равна 3,09, а не 1,96, как это было бы в обычном анализе.

В каждой клетке расположены также наблюдаемые множественные значимости. Например, *Z*-статистика 6,46 в клетке «Легковой автомобиль – пиво» практически не может быть получена случайно (вероятность получить большее значение равна нулю). Связь, характеризуемая значением Z = 2,84 в клетке «Наличие второй квартиры – водка» – под сомнением: такие и большие значения в одной из 28 клеток таблицы можно получить случайно с вероятностью 10,8 %. С точки зрения обычного анализа эта связь существенна, с точки зрения множественных сравнений не существенна.

Таблица 3.12

***Z*-статистики отклонений частот и их наблюдаемая множественная значимость (в %, 5 %-е критическое значение max $|Z_{ij}| = 3{,}09$)**

Крупная Собственность	Алкоголь и табачные изделия			
	Пиво	Табачные изделия	Водка	Вина
Легковой автомобиль	6,46 0,00%	4,24 0,1%	4,84 0,00%	3,01 6,5%
Наличие второй квартиры	3,33 2,3%	3,74 0,5%	2,84 10,8%	1,47 97,3%
Грузовой автомобиль	2,49 27,1%	0,5 100,00%	1,44 97,8%	0,28 100,00%
Мотоцикл	1,46 97,4%	2,72 15,1%	2,3 40,9%	0,95 100,00%
Трактор	0,8 100,00%	-0,3 100,00%	1,63 92,4%	0,38 100,00%
Садовый Домик	-0,3 100,00%	-0,5 100,00%	0,42 100,00%	-1,73 87,4%
Дача	2,45 29,7%	1,34 99,1%	2,29 41,8%	2,21 48,5%

3.4.7.2. Таблица средних. Молочные продукты и жилплощадь

Некоторые товары настолько общеупотребительны, что их покупает каждая семья, другие чаще приобретаются семьями с детьми, третьи товары берут для стариков, и т. п. Молодые семьи обычно имеют маленьких детей и часто нуждаются в жилплощади. Можно ли по косвенному признаку – жилплощади выяснить, какие молочные товары приобретаются семьей? Для исследования подобных вопросов в клетках таблицы для неальтернативных вопросов размещаются средние значения количественной переменной. В табл. 3.13. размещена средняя жилплощадь в пересекающихся группах семей по покупкам молочных продуктов. Эта таблица показывает, что городские семьи, покупающие кисломолочные продукты, имеют в среднем меньшую, а семьи, покупающие сухое молоко, – большую жилплощадь. Но, может быть, это не закономерность, а игра случая?

Таблица 3.13

Средняя жилплощадь в группах семей по покупкам молочных продуктов

Показатель	По всей выборке	Молочные продукты						
		Сухое молоко	Молоко	Кисло-молочные продукты	Сметана, сливки	Масло животное	Творог, сырковая масса	Сыр, брынза
Полезн. Жилплощадь кв.м.на чел.	7,7	8,2	7,7	7,3	7,6	7,7	7,6	7,7
Стандартное отклонение	4,4	5,1	4,4	4,1	4,3	4,5	4,3	4,1
Число наблюдений	2495	219	1668	784	807	870	586	485

Определить, какое смещение значимо, а какое – нет помогут множественные сравнения *Z*-статистик отклонения средних в клетках от среднего по всей совокупности (см. табл. 3.14). В ней выделена единственная значимая на 5 %-м уровне клетка, показывающая относительно малую обеспеченность жилплощадью покупателей кисломолочных продуктов (скорее всего, эти покупатели из молодых семей с детьми). Абсолютная величина ее значения (–2,87) случайно может быть перекрыта лишь с вероятностью 0,029 (наблюдаемая множественная значимость равна 2,9 %).

Таблица 3.14

***Z*-статистики отклонений средних для таблицы 3,13**
(5 %-е множественное критическое значение равно 2,69)

Показатель	Молочные продукты						
	Сухое молоко	Молоко	Кисло-молочные продукты	Сметана, сливки	Масло животное	Творог, сырковая масса	Сыр, брынза
Z-отклонения среднего	1,43	-0,58	-2,87	-0,48	0,32	-0,45	0,3
Множественная наблюдаемая значимость	66,6%	99,5%	2,9%	99,8%	100,0%	99,9%	100,0%

3.4.7.3. Душевой доход любителей сладкого и жилье. Одновременное сравнение средних по строкам таблицы

Насколько отличаются доходы потребителей сладкого внутри групп, поразному обеспеченных жильем: имеющих квартиру, свой дом, часть квартиры и др.?

Для выяснения этого изучим средние логарифмы доходов, так как для получения устойчивых результатов в таких исследованиях лучше использовать логарифм дохода.

Из табл. 3.15 видно, что обладатели отдельных квартир самые богатые, отдельного дома – чуть победнее (скорее всего, это обитатели городских окраин), а те, кто имеет часть дома или квартиры, – самые бедные. У них разные условия существования, и полезно изучить эти группы по отдельности. Это значит, что смещение средних в клетках таблицы нужно рассмотреть не по отношению к общему среднему (5,6), а по отношению к итогам по строкам (например, существенно ли выделяются по доходам среди обитателей домов (средний логарифм дохода равен 5,5) любители мороженого (средний логарифм дохода равен 5,9)?).

Таблица 3.15

Средний логарифм доходов в группах по жилищным условиям и покупкам сладкого (среднее, стд. отклонение, численность в группах)

Тип жилья	Сладости						В целом по выборке
	Мороженое	Сахар	Конфеты, шоколад	Печенье, пирожные	Варенье, джем	Мед	
Отдельная квартира	5,9 0,9 321	5,6 0,8 518	5,9 0,8 477	5,9 0,9 575	6,3 0,8 17	6 1 57	5,6 0,9 1841
Отдельный дом	5,9 0,8 22	5,8 0,8 52	5,7 0,9 44	5,8 0,8 48	, , 0	5,9 1,1 6	5,5 0,8 211
Часть квартиры	5,9 0,9 10	5,4 0,8 35	5,7 0,8 23	5,7 0,9 24	, , 0	4,8 0,7 3	5,3 0,8 96
Часть дома	6,2 0,6 5	5,3 1 17	5,8 1 7	5,6 0,7 12	, , 0	5,4 0 1	5,3 0,8 54
В целом по выборке	5,9 0,9 423	5,6 0,9 691	5,9 0,9 633	5,9 0,9 748	6,3 0,8 18	6 1 74	5,6 0,9 2444

Таким образом, мы одновременно рассматриваем *Z*-статистики для каждой группы и проводим множественные сравнения 21 смещения средних[1]. Способы определения значимости смещений в двумерной таблице и одномерной таблице средних идентичны, здесь также используется перемешивание данных по зависимой переменной.

Таблица 3.16

***Z*-статистики отклонений средних для таблицы 3,15 (5 %-е множественное критическое значение равно 3,1)**

[1] Покупателями джема и варенья оказались только жители отдельных квартир, поэтому для части клеток таблицы средние и, следовательно, *Z*-статистики их отклонений не определены.

Жилье\Сладости	Мороженое	Сахар	Конфеты, шоколад	Печенье, пирожные	Варенье, джем	Мед
Отдельная квартира	6,64 0,0%	-0,39 100,0%	9,01 0,0%	9,49 0,0%	2,7 15,0%	2,57 20,7%
Отдельный дом	2,35 33,5%	3,1 5,0%	2,03 58,5%	2,93 8,1%	,	0,9 100,0%
Часть квартиры	1,77 79,1%	0,77 100,0%	2,55 21,7%	2,21 43,8%	,	-0,8 100,0%
Часть дома	2,13 50,2%	-0,14 100,0%	1,22 99,1%	1,07 99,8%	,	,

На основании табл. 3.16 можно достоверно утверждать, что среди обитателей отдельных квартир большие доходы имеют семьи любителей мороженого, конфет и печенья с пирожными; среди жильцов отдельных домов существенно выделяются по доходам семьи у покупателей сахара (только в 5 % случаев в таблице случайно можно получить большие *Z*-статистики). В остальных клетках таблицы *Z*-статистики незначимы – либо отклонения несущественны, либо выборка маловата, чтобы делать надежные выводы.

Глава 4. СРАВНЕНИЕ СРЕДНИХ, КОРРЕЛЯЦИИ

4.1. Compare Means – простые параметрические методы сравнения средних

Параметрические методы проверки гипотез о равенстве средних нулю (нулевые гипотезы) предполагают нормальность распределения анализируемых переменных или остатков в моделях дисперсионного анализа, сравнения средних в парах групп объектов и т. д. Однако условие нормальности выборки при анализе анкетной информации выполняется весьма редко. Наиболее доступным решением проблемы является создание новых переменных путем усреднения множества независимых случайных данных. По центральной предельной теореме такие переменные имеют распределение, близкое к нормальному.

На практике эти методы все же используются для больших совокупностей данных других типов распределений при условии, что они «не слишком сильно» отклоняются от нормального распределения. «Не слишком сильно» – неопределенное понятие, обычно решение принимается при рассмотрении гистограммы распределения на фоне кривой нормального распределения.

Взгляните, например, на распределение населения по душевому доходу, рис. 4.1. Распределение имеет длинный хвост в направлении больших доходов, нормальная кривая недостаточно хорошо огибает гистограмму. Если использовать вместо этой переменной логарифм доходов, полученный ко-

мандой `COMPUTE lnv14 = ln(v14).`, то получаем более приемлемое распределение (рис. 4.2).

Основные идеи и формулы параметрических методов анализа средних и дисперсий рассматриваются в курсе математической статистики; и здесь, по ходу изложения материала, мы коротко напомним отдельные положения этой теории.

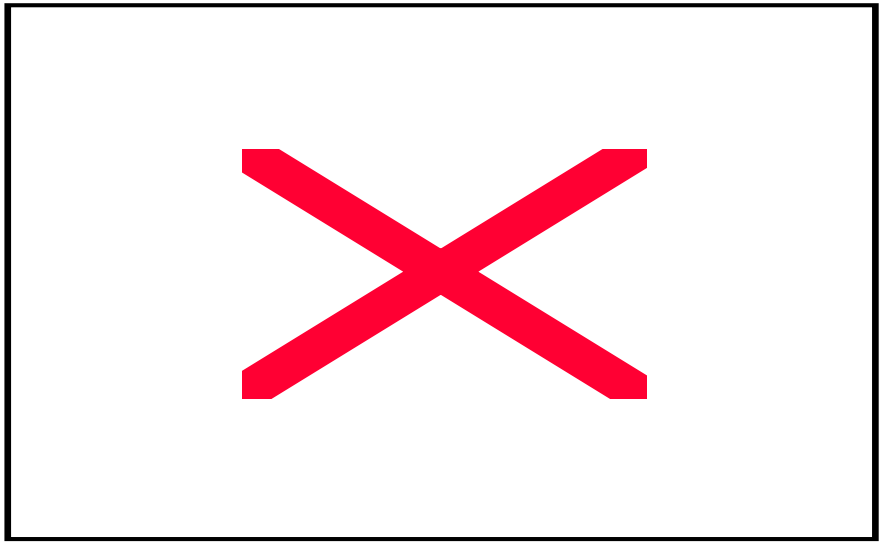


Рис. 4.1. Распределение населения по душевому доходу

4.1.1. Одновыборочный тест (One sample T-test)

Одновыборочный t-тест предназначен для проверки гипотезы о равенстве математического ожидания переменной заданной величине (в общепринятых обозначениях H_0**:** $\mu = \mu_0$). Напомним, что для проверки этой гипотезы используется статистика $t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}}$, которая распределена по закону Стьюдента с $n - 1$ степенями свободы.

Команда для проверки гипотезы выдает двусторонний доверительный интервал для μ).

Примеры применения одновыборочного t-теста.

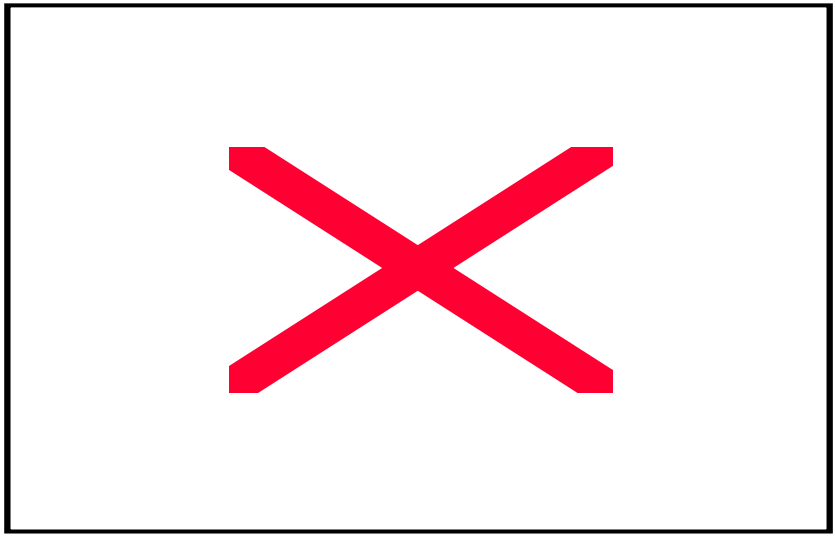


Рис. 4.2. Распределение, близкое к нормальному

***Пример* 1**. Для элиминирования влияния инфляции на измерение доходов его нормируют, измеряя в относительных единицах – числе средних или медиан. Доход, отнесенный к величине медианы, называется промедианным доходом. Оценка медианы душевых доходов населения по ранее проведенному достаточно обширному обследованию – 200 р. Если допустить, что логарифм доходов имеет нормальное распределение, то среднее логарифма промедианных доходов должно незначимо отличаться от нуля (поскольку нормальное распределение симметрично относительно математического ожидания). Проверим это:

```
   COMPUTE lnv14m = ln(v14/200).
   VARIABLE LABELS lnv14m "логарифм промедианного до-
хода".
   T-TEST /TESTVAL = 0 /VARIABLES = lnv14m /CRITERIA =
CIN (.95).
```

Таблица 4.1

Одновыборочный t-тест. Средний промедианный доход незначимо отличается от нуля

	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
LNV14M	–0,831	672	0,406	–0,017	–0,058	0,023

В нашем примере $\mu_0 = 0$ (TESTVAL = 0), отклонение среднего равно – 0,017, наблюдаемая значимость – 0,406 (почти в 40 % случаев большее отклонение от ожидаемого значения может быть получено случайно), поэтому гипотеза о равенстве нулю матожидания логарифма промедианного дохода не отклоняется. Об этом же говорит и тот факт, что 95 %-й доверительный интервал покрывает ожидаемое значение. Таким образом, по указанному параметру распределение доходов похоже на логарифмически нормальное.

***Пример* 2**. Есть предположение, что малообразованное население имеет доход, существенно меньший, чем доход более образованной его части. Это утверждение не абсолютно, а выполняется «в среднем». Мы проверим его, исследовав различие средних логарифмов доходов в указанных группах. По существу это означает сравнение средних геометрических дохода. В нашей анкете образование закодировано следующим образом:

1. Высшее;
2. Незаконченное высшее;
3. Среднее специальное;
4. ПТУ, ФЗУ;
5. 10 – 11кл;
6. 7 – 9 кл.;
7. 4 – 6 кл.;
8. Менее 4 кл.;
9. Нет образования.

Проверим предположение, воспользовавшись временной выборкой данных о респондентах, имеющих образование не выше среднего.

```
COMPUTE f = (v10 > 3).
*формирование переменной фильтра.
FILTER f.
T-TEST /TESTVAL = 0 / VARIABLES = lnv14 /CRITERIA =
CIN (.95).
FILTER OFF.
```

Таблица 4.2

Одновыборочный t-тест. Средний логарифм промедианного дохода в группе с относительно низким образованием отличается от нуля при уровне значимости 5 %

	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
LNV14	–2,0316	162	0,0438	–0,0956	–0,1886	-0,0027

4.1.2. Двухвыборочный t-тест (independent sample t-TEST)

Для сравнения средних в двух выборках необходимо выполненить процедуру `T-TEST` в следующем виде:

```
T-TEST/GROUPS V4(1,3)/VARIABLES = V9 lnV14m.
```

Подкоманда `GROUPS` указывает переменную группирования; в скобках задаются два значения этой переменной, определяющие группы. Например, приведенная команда будет выполняться только для групп объектов, у которых `V4` принимает указанные значения 1 и 3. `VARIABLES` задает сравниваемые (зависимые) переменные для выделенных групп объектов. Объекты можно также разбить на две группы, указав в параметре `GROUPS` одно значение:

```
T-TEST /GRO v9(30)/VAR V9 lnV14m.
```

В этом случае вся совокупность будет разделена на те объекты, на которых указанная переменная не больше заданного значения (`v9` ≤ 30), и те, у которых она больше (`v9` > 30).

Процедура `T-TEST` проверяет гипотезу равенства средних в двух выборках при условии, что генеральная совокупность имеет нормальное распределение. Процедура для пары групп подсчитывает средние, стандартные ошибки, статистики и их значимость. При сравнении двух выборок нас интересует, насколько случайный характер носит различие средних, т. е. отличаются ли они значимо?

В зависимости от предположения о равенстве дисперсий испльзуются разные варианты t-статистик.

Если равенство дисперсий в группах не предполагается, то для сравнения средних принято использовать статистику

$$t = \left(\overline{X}_1 - \overline{X}_2\right) \Big/ \sqrt{S_1^2 \big/ n_1 + S_2^2 \big/ n_2} \ ,$$

которая в условиях гипотезы равенства матожиданий и нормальности X имеет распределение, близкое к распределению Стьюдента.

Если заранее известно о равенстве дисперсий в группах, то предпочтительнее статистика

$$t = \left(\overline{X}_1 - \overline{X}_2\right) \Big/ \sqrt{S_p^2 \left(1/n_1 + 1/n_2\right)} \ .$$

При определении ее величины предварительно вычисляется объединенная дисперсия

$$S_p^2 = \left\{ S_1^2 \left(N_1 - 1\right) + S_2^2 \left(N_2 - 1\right) \right\} \Big/ \left(N_1 + N_2 - 2\right).$$

Из теории известно, что при условии равенства дисперсий вычисляемая величина S_p есть несмещенная оценка дисперсии, и статистика t также имеет распределение Стьюдента.

Для проверки равенства дисперсий используется статистика Ливиня, имеющая распределение Фишера.

Двусторонней наблюдаемой значимостью, вычисляемой процедурой `T-TEST`, является вероятность в условиях гипотезы равенства матожиданий случайно получить большее значение статистики t:

Sig = $P\{ \mid$ *t-теоретическое* $\mid > \mid$ *t-выборочное* $\mid \}$.

Если значимость близка к 0, делаем вывод о неслучайном характере различий средних значений в выборках.

Результат выдается в двух таблицах. В первой размещены средние и характеристики разброса в группах, во второй – результаты их сравнения.

Таблица 4.3

Т-тест, описательные статистики по группам

	V9 Возраст	N	Mean	Std. Deviation	Std. Error Mean
LNV14M	>= 30	521	0,019	0,517	0,023
	< 30	133	–0,177	0,593	0,051

Таблица 4.4

Т-тест, сравнение средних и дисперсий в группах

	Levene's Test for Equality of Variances		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
	F	Sig.						Lower	Upper
Equal variances assumed	2,47	0,1162	3,78	652	0,000	0,196	0,052	0,094	0,298
Equal variances not assumed			3,48	186,42	0,001	0,196	0,056	0,085	0,307

В табл. 4.3 и 4.4 приведен пример сравнения средних логарифмов душевых доходов в группах населения до 30 лет и старше. Статистика Ливиня в этом случае свидетельствует, что гипотеза равенства дисперсий не отвергается (sig = .1162) Поэтому для сравнения средних можно воспользоваться строкой «Equal variances assumed» – «Предполагаются равные дисперсии». Соответствующая статистика показывает, что средние различиются существенно (sig = 0,000). Впрочем, даже если бы мы не были удовлетворены статистикой Ливиня, в данном случае и без предположения равенства дис-

персий мы можем утверждать то же самое, так как (sig = 0,001). Кроме того, это подтверждает и доверительный интервал, не включающий нуля.

4.1.3. Двухвыборочный t-тест для связанных выборок (Paired sample T-TEST)

Если на одних и тех же объектах дважды измеряется некоторое свойство, то проверка значимости различия средних по измеренным переменным осуществляется этим тестом. Пример задания команды:

```
T-TEST PAIRS = x WITH y (PAIRED) /CRITERIA = CIN(.95).
```

Переменные `x` и `y` могут быть характеристиками мужа и жены при исследовании семей; по данным RLMS – измерениями, связанными с потреблением напитков в 1996 и 1998 г., и т. п. Поэтому данная процедура полезна для анализа панельных данных.

Почему же здесь нельзя воспользоваться таким же анализом, как и для двух несвязанных выборок, считая, что имеются две выборки одинакового объема?

Проверка значимости различия матожиданий x и y эквивалентна проверке гипотезы о равенстве нулю математического ожидания разности $x - y$. Дисперсия разности $x - y$ равна $D(x - y) = D(x) + D(y) - 2\,\mathrm{cov}(X, Y)$. Отсюда точность оценки матожиданий $x - y$ связана с ковариацией x и y.

Поэтому наряду с соответствующей статистикой в выдачу по этому тесту входит и коэффициент корреляции этих переменных и наблюдаемая значимость.

Для примера взгляните на выдачу, в которой сравниваются вес 1995 и 1996 г. женщин от 30 до 40 лет (в 1995 г.), табл. 4.5 – 7, данные RLMS.

Таблица 4.5

T-тест на связанных выборках, описательные статистики

	Mean	N	Std. Deviation	Std. Error Mean
AM1 Вес 1995	67,59	793	13,72	0,49
BM1 Вес 1996	68,12	793	14,22	0,50

Таблица 4.6

T-тест на связанных выборках, корреляции

	N	Correlation	Sig.
AM1 Вес 1995 & BM1 Вес 1996	793	0,914	0,0000

Таблица 4.7

T-тест на связанных выборках, сравнение средних

	Paired Differences Mean	Std. Deviation	Std. Error Mean	95 % Confidence Interval of the Difference	t	df	Sig. (2-tailed)

				Lower	Upper			
AM1 Вес 1995 & BM1 Вес 1996	–0,53	5,81	0,21	–0,93	–0,12	–2,547	792	0,011

Женщины в среднем набрали по полкилограмма веса, и этот прирост статистически значим. Значим и коэффициент корреляции – вес в целом имеет свойство сохраняться.

4.1.4. Команда MEANS – сравнение характеристик числовой переменной по группам

Процедура вычисляет одномерные статистики в группах – все описательные статистики, которые вычислялись командами DESCRIPTIVES и FREQUENCIES, а также гармоническое среднее, среднее геометрическое, проценты сумм значений переменных в группах и др. – всего 20 характеристик. Поэтому имя команды MEANS (средние) сохранилось лишь «исторически», оно пришло из ранних версий SPSS, где ее назначением, действительно, было сравнение средних. В диалоговом окне для назначения статистик используется кнопка **Options**. Проводится также одномерный дисперсионный анализ.

```
MEANS TABLES = v14 BY v11 BY v8 /CELLS MEAN STDDEV
MEDIAN COUNT /STATISTICS ANOVA.
```

В команде указывается список зависимых переменных, BY и список переменных, определяющих группы. Каждое дополнительное слово BY порождает следующий нижний уровень группирования, в диалоговом режиме слову BY соответствует кнопка **Next**.

Анализ результатов (табл. 4.8) позволяет сделать следующие выводы. Самый высокий среднемесячный доход (332 р.) имеют разведенные мужчины, при этом он значительно превосходит среднемесячный доход, полученный всеми разведенными (249 р.) и всеми мужчинами (238 р.). На втором месте по доходам находятся вдовцы (276 р.), но их всего 5 человек, поэтому цифра ненадежна. Среди женщин наиболее высокие среднемесячные доходы (226 р.) у состоящих в браке, что почти равно доходам женатых мужчин. Это естественно – ведь это же душевой доход в семье.

Мы можем сколько угодно описывать эту таблицу, но описание не будет доказательством какой-либо истины, пока оно не подтверждено статистическим выводом. Такая таблица может быть только источником гипотез о взаимосвязи, которые в дальнейшем следует проверить.

Одномерный дисперсионный анализ здесь проводится только по переменным первого уровня задания групп.

Напомним, что суть этого анализа состоит в вычислениии межгруппового квадратичныого разброса зависимой переменной $SS_в$ (*Between groups*) и внутригруппового разброса, обозначается SS_w (*Within groups*). Величина

$SS_в$ характеризует, насколько сильно отклонились от общего среднего средние между группами, а SS_w – отклонения от центров групп. Статистика $F = \frac{SS_b / k}{SS_w /(n-k)}$ в условиях гипотезы равенства средних и дисперсий распределения при нормальном распределении X в группах имеет распределение Фишера. F представляет собой в определенном смысле расстояние наблюдаемой таблицы от таблицы, в которой нет никаких зависимостей, т. е. средние в группах совпадают. Чем больше F, тем существеннее зависимость, однако сама по себе величина F ни о чем не говорит. Ответ на вопрос дает, как обычно, величина наблюдаемой значимости F-критерия: *Significance* – вероятность случайно получить значение F, большее выборочного $\mathrm{Sig} = \mathrm{P}\{F > F_{\text{выб.}}\}$.

Таблица 4.8

Среднемесячный душевой доход в семье

V11 Состояние в браке	V8 Пол	Mean	Std. Deviation	Median	N
1 женат	1 муж.	228,4	152,9	200	271
	2 жен.	225,7	140,8	200	242
	Total	227,1	147,2	200	513
2 вдовец	1 муж.	276,0	111,0	270	5
	2 жен.	192,8	112,7	155	20
	Total	209,4	115,1	168	25
3 разведен	1 муж.	331,9	230,0	295	16
	2 жен.	195,9	86,1	180	25
	Total	249,0	169,7	200	41
4 не был	1 муж.	263,3	223,0	200	41
	2 жен.	212,2	118,6	200	34
	Total	240,2	183,9	200	75
Total	1 муж.	238,4	167,8	200	333
	2 жен.	219,9	133,4	200	321
	Total	229,3	152,0	200	654

Еще раз обратим внимание на то, что в таком анализе используется предположение о нормальности распределения зависимой переменной. Не следует проводить непосредственно дисперсионный анализ переменных с существенно отличающимся от нормального распределением.

В табл. 4.9. приведена выдача одномерного дисперсионного анализа после выполнения команды

```
MEANS TABLES = lnv14m BY v11 BY v8 /STATISTICS
ANOVA.
```

Наблюдаемый уровень значимости 0,707 свидетельствует о том, что *на наших данных указанным методом связь не обнаруживается*.

Таблица 4.9

Результаты однофакторного дисперсионного анализа

		Sum of Squares	df	Mean Square	F	Sig.
LNV14M Логарифм душевого дохода * V11 Состояние в браке	Between Groups	0,40	3	0,13	0,465	0,707
	Within Groups	188,51	650	0,29		
	Total	188,92	653			

4.1.5. Одномерный дисперсионный анализ (ONEWAY)

Данная процедура позволяет проводить одномерный дисперсионный анализ, ее преимущества перед командой MEANS состоят в возможности исследования равенства дисперсий в группах, исследования полиномиальных трендов, проведения множественных сравнений:

```
ONEWAY lnv14m BY w10 /STATISTICS HOMOGENEITY
/POSTHOC = BTUKEY SCHEFFE BONFERRONI ALPHA(.05).
```

Задается тестируемая переменная, служебное слово «BY», переменная группирования. Проверка равенства дисперсий задается подкомандой /STATISTICS HOMOGENEITY, множественные сравнения – подкомандой /POSTHOC = ….

Контрасты. Контрастом называется линейная комбинация средних в группах $\sum_{i=1}^{k} a_i \bar{x}_i$, где $\sum_{i=1}^{k} a_i = 0$. С помощью контрастов можно проверять гипотезы об определенных соотношениях между математическими ожиданиями переменной в группах. В частности, если задать $a_i = -a_j = 1$, можно проверять гипотезу о равенстве *i*-го и *j*-го среднего. Можно подобрать контрасты для проверки линейного или полиномиального изменения средних (см. [7]). В условиях равенства матожиданий маловероятно существенное их отклонение от нуля.

4.1.6. Множественные сравнения

Множественные сравнения являются одной из труднейших проблем в математической статистике. В действительности при анализе данных исследователи сталкиваются с ними на каждом шагу.

Пусть, например, мы рассматриваем 100 независимых таблиц сопряженности пар переменных, отбирая среди них «интересные» для анализа, с

использованием критических значений хи-квадрат 5 %-го уровня значимости. Тогда при отсутствии связи переменных мы будем в среднем в таких испытаниях получать 5 «интересных» (значимых) таблиц, даже если связь между всеми переменными отсутствует. Таким образом, как бы ни были плохи данные, мы что-либо будем интерпретировать. Но при повторном сборе данных мы можем получить противоположные результаты. Вот что значит множественные сравнения!

Сравнение групповых средних – это одна из немногих задач, где удалось справиться с этой проблемой.

Суть задачи состоит в отборе значимых различий множества пар групп, определяемых переменной группирования. Сравнение пары средних мы научились делать с помощью процедуры `T-TEST`, и, казалось бы, можно, задавшись уровнем значимости, пропустить через этот тест все пары групп и отобрать различающиеся по заданному уровню. Однако, перебирая группы, мы перебираем множество случайных чисел и благодаря этому можем наткнуться на значимое отличие с гораздо большей вероятностью, чем при рассмотрении одной пары групп. В частности, если группы независимы и не связаны с тестируемой переменной, при 10 сравнениях по уровню значимости 0,05 мы с вероятностью $1-(1-0{,}05)^{10} = 0{,}4$ случайно получим хотя бы одно «значимое» различие. Эту проблему мы уже рассматривали в разд. 3.2.

Для пояснения механизма работы тестов множественных сравнений остановимся на 3 из 20 тестов, реализованных в SPSS.

Согласно методу Бонферрони в случае множественных сравнений назначается более строгий уровень значимости для попарных сравнений. Он определяется так: задается уровень значимости для множественных сравнений α_m и в качестве попарного уровня значимости берется $\alpha = (1/k)\alpha_m$., где k – число сравнений. Пусть A_i – событие, состоящее в том, что мы в i-м сравнении выявили существенное отличие средних; когда средние совпадают, тогда, в соответствии с заданным уровнем значимости, $P\{A_i\}<\alpha$. Ясно, что $P\{A_1 + A_2 + ... + A_k\} \le P\{A_1\} + P\{A_2\} + ... + P\{A_k\} < k\alpha = \alpha_m$, поэтому метод Бонферрони гарантирует нас от ошибки с вероятностью, не меньшей α_m. В независимых сравнениях неравенство $P\{A_1 + A_2 + ... + A_k\} < k\alpha$ будет выполняться почти точно так, как $1-(1-\alpha)^k \approx k\alpha$. Критерий несколько жестче, чем необходимо, так как средние в группах связаны их взвешенная сумма равна общему среднему.

Метод Шеффе построен на контрастах. С его помощью проверяется гипотеза равенства нулю сразу всех контрастов, не только тех, что сравнивают пары групп. В результате он часто оказывается еще строже, чем критерий Бонферрони.

Критерий Тьюки основан на одновременных доверительных интервалах разности матожиданий в группах. Этот критерий из трех рассматриваемых, пожалуй, наиболее разумен. Предположение об одновременном равенстве разностей всех групповых матожиданий – слишком сильное предположение, в критерии Тьюки такого не предполагается.

В качестве примера рассмотрим различие среднего промедианного логарифма доходов в группах по образованию, группы которого несколько укрупнены:

```
  RECODE v10 (4 5 = 4) (6 7 8 = 5) (ELSE = COPY) INTO
w10.
  VAR LAB w10 "образование".
  VALUE LAB w10 1 "Высшее" 2 "н/высш" 3 "ср. спец"
4 "среднее" 5 "ниже среднего".
  ONEWAY lnv14m BY w10 /STATISTICS DESCRIPTIVES
HOMOGENEITY /POSTHOC = BTUKEY SCHEFFE BONFERRONI
ALPHA(.05).
```

На основании полученной выдачи видим, что:

– доверительные интервалы для высшего и неполного высшего образования не пересекаются (табл. 4.10);

– дисперсии в группах различаются несущественно (см. тест Ливиня, табл. 4.11);

– в целом наблюдается связь душевого дохода с образованием (в результате дисперсионного анализа отвергается гипотеза о равенстве средних, табл. 4.12);

– выделились следующие две группы по образованию с неразличимыми средними: 2 – н/высшее, 5 – ниже среднего, 4 – среднее и 5 – ниже среднего, 4 – среднее, 3 – среднее спец., 1 – высшее (табл. 4.13);

– попарные множественные сравнения показали, что единственная пара отличающихся по средним групп – это группы с неполным высшим и респондентов с высшим образованием (наблюдаемая значимость – 0,013, табл. 4.14).

Таблица 4.10

Oneway, сравнение среднего промедианного логарифма доходов

W10 образование	N	Mean	Std. Deviation	Std. Error	95 % Confidence Interval for Mean		Minimum	Maxi-mum
					Lower Bound	Upper Bound		
1.00 Высшее	251	0,048	0,511	0,032	–0,016	0,111	–1,050	2,015
2.00 Н/высш.	37	–0,248	0,606	0,100	–0,450	–0,046	–1,386	1,099
3.00 Ср. спец.	220	0,009	0,479	0,032	–0,055	0,073	–1,386	1,740
4.00 Среднее	130	–0,093	0,619	0,054	–0,200	0,015	–2,254	1,504
5.00 Ниже сред.	33	–0,107	0,530	0,092	–0,295	0,081	–0,916	1,099
Total	671	–0,016	0,534	0,021	–0,057	0,024	–2,254	2,015

Таблица 4.11

Oneway, проверка однородности дисперсий

Levene Statistic	df1	df2	Sig.
2,282	4	666	0,059

Таблица 4.12

Oneway, обычный дисперсионный анализ

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4,187	4	1,047	3,724	0,005
Within Groups	187,202	666	0,281		
Total	191,389	670			

Таблица 4.13

Oneway, группы неразличимых средних

	W10 образование		1	2
Tukey HSD	2.00 н/высш	37	–0,248	
	5.00 ниже среднего	33	–0,107	–0,107
	4.00 среднее	130	–0,093	–0,093
	3.00 ср. спец	220		0,009
	1.00 Высшее	251		0,048
	Sig.		0,429	0,436
Scheffe	2.00 н/высш	37	–0,248	
	5.00 ниже среднего	33	–0,107	–0,107
	4.00 среднее	130	–0,093	–0,093
	3.00 ср. спец	220	0,009	0,009
	1.00 Высшее	251		0,048
	Sig.		0,093	0,579

Таблица 4.14

Oneway, множественные попарные сравнения

			Mean Difference (I-J)	Std. Error	Sig.	95 % Confidence Interval	
	(I) W10 Образование	(J) W10 Образование				Lower Bound	Upper Bound
Tukey HSD	1 Высшее	2 Н/высш.	0,296*	0,093	0,013	0,041	0,551
		3 Ср. спец.	0,039	0,049	0,934	-0,095	0,172
		4 Среднее	0,140	0,057	0,102	-0,016	0,297
		5 Ниже среднего	0,154	0,098	0,516	-0,113	0,422
	2 Н/высш.	1 Высшее	–0,296*	0,093	0,013	–0,551	–0,041
		3 Ср. спец.	–0,257	0,094	0,050	–0,514	0,000
		4 Среднее	–0,155	0,099	0,515	–0,425	0,114
		5 Ниже среднего	–0,142	0,127	0,799	–0,488	0,205
	3 Ср. спец.	1 Высшее	–0,039	0,049	0,934	–0,172	0,095
		2 Н/высш.	0,257	0,094	0,050	0,000	0,514
		4 Среднее	0,102	0,059	0,412	–0,058	0,262
		5 Ниже среднего	0,116	0,099	0,769	–0,154	0,386
	4 Среднее	1 Высшее	–0,140	0,057	0,102	–0,297	0,016
		2 Н/высш.	0,155	0,099	0,515	–0,114	0,425
		3 Ср. спец.	–0,102	0,059	0,412	–0,262	0,058
		5 Ниже среднего	0,014	0,103	1,000	–0,268	0,296
	5 Ниже ср.	1 Высшее	–0,154	0,098	0,516	–0,422	0,113
		2 Н/высш.	0,142	0,127	0,799	–0,205	0,488
		3 Ср. спец.	–0,116	0,099	0,769	–0,386	0,154
		4 Среднее	–0,014	0,103	1,000	–0,296	0,268

Следует заметить, что мы не показали здесь часть таблицы попарных сравнений с результатами для метода Бонферрони и Шеффе; результаты аналогичны, но для указанной пары групп значимость различия по Шеффе – 0,041, по Бонферрони – 0,016. Это показывает бóльшую чуствительность теста Тьюки.

4.2. CORRELATIONS – корреляции

Раздел **CORRELATIONS** содержит команды для получения парных (**Bivariate…**) и частных (**Partial…**) корреляций.

4.2.1. Парные корреляции

Команда **Bivariate…** меню производит вычисление таблицы коэффициентов Пирсона, характеризующего степень линейной связи, а также коэффициентов ранговой корреляции **BTAU** и Спирмена (*Spearman*). В синтаксисе эта команда имеет вид:

```
CORRELATIONS /VARIABLES = v9 lnv14m /PRINT = TWOTAIL
NOSIG.
```

для обычного коэффициента корреляции и

```
NONPAR CORR /VARIABLES = v10 v9 v14 /PRINT =
SPEARMAN.
```

или:

```
NONPAR CORR /VARIABLES = v10 WITH v9 v14 /PRINT =
KENDALL.
```

для ранговых корреляций.

Подкоманда /VARIABLES в этих командах указывает список переменных или два списка переменных, разделенных словом WITH. Если указывается один список переменных, то рассчитываются коэффициенты корреляции каждой переменной с каждой переменной (квадратная таблица). Если указываются два списка, разделенные служебным словом WITH, то рассчитываются коэффициенты корреляции всех переменных, расположенных слева от WITH, с переменными, расположенными справа (прямоугольная таблица). Ключевое слово WITH можно использовать только в окне синтаксиса.

Процедура CORRELATIONS выводит: r – коэффициент корреляции Пирсона; число наблюдений (объектов) в скобках и значимость коэффициента корреляции. Коэффициент корреляции Пирсона между переменными X и Y вычисляется по формуле

$$r = \sum_{i=1}^{N} \left(X_i - \overline{X}\right) \left(Y_i - \overline{Y}\right) / \left(N-1\right) S_x S_y \ .$$

Коэффициент корреляции может принимать значения от –1 до +1. При этом значимый отрицательный коэффициент корреляции позволяет принять гипотезу о наличии линейной отрицательной связи. Метод, используемый для проверки гипотезы, предполагает также двумерную нормальность распределения (*X, Y*). На практике это соответствует тому, что увеличению значения одной переменной в большинстве случаев соответствует уменьшение значения коррелирующей с ней переменной. Значимый положительный коэффициент корреляции свидетельствует о положительной

связи переменных: увеличению одной переменной соответствует увеличение другой. Чем ближе абсолютное значение r к единице, тем более линейный характер носит зависимость исследуемых переменных; близость к 0 означает отсутствие линейной связи.

Насколько полученное значение коэффициента корреляции не случайно, определяется по величине значимости (Sig. (2-tailed)) – вероятности получить большее, чем выборочное значение коэффициента корреляции. Для оценки значимости коэффициента Пирсона используется критерий $t = r \times (N-2) / (1-r^2)^{0,5}$, который в условиях нормальности и независимости переменных имеет распределение Стьюдента. Таким образом, наряду с формулировкой нулевой гипотезы здесь формулируется предположение о двумерной нормальности – довольно жесткое условие.

Для оценки значимости коэффициентов Спирмена и Кендалла используется нормальная аппроксимация этих коэффициентов. По сути, коэффициент ранговой корреляции является коэффициентом корреляции между переменными, преобразованными в ранги (или процентили), поэтому для исследования значимости с помощью этих коэффициентов не требуется делать предположения о распределении данных. Пример выдачи коэффициентов Спирмена представлен в табл. 4.15. Не обнаруживается значимой связи возраста и образования (что вполне естественно), но среднемесячный душевой доход связан с образованием (это мы уже показывали).

Таблица 4.15

Коэффициенты корреляции Спирмена (Spearman's rho)

		V9 Возраст	V14 Ср.мес. душевой доход в семье
V10 Образование	Correlation Coefficient	–,021	–,086
	Sig. (2-tailed)	,574	,026
	N	692	671

4.2.2. Частные корреляции

Пусть имеются переменные *X, Y, Z*. Что, если взаимосвязь между переменными *X* и *Y* обусловлена некоторой другой переменной *Z*? Может быть, она проявляется при условии этой переменной?

Для исследования этого вопроса применяется коэффициент частной корреляции. Вообще говоря, коэффициент корреляции *X* и *Y* должен зависеть от значений *Z*, однако известно, что в многомерной нормальной совокупности такой зависимости нет. Поэтому статистическая теория здесь разработана именно для такого случая. На практике весьма сложно доказать многомерную нормальность, и часто эту технику используют для анализа данных, не имеющих слишком больших перекосов.

Не вдаваясь в подробности вычисления, коэффициент частной корреляции можно представить как коэффициент корреляции регрессионных остатков ε_x и ε_y уравнений:

$$X = a_x + b_x \times Z + \varepsilon_x$$

$$Y = a_y + b_y \times Z + \varepsilon_y .$$

Таким образом, снимается часть зависимости, обусловленная третьей переменной, проявляется «чистая» взаимосвязь *X* и *Y*. Уравнению регрессии мы посвятим в дальнейшем специальный раздел. Здесь мы приведем пример задания частной корреляции.

Время, затраченное на покупки, и время на мытье посуды связаны положительно: чем больше человек тратит его на покупки, тем больше на посуду (табл. 4.16, RLMS, 7 волна). Может быть, это определяется тем, что человек вообще занимается домашней работой? Для проверки возьмем в качестве управляющей переменной время на уборку квартиры … и получим табл. 4.17. Оказалось, что эта связь между временными затратами на покупку продуктов и мытье посуды имеет самостоятельный смысл, так как частная корреляция по-прежнему значима, хотя уменьшилась с 0,320 до 0,256.

Таблица 4.16

Коэффициент корреляции времени приготовления пищи и закупки продуктов

		CO17A время на приготовление пищи
CO15A время на покупку продуктов	Pearson Correlation	0,3193
	Sig. (2-tailed)	0,0000
	N	3549

Таблица 4.17

Коэффициент корреляции времени приготовления пищи и закупки продуктов

Controlling for.. CO19A (время на уборку квартиры)		CO17A время на приготовление пищи
CO15A время на покупку продуктов	Pearson Correlation	0,2558
	Sig. (2-tailed)	0,0000
	N	3546

Глава 5. НЕПАРАМЕТРИЧЕСКИЕ ТЕСТЫ. КОМАНДА NONPARAMETRIC TESTS

Непараметрические тесты предназначены преимущественно для проверки статистических гипотез методами, не связанными с видом распределения совокупности. В частности, применение этих методов не требует предположения о нормальности распределения, которое необходимо для правомерного использования одномерного дисперсионного анализа, процедуры T-TEST, при определении значимости корреляций, и т. д. К средствам непараметрического анализа относятся в числе прочих методов тест хи-квадрат, служащий для проверки взаимосвязи между номинальными переменными и коэффициентами ранговой корреляции, которым мы уже уделили некоторое внимание.

Непараметрические тесты не ограничиваются таким исследованием связи пар переменных; они включают множество других методов, реализованных командой синтаксиса NPAR TESTS. В меню SPSS непараметрические тесты реализует команда **Nonparametric tests** с множеством подкоманд.

Процедура NPAR TESTS включает большую группу критериев для проверки:

– соответствия распределения выборочной совокупности заданному распределению;

– случайного характера выборки объектов;

– совпадения распределений в различных группах

– совпадения распределений в связанных выборках (например, результатов повторных измерений).

Во всех критериях допускаются асимптотические, точные оценки значимости (Exact) и оценки их методом Монте-Карло.

5.1. Одновыборочные тесты

Эти тесты служат для проверки соответствия распределения выборки заданному.

5.1.1. Тест хи-квадрат

Критерий хи-квадрат основан на статистике

$$X^2 = \sum_i \left(N_i - E_i\right)^2 / E_i ,$$

где $E_i = P_i \times N$ – ожидаемая частота *i*-го значения переменной, N_i – расчетная. Теоретическое распределение этой статистики при больших N совпадает с распределением хи-квадрат. Число степеней свободы теоретического распределения полагается равным $k - 1$, где k – число значений исследуе-

мой переменной. Эмпирическое правило говорит о том, что некорректно применять критерий, если ожидаемые частоты меньше 5, поскольку его распределение в этом случае не будет близко к теоретическому. Но использование точных методов вычисления значимости (метод Монте-Карло) позволяет избежать этого ограничения.

Пример. Пусть согласно статистическим данным 30 % трудоспособного населения имеет возраст до 30 лет, 30 % от 30 до 40 лет и 40 % свыше 40 лет. Соответствует ли выборочное распределение признака «возраст» в обследовании «Курильские острова» распределению возраста в генеральной совокупности?

```
RECODE v9 (1 THR 30 = 1)(31 THR 40 = 2)(41 THRU HI
= 3) INTO w9.
NPAR TESTS /CHISQUARE = W9 /EXPECTED 3 3 4.
```

Подкоманда `/CHISQUARE` задает тестируемую переменную; в подкоманде `/EXPECTED` задаем через пробел ожидаемые пропорции распределения.

Выполнение этих команд позволяет получить значение критерия и оценить степень соответствия нашей выборки распределению генеральной совокупности (табл. 5.1, 5.2).

Таблица 5.1

Наблюдаемые и ожидаемые частоты

	Observed N	Expected N	Residual
1	175	210	–35
2	225	210	15
3	300	280	20
Total	700		

Таблица 5.2

Статистика хи-квадрат

	W9
Chi-Square	8,333
Df	2
Asymp. Sig.	0,016

Анализируя табл. 5.1, уже по отклонениям расчетных значений от ожидаемых (см. столбец Residual), видим, что эмпирическое распределение сильно отличается от теоретического. Достаточно высокое значение критерия (Chi-Square = 8,333, табл. 5.2) малоинформативно. Ответ о совпадении

нашего распределения с теоретическим заключен в анализе наблюдаемого уровня значимости. Его малая величина (Asymp. Sig. = 0,016) показывает, что полученные отклонения значимы: вероятность получить большие значения хи-квадрат равна 1,6 %, гипотеза о соответствии выборки указанной генеральной совокупности может быть отвергнута на уровне значимости 5 %.

Таким образом, для данного случая тест показал существенное различие теоретического и эмпирического распределений.

Приведем пример применения метода статистического моделирования Монте-Карло. В этом примере производится 100 000 экспериментов по моделированию выборки из генеральной совокупности с заданными вероятностями ($p_1 = 0{,}3$, $p_2 = 0{,}3$, $p_3 = 0{,}4$):

```
   NPAR TEST /CHISQUARE = w9 /EXPECTED = 3 3 4
/METHOD = MC CIN(99) SAMPLES(100000).
```

Естественно, при такой большой выборке был получен тот же результат (табл. 5.3). Уровень значимости оценивается этим методом приближенно, на основании статистических экспериментов – чем больше экспериментов, тем точнее. Поскольку оценка значимости получена на основе случайных экспериментов, выдается доверительный интервал для уровня значимости (99 %-й по умолчанию). Точечная оценка наблюдаемого уровня значимости (Monte Carlo Sig) совпадает с асимптотической оценкой (Asymp. Sig., табл. 5.3), «оптимистическая» нижняя граница равна 0,015, «пессимистическая» верхняя – 0,017. Таким образом, во всех отношениях отклонение распределения значимо.

Таблица 5.3

Значимость критерия хи-квадрат

			W9
Chi-Square			8,333
Df			2
Asymp. Sig.			0,016
Monte Carlo Sig	Sig.		0,016
	99 % Confidence Interval	Lower Bound	0,015
		Upper Bound	0,017

5.1.2. Тест, основанный на биномиальном распределении

Проверяется гипотеза о параметре биномиального распределения H_0**:** $p = p_0$. Например, проверим по нашей выборке, действительно ли в генеральной совокупности вероятность встретить мужчину $p = 0{,}5$, а молодежь не старше 30 лет – с вероятностью $p = 0{,}3$ (см. предыдущий пример):

```
NPAR TESTS BINOMIAL(0.5) = V8(1,2).
NPAR TESTS BINOMIAL(0.3) = V9(30).
```

В скобках за ключевым словом BINOMIAL указывается вероятность «успеха». Далее следует тестируемая переменная. Если за ней в скобках следует два значения, то считается, что выборка ограничена двумя группами, соответствующими этим значениям, а успех соответствует первому значению. Если в скобках задано одно значение, то успех – это принятие переменной значения, не больше, чем указанное число. В диалоговом окне есть возможность задать как «точку разрыва», так и два кода, идентифицирующие группы объектов.

Программа подсчитывает число объектов *m***,** имеющих заданные значения (в первом случае *m* – число мужчин (код 1), во втором случае *m* – число респондентов не старше 30 лет). На основании свойств биномиального распределения подсчитывается двусторонняя наблюдаемая значимость – вероятность случайной величины в условиях биномиального распределения с параметром *P* отклониться от ожидаемого значения *np* больше, чем отклонилось выборочное значение *m*.

Наблюдаемый уровень значимости можно оценить с использованием теоремы Муавра – Лапласа, методом Монте-Карло, а также точно, по биномиальному распределению, используя возможность, представленную в SPSS в EXACT STATISTICS:

```
NPAR TEST /BINOMIAL (.50) = v8 /METHOD = EXACT
TIMER(5).
```

Таблица 5.4

Значимость критерия хи-квадрат

	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Group 1	1 муж.	362	0,508	0,5	0,708	0,708
Group 2	2 жен.	351	0,492			
Total		713	1			

В табл. 5.4 выдается расчетная 0,508 и заданная теоретическая вероятность Test Prop. = 0,5. Выборочное распределение почти совпало с заданным. Этот результат окончательно подтверждает величина двусторонней значимости: 0,708 – вероятность случайно получить значение больше полученного. Так как 70 % – это большая вероятность, мы делаем вывод, что распределение совпадает с заданным. Двусторонний тест показал незначимое отличие доли мужчин в выборке от ожидаемой доли (нулевая гипотеза не отвергается).

5.1.3. Тест Колмогорова – Смирнова

Одновыборочный тест предназначен для проверки гипотезы о распределении в генеральной совокупности. Статистика критерия – абсолютная величина разности эмпирической и теоретической функций распределения:

$$ks = \sqrt{N} \sup_x \left| F^*(x) - F(x) \right|.$$

Команда задания теста Колмогорова – Смирнова имеет вид

```
NPAR TESTS K-S(NORMAL,5,2) = X.
```

В скобках за ключевым словом `K-S` указывается предполагаемый вид распределения: `NORMAL` – нормальное; `UNIFORM` – равномерное; `POISSON` – распределение Пуассона; `EXPONENTIAL` – показательное распределение. За видом распределения в скобках можно указать его параметры: для нормального – среднее и среднеквадратичное отклонение; для равномерного – минимум и максимум; для распределения Пуассона – среднее. По умолчанию используются оценки параметров по выборочной совокупности.

Заметим, что оценка параметров по выборке дает смещение этого критерия. Поэтому ему стоит доверять только для больших выборок.

Таблица 5.5

Проверка нормальности распределения доходов с использованием критерия Колмогорова – Смирнова

		V14 Душевой доход в семье
N		673
Normal Parameters	Mean	229,11
	Std. Deviation	151,34
Most Extreme Differences	Absolute	0,187
	Positive	0,187
	Negative	–0,149
Kolmogorov – Smirnov Z		4,85
Asymp. Sig. (2-tailed)		0

В таблице результатов выдается двусторонняя значимость – вероятность в условиях гипотезы случайно превзойти выборочное значение статистики, фиксирующей отличие распределения от заданного.

Например, проверим нормальности распределения доходов командой:

```
NPAR TESTS K-S(NORMAL) = V14.
```

Поскольку двусторонняя значимость в табл. 5.5 (2-tailed) равна нулю, то можно сделать вывод, что полученная разность фиксирует существенное отличие распределения по доходам от нормального. Во многих исследованиях используется вместо дохода его логарифм, распределение которого считается близким к нормальному. Проверим нормальность логарифма доходов:

```
COMPUTE lnv14 = ln(v14).
NPAR TEST K_S(NORMAL) = w14.
```

Таблица 5.6

Проверка лог-нормальности распределения доходов

		LNV14
N		673
Normal Parameters	Mean	5,2812
	Std. Deviation	0,5344
Most Extreme Differences	Absolute	0,098
	Positive	0,098
	Negative	–0,055
Kolmogorov-Smirnov Z		2,54
Asymp. Sig. (2-tailed)		0

Значение критерия несколько уменьшилось, но существенность различия сохранилась (табл. 5.6).

Иногда бывает необходимо проверить законы распределения, не предусмотренные в NPAR TESTS. В этом случае вспомните, что распределение непрерывной случайной величины $\eta = F_{\xi}(\xi)$, где F – функция распределения ξ, равномерно на отрезке (0,1). Таким образом, воспользовавшись статистическими функциями преобразования данных SPSS, из тестируемой переменной всегда можно получить переменную, имеющую теоретически равномерное распределение, и, проверив, действительно ли ее распределение равномерно, принять или отвергнуть гипотезу о виде распределения $F_{\xi}(x)$.

5.2. Тесты сравнения нескольких выборок

Эти тесты предназначены для проверки гипотезы совпадения распределений в выборках. В отличие от *t*-теста и известных методов дисперсионного анализа, здесь не предполагается нормальность теоретического распределения.

Многие тесты основаны на поиске определенного типа противоречия с гипотезой совпадения распределений и не могут обнаружить всех отличий. Например, тест медиан проверяет совпадение только медиан. Поэтому иногда полезно воспользоваться несколькими тестами.

5.2.1. Двухвыборочный тест Колмогорова – Смирнова

Двухвыборочный тест Колмогорова – Смирнова предназначен для проверки гипотезы о совпадении распределений в паре выборок:

```
NPAR TESTS K-S = V14 BY V4(1,3).
```

В команде за ключевым словом `K-S` следует тестируемая переменная (в нашем примере – `V14`), за ней после слова `BY` указываются сравниваемые группы: переменная, определяющая эти группы, и соответствующие этим группам значения: `V4`(1,3).

Статистика критерия – абсолютная величина разности эмпирических функций распределения в указанных выборках:

$$ks = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \max_x \left| F_1(x) - F_2(x) \right|,$$ где N_1 и N_2 – объемы выборок.

В листинге выдается статистика критерия $Z = ks$ и двусторонняя значимость – вероятность случайно в условиях гипотезы превзойти выборочное значение статистики.

Пример. Сравнение распределений доходов группы, готовой отдать острова или их часть, и группы, придерживающейся твердой позиции:

```
RECODE v4(1,2 = 1)(3 = 2) INTO W4.
VAR LAB W4 "отношение к передаче островов".
VAL LAB 1 "Отдать" 2 "нет".
NPAR TEST K_S = v14 BY w4(1,2).
```

Таблица 5.7

Сравнение распределения доходов в двух группах на основе критерия Колмогорова – Смирнова

		V14 Душевой доход в семье
Most Extreme Differences	Absolute	0,05
	Positive	0,05
	Negative	–0,028
Kolmogorov – Smirnov Z		0,455
Asymp. Sig. (2-tailed)		0,986

В табл. 5.7 наблюдаемый уровень значимости велик (0,986). Поэтому приходим к заключению, что на нашей учебной выборке критерием Колмогорова – Смирнова не удалось обнаружить различие распределений по душевому доходу в группы считающих, что нужно отдать острова или их часть, и группы противников такого решения. Это не означает достоверно,

что распределения совпадают, возможны тонкие различия распределений, которые критерий не улавливает из-за малого объема данных.

5.2.2. Тест медиан

Этот тест позволяет сравнивать распределения исследуемой переменной сразу в нескольких группах. Тест весьма груб и прост.

```
NPAR TESTS MEDIAN  =  V14 BY V1(1,3).
```

Внешне задание теста похоже на задания критерия Колмогорова – Смирнова.

Задание сравниваемых групп. После слова `BY` за именем переменной в скобках указывается интервал значений. В приведенном примере сравниваются распределения в трех группах. Тестом можно сравнить также и пару групп, если в скобках вначале указать большее значение, затем меньшее (при задании `V4`(3,1) сравниваются только 1-я и 3-я группы).

Суть проверки гипотезы состоит в следующем. Значения исследуемой переменной (в нашем примере `V14`) делятся на две группы: больше медианы и меньше или равно медиане. Такое разделение можно считать заданием новой, дихотомической переменной. Вычисляется таблица сопряженности полученной дихотомической переменной и переменной, задающей группы. Далее применяется известный критерий хи-квадрат. Если величина наблюдаемой значимости критерия мала, естественно предположить, что распределение исследуемой переменной в группах различается существенно.

> ***Замечание***. Для получения дихотомии можно также навязать точку «разрыва» переменной, не совпадающую с медианой, указав в скобках за словом MEDIAN соответствующее значение.

Пример. Курильское обследование проходило в 21 городе Западной Сибири. Экспертным путем все города разделены на 4 типа: 1 – растущие, 2 – стабильные, 3 – крупные, 4 – гиганты. Типу города в наших данных соответствует переменная `TP`.

Исследуется связь доходов и типа населенного пункта:

```
NPAR TESTS MEDIAN = v14 by TP(1,4).
```

Таблица 5.8

Метод медиан. Разделение на две подвыборки

		TP тип поселения			
		Растущие	Стабильные	Крупные	Гиганты
V14 Ср. мес. душевой доход в семье	> Median	84	104	62	12
	<= Median	90	126	139	56

Таблица 5.9

Метод медиан. Значимость критерия

	V14 Ср.мес. душевой доход в семье
N	673
Median	200
Chi-Square	28,698
Df	3
Asymp. Sig.	0

Анализируя величину наблюдаемой значимости, видим, что между точкой зрения на иностранную помощь и возрастом имеется существенная связь, т. е. обнаружено значимое различие распределения доходов в группах.

5.3. Тесты для ранговых переменных

В ряде методов по имеющимся числовым значениям исследуемой переменной объектам приписываются ранги. Для вычисления рангов объекты упорядочиваются от минимального значения переменной к максимальному, и порядковые номера объектов считаются рангами. Если для некоторых объектов числовые значения переменной повторяются, то всем этим объектам приписывается единый ранг, равный среднеарифметическому значению их порядковых номеров. Об объектах, ранги которых совпадают, говорят, что они имеют связанные ранги. Наличие связанных рангов в выдаче по ранговым тестам обозначается словом «ties» (связи). Обычно выводится число связей и статистика критерия, скорректированная для связей.

В качестве примера построения рангов возьмем упорядоченную информацию об успеваемости 7 студентов.

Средний балл	3,0	3,1	4,0	4,2	4,2	4,5	4,5
Ранг	1	2	3	4,5	4,5	6	7

Первые три объекта имеют ранги 1, 2, 3; следующая пара – ранг 4,5 = (4 + 5) / 2, следующая пара – 6 и 7.

5.3.1. Двухвыборочный тест Манна – Уитни (Mann – Witney)

Критерий предназначен для сравнения распределений переменных в двух группах на основе сравнения рангов.

```
NPAR TESTS M-W = V14 BY Тр(1,4).
```

Задание теста аналогично заданию критерия Колмогорова – Смирнова (вместо ключевого слова `K-S` используется слово `M-W`).

Статистикой критерия является сумма рангов объектов в меньшей группе, хотя существует пара эквивалентных формул, обозначаемых *U* и *W*.

Можно также считать, что критерием является средний ранг в указанной группе. Если он значительно отклоняется от ожидаемой величины $(N + 1) / 2$ (или средние ранги в группах существенно различны), то обнаруживается отличие распределений.

Если гипотеза о совпадении распределений не отвергается, то это означает близость средних рангов в группах, но совпадение распределений не гарантируется (хотя бы потому, что они могут отличаться сколь угодно мало).

Авторам теста удалось показать асимптотическую нормальность статистики в условиях выборки групп из одной совокупности, на основе чего отыскивается наблюдаемая значимость критерия – вероятность случайно отклониться от среднего (ожидаемого) значения ранга больше, чем отклонилось выборочное значение статистики.

В выдаче распечатывается значения статистик U и W, а также двусторонняя значимость критерия.

Пример. Используя ранговый критерий, требуется сравнить по возрасту группу считающих, что острова нужно отдать по юридическим причинам, и группу имеющих иное мнение.

```
COUNT d2 = v6s1 TO v6s8 (2).
IF (d2>0) wd2 = 1.
IF (v4 = 1 or v4 = 2) wd2 = 2.
NPAR TEST M-W = v9 BY wd2(1,3).
```

По величине двусторонней значимости можем сделать вывод, что тест Манна – Уитни в указанных группах не обнаружил существенных различий между распределениями по возрасту (табл. 5.10 – 5.11).

Таблица 5.10

Критерий Манна – Уитни. Суммы рангов

	WD2	N	Mean Rank	Sum of Ranks
V9 Возраст	1	117	116,7	13650,5
	2	103	103,5	10659,5
	Total	220		

Таблица 5.11

Критерий Манна – Уитни. Значимость критерия

	V9 Возраст
Mann – Whitney U	5303,5
Wilcoxon W	10659,5
Z	–1,533
Asymp. Sig. (2-tailed)	0,125

5.3.2. Одномерный дисперсионный анализ Краскэла – Уоллиса (Kruskal – Wallis)

В основе сравнения средних рангов заданного числа групп лежит одномерный дисперсионный анализ, в котором вместо значений переменных используются ранги объектов исследуемой переменной.

```
NPAR TESTS K-W = V14 BY V4(1,3).
```

В условиях гипотезы равенства распределений в группах нормированный межгрупповой разброс имеет распределение, близкое к распределению хи-квадрат. В выдаче распечатывается значимость этой статистики.

Следующий пример показывает различие доходов жителей населенных пунктов разного типа.

```
NPAR TESTS K-W = v9 BY tp(1,4).
```

Таблица 5.12

Тест Краскэла – Уоллиса. Средние ранги

	TP тип поселен	N	Mean Rank
V14 Ср.мес. душевой доход в семье	1.00 растущие	174	382
	2.00 стабильные	230	365,2
	3.00 крупные	201	304,6
	4.00 гигант.	68	222,2
	Total	673	

Таблица 5.13

Тест Краскэла – Уоллиса. Значимость критерия

	V14 Ср.мес. душевой доход в семье
Chi-Square	43,702
Df	3
Asymp. Sig.	0

Тест показывает (Sig = 0), что точка зрения респондента на иностранную помощь существенно связана с типом населенного пункта, в котором он проживает (табл. 5.12 – 5.13).

5.4. Тесты для связанных выборок (Related samples)

Напомним, что связанными выборками называются совокупности повторных измерений на одних и тех же объектах. Например, доходы семьи в различных волнах панельного обследования RLMS; психологические характеристики мужа и жены и т. п.

5.4.1. Двухвыборочный критерий знаков (Sign)

Для исследования связи пары измерений X и Y рассматриваются знаки разностей $d_i = Y_i – X_i$. В случае независимости измерений и отсутствия повторов значений d_i (связей) число знаков «+» (положительных d_i) должно подчиняться биномиальному распределению с параметром $p = 0{,}5$. Именно эта гипотеза и проверяется с помощью статистики критерия – стандартизованной частоты положительных разностей.

В качестве примера по данным RLMS проверим, какой характер имели изменения веса (кг) мужчин старше 30 лет в 1994 – 1995 гг.

```
COMPUTE filter_$ = (a_age < 30 & ah5_1 = 1).
FILTER BY filter_$.
NPAR TEST / SIGN =  am1 WITH bm1 (PAIRED).
```

Таблица 5.14

Тест знаков для парных наблюдений. Частоты

Frequencies		N
BM1 вес в 1995 г. – AM1 вес в 1994 г.	Negative Differences	877
	Positive Differences	722
	Ties	350
	Total	1949

Судя по табл. 5.14, мужчины чаще худели, чем толстели, причем этот факт подтверждается отрицательным значением статистики критерия, наблюдаемая значимость которой равна 0,000118 (табл. 5.15.).

Таблица 5.15

Тест знаков для парных наблюдений. Значимость критерия

Test Statistics	
	BM1 вес в 1995г. – AM1 вес в 1994г.
Z	–3,8512
Asymp. Sig. (2-tailed)	0,000118

5.4.2. Двухвыборочный знаково-ранговый критерий Вилкоксона (Wilcoxon)

Ранжируются абсолютные величины разностей $d_i = Y_i – X_i$. Затем рассматривается сумма рангов положительных и сумма рангов отрицательных разностей. Если связь между X и Y отсутствует и распределение одинаково, то эти две суммы должны быть примерно равны. Статистика критерия – стандартизованная разность этих сумм.

По сути, это проверка, не произошло ли между измерениями событие, существенно изменившее иерархию объектов?

Обратимся к предыдущему примеру, но проверим, будет ли преобладать отрицательный ранг изменения веса мужчин старше 30 лет?

```
NPAR TEST /WILCOXON = am1 WITH bm1 (PAIRED).
```

Табл. 5.16 показывает, что преобладает уменьшение веса, что подтверждается наблюдаемой значимостью статистики критерия, равной 0,00053 (табл. 5.17).

Таблица 5.16

Знаково-ранговый тест Вилкоксона. Средние ранги

BM1 вес в 1995г. – AM1 вес в 1994 г.		N	Mean Rank	Sum of Ranks
	Negative Ranks	877	802,2	703500
	Positive Ranks	722	797,4	575700
	Ties	350		
	Total	1949		

Таблица 5.17

Знаково-ранговый тест Вилкоксона. Средние ранги

	BM1 вес в 1995 г. – AM1 вес в 1994 г.
Z	–3,46504
Asymp. Sig. (2-tailed)	0,00053

5.4.3. Критерий Фридмана (Friedman)

Имеется k переменных. На каждом объекте независимо производится их ранжировка (по строке матрицы данных), затем вычисляется средний ранг по каждой переменной (по столбцу). Если все измерения независимы и равноценны (одинаково распределены), то все эти средние должны быть приближенно равны – $(k+1)/2$ – среднему рангу в строке. Статистикой критерия является нормированная сумма квадратов отклонений средних рангов по переменным от общего среднего $(k+1)/2$, которая имеет теоретическое распределение хи-квадрат.

Таблица 5.18

Тест Фридмана. Средние ранги

	Mean Rank
AM1 вес в 1994г.	2
BM1 вес в 1995г.	2,13
CM1 вес в 1996г.	1,87

Таблица 5.19

Тест Фридмана. Значимость

N	15
Chi-Square	0,561
Df	2
Asymp. Sig.	0,755

Как ни странно, тест Фридмана, запущенный командой

```
NPAR TESTS /FRIEDMAN = am1 bm1 cm1.,
```

не показал значимых различий в измерениях веса по трем годам (см. предыдущие два примера), так как наблюдаемая значимость статистики хи-квадрат равна 0,755.

Глава 6. РЕГРЕССИОННЫЙ АНАЛИЗ

Задача регрессионного анализа состоит в построении модели, позволяющей получать оценки значений результирующей (так называемой зависимой) переменной по значениям объясняющих (так называемых независимых) показателей. Рассмотрим эту задачу в рамках самой распространенной в статистических пакетах классической модели линейной регрессии.

Специфика социологических исследований состоит в том, что очень часто необходимо изучать и предсказывать социальные события. Вторая часть данной главы посвящена логистической регрессии, целью которой является построение моделей, предсказывающих вероятности событий.

6.1. Классическая линейная модель регрессионного анализа

В линейной модели предполагается, что зависимая переменная y связана со значениями независимых показателей x^k (факторов) формулой [2]

$$y_i = B_0 + B_1 x_i^1 + ... + B_p x_i^p + \varepsilon_i .$$

Традиционные названия «зависимая» для y и «независимые» для x^k отражают не столько статистический смысл, сколько их содержательную интерпретацию.

Величина ε_i называется ошибкой регрессии. В классической модели предполагается, что регрессионные ошибки независимы и распределены

[2] Здесь x^k означает x с индексом k.

нормально с параметрами $N(0,\sigma^2)$. Кроме того, в данной модели мы рассматриваем переменные x как неслучайные значения. Такое на практике получается, когда идет активный эксперимент, в котором задают значения x (например, назначили зарплату работнику), а затем измеряют y (оценили, какой стала производительность труда). Поэтому зависимую переменную иногда называют откликом. Теория регрессионных уравнений со случайными независимыми переменными сложнее, но известно, что при большом числе наблюдений использование метода, разработанного для случайных X, корректно.

Для получения выборочных оценок b_k коэффициентов B_k регрессии минимизируется сумма квадратов ошибок регрессии:

$$\sum_i (y_i - b_0 + b_1 x_i^1 + ... + b_p x_i^p)^2 \to \min \quad .$$

Решение задачи сводится к решению системы линейных уравнений относительно b_k.

На основании оценок регрессионных коэффициентов рассчитываются оценки значений y:

$$\hat{y}_i = b_0 + b_1 x_i^1 + b_2 x_i^2 + ... + b_p x_i^p \quad .$$

По сути дела, эти оценки являются оценками математического ожидания Y при заданных значениях X.

О качестве полученного уравнения регрессии можно судить, исследовав $\hat{\varepsilon}_i = y_i - \hat{y}_i$ – оценки случайных ошибок уравнения. Оценка дисперсии случайной ошибки получается по формуле $S^2 = \left(\sum (y_i - \hat{y}_i)^2 \right) \Big/ \left(N - p - 1 \right)$.

Величина S называется стандартной ошибкой регрессии. Чем меньше величина S, тем лучше уравнение регрессии описывает независимую переменную y.

Так как мы ищем оценки b_k, используя случайные данные, то они, в свою очередь, будут представлять случайные величины. В связи с этим возникают вопросы:

1. Существует ли регрессионная зависимость? Может быть, все коэффициенты регрессии в генеральной совокупности равны нулю, оцененные их значения ненулевые только благодаря случайным отклонениям данных?

2. Существенно ли влияние на зависимую переменную отдельных независимых переменных?

В пакете вычисляются статистики, позволяющие решить эти задачи.

6.1.1. Существует ли линейная регрессионная зависимость?

Для проверки одновременного отличия всех коэффициентов регрессии от нуля проведем анализ квадратичного разброса значений зависимой переменной относительно среднего. Его можно разложить на две суммы следующим образом:

$$\sum\left(y_i - \overline{y} \right)^2 = \sum_i \left(y_i - \hat{y}_i \right)^2 + \sum_i \left(\hat{y}_i - \overline{y} \right)^2 .$$

В этом разложении обычно обозначают:

$SS_t = \sum\left(y_i - \overline{y} \right)^2$ – общую сумму квадратов отклонений;

$SS_{res} = \sum\left(y_i - \hat{y}_i \right)^2$ – сумму квадратов регрессионных отклонений;

$SS_{reg} = \sum\left(\hat{y}_i - \overline{y} \right)^2$ – разброс по линии регрессии.

Статистика $F = \dfrac{SS_{reg}/p}{SS_{res}/(N-p-1)}$ в условиях гипотезы равенства нулю регрессионных коэффициентов имеет распределение Фишера, и, естественно, по этой статистике проверяют, являются ли коэффициенты $B_1, \ldots, B_p$ одновременно нулевыми. Если наблюдаемая значимость статистики Фишера мала (например, sig F = 0,003), то это означает, что данные распределены вдоль линии регрессии и гипотеза отвергается; если значимость велика (например, Sign F = 0,12), то, следовательно, данные не связаны такой линейной связью, гипотеза не отвергается.

6.1.2. Коэффициенты детерминации и множественной корреляции

При анализе качества регрессии нужно исследовать доли объясненной и необъясненной дисперсии. Отношение SS_{reg}/SS_t представляет собой оценку доли необъясненной дисперсии. Доля дисперсии зависимой переменной $1 - \sigma^2 / \sigma_y^2$, объясненной уравнением регрессии, называется коэффициентом детерминации. В двумерном случае коэффициент детерминации совпадает с квадратом коэффициента корреляции.

Корень из коэффициента детерминации называется ***коэффициентом множественной корреляции*** (он является коэффициентом корреляции между y и $\hat{y}$). Оценкой коэффициента детерминации ($1 - \sigma^2 / \sigma_y^2$) является $R^2 = 1 - SS_{res} / SS_t$. Соответственно, величина R является оценкой коэффициента множественной корреляции. Следует иметь в виду, что R^2 является смещенной оценкой. Корректированная оценка коэффициента детерминации получается по формуле

$$R_a^2 = 1 - (SS_{res}/(N-p-1))/(SS_t/(N-1)) .$$

В этой формуле используются несмещенные оценки дисперсий регрессионного остатка и зависимой переменной.

6.1.3. Оценка влияния независимой переменной

Если переменные x независимы, то величина коэффициента b_i интерпретируется как прирост y, если x_i увеличить на единицу. Если переменные связаны, то изменение одной объясняющей переменной вызовет изменение других переменных, в результате чего изменения y будут непредсказуемы.

Можно ли по абсолютной величине коэффициента судить о роли соответствующего ему фактора в формировании зависимой переменной? То есть, если $b_1 > b_2$, будет ли x_1 важнее x_2?

Абсолютные значения коэффициентов не позволяют сделать такой вывод. Однако при небольшой взаимосвязи переменных x, если стандартизовать переменные и рассчитать уравнение регрессии для стандартизованных переменных, оценки коэффициентов регрессии позволят по их абсолютной величине судить о том, какой аргумент в большей степени влияет на функцию.

6.1.3.1. Стандартизация переменных. Бета-коэффициенты

Стандартизация переменных, т. е. замена переменных x^k на $z^k = \left(x^k - \overline{x}^k \right) \Big/ S_{xk}$ и y на $u = \left(y - \overline{y} \right) / S_y$ $u = \sum_{k=1}^{p} \beta_k z^k$, где k – порядковый номер независимой переменной.

Коэффициенты в последнем уравнении получены при одинаковых масштабах изменения всех переменных и сравнимы. Более того, если «независимые» переменные не связаны друг с другом, β-коэффициенты суть коэффициенты корреляции между x^k и y. Таким образом, в последнем случае коэффициенты β непосредственно характеризуют связь x и y.

В случае взаимосвязи переменных x могут x^k происходить странные вещи. Несмотря на связь переменных x^k и y, β-коэффициент может оказаться равным нулю, или, наоборот, его величина может оказаться больше единицы!

Взаимосвязь аргументов в правой части регрессионного уравнения называется мультиколлинеарностью. При наличии мультиколлинеарности переменных по коэффициентам регрессии нельзя судить о влиянии этих переменных на функцию.

6.1.3.2. Надежность и значимость коэффициента регрессии

Для изучения «механизма» действия мультиколлинеарности на регрессионные коэффициенты рассмотрим выражение для дисперсии отдельного регрессионного коэффициента

$$S_{b_k}^2 = \frac{S^2}{\left(1 - R_k^2\right)\ (N-1)\ S_{x_k}^2}.$$

Здесь R_k^2 – коэффициент детерминации, получаемый при построении уравнения регрессии, в котором в качестве зависимой переменной взята переменная x^k. Из выражения видно, что величина коэффициента тем неустойчивее, чем сильнее переменная x_k связана с остальными переменными (чем ближе к единице коэффициент детерминации R_k^2).

Величина $1 - R_k^2$, характеризующая устойчивость регрессионного коэффициента, называется надежностью. В англоязычной литературе она обозначается словом *tolerance.* Чем толерантность ближе к 1, тем надежнее оценка коэффициента.

Дисперсия коэффициента позволяет получить статистику для проверки его значимости:

$$t = \frac{b_k}{S_{b_k}}.$$

Эта статистика имеет распределение Стьюдента. В выдаче пакета печатается ее наблюдаемая двусторонняя значимость – вероятность случайно при нулевом регрессионном коэффициенте B_k получить значение статистики, большее по абсолютной величине, чем выборочное.

6.1.3.3. Значимость включения переменной в регрессию

При последовательном подборе переменных в SPSS предусмотрена автоматизация, основанная на значимости включения и исключения переменных. Рассмотрим, что представляет собой эта значимость.

Обозначим $R_{(k)}^2$ коэффициент детерминации, полученный при исключении из правой части уравнения переменной x^k (зависимая переменная y). При этом мы получим уменьшение объясненной дисперсии на величину $R_{ch}^2 = R^2 - R_{(k)}^2$.

Для оценки значимости включения переменной x^k используется статистика

$$F_{ch} = R_{ch}^2 (N - p - 1) \Big/ \left(1 - R^2\right),$$

имеющая распределение Фишера при нулевом теоретическом приросте R^2_{ch}. Вообще, если из уравнения регрессии исключаются q переменных, статистикой значимости исключения будет

$$F_{ch} = R^2_{ch}(N - p - 1)\Big/(q(1 - R^2)).$$

6.1.4. Пошаговая процедура построения модели

Основным критерием отбора аргументов должно быть качественное представление о факторах, влияющих на зависимую переменную, которую мы пытаемся смоделировать. В SPSS очень хорошо реализован процесс построения регрессионной модели: на машину переложена значительная доля трудностей в решении этой задачи. Возможно последовательное построение модели путем добавления и удаления переменных блоками или по отдельности. Но мы рассмотрим только работу с отдельными переменными.

По умолчанию программа включает в уравнение все заданные переменные (метод **ENTER**).

Метод включения и исключения переменных (**STEPWISE**) состоит в следующем.

Из множества факторов, заданных исследователем в качестве возможных аргументов регрессионного уравнения, отбирается один x^k, который более всего связан корреляционной зависимостью с y. Для этого рассчитываются частные коэффициенты корреляции остальных переменных с y при x^k, включенном в регрессию, и выбирается следующая переменная с наибольшим частным коэффициентом корреляции. Это равносильно следующему: вычислить регрессионный остаток переменной y; вычислить регрессионный остаток независимых переменных по регрессионным уравнениям их как зависимых переменных от выбранной переменной (т. е. устранить из всех переменных влияние выбранной переменной); найти наибольший коэффициент корреляции остатков и включить соответствующую переменную x в уравнение регрессии. Далее проводится та же процедура при двух выбранных переменных, при трех и т. д.

Процедура повторяется до тех пор, пока в уравнение не будут включены все аргументы, выделенные исследователем, удовлетворяющие критериям значимости включения.

Замечание: во избежание зацикливания процесса включения/исключения уровень значимости включения устанавливается меньше значимости исключения (например $P_{in} = 0.05$, $P_{out} = 0.1$).

6.1.5. Переменные, порождаемые регрессионным уравнением

Сохранение переменных, порождаемых регрессией, производится подкомандой SAVE. В диалоговом окне эта возможность включается одноименной кнопкой **Save**.

Благодаря полученным оценкам коэффициентов уравнения регрессии могут быть оценены математические ожидания зависимой переменной $\hat{y}$. Иногда в данных для некоторых объектов отсутствуют наблюдения для *y*, а имеются лишь значения *x*. На основании уравнения регрессии SPSS оценивает ожидаемые значения и значения ненаблюдаемых *y*.

Поскольку коэффициенты регрессии – случайные величины, линия регрессии также случайна. Поэтому предсказываемые значения $\hat{y}$ случайны и имеют некоторое стандартное отклонение $S_{\hat{y}}(x^1, ..., x^p)$, зависящее от *X*. Благодаря этому можно получить и доверительные границы для прогнозных значений регрессии (математических ожиданий $M(y)$).

Кроме того, с учетом дисперсии остатка могут быть вычислены доверительные границы значений *y* (не средних, а индивидуальных!).

Для каждого объекта может быть вычислен остаток $\hat{\varepsilon}_i = y_i - \hat{y}_i$ (оценка ε_i). Остаток полезен для изучения адекватности модели данным. Это означает, что должны быть выполнены требования независимости остатков для отдельных наблюдений, дисперсия не должна зависеть от *X*.

Для изучения отклонений от модели удобно использовать стандартизованный остаток, деленный на стандартную ошибку регрессии.

Случайность оценки коэффициентов регрессии вносит дополнительную дисперсию в регрессионный остаток, из-за этого дисперсия остатка зависит от значений независимых переменных ($S_x = S_x(x_1, \ldots, x_p)$). Стьюдентеризованный остаток – это остаток, деленный на оценку дисперсии остатка: (**Sresid** = $S_x(x_1, \ldots, x_p)$).

Таким образом, мы можем получить: (прогнозную) оценку значений зависимой переменной **Unstandardized predicted value**), ее стандартное отклонение (**S.E. of mean predictions**), доверительные интервалы для математического ожидания $M(y(x))$ и для индивидуального значения $y(x)$. В окне, включенном кнопкой **Save**, такое сохранение назначается в разделе **Prediction intervals** включением позиций **Mean** и **Individual**.

Это далеко не полный перечень переменных, порождаемых SPSS.

6.1.6. Взвешенная регрессия

Пусть прогнозируется вес ребенка в зависимости от его возраста. Ясно, что дисперсия веса для четырехлетнего ребенка будет значительно меньше, чем дисперсия веса 14-летнего юноши. Таким образом, дисперсия остатка ε_i зависит от значений *x*, а значит, условия для оценки регрессионной зави-

симости не выполнены. Проблема неоднородности дисперсии в регрессионном анализе называется проблемой гетероскедастичности.

Гетероскедастичность для сгруппированных данных может быть обнаружена с помощью сравнения дисперсий в группах (критерий Ливиня), визуально или на основании содержания задачи (как в описанном выше примере).

В SPSS имеется возможность корректно сделать соответствующие оценки за счет приписывания весов слагаемым минимизируемой суммы квадратов. Эта весовая функция должна быть равна $1/\sigma^2(x)$, где $\sigma^2(x)$ – дисперсия y как функция от x. Естественно, чем меньше дисперсия остатка на объекте, тем больший вес он будет иметь. В качестве такой функции можно использовать ее оценку, полученную при фиксированных значениях x.

Например, в приведенном примере на достаточно больших данных можно оценить дисперсию для каждой возрастной группы и вычислить необходимую весовую переменную. Увеличение влияния возрастных групп с меньшим возрастом в данном случае вполне оправданно.

В диалоговом окне назначение весовой переменной производится с помощью кнопки **WLS** (*Weighed Least Squares* – метод взвешенных наименьших квадратов).

В SPSS имеется возможность оценить весовую переменную как степенную функцию переменной, с которой, по предположению, может быть связан вес. Для этого используется команда **Weight Estimation** из раздела меню **Regression**.

6.1.7. Команда построения линейной модели регрессии

В меню это команда **Linear Regression**. В диалоговом окне команды:

– Назначаются независимые и зависимая переменные;

– Назначается метод отбора переменных. **Stepwise –** пошаговое включение/удаление переменных. **Forward** – пошаговое включение переменных. **Backward** – пошаговое исключение переменных. При пошаговом алгоритме назначаются значимости включения и исключения переменных (**Optios**). **Enter** – принудительное включение.

– Имеется возможность отбора данных, на которых будет оценена модель (**Selection**). Для остальных данных могут быть оценены значения функции регрессии, их стандартные отклонения и др.

– Имеется возможность назначения вывода статистик (**Statistics**) – доверительные коэффициенты коэффициентов регрессии, их ковариационная матрица, статистика Дарбина – Уотсона (для проверки независимости остатков) и пр.

– Задаются графики рассеяния остатков, их гистограммы (**Plots**);

– Назначается сохранение переменных (**Save**), порождаемых регрессией;

– Если используется пошаговая регрессия, назначаются пороговые значимости для включения (**PIN**) и исключения (**POUT**) переменных (**Options**);

– Если обнаружена гетероскедастичность, назначается и весовая переменная.

6.1.8. Пример построения модели

Обычно демонстрацию модели начинают с простейшего примера, и такие примеры Вы можете найти в [7]. Мы пойдем немного дальше и покажем, как получить полиномиальную регрессию.

Курильский опрос касался населения трудоспособного возраста. Как показали расчеты, в среднем меньшие доходы имеют молодые люди и люди старшего возраста. Поэтому прогнозировать доход лучше квадратичной кривой, а не простой линейной зависимостью. В рамках линейной модели это можно сделать, введя переменную квадрат возраста. Приведенное ниже задание SPSS предназначено для прогноза логарифма промедианного дохода (ранее сформированного).

```
COMPUTE  v9_2 = v9**2.
*квадрат возраста.
REGRESSION /DEPENDENT lnv14m /METHOD = ENTER v9 v9_2
/SAVE PRED MCIN ICIN.
*регрессия с сохранением предсказанных значений и
доверительных интервалов средних и индивидуальных про-
гнозных значений.
```

В табл. 6.1 показано, что уравнение объясняет всего 4,5 % дисперсии зависимой переменной (коэффициент детерминации R^2 = 0,045), скорректированная величина коэффициента равна 0,042, а коэффициент множественной корреляции равен 0,211. Много это или мало, трудно сказать, поскольку у нас нет подобных результатов на других данных, но то, что здесь есть взаимосвязь, можно определить на основании табл. 6.2.

Таблица 6.1

Общие характеристики уравнения

R	R Square	Adjusted R Square	Std. Error of the Estimate
.211	.045	.042	.5277

a) Predictors: (Constant), V9_2, V9 Возраст

b) Dependent Variable: LNV14M логарифм промедианного дохода

Результаты дисперсионного анализа уравнения регрессии показывают, что гипотеза равенства всех коэффициентов регрессии нулю должна быть отклонена.

Таблица 6.2

Дисперсионный анализ уравнения

	Sum of Squares	df	Mean Square	F	Sig.
Regression	8,484	2	4,242	15,232	,000
Residual	181,298	651	0,278		
Total	189,782	653			

Таблица 6.3

Коэффициенты регрессии

	Unstandardized Coefficients		Standardized Coefficients	T	Sig.
	B	Std. Error	Beta		
(Constant)	–1,0569	0,1888		–5,5992	0,0000
V9 Возраст	0,0505	0,0093	1,1406	5,4267	0,0000
V9_2	–0,0006	0,0001	–1,0829	–5,1521	0,0000

Регрессионные коэффициенты представлены в табл. 6.3. В соответствии с ними уравнение регрессии имеет вид

$$\text{Лог. промед. дохода} = -1{,}0569 + 0{,}0505 \times \text{возраст} - 0{,}0006 \times \text{возраст}^2.$$

Стандартная ошибка коэффициентов регрессии значительно меньше величин самих коэффициентов, их отношения – *t*–статистики по абсолютной величине больше 5. Наблюдаемая значимость статистик (Sig) равна нулю, поэтому гипотеза о равенстве коэффициентов нулю отвергается для каждого коэффициента. Стоит обратить внимание на редкую ситуацию – коэффициенты бета по абсолютной величине больше 1. Это произошло, по-видимому, из-за того, что корреляция между возрастом и его квадратом весьма велика.

Рис. 6.1 показывает линию регрессии и доверительные границы для $M(y)$ – математического ожидания y и для индивидуальных значений y. Он получается с помощью наложения полей рассеяния возраста с зависимой переменной, с переменной – прогнозом, с переменными – доверительными границами:

```
  GRAPH /SCATTERPLOT(OVERLAY) = v9 v9 v9 v9 v9 v9
WITH pre_1 lmci_1 umci_1 lici_1 uici_1 lnv14m(PAIR).
```

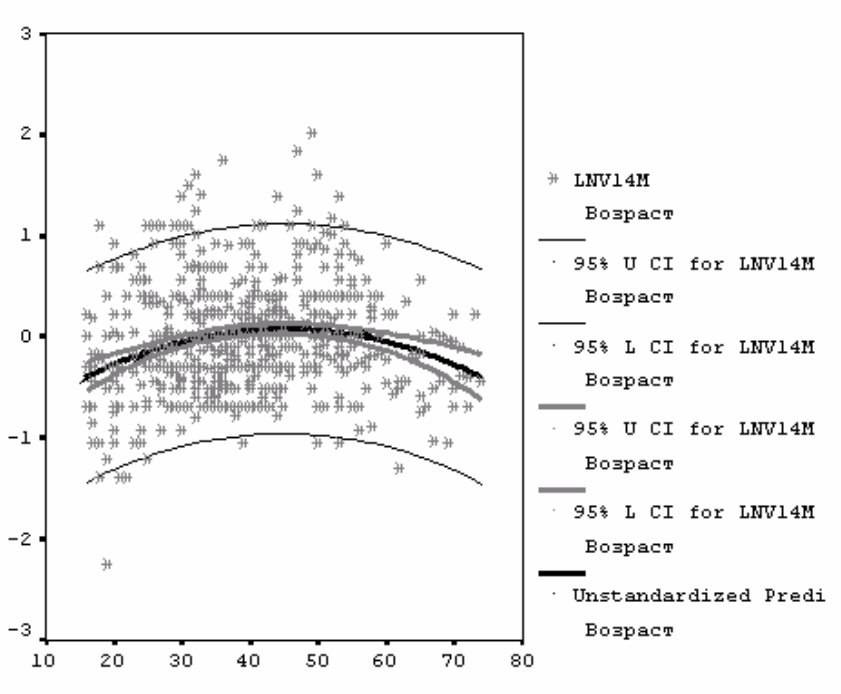


Рис. 6.1. Зависимость логарифма душевого дохода от возраста, доверительные интервалы предсказания и среднего предсказания

Границы для $M(y)$ (матожидания y) значительно уже, чем для y, так как последние должны охватывать больше 95 % точек графика.

На графике не прослеживается явной зависимости дисперсии остатка от значений независимой переменной – возраста. Некоторое суживание рассеяния данных для старших возрастов произошло, вероятно, за счет общего уменьшения плотности двумерного распределения.

6.1.9. Можно ли в регрессии использовать неколичественные переменные?

Определенно можно сказать, что неколичественные переменные не могут быть использованы в качестве зависимой переменной y. Это было бы грубейшей ошибкой; в таком случае уравнением регрессии может быть предсказан, к примеру, пол, имеющий код 1,5, или 0,5 при общепринятой кодировке пола 1 – мужчины, 2 – женщины.

В качестве независимой переменной применяются индексные переменные (в англоязычной литературе *dummy-variables*).

Например, для семейного положения в данных Курильского обследования (женат, вдов, разведен, холост) стоит ввести три индикаторные переменные: t_1, t_2 и t_3 для выделения женатых, вдовых и разведенных. Эти переменные будут равны, соответственно, 1 или 0, в зависимости от того, принадлежит или не принадлежит респондент к соответствующей группе.

Почему не 4, а 3 индексные переменные? Четвертая переменная определяется однозначно через первые три, поэтому введение ее вызвало бы коллинеарность, не позволяющую найти коэффициенты регрессии.

Ниже приведена программа, позволяющая изучить зависимость душевого дохода от возраста и семейного положения:

```
  COMPUTE lnv14m  = ln(v14/200).
  COMPUTE t1 = (v11 = 1).
  COMPUTE t2 = (v11 = 2).
  COMPUTE t3 = (v11 = 3).
  COMPUTE v9_2 = v9**2.
  *квадрат возраста.
  REGRESSION /DEPENDENT lnv14m /METHOD = ENTER v9
v9_2 t1 t2 t3 /SAVE PRED.
```

График связи возраста (V9) с предсказанным уравнением логарифмом доходов (переменная pre_2) получается командой

```
  GRAPH /SCATTERPLOT(BIVAR) = v9 WITH pre_2
/MISSING = LISTWISE
```

Он представляет собой 4 параболы (рис. 6.2). В соответствии с коэффициентами перед `t1, t2` и `t3` (см. табл. 6.4), эти параболы соответствуют – сверху вниз – группам холостяков, разведенных, женатых и вдовцов (парабола холостяков получается при `t1 = t2 = t3 = 0`).

Вероятно, полученное уравнение можно улучшить, исключив из него переменные с незначимыми коэффициентами. Поскольку индексные переменные должны быть в определенной степени взаимосвязаны, уровень наблюдаемой значимости может определяться здесь коллинеарностью, поэтому «ревизию» переменных нужно проводить осторожно, чтобы существенно не ухудшить полученного уравнения.

Из-за взаимосвязи переменных здесь нет возможности говорить о том, какая переменная больше влияет на зависимую переменную. Обратите внимание на довольно редкий эффект: β-коэффициенты для возраста и его квадрата по абсолютной величине больше 1!

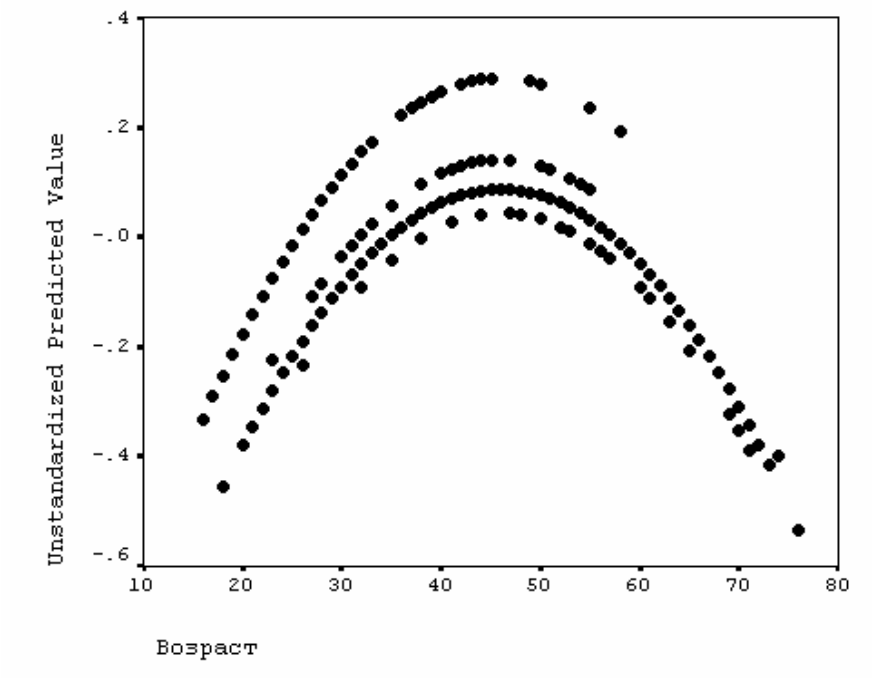


Рис. 6.2. Зависимость логарифма душевого дохода от возраста и семейного положения

Таблица 6.4

Коэффициенты регрессии с индексными переменными

	B	Std. Error	Beta	T	Sig.
(Constant)	–1,1721	0,1937		–6,0500	0,0000
V9 Возраст	0,0635	0,0105	1,4298	6,0299	
V9_2	–0,0007	0,0001	–1,3243	–5,7351	
T1 Женат	–0,2030	0,0766	–0,1540	–2,6488	0,0000
T2 Вдовец	–0,2471	0,1352	–0,0850	–1,8279	0,0000
T3 Разведен	–0,1494	0,1134	–0,0661	–1,3176	0,1881

Кроме того, модель с тремя «параллельными» параболами, вероятно, не полностью адекватна – каждая группа может иметь свою конфигурацию линии регрессии. Для учета этого в уравнении стоит использовать переменные взаимодействия. Вопросам их конструирования посвящен следующий раздел.

6.1.10. Взаимодействие переменных

Предположим, что мы рассматриваем пару индикаторных переменных: x^1 – для выделения группы женатых и x^2 – для выделения группы «начальников», а прогнозируем с помощью уравнения регрессии все тот же логарифм дохода: $y = B_0 + B_1 \times x^1 + B_2 \times x^2$.

Это уравнение моделирует ситуацию, когда действие факторов x^1 и x^2 складывается, т. е. считается, например, что женатый начальник имеет зарплату $B_1 + B_2$, неженатый начальник – B_2. Это достаточно смелое предположение, так как, скорее всего, закономерность не так груба и существует взаимодействие между факторами, в результате которого их совместный вклад имеет другую величину. Для учета такого взаимодействия можно ввести в уравнение переменную, равную произведению x^1 и x^2:

$$y = B_0 + B_1 \times x^1 + B_2 \times x^2 + B_3 \times x^1 \times x^2.$$

Произведение $x^1 \times x^2$ равно единице, если факторы действуют совместно и нулю, если какой-либо из факторов отсутствует. Аналогично можно поступить для учета взаимодействия обычных количественных переменных, а также индексных переменных с количественными.

Для получения переменных взаимодействия следует воспользоваться средствами преобразования данных SPSS.

6.2. Логистическая регрессия

Предсказания событий, исследования связи событий с теми или иными факторами с нетерпением ждут от социологов. Будем считать, что событие в данных фиксируется дихотомической переменной (0 – не произошло со-

бытие, 1 – произошло). Для построения модели предсказания можно было бы построить, например, линейное регрессионное уравнение с зависимой дихотомической переменной y, но оно будет неадекватно поставленной задаче, так как в классическом уравнении регрессии предполагается, что y – непрерывная переменная. С этой целью рассматривается логистическая регрессия. Ее целью является построение модели прогноза вероятности события $\{y = 1\}$ в зависимости от независимых переменных $x^1, \ldots, x^p$. Иначе эта связь может быть выражена в виде зависимости $P\{y = 1| x\} = F(x)$.

Логистическая регрессия выражает эту связь в виде формулы

$$P\{y = 1|x^1,...,x^p\} = \frac{e^Z}{1+e^Z}, \text{ где } Z = B_0 + B_1x^1 + \ldots + B_px^p.$$

Название «логистическая регрессия» происходит от названия логистического распределения, имеющего функцию распределения

$$F(x) = \frac{e^{(x-a)/k}}{1-e^{(x-a)/k}}.$$

Таким образом, искомая вероятность здесь ищется в виде значения этой функции распределения, в которую в качестве аргумента подставлена линейная комбинация независимых переменных.

6.2.1. Отношение шансов и логит

Отношение вероятности того, что событие произойдет, к вероятности того, что оно не произойдет, $P/(1 – P)$, называется отношением шансов.

С этим отношением связано еще одно представление логистической регрессии, получаемое за счет непосредственного задания зависимой переменной в виде $Z = Ln\,(P/(1 – P))$, где $P = P\{y = 1| X^1,\ldots, X^p\}$. Переменная Z называется ***логитом***. По сути дела, логистическая регрессия определяется уравнением регрессии $Z = B_0 + B_1x^1 + \ldots + B_px^p$.

В связи с этим отношение шансов может быть записано в следующем виде:

$$P/(1–P) = e^{B_0+B_1X^1+B_2X^2+\ldots+B_pX^p} = e^{B_0}e^{B_1X^1}\ldots e^{B_pX^p} = e^{B_0}(e^{B_1})^{X^1}\ldots(e^{B_p})^{X^p}.$$

Отсюда получается, что, если модель верна, изменение x^k на единицу при независимых $x^1, \ldots, x^p$ вызывает изменение отношения шансов в e^{B_k} раз.

6.2.2. Решение уравнения с использованием логита

Решение такого уравнения упрощенно можно представить следующим образом

1. Получаются агрегированные данные по переменным x, в которых для каждой группы, характеризуемой значениями $x_j = (x_j^1,...,x_j^p)$, подсчитывается доля объектов, соответствующих событию $\{y = 1\}$. Эта доля является оценкой вероятности $\hat{P}_j = P\{y = 1 | x_j^1,...,x_j^p\}$. В соответствии с этим для каждой группы получается значение логита Z_j.

2. На агрегированных данных оцениваются коэффициенты уравнения $Z = B_0 + B_1 x^1 + \ldots + B_p x^p$. К сожалению, дисперсия Z здесь зависит от значений x, поэтому при использовании логита применяется специальная техника оценки коэффициентов – взвешенной регрессии.

Еще одна особенность состоит в том, что в реальных данных очень часто группы по x оказываются однородными по y, поэтому оценки $\hat{P}_j$ оказываются равными 0 или 1. Таким образом, оценка логита для них не определена (для этих значений $Z = Ln(0/(1-0)) = -\infty$, $Z = Ln(1/(1-1)) = \infty$).

В некоторых статистических пакетах такие группы объектов просто-напросто отбрасываются.

В настоящее время в статистическом пакете для оценки коэффициентов используется метод максимального правдоподобия, лишенный этого недостатка. Тем не менее проблема, хотя и не в таком остром виде, остается: если оценки вероятности для многих групп оказываются равными 0 или 1, оценки коэффициентов регрессии имеют слишком большую дисперсию. Поэтому, имея в качестве независимых переменных такие признаки, как душевой доход в сочетании с возрастом, их следует укрупнить по интервалам, приписав объектам средние значения интервалов.

6.2.3. Неколичественные данные

Если в обычной линейной регрессии для работы с неколичественными переменными нам приходилось подготавливать специальные индикаторные переменные, то в реализации логистической регрессии в SPSS это делается автоматически. Для этого в диалоговом окне специально предусмотрены средства, сообщающие пакету, что ту или иную переменную следует считать категориальной. При этом, чтобы не получить линейно зависимых переменных, максимальный код ее значения (или минимальный, в зависимости от задания процедуры) не перекодируется в дихотомическую (индексную) переменную. Впрочем, средства преобразования данных позволяют не учитывать любой код значения. Имеются другие способы перекодирования категориальных (неколичественных) переменных в несколько переменных, но мы будем пользоваться только указанным способом как наиболее естественным.

6.2.4. Взаимодействие переменных

В процедуре логистической регрессии в SPSS предусмотрены средства для автоматического включения в уравнение переменных взаимодействий. В диалоговом окне в списке исходных переменных для этого следует выделить имена переменных, взаимодействия которых предполагается рассмотреть, затем переправить выделенные имена в окно независимых переменных кнопкой с текстом > a × b >.

6.2.5. Пример логистической регрессии и статистики

Процедура логистической регрессии в SPSS в диалоговом режиме вызывается из меню командой **Statistics \Regression \Binary logistic…**.

В качестве примера по данным RLMS изучим, как связано употребление спиртных напитков с зарплатой, полом, статусом (ранг руководителя), курит ли он. Для этого подготовим данные: выберем в обследовании RLMS население старше 18 лет, сконструируем индикаторы курения (`smoke`) и употребления спиртных напитков (`alcohol`) (в обследовании задавался вопрос «Употребляли ли Вы в течение 30 дней алкогольные напитки»):

```
COMPUTE filter_$ = (vozr>18).
FILTER BY filter_$.
COMPUTE smoke = (dm71 = 1).
VAL LAB smoke 1 "курит" 0 "не курит".
COMPUTE alcohol = (dm80 = 1).
VAL LAB alcohol 1 "пьет" 0 "не пьет".
```

Укрупним переменную `dj10` (зарплата на основном рабочем месте). В данном случае группы по значениям этой переменной в основном достаточно наполнены. С методической целью покажем один из способов укрупнения. Для этого вначале получаем переменную `wage`, которая содержит номера децилей по зарплате, затем среднюю зарплату по этим децилям (см. табл. 6.5). Это осуществляется приведенной ниже программой:

```
  MISSING VALUES dj6.0 (9997,9998,9999)
dj10(99997,99998,99999).
  RANK VARIABLES = dj10 (A) /NTILES (10) into wage
/PRINT = YES /TIES = MEAN .
  MEANS TABLES = dj10 BY wage /CELLS MEAN,
```

в результате которой получается

Таблица 6.5

Средняя зарплата по децилям

WAGE децили зарплаты	1	2	3	4	5	6	7	8	9	10
DJ10 зарплата за 30 дней	1,01	2,11	3,07	4,16	5,42	7,03	8,53	11,08	15,65	34,64

Полученные средние используем для формирования переменной, соответствующей укрупненной зарплате (для удобства, чтобы коэффициенты регрессии не были слишком малы, в качестве единицы ее измерения возьмем сто рублей).

```
  RECODE wage (1 = 1.01) (2 = 2.11) (3 = 3.07) (4 = 4.16)
(5 = 5.42) (6 = 7.03) (7 = 8.53) (8 = 11.08) (9 = 15.65)
(10 = 34.64).
  RECODE dj6.0 (SYSMIS = 4)(1 THRU 5 = 1)(6 thru 10 = 2)
(10 THRU HI = 3) INTO manag.
  VAR LAB manag "статус" wage "заработок".
  VAL LAB manag 4 "не начальник" 1 "шеф"
2 "начальничек" 3 "начальник".
  EXEC.
```

Далее формируем переменную `manag` («статус») из переменной `dj6.0` – «количество подчиненных».

Запускаем команду построения регрессии `LOGISTIC REGRESSION`, в которой использованы переменные `wage` – зарплата, `manag` – статус, `dh5` – пол (1 – мужчины, 2 – женщины) `smoke` – курение (1 – курит, 0 – не курит), `dh5 wage` – «взаимодействие» пола с зарплатой (для женщин значение – 0, для мужчин – совпадает с зарплатой).

```
  LOGISTIC REGRESSION VAR = alcohol /METHOD = ENTER
wage manag dh5 smoke dh5*wage /CONTRAST (dh5) =
Indicator /CONTRAST (manag) = Indicator /CONTRAST
(smoke) = Indicator /PRINT = CI(95) /CRITERIA PIN(.05)
POUT(.10) ITERATE(20) CUT(.69).
```

В выдаче программа сообщает прежде всего о перекодировании данных:

Dependent Variable Encoding:	
Original Value	Internal Value
.00	0
1.00	1

Следует обратить внимание, что зависимая переменная здесь должна быть дихотомической, и ее максимальный код считается кодом события, вероятность которого прогнозируется. Например, если Вы закодировали переменную `ALCOHOL`: (1 – употреблял, 2 – не употреблял), то будет прогнозироваться вероятность неупотребления алкоголя.

Далее идут сведения о кодировании индексных переменных для категориальных переменных; из-за их естественности здесь мы их не приводим.

Далее следуют обозначения для переменных взаимодействия, в нашем простом случае это

```
Interactions:
INT_1   DH5(1) by WAGE
```

6.2.6. Качество логистической регрессии

Далее в выдаче появляется описательная информация о качестве модели:

```
-2 Log Likelihood          3289.971
 Goodness of Fit           2830.214
 Cox & Snell - R^2            .072
 Nagelkerke - R^2             .102
```

которые означают:

– `-2 Log Likelihood` – удвоенный логарифм функции правдоподобия со знаком минус;

– `Goodness of Fit` – характеристика отличия наблюдаемых частот от ожидаемых;

– `Cox & Snell - R^2` и `Nagelkerke - R^2` – псевдокоэффициенты детерминации, полученные на основе отношения функций правдоподобия модели с константой к модели со всеми коэффициентами.

Эти коэффициенты стоит использовать при сравнении очень похожих моделей, построенных на аналогичных данных, что практически нереально, поэтому на них мы не будем останавливаться.

6.2.7. Вероятность правильного предсказания

На основе модели логистической регрессии можно строить предсказание, произойдет или не произойдет событие $\{y = 1\}$. Правило предсказания, по умолчанию заложенное в процедуру `LOGISTIC REGRESSION`, устроено по следующему принципу: если $\hat{P}_j = P\{y = 1 | x_j^1, ..., x_j^p\} > 0{,}5$, то считаем, что событие произойдет; если $\hat{P}_j = P\{y = 1 | x_j^1, ..., x_j^p\} \le 0{,}5$, то считаем, что событие не произойдет. Это правило оптимально с точки зрения минимизации числа ошибок, но очень грубо с точки зрения исследования связи. Зачастую оказывается, что вероятность события $P\{y = 1\}$ мала (значительно меньше 0,5), тогда все имеющиеся в данных сочетания x предсказывают противоположное событие, или велика (значительно больше 0,5), поэтому оказывается, что они предсказывают событие $\{y = 1\}$.

Поэтому необходима другая классификация, которая демонстрирует связь между зависимой и независимыми переменными. С этой целью стоит выделить два типа объектов:

– объекты, имеющие повышенную вероятность события $\{y = 1\}$, для которых оцененная условная вероятность $P\{y = 1|x\}$, больше безусловной оценки вероятности $P\{y = 1\}$ (доли объектов, для которых $y = 1$);

– объекты, имеющие повышенную вероятность противоположного события $\{y = 0\}$, для которых оцененная условная вероятность $P\{y = 1|x\}$ меньше оценки безусловной вероятности $P\{y = 1\}$.

В нашем случае доля объектов, для которых $y = 1$, равна 0,69. Поэтому в процедуре указан параметр `/CRITERIA CUT(.69)`. Связь между этими классификациями представлена в таблице сопряженности (рис. 6.3). Но лучше, пользуясь EXCEL или калькулятором, в этой таблице вычислить процентные соотношения.

```
 Classification Table for ALCOHOL
  The Cut Value is .69
                        Predicted
                     не пьет     пьет     Percent
  Observed
     не пьет   н         541       340      61.41%

     пьет      п         694      1244      64.19%
```

Таблица 6.6

Связь наблюдения и предсказания в логистической регрессии

Наблюдается	Предсказанный тип		Всего
	Не пьет	Пьет	
Не пьет	43,8 %	21,5 %	31,3 %

Рис. 6.3. Классификационная таблица

Пьет	56,2 %	78,5 %	68,7 %

6.2.8. Коэффициенты логистической регрессии

Основная информация содержится в таблице коэффициентов регрессии (рис. 6.4). Прежде всего, следует обратить внимание на значимость коэффициентов. Наблюдаемая значимость вычисляется на основе статистики

```
Variable             B   S.E.      Wald df   Sig     R

WAGE             .0432 .0078   30.4619  1 .0000  .0902
MANAG                          9.9788  3 .0187  .0337
 MANAG(1)        .3544 .1489   5.6637  1 .0173  .0323
 MANAG(2)        .5241 .2328   5.0673  1 .0244  .0296
```

```
 MANAG(3)          .0393 .1580    .0618  1 .8036  .0000
SMOKE              .6419 .1074  35.6956  1 .0000  .0981
DH5(1)             .8801 .1366  41.5022  1 .0000  .1062
DH5(1) by WAGE    -.0390 .0101  14.7972  1 .0001 -.0605
Constant          -.0534 .0767    .4852  1 .4861
```

Вальда. Эта статистика связана с методом максимального правдоподобия и может быть использована при оценках разнообразных параметров.

Универсальность статистики Вальда позволяет оценить значимость не только отдельных переменных, но и в целом значимость категориальных переменных, несмотря на то что они дезагрегированы на индексные переменные. Статистика Вальда имеет распределение хи-квадрат. Число степе-

Рис. 6.4. Коэффициенты логистической регрессии

ней свободы равно единице, если проверяется гипотеза о равенстве нулю коэффициента при обычной или индексной переменной, а для категориальной переменной – числу значений без единицы (т. е. числу соответствующих индексных переменных). Квадратный корень из статистики Вальда приближенно равен отношению величины коэффициента к его стандартной ошибке – так же выражается *t*-статистика в обычной линейной модели регрессии.

В нашей таблице коэффициентов почти все переменные значимы на уровне значимости 5 %. Закрыв глаза на возможное взаимодействие между независимыми переменными (коллинеарность), можно считать, что вероятность употребления алкоголя повышена при высокой зарплате, а также у руководителей различного ранга. Из-за незначимости статистики Вальда нет, правда, полной уверенности относительно повышенной вероятности для начальников, имеющих более 10 подчиненных. Курение и принадлежность к мужскому полу также повышают эту вероятность, однако, взаимодействие «мужчина – зарплата» имеет обратное действие.

В этой же таблице присутствует аналог коэффициента корреляции (R), также построенный на основании статистики Вальда. Для обычных и индексных переменных положительные значения коэффициента свидетельствуют о положительной связи переменной с вероятностью события, отрицательные – об отрицательной связи.

Кроме того, мы выдали таблицу экспонент коэффициентов e^B и их доверительные границы (см. рис. 6.5). Эта таблица выдана подкомандой `/PRINT = CI(95)` в команде задания логистической регрессии.

```
                                        95% CI for Exp(B)
  Variable              Exp(B)      Lower      Upper
```

WAGE	1.0441	1.0282	1.0603
MANAG(1)	1.4253	1.0645	1.9083
MANAG(2)	1.6889	1.0701	2.6654
MANAG(3)	1.0401	.7630	1.4177
SMOKE	1.9001	1.5393	2.3455
DH5(1)	2.4112	1.8448	3.1515
DH5(1) by WAGE	.9618	.9429	.9811

Согласно модели и полученным значениям коэффициентов при фиксированных прочих переменных принадлежность к мужскому полу увеличивает отношение шансов «пития» и «не пития» в 2,4 раза (точнее, в 1,84 – 3,15 раза), курения – в 1,9 раза (1,54 – 2,35), а прибавка к зарплате 100 рублей – на 4,4 % (2,8 – 6 %). Правда, такая прибавка мужчине одно-

Рис. 6.5. Экспоненты коэффициентов

временно уменьшает это отношение на 3,8 % (5,7 % – 1,9 %). Быть начальником низкого ранга – значит увеличить отношение шансов в 1,43 (1,06 – 1,9) раза, чем в среднем, а начальником среднего ранга – в 1,7 (1,07 – 2,67) раза.

6.2.9. О статистике Вальда

Как отмечено в документации SPSS, недостаток статистики Вальда состоит в том, что при малом числе наблюдений она может давать заниженные оценки наблюдаемой значимости коэффициентов. Для получения более точной информации о значимости переменных можно воспользоваться пошаговой регрессией, метод *forward lr* (*lr – likelyhood ratio* – отношение правдоподобия). Тогда для каждой переменной будет выдана значимость включения/исключения, полученная на основе отношения функций правдоподобия модели. Первые выводы удобнее делать на основе статистики Вальда, а потом уже уточнять результаты, если это необходимо.

6.2.10. Сохранение переменных

Программа позволяет сохранить множество переменных, среди которых наиболее полезной является, по-видимому, предсказанная вероятность.

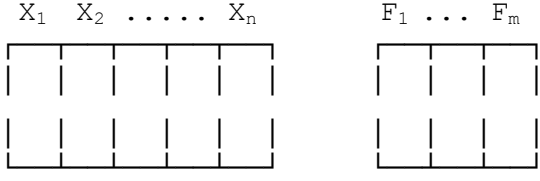
Глава 7. ИССЛЕДОВАНИЕ СТРУКТУРЫ ДАННЫХ

Хотя исследователь и руководствуется определенными гипотезами в процессе сбора данных, информация нередко представляет собой сырой материал, в котором нужно изучить структуру показателей, характеризующих объекты, а также выявить однородные группы объектов. Полезно представить эту информацию в геометрическом пространстве, лаконично отразить ее особенности в классификации объектов и переменных. Такая

работа создает предпосылки для построения типологий объектов и определения «социального пространства», в котором обозначены расстояния между объектами наблюдения.

7.1. Факторный анализ

Идея метода состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица. В основе моделей факторного анализа лежит гипотеза, что наблюдаемые переменные являются косвенными проявлениями небольшого числа скрытых (латентных) факторов. Под моделью факторного анализа понимают представление исходных переменных в виде линейной комбинации факторов.



$$X_i = \sum_{k=1} a_{ik} F_k + U_i$$

Рис. 7.1. Сжатие признакового пространства
с применением факторного анализа

Факторы F построены так, чтобы наилучшим способом (с минимальной погрешностью) представить X. В этой модели «скрытые» переменные F_k называются общими факторами, а переменные U_i – специфическими факторами («специфический» – это лишь одно из значений используемого в англоязычной литературе слова *unique*, в отечественной литературе в качестве определения U_i встречаются также слова «характерный», «уникальный»). Значения a_{ik} называются факторными нагрузками.

Обычно (хотя и не всегда) предполагается, что X_i стандартизованы ($\sigma_i = 1$, $X_i = 0$), а факторы F_1, F_2, …, F_m независимы и не связаны со специфическими факторами U_i (существуют модели, выполненные в других предположениях). Предполагается также, что факторы F_i стандартизованы.

В этих условиях факторные нагрузки a_{ik} совпадают с коэффициентами корреляции между общими факторами и переменными X_i. Дисперсия X_i раскладывается на сумму квадратов факторных нагрузок и дисперсию специфического фактора:

$$S_{x_i}^2 = H_i^2 + S_{u_i}^2, \text{ где } H_i^2 = \sum_k a_{ik}{}^2 .$$

Величина H_i^2 называется общностью, $S_{u_i}^2$ – специфичностью. Другими словами, общность представляет собой часть дисперсии переменных, объ-

ясненную факторами, специфичность – часть не объясненной факторами дисперсии.

В соответствии с постановкой задачи необходимо искать такие факторы, при которых суммарная общность максимальна, а специфичность – минимальна.

7.1.1. Метод главных компонент

Один из наиболее распространенных методов факторного анализа – метод главных компонент – состоит в последовательном поиске факторов. Вначале ищется первый фактор, который объясняет наибольшую часть дисперсии, затем не зависимый от него второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т. д. Описание всей математики построения факторов слишком сложно, поэтому для пояснения сути мы прибегнем к зрительным образам (рис. 7.2).

Геометрически это выглядит следующим образом. Для построения первого фактора берется прямая, проходящая через центр координат и облако рассеяния данных. Объектам можно сопоставить расстояния их проекций на эту прямую до центра координат, причем для одной из половин прямой (по отношению к нулевой точке) можно взять эти расстояния с отрицательным знаком. Такое построение представляет собой новую переменную, которую мы назовем осью. При построении фактора отыскивается такая ось, чтобы ее дисперсия была максимальна. Это значит, что данной осью объясняется максимум дисперсии переменных. Найденная ось после нормировки используется в качестве первого фактора. Если облако данных вытянуто в виде эллипсоида (имеет форму «огурца»), фактор совпадет с направлением, в котором вытянуты объекты, и по нему (по проекциям) с наибольшей точностью можно предсказать значения исходных переменных.

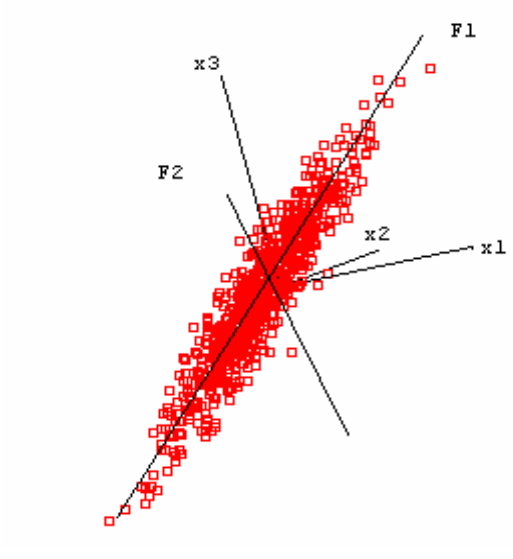


Рис. 7.2. Главные компоненты

Для поиска второго фактора опре-

деляется ось, перпендикулярная первому фактору, также объясняющая наибольшую часть дисперсии, не объясненную первой осью. После нормировки эта ось становится вторым фактором. Если данные представляют собой плоский эллипсоид («блин») в трехмерном пространстве, два первых фактора позволяют в точности описать эти данные.

Максимально возможное число главных компонент равно количеству переменных. Сколько главных компонент необходимо построить для оптимального представления рассматриваемых исходных факторов?

Обозначим λ_k объясненную главной компонентой F_k часть суммарной дисперсии совокупности исходных факторов. По умолчанию в пакете предусмотрено продолжение построения факторов до тех пор, пока $\lambda_к > 1$. Напомним, что переменные стандартизованы, и поэтому нет смысла строить очередной фактор, если он объясняет часть дисперсии, меньшую, чем приходящаяся непосредственно на одну переменную. При этом следует учесть, что $\lambda_1 > \lambda_2 > \lambda_3, \ldots$.

К сведению читателя заметим, что значения λ_k являются также собственными значениями корреляционной матрицы X_i, поэтому в выдаче они будут помечены текстом **EIGEN VALUE**, что в переводе означает «собственные значения».

Заметим, что техника построения главных компонент расходится с теоретическими предположениями о факторах: имеется $m + n$ независимых факторов (включая уникальные), полученных методом главных компонент в n-мерном пространстве, что невозможно.

7.1.2. Интерпретация факторов

Как же понять, что скрыто в найденных факторах? Основной информацией, которую использует исследователь, являются факторные нагрузки. Для интерпретации необходимо приписать фактору термин. Этот термин появляется на основании анализа корреляций фактора с исходными переменными. Например, при анализе успеваемости школьников фактор имеет высокую положительную корреляцию с оценкой по алгебре, геометрии и большую отрицательную корреляцию с оценками по рисованию – он характеризует точное мышление.

Не всегда такая интерпретация возможна. Для повышения интерпретируемости факторов добиваются большей контрастности матрицы факторных нагрузок. Метод такого улучшения результата называется методом ***вращения факторов***. Его суть состоит в следующем: если мы будем вращать координатные оси, образуемые факторами, мы не потеряем в точности, представляя данные через новые оси, и не беда, что при этом факторы не будут упорядочены по величине объясненной ими дисперсии, зато у нас появляется возможность получить более контрастные факторные нагрузки.

Вращение состоит в получении новых факторов – в виде специального вида линейной комбинации имеющихся факторов:

$$\hat{F}_i = \sum_{k=1}^{m} b_{ik} F_k \ .$$

Чтобы не вводить новые обозначения, факторы и факторные нагрузки, полученные вращением, будем обозначать их теми же символами, что и до вращения. Для достижения цели интерпретируемости существует достаточно много методов, которые состоят в оптимизации подходящей функции от факторных нагрузок. Мы рассмотрим реализуемый пакетом метод *varimax*. Этот метод состоит в максимизации «дисперсии» квадратов факторных нагрузок для переменных:

$$\sum_i \left[\sum_k a_{ik}^4 \Big/ m - \left[\sum_k a_{ik}^2 \Big/ m \right]^2 \right] \to \max .$$

Чем сильнее разойдутся квадраты факторных нагрузок к концам отрезка [0,1], тем больше будет значение целевой функции вращения, тем четче интерпретация факторов.

7.1.3. Оценка факторов

Математический аппарат, используемый в факторном анализе, в действительности позволяет не вычислять непосредственно главные оси. И факторные нагрузки до и после вращения факторов и общности вычисляются за счет операций с корреляционной матрицей. Поэтому оценка значений факторов для объектов является одной из проблем факторного анализа.

Конкретные значения факторов, полученные с помощью метода главных компонент, определяются на основе регрессионного уравнения. Известно, что для оценки регрессионных коэффициентов для стандартизованных переменных достаточно знать корреляционную матрицу переменных. Корреляционная матрица по переменным X_i и F_k определяется исходя из модели и имеющейся матрицы корреляций X_i. На ее основе факторы находятся регрессионным методом в виде линейных комбинаций исходных переменных: $F_k = \sum_i c_{ki} X_i$.

7.1.4. Статистические гипотезы в факторном анализе

В SPSS предусмотрена проверка теста Барлетта о сферичности распределения данных. В предположении многомерной нормальности распределения здесь проверяется, не диагональна ли матрица корреляций. Если гипотеза не отвергается (наблюдаемый уровень значимости велик, скажем больше 5 %), то нет смысла в факторном анализе, поскольку направления главных осей случайны. Этот тест предусмотрен в диалоговом окне фак-

торного анализа вместе с возможностью получения описательных статистик переменных и матрицы корреляций. На практике предположение о многомерной нормальности проверить весьма трудно, поэтому факторный анализ чаще применяется без такого анализа.

7.1.5. Задание факторного анализа

Задание факторного анализа может быть весьма простым. Например, достаточно задать команду FACTOR и подкоманду VARIABLES с указанием переменных и запустить команду на счет. Однако если удобнее самому управлять расчетами, то следует задать некоторые параметры.

Рассмотрим работу такой команды на агрегированном по городам файле наших учебных данных (напоминаем, что объектами этого файла являются города, в которых проводился опрос по поводу возможности передачи Японии Курильских островов, см. выше):

```
FACTOR /VARIABLES W3D1 TO W3D6 /PLOT EIGEN
/CRITERIA FACTORS (2) /SAVE REGRESSION (ALL F).
```

Команда задана для получения факторов по переменным – долям числа респондентов, указавших различные причины неподписания договора (/VARIABLES W3D1 TO W3D6): W3D1 – нет необходимости; W3D2 – традиционное недоверие; W3D3 – незаинтересованность Японии; W3D4 – разные политические симпатии; W3D5 – нежелание Японии признать границы; W3D6 – нежелание СССР рассматривать вопрос об островах.

Подкоманда /PLOT EIGEN выдает графическую иллюстрацию долей объясненной дисперсии. Подкоманда /CRITERIA FACTORS (2) задает получение 2 факторов; если этой подкоманды не будет, программа сама определит число факторов. Задавая /SAVE REGRESSION ALL F), мы получаем регрессионным методом непосредственно в активном файле оценки всех (ALL) факторов. Это будут переменные F1, F2 с заданным нами корневым именем F и добавленными к нему номерами факторов.

Рассмотрим результаты анализа. Табл. 7.1 содержит сведения об информативности полученных главных компонент. Первый фактор объясняет часть общей дисперсии, равную 2,402 (40,04 %), фактор 2 – 1,393 (23,1 %), третий – 0,853 (14,22 %) и т. д. Первые два фактора объясняют 63,25 % дисперсии, первые три – 77,47 %. Поскольку уже третья компонента объясняет менее 1 дисперсии ($\lambda_3 = 0{,}853$), рассматривается всего 2 фактора – какой смысл рассматривать факторы, объясняющие меньше дисперсии, чем переменная из исходных данных?

Матрица факторных нагрузок представлена в табл. 7.2. Мы не будем анализировать эту матрицу, но ниже проанализируем подробнее факторные нагрузки после вращения (табл. 7.3).

Таблица 7.1

Дисперсия, объясненная факторным анализом

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative, %	Total	% of Variance	Cumulative, %
1	2,402	40,04	40,04	2,40	40,04	40,04
2	1,393	23,21	63,25	1,39	23,21	63,25
3	0,853	14,22	77,47			
4	0,719	11,98	89,45			
5	0,345	5,75	95,20			
6	0,288	4,80	100,00			

Extraction Method: Principal Component Analysis.

Таблица 7.2

Матрица факторных нагрузок

	Component	
	1	2
W3D4 разные политические симпатии	0,769	0,327
W3D1 нет необходимости, отношения нормальны	–0,723	0,26
W3D3 незаинтересованность Японии	0,674	0,578
W3D2 недоверие друг другу	–0,569	–0,315
W3D5 нежелание Японии признать границы	0,527	–0,647
W3D6 нежелание СССР рассматривать вопрос	–0,481	0,605

Таблица 7.3

Матрица факторных нагрузок после вращения факторов

	Component	
	1	2
W3D3 незаинтересованность Японии	0,887	0,049
W3D4 разные политические симпатии	0,81	–0,208
W3D2 недоверие друг другу	–0,643	0,095
W3D5 нежелание Японии признать границы	0,025	–0,834
W3D6 нежелание СССР рассматривать вопрос	–0,014	0,773
W3D1 нет необходимости, отношения нормальны	–0,416	0,646

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Анализ факторных нагрузок показал, что фактор 2 существенно связан с `W3D6` – долей считающих, что договор не подписан, так как СССР не желает рассматривать вопрос об островах. Он также отрицательно связан с долей считающих, что все беды из-за непризнания границ Японией (`W3D5`). Имеется относительно небольшая положительная его связь с переменной

W3D1 – «нет необходимости, отношения нормальны». Поэтому можно условно назвать этот фактор «фактором несоветской ориентации».

Первый фактор связан с переменными W3D3 – «нет заинтересованности Японии», W3D4 «разные политические симпатии», и несколько слабее, отрицательно, с W3D2 – «недоверие друг другу». Условно его можно назвать фактором «судьбы». Конечно, в серьезных исследованиях можно было бы проверить факторы с самых разных сторон, нам же пока достаточно пояснить принцип интерпретации, который состоит в формулировке содержания фактора, ухватывающего суть явления.

Сохраненные в виде переменных подкомандой SAVE факторы могут быть использованы для исследования данных, конструирования типологий и т. д. В частности, с помощью команды GRAPH мы получили поле рассеяния наших объектов (рис. 7.3*) – городов в пространстве двух переменных – факторов. По этому графику, например, можно заключить, что жители Александровска-Сахалинского проявили в курильском опросе наиболее «несоветскую» ориентацию. Они менее всего склонны считать, что договора нет потому, что «так сложилось» в силу «недоверия» между странами и из-за разных политических симпатий.

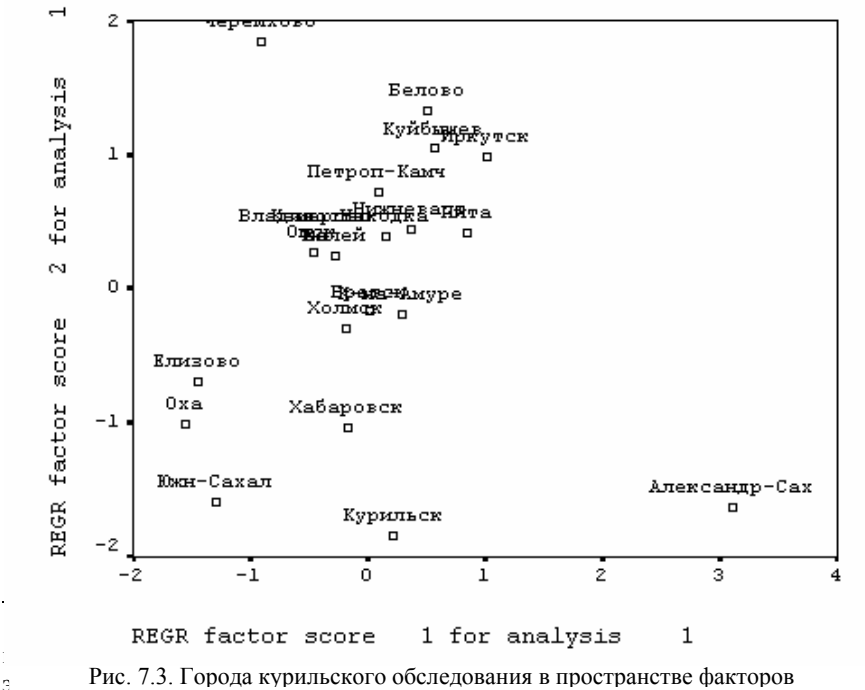


Рис. 7.3. Города курильского обследования в пространстве факторов

7.2. Кластерный анализ

Если процедура факторного анализа сжимает в малое число количественных переменных данные, описанные количественными переменными, то кластерный анализ сжимает данные в классификацию объектов. Синонимами термина *кластерный анализ* являются *автоматическая классификация объектов без учителя* и *таксономия*.

Если данные понимать как точки в признаковом пространстве, то задача кластерного анализа формулируется как выделение сгущений точек, разбиение совокупности на однородные подмножества объектов.

При проведении кластерного анализа обычно определяют расстояние на множестве объектов; алгоритмы кластерного анализа формулируют в терминах этих расстояний. Мер близости и расстояний между объектами существует великое множество. Их выбирают в зависимости от цели исследования. В частности, евклидово расстояние лучше использовать для количественных переменных, расстояние хи-квадрат – для исследования частотных таблиц, имеется множество мер для бинарных переменных.

Кластерный анализ является описательной процедурой, он не позволяет сделать никаких статистических выводов, но дает возможность провести своеобразную разведку – изучить структуру совокупности.

7.2.1. Иерархический кластерный анализ

Процедура иерархического кластерного анализа в SPSS предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют переменные, а роль переменных – столбцы.

Этот метод реализует иерархический агломеративный алгоритм. Его смысл заключается в следующем. Перед началом кластеризации все N объектов считаются отдельными кластерами, которые в ходе алгоритма объединяются. Вначале выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате количество кластеров становится равным $N - 1$. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования определяют способы вычисления расстояния между объектами и определения близости между кластерами.

Для определения расстояния между парой кластеров могут быть сформулированы различные подходы, для чего в SPSS предусмотрены методы, определяемые на основе расстояний между объектами:

– Среднее расстояние между кластерами (Between-groups linkage).

– Среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage).

– Расстояние между ближайшими соседями – ближайшими объектами кластеров (Nearest neighbour).

– Расстояние между самыми далекими соседями (Furthest neighbour).

– Расстояние между центрами кластеров (Centroid clustering), или «центроидный» метод. Недостатком этого метода является то, что центр объединенного кластера вычисляется как среднее центров объединяемых кластеров, без учета их объема.

– Метод медиан – тот же «центроидный» метод, но центр объединенного кластера вычисляется как среднее всех объектов (Median clustering).

– Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

Расстояния и меры близости между объектами. У нас нет возможности сделать полный обзор всех коэффициентов, поэтому остановимся лишь на характерных расстояниях и мерах близости для определенных видов данных.

Меры близости отличаются от расстояний тем, что они тем больше, чем более похожи объекты.

Пусть имеются два объекта $X = (X_1, \ldots, X_m)$ и $Y = (Y_1, \ldots, Y_m)$. Используя эту запись для объектов, определим основные виды расстояний, используемых в процедуре `CLUSTER`:

– Евклидово расстояние $d(X,Y) = \sqrt{\sum_{i=1}^{m}(X_i - Y_i)^2}$ (Euclidian distance).

– Квадрат евклидова расстояния $d(X,Y) = \sum_{i=1}^{m}(X_i - Y_i)^2$ (Squared Euclidian distance).

Эвклидово расстояние и его квадрат целесообразно использовать для анализа количественных данных.

– Мера близости – коэффициент корреляции

$$S(X,Y) = (\sum_{i=1}^{m} Z_{X_i} Z_{Y_i}) / (m-1),$$

где Z_{X_i} и Z_{Y_i} компоненты стандартизованных векторов X и Y. Эту меру целесообразно использовать для выявления кластеров переменных, а не объектов.

– Расстояние хи-квадрат получается на основе таблицы сопряженности, составленной из объектов X и Y,

X	X_1	...	X_m	$X.$
Y	Y_1	...	Y_m	$Y.$
$X+Y$	X_1+Y_1	...	X_m+Y_m	$X.+Y.$

которые, предположительно, являются векторами частот. Здесь рассматриваются ожидаемые значения элементов, равные $E(X_i)=X.\times(X_i+Y_i)/(X.+Y.)$ и $E(Y_i)=Y.\times(X_i+Y_i)/(X.+Y.)$, а расстояние хи-квадрат имеет вид корня из соответствующего показателя

$$d(X,Y)=\sqrt{\sum_{i=1}^{m}\frac{(X_i-E(X_i))^2}{E(X_i)}+\sum_{i=1}^{m}\frac{(Y_i-E(Y_i))^2}{E(Y_i)}}\ .$$

– Расстояние Фи-квадрат является расстоянием хи-квадрат, нормированным на число объектов в таблице сопряженности, представляемой строками X и Y, т. е. на корень квадратный из $N=X.+Y.$.

– В иерархическом кластерном анализе в SPSS также имеется несколько видов расстояний для бинарных данных (векторы X и Y состоят из нулей и единиц, обозначающих наличие или отсутствие определенных свойств объектов). Наиболее естественными из них, по-видимому, являются евклидово расстояние и его квадрат.

Стандартизация. Непосредственное использование переменных в анализе может привести к тому, что классификацию будут определять переменные, имеющие наибольший разброс значений. Поэтому применяются следующие виды стандартизации:

– Z-шкалы (Z-Scores). Из значений переменных вычитается их среднее и эти значения делятся на стандартное отклонение.

– Разброс от –1 до 1. Линейным преобразованием переменных добиваются разброса значений от –1 до 1.

– Разброс от 0 до 1. Линейным преобразованием переменных добиваются разброса значений от 0 до 1.

– Максимум 1. Значения переменных делятся на их максимум.

– Среднее 1. Значения переменных делятся на их среднее.

– Стандартное отклонение 1. Значения переменных делятся на стандартное отклонение.

– Кроме того, возможны преобразования самих расстояний, в частности, можно расстояния заменить их абсолютными значениями, это актуально для коэффициентов корреляции. Можно также все расстояния преобразовать так, чтобы они изменялись от 0 до 1.

Таким образом, работа с кластерным анализом может превратиться в увлекательную игру, связанную с подбором метода агрегирования, рас-

стояния и стандартизации переменных с целью получения наиболее интерпретируемого результата. Желательно только, чтобы это не стало самоцелью и исследователь получил действительно необходимые содержательные сведения о структуре данных.

Процесс агрегирования данных может быть представлен графически деревом объединения кластеров (Dendrogramm) либо «сосульковой» диаграммой (Icicle). Но подробнее о процессе кластеризации можно узнать по протоколу объединения кластеров (Schedule).

Пример иерархического кластерного анализа. Следующая команда осуществляет кластерный анализ по полученным нами ранее факторам на агрегированном файле курильского опроса:

```
  CLUSTER fac1_1 fac2_1 /METHOD BAVERAGE /MEASURE =
SEUCLID /ID = name /PRINT SCHEDULE CLUSTER(3,5)
/PLOT DENDROGRAM .
```

В команде указаны переменные `fac1_1` и `fac2_1` для кластеризации. По умолчанию расстояние между кластерами определяется по среднему расстоянию между объектами (Method baverage), а расстояние между объектами – как квадрат евклидова расстояния (`MEASURE = SEUCLID`). Кроме того, распечатывается протокол (`PRINT SCHEDULE`), в качестве переменных выводятся классификации из 3, 4, 5 кластеров (`CLUSTER(3,5)`) и строится дендрограмма (`PLOT DENDROGRAM`).

Разрез дерева агрегирования (рис. 7.3) вертикальной чертой на четыре части дал два кластера, состоящих из уникальных по своим характеристикам городов Александровск-Сахалинский и Черемхово; кластер из 5 городов (Оха, Елизово, Южно-Сахалинск, Хабаровск, Курильск); еще 14 городов составили последний кластер.

Естественность такой классификации демонстрирует полученное поле рассеяния данных (рис. 7.4).

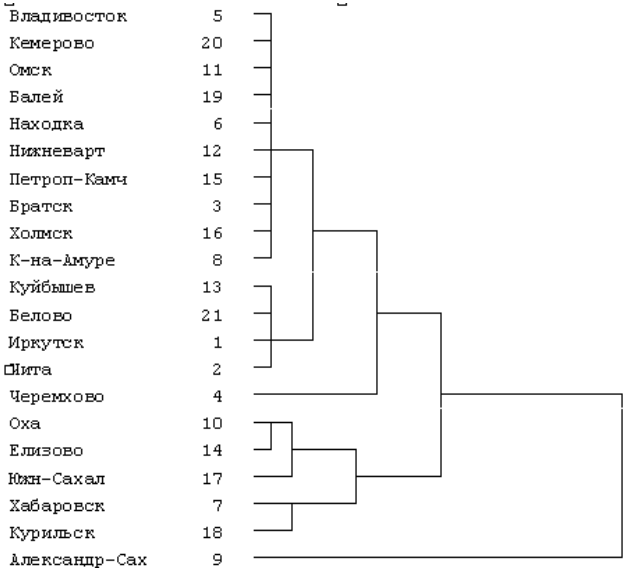


Рис. 7.3. Дендрограмма классификации

Таблица 7.4

Протокол объединения кластеров

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	20	0,0115	0	0	2
2	5	11	0,0175	1	0	3
3	5	19	0,0464	2	0	11
4	6	12	0,051	0	0	8
5	3	16	0,0549	0	0	9
6	13	21	0,0808	0	0	10
7	10	14	0,1082	0	0	14
8	6	15	0,1349	4	0	11
9	3	8	0,1538	5	0	13
10	1	13	0,2818	0	6	12
11	5	6	0,456	3	8	13
12	1	2	0,5768	10	0	16
13	3	5	0,5861	9	11	16
14	10	17	0,613	7	0	17
15	7	18	0,8098	0	0	17
16	1	3	1,5406	12	13	18
17	7	10	2,5726	15	14	19
18	1	4	3,5613	16	0	19
19	1	7	5,2217	18	17	20
20	1	9	14,9146	19	0	0

Процесс объединения подробно показан в протоколе объединения (табл. 7.4). В нем указаны стадии объединения, объединяемые кластеры (после объединения кластер принимает минимальный номер из номеров объединяемых кластеров). Далее следует расстояние между кластерами, номер стадии, на которой кластеры ранее уже участвовали в объединении; следующая стадия, где произойдет объединение с другим кластером.

На практике интерпретация кластеров требует достаточно серьезной работы, изучения разнообразных характеристик объектов для точного описания типов объектов, которые составляют тот или иной класс.

7.2.2. Быстрый кластерный анализ

Процедура иерархического кластерного анализа хороша для малого числа объектов. Ее преимущество в том, что каждый объект можно, образно говоря, пощупать руками. Но эта процедура не годится для огромных социологических данных из-за трудоемкости агломеративного алгоритма и

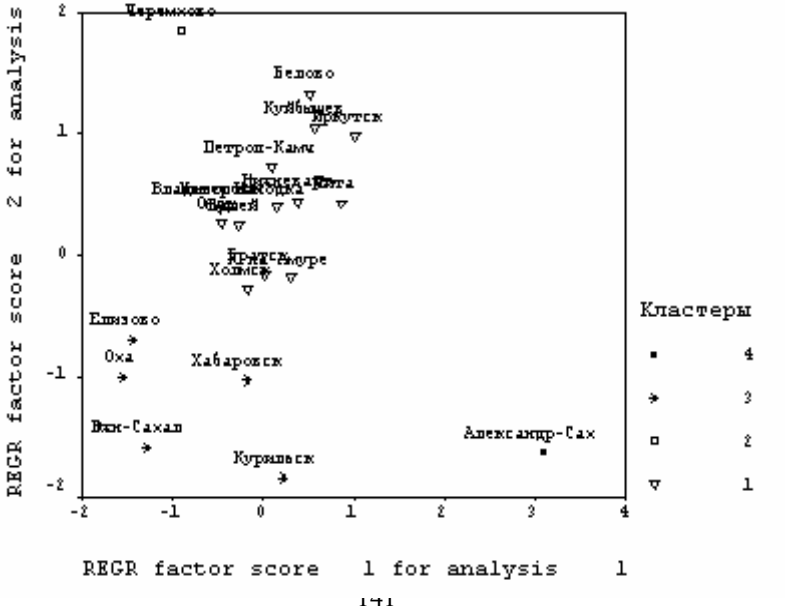


Рис. 7.4. Классификация городов

слишком больших размеров дендрограмм.

Здесь наиболее приемлем быстрый алгоритм, носящий название метода *k-средних*. Он реализуется в пакете командой `QUICK CLUSTER` или командой меню ***k*-means**.

Алгоритм заключается в следующем: выбирается заданное число *k-точек* (объектов из данных), и на первом шаге эти точки рассматривают-

ся как центры кластеров. Каждому кластеру соответствует один центр. Объекты распределяются в кластерах по такому принципу: каждый объект относится к кластеру с ближайшим к этому объекту центром. Таким образом, все объекты распределились по *k-кластерам*.

Затем заново вычисляются центры этих кластеров, которыми с этого момента считаются покоординатные средние кластеров. После этого опять перераспределяются объекты. Вычисление центров и перераспределение объектов происходит до тех пор, пока не стабилизируются центры.

Синтаксис команды:

```
  QUICK CLUSTER W3d1 TO W3D6/CRITERIA CLUSTERS(3)
/MISSING = PAIRWISE /SAVE CLUSTER(SAVCLU)
/PRINT ANOVA.
```

За именем команды располагаются переменные, по которым происходит кластеризация. Параметр `/CRITERIA CLUSTERS` задает в скобках число кластеров. Подкомандой `/SAVE CLUSTER` можно сохранить полученную классификацию в виде переменной, имя которой дается в скобках. Подкоманда `/PRINT ANOVA` позволяет провести по каждой переменной одномерный дисперсионный анализ – сравнение средних в кластерах. Последний имеет лишь описательное значение и позволяет определить переменные, которые не оказывают никакого влияния на классификацию.

Команда использует только евклидово расстояние. При этом часть переменных может иметь неопределенные значения, расстояния до центров определяются по определенным значениям. Для использования такой возможности следует употребить подкоманду `/MISSING = PAIRWISE.`

Часто переменные имеют разный диапазон изменений, так как измерены они в различных шкалах или просто из-за того, что характеризуют разные свойства объектов (например, рост и вес, килограммы и граммы). В этих условиях основное влияние на кластеризацию окажут переменные, имеющие большую дисперсию. Поэтому перед кластеризацией полезно стандартизовать переменные. К сожалению, в «быстром» кластерном анализе средства стандартизации не предусмотрены непосредственно, как в процедуре иерархического кластерного анализа.

Для этого можно использовать команду `DESCRIPTIVE`. Напомним, что подкоманда `/SAVE` в ней позволяет автоматически сохранить стандартизованные переменные. Кроме того, хорошие средства стандартизующих преобразований шкал дает команда `RANK`.

В выдаче распечатываются центры кластеров (средние значения переменных кластеризации для каждого кластера), получаемые на каждой итерации алгоритма. Однако для нас полезна лишь часть выдачи, помеченная текстом **Final centres**.

Интерпретация кластеров осуществляется на основе сравнения средних значений, выдаваемых процедурой, а также исследования сохраненной переменной средствами статистического пакета.

Пример использования QUICK CLUSTER. Для иллюстрации построим классификацию по предварительно отобранным данным городских семей по жилплощади и душевому доходу. Такая классификация может грубо, но наглядно показать различие семей по благосостоянию.

В данных, полученных из обследования RLMS 1998 г., имеются переменные: c5 – жилплощадь, приходящаяся на семью, memb – число членов семьи, df14 – суммарные денежные доходы семьи.

В ранее проведенном анализе выяснилось, что не только доходы имеют близкое к логарифмически нормальному распределение, но и жилплощадь. Для того чтобы кластерный анализ не конструировал кластеры из «выбросов» больших доходов и жилплощади, мы работаем со стандартизованными переменными «логарифм душевых доходов» и «логарифм жилплощади», приходящейся на члена семьи.

```
*вычисление логарифма жилплощади на члена семьи.
COMPUTE lns = LN(dc5/memb).
*вычисление логарифма душевого дохода.
COMPUTE lincome = LN(df14/memb).
*стандартизация переменных.
DESCRIPTIVES VARIABLES = lincome lns/SAVE .
QUICK CLUSTER zlincome zlns /MISSING = PAIRWISE
/CRITERIA = CLUSTER(3) /SAVE CLUSTER /PRINT ANOVA.
```

На основании табл. 7.5 получается следующая интерпретация полученных кластеров:

Кластер 1 – зажиточные семьи, имеющие относительно большой доход и жилплощадь.

Кластер 2 – семьи, проживающие в квартирах с небольшой площадью, но имеющие относительно высокий доход.

Кластер 3 – семьи, имеющие низкий доход и ограниченные в жилплощади.

Кластер 4 – семьи, имеющие несколько больший доход, чем в среднем, но ограниченные в жилплощади.

Таблица 7.5

Центры кластеров (Final Cluster Centers)

	Cluster			
	1	2	3	4
Zscore (LINCOME)	1,26	0,52	–1,08	–0,40
Zscore (LNS)	1,35	–0,56	–0,86	0,58

Таблица 7.6

Дисперсионный анализ в методе *k-средних*
(ANOVA, имеет только описательное значение)

	Cluster		Error		F	Sig.
	Mean Square	Df	Mean Square	df		
ZLINCOME Zscore (LINCOME)	513,006	3	.370	2440	1384,7	0,000
ZLNS Zscore (LNS)	530,153	3	.363	2491	1461,6	0,000

Дисперсионный анализ (табл. 7.6) показал, что по обеим переменным различие кластеров существенно. Но о статистической значимости переменных говорить бессмысленно, поскольку гипотеза дисперсионного анализа, по сути, – независимость групп и «зависимой» переменной, а в данном случае группы сформированы на основе значений «независимых» переменных.

Зато уж если бы различие средних по какой-либо переменной оказалось формально незначимым, переменную почти наверняка можно было бы исключить из анализа.

Полезно рассмотреть график рассеяния данных по кластерам (рис. 7.5). В нашем случае, пожалуй, не стоит говорить о выделении «сгущений» то-

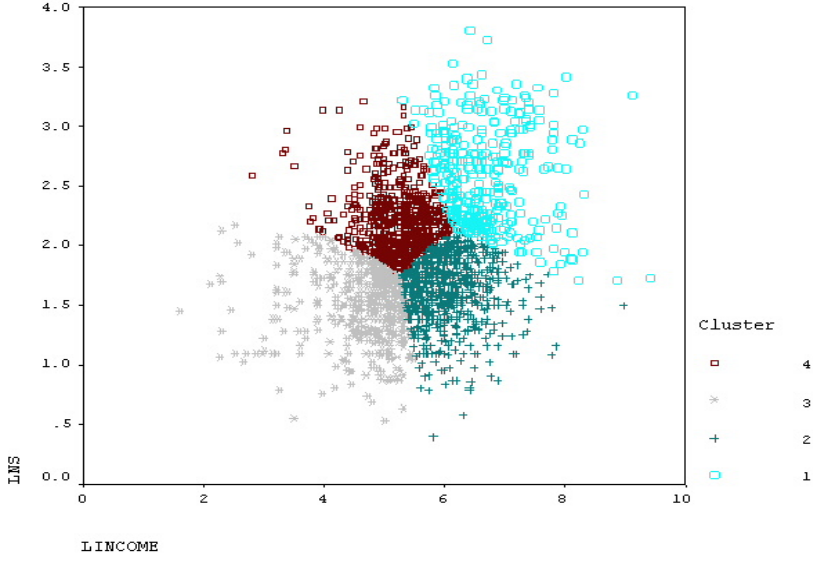


Рис. 7.5. Классификация семей по душевому доходу Lincome и жилплощади на человека LNS (в логарифмических шкалах)

чек в признаковом пространстве, скорее, программа нашла разумные границы для описания типов, выделив их в непрерывном двумерном распределении.

Имеется масса возможностей изучить и сравнить полученные классы, используя сохраненную в виде переменной классификацию, например, посмотреть какая доля семей в классах имеет автомобили, средние размеры личного подсобного хозяйства и т. п.

7.3. Многомерное шкалирование

Задача многомерного шкалирования состоит в построении переменных на основе имеющихся расстояний между объектами. В частности, если нам даны расстояния между городами, программа многомерного шкалирования должна восстановить систему координат (с точностью до поворота и единицы длины) и приписать координаты каждому городу, так чтобы зрительно карта и изображение городов в этой системе координат совпали. Близость может определяться не только расстоянием в километрах, но и другими показателями, такими как размеры миграционных потоков между городами, интенсивность телефонных звонков, а также расстояниями в многомерном признаковом пространстве. В последнем случае задача построения такой системы координат близка к задаче, решаемой факторным анализом – сжатию данных, описанию их небольшим числом переменных. Нередко требуется также наглядное представление свойств объектов. В этом случае полезно придать координаты переменным, расположить переменные в геометрическом пространстве. С технической точки зрения это всего лишь транспонирование матрицы данных. Для определенности мы будем говорить о создании геометрического пространства для объектов, специально оговаривая случаи анализа множества свойств. В социальных исследованиях методом многомерного шкалирования создают зрительный образ «социального пространства» объектов наблюдения или свойств. Для такого образа наиболее приемлемо создание двумерного пространства.

Основная идея метода состоит в приписывании каждому объекту значений координат, так, чтобы матрица евклидовых расстояний между объектами в этих координатах, умноженная на константу, оказалась близка к матрице расстояний между объектами, определенной из каких-либо соображений ранее.

Метод весьма трудоемкий и рассчитан на анализ данных, имеющих небольшое число объектов.

7.3.1. Евклидово пространство

Пусть мы определили r шкал X^1**, …,** X^r. Расстояние между парой объектов i и j определяется формулой $d_{ij} = \sqrt{\sum_{k=1}^{r}(X_i^k - X_j^k)^2}$.

Для однозначности задания шкал предполагается, что $\sum_i^n X_i^k = 0$ и $\sum_i^n \sum_k^r (X_i^k)^2 = nr$. Кроме того, по аналогии с методом главных компонент, первая шкала выбирается с наибольшей дисперсией, вторая имеет вторую наибольшую дисперсию и т. д.

7.3.2. Идея многомерного шкалирования

Первая в этом направлении работа Торгерсона (*Torgerson*, 1952, [11]) была посвящена метрическому многомерному шкалированию. Модель этого метода имеет вид $L\{S\} = D^2 + E$, где $L\{S\}$ – линейное преобразование исходной матрицы расстояний, D^2 – матрица расстояний, полученная на основе созданных шкал, E – матрица отклонений модели от исходных данных. Линейное преобразование дает матрицу преобразованных расстояний $T = L\{S\}$. Целью многомерного метрического шкалирования является поиск оптимальных шкал и линейного преобразования матрицы исходных расстояний, минимизирующих ошибку E.

Шепард и Краскэл (*Shepard*, 1962, *Kruscal*, 1964, см. ссылку в [11]) совершили существенный прорыв, разработав метод неметрического шкалирования. Суть этого метода состоит в нелинейном преобразовании расстояний. Модель неметрического шкалирования имеет вид $M\{S\} = D^2 + E$, где $M\{S\}$ – монотонное преобразование исходной матрицы расстояний. Этот метод имеет больше шансов получить действительно геометрическое пространство, метрическое шкалирование. Монотонное преобразование дает матрицу преобразованных расстояний $T = L\{S\}$.

7.3.3. Качество подгонки модели

Для измерения качества подгонки модели Такейном (Takane, 1977) был предложен показатель *S-stress* = $\left(\frac{\|E\|}{\|T\|}\right)^{1/2}$, где норма матрицы $||\ ||$ означает сумму квадратов элементов матрицы. Слово *stress* в английском языке имеет множество значений, одно из этих значений – нагрузка. Этот показатель изменяется от 0 до 1. Равенство его нулю означает точную подгонку модели, единице – полную ее бессмысленность.

Кроме того, оценить качество модели можно с помощью показателя *stress index* Краскэла, который согласно документации SPSS [11] получается с использованием матрицы не квадратов расстояний, а расстояний. Заметим, что алгоритм оптимизирует *S-stress,* не *stress index*.

Еще один показатель качества модели, *RSQ,* представляет собой квадрат коэффициента корреляции между матрицами T и E. Таким образом, так же

как в регрессионном анализе, *RSQ* может быть интерпретирован как доля дисперсии преобразованных расстояний T, объясненная матрицей расстояний D.

7.3.4. Вызов процедуры многомерного шкалирования

Вызов процедуры в диалоговом режиме осуществляется командой меню **Statistics\Scale\Multidimentional scaling**. В результате «приклеивания» команды из меню в окно синтаксиса многомерного шкалирования обычно получается целая серия команд, связанных с вычислением расстояний, сохраняемых во временных файлах, работой с несколькими матрицами одновременно и уничтожением матриц данных. Команда меню устроена достаточно удобно, но, к сожалению, в ней предусмотрена возможность сохранения полученных шкал в виде переменных исходного файла данных. Это можно сделать только в синтаксисе, дополнив сгенерированную команду `ALSCALE` подкомандой `/OUTFILE` с указанием имени файла (например, `/OUTFILE = "scale.save"`). С помощью команды **Merge files** полученные переменные можно подключить к исходному файлу данных.

По умолчанию в процедуре проводится неметрическое шкалирование, кнопкой **Model** можно переключиться на метрическое шкалирование.

7.3.5. Исходная матрица расстояний

По умолчанию в процедуре предполагается, что исходная матрица расстояний вводится в файле SPSS. Но подготовленная матрица расстояний у исследователя бывает весьма редко. Поэтому чаще используется возможность вычисления расстояний на основе имеющихся данных, которая реализуется в диалоговом окне команды в разделе **Distances** включения пункта **Create distances from data**. Здесь предусмотрен такой же широкий набор мер близости и расстояний, как и в иерархическом кластерном анализе. Их можно выбрать, воспользовавшись кнопкой **Measures** в том же разделе **Distances**, при этом можно определить, что визуализируется – матрица расстояний между объектами или переменными.

7.3.6. Пример построения шкал

В качестве примера исследуем данные по средней обеспеченности семей дорогостоящими предметами быта – электроникой, средствами транспорта и дачами (всего 9 предметов) в 38 территориальных общностях (данные RLMS, 1996 г.). В результате применения процедуры шкалирования территориальные общности должны расположиться в двумерном геометрическом пространстве, построенном исходя из расстояний по 9 переменным.

Для этого получим файл, в котором объектами будут территориальные общности, а переменными – обеспеченность семей этими предметами. Значения этих переменных – доли семей, обладающих этими предметами. Ис-

ходными данными здесь являются ответы на вопрос «Имеете ли Вы холодильник?», «Имеете ли Вы стиральную машину?» и т. д. (1 – да, 2 – нет, 9 – нет ответа) в файле анкет семьи.

Этот файл агрегируем по территориальным общностям (переменная PSU), сохранив доли семей, имеющих соответствующие предметы, в файле **property.sav**:

```
  AGGREGATE /OUTFILE = 'N:\USR\RLMS\property.SAV'
/BREAK = psu /CC9.1A 'холодильник' CC9.3A 'стиральная
машина' CC9.4A 'черно-белый телевизор' CC9.5A 'цветной
телевизор' CC9.6A 'видеомагнитофон или видеоплейер'
CC9.6.1A 'фен' CC9.7A 'легковой автомобиль' CC9.10A
'садовый домик' CC9.11A 'дача или другой дом'  = PLT
(2, CC9.1A CC9.3A CC9.4A CC9.5A CC9.6A CC9.6.1A CC9.7A
CC9.10A CC9.11A).
```

Полученный файл используется для запуска процедуры многомерного шкалирования:

```
  GET FILE 'N:\USR\RLMS\property.SAV'.
  ERASE FILE = 'J:\TEMP\spssalsc.tmp'.
  PROXIMITIES cc9.1a cc9.3a cc9.4a cc9.5a cc9.6a
cc9.6.1a cc9.7a cc9.10a cc9.11a /PRINT NONE
/MATRIX OUT('J:\TEMP\spssalsc.tmp')
  /MEASURE = EUCLID /STANDARDIZE = NONE /VIEW = CASE.
  SPLIT FILE OFF.
  ALSCAL /MATRIX =  IN('J:\TEMP\spssalsc.tmp')
/LEVEL = ORDINAL /CONDITION = MATRIX /MODEL = EUCLID
/CRITERIA = CONVERGE(.001) STRESSMIN(.005) ITER(30)
CUTOFF(0) DIMENS(2,2) /PLOT = DEFAULT ALL
/outfile = "scale.save" /PRINT = HEADER .
  ERASE FILE = 'J:\TEMP\spssalsc.tmp'.
```

Далее, переменные Dim1 и Dim2, сохраненные подкомандой /outfile = "scale.save" с помощью команды меню **Merge files**, присоединяются к нашему файлу **property.sav**.

Проблема выяснить, как же интерпретируются наши шкалы. Для интерпретации можно изучить их связь с имеющимися данными, в частности с исходными переменными, по которым строилась матрица расстояний.

В нашем примере таблица ранговых корреляций с этими переменными свидетельствует о том, что первое измерение (Dim1) характеризует уровень благосостояния жителей территориальных образований в целом, второе измерение связано с приверженностью их садоводству.

Таблица 7.7

Коэффициенты ранговой корреляции Спирмена построенных шкал с обеспеченностью предметами быта

		CC9.1A холодильник	CC9.3A стиральная машина	CC9.4A черно-белый телевизор	CC9.5A цветной телевизор	CC9.6A видеомагнитофон	CC9.6.1A фен	CC9.7A легковой автомобиль	CC9.10A садовый домик	CC9.11A дача или другой дом
DIM1		0,844	0,265	–0,820	0,950	0,773	0,929	0,426	0,426	0,659
	Sig.	0,000	0,108	0,000	0,000	0,000	0,000	0,008	0,008	0,000
DIM2		–0,112	–0,156	–0,145	0,113	0,402	0,240	0,262	–0,687	0,232
	Sig.	0,502	0,350	0,385	0,501	0,012	0,148	0,112	0,000	0,161

Наглядную картину дает непосредственное размещения объектов (у нас – территориальных общностей) на поле рассеяния в построенном геометрическом пространстве (рис. 7.6). На этом графике видим, что шкала `Dim1` имеет больший разброс, чем шкала `Dim2`, а значит, объясняет боль-

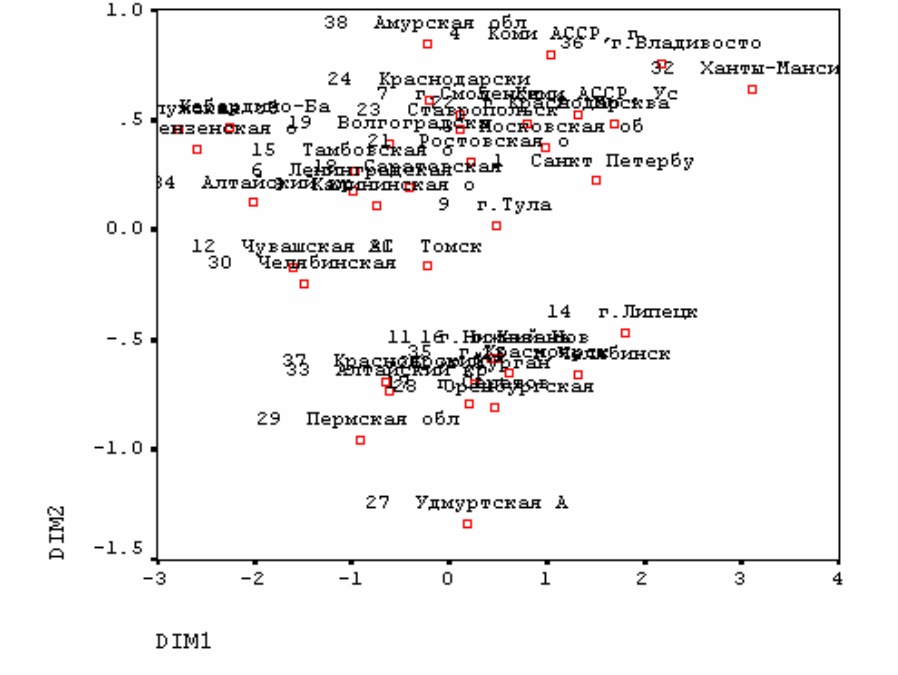


Рис.7.6 Представление объектов в сконструированном геометрическом пространстве

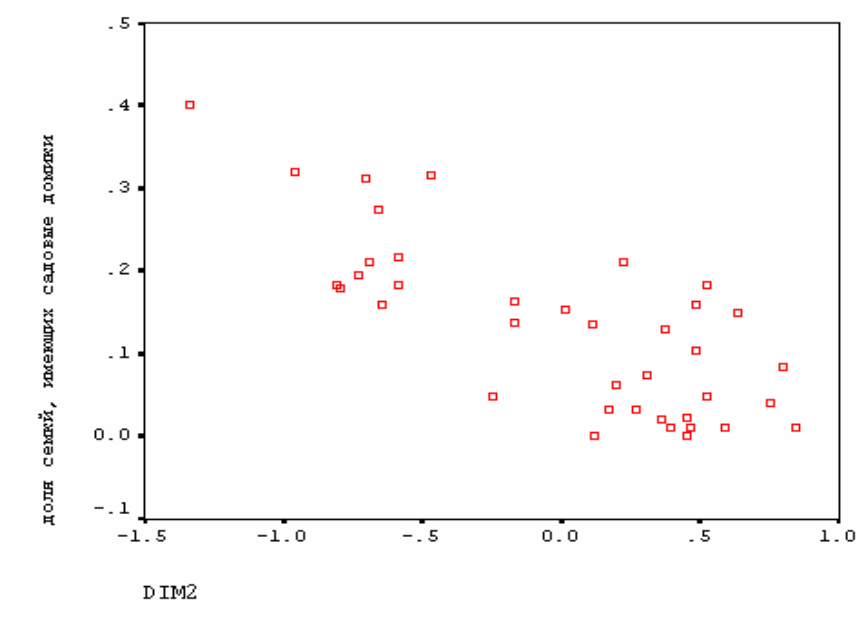


Рис. 7.7. Поле рассеяния второй шкалы, порожденной процедурой многомерного шкалирования, и доли семей, имеющих садовые домики

шую часть разброса расстояний объектов. Зримо подтверждается интерпретация первой шкалы: по разным полюсам `Dim1` стоят Ханты-Мансийский автономный округ – весьма богатый регион и Пензенская область, Кабардино-Балкария – беднейшие регионы России.

Поскольку по поводу развитости садоводства мы не имеем общедоступной информации, для проверки интерпретации второй шкалы полезно рассмотреть диаграмму рассеяния `Dim2` и доли семей, имеющих садовые домики (рис. 7.7). На этом рисунке ясно видно, что указанная выше интерпретация небезосновательна.

ЛИТЕРАТУРА

1. GREEN H.WILLIAM. Econometric Analysis. Upper Saddle River, New Jersey, 1997.

2. Handbook of Statistical Modeling for Social and Behavioral Sciences. New York and London: Plenium press, 1995.

3. АЙВАЗЯН С. А., МХИТОРЯН В. С. Прикладная статистика и основы эконометрики. М.: Издательское объединение «Юнити», 1998.

4. ТОЛСТОВА Ю. Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000.

5. SPSS для Windows. Руководство пользователя SPSS. М.: Статистические системы и сервис. 1995. Кн. 1.

6. SPSS BASE 8.0. Руководство пользователя SPSS. М.: СПСС РУСЬ. 1998.

7. SPSS SPSS BASE 8.0. Руководство по применению SPSS. М.: СПСС РУСЬ. 1998.

8. BASE 7.5. Syntax Reference Guide / SPSS. Chicago, 1997.

9. SPSS. Regression Models 9.0. Chicago, 1999.

10. SPSS. Exact tests 6.1 for windows. Chicago, 1995.

11. SPSS. Professional statistics. Chicago, 1994

12. РОСТОВЦЕВ П. С., КОСТИН В. С., ОЛЕХ А. Л. Множественные сравнения в таблицах для неальтернативных вопросов // Анализ и моделирование экономических процессов переходного периода в России. Новосибирск: ИЭиОПП СО РАН, 1999. Вып. 4. С. 148 – 164.

13. Российский мониторинг экономического положения и здоровья населения // Мир России. 1999. № 3.

Приложение 1

Анкета опроса общественного мнения

Номер анкеты .....
Город .....
Регион .....

I. Какая из предложенных точек зрения на роль иностранной помощи в развитии экономики восточных районов совпадает с Вашим мнением?

1. Нужно рассчитывать только на собственные силы и средства.
2. Иностранную помощь следует использовать, но в определенных пределах.
3. Для развития экономики восточных районов страны необходима широкомасштабная иностранная помощь.
4. Затрудняюсь сказать, не знаю.

II. Какими будут изменения в уровне жизни, труде, занятости и безопасности населения в результате организации свободных экономических зон?

1. В целом положительные.
2. Скорее положительные, чем отрицательные.
3. Скорее отрицательные, чем положительные.
4. В целом отрицательные.
5. Неоднозначные.
6. Затрудняюсь сказать, не знаю.

III. Как Вы считаете, что мешает подписать мирный договор между СССР и Японией?

1. Нет настоятельной необходимости, отношения нормальные.
2. Традиционное недоверие друг другу в результате войн.
3. Слабая экономическая заинтересованность Японии.
4. Разные политические симпатии СССР и Японии.
5. Нежелание Японии признать послевоенные границы с СССР.
6. Нежелание СССР рассматривать вопрос о спорных островах.
7. Другое.
8. Затрудняюсь сказать, не знаю.

IV. Считаете ли Вы возможным удовлетворение территориальных требований Японии?

1. Эти острова надо отдать.
2. Надо отдать часть островов.

3. Отдавать острова не надо.

4. Затрудняюсь сказать, не знаю.

V. Ответы тех, кто за передачу островов:

1. Эти острова длительное время принадлежали Японии.

2. Спорные острова не входят в Курильскую гряду и права СССР на них не распространяются.

3. Острова не представляют собой особой ценности.

4. Для обустройства и выгодного использования островов все равно никогда не хватит сил и средств.

5. Передача островов приведет к налаживанию хороших, добрососедских отношений с Японией.

6. Передача островов позволяет надеяться на решение многих экономических проблем.

7. Другое.

8. Затрудняюсь сказать, не знаю.

VI. Ответы тех, кто против передачи островов:

1. Эти острова были открыты, заселены русскими и принадлежали сначала России.

2. После войны мы лишь вернули эти острова, и наши права на них подтверждены международными юридическими документами.

3. Это уникальные места, и их потеря для страны невосполнима.

4. Не хватило сил и средств у нас, чтобы быть хорошими хозяевами – хватит у наших детей.

5. Передача островов не приведет с необходимостью к налаживанию хороших отношений с Японией.

6. Передача островов не решит наших проблем и не принесет пользы, как это показывает прошлый исторический опыт.

7. Другое.

8. Затрудняюсь сказать, не знаю.

VII. Какие варианты решения по вопросу о спорных островах Вам кажутся разумными и приемлемыми для нашей страны?

1. Сохранение принадлежности островов в СССР без изменений.

2. Создание совместной экономической зоны.

3. Упрощение или отмена визовых процедур для посещения островов японцами.

4. Ликвидация любого военного присутствия на островах.

5. Передача островов с компенсацией.

6. Продажа островов.

7. Поэтапная передача части островов в перспективе.

8. Поэтапная передача всех спорных островов в перспективе.

9. Безотлагательная передача всех спорных островов.

10. Передача вопроса о принадлежности островов на рассмотрение международного суда.

11. Другое.

12. Затрудняюсь сказать, не знаю.

VIII. Ваш пол

1. Муж. 2. Жен.

IX. Возраст .....

X. Образование

1. Высшее.
2. Незаконченное высшее.
3. Среднее специальное.
4. ПТУ, ФЗУ.
5. 10 – 11 кл.
6. 7 – 9 кл.
7. 4 – 6 кл.
8. 3 кл. и менее.
9. Не учился.

XI. Состояние в браке:

1. Женат (замужем).
2. Вдовец/вдова.
3. Разведен/а.
4. Никогда не состоял/а в браке.

XII. Кем работаете?

1. Руководитель или его заместитель.
2. Рядовой специалист с высшим и средне-специальным образованием.
3. Служащий без специального образования.
4. Рабочий высокой и средней квалификации.
5. Неквалифицированный рабочий.
6. Студент, учащийся.
7. Не работаю.

XIII. В какой сфере народного хозяйства Вы работаете?

1. Управление.
2. Юстиция и охрана общественного порядка.

3. Армия, флот.
4. Наука и высшая школа.
5. Промышленность, строительство, транспорт, связь.
6. Торговля, общественное питание, материально-техническое снабжение.
7. Коммунальное хозяйство и бытовое обслуживание.
8. Здравоохранение, просвещение, культура, искусство, печать.
9. Финансы, статистика, кредит.
10. Другое.
11. Не работаю, учусь.

XIV. Среднемесячный душевой доход в семье .....

Сколько лет проживали:
XV. В Западной Сибири .....
XVI. В Восточной Сибири .....
XVII. На Дальнем Востоке .....

Приложение 2

Переменные файла обследования общественного мнения

Неальтернативные вопросы 3, 5, 6, 7 записаны в виде списка подсказок. В дополнение к анкете введены переменные *G* – город, *S* – докодировка отрасли, *Wes* – весовая переменная, *TP* – тип города, *R* – регион.

N Номер анкеты

V1 Точка зр. на иностр. помощь
1 не нужна
2 огранич.
3 нужна для вост. р-нов
4 не знаю

V2 Последствия ЗСП
1 положит.
2 > положит.
3 >отриц.
4 отриц.
5 неоднозн.
6 не знаю

V3S1 – v3s8
1 нет необх.
2 недоверие
3 незаинт. Яп.
4 разн. полит.
5 непризн. гр.
6 нежел. СССР
7 другое
8 не знаю

V4 Возмож. удовл. территор. треб. Японии
1 отдать
2 отд. часть
3 не надо

4 не знаю

V5S1 – v5s8
1 были Япон.
2 не Куриль.
3 не ценные
4 нет сил
5 отношения
6 эк. проб.
7 другое
8 не знаю

V6S1
1 были Рос.
2 юр. док.
3 уникальн.
4 детям
5 не помож.
6 не решит
7 другое
8 не знаю

V7S1 – V7S7
1 без изм.
2 эк. зоны
3 визы
4 демилит.
5 с комп.
6 продажа
7 –> части
8 –> всех
9 –> безог.
10 м/н суд
11 другое
12 не знаю

V8 Пол
1 муж.
2 жен.

V9 Возраст

V10 Образование
1 Высшее
2 н/высш.
3 ср. спец.
4 ПТУ, ФЗУ
5 10 – 11 кл.
6 7 – 9 кл.
7 4 – 6 кл.
8 менее
9 нет

V11 Состояние в браке
1 женат
2 вдовец
3 разведен
4 не был

V12 Кем работаете
1 рук/зам.
2 спец.
3 служ.
4 кв. раб.
5 некв. раб.
6 учащийся
7 не раб.

V13 В какой сфере нар. хоз.
1 управл.
2 юстиция
3 армия
4 наука
5 пром.
6 торг.
7 БО
8 непроизв.
9 финансы
10 другое
11 не раб.

V14 ср. мес. душевой доход в семье.

V15 Сколько лет жил/а в Западной Сибири?

V16 Сколько лет жил/а в Восточной Сибири?

V17 Сколько лет жил/а на Дальнем Востоке?

G номер города
1 Иркутск
2 Чита
3 Братск
4 Черемхово
5 Владивосток
6 Находка
7 Хабаровск
8 К-на-Амуре
9 Александр-Сах.
10 Оха
11 Омск
12 Нижневарт.
13 Куйбышев
14 Елизово
15 Петроп.-Камч.
16 Холмск
17 Южн.-Сахал.
18 Курильск
19 Балей
20 Кемерово
21 Белово

S отрасль н/х.
1 эн. пром.
2 хим. пром.
3 маш. стр.
4 лес. пром.
5 пром. стр.
6 лег. пром.
7 стр-во
8 трансп., связ.
9 торг., пит.
10 МТС наб.
11 здрав.
12 ЖКХ, БО
13 нар. обр.
14 наука
15 культ.
16 управл.
17 армия, фл.
18 уч-ся
19 пенсион.
20 другое

WES

TP тип поселен.
1 растущие
2 стабильные
3 крупные
4 гиганты

R регион
1 Дальн. В.
2 Вост. Сиб.
3 Зап. Сиб.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ 3
Глава 1. Информация, обрабатываемая статистическим пакетом 5
1.1. Анкетные данные 5
1.2. Типы переменных 6
1.2.1. Типы кодирования переменных 6
1.2.2. Тип шкалы измерения переменной 7
1.2.3. Неколичественные шкалы 7
1.2.4. Количественные шкалы 7
1.2.5. Неальтернативные признаки 8
1.3. Имена и метки переменных 9
1.4. Коды неопределенных значений 10
Глава 2. Общее описание статистического пакета для социологических исследований и подготовка данных 10
2.1. Структура пакета 10
2.2. Схема организации данных, окна SPSS 11
2.3. Управление работой пакета 12
2.3.1. Основные команды меню SPSS верхнего уровня 12
2.3.2. Статусная строка 13
2.3.3. Ввод данных с экрана 14
2.4. Режим диалога и командный режим 14
2.4.1. Командный режим работы с пакетом. Основные правила написания команд на языке пакета 15
2.4.2. Порядок выполнения команд 16
2.4.3. Команды вызова GET и сохранения данных SAVE 16
2.4.4. Основные команды описания данных 17
2.5. Основные команды преобразования данных 19
2.5.1. Команды COMPUTE и IF 19
2.5.2. Команда RECODE 24
2.5.3. Команда COUNT 27
2.5.4. Условное выполнение команд 28
2.5.5. Команда RANK 29
2.5.6. Отбор подмножеств наблюдений 29
2.5.7. Команда SPLIT FILE 31
2.5.8. Взвешивание выборки WEIGHT 32
2.6. Операции с файлами 35
2.6.1. Агрегирование данных (команда AGGREGATE) 35
2.6.2. Объединение файлов (MERGE FILES) 38

Глава 3. Процедуры получения описательных статистик и таблиц сопряженности 40
3.1. Команды получения распределений и описательных статистик 40
3.1.1. FREQUENCIES – получение одномерных распределений переменных 40
3.1.2. DESCRIPTIVES – описательные статистики 47
3.1.3. EXPLORE – исследование распределений и сравнение групп объектов 48
3.2 Анализ связи между неколичественными переменными 49
3.2.1. CROSSTABS – таблицы сопряженности 49
3.3. Сложные табличные отчеты. Таблицы для неальтернативных вопросов 64
3.3.1. Работа с командой General Tables 65
3.3.2. Типичные примеры использования Multiple Response Tables 68
3.4. Множественные сравнения в таблицах для неальтернативных вопросов. Программа Typology Tables 69
3.4.1. *Z*-статистика значимости отклонения частот 69
3.4.2. *Z*-статистика отклонения средних 69
3.4.3. Как выяснить надежность результата? 70
3.4.4. Критические значения *Z*-статистики при множественных сравнениях 70
3.4.5. Статистические эксперименты 70
3.4.6. Работа с программой Typology Tables 71
3.4.7. Примеры использования программы Typology Tables 71
Глава 4. Сравнение средних, корреляции 76
4.1. Compare Means – простые параметрические методы сравнения средних 76
4.1.1. Одновыборочный тест (One sample t-test) 77
4.1.2. Двухвыборочный *t*-тест (independent sample t-test) 80
4.1.3. Двухвыборочный *t*-тест для связанных выборок (Paired sample t-test) 82
4.1.4. Команда MEANS – сравнение характеристик числовой переменной по группам 83
4.1.5. Одномерный дисперсионный анализ (ONEWAY) 85
4.1.6. Множественные сравнения 85
4.2. CORRELATIONS – корреляции 90
4.2.1. Парные корреляции 91
4.2.2. Частные корреляции 92
Глава 5. Непараметрические тесты. Команда NONPARAMETRIC TESTS 94
5.1. Одновыборочные тесты 94
5.1.1. Тест хи-квадрат 94

5.1.2. Тест, основанный на биномиальном распределении 96
5.1.3. Тест Колмогорова – Смирнова 98
5.2. Тесты сравнения нескольких выборок 99
5.2.1. Двухвыборочный тест Колмогорова – Смирнова 100
5.2.2. Тест медиан 101
5.3. Тесты для ранговых переменных 102
5.3.1. Двухвыборочный тест Манна – Уитни (Mann – Witney) 102
5.3.2. Одномерный дисперсионный анализ Краскэла – Уоллиса (Kruskal – Wallis) 104
5.4. Тесты для связанных выборок (Related samples) 104
5.4.1. Двухвыборочный критерий знаков (Sign) 105
5.4.2. Двухвыборочный знаково-ранговый критерий Вилкоксона (Wilcoxon) 105
5.4.3. Критерий Фридмана (Friedman) 106
Глава 6. Регрессионный анализ 107
6.1. Классическая линейная модель регрессионного анализа 107
6.1.1. Существует ли линейная регрессионная зависимость? 109
6.1.2. Коэффициенты детерминации и множественной корреляции 109
6.1.3. Оценка влияния независимой переменной 110
6.1.4. Пошаговая процедура построения модели 112
6.1.5. Переменные, порождаемые регрессионным уравнением 113
6.1.6. Взвешенная регрессия 113
6.1.7. Команда построения линейной модели регрессии 114
6.1.8. Пример построения модели 115
6.1.9. Можно ли в регрессии использовать неколичественные переменные? 117
6.1.10. Взаимодействие переменных 120
6.2. Логистическая регрессия 120
6.2.1. Отношение шансов и логит 121
6.2.2. Решение уравнения с использованием логита 121
6.2.3. Неколичественные данные 122
6.2.4. Взаимодействие переменных 123
6.2.5. Пример логистической регрессии и статистики 123
6.2.6. Качество логистической регрессии 125
6.2.7. Вероятность правильного предсказания 125
6.2.8. Коэффициенты логистической регрессии 126
6.2.9. О статистике Вальда 128
6.2.10. Сохранение переменных 128
Глава 7. Исследование структуры данных 128
7.1. Факторный анализ 129
7.1.1. Метод главных компонент 130
7.1.2. Интерпретация факторов 131

7.1.3. Оценка факторов 132
7.1.4. Статистические гипотезы в факторном анализе 132
7.1.5. Задание факторного анализа 133
7.2. Кластерный анализ 136
7.2.1. Иерархический кластерный анализ 136
7.2.2. Быстрый кластерный анализ 141
7.3. Многомерное шкалирование 145
7.3.1. Евклидово пространство 145
7.3.2. Идея многомерного шкалирования 146
7.3.3. Качество подгонки модели 146
7.3.4. Вызов процедуры многомерного шкалирования 147
7.3.5. Исходная матрица расстояний 147
7.3.6. Пример построения шкал 147
ЛИТЕРАТУРА 151
Приложение 1. Анкета опроса общественного мнения 152
Приложение 2. Переменные файла обследования общественного мнения 155