

Введение

Практика показывает, что в настоящее время многие российские компании, занимающиеся проведением маркетинговых исследований, а также отделы маркетинга промышленных и торговых организаций часто используют для анализа получаемых полевых данных весьма ограниченный набор аналитических инструментов, иногда даже вовсе без применения статистики. Вместе с тем именно статистический анализ позволяет вскрыть такие закономерности и внутренние связи в данных, которые невозможно выявить другими средствами. Подтверждение гипотез о наличии связи между переменными, оценка характера данных связей, оценка влияния частных параметров продукта на общее впечатление от него потребителей, сегментирование потребителей, прогнозирование изменений рыночной конъюнктуры — вот лишь некоторые задачи, с успехом решаемые с применением статистических методов анализа. На новый уровень выводит статистические методы применение специализированного программного обеспечения для анализа. Наиболее популярным в настоящее время является статистический программный комплекс SPSS.

Предлагаемое пособие имеет своей целью в доступной для понимания форме систематизировать суть основных методов проведения статистического анализа данных при помощи программного пакета SPSS версии 11-12, используемого в практике проведения маркетинговых исследований. Пособие рассчитано на аудиторию, уже имеющую определенные знания в области маркетинга, — практикующих маркетологов и аналитиков. Здесь не разъясняется суть маркетинга и роль маркетинговых исследований, а дается мощный инструментальный аппарат анализа, который можно применять на практике для повышения эффективности деятельности различных организаций. Автор имеет значительный опыт аналитической работы в данной сфере и надеется, что настоящее пособие поможет всем желающим повысить качественный уровень своей собственной работы при анализе полевых данных и написании аналитических отчетов.

Практически все книги, посвященные рассматриваемой теме, представляют собой объемные произведения, содержащие массу ненужной практикам статистической теории и/или описание редко используемых в практике маркетинговых исследований статистических методик. В данном пособии содержатся только практические сведения, причем изложение ведется последовательно, шаг за шагом: от подготовки матрицы исходных данных до применения к ней различных статистических методов. Здесь вы не найдете ни капли «воды»: только та информация, которая реально нужна для того, чтобы немедленно приступить к анализу и наиболее быстро и эффективно его провести. Вместе с тем необходимо отметить, что данное пособие не является исчерпывающим руководством по работе с SPSS. В нем содержится только та информация, которая реально поможет на практике осуществить наиболее часто применяемые методы статистического анализа. Изложение материала снабжено подробными пошаговыми иллюстрациями и конкретными примерами, облегчающими его восприятие.

Для понимания сути описываемых в настоящем пособии статистических методик необходимо прежде всего определить роль и место компьютеризованного статистического анализа в системе маркетинговых исследований. На рис. В.1 представлена принципиальная схема проведения полевого маркетингового исследования.

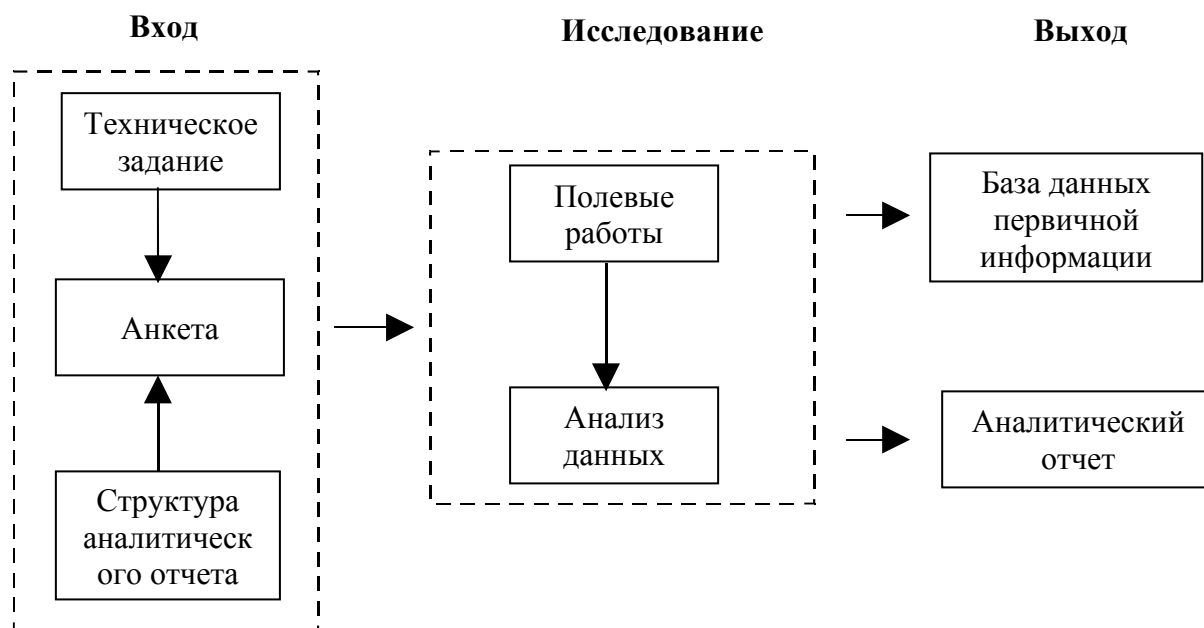


Рис. В.1. Принципиальная схема проведения полевого маркетингового исследования

В целом весь процесс проведения полевого маркетингового исследования можно условно подразделить на два этапа.

■ Подготовка материалов, необходимых для проведения исследования:

- подготовка технического задания (ТЗ);
- подготовка структуры аналитического отчета;
- формирование анкеты (на основании ТЗ и структуры отчета).

■ Проведение исследования:

- полевые работы (сбор данных, анкетирование), результатом которых является формирование базы данных первичной информации;
- анализ данных и написание аналитического отчета.

При этом, как следует из рис. В.1, основным результатом первого этапа («Вход») является анкета для опроса целевой аудитории на втором этапе («Исследование»). Второй этап имеет сразу два результата («Выход»). С одной стороны, в результате полевых работ происходит формирование базы данных первичной информации (на основании заполненных анкет), которые затем вводятся в компьютер и анализируются при помощи статистических и когнитивных методов. С другой стороны, в результате анализа данных происходит написание аналитического отчета по исследованию. Оба данных элемента — база данных и аналитический отчет — передаются заказчику (предоставляются руководству компании).

Статистический анализ данных является неотъемлемой частью практически любого серьезного полевого маркетингового исследования. Для его проведения задействуются ресурсы на всех ранее названных этапах маркетингового исследования.

1. На этапе подготовки к исследованию происходит составление анкеты, по которой затем формируется схема кодировки вопросов. Также важный вклад в процесс статистического анализа вносит составление структуры аналитического отчета, которая заранее (еще до сбора данных) позволяет определить, какие переменные будут созданы в базе данных и какие статистические процедуры будут использоваться для их анализа.

2. Когда все анкеты уже собраны и соответствующие данные введены в компьютер, исследователи приступают непосредственно к статистическому анализу. Данный этап, так же как и все маркетинговое исследование в целом, начинается с подготовки

(например, кодирования переменных) и заканчивается практически одновременно с окончанием написания аналитического отчета.

При этом основным ресурсом для проведения статистического анализа является база данных, в которой в закодированном виде содержатся заполненные анкеты по исследованию. В следующем параграфе процесс проведения статистического анализа рассматривается более детально.

Классификация основных методов статистического анализа, применяемых в маркетинговых исследованиях

Несмотря на огромное многообразие существующих статистических методов анализа данных, разработанных в рамках теории математической статистики, в практике маркетинговых исследований находит эффективное применение лишь ограниченный набор статистических инструментов. Такие ограничения отчасти связаны с небольшими, как правило, размерами выборок в большей части проводимых маркетинговых исследований, отчасти — с ограниченной сферой интересов маркетингового анализа, в котором далеко не все существующие статистические методы находят применение. Основываясь на практическом опыте, можно предложить следующую схему классификации статистических методов, используемых при анализе данных количественных маркетинговых исследований (рис. В.2). Схема классификации построена таким образом, как обычно происходит процесс анализа, начиная еще с того момента, когда заказчиком и исполнителем исследования только дописано техническое задание и составлена анкета. Как следует из представленной схемы, весь процесс статистического анализа можно разделить на два этапа: подготовительный и собственно анализ данных.

Первый этап имеет целью собрать и систематизировать информацию, необходимую для последующей обработки анкет (например, схемы кодировки вопросов), а также обеспечить исследователя данными в том виде, который наиболее подходит для конкретного вида статистического анализа. Несмотря на название данного этапа — «предварительный» — некоторые его элементы (в частности, различные манипуляции с формой представления данных) присутствуют и непосредственно в процессе статистического анализа на втором этапе (например, сортировка и отбор анкет). Таким образом, результаты первого, подготовительного этапа используются в течение всего хода статистического (и когнитивного) анализа в маркетинговых исследованиях.

На втором этапе данные, содержащиеся в исходной базе (заполненные анкеты), превращаются в коммерческую информацию: систематизируются, классифицируются, между ними производится поиск взаимосвязей. Результатом второго, основного этапа статистического анализа являются аналитические материалы

(табуляции, диаграммы и вербальные выводы), которые затем используются при написании аналитического отчета.

Рассмотрим теперь основные элементы, составляющие оба этапа статистического анализа, более подробно.

Итак, как мы видим на рис. В.2, подготовительный этап проходит в целом по линейной схеме. Первым шагом здесь является сбор материалов, необходимых для анализа. В полевых маркетинговых исследованиях к ним обычно относятся: техническое задание на исследование, анкета для опроса целевой аудитории, а также структура будущего аналитического отчета, который формируется по результатам проведенного исследования. Данный шаг обычно проводится еще до начала полевых работ (анкетирования), сразу после того, как утверждено задание на исследование. На основании перечисленных материалов вторым шагом определяются так называемые общие параметры выборки, то есть



Подготовительные этапы статистического анализа

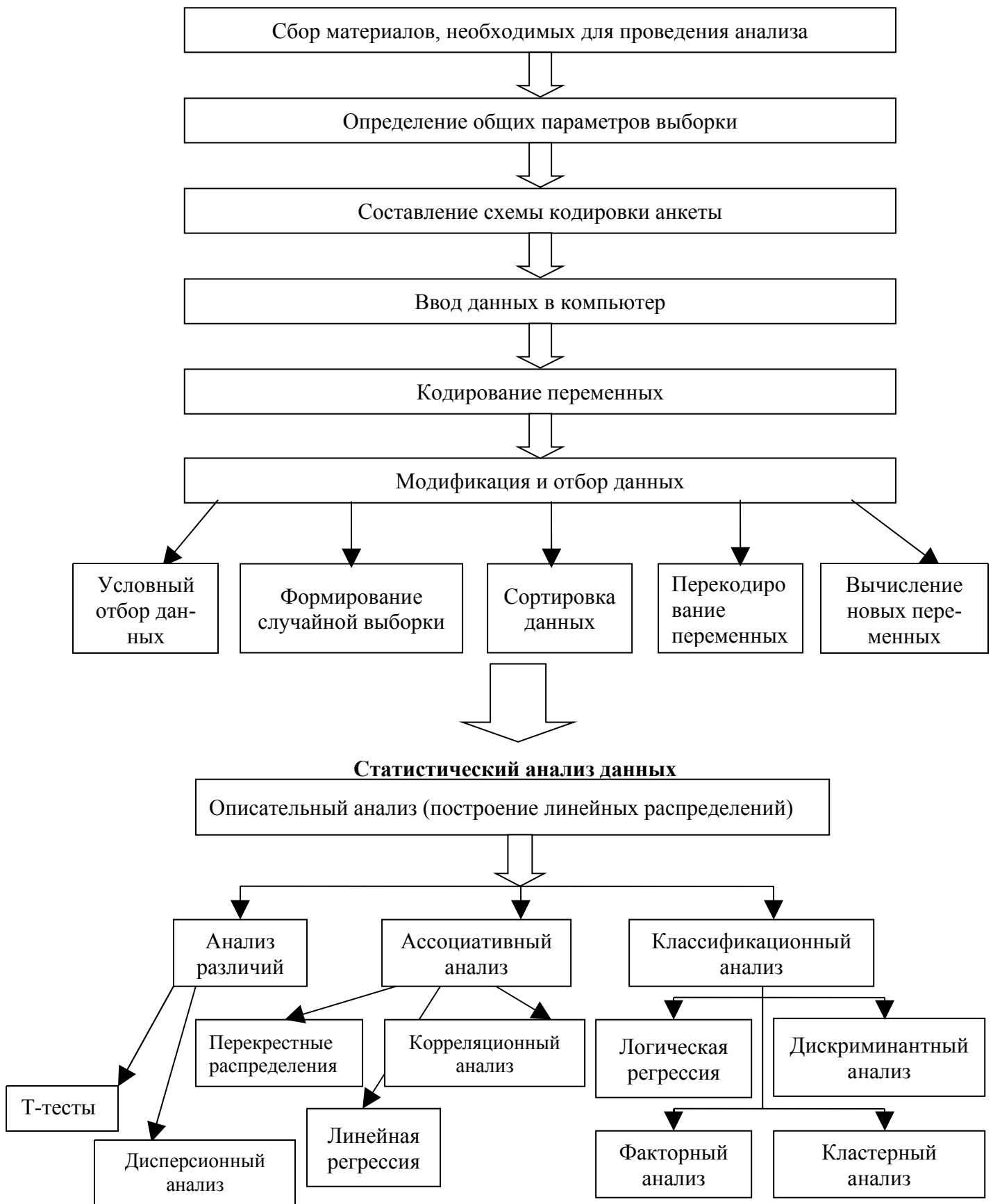


Рис. В.2. Классификация основных методов статистического анализа в процессе его проведения в маркетинговых исследованиях

устанавливается уровень доверия к результатам исследования и рассчитывается статистическая ошибка всей выборки. Необходимо отметить, что данный шаг следует уже после окончания сбора данных, когда появляется возможность точно определить реально получившийся размер выборки, а также получить информацию о сложностях, возникавших в ходе опроса. Эта информация может в дальнейшем внести определенные коррективы в ход статистического анализа. Например, если предварительно заказчиком были установлены квоты по опросу, а в реальности их выдержать не удалось, может потребоваться корректировка базы данных (скажем, удаление анкет одной целевой группы). Далее следует весьма важный шаг — составляется так называемая схема кодировки вопросов и ответов анкеты. С учетом сведений, полученных от интервьюеров, проводивших анкетирование респондентов, исследователь кодирует вопросы и ответы анкеты, формализуя их в соответствии с требованиями, предъявляемыми SPSS (см. п. 1.3). На описываемом шаге также иногда может потребоваться создание специализированной базы данных для проводимого исследования (если ввод данных осуществляется не непосредственно в SPSS, а в какую-либо другую программу — например, в Microsoft Access). Затем на основании имеющейся схемы кодировки анкеты выполняются ввод в компьютер анкет, заполненных в ходе полевых работ, и предварительное формирование базы данных в формате SPSS (создание собственного файла данных с расширением .sav). Окончательное формирование базы данных в SPSS происходит на следующем шаге, когда переменным и их значениям в полученном файле данных присваиваются вербальные метки. И на этом, собственно, заканчивается деятельность по подготовке исходного файла данных для статистического анализа. После осуществления вышеописанных пяти шагов перед исследователем оказывается полностью работоспособная база данных, содержащая все необходимые данные для проведения статистического анализа. Однако у нас остался нерассмотренным еще один существенный шаг в рамках первого, подготовительного этапа — модификация и отбор данных. Данный шаг позволяет аналитику производить предварительные (перед началом статистического анализа) манипуляции с имеющимися данными: перекодировать их, формировать условные и случайные выборки, сортировать, а также вычислять новые переменные на основании имеющихся закодированных вопросов анкеты. Действия, осуществляемые над базой данных в рамках описываемого шага, могут производиться не только непосредственно после ввода данных в компьютер, но и в продолжение всего процесса работы с ними.

Таким образом, данный подготовительный этап статистического анализа осуществляется в шесть основных шагов по линейной схеме. Следующий, основной этап статистического анализа проходит несколько по-другому. Он практически всегда начинается с общей систематизации полученных данных (наиболее часто в форме построения линейных распределений). Дальнейшие шаги статистического анализа полностью зависят от целей исследования и специфики имеющихся данных. Так, исследователю может потребоваться: установить различия между различными целевыми группами респондентов; установить взаимозависимости, существующие между переменными (вопросами анкеты); классифицировать респондентов по группам (сегментировать) на основании определенных критериев. Данные статистические методы могут использоваться как последовательно, так и параллельно: все вместе или только несколько методов (возможно, даже один).

В последующих главах настоящего пособия рассматриваются все перечисленные выше этапы статистического анализа в описанном логическом порядке — с самого начала, то есть начиная с подготовительных этапов анализа.

От издательства

Ваши замечания, предложения и вопросы отправляйте по адресу электронной почты comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

Подробную информацию о наших книгах вы найдете на веб-сайте издательства: <http://www.piter.com>.

Глава 1. Подготовительные этапы статистического анализа

В настоящей главе мы рассмотрим основные методы манипулирования с данными в SPSS. Рассматриваемые здесь действия обычно производятся перед началом статистического анализа. Мы начнем обсуждение с самого начала — то есть с того момента, когда к исследователю попадает задание на проведение маркетингового исследования. Далее по порядку будут рассмотрены все основные действия с матрицей данных.

1.1. Материалы, необходимые для проведения статистического анализа

Первым шагом при подготовке к проведению статистического анализа данных в маркетинговых исследованиях является подбор исходных материалов, в которых содержатся основные параметры проводимого исследования. Обычно эти материалы включают в себя следующие документы.

1. **Техническое задание на исследование (ТЗ)** охватывает все общие параметры исследования: цели и задачи, планируемый размер выборки, информацию о квотах, методе и месте сбора данных, а также другую полезную информацию.

2. **Структура аналитического отчета по результатам исследования** позволяет определить заранее, какие статистические процедуры понадобятся при написании аналитического отчета по исследованию.

3. **Анкета для опроса** является основой для составления схемы кодировки переменных в базе данных SPSS.

На основании ТЗ и структуры аналитического отчета исследователь должен еще до получения данных для анализа (заполненных анкет) составить план предстоящих манипуляций с анкетами респондентов: преобразования данных, статистических процедур и методик. Исследователь должен приступить к обработке анкеты сразу после ее получения, не дожидаясь окончания полевых работ: изучить ее структуру и составить перечень переменных, которые впоследствии войдут в базу данных SPSS.

Основными выходными данными на названном этапе являются:

- планируемый размер выборки;
- структура выборки (наличие и размер квот);
- вид опроса (личный, телефонный);
- информация о параметрах опроса (наличие фактов фальсификации анкет);
- схема (таблица) кодировки переменных в базе данных SPSS;
- план-схема преобразования данных;
- план-схема используемых статистических процедур.

Как вы увидите далее, эти данные являются весьма ценным ресурсом для последующего статистического анализа.

Необходимо отметить, что на рассматриваемом этапе также можно выполнять и другие действия. Так, если заполненные анкеты вводятся в компьютер при помощи специализированного программного обеспечения (например, программы Data Entry или сканерного программного комплекса), на основании имеющейся анкеты и согласно целям и задачам исследования следует сформировать соответствующие формы (для программы Data Entry) или создать шаблоны и макеты анкеты (для сканерного ввода). Только после успешного завершения этого подготовительного шага можно приступить к дальнейшим этапам.

1.2. Общие параметры выборки

Определение общих параметров выборки осуществляется после завершения поле-

вых работ (когда собраны все анкеты). Данный этап состоит из ряда взаимосвязанных шагов. Это:

- определение реального количества опрошенных респондентов;
- определение структуры выборки;
- распределение по месту опроса;
- установление доверительного уровня статистической надежности выборки;
- расчет статистической ошибки и определение репрезентативности выборки.

Первое, что должно интересовать исследователя после получения заполненных анкет, — это количество респондентов. Оно может быть либо больше, либо меньше запланированного количества анкет. При этом первый вариант лучше с точки зрения статистического анализа, но хуже с точки зрения руководства фирмы, так как дополнительные анкеты являются незапланированными расходами на оплату работы интервьюеров. Второй вариант обычно хуже и с точки зрения анализа (выборка менее представительна), и с точки зрения руководства (заказчик будет недоволен несоблюдением требований, оговоренных в ТЗ).

При оценке разницы между реальным и плановым размером выборки следует принимать в расчет разницу в статистической ошибке (см. ниже). Если она невелика (в ту или другую сторону), репрезентативность всей выборки существенно не страдает. Но если разница достаточно значима, выборка может оказаться неrepresentative. Кроме того, при определении общего размера выборки необходимо иметь в виду, что статистическая ошибка всей выборки относится только к общим распределениям. Разрезы существенно увеличивают статистическую ошибку. Поэтому еще до начала опроса следует определить, какая численность каждой из интересующих целевых групп респондентов является достаточной для построения статистически значимых заключений и выводов.

Структура выборки может быть случайной (респонденты отбирались в случайном порядке) или неслучайной (респонденты отбирались на основании заранее известных критериев, например методом квотирования). Эта информация важна при интерпретации результатов статистического анализа. Случайные выборки априори являются репрезентативными, так как на попадание/непопадание каждого респондента в выборку не влияют никакие факторы, кроме случайных. Представительность неслучайных выборок не следует из их определения. Иногда они специально делаются нерепрезентативными относительно генеральной совокупности, однако могут являться весьма представительными относительно какой-либо одной интересующей целевой группы (например, исследуется только мнение мужчин в возрасте после 40 лет).

При анализе структуры выборки необходимо также изучить фильтрационные вопросы анкеты, то есть вопросы, специально предназначенные для отсеивания не подходящих под требования выборки респондентов. Несмотря на то, что такие вопросы позволяют исключить не нужные для конкретного исследования целевые группы, знание доли исключенных категорий позволит впоследствии составить общее представление о параметрах всей генеральной совокупности.

Приведем пример. Методом телефонного опроса исследуется потребительский спрос на московском рынке творожной массы. При этом опрашиваются только лица, покупающие данный продукт, — для чего в анкету добавлен соответствующий фильтрационный вопрос. Однако в дальнейшем потребуется рассчитать емкость рынка исследуемого продукта. Решением данной задачи будет подсчет количества отсеянных респондентов (лиц, не покупающих творожную массу). Таким образом, впоследствии мы сможем определить долю покупателей творожной массы от общей численности населения Москвы.

Еще одна важная для исследователя характеристика выборки — это распределение респондентов по месту опроса (личные интервью). Позже эти данные могут помочь при определении различий между респондентами, опрошенными в разных местах. (Очевидна разница в доходах между посетителями рынков и бутиков.)

Имея в своем распоряжении указанную выше информацию, можно приступать к определению представительности (или репрезентативности) выборки. Прежде всего необ-

ходимо установить уровень доверия к результатам опроса. Обычно в маркетинговых исследованиях используется уровень доверия 95 % и 99 %. Мы рекомендуем остановиться именно на первом варианте как на наиболее релевантном по отношению к маркетинговым исследованиям.

В зависимости от выбранного доверительного уровня определяется специфическая константа z , участвующая в формуле расчета статистической ошибки выборки. Константы доверительных уровней, наиболее часто используемых в маркетинговых исследованиях, представлены в табл. 1.1.

Таблица 1.1. Константы доверительных уровней

Доверительный уровень	Константа z
90 %	$\pm 1,64$
95 %	$\pm 1,96$
99 %	$\pm 2,58$

Максимальная статистическая ошибка выборки рассчитывается по следующей формуле:

$$\Delta_{x\%} = \pm z \sqrt{\frac{pq}{n}},$$

где — статистическая константа для соответствующего доверительного уровня; $p = q = 50\%$ — вероятность наступления/ненаступления исследуемого события (то есть попадания/непопадания респондента в выборку); для случайных выборок данная вероятность равна $1/2$ или 50% ; n — размер выборки (общее количество опрошенных).

Таким образом, для выборки в 1000 респондентов и при уровне доверия к результатам опроса 95 % статистическая ошибка выборки будет равна:

$$\Delta = \pm 1,96 \sqrt{\frac{50 \cdot 50}{1000}} = \pm 1,96 \times \sqrt{2,5} \approx \pm 3,1\%.$$

Эта же статистическая ошибка используется для характеристики всех значений в выборке, выраженных в относительных величинах. То есть если в дальнейшем при построении линейных распределений по вопросам анкеты мы выясним, что 32 % респондентов покупают газеты в киосках на улице, — это будет означать, что данное значение варьируется в пределах от 28,9 % (32 % - 3,1 %) до 35,1 % (32 % + 3,1 %).

Для расчета статистической ошибки значений переменных, выраженных в абсолютных величинах, применяется другая формула. При этом ошибка варьируется в зависимости от конкретной анализируемой величины. Ее расчет основан на построении линейных распределений и показан в разделе 2.1.

1.3. Составление схемы кодировки анкеты

Схема кодировки анкеты представляет собой таблицу соответствия вопросов и вариантов ответа анкеты внутреннему представлению переменных в базе данных SPSS. Впоследствии ввод анкет в компьютер и кодирование ответов респондентов производятся согласно данной формализованной структуре. Пример таблицы кодировки представлен в табл. 1.2.

Как вы видите, различные типы вопросов анкеты кодируются в схеме кодировки (и в базе данных SPSS) по-разному. Существует три основных типа кодирования вопросов анкеты.

1. Закрытые вопросы, в которых респондент может указать только один вариант ответа (одновариантные), кодируются одной переменной (например, $q1$). Тип шкалы в дан-

ном случае может быть любым.

2. Закрытые вопросы, в которых респондент может дать несколько вариантов ответа (многовариантные), кодируются несколькими одновариантными переменными (например, q3_1, q3_2). Тип шкалы одновариантных переменных может быть только номинальным (дихотомическим).

3. Открытые вопросы, независимо от количества возможных вариантов ответа на них, кодируются одной переменной. Тип шкалы в данном случае может быть либо интервальным (для числовых данных, например q5_t), либо номинальным (для нечисловых данных, например q4_t).

Таблица 1.2. Кодировка различных типов вопросов

Вопрос анкеты	Код и тип переменной в базе данных
Номер анкеты _____	n_resr – интервальная шкала
1. Покупаете ли Вы мясные полфобриканы?	q1 – номинальная шкала
Да	Вариант ответа 1
Нет	Вариант ответа 2
2. Как часто Вы покупаете эти продукты?	q2 – порядковая шкала
Почти каждый день	вариант ответа 1
2-3 раза в неделю	вариант ответа 2
Примерно раз в неделю	вариант ответа 3
2-3 раза в месяц	вариант ответа 4
Примерно раз в месяц	вариант ответа 5
Реже раза в месяц	вариант ответа 6
3. Где Вы обычно покупаете мясные продукты? (возможно несколько ответов)	Все варианты ответа являются номинальными переменными
В магазине	q3_1
На рынке	q3_2
В супермаркете	q3_3
Другое (укажите где именно)	q3_4
	q3_4t
4. Каких производителей мясных продуктов Вы знаете?	q4_1t – номинальная шкала
5. Укажите Ваш возраст: _____ лет	q5_1t – интервальная шкала

1.4. Ввод данных в компьютер и кодирование переменных

Ввод данных в компьютер является четвертым шагом первого (подготовительного) этапа статистического анализа данных (см. рис. В.2). Он неразрывно связан со следующим шагом — кодированием переменных. В этом разделе мы последовательно рассмотрим эти две взаимосвязанные и взаимообусловленные процедуры.

1.4.1. Способы ввода данных в SPSS

Существует три основных способа формирования базы данных в формате SPSS (перечислены в порядке убывания популярности).

1. Импорт базы данных из других программных источников (Microsoft Access, Microsoft Excel, текстовых файлов и других).

2. Ввод данных непосредственно в SPSS при помощи специализированного программного обеспечения (SPSS Data Entry).

3. Ручной ввод данных в SPSS.

Теперь рассмотрим каждый способ более подробно.

1.4.1.1. Импорт данных из других источников

Данный способ создания базы данных в формате SPSS является наиболее распространенным. Чаще всего он предполагает использование SPSS в качестве вспомогательного средства для статистического анализа данных. При этом построение линейных распределений в графическом виде (диаграмм по общим распределениям) может производиться, например, в Microsoft Excel. Также данный метод применим и если у вас есть программное обеспечение для автоматически сканируемого ввода бумажных анкет в компьютер. В этом случае специализированная программа (например, ABBYY FormReader) создает особую базу данных в собственном формате (во внутреннем представлении).

Рассмотрим пример создания базы данных в SPSS при помощи перекачки данных из другой программы — Microsoft Access, как одной из наиболее распространенных систем управления базами данных (СУБД).

Чтобы осуществить импорт данных в SPSS, необходимо сформировать в соответствующей программе (из которой будет осуществляться импорт) таблицу данных, отформатированную определенным способом. Файл данных SPSS напоминает рабочую книгу Microsoft Excel (электронную таблицу). Однако SPSS, к сожалению, не обладает функциональностью электронной таблицы, и схожесть этих двух программных продуктов заканчивается на внешнем виде. Общая схема построения файла SPSS выглядит примерно так, как на рис. 1.1.

Таблица данных в сторонней программе, из которой будет осуществляться импорт, должна соответствовать именно такой схеме (заголовок переменной → значения переменной). Примеры таблиц из Microsoft Access Base.mdb, Microsoft Excel Base.xls, простого текстового файла MS DOS Base.txt и текстового файла с разделителями Base.csv представлены на рис. 1.2-1.5. Независимо от вида разделителей данных в таблицах их объединяет общая структура: заголовок переменной → данные (значение переменной). Представим, что была создана база данных Microsoft Access Base.mdb, содержащая Таблицу данных.

После того как была создана подходящая для импорта таблица данных, следует открыть SPSS и вызвать диалоговое окно импорта данных при помощи меню File ► Open Database ► New Query. Откроется мастер Database Wizard (рис. 1.6); в его окне необходимо указать источник данных, из которого будет производиться импорт данных. Выберите в списке справа База данных MS Access и щелкните на кнопке Далее.

Следует отметить, что SPSS поддерживает импорт из любых источников данных, совместимых с технологией ODBC (соответствующие драйверы для них должны быть предварительно установлены в Microsoft Windows). Например, чтобы добавить возможность импорта из базы данных Microsoft Paradox (файлы типа *.db), необходимо щелкнуть на кнопке Add Data Source в диалоговом окне Database Wizard. На экране появится стандартное окно Microsoft Windows Администратор источников данных ODBC (рис. 1.7). В этом диалоговом окне представлен список уже установленных в SPSS источников данных. Чтобы добавить новый источник, отсутствующий в данном перечне, следует щелкнуть на кнопке Добавить.

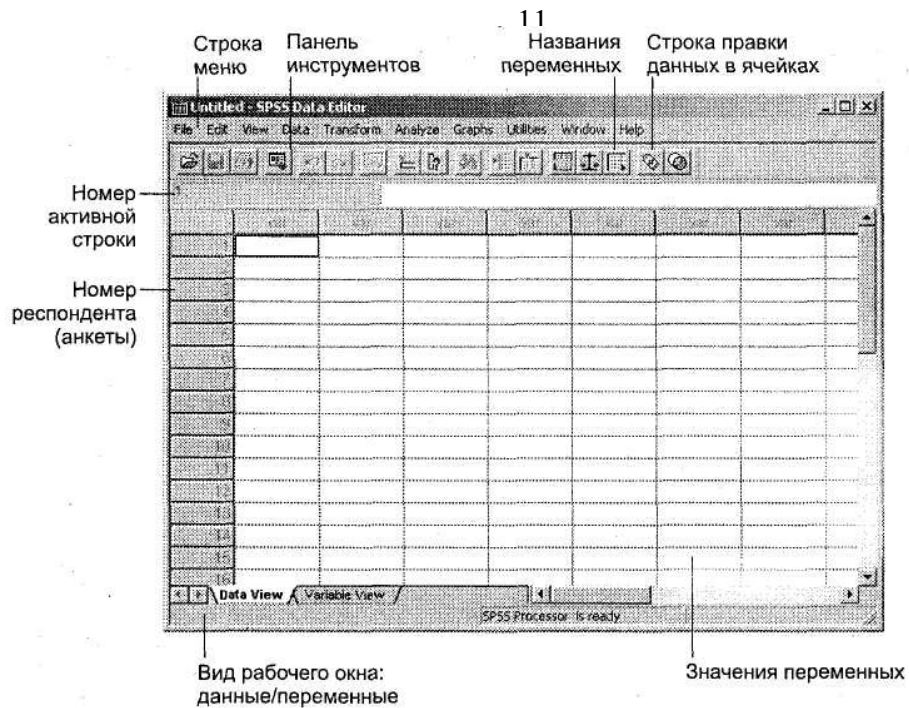


Рис 1.1 Общая схема построения файла данных SPSS

Microsoft Access - [for SPSS: таблица]

Файл Правка Вид Вставка Формат Записи Сервис Окно Справка

Введите текст

	N_групп	Kod_rev	q1	q2	q3	q4	q5	q6	
1	0	3	7	8	3	10	4		
2	0	8	7	10	4	10	6		
3	0	6	4	6	3	5	3		
4	0	8	1	3	2	10	4		
5	0	9	10	10	7	10	1		
6	0	7	6	10	1	10	5		
7	0	8	10	10	8	10	10		
8	0	7	9	8	5	10	2		
9	0	5	7	10	1	5	2		
10	0	6	6	10	1	10	3		
11	0	8	9	8	2	10	8		
12	0	5	5	4	3	4	4		
13	0	8	9	10	1	10	2		
14	0	8	8	10	1	10	8		
15	0	8	10	10	1	8	4		
16	0	4	6	4	1	3	3		
17	0	3	3	5	5	5	1		
18	0	5	4	3	2	3	3		
19	0	9	2	7	3	7	5		
20	0	5	8	5	1	5	1		
21	0	10	10	10	2	10	5		
22	0	7	8	7	2	5	5		

Запись: 14 из 1577

Режим таблицы

NUM

Рис. 1.2. Таблицы данных, подходящие для импорта в SPSS: таблица MS Access

[illegible]

Рис. 1.3. Таблицы данных, подходящие для импорта
в SPSS: лист MS Excel

Base.txt - БЛОКНОТ																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
Файл	Правка	Формат	Вид	Справка																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																									
N_RESP	KOD_REV	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62	Q63	Q64	Q65	Q66	Q67	Q68	Q69	Q70	Q71	Q72	Q73	Q74	Q75	Q76	Q77	Q78	Q79	Q80	Q81	Q82	Q83	Q84	Q85	Q86	Q87	Q88	Q89	Q90	Q91	Q92	Q93	Q94	Q95	Q96	Q97	Q98	Q99	Q100	Q101	Q102	Q103	Q104	Q105	Q106	Q107	Q108	Q109	Q110	Q111	Q112	Q113	Q114	Q115	Q116	Q117	Q118	Q119	Q120	Q121	Q122	Q123	Q124	Q125	Q126	Q127	Q128	Q129	Q130	Q131	Q132	Q133	Q134	Q135	Q136	Q137	Q138	Q139	Q140	Q141	Q142	Q143	Q144	Q145	Q146	Q147	Q148	Q149	Q150	Q151	Q152	Q153	Q154	Q155	Q156	Q157	Q158	Q159	Q160	Q161	Q162	Q163	Q164	Q165	Q166	Q167	Q168	Q169	Q170	Q171	Q172	Q173	Q174	Q175	Q176	Q177	Q178	Q179	Q180	Q181	Q182	Q183	Q184	Q185	Q186	Q187	Q188	Q189	Q190	Q191	Q192	Q193	Q194	Q195	Q196	Q197	Q198	Q199	Q200	Q201	Q202	Q203	Q204	Q205	Q206	Q207	Q208	Q209	Q210	Q211	Q212	Q213	Q214	Q215	Q216	Q217	Q218	Q219	Q220	Q221	Q222	Q223	Q224	Q225	Q226	Q227	Q228	Q229	Q230	Q231	Q232	Q233	Q234	Q235	Q236	Q237	Q238	Q239	Q240	Q241	Q242	Q243	Q244	Q245	Q246	Q247	Q248	Q249	Q250	Q251	Q252	Q253	Q254	Q255	Q256	Q257	Q258	Q259	Q260	Q261	Q262	Q263	Q264	Q265	Q266	Q267	Q268	Q269	Q270	Q271	Q272	Q273	Q274	Q275	Q276	Q277	Q278	Q279	Q280	Q281	Q282	Q283	Q284	Q285	Q286	Q287	Q288	Q289	Q290	Q291	Q292	Q293	Q294	Q295	Q296	Q297	Q298	Q299	Q300	Q301	Q302	Q303	Q304	Q305	Q306	Q307	Q308	Q309	Q310	Q311	Q312	Q313	Q314	Q315	Q316	Q317	Q318	Q319	Q320	Q321	Q322	Q323	Q324	Q325	Q326	Q327	Q328	Q329	Q330	Q331	Q332	Q333	Q334	Q335	Q336	Q337	Q338	Q339	Q340	Q341	Q342	Q343	Q344	Q345	Q346	Q347	Q348	Q349	Q350	Q351	Q352	Q353	Q354	Q355	Q356	Q357	Q358	Q359	Q360	Q361	Q362	Q363	Q364	Q365	Q366	Q367	Q368	Q369	Q370	Q371	Q372	Q373	Q374	Q375	Q376	Q377	Q378	Q379	Q380	Q381	Q382	Q383	Q384	Q385	Q386	Q387	Q388	Q389	Q390	Q391	Q392	Q393	Q394	Q395	Q396	Q397	Q398	Q399	Q400	Q401	Q402	Q403	Q404	Q405	Q406	Q407	Q408	Q409	Q410	Q411	Q412	Q413	Q414	Q415	Q416	Q417	Q418	Q419	Q420	Q421	Q422	Q423	Q424	Q425	Q426	Q427	Q428	Q429	Q430	Q431	Q432	Q433	Q434	Q435	Q436	Q437	Q438	Q439	Q440	Q441	Q442	Q443	Q444	Q445	Q446	Q447	Q448	Q449	Q450	Q451	Q452	Q453	Q454	Q455	Q456	Q457	Q458	Q459	Q460	Q461	Q462	Q463	Q464	Q465	Q466	Q467	Q468	Q469	Q470	Q471	Q472	Q473	Q474	Q475	Q476	Q477	Q478	Q479	Q480	Q481	Q482	Q483	Q484	Q485	Q486	Q487	Q488	Q489	Q490	Q491	Q492	Q493	Q494	Q495	Q496	Q497	Q498	Q499	Q500	Q501	Q502	Q503	Q504	Q505	Q506	Q507	Q508	Q509	Q510	Q511	Q512	Q513	Q514	Q515	Q516	Q517	Q518	Q519	Q520	Q521	Q522	Q523	Q524	Q525	Q526	Q527	Q528	Q529	Q530	Q531	Q532	Q533	Q534	Q535	Q536	Q537	Q538	Q539	Q540	Q541	Q542	Q543	Q544	Q545	Q546	Q547	Q548	Q549	Q550	Q551	Q552	Q553	Q554	Q555	Q556	Q557	Q558	Q559	Q560	Q561	Q562	Q563	Q564	Q565	Q566	Q567	Q568	Q569	Q570	Q571	Q572	Q573	Q574	Q575	Q576	Q577	Q578	Q579	Q580	Q581	Q582	Q583	Q584	Q585	Q586	Q587	Q588	Q589	Q590	Q591	Q592	Q593	Q594	Q595	Q596	Q597	Q598	Q599	Q600	Q601	Q602	Q603	Q604	Q605	Q606	Q607	Q608	Q609	Q610	Q611	Q612	Q613	Q614	Q615	Q616	Q617	Q618	Q619	Q620	Q621	Q622	Q623	Q624	Q625	Q626	Q627	Q628	Q629	Q630	Q631	Q632	Q633	Q634	Q635	Q636	Q637	Q638	Q639	Q640	Q641	Q642	Q643	Q644	Q645	Q646	Q647	Q648	Q649	Q650	Q651	Q652	Q653	Q654	Q655	Q656	Q657	Q658	Q659	Q660	Q661	Q662	Q663	Q664	Q665	Q666	Q667	Q668	Q669	Q670	Q671	Q672	Q673	Q674	Q675	Q676	Q677	Q678	Q679	Q680	Q681	Q682	Q683	Q684	Q685	Q686	Q687	Q688	Q689	Q690	Q691	Q692	Q693	Q694	Q695	Q696	Q697	Q698	Q699	Q700	Q701	Q702	Q703	Q704	Q705	Q706	Q707	Q708	Q709	Q710	Q711	Q712	Q713	Q714	Q715	Q716	Q717	Q718	Q719	Q720	Q721	Q722	Q723	Q724	Q725	Q726	Q727	Q728	Q729	Q730	Q731	Q732	Q733	Q734	Q735	Q736	Q737	Q738	Q739	Q740	Q741	Q742	Q743	Q744	Q745	Q746	Q747	Q748	Q749	Q750	Q751	Q752	Q753	Q754	Q755	Q756	Q757	Q758	Q759	Q760	Q761	Q762	Q763	Q764	Q765	Q766	Q767	Q768	Q769	Q770	Q771	Q772	Q773	Q774	Q775	Q776	Q777	Q778	Q779	Q780	Q781	Q782	Q783	Q784	Q785	Q786	Q787	Q788	Q789	Q790	Q791	Q792	Q793	Q794	Q795	Q796	Q797	Q798	Q799	Q800	Q801	Q802	Q803	Q804	Q805	Q806	Q807	Q808	Q809	Q810	Q811	Q812	Q813	Q814	Q815	Q816	Q817	Q818	Q819	Q820	Q821	Q822	Q823	Q824	Q825	Q826	Q827	Q828	Q829	Q830	Q831	Q832	Q833	Q834	Q835	Q836	Q837	Q838	Q839	Q840	Q841	Q842	Q843	Q844	Q845	Q846	Q847	Q848	Q849	Q850	Q851	Q852	Q853	Q854	Q855	Q856	Q857	Q858	Q859	Q860	Q861	Q862	Q863	Q864	Q865	Q866	Q867	Q868	Q869	Q870	Q871	Q872	Q873	Q874	Q875	Q876	Q877	Q878	Q879	Q880	Q881	Q882	Q883	Q884	Q885	Q886	Q887	Q888	Q889	Q890	Q891	Q892	Q893	Q894	Q895	Q896	Q897	Q898	Q899	Q900	Q901	Q902	Q903	Q904	Q905	Q906	Q907	Q908	Q909	Q910	Q911	Q912	Q913	Q914	Q915	Q916	Q917	Q918	Q919	Q920	Q921	Q922	Q923	Q924	Q925	Q926	Q927	Q928	Q929	Q930	Q931	Q932	Q933	Q934	Q935	Q936	Q937	Q938	Q939	Q940	Q941	Q942	Q943	Q944	Q945	Q946	Q947	Q948	Q949	Q950	Q951	Q952	Q953	Q954	Q955	Q956	Q957	Q958	Q959	Q960	Q961	Q962	Q963	Q964	Q965	Q966	Q967	Q968	Q969	Q970	Q971	Q972	Q973	Q974	Q975	Q976	Q977	Q978	Q979	Q980	Q981	Q982	Q983	Q984	Q985	Q986	Q987	Q988	Q989	Q990	Q991	Q992	Q993	Q994	Q995	Q996	Q997	Q998	Q999	Q1000	Q1001	Q1002	Q1003	Q1004	Q1005	Q1006	Q1007	Q1008	Q1009	Q1010	Q1011	Q1012	Q1013	Q1014	Q1015	Q1016	Q1017	Q1018	Q1019	Q1020	Q1021	Q1022	Q1023	Q1024	Q1025	Q1026	Q1027	Q1028	Q1029	Q1030	Q1031	Q1032	Q1033	Q1034	Q1035	Q1036	Q1037	Q1038	Q1039	Q1040	Q1041	Q1042	Q1043	Q1044	Q1045	Q1046	Q1047	Q1048	Q1049	Q1050	Q1051	Q1052	Q1053	Q1054	Q1055	Q1056	Q1057	Q1058	Q1059	Q1060	Q1061	Q1062	Q1063	Q1064	Q1065	Q1066	Q1067	Q1068	Q1069	Q1070	Q1071	Q1072	Q1073	Q1074	Q1075	Q1076	Q1077	Q1078	Q1079	Q1080	Q1081	Q1082	Q1083	Q1084	Q1085	Q1086	Q1087	Q1088	Q1089	Q1090	Q1091	Q1092	Q1093	Q1094	Q1095	Q1096	Q1097	Q1098	Q1099	Q1100	Q1101	Q1102	Q1103	Q1104	Q1105	Q1106	Q1107	Q1108	Q1109	Q1110	Q1111	Q1112	Q1113	Q1114	Q1115	Q1116	Q1117	Q1118	Q1119	Q1120	Q1121	Q1122	Q1123	Q1124	Q1125	Q1126	Q1127	Q1128	Q1129	Q1130	Q1131	Q1132	Q1133	Q1134	Q1135	Q1136	Q1137	Q1138	Q1139	Q1140	Q1141	Q1142	Q1143	Q1144	Q1145	Q1146	Q1147	Q1148	Q1149	Q1150	Q1151	Q1152	Q1153	Q1154	Q1155	Q1156	Q1157	Q1158	Q1159	Q1160	Q1161	Q1162	Q1163	Q1164	Q1165	Q1166	Q1167	Q1168	Q1169	Q1170	Q1171	Q1172	Q1173	Q1174	Q1175	Q1176	Q1177	Q1178	Q1179	Q1180	Q1181	Q1182	Q1183	Q1184	Q1185	Q1186	Q1187	Q1188	Q1189	Q1190	Q1191	Q1192	Q1193	Q1194	Q1195	Q1196	Q1197	Q1198	Q1199	Q1200	Q1201	Q1202	Q1203	Q1204	Q1205	Q1206	Q1207	Q1208	Q1209	Q1210	Q1211	Q1212	Q1213	Q1214	Q1215	Q1216	Q1217	Q1218	Q1219	Q1220	Q1221	Q1222	Q1223	Q1224	Q1225	Q1226	Q1227	Q1228	Q1229	Q1230	Q1231	Q1232	Q1233	Q1234	Q1235	Q1236	Q1237	Q1238	Q1239	Q1240	Q1241	Q1242	Q1243	Q1244	Q1245	Q1246	Q1247	Q1248	Q1249	Q1250	Q1251	Q1252	Q1253	Q1254	Q1255	Q1256	Q1257	Q1258	Q1259	Q1260	Q1261	Q1262	Q1263	Q1264	Q1265	Q1266	Q1267	Q1268	Q1269	Q1270	Q1271	Q1272	Q1273	Q1274	Q1275	Q1276	Q1277	Q1278	Q1279	Q1280	Q1281	Q1282	Q1283	Q1284	Q1285	Q1286	Q1287	Q1288	Q1289	Q1290	Q1291	Q1292	Q1293	Q1294	Q1295	Q1296	Q1297	Q1298	Q1299	Q1300	Q1301	Q1302	Q1303	Q1304	Q1305	Q1306	Q1307	Q1308	Q1309	Q1310	Q1311	Q1312	Q1313	Q1314	Q1315	Q1316	Q1317	Q1318	Q1319	Q1320	Q1321	Q1322	Q1323	Q1324	Q1325	Q1326	Q1327	Q1328	Q1329	Q1330	Q1331	Q1332	Q1333	Q1334	Q1335	Q1336	Q1337	Q1338	Q1339	Q134

Рис. 1.4. Таблицы данных, подходящие для импорта в SPSS: текстовый файл с фиксированными столбцами

[illegible]

Рис. 1.5 Таблицы данных, подходящие для импорта SPSS: текстовый файл с разделителями

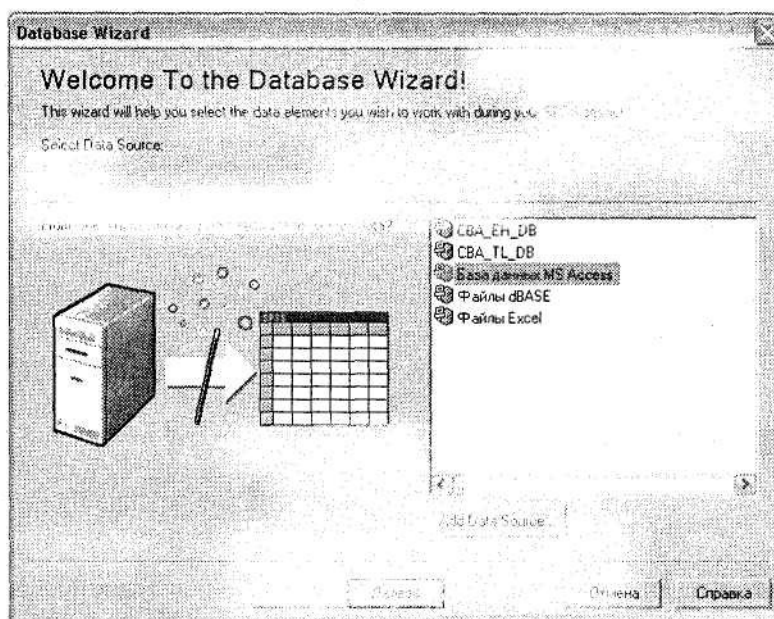


Рис. 1.5 Диалоговое окно Database Wizard

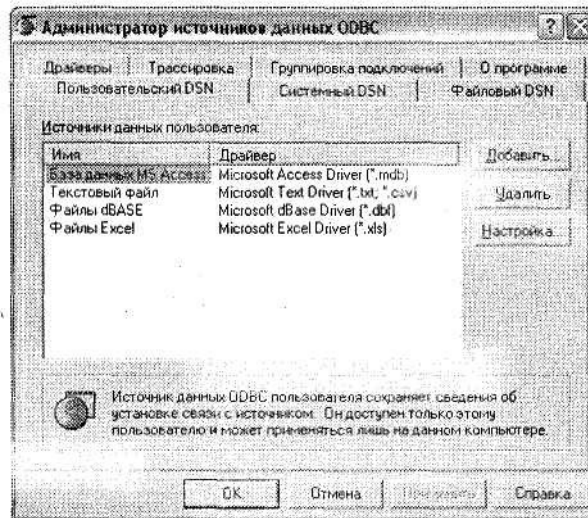


Рис. 1.7. Диалоговое окно Администратор источников данных ODBC

В открывшемся диалоговом окне Создание нового источника данных (рис. 1.8) содержится список всех источников данных, установленных в вашей системе Microsoft Windows. Кроме названий источников, в данном перечне вы можете увидеть номер версии и название файла соответствующего драйвера. Выберите драйвер Microsoft Paradox Driver (*.db) и щелкните на кнопке Готово.



Рис. 1.8. Диалоговое окно Создание нового источника данных

При этом будет открыто новое диалоговое окно Установка драйвера ODBC для Paradox (рис. 1.9). Здесь в строке Имя источника данных следует ввести то название, которое будет в дальнейшем отображаться в диалоговом окне Database Wizard в SPSS (например, База данных Paradox). В этом диалоговом окне можно установить дополнительные параметры. Чтобы вернуться в SPSS, следует закрыть все использован-

ные диалоговые окна установки источника данных ODBC. Вы увидите, что в списке доступных источников в окне Database Wizard появится база данных Paradox.

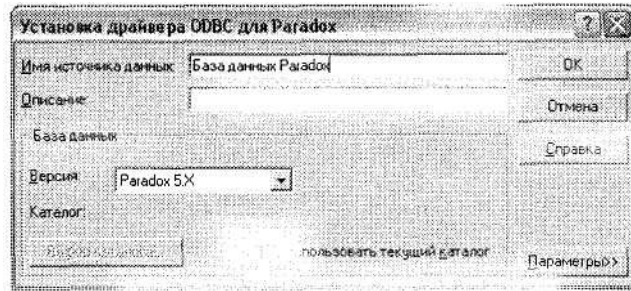


Рис. 1.9. Диалоговое окно Установка драйвера ODBC для Paradox

Вернемся к рис. 1.6. Выберите соответствующий источник данных и щелкните на кнопке Далее, после чего на экране откроется диалоговое окно ODBC Driver Login (рис. 1.10). В этом окне следует указать полный путь к базе данных, из которой будет производиться импорт таблицы (в нашем случае это C:\Base.mdb). Щелкните на кнопке OK для продолжения работы.

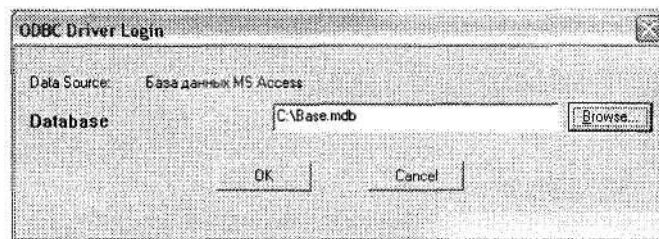


Рис. 1.10. Диалоговое окно ODBC Driver Login

Откроется новое диалоговое окно (рис. 1.11). В нем из левого списка всех таблиц, доступных в указанном источнике данных, выберите ту, которая содержит импортируемые данные (в нашем случае Таблица данных), и переместите ее в правый список. Затем щелкните на кнопке Готово, после чего в окне SPSS Data Editor появится импортированная таблица.

Следует отметить, что процедуры импорта данных для разных источников отличаются друг от друга. Однако эти различия несущественны, и поэтому мы не будем описывать все типы импорта. Как правило, для таблицы из базы данных Microsoft Access действия, показанные при помощи вышеописанных шагов, достаточны.

1.4.1.2. Ввод данных в SPSS при помощи Data Entry

Данная программа призвана упростить ввод данных в SPSS. При работе с ней генерируются пользовательские формы, содержащие поля анкеты, куда и вводятся данные. Модуль SPSS Data Entry Builder позволяет создавать формы и правила для их заполнения, а модуль SPSS Data Entry Station — вводить анкеты в компьютер в распределенном режиме (то есть с нескольких компьютеров одновременно). Детальное описание работы с программой Data Entry выходит за рамки настоящего пособия. Отметим лишь, что данная программа является самостоятельным приложением Microsoft Windows и не входит в комплект поставки SPSS. Кроме того, программные продукты SPSS достаточно дороги для большинства российских компаний, и поэтому рассматриваемый способ ввода данных не получил должного распространения в нашей стране.

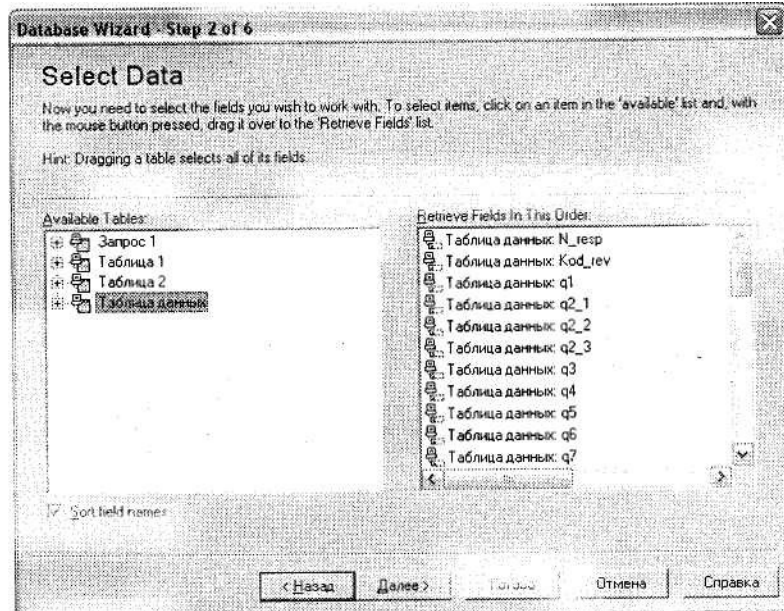


Рис. 1.11. Диалоговое окно Database Wizard, шаг 2 из 6

1.4.1.3. Ручной ввод данных в SPSS

Ручной ввод наиболее эффективен при малых размерах выборки, а также для достижения некоторых специфических целей (например, при вводе ранжированных списков в ходе расчета корреляции Спирмана; см. раздел 4.2.1). Как и в случае использования программы Data Entry, существует возможность распределенного ввода анкет с несколькими операторами. Когда все операторы закончат ввод своей части анкет, полученные базы данных сливаются в одну при помощи меню SPSS Data ► Merge files, в котором следует выбрать объект добавления анкеты (Add Cases) или переменных (Add Variables).

1.4.2. Кодирование переменных

После того как в файл SPSS помещена таблица с данными по исследованию, следует перейти к очередному этапу формирования базы данных — кодированию переменных.

Если данные вводились в SPSS методом импорта, вы увидите только имена переменных и их значения. В этом случае кодирование переменных является обязательным шагом и должно проводиться сразу после процедуры импорта. Если для

ввода данных в SPSS использовалась программа Data Entry, все переменные и их значения окажутся, скорее всего, уже закодированными (на этапе генерирования пользовательских форм). При ручном вводе картина может быть такой, как при импорте данных из других источников (если вы предварительно не производили кодирование), либо аналогичной использованию Data Entry. Тем не менее, независимо от способа ввода, на этапе кодирования необходимо произвести ревизию имеющихся переменных и меток их значений — чтобы удостовериться, что в будущем при проведении статистического анализа все используемые величины будут названы осмысленными именами.

Основное рабочее окно SPSS (см. рис. 1.1) содержит специальные вкладки для перемещения между видом файла данных (Data View) и таблицы переменных (Variable View). Кодирование переменных осуществляется на вкладке Variable View. Общий вид окна программы после щелчка на вкладке Variable View показан на рис. 1.121.

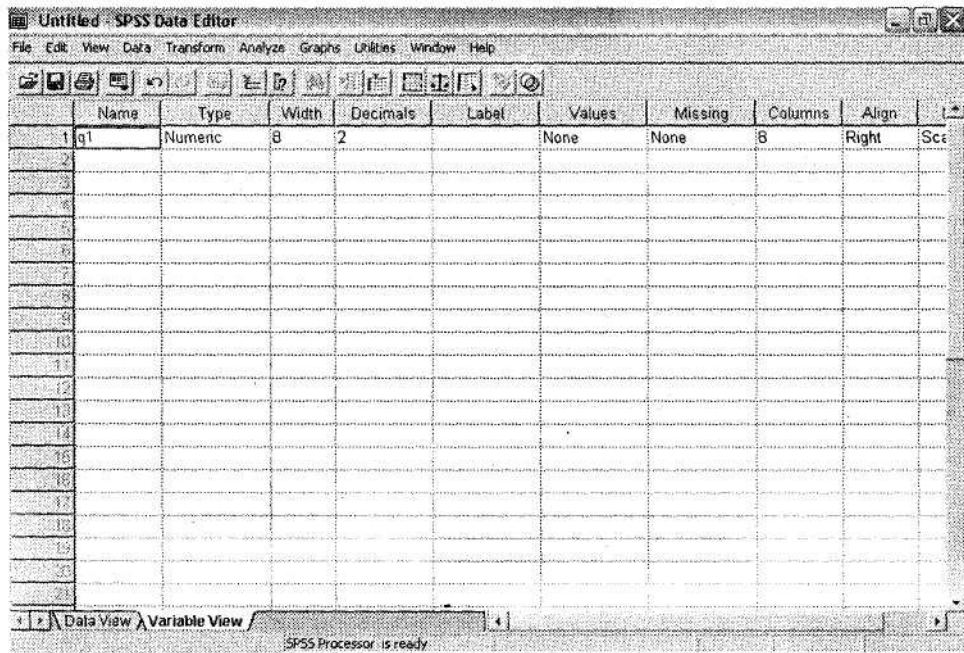


Рис. 1.12. Окно SPSS Data Editor, вкладка Variable View

Если в данную таблицу ввести какую-либо переменную (поле Name), все остальные ее поля будут заполнены автоматически значениями по умолчанию. После импорта данных из другой программы все полученные переменные будут представлены также значениями по умолчанию (сохранятся только имена переменных). Рассмотрим более детально структуру таблицы Variable View.

Первое поле таблицы Name предназначено для ввода имени переменной, которое должно состоять только из латинских букв и цифр; имя переменной не может начинаться с цифры. При импорте данных из другого источника данное поле заполняется теми значениями, которые были указаны в исходной базе данных. Все остальные поля рассматриваемой таблицы заполняются программой автоматически, причем SPSS сама определяет, к какому типу относится та или иная переменная, а в качестве меток дублирует имена переменных.

Поле Type служит для указания типа переменной. Установленный по умолчанию тип Numeric можно изменить, установив курсор в данную ячейку и щелкнув на появившейся кнопке со значком Доступные типы переменных представлены на рис. 1.13. Для некоторых из них (например, Numeric) необходимо задать количество используемых разрядов (или букв — для текстовых переменных) и цифр после запятой, а для других (например, Date) — шаблон, по которому отражаются значения.

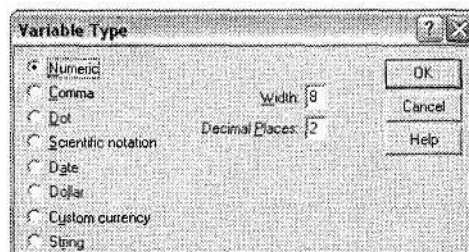


Рис. 1.13. Диалоговое окно Variable Type

Поле Width служит для указания количества разрядов (для числовых переменных) или букв (для текстовых переменных), если они не были указаны в диалоговом окне указания типа переменной. Следующее поле Decimals позволяет указать количество цифр по-

сле запятой для числовых переменных.

Поле Label служит для задания метки переменной. Данное поле важно, так как именно указанные в нем значения появляются на графиках и в таблицах при проведении всех видов статистического анализа. В анкетах, используемых при проведении маркетинговых исследований, содержатся как одновариантные вопросы (респонденты могут указать только один вариант ответа), так и многовариантные (респонденты могут указать несколько вариантов ответа). При этом если одновариантные вопросы обычно представляются одной переменной, которая может принимать столько значений, сколько имеется вариантов ответа, то многовариантные вопросы, как правило, кодируются количеством одновариантных переменных, равным числу вариантов ответа. Каждая такая одновариантная переменная всегда принимает только два значения (дихотомии) — отмечено/не отмечено, которые кодируются соответственно двумя цифрами (обычно 1 и 0). Более подробно схема работы с многовариантными переменными описана в разделе 2.2, мы отметим лишь способ кодирования различных переменных.

Так, при кодировании одновариантных переменных поле Label используется для указания формулировки вопроса анкеты (варианты ответа кодируются в другом поле). При кодировании многовариантных переменных, представленных вариантами ответа, формулировка самого вопроса не отражается в рассматриваемой таблице: кодируются только варианты ответа (дихотомические переменные).

Приведем пример. У нас есть одновариантный вопрос Укажите пол респондента — это формулировка данного вопроса, и она отражается в поле Label, а переменной присваивается имя по принципу q1. Формулировка многовариантного вопроса Что для Вас наиболее важно при выборе велосипеда? не будет фигурировать в таблице Variable View. Вместо нее будет указан набор одновариантных дихотомических переменных (по числу вариантов ответа). В поле Label будут указаны названия вариантов ответа, а в поле Name — имена переменных, кодирующие каждый из вариантов ответа (например, переменная q2_1 — Цена велосипеда; q2_2 — Качество велосипеда и т. д.).

Поле Values предназначено для указания вариантов ответа в одновариантных вопросах. Общий вид соответствующего диалогового окна представлен на рис. 1.14. Данное поле не заполняется для многовариантных переменных. В окне Value Labels в поле Value указываются числовые коды вариантов ответа, а в поле Value Label — вербальные формулировки вариантов ответа. При задании меток необходимо предлагать разумные варианты ответов, учитывая, что впоследствии именно эти названия (в том же виде) будут фигурировать на графиках и в аналитических таблицах. Например, вариант ответа на вопрос о половой принадлежности респондента следует называть не Мужской или Женский, а Мужчины или Женщины. Также при наименовании переменных и вариантов ответа следует избавляться от лишних слов, как то: предлоги в начале предложения, междометия, вводные слова. Это, с одной стороны, позволит сократить само название, что в дальнейшем облегчит его восприятие, а с другой стороны, избавит таблицы и диаграммы от массы ненужной информации. Итак, наша основная рекомендация при наименовании переменных — формализация названий.

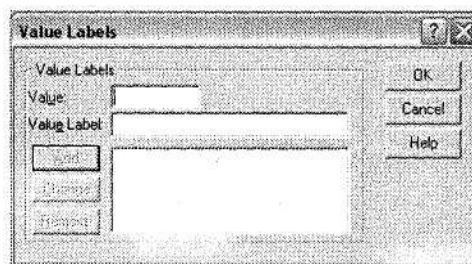


Рис. 1.14. Диалоговое окно Value Labels

Поле Missing используется редко, так как не несет существенной смысловой нагрузки. В нем можно указать, какие коды следует исключить из анализа (присвоить им статус System Missing). По умолчанию все отсутствующие значения (пропущенные одно-вариантные вопросы или неотмеченные варианты ответа многовариантных вопросов) представляются в SPSS как System Missing и отражаются для числовых переменных символом „.

Также при помощи поля Missing можно наглядно продемонстрировать разницу между различными типами пропущенных значений — типа «user missing» (значения, специально пропущенные исследователем) и типа «system missing» (значения, которые в принципе должны были присутствовать, но которых не оказалось в базе данных в связи с причинами случайного характера, — в том числе и динамически, не меняя структуры базы данных. Предположим, что для исследования нам нужны только люди с доходом свыше \$ 500. Тогда в начале анкеты мы зададим респондентам фильтрационный вопрос (закрытый): Укажите Ваш примерный среднемесячный доход в расчете на 1 члена семьи. При этом респондент может выбрать один из пяти вариантов ответа:

1. до \$500;
2. от \$ 500 до \$ 1000;
3. от \$1000 до \$1500;
4. свыше \$1500;
5. отказываюсь отвечать.

Очевидно, что для дальнейшего анализа нам подходят только те респонденты, которые указали варианты ответа 2-4. Теперь эти три варианта ответа, которые необходимы нам для построения линейных и перекрестных распределений, мы заносим в поле Values, а оставшиеся два — 1 и 5 — в поле Missing. Два последние варианта исключаются из дальнейшего анализа и будут представляться как значение System Missing. Впоследствии, если мы захотим, например, построить общее линейное распределение по всему фильтрационному вопросу (включая все категории), нужно будет просто убрать два пропущенных (в терминологии SPSS — User Missing) значения из поля Missing и добавить их в поле Values. Поле Columns служит для указания ширины столбца при отображении переменной в окне Data View. Следующее поле Align предназначено для выбора выравнивания значений переменной в столбце: по правому краю (Right), по левому краю (Left) или по центру (Center).

Поле Measure является для SPSS единственной возможностью определить тип шкалы имеющихся переменных: номинальная (Nominal), порядковая (Ordinal) или интервальная (Scale). Как показано далее в разделе 2.5 «Статистический анализ данных», важно знать, к какому типу шкалы относится та или иная переменная в базе данных. От этого во многом зависит выбор используемой статистической процедуры. Ниже приведена краткая характеристика трех типов шкалы переменных, используемых в SPSS.

1. **Номинальные переменные (Nominal)** могут принимать дискретные, не связанные друг с другом значения. Вопросы анкеты, кодируемые номинальными переменными, могут быть как закрытыми (с вариантами ответов), так и открытыми (с текстовым полем вместо прямого указания вариантов ответа). Например, вопрос анкеты Каких производителей мясных полуфабрикатов Вы знаете? с вариантами ответа Царицыно, Черкизовский, Браво и Другое будет закодирован в базе данных SPSS номинальной переменной, так как между вариантами ответа на данный вопрос не существует логического порядка, это просто названия компаний-производителей.

2. Особое место среди номинальных переменных занимают переменные, являющиеся вариантами ответа на многовариантные вопросы или имеющие только два варианта ответа. Тип шкалы данных переменных называется **дихотомическим (Dichotomous)**. Данным переменным в SPSS отводится особая роль, так

как их варианты ответа могут рассматриваться в статистических процедурах как вероятность выбора одной категории или не выбора другой. В качестве вопросов анкеты

дихотомические переменные могут кодировать как открытые, так и закрытые вопросы.

3. **Порядковые переменные (Ordinal)** кодируют такие закрытые вопросы, варианты ответа на которые подчиняются логическому числовому порядку. То есть варианты ответа на такие вопросы представляют собой связанные между собой группы значений. Например, вопрос Как часто Вы покупаете мясные полуфабрикаты? с вариантами ответа: Чаще раза в неделю, Примерно раз в неделю и Реже раза в неделю — кодируется переменной с порядковой шкалой.

4. **Интервальными (Scale)** являются переменные, не имеющие выделенных категорий. Они содержат числовые данные (например, номер анкеты в базе данных) и кодируют чаще всего открытые вопросы. Интервальные переменные (или другие типы переменных, приводимые к интервальному виду) используются практически во всех статистических процедурах. Они являются основным ресурсом для SPSS.

1.5. Модификация и отбор данных

Этап модификации и отбора данных объединяет целый ряд процедур, используемых для манипуляции с имеющимися данными: условный отбор данных, формирование случайной выборки, сортировка данных, перекодирование переменных, вычисление новых переменных и т. д. В настоящем разделе мы рассмотрим наиболее часто используемые методы автоматизированного управления переменными и их значениями в базах данных SPSS.

1.5.1. Условный отбор данных и случайная выборка

В настоящем параграфе мы рассмотрим такие методы манипуляций с данными, как отбор респондентов по определенному условию (например, выбор из всей базы данных только анкет мужчин), а также формирование случайной выборки.

1.5.1.1. Отбор анкет по условию

Часто при анализе данных в SPSS возникает необходимость отбора только тех респондентов, которые соответствуют определенным требованиям (например, имеют среднемесячный доход свыше \$ 1000). В этом случае используют условный отбор данных. Соответствующее диалоговое окно вызывается при помощи меню Data ► Select Cases.

Как вы видите на рис. 1.15, это диалоговое окно не только позволяет осуществлять условный отбор данных, но и разрешает многие другие манипуляции. При проведении маркетинговых исследований наиболее часто применяются только два параметра: If condition is specified (Условный отбор данных) и Random sample of cases (Формирование случайной выборки). По умолчанию установлен параметр All cases, что означает выбор всех без исключения респондентов.

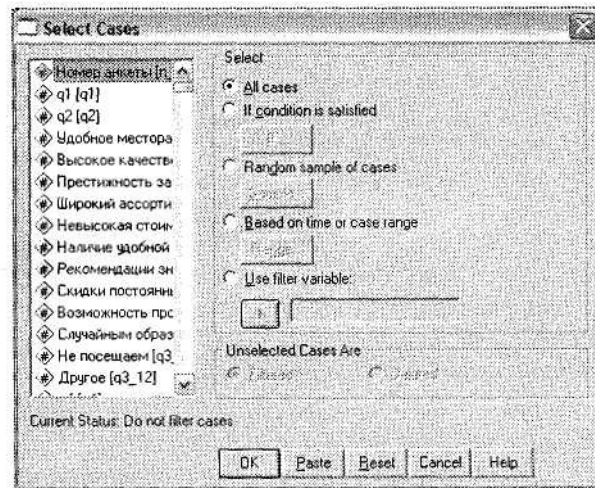


Рис. 1.15. Диалоговое окно Select Cases

Выберите параметр If condition is specified и щелкните на кнопке If. Откроется новое диалоговое окно Select Cases: If, позволяющее задать условие, согласно которому будет производиться отбор респондентов (рис. 1.16). Основная рекомендация относительно работы с данным диалоговым окном — заключайте все уравнения (название переменной и ее значение) в круглые скобки. Соблюдение данного требования весьма полезно при составлении длинных последовательностей условий.

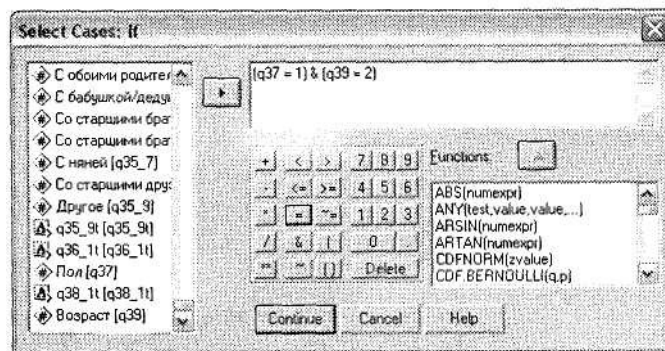


Рис.1.16. Диалоговое окно Select Cases: If

В табл. 1.3 представлена расшифровка всех логических и арифметических операторов, используемых при составлении условных выражений. Такие же операнды используются и в других диалоговых окнах, описываемых в разделе 1.5. Это стандартные операнды для составления логических выражений.

Необходимо отметить, что все логические операторы, кроме = и \neq , применимы только для числовых переменных (не для текстовых).

Помимо представленных стандартных логических операторов, существуют специальные предустановленные функции (область Functions) — при щелчке правой кнопкой мыши на любой из них появляется описание соответствующей функции.

Таблица 1.3. Стандартные логические операторы, используемые в SPSS

Арифметические		Логические	
Оператор	Значение	Оператор	Значение
+	Сложение ($x + y$)	<	меньше ($x < y$)
-	вычитание ($x - y$)	>	больше ($x > y$)
*	умножение ($x * y$)	<=	меньше или равно ($x \leq y$)
/	деление (x / y)	>=	больше или равно ($x \geq y$)
**	возведение в степень ($x ** y$)	=	равно ($x = y$)

()	приоритет вычислений	\sim	не равно ($x \sim y$)
	или ($x y$)	$\&$	и ($x \& y$)
\sim	отрицание ($\sim x$)		

В приведенном примере мы выбрали все анкеты, полученные от респондентов, являющихся мужчинами (вопрос q37, вариант ответа 1) в возрасте от 26 до 30 лет (вопрос q39, вариант ответа 2). Щелкнув на кнопке Continue и завершив операцию при помощи щелчка на кнопке ОК в главном диалоговом окне, мы увидим, что респонденты, не соответствующие данному условию, оказались исключенными из рассмотрения (их номера перечеркнуты). Можно не только временно исключить из рассмотрения респондентов, не подходящих под определенное условие, но и полностью удалить такие нерелевантные анкеты из базы данных SPSS. Для этого в диалоговом окне Select cases (рис. 1.15) необходимо заменить выбранный по умолчанию параметр Filtered (в области Unselected Cases Are) на Deleted.

1.5.1.2. Отбор анкет случайным образом

Иногда при обработке данных маркетинговых исследований возникает необходимость отбора респондентов не по конкретному условию, а случайным образом (то есть формирование случайной выборки). Эта возможность весьма полезна для уменьшения размера исходной выборки — например, для выполнения статистических процедур, предъявляющих повышенные требования к вычислительным ресурсам компьютера. Также случайная выборка применяется при проверке корректности работы некоторых статистических процедур (например, факторного анализа): сначала процедура проводится для всей выборки, а затем — для случайной выборки из n -го количества респондентов.

Для формирования случайных выборок в диалоговом окне Select Cases, (см. рис. 1.15) предусмотрен параметр Random sample of cases. Выберите этот параметр и щелкните на кнопке Sample. Открывшееся диалоговое окно (рис. 1.17) содержит два способа формирования случайной выборки: с указанием доли респондентов, которых необходимо отобрать из исходной выборки (Approximately), либо с указанием конкретного количества респондентов, которое необходимо отобрать (Exactly). При этом в последнем случае необходимо также указать в поле from the first ... cases количество респондентов, из которого следует осуществить выбор. Для формирования случайной выборки из общего числа опрошенных в данном поле следует указать совокупный размер выборки.

В нашем случае мы случайным образом отобрали 50 % респондентов из исходной выборки.

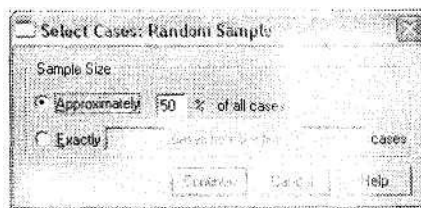


Рис. 1.17. Диалоговое окно Select Cases: Random Sample

1.5.2. Сортировка и группировка данных

Сортировка и группировка данных — наиболее часто применяющиеся операции с данными. Причем эти операции могут производиться как перед началом проведения статистического анализа, так и на других этапах работы.

1.5.2.1. Сортировка файла данных SPSS

При помощи функции сортировки в SPSS можно упорядочить значения переменных по одному или нескольким ключевым полям анкеты. Вызов диалогового окна сорти-

ровки осуществляется последовательностью меню Data ► Sort Cases.

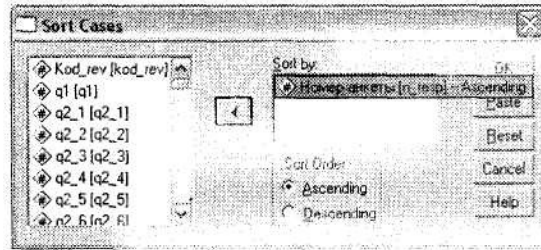


Рис. 1.18. Диалоговое окно Sort Cases

Как указано на рис. 1.18, левый список содержит все доступные в текущей базе данных переменные. В область Sort by помещаются переменные, по которым следует произвести сортировку. Порядок следования переменных в данной области соответствует порядку сортировки, то есть сначала сортировка происходит по первой переменной, затем — по второй и т. д. Группа переключателей Sort Order позволяет выбрать направление сортировки: по возрастанию (Ascending) или убыванию (Descending). При этом для каждой переменной можно выбрать свой тип сортировки.

В нашем случае мы отсортировали базу данных по возрастанию номера анкеты.

1.5.2.2. Группировка значений переменных

SPSS позволяет автоматически разделять значения интервальных переменных на заданное число групп. Разделение производится на основании процентилей, то есть образующиеся группы содержат примерно одинаковое количество значений. Результатом работы этой процедуры является новая порядковая переменная, которая содержит столько категорий, сколько было указано групп. Диалоговое окно группировки данных вызывается при помощи меню Transform ► Categorize Variables (рис. 1.19). В область Create Categories for переносятся переменные, значения которых необходимо сгруппировать. Поле Number of categories служит для указания числа групп.

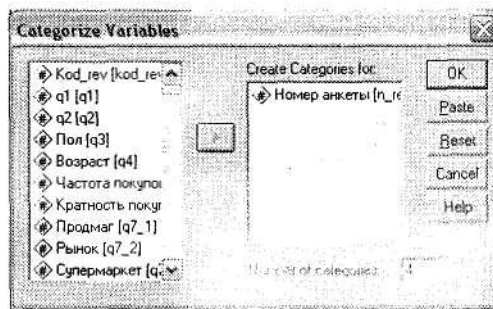


Рис. 1.19. Диалоговое окно Categorize Variables

В нашем примере мы разделили выборку по номеру анкеты на четыре примерно равных доли — по 25 %.

1.5.3. Перекодирование переменных

Перекодирование переменных служит для трансформации значений переменных с созданием или без создания новых переменных, а также для автоматического кодирования текстовых переменных для преобразования их к числовому виду.

1.5.3.1. Перекодирование внутри одной переменной

Рекомендуется производить перекодирование значений многовариантных переменных (точнее, наборов дихотомий, как было показано в разделе 1.4.2) сразу после создания базы данных. При этом все пропущенные значения (вариант не отмечено) в таких вопросах следует перекодировать из System Missing в нули. В дальнейшем это позволит использовать данные дихотомические переменные (уже с двумя вариантами ответа: 0 и 1) при проведении статистического анализа (например, при построении перекрестных распределений). Альтернативой обработки многовариантных переменных является формирование серии полноценных одновариантных переменных путем кодирования всех возможных взаимодействий между вариантами ответа на многовариантный вопрос. Очевидно, что такая методика подходит только для вопросов с небольшим количеством вариантов ответа.

Перекодирование может осуществляться как внутри одной уже существующей переменной, так и с созданием новой переменной, содержащей перекодированные значения. В последнем случае исходная переменная будет содержать неперекодированные значения, а вновь созданная — перекодированные значения.

Рассмотрим методику перекодирования внутри одной существующей переменной (без создания новой). В качестве примера возьмем случай с многовариантным вопросом **Где Вы обычно покупаете кетчуп?**, у которого есть четыре варианта ответа:

1. q2_1 — рынки;
2. q2_2 — магазины;
3. q2_3 — палатки;
4. q2_4 — другое.

При этом выбор респондентом данных пунктов закодирован в базе данных как 1, а отсутствие выбора — значением System Missing (отражается символом,). Произведем перекодирование отсутствующих значений System Missing в нули.

Вызов диалогового окна перекодировки внутри одной переменной осуществляется при помощи меню **Transform ► Recode ► Into Same Variables**. Открывшееся диалоговое окно, как и большинство других окон SPSS, в левой области содержит список всех доступных переменных, а в правой (имеющей метку **Variables**) — место для помещения перекодируемых переменных. Необходимо особо подчеркнуть, что за один цикл использования диалогового окна **Recode into Same Variables** можно перекодировать сколько угодно переменных, но только одними и теми же кодами. Иными словами, нельзя в одной переменной нули заменить на единицы, а в другой — шестерки на строки Шесть. Для этого придется сначала перекодировать первую переменную (нули на единицы), а затем вновь вернуться в диалоговое окно **Recode into Same Variables**, щелкнуть на кнопке **Reset** и затем ввести данные для перекодировки другой переменной.

В нашем случае мы собираемся перекодировать четыре переменные, имеющие одинаковые унарные шкалы, состоящие всего из одного значения 1. Поэтому в описываемом диалоговом окне можно ввести их одновременно в область **Variables** (рис. 1.20).

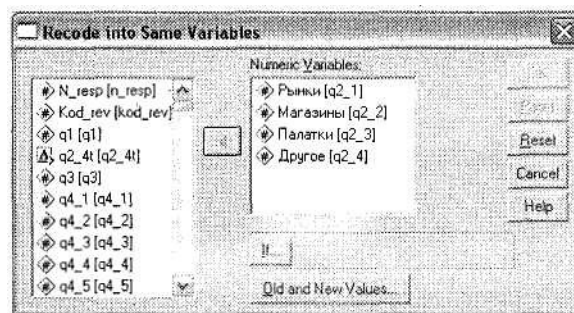


Рис. 1.20. Диалоговое окно **Recode into Same Variables**

При щелчке на кнопке If вызывается диалоговое окно, по внешнему виду и по функциям аналогичное окну Select Cases: If, представленному на рис. 1.16. Из этого окна можно производить перекодирование переменных, помещенных в область Variables, не для всех респондентов, а только для конкретных групп (например, только для мужчин).

В нашем случае мы не будем ставить никаких условий. Щелкните на кнопке Old and New Values, которая открывает диалоговое окно, позволяющее задать перекодируемые значения (рис. 1.21). Это окно разделено на две части. В левой можно указать, какие конкретно значения подлежат перекодировке, а в правой — в какие значения они будут перекодированы. Чтобы указать конкретное значение для перекодировки, введите исходное значение в левое поле Value, а конечное значение — в правое поле Value.

Для специальных значений System Missing есть специальный одноименный параметр. В нашем примере в левой области диалогового окна выберите пункт System Missing, а в правой — в поле Value введите 0. Далее щелкните на кнопке Add, чтобы добавить указанное сочетание в список перекодировки. (Необходимо особо отметить, что значения, не указанные в списке перекодировки, остаются неизменными.)

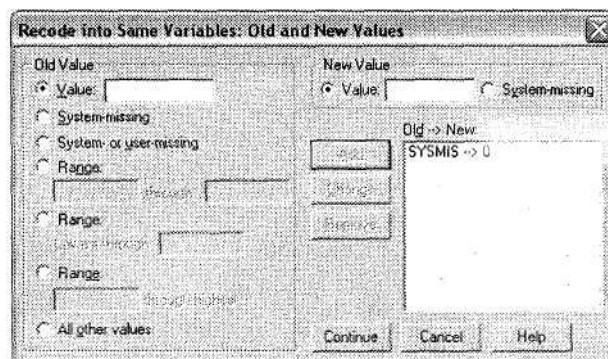


Рис. 1.21. Диалоговое окно Old and New Values

После того как были указаны все необходимые варианты перекодирования (в нашем случае — только один), следует закрыть окно щелчком на кнопке Continue и запустить процедуру перекодирования кнопкой OK. В исходной базе данных SPSS все значения System Missing в переменных q2_1 - q2_4 будут перекодированы в нули, единицы при этом сохраняются.

1.5.3.2. Перекодирование с образованием новых переменных

Рассмотрим теперь другой случай перекодирования переменных, в результате которого исходная переменная остается неизменной, а перекодированные значения отражаются в новой переменной. Данная процедура осуществляется при помощи меню Transform ► Recode ► Into Different Variables. Диалоговое окно Recode into Different Variables (рис. 1.22) аналогично окну Recode into Same Variables (рис. 1.20), только добавлена дополнительная область Output Variable, предназначенная для указания имени (Name) и метки (Label) вновь создаваемой переменной, которая будет содержать перекодированные значения.

В качестве примера мы взяли переменную q16, содержащую ответы на вопрос относительно частоты покупок респондентами плавленого сыра. При этом опрошенные должны были выбрать один из восьми вариантов:

1. каждый день;
2. 3-4 раза в неделю;
3. 1-2 раза в неделю;
4. 1-2 раза в месяц;
5. реже 1 раза в месяц;
6. 1 раз в полгода;

7. 1 раз в год;
8. затрудняюсь ответить.

После перекодирования мы должны получить переменную `q16_rec`, в которой интервалы 1, 2 и 3 будут объединены в группу с кодом 1 (Частые покупатели); интервалы 4, 5, 6 и 7 — в группу с кодом 2 (Редкие покупатели); а интервал 8 — в значения System Missing.

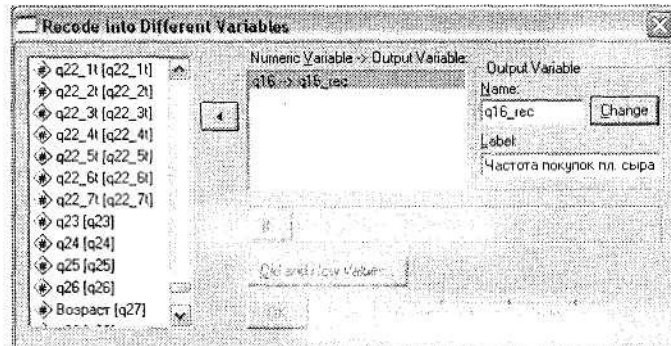


Рис. 1.22. Диалоговое окно Recode into Different Variables

Введите в соответствующие поля название и метку новой переменной. Обратите внимание, что в описываемом диалоговом окне также есть кнопка условного отбора данных If. Откройте диалоговое окно Old and New Values, щелкнув на одноименной кнопке (рис. 1.23).

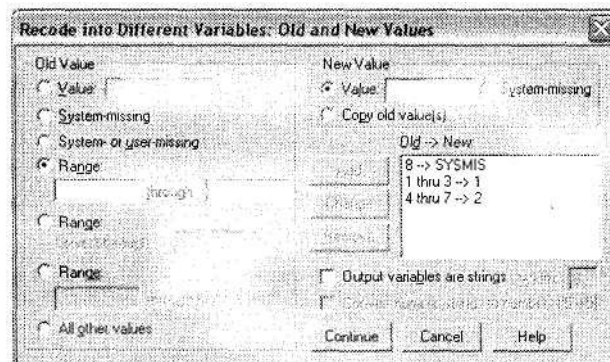


Рис. 1.23. Диалоговое окно Old and New Values

Это окно напоминает окно, представленное на рис. 1.21, но в нем также содержатся некоторые дополнительные полезные инструменты. По умолчанию значения исходной переменной, не указанные в списке перекодировки, не попадают в новую переменную. Изменить данное условие по умолчанию можно при помощи параметра Copy old value(s). Также появилась возможность конвертации числовых значений в строковые (параметр Output variables are strings). При этом изменится тип всей новой переменной; следовательно, все исходные значения должны быть перекодированы как

строковые. Существует и обратная возможность — конвертации строковых значений, похожих на цифры, в числовой вид (например, «5» в 5). Данная возможность реализуется при помощи параметра Convert numeric strings to numbers.

В нашем случае мы при помощи параметра Range перекодировали значения исходной переменной — от 1 до 3 — в 1, от 4 до 7 — в 2, а значение 8 — в System Missing. После щелчков в соответствующих диалоговых окнах на кнопках Continue и ОК будет создана новая переменная `q16_rec`, содержащая перекодированные по указанной схеме значения переменной `q16`.

1.5.3.3. Автоматическое перекодирование

Данная процедура предназначена для автоматического кодирования полей анкеты числовыми значениями типа индекс. В маркетинговых исследованиях эта процедура применяется в основном для текстовых полей в тех случаях, когда в анкете есть либо открытые вопросы (являющиеся текстовыми переменными в базе данных), либо варианты ответа Другое с дополнительным полем для указания респондентом конкретного варианта.

При выполнении процедуры одинаковые ответы из текстовых полей группируются, и им присваиваются соответствующие коды ответа (например, начиная с 1). Для того чтобы автоматическое перекодирование имело практический смысл, необходимо жестко формализовать ответы респондентов в текстовых полях. Если при заполнении анкет допускалась свободная формулировка респондентами своих ответов, следует перед вводом анкет в компьютер (или на этапе ввода) переформулировать их в соответствии с требованиями формализации. Меньшее количество различных вариантов ответа на открытый вопрос является предпочтительным, так как в дальнейшем при построении распределений большое число категорий трудно читается на графиках и в таблицах. Еще одно существенное требование к ответам респондентов на открытые вопросы — это достаточное количество респондентов в каждой группе ответов. Варианты ответов, указанные малым числом опрошенных, обычно относятся к варианту Другое.

Диалоговое окно Automatic Recode (рис. 1.24) вызывается при помощи меню Transform ► Automatic Recode. В нашем примере мы задавали респондентам вопрос Какие марки глазированных сырков Вы знаете?. После списка основных конкурентов на данном рынке в анкете был вариант ответа Другое (переменная q9_13t), в который записывались компании-производители, не вошедшие в данный перечень. Закодируем эти марки числовыми значениями (вместо текстовых полей). Для этого следует перенести из левого списка всех доступных переменных интересующую нас текстовую переменную q9_13t в область Variable ► New Name и в соответствующем поле указать новое имя вновь создаваемой числовой переменной q9_13t_n. Затем, чтобы подтвердить преобразование, необходимо щелкнуть на кнопке New Name. В группе переключателей Recode Starting from есть два параметра, позволяющие присвоить номера вариантам ответа либо по алфавиту, начиная с самого малого значения (Lowest value), либо начиная с конца упорядоченного списка вариантов ответа (Highest value).

После щелчка на кнопке ОК и выполнения указанных преобразований в базе данных будет создана новая числовая переменная (q9_13t_n) с вариантами ответа согласно списку перекодировки. Список также выводится SPSS (в окне SPSS Viewer), он показан на рис. 1.25.

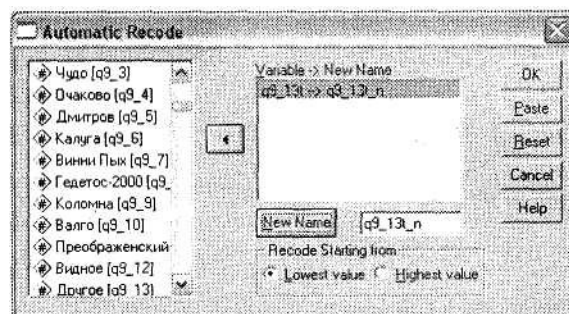


Рис. 1.24. Диалоговое окно Automatic Recode

Q9_13T	Q9_13T_N	Q9_13t
Old Value	New Value	Value Label
БИТЦЕВСКИЕ	1	БИТЦЕВСКИЕ
ВАНИЛЬНЫЕ	2	ВАНИЛЬНЫЕ
ДОМИК В ДЕРЕВНЕ	3	ДОМИК В ДЕРЕВНЕ
ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	4	ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
ИВАНЬКОВСКИЙ ЗАВОД	5	ИВАНЬКОВСКИЙ ЗАВОД
КУТЕЖ	6	КУТЕЖ
ЛЮБЫЕ С МАКОМ	7	ЛЮБЫЕ С МАКОМ
НАСТЕНЬКА	8	НАСТЕНЬКА
НОГИНСКОГО ЗАВОДА	9	НОГИНСКОГО ЗАВОДА
ПУШКИНСКИЕ	10	ПУШКИНСКИЕ

Рис. 1.25. Список перекодировки

Как видно на рисунке, список разделен на три части: слева находятся значения исходной переменной (q9_13t); в среднем столбце расположены коды, под которыми данные текстовые значения представляются в новой переменной (q9_13t_n); правый столбец дублирует левый. Теперь по вновь созданной числовой переменной можно строить графики, а также использовать ее в других статистических процедурах.

1.5.4. Вычисление новых переменных

Вычисление новых переменных — весьма полезная возможность SPSS. При помощи данной функции можно производить расчеты по формулам любой сложности, задаваемым пользователем.

1.5.4.1. Вычисление новых переменных

Кроме перекодирования переменных, SPSS позволяет создавать новые переменные, содержащие либо совершенно новые значения, либо значения, вычисленные на основании существующих переменных. Таким образом действует процедура Compute Variable, вызываемая при помощи меню Transform ► Compute (рис. 1.26).

В качестве примера мы рассчитаем годовой объем закупок сметаны на основании имеющихся данных о частоте покупок данного продукта в месяц (интервальная переменная q5) и кратности покупок (интервальная переменная q6).

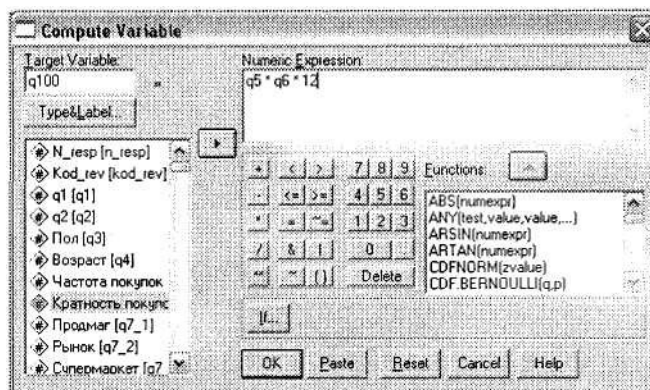


Рис. 1.26. Диалоговое окно Compute Variable

В поле Target Variable мы указали имя вновь создаваемой переменной, которая будет содержать вычисленные для каждого респондента годовые объемы покупок сметаны. Далее щелкните на кнопке Type&Label и укажите метку и ее тип (рис. 1.27). В нашем случае в качестве метки в поле Label мы указали Годовой объем закупок сметаны. Новая переменная будет содержать числовые значения, поэтому мы выбрали тип Numeric.

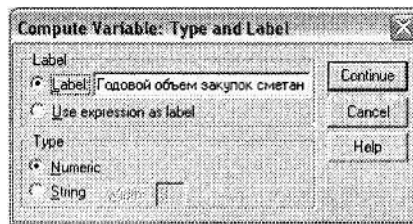


Рис. 1.27. Диалоговое окно Type and Label

После определения новой переменной в области Numeric Expression следует указать непосредственно рассчитываемое выражение. В нашем случае мы умножаем частоту покупок (q5) на кратность покупок (q6) и затем умножаем на 12 месяцев, чтобы получить объем покупок сметаны в год. После запуска процедуры вычисления новой переменной будет создана новая переменная q100, содержащая годовые объемы покупок сметаны каждым респондентом в выборке.

1.5.4.2. Подсчет значений переменных

Еще одной полезной возможностью SPSS, не рассмотренной при описании процесса модификации и отбора данных, является подсчет значений переменных (как правило, многовариантных).

Приведем пример. Предположим, у нас есть ответы респондентов на многовариантный вопрос Из каких источников Вы получаете информацию о рынке сантехники? с пятью вариантами ответа:

1. q22_1 - газеты;
2. q22_2 — журналы;
3. q22_3 — выставки;
4. q22_4 — Интернет;
5. q22_5 — другие источники.

В результате работы описываемой процедуры мы получим новую переменную q100, в которой для каждого респондента в выборке будет отражаться количество используемых источников при поиске информации о рынке сантехники.

Диалоговое окно Count Occurrences of Values within Cases, позволяющее выполнить поставленную задачу, открывается при помощи меню Transform ► Count (рис. 1.28). В полях Target Variable и Target Label следует указать соответственно имя вновь создаваемой переменной (q100) и ее метку (Количество используемых источников). В область Numeric Variables помещаются интересующие нас переменные q22_1 - q22_5, значения которых необходимо подсчитать.

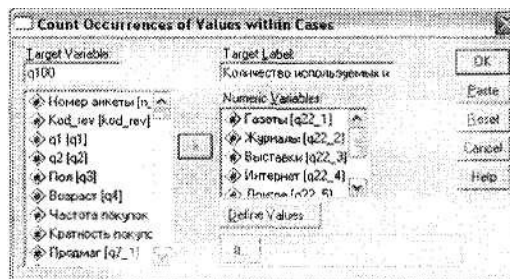


Рис. 1.28. Диалоговое окно Count Occurrences of Values within Cases

Диалоговое окно Count Occurrences of Values within Cases так же, как и многие другие окна SPSS, содержит кнопку If, позволяющую осуществить расчеты не для всех респондентов в выборке, а только для отдельных групп.

Щелкните на кнопке Define Values. Открывшееся диалоговое окно (рис. 1.29) предназначено для указания конкретных значений рассматриваемых переменных, подлежащих подсчету. Так как у нас есть пять дихотомических переменных, соответствующих вариантам ответа на многовариантный вопрос, мы указали 1 в качестве объекта подсчетов.

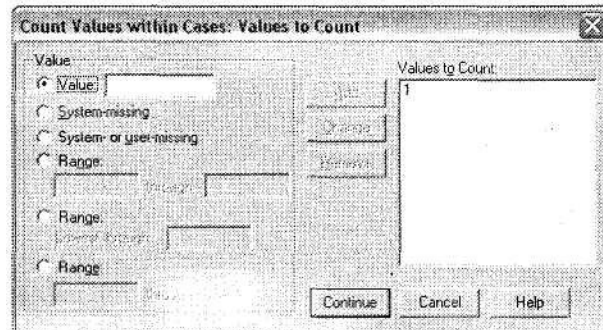


Рис. 1.29. Диалоговое окно Values to Count

Запустив процедуру, мы получим новую переменную с результатами расчетов. В дальнейшем мы можем построить по данной переменной линейное распределение (см. раздел 2), чтобы узнать, сколько респондентов используют при поиске сантехники только один, два, три, четыре или пять источников информации.

1.5.5. Коррекция нерепрезентативности выборки

В практике маркетинговых исследований случается, что собранные в ходе опроса данные не соответствуют параметрам генеральной совокупности (то есть являются нерепрезентативными). Такие ситуации возникают, если заложенные перед началом исследования социально-демографические квоты были искажены в результате нарушения методологии проведения исследования, ошибок в работе интервьюеров или недостаточного контроля проведения полевых работ.

Например, в результате проведения контрольных мероприятий после завершения основных полевых работ были выявлены многочисленные факты некорректного заполнения анкет интервьюерами или даже фальсификация анкет, вследствие чего из итоговой базы данных пришлось удалить некоторую часть анкет. Очевидно, что в этом случае социально-демографические квоты, заложенные в начале исследования и обеспечивающие соответствие параметров выборки параметрам общей генеральной совокупности (репрезентативность), скорее всего, изменятся. Это в свою очередь приведет к тому, что выводы, основанные на результатах проведенного опроса, не могут быть отнесены к генеральной совокупности. То есть мы не можем утверждать, что наши выводы действительно отражают мнение реальных потребителей. Исследование фактически теряет свой смысл.

Если полученная выборка является нерепрезентативной, применяется метод коррекции параметров выборки путем взвешивания. Приведем пример. Известно, что доля мужчин всего населения России составляет 45,5 %. В результате проведения всероссийского исследования оказалось, что доля мужчин в выборке составляет 72,1 %. Следовательно, полученная выборка является нерепрезентативной. Для устранения ошибки следует провести взвешивание, то есть скорректировать полученные значения переменной Пол (dl) на весовой коэффициент. Данный коэффициент рассчитывается для каждой социально-демографической группы по следующей формуле:

$$\Delta = \frac{\delta}{\delta'}$$

где Δ — весовой коэффициент; δ — значение исследуемого параметра в генеральной совокупности; δ' — значение исследуемого параметра в выборке.

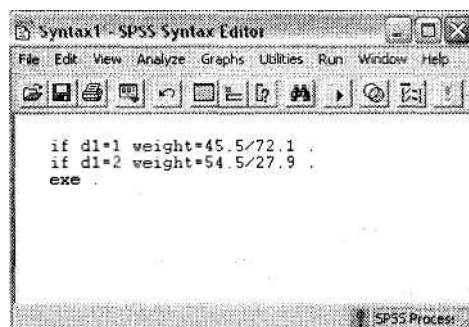
В нашем случае весовой коэффициент должен рассчитываться для двух социально-демографических групп: мужчин и женщин. (Если у вас большое количество групп, весовой коэффициент следует рассчитывать для каждой группы.) Для мужчин А будет равна: $45,5 \% / 72,1 \% \approx 0,63$. Так как у нас всего две группы, подлежащие взвешиванию (мужчины и женщины), то весовой коэффициент для женщин будет рассчитан так: $(100 \% - 45,5 \%) / (100 \% - 72,1 \%) = 54,5 \% / 27,9 \approx 1,95$. (Если у вас большое количество групп, подлежащих взвешиванию, вам нужно знать значения параметров генеральной совокупности для каждой из групп.)

Итак, на первом этапе мы получили весовые коэффициенты, которые помогут нам скорректировать полученную нерепрезентативную выборку. Теперь необходимо создать новую переменную в файле данных SPSS, которая будет содержать для каждого респондента его вес (то есть для мужчин — 0,63, а для женщин — 1,95). Проще всего перекодировать с образованием новой переменной (как было описано в разделе 1.5.3.2).

В настоящем пособии мы не описываем важный элемент SPSS — программный синтаксис. Данный элемент является альтернативой использованию диалоговых окон в SPSS. Другими словами, все то, что можно сделать при помощи мыши в диалоговых окнах (и многое другое), можно выполнить посредством программного синтаксиса. В некоторых случаях его использование является предпочтительным. В частности, в нашем примере для создания новой весовой переменной удобнее воспользоваться синтаксисом. Откройте редактор синтаксиса File ► New ► Syntax. На экране появится окно, показанное на рис. 1.30. Введем в нем следующие команды:

```
if dl=1 weight=45.5/72.1 .
if dl=2 weight=54.5/27.9 .
exe .
```

Обратите внимание, что в синтаксисе SPSS символ, отделяющий целую и дробную части числа, — всегда точка, а не запятая. Также следует внимательно относиться к точкам в конце каждой строки. Эти точки дают понять интерпретатору SPSS, что следует выполнить данную команду. Последовательность символов exe. на третьей строке запускает процедуру синтаксиса. Рекомендуется использовать не приблизительные значения весовых коэффициентов (0,63 и 1,95), а вычисляемые выражения (45.5/72.1 и 54.5/27.9); что обеспечивает точность расчетов. После того как вы введете указанные строки в редакторе синтаксиса (см. рис. 1.30), выделите их все (это очень важно) и затем нажмите Ctrl+R или на кнопке ► на панели инструментов окна синтаксиса.



В результате работы процедуры синтаксиса будет создана новая переменная weight, содержащая весовые коэффициенты для каждого респондента. Теперь осталось только провести собственно процедуру взвешивания каждого респондента на его весовой коэффициент. В этом вам поможет диалоговое окно Weight Cases (Data ► Weight Cases). В данном диалоговом окне (рис. 1.31) следует выбрать параметр Weight cases by, затем в

левом списке всех доступных переменных выбрать весовую переменную (в нашем случае weight) и перенести ее в поле Frequency variable, при щелчке на кнопке ОК база данных будет скорректирована на весовые коэффициенты, и репрезентативность данных будет восстановлена. Для отмены взвешивания следует в данном диалоговом окне установить переключатель в положение Do not weight cases.

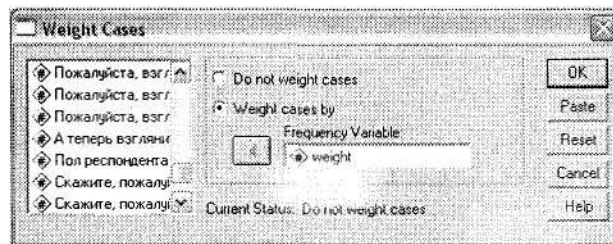


Рис. 1.31. Диалоговое окно Weight Cases

Если искажение квот в выборке произошло не только по одной социально-демографической переменной, а сразу по нескольким (например, не только по полу, но и по возрасту и уровню образования), следует сначала создать отдельные весовые переменные для каждой из искаженных социально-демографических переменных, а затем создать новую общую весовую переменную, которая будет произведением всех отдельных весовых коэффициентов (то есть для каждого респондента: вес по полу, вес по возрасту, вес по образованию).

При всей кажущейся простоте корректировки репрезентативности при помощи взвешивания следует иметь в виду, что для использования данного метода существуют серьезные ограничения. Например, часто число респондентов во взвешенной базе данных оказывается иным, чем в невзвешенной. Это происходит из-за того, что сумма весовых коэффициентов по всем респондентам не равна общему количеству респондентов. Также нужно весьма осторожно подходить к интерпретации статистических тестов по взвешенной базе. Поскольку число респондентов с определенными социально-демографическими характеристиками во взвешенной базе искусственно увеличивается (в нашем случае это доля женщин), рассчитанная статистическая значимость является некорректной. Таким образом, взвешивание рекомендуется проводить для построения общих (линейных) распределений.

Итак, в главе 1 мы подробно рассмотрели часто используемые в маркетинговых исследованиях методы манипуляции с данными. SPSS содержит массу других дополнительных возможностей, но в данном пособии мы не стали их приводить, поскольку на практике эти методы не находят широкого применения.

Глава 2 Описательный анализ и линейные распределения

Статистический анализ данных — основное предназначение SPSS (в отличие, например, от Microsoft Excel или Microsoft Access). Графическая подсистема данного программного комплекса, внешний вид создаваемых отчетов и возможности электронной таблицы оставляют желать лучшего; пользовательский интерфейс рассчитан на лиц, хорошо знакомых со статистикой. Некоторые статистические процедуры (например, множественный дисперсионный анализ по методу Фишера) вызываются исключительно при помощи программного синтаксиса (Syntax), работа с которым требует определенных навыков программирования. Но все же, несмотря на эти недостатки, в настоящее время SPSS является одной из лучших программ для проведения профессионального статистического анализа в самых различных областях человеческой деятельности: в бизнесе, психологии, медицине и т. д.

Данный раздел знакомит читателя с основными статистическими процедурами и методами статистического моделирования, наиболее часто применяемыми в маркетинговых исследованиях. Практически все описываемые статистические функции могут применяться для решения нескольких задач. В этом смысле предлагаемое общепринятое разделение методов статистического анализа на описательный анализ, анализ различий, ассоциативный и классификационный анализ весьма условно и отражает лишь общие тенденции их использования именно в маркетинговых исследованиях. Прежде чем приступить к рассмотрению статистических функций SPSS, сделаем одно существенное отступление необходимое для понимания всех последующих разделов этого пособия.

Одним из центральных понятий в статистике является *статистическая значимость* (p). Именно на основании статистической значимости в большинстве процедур SPSS проверяется практическая пригодность построенных моделей. По сути, статистическая значимость — это вероятность наступления ненаступления исследуемого события. Уровень $p \leq 0,05$ часто используется в качестве критерия установления статистической значимости. Он означает, что с вероятностью 95 % можно утверждать: исследуемое событие произошло неслучайно, то есть связано с какой-то системой. В табл. 2.1 представлен наиболее распространенный способ интерпретации различных уровней значимости в маркетинговых исследованиях.

Таблица 2.1. Интерпретация уровней значимости

Уровень статистической значимости, p	Статистическая интерпретация	Обозначение в SPSS
$p < 0,001$	Максимально значимая	***
$0,001 \leq p \leq 0,01$	Очень значимая	**
$0,01 < p \leq 0,05$	Значимая	*
$0,05 < p \leq 0,10$	Слабо значимая	
$p > 0,10$	Незначимая	

В некоторых случаях (например, t -тесты) статистическая значимость в SPSS может быть одно- (1-tailed Sig.) или двухсторонней (2-tailed Sig.). Двухсторонняя значимость показывает, отличается ли значительно среднее значение первой исследуемой переменной от среднего значения второй — без указания направления этого различия, положительно-го или отрицательного. Односторонняя значимость показывает только направление, в котором второе исследуемое среднее отличается от первого. Второй тип значимости (одно-

сторонняя) при анализе данных маркетинговых исследований используется редко, и именно двухсторонняя значимость выводится SPSS по умолчанию. Таким образом, на практике нет необходимости обращать внимание на тип значимости, выводимой SPSS: она всегда будет показывать статистическую значимость исследуемого события¹.

Целью описательного анализа является систематизация имеющихся данных. В рамках данной задачи происходит построение линейных распределений, а также характеристика переменных в различных статистических аспектах: расчет среднего, медианы, моды и т. п. Линейные (общие) распределения позволяют подсчитать количество респондентов, указавших тот или иной вариант ответа на рассматриваемый вопрос.

Построение линейных распределений обычно является первым шагом в статистическом анализе данных. При помощи линейных распределений становится возможным систематизировать ответы респондентов. В табл. 2.2 представлены основные характеристики переменных, участвующих в анализе.

Таблица 2.2. Основные характеристики переменных, участвующих в линейных распределениях

Линейные распределения

Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
-	-	Одна	Любой

2.1. Линейные распределения для одновариантных вопросов

Одновариантные вопросы являются основным ресурсом анализа при помощи SPSS. Практически все функции, реализованные в данном программном пакете, предназначены для работы только с одновариантными переменными. Анализ многовариантных переменных производится методом выделения каждого варианта ответа в отдельную одновариантную переменную и последующей работы уже с набором одновариантных переменных. Существуют табличные и графические способы построения линейных распределений по одновариантным вопросам. Ниже представлен способ, наиболее распространенный в маркетинговых исследованиях. Рассмотрим линейное распределение респондентов по возрастному признаку. Для этого предположим, что у нас есть файл данных, содержащий одновариантную переменную q4 (Возраст), имеющую порядковую шкалу, с четырьмя возможными вариантами ответа:

1. от 16 до 18 лет;
2. от 19 до 35 лет;
3. от 36 до 60 лет;
4. старше 60 лет.

Вызов диалогового окна для построения линейных распределений (также называемых частотами) осуществляется при помощи меню **Analyze ► Descriptive Statistics ► Frequencies** (рис. 2.1). В открывшемся окне в левом списке содержатся все доступные переменные, по которым можно построить линейные распределения. При помощи мыши перетащите нужные одновариантные переменные в правый список (в нашем случае — q4). При этом для анализа можно указать сразу несколько переменных.

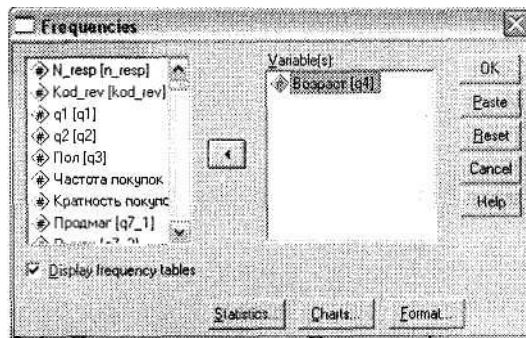


Рис. 2.1. Диалоговое окно Frequencies с выбранной для анализа переменной Возраст

В диалоговом окне Statistics, вызываемом при помощи одноименной кнопки, можно указать, какие описательные статистики, кроме относительных и абсолютных значений, необходимо рассчитать (рис. 2.2). Например, рассчитаем моду (наиболее часто встречающееся значение), выбрав соответствующий параметр. Кроме этой статистики, SPSS позволяет рассчитать другие полезные величины:

- среднее арифметическое для интервальных переменных (Mean);
- минимальное и максимальное значения (Minimum и Maximum), — а также разбить значения переменной на квантили или другие процентиля (область Percentile Values) и т. д.

Однако большинство представленных в этом диалоговом окне статистик подходит только для переменных, имеющих интервальный тип шкалы. Закрыв диалоговое окно Statistics посредством щелчка на кнопке Continue, вы вновь попадете в ос-, новное окно Frequencies.

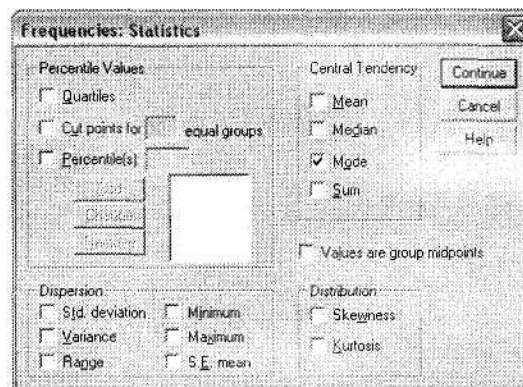


Рис. 2.2. Диалоговое окно Statistics

Необходимо сказать несколько слов относительно основных описательных статистик, показанных на рис. 2.2. Пожалуй, наиболее популярными характеристиками, используемыми для описания переменных, являются показатели группы Central Tendency (центральная тенденция): среднее арифметическое (Mean); медиана, или половина значений отрезка (Median); мода, или наиболее часто встречающееся значение (Mode); а также сумма (Sum). Имейте в виду, что данные показатели применяются неодинаково к переменным с различным типом шкалы (табл. 2.3).

Таблица 2.3. Наиболее релевантные показатели центральной тенденции для переменных с различным типом шкалы

Тип шкалы	Наиболее релевантная характеристика	Другие релевантные характеристики
Интервальная	Среднее арифметическое	Средневзвешенное, мода
Порядковая	Средневзвешенное	Мода
Номинальная	Мода	-

Из представленной таблицы видно, что наиболее релевантной описательной статистикой, характеризующей переменные с интервальной шкалой, является среднее арифметическое (Mean). Для переменных с порядковой шкалой данный показатель неприменим, так как он рассчитывается исходя из значений переменной (кодов вариантов ответа), а не самих значений интервалов.

Например, если рассчитать простое среднее по переменной Возраст (в которой возрастные группы закодированы цифрами от 1 до 4), получится 250,5 (см. рис. 2.6). Данное значение не несет в себе практически значимой нагрузки. Если же мы вместо этого рассчитаем средневзвешенное значение данной переменной по нижеприведенной формуле, мы получим реальный средний возраст респондентов: 43 года $(43 - (408 \times 48 + 321 \times 27 + 207 \times 68 + 66 \times 17) / (408 + 321 + 207 + 66))$.

$$\bar{w} = \frac{\sum_{i=1}^n f \cdot \bar{s}}{\sum_{i=1}^n f},$$

где \bar{w} — средневзвешенное значение; n — количество интервалов (вариантов ответа) в порядковой переменной; f — частота появления i -го варианта ответа; \bar{s} — среднее арифметическое значение i -го интервала.

Средняя тенденция переменных с номинальной шкалой не может быть оценена никак, кроме моды, — то есть для таких переменных можно определить только наиболее многочисленную группу. Например, по переменной Пол можно сказать, что в данном случае мужчины составляют три четверти всей выборочной совокупности респондентов.

В табл. 2.2 также видно, что интервальные переменные — наиболее гибкие относительно применения показателей центральной тенденции. Для них можно рассчитать все три рассматриваемые статистики: среднее арифметическое, средневзвешенное и моду. Порядковые переменные находятся на втором месте: с ними могут использоваться только средневзвешенное и мода. И наконец, номинальные переменные являются наименее гибкими: к ним может эффективно применяться только мода.

Теперь мы вновь возвращаемся к диалоговому окну Frequencies. Кнопка Charts вызывает одноименное диалоговое окно, которое позволяет помимо таблиц вывести диаграммы по выбранным переменным (рис. 2.3). По умолчанию SPSS не выводит диаграмм. Давайте построим круговую диаграмму (сектограмму), выбрав параметр Pie charts и указав в области Chart Values на необходимость отобразить на диаграмме не абсолютные (установлено по умолчанию), а относительные значения (Percentages). Выполнив это, закройте диалоговое окно Charts.

С помощью кнопки Format в главном диалоговом окне линейных распределений Frequencies можно указать, каким способом следует сортировать результаты в частотных

таблицах (рис. 2.4). Это можно сделать, выбрав соответствующий параметр в области Order by. При этом возможной альтернативой будет сортировка кодов вариантов ответа (в нашем случае — кодировок возрастных групп):

- по возрастанию: от 1 (16-18 лет) до 4 (старше 60 лет);
- по убыванию: от 4 до 1;
- по количеству респондентов, выбравших каждый из рассматриваемых вариантов ответа (в нашем случае — по численности четырех рассматриваемых возрастных групп).

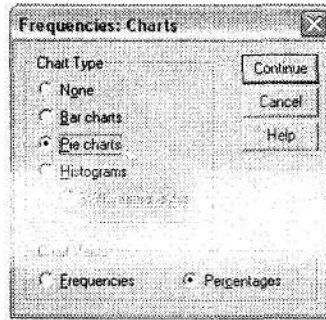


Рис. 2.3. Диалоговое окно Charts

Для иллюстрации нашего примера выберем сортировку по численности возрастных групп по убыванию Descending counts и закроем диалоговое окно Format, щелкнув на кнопке Continue.

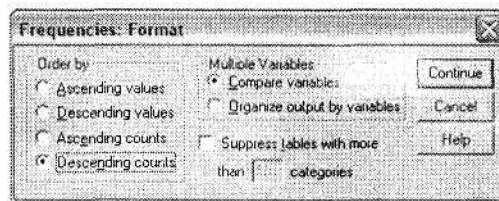


Рис. 2.4. Диалоговое окно Format

После щелчка на кнопке ОК в главном диалоговом окне Frequencies откроется окно SPSS Viewer, в котором будут представлены частотные таблицы, а также другая информация, указанная нами на подготовительном этапе.

В таблице Statistics (рис. 2.5) отражаются общие параметры линейного распределения. Здесь представлены:

- количество респондентов, ответивших на вопрос Возраст (строка Valid), — 1002 человека;
- количество анкет, в которых на данный вопрос не было получено ответа (строка Missing), — 1 человек;
- мода (строка Mode), то есть наиболее многочисленная возрастная группа респондентов (в нашем случае вариант 3: лица от 36 до 60 лет).

Следующая таблица, озаглавленная меткой анализируемой переменной (Возраст), отражает количество респондентов, которые указали тот или иной вариант ответа (столбец 2, Frequency), отсортированный по убыванию (рис. 2.6). Также в этой таблице представлен процент лиц, указавших данные варианты ответа от общего числа респондентов (столбец 3, Percent) и от числа ответивших на анализируемый вопрос Возраст (столбец 4, Valid Percent). Последний столбец 5 (Cumulative Percent)

отражает кумулятивные проценты (то есть вклад каждого варианта ответа в общую сумму). Так же как и в таблице Statistics, здесь указано общее количество ответивших (строка Valid Total) и не ответивших (строка Missing System) на данный вопрос, а также общее количество респондентов (строка Total, в нашем случае 1003).

Statistics

N	Valid	1002
	Missing	1
Mode		3

Рис. 2.5. Таблица Statistics

Возраст

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	36-60 лет	408	40,7	40,7	40,7
	19-35 лет	321	32,0	32,0	72,8
	Старше 60 лет	207	20,6	20,7	93,4
	16-18 лет	66	6,6	6,6	100,0
	Total	1002	99,9	100,0	
Missing	System	1	,1		
Total		1003	100,0		

Рис. 2.6. Таблица Возраст

На подготовительном этапе анализа мы указали на необходимость построения сектограммы по рассматриваемой переменной. Она представлена в результатах линейных распределений после таблицы Возраст (рис. 2.7). Несмотря на то, что мы прямо указали SPSS вывести на диаграмме проценты каждой возрастной группы, программа проигнорировала это указание: в построенной сектограмме указаны только названия категорий.

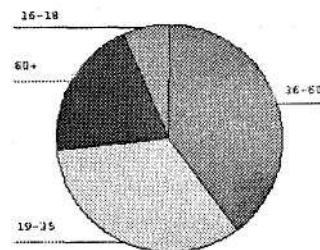


Рис. 2.7. Диаграмма Возраст

К сожалению, графическая подсистема SPSS весьма слаба и не выдерживает сравнения со средствами Microsoft Office. Поэтому рекомендуем пользоваться ею, только когда это действительно оправдано (например, в дисперсионном анализе). Во всех остальных случаях предпочтительнее копировать выводимые таблицы в Microsoft Excel и уже там строить по полученным данным диаграммы.

В рассматриваемом случае, чтобы исправить ситуацию и вывести проценты, дважды щелкните мышью по диаграмме Возраст в окне SPSS Viewer. Откроется специальное окно SPSS Chart Editor, предназначенное для редактирования простых диаграмм (simple charts)¹. В нем выберите меню Chart ► Options. Откроется диалоговое окно Pie Options, в котором следует указать параметр Percents в области Labels (рис. 2.8). Далее щелкните на кнопке ОК и закройте окно SPSS Chart Editor. В окне SPSS Viewer к построенной диаграмме будут добавлены проценты каждой возрастной группы.

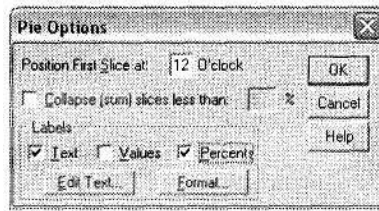


Рис. 2.8. Диалоговое окно Pie Options

Существует еще один способ построения диаграмм по линейным распределениям. Он применяется в случае, если вы уже построили частотную таблицу, но не указали на подготовительном этапе на необходимость вывести диаграмму. В такой ситуации следует дважды щелкнуть мышью на данной таблице в окне SPSS Viewer, а затем выделить тот ее столбец, по которому необходимо построить диаграмму. Например, выделите столбец Valid Percent (значения во всех четырех строках, обозначающих варианты ответа на вопрос Возраст). Затем щелкните правой кнопкой мыши и в открывшемся контекстном меню выберите пункт Create Graph ► Pie для построения сектограммы по долям каждой возрастной группы. В результате после частотной таблицы будет выведена соответствующая круговая диаграмма.

В разделе 1.2 было показано, как рассчитывается статистическая ошибка для величин, выраженных в процентах. Теперь, после того как мы изучили линейные распределения и основные описательные статистики, можно рассмотреть формулу для расчета статистической ошибки значений, выраженных в абсолютных величинах (например, средние значения). Напомним, что статистическая ошибка для данной категории величин рассчитывается для каждой из них в отдельности.

В качестве примера рассмотрим линейное распределение оценок на вопрос Оцените, пожалуйста, качество сухих строительных смесей марки X по пятибалльной шкале: от 1 (очень плохо) до 5 (отлично). При этом в диалоговом окне Statistics (см. рис. 2.2) необходимо выбрать параметры: Mean (среднее арифметическое) и Variance (дисперсия). После окончания расчетов в окне SPSS Viewer будет выведена следующая таблица (рис. 2.9).

Statistics		
N	Valid	6762
	Missing	1406
Mean		3,89
Variance		,634

Рис. 2.9. Таблица Statistics

Формула для расчета статистической ошибки величин, выраженных в абсолютных показателях, имеет следующий вид:

$$\Delta_{\bar{x}} = \pm z \frac{\sigma}{\sqrt{n}},$$

где z — статистическая константа для выбранного доверительного уровня (см. табл. 1.1); σ — дисперсия (строка Variance в таблице Statistics на рис. 2.9); n — размер выборки для данного вопроса (строка Valid в таблице Statistics на рис. 2.9).

Таким образом, для нашего случая и стандартного для маркетинговых исследований доверительного уровня в 95 % статистическая ошибка выборки будет равна:

$$\Delta_{\bar{x}} = \pm 1,96 \frac{0,634}{\sqrt{6762}} \approx \pm 0,02,$$

то есть средняя оценка качества ССС варьируется в пределах от 3,87 балла (3,89 – 0,02) до 3,91 (3,89 + 0,02).

2.2. Линейные распределения для многовариантных вопросов

Как было сказано выше (см. раздел 1.4.2), в SPSS все многовариантные вопросы рассматриваются как совокупность одновариантных переменных, обозначающих варианты ответа. Иными словами, многовариантный вопрос, содержащий три варианта ответа, в SPSS представляется как три дихотомические переменные, принимающие два значения-флага: отмечено/не отмечено.

Наиболее распространены два формата представления многовариантных переменных. В первом случае переменные, представляющие варианты ответа многовариантной переменной, принимают значение 1 (выбрано) или 0 (не выбрано); во втором случае — 1 (выбрано) или System Missing (не выбрано).

Как показывает опыт, первый способ предпочтительнее. Второй способ используется в специфических случаях (например, если необходимо использовать SPSS в качестве клиента автоматизации построения распределений при помощи программ на Sax Basic). Чтобы указать SPSS, какие переменные являются вариантами ответа для многовариантной переменной, наиболее часто используется описываемый далее способ, при котором после формирования многовариантной переменной ее можно использовать для построения линейных и перекрестных распределений.

Для иллюстрации мы построим линейное распределение по многовариантному вопросу Где Вы покупаете сметану? (q7) с вариантами ответа:

1. продмаг (q7_1);
2. рынок (q7_2);
3. супермаркет (q7_3);
4. палатка (q7_4);
5. универсам (q7_5).

Чтобы построить распределения по многовариантным вопросам, прежде всего необходимо сформировать многовариантную переменную. Это делается при помощи меню Analyze ► Multiple Response ► Define Sets. Открывшееся диалоговое окно позволяет сформировать многовариантные переменные (правый список) из общего списка доступных переменных (левый список), как показано на рис. 2.10.

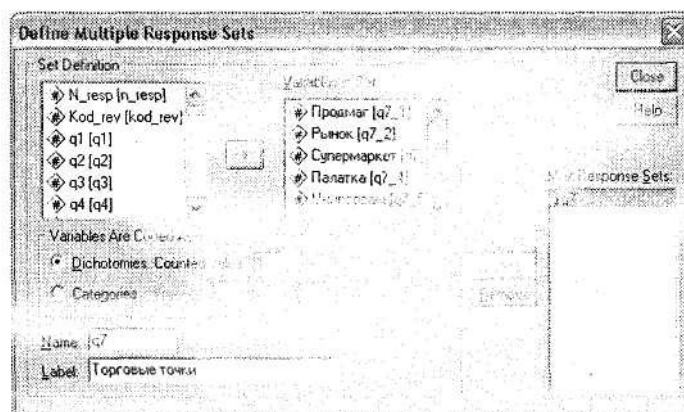


Рис. 2.10. Диалоговое окно Define Multiple Response Sets со сформированной многовариантной переменной Торговые точки

Для создания многовариантной переменной, обозначающей типы торговых точек, сначала выберите в левом списке все дихотомические переменные, кодирующие множественные варианты ответов (q7_1 — q7_5), и переместите их в правый список. Далее в области Variables Are Coded As оставьте выбранный по умолчанию параметр Dichotomies (он указывает, что переменные, обозначающие варианты ответа в многовариантном вопросе, являются дихотомическими) и в соответствующее поле введите цифру, указывающую, что вариант ответа выбран (в нашем случае 1). В поле Name введите имя для вновь создаваемой многовариантной переменной. Назовите ее q7 и присвойте метку Торговые точки (в поле Label). Затем, чтобы создать новую переменную, щелкните на кнопке Add. Обратите внимание, что к именам создаваемых многовариантных переменных добавляется префикс \$ (этим они отличаются от обычных одновариантных переменных). Теперь вы можете создать еще одну или несколько многовариантных переменных, добавляя их в соответствующий список при помощи кнопки Add. Так как в нашем случае мы собираемся анализировать только один многовариантный вопрос, завершим процесс создания новых переменных щелчком на кнопке Close.

Необходимо отметить, что SPSS не сохраняет многовариантные переменные при закрытии рабочего файла с данными. Поэтому каждый раз, когда нужно проанализировать многовариантные вопросы, вам придется снова создавать соответствующие переменные.

Мы создали многовариантную переменную для анализа и теперь можем приступить к построению линейных распределений. Для этого воспользуемся меню Analyze ► Multiple Response ► Frequencies. Следует отметить, что данное меню позволяет строить только таблицы линейных распределений (и нет возможности вывести диаграммы). В открывшемся диалоговом окне в левом списке всех доступных многовариантных переменных (в нашем случае там только одна переменная Торговые точки) выберите интересующие переменные для анализа и перенесите их в правую область Table(s) for (рис. 2.11). Для того чтобы запустить процедуру построения линейных распределений, щелкните на кнопке OK.

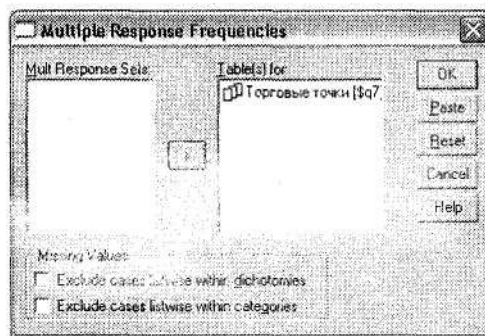


Рис. 2.11. Диалоговое окно Multiple Response Frequencies

В окне SPSS Viewer будет создана таблица с линейными распределениями (частотами) по выбранным переменным (рис. 2.12). Столбец Count содержит количество респондентов, указавших каждый из возможных вариантов ответа на многовариантный вопрос. Столбец Per of Cases показывает доли каждого варианта ответа от общего числа респондентов, ответивших на многовариантный вопрос (гистограмма). Данное число показано под таблицей (999 valid cases, то есть линейное распределение построено по 999 респондентам) и рассчитано как количество анкет, в которых выбран хотя бы один из возможных вариантов ответа на данный многовариантный вопрос. В той же строке (под таблицей) указано количество анкет, в которых не выбрано ни одного варианта ответа (4 missing cases, то есть четыре респондента не указали, в каких типах торговых точек они обычно приобретают сметану). Столбец Per of Responses показывает доли каждого варианта ответа от общего числа ответов; их сумма всегда равна 100 % (сектограмма). Суммы по каждому

столбцу анализируемой таблицы представлены в строке Total responses.

Multiple Response

Group \$Q7 Торговые точки

(Value tabulated = 1)

Dichotomy label	Name	Count	Pet of Responses	Pet of Cases
Продмаг	Q7_1	518	39,4	31,9
Рынок	Q7_2	306	23,3	30,6
Супермаркет	Q7_3	258	19,6	25,8
Палатка	Q7_4	166	12,6	16,6
Универсам	Q7_5	66	5,0	6,6
		-----	-----	-----
		Total re- sponses	314	100,0
				131,5

4 missing cases; 999
valid cases

Рис. 2.12. Таблица Multiple Response, отражающая результаты построения линейного распределения по многовариантной переменной Торговые точки

В связи с тем, что линейные распределения по многовариантным вопросам в SPSS выводятся в текстовом формате (Plain text) и не могут быть перенесены в Microsoft Excel для построения диаграмм, далее мы рассмотрим, как можно строить диаграммы по многовариантным вопросам непосредственно в SPSS.

Если вам необходимо построить гистограмму или сектограмму по многовариантному вопросу, меню Define Sets не используется. Вместо него применяется меню Graphs ► Bar (для гистограмм) или Graphs ► Pie (для сектограмм). За один раз можно построить гистограмму или сектограмму только по одной многовариантной переменной.

Итак, давайте построим гистограмму по многовариантной переменной Торговые точки (параллельно мы построим и сектограмму). Для этого воспользуемся меню Graphs ► Bar. В открывшемся диалоговом окне (рис. 2.13) необходимо указать тип гистограммы Simple (если мы строим сектограмму, данный пункт отсутствует; см. рис. 2.14), а в группе Data in Chart Are выбрать пункт Summaries of separate variables. Затем необходимо щелкнуть на кнопке Define, чтобы перейти к следующему шагу построения диаграммы.

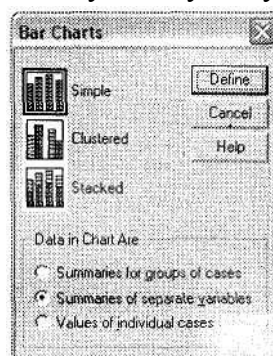


Рис. 2.13. Диалоговое окно Bar Charts с выбранными параметрами для построения гистограмм и сектограмм по многовариантной переменной

В открывшемся диалоговом окне Summaries of Separate Variables (оно одинаково и для гистограмм и для сектограмм) из левого списка всех доступных переменных, имеющихся в файле данных, переместите в правый список все варианты ответа на какой-либо один многовариантный вопрос (в нашем случае это переменные q7_1 — q7_5). как видно на рис. 2.15.

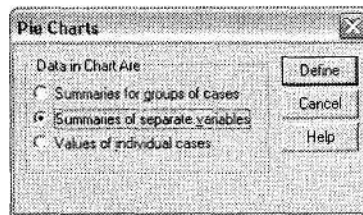


Рис. 2.14. Диалоговые окна Pie Charts с выбранными параметрами для построения гистограмм и сектограмм по многовариантной переменной

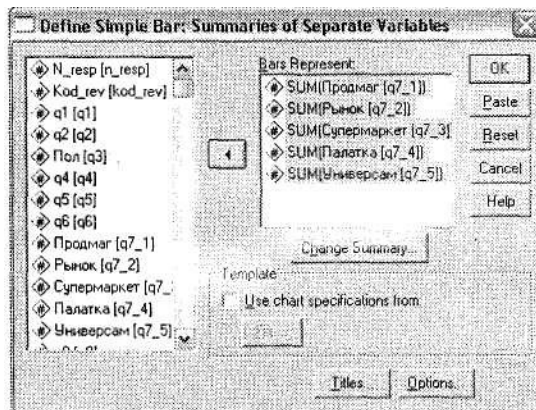


Рис. 2.15. Диалоговое окно Summaries of Separate Variables с выбранной для построения многовариантной переменной Торговые точки

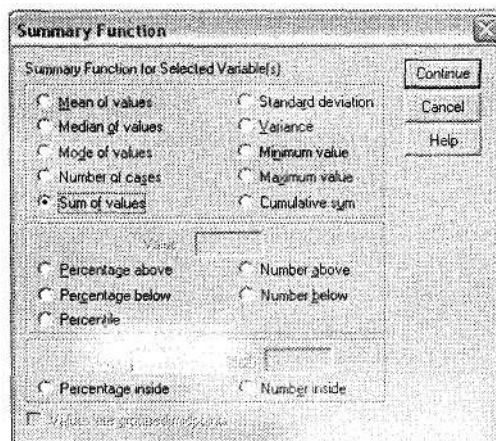


Рис. 2.16. Диалоговое окно Summary Function

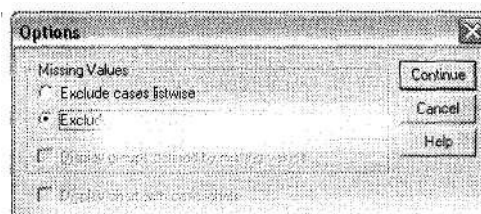


Рис. 2.17. Диалоговое окно Options

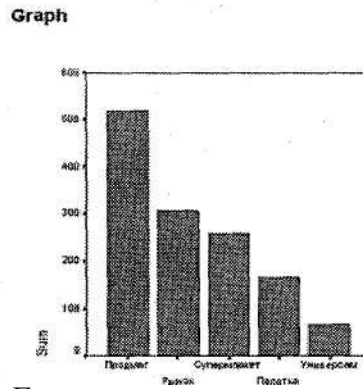


Рис. 2.18. Гистограмма по многовариантной переменной Торговые точки

Щелкните на кнопке **Change Summary** и в открывшемся диалоговом окне (рис. 2.16) выберите пункт **Sum of values**. Данный параметр указывает SPSS на необходимость построить гистограмму по суммарному количеству выбранных вариантов ответа в многовариантном вопросе. После этого закройте данное окно, щелкнув на кнопке **Continue**.

Теперь щелкните на кнопке **Options** и в открывшемся окне выберите пункт **Exclude cases variable by variable**; щелкните на **Continue** (рис. 2.17).

Щелкните на кнопке **OK** в главном диалоговом окне **Summaries of Separate Variables**, и программа выведет результаты построения гистограммы в окне **SPSS Viewer** (рис. 2.18).

Как видите, столбцы построенной гистограммы отражают абсолютное количество респондентов, указавших ту или иную торговую точку. К сожалению, SPSS не позволяет строить гистограмму по многовариантным вопросам, отражающую проценты каждого варианта ответа от общего числа респондентов (или от общего числа ответов). Чтобы отобразить на нашей гистограмме точные количества респондентов, указавших ту или иную торговую точку, следует воспользоваться схемой действий, представленной выше.

Мы рассмотрели наиболее популярный метод статистического анализа данных в маркетинговых исследованиях — построение линейных распределений. Как показывает практика, именно на этом этапе в некоторых отечественных компаниях заканчивается работа с SPSS (иногда строятся также перекрестные распределения), в то время как описательный анализ является лишь начальным этапом анализа данных.

Глава 3 Анализ различий

Цель анализа различий — выявление групп респондентов, статистически значимо различающихся между собой. Все статистические процедуры, относящиеся к группе процедур, которые позволяют выявить такие различия (t-тесты и дисперсионный анализ), сравнивают респондентов на основании средних значений переменных. Иными словами, провести различие можно на основании двух или более числовых переменных.

В практике маркетинговых исследований достаточно часто встречаются ситуации, когда в ходе предварительного анализа (на основании опыта исследователя, когнитивного или статистического анализа) появляется гипотеза о разделении всей выборочной совокупности на определенные группы на основании одного или нескольких признаков (например, при сегментировании потребителей продукта или при построении разрезов). Линейное распределение может показывать, что данные группы респондентов действительно различаются (например, мужчин в выборке в два раза больше, чем женщин). Однако визуального различия между категориями недостаточно для того, чтобы с уверенно-

стью констатировать наличие статистически значимого различия. На установление статистической значимости различий между целевыми группами респондентов и направлены процедуры, объединенные под названием «Анализ различий».

Существует два основных метода определения различий между группами: t-тесты и дисперсионный анализ. Первый метод прост в использовании, и поэтому он применяется часто (в том числе и в маркетинговых исследованиях). Однако в связи с ограничением на количество тестируемых групп (между которыми устанавливается различие) t-тесты не могут применяться для решения всех задач, возникающих при проведении маркетингового анализа. Для преодоления данного ограничения используется дисперсионный анализ, являющийся универсальной методикой для определения статистически значимых различий между любым числом групп респондентов.

3.1. Т-тесты

Т-тесты предназначены для установления различий между двумя группами респондентов. При этом сравниваются только два средних значения. SPSS предлагает три основных типа t-тестов:

- для двух независимых выборок;
- для двух зависимых выборок;
- для одной выборки.

В последующих разделах мы подробно расскажем о каждом из них, но сначала приведем основные характеристики переменных, участвующих в t-тестах (табл. 3.1).

Т-тесты для независимых выборок			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Дихотомическая интервальная	Любое	Интервальная
Т-тесты для зависимых выборок			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
-	-	Две	Интервальная
Т-тесты для одной выборки			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
-	-	Любое	Интервальная

Обратите внимание: зависимая переменная есть только для t-тестов независимых выборок. Для других видов t-тестов (зависимых выборок и одной выборки) зависимая переменная отсутствует. Это связано с тем, что в последнем случае анализу подвергается фактически одна и та же выборка респондентов. В качестве тестируемых независимых переменных во всех случаях используются только переменные с интервальной шкалой. Порядковые переменные могут использоваться только после преобразования их к интервальному виду (см. раздел 2.1).

3.1.1. Т-тесты для независимых выборок

В случае t-тестов для независимых выборок под независимыми выборками понимаются бинарные категории (то есть варианты ответа) какой-либо переменной. Например,

мужчины и женщины (вопрос Пол респондента), покупатели и не покупатели какого-либо продукта (вопрос Покупаете ли Вы данный продукт?) и т. д. То есть когда есть два уровня группирующей (зависимой) переменной и несколько независимых переменных, на основании которых и будет выполняться различие между группами зависимой переменной.

Рассмотрим методику проведения t-тестов для независимых выборок на следующем примере. Предположим, что мы оцениваем различия в частоте посещения игровых клубов между посетителями заведений марки X и других марок. Откройте диалоговое окно Independent-Samples T Test при помощи меню Analyze ► Compare Means ► Independent-Samples T Test (рис. 3.1). В область Test Variable(s) поместите переменные, являющиеся критерием для установления различий (в нашем случае это q18_i Частота посещения). Затем в поле Grouping Variable переместите переменную, которая будет являться группирующей (зависимой). В нашем случае это переменная q1_8, кодирующая категории респондентов, посещающих/не посещающих игровые залы марки X.

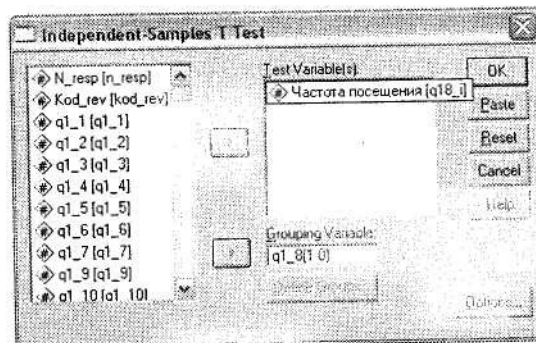


Рис. 3.1. Диалоговое окно Independent-Samples T Test

Так как данная переменная является вариантом ответа на многовариантный вопрос Какие игровые клубы Вы посещаете?, она может принимать два значения:

- 1 — посещают клубы X;
- 0 — не посещают клубы X.

Эти два значения необходимо указать в специальном диалоговом окне Define Groups, вызываемом одноименной кнопкой (рис. 3.2). Обратите внимание, что если вместо дихотомии мы имеем группирующую переменную с интервальной шкалой, это диалоговое окно позволяет установить точку отсечения Cut point, которая будет разделять все возможные значения данной переменной на две группы.

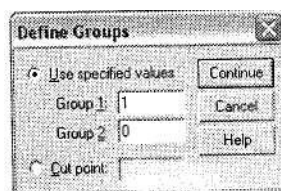


Рис. 3.2. Диалоговое окно Define Groups

С помощью кнопки Options в главном диалоговом окне рассматриваемой процедуры можно установить доверительный уровень для результатов расчета t-теста (рис. 3.3). По умолчанию установлен уровень доверия 95 %. Как было показано выше в разделе 1.2, этот уровень точности (достоверности) результатов является достаточным при проведении статистического анализа в маркетинговых исследованиях.

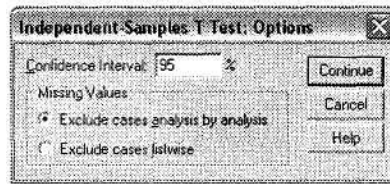


Рис. 3.3. Диалоговое окно Independent-Samples T Test: Options

После завершения процедуры расчета t-теста в окне SPSS Viewer будут отражены результаты (рис. 3.4). В первой таблице Group Statistics вы видите средние значения тестируемой переменной (частота посещения клубов) для обеих групп зависимой переменной X. Как следует из рисунка, для респондентов, посещающих игровые залы марки X, средняя частота посещения составляет 11,9 раз в месяц. Для респондентов, не посещающих данные залы, это значение равно 11,5. Вторая таблица Independent Samples Test позволяет установить статистическое различие между данными значениями.

T-Test

Group Statistics

X		N	Mean	Std. Deviation	Std. Error Mean
Частота посещения	1	49	11,9288	10,43081	1,49140
	0	526	11,5048	9,98682	,43546

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	
Частота посещения	Equal variances assumed	,382	,547	,283	573	,777	,4238	1,49745	-2,61734	3,36497
	Equal variances not assumed			,273	56,495	,786	,4230	1,55367	-2,68795	3,51559

Рис. 3.4. Результаты расчета t-теста для независимых выборок

Анализ этой таблицы начинается с определения значимости теста Ливина (Levene). Данный тест служит для тестирования гипотезы о равенстве дисперсий в тестируемых переменных. Если значение в столбце Sig. столбца Levene's Test for Equality of Variances показывает статистическую *незначимость* теста (в нашем случае — 0,547), то различие между двумя анализируемыми средними определяется из строки Equal variances assumed. В противном случае, если тест Levene статистически *значим*, различие между двумя средними определяется из строки Equal variances not assumed.

Поскольку в нашем примере тест Ливина является статистически незначимым, то определить значимость различия между двумя тестируемыми группами можно при помощи значения, находящегося на пересечении первой строки и столбца Sig. (2-tailed). Значение 0,777 говорит о том, что различие в частоте посещения игровых залов респондентами, посещающими и не посещающими клубы марки X, является статистически незначимым.

3.1.2. Т-тесты для спаренных выборок

Т-тесты для спаренных выборок применяются в случае, когда на различные вопросы отвечает одна и та же группа респондентов.

Например, пассажиры оценивают уровень и качество питания авиакомпании X и авиакомпании Y. Чтобы определить, является ли статистически значимой разница в оценке этих двух авиакомпаний, следует воспользоваться диалоговым окном Paired-Samples T Test, вызываемым при помощи меню Analyze ► Compare Means ► Paired-Samples T Test (рис. 3.5). В левом списке содержатся все доступные переменные из базы данных. Выберите из списка две переменные для тестирования. В нашем случае это q11 (Питание в авиакомпании X) и q26 (Питание в авиакомпании Y). По мере того как вы будете выбирать переменные, они будут последовательно отображаться в области Current Selections. Указав две переменные для анализа, щелкните на кнопке с символом ►, чтобы перенести переменные в область Paired Variables. Кнопка Options позволяет установить уровень доверия для производимых расчетов.

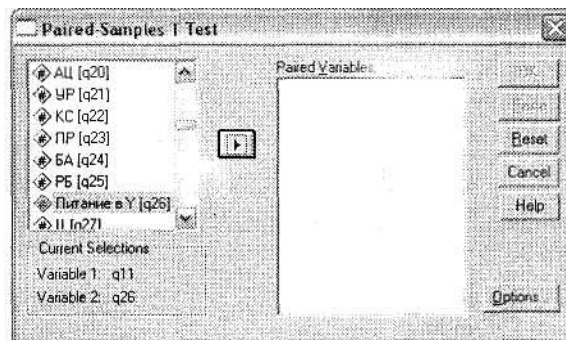


Рис. 3.5. Диалоговое окно Paired-Samples T Test

После щелчка на кнопке ОК будут произведены расчеты t-теста для анализируемых переменных; результаты теста будут отражены в окне SPSS Viewer (рис. 3.6). Как видно на рисунке, SPSS выводит на экран три таблицы. Рассмотрим их по порядку.

Итак, в первой таблице, Paired Samples Statistics, вы видите рассчитанные средние значения для обеих тестируемых переменных. Так, в нашем случае респонденты оценили питание в авиакомпании Y в среднем на 0,4 балла выше, чем в авиакомпании X.

В следующей таблице Paired Samples Correlations представлен коэффициент корреляции (Пирсона) между оценками двух анализируемых переменных. Подробно корреляционный анализ рассматривается в разделе 4.2. Здесь стоит сказать лишь, что чем ближе значение коэффициента к 1, тем сильнее линейная связь между переменными (при условии статистической значимости коэффициента). То есть чем выше уровень оценки по первой переменной, тем выше оценка второй — и наоборот. В нашем случае налицо отсутствие линейной связи между оценками питания в авиакомпании X и Y (коэффициент корреляции = 0,027 при статистической значимости 0,463).

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Питание в X	3,9	731	,974	,036
	Питание в Y	4,3	731	,787	,029

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Питание в X & Питание в Y	73	,027	,463

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Devi- tion	Std. Emor Mean	95% Confidence interval of the Difference				
					Lower	Upper			
Pair	Питание в X &	-,4	1,236	,046	-,44	-,26	-7,692	730	,000
1	Питание в Y								

Рис. 3.6. Результаты расчетов t-теста для спаренных выборок

Наконец, третья таблица, Paired Samples Test, позволяет сделать вывод о наличии/отсутствии статистически значимого различия между тестируемыми переменными, что следует из значения в столбце Sig. (2-tailed). В нашем случае различие между оценками питания в авиакомпаниях X и Y, равное 0,4 балла, является статистически значимым ($<0,001$).

3.1.3. Т-тесты для одной выборки

В результате t-теста для одной выборки можно выяснить, отличается ли значительно реальное среднее значение какой-либо переменной от стандарта. В маркетинговых исследованиях при помощи данного теста определяют, отличается ли среднее значение какого-либо параметра для определенной целевой группы респондентов от среднего значения по всей выборке.

Например, питание на борту самолетов авиакомпании X (переменная q11) всеми респондентами оценено в среднем на 4,0 балла. Вместе с тем пассажиры первого класса оценили питание несколько выше: в среднем на 4,1 балла. Возникает вопрос, является ли выявленное различие статистически значимым. То есть отличаются ли пассажиры первого класса от всех респондентов на основании уровня оценки питания на борту? Выяснить это нам поможет t-тест для одной выборки. Ниже описан механизм его проведения.

Для проведения t-теста мы должны отобрать только тех респондентов, которые летают первым классом. (Как это сделать, см. в разделе 1.5.1.1.) После этого следует воспользоваться меню Analyze ► Compare Means ► One-Sample T Test, чтобы открыть диалоговое окно One-Sample T Test (рис. 3.7). Далее перенесите из левого списка всех доступных переменных в область Test Variable(s) интересующую нас переменную q11 (Питание). В поле Test Value укажите стандартное значение, с которым мы будем сравнивать среднее тестируемой переменной. В нашем случае это 4,0. Кнопка Options позволяет указать доверительный уровень, для которого устанавливается различие.

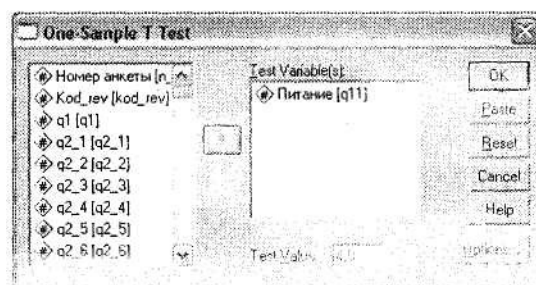


Рис. 3.7. Диалоговое окно One-Sample T Test

После того как SPSS завершит расчет t-теста, в окне SPSS Viewer появятся две таблицы с результатами (рис. 3.8).

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Питание в X	3,9	731	,974	,036
	Питание в Y	4,3	731	,787	,029

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Питание в X & Питание в Y	731	,027	,463

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)			
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference							
					Lower	Upper						
Pair 1	Питание в X - Питание в Y	-,4	1,236	,046	-,44	-,26	-7,692	730	,000			

Рис. 3.8. Результаты расчета t-теста для одной выборки

В первой таблице, One-Sample Statistics, отражены расчеты среднего значения исследуемой переменной (столбец Mean). В нашем случае данное значение отражает среднюю оценку питания пассажиров первого класса (4,1 балла). Вторая таблица, One-Sample Test, позволяет сделать вывод о статистической значимости/незначимости тестируемого различия. Как следует из значения столбца Sig. (2-tailed), различие в оценках пассажиров первого класса и всей выборочной совокупности респондентов является статистически незначимым (0,149). Разница между реальным и тестируемым значениями (в нашем случае — 0,1 балла) отражается в столбце Mean Difference.

3.2. Дисперсионный анализ

Иногда при анализе данных маркетинговых исследований достаточно сравнить только две группы респондентов, то есть установить различия между двумя категориями опрошенных. Однако часто у исследователей возникает необходимость проанализировать не две, а три или более категории респондентов. В этом случае

следует прибегнуть к использованию дисперсионного анализа, который позволяет анализировать одновременно любое число групп.

Различают одномерный (Analysis of variance, ANOVA) и многомерный (Multiple analysis of variance, MANOVA) дисперсионный анализ. Для одномерного дисперсионного анализа существует только одна зависимая переменная; для многомерного — несколько. Также в этом разделе мы рассмотрим одномерный дисперсионный анализ с повторными измерениями (ANOVARM)1.

В табл. 3.2 приведены основные характеристики переменных, участвующих в различных видах дисперсионного анализа.

Таблица 3.2. Основные характеристики переменных, участвующих в дисперсион-

ном анализе

Одномерный дисперсионный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Любой	Любое	Любой
Одномерный дисперсионный анализе повторными измерениями			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Любой	Любой	Любой
Многомерный дисперсионный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Любое	Любой	Любое	Любой

3.2.1. Одномерный дисперсионный анализ

Как было сказано выше, одномерный дисперсионный анализ исследует влияние одной или нескольких независимых переменных на одну зависимую. Одномерный дисперсионный анализ может быть однофакторным (one-way ANOVA) или многофакторным (n-way ANOVA). В первом случае есть только одна независимая переменная; во втором — несколько.

Однофакторный одномерный дисперсионный анализ можно проводить двумя способами: при помощи специальной процедуры One-way ANOVA (меню Analyze ► Compare Means ► One-way ANOVA) или посредством обобщенной линейной модели (меню Analyze ► General Linear Model ► Univariate). Второй прием является более универсальным и обладает полным объемом функциональности первого, поэтому далее мы рассмотрим только GLM (использование первого метода аналогично GLM). Необходимо отметить, что для проведения одномерного дисперсионного анализа на практике (в маркетинговых исследованиях) существует одно весьма существенное ограничение. При увеличении количества факторов (то есть независимых переменных) в модели сложность интерпретации результатов расчета возрастает многократно. Так, однофакторный анализ является наиболее простым. Его результаты понятны сразу при взгляде на итоговую таблицу. Двухфакторный анализ намного сложнее в интерпретации — чтобы понять его результаты, придется потратить много времени, разбираясь в таблицах и графиках. Для интерпретации результатов трехфакторного анализа необходимо обладать некоторым опытом в его проведении. Четырех- и мультифакторные модели в большинстве своем могут успешно интерпретироваться только квалифицированными исследователями. Таким образом, для практических целей лучше воздержаться от исследования большого числа взаимодействий между факторами и ограничиться несколькими наиболее важными. В настоящем разделе мы последовательно рассмотрим одно-, двух- и трехфакторные модели одномерного дисперсионного анализа. При этом будут использоваться следующие исходные данные:

Исследуется покупательское поведение потребителей глазированных сырков. Респонденты разделяются на целевые группы в зависимости от их пола (q3), возраста (q4) и количества членов семьи (q72). Одним из вопросов анкеты является: «Какое количество глазированных сырков в среднем Вы покупаете за одно посещение магазина?» (q6) с вариантами ответа: 1 шт., 2 шт., 3 шт., 4 шт., 5 шт., 6-7 шт., 8-10 шт. и более 10 шт. Требуется выяснить, различается ли кратность покупок глазированных сырков различными целевыми группами респондентов (половыми, возрастными и по количеству членов семьи).

Прежде всего мы проведем однофакторный одномерный дисперсионный анализ и установим, насколько значимо различается кратность покупок в различных возрастных группах респондентов (1 — младше 18 лет; 2 — 19-35 лет; 3 — 36-60 лет; 4 — старше 60 лет).

Диалоговое окно одномерного дисперсионного анализа запускается при помощи меню Analyze ► General Linear Model ► Univariate (рис. 3.9). Из левого списка всех доступных переменных переместите в поле для зависимой переменной Dependent Variable переменную q6 (Кратность покупок). Как видите, в качестве зависимой переменной в дисперсионном анализе выступает основание сегментирования респондентов по группам, то есть та переменная, которая и определяет различия между категориями независимой переменной. (Это замечание достаточно сложно осознать, так как при проведении дисперсионного анализа как бы стираются границы в трактовке зависимых и независимых переменных — по крайней мере, по сравнению с другими видами статистического анализа, например регрессионного.)

В область для независимых переменных Fixed Factor(s) поместите Возраст (q4). Обратите внимание на разницу между областями Fixed Factor(s) (факторы с фиксированными эффектами) и Random Factor(s) (факторы со случайными эффектами). Фиксированными факторами называют переменные, уровни которых охватывают все возможные состояния этой переменной. Например, пол может быть только мужской или женский, а возраст, например, младше 30 лет, от 30 до 60 лет и старше 60 лет. Случайные факторы представляют переменные, уровни которых охватывают лишь часть из всего многообразия возможных состояний. Так как в нашем случае переменная q4 (Возраст) содержит все возможные возрастные группы респондентов, мы поместили ее в область фиксированных факторов.

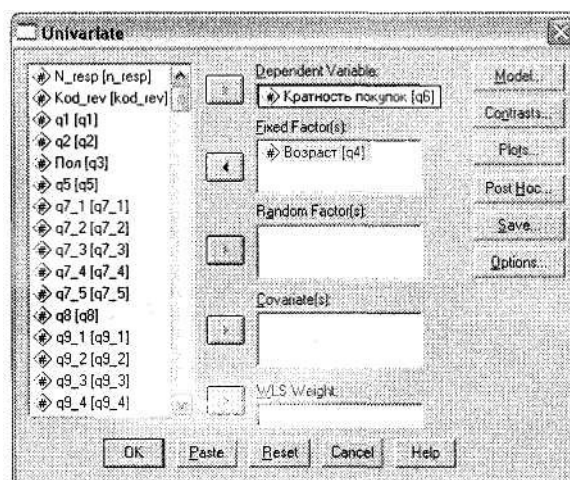


Рис. 3.9. Диалоговое окно Univariate

Если после этого вы щелкнете на кнопке ОК, то получите только одну таблицу, из которой можно узнать лишь о наличии/отсутствии значимых различий между возрастными группами. Однако останется неизвестным, какие именно группы отличаются от других.

Для того чтобы определить это, существуют дополнительные статистические тесты, задаваемые при помощи кнопки Post Hoc. Соответствующее диалоговое окно представлено на рис. 3.10. Перенесите из области Factor(s) в область Post Hoc Tests for те независимые переменные (факторы), которые необходимо подвергнуть тестированию на предмет установления различий между их группами. В нашем случае есть всего одна факторная переменная q4, которую и следует перенести в область тестирования. Далее укажите релевантные дополнительные тесты для указанной переменной. При этом, как видно на

рисунке, SPSS выводит различные тесты для равных и неравных дисперсий (Equal Variances Assumed и Equal Variances Not Assumed соответственно).

Установить равенство/неравенство дисперсий позволяет тест Levene, вывод которого на экран мы покажем ниже. В общем случае мы не знаем, равны ли дисперсии и, соответственно, какую группу статистических тестов следует использовать. Поэтому рекомендуется сразу вывести тесты для равных и неравных дисперсий, чтобы сократить количество итераций при проведении дисперсионного анализа. SPSS предлагает много различных дополнительных тестов, помогающих определить различия между группами исследуемых переменных. Однако использовать их все нецелесообразно. Мы рекомендуем ограничиться наиболее популярным и универсальным тестом Scheffe для равных дисперсий и тестом Tamhane's T2 — для неравных дисперсий. Теперь можно закрыть описываемое диалоговое окно щелчком на кнопке Continue.

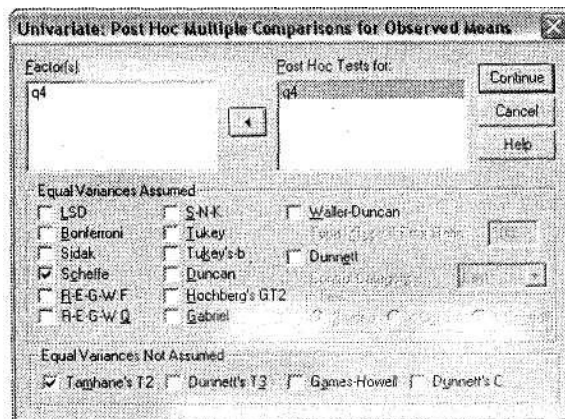


Рис. 3.10. Диалоговое окно Univariate: Post Hoc Multiple Comparisons for Observed Means

Выше мы упомянули о специальном тесте, позволяющем установить равенство/неравенство дисперсий. На необходимость проведения данного теста (так же как и многих других) можно указать в диалоговом окне Options, вызываемом одноименной кнопкой в главном диалоговом окне Univariate (рис. 3.11). Для однофакторного дисперсионного анализа можно ограничиться только одним тестом Levene на равенство дисперсий (параметр Homogeneity tests).

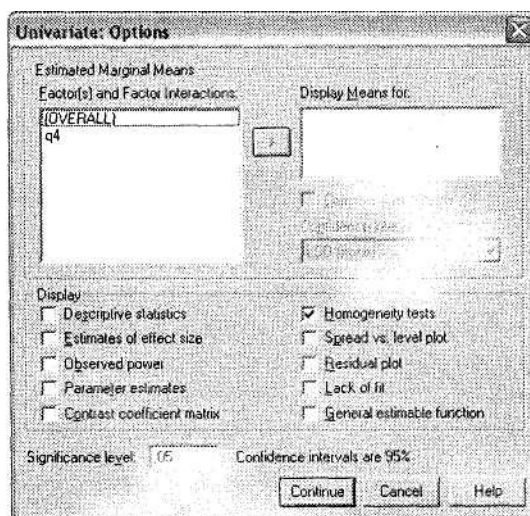


Рис. 3.11. Диалоговое окно Univariate: Options

Следует отметить, что если исследуемая независимая переменная имеет всего две категории (дихотомия), апостериорные тесты для нее не проводятся. Установить направление различия между категориями позволяет вывод средних значений зависимой переменной в каждой из двух категорий. Для этого перенесите исследуемую независимую дихотомическую переменную из области Factor(s) and Factor Interactions

в область Display Means for. В нашем случае единственная независимая переменная Возраст имеет больше двух категорий (4), и поэтому специально выводить для нее средние значения нет смысла (они будут выведены в таблице Homogenous Subsets).

Остальные кнопки главного диалогового окна Univariate предназначены для многофакторного анализа, рассматриваемого ниже. Теперь щелкните на кнопке О К, чтобы запустить процедуру дисперсионного анализа. В окне SPSS Viewer будут выведены результаты расчетов.

Первой практически значимой таблицей является результат теста на равенство дисперсий зависимой и независимых переменных Levene's Test of Equality of Error Variances (рис. 3.12). В столбце Sig. данной таблицы содержится единственное интересующее нас значение — это статистическая значимость тестовой статистики F. Если значение в данном столбце показывает незначимость F — значит, дисперсии равны, и в дальнейшем мы будем анализировать результаты расчета теста Scheffe (предполагающего равенство дисперсий). В противном случае, если F-статистика значима, — дисперсии не равны, и при анализе различий между группами следует использовать тест Tamhane's T2 (предполагающий неравенство дисперсий). Как вы видите на рисунке, статистика F незначима (Sig. = 0,433) — и, следовательно, можно сделать вывод о равенстве дисперсий.

Levene's Test of Equality of Error Variances^a

Dependent Variable: Кратность покупок

F	df1	df2	Sig.
,916	3	998	,433

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Q4

Рис. 3.12. Таблица Levene's Test of Equality of Error Variances

Следующая таблица — это Tests of Between-Subjects Effects (рис. 3.13). Данная таблица является центральной в выводимых результатах дисперсионного анализа и показывает наличие/отсутствие значимых различий между категориями исследуемых переменных. Первое, на что следует обратить внимание при анализе описываемой таблицы, — это величина R2, отражающая долю совокупной дисперсии в зависимой переменной, описываемой статистической моделью. Другими словами, это та часть вариации зависимой переменной, которую можно объяснить на основании независимой переменной. Естественно, что чем меньше независимых переменных, тем меньше величина R2, и наоборот.

Так, в нашем случае есть только одна независимая переменная q4 (Возраст), и при этом R2 весьма мала (0,019). Для дисперсионного анализа значения R2 можно просто проигнорировать, так как они не важны для практического использования полученной модели'. Второе, на что обращают внимание исследователи при интерпретации таблицы Tests of Between-Subjects Effects, — это собственно значимость различия между группами независимой переменной. Этот вывод следует из значения на пересечении строки, содержащей соответствующую независимую переменную, и столбца Sig.. Как вы видите на рисунке, имеет место статистически высоко значимое различие между различными возрастными группами респондентов по кратности покупок глазированных сырков (значимость

F-статистики у переменной $q4 < 0,001$). Обратите внимание, что если тест Levene выявил факт неравенства дисперсий независимых и зависимых переменных, следует поднять порог значимости со стандартного значения 0,05 до 0,01.

Tests of Between-Subjects Effects

Dependent Variable: Кратность покупок

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	54,121 ^a	3	18,040	6,602	,000
Intercept	12184,400	1	12184,400	4458,853	,000
Q4	54,121	3	18,040	6,602	,000
Error	2727,166	998	2,733		
Total	22129,000	1002			
Corrected Total	2781,286	1001			

a. R Squared = ,019 (Adjusted R Squared = ,017)

Рис. 3.13. Таблица Tests of Between-Subjects Effects

После того как мы установили наличие статистически значимого различия между возрастными группами респондентов на основании кратности покупок сырков, необходимо определить, какие из четырех имеющихся возрастных групп отличаются от остальных и каким образом (в большую или в меньшую сторону).

Давайте сделаем это при помощи таблицы Multiple Comparisons, представленной на рис. 3.14. При интерпретации данной таблицы прежде всего вспомните результаты теста Levene. Так, в нашем случае на основании данного теста дисперсии оказались равными, и поэтому в данной таблице мы будем рассматривать только ту ее часть, в которой приведены расчеты по методу Scheffe (напомним, что тест Tamhane мы бы применяли только если бы дисперсии были неравны).

Итак, в первой части таблицы (Scheffe) мы видим сравнение различий между каждой из четырех возрастных категорий с остальными категориями. На основе этих данных и определяются та или те группы, которые значимо отличаются от других. Так, из столбца Sig. (статистическая значимость) мы видим, что только группа респондентов старше 60 лет статистически значимо отличается от всех остальных. Остальные целевые группы не отличаются друг от друга. При этом из столбца Mean Difference можно видеть, насколько отличается среднее значение той или иной группы от среднего значения других групп (звездочками отмечены значимые различия при 95%-ном доверительном уровне)¹.

Наконец, в последней таблице Homogeneous Subsets (рис. 3.15) представлена однозначная картина различий между группами независимой переменной. Здесь все возрастные группы разделены на две категории на основании различий в кратности покупок. В первую категорию входит целевая группа респондентов старше 60 лет; во вторую — все остальные возрастные группы (то есть респонденты младше 60 лет). Если бы оказалось, что статистически значимых различий в кратности покупок глазированных сырков различными возрастными группами респондентов не наблюдается, все группы независимой переменной были бы отнесены к одной категории (Subset был бы только 1). Иногда возникает ситуация, при которой одна и та же группа респондентов может относиться сразу к нескольким группам. В таком случае следует поднять порог значимости со стандартных 0,05, скажем, до 0,01 (или любого другого значения).

Multiple Comparisons

Dependent Variable: Кратность покупок

			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
(h) Возраст	(j) Возраст	Lower Bound				Upper Bound	
Schaeffe	Младше 18 лет	От 18 до 35 лет	,23	,223	,781	-,39	,86
		От 36 до 60 лет	,21	,219	,817	-,40	,83
		Старше 60 лет	,76*	,234	,014	,11	1,41
	От 18 до 35 лет	Младше 18 лет	-,23	,223	,781	-,86	,39
		От 36 до 60 лет	-,02	,123	,999	-,37	,33
		Старше 60 лет	,53*	,147	,005	,12	,94
	От 36 до 60 лет	Младше 18 лет	-,21	,219	,817	-,83	,40
		От 18 до 35 лет	,02	,123	,999	-,33	,37
		Старше 60 лет	,55*	,141	,002	,15	,94
	Старше 60 лет	Младше 18 лет	-,76*	,234	,014	-1,41	-,11
		От 18 до 35 лет	-,53*	,147	,005	-,94	-,12
		От 36 до 60 лет	-,55*	,141	,002	-,94	-,15
Tamhane	Младше 18 лет	От 18 до 35 лет	,23	,223	,883	-,37	,83
		От 36 до 60 лет	,21	,217	,911	-,37	,80
		Старше 60 лет	,76*	,234	,009	,13	1,39
	От 18 до 35 лет	Младше 18 лет	-,23	,223	,883	-,83	,37
		От 36 до 60 лет	-,02	,123	1,000	-,35	,30
		Старше 60 лет	,53*	,152	,003	,13	,93
	От 36 до 60 лет	Младше 18 лет	-,21	,217	,911	-,80	,37
		От 18 до 35 лет	,02	,123	1,000	-,30	,35
		Старше 60 лет	,55*	,142	,001	,17	,92
	Старше 60 лет	Младше 18 лет	-,76*	,234	,009	-1,39	-,13
		От 18 до 35 лет	-,53*	,152	,003	-,93	-,13
		От 36 до 60 лет	-,55*	,142	,001	-,92	-,17

Based on observed means.

*. The mean difference is significant at the .05 level.

Рис. 3.14. Таблица Multiple Comparisons

Также из рассматриваемой таблицы можно сделать вывод о направлении различия между выделенными категориями. Так, в нашем случае мы можем заключить, что респонденты старше 60 лет покупают глазированные сырки в меньших объемах, чем респонденты младше 60 лет. В точности определить размер или величину различия можно, только если в качестве зависимой переменной выступает интервальная переменная. Так как у нас переменная q6 Кратность покупок относится к порядковой шкале, мы не можем сделать точный вывод о величине различия. Если стоит такая задача, можно преобразовать зависимую порядковую переменную к интервальному виду (например, при помощи перекодирования кодов групп в средние значения данных групп: 1 (от 16 до 18 лет) —> 17 и пересчитать дисперсионный анализ. Это даст хотя бы приблизительную оценку величины различия. Нам достаточно только установленной статистической значимости (то есть существования) различия и его направления (респонденты старше 60 лет покупают меньше сырков, чем более молодые).

Homogeneous Subsets

Кратность покупок

Возраст	N	Subset	
		1	2
Scheffe ^{a,1} Старше 60 лет	207	3,95	
От 19 до 35 лет	321		4,48
От 36 до 60 лет	408		4,50
Младше 18 лет	66		4,71
Sig.		1,000	,672

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 2,733.

a. Uses Harmonic Mean Sample Size = 156,564.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = ,05.

Рис. 3.15. Таблица Homogeneous Subsets

Рассмотрим теперь ситуацию, когда необходимо исследовать сразу две независимые переменные (и взаимодействия между ними), то есть выполнить двухфакторный одномерный дисперсионный анализ.

Исходные данные останутся такими же, как в предыдущем примере, однако теперь мы будем устанавливать различие в кратности покупок сырков возрастными и половыми группами (переменная q3). Для этого вновь откроем диалоговое окно Univariate (рис. 3.9) и добавим в область для фиксированных факторов (независимых переменных с фиксированными эффектами) переменную Пол. При проведении многофакторного анализа (двухфакторной и более) кнопка Model позволяет задать исследование либо всех возможных взаимодействий между независимыми переменными (в нашем случае будет установлено различие не только между четырьмя возрастными и двумя половыми группами по отдельности, но и между каждой половозрастной группой), либо только каких-то конкретных взаимодействий. В диалоговом окне Model можно задать и другие значения, но для большинства задач маркетинговых исследований достаточно оставлять все эти значения по умолчанию. Иными словами, кнопкой Model лучше не пользоваться. То же самое касается и кнопки Contrasts (исследование взаимодействий между уровнями независимых переменных), а также кнопки Save, позволяющей сохранять некоторые значения. В большинстве практических случаев, встречающихся в маркетинговых исследованиях, при проведении дисперсионного анализа вам не потребуется ничего сохранять. При проведении многофакторного дисперсионного анализа в диалоговом окне Post Hoc (рис. 3.10) следует добавить к списку исследуемых переменных все независимые факторы, кроме дихотомических. В нашем случае переменная Пол является

дихотомической, так что добавлять ее в область Post Hoc Tests for (дополнительно к переменной Возраст) не следует. Таким образом, все параметры этого диалогового окна останутся неизменными по сравнению с предыдущим примером.

В диалоговом окне Options (рис. 3.11) необходимо добавить дихотомическую переменную q3 (Пол), а также ее взаимодействие с переменной q4 (Возраст) — q3*q4 — в область Display Means for, что позволит вывести средние значения по каждой группе муж-

чин и женщин при определении направления различия между ними. После этого можно запускать процедуру дисперсионного анализа на выполнение.

В окне SPSS Viewer будут выведены результаты расчетов. Они будут отличаться от результатов предыдущего примера. Во-первых, как видно из рис. 3.16, тест Levene теперь является значимым (Sig. = 0,033), из чего следует вывод о неравенстве дисперсий.

Levene's Test of Equality of Error Variances^a

Dependent Variable: Кратность покупок

F	df1	df2	Sig.
2,189	7	991	,033

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Q3+Q4+Q3 * Q4

Рис. 3.16. Таблица Levene's Test of Equality of Error Variances

Во-вторых, в таблице Tests of Between-Subjects Effects появились результаты расчета значимости F-статистики для переменной Пол (q3), а также для взаимодействия q3*q4. Как видно из рис. 3.17, мужчины и женщины не имеют статистически значимых различий по кратности покупок глазированных сырков. То же относится и к взаимодействию q3*q4: оно не является статистически значимым. При этом, несмотря на неравенство дисперсий (порог значимости возрос до 0,01), переменная q4 (Возраст) сохранила свое значимое влияние на зависимую переменную (Sig. = 0,011), то есть возрастные группы по-прежнему различаются по кратности покупок сырков. Необходимо также отметить, что с добавлением переменной q3 доля совокупной дисперсии в зависимой переменной, объясняемая построенной моделью, несколько возросла (R2 = 0,022).

После таблицы Tests of Between-Subjects Effects следуют расчеты средних значений для дихотомической переменной q3 (Пол) и для взаимодействия q3 x q4 (рис. 3.18). В нашем случае ни переменная q3, ни ее взаимодействие с q4 не являются статистически значимыми, поэтому данные таблицы бесполезны. Однако если бы переменная Пол была значима (то есть различие между мужчинами и женщинами существовало), на основании первой таблицы можно было бы сделать заключение о том, какая именно половая группа покупает больше сырков.

Так, если предположить, что влияние переменной Пол статистически значимо, из рис. 3.18 можно было бы заключить, что женщины покупают глазированные сырки в больших объемах по сравнению с мужчинами. То же можно сказать и относительно второй таблицы (Пол x Возраст). Случается, что по результатам таблицы Tests of Between-Subjects Effects некая переменная оказывается незначимой, однако в таблице Multiple Comparisons отдельные уровни этой переменной значимо отличаются друг от друга. В такой ситуации все равно следует признать рассматриваемую переменную незначимой и в дальнейшем игнорировать связанные с ней апостериорные тесты.

Tests of Between-Subjects Effects

Dependent Variable: Кратность покупок

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	61,992 ^a	7	8,856	3,235	,002
Intercept	8890,961	1	8890,961	3247,633	,000
Q3	1,086	1	1,086	,397	,529
Q4	30,717	3	10,239	3,740	,011
Q3 * Q4	6,130	3	2,043	,746	,525
Error	2713,035	991	2,738		
Total	22084,000	999			
Corrected Total	2775,027	998			

a. R Squared = ,022 (Adjusted R Squared = ,015)

Рис. 3.17. Таблица Tests of Between-Subjects Effects

Estimated Marginal Means

1. Пол

Dependent Variable: Кратность покупок

Пол	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Мужчины	4,359	,132	4,100	4,618
Женщины	4,457	,080	4,299	4,614

2. Пол * Возраст

Dependent Variable: Кратность покупок

Пол	Возраст	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Мужчины	Младше 18 лет	4,480	,331	3,831	5,129
	От 19 до 35 лет	4,617	,184	4,257	4,978
	От 36 до 60 лет	4,370	,225	3,929	4,812
	Старше 60 лет	3,969	,292	3,395	4,543
Женщины	Младше 18 лет	4,925	,262	4,412	5,438
	От 19 до 35 лет	4,433	,107	4,222	4,643
	От 36 до 60 лет	4,520	,088	4,347	4,692
	Старше 60 лет	3,949	,125	3,703	4,194

Рис. 3.18. Таблицы Estimated Marginal Means

Завершают вывод результатов двухфакторного анализа таблицы с расчетами апостериорных тестов. В нашем случае они практически такие же, как в предыдущем примере, поскольку переменная Возраст сохранила свою значимость (см. рис. 3.14 и 3.15). Однако при интерпретации таблицы Multiple Comparisons следует помнить о неравенстве дисперсий. Поэтому значимость различий между отдельными воз-

растными группами надо устанавливать на основании второй части таблицы Tamhane.

Итак, мы рассмотрели одно- и двухфакторный одномерный дисперсионный анализ. Далее мы поговорим более подробно о трехфакторном дисперсионном анализе. На его примере мы рассмотрим построение графиков и методы их использования с целью облегчения интерпретации значимых взаимодействий между переменными.

Теперь мы будем использовать все четыре переменные из исходного условия задачи (см. выше), то есть проанализируем различия в кратности покупки глазированных сырков анализируемыми целевыми группами респондентов (половыми, возрастными и по количеству членов семьи). Откройте диалоговое окно Univariate и добавьте в список независимых переменных (область Fixed Factor(s)) еще одну переменную q72 (Количество членов семьи).

Здесь необходимо сделать одно важное отступление. Время проведения расчетов в дисперсионном анализе (как одномерном, так и многомерном) при добавлении каждого нового фактора существенно возрастает. Если при этом зависимая переменная содержит достаточно большое количество уровней, расчеты могут затянуться на весьма длительное время. Исследователям-практикам следует знать об одной существенной особенности SPSS: скорость ее работы лимитируется тактовой частотой основного микропроцессора и объемом оперативной памяти (скорость работы жесткого диска не играет существенной роли). SPSS может использовать в своей работе только один процессор, то есть если у вас в компьютере установлено два и более процессора, для SPSS это не будет иметь никакого значения. Поэтому при работе с данной программой мы настоятельно рекомендуем использовать мощные машины с высокопроизводительным процессором и достаточным объемом оперативной памяти. К сожалению, в настоящее время не все отечественные компании имеют возможность приобретать мощные компьютеры. Предлагаем следующий выход. В главном диалоговом окне Univariate есть кнопка Model, которая, как мы сказали выше, в маркетинговых исследованиях используется редко, поскольку при проведении дисперсионного анализа не требуется анализировать сразу много (четыре и более) факторов и, следовательно, скорость работы программы будет приемлемой. Однако если в анализ приходится включать четыре и более независимых переменных, придется воспользоваться кнопкой Model. Щелкните на ней — и вы увидите одноименное диалоговое окно, показанное на рис. 3.19. По умолчанию в SPSS выбрана полнофакторная модель дисперсионного анализа Full factorial, где исследуется влияние на зависимую переменную:

1. всех независимых переменных по отдельности;
2. всех возможных взаимодействий между независимыми переменными.

Именно на расчеты, связанные со вторым пунктом, и тратится основное время. Поэтому при ограничениях, налагаемых аппаратным обеспечением компьютера, следует отказаться от использования полнофакторных моделей в пользу определяемых пользователем (Custom). Если ограничения жесткие, можно выполнить только исследования влияния независимых переменных на зависимую по отдельности (в терминологии SPSS, Main effects)¹.

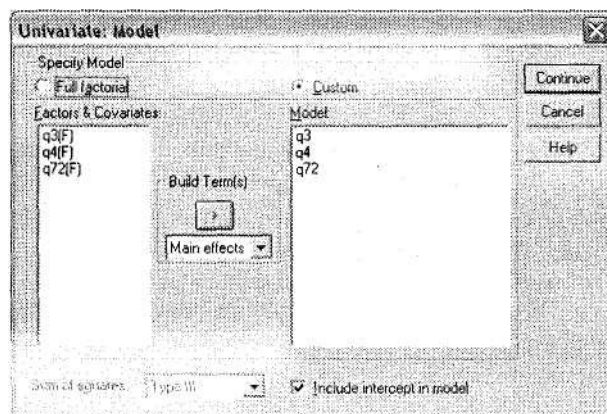


Рис. 3.19. Диалоговое окно Univariate: Model

В данном диалоговом окне в левом списке содержатся все выбранные для анализа независимые переменные. Чтобы определить пользовательскую модель, в левом списке Factors & Covariates выберите переменные, которые будут включены в итоговую пользовательскую модель. Затем из раскрывающегося списка Build Term(s) выберите тот или иной тип взаимодействия между переменными. И наконец, щелкните на соответствующей кнопке, чтобы перенести сформированную пользовательскую модель в правый список Model.

Если вы хотите рассмотреть только влияние факторных переменных по отдельности, выполните действия, показанные на рис. 3.19. Выберите все независимые переменные в левом списке, тип модели Main effects и перенесите эти переменные в правую область. Другими видами моделей являются:

- Interaction — исследование всех видов взаимодействий между выбранными переменными;

- AN 2-, 3-, 4-, 5-way — исследование только взаимодействий соответственно второго ($q1*q2$), третьего ($q1*q2*q3$), четвертого ($q1*q2*q3*q4$) и пятого ($q1*q2*q3*q4*q5$) порядков.

Обратите внимание, что одновременно можно сформировать в правом списке Model сколько угодно различных моделей, подбирая только основные, необходимые вам взаимодействия факторов.

Для иллюстрации решения задачи (выполнение трехфакторного дисперсионного анализа) не будем задавать пользовательские модели, а воспользуемся полнофакторной моделью, установленной по умолчанию. В диалоговом окне Model есть еще два не рассмотренных ранее параметра: Sum of squares и Include intercept in model. Первый параметр позволяет задать тип формулы для расчета суммы квадратов (тестовой величины, на основании которой и производится расчет статистической значимости различий). В маркетинговых исследованиях рекомендуется использовать тип III, установленный по умолчанию. Второй параметр служит для указания на необходимость включить в итоговую модель расчеты значимости отрезка значений. Данный параметр также можно всегда оставлять установленным по умолчанию.

Вернемся к описанию решения поставленной задачи. Мы добавили в соответствующие поля главного диалогового окна Univariate одну зависимую переменную и сразу три независимые. При помощи кнопок Post Hoc и Options необходимо выбрать те же параметры, которые мы выбирали для одно- и двухфакторного анализа. В результате останется не рассмотренной одна важная кнопка в главном диалоговом окне Plots, позволяющая указать параметры для построения графиков. Эту кнопку следует использовать в тех ситуациях, когда обнаружено статистически значимое взаимодействие между факторами.

Для того чтобы построить график взаимодействия факторов, сначала мы должны провести дисперсионный анализ по обычной схеме (без графиков) и выяснить, есть ли значимые взаимодействия. После щелчка на кнопке ОК в окне SPSS Viewer будут выведены результаты расчетов для трехфакторного одномерного дисперсионного анализа. Нет смысла приводить их здесь — в них нет ничего для вас нового. Вместо этого давайте посмотрим, как интерпретировать значимые взаимодействия между факторами.

Существует два основных способа интерпретации взаимодействий:

- в табличной форме — по результатам апостериорных тестов;
- в графической форме — по построенным графикам взаимодействий.

Графическая форма представления результатов зачастую более предпочтительна по сравнению с табличной, особенно при анализе взаимодействий трех и более уровней. На рис. 3.20 показано диалоговое окно Profile Plots. Для того чтобы построить график по двухуровневому взаимодействию, из левого списка всех независимых переменных (область Factors) выберите переменную, категории которой будут располагаться по оси абсцисс (горизонтальной), и поместите ее в поле Horizontal Axis. Далее выберите пере-

менную, значения каждой категории которой будут отображаться на графике в виде отдельных линий (пример см. ниже), и поместите ее в поле *Separate Lines*.

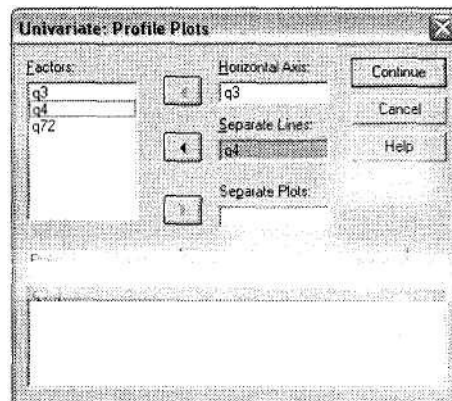


Рис. 3.20. Диалоговое окно *Univariate: Profile Plots*

Для иллюстрации процесса построения графиков предположим, что по результатам трехфакторного дисперсионного анализа была установлена статистическая значимость взаимодействия между переменными q3 (Пол) и q4 (Возраст). В окне *Profile Plots* мы поместили переменную с наименьшим числом категорий q3 в поле *Horizontal Axis*, а переменную q4 — в поле *Separate Lines*. Теперь щелкните на кнопке *Add*, чтобы подтвердить построение графика с заданными параметрами. Таким способом можно задать вывод сразу нескольких графиков.

После того как SPSS завершит расчеты, связанные с дисперсионным анализом, в окне *SPSS Viewer* после таблиц появится заданный график. В нашем примере он будет выглядеть так, как показано на рис. 3.21.

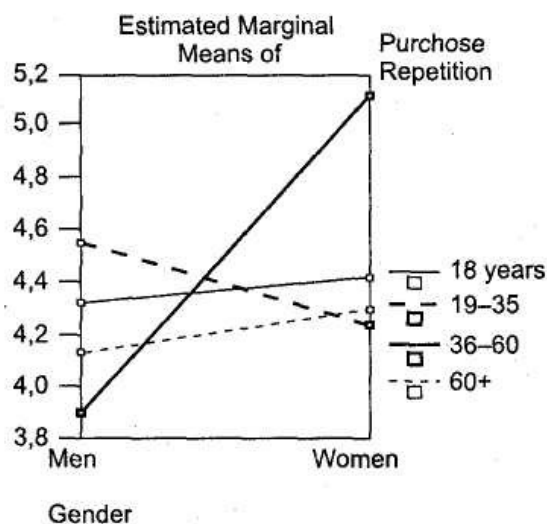


Рис. 3.21. График взаимодействия q3*q4

По оси ординат здесь (вертикальная ось) располагаются средние значения кратности покупок глазированных сырков каждой из рассматриваемых половозрастных групп. При этом на рисунке видно, что в возрастных группах от 36 до 60 лет и старше 60 лет кратность покупок сырков мужчинами и женщинами практически не различается (соответствующие линии близки к параллели), тогда как в других возрастных группах различие между мужчинами и женщинами выражено достаточно существенно (соответствующие линии перпендикулярны). Так, мужчины младше 18 лет характеризуются существенно

меньшей кратностью покупок сырков, чем женщины младше 18 лет. Мужчины в возрасте до 18 лет имеют наименьшую кратность покупок и по сравнению со всеми другими половозрастными группами. Мужчины в возрасте 19-35 лет характеризуются наивысшей кратностью покупок сырков среди всех возрастных групп мужчин. Можно заметить, что ситуация с женщинами в двух рассматриваемых возрастных группах диаметрально противоположная. Мужчины младше 18 лет имеют наименьшую кратность покупок; женщины младше 18 лет — наивысшую. Мужчины от 19 до 35 лет имеют наивысшую кратность покупок; женщины 19-35 лет — наименьшую.

Таким образом, вы видите, что графики в дисперсионном анализе являются весьма ценным ресурсом для построения заключений и выводов. Еще одним направлением интерпретации является кластеризация респондентов на основании их средних показателей (например, кратности покупок). Так, в нашем примере на основании кратности покупок можно разделить всех респондентов на следующие целевые сегменты:

1. мужчины младше 18 лет характеризуются наименьшей кратностью покупок сырков;
2. мужчины старше 36 лет и женщины старше 19 лет характеризуются средней кратностью покупок сырков;
3. мужчины от 19 до 35 лет и женщины младше 18 лет характеризуются наивысшей кратностью покупок сырков.

В целом общая схема интерпретации графиков в дисперсионном анализе состоит из двух этапов. Сначала следует определить категории респондентов, отличающиеся и не отличающиеся друг от друга. При этом интерпретация графиков всегда происходит только по двум переменным (представленным по горизонтальной оси и в виде отдельных линий). Для установления различия следует смотреть на форму данных линий. Если две (или более) линии близки к параллели, следовательно, различия между данными категориями минимальны (незначимы). В противном случае, если линии пересекаются, следует признать различие между ними существенным (значимым).

Наиболее простым для интерпретации случаем является ситуация, в которой по горизонтальной оси располагается дихотомическая переменная (например, переменная Пол). Если линии на отрезке между двумя категориями данной переменной не пересекаются — различий нет; если пересекаются — различия есть. На рис. 3.22 представлен пример максимальных различий (линии пересекаются под прямым углом); на рис. 3.23 — минимальных (линии параллельны).



Рис. 3.22. Значимые различия между взаимодействиями факторов



Рис. 3.23. Незначимые различия между взаимодействиями факторов

Можно сформулировать следующие рекомендации по построению графиков в дисперсионном анализе.

1. Для горизонтальной оси лучше выбирать дихотомические вопросы.
2. Если дихотомических переменных нет, следует выбрать переменную с наименьшим четным количеством категорий и перекодировать данные категории в дихотомию. Для горизонтальной оси следует выбирать именно данную (уже дихотомическую) переменную. Данный способ работает далеко не всегда, ведь часто различия между взаимодействиями факторов находятся именно в тех категориях, которые будут перекодированы (сокращены).

При исследовании трехуровневых взаимодействий ($q_1 \times q_2 \times q_3$) переменную с наименьшим числом категорий (лучше дихотомическую) следует поместить в поле *Separate Plots* в диалоговом окне *Univariate* (например, q_1), а для остальных двух исследуемых переменных (например, q_2 и q_3) — следовать вышеописанным правилам. Это будет означать, что в результате будут построены отдельные графики по каждой категории переменной q_1 , где будут показаны двухуровневые взаимодействия переменных q_2 и q_3 .

В заключение настоящего раздела необходимо особо отметить, что графики взаимодействий могут эффективно применяться только при числе взаимодействий 2 ($q_1 \times q_2$) или 3 ($q_1 \times q_2 \times q_3$). При взаимодействиях первого уровня (q_1) мы говорим уже не о взаимодействиях как таковых, а о главных эффектах (*Main effects*), то есть о влиянии на зависимую переменную только каждого фактора в отдельности. В таком случае различия между конкретными группами независимой переменной определяются исходя из результатов апостериорных тестов. При числе взаимодействий более трех сохраняется возможность разбиения данного взаимодействия на несколько взаимодействий второго или третьего уровней и построения затем серии графиков. Однако в этом случае интерпретация данных графиков является практически неразрешимой задачей.

3.2.2. Одномерный дисперсионный анализ с повторными измерениями

Одномерный дисперсионный анализ с повторными измерениями (ANOVARM) является расширением одномерного дисперсионного анализа (ANOVA). Цель его заключается в анализе различий между ответами одних и тех же респондентов на одни и те же вопросы в несколько приемов, то есть в течение ряда дискретных временных промежутков.

В качестве примера можно привести панельные исследования, когда одни и те же респонденты (потребители какого-либо продукта) отвечают на одни и те же вопросы через

определенные интервалы времени (скажем, каждый квартал). Одной из основных целей дисперсионного анализа в рассматриваемом случае будет оценка влияния на ответы респондентов временного фактора. Таким образом, в частности, можно установить уровень лояльности к продуктам различных марок: если с течением времени средние оценки продукта марки X существенно не меняются/возрастают/убывают, следовательно, и отношение респондентов к данной марке сохраняется на прежнем уровне/улучшается/ухудшается. Иными словами, дисперсионный анализ с повторными измерениями может применяться для оценки значимости тенденций.

В маркетинговых исследованиях этот тип статистического анализа находит весьма разнообразные применения. Он может применяться не только в процессе анализа баз данных по маркетинговым исследованиям, но и в процессе сбора анкет — для контроля работы интервьюеров. Например, если опрос производится каждый

день в течение недели в одних и тех же местах, можно анализировать средние значения основных переменных, во-первых, по дням недели, а во-вторых, по каждому интервьюеру. Если будут выявлены существенные различия в анкетах интервьюеров, то высока вероятность фальсификации (тем интервьюером, анкеты которого наиболее сильно отличаются от остальных).

Необходимо сделать важное отступление. Дело в том, что некоторые источники иногда относят анализ с повторными измерениями к одномерному, а иногда — к многомерному дисперсионному анализу. В справочной системе SPSS не указана явно принадлежность ANOVARМ к одной или другой группе статистических методов. По сути расчетов ANOVARМ близок к многомерному дисперсионному анализу, поскольку в качестве зависимой переменной выступают сразу несколько переменных, кодирующих ряд временных периодов. Но так как основная задача данного пособия — объяснение практических приемов работы с SPSS для эффективного применения этого программного продукта в маркетинговых исследованиях, мы отдаем предпочтение семантическому толкованию статистических терминов. Зависимые переменные в ANOVARМ по смыслу (с точки зрения исследователя) представляют собой фактически одну и ту же переменную, только измеренную многократно. В этой трактовке следует скорее говорить о специфической форме одномерного дисперсионного анализа, в котором зависимая переменная представлена набором подпеременных (точно так же, как при кодировании многовариантных вопросов; см. раздел 1.4.2). Таким образом, мы придерживаемся точки зрения тех авторов, которые считают ANOVARМ видом одномерного дисперсионного анализа (ANOVA).

Итак, в качестве иллюстрации использования одномерного дисперсионного анализа с повторяющимися измерениями рассмотрим следующий пример. Проводится исследование мнений респондентов относительно одежды марки X. Одним из вопросов анкеты является следующий: Поставьте оценку одежды марки X по пятибалльной шкале (от 1 — очень плохо до 5 — отлично). Респонденты разделяются на группы по полу и возрасту. Исследование проводится с частотой раз в квартал в течение года. В результате в итоговой базе данных получены три переменные: q18, q19 и q20, отражающие уровень оценки респондентами одежды марки X в первом, втором и третьем кварталах, а также две переменные, указывающие пол (q80) и возраст (q74) опрошенных. Требуется установить, как меняется общая картина восприятия респондентами одежды марки X в течение одного года. Поставленная задача легко решается методом одномерного дисперсионного анализа с повторяющимися измерениями. Откройте диалоговое окно Repeated Measures Define Factor(s) при помощи меню Analyze ► General Linear Model ► Repeated Measures (рис. 3.24). Это диалоговое окно предназначено для формирования временных факторов, то есть определения составных переменных, описывающих эти факторы. У нас есть три временных интервала (квартала), поэтому в поле Within-Subject Factor Name напишите название этой составной переменной: кварталы, а в поле Number of Levels — число временных периодов, когда производились измерения (3 квартала). После этого щелкните на кнопке Add, чтобы добавить новую составную переменную в список. Таким способом можно за-

дать сразу несколько составных временных переменных, однако в маркетинговых исследованиях в большинстве случаев ограничиваются только одной.

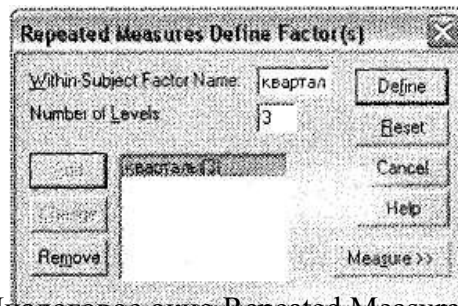


Рис. 3.24. Диалоговое окно Repeated Measures Define Factors

Кнопка Measure служит для задания дополнительных измерений временных переменных, но в маркетинговых исследованиях эта функция обычно не используется.

Щелкните на кнопке Define, и откроется новое диалоговое окно Repeated Measures (рис. 3.25), похожее (как по внешнему виду, так и по своим функциям) на окно Univariate. В этом окне в левом списке всех доступных переменных выберите те, в которых закодированы оценки респондентов в каждый из временных промежутков (в нашем случае — q18, q19, q20), и последовательно (то есть в порядке возрастания периодов) перенесите их в область Within-Subjects Variables (кварталы).

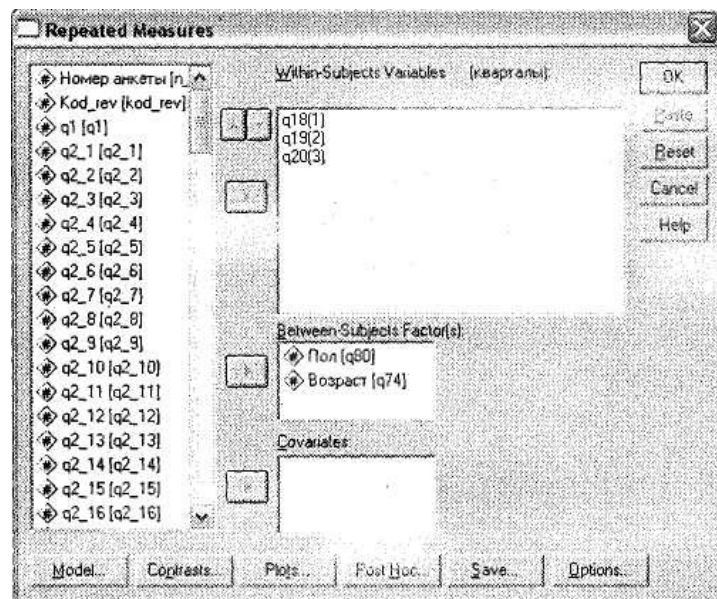


Рис. 3.25. Диалоговое окно Repeated Measures

Теперь в рассматриваемой области определена составная временная переменная, описывающая оценки респондентами одежды марки X в каждый из трех рассматриваемых кварталов. Таким образом, область Within-Subjects Variables (кварталы) является аналогом области Dependent Variable в одномерном дисперсионном анализе, только зависимая переменная в нашем случае как бы распадается на три подпеременные, вместе составляющие одно целое. Далее в область Between-Subjects Factor(s) поместите те переменные, которые служат основаниями для различения оценок. В нашем случае это демографические характеристики респондентов: пол (q80) и возраст (q74).

Итак, вы задали все переменные для исследования и можете использовать кнопки, расположенные в нижней части этого диалогового окна, — так же, как вы делали это при

одномерном дисперсионном анализе (см. раздел 3.2.1). В окне Post Hoc задайте апостериорные тесты Scheffe (для равных дисперсий) и Tukey (для неравных дисперсий) для переменных, имеющих более двух категорий (в нашем случае это только q74 — Возраст). В окне Options выберите параметр Homogeneity Tests и в соответствующее поле поместите переменные с двумя категориями, для которых следует рассчитать средние значения (q80 — Пол и все взаимодействия, в которых она участвует). Остальные диалоговые окна аналогичны рассмотренным для одномерного дисперсионного анализа, поэтому мы не приводим их второй раз.

В результате мы выясняем, какой из трех факторов — пол, возраст или время (кварталы) — определяет различия в оценках одежды марки X. Запустив программу на исполнение щелчком на кнопке ОК, в окне SPSS Viewer вы увидите результаты дисперсионного анализа. В целом они аналогичны результатам, отображаемым при одномерном дисперсионном анализе, однако данные результаты значительно обширнее и содержат несколько дополнительных таблиц. Так как настоящее пособие посвящено сугубо практическим задачам использования SPSS в маркетинговых исследованиях, мы рассмотрим только ту часть результатов, которая необходима на практике.

Итак, первое, что должно привлечь ваше внимание, — это таблица Box's Test of Equality of Covariance Matrices (рис. 3.26). Тестовая статистика Box показывает, существуют ли статистически значимые различия в оценках респондентов в каждом из анализируемых периодов. В нашем случае мы видим высокую значимость (Sig. < 0,001), свидетельствующую о том, что оценки респондентами одежды марки X существенно меняются от квартала к кварталу.

Box's Test of Equality of Covariance Matrices^a

Box's M	188,218
F	2,085
df1	90
df2	1,5E+07
Sig.	,000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a.

Design: Intercept+Q80+Q74+Q80 * Q74

Within Subjects Design: КВАРТАЛЫ

Рис. 3.26. Таблица Box's Test of Equality of Covariance Matrices

После анализа результатов теста Box мы смотрим на следующую важную таблицу — Multivariate Tests (рис. 3.27), позволяющую сделать выводы о том, в какой степени выявленные различия определяются влиянием временного фактора, а также взаимодействием этого фактора с другими переменными, включенными

в анализ. Так, в нашем случае мы видим, что непосредственно временной фактор (кварталы) в значительной степени определяет различия в исследуемых оценках (Sig. < 0,001). Сочетание эффектов времени и пола (Кварталы x q80), а также времени и возраста респондентов (Кварталы x q74) с высокой вероятностью определяют различия в оценках одежды (Sig. = 0,002 и 0,024). А вот тройственное взаимодействие всех анализируемых величин в совокупности не оказывает никакого влияния на изучаемую разницу в оценках (Sig. = 0,935). Обратите внимание на то, что при интерпретации таблицы Multivariate Tests

следует оценивать значимость того или иного фактора всегда на основании теста Pillai's Trace. Именно этот тест статистической значимости является наиболее надежным (робастным).

Multivariate Tests ^a						
Effect		Value	F	Hypothesis df	Error df	Sig.
КВАРТАЛЫ	Pillai's Trace	,212	1056,642 ^a	2,000	7872,000	,000
	Wilks' Lambda	,788	1056,642 ^a	2,000	7872,000	,000
	Hotelling's Trace	,268	1056,642 ^a	2,000	7872,000	,000
	Roy's Largest Root	,268	1056,642 ^a	2,000	7872,000	,000
КВАРТАЛЫ * Q80	Pillai's Trace	,002	6,337 ^a	2,000	7872,000	,002
	Wilks' Lambda	,998	6,337 ^a	2,000	7872,000	,002
	Hotelling's Trace	,002	6,337 ^a	2,000	7872,000	,002
	Roy's Largest Root	,002	6,337 ^a	2,000	7872,000	,002
КВАРТАЛЫ * Q74	Pillai's Trace	,003	1,874	14,000	15746,000	,024
	Wilks' Lambda	,997	1,875 ^a	14,000	15744,000	,024
	Hotelling's Trace	,003	1,875	14,000	15742,000	,024
	Roy's Largest Root	,003	3,178 ^b	7,000	7873,000	,002
КВАРТАЛЫ * Q80 * Q74	Pillai's Trace	,001	,500	14,000	15746,000	,935
	Wilks' Lambda	,999	,500 ^a	14,000	15744,000	,935
	Hotelling's Trace	,001	,499	14,000	15742,000	,935
	Roy's Largest Root	,001	,770 ^b	7,000	7873,000	,613

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c.

Design: Intercept+Q80+Q74+Q80 * Q74

Within Subjects Design: КВАРТАЛЫ

Рис. 3.27. Таблица Multivariate Tests

Мы ответили на два основных вопроса:

1. изменяются ли статистически значимо оценки респондентами одежды марки X?
2. чем определяются эти различия: только влиянием временного фактора или также влиянием независимых переменных (пола и возраста)?

В результате анализа мы смогли утвердительно ответить на оба вопроса: различия в оценках есть, и они определяются как временем, так и его взаимодействием с полом и возрастом. Дальнейший анализ будет направлен на исследование влияния независимых переменных и их взаимодействий по отдельности на оценки респондентов.

Следующие три таблицы: — Mauchly's Test of Sphericity, Tests of Within-Subjects Effects и Tests of Within-Subjects Contrasts — обычно пропускаются, так как они не позволяют сделать никаких новых выводов и лишь подтверждают представленные выше результаты. После трех таблиц следуют результаты одномерного дисперсионного анализа для независимых переменных, для которых не производятся повторные измерения, знакомые вам по разделу 3.2.1.

Таблица Levene's Tests of Equality of Error Variances (рис. 3.28) позволяет определить однородность дисперсий в каждый из исследуемых промежутков времени. Так, в нашем случае мы видим, что во всех трех исследуемых кварталах дисперсии однородны (Sig. > 0,05).

Levene's Test of Equality of Error Variances^a

	F	df1	df2	Sig.
Оценка одежды марки X (I квартал)	5,247	15	7873	,079
Оценка одежды марки X (II квартал)	1,352	15	7873	,162
Оценка одежды марки X (III квартал)	2,709	15	7873	,101

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a.

Design: Intercept+Q80+Q74+Q80 * Q74

Within Subjects Design: КВАРТАЛЫ

Рис. 3.28. Таблица Levene's Tests of Equality of Error Variances

Из таблицы Tests of Between-Subjects Effects (рис. 3.29) мы видим, что и пол, и возраст респондентов с весьма высокой вероятностью определяют различия в оценках одежды марки X (Sig. < 0,001), а вот их взаимодействие — нет (Sig. = 0,058).

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	276674,706	1	276674,706	215298,4	,000
Q80	41,593	1	41,593	32,367	,000
Q74	59,918	7	8,560	6,661	,000
Q80 * Q74	17,518	7	2,503	1,947	,058
Error	10117,402	7873	1,285		

Рис. 3.29. Таблица Tests of Between-Subjects Effects

Теперь нам осталось определить, как именно различаются оценки под влиянием выявленных значимых факторов и их взаимодействий. Во-первых, мы определили, что на различии в оценках респондентов в каждый из трех анализируемых периодов оказывают влияние пяти эффектов:

- временной фактор (кварталы);
- взаимодействие времени с полом;
- взаимодействие времени с возрастом;
- пол;
- возраст.

К сожалению, провести апостериорные тесты для временной переменной SPSS не позволяет, поэтому при определении различий между группами временной переменной приходится ориентироваться исключительно на средние значения (оценки). Возраст является единственной переменной, для которой можно провести стандартные апостериорные тесты (см. далее). Для остальных значимых взаимодействий выводятся средние значе-

ния: оценки одежды марки X в каждой рассматриваемой категории респондентов (см. рис. 3.28). Кроме таблиц для данных взаимодействий целесообразно вывести и графики. Это облегчит интерпретацию и позволит наглядно определить различия между категориями респондентов.

MEASURE_1

Возраст	N	Subset	
		1	2
Scheffe ^{a,1} 31-35 лет	1307	3,76	
36-40 лет	1343	3,77	
41-45 лет	1169	3,77	
Старше 55 лет	585	3,78	
51-55 лет	443	3,79	
46-50 лет	819	3,80	
25-30 лет	1440	3,83	
До 25 лет	783		3,97
Sig.		,651	1,000

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = ,428.

a. Uses Harmonic Mean Sample Size = 839,942.

b. Alpha = ,05.

Рис. 3.30. Таблица Homogeneous Subsets (вторая часть: MEASURE_1)

Завершают вывод результатов одномерного дисперсионного анализа с повторяющимися измерениями таблицы апостериорных тестов для переменных с числом категорий более двух¹. В нашем случае это две таблицы для переменной Возраст: Multiple Comparisons и Homogeneous Subsets. Первую таблицу мы не приводим из-за ее большого размера, вместо этого приведена дублирующая ее вторая таблица, показывающая однородные группы респондентов по оценкам одежды марки X

(рис. 3.30). Из таблицы вы видите, что наивысший уровень оценок достигается в возрастной группе респондентов (в среднем 4,0 балла) младше 25 лет. Респонденты старше 25 лет склонны оценивать одежду марки X несколько ниже (в среднем на 3,8 балла).

3.2.3. Многомерный дисперсионный анализ

Многомерный дисперсионный анализ является дальнейшим расширением одномерного дисперсионного анализа (после рассмотренного в разделе 3.2.2 ANOVARM), предназначенным для одновременного анализа сразу нескольких зависимых и независимых переменных. Процесс проведения многомерного анализа аналогичен рассмотренному выше обычному одномерному дисперсионному анализу, за исключением того, что в данном случае в область для зависимых переменных можно поместить сразу несколько переменных, а при интерпретации приходится анализировать сразу несколько различий (во всех зависимых переменных).

Давайте рассмотрим процесс проведения многомерного дисперсионного анализа на примере, аналогичном приведённому в разделе 3.2.1 для обычного одномерного дисперсионного анализа, — но в качестве зависимых переменных мы будем рассматривать не только кратность покупок глазированных сырков, но и частоту покупок. В качестве независимых переменных мы возьмем также две переменные: возраст респондентов и количество членов их семей.

Откройте диалоговое окно Multivariate при помощи меню Analyze ► General Linear Model ► Multivariate. Как вы видите на рис. 3.31, оно аналогично окну Univariate. По-

местите две зависимые переменные: q5 (Частота покупок) и q6 (Кратность покупок) в область для зависимых переменных Dependent Variables, а переменные q4 (Возраст) и q72 (Количество членов семьи) — в область для независимых переменных Fixed Factor(s). После этого так же, как для одномерного дисперсионного анализа в окне Post Hoc, задайте вывод тестов Scheffe и Tuhale для обеих независимых переменных, а в окне Options отметьте параметр Homogeneity Tests. После этого можно начать расчеты, щелкнув на кнопке ОК.

В окне SPSS Viewer появятся результаты многомерного дисперсионного анализа. Первой таблицей, которая должна привлечь ваше внимание, является Box's Test of Equality of Covariance Matrices, представленная на рис. 3.32. В отличие от одномерного дисперсионного анализа с повторяющимися измерениями, здесь тест Вох должен быть незначимым (как в нашем случае, Sig. = 0,131), так как неравенство дисперсий исследуемых зависимых переменных в многомерном анализе не является положительным фактом. И напротив, равенство дисперсий зависимых переменных является одним из основных условий проведения многомерного дисперсионного анализа¹.

Таблица Multivariate Tests позволяет сделать выводы относительно влияния независимых переменных в отдельности, а также их взаимодействий на зависимые переменные в целом. Поскольку с практической точки зрения влияние не несет никакой смысловой нагрузки, данная таблица обычно не рассматривается.

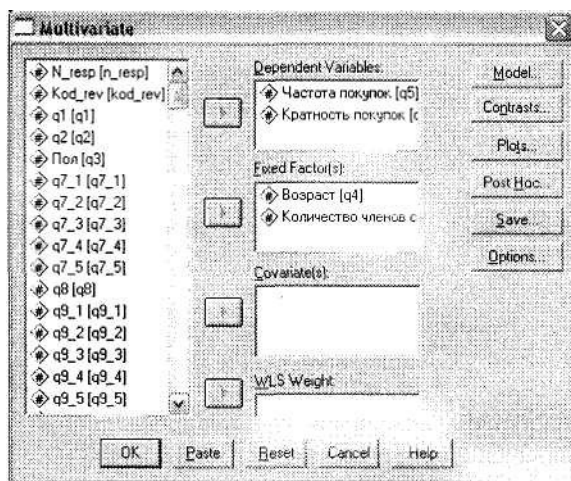


Рис. 3.31. Диалоговое окно Multivariate

Box's Test of Equality of Covariance Matrices^a

Box's M	68,638
F	1,219
df1	54
df2	16996,857
Sig.	,131

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+Q4+Q72+Q4 * Q72

Рис. 3.32. Таблица Box's Test of Equality of Covariance Matrices

Следующей важной таблицей является тест Levene на равенство дисперсий зависимых переменных. Как мы помним из описания одномерного дисперсионного анализа, от факта равенства/неравенства дисперсий в дальнейшем зависит выбор конкретного апостериорного теста: Scheffe или Tukey. Как вы видите на рис. 3.33, в нашем случае дисперсии равны у обеих зависимых переменных, поэтому далее мы будем опираться на результаты теста Scheffe.

Таблица Tests of Between-Subjects Effects (рис. 3.34) позволяет установить, как каждый эффект влияет на каждую зависимую переменную в отдельности. В отличие от таблицы Multivariate Tests, рассматриваемая таблица позволяет выяснить, на какую конкретно зависимую переменную влияет та или иная независимая переменная и их комбинации. В нашем случае мы видим, что частота покупок определяет различия между категориями переменной q4 Возраст (Sig. = 0,045), а кратность покупок — в категориях переменной q72 Количество членов семьи (Sig. < 0,001).

Levene's Test of Equality of Error Variances^a

	F	df1	df2	Sig.
Частота покупок	1,651	18	977	,142
Кратность покупок	1,172	18	977	,277

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Q4+Q72+Q4 * Q72

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Частота покупок	68,702 ^a	18	3,817	2,719	,000
	Кратность покупок	267,071 ^b	18	14,837	5,825	,000
Intercept	Частота покупок	2248,196	1	2248,196	1601,856	,000
	Кратность покупок	6810,265	1	6810,265	2673,592	,000
Q4	Частота покупок	11,364	3	3,788	2,699	,045
	Кратность покупок	3,223	3	1,074	,422	,737
Q72	Частота покупок	5,972	4	1,493	1,064	,373
	Кратность покупок	114,751	4	28,688	11,262	,000
Q4 * Q72	Частота покупок	27,264	11	2,479	1,766	,056
	Кратность покупок	12,334	11	1,121	,440	,938
Error	Частота покупок	1371,214	977	1,403		
	Кратность покупок	2488,647	977	2,547		
Total	Частота покупок	8092,000	996			
	Кратность покупок	21982,000	996			
Corrected Total	Частота покупок	1439,916	995			
	Кратность покупок	2755,719	995			

a. R Squared = ,048 (Adjusted R Squared = ,030)

b. R Squared = ,097 (Adjusted R Squared = ,080)

Рис. 3.34. Таблица Tests of Between-Subjects Effects

И наконец, последнее, что важно при практической интерпретации результатов многомерного дисперсионного анализа: какие группы каждой из рассматриваемых независимых переменных различаются на основании средних значений зависимых переменных.

Это позволяют определить апостериорные тесты (в нашем случае Scheffe). Они рассчитываются для каждой комбинации зависимая переменная/ независимая переменная для всех значений индексов i . Эти таблицы по своему виду аналогичны рассмотренным в предыдущих разделах, посвященных дисперсионному анализу.

Мы не приводим полностью результаты апостериорных тестов из-за их большого объема.

На рис. 3.35 представлены результирующие таблицы Homogeneous Subsets, по которым можно сделать выводы относительно различий между отдельными категориями независимых переменных на основании обеих рассматриваемых зависимых переменных. Также в этих таблицах вы видите однородные кластеры респондентов, различающиеся частотой и кратностью покупок глазированных сырков.

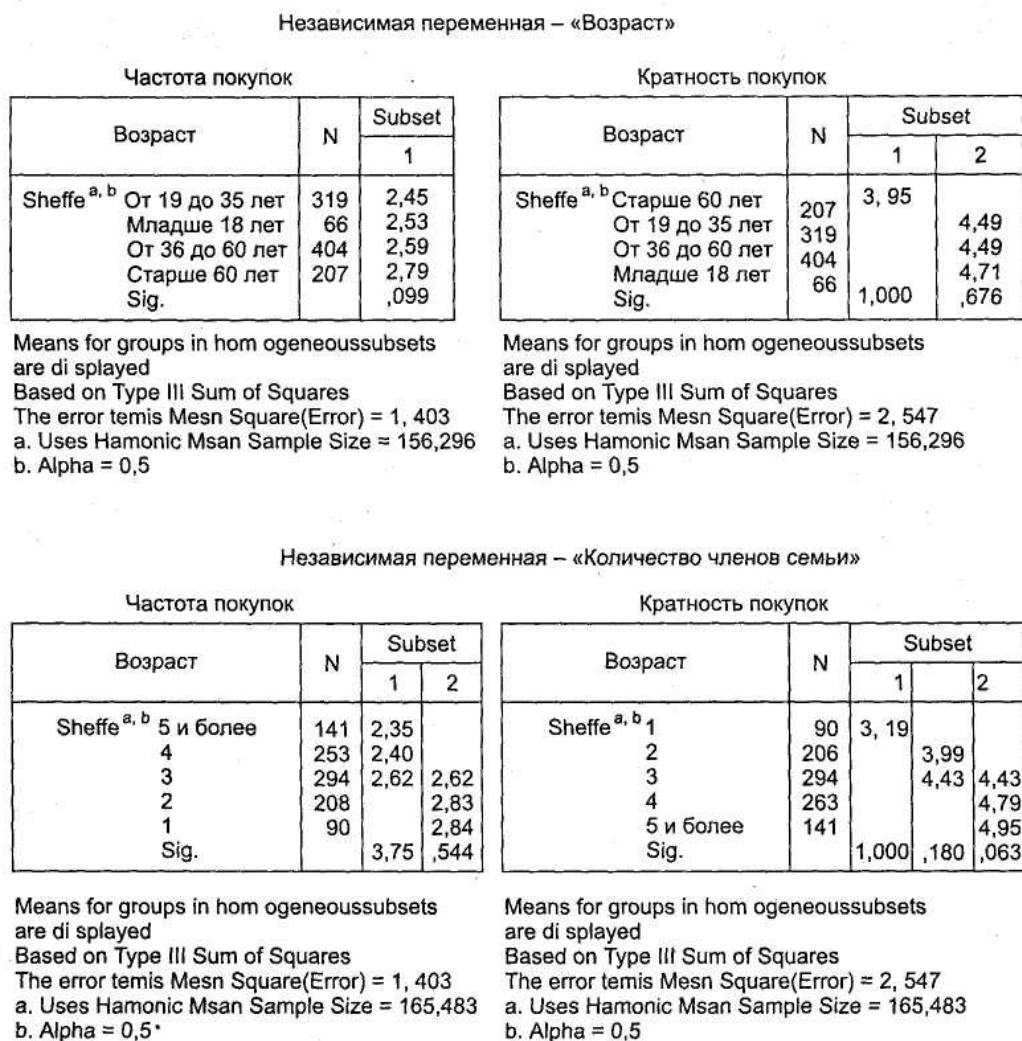


Рис. 3.35. Таблицы Homogeneous Subsets для переменных Возраст и Количество членов семьи

Итак, в данной главе мы рассмотрели статистические методы, применяемые для анализа различий между целевыми группами респондентов. Несмотря на то что данные методы (особенно обобщенная линейная модель) достаточно сложны для изучения, их применение позволяет поднять аналитическую работу на существенно более высокий уровень.

Глава 4 Ассоциативный анализ

Ассоциативный анализ служит для выявления связей между переменными. Применительно к маркетинговым исследованиям данная группа статистических процедур позволяет ответить на вопросы типа:

- Влияет ли на частоту посещения магазина уровень доходов покупателей?
- Как связаны между собой пол респондентов и желание купить мотоцикл?
- Как влияет на покупательское поведение потребителей сухих строительных смесей род занятий респондентов?

То есть при помощи ассоциативного анализа становится возможным анализировать вопросы анкеты не только по отдельности, а в зависимости от других вопросов. Этот вид анализа иногда называют построением разрезов, поскольку он позволяет определить не только наличие связи между вопросами анкеты, но и силу связи между переменными и то, каким образом ведет себя одна переменная при изменении другой (возрастает или убывает).

В процессе ассоциативного анализа выявляются следующие типы зависимостей.

■ **Немонотонные** зависимости свидетельствуют только о наличии определенной связи между двумя переменными, но не позволяют судить о направлении или силе связи. Пример немонотонной зависимости: мужчины в основном покупают рыбные консервы в продовольственных магазинах, а женщины — на рынках.

■ **Монотонные** зависимости — это зависимости, по которым можно узнать не только наличие, но и направление связи. Пример монотонной зависимости: мужчины покупают пиво чаще, чем женщины. Монотонные зависимости бывают двух видов:

- возрастающие — первая переменная возрастает при возрастании второй;
- убывающие — первая переменная убывает при возрастании второй.

■ **Линейные** зависимости характеризуются уравнением функции $y = a + b \cdot x$ (график линейной функции). Связь между двумя переменными в данном случае является линейной, то есть на основании этой зависимости мы можем сказать, насколько изменится одна переменная при изменении второй.

■ **Нелинейные**. Примерами нелинейных связей между двумя переменными являются: экспоненциальная, логарифмическая, степенная, полиномиальная зависимости — то есть в данном случае связь присутствует и изменяется по какому-либо известному математическому закону.

Зависимости, выявленные в результате ассоциативного анализа, можно охарактеризовать тремя аспектами.

- По наличию — определенная (систематическая) связь между двумя переменными есть.
- По направлению — связь является убывающей или возрастающей.
- По силе — можно определить, насколько тесно связаны между собой две переменные, то есть насколько значима данная зависимость.

Между переменными с номинальной шкалой может быть установлена только немонотонная зависимость, характеризующаяся только наличием связи. Для переменных, имеющих порядковую или интервальную шкалу, данное ограничение не действует — для них можно определить и направление, и силу связи.

4.1. Перекрестные распределения и χ^2

Перекрестные распределения служат для выявления различных типов зависимостей между двумя и более переменными. Например, если требуется установить, где покупают сгущенное молоко мужчины и женщины, следует воспользоваться таблицами перекрестных распределений (таблицами сопряженности, или кросстабуляции). На основании

перекрестных распределений можно установить не только наличие зависимости (немонотонной или монотонной) между переменными, но, в большинстве случаев, ее тип (линейная или нелинейная) и направление (возрастающая или убывающая)¹. Установленная при помощи перекрестного распределения зависимость может оказаться незначимой из-за малого размера выборки или по другим причинам. Статистическую значимость выявленной зависимости позволяет определить критерий χ^2 .

В табл. 4.1 представлены основные характеристики переменных, участвующих в анализе.

Несмотря на то что перекрестные табуляции можно строить по переменным, имеющим любой тип шкалы, необходимо иметь в виду, что большое количество категорий (вариантов ответа) анализировать трудно. Даже если анализ выявит значимую зависимость, при наличии большого числа категорий переменных исследователю будет сложно понять, каким именно образом связаны данные переменные.

Таблица 4.1. Основные характеристики переменных, участвующих в перекрестных распределениях

Перекрестные распределения			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
От двух до десяти	Любой	От двух до десяти	Любой

Также следует отметить, что наибольшую эффективность кросстабуляционный анализ показывает на номинальных и порядковых переменных. Для интервальных переменных больше подходит корреляционный анализ, рассматриваемый в разделе 4.2.

И наконец, последним ограничением применения перекрестных распределений для анализа зависимостей между переменными является тот факт, что различные статистические тесты (такие как χ^2) могут быть использованы только при анализе одновариантных переменных. Статистические тесты, применяемые для анализа зависимостей, предназначены только для двух переменных. При наличии дополнительных слоев или уровней кросстабуляционной таблицы статистический анализ производится для каждого уровня отдельно, при этом на каждом уровне он работает только с двумя переменными. Для многовариантных переменных SPSS содержит возможность отдельного построения кросстабуляции — выявить наличие и направление связи в данном случае можно только визуально.

Далее в этой главе мы покажем, как строить перекрестные распределения и анализировать зависимости для одновариантных и многовариантных переменных.

4.1.1. Перекрестные распределения для одновариантных вопросов и χ^2

Давайте рассмотрим перекрестные распределения для одновариантных вопросов на следующем примере.

ПРИМЕР-----

Исходные данные:

В результате маркетингового исследования, посвященного исследованию потребительских предпочтений посетителей развлекательного центра, оказалось, что средняя частота посещения центра составляет приблизительно 12 раз в месяц. Также были получены данные о распределении среди посетителей центра мужчин и женщин различных возрастных групп. В ходе подготовительного этапа анализа были сформированы, в частности, три одновариантные переменные:

- 1) частота посещения центра (q25);
- 2) возраст респондентов (q18);

3) пол респондентов (q23). Требуется:

1. Построить перекрестное распределение частоты посещения развлекательного центра в разрезе возраста и пола респондентов. Рассчитать среднюю частоту посещения центра различными целевыми группами потребителей.

2. Определить, влияет ли на частоту посещения центра возраст потребителей. Установить статистическую значимость зависимости между частотой посещения и возрастом.

Из условия первой задачи следует, что мы должны построить перекрестное распределение сразу по трем переменным: q25 в зависимости от q18 и q23 (то есть трехуровневое). Для решения задачи воспользуемся меню Analyze ► Descriptive Statistics ► Crosstabs. В открывшемся диалоговом окне (рис. 4.1) из левого списка, содержащего все доступные переменные, выберите те, которые будут расположены в строках результирующей таблицы, и те, которые будут расположены в столбцах. Поместите зависимую переменную q25 Частота посещения в область Rows (варианты ответа на вопрос о частоте посещения будут расположены в строках таблицы), а независимую переменную q18 Возраст — в область Columns (возрастные группы будут расположены в столбцах таблицы). Осталась еще одна независимая переменная q23 Пол. Поместите ее в область Layer (уровень или слой таблицы).

Обратите внимание, что всегда, когда обратное не обусловлено задачами исследования, рекомендуется размещать переменные с малым количеством вариантов ответа в слоях. Это позволит уменьшить размерность результирующей таблицы. Мы можем задать и большее количество измерений таблицы, щелкая на кнопке Next в области Layer и добавляя релевантные переменные. Максимальное количество слоев, которое можно задать, щелкая на кнопке Next, — 8. Следовательно, максимально возможное количество измерений перекрестной таблицы по одновариантным вопросам — $10(10 = 8 \text{ слоев} + 1 \text{ строковая переменная} + 1 \text{ столбцовая переменная})$.

В диалоговом окне Crosstabs в область каждого измерения (Rows, Columns, Layer) можно поместить сразу несколько переменных. Максимальное число переменных, которые можно поместить в области Rows и Columns, — 76; для каждого из восьми возможных уровней Layer — 6. Если задано по одной переменной в строке и столбце (как в нашем случае), все дополнительно указанные слои будут отображаться в одной и той же таблице. Ситуация будет отличаться, если мы укажем несколько переменных для строк, столбцов и слоев в одних и тех же областях (не щелкая на кнопке Next для задания нескольких слоев) перекрестной таблицы. В этом случае будут построены отдельные таблицы для каждой пары строковых и столбцовых переменных.

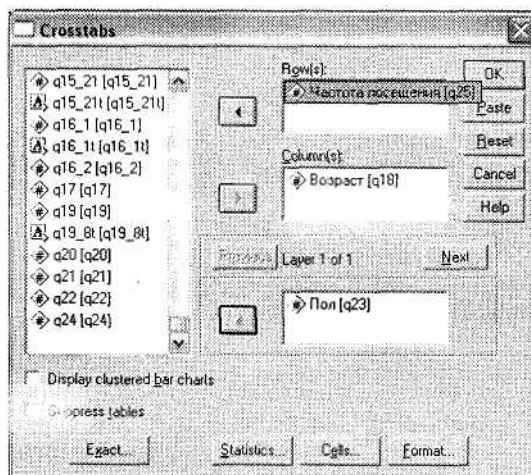


Рис. 4.1. Диалоговое окно Crosstabs

Теперь, когда вы указали все переменные для анализа, для построения перекрестных распределений можно щелкнуть на кнопке ОК. Однако сначала давайте рассмотрим некоторые другие полезные функции диалогового окна Crosstabs. Щелкните на кнопке Cells. Отрывшееся диалоговое окно Cell Display (рис. 4.2) предназначено для задания значений, выводимых в кросстабуляционной таблице. По умолчанию SPSS в каждой ячейке таблицы выводит только количество респондентов (параметр Observed). Область Percentages позволяет организовать вывод в ячейках таблицы процентов по строкам (Rows), столбцам (Columns), а также от общего числа респондентов, ответивших одновременно на все вопросы, по которым строится перекрестное распределение (Частота посещения, Возраст и Пол) (Total).

Чтобы проиллюстрировать наш пример (расчет средних частот покупки), выведем проценты по вопросу Частота посещения внутри каждой возрастной и половой группы респондентов, отметив параметр Columns и проценты по всем возрастным группам в целом (Total). Также оставим выбранный по умолчанию вывод наблюдаемых частот (Observed). После этого можно закрыть окно Cell Display, щелкнув на кнопке Continue.

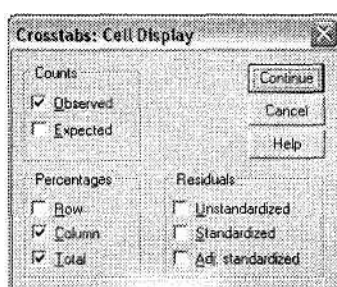


Рис. 4.2. Диалоговое окно Cell Display
Crosstabs

Следующее диалоговое окно, которое мы рассмотрим, — это Table Format, вызываемое при помощи кнопки Format (рис. 4.3). В нем можно выбрать тип сортировки вариантов ответа строковой переменной: возрастающая или убывающая (по алфавиту). Оставьте выбранный по умолчанию вариант Ascending (возрастающая) и щелкните на кнопке Continue, чтобы закрыть окно. После этого запустите процедуру построения перекрестных распределений, щелкнув на кнопке ОК в главном диалоговом окне Crosstabs. В главном диалоговом окне процедуры есть и другие полезные функции: мы рассмотрим их ниже.

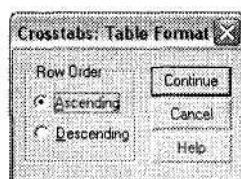


Рис. 4.3. Диалоговое окно Table Format

После этого в окне SPSS Viewer будет выведена требуемая таблица перекрестного распределения (рис. 4.4). В ячейках данной таблицы находятся искомые частоты

посещения развлекательного центра каждой из анализируемых целевых групп опрошенных. Например, первая ячейка показывает, что 5 (строка Count) респондентов-мужчин в возрасте от 18 до 25 лет посещают развлекательный центр каждый день. Это составляет 8,1% (подстрока % within Возраст) от общего количества мужчин в возрасте от 18 до 25 лет, ответивших на три предложенных вопроса, или 1,5% (подстрока % of Total) от общего числа мужчин, ответивших на вопросы (это число 333, оно представлено на пересечении строки и столбца Total в первой части таблицы Мужчины).

Строка Total показывает, сколько всего мужчин из каждой возрастной группы ответили на вопрос о частоте посещения центра (в нашем случае 62 респондента-мужчины в возрасте от 18 до 25 лет). Столбец Total показывает, сколько всего мужчин, посещающих развлекательный центр с различной частотой, ответили на вопрос о возрасте (в нашем случае 15 респондентов-мужчин, посещающих центр каждый день).

Вторая часть таблицы Женщины построена аналогичным образом. Как вы видите, 15,8% женщин в возрасте от 41 до 45 лет посещают развлекательный центр 1-2 раза в месяц.

Частота посещения * Возраст * Пол Crosstabulation												
Пол				Возраст							Total	
				От 18 до 25 лет	От 26 до 30 лет	От 31 до 35 лет	От 36 до 40 лет	От 41 до 45 лет	От 46 до 50 лет	От 51 до 55 лет		Старше 55 лет
Мужчины	Частота посещения	Каждый день	Count	6			2	4	1	1	2	16
			% within Возраст	8,1%			4,9%	10,0%	2,9%	3,7%	3,3%	4,5%
			% of Total	1,9%			,6%	1,2%	,3%	,3%	,6%	4,5%
	Total		Count	62	37	31	41	40	35	27	60	333
			% within Возраст	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
			% of Total	18,6%	11,1%	9,3%	12,3%	12,0%	10,5%	8,1%	18,0%	100,0%
Женщины	Частота посещения	Каждый день	Count	1		3	3	1	2		1	11
			% within Возраст	1,8%		7,7%	7,0%	2,6%	4,4%		1,3%	2,9%
			% of Total	,3%		,8%	,8%	,3%	,5%		,3%	2,9%
		1-2 раза в месяц	Count	3	3	1	3	6	1	1	8	26
			% within Возраст	5,3%	6,4%	2,6%	7,0%	15,8%	2,2%	3,1%	10,3%	6,9%
			% of Total	,8%	,8%	,3%	,8%	1,6%	,3%	,3%	3,1%	6,9%
	Total		Count	67	47	39	43	38	45	31	77	378
			% within Возраст	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
			% of Total	15,1%	12,4%	10,3%	11,4%	10,1%	11,9%	8,5%	20,4%	100,0%

Рис. 4.4. Таблица перекрестного распределения по трем вопросам: Частота посещения, Возраст и Пол

На основании представленной таблицы перекрестного распределения вы можете рассчитать вручную средневзвешенные частоты посещения респондентами развлекательного центра в зависимости от их пола и возраста. Для этого скопируйте анализируемую таблицу в Microsoft Excel, щелкнув на ней правой кнопкой мыши в окне SPSS Viewer и выбрав пункт Copy (не Copy Objects !). Окончательный вид полученного распределения представлен в табл. 4.2.

Таблица 4.2. Средневзвешенные частоты посещения развлекательного центра в зависимости от пола и возраста респондентов (раз в месяц)

Пол	Возраст									
	От 18 до 25 лет	От 26 до 30 лет	От 31 до 35 лет	От 36 до 40 лет	От 41 до 45 лет	От 46 до 50 лет	От 51 до 55 лет	Старше 55		
Мужчины	12	12	12	12	13	13	9	10		
Женщины	11	12	14	12	10	12	11	12		

Из представленной таблицы следует, что средняя частота посещения развлекательного центра составляет 12 раз в месяц.

тельного центра различными половозрастными группами респондентов несколько различается. Однако, исходя только из визуальных предположений, нельзя утверждать то, что частота посещения центра действительно зависит от пола и возраста. Для этого любая выявленная закономерность должна удовлетворять условию статистической значимости. Определить, значима ли выявленная нами зависимость, позволяют статистические тесты, выполняемые при построении перекрестных распределений.

Далее мы покажем, как решается второй пункт нашей задачи (условие см. выше), то есть как ответить на вопрос: «Действительно ли существует статистически значимая зависимость между тремя анализируемыми переменными или показанные в табл. 4.2 различия в частотах посещения центра вызваны влиянием случайных факторов (то есть как таковой зависимости нет)?».

Выявить статистическую значимость зависимостей между переменными позволяют критерий χ^2 и сопутствующие тесты. Исследуем нашу зависимость между частотой посещения развлекательного центра, полом и возрастом респондентов на предмет статистической значимости. Для этого вновь откройте диалоговое окно Crosstabs. В этом окне остались две не рассмотренные нами кнопки: Exact и Statistics — именно они позволяют исследовать значимость перекрестных распределений. По умолчанию SPSS определяет статистическую значимость только на основании асимптотического метода. Одной из разновидностей данного метода и является χ^2 . Данный критерий используется наиболее часто в маркетинговых исследованиях. Однако применение асимптотического критерия χ^2 накладывает на данные, содержащиеся в анализируемой перекрестной таблице, существенные ограничения, которые подробно описаны ниже.

Так, важнейшим требованием к исследуемым данным является достаточно большие значения в ячейках таблицы. При наличии небольших по размеру выборок или при построении разрезов третьего и более уровня данное условие является недостижимым. Исходя из опыта анализа данных в маркетинговых исследованиях, можно утверждать, что подобные ситуации встречаются достаточно часто. В связи с этим в случае несоответствия имеющихся данных общепринятому критерию χ^2 следует воспользоваться другими статистическими методами.

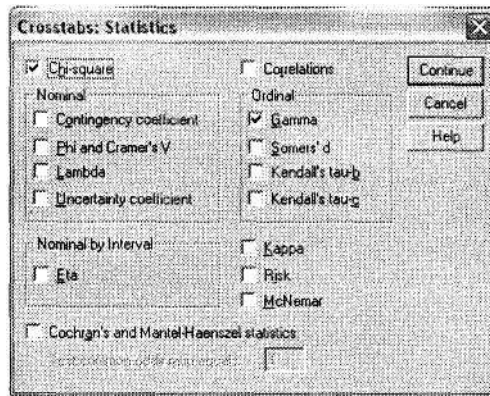
Сначала на примере перекрестного распределения по трем переменным рассмотрим использование наиболее популярного статистического метода установления статистической значимости зависимостей — критерия χ^2 . Для того чтобы организовать наряду с перекрестной таблицей вывод соответствующих статистик, в главном диалоговом окне Crosstabs щелкните на кнопке Statistics (рис. 4.5). В открывшемся диа-

логовом окне выберите параметр Chi-square χ^2). Это позволит впоследствии определить, имеется ли определенная связь между исследуемыми переменными.

При анализе зависимостей, кроме обнаружения наличия связи, также можно определить, насколько сильно выражена данная зависимость (установить силу связи). Сделать это позволяют релевантные статистические тесты, применяемые отдельно для каждого из трех типов переменных, участвующих в анализе. Для номинальных переменных следует применять один из тестов, представленных в области Nominal. Наиболее универсальным и часто применяемым методом является V Cramer's. Для порядковых переменных следует применять один из методов, представленных в области Ordinal. Мы рекомендуем использовать наиболее универсальный метод: Gamma. Теоретически этот же метод можно применять и для интервальных переменных, однако все же для них рекомендуется использовать более релевантную процедуру корреляционного анализа.

Далее рассмотрим, как применять перечисленные статистические методы на примере нашей задачи с двумя порядковыми переменными Частота посещения развлекательного центра и Возраст. Для этого выберите параметр Gamma и закройте описываемое окно, щелкнув на кнопке Continue. Запустите процедуру построения перекрестных распределений, щелкнув на кнопке ОК в главном диалоговом окне Crosstabs.

Рис. 4.5. Диалоговое окно Statistics



В окне SPSS Viewer появится уже рассмотренная выше таблица перекрестного распределения трех переменных: Частота посещения, Возраст и Пол. Но, в отличие от предыдущего случая, ниже будут отображены две таблицы, из которых можно узнать о наличии, силе и направлении (только для порядковых и интервальных переменных) связи между анализируемыми переменными. Рассмотрим их по порядку.

В первой таблице, Chi-Square Tests, выводятся результаты расчета критерия χ^2 (строка Pearson Chi-Square) и некоторых других статистик (рис. 4.6). Необходимо отметить, что расчет всех статистических процедур производится для каждого варианта переменной, расположенной в слоях (в нашем случае Пол) по отдельности (то есть отдельно для целевых групп мужчин и женщин). Данное обстоятельство уже было отмечено выше.

В нашем примере для респондентов-мужчин величина критерия χ^2 — 56,048, однако для практических целей важна не столько сама величина, сколько ее значимость, представленная в столбце Asymp. Sig. (2-sided). Именно из условия статистической значимости критерия χ^2 следует статистическая значимость всей зависимости. В нашем примере значимость анализируемого критерия и для мужчин, и для женщин достаточно высока (0,001 и 0,014 соответственно), что позволяет сделать предварительный вывод о том, что между частотой посещения развлекательного центра и возрастом для каждой половой группы респондентов существует определенная статистически значимая зависимость. Тем не менее одной значимости критерия χ^2 недостаточно, чтобы с уверенностью утверждать о наличии значимой зависимости между тремя анализируемыми переменными. Для этого необходимо, чтобы выполнялись следующие два критерия.

Процент ячеек, в которых ожидаемые значения¹ (Expected counts) меньше или равны 5, должен быть менее или равным 20 %. Это значение отображается в примечании «а» в первой строке после таблицы Chi-Square Tests. На практике приемлемая доля ожидаемых частот меньше 5 может отклоняться от 20 % (в пределах +5 %). При наличии ярко выраженной зависимости можно считать такую зависимость статистически значимой. Также всегда необходимо иметь в виду практические соображения (и это относится ко всем без исключения статистическим процедурам). Если ожидаемые частоты меньше 5 у переменных, представляющих малую практическую значимость для исследователя, — значит, можно не принимать в расчет рассматриваемый критерий и признать зависимость значимой по практическим соображениям. Как видно на рис. 4.58, в нашем случае 55 % ячеек имеют ожидаемые значения меньше 5 (при этом минимальное ожидаемое значение 0,32). Следовательно, несмотря на то что критерий χ^2 является статистически значимым, он не удовлетворяет рассматриваемому дополнительному условию.

Суммы по строкам и столбцам должны быть больше 0. В нашем случае данное условие удовлетворяется.

Еще одной не рассмотренной статистикой в таблице Chi-Square Tests является тест Mantel-Haenszel (строка Linear-by-Lf near Association). Его значимость позволяет сделать вывод о наличии линейной зависимости между неноминальными переменными. Если величина данного теста статистически значима, следовательно, между строковой и столб-

цовой переменными есть линейная зависимость. В нашем случае (рис. 4.6) линейная зависимость между возрастом и частотой посещения развлекательного центра существует только в целевой группе респондентов-женщин. Про мужчин подобное сказать нельзя.

После того как мы установили наличие зависимости между тремя анализируемыми переменными (при этом между возрастом и частотой посещения для респондентов-женщин существует и линейная зависимость), можно приступить к анализу таблицы Symmetric Measures (рис. 4.7), чтобы определить силу выявленной связи.

Chi-Square Tests

Пол		Value	df	Asymp. Sig. (2-sided)
Мужчины	Pearson Chi-Square	56,048 ^a	28	,001
	Likelihood Ratio	56,557	28	,001
	Linear-by-Linear Association	3,532	1	,060
	N of Valid Cases	333		
Женщины	Pearson Chi-Square	46,844 ^b	28	,014
	Likelihood Ratio	47,776	28	,011
	Linear-by-Linear Association	5,500	1	,019
	N of Valid Cases	378		

a. 22 cells (55,0%) have expected count less than 5. The minimum expected count is ,32.

b. 23 cells (57,5%) have expected count less than 5. The minimum expected count is ,17.

Рис. 4.6. Таблица Chi-Square Tests

Для порядковых переменных (как в нашем случае) определить силу связи позволяет критерий Gamma. Этот показатель может варьироваться в интервале от -1 (максимально разнонаправленная зависимость) до 1 (полная зависимость); значение 0 показывает полное отсутствие зависимости. Значение критерия Gamma представлено в столбце Value таблицы Symmetric Measures. В нашем случае в группе респондентов-мужчин имеется лишь весьма слабая положительная зависимость ($\text{Gamma} = 0,080$). Столбец Approx. Sig. свидетельствует о том, что данная зависимость еще и статистически незначима. Обратная ситуация в группе респондентов-женщин: для них установлена слабая, но статистически значимая положительная зависимость между возрастом и частотой посещения развлекательного центра.

Symmetric Measures

Пол			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Мужчины	Ordinal by Ordinal	Gamma	,080	,067	1,193	,233
	N of Valid Cases		333			
Женщины	Ordinal by Ordinal	Gamma	,148	,057	2,592	,010
	N of Valid Cases		378			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Рис. 4.7. Таблица Symmetric Measures для порядковых переменных

Если в перекрестном анализе участвуют номинальные переменные, силу (но не направление) связи позволяет определить критерий Cramer's V. Отображение этого критерия можно установить в диалоговом окне Statistics при помощи параметра Phi and Cramer's V (см. рис. 4.5).

Давайте рассчитаем данный критерий для наших переменных. Результаты расчетов представлены на рис. 4.8. В целом, критерий Cramer's V может варьироваться в пределах от 0 до 1, где 0 показывает отсутствие связи между исследуемыми переменными, а 1 — полную зависимость. В нашем случае и для мужчин, и для женщин есть статистически значимые (как показывает столбец Approx. Sig.) слабые зависимости (для мужчин Cramer's V = 0,205; для женщин = 0,176). Необходимо отметить, что значение 1 для теста Cramer's V является практически недостижимым, поэтому значения 0,8-0,9 следует считать весьма высокими.

Symmetric Measures				
Пол			Value	Approx. Sig.
Мужчины	Nominal by	Phi	,410	,001
	Nominal	Cramer's V	,205	,001
	N of Valid Cases		333	
Женщины	Nominal by	Phi	,352	,014
	Nominal	Cramer's V	,178	,014
	N of Valid Cases		378	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Рис. 4.8. Таблица Symmetric Measures для номинальных переменных (пример)

Итак, мы определили, что между тремя анализируемыми переменными — возрастом, полом и частотой посещения респондентами развлекательного центра — есть слабые, но статистически значимые зависимости. Вместе с тем было установлено, что больше половины (55 %) ячеек в перекрестной таблице имеют ожидаемые частоты меньше 5 — из чего следует вывод о неприменимости теста χ^2 и сопутствующих асимптотических тестов (Gamma и Cramer's V) в нашем случае. В принципе мы ответили на второй пункт задачи (условие см. выше) и можем сказать, что различия, выявленные в ходе перекрестного анализа (см. табл. 4.2), действительно имеют место и являются статистически значимыми. Однако добросовестный аналитик в такой ситуации все же попытается доказать истинность сделанных выводов.

Когда анализируемые данные не удовлетворяют требованиям, предъявляемым асимптотическими методами (как, например, в нашем случае χ^2), есть другая возможность установить статистическую значимость исследуемой зависимости. Это позволяет сделать точные (Exact) тесты.

Откройте главное диалоговое окно перекрестного анализа Crosstabs (см. рис. 4.1), щелкнув на кнопке Exact. В появившемся диалоговом окне Exact Tests (рис. 4.9) по умолчанию установлен расчет только асимптотических критериев. Данное диалоговое окно позволяет провести расчеты по двум неасимптотическим методам: Monte-Carlo и Exact, причем последний метод не рекомендуется использовать в практических целях, так как он занимает много времени. Для практических целей следует применять метод Monte-Carlo с установленным по умолчанию количеством выборок (10 000). Доверительный уровень 99 % практически всегда является слишком высоким, поэтому измените его на 95 %, что соответствует доверительному уровню при расчете статистической ошибки выборки для маркетинговых исследований (см. раздел 1.2). Все остальные параметры диалогового окна

Crosstabs аналогичны указанным в предыдущем примере. Теперь можно запустить процедуру построения перекрестных распределений.

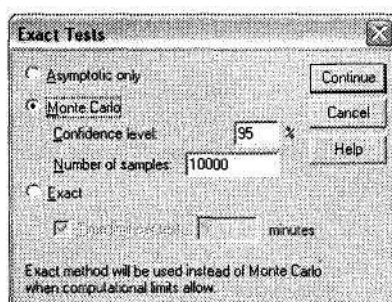


Рис. 4.9. Диалоговое окно Exact Tests

После завершения всех необходимых расчетов в окне SPSS Viewer будут выведены результаты. Их структура аналогична рассмотренной выше, за исключением того, что таблицы Chi-Square Tests и Symmetric Measures расширены за счет результатов теста Monte-Carlo. Единственным практическим результатом данного теста является рассчитанная статистическая значимость критериев, указанных в диалоговом окне Statistics (см. рис. 4.5).

На рис. 4.10 представлена таблица Chi-Square Tests с результатами теста Monte-Carlo. Искомые значения статистической значимости представлены в столбце Monte Carlo Sig. (2-sided) в подстолбце Sig.. В подстолбцах Lower Bound и Upper Bound показаны, соответственно, нижний и верхний пределы, в которых варьируется значение статистической значимости Sig.. Так, в нашем случае критерий χ^2 действительно свидетельствует о наличии статистически значимой зависимости между полом, возрастом и частотой посещения развлекательного центра — это следует из весьма высокой значимости теста Monte-Carlo (0,001 — для мужчин и 0,012 — для женщин). В 95 % случаев данное значение не выходит за рамки статистической значимости (например, для мужчин оно варьируется в пределах от 0,001 до 0,002). Также из таблицы мы видим, что выявленная связь является линейной только для целевой группы респондентов-женщин. Таким образом, для нашего случая все предварительные выводы, сделанные нами в таблице Chi-Square Tests, подтвердились результатами теста Monte-Carlo.

Chi-Square Tests										
Пол		Value	df	Asymp. Sig (2-sided)	Monte Carlo Sig. (2-sided)			Monte Carlo Sig. (1-sided)		
					Sig.	95% Confidence Interval		Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound		Lower Bound	Upper Bound
Мужчины	Pearson Chi-Square	56,040 ^a	28	,001	,001 ^b	,001	,002			
	Likelihood Ratio	56,557	28	,001	,001 ^b	,000	,002			
	Fisher's Exact Test	42,855			,006 ^b	,004	,008			
	Linear-by-Linear Association	3,532 ^c	1	,060	,000 ^b	,055	,065	,029 ^b	,025	,032
	N of Valid Cases	333								
Женщины	Pearson Chi-Square	46,844 ^a	28	,275	,012 ^b	,010	,014			
	Likelihood Ratio	47,776	28	,630	,007 ^b	,005	,008			
	Fisher's Exact Test	44,250			,002 ^b	,001	,003			
	Linear-by-Linear Association	5,500 ^d	1	,917	,020 ^b	,017	,023	,009 ^b	,007	,011
	N of Valid Cases	378								

a. 22 cells (55,0%) have expected count less than 5. The minimum expected count is ,32.

b. Based on 10000 sampled tables with starting seed 2000000.

c. The standardized statistic is 1,879.

d. The standardized statistic is 2,347.

Рис. 4.10. Таблица Chi-Square Tests с результатами теста Monte-Carlo

Теперь рассмотрим таблицу Symmetric Measures (рис. 4.11), на основании которой мы сделали выводы о силе выявленной зависимости. Результаты теста Monte-Carlo и в данном случае подтверждают выводы асимптотического метода: между частотой посещения центра и возрастом в целевой группе респондентов-женщин выявлена слабая статистически значимая зависимость. Для мужчин зависимость статистически незначима.

Symmetric Measures									
			Value	Asymp. Std. Error ^a	Approx. χ^2	Approx. Sig.	Monte Carlo Sig.		
							Sig.	95% Confidence Interval	
								Lower Bound	Upper Bound
Мужчины	Ordinal by Ordinal Gamma		,080	,067	1,193	,233	,102 ^c	,174	,169
	N of Valid Cases		333						
Женщины	Ordinal by Ordinal Gamma		,140	,057	2,592	,010	,013 ^c	,011	,016
	N of Valid Cases		378						

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on 10000 sampled tables with starting seed 2000000.

Рис. 4.11. Таблица Symmetric Measures с результатами теста Monte-Carlo

Таким образом, мы выяснили, что между частотой посещения развлекательного центра и возрастом респондентов-женщин существует статистически значимая зависимость, характеризующаяся слабой положительной линейностью. Для респондентов-мужчин возраст и частота посещения центра также связаны статистически значимой зависимостью, однако сделать точный вывод о характере данной зависимости не представляется возможным.

Вернемся к табл. 4.2 и покажем, как интерпретировать представленные в ней данные. На основании проведенных расчетов можно утверждать, что мужчины в возрасте старше 51 года посещают развлекательный центр реже всего (примерно 2 раза в неделю). Наиболее частыми посетителями развлекательного центра являются мужчины в возрасте младше 50 лет (примерно 3 раза в месяц). В целевой группе женщин можно выделить три группы. Наиболее частыми посетителями являются женщины в возрасте 31-35 лет (примерно 4 раза в неделю). Среднюю группу (примерно 3 раза в неделю) составляют женщины младше 30 лет, от 36 до 40 лет и старше 46 лет. И наконец, группу респондентов-женщин, посещающих центр реже всего, составляет возрастная группа от 41 до 45 лет.

4.1.2. Перекрестные распределения для многовариантных вопросов

Как уже было сказано выше (см. раздел 3.2), все статистические процедуры применимы только для одновариантных вопросов. На практике установить статистическую зависимость в многовариантных вопросах можно только двумя способами.

■ Визуально. В этом случае аналитик должен самостоятельно (на основании опыта или опираясь на другие данные, выявленные в ходе исследования) попытаться сделать заключение о значимости различий между двумя переменными. Например, если мужчины покупают сметану в упаковке в 4 раза чаще, чем женщины, и при этом число респондентов, ответивших на данный вопрос, достаточно велико (скажем, 100 человек), можно сделать вывод о статистической значимости данного различия.

■ Можно рассматривать многовариантный вопрос как несколько дихотомических переменных с вариантами ответа «есть/нет» и строить по ним стандартные перекрестные распределения (при помощи процедуры Crosstabs). На практике в подавляющем большинстве случаев именно данный способ является оптимальным. Тем не менее необходимо отметить, что дихотомические переменные, являющиеся вариантами ответа на многовариантный вопрос, могут принимать участие даже в корреляционном анализе в качестве порядковых переменных (см. раздел 4.2).

Кроме существенных ограничений при установлении статистических зависимо-

стей между многовариантными переменными, их анализ осложнен также и тем, что результаты перекрестных распределений по многовариантным вопросам SPSS выводит только в виде простого текста (plain text)1.

Ниже мы проиллюстрируем процесс построения перекрестных распределений по многовариантным переменным на примере двух многовариантных вопросов из маркетингового исследования московского рынка сметаны. Первый вопрос Где Вы покупаете сметану? (q7) с вариантами ответа:

- продмаг (q7_1);
- рынок (q7_2);
- супермаркет (q7_3);
- палатка (q7_4);
- универсам (q7_5).

Второй вопрос Какую сметану Вы предпочитаете? с вариантами ответа:

- в упаковке (q16_1);
- развесную (q16_2).

Как было сказано выше в разделе 2.2.2, чтобы строить распределения (линейные или перекрестные) по многовариантным переменным, сначала их нужно сформировать. Мы не будем возвращаться к процедуре создания многовариантных переменных при помощи меню Analyze ► Multiple Response ► Define Sets; этот процесс описан в разделе 2.2.2. Давайте исходить из того, что вы самостоятельно сформировали две многовариантные переменные, назовем их q7 (Место покупки сметаны) и q16 (Наиболее предпочтительная для респондентов упаковка сметаны). Теперь можно заняться построением перекрестного распределения по этим вопросам, то есть ответить на вопрос: «Зависят ли предпочтения респондентов в отношении сметаны (упакованной или развесной) от места совершения покупки?».

Построение перекрестного распределения по многовариантным вопросам осуществляется при помощи меню Analyze ► Multiple Response ► Crosstabs. В открывшемся диалоговом окне (рис. 4.12) слева вы видите два списка переменных. В верхнем находятся все доступные переменные из файла данных (включая и дихотомические переменные — варианты ответа на анализируемые многовариантные вопросы). Нижний список содержит только сформированные нами многовариантные переменные (\$q7 и \$q16). В перекрестном анализе могут принимать участие как

многовариантные переменные, так и другие доступные одновариантные переменные. Как для кросстабуляций (см. раздел 4.1.1), для перекрестных таблиц можно задать несколько измерений (максимум три) при помощи введения одного дополнительного слоя (область Layer). Имейте в виду, что при построении перекрестных таблиц, переменные, находящиеся в областях Row(s), Column(s) и Layer(s), перекрещиваются по тройкам последовательно.

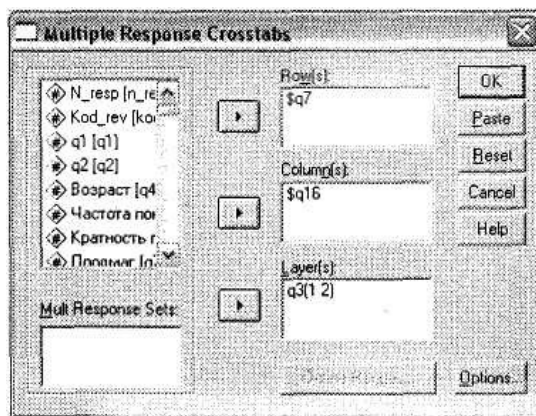


Рис. 4.12. Диалоговое окно Multiple Response Crosstabs

Итак, поместите в область Row(s) переменную Место покупки сметаны (q7), а в область Column(s) — переменную Предпочтения сметаны (q16). В область Layer(s) поместите переменную Пол (q3).

Как вы поняли, мы будем рассматривать трехмерное перекрестное распределение. Обратите внимание на то, что при внесении в одну из трех областей переменной из верхнего левого списка (всех доступных переменных в базе данных) после имени этой переменной появляется строка символов вида (? ?) и становится доступной кнопка Define Ranges. Это подсказывает нам, что следует ввести границы изменения одновариантной переменной. Выделите переменную q3 в поле Layer(s) и щелкните на кнопке Define Ranges.

На экране появится новое диалоговое окно Define Variable Ranges (рис. 4.13). В нем в соответствующих полях следует указать минимальное Minimum и максимальное Maximum значения, которые может принимать данная переменная. В нашем случае пол респондентов может быть либо мужским (код 1), либо женским (код 2). Поэтому введите 1 в качестве минимального значения, а 2 — в качестве максимального и щелкните на кнопке Continue для того, чтобы закрыть это диалоговое окно.

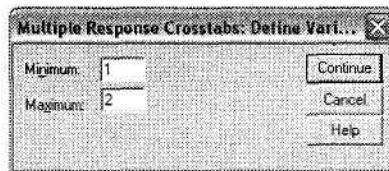


Рис. 4.13. Диалоговое окно Define Variable Ranges

Необходимо отметить, что переменные, участвующие в рассматриваемом статистическом анализе, для которых указываются интервалы допустимых значений, должны принимать только целые значения (дробные SPSS будет игнорировать). Это связано с ограничением при использовании в кросстабуляциях по многовариантным вопросам переменных с интервальной шкалой. Такие переменные могут использоваться, только если они принимают целые значения.

Щелкните на кнопке Options. Открывшееся диалоговое окно (рис. 4.14) позволяет указать, нужно ли выводить проценты (по строкам — Row, по столбцам — Column или общие — Total), а также определить, что является базой для расчета процентов: количество респондентов (Cases) или количество ответов на вопрос (Responses)¹.

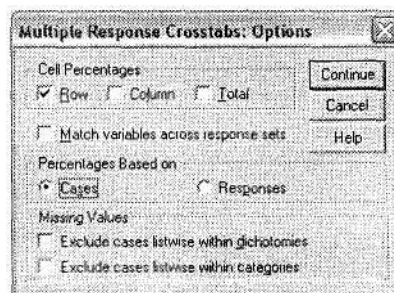


Рис. 4.14. Диалоговое окно Options

Давайте выведем проценты по строкам (то есть доли респондентов, предпочитающих разный вид сметаны в каждом из пяти рассматриваемых типов торговых точек). Оставьте выбранный по умолчанию параметр Cases в области Percentages Based on — это позволит вам рассчитать проценты от общего числа респондентов (гистограмма), а не от количества ответов на вопрос (сектограмма). Щелкните на кнопке Continue для того, чтобы закрыть диалоговое окно, и запустите процедуру построения перекрестного распределения при помощи щелчка на кнопке О К в главном диалоговом окне программы.

В окне SPSS Viewer будет выведена перекрестная таблица с результатами расчетов. Обратите внимание, что таблица разбита на две части: первая содержит результаты построения перекрестного распределения предпочтений сметаны и места покупки для мужчин (рис. 4.15), а вторая — для женщин (рис. 4.16). Таким образом, можно сказать, что собственно построения перекрестного распределения по трем заданным переменным (включая переменную Пол) не происходит.

Переменная, указанная в качестве слоя (Layer), не отображается в таблице. Вместо этого ее значение (для каждого из вариантов ответа, в нашем случае — мужчины и женщины) отображается в верхней части каждой кросстабуляции как текст Category = 1 Мужчины (для мужчин) и Category = 2 Женщины (для женщин).

*** C R O S S T A B U L A T I O N ***

\$Q7 (tabulating 1) Место покупки сметаны
by \$Q16 (tabulating 1) Предпочтения сметаны
by Q3 Пол
Category = 1 Мужчины

		\$Q16		
	Count Row pct	Сметана	Развесна	Row Total
		в упаков	а сметан	
		ке	а	
		Q16_1	Q16_2	
		Q16_1	Q16_2	
\$Q7				
Продукт	Q7_1	71	23	94
		75,5	24,5	51,9
Рынок	Q7_2	33	9	42
		78,6	21,4	23,2
Супермаркет	Q7_3	35	12	47
		74,5	25,5	26,0
Палатка	Q7_4	26	6	32
		81,3	18,8	17,7
Универсам	Q7_5	11	3	14
		78,6	21,4	7,7
Column		143	38	181
Total		79,0	21,0	100,0

Percents and totals based on respondents

Рис. 4.15. Таблица Crosstabulation для мужчин

*** CROSSTABULATION ***

\$Q7 (tabulating 1) Место покупки сметаны
 by \$Q16 (tabulating 1) Предпочтения сметаны
 by Q3 Пол
 Category = 2 Женщины

		\$Q16		
		Count	Сметана Развесна	
		Row pct	в упаков я сметан	Row
			ре а	Total
			Q16_1	Q16_2
\$Q7	Q7_1	276	115	391
	Продукт	70,6	29,4	51,5
	Q7_2	196	52	248
	Рынок	79,0	21,0	32,7
	Q7_3	124	69	193
	Супермаркет	64,2	35,8	25,4
	Q7_4	102	25	127
	Палатка	80,3	19,7	16,7
	Q7_5	35	14	49
	Универсам	71,4	26,6	6,5
Column		564	195	759
Total		74,3	25,7	100,0

Percents and totals based on respondents

940 valid cases; 63 missing cases

Рис. 4.16. Таблица Crosstabulation для женщин

В нижней части под всеми таблицами расположены две строки, содержащие информацию об общих параметрах построения перекрестных распределений. Так, в нашем случае мы видим, что все проценты, представленные в таблицах, рассчитаны от общего числа респондентов (Percents and totals based on respondents). Во второй строке отражаются:

- количество результативных анкет (то есть анкет, в которых респонденты ответили на три вопроса) — 940 valid cases;
- количество анкет, не включенных в анализ (респонденты не дали ответа хотя бы на один из трех вопросов), — 63 missing cases.

Общий размер выборки равен сумме результативных и исключенных анкет: 1003

= 940 + 63. В таблицах приведены результаты построения перекрестного распределения предпочтений респондентов по типу сметаны в зависимости от места покупки. Необходимо отметить, что проценты в ячейках таблицы отражают доли покупателей, предпочитающих сметану в упаковке и развесную для каждого из рассматриваемых мест покупки. Например, 75,5 % мужчин, покупающих сметану в продовольственных магазинах, предпочитают сметану в упаковке, а 24,5 % — развесную¹.

Проценты в строке Column Total отражают доли респондентов, предпочитающих сметану в упаковке или развесную, от общего числа респондентов (в нашем случае мужского или женского пола), ответивших на рассматриваемые вопросы. Например, 79 % мужчин, ответивших на рассматриваемые вопросы, предпочитают упакованную сметану, а 21 % — развесную.

Проценты в столбце Row Total отражают доли респондентов, покупающих сметану в различных торговых точках. На рис. 4.15 вы видите, что 51,9 % мужчин, ответивших на рассматриваемые вопросы, покупают сметану в продовольственных магазинах. Значения на пересечении строки Column Total и столбца Row Total показывают общее количество респондентов мужского пола, ответивших на вопросы о предпочтениях сметаны и месте покупки (как и всегда, в абсолютных и относительных величинах). В нашем случае на рассматриваемые вопросы ответил 181 мужчина. Обратите внимание, что длинные таблицы, выводимые в виде текста, могут по умолчанию не отражаться полностью в окне SPSS Viewer. Чтобы убедиться, что вы видите таблицу целиком, дважды щелкните мышью на ней. Откроется специальная область с возможностью прокрутки, в которой вы можете увидеть все построенные таблицы.

4.2. Корреляционный анализ

Корреляционный анализ предназначен для выявления наличия, а также определения направления и силы линейной связи между несколькими переменными, имеющими интервальный или порядковый тип шкалы. Необходимо отметить, что дихотомические переменные также могут принимать участие в корреляционном анализе. С точки зрения SPSS они рассматриваются как порядковые переменные.

В табл. 4.3 представлены основные характеристики переменных, участвующих в анализе.

Таблица 4.3. Основные характеристики переменных, участвующих в корреляционном анализе

Корреляционный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
—	—	Любое	Интервальная
			Порядковая
			Дихотомическая

Наличие, направление и силу линейной связи отражают коэффициенты корреляции. Они варьируются от -1 до +1.

- -1 соответствует абсолютно разнонаправленной зависимости (с возрастанием одной переменной другая убывает);

- +1 отражает полное соответствие между переменными (то есть они, по сути, являются одним и тем же);

- 0 показывает полное отсутствие всякой связи.

Для удобства интерпретации корреляций применяются семантические интервалы, причем при анализе данных в маркетинговых исследованиях обычно используются следующие градации (табл. 4.4).

Таблица 4.4. Градации коэффициентов корреляции

Значение коэффициента корреляции	Характеристика силы линейной связи
От $\pm 0,81$ до $\pm 1,00$	Сильная
От $\pm 0,61$ до $\pm 0,80$	Умеренная (средняя)
От $\pm 0,41$ до $\pm 0,60$	Слабая
От $\pm 0,21$ до $\pm 0,40$	Очень слабая
От $\pm 0,00$ до $\pm 0,20$	Нет корреляции

Существует два основных типа коэффициентов корреляции, рассчитываемых в зависимости от вида шкалы переменных, участвующих в анализе.

1. Для переменных с интервальной шкалой применяется коэффициент корреляции Пирсона. Он позволяет охарактеризовать линейную связь между двумя переменными по указанным параметрам (табл. 4.4): наличию (есть/нет), направлению (убывает/возрастает) и силе (очень слабая/слабая/умеренная/сильная).

2. Если хотя бы одна из пары исследуемых переменных имеет порядковую или дихотомическую шкалу, используются ранговые коэффициенты корреляции Спирмана или Кендала. Чаще всего эти коэффициенты применяются в маркетинговых исследованиях в тех случаях, когда необходимо установить степень соответствия двух ранжированных списков. Например, если имеются схемы выбора какого-либо продукта различными целевыми группами респондентов (в виде ранжированных по важности параметров) и необходимо установить, насколько точно они соответствуют друг другу (или различаются).

Ниже мы рассмотрим перечисленные типы коэффициентов корреляции более подробно на практических примерах из маркетинговых исследований.

4.2.1. Исследование линейных корреляций по Пирсону, Спирману и Кендалу

Сначала мы рассмотрим пример применения коэффициента корреляции Пирсона. Предположим, что у нас есть ответы респондентов на следующие два вопроса. Каков Ваш среднемесячный доход в расчете на одного члена семьи? с вариантами ответа:

- до \$100;
- от \$ 100 до \$ 300;
- от \$ 300 до \$ 600;
- от \$ 600 до \$ 1000;
- от \$ 1000 до \$ 1500;
- свыше \$1500.

Как часто Вы посещаете рестораны? с вариантами ответа:

- более 1 раза в день;
- примерно 1 раз в день;
- 2-3 раза в неделю;
- примерно 1 раз в неделю;
- 2-3 раза в месяц;
- примерно 1 раз в месяц;
- реже 1 раза в месяц.

В результате ввода в компьютер заполненных анкет респондентов были получены две переменные: q3 (первый вопрос) и q28 (второй вопрос). Необходимо установить, зависит ли частота посещения ресторанов от дохода респондентов, и если да, то каким образом. В связи с тем, что в ходе опроса при ответе на каждый вопрос респондентам предлагалось на выбор несколько вариантов ответа, тип шкалы у полученных переменных получился порядковым (в файле данных есть только коды ответов, но не сами числовые

значения, отражающие частоту посещения ресторана или уровень дохода).

Далее мы рассмотрим не только как использовать коэффициент корреляции Пирсона, но также как использовать данный коэффициент для анализа квазипорядковых переменных. Дело в том, что некоторые переменные, хотя они и закодированы как порядковые, по сути являются интервальными (как в нашем случае). Это делается специально, чтобы, с одной стороны, увеличить долю респондентов, ответивших на вопрос, а с другой стороны, уменьшить число возможных ошибок при вводе в компьютер текстовых полей (для открытых вопросов). Интервалы также полезны при анализе, поскольку нет необходимости кодировать текстовые (или интервальные) переменные, а можно сразу увидеть группы (интервалы) значений. Практика показывает, что подобное составление анкет для маркетинговых исследований является стандартным, поэтому корреляционный анализ редко проводится на изначально интервальных переменных (текстовые поля анкеты).

Для описываемых квазипорядковых переменных следует применять именно коэффициент корреляции Пирсона. Использование коэффициентов Спирмана или Кендала в этом случае является некорректным. Более подробно эти два коэффициента представлены ниже; пока же в общих чертах о них можно сказать следующее. Коэффициенты Спирмана или Кендала показывают только степень соответствия порядка следования вариантов ответа в ранжированных списках (есть отсутствие инверсий). При этом корреляции по Спирману и Кендалу используются в основном, когда элементы ранжированных списков представлены мнемоническими, а не числовыми константами. Таким образом, данные коэффициенты не помогут нам в характеристике зависимости между частотой посещения ресторанов и доходом респондентов. Однако в нашем случае нельзя применять и коэффициент корреляции Пирсона, так как в этом случае анализировались бы коды интервалов (1-6 — в первом вопросе и 1-7 — во втором), а не действительные ответы респондентов на вопросы¹.

Итак, сначала мы должны преобразовать имеющиеся у нас порядковые переменные к интервальному виду. Лучше всего сделать это при помощи замены кодов интервалов (1-6) на средние значения данных интервалов. Например, среднее значение для интервала 3 в переменной q3 — это \$ 450 ($450 = (300 + 600) / 2$). Преобразовав обе переменные к данному виду, мы получим следующие интервальные переменные q3_i и q28_i (табл. 4.5)².

Таблица 4.5. Схема перекодировки порядковых переменных (q3 и q28) в интервальные (q3_i и q28_i)

Порядковые переменные		Интервальные переменные	
Каков Ваш среднемесячный доход в расчете на одного члена семьи?			
до \$ 100		\$50	
от \$ 100 до \$ 300		\$200	
от \$ 300 до \$ 600		\$450	
от \$ 600 до \$ 1000		\$ 800	
от \$ 1000 до \$ 1500		\$ 1250	
свыше \$ 1500		\$ 1750	
Как часто Вы посещаете рестораны?			
более 1 раза в день		60 раз в месяц	
примерно 1 раз в день		30 раз в месяц	
2-3 раза в неделю		10 раз в месяц	
примерно 1 раз в неделю		4 раза в месяц	
2-3 раза в месяц		2,5 раза в месяц	
примерно 1 раз в месяц		1 раз в месяц	
реже 1 раза в месяц		0,5 раза в месяц	

Теперь мы можем приступить непосредственно к корреляционному анализу (описанию зависимости между частотой посещения ресторанов и уровнем дохода). Для этого выберите пункт меню **Analyze ► Correlate ► Bivariate**. В открывшемся диалоговом окне (рис. 4.17) выберите в левом списке всех доступных переменных две интересующие нас ($q3_i$ и $q28_i$) и перенесите их в область **Variables**. Остальные параметры в этом диалоговом окне, установленные по умолчанию, следует оставить неизменными: вывод коэффициентов корреляции Пирсона (параметр **Pearson** в области **Correlation Coefficients**) и статистической значимости коэффициентов (параметр **Two-tailed** в области **Test of Significance**). Кнопка **Options** не предлагает исследователю каких-либо существенных параметров. Чтобы запустить процедуру построения корреляционной таблицы, щелкните на кнопке **OK**.

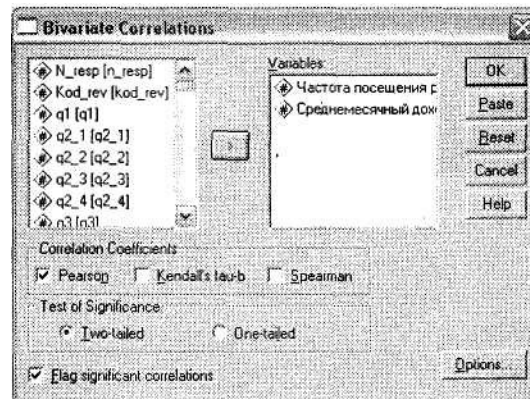


Рис. 4.17. Диалоговое окно **Bivariate Correlations** (корреляция Пирсона)

В окне **SPSS Viewer** появится таблица **Correlations** с результатами расчетов коэффициента корреляции Пирсона и статистической значимости данного коэффициента. Как видно из рис. 4.18, в нашем случае коэффициент корреляции Пирсона между двумя исследуемыми переменными ($q3_i$ и $q28_i$) равен +0,665, а его статистическая значимость меньше 0,001. Следовательно, можно сделать вывод о том, что между среднемесячным доходом респондентов и частотой посещения ими ресторанов существует статистически значимая умеренная (средняя) линейная возрастающая зависимость. То есть частота посещения ресторанов в достаточно высокой степени (коэффициент Пирсона = 0,7) зависит от уровня доходов потребителей, причем при росте среднемесячного дохода частота посещения ресторанов линейно возрастает.

Существует возможность проводить корреляционный анализ сразу для нескольких переменных. Для этого необходимо поместить эти переменные в область **Variables** диалогового окна **Bivariate Correlations**. В таблице **Correlations** будут показаны коэффициенты корреляции для каждой пары исследуемых переменных.

Теперь рассмотрим процедуру проведения корреляционного анализа при помощи ранговых коэффициентов Спирмана и Кендала. В данных методах одна переменная (эталонная) представлена в виде ранжированной последовательности мнемонических категорий, а другой переменной присваиваются ранговые места. Корреляционные коэффициенты рассчитываются исходя из количества инверсий, то есть числа нарушений порядка следования рангов по сравнению с первым рядом. В большинстве случаев рекомендуется применять коэффициент корреляции Спирмана. Использование коэффициента Кендала оправдано только в том случае, когда в структуре данных имеются выбросы.

Correlations

		Среднемесячный доход на 1 члена семьи	Частота посещения ресторанов
Среднемесячный доход на 1 члена семьи	Pearson Correlation	1	,665*
	Sig. (2-tailed)	.	,000
	N	81	81
Частота посещения ресторанов	Pearson Correlation	,665*	1
	Sig. (2-tailed)	,000	.
	N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 4.18. Таблица Correlations (корреляция Пирсона)

В практике маркетинговых исследований наиболее часто коэффициенты корреляции Спирмана применяются для анализа не всей выборочной совокупности респондентов (базы данных в целом), а агрегированных ранжированных перечней, полученных в результате других преобразований¹. Приведем пример. Предположим, что в результате опроса посетителей магазинов одежды были получены ответы на следующие два вопроса. Какие факторы для Вас наиболее важны при выборе одежды? с вариантами ответа:

- Высокое качество одежды.
- Доступные цены.
- Широта ассортимента одежды.
- Близость к дому или работе.
- Высокое качество обслуживания.
- Красивый интерьер магазина.

Оцените, пожалуйста, следующие характеристики данного магазина одежды (в котором происходит опрос) по пятибалльной шкале (от 1 — очень плохо до 5 — отлично) с вариантами ответа:

- Высокое качество одежды.
- Доступные цены.
- Широта ассортимента одежды.
- Близость к дому или работе.
- Высокое качество обслуживания.
- Красивый интерьер магазина.
- Ваша общая оценка работы данного магазина.

Над результатами второго вопроса был проведен множественный линейный регрессионный анализ. Анализировалось влияние оценок частных параметров всех исследованных магазинов одежды на их общую оценку. В разделе 4.3 подробно рассматривается процедура линейного регрессионного анализа, позволяющая, в частности, построить ранжированный перечень частных параметров по силе их влияния на общую оценку.

Таким образом, были получены два ранжированных списка с одинаковыми категориями: две схемы выбора магазина одежды. Затем оба списка были введены в SPSS под кодами, представленными выше: от 1 (наиболее важный фактор) до 6 (наименее важный фактор) (рис. 4.19). На рис. 4.20 представлены данные списки в мнемонической форме. Первый список представлен в переменной `sc_1`; второй — в `sc_2`.

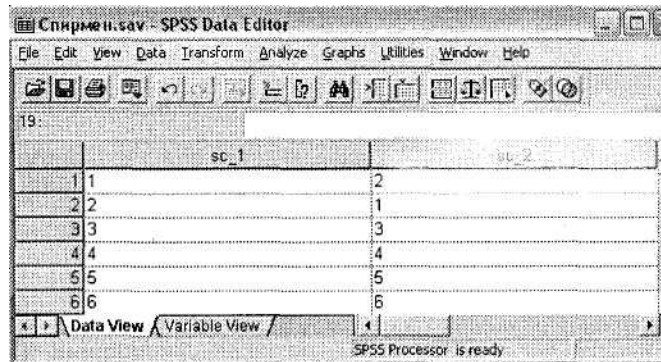


Рис. 4.19. Окно SPSS Data Editor с двумя ранжированными перечнями наиболее значимых для респондентов факторов выбора магазинов одежды

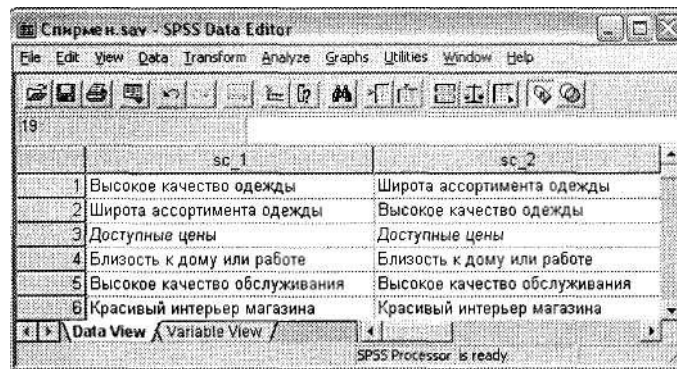


Рис. 4.20. Окно SPSS Data Editor с двумя ранжированными перечнями наиболее значимых для респондентов факторов выбора магазинов одежды

Как вы видите на рис. 4.20, две схемы выбора, составленные на основании прямого метода (вопрос 1) и на основании регрессионного анализа (вопрос 2), соответствуют друг другу не полностью, различаясь в порядке следования первой и второй категорий. Проанализируем эти схемы выбора магазинов одежды на предмет соответствия при помощи коэффициента корреляции Спирмана.

Для этого снова откройте диалоговое окно Bi variate Correlations, выбрав пункт меню Analyze ► Correlate ► Bivariate. Перенесите две интересующие нас переменные — Схема №1 (составленная по вопросу 1) и Схема №2 (составленная по вопросу 2) — из левого списка всех доступных переменных в область Variables (рис. 4.21). Отмените вывод корреляции Пирсона и вместо него выберите параметр Spearman (корреляция Спирмана). После этого начните расчет при помощи щелчка на кнопке ОК.

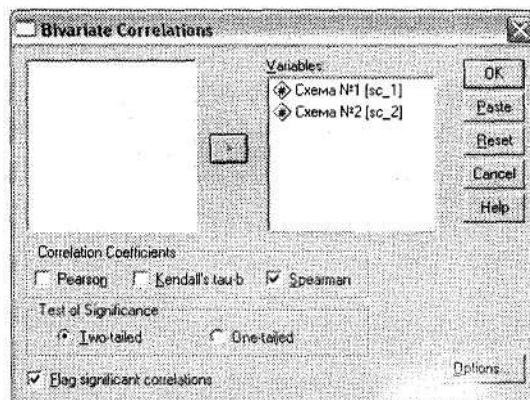


Рис. 4.21. Диалоговое окно Bivariate Correlations (корреляция Спирмана) одежды

В окне SPSS Viewer появится таблица Correlations с результатами расчета коэффициента ранговой корреляции (Спирмана) по двум анализируемым переменным. Как следует из рис. 4.22, две рассматриваемые схемы выбора различаются несущественно. Данный вывод можно сделать из сильной корреляции между переменными sc_1 и sc_2 (коэффициент корреляции Спирмана = 0,9), характеризующейся весьма высокой статистической значимостью (0,005).

		Correlations		
		Возраст	Количество членов семьи	Доход
Возраст	Pearson Correlation	1	,250	,794
	Sig. (2-tailed)		,050	,010
	N	653	643	646
Количество членов семьи	Pearson Correlation	,250	1	,402
	Sig. (2-tailed)	,050		,046
	N	643	653	653
Доход	Pearson Correlation	,794	,402	1
	Sig. (2-tailed)	,010	,046	
	N	646	653	662

Рис. 4.22. Таблица Correlations (корреляция Спирмана)

В заключение напомним, что ранговый коэффициент корреляции Спирмана (в отличие от Кендала) может применяться и в качестве аналога корреляции Пирсона при исследовании зависимостей между переменными, не приводимыми к интервальному виду и потому не являющимися ранжированными списками. В качестве примера можно привести гипотетический случай, рассмотренный выше, когда анализируется влияние пола респондентов (дихотомическая шкала) на уровень образования (порядковая по сути, но номинальная по виду шкала).

4.2.2. Частные корреляции. Выявление ложных корреляций

На практике иногда возникают ситуации, когда в результате корреляционного анализа обнаруживаются логически необъяснимые, противоречащие объективному опыту исследователя корреляции между двумя переменными (например, оказывается, что между уровнем дохода респондентов и количеством детей в семье существует статистически значимая зависимость). В этом случае говорят о так называемой ложной корреляции, исследовать которую помогают частные коэффициенты корреляции.

Рассмотрим процедуру исследования частных корреляций на следующем примере из маркетингового исследования поведения посетителей залов игровых автоматов. В результате обработки анкет респондентов были, в частности, получены три интервальные переменные:

- q47 — возраст;
- q49 — количество членов семьи;
- q50 — среднемесячный доход на 1 члена семьи.

Над данными переменными был проведен корреляционный анализ (Пирсона), который выявил логически необъяснимую, но статистически значимую зависимость между переменными: Доход и Количество членов семьи (рис. 4.23).

Correlations

			Схема №1	Схема №2
Spearman's rho	Схема №1	Correlation Coefficient	1,000	,943**
		Sig. (2-tailed)	,	,005
		N	6	6
	Схема №2	Correlation Coefficient	,943**	1,000
		Sig. (2-tailed)	,005	,
		N	6	6

** . Correlation is significant at the .01 level (2-tailed).

Рис. 4.23. Коэффициенты корреляции (Пирсона) для трех переменных: возраст, уровень доходов и количество членов семьи

Как видно из таблицы, обе рассматриваемые переменные коррелируют с третьей переменной Возраст. В такой ситуации корреляция между уровнем дохода респондентов и численностью их семей может объясняться влиянием третьей переменной: возраста респондентов. То есть связанными (коррелирующими), на самом деле, являются пары возраст/уровень дохода и возраст/количество членов семьи. Проверим данную гипотезу при помощи частных коэффициентов корреляции.

Откройте диалоговое окно Partial Correlations (меню Analyze ► Correlate ► Partial). В левом списке всех доступных переменных выберите переменные, между которыми обнаружена странная корреляция (q50 Доход и q49 Количество членов семьи), и поместите их в область Variables. Переменную, с которой коррелируют обе исследуемые переменные (q47 Возраст), поместите в область Controlling for (рис. 4.24). В этом диалоговом окне больше ничего не изменяйте — просто запустите программу на исполнение, щелкнув на кнопке ОК.

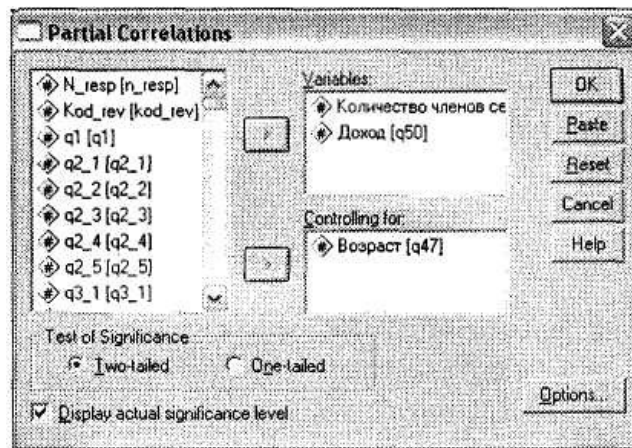


Рис. 4.24. Диалоговое окно Partial Correlations

В окне SPSS Viewer появятся результаты расчетов частных коэффициентов корреляции (рис. 4.25). В данной таблице первая строка каждой ячейки содержит коэффициент корреляции Пирсона, а третья — статистическую значимость данного коэффициента. Из таблицы вы видите, что между количеством членов семьи (q49) и уровнем дохода (q50) больше не наблюдается статистически значимой корреляции ($P = 0,520$), а коэффициент Пирсона сильно уменьшился (0,0256). Следовательно, корреляция, представленная на рис. 4.23, объясняется влиянием третьей переменной Возраст и, таким образом, является ложной.

- - - P A R T I A L C O R R E L A T I O N C O E F F I C I E N T S - - -

Controlling for... Q47

	Q49	Q50
Q49	1,0000 (0) P= ,	,0256 (634) P= ,520
Q50	,0256 (634) P= ,520	1,0000 (0) P= ,

(Coefficient / (D.F.) / 2-tailed Significance)

Рис. 4.25. Таблица Partial Correlation Coefficients

4.3. Линейный регрессионный анализ и статистическое прогнозирование

Линейная регрессия является наиболее часто используемым видом регрессионного анализа. Ниже перечислены три основные задачи, решаемые в маркетинговых исследованиях при помощи линейного регрессионного анализа.

1. Определение того, какие частные параметры продукта оказывают влияние на общее впечатление потребителей от данного продукта. Установление направления и силы данного влияния. Расчет, каким будет значение результирующего параметра при тех или иных значениях частных параметров. Например, требуется установить, как влияет возраст респондента и его среднемесячный доход на частоту покупок глазированных сырков.

2. Выявление того, какие частные характеристики продукта влияют на общее впечатление потребителей от данного продукта (построение схемы выбора продукта потребителями). Установление соотношения между различными частными параметрами по силе и направлению влияния на общее впечатление. Например, имеются оценки респондентами двух характеристик мебели производителя X — цены и качества, — а также общая оценка мебели данного производителя. Требуется установить, какой из двух параметров является наиболее значимым для покупателей при выборе производителя мебели и в каком конкретном соотношении находится значимость для покупателей данных двух факторов (параметр Цена в x раз более значим для покупателей при выборе мебели, чем параметр Качество).

3. Графическое прогнозирование поведения одной переменной в зависимости от изменения другой (используется только для двух переменных). Как правило, целью проведения регрессионного анализа в данном случае является не столько расчет уравнения, сколько построение тренда (то есть аппроксимирующей кривой, графически показывающей зависимость между переменными). По полученному уравнению можно предсказать, каким будет значение одной переменной при изменении (увеличении или уменьшении) другой. Например, требуется установить характер зависимости между долей респондентов, осведомленных о различных марках глазированных сырков, и долей респондентов,

покупающих данные марки. Также требуется рассчитать, насколько возрастет доля покупателей сыров марки x при увеличении потребительской осведомленности на 10 % (в результате проведения рекламной кампании).

В зависимости от типа решаемой задачи выбирается вид линейного регрессионного анализа. В большинстве случаев (1 и 2) применяется множественная линейная регрессия, в которой исследуется влияние нескольких независимых переменных на одну зависимую. В случае 3 применима только простая линейная регрессия, в которой участвуют только одна независимая и одна зависимая переменные. Это связано с тем, что основным результатом анализа в случае 3 является линия тренда, которая может быть логически интерпретирована только в двухмерном пространстве. В общем случае результатом проведения регрессионного анализа является построение уравнения регрессии вида: $y = a + B_1x_1 + B_2x_2 + \dots + B_nx_n$, позволяющего рассчитать значение зависимой переменной при различных значениях независимых переменных.

В табл. 4.6 представлены основные характеристики переменных, участвующих в анализе.

Таблица 4.6. Основные характеристики переменных, участвующих в линейном регрессионном анализе

Линейная регрессия			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Интервальная Порядковая	Любое	Интервальная Порядковая Дихотомическая

В связи с тем что и множественная и простая регрессии строятся в SPSS одинаковым способом, рассмотрим общий случай множественной линейной регрессии как наиболее полно раскрывающий суть описываемого статистического метода. Давайте рассмотрим, как построить линию тренда с целью статистического прогнозирования.

Исходные данные:

В ходе опроса респондентов, летающих одним из трех классов (первым, бизнес- или эконом-классом), просили оценить по пятибалльной шкале — от 1 (очень плохо) до 5 (отлично) — следующие характеристики сервиса на борту самолетов авиакомпании X: комфортабельность салона, работа бортпроводников, питание во время полета, цена билетов, спиртные напитки, дорожные наборы, аудиопрограммы, видеопрограммы и пресса. Также респондентам предлагалось поставить общую (итоговую) оценку обслуживания на борту самолетов данной авиакомпании.

Для каждого класса полета требуется:

- 1) Выявить наиболее значимые для респондентов параметры обслуживания на борту.
- 2) Установить, какое влияние оказывают оценки частных параметров обслуживания на борту на общее впечатление авиапассажиров от полета.

Откройте диалоговое окно Linear Regression при помощи меню Analyze ► Regression ► Linear. Из левого списка выберите зависимую переменную для анализа. Это будет Общая оценка сервиса на борту. Поместите ее в область Dependent. Далее в левом списке выберите независимые переменные для анализа: частные параметры сервиса на борту — и поместите их в область Independent(s).

Существует несколько методов проведения регрессионного анализа: enter, stepwise, forward и backward. Не вдаваясь в статистические тонкости, проведем регрессионный анализ посредством пошагового метода backward как наиболее универсального и релевантного для всех примеров из маркетинговых исследований.

Так как задача анализа содержит требование провести регрессионный анализ в разрезе трех классов полета, выберите в левом списке переменную, обозначающую класс (q5) и перенесите ее в область Selection Variable. Затем щелкните на кнопке Rule, чтобы задать конкретное значение данной переменной для регрессионного анализа. Следует отметить, что за одну итерацию можно построить регрессию только в разрезе какого-то одного класса полета. В дальнейшем следует повторить все этапы сначала по количеству классов (3), каждый раз выбирая следующий класс.

Если нет необходимости проводить регрессионный анализ в каком-либо разрезе, оставьте поле Selection Variable пустым.

Итак, на экране открылось диалоговое окно Set Rule, в котором вы должны указать, для какого именно класса полета вы хотите построить регрессионную модель. Выберите экономический класс, закодированный как 3 (рис. 4.26).

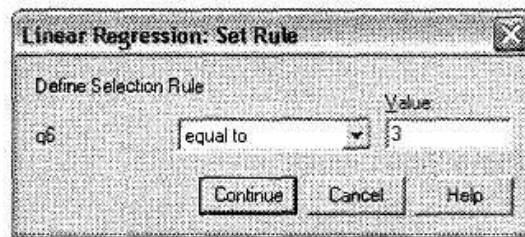


Рис. 4.26. Диалоговое окно Set Rule

В более сложных случаях, когда требуется построить регрессионную модель в разрезе трех и более переменных, следует воспользоваться условным отбором данных (см. раздел 1.5.1). Например, если кроме класса полета есть еще и необходимость раздельного построения регрессионной модели для респондентов (мужчин и женщин), необходимо перед открытием диалогового окна Linear Regression произвести условный отбор анкет респондентов, являющихся мужчинами. Далее проводится регрессионный анализ по описываемой схеме. Для построения регрессии для женщин следует повторить все этапы сначала: вначале выбрать только анкеты респондентов-женщин и затем уже для них построить регрессионную модель.

Щелкните на кнопке Continue в диалоговом окне Set Rule — вы вновь вернетесь к основному диалоговому окну Linear Regression. Последним шагом перед запуском процедуры построения регрессионной модели является выбор пункта Collinearity Diagnostics в диалоговом окне, появляющемся при щелчке на кнопке Statistics (рис. 4.27). Установление требования провести диагностику наличия коллинеарности между независимыми переменными позволяет избежать эффекта мульти-коллинеарности, при котором несколько независимых переменных могут иметь настолько сильную корреляцию, что в регрессионной модели обозначают, в принципе, одно и то же (это неприемлемо).

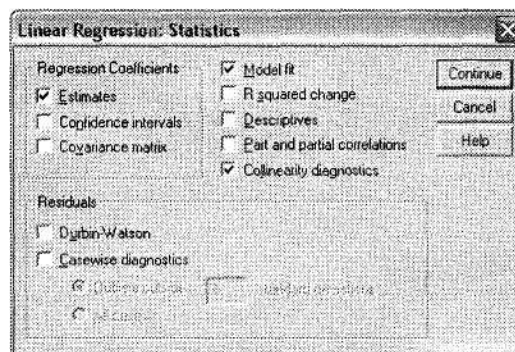


Рис. 4.26. Диалоговое окно Set Rule

Теперь основное диалоговое окно Linear Regression примет вид, показанный на рис. 4.28. Щелчок на кнопке О К приведет к запуску процедуры построения линейной регрессии.

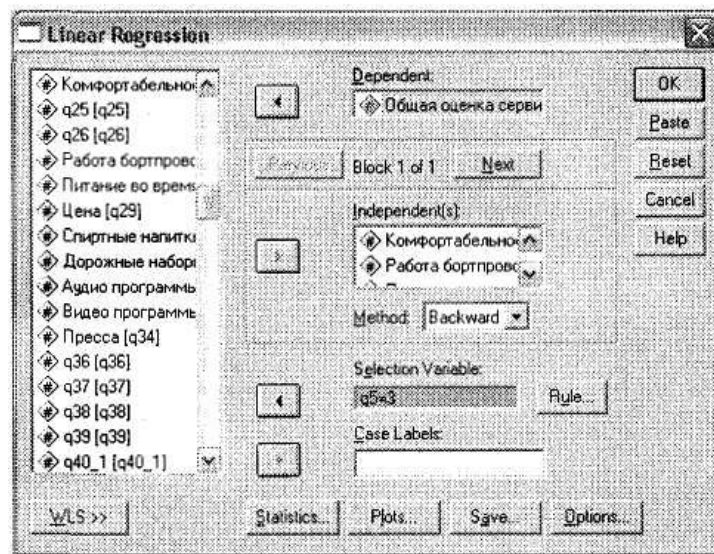


Рис. 4.28. Диалоговое окно Linear Regression

Рассмотрим основные элементы отчета о построении регрессионной модели (окно SPSS Viewer), содержащие наиболее значимые для исследователя данные. Необходимо отметить, что все таблицы, представленные в отчете Output, содержат несколько блоков, соответствующих количеству шагов SPSS при построении модели. На каждом шаге при использовании методе backward из полного списка независимых переменных, введенных в модель изначально, при помощи наименьших частных коэффициентов корреляции последовательно исключаются переменные — до тех пор, пока соответствующий коэффициент регрессии не оказывается незначимым ($\text{Sig} > 0,05$). В нашем примере таблицы состоят из трех блоков (регрессия строилась в три шага). При интерпретации результатов регрессионного анализа следует обращать внимание только на последний блок (в нашем случае 3).

Первое, на что следует обратить внимание, — это таблица ANOVA (рис. 4.29). На третьем шаге статистическая значимость (столбец Sig) должна быть меньше или равна 0,05.

Затем следует рассмотреть таблицу Model Summary, содержащую важные сведения о построенной модели (рис. 4.30). Коэффициент детерминации R является характеристикой силы общей линейной связи между переменными в регрессионной модели. Он показывает, насколько хорошо выбранные независимые переменные способны определять поведение зависимой переменной. Чем выше коэффициент детерминации (изменяющийся в пределах от 0 до 1), тем лучше выбранные независимые переменные подходят для определения поведения зависимой переменной. Требования к коэффициенту R такие же, как к коэффициенту корреляции (см. табл. 4.4): в общем случае он должен превышать хотя бы 0,5. В нашем примере $R = 0,66$, что является приемлемым показателем.

ANOVA^{a,e}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	381,673	9	42,400	242,690	,010 ^a
	Residual	487,486	2847	,175		
	Total	979,172	2856			
2	Regression	381,581	8	47,699	273,013	,010 ^b
	Residual	497,581	2848	,175		
	Total	879,172	2856			
3	Regression	381,156	7	54,451	311,493	,010 ^c
	Residual	498,016	2849	,175		
	Total	879,172	2856			

- a. Predictors: (Constant), Пресса, Работа бортпроводников, Цена, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы, Аудиопрограммы
- b. Predictors: (Constant), Пресса, Работа бортпроводников, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы, Аудиопрограммы
- c. Predictors: (Constant), Пресса, Работа бортпроводников, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы
- d. Dependent Variable: Общая оценка сервиса на борту
- e. Selecting only cases for which Класс полета = Экономический класс

Рис. 4.29. Таблица ANOVA

Также важной характеристикой регрессионной модели является коэффициент R^2 , показывающий, какая доля совокупной вариации в зависимой переменной описывается выбранным набором независимых переменных. Величина R^2 изменяется от 0 до 1. Как правило, данный показатель должен превышать 0,5 (чем он выше, тем показательнее построенная регрессионная модель). В нашем примере $R^2 = 0,43$ — это значит, что регрессионной моделью описано только 43 % случаев (дисперсии в итоговой оценке полета). Таким образом, при интерпретации результатов регрессионного анализа следует постоянно иметь в виду существенное ограничение: построенная модель справедлива только для 43 % случаев.

Третьим практически значимым показателем, определяющим качество регрессионной модели, является величина стандартной ошибки расчетов (столбец Std. Error of the Estimate). Данный показатель варьируется в пределах от 0 до 1. Чем он меньше, тем надежнее модель (в общем случае показатель должен быть меньше 0,5). В нашем примере ошибка составляет 0,42, что является завышенным, но в целом приемлемым результатом.

На основании таблиц ANOVA и Model Summary можно судить о практической пригодности построенной регрессионной модели. Учитывая, что ANOVA показывает весьма высокую значимость (менее 0,001), коэффициент детерминации превышает 0,6, а стандартная ошибка расчетов меньше 0,5, можно сделать вывод о том, что с учетом ограничения модель описывает 43 % совокупной дисперсии, то есть построенная регрессионная модель является статистически значимой и практически приемлемой.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	Класс полета = Экономический класс (Selected)			
1	,659 ^a	,434	,432	,419
2	,659 ^b	,434	,432	,418
3	,658 ^c	,434	,432	,418

a. Predictors: (Constant), Пресса, Работа бортпроводников, Цена, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы, Аудио-программы

b. Predictors: (Constant), Пресса, Работа бортпроводников, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы, Аудио-программы

c. Predictors: (Constant), Пресса, Работа бортпроводников, Комфортабельность салона, Спиртные напитки, Питание во время полета, Дорожные наборы, Видеопрограммы

Рис. 4.30. Таблица Model Summary

После того как мы констатировали приемлемый уровень качества регрессионной модели, можно приступить к интерпретации ее результатов. Основные практические результаты регрессии содержатся в таблице Coefficients (рис. 4.31). Под таблицей вы можете видеть, какая переменная была зависимой (общая оценка сервиса на борту) и для какого класса полета происходило построение регрессионной модели (эконом-класс). В таблице Coefficients практически значимыми являются четыре показателя: VIF, Beta, B и Std. Error. Рассмотрим последовательно, как их следует интерпретировать.

1. Работа бортпроводников	21
2. Комфортабельность салона	21
3. Пресса	16
4. Дорожные работы	12
5. Видеопрограммы	10
6. Питание во время полета	9

Рис. 4.31. Таблица Coefficients

Прежде всего необходимо исключить возможность возникновения ситуации мультиколлинеарности (см. выше), при которой несколько переменных могут обозначать почти одно и то же. Для этого необходимо посмотреть на значение VIF возле каждой независимой переменной. Если величина данного показателя меньше 10 — значит, эффекта мультиколлинеарности не наблюдается и регрессионная модель приемлема для дальнейшей интерпретации. Чем выше этот показатель, тем более связаны между собой переменные. Если какая-либо переменная превышает значение в 10 VIF, следует пересчитать ре-

грессию без этой независимой переменной. В данном примере автоматически уменьшится величина R^2 и возрастет величина свободного члена (константы), однако, несмотря на это, новая регрессионная модель будет более практически приемлема, чем первая.

В первом столбце таблицы Coefficients содержатся независимые переменные, составляющие регрессионное уравнение (удовлетворяющие требованию статистической значимости). В нашем случае в регрессионную модель входят все частные характеристики сервиса на борту самолета, кроме аудиопрограмм. Исключенные переменные содержатся в таблице Excluded Variables (здесь не приводится). Итак, мы можем сделать первый вывод о том, что на общее впечатление авиапассажиров от полета оказывают влияние семь параметров: комфортабельность салона, работа бортпроводников, питание во время полета, спиртные напитки, дорожные наборы, видеопрограммы и пресса.

После того, как мы определили состав параметров, формирующих итоговое впечатление от полета, можно определить направление и силу влияния на него каждого частного параметра. Это позволяет сделать столбец Beta, содержащий стандартизированные β - коэффициенты регрессии. Данные коэффициенты также дают возможность сравнить силу влияния параметров между собой. Знак (+ или -) перед β -коэффициентом показывает направление связи между независимой и зависимой переменными. Положительные β -коэффициенты свидетельствуют о том, что возрастание величины данного частного параметра увеличивает зависимую переменную (в нашем случае все независимые переменные ведут себя подобным образом). Отрицательные коэффициенты означают, что при возрастании данного частного параметра общая оценка снижается. Как правило, при определении связи между оценками параметров это свидетельствует об ошибке и означает, например, что выборка слишком мала.

Например, если бы перед β - коэффициентом параметра работы бортпроводников стоял знак -, его следовало бы интерпретировать следующим образом: чем хуже работают бортпроводники, тем лучше становится общее впечатление пассажиров от полета. Такая интерпретация является бессмысленной и не отражающей реального положения вещей, то есть ложной. В таком случае лучше пересчитать регрессию без данного параметра; тогда доля вариации в итоговой оценке, описываемой исключенным параметром, будет отнесена на счет константы (увеличивая ее). Соответственно уменьшится и процент совокупной дисперсии, описываемой регрессионной моделью (величина R^2). Однако это позволит восстановить семантическую релевантность.

Еще раз подчеркнем, что сделанное замечание справедливо для нашего случая (оценки параметров). Отрицательные β - коэффициенты могут быть верными и отражать семантические реалии в других случаях. Например, когда уменьшение дохода респондентов приводит к увеличению частоты покупок дешевых товаров. В таблице вы видите, что в наибольшей степени на общее впечатление пассажиров от полета влияют два параметра: работа бортпроводников и комфортабельность салона (β - коэффициенты по 0,21). Напротив, в наименьшей степени формирование итоговой оценки сервиса на борту происходит за счет впечатления от обслуживания спиртными напитками (0,08). При этом два первых параметра оказывают почти в три раза более сильное влияние на итоговую оценку полета, чем

спиртные напитки. На основании стандартизированных (3-коэффициентов регрессии) можно построить рейтинг влияния частных параметров сервиса на борту на общее впечатление авиапассажиров от полета, разделив их на три группы по силе влияния:

- наиболее значимые параметры;
- параметры, имеющие среднюю значимость;
- параметры, имеющие низкую значимость для респондентов (рис. 4.32).

В крайнем правом столбце содержатся β - коэффициенты, умноженные на 100, — для облегчения сравнения параметров между собой.

Coefficients ^{ab}							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
3	(Constant)	,779		9,494	,000		
	Комфортабельность салона	,202	,200	12,698	,000	,740	1,351
	Работа бортпроводников	,204	,215	12,977	,000	,725	1,390
	Питание во время полета	7,981E-02	,097	5,909	,000	,735	1,361
	Спиртные напитки	6,650E-02	,081	4,772	,000	,698	1,433
	Дорожные наборы	9,655E-02	,117	6,506	,000	,613	1,633
	Видео программы	7,826E-02	,103	5,591	,000	,582	1,718
	Пресса	,117	,157	8,803	,000	,626	1,597

a. Dependent Variable: Общая оценка сервиса на борту
b. Selecting only cases for which Класс полета = Экономический класс

Рис. 4.32. Рейтинг значимости параметров сервиса на борту

Данный рейтинг также можно интерпретировать и как рейтинг значимости для респондентов различных параметров сервиса на борту (в общем случае — схема выбора). Так, наиболее важными факторами являются первые два (1-2); среднюю значимость для пассажиров имеют следующие три параметра (3-5); относительно малое значение имеют последние два фактора (6-7).

Регрессионный анализ позволяет выявить истинные, глубинные мотивы респондентов при формировании общего впечатления о каком-либо продукте. Как показывает практика, такого уровня приближения нельзя достичь обычными методами — например, просто спросив респондентов: Какие факторы из нижеперечисленных оказывают наибольшее влияние на Ваше общее впечатление от полета самолетами нашей авиакомпании?. Кроме того, регрессионный анализ позволяет достаточно точно оценить, насколько один параметр более-менее значим для респондентов, чем другой, и на этом основании классифицировать параметры на критические, имеющие среднюю значимость и малозначимые.

Столбец В таблицы Coefficients содержит коэффициенты регрессии (нестандартизированные). Они служат для формирования собственно регрессионного уравнения, по которому можно рассчитать величину зависимой переменной при разных значениях независимых.

Особая строка Constant содержит важную информацию о полученной регрессионной модели: значение зависимой переменной при нулевых значениях независимых переменных. Чем выше значение константы, тем хуже подходит выбранный перечень независимых переменных для описания поведения зависимой переменной. В общем случае считается, что константа не должна быть наибольшим коэффициентом в регрессионном уравнении (коэффициент хотя бы при одной переменной должен быть больше константы). Однако в практике маркетинговых исследований часто свободный член оказывается больше всех коэффициентов вместе взятых. Это связано в основном с относительно малыми размерами выборок, с которыми приходится работать маркетологам, а также с неаккуратным заполнением анкет (некоторые респонденты могут не поставить оценку каким-либо параметрам). В нашем случае величина константы меньше 1, что является весьма хорошим результатом.

Итак, в результате построения регрессионной модели можно сформировать следующее регрессионное уравнение:

$$СБ = 0,78 + 0,20К + 0,20Б + 0,08ПП + 0,07С + 0Д0Н + 0,08В + 0Д2П, \text{ где}$$

- СБ — общая оценка сервиса на борту;
- К — комфортабельность салона;

- Б — работа бортпроводников;
- ПП — питание во время полета;
- С — спиртные напитки;
- Н — дорожные наборы;
- В — видеопрограмма;
- П — пресса.

Последний показатель, на который целесообразно обращать внимание при интерпретации результатов регрессионного анализа, — это стандартная ошибка, рассчитываемая для каждого коэффициента в регрессионном уравнении (столбец Std. Error). При 95%-ном доверительном уровне каждый коэффициент может отклоняться от величины В на $\pm 2 \times \text{Std.Error}$. Это означает, что, например, коэффициент при параметре Комфортабельность салона (равный 0,202) в 95 % случаев может отклоняться от данного значения на $\pm 2 \times 0,016$ или на $\pm 0,032$. Минимальное значение коэффициента будет равно $0,202 - 0,032 = 0,17$; а максимальное — $0,202 + 0,032 = 0,234$. Таким образом, в 95 % случаев коэффициент при параметре «комфортабельность салона» варьируется в пределах от 0,17 до 0,234 (при среднем значении 0,202). На этом интерпретация результатов регрессионного анализа может считаться завершённой. В нашем случае следует повторить все шаги еще раз: сначала для бизнес -, потом для эконом-класса.

Теперь давайте рассмотрим другой случай, когда необходимо графически представить зависимость между двумя переменными (одной зависимой и одной независимой) при помощи регрессионного анализа. Например, если мы примем итоговую оценку полета авиакомпанией X в 2001 г. за зависимую переменную S_{2001} , а тот же показатель в 2000 г. — за независимую переменную S_{2000} , то для построения уравнения тренда (или регрессионного уравнения) нужно будет определить параметры соотношения $S_{2001} = a + b \times S_{2000}$. Построив данное уравнение, также можно построить регрессионную прямую и, зная исходную итоговую оценку полета, спрогнозировать величину данного параметра на следующий год.

Эту операцию следует начать с построения регрессионного уравнения. Для этого повторите все вышеописанные шаги для двух переменных: зависимой Итоговая оценка 2001 и независимой Итоговая оценка 2000. Вы получите коэффициенты, при помощи которых можно в дальнейшем строить линию тренда (как в SPSS, так и любыми другими средствами). В нашем случае полученное регрессионное уравнение имеет вид: $S_{2001} = 0,18 + 0,81 \times S_{2000}$. Теперь построим уравнение линии тренда в SPSS.

Диалоговое окно Linear Regression имеет встроенное средство для построения графиков — кнопку Plots. Однако это средство, к сожалению, не позволяет на одном графике построить две переменные: S_{2001} и S_{2000} . Для того чтобы построить тренд, необходимо использовать меню Graphs ► Scatter. На экране появится диалоговое окно Scatterplot (рис. 4.32), которое служит для выбора типа диаграммы. Выберите вид Simple. Максимально возможное число независимых переменных, которое можно изобразить графически, — 2. Поэтому при необходимости графического построения зависимости одной переменной (зависимой) от двух независимых (например, если бы в нашем распоряжении были данные не по двум, а по трем годам), в окне Scatterplot следует выбрать 3-D. Схема построения трехмерной диаграммы рассеяния не имеет существенных отличий от описываемого способа построения двумерной диаграммы.

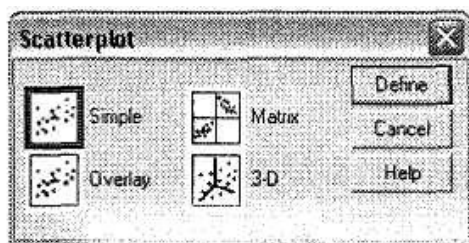


Рис. 4.33. Диалоговое окно Scatterplot

После щелчка на кнопке Define на экране появится новое диалоговое окно, представленное на рис. 4.34. Поместите в поле Y Axis зависимую переменную (Итоговая оценка 2001), а в поле X Axis — независимую (Итоговая оценка 2000). Щелкните на кнопке OK, что приведет к построению диаграммы рассеяния.

Для того чтобы построить линию тренда, дважды щелкните мышью на полученной диаграмме; откроется окно SPSS Chart Editor. В этом окне выберите пункт меню Chart ► Options; далее пункт Total в области Fit Line; щелкните на кнопке Fit Options. Откроется диалоговое окно Fit Line, выберите в нем тип аппроксимирующей линии (в нашем случае Linear regression) и пункт Display R-square in legend. После закрытия окна SPSS Chart Editor в окне SPSS Viewer появится линейный тренд, аппроксимирующий наши наблюдения по методу наименьших квадратов. Также на диаграмме будет отражаться величина R^2 , которая, как было сказано выше, обозначает долю совокупной вариации, описываемой данной моделью (рис. 4.35). В нашем примере она равна 53 %.

С линейным регрессионным анализом связано множество интегральных показателей, рассчитываемых на основании коэффициентов регрессии (чаще всего стандартизированных). В качестве примера приведем расчет коэффициента потребительской привлекательности продукта/марки (Consumer Attractiveness), или коэффициента СА.

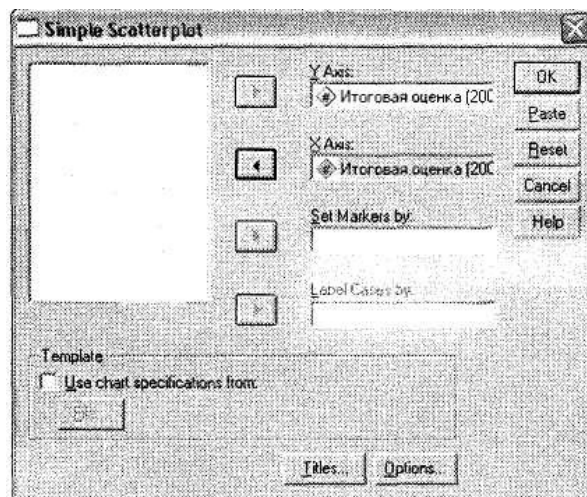


Рис. 4.34. Диалоговое окно Simple Scatterplot

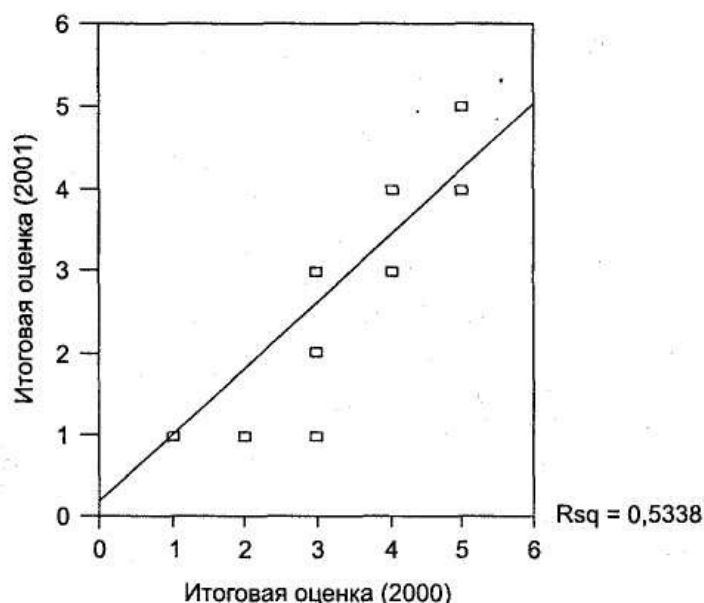


Рис. 4.35. Диаграмма Scatterplot с построенной линией тренда

Этот коэффициент вводится в маркетинговых исследованиях для удобства сравнения привлекательности для респондентов анализируемых продуктов/марок. В анкете должны присутствовать вопросы типа **Оцените представленные параметры продукта/марки X**, в которых респондентам предлагается дать свои оценки частным параметрам продукта или марки X, скажем, по пятибалльной шкале (от 1 — очень плохо до 5 — отлично). В конце списка оцениваемых частных параметров респонденты должны поставить итоговую оценку продукту/марке X. При анализе полученных в ходе опроса ответов респондентов на основании оценок респондентов формируются:

- матрица средневзвешенных оценок по параметрам продукта/марки;
- список стандартизированных β - коэффициентов регрессии (оценка влияния частных параметров продукта/марки X на его/ее общую оценку).

Далее коэффициент СА рассчитывается по следующей формуле:

$$CA = \frac{\sum_{i=1}^n s_i m_i}{n},$$

где n — число параметров, формирующих итоговую оценку продукта или марки:

$$n = \{ \Pi_1 + \Pi_2 + \dots + \Pi_k \}$$

s_i - — значимость для респондентов параметра с индексом i (стандартизированный β -коэффициент регрессии, оценивающей влияние частных параметров на общую оценку продукта/марки, подробнее см. выше); m_i — уровень средневзвешенной оценки продукта/марки по параметру с индексом i (при наличии пятибалльной шкалы):

- | | |
|------------|---|
| $m_i = 2$ | при высоком уровне оценки (средневзвешенный балл $\geq 4,5$) |
| $m_i = 1$ | при среднем уровне оценки (средневзвешенный балл $\geq 4,0$ и $< 4,5$) |
| $m_i = -1$ | при низком уровне оценки (средневзвешенный балл $\geq 3,0$ и $< 4,0$) |
| $m_i = -2$ | при неудовлетворительной оценке (средневзвешенный балл $< 3,0$) |

Рассчитанный для каждого конкурирующего продукта/марки коэффициент СА показывает его/ее относительную позицию в структуре потребительских предпочтений. Данный интегральный показатель учитывает уровень оценок по каждому параметру, скорректированный на их значимость. При этом он может изменяться в пределах от -1 (наихудшая относительная позиция среди всех рассматриваемых продуктов/марок) до 1 (наилучшее положение); 0 означает, что данный продукт/ марка ничем особенным не выделяется в глазах респондентов.

Итогом расчета коэффициента СА является рейтинг конкурентов по данному показателю. На основании рейтинга можно сделать важные выводы относительно лидерства и аутсайдерства конкретных продуктов/марок на потребительском рынке.

Мы завершаем рассмотрение ассоциативного анализа. Данная группа статистических методов применяется в отечественных компаниях в настоящее время достаточно широко (особенно это касается перекрестных распределений). Вместе с тем хотелось бы подчеркнуть, что только лишь перекрестными распределениями ассоциативные методы не ограничиваются. Для проведения действительно глубокого анализа следует расширить спектр применяемых методик за счет методов, описанных в настоящей главе.

Глава 5. Классификационный анализ

Цель классификационного анализа — классификация респондентов и/или переменных по определенным целевым группам. Наиболее распространенными примерами использования классификационного анализа в маркетинговых исследованиях являются:

- сегментирование респондентов по заранее известным (логистическая регрессия и дискриминантный анализ) или не известным (факторный и кластерный анализ) целевым группам;

- классификация переменных по макрокатегориям, то есть сокращение их числа до нескольких значимых групп (факторный и кластерный анализ).

Далее в разделе мы рассмотрим эти статистические методики в указанном порядке, а также приведем примеры задач из практики маркетинговых исследований, решаемых с помощью классификационного анализа.

5.1. Логистическая регрессия и дискриминантный анализ

Логистическая регрессия и дискриминантный анализ применяются в том случае, когда необходимо классифицировать (сегментировать) респондентов по целевым группам, которые, в свою очередь, представлены уровнями (вариантами ответа) одной одновариантной переменной.

Примером задачи, решаемой при помощи этих статистических методов, может служить задача классифицировать респондентов по двум группам — покупающие горчицу и не покупающие горчицу — на основании их социально-демографических характеристик (пол, возраст, доход, количество членов семьи и т. п.). Как вы видите, в процедурах логистической регрессии и дискриминантного анализа присутствуют переменные — критерии сегментирования и одна переменная, кодирующая целевые группы, на которые следует разделить респондентов на основании критериев сегментирования.

Необходимо отметить, что спектр возможностей применения логистической регрессии уже, чем для дискриминантного анализа, поэтому использование дискриминантного анализа в качестве универсального метода предпочтительнее. Более того, рекомендуется всегда начинать классификационное исследование именно с дискриминантного анализа, а не с логистической регрессии, — и применять последнюю в случае неуверенности в результатах дискриминантного анализа. Это связано, в частности, с тем, что при применении методов логистической регрессии еле дует четко представлять, какой тип имеют зависимая и независимые переменные и, исходя из этого, выбирать одну из трех возможных процедур логистической регрессии: бинарную, мультиномиальную или порядковую. При дискриминантном анализе мы всегда имеем дело только с одной статистической процедурой, в которой принимают участие одна категориальная зависимая переменная и несколько независимых переменных с любым типом шкалы. Таким образом, дискриминантный анализ является более универсальной методикой (что особенно важно для исследователей, имеющих незначительный опыт в статистическом анализе данных).

В разделах 5.1.1 и 5.1.2 мы на конкретных примерах покажем, как можно использовать процедуры логистической регрессии и дискриминантного анализа в маркетинговых исследованиях. При этом мы увидим, что, несмотря на преимущества универсального дискриминантного анализа, логистическая регрессия в некоторых случаях дает наивысшую четкость классификации.

5.1.1. Бинарная и мультиномиальная логистические регрессии

В настоящем разделе мы рассмотрим два основных типа логистической регрессии — бинарную и мультиномиальную, а также дадим общий обзор порядковой логистиче-

ской регрессии. Цель статистического анализа при применении методов логистической регрессии — определить вероятность того, что тот или иной респондент (на основании определенных характеристик) попадет в ту или иную целевую группу. На практике описываемые методы, согласно значениям одной или нескольких независимых переменных (факторов), позволяют классифицировать респондентов по двум (бинарная) или более (мультиномиальная) группам, которые выражаются уровнями (вариантами ответа) какой-либо одной переменной.

Например, имеются ответы респондентов на вопрос Интересно ли Вам предложение о покупке земельного участка недалеко от Москвы? с вариантами ответа Да и Нет. Требуется выяснить, какие факторы в наибольшей степени определяют решение потенциальных покупателей о приобретении земельного участка. Для этого респондентам задается ряд вопросов с просьбой указать, какие элементы инфраструктуры им необходимы на данном участке, какое расстояние от Москвы является для них оптимальным, каков должен быть размер данного участка, должен ли на участке быть дом и т. п. Используя в данном случае метод бинарной логистической регрессии, можно классифицировать всех респондентов по двум целевым группам: заинтересованные в покупке земельного участка (потенциальные покупатели) и не заинтересованные. Также для каждого респондента в выборке будет рассчитана вероятность попадания в ту или иную группу.

Различие между рассматриваемыми двумя методами логистической регрессии заключается в количестве категорий и типе зависимой переменной, а также типе независимых переменных. Так, в случае бинарной логистической регрессии исследуется зависимость дихотомической переменной от одной или нескольких независимых переменных, имеющих любой тип шкалы. Мультиномиальная логистическая регрессия является разновидностью бинарной, в которой зависимая переменная имеет более двух категорий. Независимые переменные должны относиться либо к номинальной, либо к порядковой шкале.

Еще в версии SPSS 11-12 был введен новый метод логистической регрессии: порядковая. Он используется в том случае, когда зависимая переменная относится к порядковой шкале. Причем независимые переменные должны быть либо номинальными, либо порядковыми. Мультиномиальный логистический регрессионный анализ является наиболее универсальным и, в целом, способен заменить собой два других метода. Однако наиболее качественное приближение статистических моделей может быть достигнуто только при использовании именно трех описываемых методов: для каждого случая — свой. В табл. 5.1 систематизированы основные характеристики переменных, участвующих в рассматриваемых трех типах логистического регрессионного анализа.

Таблица 5.1. Основные характеристики переменных, участвующих в анализе

Бинарная логистическая регрессия			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Она	Дихотомическая	Любое	Любой
Мультиномиальная логическая регрессия			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Номинальная Порядковая	Любое	Номинальная Порядковая
Порядковая логистическая регрессия			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Порядковая	Любое	Номинальная Порядковая

Необходимо отметить, что ранее в SPSS отсутствовала стандартная возможность проведения специализированного логистического регрессионного анализа для зависимых переменных с порядковой шкалой. Для любых переменных с числом категорий больше двух применялся мультиномиальный регрессионный анализ. Дело в том, что недавно введенная в практику анализа порядковая логистическая регрессия имеет некоторые особенности, учитывающие именно специфику порядковой шкалы (связанных упорядоченных категорий). Однако в настоящем пособии порядковая логистическая регрессия не рассматривается отдельно — в первую очередь из-за того, что она не обладает какими-либо существенными преимуществами над мультиномиальным методом. Вы можете спокойно применять мультиномиальную регрессию и в случае номинальной, и в случае порядковой зависимой переменной. Если вы все же решите провести порядковый логистический регрессионный анализ, вы без труда в нем разберетесь, так как данный процесс практически не отличается от построения мультиномиальной логистической регрессии.

Далее мы рассмотрим примеры проведения статистического анализа с использованием логистической регрессии отдельно для бинарной и мультиномиальной логистической регрессии.

Начнем с наиболее простого случая — бинарной логистической регрессии. Предположим, в ходе маркетингового исследования проводится оценка востребованности выпускников одного из московских вузов. В анкете респондентам в числе прочих задаются три вопроса:

- Работаете ли вы? (q1);
- В каком году Вы окончили вуз? (q21);
- Каков был Ваш средний балл при выпуске из вуза? (aver), а также уточняется пол опрошенных (q22).

В ходе логистического анализа мы оценим влияние независимых переменных q21, q22 и aver на зависимую переменную q1. Другими словами, мы попытаемся предсказать трудоустройство выпускников вуза на основании пола, года окончания вуза и среднего балла, полученного за годы обучения.

Для того чтобы задать параметры построения регрессионной модели при помощи бинарного логистического метода, воспользуемся меню Analyze ► Regression ► Binary Logistic. В открывшемся диалоговом окне Logistic Regression (рис. 5.1) выберите в левом списке всех доступных переменных зависимую (в нашем случае q1) и поместите ее в поле Dependent. Затем в область Covariates поместите исследуемые независимые переменные (q21, q22, aver) и выберите метод их включения в регрессионный анализ. При числе независимых переменных больше двух следует выбрать не установленный по умолчанию метод одновременного включения всех переменных (Enter), а один из пошаговых. Наиболее часто используемым пошаговым методом является Backward:LR. Кнопка Select позволяет включить в анализ не всех респондентов из выборочной совокупности, а только отдельную целевую группу.

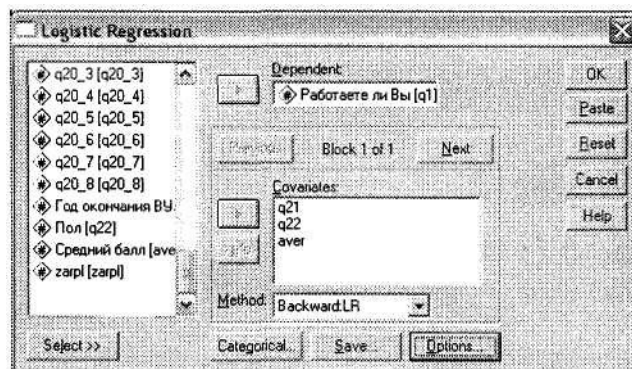


Рис. 5.1. Диалоговое окно Logistic Regression

Кнопкой Categorical следует воспользоваться, если в качестве одной из независимых переменных выступает номинальная переменная с числом категорий больше двух. В данном случае в диалоговом окне Define Categorical Variables (рис. 5.2) следует поместить в область Categorical Covariates такую переменную (в нашем случае таких переменных нет). Далее следует выбрать в раскрывающемся списке Contrast пункт Deviation и щелкнуть на кнопке Change. В результате из каждой номинальной переменной будет создано несколько дихотомических переменных (по числу категорий исходной переменной).

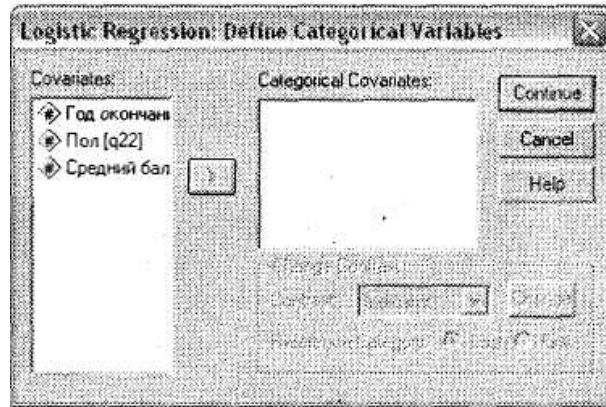


Рис. 5.2. Диалоговое окно Define Categorical Variables

При помощи кнопки Save в главном диалоговом окне анализа (рис. 5.3) можно задать создание новых переменных, содержащих значения, рассчитанные в ходе регрессионного анализа. Так давайте создадим две новые переменные, содержащие:

- принадлежность к определенной группе классификации (параметр Group membership);
- вероятность попадания респондента в каждую из двух рассматриваемых групп (параметр Probabilities).

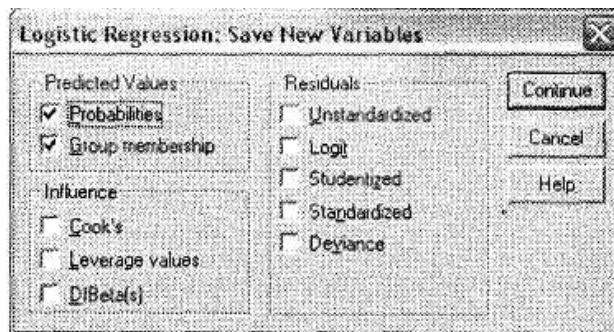


Рис. 5.3. Диалоговое окно Save New Variables

Кнопка Options не предоставляет исследователю никаких важных возможностей, поэтому ее можно не использовать. После щелчка на кнопке О К в главном диалоговом окне Logistic Regression в окне SPSS Viewer будут выведены результаты бинарного логистического регрессионного анализа.

Далее мы рассмотрим наиболее существенные для маркетингового анализа результаты. В таблице Omnibus Tests of Model Coefficients отображаются результаты оценки качества приближения статистической модели (рис. 5.4). Поскольку мы задали пошаговый метод, мы должны смотреть на результаты последнего шага (в нашем случае Step 2). Положительным результатом считается возрастание величины Chi-square при переходе на каждый следующий шаг (строка Step) при высоком уровне значимости (Sig. < 0,05). Качество всей модели оценивается на основании статистической значимости в строке Model. В нашем случае на втором шаге получена отрицательная величина Chi-square,

однако она не является значимой (Sig. = 0,913), к тому же общая значимость всей модели весьма высока (Sig. < 0,001). Поэтому построенную модель следует признать значимой и практически пригодной.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	25,005	3	,000
	Block	25,005	3	,000
	Model	25,005	3	,000
Step 2 ^a	Step	-,012	1	,913
	Block	24,993	2	,000
	Model	24,993	2	,000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Рис. 5.4. Таблица Omnibus Tests of Model Coefficients

Следующая таблица Model Summary (рис. 5.5) позволяет оценить долю совокупной дисперсии, описываемой построенной моделью (величина R Square). Рекомендуется использовать величину Nagelkerke. В нашем случае эта величина мала (лишь 6 %). Положительным результатом можно считать величину Nagelkerke R Square, превышающую 0,50.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	611,674	,038	,061
2	611,686	,038	,061

Рис. 5.5. Таблица Model Summary

Далее следуют результаты классификации (таблица Classification Table, рис. 5.6), в которой реально наблюдаемые показатели принадлежности к той или иной из двух исследуемых групп сопоставляются с предсказанными на основе логистической регрессионной модели. В нашем случае из строки Overall Percentage мы видим, что построенная модель позволяет корректно классифицировать 80,4 % респондентов. Также можно сделать соответствующие выводы о корректности классификации для каждой из двух рассматриваемых групп.

Из следующей таблицы (рис. 5.7) можно выяснить статистическую значимость независимых переменных, включенных в анализ (в нашем случае q22 и aver), а также нестандартизированные регрессионные коэффициенты, являющиеся коэффициентами регрессионной функции. На основании этих коэффициентов (включая константу Constant) вы можете спрогнозировать принадлежность к определенной группе каждого конкретного респондента в выборке. Это делается следующим образом.

Classification Table^a

Observed			Predicted		
			Работаете ли Вы		Percentage Correct
			Работают	Не работают	
Step 1	Работаете ли Вы	Работают	518	0	100,0
		Не работают	126	0	,0
	Overall Percentage				80,4
Step 2	Работаете ли Вы	Работают	518	0	100,0
		Не работают	126	0	,0
	Overall Percentage				80,4

a. The cut value is ,500

Рис. 5.6. Таблица Classification Table

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Q21	,013	,119	,012	1	,913	1,013
	Q22	,990	,238	17,343	1	,000	2,691
	AVER	-,758	,233	10,635	1	,001	,468
	Constant	,254	1,048	,059	1	,808	1,289
Step 2	Q22	,990	,238	17,365	1	,000	2,692
	AVER	-,758	,233	10,615	1	,001	,469
	Constant	,280	1,021	,075	1	,784	1,323

a. Variable(s) entered on step 1: Q21, Q22, AVER.

Рис. 5.7. Таблица Variables in the Equation

Например, выпускник вуза получил средний балл 3,3 (aver = 3,3); это женщина (q22 = 2). В таком случае уравнение регрессии будет выглядеть следующим образом:

$$z = 0,280 - 0,758 \times 3,3 + 0,990 \times 2 = -1,798,$$

а вероятность для рассматриваемого респондента оказаться в одной из анализируемых групп классификации (это всегда группа зависимой переменной, имеющая больший код, в нашем случае 2 — Не работают) будет рассчитываться по формуле:

$$p = \frac{1}{1 + e^z} \approx 0,68$$

Таким образом, женщина со средним баллом 3,3 имеет достаточно высокие шансы оказаться безработной (68 %).

Теперь рассмотрим пример проведения мультиномиальной логистической регрессии. В качестве исходных данных мы будем использовать три независимые переменные из предыдущего примера, а в качестве зависимой — переменную q24 Заработная плата с пятью категориями, кодирующими интервалы зарплаты.

Откройте диалоговое окно Multinomial Logistic Regression при помощи меню Analyze ► Regression ► Multinomial Logistic (рис. 5.8). В поле для зависимой переменной поместите переменную q24, а в область для независимых переменных — q21, q22 и aver.

Кнопка Model позволяет задать конкретный тип модели (полнофакторная, основные эффекты или пользовательская), однако для маркетинговых исследований мы советуем ничего не менять в окне Model.

При помощи кнопки Statistics вызывается одноименное диалоговое окно (рис. 5.9). В нем следует оставить выбранные по умолчанию три параметра: Summary statistics, Likelihood ratio test и Parameter estimates, а также выбрать еще один пункт — Cell Probabil-

ities.

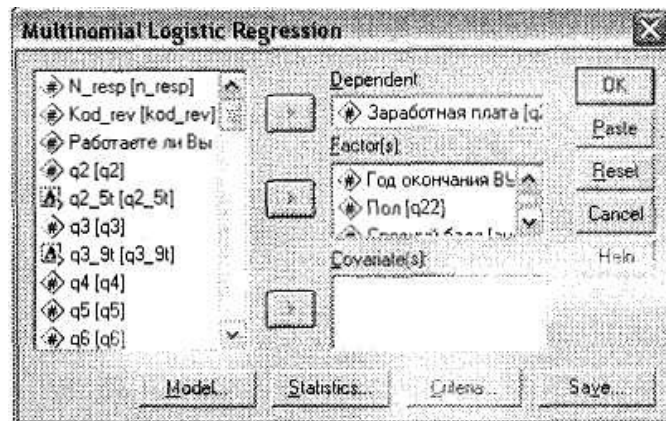


Рис. 5.8. Диалоговое окно Multinomial Logistic Regression

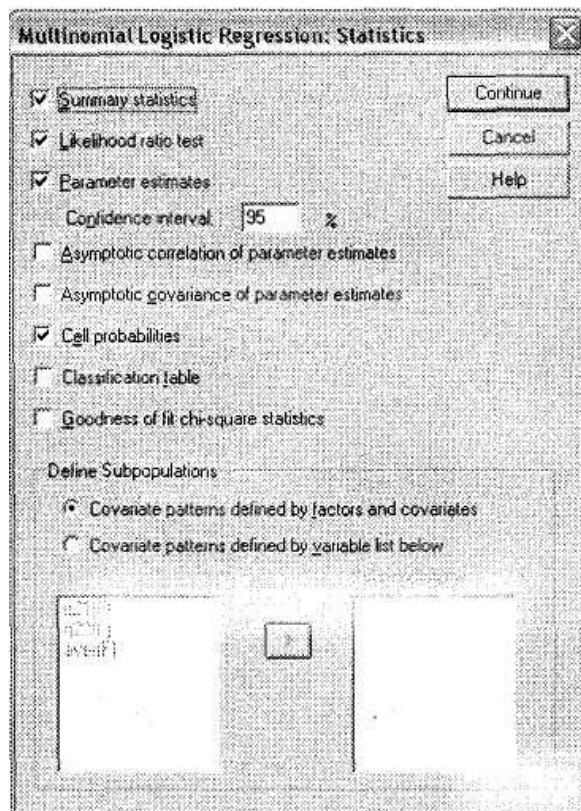


Рис. 5.9. Диалоговое окно Statistics

Кнопка Criteria не предоставляет маркетологам существенных для решения их задач функций, поэтому используется редко.

При помощи кнопки Save (рис. 5.10) можно задать новые переменные, содержащие принадлежность к определенной классификационной группе (параметр Predicted category) и вероятность попадания в данные категории (параметр Predicted probabilities membership).

После щелчка на кнопке OK в главном диалоговом окне Multinomial Logistic Regression в окне SPSS Viewer появятся результаты расчетов. Первая таблица, содержащая важные для нас сведения, — это Model Fitting Information, показанная на рис. 5.11. Высокая статистическая значимость построенной модели (Sig. < 0,001) свидетельствует о ее высоком качестве и пригодности для решения практических задач.

Вторая значимая таблица Pseudo R-Square предоставляет возможность оценить долю совокупной дисперсии в зависимой переменной, объясняемой выбранными для анализа независимыми переменными (по тесту Nagelkerke). В нашем случае построенная модель объясняет 15 % совокупной дисперсии (рис. 5.12).

Таблица Likelihood Ratio Tests (рис. 5.13) позволяет сделать выводы относительно статистической значимости каждой из зависимых переменных, входящих в построенную модель. В нашем случае все три исследуемые переменные оказывают весьма значимое влияние на зависимую переменную (Sig. < 0,05).

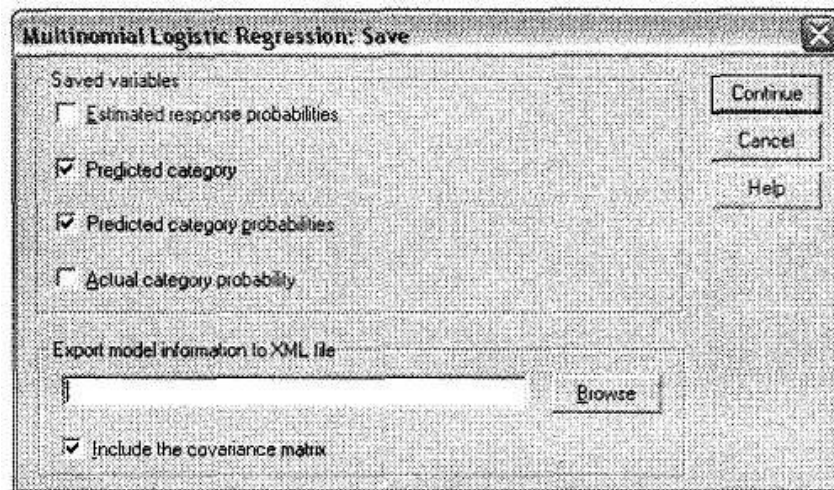


Рис. 5.10. Диалоговое окно Save

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	234,955			
Final	166,266	68,689	20	,000

Рис. 5.11. Таблица Model Fitting Information

Pseudo R-Square

Cox and Snell	,142
Nagelkerke	,152
McFadden	,056

Рис. 5.12. Таблица Pseudo R-Square

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	166,266 ^a	,000	0	.
Q21	193,430	27,164	8	,001
Q22	191,157	24,891	4	,000
AVER	191,216	24,950	8	,002

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

- a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Рис. 5.13. Таблица Likelihood Ratio Tests

Parameter Estimates								95% Confidence Interval for Exp(B)	
Заработная плата		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
До \$400	Intercept	2,732	,495	30,512	1	,000			
	[Q21=1]	-1,115	,642	3,015	1	,083	,328	9,307E-02	1,155
	[Q21=2]	-1,128	,571	3,903	1	,048	,324	,106	,991
	[Q21=3]	0 ^a	.	.	0
	[Q22=1]	-1,467	,509	8,312	1	,004	,231	8,504E-02	,625
	[Q22=2]	0 ^a	.	.	0
	[AVER=3]	17,785	1,014	307,403	1	,000	5,3E+07	7250271,789	388544940,6
	[AVER=4]	1,168	,605	3,730	1	,053	3,215	,983	10,515
	[AVER=5]	0 ^a	.	.	0
От \$400 до \$800	Intercept	2,574	,493	27,254	1	,000			
	[Q21=1]	-,686	,624	1,140	1	,286	,514	,151	1,746
	[Q21=2]	-1,080	,564	3,526	1	,060	,346	,115	1,047
	[Q21=3]	0 ^a	.	.	0
	[Q22=1]	-,802	,494	2,636	1	,104	,448	,170	1,181
	[Q22=2]	0 ^a	.	.	0
	[AVER=3]	17,814	,991	316,039	1	,000	4,5E+07	6401742,888	311206836,6
	[AVER=4]	1,245	,596	4,366	1	,037	3,473	1,080	11,163
	[AVER=5]	0 ^a	.	.	0
От \$800 до \$800	Intercept	1,758	,515	11,869	1	,001			
	[Q21=1]	-,234	,645	,132	1	,717	,791	,223	2,803
	[Q21=2]	-,859	,596	2,078	1	,149	,423	,132	1,382
	[Q21=3]	0 ^a	.	.	0
	[Q22=1]	-,362	,515	,493	1	,483	,696	,254	1,912
	[Q22=2]	0 ^a	.	.	0
	[AVER=3]	-1,957	,000	.	1	.	,141	,141	,141
	[AVER=4]	,580	,622	,869	1	,351	1,786	,528	6,050
	[AVER=5]	0 ^a	.	.	0
От \$800 до \$1000	Intercept	-,071	,638	,012	1	,912			
	[Q21=1]	1,174	,733	2,568	1	,109	3,235	,770	13,598
	[Q21=2]	,035	,719	,002	1	,961	1,036	,253	4,241
	[Q21=3]	0 ^a	.	.	0
	[Q22=1]	,176	,582	,092	1	,762	1,193	,382	3,729
	[Q22=2]	0 ^a	.	.	0
	[AVER=3]	19,894	,000	.	1	.	4,4E+08	436511812,6	436511812,6
	[AVER=4]	,119	,710	,028	1	,867	1,126	,280	4,525
	[AVER=5]	0 ^a	.	.	0

a. This parameter is set to zero because it is redundant.

Рис. 5.14. Таблица Parameter Estimates

Следующая таблица, Parameter Estimates (рис. 5.14), отражает нестандартизированные регрессионные коэффициенты, на основании которых происходит построение регрессионного уравнения. Также для каждого сочетания анализируемых переменных рассчитана статистическая значимость их влияния на зависимую переменную. В дальнейшем рассчитать вероятность попадания того или иного респондента в одну из исследуемых групп зависимой переменной можно по вышеприведенной формуле (показана при обсуждении бинарной логистической регрессии).

Однако в маркетинговых исследованиях чаще всего возникает необходимость классифицировать по группам не отдельных респондентов, а целые целевые группы. Для этого служит таблица Observed and Predicted Frequencies, представленная на рис. 5.15. В столбце Percentage ► Predicted показаны вероятности попадания каждой исследуемой целевой группы респондентов в ту или иную категорию зависимой переменной. Так, например, мы видим, что 20 % мужчин, окончивших ВУЗ в 2001 г. и получивших средний балл 3,0, зарабатывают до \$ 400 в месяц.

Observed and Predicted Frequencies								
Средний балл	Пол	Год окончания вуза	Зарплата	Frequency			Percentage	
				Observed	Predicted	Pearson Residual	Observed	Predicted
3	Мужчины	2001 г.	До \$ 400	1	.802	.247	25,0 %	20,1 %
			От \$ 400 до \$ 600	2	1,123	.976	50,0 %	28,1 %
			От \$ 600 до \$ 800	0	.000	.000	.0 %	.0 %
			От \$ 800 до \$ 1000	1	2,075	-1,076	25,0 %	51,8 %
	Женщины	2000 г.	Более \$ 1000	0	.000	.000	.0 %	.0 %
			До \$ 400	1	.207	1,539	100,0 %	29,7 %
			От \$ 400 до \$ 600	0	.229	-.544	.0 %	22,9 %
			От \$ 600 до \$ 800	0	.000	.000	.0 %	.0 %
			От \$ 800 до \$ 1000	0	.475	-.951	.0 %	47,5 %
			Более \$ 1000	0	.000	.000	.0 %	.0 %
		2001 г.	До \$ 400	0	.001	-1,281	.0 %	46,1 %
			От \$ 400 до \$ 600	0	.649	-.980	.0 %	32,4 %
			От \$ 600 до \$ 800	0	.000	.000	.0 %	.0 %
			От \$ 800 до \$ 1000	2	.460	2,823	100,0 %	22,5 %
			Более \$ 1000	0	.000	.000	.0 %	.0 %

Рис. 5.15. Таблица Observed and Predicted Frequencies

5.1.2. Дискриминантный анализ

Дискриминантный анализ является более универсальной статистической процедурой по сравнению с рассмотренными выше методами логистической регрессии. Основным результатом проведения дискриминантного анализа являются (также как для логистической регрессии) рассчитанные вероятности попадания каждого респондента в ту или иную группу, а также переменная, кодирующая принадлежность их к данным группам. Наряду с этой информацией по результатам дискриминантного анализа можно составить уравнение дискриминантной функции.

В табл. 5.2 приведены основные характеристики переменных, участвующих в дискриминантном анализе.

Таблица 5.2. Основные характеристики переменных, участвующих в анализе

Дискриминантный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Одна	Номинальная	Любое	Любой
	Порядковая		

При выборе зависимой переменной для дискриминантного анализа следует помнить, что увеличение числа категорий в ней практически всегда влечет уменьшение качества статистической модели, то есть ее точности и надежности. Поэтому рекомендуется использовать в качестве зависимых переменные с малым количеством категорий (или преобразовывать существующие переменные к данному виду).

Для описания процесса проведения дискриминантного анализа применим следующие исходные данные. Проводится маркетинговое исследование потенциального спроса на услуги нового развлекательного комплекса. Респонденты в ходе опроса отвечают на вопрос Будете ли Вы посещать новый комплекс? (q26) с вариантами ответа Да и Нет. В качестве независимых переменных, характеризующих респондентов, выделены:

- возраст (q18);
- род занятий (q19);
- среднемесячный доход (q20);
- количество членов семьи (q21);

- среднемесячные расходы на досуг (q22);
- пол (q23).

В результате дискриминантного анализа мы разделим респондентов на посетителей и не посетителей нового центра на основании выделенных социально-демографических характеристик опрошенных.

Откройте диалоговое окно Discriminant Analysis при помощи меню Analyze ► Classify ► Discriminant (рис. 5.16). Поместите переменную q26 в поле для зависимых переменных Grouping Variable, а анализируемые независимые переменные — в область Independents. Выберите пошаговый метод ввода независимых переменных в модель (параметр Use stepwise method).

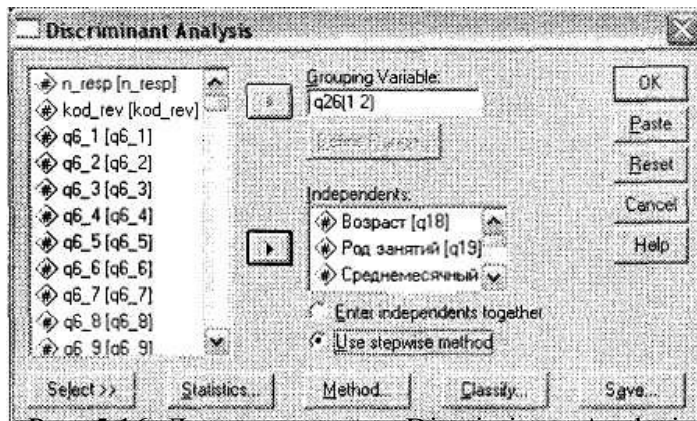


Рис. 5.16. Диалоговое окно Discriminant Analysis

Далее щелкните на кнопке Define Range для определения границ изменения зависимой переменной q26 (рис. 5.17). В нашем случае минимальным значением (Minimum) является 1, а максимальным (Maximum) — 2.

При помощи диалогового окна Statistics, активизируемого одноименной кнопкой, следует задать вывод результатов одномерного дисперсионного анализа (параметр

Univariate ANOVA), теста Box (параметр Box's M), а также нестандартизованных коэффициентов регрессии (параметр Unstandardized) (рис. 5.18).

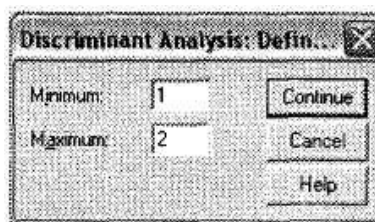


Рис. 5.17. Диалоговое окно Define Range

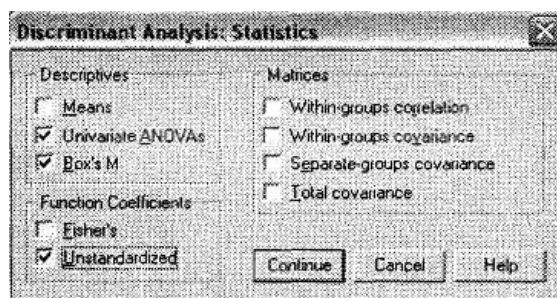


Рис. 5.18. Диалоговое окно Statistics

В следующем диалоговом окне, Stepwise Method, вызываемом при помощи кнопки Method, следует выбрать параметр Use probability of F (рис. 5.19). Активизация данного

параметра позволяет проводить введение переменных в регрессионную модель более гибко по сравнению с абсолютным значением F-статистики (параметр, выбранный по умолчанию).

В следующем диалоговом окне, Classification, нас интересует только один параметр — Summary Table (рис. 5.20),

Наконец, при помощи кнопки Save можно создать в исходном файле данных новые переменные, содержащие для каждого респондента в выборке прогнозируемую принадлежность к группе (параметр Predicted group membership) и вероятность попадания каждого респондента в данные группы (параметр Probabilities of group membership; см. рис. 5.21).

После выполнения вышеописанных шагов щелкните на кнопке ОК, чтобы запустить программу дискриминантного анализа на исполнение. После окончания расчетов в окне SPSS Viewer будут выведены результаты расчетов.

Первой важной для нас таблицей является Tests of Equality of Group Means (рис. 5.22). Она показывает, насколько значимо выбранные независимые переменные разделяют выборочную совокупность респондентов на исследуемые группы. В нашем случае получены весьма значимые результаты для всех исследуемых переменных ($\text{Sig.} < 0,05$). Это свидетельствует о том, что на их основании исследуемые группы зависимой переменной существенно различаются.

Следующая таблица, Test Results, показывает результаты теста Вох на значимость различия между категориями исследуемой зависимой переменной (рис. 5.23). В нашем случае данный тест показывает весьма высокую вероятность того, что данные различия являются статистически значимыми ($\text{Sig.} < 0,001$).

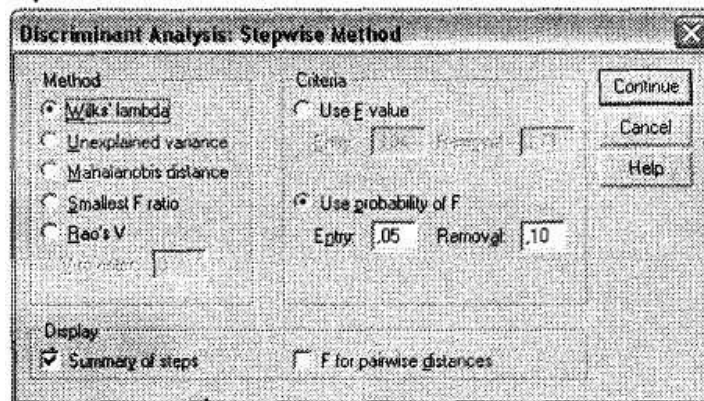


Рис. 5.19. Диалоговое окно Stepwise Method

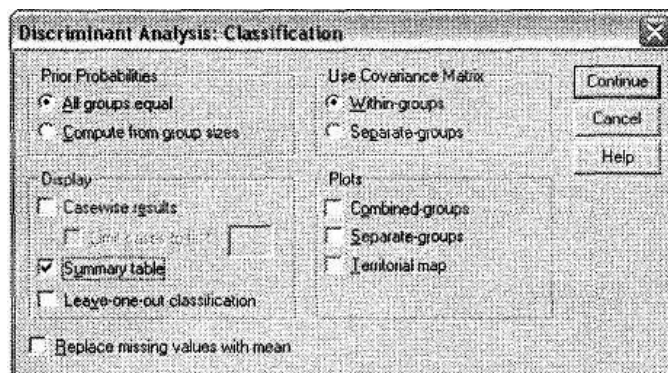


Рис. 5.20. Диалоговое окно Classification

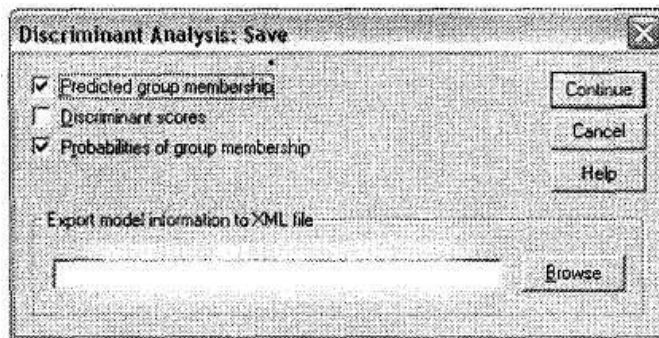


Рис. 5.21. Диалоговое окно Save

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
Возраст	,836	172,258	1	881	,000
Род занятий	,882	117,782	1	881	,000
Среднемесячный доход	,899	98,492	1	881	,000
Количество членов семьи	,917	79,220	1	881	,000
Среднемесячные расходы на досуг	,879	120,881	1	881	,000
Пол	,993	6,406	1	881	,012

Рис. 5.22. Таблица Tests of Equality of Group Means

Test Results		
Box's M		219,013
F	Approx.	21,734
	df1	10
	df2	633678,8
	Sig.	,000

Tests null hypothesis of equal population covariance matrices

Рис. 5.23. Таблица Test Results

Таблица Variables in the Analysis показывает, какие независимые переменные оказались включенными в итоговую дискриминантную модель на последнем шаге анализа (напомним, что мы выбрали пошаговый метод включения переменных в модель). В нашем случае последним шагом является шаг 4. На четвертом шаге у нас остались четыре независимые переменные из шести (рис. 5.24).

Variables in the Analysis

Step		Tolerance	Sig. of F to Remove	Wilks' Lambda
1	Возраст	1,000	,000	
2	Возраст	,935	,000	,882
	Род занятий	,935	,000	,836
3	Возраст	,887	,000	,838
	Род занятий	,926	,000	,811
	Количество членов семьи	,925	,000	,791
4	Возраст	,777	,000	,795
	Род занятий	,882	,000	,785
	Количество членов семьи	,925	,000	,777
	Среднемесячные расходы на досуг	,793	,000	,774

Рис. 5.24. Таблица Variables in the Analysis

Таблица Eigenvalues позволяет оценить качество разделения респондентов на заданные группы зависимой переменной (рис. 5.25). Соответствующий вывод можно сделать исходя из корреляционного коэффициента (столбец Canonical Correlation). В нашем случае данный коэффициент примерно равен 0,5, что свидетельствует о неудовлетворительном результате.

Еще одним важным показателем в этой таблице является собственное значение дискриминантной функции (столбец Eigenvalue). В общем случае большие значения Eigenvalues указывают на высокую точность подобранной дискриминантной функции. В нашем случае рассматриваемое собственное значение весьма мало, что является негативным фактом. Необходимо отметить, что при наличии у зависимой переменной более двух категорий в ходе дискриминантного анализа строится несколько дискриминантных функций (по количеству категорий зависимой переменной минус 1).

Следующая таблица (рис. 5.26) также позволяет оценить качество приближения дискриминантной модели. В нашем случае статистическая значимость (Sig. < 0,001)

указывает на существенные различия между средними значениями дискриминантных функций в двух исследуемых группах зависимой переменной.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,316 ^a	100,0	100,0	,490

a. First 1 canonical discriminant functions were used in the analysis.

Рис. 5.25. Таблица Eigenvalues

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,760	241,179	4	,000

Рис. 5.26. Таблица Wilks' Lambda

Следующие две таблицы (рис. 5.27 и 5.28) позволяют оценить, насколько отдельные независимые переменные, применяемые в дискриминантной функции, коррелируют с ее стандартизированными коэффициентами. В первой таблице приводятся стандартизированные коэффициенты, а во второй — корреляционные коэффициенты. При помощи стандартизированных коэффициентов, кроме всего прочего, можно непосредственно сравнивать относительный вклад каждой независимой переменной в различение двух исследуемых групп. Например, мы видим, что возраст респондентов влияет на их желание/нежелание посещать новый центр в 1,3 раза сильнее, чем род занятий.

Далее следуют коэффициенты дискриминантной функции (нестандартизированные), на основании которых и строится дискриминантное уравнение, по форме похожее на уравнение регрессии (рис. 5.29). Это просто множители при соответствующих переменных. С учетом константы уравнение дискриминантной функции имеет вид:

$$Z = -0,845 + 0,207 \times \text{Возраст} + 0,198 \times \text{Род_занятий} - 0,289 \times \text{Кол-во_членов_семьи} - 0,285 \times \text{Среднемесячные_расходы_на_досуг}$$

Теперь на основании данного уравнения можно рассчитать вероятность, с которой та или иная социально-демографическая целевая группа респондентов будет посещать новый центр. Подставив в дискриминантное уравнение соответствующие значения, можно сделать вывод о том, что студенты в возрасте 20 лет, проживающие одни и расходующие на свой досуг \$ 50 в месяц, скорее всего, будут посещать новый развлекательный центр (вероятность 79 %).

Таблица, представленная на рис. 5.30, показывает средние значения дискриминантной функции в каждой анализируемой группе зависимой переменной.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Возраст	,483
Род занятий	,387
Количество членов семьи	-,310
Среднемесячные расходы на досуг	-,307

Рис. 5.27. Таблица Standardized Canonical Discriminant Function Coefficients

Structure Matrix

	Function
	1
Возраст	,787
Среднемесячные расходы на досуг	-,659
Род занятий	,651
Количество членов семьи	-,534
Среднемесячный доход ^a	-,527
Пол ^a	,171

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Рис. 5.28. Таблица Structure Matrix

Canonical Discriminant Function Coefficient

	Function
	1
Возраст	,207
Род занятий	,198
Количество членов семьи	-,289
Среднемесячные расходы на досуг	-,285
(Constant)	-,845

Unstandardized coefficients

Рис. 5.29. Таблица Canonical Discriminant Function Coefficients

Functions at Group Centroids

	Function
	1
Готовность посещать магазин	
Собираются посещать	-,306
Не собираются посещать	1,031

Unstandardized canonical discriminant functions evaluated at group means

Рис. 5.30. Таблица Functions at Group Centroids

Завершает вывод результатов дискриминантного анализа таблица Classification Results, в последней строке которой содержится информация о точности построенной модели (рис. 5.31). В нашем случае мы видим, что 77,7 % респондентов были корректно отнесены к одной из двух исследуемых групп (77,7% of original grouped cases correctly classified). Результаты оценки корректности классификации варьируются в пределах от 50 % до 100 %, поэтому полученный нами результат — примерно 78 % — можно считать удовлетворительным.

Classification Results^a

			Predicted Group Membership		Total
			Собираются посещать	Не собираются посещать	
Original	Count	Готовность посещать магазин			
		Собираются посещать	603	176	779
		Не собираются посещать	49	179	228
		Ungrouped cases	3	0	3
%		Собираются посещать	77,4	22,6	100,0
		Не собираются посещать	21,5	78,5	100,0
		Ungrouped cases	100,0	,0	100,0

a. 77,7% of original grouped cases correctly classified.

Рис. 5.31. Таблица Classification Results

5.2. Факторный и кластерный анализ

Кластерный и факторный анализы преследуют ту же цель, что и рассмотренные в предыдущем разделе методы логистической регрессии и дискриминантного анализа: классифицировать переменные и/или категории респондентов по однородным группам (сегментам, кластерам). Однако между этими методами существует одно серьезное различие.

При дискриминантном анализе и логистической регрессии у нас заранее есть некая зависимая (результатирующая) переменная с двумя или более вариантами ответа (уровнями, категориями). Задача анализа в данном случае состоит в классификации имеющихся категорий респондентов (возрастных, половых и других) по этим уровням результирующей переменной. Эти два статистических метода позволяют сегментировать выборку на заранее известные целевые группы. При кластерном и факторном анализе ситуация иная: кластеры (сегменты, категории), на которые следует разделить выборку, заранее не известны. Задачей статистического анализа в данном случае будет не только формирование максимально однородных сегментов, но и выделение кластеров, по которым будет производиться сегментирование. Приведем пример релевантного задания для факторного (кластерного) анализа.

Исходные данные:

Респондентам (пассажирам международных рейсов авиакомпании X) в ходе опроса предлагалось 24 утверждения, по которыми они должны были выразить степень своего согласия либо несогласия по десятибалльной шкале — от 1 (совершенно не согласен) до 10 (абсолютно согласен). Предложенные утверждения описывают текущую конкурентную позицию рассматриваемой компании на международном рынке авиаперевозок. В результате опроса и последующих подготовительных этапов к статистическому анализу (см. раздел 3) был получен массив из 24 одновариантных переменных (q1-q24) с кодами ответов соответственно от 1 до 10 (интервальная шкала).

q1. Авиакомпания X обладает репутацией компании, превосходно обслуживающей пассажиров.

q2. Авиакомпания X может конкурировать с лучшими авиакомпаниями мира.

q3. Я верю, что у авиакомпании X есть перспективное будущее в мировой авиации.

q4. Я знаю, какой будет стратегия развития авиакомпании X в будущем.

q5. Я горжусь тем, что работаю в авиакомпании X.

q6. Внутри авиакомпании X хорошее взаимодействие между подразделениями.

q7. Каждый сотрудник авиакомпании прикладывает все усилия для того, чтобы обеспечить ее успех.

q8. Сейчас авиакомпания X быстро улучшается.

q9. Нам предстоит долгий путь, прежде чем мы сможем претендовать на то, чтобы называться авиакомпанией мирового класса.

q10. Авиакомпания X действительно заботится о пассажирах.

q11. Среди сотрудников авиакомпании имеет место высокая степень удовлетворенности работой.

q12. Я верю, что менеджеры высшего звена прикладывают все усилия для достижения успеха авиакомпании.

q13. Мне нравится, как в настоящее время авиакомпания X представлена визуально широкой общественности (в плане цветовой гаммы и фирменного стиля).

q14. Авиакомпания X — лицо России.

q15. Мы выглядим «вчерашним днем» по сравнению с другими авиакомпаниями.

q16. Обслуживание авиакомпании X является последовательным и узнаваемым во всем мире.

q17. Я бы не хотел, чтобы авиакомпания X менялась.

q18. Авиакомпания X необходимо меняться для того, чтобы использовать в полной мере имеющийся потенциал.

q19. Я думаю, что авиакомпании X необходимо представить себя в визуальном плане более современно.

q20. Изменения в авиакомпании X будут позитивным моментом. q21. Авиакомпания X — эффективная авиакомпания.

q22. Я бы хотел, чтобы имидж авиакомпании X улучшился с точки зрения иностранных пассажиров.

q23. Авиакомпания X — лучше, чем многие о ней думают.

q24. Важно, чтобы люди во всем мире знали, что мы — российская авиакомпания. Требуется:

Выявить схожие (то есть тесно коррелирующие между собой) утверждения и разделить их на несколько однородных групп, описывающих различные аспекты (макропараметры)

конкурентной позиции авиакомпании X на рынке. Другими словами, выделить группы схожих по значению параметров авиакомпании, характеризующих ее состояние на рынке с различных сторон.

Данную задачу невозможно решить методами логистической регрессии или дискриминантного анализа, так как у нас нет зависимой (результатирующей) переменной: есть только массив, на первый взгляд, независимых равнозначных параметров. Поставленную цель можно достичь при помощи либо факторного анализа, либо кластерного. Однако прежде, чем мы приступим к решению, следует сказать несколько слов об основных характерных чертах и различиях между этими двумя статистическими методами, предназначенными для решения схожих задач.

Факторный анализ позволяет разделить массив переменных на малое число групп, которые называются факторами. Классификация производится на основании критерия корреляции между переменными. В один фактор объединяются несколько переменных, тесно коррелирующих между собой и не коррелирующих или слабо коррелирующих с другими переменными, составляющими другие факторы. Таким образом, в результате факторного анализа мы получаем из несистематизированного массива данных несколько макропеременных, описывающих различные характеристики продукта компании (или другого исследуемого объекта). Основная сложность при проведении факторного анализа заключается в необходимости рационально интерпретировать полученные макрокатегории с точки зрения здравого смысла (применительно к целям и специфике конкретного исследования). Данная проблема не имеет универсального решения и подлежит отдельному анализу в каждом конкретном случае. Ниже мы продемонстрируем пример интерпретации результатов факторного анализа. Именно сложность интерпретации результатов является существенным ограничением рассматриваемой статистической методики, так как из-за невозможности логического описания полученных категорий иногда приходится вообще отказаться от ее использования.

Еще одним ограничением применения факторного анализа является ситуация, когда одна и та же переменная относится сразу к двум или более факторам, то есть переменную нельзя однозначно классифицировать. В таком случае следует либо отказаться от использования факторного анализа и попытаться применить другие статистические методики (например, кластерный анализ), либо заново пересчитать факторную модель без данной переменной, а затем вручную отнести неоднозначную переменную к тому или иному фактору на основании логических соображений.

Далее приведены основные примеры использования факторного анализа в маркетинговых исследованиях.

Сегментирование рынка. Факторный анализ применяется для выявления агрегатных переменных, являющихся основанием для сегментирования потребителей. Например, потребители плавленых сыров могут характеризоваться различной степенью значимости, которую они видят в исследуемых характеристиках данного продукта (респондентов просят оценить по пятибалльной шкале важность нескольких характеристик плавленых сыров: срок хранения, калорийность, процент жирности и т. д.). Здесь факторный анализ позволит выявить целевые сегменты потребителей на основании значимости для них различных групп факторов:

- покупатели, ориентирующиеся при выборе плавленого сыра преимущественно

на ценовые факторы (стоимость, скидки);

- покупатели, ориентирующиеся на качество исследуемого продукта (срок хранения, состав ингредиентов, вкус);
- покупатели, выбирающие сыр в основном по внешнему виду (дизайн упаковки).

В целом следует отметить, что в настоящее время все большую популярность среди исследователей приобретают методы сегментирования потребителей на основании их психографических характеристик. Для этого в анкету включается достаточно большое количество высказываний (порядка 100-150), характеризующих различные стороны жизненного стиля респондентов. Респонденты должны выразить свое согласие или несогласие с данными высказываниями по шкале Лайкерта (согласен — скорее согласен — ни то ни другое — скорее не согласен — не согласен). В дальнейшем на основании ответов респондентов формируются однородные целевые сегменты (обычно порядка 10).

Изучение продукта и бенчмаркинг продукта. В данном случае факторный анализ помогает выявить агрегатные параметры продукта, влияющие на выбор потребителя. Например, различные марки шоколадных конфет могут быть оценены по следующим макрокатегориям: качество (ингредиенты, вкус), полезность для здоровья (наличие сахара, калорийность) и цена.

Рекламные и медиа-исследования. Факторный анализ может использоваться для выявления скрытых мотивов поведения потребителей при восприятии рекламы.

Ценообразование. Факторный анализ используется для выявления особенностей поведения потребителей, чувствительных к цене. Например, данная категория респондентов может характеризоваться повышенным вниманием к ценовым факторам при выборе продукта, низкими доходами, большой численностью семьи и т. д.

Кластерный анализ является аналогом факторного анализа в том-смысле, что он так же, как и факторный анализ, позволяет выделить факторы (кластеры), объединяющие статистически схожие переменные. Однако в данном случае переменные классифицируются не на основании степени тесноты корреляционной связи, а на основании более сложных статистических процедур (наиболее часто используется метод исследования расстояний между переменными в кластерах). Ниже мы продемонстрируем действие обеих анализируемых статистических методик на одном массиве данных (см. выше).

Несмотря на имеющуюся возможность классифицировать переменные кластерный анализ чаще всего применяется для кластеризации групп респондентов (то есть уровней или категорий переменных)¹. Данная возможность позволяет, например,

провести пробное (при неизвестных целевых группах) сегментирование целевых покупателей какого-либо продукта. Сформированные в результате кластерного анализа целевые группы респондентов обладают схожим поведением (то есть взаимозависимостями) своих характеристик. В качестве примера успешной кластеризации можно привести разбиение респондентов на две группы:

- женщины в возрасте старше 45 лет;
- все мужчины и женщины младше 45 лет.

При использовании рассматриваемой статистической методики для кластеризации респондентов можно совмещать кластерный и факторный анализ, причем в данном случае факторный анализ будет предшествовать кластерному. Часто это делается для того, чтобы сократить количество переменных, участвующих в кластерном анализе (при большом числе этих переменных). Так, можно сначала выделить среди большого числа переменных макропараметры, а затем сегментировать респондентов уже на основании данных факторов.

Теперь у вас сложилось общее представление о методах факторного и кластерного анализа, и мы можем приступить к описанию их практического применения. Воспользуемся условием задачи про анализ текущей конкурентной позиции авиакомпании X. Эта задача поможет нам также сравнить действие данных статистических методик на од-

ной и той же выборке. Для описания кластерного анализа, применяемого для кластеризации респондентов, мы будем использовать другой пример из практики маркетинговых исследований. Так как для классификации переменных факторный анализ все же применяется чаще, чем кластерный (это сложилось исторически и, кроме того, оправдано меньшими усилиями, затрачиваемыми на проведение факторного анализа), в разделе 5.2.2 основное внимание будет уделено описанию действия кластерного анализа для классификации респондентов (выделения целевых групп потребителей). Сравнение действия двух статистических методов при классификации переменных мы предложим уже в заключении раздела 5.2.2.

В качестве примеров практического применения кластерного анализа в маркетинговых исследованиях можно указать все те же случаи, что и при факторном анализе (если кластерный анализ используется для классификации переменных). В случае применения кластерного анализа для классификации конкретных групп респондентов он предоставляет исследователю гораздо более гибкие возможности и в большем числе областей маркетинговых исследований по сравнению с факторным анализом. Это преимущество кластерного анализа обусловлено тем, что он анализирует не переменные в целом, а конкретные категории респондентов (например, различные половозрастные, доходные и другие группы покупателей). Таким образом, можно сделать важный вывод относительно факторного и кластерного анализов. Целью факторного анализа является сокращение числа переменных, участвующих в анализе (выделение релевантных макрокатегорий переменных), а целью кластерного — классификация респондентов на целевые группы на основании их существенных характеристик.

Из всего сказанного становится понятно, почему оба типа статистического анализа иногда используются в паре: факторный анализ определяет состав макропеременных (например, для сегментирования потребителей), а кластерный на основании выделенных существенных характеристик респондентов производит формирование целевых сегментов. Применение факторного и кластерного анализов в паре оправдано в основном в тех случаях, когда изначально респонденты оцениваются по большому числу параметров и проведение кластерного анализа непосредственно над данным (большим) набором переменных представляется затруднительным или даже практически невозможным. Отметим, что для проведения факторного и кластерного анализов в паре следует сначала провести факторный анализ, сохранив полученные факторные рейтинги, а затем проводить кластерный анализ на основании полученных групп переменных. Более подробно парное использование факторного и кластерного анализов будет показано в разделе 5.2.2.

В табл. 5.3 представлены основные характеристики переменных, участвующих в факторном и кластерном анализах.

Таблица 5.3. Основные характеристики переменных, участвующих в факторном и кластерном анализах

Факторный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Нет	-	Любое	Любой
Кластерный анализ			
Зависимые переменные		Независимые переменные	
Количество	Тип	Количество	Тип
Нет	-	Любое	Любой

5.2.1. Факторный анализ

Итак, из условия представленной выше задачи следует, что у нас есть массив данных, состоящий из 24 независимых переменных (утверждений), в различных аспектах описывающих текущее состояние авиакомпании X на международном рынке авиаперевозок. Основной задачей проводимого факторного анализа является группировка схожих по смыслу утверждений в макрокатегории с целью сократить число переменных и оптимизировать структуру данных.

При помощи меню Analyze ► Data Reduction ► Factor вызовите окно Factor Analysis. Перенесите из левого списка в правый переменные для анализа (q1-q24), как показано на рис. 5.32. Поле Selection Variable позволяет выбрать переменную, в разрезе которой будет проводиться анализ (например, класс полета). В нашем случае оставьте это поле Пустым.

Щелкните на кнопке Descriptives и в открывшемся диалоговом окне (рис. 5.33) выберите пункт КМО and Barlett's test of sphericity. Это позволит определить, насколько имеющиеся данные пригодны для факторного анализа. Окно Descriptives позволяет вывести и другие необходимые описательные статистики. Однако в большинстве примеров из маркетинговых исследований эти возможности, как правило, не используются.

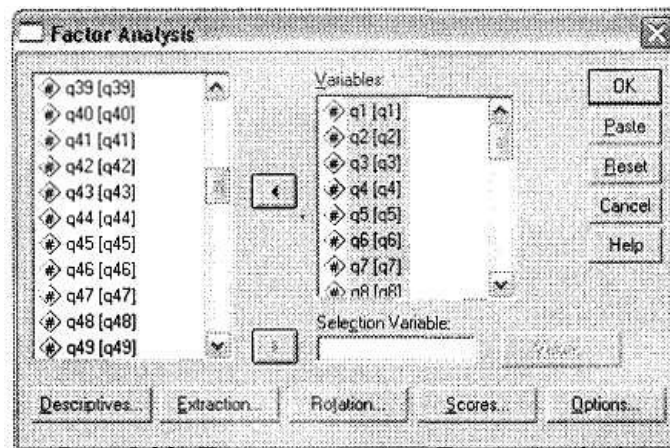


Рис. 5.32. Диалоговое окно Factor Analysis

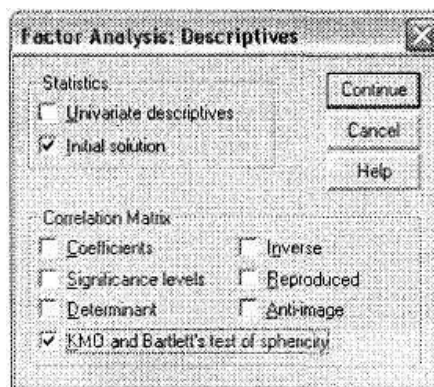


Рис. 5.33. Диалоговое окно Descriptives

Закройте окно Descriptives, щелкнув на кнопке Continue. Далее откройте окно Extraction (рис. 5.34), щелкнув на соответствующей кнопке в главном диалоговом окне Factor Analysis. Это окно предназначено для выбора метода формирования факторной модели; выполните в нем следующие действия.

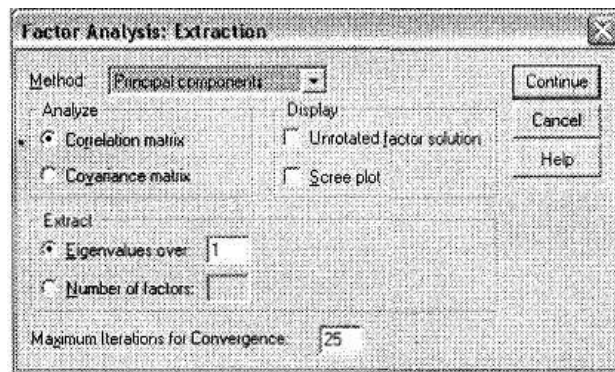


Рис. 5.34. Диалоговое окно Extraction

Во-первых, в поле Method выберите метод извлечения (формирования) факторов. Общая рекомендация по выбору метода состоит в следующем. Необходимо выбирать тот метод извлечения факторов, который позволяет однозначно классифицировать как можно больше переменных. Таким образом, основные соображения здесь — число классифицированных факторов и однозначность классификации (то есть каждая переменная должна принадлежать только одному фактору). Как вы увидите ниже, установленный по умолчанию в SPSS метод Principal components в нашем случае позволяет однозначно классифицировать 22 переменные из 24 имеющихся (92 %), что является весьма хорошим показателем. На основании имеющегося опыта автор может утверждать, что хорошим результатом факторного анализа является доля однозначно классифицированных переменных не менее 90 %. Выберите метод Principal components. Данный метод является наиболее подходящим для решения большинства задач маркетинговых исследований при помощи факторного анализа.

Во-вторых, укажите количество образуемых факторов (группа Extract). По умолчанию установлен метод определения количества извлекаемых факторов на основании значений характеристических чисел (Eigenvalues over). Не вдаваясь в статистические тонкости, отметим, что характеристические числа используются SPSS для определения количественного и качественного состава извлекаемых факторов. При предустановленном значении данного показателя, равном 1, количество образуемых факторов будет равно количеству переменных, значение характеристических чисел для которых больше или равно 1.

Также существует возможность вручную указать программе, сколько факторов необходимо извлекать (Number of factors). Эта возможность предусмотрена в SPSS для того, чтобы при слишком большом количестве переменных с характеристическим числом больше 1 вручную сократить число факторов. Большое число факторов трудно интерпретировать, поэтому если методом характеристических чисел не удастся извлечь приемлемое для интерпретации число факторов (чем меньше, тем лучше), следует самостоятельно указать программе число факторов. Эта задача решается аналитиком в каждом конкретном случае индивидуально. В качестве одного из вариантов решения можно рекомендовать увеличить число eigenvalue с предустановленного значения 1, скажем, до 1,5 или более. Это поможет, если получено большое число факторов с характеристическим числом, приблизительно равным 1, и несколько (2-3 и более) факторов — с характеристическим числом более 1,5 или другого значения. Также при ручном определении количества факторов аналитик может принять релевантное решение, основываясь на своем опыте или на каких-либо иных предположениях. И наконец, необходимо отметить, что при ручном указании числа извлекаемых факторов иногда количество однозначно классифицированных переменных оказывается меньше, чем при методе экстракции по величине характеристических чисел. Однако данный негативный момент нивелируется возросшей наглядностью результатов факторного анализа — ведь это позволяет освободиться от факторов, в которых нет переменных со значимым коэффициентом корреляции (в нашем случае 0,5).

Закройте диалоговое окно Extraction, щелкнув на кнопке Continue. Выберите тип ротации матрицы коэффициентов (кнопка Rotation в главном диалоговом окне Factor Analysis). Ротация коэффициентной матрицы производится для того, чтобы максимально приблизить факторную модель к идеалу: возможности однозначно классифицировать все переменные. В диалоговом окне Rotation (рис. 5.35) выберите конкретный метод ротации. В большинстве случаев наиболее приемлемым вариантом является метод Varimax. Он облегчает интерпретацию факторов, минимизируя количество переменных с высокими факторными нагрузками. Выберите этот тип ротации и закройте диалоговое окно, щелкнув на кнопке Continue.

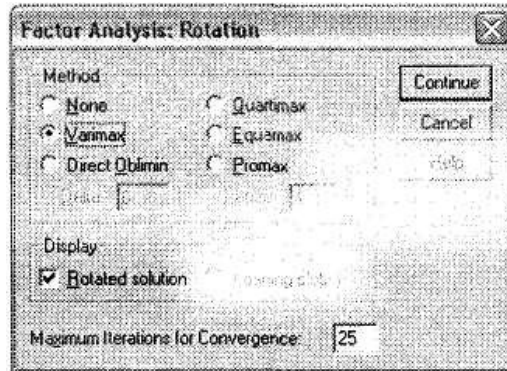


Рис. 5.35. Диалоговое окно Rotation

Далее откройте диалоговое окно Factor Scores (рис. 5.36), щелкнув на кнопке Scores. Это окно служит для создания в исходном файле данных новых переменных, которые в дальнейшем позволят отнести каждого респондента к определенной группе (фактору). Число вновь создаваемых переменных равно числу извлеченных факторов. Ниже мы покажем, каким образом использовать данные переменные. Выберите в диалоговом окне Factor Scores параметр Save as variables, а в качестве метода определения значений для этих новых переменных — регрессионную модель Regression. После этого закройте диалоговое окно, щелкнув на кнопке Continue.

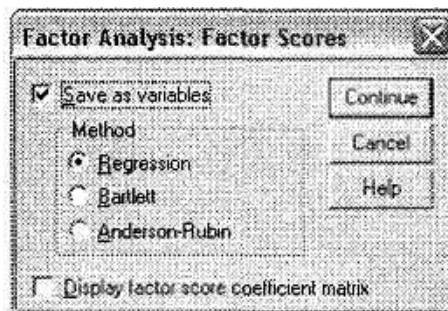


Рис. 5.36. Диалоговое окно Factor Scores

Последним этапом перед запуском процедуры факторного анализа является выбор некоторых дополнительных параметров (кнопка Options). В открывшемся диалоговом окне (рис. 5.37) выберите два пункта: Sorted by size и Suppress absolute values less than. Первая опция позволяет вывести переменные, входящие в каждый фактор, в порядке убывания их факторных коэффициентов (величины вклада переменной в формирование фактора). Вторая оказывается весьма полезна, так как облегчает задачу однозначной интерпретации полученных факторов. Указанное в соответствующем поле значение данного параметра (в нашем случае 0,5) отсекает переменные с факторными коэффициентами менее данного значения. Это позволяет упростить ротированную матрицу факторов, поскольку из нее исчезают незначимые переменные, входящие в каждый извлеченный фактор. Если

вы не задействуете данный параметр, для каждой переменной будет отображен факторный коэффициент по каждому фактору, что излишне перегрузит факторную модель и затруднит ее восприятие исследователями.

Параметр Suppress absolute values less than вводится, чтобы облегчить практическую интерпретацию результатов факторного анализа. Так как факторные коэффициенты в результирующей ротированной матрице коэффициентов являются коэффициентами корреляции между соответствующими переменными и факторами, в большинстве практических случаев целесообразно устанавливать начальное значение отсечения незначимых переменных на уровне 0,5. Если в результате факторного анализа окажется, что число классифицированных переменных менее приемлемого (например, если структура данных не вполне подходит для факторного анализа; см. ниже), можно пересчитать факторную модель с меньшим значением отсечения (например, 0,4). В обратной ситуации, если переменная входит в несколько факторов, можно предложить повысить уровень экстракции с 0,5 до 0,6. Это позволит устранить переменные, входящие сразу в несколько факторов, увеличив практическую пригодность результатов факторного анализа.

Итак, указав все необходимые параметры в окне Options, закройте его (кнопка Continue) и запустите процедуру факторного анализа при помощи щелчка на кнопке ОК в главном диалоговом окне Factor Analysis.

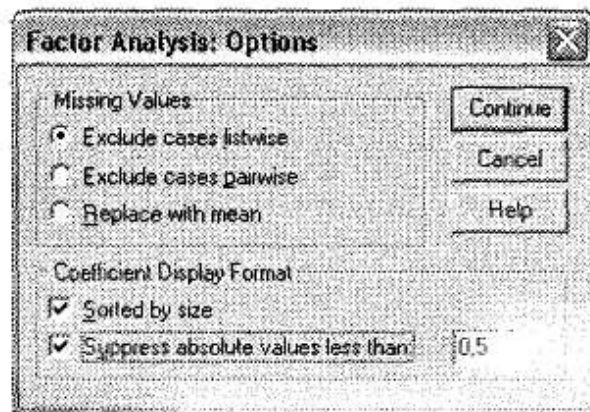


Рис. 5.37. Диалоговое окно Options

После того как программа произведет все необходимые расчеты, откроется окно SPSS Viewer с результатами построения факторной модели. Первое, что нас интересует, — это пригодность имеющихся данных для факторного анализа в целом. Посмотрим на таблицу KMO and Barlett's Test (рис. 5.38). В ней есть два интересующих нас показателя: тест КМО и значимость теста Barlett. Результаты теста КМО позволяют сделать вывод относительно общей пригодности имеющихся данных для факторного анализа, то есть насколько хорошо построенная факторная модель описывает структуру ответов респондентов на анализируемые вопросы. Результаты данного теста варьируются в интервале от 0 (факторная модель абсолютно неприменима) до 1 (факторная модель идеально описывает структуру данных). Факторный анализ следует считать пригодным, если КМО находится в пределах от 0,5 до 1. В нашем случае этот показатель равен 0,9, что является весьма хорошим результатом.

Barlett's test of sphericity проверяет гипотезу о том, что переменные, участвующие в факторном анализе, некоррелированы между собой. Если данный тест дает положительный результат (переменные некоррелированы), факторный анализ следует признать непригодным и использовать другие статистические методы (например, кластерный анализ). Статистикой, определяющей пригодность факторного анализа по тесту Barlett, является значимость (строка Sig.). При приемлемом уровне

значимости (ниже 0,05) факторный анализ считается пригодным для анализа исследуемой выборочной совокупности. В нашем случае рассматриваемый тест показывает

весьма низкую значимость (менее 0,001), из чего следует вывод о применимости факторного анализа.

Итак, на основании тестов КМО и Barlett мы пришли к выводу, что имеющиеся у нас данные практически идеально подходят для исследования при помощи факторного анализа.

KMO and Barlett's Test			
Kaiser-Meyer-Okin Measure of Sampling Adequacy.			.904
Barlett's Test of Sphericity	Approx. Chi-Square		10597,129
	Df		276
	Sig.		.000

Рис. 5.38. Таблица КМО and Barlett s Test

Следующим шагом в интерпретации результатов факторного анализа является рассмотрение результирующей ротированной матрицы факторных коэффициентов: таблицы Rotated Component Matrix (рис. 5.39). Данная таблица является основным результатом факторного анализа. В ней отражаются результаты классификации переменных по факторам. В нашем случае при помощи автоматического метода определения количества факторов (на основании характеристических чисел больше 1) была построена практически приемлемая факторная модель, в которой 22 из 24 переменных удалось однозначно классифицировать по небольшому числу факторов (5). Данный результат может считаться хорошим.

С неклассифицированными переменными можно поступить следующим образом. Необходимо просто пересчитать факторную модель, удалив в диалоговом окне Options ранее установленное значение отсечения 0,5. Далее будет построена факторная матрица (рис. 5.40), в которой аналитику предстоит самостоятельно определить принадлежность неклассифицированных переменных к тому или иному фактору на основании критерия наибольшего коэффициента корреляции между переменными и пятью факторами. В нашем случае вы видите, что переменная q16 в наибольшей степени коррелирует с фактором 1 (факторный коэффициент 0,468) и, следовательно, должна быть отнесена к данному фактору, а переменная q24 — с фактором 4 (0,474).

После того как мы однозначно классифицировали все переменные, вернемся к таблице на рис. 5.40. Мы получили пять групп переменных (факторов), описывающих текущую конкурентную позицию авиакомпании X с пяти различных сторон. Вот эти группы.

Фактор 1

q2. Авиакомпания X может конкурировать с лучшими авиакомпаниями мира. q3. Я верю, что у авиакомпании X есть перспективное будущее в мировой авиации. q23. Авиакомпания X — лучше, чем многие о ней думают. q14. Авиакомпания X — лицо России.

Rotated Component Matrix ^a					
	Component				
	1	2	3	4	5
q2	.727				
q3	.659				
q23	.643				
q14	.606				
q10	.585				
q1	.593				
q21	.592				
q5	.584				
q16					
q12		.744			
q11		.711			
q6		.694			
q8		.691			
q7		.635			
q4		.524			
q17			-.728		
q20			.597		
q18			.574		
q9				.698	
q22				.525	
q24					
q19					.737
q13					-.690
q15					.504

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 13 iterations.

Рис. 5.39. Таблица Rotated Component Matrix

q10. Авиакомпания X действительно заботится о пассажирах.

q1. Авиакомпания X обладает репутацией компаний, превосходно обслуживающей пассажиров.

q21. Авиакомпания X — эффективная авиакомпания. q5. Я горжусь тем, что работаю в авиакомпании X.

q16. Обслуживание авиакомпании X является последовательным и узнаваемым во всем мире.

Фактор 2

q12. Я верю, что менеджеры высшего звена прикладывают все усилия для достижения успеха авиакомпании.

q11. Среди сотрудников авиакомпании имеет место высокая степень удовлетворенности работой.

q6. Внутри авиакомпании X хорошее взаимодействие между подразделениями.

Rotated Component Matrix ^a					
	Component				
	1	2	3	4	5
q2	.727	.164	-.688E-02	-.150	1.029E-02
q3	.659	.259	.236	6.727E-02	-.127
q23	.643	.160	3.438E-02	-.239E-02	6.684E-03
q14	.606	.262	-.926E-02	.165	-.211
q10	.585	.407	-.102	4.193E-02	1.161E-02
q1	.593	.331	-.261	-.105	6.363E-02
q21	.592	.395	-.125E-02	-.270E-02	5.795E-02
q5	.584	.355	.109	4.398E-02	-.118
q16	.468	.219	-.330	.256	-.985E-02
q12	.195	.744	.143	-.553E-02	-.915E-02
q11	.206	.711	-.113	1.653E-02	-.535E-02
q6	.210	.694	-.102	-.338E-02	5.134E-02
q8	.312	.691	8.671E-02	-.493E-02	-.492E-02
q7	.199	.635	-.134	2.307E-02	7.782E-02
q4	.135	.524	3.831E-02	5.889E-03	-.708E-02
q17	.198	.126	-.728	1.102E-02	3.13E-02
q20	.152	9.380E-02	.597	.204	.300
q18	1.869E-02	2.201E-02	.574	.482	.187
q9	-.237	-.112	2.379E-02	.608	2.124E-02
q22	.177	3.011E-02	.355	.525	2.520E-02
q24	.446	-.243E-02	7.528E-02	.474	5.995E-02
q19	.133	.109	.281	.235	.737
q13	.228	.322	-.901E-02	.220	-.690
q15	-.418	9.389E-03	-.124	.400	.504

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 13 iterations.

Рис. 5.40. Таблица Rotated Component Matrix, содержащая все факторные коэффициенты

q8. Сейчас авиакомпания X быстро улучшается.

q7. Каждый сотрудник авиакомпании прикладывает все усилия для того, чтобы обеспечить ее успех.

q4. Я знаю, какой будет стратегия развития авиакомпании X в будущем.

Фактор 3

q17. Я бы не хотел, чтобы авиакомпания X менялась.

q20. Изменения в авиакомпании X будут позитивным моментом.

q18. Авиакомпании X необходимо меняться для того, чтобы использовать в полной мере имеющийся потенциал.

Фактор 4

q9. Нам предстоит долгий путь, прежде чем мы сможем претендовать на то, чтобы называться авиакомпанией мирового класса.

q22. Я бы хотел, чтобы имидж авиакомпании X улучшился с точки зрения иностранных пассажиров.

q24. Важно, чтобы люди во всем мире знали, что мы — российская авиакомпания.

Фактор 5

q19. Я думаю, что авиакомпания X необходимо представить себя в визуальном плане более современно.

q13. Мне нравится, как в настоящее время авиакомпания X представлена визуально широкой общественности (в плане цветовой гаммы и фирменного стиля).

q15. Мы выглядим «вчерашним днем» по сравнению с другими авиакомпаниями.

Наиболее сложной задачей при проведении факторного анализа является интерпретация полученных факторов. Здесь не существует какого-либо универсального решения: в каждом конкретном случае, аналитик использует имеющийся практический опыт для того, чтобы понять, почему факторная модель относит ту или иную переменную к данному конкретному фактору. Бывают случаи (особенно при малом числе хорошо формализованных переменных), когда образованные факторы являются очевидными и различия между переменными видны невооруженным глазом. В такой ситуации можно обойтись без факторного анализа и разбить переменные на группы вручную. Однако эффективность и мощь факторного анализа проявляются в сложных и нетривиальных случаях, когда переменные нельзя заранее классифицировать, а их формулировки запутаны. Тогда большой исследовательский интерес будет вызывать классификация переменных именно на основании мнений респондентов, что позволит выявить то, как сами опрошенные поняли тот или иной вопрос.

Приводим рекомендации, которые помогут вам при затруднении интерпретировать результаты факторного анализа.

Когда это возможно и приемлемо для целей исследования, следует формализовать переменные до проведения факторного анализа. Это позволит аналитику заранее сделать предположения о разделении совокупности имеющихся переменных на группы. Задача исследователя при интерпретации результатов факторной матрицы в данном случае упростится, так как он уже не будет начинать «с чистого листа». Его задача сведется к проверке ранее выдвинутых гипотез о принадлежности той или иной переменной к конкретной группе.

Иногда возникают случаи, когда переменная, отнесенная SPSS к конкретному фактору, логически никак не связана с остальными переменными, составляющими тот же фактор. Можно пересчитать факторную модель без отсека незначимых коэффициентов (как в примере на рис. 5.40) и посмотреть, с каким еще фактором данная нелогичная переменная коррелирует практически с той же силой, как с фактором, к которому она была отнесена автоматически. Например, переменная Z имеет коэффициент корреляции с фактором 1, равный 0,505, а с фактором 2 она коррелирует с коэффициентом 0,491. SPSS автоматически относит данную переменную к тому фактору, с которым выявлена наибольшая корреляция, не учитывая при этом, что с другим фактором данная переменная коррелирует практически с той же силой. Именно в такой ситуации (при небольшой разнице в коэффициентах корреляции) можно попробовать отнести переменную Z к фактору 2, и если это окажется логичным, рассматривать ее в группе переменных из второго фактора.

Можно вручную сократить число извлекаемых факторов, что облегчит задачу исследователя при интерпретации результатов факторного анализа. Однако необходимо иметь в виду, что такое сокращение снизит гибкость факторной модели и даже может привести к ситуации, когда переменные будут ложно разделены на неверные, с практической точки зрения, группы. Также снижение числа извлекаемых факторов неизбежно снизит и долю однозначно классифицированных факторов.

В качестве варианта предыдущего решения можно предложить объединить два или более факторов с небольшими количествами входящих в них переменных. Такая группировка, с одной стороны, позволит снизить число интерпретируемых факторов, а с другой — облегчит понимание малочисленных факторов.

Если исследователь зашел в тупик и никакие средства не помогают объяснить

принадлежность той или иной переменной к конкретному фактору, остается применить другую статистическую процедуру (например, кластерный анализ).

Вернемся к нашим пяти факторам. Задача их описания и объяснения представляется не очень сложной. Так, можно заметить, что утверждения, входящие в первый фактор (q2, q3, q23, q14, q10, q1, q21, q5 и q16), являются общими, то есть касаются всей авиакомпании и описывают отношение к ней со стороны авиапассажиров. Единственное исключение составила переменная q5, имеющая отношение скорее ко второму фактору. Коэффициент корреляции с фактором 2 — 0,355 (см. рис. 5.40), что позволяет отнести его в данную группу из соображений логики. Фактор 2 (q12, q11, q6, q8, q7 и q4) описывает отношение к авиакомпании X со стороны сотрудников. Третий фактор (q17, q20 и q18) описывает отношение респондентов к изменениям в авиакомпании (в него попали все утверждения, имеющие корень «мен» — от слова «изменение»). Четвертый фактор (q9, q22 и q24) описывает отношение респондентов к имиджу авиакомпании. Наконец, пятый фактор (q19, q13 и q15) объединяет утверждения, характеризующие отношение респондентов к визуальному образу авиакомпании X.

Таким образом, мы получили пять групп утверждений, описывающих текущую конкурентную позицию компании X на международном рынке авиаперевозок. На основании проведенного интерпретационного (семантического) анализа можно присвоить данным группам (факторам) следующие определения.

- Фактор 1 характеризует общее положение авиакомпании X в глазах ее клиентов.
- Фактор 2 характеризует внутреннее состояние авиакомпании X с точки зрения ее сотрудников.
- Фактор 3 характеризует изменения, происходящие в авиакомпании X.
- Фактор 4 характеризует имидж авиакомпании X.
- Фактор 5 характеризует визуальный образ авиакомпании X.

После того как мы успешно интерпретировали все полученные факторы, можно считать факторный анализ завершенным и удавшимся. Далее мы покажем, как можно использовать результаты факторного анализа для построения разрезов.

Вспомним о том, что мы сохранили факторные рейтинги (то есть принадлежность каждого респондента к определенному фактору) в исходном файле данных в виде новых переменных. Эти переменные имеют имена типа: facX_Y, где X — это номер фактора, а Y — порядковый номер факторной модели. Если мы строили факторную модель дважды и в результате в первый раз было извлечено три фактора, а во второй — два, имена переменных будут следующими:

- fac1_1, fac2_1, fac3_1 (для трех факторов из первой построенной модели);
- fac1_2, fac2_2 (для двух факторов из второй модели).

В нашем случае будет создано пять новых переменных (по числу извлеченных факторов). Эти факторные рейтинги в дальнейшем могут использоваться, например, для построения разрезов. Так, если необходимо выяснить, каким образом респонденты — мужчины и женщины — оценивают различные стороны деятельности авиакомпании X, это можно сделать при помощи анализа факторных рейтингов.

Наиболее частый способ использования факторных рейтингов в дальнейших расчетах — это ранжирование и последующее разделение вновь созданных переменных, обозначающих извлеченные факторы, на четыре квартиля (25%-проценти-ля). Такой подход позволяет создать новые переменные с порядковой шкалой, описывающие четыре уровня каждого фактора. В нашем случае для утверждений, составляющих фактор 2, такими уровнями будут: не согласен (состояние внутренних дел компании не удовлетворяет сотрудников), скорее не согласен (оценка внутренней ситуации в компании ниже среднего), скорее согласен (оценка выше среднего), согласен (оценка отлично).

Чтобы создать переменные, по которым далее будут группироваться респонденты, вызовите меню Transform ► Rank Cases. В открывшемся диалоговом окне (рис. 5.41)

из левого списка выберите переменную, содержащую факторные рейтинги для фактора 2 (fac2_1), и поместите ее в поле Variables. Далее в области Assign Rank I to выберите пункт Smallest value, в нашем случае это означает, что первую группу (не согласен) составят респонденты, оценивающие состояние внутренних дел авиакомпании как плохое. Соответственно группы 2, 3 и 4 будут определены для категорий скорее не согласен, скорее согласен и согласен соответственно.

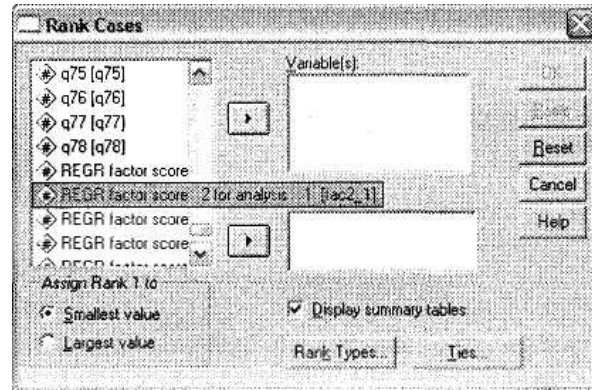


Рис. 5.41. Диалоговое окно Rank Cases

Щелкните на Rank Types ► Types, отмените установленный по умолчанию параметр Rank и вместо него выберите Ntiles с предустановленным числом групп, равным 4 (рис. 5.42). Щелкните на кнопке Continue и затем в главном диалоговом окне на ОК. Данная процедура создаст в файле данных новую переменную nfac2_1 (2 означает второй фактор), распределяющую респондентов на четыре группы.

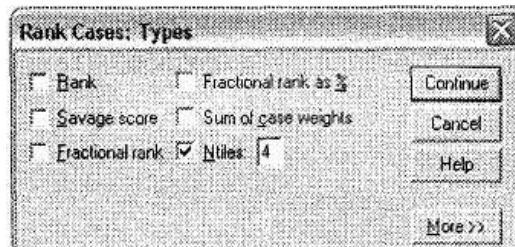


Рис. 5.42. Диалоговое окно Types

Все респонденты в выборке характеризуются положительным, скорее положительным, скорее отрицательным или отрицательным отношением к текущему состоянию дел в авиакомпании X. Для повышения наглядности рекомендуется присвоить метки каждому из выделенных четырех уровней; можно переименовать и саму переменную. Теперь вы можете проводить перекрестный анализ при помощи новой порядковой переменной, а также строить другие статистические модели, предусмотренные в SPSS. Ниже будет показано, как использовать результаты построения факторной модели в кластерном анализе.

Для иллюстрации возможностей практического использования новой переменной проведем перекрестный анализ влияния пола респондентов на их оценку текущего состояния дел в авиакомпании X (рис. 5.43). Как следует из представленной таблицы, респонденты-мужчины в целом склонны ставить более низкие оценки рассматриваемому параметру авиакомпании по сравнению с женщинами. Так, в структуре оценок очень плохо, плохо и удовлетворительно доля мужчин преобладает; в оценках очень хорошо, напротив, преобладают женщины. При переходе в каждую следующую (более высокую) категорию оценок доля мужчин равномерно убывает, а доля женщин, соответственно, возрастает. Тест χ^2 показывает, что выявленная зависимость является статистически значимой.

Пол респондента * Оценка состояния дел внутри компании Crosstabulation

			Оценка состояния дел внутри компании				Total
			Очень плохо	Плохо	Удовлетворительно	Очень хорошо	
Пол респондента Мужчины	Count		153	151	140	98	542
	% within Пол респондента		28,2%	27,9%	25,0%	18,1%	100,0%
Женщины	Count		156	163	172	217	708
	% within Пол респондента		22,0%	23,0%	24,3%	30,6%	100,0%
Total	Count		309	314	312	315	1250
	% within Пол респондента		24,7%	25,1%	25,0%	25,2%	100,0%

Рис. 5.43. Перекрестное распределение: влияние пола респондентов на их оценку текущего состояния дел в авиакомпании X

5.2.2. Иерархический кластерный анализ

В статистике существует два основных типа кластерного анализа (оба представлены в SPSS): иерархический и осуществляемый методом k-средних. В первом случае автоматизированная статистическая процедура самостоятельно определяет оптимальное число кластеров и ряд других параметров, необходимых для кластерного

анализа. Второй тип анализа имеет существенные ограничения по практической применимости — для него необходимо самостоятельно определять и точное количество выделяемых кластеров, и начальные значения центров каждого кластера (центроиды), и некоторые другие статистики. При анализе методом k-средних данные problems решаются предварительным проведением иерархического кластерного анализа и затем на основании его результатов расчетом кластерной модели по методу k-средних, что в большинстве случаев не только не упрощает, а наоборот, усложняет работу исследователя (в особенности неподготовленного).

В целом можно сказать, что в связи с тем, что иерархический кластерный анализ весьма требователен к аппаратным ресурсам компьютера, кластерный анализ по методу k-средних введен в SPSS для обработки очень больших массивов данных, состоящих из многих тысяч наблюдений (респондентов), в условиях недостаточной мощности компьютерного оборудования¹. Размеры выборок, используемых в маркетинговых исследованиях, в большинстве случаев не превышают четыре тысячи респондентов. Практика маркетинговых исследований показывает, что именно первый тип кластерного анализа — иерархический — рекомендуется для использования во всех случаях как наиболее релевантный, универсальный и точный. Вместе с тем необходимо подчеркнуть, что при проведении кластерного анализа важным является отбор релевантных переменных. Данное замечание очень существенно, так как включение в анализ нескольких или даже одной нерелевантной переменной способно привести к неудаче всей статистической процедуры.

Описание методики проведения кластерного анализа мы проведем на следующем примере из практики маркетинговых исследований.

Исходные данные:

В ходе исследования было опрошено 745 авиапассажиров, летавших одной из 22 российских и зарубежных авиакомпаний. Авиапассажиров просили оценить по пятибалльной шкале — от 1 (очень плохо) до 5 (отлично) — семь параметров работы наземного персонала авиакомпаний в процессе регистрации пассажиров на рейс: вежливость, профессионализм, оперативность, готовность помочь, регулирование очереди, внешний вид, работа персонала в целом.

Требуется:

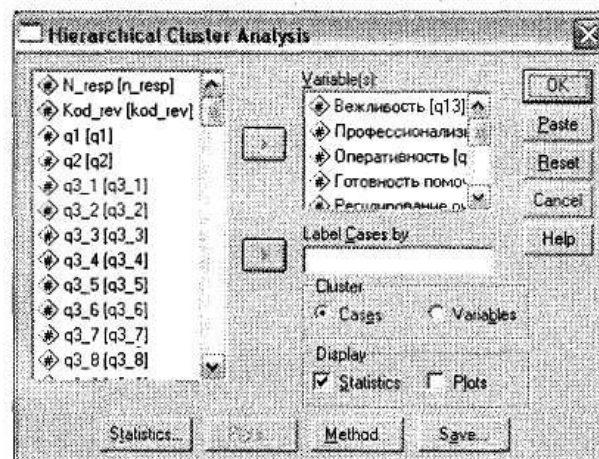
Сегментировать исследуемые авиакомпании по уровню воспринимаемого авиапассажирами качества работы наземного персонала.

Итак, у нас есть файл данных, который состоит из семи интервальных переменных, обозначающих оценки качества работы наземного персонала различных авиакомпаний (q13-q19), представленные в единой пятибалльной шкале. Файл данных содержит одновариантную переменную q4, указывающую выбранные респондентами авиакомпании (всего 22 наименования). Проведем кластерный анализ и определим, на какие целевые группы можно разделить данные авиакомпании.

Иерархический кластерный анализ проводится в два этапа. Результат первого этапа — число кластеров (целевых сегментов), на которые следует разделить исследуемую выборку респондентов. Процедура кластерного анализа как таковая не

может самостоятельно определить оптимальное число кластеров. Она может только подсказать искомое число. Поскольку задача определения оптимального числа сегментов является ключевой, она обычно решается на отдельном этапе анализа. На втором этапе производится собственно кластеризация наблюдений по тому числу кластеров, которое было определено в ходе первого этапа анализа. Теперь рассмотрим эти шаги кластерного анализа по порядку.

Процедура кластерного анализа запускается при помощи меню Analyze ► Classify ► Hierarchical Cluster. В открывшемся диалоговом окне из левого списка всех имеющихся в файле данных переменных выберите переменные, являющиеся критериями сегментирования. В нашем случае их семь, и обозначают они оценки параметров работы наземного персонала q13-q19 (рис. 5.44). В принципе указания совокупности критериев сегментирования будет вполне достаточно для выполнения первого этапа кластерного анали-



за.

По умолчанию кроме таблицы с результатами формирования кластеров, на основании которой мы определим их оптимальное число, SPSS выводит также специальную перевернутую гистограмму icicle, помогающую, по замыслу создателей программы, определить оптимальное количество кластеров; вывод диаграмм осуществляется кнопкой Plots (рис. 5.45). Однако если оставить данный параметр установленным, мы потратим много времени на обработку даже сравнительно небольшого файла данных. Кроме icicle в окне Plots можно выбрать более быструю линейчатую диаграмму Dendrogram. Она представляет собой горизонтальные столбики, отражающие процесс формирования кластеров. Теоретически при небольшом (до 50-100) количестве респондентов данная диаграмма действительно помогает выбрать оптимальное решение относительно требуемого числа кластеров. Однако практически во всех примерах из маркетинговых исследований размер выборки превышает это количество. Дендограмма относительно бесполезна, так как даже при относительно небольшом числе наблюдений представляет собой очень длинную

последовательность номеров строк исходного файла данных, соединенных между собой горизонтальными и вертикальными линиями. Большинство учебников по SPSS содержат примеры кластерного анализа именно на таких искусственных, малых выборках. В настоящем пособии мы показываем, как наиболее эффективно работать с SPSS в практических условиях и на примере реальных маркетинговых исследований.

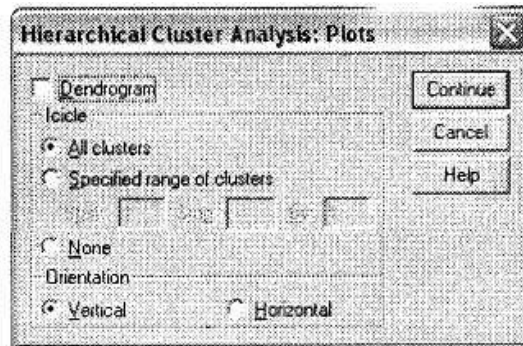


Рис. 5.45. Диалоговое окно Plots

Как мы установили, для практических целей ни Icicle, ни Dendrogram не пригодны. Поэтому в главном диалоговом окне Hierarchical Cluster Analysis рекомендуется не выводить диаграммы, отменив выбранный по умолчанию параметр Plots в области Display, как показано на рис. 5.44. Теперь все готово для выполнения первого этапа кластерного анализа. Запустите процедуру, щелкнув на кнопке ОК.

Через некоторое время в окне SPSS Viewer появятся результаты. Как было сказано выше, единственным значимым для нас итогом первого этапа анализа будет таблица Average Linkage (Between Groups), представленная на рис. 5.46. На основании этой таблицы мы должны определить оптимальное число кластеров. Необходимо заметить, что единого универсального метода определения оптимального числа кластеров не существует. В каждом конкретном случае исследователь должен сам определить это число.

Исходя из имеющегося опыта, автор предлагает следующую схему данного процесса. Прежде всего, попробуем применить наиболее распространенный стандартный метод для определения числа кластеров. По таблице Average Linkage (Between Groups) следует определить, на каком шаге процесса формирования кластеров (колонка Stage) происходит первый сравнительно большой скачок коэффициента агломерации (колонка Coefficients). Данный скачок означает, что до него в кластеры объединялись наблюдения, находящиеся на достаточно малых расстояниях друг от друга (в нашем случае респонденты со схожим уровнем оценок по анализируемым параметрам), а начиная с этого этапа происходит объединение более далеких наблюдений.

В нашем случае коэффициенты плавно возрастают от 0 до 7,452, то есть разница между коэффициентами на шагах с первого по 728 была мала (например, между 728 и 727 шагами — 0,534). Начиная с 729 шага происходит первый существенный скачок коэффициента: с 7,452 до 10,364 (на 2,912). Шаг, на котором происходит первый скачок коэффициента, — 729. Теперь, чтобы определить оптимальное ко-

личество кластеров, необходимо вычесть полученное значение из общего числа наблюдений (размера выборки). Общий размер выборки в нашем случае составляет 745 человек; следовательно, оптимальное количество кластеров составляет $745 - 729 = 16$.

Average Linkage (Between Groups)

Agglomeration Schedule

Step	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	743	745	,000	0	0	2
2	1	743	,000	0	1	7
3	737	742	,000	0	0	7
4	740	741	,000	0	0	5
5	782	740	,000	0	4	10
6	734	738	,000	0	0	10
7	1	737	,000	2	3	9
8	735	736	,000	0	0	9
9	1	735	,000	7	8	13

726	12	146	6,004	721	710	728
727	1	5	6,015	606	725	734
728	12	222	7,452	726	715	729
729	12	160	10,364	728	0	730
730	12	165	13,412	729	0	732
731	125	166	14,000	677	0	732
732	12	128	15,714	750	731	735
733	12	307	17,664	732	609	734
734	1	12	21,011	727	702	0

Рис. 5.46. Таблица Average Linkage (Between Groups)

Мы получили достаточно большое число кластеров, которое в дальнейшем будет сложно интерпретировать. Поэтому теперь следует исследовать полученные кластеры и определить, какие из них являются значимыми, а какие нужно попытаться сократить. Данная задача решается на втором этапе кластерного анализа.

Откройте главное диалоговое окно процедуры кластерного анализа (меню Analyze ► Classify ► Hierarchical Cluster). В поле для анализируемых переменных у нас уже есть семь параметров. Щелкните на кнопке Save. Открывшееся диалоговое окно (рис. 5.47) позволяет создать в исходном файле данных новую переменную, распределяющую респондентов на целевые группы. Выберите параметр Single Solution и укажите в соответствующем поле необходимое количество кластеров — 16 (определено на первом этапе кластерного анализа). Щелкнув на кнопке Continue, вернитесь в главное диалоговое окно, в котором щелкните на кнопке ОК, чтобы запустить процедуру кластерного анализа.

Прежде чем продолжить описание процесса кластерного анализа, необходимо привести краткое описание других параметров. Среди них есть как полезные возможности, так и фактически лишние (с точки зрения практических маркетинговых исследований). Так, например, главное диалоговое окно Hierarchical Cluster Analysis содержит поле Label Cases by, в которое при желании можно поместить текстовую переменную, идентифицирующую респондентов. В нашем случае для этих целей может служить переменная q4, кодирующая выбранные респондентами авиакомпания. На практике сложно придумать рациональное объяснение использованию поля Label Cases by, поэтому можно спокойно всегда оставлять его пустым.

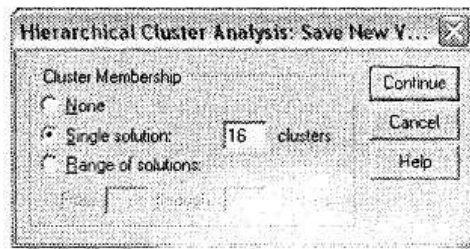


Рис. 5.47. Диалоговое окно создания новой переменной

Нечасто при проведении кластерного анализа используется диалоговое окно Statistics, вызываемое одноименной кнопкой в главном диалоговом окне. Оно позволяет организовать вывод в окне SPSS Viewer таблицы Cluster Membership, в которой каждому респонденту в исходном файле данных сопоставляется номер кластера. Данная таблица при достаточно большом количестве респондентов (практически во всех примерах маркетинговых исследований) становится совершенно бесполезной, так как представляет собой длинную последовательность пар значений «номер респондента/номер кластера», в таком виде не поддающуюся интерпретации. Технически цель кластерного анализа всегда состоит в образовании в файле данных дополнительной переменной, отражающей разделение респондентов на целевые группы (при помощи щелчка на кнопке Save в главном диалоговом окне кластерного анализа). Эта переменная в совокупности с номерами респондентов и есть таблица Cluster Membership. Единственный практически полезный параметр в окне Statistics — вывод таблицы Average Linkage (Between Groups), однако он уже установлен по умолчанию. Таким образом, использование кнопки Statistics и вывод отдельной таблицы Cluster Membership в окне SPSS Viewer является нецелесообразным.

Про кнопку Plots уже было сказано выше: ее следует деактивизировать, отменив параметр Plots в главном диалоговом окне кластерного анализа.

Кроме этих редко используемых возможностей процедуры кластерного анализа, SPSS предлагает и весьма полезные параметры. Среди них прежде всего кнопка Save, позволяющая создать в исходном файле данных новую переменную, распределяющую респондентов по кластерам. Также в главном диалоговом окне существует область для выбора объекта кластеризации: респондентов или переменных. Об этой возможности говорилось выше в разделе 5.4. В первом случае кластерный анализ используется в основном для сегментирования респондентов по некоторым критериям; во втором цель проведения кластерного анализа аналогична факторному анализу: классификация (сокращение числа) переменных.

Как видно из рис. 5.44, единственной не рассмотренной возможностью кластерного анализа является кнопка выбора метода проведения статистической процедуры Method. Эксперименты с данным Параметром позволяют добиться большей точности при определении оптимального числа кластеров. Общий вид этого диалогового окна с параметрами, установленными по умолчанию, представлен на рис. 5.48.

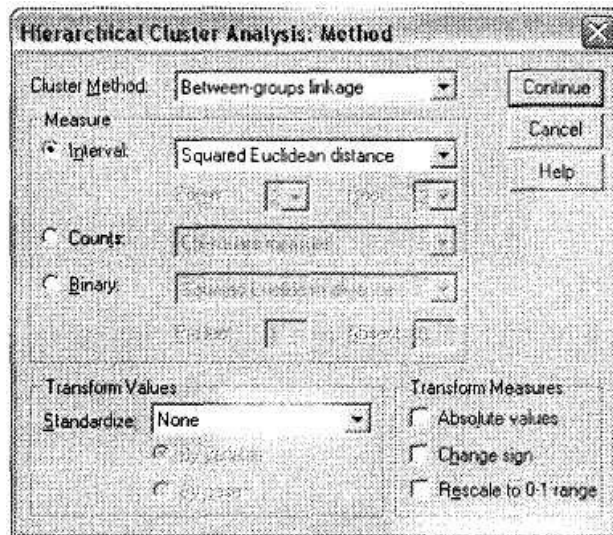


Рис. 5.48. Диалоговое окно Method

Первое, что устанавливается в данном окне, — это метод формирования кластеров (то есть объединения наблюдений). Среди всех возможных вариантов статистических методик, предлагаемых SPSS, следует выбирать либо установленный по умолчанию метод *Between-groups linkage*, либо процедуру *Ward* (*Ward's method*). Первый метод используется чаще ввиду его универсальности и относительной простоты статистической процедуры, на которой он основан. При использовании этого метода расстояние между кластерами вычисляется как среднее значение расстояний между всеми возможными парами наблюдений, причем в каждой итерации принимает участие одно наблюдение из одного кластера, а второе — из другого. Информация, необходимая для расчетов расстояния между наблюдениями, находится на основании всех теоретически возможных пар наблюдений. Метод *Ward* более сложен для понимания и используется реже. Он состоит из множества этапов и основан на усреднении значений всех переменных для каждого наблюдения и последующем суммировании квадратов расстояний от вычисленных средних до каждого наблюдения. Для решения практических задач маркетинговых исследований мы рекомендуем всегда использовать метод *Between-groups linkage*, установленный по умолчанию.

После выбора статистической процедуры кластеризации следует выбрать метод для вычисления расстояний между наблюдениями (область *Measure* в диалоговом окне *Method*). Существуют различные методы определения расстояний для трех типов переменных, участвующих в кластерном анализе (критериев сегментирования). Эти переменные могут иметь интервальную (*Interval*), номинальную (*Counts*) или дихотомическую (*Binary*) шкалу. Дихотомическая шкала (*Binary*) подразумевает только переменные, отражающие наступление/ненаступление какого-либо события (купил/не купил, да/нет и т. д.). Другие типы дихотомических переменных (например, мужчина/женщина) следует рассматривать и анализировать как номинальные (*Counts*).

Наиболее часто используемым методом определения расстояний для интервальных переменных является квадрат евклидова расстояния (*Squared Euclidean Distance*), устанавливаемый по умолчанию. Именно этот метод зарекомендовал себя в маркетинговых исследованиях как наиболее точный и универсальный. Однако для дихотомических переменных, где наблюдения представлены только двумя значениями (например, 0 и 1), данный метод не подходит. Дело в том, что он учитывает только взаимодействия между наблюдениями типа: $X = 1, Y = 0$ и $X = 0, Y = 1$ (где X и Y — переменные) и не учитывает другие типы взаимодействий. Наиболее комплексной мерой расстояния, учитывающей все важные типы взаимодействий между двумя дихотомическими переменными, является метод Лямбда (*Lambda*). Мы рекомендуем применять именно данный метод ввиду его универсальности. Однако существуют и другие методы, например *Shape*, *Hamann* или *An-*

derbergs's D.

При указании метода определения расстояний для дихотомических переменных в соответствующем поле необходимо указать конкретные значения, которые могут принимать исследуемые дихотомические переменные: в поле Present — кодировку ответа Да, а в поле Absent — Нет. Названия полей присутствуют и отсутствуют ассоциированы с тем, что в группе методов Binary предполагается использовать только дихотомические переменные, отражающие наступление/ненаступление какого-либо события. Для двух типов переменных Interval и Binary существует несколько методов определения расстояния. Для переменных с номинальным типом шкалы SPSS предлагает всего два метода: χ^2 (Chi-square measure) и ϕ^2 (Phi-square measure). Мы рекомендуем использовать первый метод как наиболее распространенный.

В диалоговом окне Method есть область Transform Values, в которой находится поле Standardize. Данное поле применяется в том случае, когда в кластерном анализе принимают участие переменные с различным типом шкалы (например, интервальные и номинальные). Для того чтобы использовать эти переменные в кластерном анализе, следует провести стандартизацию, приводящую их к единому типу шкалы — интервальному. Самым распространенным методом стандартизации переменных является 2-стандартизация (Zscores): все переменные приводятся к единому диапазону значений от -3 до +3 и после преобразования являются интервальными.

Так как все оптимальные методы (кластеризации и определения расстояний) установлены по умолчанию, целесообразно использовать диалоговое окно Method только для указания типа анализируемых переменных, а также для указания необходимости произвести 2-стандартизацию переменных.

Итак, мы описали все основные возможности, предоставляемые SPSS для проведения кластерного анализа. Вернемся к описанию кластерного анализа, проводимого с целью сегментирования авиакомпаний. Напомним, что мы остановились на шестнадцатикластерном решении и создали в исходном файле данных новую переменную clul6_1, распределяющую все анализируемые авиакомпании по кластерам.

Чтобы установить, насколько верно мы определили оптимальное число кластеров, построим линейное распределение переменной clul6_1 (меню Analyze ► Descriptive Statistics ► Frequencies). Как видно на рис. 5.49, в кластерах с номерами 5-16 число респондентов составляет от 1 до 7. Наряду с вышеописанным универсальным методом определения оптимального количества кластеров (на основании разности между общим числом респондентов и первым скачком коэффициента агломерации) существует также дополнительная рекомендация: размер кластеров должен быть статистически значимым и практически приемлемым. При нашем размере выборки такое критическое значение можно установить хотя бы на уровне 10. Мы видим, что под данное условие попадают лишь кластеры с номерами 1-4. Поэтому

теперь необходимо пересчитать процедуру кластерного анализа с выводом четырехкластерного решения (будет создана новая переменная du4_1).

Average Linkage (Between Groups)					
		Frequency	Percent	Valid Percent	Cummulative Percent
Valid	1	454	60,9	61,8	61,8
	2	195	26,2	26,5	88,3
	3	22	3,0	3,0	91,3
	4	32	4,3	4,4	95,6
	5	7	,9	1,0	96,6
	6	2	,3	,3	96,9
	7	4	,5	,5	97,4
	8	1	,1	,1	97,6
	9	1	,1	,1	97,7
	10	1	,1	,1	97,8
	11	5	,7	,7	98,5
	12	1	,1	,1	98,6
	13	2	,3	,3	98,9
	14	2	,3	,3	99,2
	15	2	,3	,3	99,5
	16	4	,5	,5	100,0
	Total	735	98,7	100,0	
Missing System		10	1,3		
Total		745	100,0		

Рис. 5.49. Линейное распределение для 16-кластерного решения

Построив линейное распределение по вновь созданной переменной du4_1, мы увидим, что только в двух кластерах (1 и 2) число респондентов является практически значимым. Нам необходимо снова перестроить кластерную модель — теперь для двухкластерного решения. После этого построим распределение по переменной du2_1 (рис. 5.50). Как вы видите из таблицы, двухкластерное решение имеет статистически и практически значимое число респондентов в каждом из двух сформированных кластеров: в кластере 1 — 695 респондентов; в кластере 2 — 40. Итак, мы определили оптимальное число кластеров для нашей задачи и провели собственно сегментирование респондентов по семи избранным критериям. Теперь можно считать основную цель нашей задачи достигнутой и приступать к завершающему этапу кластерного анализа — интерпретации полученных целевых групп (сегментов).

Average Linkage (Between Groups)					
		Frequency	Percent	Valid Percent	Cummulative Percent
Valid	1	695	93,3	94,6	94,6
	2	40	5,4	5,4	100,0
	Total	735	98,7	100,0	
System Missing		10	1,3		
Total		745	100,0		

Рис. 5.50. Численность кластеров (решение для 2 кластеров)

Полученное решение несколько отличается от тех, которые вы, может быть, видели в учебных пособиях по SPSS. Даже в наиболее практически ориентированных учебниках приведены искусственные примеры, где в результате кластеризации получаются иде-

альные целевые группы респондентов. В некоторых случаях (5) авторы даже прямо указывают на искусственное происхождение примеров. В настоящем пособии мы применим в качестве иллюстрации действия кластерного анализа реальный пример из практического маркетингового исследования, не отличающийся идеальными пропорциями. Это позволит нам показать наиболее распространенные трудности проведения кластерного анализа, а также оптимальные методы их устранения.

Перед тем как приступить к интерпретации полученных кластеров, давайте подведем итоги. У нас получилась следующая схема определения оптимального числа кластеров.

- На этапе 1 мы определяем количество кластеров на основании математического метода, основанного на коэффициенте агломерации.

- На этапе 2 мы проводим кластеризацию респондентов по полученному числу кластеров и затем строим линейное распределение по образованной новой переменной (clul6_1). Здесь также следует определить, сколько кластеров состоят из статистически значимого количества респондентов. В общем случае рекомендуется устанавливать минимально значимую численность кластеров на уровне не менее 10 респондентов.

- Если все кластеры удовлетворяют данному критерию, переходим к завершающему этапу кластерного анализа: интерпретации кластеров. Если есть кластеры с незначимым числом составляющих их наблюдений, устанавливаем, сколько кластеров состоят из значимого количества респондентов.

- Пересчитываем процедуру кластерного анализа, указав в диалоговом окне Save число кластеров, состоящих из значимого количества наблюдений.

- Строим линейное распределение по новой переменной.

Такая последовательность действий повторяется до тех пор, пока не будет найдено решение, в котором все кластеры будут состоять из статистически значимого числа респондентов. После этого можно переходить к завершающему этапу кластерного анализа — интерпретации кластеров.

Необходимо особо отметить, что критерий практической и статистической значимости численности кластеров не является единственным критерием, по которому можно определить оптимальное число кластеров. Исследователь может самостоятельно, на основании имеющегося у него опыта предложить число кластеров (условие значимости должно удовлетворяться). Другим вариантом является довольно распространенная ситуация, когда в целях исследования заранее ставится условие сегментировать респондентов по заданному числу целевых групп. В этом случае необходимо просто один раз провести иерархический кластерный анализ с сохранением требуемого числа кластеров и затем пытаться интерпретировать то, что получится.

Для того чтобы описать полученные целевые сегменты, следует воспользоваться процедурой сравнения средних значений исследуемых переменных (кластерных центроидов). Мы сравним средние значения семи рассматриваемых критериев сегментирования в каждом из двух полученных кластеров.

Процедура сравнения средних значений вызывается при помощи меню Analyze ► Compare Means ► Means. В открывшемся диалоговом окне (рис. 5.51) из левого списка выберите семь переменных, избранных в качестве критериев сегментирования (ql3-ql9), и перенесите их в поле для зависимых переменных Dependent List. Затем переменную cШ2_1, отражающую разделение респондентов на кластеры при окончательном (двухкластерном) решении задачи, переместите из левого списка в поле для независимых переменных Independent List. После этого щелкните на кнопке Options.

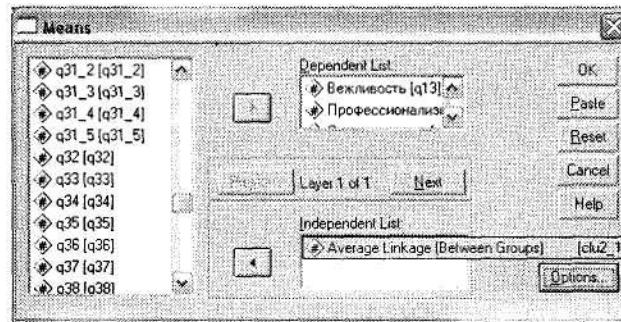


Рис. 5.51. Диалоговое окно Means

Откроется диалоговое окно Options, выберите в нем необходимые статистики для сравнения кластеров (рис. 5.52). Для этого в поле Cell Statistics оставьте только вывод средних значений Mean, удалив из него другие установленные по умолчанию статистики. Закройте диалоговое окно Options щелчком на кнопке Continue. Наконец, из главного диалогового окна Means запустите процедуру сравнения средних значений (кнопка OK).

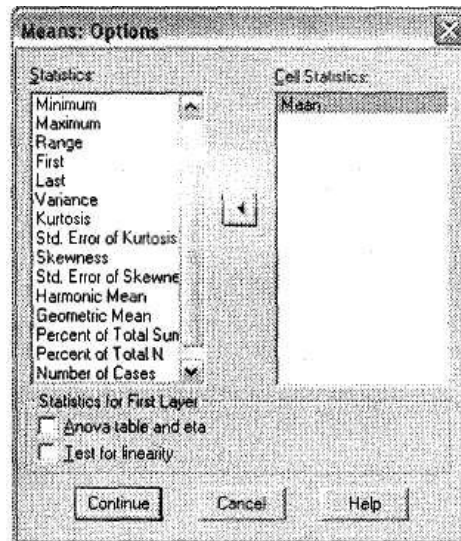


Рис. 5.52. Диалоговое окно Options

В открывшемся окне SPSS Viewer появятся результаты работы статистической процедуры сравнения средних значений. Нас интересует таблица Report (рис. 5.53). Из нее можно увидеть, на каком основании SPSS разделила респондентов на два кластера. Таким критерием в нашем случае служит уровень оценок по анализируемым параметрам. Кластер 1 состоит из респондентов, для которых средние оценки по всем критериям сегментирования находятся на сравнительно высоком уровне (4,40 балла и выше). Кластер 2 включает респондентов, оценивших рассматриваемые критерии сегментирования достаточно низко (3,35 балла и ниже). Таким образом, можно сделать вывод о том, что 93,3 % респондентов, сформировавшие кластер 1, оценили анализируемые авиакомпании по всем параметрам в целом хорошо; 5,4 % — достаточно низко; 1,3 % — затруднились ответить (см. рис. 5.50). Из рис. 5.53 можно также сделать вывод о том, какой уровень оценок для каждого из рассматриваемых параметров в отдельности является высоким, а какой — низким (причем данный вывод будет сделан со стороны респондентов, что позволяет добиться высокой точности классификации). Из таблицы Report можно видеть, что для переменной Регулирование очереди высоким считается уровень средней оценки 4,40, а для параметра Внешний вид — 4.72.

Report

Mean

Average Linkage (Between Groups)	Вежливость	Профессионализм	Оперативность	Готовность помочь	Регулирование очереди	Внешний вид	Работа персонала в целом
1	4,68	4,67	4,50	4,62	4,40	4,72	4,65
2	3,30	3,23	3,10	3,10	2,70	3,35	3,05
Total	4,80	4,59	4,50	4,54	4,31	4,84	4,58

Рис. 5.53. Сравнение средних для двух выделенных кластеров

Может оказаться, что в аналогичном случае по параметру X высокой оценкой считается 4,5, а по параметру Y — только 3,9. Это не будет ошибкой кластеризации, а напротив, позволит сделать важный вывод относительно значимости для респондентов рассматриваемых параметров. Так, для параметра Y уже 3,9 балла является хорошей оценкой, тогда как к параметру X респонденты предъявляют более строгие требования.

Мы идентифицировали два значимых кластера, различающиеся по уровню средних оценок по критериям сегментирования. Теперь можно присвоить метки полученным кластерам: для 1 — Авиакомпания, удовлетворяющие требованиям респондентов (по семи анализируемым критериям); для 2 — Авиакомпания, не удовлетворяющие требованиям респондентов. Теперь можно посмотреть, какие конкретно авиакомпании (закодированные в переменной q4) удовлетворяют требованиям респондентов, а какие — нет по критериям сегментирования. Для этого следует построить перекрестное распределение переменной q4 (анализируемые авиакомпании) в зависимости от кластеризующей переменной clu2_1. Результаты такого перекрестного анализа представлены на рис. 5.54.

По этой таблице можно сделать следующие выводы относительно членства исследуемых авиакомпаний в выделенных целевых сегментах.

Авиакомпания * Average Linkage (Between Groups)		Average Linkage (Between Groups)		Total
		Авиакомпания, удовлетворяющая респондентов	Авиакомпания, не удовлетворяющая респондентов	
Авиакомпания для сравнения	Внуковские авиалинии	5,5%		6,2%
	Домодедовские авиалинии	0,0%	10,2%	0,5%
	Пулково	6,8%	36,4%	8,2%
	Сибирь	3,7%	12,1%	4,1%
	Уральские авиалинии	2,1%	3,0%	2,1%
	Самарские авиалинии	,3%	3,0%	,4%
	Трансаэро	11,4%	9,1%	11,3%
	KrasAir	1,0%	9,1%	2,3%
	American Airlines	5,0%		4,8%
	Continental	1,6%		1,6%
	Delta Airlines	9,5%		9,0%
	Air France	7,0%		6,6%
	Alitalia	2,5%		2,4%
	Austrian Airlines	,6%		,6%
	British Airways	4,9%		4,7%
	Finnair	,9%	9,1%	1,3%
	Swissair	1,2%		1,1%
	KLM	10,4%		9,9%
	Lufthansa	10,5%		10,0%
	SAS	1,9%		1,0%
	Korean Airlines	,9%		,8%
	Japan Airlines	1,2%		1,1%
	24	1,2%		1,1%
	Total	100,0%	100,0%	100,0%

Рис. 5.54. Членство авиакомпаний в кластерах

1. Авиакомпании, полностью удовлетворяющие требованиям всех клиентов по параметру работы наземного персонала (входят только в один первый кластер):

- Внуковские авиалинии;
- American Airlines;
- Continental;
- Delta Airlines;
- Air France;
- Alitalia;
- Austrian Airlines;
- British Airways;
- Swiss Air;
- KLM;
- Lufthansa;
- SAS;
- Korean Airlines;
- Japan Airlines.

2. Авиакомпании, удовлетворяющие требованиям большинства своих клиентов по параметру работы наземного персонала (большая часть респондентов, летающих данными авиакомпаниями, удовлетворены работой наземного персонала):

- Трансаэро.

3. Авиакомпании, не удовлетворяющие требованиям большинства своих клиентов по параметру работы наземного персонала (большая часть респондентов, летающих данными авиакомпаниями, не удовлетворены работой наземного персонала):

- Домодедовские авиалинии;
- Пулково;
- Сибирь;
- Уральские авиалинии;
- Самарские авиалинии;
- KrasAir;
- Finnair.

Таким образом, получено три целевых сегмента авиакомпаний по уровню средних оценок, характеризующиеся различной степенью удовлетворенности респондентов работой наземного персонала:

1. наиболее привлекательные для пассажиров авиакомпании по уровню работы наземного персонала (14);
2. скорее привлекательные авиакомпании (1);
3. скорее непривлекательные авиакомпании (7).

Мы успешно завершили все этапы кластерного анализа и сегментировали авиакомпании по семи выделенным критериям.

Теперь приведем описание методики кластерного анализа в паре с факторным. Используем условие задачи из раздела 5.2.1 (факторный анализ). Как уже было сказано, в задачах сегментирования при большом числе переменных целесообразно предварять кластерный анализ факторным. Это делается для сокращения количества критериев сегментирования до наиболее значимых. В нашем случае в исходном файле данных у нас есть 24 переменные. В результате факторного анализа нам удалось сократить их число до 5. Теперь это число факторов может эффективно применяться для кластерного анализа, а сами факторы — использоваться в качестве критериев сегментирования.

Если перед нами стоит задача сегментировать респондентов по их оценке различных аспектов текущей конкурентной позиции авиакомпании X, можно провести иерархический кластерный анализ по выделенным пяти критериям (переменные `nfac1_1-nfac5_1`). В нашем случае переменные оценивались по разным шкалам. Например, оценка 1 для утверждения Я бы не хотел, чтобы авиакомпания менялась и такая же оценка утверждению Из-

менения в авиакомпании будут позитивным моментом диаметрально противоположны по смыслу. В первом случае 1 балл (совершенно не согласен) означает, что респондент приветствует изменения в авиакомпании; во втором случае оценка в 1 балл свидетельствует о том, что респондент отвергает изменения в авиакомпании. При интерпретации кластеров у нас неизбежно возникнут трудности, так как такие противоположные по смыслу переменные могут

попасть в один и тот же фактор. Таким образом, для целей сегментирования рекомендуется сначала привести в соответствие шкалы исследуемых переменных, а затем пересчитать факторную модель. И уже далее проводить кластерный анализ над полученными в результате факторного анализа переменными-факторами. Мы не будем снова подробно описывать процедуры факторного и кластерного анализа (это было сделано выше в соответствующих разделах). Отметим лишь, что при такой методике в результате у нас получилось три целевые группы авиапассажиров, различающихся по уровню оценок выделенным факторам (то есть группам переменных): низшая, средняя и высшая.

Весьма полезным применением кластерного анализа является разделение на группы частотных таблиц. Предположим, у нас есть линейное распределение ответов на вопрос Какие марки антивирусов установлены в Вашей организации?. Для формирования выводов по данному распределению необходимо разделить марки антивирусов на несколько групп (обычно 2-3). Чтобы разделить все марки на три группы (наиболее популярные марки, средняя популярность и непопулярные марки), лучше всего воспользоваться кластерным анализом, хотя, как правило, исследователи разделяют элементы частотных таблиц на глаз, основываясь на субъективных соображениях. В противоположность такому подходу кластерный анализ позволяет научно обосновать выполненную группировку. Для этого следует ввести значения каждого параметра в SPSS (эти значения целесообразно выражать в процентах) и затем выполнить кластерный анализ для этих данных. Сохранив кластерное решение для необходимого количества групп (в нашем случае 3) в виде новой переменной, мы получим статистически обоснованную группировку.

Заключительную часть этого раздела мы посвятим описанию применения кластерного анализа для классификации переменных и сравнения его результатов с результатами факторного анализа, проведенного в разделе 5.2.1. Для этого мы вновь воспользуемся условием задачи про оценку текущей позиции авиакомпании X на рынке авиаперевозок. Методика проведения кластерного анализа практически полностью повторяет описанную выше (когда сегментировались респонденты).

Итак, в исходном файле данных у нас есть 24 переменные, описывающие отношение респондентов к различным аспектам текущей конкурентной позиции авиакомпании X. Откройте главное диалоговое окно Hierarchical Cluster Analysis и поместите 24 переменные (q1-q24) в поле Variable(s), рис. 5.55. В области Cluster укажите, что вы классифицируете переменные (отметьте параметр Variables). Вы увидите, что кнопка Save стала недоступна — в отличие от факторного, в кластерном анализе нельзя сохранить факторные рейтинги для всех респондентов. Откажитесь от вывода диаграмм, деактивизировав параметр Plots. На первом этапе вам не нужны другие параметры, поэтому просто щелкните на кнопке О К, чтобы запустить процедуру кластерного анализа.

В окне SPSS Viewer появилась таблица Agglomeration Schedule, по которой мы определили оптимальное число кластеров описанным выше методом (рис. 5.56). Первый скачок коэффициента агломерации наблюдается на 20 шаге (с 18834,000 до 21980,967). Исходя из общего числа анализируемых переменных, равного 24, можно вычислить оптимальное число кластеров: $24 - 20 = 4$.

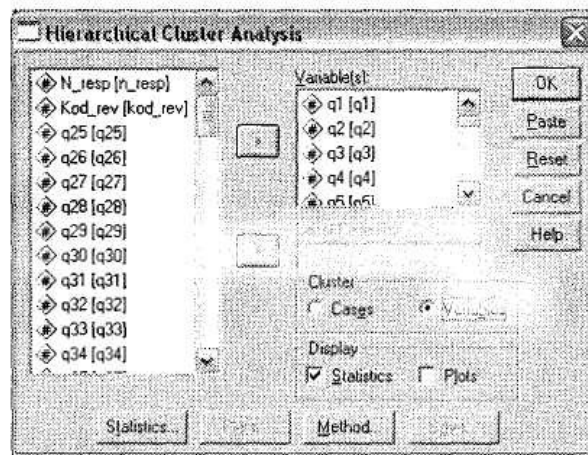


Рис. 5.55. Диалоговое окно Hierarchical Cluster Analysis при кластеризации переменных

Stage	Cluster Combined		Coefficient	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	22	24	4721,000	0	0	2
2	18	22	7556,500	0	1	16
3	1	10	7874,000	0	0	6
4	3	5	9112,000	0	0	10
5	8	11	9610,000	0	0	7
6	1	2	9670,000	3	0	8
7	8	12	10455,000	5	0	9
8	1	21	10871,667	6	0	11
9	6	8	11345,000	0	7	13
10	3	14	11972,000	4	0	12
11	1	16	12127,250	8	0	14
12	3	23	12408,333	10	0	14
13	8	7	13106,250	9	0	17
14	1	3	13979,850	11	12	16
15	19	20	14408,000	0	0	16
16	18	19	15684,333	2	15	19
17	4	6	16755,200	0	13	21
18	1	13	17088,444	14	0	20
19	9	18	18834,000	0	16	20
20	1	9	21980,967	18	19	23
21	4	17	22526,000	17	0	22
22	4	15	27274,429	21	0	23
23	1	4	30669,563	20	22	0

Рис. 5.56. Таблица Agglomeration Schedule

При классификации переменных практически и статистически значимым является кластер, состоящий всего из одной переменной. Поэтому, поскольку мы получили приемлемое число кластеров математическим методом, проведение дальнейших проверок не требуется. Вместо этого снова откройте главное диалоговое окно кластерного анализа (все данные, использованные на предыдущем этапе, сохранились) и щелкните на кнопке Statistics, чтобы организовать вывод классификационной таблицы. Вы увидите одноименное диалоговое окно, где необходимо указать число кластеров, на которое необходимо разделить 24 переменные (рис. 5.57). Для этого выберите параметр Single solution и в соответствующем поле укажите требуемое число кластеров: 4. Теперь закройте диалоговое окно Statistics щелчком на кнопке Continue и из главного окна кластерного анализа запустите

процедуру на выполнение.

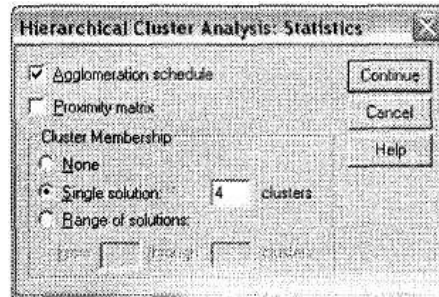


Рис. 5.57. Диалоговое окно Statistics

В результате в окне SPSS Viewer появится таблица Cluster Membership, распределяющая анализируемые переменные на четыре кластера (рис. 5.58).

Cluster Membership	
Case	4 Clusters
q1	1
q2	1
q3	1
q4	2
q5	1
q6	2
q7	2
q8	2
q9	1
q10	1
q11	2
q12	2
q13	1
q14	1
q15	3
q16	1
q17	4
q18	1
q19	1
q20	1
q21	1
q22	1
q23	1
q24	1

Рис. 5.58. Таблица Cluster Membership

По данной таблице можно отнести каждую рассматриваемую переменную в определенный кластер следующим образом.

Кластер 1

q1. Авиакомпания X обладает репутацией компании, превосходно обслуживающей пассажиров.

- q2. Авиакомпания X может конкурировать с лучшими авиакомпаниями мира.
- q3. Я верю, что у авиакомпании X есть перспективное будущее в мировой авиации.
- q5. Я горжусь тем, что работаю в авиакомпании X.
- q9. Нам предстоит долгий путь, прежде чем мы сможем претендовать на то, чтобы называться авиакомпанией мирового класса.
- q10. Авиакомпания X действительно заботится о пассажирах.
- q13. Мне нравится, как в настоящее время авиакомпания X представлена визуально широкой общественности (в плане цветовой гаммы и фирменного стиля).
- q14. Авиакомпания X — лицо России.
- q16. Обслуживание авиакомпании X является последовательным и узнаваемым во всем мире.
- q18. Авиакомпании X необходимо меняться для того, чтобы использовать в полной мере имеющийся потенциал.
- q19. Я думаю, что авиакомпании X необходимо представить себя в визуальном плане более современно.
- q20. Изменения в авиакомпании X будут позитивным моментом. q21. Авиакомпания X — эффективная авиакомпания.
- q22. Я бы хотел, чтобы имидж авиакомпании X улучшился с точки зрения иностранных пассажиров.
- q23. Авиакомпания X — лучше, чем многие о ней думают.
- q24. Важно, чтобы люди во всем мире знали, что мы — российская авиакомпания.

Кластер 2

- q4. Я знаю, какой будет стратегия развития авиакомпании X в будущем.
- q6. В авиакомпании X хорошее взаимодействие между подразделениями.
- q7. Каждый сотрудник авиакомпании прикладывает все усилия для того, чтобы обеспечить ее успех.
- q8. Сейчас авиакомпания X быстро улучшается.
- q11. Среди сотрудников авиакомпании имеет место высокая степень удовлетворенности работой.
- q12. Я верю, что менеджеры высшего звена прикладывают все усилия для достижения успеха авиакомпании.

Кластер 3

- q15. Мы выглядим «вчерашним днем» по сравнению с другими авиакомпаниями.

Кластер 4

- q17. Я бы не хотел, чтобы авиакомпания X менялась.

Сравнив результаты факторного (раздел 5.2.1) и кластерного анализов, вы увидите, что они существенно различаются. Кластерный анализ не только предоставляет существенно меньшие возможности для кластеризации переменных (например, отсутствие возможности сохранять групповые рейтинги) по сравнению с факторным анализом, но и выдает гораздо менее наглядные результаты. В нашем случае, если кластеры 2, 3 и 4 еще поддаются логической интерпретации¹, то кластер 1 содержит совершенно разные по смыслу утверждения. В данной ситуации можно либо попытаться описать кластер 1 как есть, либо перестроить статистическую модель с другим числом кластеров. В последнем случае для поиска оптимального числа кластеров, поддающихся логическому описанию, можно воспользоваться параметром Range of solutions в диалоговом окне Statistics (см. рис. 5.57), указав в соответствующих полях минимальное и максимальное число кластеров (в нашем случае 4 и 6 соответственно). В такой ситуации SPSS перестроит таблицу Cluster Membership для каждого числа кластеров. Задача аналитика в данном случае — попытаться подобрать такую классификационную модель, при которой все кластеры будут

интерпретироваться однозначно. С целью демонстрации возможностей процедуры кластерного анализа для кластеризации переменных мы не будем перестраивать кластерную модель, а ограничимся лишь сказанным выше.

Необходимо отметить, что, несмотря на кажущуюся простоту проведения кластерного анализа по сравнению с факторным, практически во всех случаях из маркетинговых исследований факторный анализ оказывается быстрее и эффективнее кластерного. Поэтому для классификации (сокращения) переменных мы настоятельно рекомендуем использовать именно факторный анализ и оставить применение кластерного анализа для классификации респондентов.

Классификационный анализ является, пожалуй, одним из наиболее сложных, с точки зрения неподготовленного пользователя, статистических инструментов. С этим связана его весьма малая распространенность в маркетинговых компаниях. Вместе с тем именно данная группа статистических методов является и одной из наиболее полезных для практиков в области маркетинговых исследований.

Заключение

SPSS — это мощный современный аналитический инструмент, при помощи которого можно проводить любой тип анализа данных в маркетинговых исследованиях. И в то время как построение математических моделей обычно считается прерогативой ученых, статистический анализ данных — это задача исследователей. Исследования являются неотъемлемой частью работы маркетологов, поэтому до тех пор, пока будут существовать вопросы, подлежащие решению при помощи маркетингового анализа, будет существовать и потребность в проведении статистического анализа данных. Цель настоящего пособия — сделать проведение статистического анализа понятным каждому исследователю.

Компания SPSS Inc. разработала целый ряд руководств, в которых подробно описываются выпущенные ею программные продукты. В сумме они насчитывают 5000 страниц, где вы найдете ответ на любой вопрос, связанный с работой практикующего исследователя. Однако все эти пособия написаны преимущественно на английском языке, что является препятствием для многих отечественных маркетологов. Руководства содержат весь объем информации о возможностях SPSS, включая информацию, которая используется редко либо вообще неприменима к маркетинговым исследованиям. В настоящем же пособии страниц гораздо меньше, и наша задача состояла в том, чтобы рассказать о ключевых моментах применения SPSS именно в маркетинговых исследованиях, с учетом специфических особенностей данного вида профессиональной деятельности. Несмотря на небольшой объем настоящего руководства, оно охватывает 95 % всех используемых на практике статистических методик. Для изучения оставшихся 5 % методов, редко применяемых в маркетинговых исследованиях, мы рекомендуем обратиться к оригинальным руководствам по SPSS.

Приложение. 12 полезных советов

Напоследок хочется рассказать о некоторых полезных свойствах программы SPSS, которые существенно облегчат работу с ней.

1. Несколько файлов (баз данных) SPSS можно объединять, добавляя при этом либо новые переменные, либо новых респондентов.

Чтобы добавить поля (переменные) в базу данных SPSS, подготовьте два файла данных (за один цикл можно объединить только два файла). В обоих файлах — реципиенте (база данных, в которую следует добавить переменные) и доноре (база данных, в которой содержатся добавляемые переменные) — необходимо, во-первых, проследить, чтобы имена добавляемых переменных не повторяли имя файла-реципиента; во-вторых, создать ключевое поле, то есть переменную, уникальным образом идентифицирующую респондентов. Обычно эту роль берет на себя номер анкеты. Отсортируйте оба файла по этой переменной (одинаковым образом: по возрастанию или убыванию). При помощи меню **Data ► Merge Files ► Add Variables** в открывшемся диалоговом окне выберите эту ключевую переменную; затем выберите параметр **Match case on key variables in sorted files**; поместите ключевую переменную в поле **Key Variables**. Щелкните на кнопке **OK**, и в файл-реципиент будут добавлены новые переменные из файла-донора (после всех существующих переменных).

Добавление респондентов происходит следующим образом. Убедитесь, что оба файла (реципиент и донор) содержат одинаковые переменные (по имени и типу). Откройте диалоговое окно добавления респондентов при помощи меню **Data ► Merge Files ► Add Cases**. В нем будут автоматически отобраны и помещены в файл-реципиент только одинаковые переменные. После щелчка на кнопке **OK** изменения вступят в силу: новые респонденты будут добавлены в конец рабочего файла.

2. Построенные диаграммы можно изменять, дважды щелкнув на них мышью в окне SPSS Viewer. Простые диаграммы, как будет показано в п. 3, содержат лишь базовые возможности форматирования диаграмм (в специальном окне SPSS Chart Editor), тогда как интерактивные диаграммы предоставляют значительный набор средств, аналогичных MS Microsoft Excel.

3. Графическая подсистема SPSS позволяет строить обычные (Simple) и интерактивные (Interactive) диаграммы. Вторые отличаются от первых более широкими возможностями форматирования. Однако какой бы тип диаграммы вы ни выбрали, они все равно не будут иметь такого же привлекательного вида, как диаграммы в Microsoft Excel. Диаграммы, выводимые в качестве дополнительного параметра в различных статистических процедурах (меню **Analyze**), — это только обычные диаграммы. Они предназначены исключительно для использования в процессе анализа данных аналитиками и не подходят для презентаций. Обычные диаграммы можно строить и отдельно от статистических процедур — при помощи меню **Graphs**. При этом если, скажем, в Microsoft Excel все диаграммы могут быть «на лету» преобразованы одна в другую, то в SPSS однажды построенная диаграмма может менять только элементы форматирования. Наиболее часто используемые виды диаграмм: **Bar** (гистограмма), **Line** (график), **Pie** (сектограмма) и **Scatter** (точечная). Интерактивные диаграммы доступны посредством меню **Graphs ► Interactive**, которое также содержит четыре типа наиболее часто используемых видов диаграмм. Обычные и интерактивные диаграммы могут быть как плоскими, так и объемными.

4. Таблицы в окне SPSS Viewer можно изменять, дважды щелкнув на них мышью. Далее выберите в меню **Pivot** пункт **Pivoting Traus**. Откроется дополнительное окно, с помощью которого можно поменять местами столбцы, ряды и уровни таблицы.

5. После создания таблиц линейных или перекрестных распределений на их основе можно строить различные диаграммы. Дважды щелкните на таблице мышью, чтобы

открыть ее. Затем выделите требуемые числовые значения (без названий переменных и вариантов ответа) и щелкните правой кнопкой мыши. В появившемся контекстном меню выберите Create Graph и в нем — требуемый тип диаграммы. После этого, например, будет построена интерактивная диаграмма.

6. Меню Analyze ► Custom Tables предоставляет доступ к диалоговым окнам, предназначенным для построения одно- и многомерных таблиц. При помощи этих окон вы можете создавать более презентабельные таблицы, чем Frequencies или Crosstabs. Мы рекомендуем использовать диалоговое окно Multiple Response Tables для работы с многовариантными переменными (вместо стандартной процедуры Analyze ► Multiple Response).

7. Часто при работе с SPSS возникает необходимость скопировать результаты работы программы из окна SPSS Viewer в Microsoft Word или Microsoft Excel. Для того чтобы скопировать диаграмму, выделите ее, щелкнув на ней правой кнопкой мыши, и в открывшемся контекстном меню выберите пункт Copy. Таблицы копируются методами, различными для Microsoft Word и Microsoft Excel. Так, чтобы скопировать таблицу в Microsoft Excel, выделите ее правой кнопкой мыши и в открывшемся меню выберите пункт Copy. После этого вставка в Microsoft Excel производится обычным способом. В Microsoft Word вы можете вставить таблицу, во-первых, в виде рисунка (метафайла) — выделите ее при помощи правой кнопки мыши и выберите пункт Copy Objects (при этом таблицу нельзя изменять) — и, во-вторых, в виде собственно таблицы. Однако если вы просто скопируете и вставите ее в Microsoft Word, таблица потеряет оформление и может стать нечитаемой. Мы рекомендуем вставлять таблицу в Microsoft Word, предварительно скопировав ее в Microsoft Excel.

8. В любых диалоговых окнах SPSS, так же как и в окне SPSS Viewer, вы можете получить справку по конкретным элементам (кнопкам, полям, статистикам) — для этого щелкните на них правой кнопкой мыши. Чтобы уточнить смысл какой-либо статистики, представленной в объектах SPSS Viewer, сначала нужно открыть их двойным щелчком мыши, а затем при помощи правой кнопки мыши получить информацию об интересующей статистике.

9. Весьма мощным средством работы с SPSS для опытных исследователей является командный синтаксис (Syntax). С его помощью можно, во-первых, автоматизировать повторяющиеся операции (например, построение 30 регрессий), а во-вторых, получать доступ к статистическим процедурам, не выведенным в основное меню программы (например, MANOVA). Краткое описание командного синтаксиса заняло бы много страниц. Тем не менее даже начинающие аналитики, не имеющие опыта работы с ним, могут достаточно эффективно использовать командный синтаксис, изучая автоматически генерируемые при работе с меню команды. Для того чтобы увидеть внутреннюю команду синтаксиса при работе с какой-либо статистической процедурой, следует предварительно выбрать в меню Edit ► Options на вкладке Viewer параметр Display commands in the log. После этого все ваши действия будут автоматически отображаться в окне SPSS Viewer в виде простого текста, который можно скопировать в окно Syntax (вызывается при помощи меню File ► New ► Syntax).

10. Не следует путать программный синтаксис (Syntax) со встроенным языком программирования SPSS (Script). Окно программирования открывается при помощи меню File ► New ► Script. Язык программирования SPSS похож на Microsoft Visual Basic for Applications (VBA), однако он содержит отдельные функции, специфичные для работы со структурой базы данных формата SPSS. Встроенный язык программирования весьма беден на визуальные средства интерактивного пользовательского интерфейса, однако он может с успехом применяться в качестве клиента автоматизации, то есть для интегрирования различных приложений, поддерживающих VBA (например, все приложения Microsoft Office) с SPSS. При помощи этого языка можно, например, строить графики в Microsoft Excel или формировать демонстрационные отчеты в Microsoft Word.

11. Командный синтаксис SPSS обладает многими возможностями полноценного макроязыка. В нем есть переменные, циклы, условные операции и т. д. Однако в некоторых случаях языка синтаксиса оказывается недостаточно. Мы рекомендуем использовать командный синтаксис для операций с матрицей данных, то есть с анкетами респондентов, находящимися в окне Data View. Иными словами, проводить такие операции, как чистка базы данных (корректировка пропущенных значений, логической структуры ответов и т. п.), формирование исходного списка переменных в окне Variable View, «подвешивание» меток переменных, а также операции с отдельными ячейками данных (например, копирование-вставка из других программ). Для операций с результатами расчетов (таблицами, результатами статистических тестов и т. д.), расположенными в окне SPSS Viewer, рекомендуется использовать другой встроенный язык программирования SPSS — язык скриптов. Практика показывает, что большинство компаний, занимающихся маркетинговыми исследованиями, производят обработку таблиц, построенных в SPSS, в других программах (чаще всего в MS Excel). Ниже мы покажем, как при помощи языка скриптов SPSS автоматизировать процесс переноса таблиц из окна SPSS Viewer в MS Excel (для построения диаграмм) и в MS Word.

Мы уже не раз упоминали о слабости графической подсистемы SPSS. В связи с этим исследователи строят диаграммы в MS Excel, копируя их из окна SPSS

Viewer. Этот процесс может стать «узким местом» всего исследования, так как при большом объеме таблиц с линейными и перекрестными распределениями процесс построения диаграмм занимает весьма значительный период времени. Давайте посмотрим, как можно легко и быстро автоматизировать данный процесс. Итак, предположим, что у нас есть 100 таблиц с линейными распределениями по различным вопросам анкеты. Все эти таблицы находятся в окне SPSS Viewer. Откройте редактор скриптов SPSS при помощи меню File ► New ► Script. Появится диалоговое окно Use Starter Script, которое предлагает использовать текст уже написанной программы в качестве шаблона для нашего скрипта. Мы будем создавать скрипт самостоятельно, поэтому просто щелкните на кнопке Отмена. Появится окно редактора скриптов SPSS, содержащее полноценную среду разработки (IDE). Слева вы увидите две вкладки — 1 и 2. Мы будем писать скрипт1 на установленной по умолчанию вкладке 1. Скрипты в SPSS пишутся на VBA-совместимом языке Sax Basic. Его возможности в целом более ограничены по сравнению с VBA (а средства разработки диалоговых окон не выдерживают никакой критики). В окне редактора скриптов SPSS по умолчанию введены начальная и конечная строки программы:

```
Sub Main
```

```
End Sub
```

Пользователь не должен изменять эти строки. Весь текст программы записывается между данными двумя строками2. Для переноса всего содержимого окна SPSS Viewer в MS Word следует ввести в редакторе скриптов следующий текст (листинг П. 1).

```
Set objWD = CreateObject("Word.Application")
objWD.Visible = True
objWD.Documents.Add
Dim objOutputDoc As ISpssOutputDoc
Set objOutputDoc = objSpssApp.GetDesignatedOutputDoc
Dim objItems As ISpssItems
Set objItems = objOutputDoc.Items
For i = 1 To objItems.Count - 1
    objOutputDoc.ClearSelection
    Set objItem = objItems.GetItem(i)
    intItemType = objItem.SPSSType
    If intItemType = SPSSPivot Then
        objItem.Selected = True
        objOutputDoc.Copy
        objWD.Selection.PasteAndFormat (0)
    End If
Next
```

Листинг П.1. Скрипт для переноса таблиц из SPSS Viewer в MS Word

После того как вы введете этот текст в окно редактора скриптов SPSS, запустите его на выполнение при помощи щелчка на кнопке с символом ► на панели инструментов или просто нажав F5. В результате выполнения данной программы будет создан новый документ MS Word, содержащий все таблицы из окна SPSS Viewer.

Для того чтобы перенести все таблицы из окна SPSS Viewer в MS Excel, нужно ввести между начальной и конечной строками программы в редакторе синтаксиса следующий текст (листинг П.2).

Листинг П.2. Скрипт для переноса таблиц из SPSS Viewer в MS Excel

```
Set objXL = CreateObject("Excel.Application")
objXL.Visible = True
objXL.Workbooks.Add
Dim objOutputDoc As ISpssOutputDoc
Set objOutputDoc = objSpssApp.GetDesignatedOutputDoc
Dim objItems As ISpssItems
Set objItems = objOutputDoc.Items
For i = 0 To objItems.Count - 1
    objOutputDoc.ClearSelection
    Set objItem = objItems.GetItem(i)
    intItemType = objItem.SPSSType
    If intItemType = SPSSPivot Then
        objItem.Selected = True
        objOutputDoc.Copy
        objXL.Workbooks(1).Sheets.Add
        objXL.Workbooks(1).Sheets(1).Select
        objXL.Workbooks(1).Sheets(1).Paste
    End If
Next
```

В результате выполнения данной программы будет создана новая рабочая книга MS Excel, в которой на каждой вкладке будет одна таблица из окна SPSS Viewer. Далее можно написать макрос (например, на VBA) для построения диаграмм на основании таблиц, содержащихся в MS Excel.

12. И наконец, последнее. Согласно статистической теории, чтобы сделать возможным применение большинства статистических процедур, данные должны подчиняться закону нормального распределения. Если это не так, теоретически вместо стандартных тестов следует проводить непараметрические тесты. На практике (в маркетинговых исследованиях) данные оказываются нормально распределены редко. Более того, многие исследователи просто игнорируют данный теоретический постулат, считая данные, не подчиняющиеся нормальному распределению, выбросами (случайными значениями). Данная техника действительно оправдывает себя во многих примерах маркетинговых исследований, когда от абсолютной точности построенных статистических моделей ровным счетом ничего не зависит. Ведь исследователей в большинстве случаев интересует лишь общее направление различий, связей и т. п. В этом и заключается специфика маркетинговых исследований: нас не интересует, как ведет себя каждый респондент в выборке, — нам интересно знать, как ведут себя целевые группы. В частности, по этой причине в настоящем пособии мы не приводили проверку данных на нормальное распределение в качестве обязательного предварительного этапа статистического анализа. Если вас все же заинтересует форма распределения данных, это легко выяснить при помощи критерия Колмогорова-Смирнова. Откройте соответствующее диалоговое окно при помощи меню Analyze ► Nonparametric Tests ► 1-Sample K-S. На рис. П.1 показан общий вид данного

окна.

Для того чтобы протестировать какую-либо переменную на нормальность распределения, перенесите ее из левого списка всех доступных переменных в область для тестируемых переменных Test Variable List. Затем выберите тип распределения, на соответствие которому вы собираетесь проводить тестирование. По умолчанию выбран только тест на нормальное распределение (Normal). Также можно провести тест на распределение Пуассона (Poisson), равномерное (Uniform) и экспоненциальное распределение (Exponential). Как в тесте %2 (см. раздел 4.1), тесты на вид распределения можно проводить асимптотическими методами и точными методами (Exact). Точные методы могут применяться в тех случаях, когда асимптотические методы неприменимы (например, при малых размерах выборки). Мы рекомендуем всегда проверять результаты асимптотических методов при помощи точных. Вывод точных тестов на вид распределения осуществляется при помощи кнопки Exact. Вид соответствующего диалогового окна аналогичен виду окна для теста %2. Напомним, что в нем следует выбрать параметр Monte-Carlo и указать доверительный уровень 95 %.

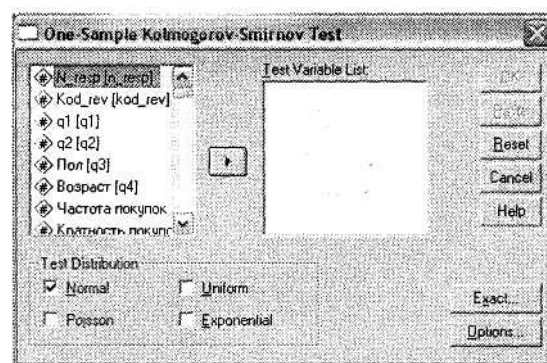


Рис. П.1. Диалоговое окно One-Sample Kolmogorov-Smirnov Test

One-Sample Kolmogorov-Smirnov Test			
			q5
N			1002
Normal Parameters a,b	Mean		
	Std. Deviation		
Most Extreme Differences	Absolute		
	Positive		
	Negative		
Kolmogorov-Smirnov Z			7,641
Asymp. Sig. (2-tailed)			,000
Monte Carlo Sig. (2-tailed)	Sig.		
	95% Confidence Interval	Lower Bound	,000
		Upper Bound	,000

a. Test distribution is Normal.

b. Calculated from data.

c. Based on 10000 sampled tables with starting seed 2000000.

Рис. П.2. Таблица One-Sample Kolmogorov-Smirnov Test

В результатах данного теста (окно SPSS Viewer) наше внимание должна привлечь значимость тестовой характеристики: асимптотическая (строка Asymp. Sig. (2-tailed)) и точная (строка Monte Carlo Sig.). На рис. П.2 представлен общий вид выводимых результатов при тесте на нормальное распределение. Так как нулевая (исходная) гипотеза для тестирования состоит в наличии нормального распределения, вероятность (статисти-

ческая значимость) менее 0,05 означает, что исследуемая переменная не подчиняется закону нормального распределения. Таким образом, значимая тестовая величина означает отсутствие, а незначимая — наличие исследуемого распределения.

Литература

1. Bums A. C, Bush R. F. Marketing research. New Jersey: Prentice-Hall Inc., 2000. 700 p.
2. George D., Mallery P. SPSS for Windows step by step: A simple guide and reference. Needham Heights: A Pearson Education Company, 2001. 380 p.
3. Malhotra N. K., Birks D. F. Marketing research. An applied approach. Essex: Pearson Education Ltd., 2000. 734 p.
4. SPSS Help.
5. БююльА., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб.: ДиаСофтЮП, 2002. 608 с.
6. Черчилль Г. А. Маркетинговые исследования. СПб.: Питер, 2001. 752 с.
7. Электронный учебник StatSoft. StatSoft Inc., 1998.
8. Field A. Discovering Statistics Using SPSS for Windows: Advanced Techniques for Beginners. Sage Publications, 2000. 512 p.
9. Einspruch E. L. Next Steps With SPSS. Sage Publications, 2003. 184 p.
10. Halley F., ZainoJ., Babbie E. Adventures in Social Research : Data Analysis Using SPSS 11.0/11.5 for Windows. Pine Forge Press, 2003. 544 p.
11. KinnearP. R., Gray C D. SPSS for Windows Made Simple: Release 10. Psychology Pr, 2000. 432 p.
12. Bums A. C, Bush R. F. Marketing Research and SPSS 11.0, Fourth Edition. Prentice Hall, 2002. 688 p.
13. Pavkov T. W., Pierce K.A. Ready, Set, Go! A Student Guide to SPSS® 11.0 for Windows®. McGraw-Hill Humanities; Social Sciences; Languages, 2002; 96 p.
14. Norusis M.J. SPSS 11.0 Guide to Data Analysis. Prentice Hall, 2002. 637 p.
15. Green S. B., SalkindN.J. Using SPSS for the Windows and Macintosh: Analyzing and Understanding Data (3rd Edition). Prentice Hall, 2002. 443 p.
16. PallantJ. SPSS Survival Manual: A Step By Step Guide to Data Analysis Using SPSS for Windows (Version 10). Open Univ Pr, 2001. 304 p.
17. Frankfort-Nachmias C, Leon-Guerrero A. Social Statistics for a Diverse Society With SPSS Student Version 11.0. Pine Forge Press, 2002. 672 p.