БАКАЛАВРИАТ

Г. Моосмюллер, Н.Н. Ребик

МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ С SPSS

УЧЕБНОЕ ПОСОБИЕ

ВЫСШЕЕ ОБРАЗОВАНИЕ – БАКАЛАВРИАТ
серия основана в 1996 г.

Г. МООСМЮЛЛЕР
Н.Н. РЕБИК

МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ С SPSS

Второе издание

УЧЕБНОЕ ПОСОБИЕ

Допущено Советом Учебно-методического объединения вузов России по образованию в области менеджмента в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности «Маркетинг»

Соответствует Федеральному государственному образовательному стандарту 3-го поколения

Москва
ИНФРА-М
2015

УДК 339.138(075.8)
ББК 65.290-2я73
М74

ФЗ № 436-ФЗ | Издание не подлежит маркировке в соответствии с п. 1 ч. 4 ст. 11

Рецензент:

Азоев Г.Л., д.э.н., профессор,
директор Института маркетинга ГУУ

Моосмюллер Г., Ребик Н.Н.

М74 Маркетинговые исследования с *SPSS*: Учеб. пособие. – 2-е изд. – М.: ИНФРА-М, 2015. – 200 с. – (Высшее образование: Бакалавриат).

ISBN 978-5-16-004240-4 (print)
ISBN 978-5-16-101563-6 (online)

В пособии подробно описаны основные методы статистического анализа, применяемые при обработке маркетинговой информации с использованием программного комплекса *SPSS*. Приводятся детальные инструкции пользования программой, показано, как проводить поэтапную интерпретацию результатов анализа.

Для студентов, обучающихся по специальности «Маркетинг» по дисциплинам «Маркетинговые исследования» и «Технологии маркетинговых исследований», для студентов-магистров и слушателей программы *MBA*. Может быть рекомендовано для студентов, обучающихся по специальностям «Статистика» и «Социология».

ББК 65.290-2я73

ISBN 978-5-16-004240-4 (print)
ISBN 978-5-16-101563-6 (online)

ВВЕДЕНИЕ

Данная работа посвящена основным методам статистического анализа, применяемым при обработке маркетинговой информации с использованием программного комплекса *SPSS* (версия 13.0) для *Windows*.

SPSS (***Statistical Package for Social Sciences*** или в новой интерпретации – ***Superior Performing Software Systems***) – система (программный пакет) статистической обработки информации, которая предоставляет пользователю широкие возможности преобразования и анализа данных, а также наглядного представления полученных результатов.

Подробное описание структуры редактора данных, детальные инструкции по использованию *SPSS*, поэтапная интерпретация результатов анализа, содержащиеся в этой книге, предназначены для начинающих пользователей программы.

Наше пособие существенно отличается от многих учебных пособий по *SPSS* тем, что в нем инструкции для пользователей объединены с подробным описанием механизма действия применяемых методов анализа.

Особенность представления основных методов статистического анализа в данной работе заключается в отсутствии формул расчета статистических показателей. Механизм действия различных методов анализа описан с помощью рисунков и графиков.

Применение каждого из представленных в книге методов анализа иллюстрируется примерами из практики Института исследования рынка *CenTouris* (производная от словосочетания «центр туризма»), который специализируется на исследованиях туристического рынка Восточной Баварии.

Данное учебное пособие подготовлено по материалам лекций, которые читаются в Университете г. Пассау, и по программе *MBA* специальности «Маркетинг» в Государственном университете управления (Москва).

1. ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА В МАРКЕТИНГОВЫХ ИССЛЕДОВАНИЯХ

1.1. ФОРМИРОВАНИЕ СТАТИСТИЧЕСКОЙ ВЫБОРКИ

В ходе масштабных маркетинговых исследований сбор информации по каждому респонденту, представляющему интерес для исследователей, сопряжен со значительными затратами времени и средств и поэтому является нереальным или экономически нецелесообразным.

Например, если объектом исследования являются способы проведения досуга студентами города Москвы, то собрать информацию о каждом студенте проблематично. В этом случае из ***общей (генеральной) совокупности*** (из числа лиц, интересующих исследователей) производится ***статистическая выборка*** с целью определения круга лиц для участия в проведении исследований.

Основным требованием, предъявляемым к статистической выборке, является ее ***репрезентативность***. Статистическая выборка считается репрезентативной (представительной), если она представляет собой «уменьшенную копию» генеральной совокупности и, следовательно, по данным, собранным в рамках статистической выборки, можно судить о генеральной совокупности в целом.

Существуют различные виды статистической выборки, которые отличаются по способу ее формирования, т.е. по технике проведения отбора. Различают случайную и эмпирическую выборки (табл. 1.1).

Случайная выборка характеризуется тем, что каждый элемент генеральной совокупности имеет шанс (отличный от нуля) оказаться в статистической выборке. При этом возможно рассчитать вероятность, с которой каждый элемент генеральной совокупности может оказаться в выборке.

Таблица 1.1

Виды статистической выборки

№ п/п	Вид выборки	Техника осуществления отбора
1	Случайная выборка (*высокая степень репрезентативности*): • простая случайная выборка • взвешенная случайная выборка • региональная (клюмпенная) выборка	Отбор респондентов производится случайным образом. По каждому элементу генеральной совокупности имеется одинаковая возможность собрать информацию
2	Эмпирическая выборка (*низкая степень репрезентативности*): • простая выборка • квотированная выборка	Отбор респондентов производится случайным образом из числа элементов генеральной совокупности, по которым имеется возможность собрать информацию

Существует несколько видов случайной выборки в зависимости от метода ее формирования (*Schmalen*, 2002. S. 390):

1. ***Простая случайная выборка*** предполагает, что все элементы генеральной совокупности имеют **равные** шансы оказаться в статистической выборке. Выбор производится по принципу лотереи. Элементы выборки извлекаются непосредственно из генеральной совокупности. Достоинство данного метода формирования выборки состоит в том, что не требуется знания структуры генеральной совокупности.

2. ***Взвешенная случайная выборка*** используется в том случае, если существует необходимость учитывать разделение генеральной совокупности на группы (слои). При этом известна структура генеральной совокупности (доли отдельных групп).

Статистическая выборка проводится случайным образом отдельно в каждой группе генеральной совокупности с сохранением пропорций соотношения размеров этих групп.

Например, в числе студентов, представляющих собой генеральную совокупность, 47% составляют юноши и 53% – девушки. При формировании взвешенной случайной выборки размером в 100 человек должны быть отобраны 47 юношей и 53 девушки (рис. 1.1). В результате этого, хотя отбор респондентов производится случайно, статистическая выборка имеет структуру, идентичную структуре генеральной совокупности, что повышает степень ее репрезентативности.

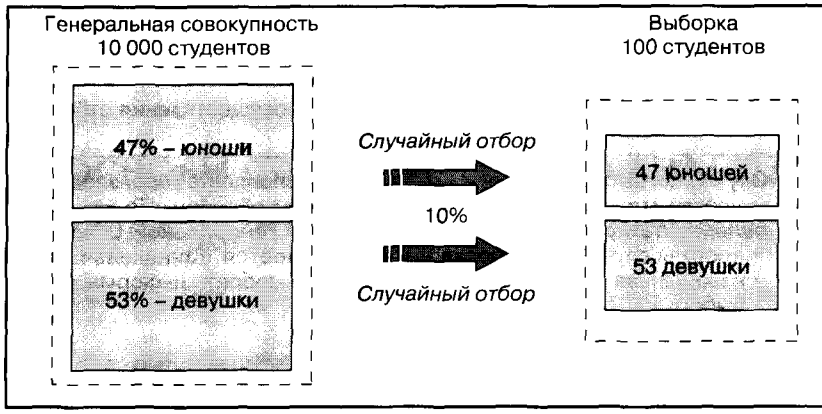


Рис. 1.1. Взвешенная случайная выборка

В качестве недостатков этого метода формирования статистической выборки следует отметить необходимость знания структуры генеральной совокупности и сложность организации сбора информации на практике.

3. ***Клюмпенная выборка*** используется также в том случае, если генеральная совокупность разделена на группы (клюмпены). Из общего числа клюмпенов случайным образом выбирается один, который используется как статистическая выборка. **Все** элементы клюмпена становятся элементами статистической выборки.

Этот метод формирования выборки часто называется «региональным»: генеральная совокупность – страна (город), выборка – республика (район города) (рис. 1.2). Например, если в качестве генеральной совокупности выступают все студенты города Москвы, то для формирования клюмпенной выборки случайным образом может быть выбран один из столичных вузов.

Достоинство клюмпенной выборки состоит в более простой организации процесса сбора информации и снижении затрат (экономия на транспортных расходах).

Основным недостатком данного метода формирования статистической выборки является клюмпенный эффект, который состоит в том, что клюмпены могут существенно отличаться друг от друга по структуре, что обусловливает низкую степень репрезентативности клюмпенной выборки. Едва ли по данным, собранным при участии студентов только одного московского вуза, можно судить обо всех студентах города Москвы.

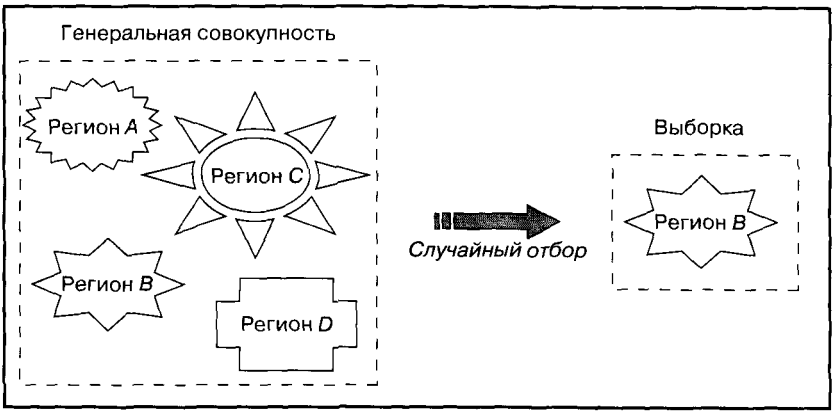


Рис. 1.2. Региональная (клюмпенная) выборка

При формировании круга респондентов для проведения маркетинговых исследований использование случайной выборки не всегда возможно или целесообразно. Например, при сборе информации посредством наблюдения не всегда возможно заранее четко определить круг людей, которые окажутся в поле зрения наблюдателя.

Формирование случайной статистической выборки предполагает возможность сбора информации по каждому элементу генеральной совокупности. Однако такая возможность не всегда существует на практике. Например, проведение исследований на территории вуза требует получения согласия его администрации. Одно это обстоятельство может стать серьезным препятствием быстрому и оперативному сбору информации.

На практике часто применяют эмпирическую выборку, когда в круг респондентов для сбора информации включается каждый «первый встречный», согласный принять участие в исследовании (при проведении наблюдения такое согласие не всегда является необходимым условием). В этом случае возможно также использование квотированной выборки, когда структура неэмпирической выборки определена заранее (например, 50% женщин и 50% мужчин).

Эмпирическая выборка характеризуется низкой степенью репрезентативности. Результаты исследований при использовании эмпирической выборки зависят от места и времени сбора информации. Например, при изучении досуга студентов города

Москвы результаты исследования будут определяться тем, где происходит сбор информации – у входа в ночной клуб или в библиотеку.

Статистическая выборка не используется при проведении качественных маркетинговых исследований, например исследований в форме экспертных опросов или фокус-групп. В этих случаях круг респондентов для проведения маркетинговых исследований формируется при помощи целенаправленной выборки.

При осуществлении целенаправленной выборки для участия в исследовании отбираются респонденты, которые могут предоставить наиболее точную и полную информацию (формирование экспертной группы), при участии которых можно организовать наиболее плодотворную дискуссию (формирование фокус-группы). В данном случае из числа потенциально возможных респондентов выбираются те, которые обладают наиболее ценной информацией и готовы поделиться ею для проведения исследований.

При формировании статистической выборки следует решить следующие вопросы:

1. Определить генеральную совокупность.
2. Определить размер выборки.
3. Выбрать метод формирования выборки.

Определение генеральной совокупности позволяет ответить на вопрос: «Из каких потенциальных респондентов следует производить выборку?» Это не всегда является очевидным. Например, кого следует привлекать для сбора информации при изучении вопросов семейного отдыха: жен, мужей, других членов семьи, работников туристических фирм или, может быть, всех вместе? Чтобы ответить на этот вопрос, исследователям необходимо решить, какого типа информация им нужна и кто ею, скорее всего, обладает (Янкевич, Безрукова, 2002. С.111).

Размер выборки определяется экономической целесообразностью сбора информации. Увеличение размера выборки способствует повышению репрезентативности и, следовательно, точности результатов исследования, однако это сопряжено с дополнительными затратами. В этом случае необходимо взвешивать экономическую ценность получаемой информации и затраты, связанные с ее сбором.

Сбор первичной информации в рамках статистической выборки осуществляется в форме проведения опроса, наблюдения или эксперимента.

1.2. ОСНОВНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

1.2.1. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ – метод ***классификации объектов*** по заданным признакам. Задача кластерного анализа состоит в формировании групп:

- однородных внутри (условие внутренней гомогенности);
- четко отличных друг от друга (условие внешней гетерогенности).

Целью кластерного анализа в маркетинге является определение целевых групп потребителей, для которых было бы целесообразно разработать специальное торговое предложение, т.е. уникальную комбинацию инструментов маркетинга.

Пример. Курильщики сигар, возраст и уровень доходов которых известны, исследуются на предмет возможности их разделения на однородные группы (кластеры) (рис. 1.3).

В варианте *B* однородные кластеры не выявлены. Следовательно, целенаправленная дифференциация торгового предложения невозможна.

В варианте *A* выявлены две однородные группы курильщиков сигар: «старые и бедные», «молодые и богатые», которых можно считать двумя целевыми группами потребителей. В этом случае целесообразно разработать два специальных торговых предложения – уникальных по цене, уровню качества продукции, упаковке, системе продвижения товара и т.д. (*Schmalen*, 2003. S. 401).

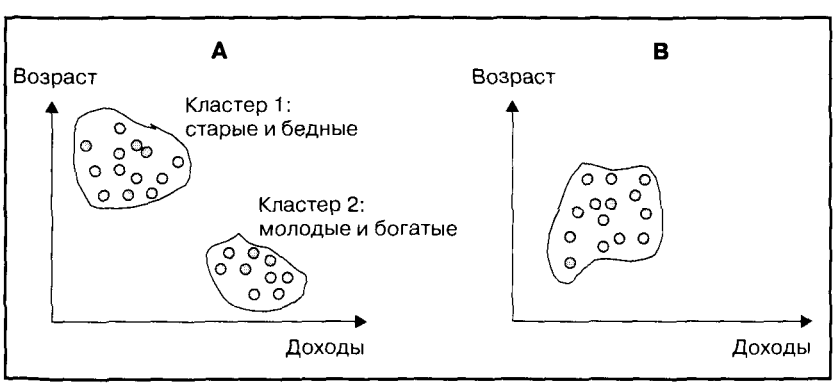


Рис. 1.3. Кластерный анализ

Элементы, включаемые в один и тот же кластер, имеют разную степень схожести (уровень отличия друг от друга). Техника кластерного анализа заключается в ***выявлении уровня схожести*** всех исследуемых элементов и последовательном ***объединении элементов в порядке возрастания уровня различия*** между ними. Число выявленных кластеров зависит от заданного уровня схожести (различия) элементов, включаемых в один кластер.

Техника кластерного анализа может быть проиллюстрирована ***дендограммой***, составляемой при помощи статистической компьютерной программы, в том числе *SPSS* (рис. 1.4).

На рис. 1.4 изображен результат кластерного анализа 18 предприятий розничной торговли, которые предлагают в качестве «особого предложения» (товары со скидками) один и тот же набор продуктов (примерно 50 наименований): молочные продукты, чистящие средства, косметика и т.д.

Целью кластерного анализа в данном случае является ответ на вопрос: возможно ли разделение исследуемых предприятий розничной торговли на кластеры в зависимости от их ценовой политики в плане формирования «особых предложений»?

В результате проведения кластерного анализа было выявлено три кластера: *A*, *B* и *C* (рис. 1.4). Предприятия розничной торговли 6, 18, 16, 1, 5, 15 (кластер *A*), так же как и 12, 2, 9, 17, 10 (кластер *C*), проводят одинаковую ценовую политику при формировании «особых предложений» (это, в частности, магазины торговых сетей *EDEKA* и *REWE*).

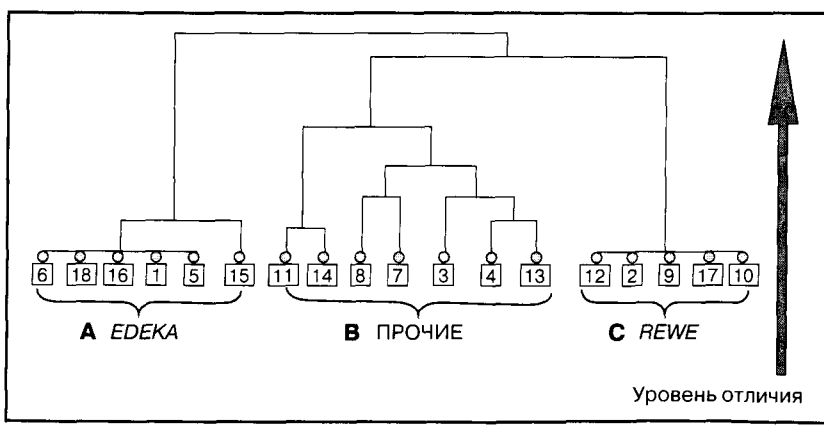


Рис. 1.4. Компьютерная дендограмма (кластерный анализ)

Предприятия розничной торговли, вошедшие в кластер *B* («Прочие»), не имеют одинаковой ценовой политики, но, тем не менее, их «особые предложения» имеют схожую ценовую структуру. Их можно объединить в одну группу только при задании определенного допустимого уровня их отличия друг от друга (*Schmalen*, 19. S. 402).

При повышении допустимого уровня отличия исследуемых элементов (снижении требований к однородности кластера) возможно объединение кластеров *B* и *C*, а затем присоединения к ним кластера *A*.

1.2.2. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Дискриминантный анализ проводится с целью ***выявления различий*** между исследуемыми группами. Например, могут быть исследованы группы потребителей конкурирующих товаров (или покупатели конкурирующих брендов) на предмет того, существуют ли различия между исследуемыми группами по заданным признакам. Иными словами, цель анализа – выяснить, можно ли составить «типичный портрет покупателя» для каждой исследуемой группы по заданным характеристикам.

Пример. Владельцев *BMW* и *VW*, возраст и доходы которых известны, исследуют на предмет того, можно ли разделить их (дискриминация) на две группы – «типичных владельцев *BMW*» и «типичных владельцев *VW*», так, чтобы группы владельцев характеризовались определенным уровнем дохода и возрастом (рис. 1.5).

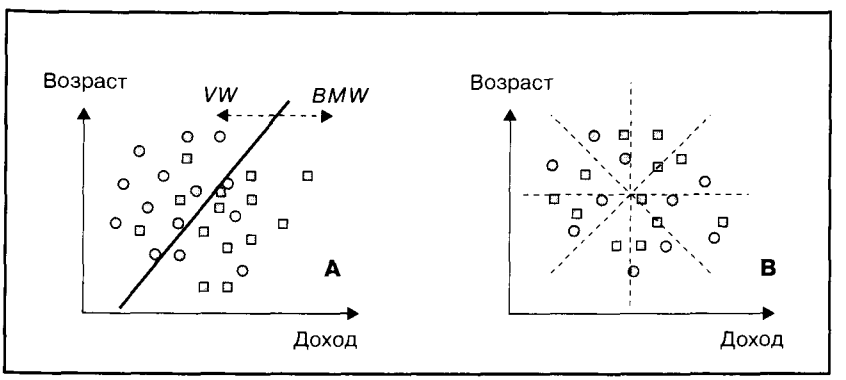


Рис. 1.5. Дискриминантный анализ

На рис. 1.5 в системе координат заданных характеристик отмечены сочетания возраста и дохода каждого исследуемого владельца автомобилей (*BMW* и *VW*).

В ходе дискриминантного анализа предпринимается попытка разделить существующие группы автовладельцев по возрасту и уровню дохода при помощи ***дискриминантной линии***. Дискриминантная линия должна быть проведена таким образом, чтоб комбинации характеристик владельцев автомобилей разных марок оказались расположенными по разные стороны линии и возможных пересечений было бы как можно меньше. В этом случае можно составить портрет «типичного владельца автомобиля определенной марки» по заданным характеристикам.

В варианте *B* возможны различные положения дискриминантной линии, при которых число пересечений будет в равной степени многочисленным. В данном случае невозможно разделить владельцев *BMW* и *VW* по уровню дохода и возрасту, т.е. не существует «портрета типичного владельца» *BMW* или *VW*.

В варианте *A* большая часть комбинаций уровней дохода и возраста владельцев *VW* лежит слева от дискриминантной линии, а владельцев *BMW* – справа. Это говорит о том, что владельцы *BMW* характеризуются более высоким уровнем дохода и относительно молоды по сравнению с владельцами *VW* (*Schmalen*, 2002. S. 403).

Характеристики «типичного потребителя», выявленные в результате проведения дискриминантного анализа, используются при прогнозировании поведения покупателей. Руководствуясь выявленными характеристиками «типичного покупателя», можно спрогнозировать, в пользу какого именно товара будет принято решение о покупке. В нашем примере (см. рис. 1.5) молодого человека с высоким уровнем дохода, желающего приобрести автомобиль, можно рассматривать как потенциального владельца *BMW*.

Если кластерный анализ выявляет возможность разбиения совокупности респондентов на группы, то дискриминантный анализ выявляет возможность установления различий уже существующих групп респондентов.

В настоящее время на практике для прогнозирования поведения потребителей используется более совершенный статистический метод – ***логистической регрессии***. Этот метод позволяет не только ответить на вопрос, какой именно товар потребитель выберет скорее всего, но и определить вероятность, с которой потребитель выберет тот или иной товар.

1.2.3. РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ – метод выявления ***статистической зависимости*** между исследуемыми переменными. На основе анализа эмпирических данных (данных, собранных в ходе проведения исследования) описывается не только сам факт существования статистической зависимости, но также описывается и математическая формула функции зависимости исследуемых переменных.

Современная техника регрессионного анализа позволяет описывать функции зависимости исследуемых переменных различных видов. Самая простая – линейная функция, определяемая при помощи линейного регрессионного анализа.

Стандартная модель простой линейной регрессии имеет вид

$$Y = a + b \cdot X,$$

где X – независимая переменная (фактор, влияющий на объект исследования);

Y – зависимая переменная (объект исследования);

a, b – постоянные величины (параметры модели).

Определение параметров модели (a, b) осуществляется путем применения ***метода наименьших квадратов***. Регрессионная линия должна быть проведена в «облаке эмпирических значений» таким образом, чтобы сумма квадратов вертикальных и горизонтальных расстояний от каждой точки до регрессионной линии была бы минимальной (рис. 1.6).

На рис. 1.6 показана технология выявления зависимости между исследуемыми переменными: уровнем дохода населения (независимая переменная X) и объемом оборота розничной торговли (за-

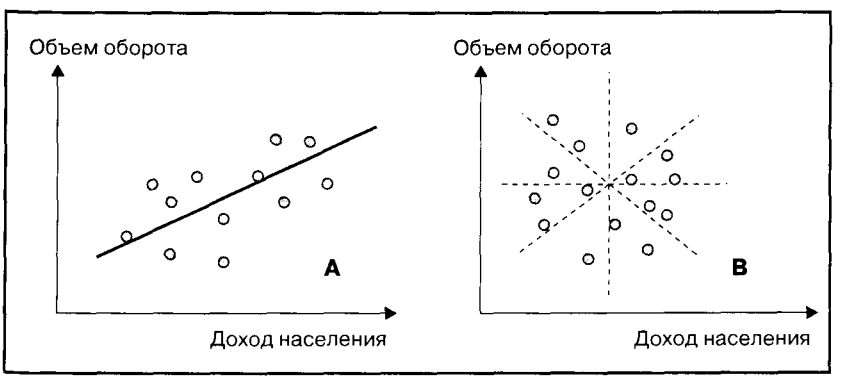


Рис. 1.6. Линейный регрессионный анализ

висимая переменная Y). В варианте *B* существует много возможностей проведения регрессионной линии, когда сумма квадратов расстояний от точек эмпирических значений до регрессионной линии будет примерно одинаковой. Возникает так называемый эффект пропеллера. В этом случае линейная зависимость между исследуемыми переменными отсутствует.

В варианте *A* можно найти наилучший вариант положения регрессионной линии при помощи метода наименьших квадратов. В этом случае действительно существует прямая линейная зависимость между уровнем доходов населения и объемом розничной торговли (*Schmalen*, 2002. S. 405).

Результаты регрессионного анализа используются для составления прогнозов изменения количественных переменных путем перенесения выявленных тенденций на будущие периоды.

В рассматриваемом примере (см. рис. 1.6 – вариант *A*) между уровнем дохода населения (X) и объемом торгового оборота (Y) существует линейная зависимость $Y = a + b \cdot X$. Если существует достаточно надежный прогноз относительно роста доходов населения (X), тогда исходя из данных прогноза (X) и регрессионной зависимости ($Y = a + b \cdot X$) можно составить прогноз роста объемов оборота розничной торговли (Y).

Использование регрессионного анализа в прогнозировании сопряжено с рядом проблем.

Во-первых, исходя из наличия достаточно устойчивой ***статистической зависимости*** не всегда можно делать выводы о существовании ***каузальной*** (причинно-следственной) ***взаимосвязи***. В нашем примере результаты регрессионного анализа не доказывают того, что растущий уровень доходов населения является причиной роста объемов оборота розничной торговли.

Во-вторых, результаты регрессионного анализа могут быть использованы для построения прогнозов только в случае верности «гипотезы стабильности во времени», т.е. если не происходит никаких структурных изменений. ***Гипотеза стабильности во времени*** предполагает изменение во времени только исследуемых переменных, все прочие величины являются постоянными. В приведенном выше примере рассматривается влияние уровня дохода на оборот розничной торговли. Предполагается, что степень влияния прочих факторов (например, цены, склонности потребителей к накоплению и т.д.) остается неизменной.

На практике результаты регрессионного анализа используются для составления прогнозов, как правило, в сочетании с опросами

экспертов. Такая комбинация количественных и качественных методов маркетинговых исследований соединяет точность математических расчетов со знаниями и интуицией экспертов.

1.2.4. ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ – метод, который позволяет сгруппировать большое число переменных (факторов, влияющих на предмет исследования) и свести их к минимальному числу «обобщающих факторов». Группировка данных производится по принципу:

- переменные, имеющие между собой высокую степень корреляции (тесную взаимосвязь), объединяются в один фактор;
- переменные, отнесенные к разным «обобщающим факторам», имеют между собой низкую степень корреляции (слабую взаимосвязь).

Факторный анализ производится в том случае, если существует огромный массив данных, который необходимо уменьшить («сжать») для проведения дальнейших исследований.

Например, существует база данных по результатам опроса, в ходе которого туристы, отдыхающие в курортной зоне «Баварский лес», оценивали эту курортную зону. Респонденты оценивали степень важности для них каждого из 13 предложенных мотивов выбора места отдыха (табл. 1.2).

Предположим, исследователям необходимо провести кластерный анализ туристов, отдыхающих в курортной зоне «Баварский лес», по таким характеристикам, как гражданство, уровень дохода и мотив выбора места отдыха. Проведение кластерного анализа затруднительно из-за больших размеров массива данных, содержащего информацию о мотивах проведения отпуска в «Баварском лесу», и из-за ограничений мощности вычислительной техники. Для удобства проведения кластерного анализа необходимо уменьшить объем («сжатие») данных при помощи факторного анализа.

В ходе факторного анализа осуществляется попарное сравнение исследуемых переменных с целью определения их схожести друг с другом, а также определяется число «группирующих факторов». В табл. 1.2 представлены результаты факторного анализа в рассматриваемом примере. Заданные 13 мотивов выбора места отдыха объединены в 4 фактора, определяющих выбор туристов в пользу «Баварского леса»:

1) гостеприимство по приемлемым ценам;
2) общение с природой;

3) специальное предложение Восточной Баварии;

4) культурная программа.

Также в табл. 1.2 представлены коэффициенты корреляции, которые характеризуют степень взаимосвязи между группируемыми переменными и группирующими факторами. Значения коэффициентов корреляции изменяется от –1 до +1.

Значение коэффициента корреляции, близкое к нулю, указывает на низкую степень взаимосвязи. Например, национальный колорит и самобытность «Баварского леса» (фактор «Специальное предложение Восточной Баварии») не обусловливается приемлемым уровнем цен (коэффициент корреляции 0,00055).

Отрицательное значение коэффициента корреляции указывает на существование обратной взаимосвязи. Например, приемлемые цены слабо отрицательно влияют на привлекательность «Баварского леса» с точки зрения общения с природой (коэффи-

Таблица 1.2

Результаты факторного анализа, проводимого при оценке курортной зоны «Баварский лес»

(адаптировано по: (*Schmalen*, 2002. S. 404))

Мотив выбора места отдыха (характеристики объекта исследования)	Обобщающий фактор			
	1	2	3	4
Искусство/ Достопримечательности	0,06360	0,04797	0,05432	0,83043
Лес/ Пеший туризм	–0,08891	0,81047	0,06419	0,02471
Ландшафт	0,00185	0,80955	0,01390	0,05039
Климат	0,29657	0,43442	0,28385	–0,15208
Национальный парк «Баварский лес»	–0,05132	0,13756	0,91019	0,00225
Заповедник	0,03718	–0,23299	0,86925	0,06811
Благотворная тишина	0,16038	–0,62086	0,16311	–0,01087
Старые города на Дунае	0,06461	–0,03917	0,07644	0,84183
Выгодные покупки изделий из стекла	0,33185	–0,04556	0,45635	0.17579
Приемлемые цены	0,72846	–0,01297	0,00055	0,05118
Вкусная еда	0,79633	0,04174	0,04779	0,13564
Гостеприимство	0,77615	0,24557	0,06791	0,11821
Комфорт отдыха с детьми	0,64865	–0,01099	0,06586	–0,09644

циент корреляции 0,01297). Это объясняется тем, что приемлемые цены привлекают множество туристов, что не способствует созданию атмосферы общения с природой.

Значение коэффициента корреляции, близкое к –1, указывает на наличие сильной обратной взаимосвязи. Такие случаи в рассматриваемом примере отсутствуют. Если значение коэффициента корреляции близко к +1, это свидетельствует о существовании плотной прямой взаимосвязи. Например, возможность заниматься пешим туризмом в лесу во многом определяет привлекательность рассматриваемого региона для тех, кто ценит общение с природой (коэффициент корреляции 0,81047) (см. табл. 1.2).

Характеристики объекта исследования объединяются в один обобщающий фактор при наличии высокой степени корреляции – как позитивной, так и негативной (в рассматриваемом примере встречается только сильная позитивная корреляция). Например, приемлемые цены, вкусная еда, гостеприимство и комфорт отдыха с детьми обобщаются в один фактор привлекательности курорта – «Гостеприимство по приемлемым ценам».

При допуске определенной потери информации (в данном случае 30%) впоследствии анализируются не 13 факторов, а только четыре. Такое «сжатие» данных существенно упрощает дальнейшее проведение исследования без существенной потери информации.

Факторный анализ целесообразно проводить только в том случае, если он *предшествует* применению других методов статистического анализа.

На практике факторный анализ всегда применяется в комбинации с другими статистическими методами обработки информации. Его можно охарактеризовать как вспомогательный метод, позволяющий упростить исследования путем сокращения анализируемой информации.

1.2.5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ – метод, при помощи которого исследуется влияние одной или нескольких независимых переменных на одну или несколько зависимых переменных.

Например, один и тот же продукт продается в нескольких регионах в упаковке разных типов (табл. 1.3). На основе данных объема продаж, сгруппированных по указанным признакам,

нужно определить, имеют ли существенное влияние на результаты продаж:

- регион и тип упаковки (основной эффект);
- комбинация этих факторов (интерактивный эффект).

Таблица 1.3

Дисперсионный анализ (зависимые и независимые переменные)

Независимая переменная № 1 (категориальный фактор)	Независимая переменная № 2 (категориальный фактор)		
	Регион I	Регион II	Регион III
	Показатели объема продаж (тыс. шт.) (зависимая переменная)		
Тип упаковки *A*	3567	5673	6478
Тип упаковки *B*	4567	2567	3569
Тип упаковки *C*	7856	4769	4736

Возможно, что исследуемые факторы влияют на объект исследования только в сочетании друг с другом. Например, упаковка, предназначенная для помещения в микроволновую печь, может способствовать значительному увеличению объемов продаж только в крупных городах.

Различают несколько видов дисперсионного анализа – в зависимости от числа исследуемых переменных (табл. 1.4).

Таблица 1.4

Виды дисперсионного анализа

Вид анализа	Число независимых переменных	Число зависимых переменных
Одномерный дисперсионный анализ (*ANOVA*: *Analysis of Variance*):		
• однофакторный (*one-way ANOVA*)	1	1
• двухфакторный	2	1
• трехфакторный	3	1
• ...	...	...
• многофакторный (*n-way ANOVA*)	Несколько	1
Многомерный дисперсионный анализ (*MANOVA*: *Multiple Analysis of Variance*)	1 или несколько	несколько

Пример постановки вопроса однофакторного дисперсионного анализа: влияет ли тип рекламы (плакаты, объявления в сред-

ствах массовой информации и др.) на число посетителей в кинотеатре?

Пример постановки вопроса двухфакторного дисперсионного анализа (см. табл. 1.3): влияет ли регион и тип упаковки на объем продаж определенного товара?

Пример постановки вопроса многомерного дисперсионного анализа: влияют ли регион и тип упаковки на объем продаж и число жалоб потребителей определенного товара?

В основе техники проведения дисперсионного анализа лежит ***сравнение средних величин в разных группах***. Например, для того чтобы определить, влияет ли пол студента на успеваемость, необходимо сравнить среднюю успеваемость юношей и девушек. Если средняя успеваемость девушек отличается от средней успеваемости юношей, то можно утверждать, что пол студента влияет на успеваемость, и наоборот.

Приведенный пример сравнения средних величин в двух группах (юношей и девушек) осуществляется при помощи ***T-теста***. *T*-тест является частным случаем дисперсионного анализа, в ходе которого осуществляется сравнение средних величин в нескольких группах.

Свое название дисперсионный анализ получил благодаря одному из условий сравнения средних величин в разных группах: дисперсии исследуемых величин в разных группах должны быть равны.

Дисперсия – показатель, характеризующий рассеяние значений количественного признака вокруг своего среднего значения. Подробно техника сравнения средних величин будет рассмотрена в главе 5 «Сравнение средних величин в *SPSS*».

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Чем обусловливается необходимость использования статистической выборки при проведении масштабных маркетинговых исследований?
2. В чем заключается основное требование, предъявляемое к статистической выборке?
3. Назовите основные методы формирования статистической выборки, их достоинства и недостатки.
4. Какие виды статистической выборки отличаются наиболее высокой степенью репрезентативности и почему?
5. Назовите основные задачи, решаемые в ходе формирования статистической выборки.

6. Назовите основные методы статистического анализа, применяемые в маркетинговых исследованиях, и их виды.
7. Назовите методы статистического анализа, при помощи которых можно найти ответы на следующие вопросы:

   а) влияет ли цвет упаковки товара и место его расположения в торговом зале на объем продаж;

   б) возможно ли разделить постоянных клиентов магазина на группы, используя в качестве критериев разделения объемы совершаемых покупок и частоту посещения магазина;

   в) насколько увеличится объем продаж товара при увеличении расходов на рекламу на 10% при условии постоянства цены на данный товар;

   г) по каким социально-демографическим признакам отличаются люди, приобретающие и не приобретающие товар *X*.

2. ФОРМИРОВАНИЕ ИСХОДНОЙ БАЗЫ ДАННЫХ В *SPSS*

2.1. СТРУКТУРА РЕДАКТОРА ДАННЫХ

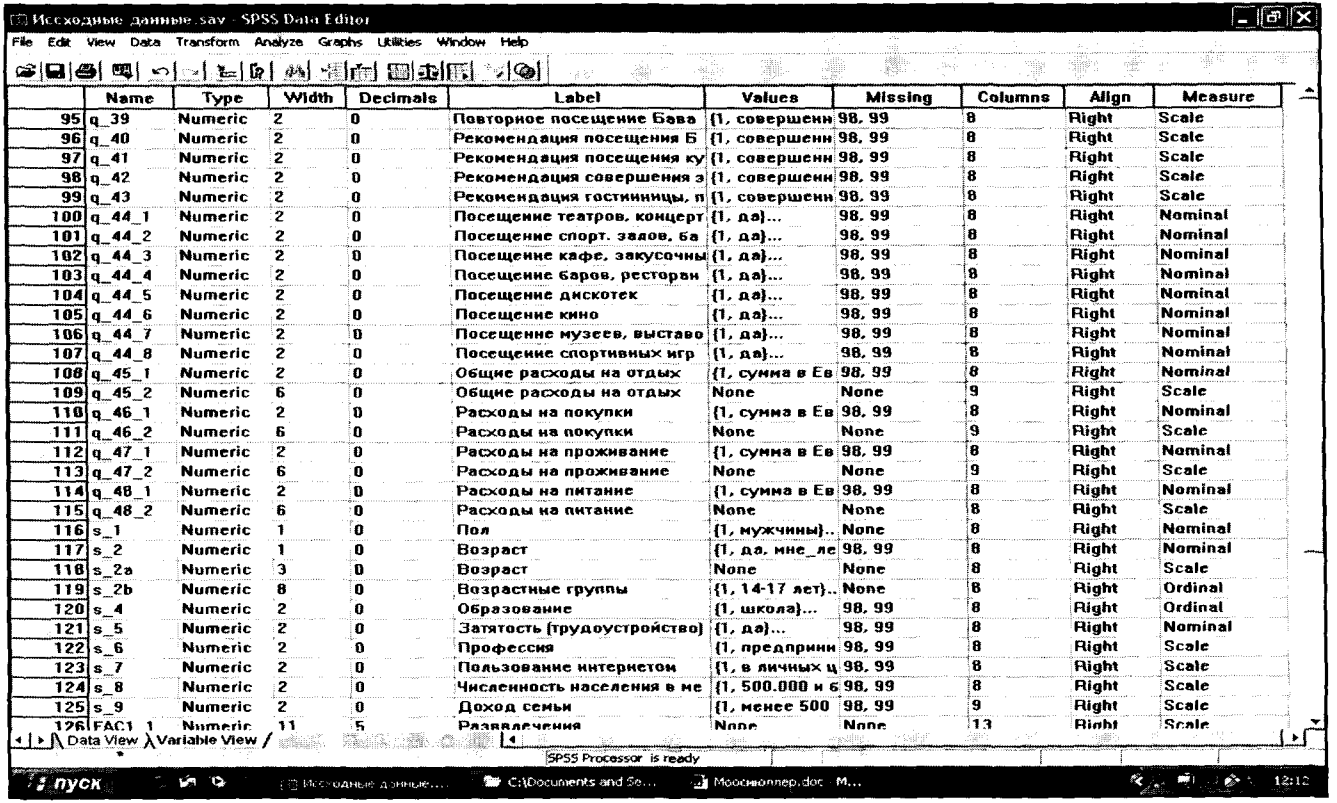
Файл исходной базы данных для проведения статистического анализа в *SPSS* формируется в редакторе данных (*Data Editor*). Редактор данных имеет две вкладки: «Свойства переменных» (*Variable View*) и «Значения переменных» (*Date View*). Данные вкладки представляют собой таблицы, содержащие информацию о данных, собранных для проведения анализа.

Во вкладке «*Variable View*» представлена таблица с данными, описывающими свойства переменных. Каждая строка отображает переменную (вопрос анкеты), каждый столбец – ее свойства (рис. 2.1).

В столбце «*Name*» таблицы «Свойства переменных» указывается имя переменной – как правило, это номер вопроса в анкете. Например, в базе данных, представленной в табл. 2.1, переменная «пол» имеет название «s_1», поскольку в разделе анкеты «социально-демографические признаки» вопрос о поле респондента находился на первом месте.

Имена переменных могут содержать буквы латинского алфавита и цифры, а также некоторые символы: @, $, _, #. В сумме число знаков не должно превышать «8». Не допускаются пробелы и буквы других алфавитов. Имя переменной должно начинаться с буквы и не может заканчиваться знаком подчеркивания «_».

В столбце «*Type*» таблицы «Свойства переменных» указывается тип переменной; новые, созданные, переменные по умолчанию являются числовыми (*Numeric*). Если требуется изменить тип переменной, следует подвести курсор в соответствующую ячейку таблицы, и при нажатии кнопки мыши на экране появится диалоговое окно «Тип переменной» (*Variable Type*) (рис. 2.2).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
95	q_39	Numeric	2	0	Повторное посещение Бава	{1, совершенн	98, 99	8	Right	Scale
96	q_40	Numeric	2	0	Рекомендация посещения Б	{1, совершенн	98, 99	8	Right	Scale
97	q_41	Numeric	2	0	Рекомендация посещения ку	{1, совершенн	98, 99	8	Right	Scale
98	q_42	Numeric	2	0	Рекомендация совершения з	{1, совершенн	98, 99	8	Right	Scale
99	q_43	Numeric	2	0	Рекомендация гостиницы, п	{1, совершенн	98, 99	8	Right	Scale
100	q_44_1	Numeric	2	0	Посещение театров, концерт	{1, да}...	98, 99	8	Right	Nominal
101	q_44_2	Numeric	2	0	Посещение спорт. залов, ба	{1, да}...	98, 99	8	Right	Nominal
102	q_44_3	Numeric	2	0	Посещение кафе, закусочны	{1, да}...	98, 99	8	Right	Nominal
103	q_44_4	Numeric	2	0	Посещение баров, ресторан	{1, да}...	98, 99	8	Right	Nominal
104	q_44_5	Numeric	2	0	Посещение дискотек	{1, да}...	98, 99	8	Right	Nominal
105	q_44_6	Numeric	2	0	Посещение кино	{1, да}...	98, 99	8	Right	Nominal
106	q_44_7	Numeric	2	0	Посещение музеев, выставо	{1, да}...	98, 99	8	Right	Nominal
107	q_44_8	Numeric	2	0	Посещение спортивных игр	{1, да}...	98, 99	8	Right	Nominal
108	q_45_1	Numeric	2	0	Общие расходы на отдых	{1, сумма в Ев	98, 99	8	Right	Nominal
109	q_45_2	Numeric	6	0	Общие расходы на отдых	None	None	9	Right	Scale
110	q_46_1	Numeric	2	0	Расходы на покупки	{1, сумма в Ев	98, 99	8	Right	Nominal
111	q_46_2	Numeric	6	0	Расходы на покупки	None	None	9	Right	Scale
112	q_47_1	Numeric	2	0	Расходы на проживание	{1, сумма в Ев	98, 99	8	Right	Nominal
113	q_47_2	Numeric	6	0	Расходы на проживание	None	None	9	Right	Scale
114	q_48_1	Numeric	2	0	Расходы на питание	{1, сумма в Ев	98, 99	8	Right	Nominal
115	q_48_2	Numeric	6	0	Расходы на питание	None	None	8	Right	Scale
116	s_1	Numeric	1	0	Пол	{1, мужчины}..	None	8	Right	Nominal
117	s_2	Numeric	1	0	Возраст	{1, да, мне_ле	98, 99	8	Right	Nominal
118	s_2a	Numeric	3	0	Возраст	None	None	8	Right	Scale
119	s_2b	Numeric	8	0	Возрастные группы	{1, 14-17 лет}..	None	8	Right	Ordinal
120	s_4	Numeric	2	0	Образование	{1, школа}...	98, 99	8	Right	Ordinal
121	s_5	Numeric	2	0	Затятость (трудоустройство)	{1, да}...	98, 99	8	Right	Nominal
122	s_6	Numeric	2	0	Профессия	{1, предприни	98, 99	8	Right	Scale
123	s_7	Numeric	2	0	Пользование интернетом	{1, в личных ц	98, 99	8	Right	Scale
124	s_8	Numeric	2	0	Численность населения в ме	{1, 500.000 и б	98, 99	8	Right	Scale
125	s_9	Numeric	2	0	Доход семьи	{1, менее 500	98, 99	9	Right	Scale
126	FAC1_1	Numeric	11	5	Развлечения	None	None	13	Right	Scale

Рис. 2.1. Редактор данных: вкладка «Свойства переменных» (*Variable View*)

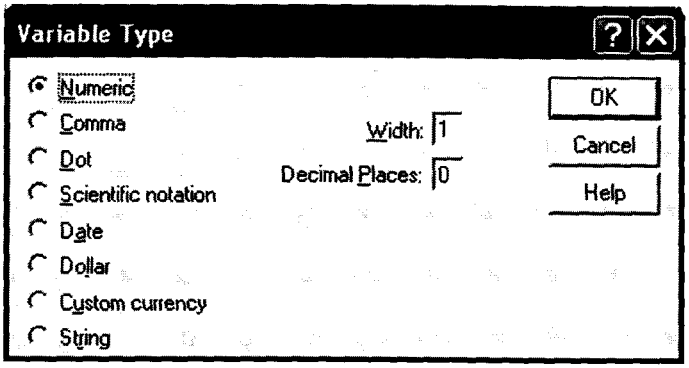


Рис. 2.2. Диалоговое окно «Тип переменной»

В диалоговом окне «Тип переменной» возможен выбор формата записи значений переменной:

Comma (например: 43,675.67);

Dot (например: 43.675,67);

Scientific notation (например: 43E+0,4);

Dollar (например: $43,675).

Аналогичным образом можно выбрать текстовую переменную (*String*). Однако применение текстовых переменных в *SPSS* практически невозможно, поскольку с ними нельзя производить никаких арифметических операций и рассчитывать какие-либо статистические показатели.

В поле «*Width*» диалогового окна «Тип переменной» (см. рис. 2.2) указывается число знаков, используемых для кодировки переменной. Например, для кодировки переменной «пол» используется только один знак («1» – «мужчины» или «2» – «женщины»).

Число знаков, используемых для кодировки переменной, можно также указать в столбце «*Width*» («Формат столбца») таблицы «Свойства переменных» (см. табл. 2.1).

В поле «*Decimal Places*» диалогового окна «Тип переменной» указывается число знаков после запятой при записи значений переменной. Например, для переменной «пол» в поле «*Decimal Places*» указывается значение 0. Ответы респондентов в данном случае заносятся в базу данных в виде целых чисел («1» – «мужчины» или «2» – «женщины»).

Число знаков после запятой при записи значений переменной можно также указать в столбце «*Decimals*» («Десятичные разряды») таблицы «Свойства переменных».

В столбце «*Label*» таблицы «Свойства переменных» указываются метки переменных. Метка – название, позволяющее описать переменную более подробно, чем имя переменной, она может содержать до 256 символов. В качестве этих символов могут выступать также буквы русского алфавита.

При задании меток переменных часто используются формулировки вопросов, содержащихся в анкете. Например, в качестве метки переменной «пол» в редакторе данных может быть введена фраза: «Укажите, пожалуйста, свой пол». Однако следует помнить, что метка переменной будет отображаться во всех графиках и таблицах, представляющих результаты статистического анализа. Поэтому рекомендуется использовать более лаконичные метки для наглядности представления результатов анализа.

В столбце «*Values*» таблицы «Свойства переменных» (см. рис. 2.1) отображаются значения меток переменных. Если в поле «*Label*» указывается вопрос анкеты, то в поле «*Values*» указываются коды возможных вариантов ответа на этот вопрос.

Для заполнения поля «*Values*» необходимо произвести кодировку вариантов ответа. При подведении курсора к соответствующей ячейке таблицы и нажатии клавиши мыши на экране компьютера появляется диалоговое окно «Значение меток переменных» (*Value Labels*) (рис. 2.3). В диалоговом окне «Значение меток переменных» в поле «*Value*» указываются числовые коды вариантов ответа, а в поле «*Value Label*» – их вербальные формулировки.

При задании вербальных формулировок следует учитывать, что они будут фигурировать впоследствии в графиках и аналити-

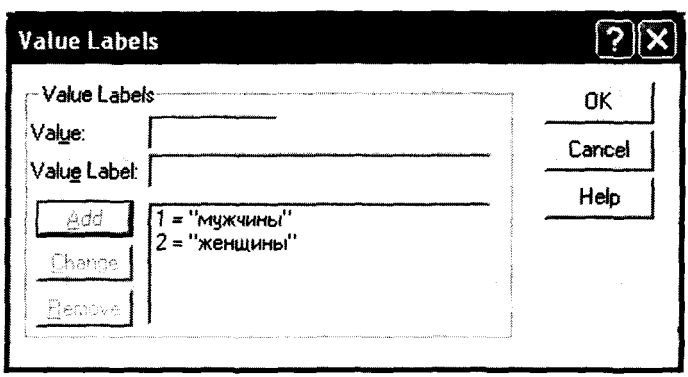


Рис. 2.3. Диалоговое окно «Значение меток переменных»

ческих таблицах. Например, ответ на вопрос о половой принадлежности респондента должен быть не «мужской» («женский»), а «мужчины» («женщины»).

Процедура кодировки производится поэтапно по каждому варианту ответа. В рассматриваемом примере кодировки переменной «пол», сначала в поле «*Value*» указывается числовой код «1», а в поле «*Value Label*» – вербальный вариант ответа «мужчины». После нажатия кнопки «*Add*» эти данные переносятся в большое поле диалогового окна «Значение меток переменных». Затем подобным образом кодируется вариант ответа «женщины». После нажатия кнопки «*OK*» диалоговое окно «Значение меток переменных» закрывается, а указанные в нем данные заносятся в столбец «*Values*» таблицы «Свойства переменных».

В столбце «*Missing*» («Пропущенные значения») рис. 2.1 «Свойства переменных» следует указать, какие коды вариантов ответов следует исключить из анализа.

В *SPSS* допускаются два вида пропущенных значений:

- Пропущенные значения, определяемые системой (*System-defined missing values*). Если в матрице данных есть незаполненные ячейки, система *SPSS* самостоятельно идентифицирует их как пропущенные значения. Отсутствие ответа отражается в исходном файле данных в виде запятой.
- Пропущенные ответы, задаваемые пользователем (*User-defined missing values*). Например, среди вариантов ответа на поставленный вопрос можно закодировать отсутствие определенного ответа («98» – «не знаю», «99» – «нет данных») и затем в поле «*Missing*» указать эти коды, чтобы исключить соответствующие варианты ответа из анализируемых данных.

При подведении курсора к соответствующей ячейке столбца «*Missing*» и нажатии кнопки мыши открывается диалоговое окно «Пропущенные значения» (рис. 2.4).

По умолчанию в диалоговом окне «Пропущенные значения» отмечается команда «*No missing values*». Это означает, что пропущенных значений нет, а все варианты ответа на вопрос рассматриваются как допустимые.

Если бы нужно было указать коды вариантов ответа, исключаемых из процедуры анализа, то следовало бы выбрать команду «*Discrete missing values*» и в соответствующих ячейках указать коды «98» и «99» («98» – «не знаю», «99» – «нет данных»). Для одной переменной можно задать до трех пропущенных значений.

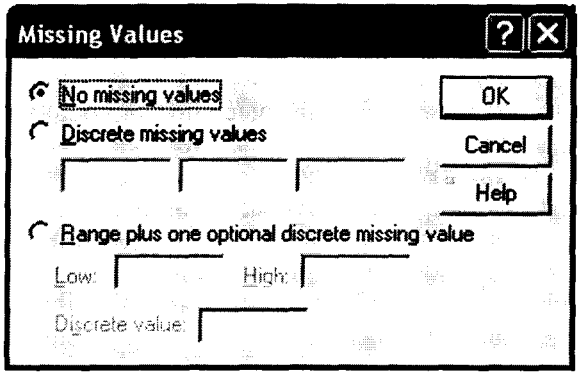


Рис. 2.4. Диалоговое окно «Пропущенные значения»

Существует еще один вариант задания пропущенных значений: «*Range plus one optional discrete missing value*» («Диапазон плюс единичное пропущенное значение»). Эта команда применялась бы в случае, если бы, например, при заданных значениях переменной «возраст» нужно было бы исключить из исследований респондентов от 20 до 40 лет, а также лиц в возрасте 55 лет.

В рассматриваемом примере описания свойств переменной «пол» достаточно сложно представить, чтобы кто-то из респондентов затруднился ответить или не захотел отвечать на вопрос о своей половой принадлежности. Поэтому в поле «*Missing*» таблицы «Свойства переменных» отсутствуют какие-либо коды вариантов ответа.

В столбце «*Columns*» («Столбцы») таблицы «Свойства переменных» указывается ширина столбца, содержащего значения соответствующей переменной в таблице другой вкладки редактора данных: «Значения переменных» (*Date View*) (рис. 2.5). По умолчанию ширина столбца задается «8».

В столбце «*Alignment*» («Выравнивание») таблицы «Свойства переменных» задается положение кодов ответов в таблице «Значения переменных» во вкладке редактора данных «*Date View*». Они могут быть выровнены по правому краю (*Right*), по левому краю (*Left*) или по центру (*Center*). По умолчанию задается выравнивание по правому краю. Если нужно изменить порядок выравнивания, то следует подвести курсор к соответствующей ячейке столбца «*Alignment*», и при нажатии клавиши мыши на экране появится меню, содержащее три вышеперечисленных варианта выравнивания данных, из которых следует выбрать желаемый.

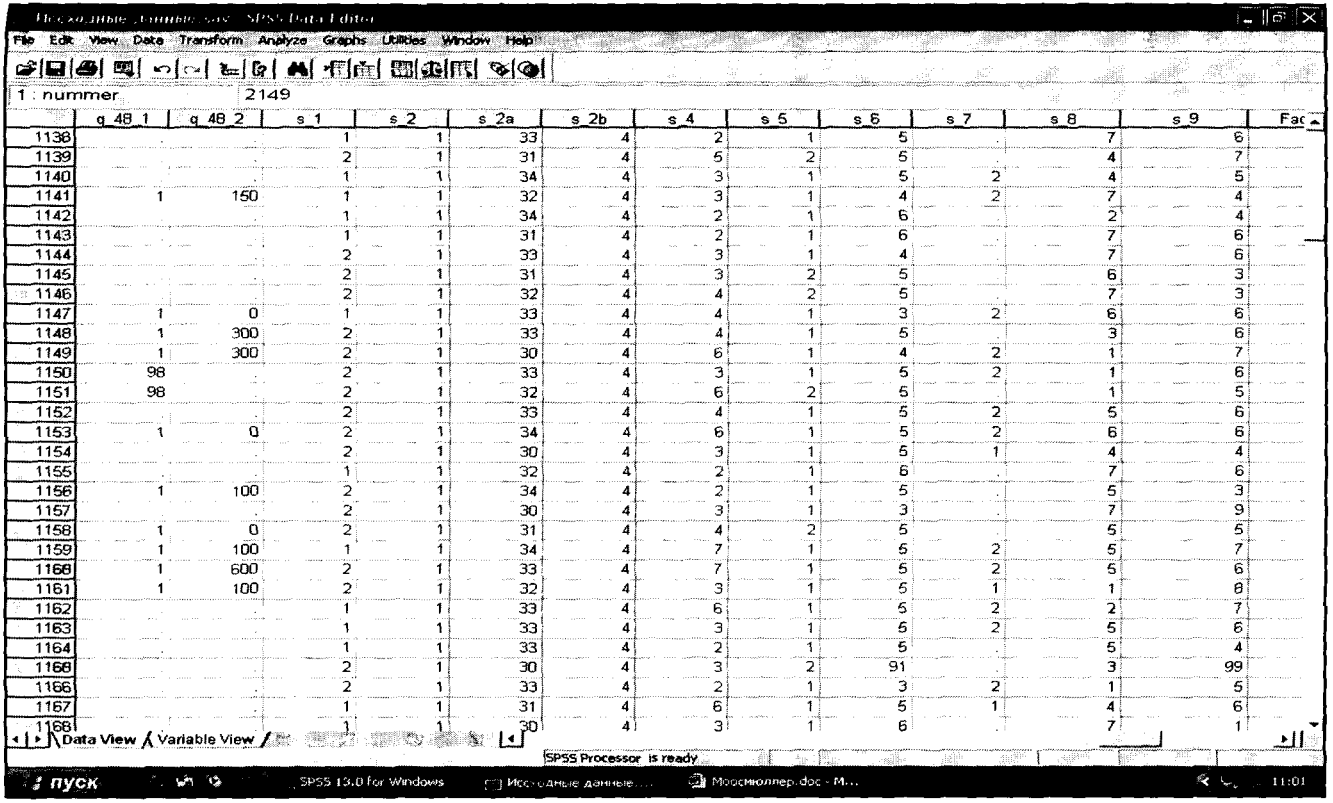


Рис. 2.5. Редактор данных: вкладка «Значения переменных» (*Data View*)

В столбце «*Measure*» («Шкала измерения») таблицы «Свойства переменных» указывается тип шкалы, по которой измеряется переменная. По умолчанию задается метрическая шкала (*Scale*). В случае необходимости тип шкалы можно изменить. Для этого следует подвести курсор в соответствующую ячейку столбца «*Measure*» и нажать клавишу мыши, после чего на экране появится меню из трех типов шкалы измерения (рис. 2.6).

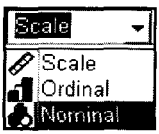


Рис. 2.6. Меню выбора типа шкалы измерения переменной

В зависимости от вида переменной следует выбрать один из трех типов шкалы измерения: метрическую (*Scale*), порядковую (*Ordinal*) или номинальную (*Nominal*). Поскольку переменная «пол» измеряется по номинальной шкале, то при заполнении таблицы «Свойства переменных» в строке этой переменной в столбце «*Measure*» выбирается тип шкалы измерения «*Nominal*». (Более подробно этот вопрос будет рассмотрен в п. 2.3 «Типы шкал измерения переменных».) После того как заполнена таблица «Свойства переменных» во вкладке редактора данных «*Variable View*», следует открыть другую вкладку редактора данных – «*Date View*».

Во вкладке редактора данных «*Date View*» представлена таблица с данными, описывающими значения переменных. Каждый столбец отображает переменную (вопрос анкеты), каждая строка – отдельное наблюдение (объект сбора информации) (см. рис. 2.5). В качестве объектов сбора информации могут выступать люди, предприятия, продукты, бренды и т.д.

На рис. 2.5 представлен фрагмент таблицы, содержащей значения переменных, описанных в таблице «Свойства переменных».

Из данных таблицы «Свойства переменных» (см. табл. 2.1) известно, что переменная с именем «s_1» имеет метку «Пол». Метка переменной «пол» имеет два значения: «мужчины» (код «1») и женщины (код «2»). В столбце «s_1» таблицы «Значения переменных» (см. рис. 2.5) содержатся закодированные ответы респондентов на вопрос об их половой принадлежности: «1» или «2». Так, по дан-

ным этой таблицы известно, что респондент в строке 1143 – мужчина, а респондент в строке 1144 – женщина.

Из данных таблицы «Свойства переменных» также известно, что переменная с именем «s_1a» имеет метку «Возраст». Эта переменная не имеет кодировки (в столбце «*Values*» отсутствуют значения меток переменных). В столбце «s_1a» таблицы «Значения переменных» содержатся незакодированные ответы респондентов на вопрос об их возрасте. Так, по данным этой таблицы известно, что респондент в строке 1143 – мужчина в возрасте 31 года, а респондент в строке 1144 – женщина в возрасте 33 лет.

2.2. ВИДЫ КОДИРОВКИ

В *SPSS* существует два основных вида кодировки данных: категориальная и дихотомическая. ***Категориальная кодировка*** предполагает несколько вариантов ответа на поставленный вопрос, т.е. метка переменной может принимать несколько значений. ***Дихотомическая кодировка*** предполагает только два варианта ответа на поставленный вопрос, т.е. метка переменной может принимать только два значения («да» или «нет»).

Вид кодировки переменных определяется типом вопроса анкеты. Вопросы бывают ***открытые*** (без заданных вариантов ответа) и ***закрытые*** (с заданными вариантами ответа). Закрытые вопросы, в свою очередь, бывают одновариантные (альтернативные) и многовариантные (безальтернативные).

Одновариантные (*альтернативные*) вопросы предполагают возможность выбора только одного из предложенных вариантов ответа. ***Многовариантные*** (*безальтернативные*) вопросы предоставляют возможность выбрать несколько из предложенных вариантов ответа.

Основное правило создания исходного файла данных в *SPSS* состоит в том, что создаваемые переменные должны быть одновариантными, т.е. одна переменная должна иметь одну метку. В этой связи при занесении в файл *SPSS* данных по ответам на один многовариантный вопрос создается несколько одновариантных переменных.

При занесении в файл *SPSS* данных по одновариантному вопросу создается одна переменная, имеющая одну метку. Метка создаваемой переменной может иметь несколько значений. В этом случае применяется категориальная кодировка данных (табл. 2.1).

Таблица 2.1

Категориальная кодировка данных.
Вопрос анкеты: «Какой продукт Вы покупаете?»

Респонденты	Значения метки переменной «Приобретаемый продукт»: Продукт *A* – «1», Продукт *B* – «2», Продукт *C* – «3»
Респондент 1	2
Респондент 2	1
Респондент 3	3
Респондент 4	1
...	...

При использовании категориальной кодировки данных все респонденты, участвующие в исследовании, могут быть поделены на категории относительно выбранного ими варианта ответа. Например, относительно приобретаемого продукта все респонденты могут быть поделены на три категории: «приобретающие продукт *A*», «приобретающие продукт *B*» и «приобретающие продукт *C*».

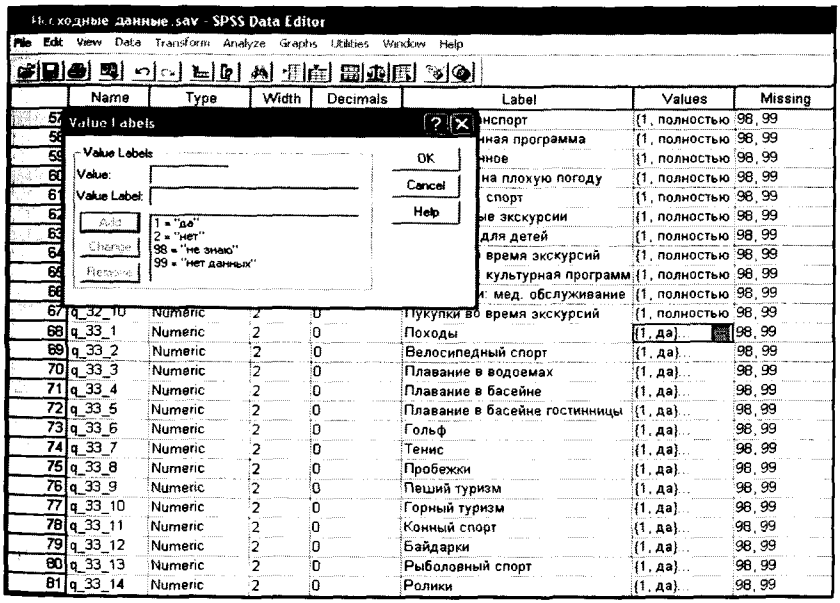
При занесении в файл *SPSS* данных по многовариантному вопросу создается несколько переменных, каждая переменная имеет свою метку. Метки создаваемых переменных могут иметь только два значения. В этом случае применяется дихотомическая кодировка данных (табл. 2.2).

Таблица 2.2

Дихотомическая кодировка данных.
Вопрос анкеты: «Какой продукт Вы покупаете?»

Респонденты	Переменные (значения меток переменных: покупаю – «1», не покупаю – «0»)		
	Продукт *A*	Продукт *B*	Продукт *C*
Респондент 1	1	1	0
Респондент 2	0	1	1
Респондент 3	1	0	1
Респондент 4	1	1	1
...			

В рис. 2.7 и 2.8 представлены фрагменты вкладок редактора данных «*Variable View*» и «*Data View*», которые иллюстрируют представление в исходном файле данных *SPSS* одного многовариантного вопроса в виде нескольких одновариантных переменных.



	Name	Type	Width	Decimals	Label	Values	Missing
57					нспорт	{1, полностью	98, 99
58					нная программа	{1, полностью	98, 99
59					нное	{1, полностью	98, 99
60					на плохую погоду	{1, полностью	98, 99
61					спорт	{1, полностью	98, 99
62					ые экскурсии	{1, полностью	98, 99
63					для детей	{1, полностью	98, 99
64					время экскурсий	{1, полностью	98, 99
65					культурная программ	{1, полностью	98, 99
66					мед. обслуживание	{1, полностью	98, 99
67	q_32_10	Numeric	2	0	Покупки во время экскурсий	{1, полностью	98, 99
68	q_33_1	Numeric	2	0	Походы	{1, да}	98, 99
69	q_33_2	Numeric	2	0	Велосипедный спорт	{1, да}	98, 99
70	q_33_3	Numeric	2	0	Плавание в водоемах	{1, да}	98, 99
71	q_33_4	Numeric	2	0	Плавание в басейне	{1, да}	98, 99
72	q_33_5	Numeric	2	0	Плавание в басейне гостинницы	{1, да}	98, 99
73	q_33_6	Numeric	2	0	Гольф	{1, да}	98, 99
74	q_33_7	Numeric	2	0	Тенис	{1, да}	98, 99
75	q_33_8	Numeric	2	0	Пробежки	{1, да}	98, 99
76	q_33_9	Numeric	2	0	Пеший туризм	{1, да}	98, 99
77	q_33_10	Numeric	2	0	Горный туризм	{1, да}	98, 99
78	q_33_11	Numeric	2	0	Конный спорт	{1, да}	98, 99
79	q_33_12	Numeric	2	0	Байдарки	{1, да}	98, 99
80	q_33_13	Numeric	2	0	Рыболовный спорт	{1, да}	98, 99
81	q_33_14	Numeric	2	0	Ролики	{1, да}	98, 99

Рис. 2.7. Фрагмент вкладки «*Variable View*». Дихотомическая кодировка ответов на многовариантный вопрос

В рассматриваемом примере (см. рис. 2.7, 2.8) представлены данные о туристах, отдыхающих в курортной зоне Восточной Баварии «Баварский лес». Эти данные содержат ответы на вопрос анкеты № 33: «Каким спортом Вы занимаетесь на отдыхе?». В качестве возможных вариантов ответа на этот вопрос предлагаются 14 видов спортивных занятий, наиболее популярных в Восточной Баварии: «туристические походы», «велосипедный спорт», «плавание в природных водоемах», «плавание в бассейне» и т.д. Респонденты могут выбрать несколько из предлагаемых вариантов ответов.

Как видно из данных, представленных в табл. 2.5, вопрос анкеты № 33 представлен в файле данных *SPSS* в виде 14 одновариантных переменных с именами «q_33_1», «q_33_2» ... «q_33_14». Каждая переменная имеет собственную метку, обозначающую вид спортивного занятия: «Походы», «Велосипедный спорт» ... «Ролики». Все 14 переменных имеют дихотомическую кодировку ответов.

Каждая метка переменной имеет четыре значения (см. рис. 2.7), два их которых («98» – «не знаю», «99» – «нет данных») исключа-

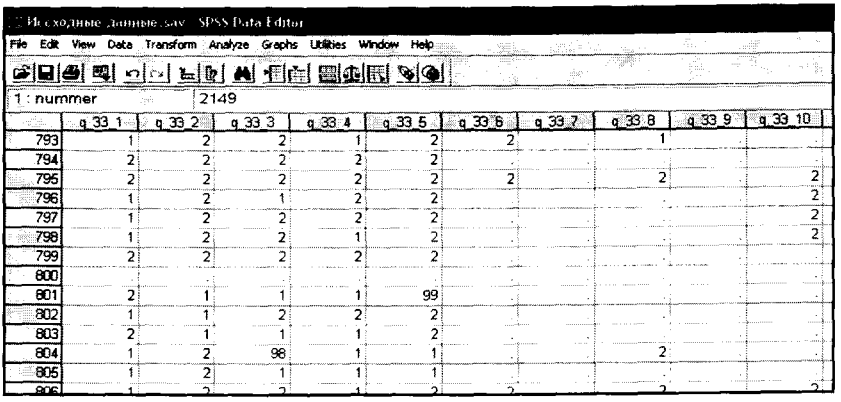


Рис. 2.8. Фрагмент вкладки «*Data View*». Дихотомическая кодировка ответов на многовариантный вопрос

ются. Коды этих значений меток переменных указаны в столбце «*Missing*» как пропущенные данные. Значимыми для проведения исследования являются только два варианта ответа: «да» и «нет», которые указывают на то, занимается ли респондент соответствующим видом спорта.

По данным фрагмента вкладки редактора данных «*Data View*», представленного на рис. 2.8, респондент в строке 793 совершает туристические походы, плавает только в бассейне и делает пробежки, поскольку переменные «q_33_1», «q_334» и «q_33_8» имеют положительный ответ («1» – «да»).

Среди закодированных вариантов ответов встречаются цифры «98» и «99». Они обозначают, что респондент затруднился или не захотел отвечать на поставленный вопрос (см. рис. 2.8). Этим данным присваивается значение «*user missing*». Они исключаются из исследований по заданным условиям (столбец «*Missing*», рис. 2.7).

Некоторые ячейки таблицы, представленной на рис. 2.8, оказались пустыми. Данным вариантам ответа присваивается значение «*system missing*». Это данные, которые должны были присутствовать, но их не оказалось в базе данных в связи с причинами случайного характера. Такие данные автоматически исключаются из исследований.

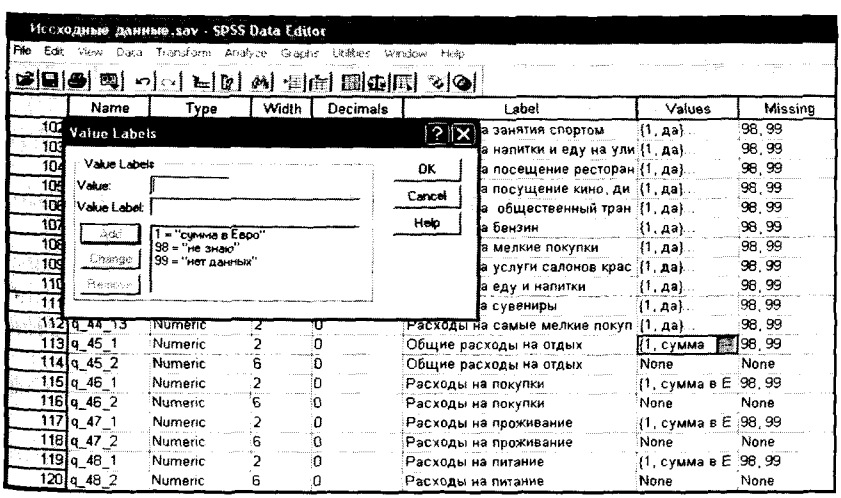
Обобщая вышеизложенное, следует еще раз отметить, что многовариантные вопросы представляются в файле данных *SPSS* в виде нескольких переменных, каждая из которых представляет

собой возможный вариант ответа. Однако возможны случаи, когда одновариантный вопрос представляется в виде двух переменных в файле данных *SPSS*. Такое возможно в случае применения так называемой двойной записи.

Применение двойной записи проиллюстрировано на рис. 2.9 и 2.10. В этих таблицах представлены фрагменты файла данных *SPSS*, содержащего информацию о туристах, отдыхающих в курортной зоне Восточной Баварии «Баварский лес». Ряд вопросов анкеты представлен в файле данных в виде двух переменных:

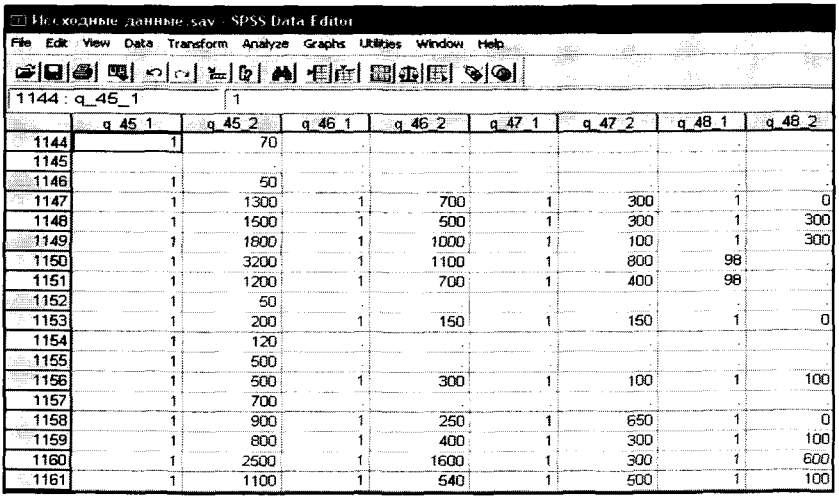
- вопрос № 45: «Какую сумму Вы тратите в целом на отдых?»;
- вопрос № 46: «Какую сумму Вы тратите на крупные покупки (одежда, обувь, спортивный инвентарь и т.п.) во время отдыха?»;
- вопрос № 47: «Какую сумму Вы тратите на проживание в гостинице/пансионе (включая обслуживание) во время отдыха?»;
- вопрос № 48: «Какую сумму Вы тратите на питание (посещение кафе и ресторанов, покупки продуктов в магазине) во время отдыха?».

Вопрос анкеты № 45 представлен в файле данных в виде двух переменных с именами «q_45_1» и «q_45_2» (см. рис. 2.9). Обе переменные имеют одинаковую метку «Общие расходы на отдых». Переменная с именем «q_45_1» имеет три значения метки



	Name	Type	Width	Decimals	Label	Values	Missing
102					а занятия спортом	{1, да}	98, 99
103					а напитки и еду на ули	{1, да}	98, 99
104					а посещение ресторан	{1, да}	98, 99
105					а посущение кино, ди	{1, да}	98, 99
106					а общественный тран	{1, да}	98, 99
107					а бензин	{1, да}	98, 99
108					а мелкие покупки	{1, да}	98, 99
109					а услуги салонов крас	{1, да}	98, 99
110					а еду и напитки	{1, да}	98, 99
111					а сувениры	{1, да}	98, 99
112	q_44_13	Numeric	2	0	Расходы на самые мелкие покуп	{1, да}	98, 99
113	q_45_1	Numeric	2	0	Общие расходы на отдых	{1, сумма	98, 99
114	q_45_2	Numeric	6	0	Общие расходы на отдых	None	None
115	q_46_1	Numeric	2	0	Расходы на покупки	{1, сумма в Е	98, 99
116	q_46_2	Numeric	6	0	Расходы на покупки	None	None
117	q_47_1	Numeric	2	0	Расходы на проживание	{1, сумма в Е	98, 99
118	q_47_2	Numeric	6	0	Расходы на проживание	None	None
119	q_48_1	Numeric	2	0	Расходы на питание	{1, сумма в Е	98, 99
120	q_48_2	Numeric	6	0	Расходы на питание	None	None

Рис. 2.9. Фрагмент вкладки «*Variable View*». Двойная запись данных



	q_45_1	q_45_2	q_46_1	q_46_2	q_47_1	q_47_2	q_48_1	q_48_2
1144	1	70	.	.	.	.	.	.
1145	.	.	.	.	.	.	.	.
1146	1	50	.	.	.	.	.	.
1147	1	1300	1	700	1	300	1	0
1148	1	1500	1	500	1	300	1	300
1149	1	1800	1	1000	1	100	1	300
1150	1	3200	1	1100	1	800	98	.
1151	1	1200	1	700	1	400	98	.
1152	1	50	.	.	.	.	.	.
1153	1	200	1	150	1	150	1	0
1154	1	120	.	.	.	.	.	.
1155	1	500	.	.	.	.	.	.
1156	1	500	1	300	1	100	1	100
1157	1	700	.	.	.	.	.	.
1158	1	900	1	250	1	650	1	0
1159	1	800	1	400	1	300	1	100
1160	1	2500	1	1600	1	300	1	600
1161	1	1100	1	540	1	500	1	100

Рис. 2.10. Фрагмент вкладки «*Data View*». Двойная запись данных

переменной: «1» – «сумма в евро», «98» – «не знаю», «99» – нет данных. Переменная с именем «q_45_2» не имеет значений метки переменной (в столбце «*Label*» стоит отметка «*None*») (см. рис. 2.7).

На рис. 2.10 представлен фрагмент файла данных *SPSS* во вкладке редактора данных «*Data View*». В столбцах таблицы «Значения переменных» (*Data View*) представлены ответы респондентов на вопросы анкеты № 45, 46, 47 и 48. Например, если в определенной строке в столбце «q_45_1» стоит числовой код «1», это означает, что респондент назвал определенную сумму, которую он тратит на проведение отдыха. В этой же строке в столбце «q_45_2» указывается названная сумма.

Двойная запись данных применяется в том случае, если необходима обработка очень большого объема информации. В рассматриваемом примере исходный файл данных *SPSS* был создан на базе проведения опроса туристов, отдыхающих в курортной зоне «Баварский лес», в ходе которого было опрошено 6396 респондентов. Каждый из респондентов в качестве ответа о размерах своих расходов во время проведения отпуска мог назвать шестизначное число (см. рис. 2.9: в столбце «*Width*» отметка «6»).

Порядок кодировки переменных определяется типом шкалы их измерения. Типы шкал измерения переменных подробно рассматриваются в следующем разделе.

2.3. ТИПЫ ШКАЛ ИЗМЕРЕНИЯ ПЕРЕМЕННЫХ

Для работы с данными в *SPSS* важно знать, по шкале какого типа измеряются исследуемые переменные. Это необходимо для выбора метода анализа данных и определения возможности расчета статистических показателей (табл. 2.3).

Существует четыре типа шкал измерения переменных:

1) номинальная шкала;

2) порядковая шкала;

3) интервальная шкала;

4) относительная шкала.

Таблица 2.3

Примеры переменных, измеряемых по шкалам разных типов

Шкала	Переменная	Значения переменной
Номинальная	Пол (дихотомическая переменная)	• «1» = мужской • «2» = женский
	Производитель продукта «*X*»	• «1» = производитель *A* • «2» =производитель *B* • «3» =производитель *C*
Порядковая	Класс полета	• «1» = первый класс • «2» = бизнес-класс • «3» =эконом-класс
	Категории потребителей по уровню дохода	• «1» = до 1000 евро • «2» = от 1001 до 3000 евро • «3» = свыше 3000 евро
Интервальная	Коэффициент интеллекта (*IQ*)	...«120»...
Относительная	Уровень дохода	... «2100» евро ...

Номинальная шкала характеризуется самым низким уровнем измерения переменных. Все значения переменной, измеряемой по номинальной шкале, находятся на одном уровне. По этой шкале измеряются, как правило, качественные характеристики объекта исследования. Между значениями переменной, измеряемой по номинальной шкале, не существует логического порядка. Например, в качестве ответа на вопрос анкеты: «Какого производителя продукта «*X*» вы предпочитаете?» – может быть предложено несколько вариантов: «Производитель *A*», «Производитель *B*», «Производитель *C*» и т.д. В этом случае, с точки зрения исследователей, все предложенные производители являются рав-

нозначными. Числовые коды («1», «2», «3»...) могут присваиваться значениям метки переменной в любом порядке.

Переменные, измеряемые по номинальной шкале и имеющие всего два значения (например, «мужчины» и «женщины»), называются ***дихотомическими.***

Порядковая шкала является второй по уровню измерения переменных. Значения переменной, измеряемой по порядковой шкале, не являются равнозначными, они находятся на разных уровнях по отношению друг к другу и подчиняются логическому числовому порядку.

Порядковая шкала характеризуется низким уровнем измерения переменных, поскольку является шкалой с неравными интервальными отрезками. Совершенно четко можно утверждать, что уровень обслуживания авиапассажиров первого класса выше, чем бизнес-класса, но насколько именно, неизвестно. Также разница в обслуживании между первым и бизнес-классом, между бизнес- и эконом-классом может быть различной (см. табл. 2.3).

Низкий уровень измерения переменных по порядковой шкале можно проиллюстрировать на примере переменной «Категории потребителей по уровню дохода». Потребители примерно с одинаковым уровнем дохода (например, 950 и 1050 евро) оказываются в разных категориях, а потребители с существенной разницей по уровню дохода (например, 1050 и 2950 евро) оказываются в одной категории.

Интервальная шкала является третьей по уровню измерения переменных. В отличие от порядковой шкалы она является шкалой с равными интервальными отрезками. Это позволяет осуществлять количественное сравнение значений переменной, т.е. можно определить, насколько одно значение больше или меньше (лучше или хуже, длиннее или короче и т.д.) другого.

Характерной чертой интервальной шкалы является отсутствие «естественного нуля», т.е. исходная точка измерения является относительной. Примерами интервальной шкалы являются шкала Цельсия и календарь. По шкале Цельсия за «0» принята температура замерзания воды, однако за «0» можно было принять любую другую температуру. Существуют также различные календари с одинаковым количеством дней в году, но разным временем начала года.

В маркетинговых исследованиях очень часто используется рейтинговая шкала, когда респондентам предлагается оценить

по балльной шкале (например, от 1 до 7 баллов) утверждение, продукт, бренд и т.п. Строго говоря, рейтинговая шкала является порядковой, поскольку балльные оценки субъективны. Одинаковые балльные оценки в действительности отображают разный уровень измеряемой переменной. Например, студенты, получившие одинаковые оценки на экзамене, в действительности могут иметь разный уровень знаний.

Очень часто при проведении исследований шкала балльных оценок рассматривается как интервальная. В основе этого лежит предположение, что интервальные отрезки шкалы балльных оценок одинаковы. Это дает возможность рассчитать среднее значение переменной (например, средний балл успеваемости студентов). Расчет средней величины (среднеарифметической) для показателя, измеряемого по порядковой шкале, невозможен. Например, не существует показателя «средний класс» полета (см. табл. 2.3).

Относительная шкала характеризуется самым высоким уровнем измерения переменных. Ее основное отличие от интервальной шкалы заключается в существовании «естественного нуля», который можно интерпретировать как отсутствие значения переменной. Например, если заработная плата равна нулю, это значит, что ее не выплачивают.

По относительной шкале измеряются количественные характеристики. Это могут быть как физические характеристики (объем, вес, скорость и пр.), так и экономические характеристики (доход, издержки, цена и пр.).

Относительная *шкала получила* свое название благодаря возможности сравнения значений переменной по отношению друг к другу, что невозможно при использовании интервальной шкалы измерения. Например, нельзя сказать, что человек, у которого коэффициент интеллекта (*IQ*) равен 160, в два раза умнее человека у которого этот показатель составляет 80. Но можно сказать, что заработная плата 1000 евро в два раза больше заработной платы 2000 евро.

При выборе типа шкалы измерения переменных в *SPSS* (столбец «*Measure*» во вкладке редактора данных «*Variable View*») интервальная шкала и шкала отношений объединяются в один вид – метрическую шкалу (*Scale*).

При построении в *SPSS* интерактивных графиков номинальная (*Nominal*) и порядковая (*Ordinal*) шкалы объединяются в «категориальный» тип (табл. 2.4).

Таблица 2.4

Типы шкал измерения переменных

Шкала		Характеристики
Категориальная	Номинальная (*Nominal*)	Служит для классификации качественных показателей. Все значения измеряемой переменной равнозначны
	Порядковая (*Ordinal*)	Служит для построения значений измеряемой переменной в определенной последовательности. Шкала с неравными интервальными отрезками
Метрическая (Scale)	Интервальная	Шкала с равными интервальными отрезками и условной точкой отсчета
	Относительная	Шкала с равными интервальными отрезками и безусловной точкой отсчета

Чем выше уровень измерения переменной, тем богаче ее информационная содержательность и тем больше возможностей осуществления расчетов и определения статистических показателей.

Числовые коды («1», «2», «3»...) значений метки переменной, измеряемой по номинальной или порядковой шкале, не могут рассматриваться как числа, они представляют собой лишь некие числовые символы. Поскольку они не являются числами, с ними нельзя производить никаких арифметических операций (сложение, вычитание, деление, умножение).

Что касается статистических показателей, характеризующих распределение величины, измеряемой по номинальной шкале, можно провести частотный анализ (*Frequencies*) и определить моду (*Mode*). Частоты показывают, например, сколько респондентов предпочитают того или иного производителя продукта «*X*». Мода обозначает самую многочисленную группу респондентов, предпочитающих определенного производителя продукта «*X*».

Для переменных, измеряемых по порядковой шкале, кроме вышеуказанных статистических показателей можно определить медиану и средневзвешенное. Значения меток переменной, измеряемой по интервальной шкале, рассматриваются как числа. С ними можно производить такие арифметические операции, как сложение и вычитание.

Что касается возможности расчета статистических показателей, характеризующих распределение переменной, измеряемой по интервальной шкале, кроме моды и медианы можно также определить стандартное отклонение (*Std. deviation*) и среднеариф-

метическое (*Mean*). (Средневзвешенное значение переменных с интервальной шкалой равно среднему арифметическому.)

При расчете статистических показателей, характеризующих распределение переменной, измеряемой по интервальной шкале, не рассчитывается такой показатель, как сумма (*Sum*). Например, не рассчитывается «суммарный коэффициент интеллекта» для группы студентов, такого показателя не существует.

Значения меток переменной, измеряемой по шкале отношений, выражаются в числах, с ними можно производить любые арифметические операции. Также можно определять любые статистические показатели, характеризующие распределение переменной.

Возможна трансформация имеющихся данных, измеряемых по шкале более высокого уровня, в данные, измеряемые по шкале более низкого уровня, но не наоборот. Например, значения переменной «Уровень дохода», измеряемой по относительной шкале, можно трансформировать в значения переменной «Категории потребителей по уровню дохода», измеряемой по порядковой шкале (см. табл. 2.3). Подобная трансформация данных, производимая в целях упрощения процедуры анализа и наглядности представления результатов, неизбежно связана с частичной потерей информации и снижением точности расчетов.

На практике, в том числе при применении *SPSS*, различие между переменными, измеряемыми по интервальной и относительной шкалам, обычно несущественно.

Во многих учебниках по SPSS метрические переменные (Scale) определяются как интервальные.

Тип шкалы измерения переменных определяет возможность применения того или иного метода анализа данных. Все методы статистического анализа делятся на две группы:

1) методы оценки связи между переменными;
2) методы выявления структуры данных.

Методы выявления структуры данных характеризуются тем, что исходные данные для проведения анализа не содержат информации (предположений) о существовании взаимосвязей между исследуемыми переменными. К таким методам относятся, например, кластерный и факторный анализ.

Методы оценки связи между переменными устанавливают влияние одной или нескольких независимых переменных на одну или несколько зависимых переменных. С точки зрения теории статистики существуют правила применения того или иного ме-

тода оценки связи между переменными в зависимости от типа шкалы их измерения (табл. 2.5).

Таблица 2.5

Методы оценки связи между переменными и типы шкал измерения переменных
(Backhaus, Erichson, Plinke, Weiber, 2000.S. XXII)

		Независимые переменные	
		Метрическая шкала	Номинальная шкала
Зависимые переменные	Метрическая шкала	Регрессионный анализ	Дисперсионный анализ
	Номинальная шкала	Дискриминантный анализ	Таблицы сопряженности

Применение некоторых основных методов статистического анализа в *SPSS* будет более подробно рассмотрено в следующих подразделах.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что представляют собой таблицы, содержащиеся во вкладках редактора данных *SPSS* «Свойства переменных» (*Variable View*) и «Значения переменных» (*Data View*)?
2. Каким образом осуществляется процедура занесения в исходный файл данных *SPSS* меток переменных?
3. Чем отличаются пропущенные значения, определяемые системой (*system-defined missing values*) от пропущенных значений, задаваемых пользователем программы (*user-defined missing values*)?
4. Какие три типа шкал измерения переменных используются в *SPSS* и каким образом задается тип шкалы измерения переменной при формировании исходного файла данных?
5. Чем отличаются дихотомическая и категориальная кодировка данных?
6. Почему при занесении в исходный файл данных *SPSS* ответов на многовариантные (безальтернативные) вопросы необходимо использовать дихотомическую кодировку данных?
7. С какой целью и в каких случаях применяется двойная запись данных при создании исходного файла *SPSS*?
8. По шкале какого типа измеряются следующие переменные:
   а) частота приобретения товара «*X*»:
      - реже 1-го раза в неделю;
      - 1–3 раза в неделю;
      - чаще 3-х раз в неделю;

б) семейное положение:
- замужем/ женат;
- не замужем/ холост;
- разведена/ разведен;

в) оценка уровня сервисного обслуживания:
- очень высокая;
- высокая;
- средняя;
- низкая;
- очень низкая;

г) возраст (23 года, 24 года, 32 года, 57 лет)?

9. Как отличаются друг от друга переменные, измеряемые по разным типам шкал, относительно возможности произведения арифметических операций и расчета статистических показателей?

3. ЧАСТОТНЫЙ АНАЛИЗ

3.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В *SPSS*

Частотный анализ применяется при проведении описательных маркетинговых исследований. Его целью является описание объекта исследования, при этом, как правило, используется вопрос: «Как часто?». Например, требуется описать социальный портрет типичного потребителя определенного товара или марки. В этом случае следует выяснить, как часто среди потребителей данного продукта или марки встречаются люди определенного возраста, дохода, семейного положения, профессии и т.д. В случае если требуется описать поведение потребителей определенной целевой группы, необходимо проанализировать, как часто респонденты посещают определенный магазин в определенное время, приобретают определенные товары по определенным ценам и т.д.

Свое название частотный анализ получил в силу того, что в ходе его проведения анализируется частота наступления определенного события (частота выбора определенного варианта ответа).

Частотный анализ является одномерным видом анализа, т.е. при его проведении анализируется только одна переменная. Этим он отличается от других видов анализа, проводимых с помощью *SPSS*, поскольку в них участвуют как минимум две переменные с целью выявления и (или) описания взаимосвязи между ними.

В частотном анализе могут участвовать все виды переменных относительно шкалы измерения — номинальные, порядковые и метрические. Однако следует отметить, что при использовании метрических переменных возникают проблемы с интерпретацией результатов анализа. Частотные таблицы, построенные для метрических переменных, не дают наглядного представления о частоте наступления событий. Также существует сложность с географическим представлением результатов частотного анализа с участием метрических переменных.

Пример. В ходе проведения исследования туристического рынка в курортной зоне Германии «Баварский лес» требуется определить структуру всей совокупности отдыхающих относительно формы размещения (проживания) (см. рис. 3.1).

Для проведения такого анализа из всех вопросов анкеты, которая была использована для опроса туристов, выбирается вопрос № 3 «Укажите форму своего размещения (проживания) в курортной зоне "Баварский лес"».

В файле данных *SPSS*, содержащем результаты опроса туристов, вопрос № 3 представлен в виде переменной с именем «q_3» и меткой «Форма проживания» (рис. 3.2).

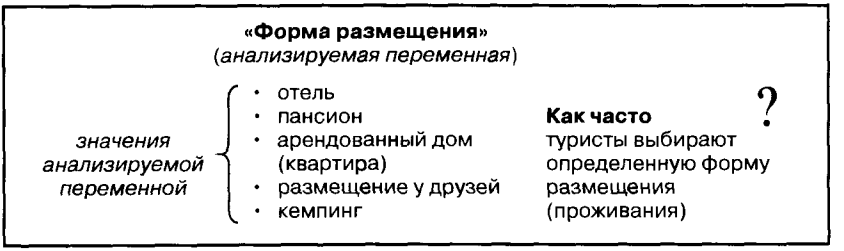


Рис. 3.1. Частотный анализ. Постановка цели исследования

Поскольку переменная «Форма проживания» является номинальной переменной, в столбце «*Values*» таблицы «Свойства переменных» указываются значения метки переменной с их числовыми кодами (см. рис. 3.2).

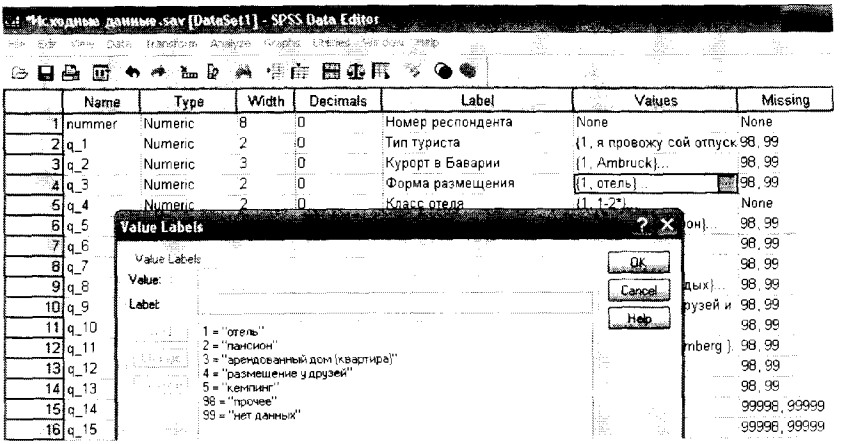


Рис. 3.2. Фрагмент вкладки «*Variable View*» («Свойства переменных»)

Из данных, представленных на рис. 3.2 видно, что группа туристов, выбравших форму размещения, отличную от пяти заявленных и обозначенную в базе данных как «прочее» (числовой код «98»), не участвует в проведении анализа. Такие ответы наряду с ответами «нет данных» (числовой код «99») обозначены как пропущенные значения в столбце «Missing» (см. рис. 3.2).

На рис. 3.3 представлен фрагмент таблицы «Значения переменных», содержащий данные об ответах респондентов на вопрос анкеты № 3 (столбец «q_3»). Респондент в строке 95 проживает в отеле (числовой код «1»), в строке 97 стоит числовой код арендованного дома (квартиры) («3»), а в строке 98 содержатся данные о туристе, проживающем в пансионе (числовой код «2»).

*Исходные данные.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : nummer 2149

	nummer	q_1	q_2	q_3	q_4	q_5	q
94	5060	1	4	1	3	3	
95	3269	1	4	1	3	4	
96	1067	2	4	1	2	3	
97	3587	2	4	3	99	.	
98	1364	1	4	2	99	.	
99	649	1	4	1	2	3	
100	3596	2	4	2	3	3	

Рис. 3.3. Фрагмент вкладки *«Data View»* («Значения переменных»)

В ходе проведения исследования было опрошено 6396 туристов. Просмотр ответов всех опрошенных по базе данных в *SPSS* не дает возможность понять, как часто туристы выбирают ту или иную форму размещения (проживания) на отдыхе. Получить обобщенную, обозримую картину ответов на этот вопрос позволяет таблица одномерного частотного распределения.

В ходе выполнения частотного анализа решаются следующие задачи:

1) построение частотных таблиц (частотных распределений);

2) графическое представление распределения анализируемой величины;

3) получение статистических показателей, характеризующих распределение анализируемой переменной.

3.2. КОМАНДЫ *SPSS* НА ВЫПОЛНЕНИЕ ЧАСТОТНОГО АНАЛИЗА

При выборе меню «Analyze> Descriptive Statistics> Frequencies» (рис. 3.4) на экране компьютера появляется диалоговое окно «Частотный анализ» (*«Frequencies»*) (рис. 3.5).

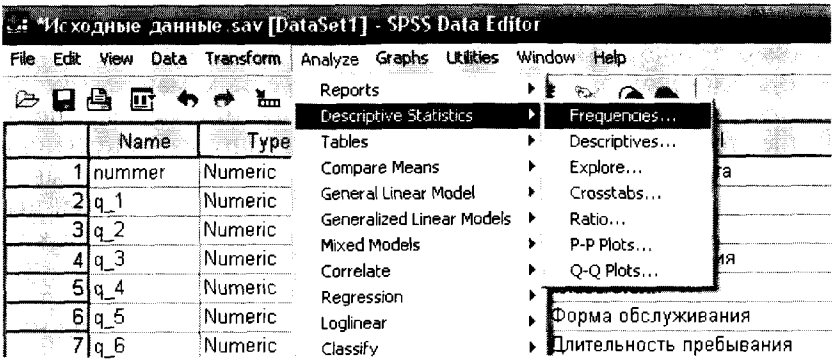


Рис. 3.4. Выбор в меню процедуры «Частотный анализ»

В левом поле диалогового окна «Частотный анализ» указываются метки всех переменных, занесенных в базу данных. Из данного списка выбирается метка тестируемой переменной (в рассматриваемом примере это «Форма размещения») и переносится в поле *«Variable(s)»* (см. рис. 3.5).

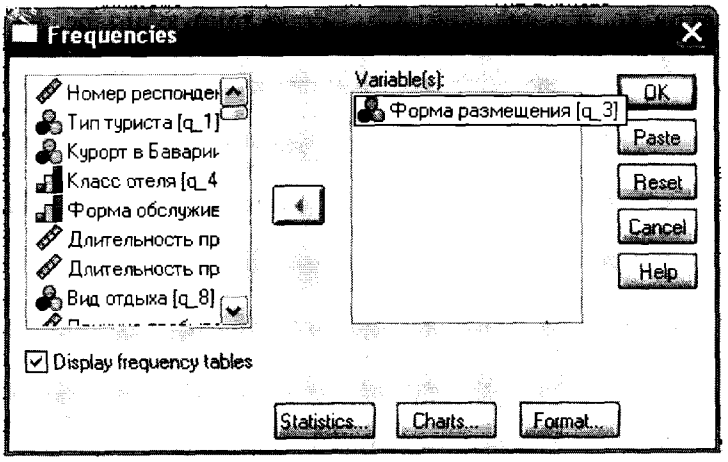


Рис. 3.5. Диалоговое окно «Частотный анализ»

Следует отметить, что в поле «*Variable(s)*» можно указывать не одну, а несколько переменных. Не стоит забывать, что частотный анализ является одномерным, т.е. в нем может участвовать только одна переменная. При задании команды выполнения частотного анализа для нескольких переменных указанные переменные будут анализироваться одновременно отдельно друг от друга. В этом случае в окне *SPSS «Output»*, содержащем результаты анализа, будут представлены несколько таблиц частотного распределения отдельно для каждой анализируемой переменной.

В диалоговом окне «Частотный анализ» есть команда «*Display frequency tables*» (см. рис. 3.5). На рис. 3.5 стоит отметка на выполнение этой команды, в результате чего после запуска выполнения процедуры частотного анализа на экран компьютера будет выведена таблица частотного распределения. Построение таблицы требуется не всегда, например, когда нужно только графическое представление частотного распределения.

В главном диалоговом окне «Частотный анализ» имеются три кнопки «*Statistics...*», «*Charts...*», «*Format...*» (см. рис. 3.5), при нажатии которых открываются одноименные вспомогательные диалоговые окна.

При нажатии кнопки «*Statistics...*» открывается диалоговое окно, в котором можно задать команды на расчет различных статистических показателей. Структура этого диалогового окна будет рассмотрена далее в п. 3.4 «Основные статистические показатели, используемые при проведении частотного анализа».

При нажатии кнопки «*Charts...*» (см. рис. 3.5) открывается диалоговое окно «Диаграммы» (*Charts*), позволяющее задать команды на построение диаграмм, иллюстрирующих результаты анализа (рис. 3.6).

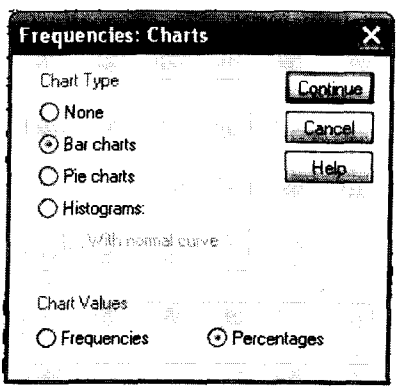


Рис. 3.6. Диалоговое окно «Диаграммы»

В случае если анализируемые величины являются номинальными или порядковыми (см. раздел «Типы шкал измерения переменных»), в диалоговом окне «Диаграммы» можно выбрать построение столбиковых или круговых диаграмм (*Bar charts* или *Pie charts*). При этом в поле «Значения диаграммы» (*Chart Values*) следует указать, как будет построена диаграмма: по данным частот в абсолютном или процентном выражении (*Frequencies* или *Percentages*). В рассматриваемом примере выбрано построение столбиковой диаграммы, иллюстрирующей разделение туристов по форме размещения (проживания), в процентном выражении (см. рис. 3.6). Результат выполнения этой команды будет представлен далее.

В случае анализа метрических величин, а также номинальных или порядковых с большим количеством возможных вариантов ответов (более 10—15) в диалоговом окне «Диаграммы» можно выбрать построение гистограмм (*Histograms*). На рис. 3.7 представлена гистограмма распределения метрической величины «Возраст». Из данных гистограммы видно, что среди опрошенных туристов молодых людей в возрасте 20 лет всего около 20 человек, а людей в возрасте 50 лет — более 300 (см. рис. 3.7).

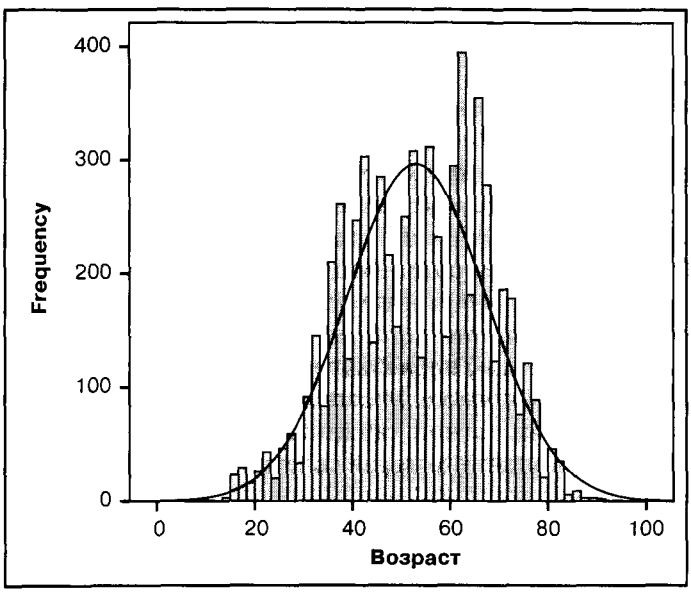


Рис. 3.7. Гистограмма распределения возраста туристов, отдыхающих в курортной зоне «Баварский лес»

При построении гистограммы существует возможность наложения на нее кривой нормального распределения (*With normal curve*) (см. рис. 3.6). Это позволяет наглядно сопоставить распределение анализируемой величины с нормальным распределением (см. рис. 3.7). Характеристики и иллюстрация нормального распределения будут рассмотрены далее в разделе «Меры средней тенденции».

При нажатии кнопки *«Format...»* (см. рис. 3.5) в главном диалоговом окне «Частотный анализ» открывается диалоговое окно, которое позволяет форматировать таблицу частотного распределения, например расположить данные в строках таблицы по возрастанию (убыванию) числовых кодов меток переменной или количества респондентов, имеющих одинаковые ответы.

При нажатии кнопки *«OK»* в диалоговом окне «Частотный анализ» (см. рис. 3.5) запускается процедура выполнения частотного анализа.

3.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ЧАСТОТНОГО АНАЛИЗА

После запуска процедуры выполнения частотного анализа в *SPSS* на экран компьютера выводятся результаты анализа. Основным результатом является таблица частотного распределения анализируемой величины (см. табл. 3.1).

Колонка *Frequency* (частота) табл. 3.1 содержит частоты, т.е. то количество респондентов, которые выбрали тот или иной вариант ответа. Таким образом, из табл. 3.1 видно, что 1755 из опрошенных туристов проживает в отеле, 2868 остановились в пансионе, 1013 респондентов снимают дом или квартиру и т.д.

Делать выводы о том, много или мало респондентов выбрали тот или иной вариант ответа, опираясь на значения колонки *Frequency*, невозможно, поскольку приходится все время соотносить эти числа с общим количеством опрошенных. Поэтому для анализа данных частотной таблицы удобнее использовать колонку *Percent* (процент), которая содержит процентные значения для каждой из частот.

Из данных табл. 3.1 видно, что большинство туристов, участвующих в опросе, проживает в пансионе (44,8% от числа общего числа опрошенных). Самую маленькую часть составляют туристы, которые разместились у друзей, это всего 1,6% от числа опрошенных.

Таблица 3.1

Таблица частотного распределения переменной «Форма размещения»

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	отель	1755	27,4	29,1	29,1
	пансион	2868	44,8	47,5	76,6
	арендованный дом (квартира)	1013	15,8	16,8	93,4
	размещение у друзей	102	1,6	1,7	95,1
	кемпинг	297	4,6	4,9	100,0
	Total	6035	94,4	100,0	
Missing	прочее, нет данных	27	,4		
	System	334	5,2		
	Total	361	5,6		
Total		6396	100,0		

Колонка *Valid Percent* (действительный процент) связана с такой важной характеристикой, как «Отсутствие ответа». При проведении опроса некоторые респонденты по разным причинам воздерживаются от ответа на поставленный вопрос. В этом случае пропущенные ответы кодируются (в рассматриваемом примере это значения 98 — «прочее» и 99 — «нет данных») и заносятся в базу данных. Как видно из табл. 3.1, количество таких пропущенных ответов составляет 27.

Кроме пропущенных ответов, задаваемых пользователем, существуют пропущенные значения, определяемые системой. В рассматриваемом примере количество таких пропущенных ответов составляет 334. Подробно виды и сущность пропущенных значений были рассмотрены в п. 2.1 «Структура редактора данных». Всего же количество пропущенных ответов в рассматриваемом примере составляет 361 (см. табл. 3.1). Таким образом, при общем количестве опрошенных туристов 6396 в базе данных существует только 5035 действительных значений анализируемой переменной «Форма размещения» (см. табл. 3.1).

В колонке *Valid Percent* табл. 3.1 содержатся данные о процентном соотношении туристов, выбравших ту или иную форму размещения (проживания) на отдыхе, и общего количества респондентов, ответивших на этот вопрос. Самую большую долю составляют туристы, проживающие в пансионе (47,5% ответивших на вопрос). Самую маленькую группу образуют туристы, разместившиеся у друзей (1,7% ответивших на вопрос).

Вопрос о том, какой из показателей — процент опрошенных или процент ответивших на вопрос — необходимо использовать для выявления определенных закономерностей, некорректен. Оба показателя несут некую информацию и используются, как правило, одновременно. В рассматриваемом примере анализ обоих показателей дает один и тот же результат: структура респондентов относительно формы размещения (проживания) на отдыхе в обоих случаях одинакова (см. табл. 3.1).

В колонке *Cumulative Percent* (накопительный процент) последовательно суммируются процентные соотношения групп респондентов, ответивших на вопрос (действительные проценты). Информация, содержащаяся в этой колонке, используется при определении таких показателей, как «медиана» и «квартили». Более подробно вопрос об определении статистических характеристик будет рассмотрен в параграфе «Основные статистические показатели, используемые при проведении частотного анализа».

Графическим представлением результатов анализа в рассматриваемом примере является столбиковая диаграмма (рис. 3.8).

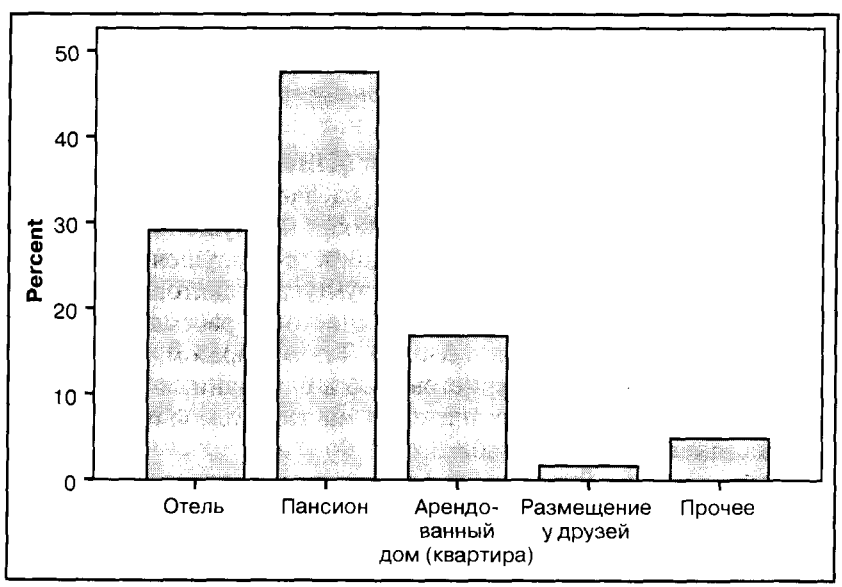


Рис. 3.8. Формы размещения (проживания) туристов в курортной зоне «Баварский лес»

Построенная диаграмма наглядно описывает данные, содержащиеся в таблице частотного распределения (см. табл. 3.1). Самой многочисленной группой является группа туристов, проживающих в пансионе (47,5%). На втором месте находится группа туристов, проживающих в отеле (29,1%). Третьей по численности респондентов является группа туристов, арендующих для отдыха дом или квартиру (16,8%).

На четвертом месте (4,9%) находится группа «прочее», объединяющая различные формы размещения на отдыхе, не выделенные в отдельные группы, например туристы, проживающие на теплоходе, в палатке, кемпинге и пр.

Самой непопулярной формой размещения на отдыхе является размещение у друзей (1,7%).

3.4. ОСНОВНЫЕ СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ, ИСПОЛЬЗУЕМЫЕ ПРИ ПРОВЕДЕНИИ ЧАСТОТНОГО АНАЛИЗА

Основные статистические характеристики одномерных частотных распределений используются также при выполнении в *SPSS* других видов анализа, поэтому они заслуживают подробного рассмотрения в отдельном разделе данного учебного пособия.

Задание на расчет статистических характеристик можно дать в диалоговом окне «Частотный анализ: Статистические показатели» (см. рис. 3.8), которое открывается при нажатии кнопки *«Statistics...»* в главном диалоговом окне «Частотный анализ» (см. рис. 3.5). Набор статистических показателей, представленных в данном диалоговом окне, состоит из четырех блоков (рис. 3.9):

- меры центральной тенденции (*Central Tendency*);
- меры разброса (*Dispersion*);
- процентили (*Persentile Values*);
- характеристики распределений (*Disribution*).

Возможность и (или) целесообразность расчета тех или иных статистических показателей, представленных в диалоговом окне «Частотный анализ: Статистические показатели», зависит от вида анализируемых переменных (см. п. 23 «Типы шкал измерения переменных»). Так, например, расчет показателя «Сумма» (*Sum*) возможен только для метрических переменных.

Возможность расчета других мер центральной тенденции, называемых мерами средней тенденции: моды (*Mode*), медианы (*Median*), средней арифметической (*Mean*) — также зависит от уровня измерения переменной (см. далее п. 3.4.1 «Меры средней тенденции»).

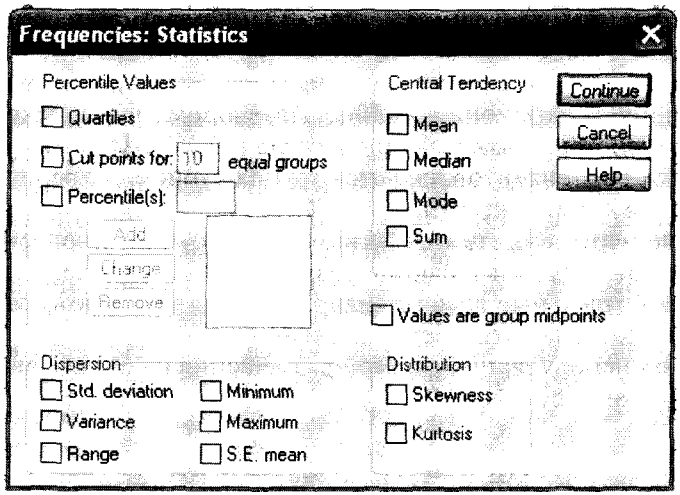


Рис. 3.9. Диалоговое окно «Частотный анализ: Статистические показатели»

Меры разброса (*Dispersion*) рассчитываются только для метрических переменных (см. рис. 3.9). Процентили (*Persentile Values*) (см. рис. 3.9) используются при расчете мер разброса для порядковых переменных, поэтому они будут подробно рассмотрены в п. 3.4.2 «Меры разброса».

Рассмотрим более подробно каждую группу статистических показателей.

3.4.1. МЕРЫ СРЕДНЕЙ ТЕНДЕНЦИИ

Как уже было отмечено, возможность расчета мер средней тенденции зависит от уровня измерения переменной. Вопрос о возможности расчета различных показателей и произведения различных математических операций для переменных с разными уровнями измерения уже рассматривался в п. 2.3 «Типы шкал измерения переменных». Рассмотрим этот вопрос более подробно относительно мер средней тенденции (табл. 3.2).

Таблица 3.2

Возможности использования различных мер средней тенденции для шкал различного типа (Крыштановский, 2007. С. 25)

№ п/п	Уровень измерения	Допустимые меры средней тенденции
1	Номинальный	Мода
2	Порядковый	Мода, медиана
3	Метрический	Мода, медиана, среднее арифметическое

Для номинальных переменных мерой средней тенденции может выступать только мода.

Мода — это наиболее часто встречающееся значение анализируемой переменной. В примере, рассматриваемом в главе 3 «Частотный анализ», модой является значение 2 — это числовой код значения «пансион» анализируемой переменной «форма размещения». Большинство опрошенных туристов (47,5%) проживает в пансионе (см. табл. 3.1).

Для порядковых переменных в качестве меры средней тенденции кроме моды возможно также определение медианы.

Медиана является точкой на шкале, которая делит всю совокупность опрошенных — тех, кто отметил градации меньше этой точки (либо равные ей), и тех, кто отметил градации больше этой точки (Крыштановский, 2007. С. 29).

В примере, рассматриваемом в разделе «Частотный анализ», анализируемая переменная «форма размещения» является номинальной, поэтому определение медианы в данном случае невозможно. Пример, иллюстрирующий показатели «мода» и «процентили», будет представлен далее в п. 3.4.2 «Меры разброса».

Для метрических переменных в качестве мер средней тенденции кроме моды и медианы возможно использование среднего арифметического.

Среднее арифметическое — это сумма всех частот, поделенная на количество событий (значений анализируемой величины).

При нормальном распределении анализируемой величины (рис. 3.10) мода, медиана и среднее арифметическое равны.

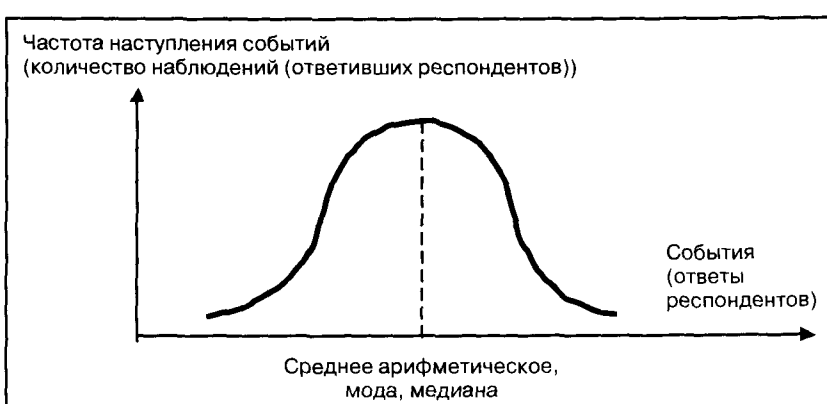


Рис. 3.10. График нормального распределения анализируемой величины

Как видно из данных графика на рис. 3.10, при нормальном распределении среднее значение анализируемой величины встречается наиболее часто, т.е. является модой.

Также из графика на рис. 3.10 видно, что частота наступления событий, расположенных на равноудаленном расстоянии левее и правее среднего значения, всегда одна и та же. Таким образом, среднее значение является той точкой на шкале ответов (значений анализируемой переменной), которая делит всю совокупность респондентов на две равные части, т.е. является медианой.

3.4.2. МЕРЫ РАЗБРОСА

При анализе показателей средней тенденции необходимо также рассмотреть показатели мер разброса. Значение и суть этих показателей нагляднее всего можно объяснить на примере мер разброса для среднего арифметического.

Рассмотрим пример вычисления средней заработной платы (табл. 3.3).

Таблица 3.3

Данные о средней заработной плате, руб.

№ п/п	Значение заработной платы	Среднее значение	Расхождение реальной заработной платы и среднего значения
1	37 500	35 000	1500
2	32 500	35 000	-2500
3	37 500	35 000	2500
4	35 500	35 000	500
5	34 000	35 000	-1000

В теории статистики расхождения реальных (эмпирических) данных и среднего значения называются остатками. Остатки характеризуют правомерность подмены эмпирических данных одним числом — средним значением.

Такая подмена является на 100% правомерной, если все эмпирические данные имеют одинаковое значение. Например, если заработная плата всех респондентов одинакова, показатель «среднее значение» полностью описывает (характеризует) уровень заработной платы в целом. Как правило, на практике такое встречается редко.

Таким образом, при использовании показателя «среднее значение» необходим анализ остатков. Остатки показывают, как точно (в какой мере) показатель «среднее значение» характеризует ана-

лизируемую переменную. Очевидно, чем меньше остатки, тем точнее результаты анализа, основанного на показателе «среднее значение».

При анализе остатков рассчитываются показатели, называемые мерами разброса: дисперсия, среднеквадратичное отклонение и стандартная ошибка среднего.

Дисперсия представляет собой сумму квадратов остатков, деленную на количество наблюдений:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \qquad (1)$$

где x_i — значение переменной x для i-го респондента; $\bar{x}$ — среднее значение переменной; n — количество респондентов.

Остатки могут быть как положительными, так и отрицательными, т.е. реальные значения могут быть как меньше, так и больше среднего. Поэтому простое суммирование остатков не дает представления о масштабах отклонений реальных значений от среднего. Сумма остатков может оказаться равной нулю. Для решения этой проблемы при расчете дисперсии остатки возводятся в квадрат.

Недостатком дисперсии является то, что этот показатель трудно оценить. В результате возведения в квадрат остатков при расчете дисперсии рассчитанное значение оказывается несоразмерно большим. В примере, представленном в табл. 3.3, средняя заработная плата составляет 35 000 руб., а дисперсия — 4 000 000. Сложно чисто интуитивно определить, много это или мало. Это значение не дает ответа на поставленный вопрос: как точно (в какой мере) показатель «среднее значение» характеризует анализируемую переменную?

Ввиду сложности оценки дисперсии используется другой показатель — стандартное отклонение.

Стандартное отклонение — это корень квадратный из дисперсии:

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

В примере, представленном в табл. 3.3, стандартное отклонение составляет 2000. Следовательно, средняя заработная плата

составляет 35 000 руб. ± 2000. Из чего можно сделать вывод, что среднее значение достаточно точно описывает реальный уровень заработной платы, поскольку стандартное отклонение весьма незначительно в сравнении со средним значением.

Как было отмечено в п. 1.1 «Формирование статистической выборки», важной характеристикой является репрезентативность выборки. Она определяет правомерность переноса результатов анализа выборки на генеральную совокупность. Например, в ходе анализа туристического рынка в курортной зоне «Баварский лес» было опрошено 6396 туристов (см. табл. 3.1). В какой мере выводы, построенные на результатах анализа собранных данных, применимы к туристическому рынку в данном регионе в целом? Частично ответ на этот вопрос дает показатель «стандартная ошибка среднего».

Стандартная ошибка среднего показывает соотношение арифметического среднего с генеральным математическим ожиданием. Иными словами, как среднее значение анализируемой величины, рассчитанное по данным опроса респондентов (выборки), соотносится со средним значением этой величины для целевой группы в целом (генеральной совокупности).

Стандартная ошибка среднего ($с.о.\bar{x}$) рассчитывается по формуле

$$с.о.\bar{x} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n}. \qquad (3)$$

Генеральное математическое ожидание (среднее значение для генеральной совокупности) с вероятностью 95% лежит в интервале ($\bar{x} \pm 2 \times с.о.\bar{x}$), т.е. среднее арифметическое плюс-минус удвоенная величина стандартной ошибки среднего.

По данным табл. 3.3, стандартная ошибка среднего (рассчитанная по формуле (3) составляет 894. Таким образом, можно утверждать, что с вероятностью 95% математическое ожидание заработной платы должно лежать в интервале от 33 712 до 37 288 руб. ($35\,000 \pm 2 \times 894$).

Подводя итог, необходимо подчеркнуть, что для интерпретации показателя «среднее арифметическое» необходимо использовать хотя бы один из показателей мер разброса (дисперсию, стандартное отклонение или стандартную ошибку среднего).

Как уже было отмечено в разделе «Меры средней тенденции», для номинальных переменных мерой средней тенденции может

выступать только мода. *Мода не имеет какого-то показателя разброса. Определенной характеристикой может выступать лишь процентное значение модальной величины.*

Например, в результате проведения частотного анализа переменной «форма размещения» (см. табл. 3.1) выяснилось, что большинство туристов, ответивших на этот вопрос, проживает в пансионе (47,5%). Тот факт, что на остальные четыре варианта ответа («отель», «арендованный дом (квартира)», «размещение у друзей» и «прочее») приходится 52,5%, может указывать на разброс значений. Однако данное указание достаточно слабо, поскольку не показывает, как именно разбросаны данные по другим вариантам.

На порядковом уровне мерой средней тенденции является медиана (см. п. 3.4.1 «Меры средней тенденции»). Наиболее распространенным показателем, характеризующим разброс значений переменной, измеренной на порядковом уровне, является квартильное отклонение.

Квартильное отклонение — это разница между третьим и первым квартилями. Чтобы понять смысл этого показателя, необходимо рассмотреть понятие квартилей.

Если медиана делит всю совокупность респондентов на две равные части, то квартили — на четыре.

Первый квартиль — точка на шкале ответов, значения меньше (либо равные) которой выбрали 25% опрошенных. Второй квартиль — точка на шкале ответов, значения меньше (либо равные) которой выбрали 50% опрошенных. Второй квартиль совпадает с медианой. Третий квартиль — точка на шкале ответов, значения меньше (либо равные) которой выбрали 75% опрошенных.

Рассмотрим понятие квартилей на примере частотного анализа порядковой величины «доход семьи». Данная величина является порядковой, а не метрической, поскольку варианты ответов представлены в виде групп, сформированных по уровню дохода семьи (табл. 3. 4).

По данным табл. 3.4 можно утверждать, что медиана в данном случае равна 6. Доход семьи «от 2301 до 2800 евро» делит всю совокупность респондентов на две равные части (50% лежит в интервале между 44,1 и 62,7%). Первый квартиль имеет значение 5. Доход семьи «до 2 300 евро» имеют как минимум 25% респондентов (25% лежит в интервале от 23,1 до 44,1%). Третий квартиль имеет значение 7. Доход семьи «до 3300 евро» имеют как минимум 75% респондентов (75% лежит в интервале от 62,7 до 77,6%) (см. табл. 3.4).

Таблица 3.4

Таблица частотного распределения переменной «доход семьи»

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	менее 500 Евро	86	1,3	1,5	1,5
	от 501 до 900 Евро	137	2,1	2,4	3,9
	от 901 до 1 250 Евро	259	4,0	4,5	8,3
	от 1 251 до 1 800 Евро	856	13,4	14,8	23,1
	от 1 801 до 2 300 Евро	1213	19,0	21,0	44,1
	от 2 301 до 2 800 Евро	1076	16,8	18,6	62,7
	от 2 801 до 3 300 Евро	863	13,5	14,9	77,6
	от 3 301 до 3 800 Евро	547	8,6	9,5	87,1
	от 3 801 Евро и выше	748	11,7	12,9	100,0
	Total	5785	90,4	100,0	
Missing	10 нет данных	611	9,6		
Total		6396	100,0		

В рассматриваемом примере (см. табл. 3.4) анализируемая переменная «доход семьи» имеет в целом девять градаций. Квартильное отклонение со значением 2 (7 – 5 = 2) может рассматриваться как не очень большое. В данном случае медиана достаточно точно отображает среднюю тенденцию переменной «доход семьи», поскольку немного респондентов имеют доход семьи, существенно отличающийся от медианы.

В продолжение процесса разделения совокупности респондентов на две (при помощи медианы) либо на четыре (при помощи квартилей) равные части можно поставить задачу разделения на 5, 10 и вообще на любое количество равных частей. Применительно к разделению на пять частей (*квинтильное разделение*) и на 10 частей (*децильное разделение*) в качестве мер разброса используются *квинтильное* и *децильное отклонение*.

3.4.3. ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЙ

В диалоговом окне «Частотный анализ: Статистические характеристики» в группе «Характеристики распределений» (*Disribution*) можно выбрать расчет коэффициента асимметрии (*Skewness*) и коэффициента вариации, или эксцесса (*Kurtosis*) (см. рис. 3.9). Эти

рактеристики показывают, насколько поведение распределения анализируемой величины соответствует нормальному распределению (см. рис. 3.10).

Коэффициент асимметрии (*Skewness*) — это мера отклонения распределения анализируемой величины от симметричного распределения. Симметричным является нормальное распределение, у которого на одинаковом удалении от среднего значения по обе стороны находится одинаковое количество респондентов, выбравших определенный ответ (значение анализируемой величины) (см. рис. 3.10).

Если анализируемая величина подчиняется нормальному распределению, то коэффициент асимметрии равен нулю. Если вершина асимметричного распределения сдвинута к меньшим значениям, то говорят о положительной асимметрии, в противном случае — об отрицательной.

Коэффициент вариации, или *эксцесс* (*Kurtosis*), указывает на то, является ли распределение анализируемой величины более пологим или более крутым по отношению к кривой нормального распределения.

Если анализируемая величина подчиняется нормальному распределению, то коэффициент вариации (эксцесс) равен нулю. При высоком значении коэффициента вариации (эксцесса) кривая распределения анализируемой величины является более пологой, чем кривая нормального распределения, в противном случае — более крутой.

Если значения коэффициентов асимметрии и вариации существенно отличаются от нуля, то гипотезу о том, что данные взяты из нормально распределенной генеральной совокупности, следует отвергнуть.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите цель проведения частотного анализа. Поясните происхождение его названия.
2. Какие требования предъявляются к переменным, участвующим в частотном анализе, относительно шкалы измерения переменных?
3. Что является результатом проведения частотного анализа в *SPSS*? Какие возможности предоставляет *SPSS* относительно графического представления результатов частотного анализа?
4. Какие показатели могут быть рассчитаны в частотной таблице при помощи *SPSS*? Охарактеризуйте практический смысл показателей «процент опрошенных» и «процент ответивших».

5. Перечислите меры средней тенденции. Каковы возможности использования мер средней тенденции для номинальных, порядковых и метрических переменных?
6. Что представляет собой нормальное распределение анализируемой величины? Как соотносятся между собой меры средней тенденции при нормальном распределении?
7. Как соотносятся между собой показатели «дисперсия» и «квадратное отклонение»? Чем обусловлено предпочтение применения на практике показателя «квадратное отклонение»?
8. Что характеризует показатель «стандартная ошибка среднего»?
9. Обоснуйте необходимость использования мер разброса (дисперсии, квадратного отклонения или стандартной ошибки среднего) при анализе показателя «среднее арифметическое».
10. Какой показатель выступает в качестве меры разброса для меры средней тенденции на порядковом уровне (для медианы)?
11. Что характеризуют показатели «коэффициент асимметрии» и «коэффициент вариации»?

4. ТАБЛИЦЫ СОПРЯЖЕННОСТИ

4.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В *SPSS*

Таблицы сопряженности выступают развитием частотного анализа. Если частотный анализ является одномерным анализом (позволяет анализировать только одну переменную), то таблицы сопряженности представляют собой двух- или многомерный частотный анализ. Это значит, что в результате построения таблиц сопряженности можно анализировать зависимость двух и более переменных.

Как правило, таблицы сопряженности используют для выявления взаимосвязи двух переменных: номинальных и (или) порядковых (см. п. 2.3 «Типы шкал измерения переменных»). Количество значений (вариантов ответов) анализируемых величин не должно быть более 8—10, иначе построенная таблица будет плохо обозримой, т.е. не даст наглядного представления о результатах анализа.

Пример. В ходе проведения исследования туристического рынка в курортной зоне Германии «Баварский лес» требуется определить взаимосвязь между двумя переменными — «сопровождающие лица» и «форма размещения» (рис. 4.1). Обе переменные являются номинальными (см. п. 2.3 «Типы шкал измерения переменных»).

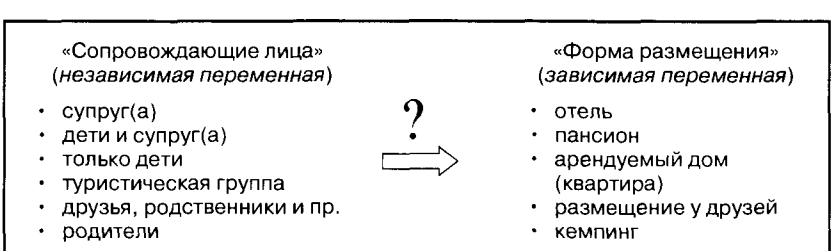


Рис. 4.1. Таблицы сопряженности: постановка цели исследования

Для проведения такого анализа из всех вопросов анкеты, которая была использована для опроса туристов, выбираются:

1) вопрос № 3 «Укажите форму своего размещения (проживания) в курортной зоне "Баварский лес"»;

2) вопрос № 17 «С кем Вы отдыхаете в курортной зоне "Баварский лес"».

В файле данных *SPSS*, сформированном по результатам опроса туристов, вопрос № 3 представлен в виде переменной с меткой «q_3» и именем «форма размещения». Поскольку эта переменная участвовала в примере частотного анализа, ее представление в файле данных *SPSS* подробно рассмотрено в главе 3 «Частотный анализ». Вопрос № 17 представлен в виде переменной с меткой «q_17» и именем «сопровождающие лица» (рис. 4.2).

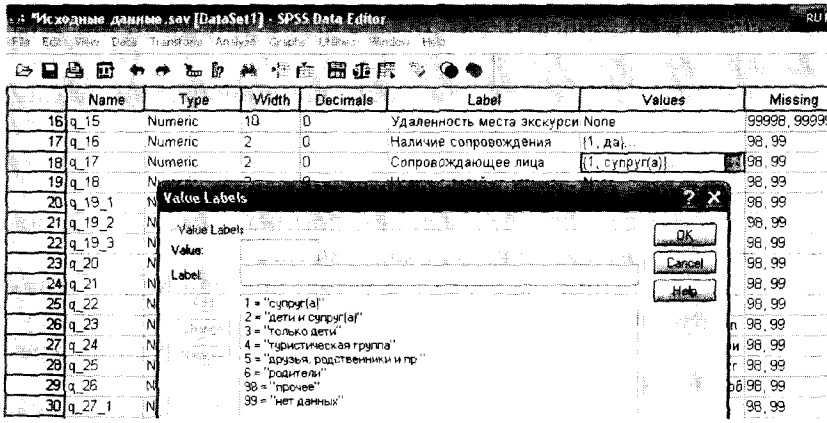
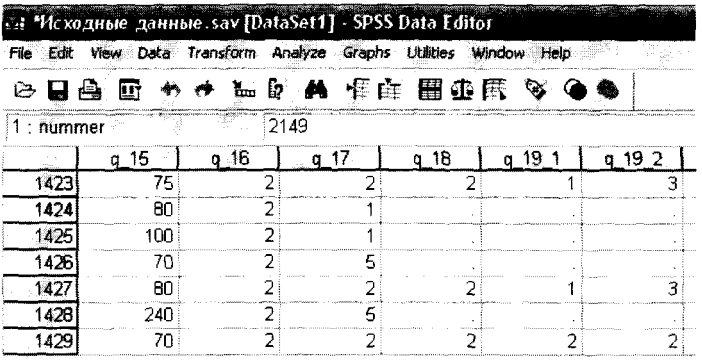


Рис. 4.2. Фрагмент вкладки «*Variable View*» («Свойства переменных»)

Туристы, имеющие других сопровождающих лиц, объединенные в группу «прочее» (числовой код «98»), не участвуют в анализе. Такие ответы, как «нет данных» (числовой код «99»), обозначены как пропущенные значения в столбце «*Missing*» (см. рис. 4.2). Причем следует отметить, что в ответы «нет данных» кроме прочих занесены ответы туристов, отдыхающих без сопровождающих лиц.

Числовые коды значений переменной «сопровождающие лица» содержатся в таблице «Значения переменных» вкладки редактора данных «*Data View*» (рис. 4.3).

Из данных, представленных в колонке«q_17» на рис. 4.3, видно, что респондент в строке 1423 отдыхает с женой (мужем) и детьми (числовой код «2»), респонденты в строках 1424 и 1425 — с супругами (числовой код «1»), а респондент в строке 1426 — с друзьями или родственниками (числовой код «5»).



	q_15	q_16	q_17	q_18	q_19_1	q_19_2
1423	75	2	2	2	1	3
1424	80	2	1	.	.	.
1425	100	2	1	.	.	.
1426	70	2	5	.	.	.
1427	80	2	2	2	1	3
1428	240	2	5	.	.	.
1429	70	2	2	2	2	2

Рис. 4.3. Фрагмент вкладки *«Data View»* («Значения переменных»)

Как уже было отмечено, в ходе проведения анализа следует выяснить, влияет ли тот факт, с кем будет отдыхать респондент, на выбор формы размещения (проживания) на отдыхе. В данном случае переменная «сопровождающие лица» выступает в качестве независимой (причинной) переменной, а «форма размещения» — в качестве зависимой (на которую оказывается влияние).

«Вопрос о том, какая из переменных является причинной, т.е. оказывает влияние, а какая меняется вследствие этой причины, не может быть решен не только с помощью анализа таблиц, но и любым другим формально-статистическим методом» (Крыштановский, 2007. С. 45).

Таким образом, мы выбираем, какая из анализируемых переменных будет зависимой, а какая — независимой, исходя из здравого смысла. Здравый смысл подсказывает, что в большинстве случаев люди сначала решают, с кем они будут отдыхать, и только потом выбирают форму размещения (проживания) на отдыхе. Хотя возможны и другие ситуации, когда люди ориентируются на тот факт, в какое время вообще возможен их отдых, или они покупают дешевую «горящую» путевку. В этих случаях вопрос о том, с кем они будут отдыхать, может стоять не на первом месте.

Для выявления взаимосвязи между исследуемыми переменными с помощью таблиц сопряженности возможны следующие действия:

1) визуальный анализ построенной таблицы сопряженности;

2) анализ остатков (разницы между наблюдаемыми и ожидаемыми частотами);

3) расчет коэффициента «хи-квадрат» и других коэффициентов связи.

В данном учебном пособии последовательно будут рассмотрены все процедуры анализа, перечисленные выше. Анализ наблюдаемых и ожидаемых частот, а также расчет коэффициентов связи будут рассмотрены в отдельных параграфах.

4.2. КОМАНДЫ SPSS НА ПОСТРОЕНИЕ ТАБЛИЦ СОПРЯЖЕННОСТИ

При выборе меню «*Analyze*> *Descriptive Statistics*> *Crosstabs*» (рис. 4.4) на экране компьютера появляется диалоговое окно «Таблицы сопряженности» (*Crosstabs*) (рис. 4.5).

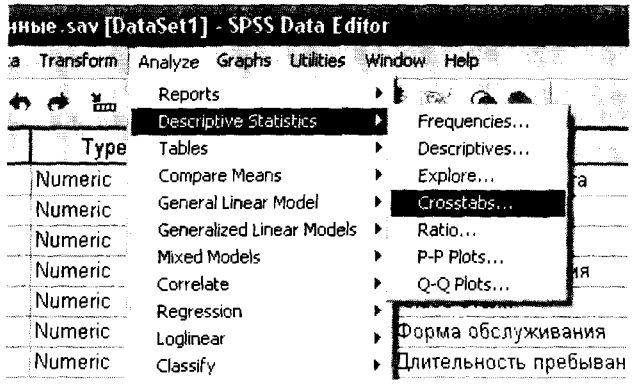


Рис. 4.4. Выбор в меню процедуры «Таблицы сопряженности»

В левом поле диалогового окна «Таблицы сопряженности» указываются метки всех переменных, занесенных в базу данных. Из данного списка следует выбрать метки исследуемых переменных и переместить их в поля «Строки» (*Rows*) и «Столбцы» (*Column(s)*) (см. рис. 4.5). Таким образом задается структура будущей таблицы.

Рекомендуется независимую (причинную) переменную разместить в строках таблицы, а зависимую (на которую оказывается влияние) — в столбцах. Это правило существует только для удобства анализа таблицы. Его выполнение (или невыполнение) никак не влияет на точность статистических расчетов. В рассматриваемом примере в поле «Строки» (*Rows*) перенесена независимая переменная «сопровождающие лица», а в поле «Столбцы» (*Column(s)*) — зависимая переменная «Форма размещения» (см. рис. 4.5).

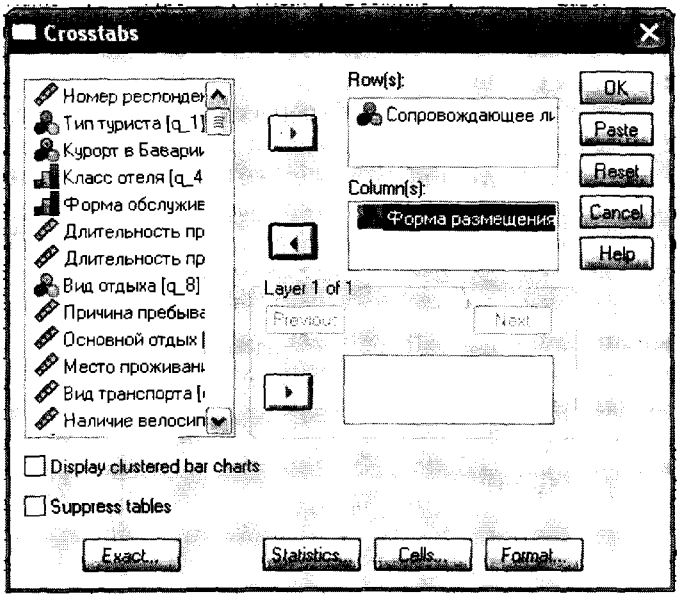


Рис. 4.5. Диалоговое окно «Таблицы сопряженности» (*Crosstabs*)

В диалоговом окне «Таблицы сопряженности» (см. рис. 4.5) можно задать несколько исследуемых переменных. Для каждого сочетания двух переменных будет создана таблица сопряженности. Например, если в поле «Строки» (*Rows*) заданы три переменные, а в поле «Столбцы» (*Column(s)*) — две, то мы получим шесть ($3 \times 2 = 6$) таблиц сопряженности.

В поле *«Layer»* диалогового окна «Таблицы сопряженности» (см. рис. 4.5) можно задать так называемые слои, т.е. переменные, разделяющие совокупность респондентов на дополнительные группы. Если в рассматриваемом примере в поле *«Layer»* задать переменную «пол», то будут созданы две таблицы сопряженности отдельно для мужчин и женщин. Каждая из таблиц будет демонстрировать, как влияет тот факт, с кем отдыхает турист, на выбор формы размещения (проживания) на отдыхе.

Можно выбрать другие уровни слоев. Каждый последующий уровень делит совокупность респондентов на меньшие группы и увеличивает количество таблиц сопряженности. Переходить от одного слоя к другому можно при помощи кнопок *«Next»* (следующий) и *«Previous»* (предыдущий) (см. рис. 4.5).

При нажатии кнопки *«Statistics»* (статистические характеристики) в главном диалоговом окне «Таблицы сопряженности»

(см. рис. 4.5) открывается вспомогательное диалоговое окно, в котором можно задать команды на выполнение теста «хи-квадрат» и расчет других статистических характеристик. Структура этого диалогового окна будет подробно рассмотрена далее в п. 4.5 «Коэффициент "хи-квадрат" и другие коэффициенты связи».

При нажатии кнопки *«Cells»* (ячейки) в главном диалоговом окне «Таблицы сопряженности» открывается одноименное вспомогательное диалоговое окно (рис. 4.6).

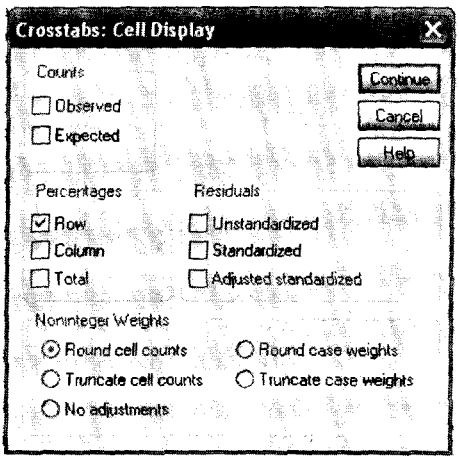


Рис. 4.6. Диалоговое окно «Таблицы сопряженности: Ячейки таблицы» (задание на расчет частот в процентном выражении)

Значение в каждой ячейке таблицы — количество наблюдений (частота, с которой встречается тот или иной ответ из возможных ответов респондентов).

В диалоговом окне «Таблицы сопряженности: Ячейки таблицы» можно задать расчет частот (значений, представленных в ячейках таблицы) в абсолютном (*Counts*) или процентном (*Percentages*) выражении (см. рис. 4.6).

Если выбирается расчет частот в абсолютном выражении, то можно задать расчет наблюдаемых (*Observed*) и ожидаемых (*Expected*) частот (см. рис. 4.6). Подробно этот вопрос будет рассмотрен далее в п. 4.4 «Анализ наблюдаемых и ожидаемых частот таблиц сопряженности».

Наиболее удобно проводить визуальный анализ данных таблицы сопряженности, если данные ячеек таблицы представлены в процентном выражении. При этом возможны следующие варианты расчета значений в ячейках таблицы (см. рис. 4.6):

- по строкам (*Row*) — количество наблюдений в каждой ячейке отнесено к сумме по строкам (100%);
- по столбцам (*Column*) — количество наблюдений в каждой ячейке отнесено к сумме по столбцам (100%);
- полные (*Total*) — количество наблюдений в каждой ячейке отнесено к общей сумме наблюдений (ответов респондентов) (100%).

Выбор того или иного варианта расчета частот в процентном выражении зависит от цели анализа и структуры таблицы. Если значения независимой переменной располагаются по строкам, а зависимой — по столбцам таблицы, то частоты в процентном отношении рассчитываются по строкам. В противоположном случае частоты в процентном отношении рассчитываются по столбцам. В рассматриваемом примере значения независимой переменной «сопровождающие лица» будут располагаться по строкам, поэтому выбирается расчет частот в процентном выражении по строкам (см. рис. 4.6).

При нажатии кнопки «*Format*» (Формат) в главном диалоговом окне «Таблицы сопряженности» (см. рис. 4.5) открывается вспомогательное диалоговое окно «Таблицы сопряженности: Формат таблицы» (рис. 4.7), в котором можно изменить порядок построения значений анализируемой переменной в строках таблицы.

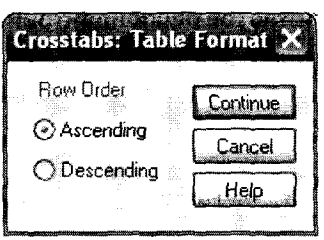


Рис. 4.7. Диалоговое окно «Таблицы сопряженности: Формат таблицы»

По умолчанию задается построение значений анализируемой величины в строках таблицы по возрастанию (*Ascending*), можно задать противоположный порядок построения — по убыванию (*Descending*). В рассматриваемом примере используется первый вариант построения значений по строкам (см. рис. 4.7). Это означает, что в первой строке таблицы будет располагаться значение переменной «сопровождающие лица» с наименьшим числовым кодом («1» — «супруг(а)»), а в последней строке — с наибольшим числовым кодом («6» — «родители»).

В главном диалоговом окне «Таблицы сопряженности» можно также задать графическое представление таблицы. Для этого нужно отметить команду *«Display clustered bar charts»* («Вывод на экран компьютера двухмерной столбиковой диаграммы») (см. рис. 4.5).

При нажатии кнопки «OK» в главном диалоговом окне «Таблицы сопряженности» (см. рис. 4.5) запускается процедура выполнения анализа.

4.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ПОСТРОЕНИЯ ТАБЛИЦ СОПРЯЖЕННОСТИ

После запуска процедуры выполнения анализа «Таблицы сопряженности» в *SPSS* на экран компьютера выводятся результаты анализа, которые объединены в две таблицы — «Обобщенные данные анализируемых случаев (ответов респондентов)» и собственно «Таблица сопряженности анализируемых переменных». Первая содержит данные о количестве пропущенных и действительных ответов, а также об общем количестве ответов (табл. 4.1).

Таблица 4.1

Обобщенные данные анализируемых случаев (ответов респондентов)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Сопровождающее лица* Форма размещения	5516	86,2%	880	13,8%	6396	100,0%

Согласно табл. 4.1 в рассматриваемом примере в проведении исследования участвовало 6396 респондентов. Из них 5516 (86,2%) указали форму своего размещения (проживания на отдыхе) и дали данные о сопровождающих их лицах. Сравнительно небольшое количество — 880 респондентов (13,8%) не дали ответа на поставленные вопросы или же выбрали вариант ответа «прочее», которые объединяют все ответы, не представленные в исследуемых вариантах ответов. К числу этих 880 респондентов также относятся туристы, отдыхающие без сопровождающих лиц.

Процент действительных значений (86,2%) весьма высок (см. табл. 4.1), что положительно характеризует качество исследований.

Непосредственно результаты анализа представлены в таблице сопряженности исследуемых переменных (табл. 4.2).

Таблица 4.2

Таблица сопряженности переменных «сопровождающие лица» и «форма размещения» (частоты в процентном выражении)

Сопровождающее лица * Форма размещения Crosstabulation

% within Сопровождающее лица

		Форма размещения					Total
		отель	пансион	арендованный дом (квартира)	размещение у друзей	кемпинг	
Сопровождающее лица	супруг(а)	30,4%	48,5%	17,6%	1,8%	1,7%	100,0%
	дети и супруг(а)	16,4%	52,9%	17,1%	,6%	12,9%	100,0%
	только дети	19,3%	40,4%	20,5%	,6%	19,3%	100,0%
	туристическая группа	68,5%	17,8%	4,1%		9,6%	100,0%
	друзья, родственники и пр.	33,9%	47,5%	11,5%	2,0%	5,2%	100,0%
	родители	20,0%	55,0%	20,0%		5,0%	100,0%
Total		28,1%	48,6%	16,6%	1,5%	5,2%	100,0%

Выявление взаимосвязи между исследуемыми переменными осуществляется путем визуального анализа таблицы сопряженности. *Между анализируемыми переменными не было бы никакой взаимосвязи, если бы частоты, выраженные в процентах, в разных столбцах были бы одинаковыми* (см. табл. 4.2). В рассматриваемом примере наблюдаются некоторые существенные отклонения от такого равенства.

Согласно данным табл. 4.2 проживание в пансионе — наиболее популярная форма размещения на отдыхе среди туристов, отдыхающих в курортной зоне «Баварский лес» (около 50%). Это справедливо почти для всех туристов, независимо от того, с кем они отдыхают. Однако у туристов, отдыхающих в составе туристической группы, этот вид размещения не так популярен (всего 17,8%).

Большинство туристов, отдыхающих в составе туристической группы, проживает в отеле (68,5%). Данный вид размещения на отдыхе не так сильно популярен у других групп, независимо от того, кто сопровождает их на отдыхе (16—30%).

Из табл. 4.2 также видно, что среди туристов, отдыхающих с детьми, сравнительно популярен кемпинг. Доля туристов, проживающих в кемпинге, составляет 19,3% для тех, кто отдыхает только с детьми, и 12,9% для тех, кто отдыхает всей семьей. Для туристов, отдыхающих без детей, этот процент существенно ниже.

Таким образом, в *результате визуального анализа частотной таблицы выявляется некая взаимосвязь между исследуемыми переменными.* Можно утверждать, что выбор формы размещения на отдыхе зависит от того, кто сопровождает отдыхающих.

Более тщательно исследовать существование зависимости между исследуемыми переменными позволяет анализ наблюдаемых и ожидаемых частот таблицы сопряженности.

4.4. АНАЛИЗ НАБЛЮДАЕМЫХ И ОЖИДАЕМЫХ ЧАСТОТ ТАБЛИЦ СОПРЯЖЕННОСТИ

Доказать существование зависимости между исследуемыми переменными можно от обратного. Другими словами, следует доказать отсутствие зависимости. Это можно сделать при помощи правила теории вероятности: «Два события считаются независимыми в том случае, если вероятность того, что они произойдут одновременно, равна произведению вероятностей того, что произойдет каждое из них».

Это правило можно объяснить на примере подбрасывания монет. Если выпадение «орла» является случайным (монета не деформирована, подбрасывающий ее человек не владеет специальной техникой), то вероятность этого события составляет 0,5. При одновременном подбрасывании двух монет вероятность одновременного выпадения двух «орлов» составит 0,25 (0,5 × 0,5). (Крыштановский, 2007. С. 48)

Таким образом, *в случае отсутствия зависимости между исследуемыми переменными на основе теории вероятности можно просчитать частоту, с которой встречаются те или иные ответы респондентов. Рассчитанные таким образом частоты называются ожидаемыми.*

Выявление зависимости между исследуемыми переменными проводится на основе данных, полученных в результате опроса

респондентов. Данные содержат информацию о том, как часто встречаются те или иные ответы на определенные вопросы. Такие частоты называются *наблюдаемыми*.

Существование или отсутствие зависимости между исследуемыми переменными выявляется путем сравнения наблюдаемых и ожидаемых частот. *Если наблюдаемые и ожидаемые частоты равны, то исследуемые переменные полностью независимы друг от друга.*

На практике редко встречаются случаи, когда правила теории вероятности выполняются с математической точностью. Например, если подкинуть одновременно две монеты 20 раз подряд, не следует ожидать, что одновременное выпадение двух «орлов» произойдет ровно 5 раз ($20 \times 0{,}25$). Возможны отклонения случайного характера. Поэтому при сравнении наблюдаемых и ожидаемых частот следует анализировать лишь то, насколько серьезно расходятся значения этих величин.

В *SPSS* задание на расчет наблюдаемых (*Observed*) и ожидаемых (*Expected*) частот дается в диалоговом окне «Таблицы сопряженности: Ячейки таблицы». Это делается в группе команд «*Counts*» (расчет частот в абсолютном выражении) (рис. 4.8).

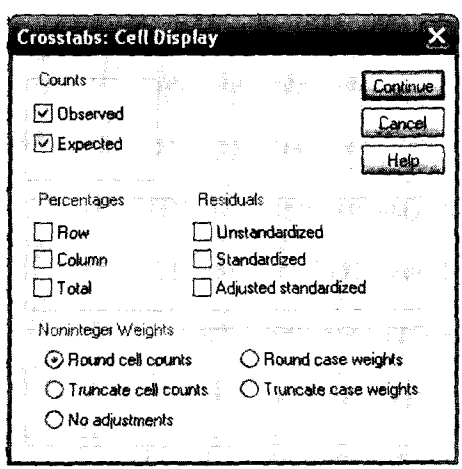


Рис. 4.8. Диалоговое окно «Таблицы сопряженности: Ячейки таблицы» (задание на расчет наблюдаемых и ожидаемых частот)

В диалоговом окне «Таблицы сопряженности: Ячейки таблицы» можно дать задание на расчет остатков (*Residuals*). Остатки характеризуют величину отклонения значений наблю-

даемых и ожидаемых частот. Возможно три варианта расчета остатков (см. рис. 4.8):

- *«Unstandardized»* (ненормированные) — разность наблюдаемых и ожидаемых частот;
- *«Standardized»* (нормированные) — разность наблюдаемых и ожидаемых частот, деленная на корень квадратный из ожидаемой частоты;
- *«Adjusted standardized»* (уточненные нормированные) — нормированные остатки вычисляются с учетом сумм по строкам и столбцам.

Остатки используются для визуального анализа таблицы. Результаты выполнения заданий в диалоговом окне «Таблицы сопряженности: Ячейки таблицы» (см. рис. 4.8) представлены в табл. 4.3.

Ожидаемая частота в таблице сопряженности рассчитывается как произведение сумм соответствующих строки и столбца, деленное на общую сумму частот. Например, ожидаемая частота в ячейке второй строки второго столбца составляет 558,7 ($1149 \times 2682/5516 = 558{,}7$). Это значение получилось в результате умножения суммы по второй строке (1149) на сумму по второму столбцу (2682) и деления этого произведения на общую сумму частот (5516) (см. табл. 4.3).

По данным табл. 4.3 видно, что в некоторых ячейках таблицы имеются существенные различия между наблюдаемыми и ожидаемыми частотами. Так, среди респондентов, отдыхающих с супругами и проживающих в кемпинге, фактически оказалось 59 туристов. Если бы не существовало связи между формой размещения на отдыхе и тем фактом, с кем отдыхает турист, то таких респондентов должно было бы быть 173 (почти в три раза больше). Такая же картина наблюдается в группе респондентов, живущих в кемпинге и отдыхающих с семьями или только с детьми.

Существенные различия наблюдаемых и ожидаемых частот выявлены у туристов, отдыхающих в составе туристической группы. Число таких туристов, проживающих в пансионе, фактически составляет 13 человек. Если бы не существовало связи между исследуемыми переменными, таких туристов должно было бы быть 36 человек (почти в три раза больше) (см. табл. 4.3).

Таким образом, в результате визуального анализа различия наблюдаемых и ожидаемых частот прослеживается наличие связи между исследуемыми переменными. И мы можем утверждать, что выбор формы размещения на отдыхе зависит от того, кто сопровождает отдыхающих.

Таблица 4.3

Таблица сопряженности переменных «сопровождающие лица» и «форма размещения» (наблюдаемые и ожидаемые частоты)

Сопровождающее лица * Форма размещения Crosstabulation

			Форма размещения					Total
			отель	пансион	арендованный дом (квартира)	размещение у друзей	кемпинг	
Сопровождающее лица	супруг(а)	Count	1017	1620	589	59	58	3343
		Expected Count	940,0	1625,4	554,5	49,7	173,3	3343,0
	дети и супруг(а)	Count	189	608	197	7	148	1149
		Expected Count	323,1	558,7	190,6	17,1	59,6	1149,0
	только дети	Count	32	67	34	1	32	166
		Expected Count	46,7	80,7	27,5	2,5	8,6	166,0
	туристическая группа	Count	50	13	3	0	7	73
		Expected Count	20,5	35,5	12,1	1,1	3,8	73,0
	друзья, родственники и пр.	Count	259	363	88	15	40	765
		Expected Count	215,1	372,0	126,9	11,4	39,7	765,0
	родители	Count	4	11	4	0	1	20
		Expected Count	5,6	9,7	3,3	,3	1,0	20,0
Total		Count	1551	2682	915	82	286	5516
		Expected Count	1551,0	2682,0	915,0	82,0	286,0	5516,0

Однако анализ наблюдаемых и ожидаемых частот не дает четкого ответа, насколько значительны (статистически значимы) выявленные отличия. Это можно пояснить с помощью приведенного примера с монетами. В случае одновременного выпадения «орлов» 17 раз при подбрасывании двух монет 20 раз подряд можно утверждать, что это происходит не случайно (например, монета деформирована или бросающий ее человек применяет специальную технику). Если же одновременное выпадение «орлов» наблюдалось 6 или 7 раз (при ожидаемой частоте 5), можно ли утверждать, что это произошло не случайно? То есть являются ли данные отклонения статистически значимыми?

Статистическая значимость отличий наблюдаемых и ожидаемых частот, позволяющая точно выявить наличие связи между исследуемыми переменными, определяется при помощи коэффициента «хи-квадрат» и других коэффициентов связи.

4.5. КОЭФФИЦИЕНТ «ХИ-КВАДРАТ» И ДРУГИЕ КОЭФФИЦИЕНТЫ СВЯЗИ

Коэффициент «хи-квадрат» используется для проверки гипотезы о независимости исследуемых переменных. Этот коэффициент рассчитывается с учетом ожидаемых и наблюдаемых частот (см. формулу (4)):

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \qquad (4)$$

где O_i — наблюдаемые частоты; E_i — ожидаемые частоты; n — число клеток в таблице.

Если $\chi^2 = 0$, можно однозначно говорить о полном совпадении ожидаемых и наблюдаемых частот, т.е. о полной независимости исследуемых переменных.

Если $\chi^2 > 0$ (меньше нуля он быть не может, поскольку разность частот возводится в квадрат, см. формулу (4)), следует использовать значение величины *Significance*, которое рассчитывается в *SPSS* автоматически с расчетом коэффициентов связи. Если значение *Significance* составит менее 0,05, гипотезу о независимости переменных следует отвергнуть, т.е. исследуемые переменные являются зависимыми.

Важность применения коэффициента «хи-квадрат» состоит в его универсальности. При использовании этого метода не суще-

ствует никаких требований к уровню измерения исследуемых переменных (см. раздел «Типы шкал измерения переменных»).

Коэффициенты связи, основанные на «хи-квадрат», появились в результате развития вышеизложенного метода проверки зависимости переменных. Это *коэффициенты сопряженности Пирсона и Крамера.* Значения этих коэффициентов лежат в интервале от нуля до единицы (в отличие от «хи-квадрат»), что делает процедуру анализа более удобной, хотя значение этих коэффициентов, близкое к единице, не свидетельствует о наличии зависимости между исследуемыми переменными. Это говорит лишь об уменьшении того уровня значимости (*Significance*), на котором отвергается гипотеза о независимости признаков.

Коэффициенты связи, основанные на прогнозе, используются также для проверки гипотезы о независимости исследуемых переменных. К таким коэффициентам относится λ (*лямбда*) и τ (*тау*) *Гудмена—Краскэла.* В основе применения этих коэффициентов лежит соображение, что если знание значений одной переменной улучшает предсказание значений другой переменной, эти две переменные взаимосвязаны.

Значения коэффициентов связи, основанных на прогнозе (в отличие от «хи-квадрат»), имеют практический смысл. Они характеризуют, насколько можно улучшить предсказание для одной переменной в ситуации знания значения другой по сравнению с незнанием, хотя для выявления взаимосвязи между переменными значения этих коэффициентов бессмысленны. Так же, как и в случае вычисления коэффициента «хи-квадрат», следует произвести оценку уровня значимости, на котором может быть отклонена гипотеза о независимости признаков. Если значение *Significance* составит менее 0,05, исследуемые переменные являются зависимыми.

Коэффициенты ранговой корреляции являются также коэффициентами связи, но применяются только для порядковых переменных. К ним относятся ρ (ро) *Спирмена,* τ *Кендэла,* γ *Гудмена—Краскэла.*

В отличие от всех перечисленных коэффициентов связи, коэффициенты ранговой корреляции фиксируют не только наличие либо отсутствие связи, но и в случае ее наличия — ее направление. Значение этих коэффициентов лежат в интервале от −1 до +1.

Недостатком использования ранговых коэффициентов корреляции является то, что они фиксируют наличие только однонаправленной, монотонной формы зависимости. Если значение

коэффициента ранговой корреляции близко к +1 (или −1), это свидетельствует о силе прямой (или обратной), но только монотонной формы зависимости.

Если значение рангового коэффициента корреляции равно нулю, это говорит не об отсутствии взаимосвязи переменных, а лишь об отсутствии монотонной связи. Для выявления наличия немонотонной связи необходимо дополнительно использовать коэффициент «хи-квадрат».

Коэффициент корреляции Пирсона является коэффициентом связи только для метрических переменных и выявляет только линейную форму зависимости (одну из форм монотонной зависимости) между исследуемыми переменными.

Значение коэффициента корреляции Пирсона, близкое к +1 (или −1), свидетельствует о силе прямой (или обратной), но только линейной формы зависимости.

Если значение коэффициента корреляции Пирсона равно нулю, это говорит не об отсутствии взаимосвязи переменных, а лишь об отсутствии линейной связи. Для выявления наличия нелинейной связи необходимо дополнительно использовать коэффициент «хи-квадрат».

Для задания команды расчета коэффициентов связи в *SPSS* в главном диалоговом окне «Таблицы сопряженности» (*Crosstabs*) (см. рис. 4.5) следует нажать кнопку «Статистические показатели» (*Statistics*). В результате этого открывается одноименное вспомогательное диалоговое окно (рис. 4.9).

В меню диалогового окна «Таблицы сопряженности: Статистические показатели» можно задать команды на расчет различных коэффициентов связи для всех типов шкал измерения переменных. Например, в результате отметки команды «Коэффициенты корреляции» (*Correlations*) (см. рис. 4.9) будет произведен расчет коэффициентов корреляции Пирсона и Спирмена.

В начале этой главы было отмечено, что данный вид анализа, как правило, не используется для метрических переменных и порядковых переменных с большим количеством градаций (возможных вариантов ответов). Построение таблиц сопряженности для таких переменных бессмысленно, поскольку они будут не пригодны для визуального анализа, но расчет коэффициентов может быть уместен.

В рассматриваемом примере задается только команда на расчет коэффициента «хи-квадрат» (*Chi-square*) (см. рис. 4.9). Результаты выполнения этой команда представлены в табл. 4.4.

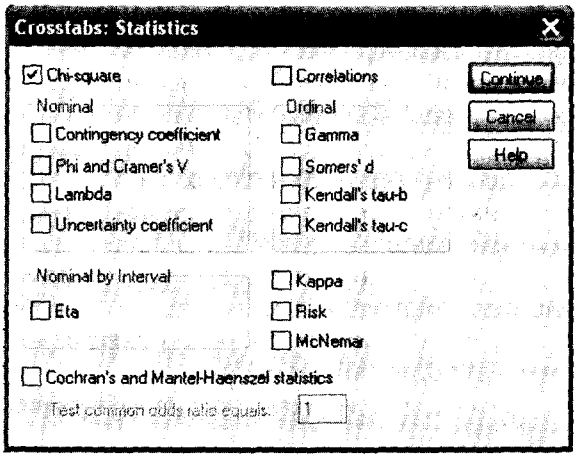


Рис. 4.9. Диалоговое окно «Таблицы сопряженности: Статистические показатели» (*Crosstabs*: *Statistics*)

Таблица 4.4

Результаты расчета коэффициента «хи-квадрат»

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	447,854(a)	20	,000
Likelihood Ratio	418,761	20	,000
Linear-by-Linear Association	37,985	1	,000
N of Valid Cases	5516		

a 6 cells (20,0%) have expected count less than 5. The minimum expected count is ,30.

Следует отметить, что существуют ограничения применения коэффициента «хи-квадрат». Его нельзя использовать, если в таблице сопряженности больше 20% клеток, в которых ожидаемая частота менее 5 и (или) есть клетки, в которых ожидаемая частота менее 1. В рассматриваемом примере необходимые требования не соблюдаются. Внизу табл. 4.4 существует надпись: «6 клеток (более 20%) содержат ожидаемые частоты со значениями менее 5. Значение минимальной ожидаемой частоты составляет 0,3» (см. табл. 4.3).

Для того чтобы стало возможным применение коэффициента «хи-квадрат», следует исключить из исследования одно или не-

сколько значений исследуемых переменных, частота выявления которых незначительна.

Если мы посмотрим на таблицу сопряженности исследуемых переменных, содержащую данные в абсолютном выражении (см. табл. 4.3), то увидим, что количество туристов, отдыхающих с родителями, составляет всего 20 человек из 5516 опрошенных. Следовательно, именно это значение переменной «сопровождающие лица» ответственно за появление слишком маленьких значений в таблице сопряженности. Следует исключить из исследований значение «Родители» переменной «сопровождающие лица». Подобное исключение неизбежно приведет к частичной, но незначительной потере информации.

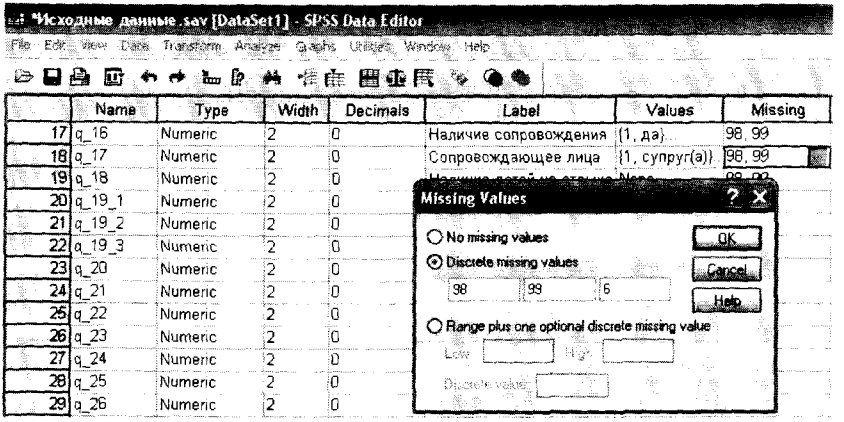


Рис. 4.10. Фрагмент вкладки «Variable View» («Свойства переменных»), диалоговое окно «Missing Values» («Пропущенные значения»)

Для исключения из исследования каких-либо данных в столбце «Пропущенные значения» (*Missing*) таблицы «Свойства переменных» следует щелкнуть курсором по точкам «...» (рис. 4.10). В результате этого на экране компьютера появляется диалоговое окно «Пропущенные значения» (*Missing Values*). (Возможности данного диалогового окна были рассмотрены в п. 2.1 «Структура редактора базы данных».) В рассматриваемом случае в поле «Дискретные пропущенные значения» (*Discrete missing values*) следует указать числовой код «6» значения «Родители» переменной «сопровождающие лица».

После запуска описанной процедуры расчета коэффициента «хи-квадрат» на экране компьютера появятся результаты анализа (табл. 4.5 и 4.6).

Таблица 4.5

Сводная таблица анализируемых случаев

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Сопровождающие лица * Форма размещения	5496	85,9%	900	14,1%	6396	100,0%

По данным табл. 4.5 видно, что на сей раз в анализе участвует только 5496 опрошенных (5516 – 20 = 5496), т.е. за вычетом 20 туристов, отдыхающих с родителями. Эти туристы оказываются среди 900 респондентов, участвовавших в опросе, но не участвующих в конкретном проводимом анализе данных.

Непосредственно результаты расчета коэффициента «хи-квадрат» представлены в табл. 4.6.

Таблица 4.6

Результаты расчета коэффициента «хи-квадрат»

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	446,603(a)	16	,000
Likelihood Ratio	417,347	16	,000
Linear-by-Linear Association	39,242	1	,000
N of Valid Cases	5496		

a 3 cells (12,0%) have expected count less than 5. The minimum expected count is 1,09.

Внизу табл. 4.6 находится надпись: «3 клетки (12%) содержат ожидаемые частоты со значениями менее 5. Значение минимальной ожидаемой частоты составляет 1,09». Следовательно, описанные выше ограничения применения коэффициента «хи-квадрат» соблюдаются.

Значение коэффициента «хи-квадрат» составляет 446,603 (см. табл. 4.6), что само по себе ни о чем не говорит. Значение величины *Significance* составляет 0,00 (см. табл. 4.6). Это означает, что гипотеза о независимости переменных может быть отклонена с вероятностью ошибки 0,00. Взаимосвязь между переменными «сопровождающие лица» и «форма размещения» доказана. Таким

образом, можно утверждать, что *присутствие сопровождающих лиц различного рода влияет на выбор формы размещения (проживания) на отдыхе.*

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем состоит сходство и различие между частотным анализом и таблицами сопряженности?
2. Какова цель построения таблиц сопряженности?
3. Каким требованиям относительно шкал измерения переменных должны соответствовать переменные, участвующие в построении таблиц сопряженности?
4. Охарактеризуйте использование данных таблиц сопряженности в процентном выражении. В каких случаях следует выбирать расчет процентного соотношения по столбцам, а в каких — по колонкам?
5. Что представляют собой наблюдаемые и ожидаемые частоты, содержащиеся в ячейках таблицы сопряженности?
6. Что характеризуют «остатки» (т.е. разность между наблюдаемыми и ожидаемыми частотами)?
7. Для чего служит коэффициент «хи-квадрат»? Каковы достоинства и недостатки его использования?
8. Что гласит исходная гипотеза, проверяемая в ходе применения коэффициента «хи-квадрат»?
9. Каковы необходимые условия применения коэффициента «хи-квадрат»? Какие действия необходимо предпринять в случае невыполнения этих условий?
10. Перечислите коэффициенты связи, рассчитываемые в *SPSS* для переменных разных уровней измерения (номинальных, порядковых и метрических).

5. СРАВНЕНИЕ СРЕДНИХ ВЕЛИЧИН В *SPSS*

Методы сравнения средних величин часто используются в маркетинговых исследованиях для выявления взаимосвязи между исследуемыми переменными. К таким методам относятся *T*-тесты и дисперсионный анализ.

В ходе проведения *T*-теста или дисперсионного анализа проверяется исходная (нулевая) гипотеза о равенстве сравниваемых средних величин, которая представляет собой утверждение: «Взаимосвязи между исследуемыми величинами не существует». Например, исходная (нулевая) гипотеза предполагает равенство среднего балла успеваемости студентов – юношей и девушек, что свидетельствует о том, что пол студента не влияет на его успеваемость. По результатам проведения анализа данная гипотеза должна быть подтверждена или опровергнута.

Основным результатом *T*-теста или дисперсионного анализа, выдаваемого *SPSS*, является величина «*Significance*» («Значимость»). Она характеризует вероятность ошибки, с которой может быть отклонена исходная (нулевая) гипотеза. Если вероятность ошибки мала, исходная гипотеза может быть отклонена, т.е. ее можно считать неверной (рис. 5.1).

Результаты проверки исходной (нулевой) гипотезы определяются доверительным интервалом (*Confidence Interval*), который задается при формировании задания на проведение *T*-теста или дисперсионного анализа в *SPSS*. По умолчанию устанавливается доверительный интервал 95%.

SPSS предоставляет возможность сравнения средних величин (*Compare Means*) при помощи различных методов (рис. 5.2).

При запуске процедуры «*Средние величины*» (*Means*) (см. рис. 5.2) определяются средние величины зависимых переменных в разных группах, сформированных по разным признакам (независимым переменным), а также вычисляются различные статистические показатели распределения зависимых переменных в группах (например, дисперсия, стандартное отклонение и др.).

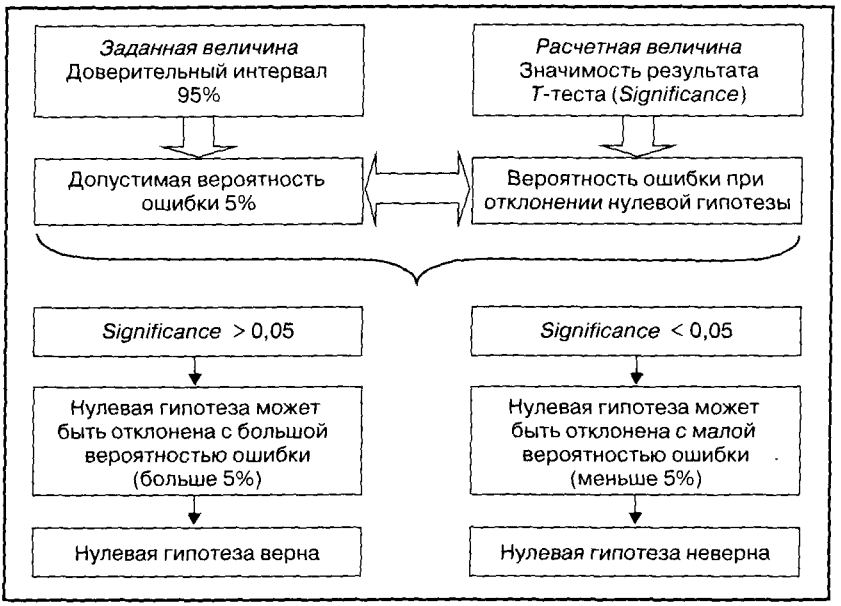


Рис. 5.1. *T*-тест: проверка исходной (нулевой) гипотезы

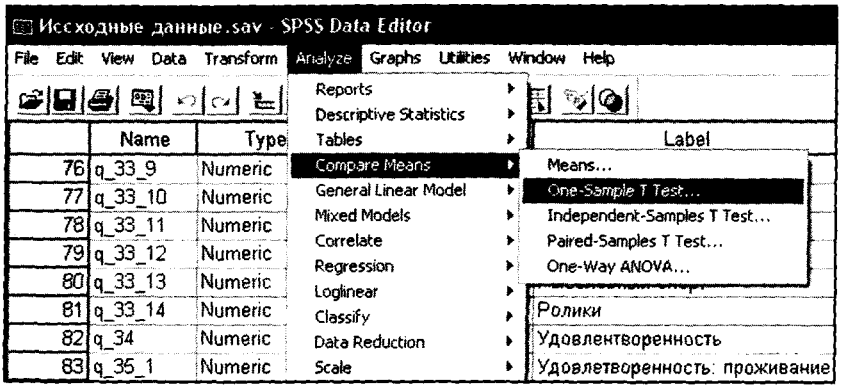


Рис. 5.2. Открытие меню «Сравнение средних величин»

Сравнение двух средних величин осуществляется при помощи *T*-теста. При этом ***T-тест для одной выборки*** (*One-Sample T-test*) используется для сравнения средней величины тестируемого признака в выборке с заданной стандартной величиной.

Например, с помощью этого теста можно определить, отличается ли средняя цена на определенный товар в заданной выборке предприятий торговли от средней цены, указанной каким-либо официальным источником.

T-тест для независимых выборок (*Independent Samples T-test*) (см. рис. 5.2) производится для сравнения средних величин тестируемого признака в двух группах. При этом каждый респондент может оказаться только в одной из двух групп, например: мужчины и женщины, семейные и несемейные, покупающие или не покупающие товар «*X*» и т.д.

T-тест для спаренных выборок (*Paired-Samples T-test*) (см. рис. 5.2) применяется для сравнения средних величин в двух группах, но при этом один и тот же респондент может одновременно оказаться в разных группах. Например, респонденты, покупающие товар *A* и товар *B*.

Для сравнения средних величин в трех и более группах применяется ***однофакторный дисперсионный анализ*** (*One-Way-ANOVA*) (см. рис. 5.2).

В данной работе подробно рассматривается *T*-тест для независимых выборок и однофакторный дисперсионный анализ.

5.1. T-ТЕСТ ДЛЯ НЕЗАВИСИМЫХ ВЫБОРОК

5.1.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В SPSS

T-тест проводится с целью определения влияния одной (независимой) переменной на другую (зависимую) переменную. По типу шкалы измерения зависимая переменная должна быть метрической, а независимая, как правило, является дихотомической (см. п. 2.3 «Типы шкал измерения переменных»).

Переменная дихотомического типа может принимать только два значения; следовательно, при помощи этой переменной можно разбить все анализируемые данные на две группы. Именно поэтому независимая переменная называется группирующей, а зависимая переменная – тестируемой.

Например, следует определить, влияет ли пол туриста на его удовлетворенность местом отдыха (рис. 5.3).

Независимая переменная «пол» является дихотомической, она разделяет всех респондентов на две группы: «мужчины» и «жен-

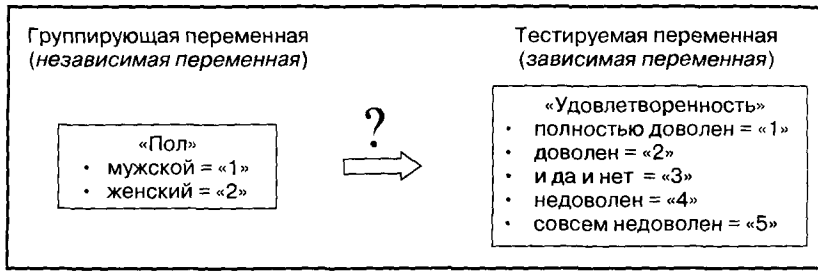


Рис. 5.3. *T*-тест для независимых выборок: постановка цели исследования

щины». В файле данных *SPSS*, сформированном по результатам опроса туристов, отдыхающих в курортной зоне «Баварский лес», эта переменная имеет имя «s_1» и метку переменной «Пол» (см. рис. 2.1, 2.3, 2.6).

В качестве зависимой переменной выступает переменная, представляющая вопрос анкеты № 34 «Оцените по 5-балльной шкале степень своей удовлетворенности местом отдыха в целом». В файле данных *SPSS*, сформированном по результатам опроса туристов, отдыхающих в курортной зоне «Баварский лес», эта переменная имеет имя «q_34» и метку переменной «Удовлетворенность» (рис. 5.4).

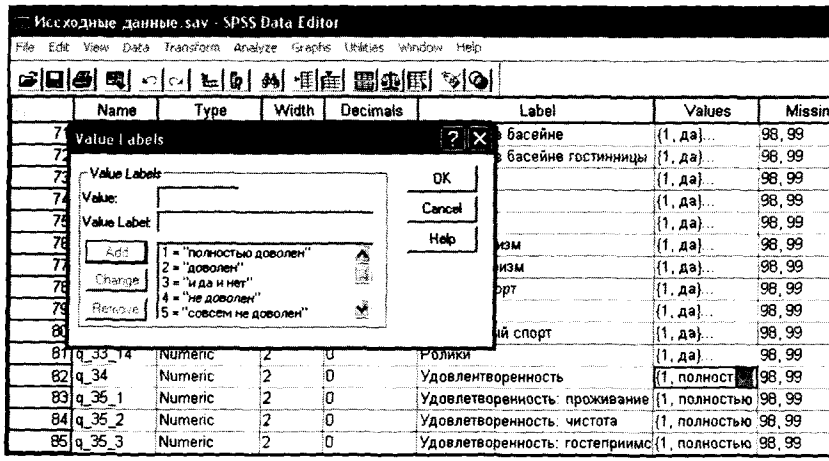


Рис. 5.4. Фрагмент вкладки «*Variable View*» («Свойства переменных»)

В столбце «*Values*» таблицы «Свойства переменных» отображаются значения метки переменной «Удовлетворенность» и их числовые коды.

Числовые коды значений переменной «Удовлетворенность» содержатся в таблице «Значения переменных» вкладки редактора данных «*Data View*» (рис. 5.5).

Из данных, представленных на рис. 5.5, видно, что респондент в строке 3525 доволен местом отдыха, респондент в строке 3526 – очень доволен, респондент в строке 3529 – доволен и недоволен одновременно, а респондент в строке 3530 затруднился ответить на поставленный вопрос.

Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

133 : nummer | 1340

	q_33_14	q_34	q_35_1	q_35_2	q_35_3	q
3525	.	2	1	1	1	
3526	.	1	1	1	1	
3527	.	2	2	2	.	
3528	.	1	1	1	1	
3529	.	3	3	2	2	
3530	.	98	1	1	.	
3531	.	2	2	1	2	
3532	.	3	2	2	.	

Рис. 5.5. Фрагмент вкладки «*Data View*» («Значения переменных»)

Для того чтобы выяснить, влияет ли пол туриста на степень его удовлетворенности местом отдыха, необходимо сравнить средний уровень удовлетворенности мужчин и женщин. Если данные показатели существенно отличаются друг от друга, то можно судить о наличии вышеуказанной взаимосвязи.

T-тест позволяет проверить верность гипотезы, согласно которой средние величины тестируемого показателя в двух группах равны. В рассматриваемом примере исходная (нулевая) гипотеза формулируется следующим образом:

«Мужчины и женщины в одинаковой степени довольны (недовольны) местом отдыха, т.е. пол туриста не влияет на его удовлетворенность местом отдыха».

В ходе проведения *T*-теста исходная гипотеза должна быть подтверждена или опровергнута. Для выполнения этого задания последовательно решаются две задачи:

- Проверяются условия равенства дисперсий тестируемой переменной в двух сравниваемых группах.
- Выявляются взаимосвязи между исследуемыми переменными, т.е. проверяются неравенства средних значений тестируемой переменной в двух сравниваемых группах.

5.1.2. КОМАНДЫ *SPSS* НА ВЫПОЛНЕНИЕ *T*-ТЕСТА ДЛЯ НЕЗАВИСИМЫХ ВЫБОРОК

При помощи выбора меню «*Analyze* > *Compare Means* > *Independent Samples T-test*» (см. рис. 5.2) открывается диалоговое окно «*T*-тест для независимых выборок» (рис. 5.6).

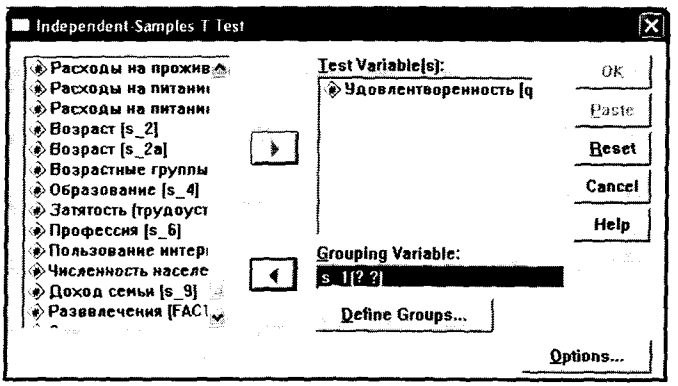


Рис. 5.6. Диалоговое окно «*T*-тест для независимых выборок».

В левом поле окна указываются метки всех переменных, занесенных в базу данных. Из данного списка выбирается метка тестируемой переменной (в рассматриваемом примере это «Удовлетворенность») и переносится в поле «*Test Variable(s)*». Далее из списка выбирается метка группирующей переменной (в рассматриваемом примере это «Пол») и переносится в поле «*Grouping Variable*».

Группирующая переменная делит всех респондентов на две группы: мужчины («1») и женщины («2»). Кодировку этих групп необходимо указать во вспомогательном диалоговом окне «Обозначенные группы» (*Define Groups*), которое открывается на-

жатием одноименной кнопки в диалоговом окне «*T*-тест для независимых выборок» (рис. 5.6 и 5.7).

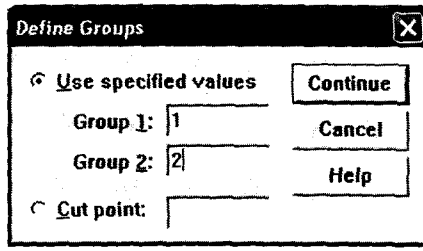


Рис. 5.7. Диалоговое окно «Обозначенные группы»

Группирующая переменная может быть не только дихотомической (т.е. принимающей только два значения), но и метрической, например переменная «возраст». В этом случае группы обозначаются при помощи указания «точки разрыва» (*Cut point*), например старше или младше указанного возраста.

Путем нажатия кнопки «Продолжение» (*Continue*) (см. рис. 5.7) снова активируется главное диалоговое окно «*T*-тест для независимых выборок» (см. рис. 5.6). При нажатии в этом окне кнопки «Опции» (*Options*) открывается одноименное диалоговое окно (рис. 5.8).

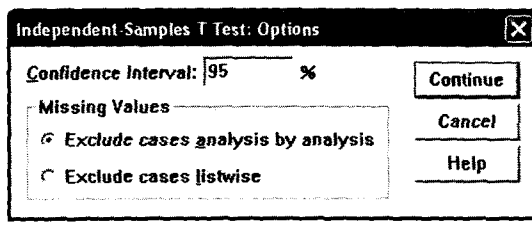


Рис. 5.8. Диалоговое окно «*T*-тест для независимых выборок: Опции»

В данном окне устанавливается «доверительный интервал» (*Confidence Interval*), который по умолчанию задается в размере 95%. Доверительный интервал определяет допустимую вероятность ошибки в случае отклонения исходной (нулевой) гипотезы (см. рис. 5.1).

После того как задан доверительный интервал путем нажатия кнопки «Продолжение» (*Continue*) осуществляется возврат в главное диалоговое окно «*T*-тест для независимых выборок» (см. рис. 5.6), где нажатием кнопки «*OK*» запускается процедура проведения *T*-теста.

5.1.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ *T*-ТЕСТА ДЛЯ НЕЗАВИСИМЫХ ВЫБОРОК

После запуска процедуры выполнения *T*-теста в *SPSS* на экран компьютера выводятся результаты анализа, которые объединены в две таблицы. Результаты *T*-теста для независимых выборок в рассматриваемом примере представлены в табл. 5.1 и 5.2. (Там, где данные в таблицах соответствуют пакету *SPSS*, ноль перед запятой не приводится – как в программном пакете. – *Примеч. ред.*)

Таблица 5.1

Результаты *T*-теста: статистические показатели в группах

Group Statistics

	Пол	N	Mean	Std. Deviation	Std. Error Mean
Удовлетворенность	Мужчины	3140	1,56	,541	,010
	Женщины	2823	1,55	,561	,011

Как видно из данных табл. 5.1, всего в исследованиях приняли участие 3140 мужчин и 2823 женщины. Средняя удовлетворенность местом отдыха (*Mean*) у мужчин составила 1,56, а у женщин – 1,55. На первый взгляд разница между средней удовлетворенностью местом отдыха у мужчин и женщин незначительная, однако неизвестно, является ли это отличие значимым с точки зрения статистики. На этот вопрос отвечают результаты *T*-теста, представленные в табл. 5.2.

Таблица 5.2

Результаты *T*-теста для независимых выборок

		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	95% Confidence Interval of the Difference	
									Lower	Upper
Удовлетворенность	Equal variance assumed	7,377	,007	1,060	5961	,289	,015	,014	–,013	,043
	Equal variance not assumed			1,058	6842,789	,290	,015	,014	–,013	,043

В результирующей таблице (см. табл. 5.2) представлено два варианта результатов *T*-теста в зависимости от равенства дисперсий распределения тестируемой переменной в разных группах:

1) дисперсии равны (*Equal variance assumed*);

2) дисперсии не равны (*Equal variance not assumed*).

Дисперсия характеризует «рассеивание» значений тестируемой переменной вокруг ее среднего значения, что является важным показателем при сравнении средних величин.

В зависимости от выполнения условия равенства дисперсий в ходе анализа результатов *T*-теста следует выбрать одну из строк табл. 5.2. Равенство дисперсий проверяется при помощи теста Ливина (*Levene's Test*)[1].

При проведении теста Ливина проверяется следующая гипотеза: «Дисперсии распределения тестируемой величины в разных группах равны». Верность этой гипотезы определяется в зависимости от величины «*Significance*» («Значимость»), которая в данном примере составляет 0,007 (см. табл. 5.2). Это означает, что исходная гипотеза может быть отклонена с вероятностью ошибки 0,7%, что существенно ниже допустимого уровня (5%). Следовательно, исходная гипотеза может быть отклонена.

Результаты теста Ливина показывают, что распределение тестируемой переменной в сравниваемых группах имеет разную дисперсию. При выборе одного из двух вариантов результата *T*-теста в табл. 5.2 следует выбрать вторую строку «*Equal variance not assumed*» («Равная вероятность не принимается»).

T-тест проверяет верность гипотезы: «Средние величины в двух группах равны». Верность этой гипотезы проверяется в зависимости от показателя «*Significance* (2-*tailed*)». В нашем примере этот показатель составляет 0,29 (см. табл. 5.2), что означает, что

[1] Тест Ливина (*Levene's Test*) – тест на равенство дисперсий в исследуемых группах. Дисперсия характеризует рассеяние значений переменной вокруг ее среднего значения. При сравнении средних величин в исследуемых группах необходимо принимать во внимание равенство или неравенство дисперсий.

Тест Ливина проверяет исходную (нулевую) гипотезу: «Дисперсии в исследуемых группах равны». Результат теста Ливина определяется значением показателя *Significance* («Значимость»). Если значение *Significance* не превышает 0,05, это означает, что вероятность ошибки при отклонении нулевой гипотезы составляет менее 5% (т.е. ниже допустимого уровня при доверительном интервале 95%). В этом случае исходная гипотеза может быть отклонена, т.е. она неверна. Значение *Significance* менее 0,05 доказывает ошибочность исходной гипотезы и статистическую значимость различия дисперсий в исследуемых группах.

исходная гипотеза может быть отклонена с вероятностью ошибки 29%, а это выше допустимого уровня (5%). Следовательно, исходная гипотеза не может быть отклонена, т.е. сравниваемые средние величины равны с точки зрения статистики.

Исходя из вышеизложенного можно сделать вывод о том, что разница между средним уровнем удовлетворенности местом отдыха у мужчин и женщин (1,56 и 1,55) не является статистически значимой.

Таким образом, в результате проведения Т-теста доказано отсутствие взаимосвязи между исследуемыми переменными, т.е. пол туриста не влияет на его удовлетворенность местом отдыха.

5.2. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

5.2.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В SPSS

Однофакторный дисперсионный анализ проводится с целью определения влияния одной (независимой) переменной на другую (зависимую) переменную. С точки зрения чистоты статистических расчетов независимая переменная должна быть номинальной или порядковой, а зависимая – метрической (см. п. 2.3 «Типы шкал измерения переменных»).

Номинальная или порядковая переменная может принимать несколько значений (более двух); следовательно, при помощи этой переменной можно разделить все анализируемые данные на несколько групп (категорий). Именно поэтому независимая переменная называется ***категориальным фактором***.

Например, следует определить, влияет ли удаленность места отдыха от места жительства туриста на выбор транспорта для проезда до места отдыха.

Для проведения такого исследования из всех вопросов анкеты, которая была использована для опроса туристов, отдыхающих в курортной зоне «Баварский лес», выбираются два вопроса:

- Вопрос № 12: «Укажите, каким видом транспорта Вы пользовались, для того чтобы добраться до места отдыха». (В качестве вариантов ответа на выбор предлагается 10 видов транспорта.)
- Вопрос № 14: «Укажите расстояние (в километрах), которое Вам пришлось преодолеть, чтобы добраться от места жительства до места отдыха».

Вопрос № 12 *является номинальной* переменной, а вопрос № 14 – метрической (см. п. 2.3 «Типы шкал измерения переменных»). В соответствии с требованиями, предъявляемыми к переменным, участвующим в дисперсионном анализе, вопрос № 12 («Вид транспорта») должен выступать в качестве независимой переменной, а вопрос № 14 («Удаленность места отдыха») – в качестве зависимой.

При указанном выше распределении ролей зависимой и независимой переменных должна поменяться формулировка вопроса исследования. Вместо исходной формулировки «Влияет ли удаленность места отдыха на выбор вида транспорта» следует использовать другую формулировку: «Существует ли взаимосвязь между удаленностью места отдыха и выбором вида транспорта» (рис. 5.9).

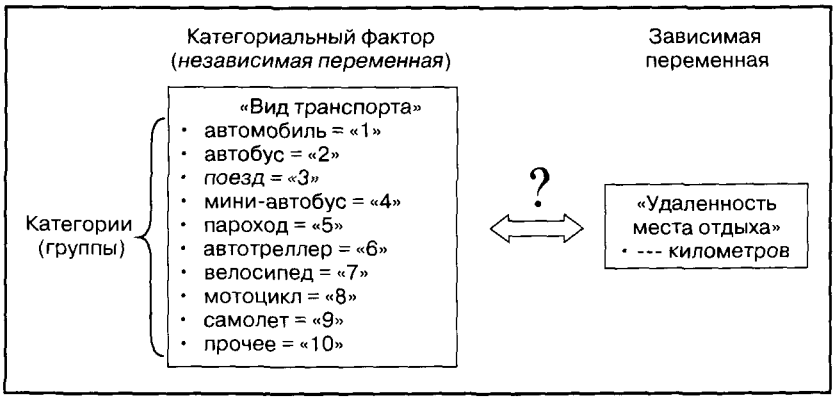


Рис. 5.9. Однофакторный дисперсионный анализ.
Постановка цели исследования

В файле данных *SPSS*, содержащем результаты опроса туристов, вопрос № 12 представлен в виде переменной с именем «q_12» и меткой «Вид транспорта». Вопрос № 14 представлен в виде переменной с именем «q_14» и меткой «Удаленность места отдыха» (рис. 5.10).

Поскольку переменная «Вид транспорта» является номинальной, в столбце «*Values*» таблицы «Свойства переменных» указываются значения метки переменной с их числовыми кодами. Переменная «Удаленность места отдыха» является метрической, поэтому значения метки переменной в столбце «*Values*» отсутствуют.

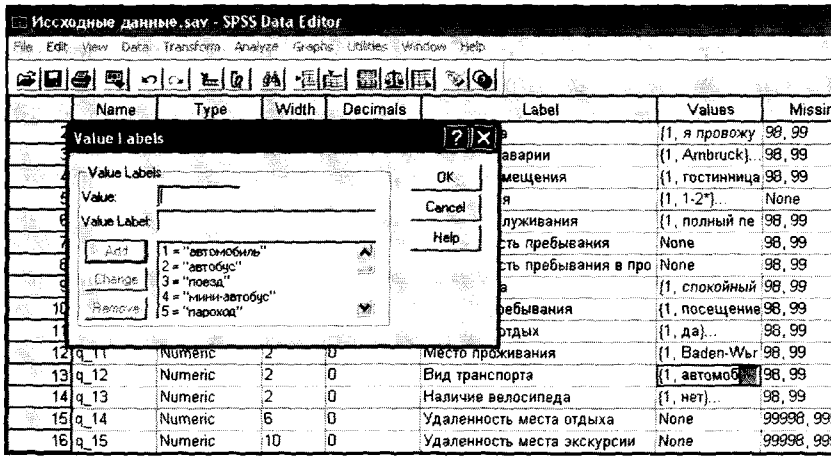
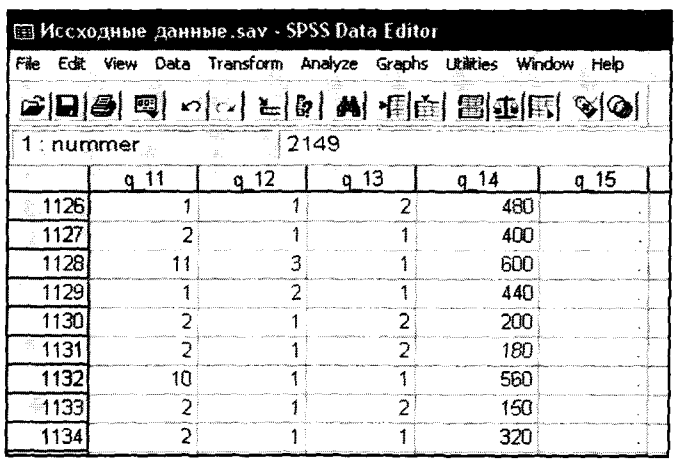


Рис. 5.10. Фрагмент вкладки «*Variable View*» («Свойства переменных»)

Фрагмент таблицы «Значения переменных» содержит данные об ответах респондентов на вопрос анкеты № 12 (столбец «q_12») и вопрос № 14 (столбец «q_14»). Респондент в строке 1127 прибыл на место отдыха на автомобиле и преодолел расстояние 400 км. Респондент в строке 1128 приехал на поезде и при этом преодолел расстояние 600 км. Респондент в строке 1129 приехал к месту отдыха на автобусе за 440 км (рис. 5.11).



	q_11	q_12	q_13	q_14	q_15
1126	1	1	2	480	.
1127	2	1	1	400	.
1128	11	3	1	600	.
1129	1	2	1	440	.
1130	2	1	2	200	.
1131	2	1	2	180	.
1132	10	1	1	560	.
1133	2	1	2	150	.
1134	2	1	1	320	.

Рис. 5.11. Фрагмент вкладки «*Data View*» («Значения переменных»)

Для того чтобы выявить взаимосвязь между удаленностью места отдыха и выбором вида транспорта, необходимо сравнить среднее расстояние, преодолеваемое туристами при помощи различных видов транспорта. Если сравниваемые расстояния существенно отличаются друг от друга, то можно судить о наличии вышеуказанной взаимосвязи.

Однофакторный дисперсионный анализ проверяет верность гипотезы, согласно которой средние величины более чем в двух группах равны. В рассматриваемом примере исходная (нулевая) гипотеза принимает следующую формулировку: *«Туристы, использующие различные виды транспорта, преодолевают в среднем одинаковое расстояние от места проживания до места отдыха»*. Иными словами: *«Не существует связи между выбором вида транспорта и дальностью преодолеваемого расстояния»*.

В результате однофакторного дисперсионного анализа исходная (нулевая) гипотеза должна быть подтверждена или опровергнута. В ходе анализа последовательно решаются три задачи:

- Проверяются условия равенства дисперсий зависимой переменной в нескольких сравниваемых группах (категориях).
- Выявляются взаимосвязи между исследуемыми переменными, т.е. проверяются неравенства средних значений зависимой переменной в нескольких сравниваемых группах (категориях).
- В случае выявления взаимосвязи определяется, каким образом именно категории (группы) обусловливают данную взаимосвязь, т.е. в каких именно группах (категориях) средние значения зависимой переменной не равны.

5.2.2. КОМАНДЫ SPSS НА ВЫПОЛНЕНИЕ ОДНОФАКТОРНОГО ДИСПЕРСИОННОГО АНАЛИЗА

При выборе меню «*Analyze* > *Compare Means* > *One-Way-ANOVA*» (см. рис. 5.2) открывается диалоговое окно «Однофакторный дисперсионный анализ» (рис. 5.12).

В левом поле окна указываются метки всех переменных, занесенных в базу данных. Из списка всех переменных выбирается категориальный фактор, т.е. переменная, разбивающая данные на группы (категории) (в рассматриваемом примере – «Вид транспорта»), и переносится в поле «*Factor*».

Далее из списка переменных выбирается другая исследуемая переменная – «зависимая переменная» (в рассматриваемом примере – «Удаленность места отдыха») – и переносится в поле «*Dependent List*».

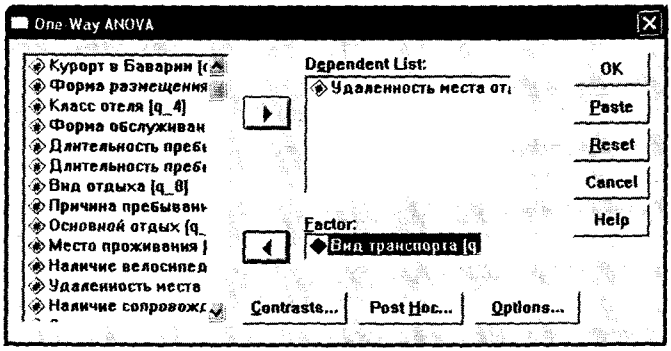


Рис. 5.12. Диалоговое окно «Однофакторный дисперсионный анализ»

Прежде чем нажать кнопку «OK» и запустить процесс выполнения анализа, следует задать дополнительные команды во вспомогательном диалоговом окне «Опции», которое открывается путем нажатия одноименной кнопки в главном диалоговом окне «Однофакторный дисперсионный анализ» (рис. 5.12 и 5.13).

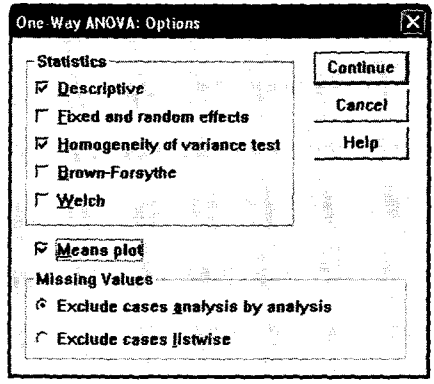


Рис. 5.13. Диалоговое окно «Опции»

В диалоговом окне «Опции» путем отметки напротив команды «*Descriptive*» (см. рис. 5.13) можно задать вывод на экран статистических показателей, описывающих распределение переменной «Удаленность места отдыха» в разных группах респондентов, сформированных в зависимости от используемого ими вида транспорта. Результат выполнения этой команды будет представлен далее в табл. 5.3.

В ходе выполнения однофакторного дисперсионного анализа следует проверить равенство дисперсий распределения зависимой переменной в сравниваемых группах. Равенство дисперсий проверяется при помощи теста Ливина. Для задания на выполнение данного теста в диалоговом окне «Опции» следует сделать отметку напротив команды «*Homogeneity of variance test*» (см. рис. 5.13). Результаты проведения теста Ливина будут представлены далее в табл. 5.4.

Также в рассматриваемом примере в диалоговом окне «Опции» поставлена отметка напротив команды «*Means Plot*». В результате выполнения этой команды на экран компьютера выводится график средних значений зависимой переменной в разных группах (категориях). Этот график будет представлен ниже в рассматриваемом примере среди результатов анализа (рис. 5.15).

При нажатии кнопки «*Continue*» в диалоговом окне «Опции» осуществляется возврат в главное диалоговое окно «Однофакторный дисперсионный анализ» (см. рис. 5.12).

В результате проведения однофакторного дисперсионного анализа согласно описанному выше заданию, т.е. перечню команд, решаются две задачи:

- Проверяются равенства дисперсий значений зависимой переменной в сравниваемых группах (категориях).
- Выявляются взаимосвязи между исследуемыми переменными, т.е. приводится доказательство неравенства средних значений зависимой переменной в сравниваемых группах (категориях).

Решение этих задач не является окончательным результатом однофакторного дисперсионного анализа. Если взаимосвязь между исследуемыми переменными существует, то следует выяснить, какие именно категории обусловливают существование этой связи. Иными словами, если выясняется, что средние значения зависимой переменной в сравниваемых группах (категориях) не равны, то необходимо выяснить – в каких именно. Для решения этой задачи проводятся дополнительные исследования (*Post Hoc*).

Для выяснения того, в каких же именно группах (категориях) средние значения зависимой переменной в наибольшей степени отличаются друг от друга, проводится многовариантное сравнение этих значений (*Multiple Comparison*). Данный процесс осуществляется путем проведения так называемых апостериорных тестов.

Команда на проведение апостериорного теста задается в дополнительном диалоговом окне «Последующие многовариант-

ные сравнения» (рис. 5.14), которое открывается путем нажатия кнопки «*Post Hoc*» в главном диалоговом окне «Однофакторный дисперсионный анализ».

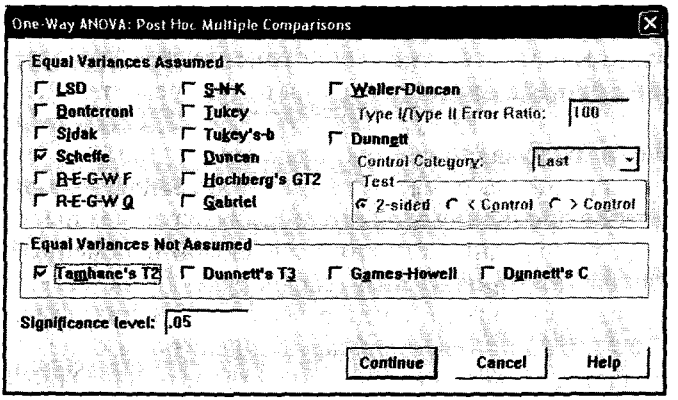


Рис. 5.14. Диалоговое окно «Последующие многовариантные сравнения»

В диалоговом окне «Последующие многовариантные сравнения» представлены различные апостериорные тесты, позволяющие выявить группы, ответственные за результаты дисперсионного анализа.

Правильный выбор апостериорного теста возможен только после проведения теста Ливина на равенство дисперсий. Некоторые тесты могут применяться только в том случае, если дисперсии зависимой переменной в сравниваемых группах (категориях) равны (*Equal Variances Assumed*). Другие тесты, наоборот, применяются в том случае, если дисперсии значений зависимой переменной в разных группах (категориях) не равны (*Equal Variances Not Assumed*).

В целях сокращения количества итераций рекомендуется задать команды на проведение сразу двух тестов: на случай равенства и неравенства дисперсий зависимой переменной в сравниваемых группах. При интерпретации результатов анализа будет выбран нужный вариант.

В рассматриваемом примере в диалоговом окне «Последующие многовариантные сравнения» заданы команды на проведение теста «*Scheffe*» и «*Tamhane*», которые наиболее часто используются на практике. После нажатия кнопки «*Continue*» в этом же

алоговом окне осуществляется возврат в главное диалоговое но «Однофакторный дисперсионный анализ».

Путем нажатия кнопки «*OK*» в главном диалоговом окне запускается процедура выполнения однофакторного дисперсионного анализа.

5.2.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ОДНОФАКТОРНОГО ДИСПЕРСИОННОГО АНАЛИЗА

Среди результатов выполнения задания (перечня команд), описанного выше, в первую очередь на экран компьютера выводится таблица «*Descriptives*». Она содержит различные статистические показатели, описывающие распределение зависимой переменной в разных группах (категориях). В рассматриваемом примере зависимой переменной является расстояние, преодолеваемое туристами от места проживания до места отдыха. Группами (категориями) являются группы туристов, пользующихся различными видами транспорта для проезда до места отдыха (табл. 5.3).

Таблица 5.3

Статистические показатели, описывающие распределение зависимой переменной («Удаленность места отдыха»)

Descriptives

Вид транспорта	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Автомобиль	5053	481,96	198,242	2,789	476,50	487,43	35	2500
Автобус	148	543,35	249,048	20,472	502,89	583,81	70	2000
Поезд	281	518,14	234,878	14,012	490,56	545,72	65	1000
Мини-автобус	96	494,68	532,545	54,353	386,77	602,58	60	5000
Пароход	1	670,00	–	–	–	–	670	670
Автотреллер	108	497,04	245,210	23,595	450,26	543,81	100	1000
Велосипед	10	352, 00	309,365	97,830	130,69	573,31	90	1170
Мотоцикл	22	569,95	355,116	75,711	412,50	727,40	200	1780
Самолет	24	3194,79	2845,452	580,825	1993,26	4396,32	550	9000
Прочее	17	264,82	322,801	78,291	98,85	430,79	35	1301
Всего	5760	496,61	329,542	4,342	488,09	505,12	35	9000

Статистические показатели, отображенные в таблице «*Descriptives*»:

- число наблюдений (респондентов) в одной группе (*N*);
- среднее значение (*Mean*);
- стандартное отклонение (*Std. Deviation*);
- стандартная ошибка (*Std. Error*);
- доверительный интервал (*Confidence Interval*);
- минимальное и максимальное значения, т.е. самое короткое и самое длинное расстояние, преодолеваемое туристами с помощью определенного вида транспорта (*Minimum*, *Maximum*).

Таблица «*Descriptives*» дает лишь общее представление о распределении значений зависимой переменной в разных группах (категориях). Основная задача данного этапа исследований – проверка практической значимости сформированных групп (категорий).

Как видно из данных табл. 5.3, из числа всех респондентов, давших ответы на вопросы анкеты № 12 и № 14 (5760 человек), только один воспользовался пароходом, чтобы добраться до места отдыха. Данная группа («туристы, выбирающие пароход») должна быть исключена из исследований, поскольку она состоит только из одного респондента и является практически незначимой.

Следующим шагом представления результатов однофакторного дисперсионного анализа является представление результатов проверки равенства дисперсий в сравниваемых группах (категориях), т.е. результатов теста Ливина (табл. 5.4).

Таблица 5.4

Результаты теста Ливина

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
756,669[a]	8	5750	,000

[a] Группы с единичным результатом не учитываются при расчетах для проверки равенства дисперсий «Удаленность места отдыха».

Тест Ливина позволяет проверить верность гипотезы: «Дисперсии в рассматриваемых группах равны». Значение расчетного показателя «*Significance*» в данном случае равно 0,000. Это означает, что исходная гипотеза может быть отклонена с вероятностью ошибки 0%, т.е. гипотеза неверна. Это доказывает, что дисперсии зависимой величины «Удаленность места отдыха» в сравнива-

емых группах (категориях) туристов, воспользовавшихся разными видами транспорта, не равны.

После результатов проверки равенства дисперсий в сравниваемых группах на экран выводятся результаты однофакторного дисперсионного анализа (*ANOVA: Analysis of Variance*) (табл. 5.5).

Таблица 5.5

Результаты однофакторного дисперсионного анализа

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	177533022,851	9	19725891,43	253,245	,000
Within Groups	447881263,272	5750	77892,394		
Total	625414286,122	5759			

Как отмечалось ранее, однофакторный дисперсионный анализ позволяет проверить вероятность исходной (нулевой) гипотезы о равенстве средних величин более чем в двух группах. В рассматриваемом примере нулевая гипотеза представляет собой утверждение: «Туристы, пользующиеся различными видами транспорта, в среднем преодолевают равные расстояния от места проживания до места отдыха».

Равенство средних значений зависимой переменной в разных группах (категориях) свидетельствует об отсутствии взаимосвязи между исследуемыми переменными. В рассматриваемом примере исходная (нулевая) гипотеза может также быть представлена в виде утверждения: «Не существует взаимосвязи между удаленностью места отдыха и выбором вида транспорта».

Верность исходной (нулевой) гипотезы проверяется при помощи расчетной величины «*Significance*», которая в рассматриваемом примере составляет 0,000. Это означает, что исходная гипотеза может быть отклонена с вероятностью ошибки 0%, т.е. она неверна.

Из результатов анализа данных, представленных в табл. 5.5, можно сделать вывод, что *средняя дальность проезда до места отдыха является различной в отдельных группах туристов, выделенных по виду используемого транспорта. Это доказывает существование взаимосвязи между исследуемыми переменными, т.е. удаленность места отдыха влияет на выбор вида транспорта.*

Представленные выше выводы не являются окончательным результатом однофакторного дисперсионного анализа. Для получения более точных результатов исследования необходимо выяснить, в каких именно группах отличия средней длительности проезда наиболее значительны. Это возможно путем проведения дополнительных исследований, а именно апостериорного теста.

В рассматриваемом примере была задана команда на проведение тестов «*Scheffe*» и «*Tamhane*» (см. рис. 5.14). Поскольку в результате проведения теста Ливина было выявлено неравенство дисперсий, в рассматриваемом примере значимыми являются только результаты теста «*Tamhane*».

Апостериорные тесты могут проводиться только в том случае, если сравниваемые группы (категории) включают в себя как минимум два наблюдения. В рассматриваемом примере только один из опрошенных туристов воспользовался пароходом для того, чтобы добраться до места отдыха. Для проведения апостериорного теста сначала необходимо исключить данную группу (категорию) из исследований.

Для исключения из исследований группы туристов, воспользовавшихся пароходом, во вкладке редактора данных «Свойства переменных» (*Variable View*) в столбце «*Missing*» следует указать числовой код значения метки «Пароход» – «5» (см. рис. 5.10), после чего следует пересчитать результаты анализа, повторив все операции по заданию команд на выполнение анализа, описанные в предыдущем разделе.

Среди результатов повторного анализа на экране появятся результаты теста «*Tamhane*» (табл. 5.6). Здесь представлены результаты попарного сравнения средней дальности проезда до места отдыха для разных групп туристов, сформированных по виду используемого транспорта. Сначала сравнивается средняя дальность проезда до места отдыха на автомобиле со средней дальностью проезда другими видами транспорта, затем средняя дальность проезда до места отдыха на автобусе и т.д. Пары, характеризующиеся значительным различием средних величин, обозначаются звездочкой (*).

Как видно из результатов теста «*Tamhane*», среднее расстояние до места отдыха, преодолеваемое на самолете, существенно отличается от среднего расстояния проезда до места отдыха другими видами транспорта.

Таблица 5.6

Фрагмент таблицы с результатами теста «*Tamhane*»

Multiple Comparsions

(I) Вид транспорта	(J) Вид транспорта	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Автомобиль	Автобус	–61,387	20,661	,117	–128,52	5,74
	Поезд	–36,174	14,286	,349	–82,17	9,82
	Мини-автобус	–12,712	54,424	1,000	–191,53	166,11
	Автотреллер	–15,072	23,760	1,000	–92,83	62,69
	Велосипед	129,965	97,870	1,000	–313,22	573,15
	Мотоцикл	–87,990	75,762	1,000	–365,98	190,00
	Самолет	–2712,827*	580,832	,004	–4817,75	–607,91
	Прочее	217,141	78,340	,389	–84,09	518,38
Автобус	Автомобиль	61,387	20,661	,117	–5,74	128,52
	Поезд	25,213	24,808	1,000	–54,70	105,13
	Мини-автобус	48,674	58,080	1,000	–140,91	238,26
	Автотреллер	46,314	31,238	,996	–54,54	147,17
	Велосипед	191,351	99,949	,959	–247,33	630,03
	Мотоцикл	–26,603	78,430	1,000	–309,01	255,80
	Самолет	–2651,440*	581,186	,005	–4756,96	–545,92
	Прочее	278,528	80,923	,098	–25,25	582,31
Поезд	Автомобиль	36,174	14,286	,349	–9,82	82,17
	Автобус	–25,213	24,808	1,000	–105,13	54,70
	Мини-автобус	23,462	56,130	1,000	–160,33	207,25
	Автотреллер	21,102	27,442	1,000	–67,76	109,96
	Велосипед	166,139	98,828	,992	–274,77	607,05
	Мотоцикл	–51,816	76,997	1,000	–331,75	228,12
	Самолет	–2676,653*	580,994	,004	–4781,85	–571,46
	Прочее	253,315	79,535	,177	–48,96	555,59
Мини-автобус	Автомобиль	12,712	54,424	1,000	–166,11	191,53

Аналогичные выводы можно сделать, руководствуясь графическим изображением зависимости средней дальности проезда до места отдыха и видом используемого транспорта (рис. 5.15). Построение данного графика было задано с помощью отметки напротив команды «График средних величин» (*Means plot*) в диалоговом окне «Опции» (см. рис. 5.13).

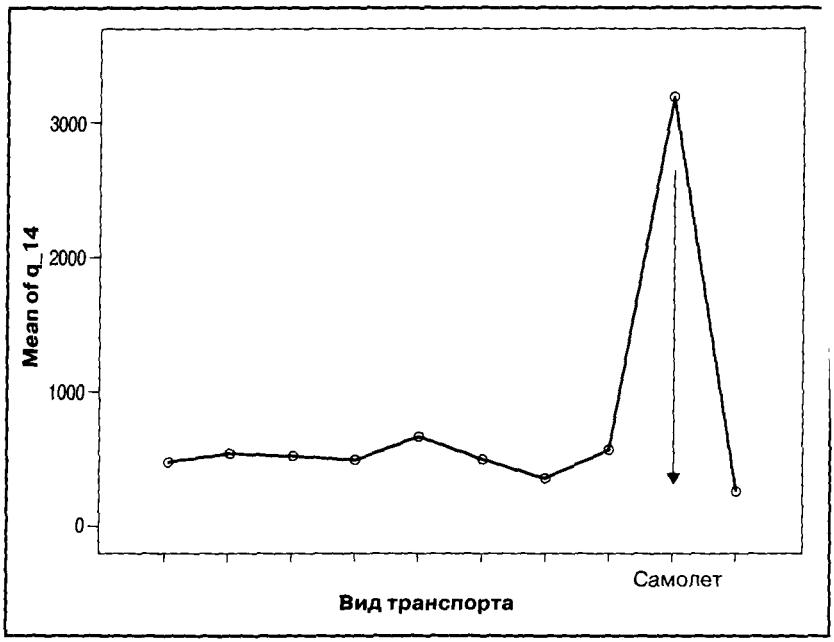


Рис. 5.15. Зависимость вида транспорта от средней дальности проезда

Графическое представление зависимости выбора вида транспорта от средней дальности проезда показывает, что для преодоления больших расстояний (в среднем 3000 км) туристы выбирают самолет. Средние расстояния, преодолеваемые с помощью других видов транспорта (поезд, автобус, автомобиль и др.), не намного отличаются друг от друга (примерно 500–700 км).

В итоге можно сделать вывод, что удаленность места отдыха от места жительства туриста в целом влияет на выбор вида транспорта. Если турист должен преодолеть длинный путь (более 1000 км), то он, скорее всего, выберет самолет. Если же расстояние до места отдыха менее 1000 км, то турист воспользуется другим транспортом, при этом дальность переезда не определяет выбор вида транспорта.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие методы статистического анализа основываются на сравнении средних величин?

2. Как звучит исходная (нулевая) гипотеза, проверяемая в ходе сравнения средних величин, и при помощи какого показателя определяется ее верность?
3. Каковы цели проведения и возможности применения результатов *T*-тестов и дисперсионного анализа?
4. Какие требования предъявляются к переменным, участвующим в проведении *T*-тестов и дисперсионного анализа, относительно типов шкал измерения переменных?
5. Назовите основные виды *T*-тестов и дисперсионного анализа и укажите, в чем состоит различие между ними.
6. Для чего проводится тест Ливина и как его результаты используются при интерпретации результатов *T*-теста?
7. Для чего и каким образом производится проверка практической значимости исходных данных однофакторного дисперсионного анализа?
8. Каким образом производится исключение из исследований, проводимых в *SPSS*, исходных данных для однофакторного дисперсионного анализа, которые оказались практически незначимыми?
9. Как влияют результаты теста Ливина на ход проведения однофакторного дисперсионного анализа?
10. Для чего при проведении однофакторного дисперсионного анализа производятся апостериорные тесты (*Post Hoc Multiple Comparisons*)?
11. С какой целью и при помощи какой команды *SPSS* строится график средних величин при проведении однофакторного дисперсионного анализа?

6. ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ В *SPSS*

Регрессионный анализ служит для выявления влияния одной или нескольких независимых переменных на одну зависимую переменную. С точки зрения чистоты статистических расчетов в регрессионном анализе могут участвовать лишь метрические, т.е. количественные, переменные (см. п. 2.3 «Типы шкал измерения переменных»). Однако в некоторых учебниках по *SPSS* указывается, что в регрессионном анализе могут участвовать как метрические, так и порядковые переменные.

Дихотомические переменные (имеющие только два значения: высокий/низкий, близкий/далекий и т.п.) могут рассматриваться как метрические. В случае необходимости использовать в регрессионном анализе номинальные переменные их следует разложить на дихотомические переменные (см. п. 2.2 «Виды кодировки данных»).

Регрессионный анализ позволяет не только сделать вывод о существовании взаимосвязи между исследуемыми переменными, но и дать математическое описание зависимости между ними.

Современные методы статистического анализа позволяют давать математическое описание зависимости переменных, выраженных в функциях различных видов. Техника регрессионного анализа, позволяющая выявлять и описывать взаимосвязи в виде линейных функций, называется ***линейным регрессионным анализом*** (см. п. 1.2 «Основные виды статистического анализа»).

Для выявления и описания линейной зависимости между объектом исследования (зависимой переменной) и одним фактором, возможно влияющим на него (независимой переменной), используется ***простая линейная регрессия***. Регрессионная модель (регрессионное уравнение) в этом случае имеет вид

$$y = a + bx,$$

где y – зависимая переменная;
x – независимая переменная;
a – свободный член (константа);
b – коэффициент регрессии.

Для выявления и описания линейной зависимости между объектом исследования (зависимой переменной) и несколькими факторами, возможно на него влияющими (независимыми переменными), используется ***множественная линейная регрессия***. Регрессионная модель (регрессионное уравнение) в этом случае имеет вид

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n.$$

Результатом регрессионного анализа является регрессионная модель (регрессионное уравнение), а именно – определение ***свободного члена*** (*a*) и ***коэффициентов регрессии*** (*b*).

Также в ходе регрессионного анализа определяются ***стандартизированные коэффициенты регрессии*** (*Beta*). Данные коэффициенты позволяют судить о значении соответствующих независимых переменных (*x*), т.е. о степени влияния на зависимую переменную (*y*).

Результатом регрессионного анализа является не только регрессионная модель, но также расчет ряда показателей, характеризующих статистическую значимость и практическую применимость построенной модели. Среди таких показателей в качестве основных можно выделить:

- ***Коэффициент детерминации*** (***R***) – является характеристикой общей силы линейной связи между переменными в регрессионной модели. Значения коэффициента находятся в интервале от нуля до единицы. Чем ближе значение коэффициента к единице, тем плотнее линейная взаимосвязь, описанная в регрессионной модели. В общем случае он должен превышать 0,5.
- ***Коэффициент R-квадрат*** (***R Square***) – показывает, какая доля совокупной вариации в зависимой переменной описывается независимой переменной. Значения коэффициента лежат в интервале от нуля до единицы. Как правило, данный показатель должен превышать 0,5. Если он равен 0,5, это говорит о том, что регрессионная модель описывает 50% случаев, т.е. она справедлива только для 50% исходных данных.

Важной частью регрессионного анализа является анализ остатков, т.е. отклонений наблюдаемых значений от теоретически ожидаемых. Остатки должны появляться случайно (не систематически) и подчиняться случайному распределению. Проверка наличия систематических связей между остатками может быть произведена при помощи теста Дарбина–Уотсона (*Durbin–Watson*) на автокорреляцию. Этот тест позволяет рассчитать коэффициент, значение которого варьирует от 0 до 4. Если значение дан-

ного коэффициента близко к 2, это означает, что автокорреляция отсутствует.

Линейный регрессионный анализ проводится в *SPSS* с помощью меню «*Analyze* > *Regression* > *Linear...*» (рис. 6.1).

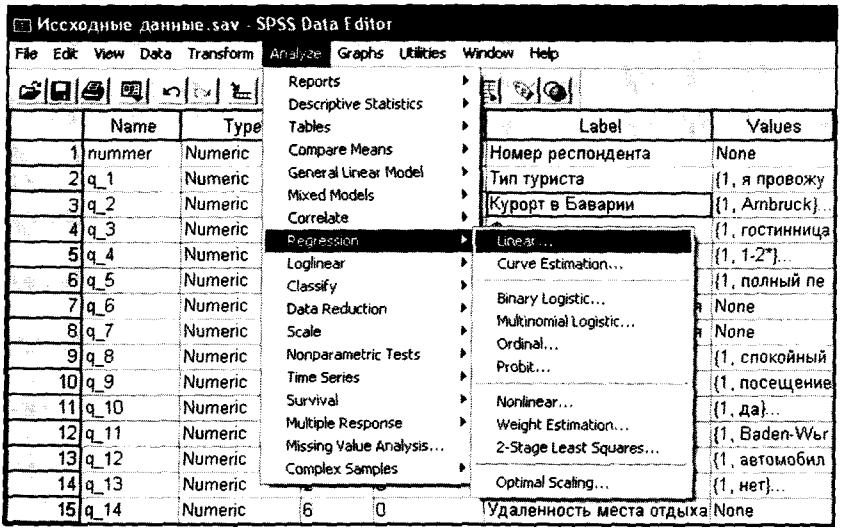


Рис. 6.1. Выбор в меню процедуры «Линейная регрессия»

После выбора меню, представленного на рис. 6.1, открывается диалоговое окно «Линейная регрессия», при помощи которого формируется задание на проведение анализа в *SPSS*. Набор команд данного задания обусловливается тем, какой именно вид линейной регрессии используется – простая или множественная линейная регрессия.

6.1. ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

6.1.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В *SPSS*

Как отмечалось ранее, простая линейная регрессия используется для выявления и описания линейной зависимости между двумя исследуемыми переменными – зависимой и независимой.

Например, следует определить, в какой зависимости находятся такие переменные, как сумма общих расходов туристов на

проведение отпуска и сумма, уплачиваемая туристами за проживание в отеле или пансионе (включая обслуживание) (рис. 6.2).

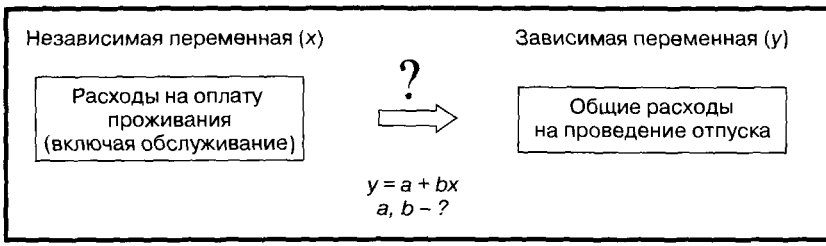


Рис. 6.2. Простая линейная регрессия. Постановка цели исследования

Для анализа используются данные опроса туристов, отдыхающих в курортной зоне «Баварский лес». Для анализа из всех вопросов анкеты выбраны два:

- вопрос № 45: «Какую сумму денег Вы тратите в целом на отдых?»;
- вопрос № 47: «Какую сумму денег Вы тратите во время отдыха на проживание в гостинице/пансионе (включая обслуживание)?».

При занесении информации по ответам на данные вопросы в файл данных *SPSS* была использована двойная запись переменных (см. п. 2.2 «Виды кодировки данных»). Каждый из этих вопросов представлен в файле данных *SPSS* в виде двух переменных (рис. 6.3).

Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values
111	q_44_12	Numeric	2	0	Расходы на сувениры	{1, да}...
112	q_44_13	Numeric	2	0	Расходы на самые мелкие	{1, да}...
113	q_45_1	Numeric	2	0	Общие расходы на отдых	{1, сумма в E
114	q_45_2	Numeric	6	0	Общие расходы на отдых	None
115	q_46_1	Numeric	2	0	Расходы на покупки	{1, сумма в E
116	q_46_2	Numeric	6	0	Расходы на покупки	None
117	q_47_1	Numeric	2	0	Расходы на проживание	{1, сумма в E
118	q_47_2	Numeric	6	0	Расходы на проживание	None
119	q_48_1	Numeric	2	0	Расходы на питание	{1, сумма в E
120	q_48_2	Numeric	6	0	Расходы на питание	None

Рис. 6.3. Фрагмент вкладки «*Variable View*» («Свойства переменных»)

Вопрос № 45 представлен в виде двух переменных с именами «q_45_1» и «q_45_2» и одинаковыми метками «Общие расходы на отдых». Вопрос № 47 представлен в виде двух переменных с именами «q_47_1» и «q_47_2» и одинаковыми метками «Расходы на проживание».

Переменные с именами «q_45_1» и «q_47_1» являются номинальными. Значения меток этих переменных имеют числовые коды («1» – «сумма в евро», «98» – «не знаю», «99» – «нет данных»). Они указывают на то, назвал ли респондент точную сумму в качестве ответа на поставленный вопрос или нет.

Переменные с именами «q_45_2» и «q_47_2» являются метрическими. В столбце «*Values*» таблицы «Свойства переменных» отсутствуют значения меток переменных (см. рис 6.3). Значения этих переменных выражаются в конкретных денежных суммах и отображаются в столбцах «q_45_2» и «q_47_2» таблицы «Значения переменных» (рис. 6.4).

Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : nummer 2149

	q_45_1	q_45_2	q_46_1	q_46_2	q_47_1	q_47_2
2194	1	2200	1	1200	1	800
2195	1	2400	1	1300	1	700
2196	1	1600	1	700	1	700
2197	1	1400	1	1022	1	300
2198	1	2400	1	1920	1	280
2199	98	.	1	820	1	600
2200	1	1600	1	680	1	600
2201	1	1200	1	500	1	600
2202	1	1200	1	400	1	600

Рис. 6.4. Фрагмент вкладки «*Data View*» («Значения переменных»)

Фрагмент таблицы «Значения переменных» файла данных *SPSS*, представленный на рис. 6.4, содержит информацию о расходах респондентов во время отдыха. Респондент в строке 2194 потратил на отдых в целом 2200 евро, на оплату гостиницы/пансиона – 800 евро. Респондент в строке 2197 потратил на отдых в целом 1400 евро, а на оплату гостиницы/пансиона – 300 евро и т.д.

Преимуществом простой линейной регрессии является возможность представить результаты анализа графически. Регрес-

сионное уравнение, представляющее собой линейную функцию с одной переменной, может быть представлено в виде линейного графика в двухмерной системе координат.

Линейный регрессионный анализ применяется для графического прогнозирования поведения одной переменной в зависимости от изменения другой. Как правило, целью регрессионного анализа в данном случае является построение тренда, т.е. линейного графика, отображающего зависимость между переменными.

Исходя из полученного уравнения регрессии можно предсказать, каким будет значение одной переменной при изменении другой. В рассматриваемом примере регрессионная модель позволит спрогнозировать, как будут расти (уменьшаться) общие расходы на проведение отпуска при увеличении (уменьшении) расходов на проживание, включая обслуживание.

6.1.2. КОМАНДЫ *SPSS* НА ВЫПОЛНЕНИЕ ПРОСТОГО РЕГРЕССИОННОГО АНАЛИЗА

Запрос на выполнение линейного регрессионного анализа осуществляется в одноименном диалоговом окне (*Linear Regression*) (рис. 6.5), которое открывается при выборе в меню «*Analyze* > > *Regression* > *Linear...*» (см. также рис. 6.1).

В левом поле диалогового окна «Линейная регрессия» (*Linear Regression*) представлен список меток всех переменных, занесенных в базу данных. Из этого списка выбирается зависимая переменная и переносится в поле «*Dependent*». В рассматриваемом примере это переменная, имеющая метку «Общие расходы на отдых». В базе данных такую метку имеют две переменные. В данном случае выбирается переменная, значения которой представляют собой денежные суммы в евро.

Из списка всех существующих в базе данных меток переменных также выбирается независимая переменная и переносится в поле «*Independent(s)*». В рассматриваемом примере это переменная, имеющая метку «Расходы на проживание». В базе данных такую метку имеют две переменные. В данном случае выбирается переменная, значения которой представляют собой денежные суммы в евро.

В диалоговом окне «Линейная регрессия» в поле «Метод» (*Method*) следует указать метод включения переменных в регрессионную модель. По умолчанию задается метод «*Enter*». В случае осуществления простого регрессионного анализа рекомендуется использовать именно этот метод.

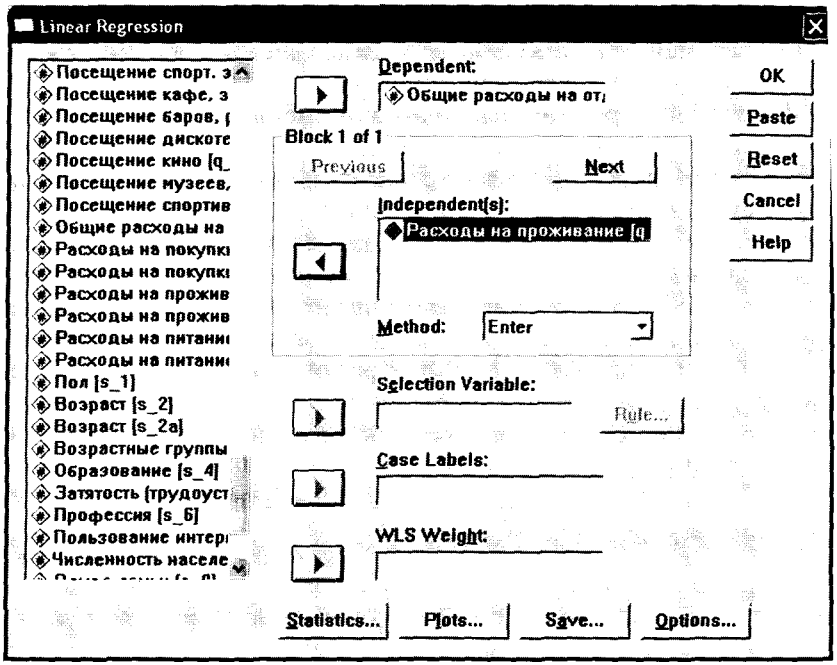


Рис. 6.5. Диалоговое окно «Линейная регрессия». Запрос выполнения простой регрессии

Если бы нужно было построить регрессионные модели для отдельных групп респондентов, то в поле «*Selection Variable*» диалогового окна «Линейная регрессия» нужно было бы указать переменную, по которой производится отбор респондентов в исследуемые группы. Например, если бы нужно было построить две регрессионные модели для мужчин и для женщин, то из общего списка переменных в поле «*Selection Variable*» следовало бы переместить метку переменной «Пол».

В главном диалоговом окне «Линейная регрессия» имеются четыре кнопки («*Statistics*», «*Plots*», «*Save*» и «*Options*»), при нажатии которых открываются вспомогательные диалоговые окна.

При нажатии кнопки «*Statistics*» на экране появляется одноименное диалоговое окно «Статистические показатели» (рис. 6.6). В нем задаются команды на расчет различных статистических показателей. В рассматриваемом примере поставлена отметка напротив команды «*Estimates*». В результате выполнения данной команды рассчитываются коэффициент детерминации (R), коэффи-

циент R-квадрат (*R Square*) и некоторые другие статистические показатели, необходимые для оценки качества построенной регрессионной модели.

Также в диалоговом окне «Статистические показатели» можно задать вычисление показателей, используемых для анализа остатков (*Residuals*). В рассматриваемом примере задается команда на выполнение теста Дарбина–Уотсона[1].

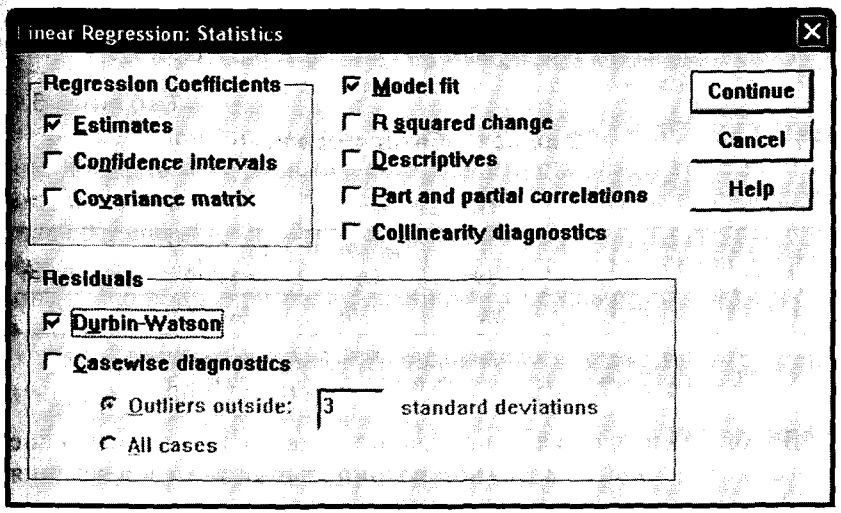


Рис. 6.6. Диалоговое окно «Статистические показатели»

При нажатии кнопки «*Continue*» в диалоговом окне «Статистические показатели» данное окно закрывается и осуществляется возврат в главное диалоговое окно «Линейная регрессия». После

[1] **Тест Дарбина–Уотсона (*Durbin–Watson Test*)** – тест на автокорреляцию. Автокорреляция выражается в наличии систематических связей между остатками.

В соответствии с теорией статистики уравнение простой регрессии (регрессионная модель) имеет вид

$y = a + bx + \varepsilon$,

где ε – остатки (отклонения наблюдаемых значений от теоретически ожидаемых).

Остатки должны появляться случайно, т.е. не систематически. Для проверки этого условия проводится тест Дарбина–Уотсона. В ходе проведения этого теста рассчитывается коэффициент, значение которого лежит в диапазоне от 0 до 4. Если значение коэффициента близко к среднему (т.е. к 2), это означает, что автокорреляция отсутствует, т.е. остатки появляются случайным образом.

нажатия кнопки «*OK*» в диалоговом окне «Линейная регрессия» запускается процедура выполнения простого регрессионного анализа.

6.1.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ПРОСТОГО РЕГРЕССИОННОГО АНАЛИЗА

В качестве результатов линейного регрессионного анализа *SPSS* выводит на экран компьютера три таблицы: «*Model Summary*», «*ANOVA*» и «*Coefficients*» (табл. 6.1, 6.2 и 6.3).

Таблица 6.1

Сводная таблица модели

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,605[a]	,366	,366	587,685	1,874

[a] Predictors – влияющие переменные (константа): расходы на проживание.
[b] Dependent Variable – зависимая переменная: общие расходы на отдых.

В табл. 6.1 представлены основные показатели, оценивающие качество линейной модели, построенной в результате проведения регрессионного анализа.

В рассматриваемом примере значение коэффициента детерминации R составляет 0,605 (>0,5), что свидетельствует о наличии тесной линейной взаимосвязи между суммой общих расходов на проведения отпуска и суммой, уплачиваемой туристами за проживание в гостинице или пансионе.

Коэффициент R-квадрат (*R Square*) в рассматриваемом примере составляет всего 0,366. Это означает, что построенная регрессионная модель описывает только 36,6% случаев, когда увеличение суммы оплаты за проживание в гостинице или пансионе влечет за собой увеличение общих расходов на проведение отпуска. Это необходимо учитывать при применении результатов анализа в прогнозировании расходов туристов.

Значение теста Дарбина–Уотсона на автокорреляцию в рассматриваемом примере составляет 1,874 (см. табл. 6.1), т.е. близко к 2. Это говорит об отсутствии систематических связей между остатками, т.е. между отклонениями наблюдаемых (эмпирических) значений от теоретически ожидаемых (расчетных).

Таблица 6.2

Результаты теста однофакторного дисперсионного анализа

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Регрессия	193625404,593	1	193625404,6	560,625	,000[a]
	Остатки	335012199,316	970	345373,401		
	Всего	528637603,910	971			

[a] Predictors – влияющие переменные (константа): расходы на проживание.
[b] Dependent Variable – зависимая переменная: общие расходы на отдых.

В последнем столбце таблицы «*ANOVA*» (см. табл. 6.2) значение показателя «Статистическая значимость» (*Sig.*) должно быть меньше или равно 0,5. В рассматриваемом примере этот показатель составляет ноль. Это свидетельствует о том, что регрессионная модель, построенная на основе данных респондентов, попавших в выборку, справедлива для всей генеральной совокупности в целом.

Результаты регрессионного анализа, описывающие построенную регрессионную модель, представлены в табл. 6.3.

Таблица 6.3

Результаты регрессионного анализа. Коэффициенты модели

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	642,273	31,526		20,373	,000
	Расходы на проживание	1,596	,067	,065	23,678	,000

[a] Dependent Variable – зависимая переменная: общие расходы на отдых.

В столбце «*B*» таблицы «Коэффициенты» представлены параметры построенной регрессионной модели. В рассматриваемом примере уравнение регрессии имеет вид $y = 642{,}273 + 1{,}596x$.

Величина «*Constant*» показывает значение зависимой переменной при нулевом значении независимой переменной. Построенная регрессионная модель в рассматриваемом примере показывает, что если турист не тратит никаких денег за проживание в отеле или пансионе (например, если он остановился у дру-

зей или живет в палатке), то его общие расходы на проведение отпуска в среднем составят 642,273 евро.

В следующем столбце табл. 4.3 представлены стандартные ошибки (*Std. Error*). При доверительном интервале 95% каждый коэффициент может отклоняться от средней величины на $\pm 2 \times$ $\times$ *Std.Error*. Например, сумма общих расходов на проведение отпуска при нулевых затратах на проживание в гостинице или пансионе может отклоняться от среднего значения (642,273 евро) на $\pm 2 \cdot 31{,}526$, т.е. на $\pm$ 63,052 евро.

Значение коэффициента регрессии независимой переменной «Затраты на проживание в гостинице или пансионе» в построенной модели составляет 1,596. Это означает, что увеличение затрат на проживание в отеле или пансионе на 1 евро влечет за собой увеличение суммы общих затрат на проведение отпуска на 1,596 евро.

6.1.4. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ПРОСТОЙ РЕГРЕССИОННОЙ МОДЕЛИ В SPSS

Как уже было отмечено выше, основным достоинством линейной регрессии является возможность наглядного представления результатов анализа в виде линейного графика в двухмерной системе координат. Задание на построение такого графика осуществляется в *SPSS* в опции «*Graphs*».

При выборе меню «*Graphs* > *Scatter*» на экране появляется диалоговое окно «*Scatte/Dot*» («Диаграмма рассеяния»), в котором следует выбрать тип диаграммы. В данном случае следует выбрать диаграмму «*Simple Scatter*» (рис. 6.7).

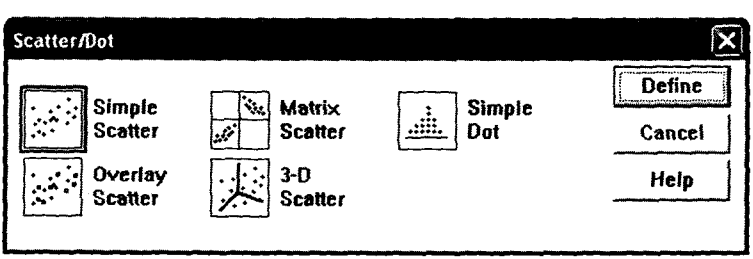


Рис. 6.7. Диалоговое окно «Диаграмма рассеяния»

Путем нажатия кнопки «*Define*» в диалоговом окне «*Scatte/Dot*» («Диаграмма рассеяния») на экране компьютера появляется но-

вое диалоговое окно «*Simple Scatterplot*» («Простая диаграмма рассеяния») (рис. 6.8).

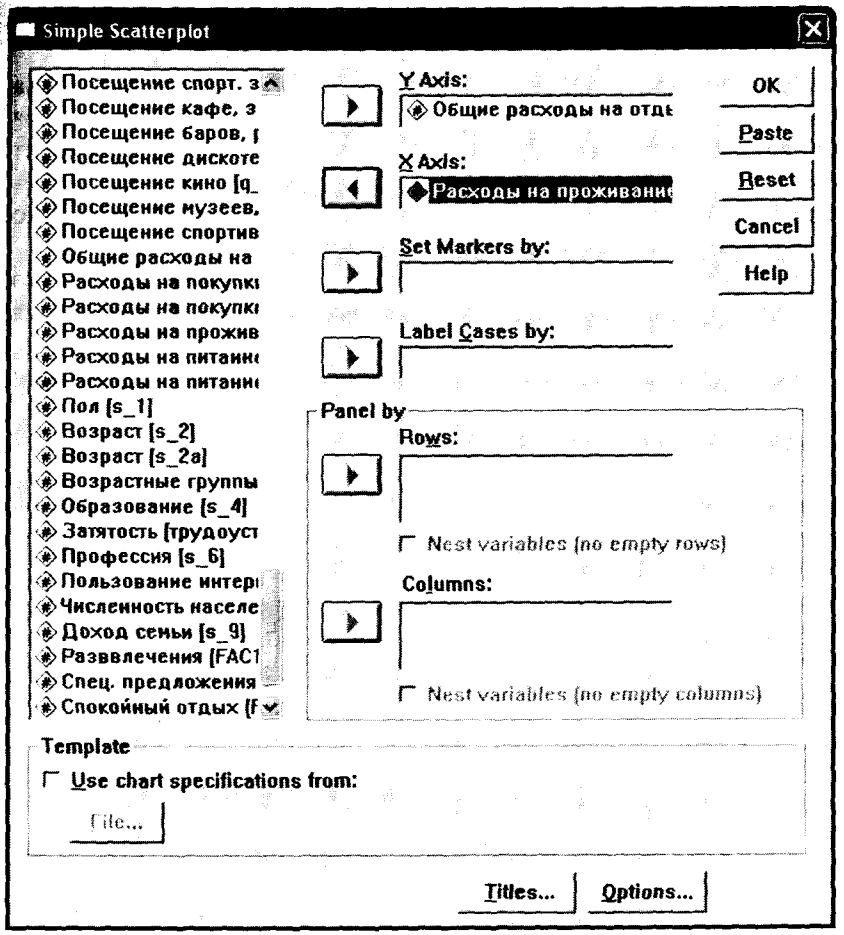


Рис. 6.8. Диалоговое окно «Простая диаграмма рассеяния»

В левом поле диалогового окна «Простая диаграмма рассеяния» указываются метки всех переменных, содержащихся в исходном файле данных *SPSS*. Из списка меток всех переменных следует выбрать метку зависимой переменной и перенести ее в правое поле окна «*Y Axis*». В рассматриваемом примере это метка переменной «q_45_2» – «Общие расходы на отдых».

Далее из списка всех переменных, представленных в левом поле окна «Простая диаграмма рассеяния», следует выбрать метку независимой переменной и перенести ее в правое поле окна «*X Axis*». В рассматриваемом примере это метка переменной «q_47_2» – «Расходы на проживание».

Нажав кнопку «*OK*» в диалоговом окне «Простая диаграмма рассеяния», мы закрываем данное окно, и на экране компьютера появляется диаграмма рассеяния. К данному рисунку следует подвести курсор мыши и нажать кнопку мыши два раза. В результате этой операции на экране появится диалоговое окно «*Chart Editor*» («Редактор диаграмм») (рис. 6.9).

В диалоговом окне «*Chart Editor*» следует выбрать меню «*Elements* > *Fit Line at Total*», в результате чего на экране появится новое диалоговое окно «*Properties*» («Свойства») (см. рис. 6.9).

Во вкладке «*Fit Line*» («Приближенная линия») диалогового окна «*Properties*» («Свойства») следует отметить линейный вид графика – «*Linear*» (рис. 6.10).

После нажатия кнопки «*Close*» в диалоговом окне «Свойства» (рис. 6.10) данное окно закрывается. На рисунке построенной ранее диаграммы рассеяния (см. рис. 6.10 – без линии тренда) появляется линия, отображающая линейную регрессионную модель. Следует отвести курсор мыши от рисунка и нажать клавишу мыши, в результате чего закроется диалоговое окно «*Chart Editor*» («Редактор диаграмм») и на экране останется только построенный график (рис. 6.11).

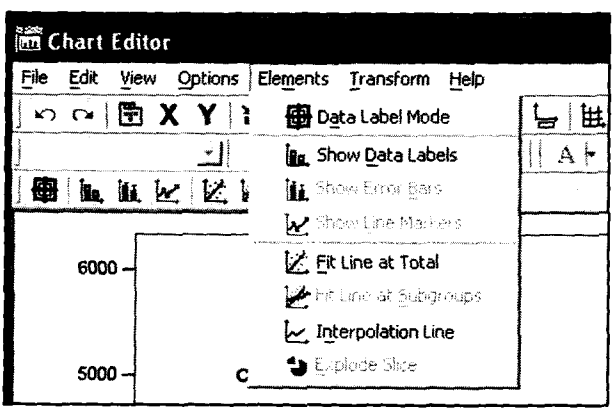


Рис. 6.9. Выбор меню «Приближенная линия»

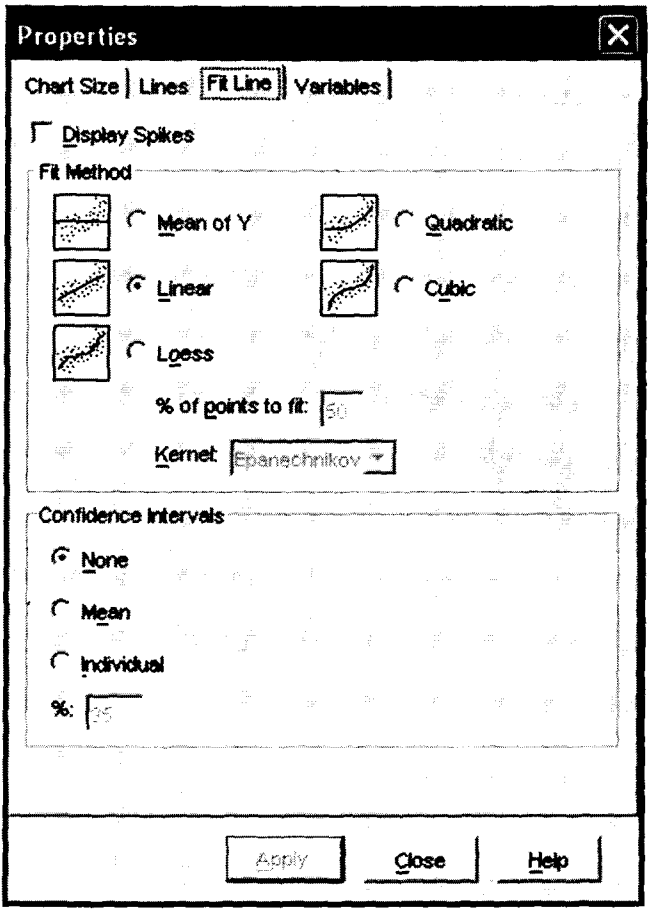


Рис. 6.10. Диалоговое окно «Свойства»

На рис. 6.11 представлено графическое изображение регрессионной модели $y = 642{,}273 + 1{,}596x$. Используя эту модель, можно прогнозировать, как будут изменяться общие расходы на отдых при изменении расходов на проживание в гостинице/пансионе для туристов, отдыхающих в курортной зоне «Баварский лес».

Построенная нами регрессионная модель описывает только 36,6% всех данных, полученных в результате опроса туристов. Это говорит о том, что вероятность ошибки при использовании данной регрессионной модели в целях прогнозирования достаточно велика.

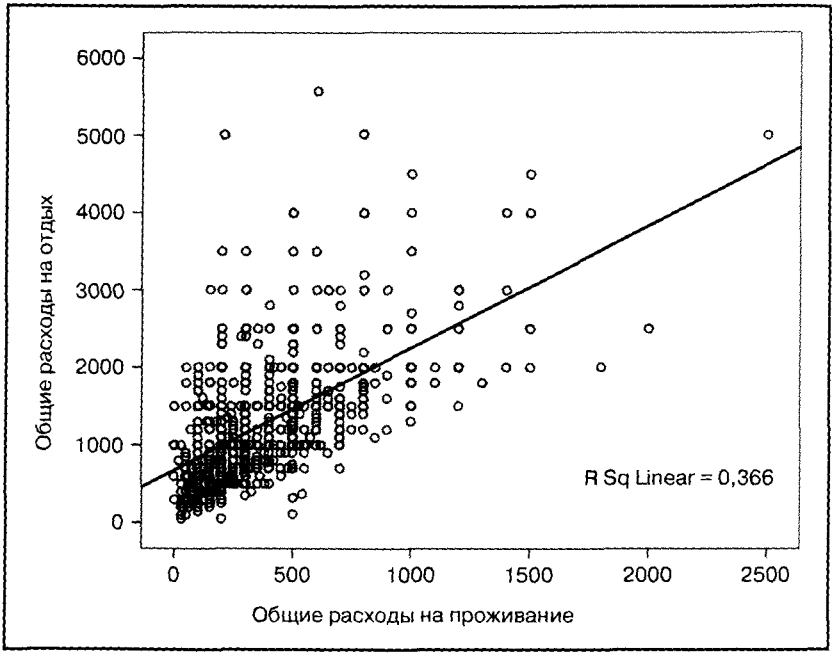


Рис. 6.11. Графическое представление простой регрессионной модели

6.2. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

6.2.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В *SPSS*

Как отмечалось ранее, множественная линейная регрессия используется для выявления и описания линейной зависимости между одной зависимой переменной и несколькими независимыми переменными. Например, следует определить, в какой зависимости между собой находятся общие расходы туристов на проведение отдыха и следующие статьи расходов:

- на покупки (одежды, обуви, галантерейных товаров, украшений, фотоаппаратов и т.д.);
- на проживание в отеле или пансионе (включая расходы на обслуживание);
- на питание (покупки продуктов в магазинах, посещение кафе и ресторанов).

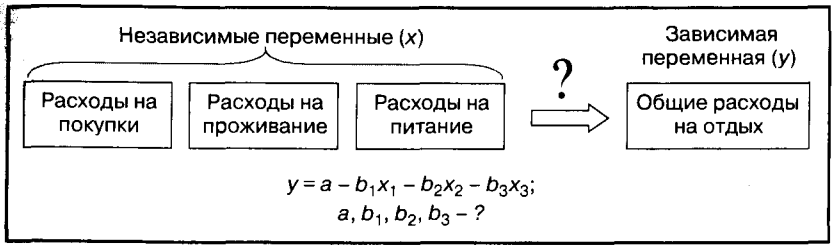


Рис. 6.12. Множественная регрессия.
Постановка вопроса исследования

Для проведения анализа используются данные опроса туристов, отдыхающих в курортной зоне «Баварский лес». Для проведения анализа из всех вопросов анкеты выбраны четыре:

- вопрос № 45: «Какую сумму денег Вы тратите в целом на отдых?»;
- вопрос № 46: «Какую сумму денег Вы тратите во время отдыха на покупки (такие, как одежда, обувь, галантерея, украшения, фотоаппараты и т.д.)?»;
- вопрос № 47: «Какую сумму денег Вы тратите во время отдыха на проживание в гостинице/пансионе (включая расходы на обслуживание)?»;
- вопрос № 8: «Какую сумму денег Вы тратите во время отдыха на питание (т.е. на покупку продуктов в магазинах, посещение кафе и ресторанов)?».

При занесении информации по ответам на данные вопросы в файл данных *SPSS* была использована двойная запись переменных (см. п. 2.2 «Виды кодировки данных»). Каждый вопрос представлен в файле данных *SPSS* в виде двух переменных. Представление этих переменных в файле данных *SPSS* проиллюстрировано в п. 6.1 «Простая линейная регрессия» (см. рис. 6.3 и 6.4).

Множественная линейная регрессия отличается от простой линейной регрессии рядом особенностей. Первая особенность состоит в невозможности графического изображения множественной регрессионной модели, что, конечно, наносит ущерб наглядности представления результатов анализа (см. рис. 6.12).

Другой особенностью множественной линейной регрессии является то, что переменные, объявленные независимыми, могут сами коррелировать между собой, т.е. возможно существование причинно-следственных связей между ними. В этом случае возникает эффект мультиколлинеарности.

Эффект мультиколлинеарности заключается в том, что независимые переменные, включенные в регрессионную модель, обозначают в принципе одно и то же. Например, в качестве зависимой переменной объявлена заработная плата выпускника университета, а в качестве независимых переменных – средний балл успеваемости во время учебы в университете и индекс интеллекта. Поскольку успеваемость студента во многом определяется уровнем интеллекта, в данной регрессионной модели возможно появление ложных корреляций.

Одним из условий построения множественной регрессионной модели является отсутствие или низкая степень корреляции между независимыми переменными. Для проверки соблюдения этого условия при проведении регрессионного анализа необходимо сначала производить диагностику наличия коллинеарности между независимыми переменными.

6.2.2. КОМАНДЫ *SPSS* НА ВЫПОЛНЕНИЕ МНОЖЕСТВЕННОГО РЕГРЕССИОННОГО АНАЛИЗА

Запрос на выполнение линейного регрессионного анализа осуществляется в одноименном диалоговом окне (*Linear Regression*) (рис. 6.13), которое открывается при выборе в меню «*Analyze* > *Regression* > *Linear...*» (см. рис. 6.1).

В левом поле диалогового окна «Линейная регрессия» (*Linear Regression*) представлен список меток всех переменных, занесенных в базу данных. Из этого списка выбирается зависимая переменная и переносится в поле «*Dependent*». В рассматриваемом примере это будет переменная, имеющая метку «Общие расходы на отдых», значения которой представляют собой денежные суммы в евро.

Из списка существующих в базе данных меток переменных также выбираются независимые переменные и переносятся в поле «*Independent(s)*». В рассматриваемом примере это переменные, имеющие метки: «Расходы на покупки», «Расходы на проживание» и «Расходы на питание». Значения отобранных переменных – денежные суммы в евро.

Далее в диалоговом окне «Линейная регрессия» в поле «Метод» (*Method*) следует указать метод включения переменных в регрессионную модель. Из четырех методов, предлагаемых *SPSS* («*Enter*», «*Stepwise*», «*Forward*» и «*Backward*»), следует выбрать один. По умолчанию задается метод «*Enter*».

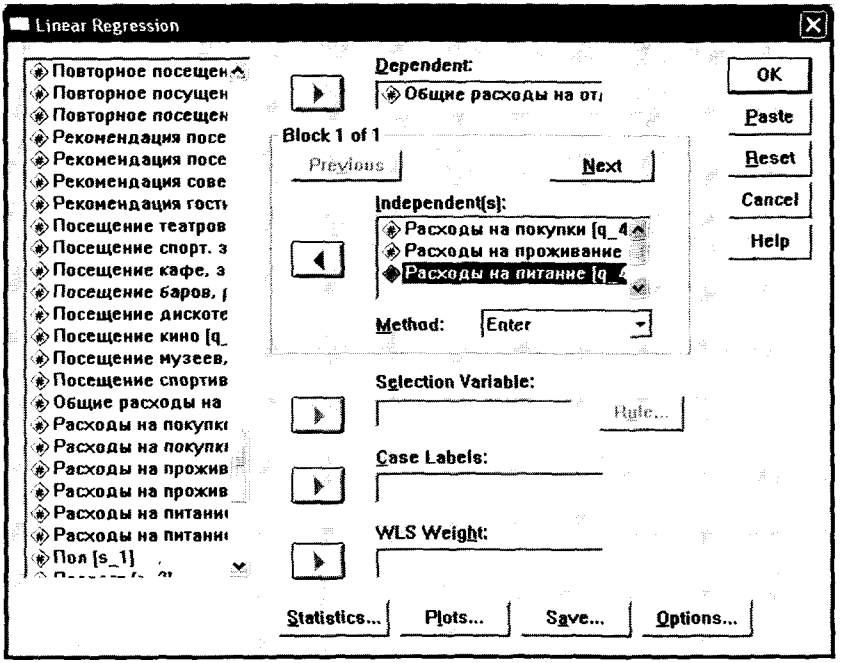


Рис. 6.13. Диалоговое окно «Линейная регрессия». Запрос выполнения множественной регрессии

Метод «*Enter*» предполагает одновременную обработку всех независимых переменных, выбранных для анализа, он рекомендуется для случая простого регрессионного анализа с одной независимой переменной. Для множественного регрессионного анализа рекомендуется выбрать один из пошаговых методов, которые предполагают поэтапное включение независимых переменных в регрессионную модель. В нашем примере выбран метод «*Stepwise*».

При нажатии кнопки «*Statistics*» в главном диалоговом окне «Линейная регрессия» (см. рис. 6.13) на экране появляется одноименное диалоговое окно (рис. 6.14).

Диалоговое окно «Статистические показатели» уже было представлено в п. 6.1 «Простая линейная регрессия» на рис. 6.6. В данном окне задаются команды на расчет статистических показателей, характеризующих качество построенной регрессионной модели.

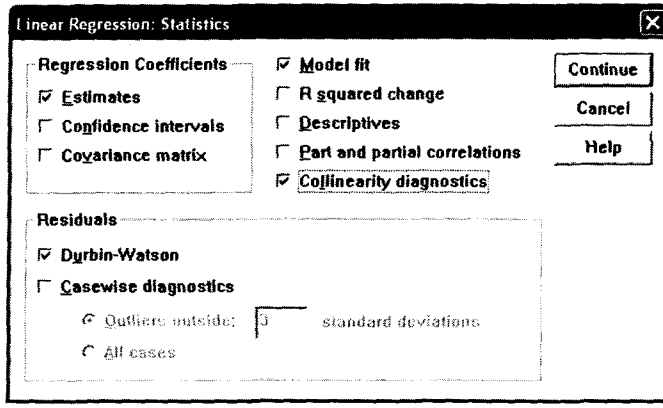


Рис. 6.14. Диалоговое окно «Статистические показатели»

В рассматриваемом примере в диалоговом окне «Статистические показатели» отмечена команда «*Estimates*», при выполнении которой рассчитываются коэффициент детерминации (*R*), коэффициент *R*-квадрат (*R Square*). Так же отмечена команда «*Durbin–Watson*», при выполнении которой производится тест Дарбина–Уотсона на автокорреляцию. Эти команды уже были подробно рассмотрены в п. 6.1 «Простая линейная регрессия».

Особенность множественного регрессионного анализа состоит в необходимости проверить наличие взаимосвязей между независимыми переменными. Для этого проводится диагностика коллинеарности, которая задается отметкой напротив команды «*Collinearity diagnostics*» в диалоговом окне «Статистические показатели».

Нажав кнопку «Продолжение» (*Continue*) в диалоговом окне «Статистические показатели» (рис. 6.14) мы закрываем данное окно и возвращаемся в главное диалоговое окно «Линейная регрессия» (см. рис. 6.13). После нажатия кнопки «*OK*» в главном диалоговом окне «Линейная регрессия» запускается процедура выполнения анализа.

6.2.3. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ МНОЖЕСТВЕННОГО РЕГРЕССИОННОГО АНАЛИЗА

В качестве результатов линейного регрессионного анализа *SPSS* выводит на экран компьютера три таблицы: «*Model Summary*», «*ANOVA*» и «*Coefficients*» (табл. 6.4, 6.5 и 6.6).

Поскольку при формировании задания на выполнение анализа был выбран пошаговый метод включения независимых переменных в регрессионную модель «*Stepwise*», то при представлении результатов анализа формируется несколько регрессионных моделей. В рассматриваемом примере таких моделей три (по числу независимых переменных). В соответствии с целями исследования основным результатом анализа является третья регрессионная модель, включающая все три независимые переменные (табл. 6.4).

Таблица 6.4

Сводная таблица модели

Model Summary[d]

Model	R	R Square	Ajusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,789[a]	,623	,622	444,996	
2	,889[b]	,790	,789	332,301	
3	,937[c]	,879	,878	252,373	1,930

[a] Predictors – влияющие переменные (константа): расходы на покупки.
[b] Predictors – влияющие переменные (константа): расходы на покупки, расходы на проживание.
[c] Predictors – влияющие переменные (константа): расходы на покупки, расходы на проживание, расходы на питание.
[d] Dependent Variable – зависимая переменная: общие расходы на отдых.

В сводной таблице модели представлены показатели, характеризующие качество построенных регрессионных моделей. В соответствии с целями исследования основным результатом нашего анализа является третья регрессионная модель, включающая все три независимые переменные.

В нашем примере значение коэффициента детерминации (*R*) составляет 0,937 (возможные значения от нуля до единицы), что свидетельствует о наличии плотной линейной взаимосвязи между суммой общих расходов на отпуск и суммами, расходуемыми туристами на текущие покупки, проживание и питание.

Коэффициент *R*-квадрат (*R Square*) составляет 0,879. Это означает, что наша регрессионная модель описывает 87,9% случаев, т.е. ответов респондентов о структуре их расходов на отпуск.

Показатели коэффициента детерминации и коэффициента *R*-квадрат для первых двух моделей ниже, чем для третьей модели

(см. табл. 6.4). Также значения стандартной ошибки расчетов для первых двух моделей выше, чем для третьей. Это доказывает целесообразность включения в регрессионную модель всех трех независимых переменных.

Сводная таблица модели представляет также результат теста Дарбина–Уотсона на автокорреляцию, значение которого должно быть приближено к 2, что свидетельствует об отсутствии системных связей между остатками, т.е. между отклонениями эмпирических (наблюдаемых) значений от теоретически ожидаемых (расчетных). В рассматриваемом примере значение этого показателя составляет 1,930, что является очень хорошим результатом.

Таблица 6.5

Результаты регрессионного анализа. Таблица *ANOVA*

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Регрессия	262222213,494	1	262222213,5	1324,213	,000[a]
	Остатки	159011004,084	803	198021,176		
	Всего	421233217,578	804			
2	Регрессия	332673124,355	2	166336562,2	1506,344	,000[b]
	Остатки	88560093,223	802	110424,056		
	Всего	421233217,578	804			
3	Регрессия	370215808,261	3	123405269,4	1937,527	,000[c]
	Остатки	51017409,317	801	63692,146		
	Всего	421233217,578	804			

[a] Predictors – влияющие переменные (константа): расходы на покупки.
[b] Predictors – влияющие переменные (константа): расходы на покупки, расходы на проживание.
[c] Predictors – влияющие переменные (константа): расходы на покупки, расходы на проживание, расходы на питание.
[d] Dependent Variable – зависимая переменная: общие расходы на отдых.

В последнем столбце таблицы «*ANOVA*» (см. табл. 6.5) значение показателя «Статистическая значимость» (*Sig.*) должно быть меньше или равно 0,05. В нашем примере для всех трех моделей этот показатель составляет 0,000. Это свидетельствует о том, что регрессионные модели, построенные на основе данных респондентов, попавших в выборку, справедливы для всей генеральной совокупности в целом.

В табл. 6.6 представлены параметры моделей, построенных в результате линейного регрессионного анализа. В рассматриваемом примере результатом анализа является третья регрессионная модель, включающая все независимые переменные.

Таблица 6.6

Коэффициенты множественной линейной регрессионной модели

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	425,294	27,603		15,407	,000		
	Расходы на покупки	1,438	,040	,789	36,390	,000	1,000	1,000
2	(Constant)	132,908	23,641		5,622	,000		
	Расходы на покупки	1,230	,031	,675	40,143	,000	,928	1,078
	Расходы на проживание	1,095	,043	,425	25,259	,000	,928	1,078
3	(Constant)	44,283	18,322		2,417	,016		
	Расходы на покупки	1,118	,024	,613	47,138	,000	,893	1,120
	Расходы на проживание	,944	,034	,366	28,176	,000	,896	1,116
	Расходы на питание	1,016	,042	,313	24,278	,000	,907	1,102

[a] Dependent Variable – зависимая переменная: общие расходы на отдых.

Интерпретация результатов таблицы начинается с рассмотрения статистических показателей, характеризующих коллинеарность (наличие взаимосвязи) между независимыми переменными регрессионной модели (*Collinearity Statistics*). Значение показателя «*Tolerance*» должно превышать 0,1, а значение показателя «*VIF*» должно быть менее 10. В рассматриваемом примере значение «*Tolerance*» составляет 0,907, а «*VIF*» – 1,102, что свидетельствует о невозможности возникновения нежелательного эффекта мультиколлинеарности.

Стандартизированные коэффициенты регрессии (*Beta*) показывают относительную значимость независимых переменных, включенных в регрессионную модель. Иными словами, они по-

казывают, как сильно влияют исследуемые факторы (независимые переменные) на итоговую величину (зависимую переменную).

В рассматриваемом примере наибольшей значимостью обладает первая независимая переменная (*Beta* = 0,613). *Это означает, что расходы на крупные покупки могут почти в два раза увеличить сумму общих расходов на отдых по сравнению с расходами на проживание* (*Beta* = 0,366) *и питание* (*Beta* = 0,313).

Результаты анализа можно объяснить тем, что расходы на питание и проживание в отеле/пансионе во время отдыха являются запланированными. Изменение этих расходов не ведет к резкому изменению расходов на отдых в целом. Что касается расходов на такие крупные покупки, как одежда, обувь, фотоаппаратура, спортивное снаряжение и т.п., то они, как правило, не являются запланированными. Туристы, отправляясь на отдых в курортную зону «Баварский лес», не планируют крупных покупок, поскольку этот регион не отличается низкими ценами. Именно поэтому совершение крупных покупок способно привести к резкому увеличению расходов на отдых.

В табл. 6.6 представлены также нестандартизированные коэффициенты регрессии (*B*). Они являются наиболее важными показателями результатов анализа, поскольку используются для построения регрессионной модели (регрессионного уравнения).

Следует отметить, что постоянный член регрессионного уравнения (*Constant*) в данном случае имеет достаточно большую величину (44,286). Это свидетельствует о том, что включенные в уравнение независимые переменные не в полной мере описывают зависимую переменную. В нашем примере это означает, что среди расходов на отпуск кроме затрат на покупки, оплаты проживания и расходов на питание существуют другие важные статьи затрат, например затраты на транспорт.

Результатом линейного регрессионного анализа является модель линейной регрессии (регрессионное уравнение)

$$y = 44{,}283 + 1{,}118x_1 + 0{,}944x_2 + 1{,}016x_3,$$

где y – общие расходы туристов на проведение отдыха;

x_1 – расходы на покупки (одежды, обуви, галантерейных товаров, украшений, фотоаппаратуры и т.д.);

x_2 – расходы на проживание в отеле или пансионе (включая расходы на обслуживание);

x_3 – расходы на питание (покупки продуктов в магазинах, посещение кафе и ресторанов).

Регрессионная модель является универсальной, поскольку описывает 87,9% случаев, т.е. ответов респондентов о структуре их расходов на отпуск. Она может быть использована специалистами по маркетингу при решении вопросов ценообразования в исследуемой курортной зоне.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите цели проведения и возможности использования результатов регрессионного анализа.
2. Какие требования предъявляются к переменным, участвующим в проведении регрессионного анализа, в отношении типов шкал измерения?
3. Как выглядит математическое описание регрессионной модели для простой и множественной линейной регрессии?
4. Что характеризуют коэффициент детерминации и коэффициент *R*-квадрат, рассчитываемые при проведении регрессионного анализа?
5. Как можно интерпретировать результаты, если значение коэффициента детерминации составляет 0,708, а коэффициента *R*-квадрат – 0,623?
6. С какой целью в ходе проведения регрессионного анализа производится тест Дарбина–Уотсона? Как можно интерпретировать результаты, если значение этого показателя составляет 1,487?
7. С какой целью в ходе проведения регрессионного анализа производится тест «*ANOVA*»? Как следует интерпретировать результаты, если величина «*Significance*» («Значимость») по результатам этого теста составляет 0,03?
8. Для чего служат стандартизированные (*Beta*) и нестандартизированные (*B*) коэффициенты регрессии?
9. Какие команды *SPSS* используются для построения диаграммы рассеяния и тренда, иллюстрирующего результаты простой линейной регрессии?
10. В чем заключается особенность представления результатов множественного регрессионного анализа при использовании пошаговых методов включения переменных в регрессионную модель?
11. В чем заключается эффект мультиколлинеарности при проведении множественного регрессионного анализа и по каким показателям определяется возможность возникновения этого эффекта?

7. ФАКТОРНЫЙ АНАЛИЗ В *SPSS*

7.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В *SPSS*

Факторный анализ позволяет разделить массив переменных на малое число групп, которые называются факторами. В один фактор объединяются несколько переменных, имеющих плотную корреляцию между собой и слабую корреляцию с переменными, объединяемыми другими факторами.

Основной задачей факторного анализа является группировка схожих по смыслу утверждений в макрокатегории (факторы) с целью сократить число переменных и упростить процедуру анализа существующей базы данных.

Факторный анализ, как правило, предшествует другим видам анализа, например кластерному. Для примера рассмотрим факторный анализ с целью сокращения массива данных, содержащего информацию об интересах туристов, отдыхающих в курортной зоне «Баварский лес». Фрагмент анкеты, при помощи которой была собрана данная информация, представлен в табл. 7.1.

Например, нужно провести кластерный анализ, позволяющий выделить однородные по возрасту и интересам (мотивам проведения времени на отдыхе) группы туристов. Собранный массив данных, содержащий информацию об интересах туристов, состоит из 12 переменных, каждая из которых может принимать не менее 5 значений (не считая ответов «затрудняюсь ответить» и «не хочу отвечать»). В целях упрощения процедуры кластерного анализа целесообразно провести факторный анализ, который позволит сократить число исследуемых переменных и оптимизировать структуру данных (рис. 7.1).

Таблица 7.1

Вопрос № 27
«Что для Вас самое важное во время отдыха?»

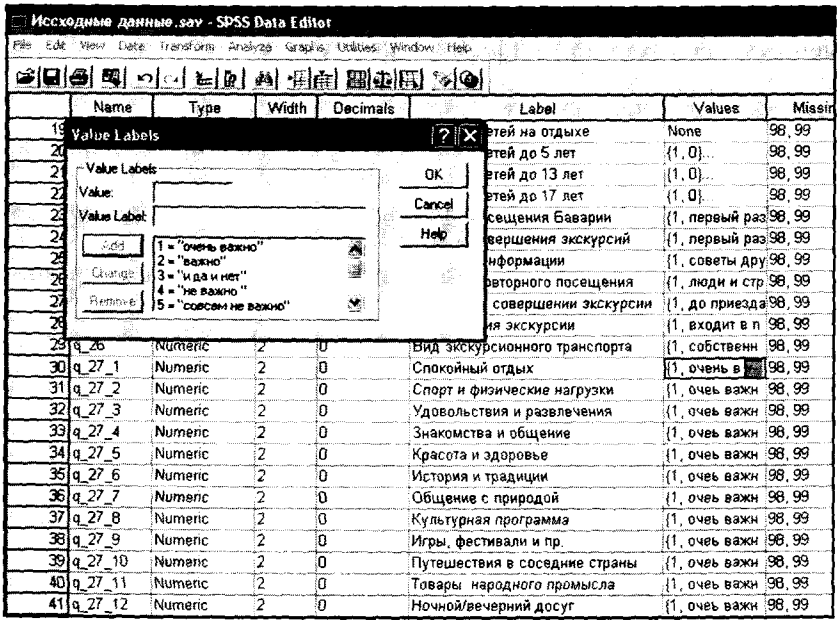
№ п/п	Оцените по 5-балльной шкале, насколько точно подходят для Вас следующие утверждения:	Балльная оценка: «1» – полностью подходит «2» – подходит «3» – и да и нет «4» – не подходит «5» – совсем не подходит
1	Я хочу расслабиться и просто отдохнуть	
2	Я хочу заниматься спортом и закалять организм физическими нагрузками	
3	Я ищу как можно больше разнообразных развлечений и удовольствий	
4	Мне важен круг общения и возможность знакомства с людьми	
5	Я использую отпуск для того, чтобы позаботиться о красоте и здоровье моего тела	
6	Я интересуюсь национальными традициями и историей Баварии	
7	Во время отдыха мне особенно важно общение с природой (лес, горы, водоемы и т.п.)	
8	Я уделяю особое внимание знакомству с достопримечательностями и культурной программе	
9	Меня привлекает возможность посетить фестивали, спортивные игры и другие мероприятия	
10	Отдых в Восточной Баварии дает мне возможность совершить экскурсии в Австрию, Чехию и другие соседние страны	
11	У меня есть возможность выгодно купить предметы народного промысла непосредственно у производителей	
12	Мне важно, чтобы я мог пойти куда-нибудь вечером и (или) ночью	

Рис. 7.1. Факторный анализ. Постановка цели исследования

При занесении в файл *SPSS* данных по ответам респондентов на вопрос анкеты № 27 создается 12 переменных, каждая из которых представляет собой мотив проведения времени на отдыхе. Данные переменные являются интервальными, поскольку их значениями являются балльные оценки их важности для респондентов (рис. 7.2).

Названия 12 переменных, содержащих информацию об ответах респондентов на вопрос анкеты № 27, указаны в первом столбце таблицы «Свойства переменных» (*Name*): от «q_27_1» до «q_27_12». В столбце «*Label*» указаны метки переменных, первая из которых – «Спокойный отдых», а последняя – «Ночной/ вечерний досуг».

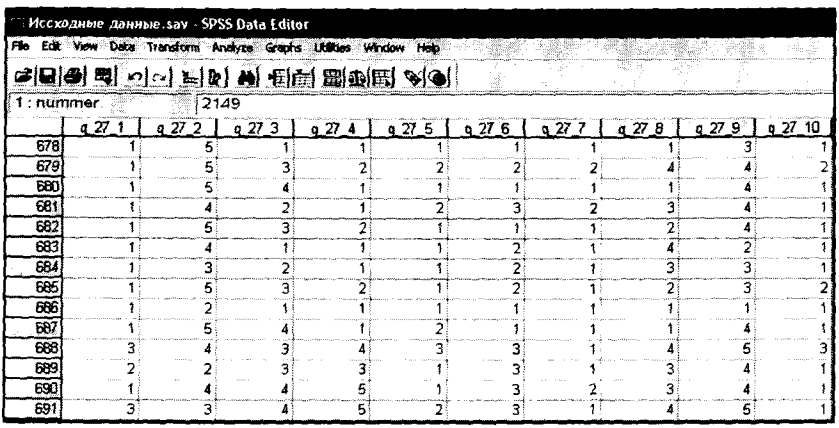
В столбце «*Values*» задаются значения меток переменных. При подведении курсора в соответствующую ячейку таблицы и нажатии кнопки мыши на экране появляется диалоговое окно «Значения меток переменных» (*Value Labels*). В данном диалоговом окне осуществляется кодировка ответов респондентов: «1» – «очень важно», «2» – «важно», «3» – «и да и нет», «4» – «неважно», «5» – «совсем не важно». В столбце «*Missing*» указывается кодировка пропущенных ответов: «98» – «затрудняюсь ответить», «99» – «ответа нет».



	Name	Type	Width	Decimals	Label	Values	Missin
19					етей на отдыхе	None	98, 99
20					етей до 5 лет	{1, 0}..	98, 99
21					етей до 13 лет	{1, 0}..	98, 99
22					етей до 17 лет	{1, 0}..	98, 99
23					сещения Баварии	{1, первый раз	98, 99
24					вершения экскурсий	{1, первый раз	98, 99
25					нформации	{1, советы дру	98, 99
26					вторного посещения	{1, люди и стр	98, 99
27					совершении экскурсии	{1, до приезда	98, 99
28					ия экскурсии	{1, входит в п	98, 99
29	q_26	Numeric	2	0	Вид экскурсионного транспорта	{1, собственн	98, 99
30	q_27_1	Numeric	2	0	Спокойный отдых	{1, очень в	98, 99
31	q_27_2	Numeric	2	0	Спорт и физические нагрузки	{1, очень важн	98, 99
32	q_27_3	Numeric	2	0	Удовольствия и развлечения	{1, очень важн	98, 99
33	q_27_4	Numeric	2	0	Знакомства и общение	{1, очень важн	98, 99
34	q_27_5	Numeric	2	0	Красота и здоровье	{1, очень важн	98, 99
35	q_27_6	Numeric	2	0	История и традиции	{1, очень важн	98, 99
36	q_27_7	Numeric	2	0	Общение с природой	{1, очень важн	98, 99
37	q_27_8	Numeric	2	0	Культурная программа	{1, очень важн	98, 99
38	q_27_9	Numeric	2	0	Игры, фестивали и пр.	{1, очень важн	98, 99
39	q_27_10	Numeric	2	0	Путешествия в соседние страны	{1, очень важн	98, 99
40	q_27_11	Numeric	2	0	Товары народного промысла	{1, очень важн	98, 99
41	q_27_12	Numeric	2	0	Ночной/вечерний досуг	{1, очень важн	98, 99

Рис. 7.2. Фрагмент вкладки «Свойства переменных» (*Variable View*)

Числовые коды («1», «2»... «5», «98», «99») значений меток переменных, используемые для кодировки ответов респондентов, заносятся в таблицу вкладки «Значения переменных» (*Data View*) редактора данных *SPSS* (рис. 7.3).



	q_27_1	q_27_2	q_27_3	q_27_4	q_27_5	q_27_6	q_27_7	q_27_8	q_27_9	q_27_10
678	1	5	1	1	1	1	1	1	3	1
679	1	5	3	2	2	2	2	4	4	2
680	1	5	4	1	1	1	1	1	4	1
681	1	4	2	1	2	3	2	3	4	1
682	1	5	3	2	1	1	1	2	4	1
683	1	4	1	1	1	2	1	4	2	1
684	1	3	2	1	1	2	1	3	3	1
685	1	5	3	2	1	2	1	2	3	2
686	1	2	1	1	1	1	1	1	1	1
687	1	5	4	1	2	1	1	1	4	1
688	3	4	3	4	3	3	1	4	5	3
689	2	2	3	3	1	3	1	3	4	1
690	1	4	4	5	1	3	2	3	4	1
691	3	3	4	5	2	3	1	4	5	1

Рис. 7.3. Фрагмент вкладки «Значения переменных» (*Data View*)

На рис. 7.3 представлен фрагмент таблицы «Свойства переменных», содержащий информацию об ответах респондентов на вопрос анкеты № 27: «Что для Вас наиболее важно во время отдыха?»

Из данных, представленных на рис. 7.3, видно, что для респондента в строке «681» очень важны («1»): спокойный отдых («q_27_1»), возможность знакомства и общения («q_27_4») и путешествия в соседние с Германией страны («q_27_10»). Для этого же респондента не важны («4»): спорт и физические нагрузки («q_27_2»), игры и фестивали («q_27_9»), товары народного промысла («q_27_11») и ночной/вечерний досуг («q_27_12») и т.д. В ходе выполнения факторного анализа решаются следующие задачи:

- оценивается пригодность исходных данных для проведения факторного анализа;
- выявляются корреляционные взаимосвязи между переменными исходного массива;
- определяется оптимальное число факторов (компонентов факторной модели), т.е. групп, на которые может быть разделен существующий массив переменных;
- разделяется существующий массив переменных на группы на основании значений коэффициентов корреляции;
- интерпретируются результаты, т.е. производится подбор названий созданным переменным (факторам).

Из перечисленных задач последняя является наиболее сложной. Ее решение представляет собой одну из ключевых проблем факторного анализа и требует творческого подхода.

Другой существенной проблемой факторного анализа является частичная потеря информации в ходе «сжатия» исходного массива переменных. В рассматриваемом примере 12 целей отдыха туристов не являются альтернативными. Например, туристы, заботящиеся о красоте и здоровье своего тела, могут приветствовать физические нагрузки и занятия спортом. Если занятия спортом и заботу о красоте тела объединить в одну группу, признав таким образом, что это одно и то же, произойдет частичная потеря информации. Это будет означать, что туристы, предпочитающие спокойный отдых, не заботятся о красоте и здоровье своего тела, например, когда они загорают на солнце или когда им делают массаж.

Одним из важнейших условий проведения факторного анализа является минимизации частичной потери информации, которая в любом случае неизбежна.

7.2. КОМАНДЫ SPSS НА ВЫПОЛНЕНИЕ ФАКТОРНОГО АНАЛИЗА

Для начала формирования задания на проведение в *SPSS* процедуры факторного анализа следует выбрать меню «*Analyze* > *Data reduction* > *Factor...*» (рис. 7.4).

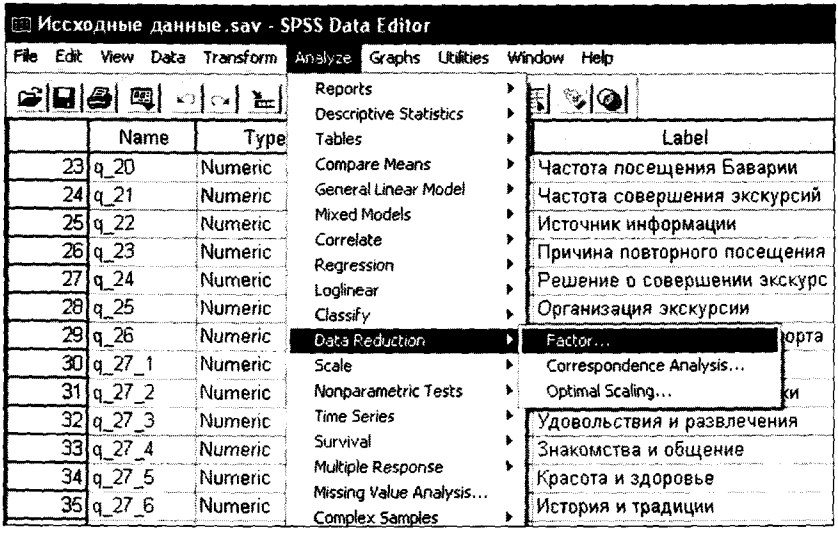


Рис. 7.4. Выбор в меню *SPSS* процедуры «Факторный анализ»

При выборе меню, представленного на рис. 7.4, открывается диалоговое окно «Факторный анализ» (*Factor Analysis*) (рис. 7.5).

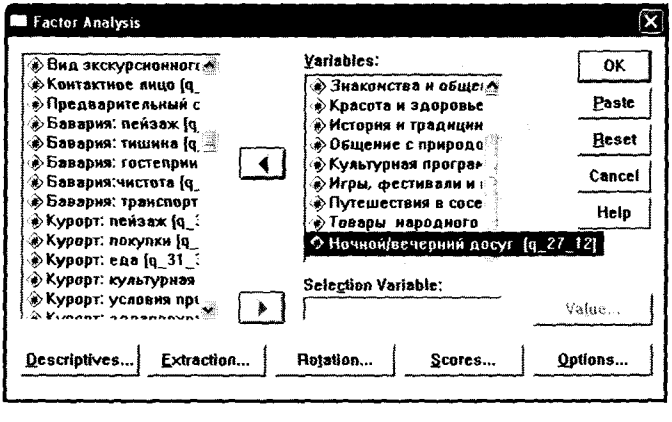


Рис. 7.5. Диалоговое окно «Факторный анализ»

В правом поле диалогового окна представлен список меток всех переменных, занесенных в базу данных. Из этого списка следует выбрать массив переменных, участвующих в факторном анализе, и перенести его в поле «*Variables*». В нашем примере это переменные, содержащие информацию об ответах респондентов на вопрос анкеты № 27: «Что для Вас самое важное во время отдыха?»

В поле «*Selection Variable*» могут быть указаны переменные для разбивки анализируемых данных на подгруппы. Например, если в поле «*Selection Variable*» перенести переменную «пол», факторный анализ будет проводиться по двум массивам данных – отдельно для мужчин и женщин. В рассматриваемом примере разделения анализируемых данных на подгруппы не производится.

При выборе команд для проведения анализа в первую очередь необходимо задать команду проверки пригодности существующего массива данных для проведения факторного анализа. Для этого нужно открыть диалоговое окно «Описательные статистические показатели» (*Descriptives*) (рис. 7.6) путем нажатия одноименной кнопки в диалоговом окне «Факторный анализ» (*Factor Analysis*) (см. рис. 7.5).

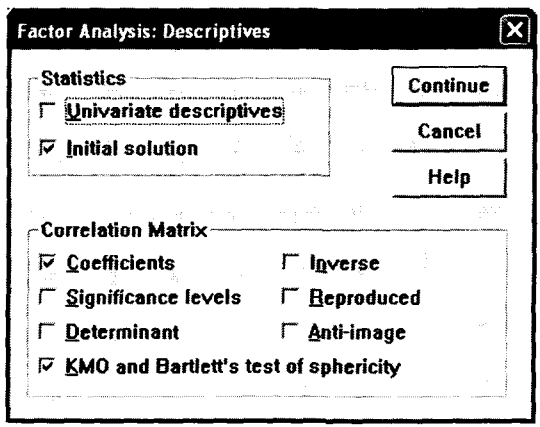


Рис. 7.6. Диалоговое окно «Описательные статистические показатели»

В диалоговом окне «Описательные статистические показатели» следует сделать отметку напротив команды «*KMO and Bartlett's test of sphericity*». Таким образом, делается заявка на проведение тестов «*KMO*» и «*Bartlett*», которые проверяют пригод-

ность исходных данных для проведения факторного анализа[1]. Результаты этих тестов будут представлены далее (см. рис. 7.10).

Также в диалоговом окне «Описательные статистические показатели» в поле «*Correlation Matrix*» следует сделать отметку напротив команды «*Coefficients*». Это означает, что в корреляционной матрице будут представлены коэффициенты корреляции между исследуемыми переменными.

Нажав кнопку «*Continue*» в диалоговом окне «Описательные статистические показатели», мы закрываем данное окно и вновь активизируем главное диалоговое окно «Факторный анализ» (см. рис. 7.5).

Далее необходимо задать условия определения количества факторов, т.е. групп, на которые будет делиться исходный массив переменных. Для этого необходимо открыть диалоговое окно «Извлечение» («*Extraction*») (рис. 7.7) путем нажатия одноименной кнопки в диалоговом окне «Факторный анализ» (*Factor Analysis*) (см. рис. 7.5).

---

[1] **Тесты «*KMO*» и «*Bartlett*»** проводятся с целью оценки пригодности исходных данных для факторного анализа, в результате проведения которого переменные исходного массива данных объединяются в группирующие факторы (компоненты факторной модели).

Тест «*Bartlett*» проводится с целью подтвердить наличие корреляционных связей между переменными исходного массива данных. Группировка переменных исходного массива данных в компоненты факторной модели производится на основе показателя плотности корреляционных связей между ними. Отсутствие таких связей делает невозможным проведение факторного анализа.

Тест «*Bartlett*» проверяет гипотезу об отсутствии корреляционных связей между переменными исходного массива данных. Верность гипотезы определяется с помощью показателя «*Significance*» («Значимость»). Если значение «*Significance*» меньше (больше) 0,05, то вероятность ошибки при отклонении исходной гипотезы не превышает (или превышает) допустимый уровень 5% (при доверительном интервале 95%), т.е. исходная гипотеза неверна (или верна). Опровержение исходной гипотезы свидетельствует о наличии корреляционных связей между переменными исходного массива данных и возможности проведения факторного анализа.

Тест «*KMO*» позволяет оценить, в какой мере построенная факторная модель полно описывает структуру ответов респондентов на вопросы анкеты, представляющие исходные переменные. Значение теста варьирует от нуля (факторная модель абсолютно неприемлема) до единицы (факторная модель идеально описывает исходную структуру данных). Результаты факторного анализа могут считаться действительными, если значение теста «*KMO*» превышает 0,05. Это свидетельствует о приемлемости построенной факторной модели и, следовательно, о пригодности исходного массива данных для проведения факторного анализа.

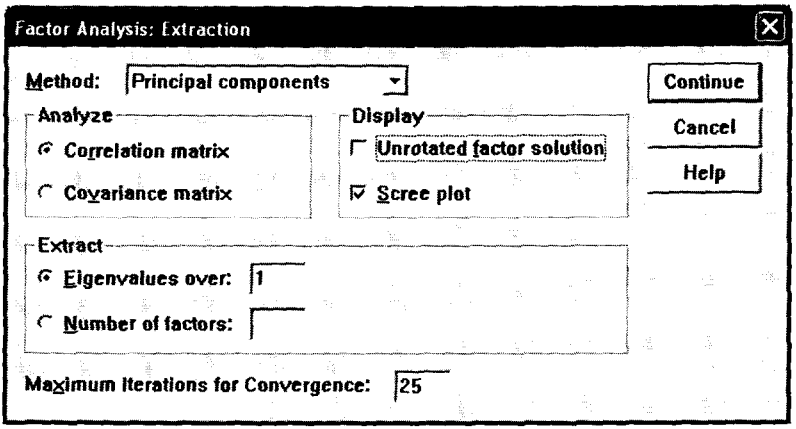


Рис. 7.7. Диалоговое окно «Извлечение»

В диалоговом окне «Извлечение» задается метод извлечения (формирования) факторов. В нашем примере выбирается наиболее распространенный метод «Основных компонентов» (*Principal components*). Также задается выведение на экран корреляционной матрицы (*Correlation matrix*). Результаты выполнения этой команды будут представлены далее (см. рис. 7.11).

Число факторов может быть задано «вручную» путем указания нужного числа в поле «*Number of factors*» (см. рис. 7.7). Однако для проведения качественного факторного анализа число факторов должно быть определено при помощи расчета «характеристических чисел» (*Eigenvalues*). В нашем примере задается условие: значение «характеристических чисел» должно быть больше «1» (*Eigenvalues over* 1). На основании выполнения этого условия будет определяться количество факторов. Результаты выполнения этой команды будут представлены далее (см. рис. 7.12).

Число факторов может быть также определено графическим способом. Для этого на экран нужно вывести график «характеристических чисел», который запрашивается путем отметки команды «*Scree plot*» (см. рис. 7.7). Результаты выполнения этой команды будут представлены далее (см. рис. 7.10).

Путем нажатия кнопки «*Continue*» в диалоговом окне «Извлечение» (см. рис. 7.7) данное окно закрывается и активизируется вновь главное диалоговое окно «Факторный анализ» (см. рис. 7.5).

Следующим шагом выполнения факторного анализа является ротация матрицы коэффициентов. Метод ротации можно вы-

брать в диалоговом окне «Ротация» (*Rotation*) (рис. 7.8), которое открывается при нажатии одноименной кнопки в главном диалоговом окне «Факторный анализ» (*Factor Analysis*) (см. рис. 7.5):

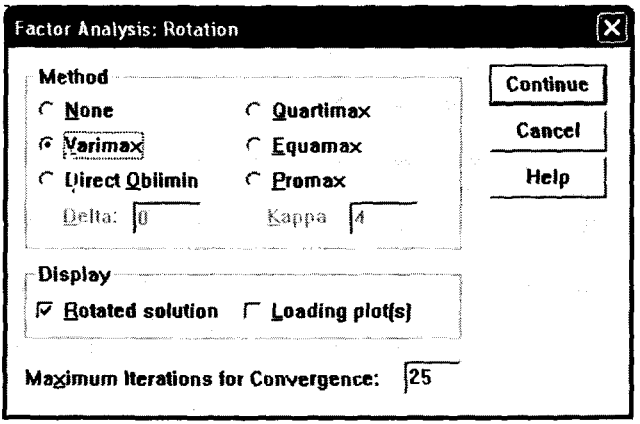


Рис. 7.8. Диалоговое окно «Ротация»

Как показано на рис. 7.8, в нашем примере в качестве метода ротации выбран наиболее распространенный «*Varimax*». Также задана команда вывода на экран ротированных значений матрицы коэффициентов (*Rotated solution*). Результаты выполнения этой команды будут представлены в табл. 7.5. После нажатия кнопки «*Continue*» в диалоговом окне «Ротация» (см. рис. 7.8) данное окно закрывается и снова активизируется главное диалоговое окно «Факторный анализ» (см. рис. 7.5).

Перед запуском процедуры факторного анализа необходимо запросить создание новых переменных в исходной базе данных. В рассматриваемом примере каждая новая переменная будет представлять собой цель отдыха туристов, объединяющая в себе несколько целей отдыха из их общего числа, заданного изначально. Создание новых переменных запрашивается в диалоговом окне «Значения факторов» (*Factor Scores*) (рис. 7.9), которое открывается при нажатии одноименной кнопки в диалоговом окне «Факторный анализ» (*Factor Analysis*) (см. рис. 7.5).

При отметке команды «Сохранить как переменную» (*Save as variable*), вновь созданные обобщающие факторы будут сохраняться в базе данных как новые переменные. В качестве метода расчета значений для этих новых переменных в рассматриваемом примере выбирается регрессионная модель (*Regression*) (рис. 7.9).

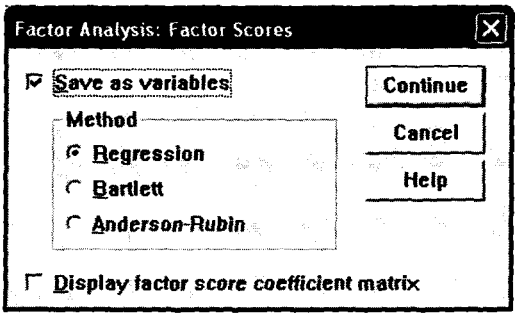


Рис. 7.9. Диалоговое окно «Значения факторов»

После нажатия кнопки «*Continue*» в диалоговом окне «Значения факторов» данное окно закрывается и происходит возврат в главное диалоговое окно «Факторный анализ» (см. рис. 7.5).

Запуск процедуры выполнения факторного анализа осуществляется нажатием кнопки «*OK*» в диалоговом окне «Факторный анализ» (*Factor Analysis*).

7.3. ОЦЕНКА ПРИГОДНОСТИ ИСХОДНЫХ ДАННЫХ ДЛЯ ВЫПОЛНЕНИЯ ФАКТОРНОГО АНАЛИЗА

В первой таблице, выводимой на экран компьютера после запуска процедуры факторного анализа, содержатся результаты тестов «*KMO*» и «*Bartlett*» (табл. 7.2).

Таблица 7.2

Результаты тестов «*KMO*» и «*Bartlett*»

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,668
Bartlett's Test of Sphericity	Approx. Chi-Square	7063,154
	df	66
	Sig.	,000

Результаты теста «*KMO*» позволяют сделать вывод об общей пригодности имеющихся данных для факторного анализа. Он позволит оценивать, насколько полно построенная факторная модель описывает структуру ответов респондентов на вопросы анкеты, представляющие исследуемые переменные. Результаты данного теста варьируются в интервале от нуля (факторная модель абсолютно неприменима) до единицы (факторная модель идеально описывает структуру данных). Результаты факторного анализа могут считаться действительными, если значение теста «*KMO*» более 0,5. В рассматриваемом примере значение этого теста 0,668 (табл. 7.2), что свидетельствует о приемлемости построенной факторной модели.

Тест «*Bartlett*» проверяет гипотезу, согласно которой между переменными, участвующими в факторном анализе, не существует корреляционной зависимости. В рассматриваемом примере это означает, что заданные 12 целей отдыха туристов никак не связаны между собой, и поэтому их группировка с целью уменьшения числа целей отдыха невозможна.

Из данных табл. 7.2 видно, что значимость теста «*Bartlett*» (*Sig.*) составляет 0,000. Это означает, что исходная гипотеза может быть отклонена с вероятностью ошибки 0,000, т.е. она неверна, а также свидетельствует о том, что корреляционные связи между переменными исходного массива существуют и возможна их группировка на основании тесноты корреляции.

Исходя из результатов тестов «*KMO*» и «*Bartlett*» в рассматриваемом примере можно сделать вывод о пригодности исходных данных нашего примера для проведения факторного анализа.

7.4. ВЫЯВЛЕНИЕ КОРРЕЛЯЦИОННОЙ ЗАВИСИМОСТИ МЕЖДУ ПЕРЕМЕННЫМИ ИСХОДНОГО МАССИВА

В качестве результатов факторного анализа на экран компьютера выводится матрица коэффициентов корреляции, характеризующих плотность связи между переменными исходного массива данных (табл. 7.3).

В таблице «Матрица коэффициентов корреляции» представлены результаты попарного сравнения переменных исходного

Таблица 7.3

Матрица коэффициентов корреляции

Correlation Matrix

		Спокойный отдых	Спорт и физические нагрузки	Удовольствия и развлечения	Знакомства и общение	Красота и здоровье	История и традиции	Общение с природой	Культурная программа	Игры, фестивали и пр.	Путешествия в соседние страны	Товары народного промысла	Ночной/вечерний досуг
Correlation	Спокойный отдых	1,000	–,066	–,004	,070	,290	,060	,190	–,040	–,008	,090	,120	–,084
	Спорт и физические нагрузки	–,066	1,000	,032	–,021	,185	,001	,177	–,046	,026	–,021	,012	,114
	Удовольствия и развлечения	–,004	,032	1,000	,425	–,017	,036	,005	,158	,186	,125	,120	,361
	Знакомства и общение	,070	–,021	,425	1,000	,093	,181	,077	,144	,197	,137	,176	,295
	Красота и здоровье	,290	,185	–,017	,093	1,000	,083	,154	–,067	,077	,021	,-77	,021
	История и традиции	,060	,001	,036	,181	,083	1,000	,213	,300	,152	,126	,249	,037
	Общение с природой	,190	,177	,005	,077	,154	,213	1,000	,116	,036	,130	,193	–,100
	Культурная программа	–,040	–,046	,158	,144	–,067	,300	,116	1,000	,198	,160	,200	,108
	Игры, фестивали и пр.	–,008	,026	,186	,197	,077	,152	,036	,198	1,000	,124	,167	,084
	Путешествия в соседние страны	,090	–,021	,125	,137	,021	,126	,130	,160	,124	1,000	,271	,069
	Товары народного промысла	,120	,012	,120	,176	,077	,249	,193	,200	,167	,271	1,000	,069
	Ночной/вечерний досуг	–,084	,114	,361	,295	,021	,037	–,100	,108	,229	,084	,069	1,000

массива. Когда одна переменная сравнивается сама с собой, коэффициент корреляции принимает максимальное значение («1»), что свидетельствует об абсолютной идентичности. Значение коэффициента корреляции между третьей переменной («удовольствия и развлечения») и четвертой переменной («круг общения и знакомства») составляет 0,425. Это свидетельствует о достаточно высокой степени взаимосвязи между переменными и является основанием для объединения их в одну группу при построении факторной модели.

Значение коэффициента корреляции между второй переменной («спорт и физические нагрузки») и шестой переменной («история и традиции») составляет 0,001 (см. табл. 7.3). Это свидетельствует о слабой взаимосвязи между переменными и является основанием для их распределение в разные группы.

Положительное значение коэффициента корреляции свидетельствует о прямой взаимосвязи между исследуемыми переменными. Например, значение коэффициента корреляции переменных «1» («спокойный отдых») и «5» («красота и здоровье») положительное (0,290). Это говорит о том, что спокойный отдых в некоторой степени способствует заботе о красоте и здоровье.

Следует также отметить, что значение коэффициента корреляции между переменными «2» («спорт и физические нагрузки») и «5» («красота и здоровье») составляет всего 0,185 (< 0,290). Из этого следует, что, по мнению респондентов, занятия спортом в меньшей степени способствуют заботе о красоте и здоровье, чем спокойный отдых.

Отрицательное значение коэффициента корреляции свидетельствует об обратной взаимосвязи между исследуемыми переменными. Например, значение коэффициента корреляции переменных «1» («спокойный отдых») и «2» («спорт и физические нагрузки») отрицательное (−0,066). Это говорит о том, что, по мнению респондентов, занятия спортом в некоторой степени мешают спокойному отдыху.

Какое именно значение коэффициента корреляции является основанием для объединения исследуемых переменных в одну группу (в один фактор), зависит от числа групп (факторов), построенных в факторной модели. Чем больше число групп (факторов), тем выше должны быть значения коэффициентов корреляции переменных, группируемых под одним фактором.

7.5. ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО ЧИСЛА КОМПОНЕНТОВ ФАКТОРНОЙ МОДЕЛИ

Как уже было отмечено выше, число групп (компонентов) факторной модели определяется при помощи расчета «характеристических чисел» (*Eigenvalues*). Эти показатели характеризуют полноту отображения исходной информации в построенной факторной модели.

Значения этих показателей содержатся в таблице «*Total Variance Explained*», которая выводится на экран компьютера среди прочих результатов факторного анализа (табл. 7.4).

Таблица 7.4

Определение числа компонентов факторной модели

Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,345	19,544	19,544	1,885	15,710	15,710
2	1,600	13,336	32,881	1,859	15,496	31,205
3	1,304	10,865	43,745	1,413	11,778	42,983
4	1,103	9,191	52,936	1,194	9,953	52,936
5	,929	7,740	60,677			
6	,882	7,351	68,028			
7	,741	6,178	74,206			
8	,716	5,967	80,173			
9	,676	5,632	85,805			
10	,623	5,191	90,996			
11	,565	4,704	95,700			
12	,516	4,300	100,000			

Extraction Method: Principal Component Analysis.

В первом столбце табл. 7.4 (*Component*) указывается число компонентов различных вариантов факторной модели. В четвертом столбце этой таблицы (*Cumulative, %*) показан процент информации, сохраненной в процессе группировки исходного массива переменных с помощью факторной модели. Например, если число факторов в факторной модели равно числу переменных исходного массива (в нашем примере 12), т.е. группировка переменных не производится, исходная информация будет сохранена на 100%.

Во втором столбце таблицы (*Total*) указываются значения «характеристических чисел» (*Eigenvalues*). В рассматриваемом примере

было задано условие: значение «характеристических чисел» должно быть больше единицы (*Eigenvalues over* 1) (см. рис. 7.7). Максимальное значение компонентов факторной модели, в которой данный показатель превышает единицу, составляет 4. Это означает, что оптимальное число групп (факторов) в факторной модели составляет 4.

Как видно из данных, представленных в табл. 7.4, факторная модель, состоящая из 4-х факторов, сохраняет лишь 52,936% исходной информации. Как отмечалось ранее, при группировке исходного массива переменных потеря информации неизбежна. При построении факторной модели следует стремиться к минимизации потерь информации.

Сохранение информации всего лишь на 52,936% является не очень хорошим показателем. Однако, принимая во внимание, что в ходе факторного анализа число переменных сократится в 3 раза (с 12 до 4), а потеря информации составит менее 48%, применение построенной факторной модели следует считать целесообразным.

В ходе формирования задания на проведение факторного анализа также было запрошено построение графика «*Screen plot*» (см. рис. 7.7), с помощью которого можно также определить оптимальное число групп. Результаты выполнения этой команды представлены на рис. 7.10.

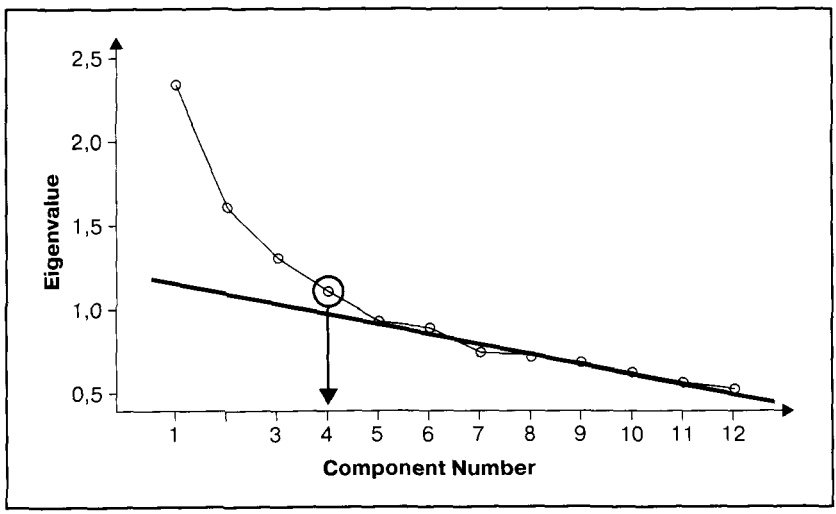


Рис. 7.10. Графическое определение количества компонентов факторной модели

На рис. 7.10 представлен график, отображающий зависимость между «характеристическими числами» (*Eigenvalues*) и числом компонентов факторной модели (*Component Number*). При изменении количества факторов с 5 до 12 данный график представляет собой практически линейную функцию, а при уменьшении числа факторов с 5 до 4 происходит «перелом» графика. Это означает, что оптимальное число компонентов факторной модели (факторов) равно 4.

Таким образом, результаты графического метода определения числа факторов подтвердили результаты расчетного метода (см. табл. 7.4). В результате применения обоих методов оптимальное число компонентов факторной модели составило 4.

7.6. ПОСТРОЕНИЕ ФАКТОРНОЙ МОДЕЛИ И ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

Следующим шагом в представлении результатов факторного анализа является ротированная матрица компонентов (табл. 7.5).

Таблица 7.5

Ротированная матрица компонентов факторной модели

***Rotated Component Matrix*[a]**

	Component			
	1	2	3	4
Спокойный отдых	–,041	,095	,809	–,191
Спорт и физические нагрузки	,078	–,047	,002	,886
Удовольствия и развлечения	,762	,058	,013	–,064
Знакомства и общение	,683	,193	,197	–,102
Красота и здоровье	,083	–,011	,676	,365
История и традиции	,006	,684	–,006	,098
Общение с природой	–,165	,481	,322	,383
Культурная программа	,143	,649	–,295	–,039
Игры, фестивали и пр.	,428	,310	–,066	,105
Путешествия в соседние страны	,161	,467	,143	–,178
Товары народного промысла	,130	,609	,178	–,037
Ночной/вечерний досуг	,742	–,047	–,129	,150

Метод извлечения: анализ главных компонентов.
Метод ротации: «*Varimax*» с нормализацией Кайзера.
[a] Ротация получена за 5 итераций.

В табл. 7.5 представлены коэффициенты корреляции, характеризующие связи между переменными исходного массива данных и компонентами построенной факторной модели (факторами). Согласно общему правилу проведения факторного анализа в одну группу (под одним фактором) собираются переменные исходного массива, имеющие наиболее тесную связь (самое большое значение коэффициента корреляции) с данным компонентом факторной модели.

В табл. 7.5 отмечены максимальные значения коэффициентов корреляции, свидетельствующие о наиболее тесной взаимосвязи переменных исходного массива с компонентами факторной модели. На основе этих данных производится группировка переменных исходного массива, представленная в табл. 7.6.

Таблица 7.6

Факторный анализ.
Группировка переменных исходного массива данных

Компоненты факторной модели	Переменные исходного массива	Коэффициенты корреляции
Фактор «1»	Удовольствия и развлечения	0,762
	Знакомства и общение	0,683
	Игры, фестивали и т.п.	0,428
	Ночной/ вечерний досуг	0,742
Фактор «2»	История и традиции	0,684
	Общение с природой	0,481
	Культурная программа	0,649
	Путешествия в соседние страны	0,467
	Товары народного промысла	0,609
Фактор «3»	Спокойный отдых	0,809
	Красота и здоровье	0,676
Фактор «4»	Спорт и физические нагрузки	0,886

Следующим шагом факторного анализа является интерпретация результатов, т.е. определение названия каждого фактора (компонента факторной модели). Название фактора подбирается специалистами, проводящими исследование, исходя из логики и названий переменных, объединенных этим фактором.

В нашем примере были подобраны следующие названия компонентов факторной модели:

- *Фактор «1» – «Развлечения».*
- *Фактор «2» – «Специальные предложения Восточной Баварии».*
- *Фактор «3» – «Спокойный отдых».*
- *Фактор «4» – «Спорт».*

Первый фактор получил название «Развлечения», поскольку он объединяет переменные исходного массива, так или иначе связанные с развлечениями и увеселительными мероприятиями.

Переменные, объединенные фактором «2», связаны с культурной программой отдыха. Лишь одна переменная – «общение с природой» – кажется случайно попавшей в эту группу. Однако ее можно интерпретировать таким образом, что для туристов, ценящих общение с природой, важны такие мотивы выбора места отдыха, как уникальный ландшафт, лесные массивы и водоемы Восточной Баварии. При такой интерпретации переменная «общение с природой» близка по смыслу переменным, связанным с уникальной культурной программой отдыха в Восточной Баварии. Таким образом, фактор «2» получил название «Специальные предложения Восточной Баварии».

Фактор «3», объединяющий переменные «спокойный отдых» и «красота и здоровье», получил название «Спокойный отдых». Переменная «спорт и физические нагрузки» оказалась единственной переменной в группе «4» – «Спорт».

Как отмечалось ранее, при построении факторной модели неизбежна частичная потеря информации. Потеря информации особо ощутима в случае, если отдельные переменные исходного массива данных имеют высокие значения коэффициентов корреляции сразу с несколькими факторами.

Например, переменная «красота и здоровье» имеет достаточно высокий коэффициент корреляции с фактором «Спорт» (0,676) и фактором «Спокойный отдых» (0,365) (см. табл. 5.5). Это говорит о том, что туристы, занимающиеся спортом, достаточно большое внимание уделяют заботе о красоте и здоровье, но все же в меньшей степени, чем туристы, предпочитающие спокойный отдых. Построенная функциональная модель ведет к частичной потере информации, поскольку предполагает, что забота о красоте и здоровье важна только для туристов, предпочитающих спокойный отдых, и не важна для других туристов.

Переменная «общение с природой» также имеет высокие коэффициенты корреляции с факторами «Специальные предложения Восточной Баварии» (0,481), «Спокойный отдых» (0,322) и «Спорт» (0,383) (см. табл. 7.5). Это свидетельствует о том, что общение с природой важно как для туристов, занимающихся спортом, так и для тех, кто предпочитает спокойный отдых. Факторная модель связана с частичной потерей информации, поскольку предполагает, что общение с природой важно только для

туристов, интересующихся специальными предложениями Восточной Баварии.

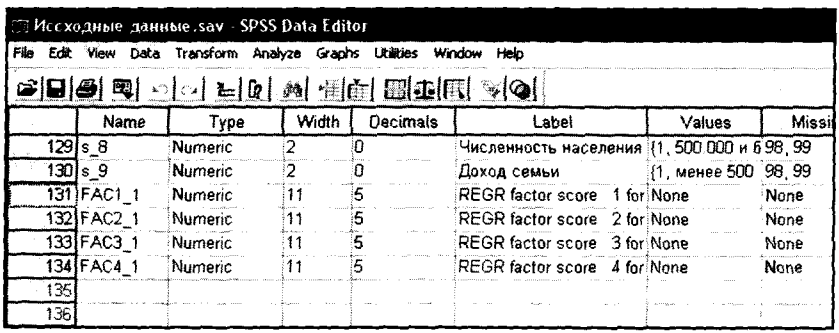
Иллюстрация результатов факторного анализа представлена на рис. 7.11.

Рис. 7.11. Иллюстрация результатов факторного анализа

Несмотря на то что факторная модель ведет к существенной потере информации исходного массива данных (почти 47%), применение данной модели является весьма целесообразным. Как уже было отмечено выше, при потере информации менее чем наполовину, число переменных исходного массива уменьшается в 3 раза.

7.7. СОХРАНЕНИЕ КОМПОНЕНТОВ ФАКТОРНОЙ МОДЕЛИ В КАЧЕСТВЕ НОВЫХ ПЕРЕМЕННЫХ БАЗЫ ДАННЫХ

В ходе формирования задания на выполнение факторного анализа была задана команда на создание новых переменных в базе данных. Результатом выполнения этой команды является автоматический перенос компонентов построенной факторной модели (см. рис. 7.11) в базу данных как новых переменных (рис. 7.12).


Исходные данные.sav - SPSS Data Editor

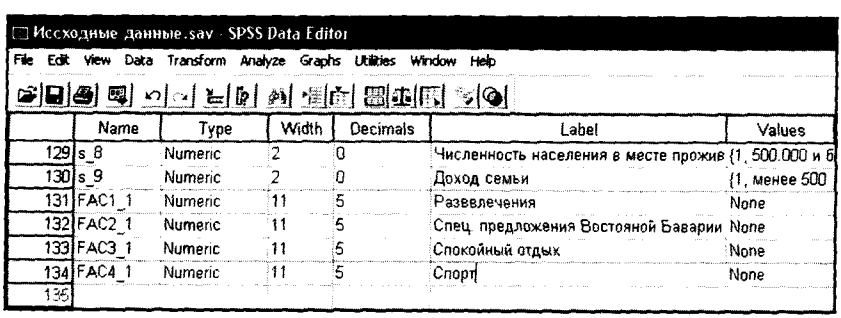
File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missi
129	s_8	Numeric	2	0	Численность населения	{1, 500 000 и б	98, 99
130	s_9	Numeric	2	0	Доход семьи	{1, менее 500	98, 99
131	FAC1_1	Numeric	11	5	REGR factor score 1 for	None	None
132	FAC2_1	Numeric	11	5	REGR factor score 2 for	None	None
133	FAC3_1	Numeric	11	5	REGR factor score 3 for	None	None
134	FAC4_1	Numeric	11	5	REGR factor score 4 for	None	None
135							
136							

Рис. 7.12. Фрагмент вкладки «Свойства переменных» (*Variable View*)

При сохранении компонентов построенной факторной модели как новых переменных в файле данных *SPSS* в столбце «*Label*» таблицы «Свойства переменных» отображается номер компонента факторной модели (см. рис. 7.12).

В ходе интерпретации результатов факторного анализа каждый компонент построенной факторной модели получил свое название, которое следует занести в исходный файл данных *SPSS* в столбец «*Label*» таблицы «Свойства переменных» (рис. 7.13).


Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values
129	s_8	Numeric	2	0	Численность населения в месте прожив	{1, 500 000 и б
130	s_9	Numeric	2	0	Доход семьи	{1, менее 500
131	FAC1_1	Numeric	11	5	Развлечения	None
132	FAC2_1	Numeric	11	5	Спец. предложения Востояной Баварии	None
133	FAC3_1	Numeric	11	5	Спокойный отдых	None
134	FAC4_1	Numeric	11	5	Спорт	None
135						

Рис. 7.13. Фрагмент вкладки «Свойства переменных» (*Variable View*)

При сохранении компонентов факторной модели как новых переменных в базе данных компьютер автоматически вычисляет значения новых переменных (рис. 7.14).

Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : nummer 2149

	s_9	FAC1_1	FAC2_1	FAC3_1	FAC4_1
460	4	,02443	-1,63184	,14764	-,57967
461	5	,25598	-,37368	-,86674	-,09097
462	8	1,36906	-1,20395	-,62135	,37462
463	99	-,67613	-,53421	,09530	-,77137
464	4	1,31253	-,56986	,43413	,26466
465	3	,42697	,75549	-,52051	,53864
466	6	,21270	-1,44238	-,01936	,47291
467	2	1,24448	,08720	-,42548	1,11570
468	4	1,09471	-,81380	-,79341	-,58547

Рис. 7.14. Фрагмент вкладки «Значения переменных» (*Data View*)

Значения новых переменных (факторов), представленные на рис. 7.14, не совпадают с кодировкой переменных исходного массива. Ответы респондентов на вопрос анкеты «Что для Вас самое важное во время отдыха?» варьировались от «1» («очень важно») до «5» («совсем не важно»).

При вычислении средних балльных оценок значения компонента факторной модели производится трансформация балльных оценок данных из интервала «от 1 до 5» в интервал «от –2 до 2».

Значение новых переменных (см. рис. 7.14) следует интерпретировать таким образом: чем больше отрицательное значение новой переменной, тем она важнее; чем больше положительное значение новой переменной, тем она менее важна.

Новые переменные могут быть использованы для дальнейших исследований. В рассматриваемом примере факторный анализ произведен с целью сокращения исходного массива данных для дальнейшего проведения кластерного анализа.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите цель проведения и возможности использования результатов факторного анализа.
2. Что представляет собой результирующая факторная модель?

3. Какие преобразования происходят с исходным массивом данных в результате проведения факторного анализа?
4. Какие задачи решаются в ходе проведения факторного анализа?
5. В чем заключается сложность факторного анализа и какие проблемы неизбежно возникают в ходе его выполнения?
6. С какой целью в ходе выполнения факторного анализа производятся тесты «*KMO*» и «*Bartlett*», как следует интерпретировать результаты, если значение теста «*KMO*» составляет 0,742, а значение величины «*Significance*» («Значимость») по результатам теста «*Bartlett*» – 0,02?
7. Что представляет собой матрица коэффициентов корреляции, выводимая в *SPSS* на экран компьютера среди результатов факторного анализа, какие выводы можно сделать на основе данных этой таблицы?
8. Как осуществляется определение оптимального количества компонентов факторной модели расчетным и графическим способами?
9. Как можно интерпретировать результаты определения оптимального количества компонентов факторной модели, если в таблице «*Total Variance Explained*» минимальное значение характеристических чисел, превышающее единицу, находится в пятой строке, т.е. значение в столбце «*Component*» составляет 5, а в столбце «*Cumulative %*» – 74,206?
10. Что представляет собой ротированная матрица компонентов факторной модели, выводимая *SPSS* на экран компьютера?
11. Каким образом данные этой таблицы используются для построения факторной модели?
12. Каким образом осуществляется подбор названий компонентов факторной модели, построенной в результате проведения факторного анализа?
13. С какой целью и каким образом компоненты факторной модели сохраняются в качестве новых переменных в исходном файле данных *SPSS*?

8. ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ В SPSS

8.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В SPSS

Кластерный анализ производится с целью выделения однородных групп (кластеров) из исследуемой совокупности объектов (потребителей, продуктов, брендов и т.п.). Формируемые группы (кластеры) должны быть однородными (гомогенными) внутри и разнородными (гетерогенными) по отношению друг к другу по заданным характеристикам.

Например, из всей совокупности туристов, отдыхающих в курортной зоне «Баварский лес», требуется выделить группы, однородные по возрасту и интересам (мотивам проведения времени на отдыхе). Для анализа используется выборка в размере 6396 человек. Следовательно, в проведении кластерного анализа участвуют 6396 объектов и две переменные, по которым будет производиться разделение объектов на однородные группы (кластеры): возраст и интересы (мотивы проведения времени на отдыхе) туристов (табл. 8.1).

Таблица 8.1

Структура исходного массива данных для проведения кластерного анализа

№ п/п	Объекты исследования (туристы)	Характеристики объектов (переменные, по которым производится разделение на кластеры)	
		Возраст	Интересы (мотивы поведения)
1	Турист № 1		
2	Турист № 2		
...	...		
6396	Турист № 6396		

В ходе факторного анализа (см. предыдущий раздел) были выделены 4 группы интересов (4 мотива поведения) туристов: «Развлечения», «Специальные предложения Восточной Баварии», «Спокойный отдых» и «Спорт».

Каждая из перечисленных групп интересов представлена в базе данных как самостоятельная переменная (см. рис. 8.13 и 8.14). Это означает, что каждый турист во время отдыха руководствуется не каким-то одним из четырех мотивов поведения, а всеми сразу. Различия между туристами состоят лишь в том, что перечисленные интересы (мотивы поведения) важны для каждого в разной степени. Следовательно, в ходе кластерного анализа следует выделить группы туристов, однородные по возрасту и по структуре их интересов.

Поскольку характеристика «интересы туристов» является множественной переменной, т.е. в базе данных она представлена в виде 4 переменных, то исходный массив данных (см. табл. 8.1) является трехмерным. Метод кластерного анализа позволяет обрабатывать лишь двухмерные массивы данных (объекты и их характеристики). Для того чтобы проведение кластерного анализа стало возможным, необходимо преобразовать структуру исходного массива данных (табл. 8.2).

Таблица 8.2

Преобразованная структура исходного массива данных для проведения кластерного анализа

№ п/п	Объекты исследования (возрастные группы туристов)	Интересы туристов (переменные, по которым производится разделение на кластеры)			
		Развлечения	Специальные предложения Восточной Баварии	Спокойный отдых	Спорт
1	17–18 лет				
2	19–24 года				
3	25–29 лет				
4	30–34 года				
...					
11	65–70 лет				

В табл. 8.2 вносятся оценки туристами степени, в какой они руководствуются теми или иными интересами при проведении времени на отдыхе. Данные оценки являются средними по каждой возрастной группе.

Разделение возрастных групп на категории (например, от 25 до 29 лет) было произведено в целях сокращения числа объектов

исследования. В проведении исследований участвовали туристы в возрасте от 17 до 70 лет. Если бы в качестве объектов исследования были взяты возрастные группы, объединяющие только туристов определенного возраста (например, 17 лет, 18 лет... 44 года и т.д.), то число объектов исследования составило бы 63 (70 – 17). Такое большое число объектов исследования существенно затрудняет интерпретацию результатов кластерного анализа. Разделение возрастных групп на категории привело к сокращению числа объектов исследования (возрастных групп туристов) с 63 до 11.

Иллюстрация постановки цели кластерного анализа в нашем примере представлена на рис. 8.1. Для проведения кластерного анализа в *SPSS* создается новый файл данных (рис. 8.2 и 8.3).

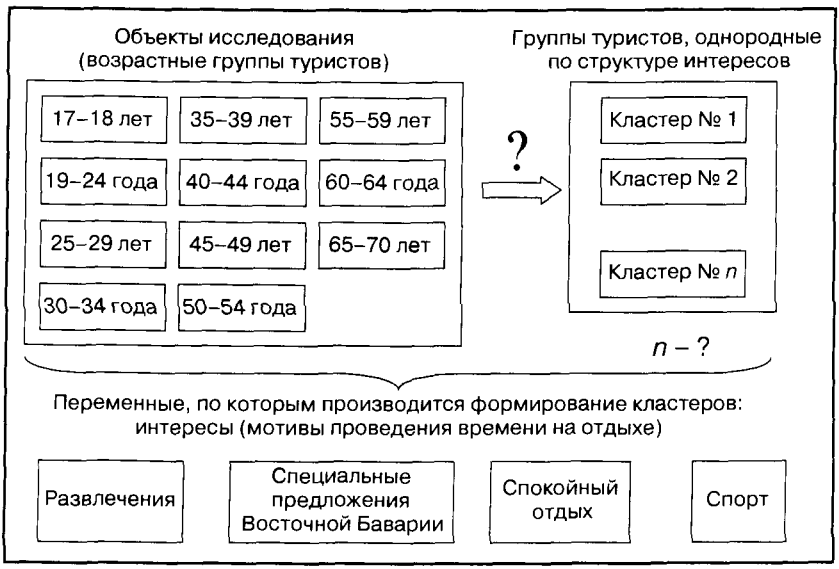
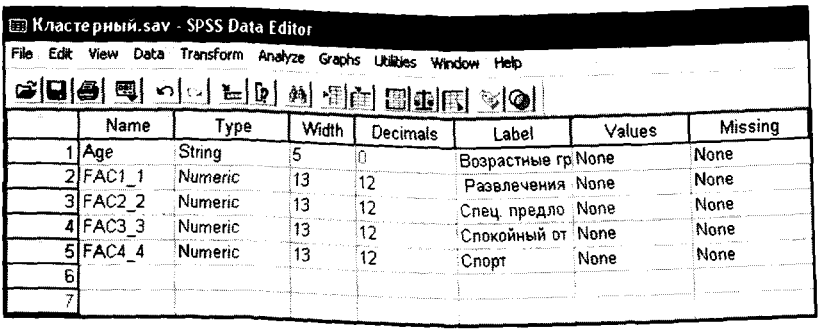


Рис. 8.1. Иллюстрация постановки цели исследования кластерного анализа

На рис. 8.2 представлен фрагмент исходного файла данных, состоящего из 5 переменных. Первая переменная с именем «*Age*» и меткой «Возрастные группы» является текстовой переменной, об этом есть соответствующая запись (*String*) в столбце «*Type*». Со значениями этой переменной нельзя будет производить никаких арифметических операций.



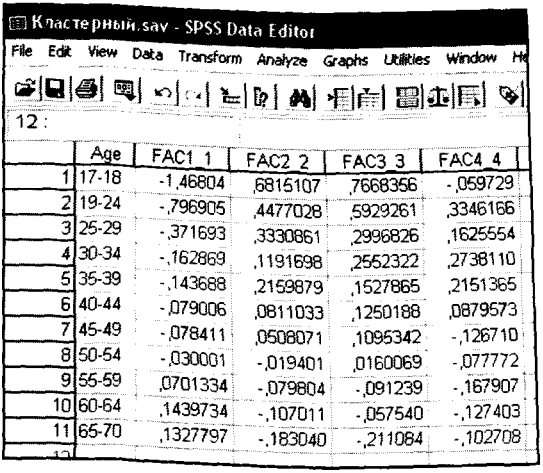
Кластерный.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing
1	Age	String	5	0	Возрастные гр	None	None
2	FAC1_1	Numeric	13	12	Развлечения	None	None
3	FAC2_2	Numeric	13	12	Спец. предло	None	None
4	FAC3_3	Numeric	13	12	Спокойный от	None	None
5	FAC4_4	Numeric	13	12	Спорт	None	None
6							
7							

Рис. 8.2. Фрагмент вкладки «Свойства переменных» (*Variable View*)

Четыре переменные с именами «FAC1_1», «FAC2_1», «FAC3_1» и «FAC4_1» являются компонентами факторной модели, построенной в результате проведения факторного анализа (см. предыдущий раздел). Значения этих переменных представляют собой усредненные балльные оценки важности для туристов каждой возрастной группы следующих интересов: «Развлечения», «Специальные предложения Восточной Баварии», «Спокойный отдых» и «Спорт» (рис. 8.3).



Кластерный.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window He

12 :

	Age	FAC1_1	FAC2_2	FAC3_3	FAC4_4
1	17-18	-1,46804	,6815107	,7668356	-,059729
2	19-24	-,796905	,4477028	,5929261	,3346166
3	25-29	-,371693	,3330861	,2996826	,1625554
4	30-34	-,162869	,1191698	,2552322	,2738110
5	35-39	-,143688	,2159879	,1527865	,2151365
6	40-44	-,079006	,0811033	,1250188	,0879573
7	45-49	-,078411	,0508071	,1095342	-,126710
8	50-54	-,030001	-,019401	,0160069	-,077772
9	55-59	,0701334	-,079804	-,091239	-,167907
10	60-64	,1439734	-,107011	-,057540	-,127403
11	65-70	,1327797	-,183040	-,211084	-,102708

Рис. 8.3. Фрагмент вкладки «Значения переменных» (*Data View*)

Как было описано в предыдущем подразделе, при проведении опроса респондентам предлагалось оценить 12 мотивов проведения времени на отдыхе по 5-балльной шкале («1» – «очень важно»

и «5» – «совсем не важно»). В результате проведения факторного анализа 12 переменных исходного массива данных были сгруппированы в 4 переменные, в ходе проведения анализа произошла трансформация значений переменных. Средняя оценка (3) была приравнена к нулю. Именно поэтому средние значения оценки значения для туристов четырех мотивов поведения, представленные на рис. 6.3, варьируют от –2 до 2. Чем больше отрицательное значение переменной, тем она важнее; чем больше положительное значение переменной, тем она менее важна.

После того как сформирована база данных в *SPSS*, следует перейти непосредственно к заданию набора команд на выполнение кластерного анализа.

8.2. КОМАНДЫ SPSS НА ВЫПОЛНЕНИЕ ИЕРАРХИЧЕСКОГО КЛАСТЕРНОГО АНАЛИЗА

Кластерный анализ является одним из видов классификационного анализа. Для задания команд на выполнение кластерного анализа в *SPSS* в меню различных видов анализа (*Analyze*) следует выбрать «Классификационный анализ» (*Classify*) (рис. 8.4).

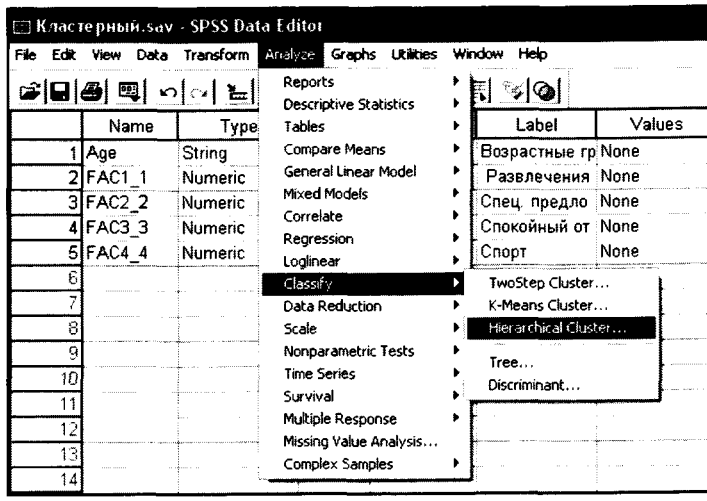


Рис. 8.4. Выбор в меню *SPSS* процедуры «Иерархический кластерный анализ»

«Классификационный анализ», в свою очередь, имеет собственное меню, содержащее различные виды классификационного анализа, в том числе три вида кластерного анализа. В рассматриваемом примере применяется иерархический кластерный анализ, наиболее часто применяемый на практике.

Иерархический кластерный анализ отличается от других видов кластерного анализа тем, что алгоритм его проведения является многоступенчатым. Алгоритм иерархического кластерного анализа может быть дивизионным или агломеративным.

Дивизионный алгоритм проведения иерархического кластерного анализа предполагает, что все объекты исследования в начале объединены в один кластер, который поэтапно делится на более мелкие кластеры. *Агломеративный* алгоритм, напротив, предполагает, что все объекты исследования вначале рассматриваются как мелкие кластеры, которые затем объединяются в более крупные. На практике чаще всего используются агломеративные методы формирования кластеров.

В результате выбора меню «*Analyze*> *Classify*> *Hierarchical Cluster*» на экране появится диалоговое окно «Иерархический кластерный анализ» (*Hierarchical Cluster Analyze*) (рис. 8.5).

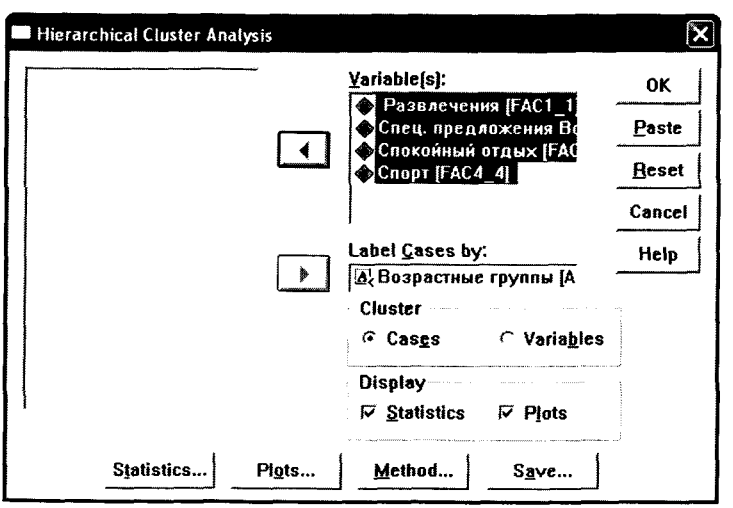


Рис. 8.5. Диалоговое окно «Иерархический кластерный анализ»

В левом поле открывшегося диалогового окна «Иерархический кластерный анализ» представлен список пяти переменных исход-

ного массива данных. Из них следует выбрать переменные, по которым будет производиться формирование кластеров, и перенести их в правое поле «*Variable(s)*». В рассматриваемом примере – это переменные, характеризующие интересы (мотивы поведения) туристов: «Развлечения», «Специальные предложения Восточной Баварии», «Спокойный отдых» и «Спорт».

Также из списка всех переменных исходной базы данных следует выбрать переменную, значения которой являются объектами исследования, и перенести ее в правое поле «*Label Cases by*». В рассматриваемом примере это переменная «возрастные группы».

В поле «*Cluster*» следует выбрать один из двух предлагаемых вариантов: «*Cases*» или «*Variables*» (см. рис. 8.5). В нашем примере выбран вариант «*Cases*». Это означает, что в ходе кластерного анализа будут классифицироваться (собираться в кластеры) возрастные группы туристов, а не их интересы (мотивы поведения).

В диалоговом окне «Иерархический кластерный анализ» также есть четыре кнопки, нажав которые открываются вспомогательные диалоговые окна: «*Statistics*», «*Plots*», «*Method*» и «*Save*».

При нажатии кнопки «*Statistics*» на экране появляется одноименное диалоговое окно «Статистические показатели» (рис. 8.6).

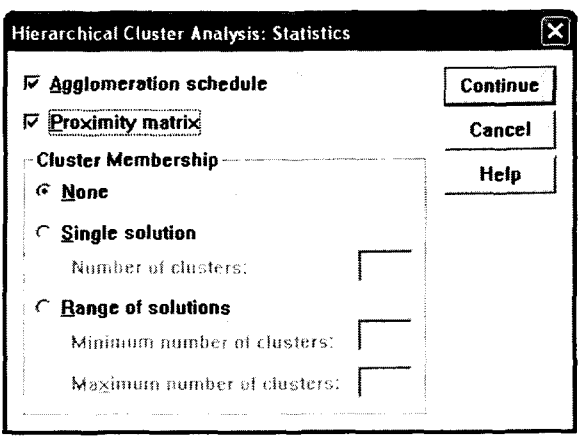


Рис. 8.6. Диалоговое окно «Статистические показатели»

Во вспомогательном диалоговом окне «Статистические показатели» отмечены команды «*Agglomeration schedule*» и «*Proximity matrix*» (см. рис. 8.6). После запуска процедуры выполнения

кластерного анализа данные команды позволяют вывести на экран в качестве результатов анализа таблицу, содержащую результаты сравнения объектов исследования (*Proximity matrix*), и таблицу, отображающую алгоритм формирования кластеров (*Agglomeration schedule*). Путем нажатия кнопки «*Continue*» осуществляется возврат в главное диалоговое окно «Иерархический кластерный анализ».

После нажатия кнопки «*Plots*» в главном диалоговом окне «Иерархический кластерный анализ» на экране появляется одноименное вспомогательное диалоговое окно «Диаграммы» (рис. 8.7).

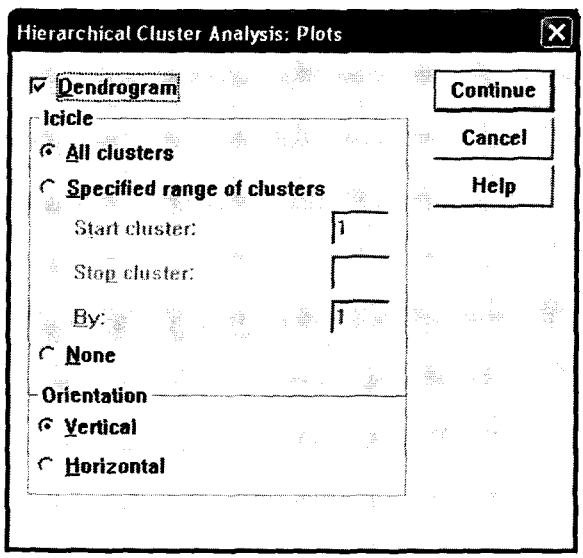


Рис. 8.7. Диалоговое окно «Диаграммы»

В диалоговом окне «Диаграммы» представлены команды на построение различных графиков и диаграмм, описывающих процедуру формирования кластеров. В данном диалоговом окне отмечена команда «*Dendogram*». После запуска процедуры выполнения кластерного анализа данная команда выводит на экран дендограмму, которая является графическим отображением выполнения алгоритма формирования кластеров. Путем нажатия кнопки «*Continue*» (см. рис. 8.7) осуществляется воз-

врат в главное диалоговое окно «Иерархический кластерный анализ» (см. рис. 8.5).

При нажатии кнопки «*Method*» в главном диалоговом окне «Иерархический кластерный анализ» (см. рис. 8.5) на экране появляется одноименное вспомогательное диалоговое окно «Методы» (рис. 8.8).

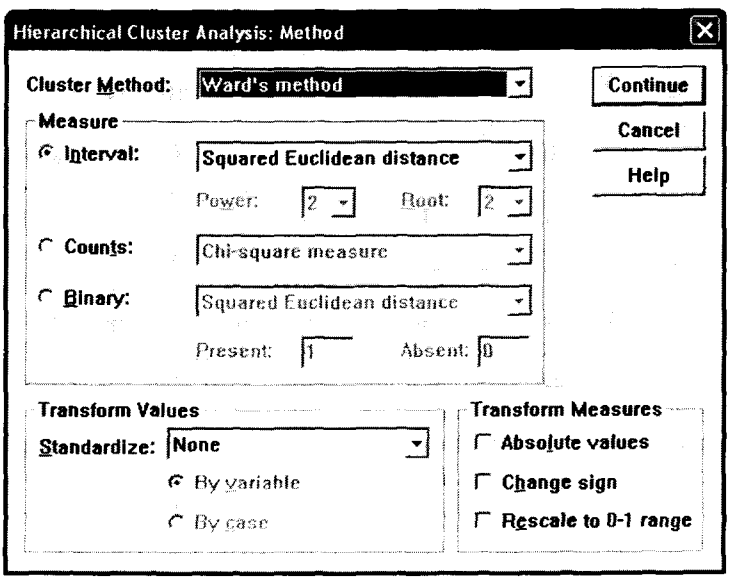


Рис. 8.8. Диалоговое окно «Методы» (*Method*)

В поле «*Cluster Method*» вспомогательного диалогового окна «Методы» из списка, предлагаемого *SPSS*, следует выбрать метод формирования кластеров. В рассматриваемом примере выбран метод «*Ward*».

В поле «*Measure*» из списка возможных вариантов следует выбрать показатель, который будет использоваться в целях определения степени схожести (различия) объектов исследования. Выбор этого показателя зависит от типа переменных, участвующих в кластерном анализе в качестве критериев сегментации. Данные переменные могут быть интервальными (*Interval*), номинальными (*Counts*) или дихотомическими (*Binary*).

В рассматриваемом примере переменные, по которым совокупность объектов исследования разделяется на кластеры, являются интервальными, поскольку респонденты в ходе опроса давали балльные оценки значимости для них различных мотивов проведения времени на отдыхе. Поэтому в поле «*Measure*» диалогового окна «*Method*» отмечается тип переменной «*Interval*». В качестве показателя, характеризующего степень схожести (различия) объектов исследования, выбирается квадрат евклидова расстояния (*Squared Euclidean Distance*).

Путем нажатия кнопки «*Continue*» в диалоговом окне «*Method*» осуществляется возврат в главное диалоговое окно «Иерархический кластерный анализ» (см. рис. 8.5).

В диалоговом окне «Иерархический кластерный анализ» имеется кнопка «*Save*», при нажатии которой активизируется одноименное диалоговое окно. В этом окне представлены команды, позволяющие сохранить результаты кластерного анализа как новые переменные в исходной базе данных. В результате выполнения этих команд после запуска процедуры выполнения кластерного анализа создается новая переменная, значения которой представляют собой номера кластеров, к которым относится тот или иной объект исследования.

Запуск процедуры выполнения иерархического кластерного анализа осуществляется путем нажатия кнопки «*OK*» в главном диалоговом окне «Иерархический кластерный анализ» (см. рис. 8.5).

8.3. СРАВНЕНИЕ ОБЪЕКТОВ ИССЛЕДОВАНИЯ

Среди данных, выдаваемых *SPSS* в качестве результатов кластерного анализа, в первую очередь на экран выводится таблица, содержащая результаты сравнения объектов исследования. Первоочередность представления этих данных в качестве результатов обусловливается агломеративным алгоритмом иерархического кластерного анализа (рис. 8.9).

В нашем примере в качестве показателя, характеризующего степень сходства (различия) объектов исследования, был выбран квадрат евклидова расстояния (*Squared Euclidean Distance*) (см. рис. 8.8). Чем меньше этот показатель, тем больше сходство сравниваемой пары объектов исследования (табл. 8.3).

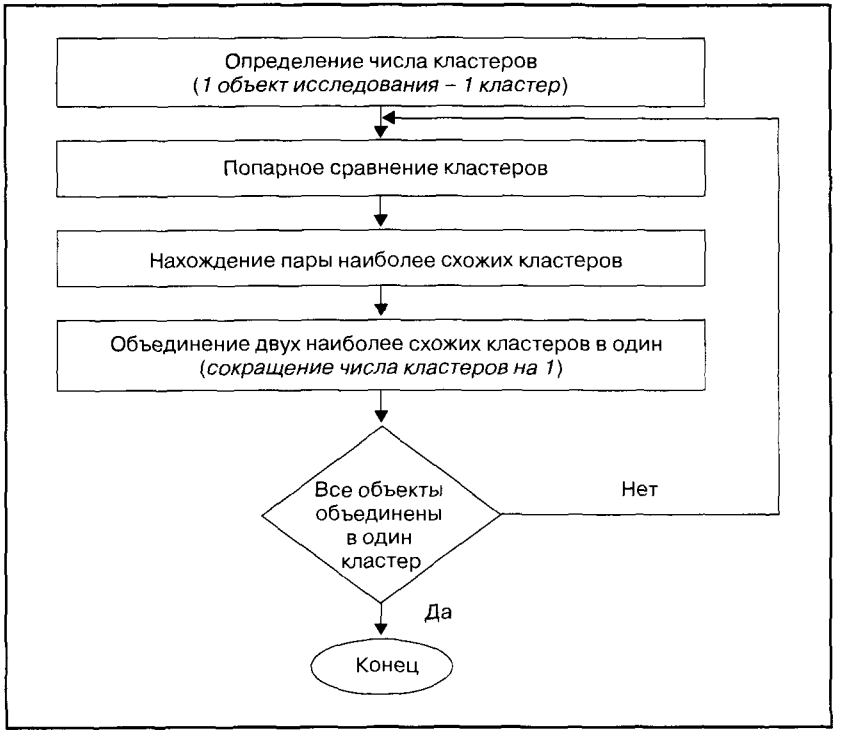


Рис. 8.9. Агломеративный алгоритм иерархического кластерного анализа

Данные табл. 8.3 показывают, в какой степени схожи (различны) между собой разные возрастные категории туристов по структуре их интересов (мотивов проведения времени на отдыхе). Наиболее схожими относительно структуры их интересов являются возрастные категории туристов «9» (55–59 лет) и «10» (60–64 года). Квадрат евклидова расстояния между этими группами составляет всего 0,009 и является минимальным из всех прочих значений этого показателя. Следовательно, данные возрастные категории туристов должны быть объединены в один кластер.

Для определения очередности последующего объединения объектов исследования в кластеры необходимо заново определить квадрат евклидова расстояния между вновь созданным кластером и прочими кластерами.

Таблица 8.3

Результаты сравнения объектов исследования

Proximity Matrix

Case	Squared Euclidean Distance										
	1:17–18	2:19–24	3:25–29	4:30–34	5:35–39	6:40–44	7:45–49	8:50–54	9:55–59	10:60–64	11:65–70
1:17–18	,000	,691	1,591	2,393	2,423	2,724	2,765	3,123	3,694	3,905	4,268
2:19–24	,691	,000	,310	,628	,688	,930	1,120	1,309	1,751	1,830	2,100
3:25–29	1,591	,310	,000	,104	,090	,185	,286	,379	,628	,671	,852
4:30–34	2,393	,628	,104	,000	,024	,060	,193	,218	,409	,404	,538
5:35–39	2,423	,688	,090	,024	,000	,039	,150	,173	,339	,349	,469
6:40–44	2,724	,930	,185	,060	,039	,000	,047	,052	,160	,165	,264
7:45–49	2,765	1,120	,286	,193	,150	,047	,000	,018	,081	,102	,203
8:50–54	3,123	1,309	,379	,218	,173	,052	,018	,000	,033	,046	,105
9:55–59	3,694	1,751	,628	,409	,339	,160	,081	,033	,000	,009	,033
10:60–64	3,905	1,830	,671	,404	,349	,165	,102	,046	,009	,000	,030
11:65–70	4,268	2,100	,852	,538	,469	,264	,203	,105	,033	,030	,000

This is a dissimilarity matrix

Результаты расчета квадратов евклидова расстояния для каждого этапа формирования кластеров не выводятся на экран компьютера. Среди данных, выводимых на экран в качестве результатов кластерного анализа, предоставляются лишь результаты сравнения кластеров на этапе, когда каждый объект исследования рассматривается как кластер.

Данные табл. 8.3 не предоставляют сведений об очередности формирования кластеров. Она дает лишь общее представление о сходстве (различии) объектов исследования. По данным этой таблицы можно сделать лишь приблизительные выводы о том, какие из объектов исследования окажутся объединенными в один кластер.

8.4. ПОРЯДОК ФОРМИРОВАНИЯ КЛАСТЕРОВ

В качестве результатов проведения кластерного анализа в *SPSS* после таблицы с результатами сравнения объектов исследования на экран выводится таблица «График агломерации» (*Agglomeration Schedule*) (табл. 8.4).

Таблица 8.4

График агломерации

Стадия формирования кластера	Объединяемые кластеры		Coefficients	Этап, на котором был сформирован объединяемый кластер		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	10	,004	0	0	4
2	7	8	,014	0	0	7
3	4	5	,026	0	0	5
4	9	11	,045	1	0	7
5	4	6	,074	3	0	6
6	3	4	,159	0	5	9
7	7	9	,258	2	4	9
8	1	2	,603	0	0	10
9	3	7	1,224	6	7	10
10	1	3	4,196	8	9	0

Таблица 8.4 «График агломерации» описывает порядок построения кластеров. В столбце «*Stage*» указываются номера строк.

Каждая строка представляет собой этап (шаг) процесса формирования кластеров. Последняя строка таблицы «График агломерации» описывает последний этап этого процесса, когда все объекты исследования объединяются в один кластер.

Число строк в таблице «График агломерации» всегда на единицу меньше числа объектов исследования. В рассматриваемом примере объектами исследования являются 11 возрастных категорий туристов, и число шагов их поэтапного объединения в один кластер составляет 10.

В столбце «*Cluster Combined*» указывается, какие именно кластеры объединяются в один на очередном этапе формирования кластеров. В столбце «*Coefficients*» указываются значения того показателя, на основании которого устанавливается очередность поэтапного объединения объектов исследования в один кластер. То, какой именно показатель используется для этих целей, зависит от выбранного метода формирования кластеров. В нашем примере был выбран метод «*Ward*».

Основной принцип метода «*Ward*» заключается в том, что в первую очередь должны объединяться те кластеры, объединение которых в наименьшей степени способствует увеличению гетерогенности (разнородности) внутри формируемых кластеров.

В столбце «*Coefficients*» указываются значения коэффициента, характеризующего степень гетерогенности (разнородности) формируемых кластеров. На начальном (нулевом) этапе формирования кластеров, когда каждый объект исследования рассматривается как кластер, все кластеры являются абсолютно гомогенными (однородными). Коэффициент, характеризующий степень их гетерогенности, равен нулю.

Гетерогенность кластеров повышается по мере их объединения в более крупные. На первом этапе при объединении кластеров «9» и «10» гетерогенность вновь созданного кластера характеризуется значением коэффициента 0,004 (см. рис. 8.10).

На последнем (десятом) этапе при объединении всех объектов исследования в один кластер гетерогенность созданного кластера характеризуется значением коэффициента 4,196.

Применение метода «*Ward*» обеспечивает минимально возможное увеличение степени гетерогенности формируемых кластеров в процессе объединения мелких кластеров в более крупные.

В столбце «*Next Stage*» указывается номер этапа формирования кластеров, когда новый кластер будет объединяться с другими.

Например, на первом этапе при объединении кластеров «9» и «10» создается новый кластер, ему присваивается номер «9». Созданный кластер «9» будет объединяться с кластером «11» на четвертом этапе формирования кластеров, о чем есть соответствующая отметка в столбце «*Next Stage*» (см. табл. 8.4).

В столбце «*Stage Cluster First Appears*» указываются этапы (строки), на которых были сформированы объединяемые кластеры. Например, при объединении кластеров «9» и «11» указывается, что кластер «9» был сформирован на первом, а кластер «11» – на нулевом этапе формирования кластеров.

Таким образом, таблица «График агломерации» достаточно подробно описывает очередность формирования кластеров, начиная с нулевой стадии, когда каждый объект исследования рассматривается как кластер, и заканчивая созданием кластера, объединяющего все объекты исследования.

8.5. ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО КОЛИЧЕСТВА ФОРМИРУЕМЫХ КЛАСТЕРОВ

Компьютерная программа *SPSS* не дает ответа на вопрос, какое число формируемых кластеров является оптимальным. Это должны решать специалисты, проводящие исследование. При решении этой задачи необходимо учитывать два аспекта:

1. В процессе формирования кластеров их число становится все меньше, а количество объектов исследования, входящих в один кластер, – все больше.
2. С увеличением числа объектов, объединяемых в один кластер, растет гетерогенность формируемого кластера.

Оптимальным является такое число кластеров, при котором сформированные кластеры:

- с одной стороны, объединяют в себе как можно больше объектов исследования;
- с другой стороны, являются возможно менее гетерогенными внутри.

Решение относительно оптимального числа формируемых кластеров принимается на основании данных таблицы «График агломерации».

Для определения оптимального числа формируемых кластеров используется критерий «*Ellbow*»: строится график зависи-

мости числа формируемых кластеров и значений коэффициента, характеризующего степень их гетерогенности (рис. 8.10).

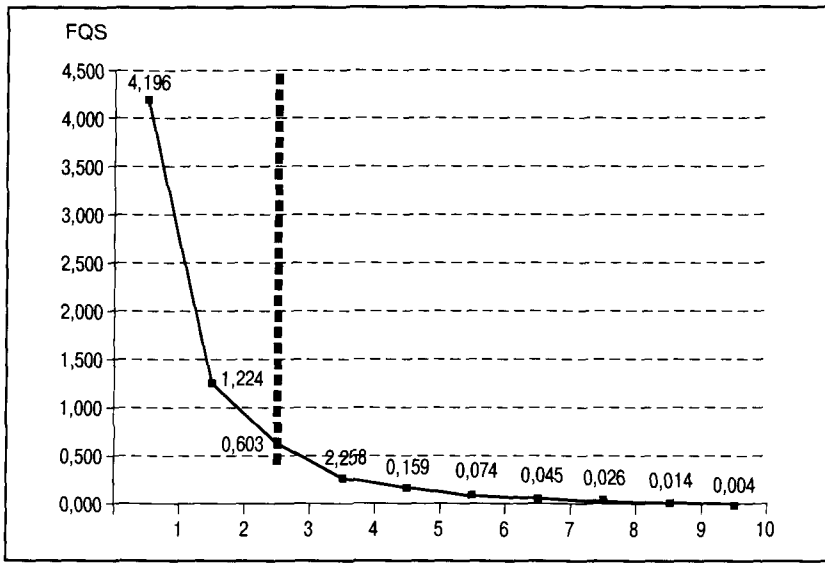


Рис. 8.10. Критерий «*Ellbow*»: определение оптимального числа кластеров

Из данных на графике, представленном на рис. 8.10, видно, что при сокращении числа кластеров с 3 до 2 происходит резкое увеличение гетерогенности кластеров (с 0,603 до 1,224). Из этого следует, что 3 является оптимальным числом кластеров, т.е. в результате проведения кластерного анализа объекты исследования должны быть объединены в три кластера. Именно такое решение обеспечит создание сравнительно однородных кластеров, объединяющих достаточно большое число объектов исследования.

8.6. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА

Результаты кластерного анализа нагляднее всего представляются в виде дендограммы (рис. 8.11).

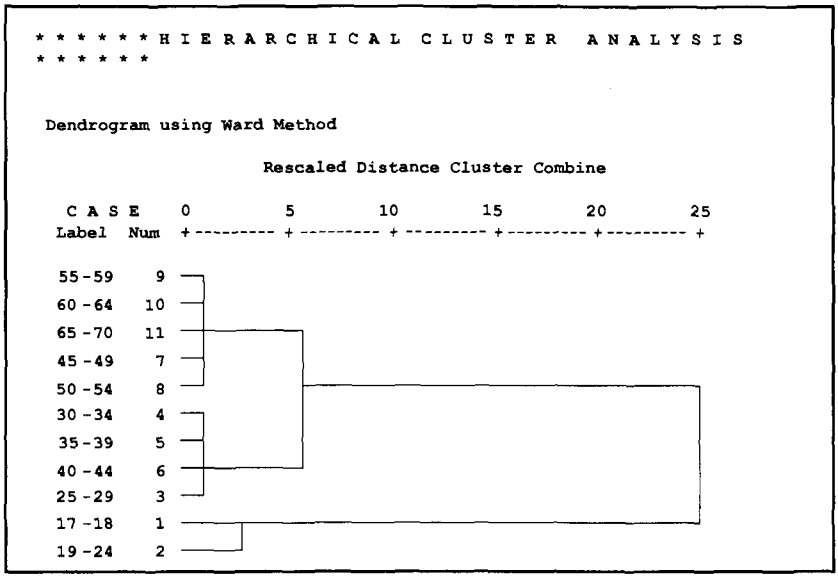


Рис. 8.11. Результат кластерного анализа: дендограмма

Дендограмма является графическим изображением таблицы «График агломерации» (см. табл. 8.4).

При построении дендограммы *SPSS* нормирует значения коэффициента, характеризующего степень гетерогенности формируемых кластеров, по шкале от нуля до 25. В рассматриваемом примере значению шкалы дендограммы 25 (см. рис. 8.11) соответствует значение коэффициента 4,196 в последней строке таблицы «График агломерации» (см. табл. 8.4).

Дендограмма иллюстрирует увеличение разнородности кластеров по мере их укрупнения. Максимальное значение шкалы дендограммы 25 характеризует максимально возможную степень гетерогенности кластеров, когда все объекты исследования объединены в один кластер.

Если объекты исследования разделить на два кластера: «17–24 года» и «25–70 лет», то данные кластеры будут значительно более разнородны. Степень их разнородности по шкале дендограммы понизится примерно до 7.

В качестве оптимального числа формируемых кластеров в рассматриваемом примере было определено число 3 (см. предыдущий раздел). Окончательным результатом кластерного анализа

является разделение 11 возрастных групп туристов на три кластера:

кластер 1: туристы 17–24 лет;

кластер 2: туристы 25–44 лет;

кластер 3: туристы 45–70 лет.

Как видно из дендограммы, кластеры «2» и «3», т.е. возрастные группы туристов «25–44 года» и «45–70 лет», являются более однородными по структуре интересов (мотивов проведения времени на отдыхе) по сравнению с возрастной группой «17–24 года» (см. рис. 8.11).

После кластерного анализа можно проводить дополнительные исследования, в ходе которых оцениваются особенности выделенных кластеров. В нашем примере можно выяснить, какие именно интересы туристов (мотивы проведения времени на отдыхе) являются наиболее важными для каждого сформированного кластера.

Также для выявления отличительных особенностей сформированных кластеров можно провести впоследствии дискриминантный анализ. С помощью дискриминантного анализа, например, можно выяснить, отличаются ли друг от друга туристы, оказавшиеся в разных кластерах, по каким-либо социально-демографическим признакам (кроме возраста, поскольку эта переменная лежит в основе формирования кластеров).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какова цель проведения и возможности использования результатов кластерного анализа?
2. Какие требования предъявляются к переменным, участвующим в проведении кластерного анализа, относительно типов шкал измерения переменных?
3. Почему и в каких случаях при проведении кластерного анализа необходимо преобразование структуры исходного массива данных?
4. Чем отличается иерархический кластерный анализ от других видов кластерного анализа?
5. В чем состоит отличие между дивизионным и агломеративным алгоритмом иерархического кластерного анализа?
6. Для чего при использовании метода формирования кластеров «*Ward*» служит показатель «Квадрат евклидова расстояния» и как следует интерпретировать его значения?
7. Что представляет собой таблица «График агломерации», выводимая в *SPSS* для результатов иерархического кластерного анализа?

8. Какие данные содержатся в столбцах «*Stage*», «*Cluster Combined*», «*Coefficients*» и «*Next Stage*» этой таблицы?
9. Какие ориентиры существуют для определения оптимального количества формируемых кластеров, что представляет собой критерий «*Ellbow*»?
10. Что представляет собой дендограмма, выводимая в *SPSS* на экран компьютера среди результатов кластерного анализа?

9. ДИСКРИМИНАНТНЫЙ АНАЛИЗ В SPSS

9.1. ПОСТАНОВКА ЦЕЛИ ИССЛЕДОВАНИЯ И ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ В SPSS

Дискриминантный анализ – анализ различий заранее заданных групп объектов исследования (потребителей, товаров, брендов и т.п.). Переменная, разделяющая совокупность объектов исследования на группы, называется *группирующей*.

С помощью дискриминантного анализа изучаются различия между двумя или более группами по определенным признакам. Признаки, используемые для выявления различий между группами, называются ***дискриминационными переменными***.

С точки зрения теории статистики группирующая переменная должна быть номинальной, т.е. измеряться по номинальной шкале, а зависимые переменные – метрическими (см. п. 2.3 «Типы шкал измерения переменных»). Соблюдение этого условия обеспечивает высокую точность статистических расчетов. Однако на практике при использовании *SPSS* допускается, что группирующая переменная может быть номинальной или порядковой, а дискриминационные переменные могут измеряться по шкале любого типа.

Результатом дискриминантного анализа является построение дискриминантной модели (дискриминантной функции), которая имеет вид

$$d = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n,$$

где d – группирующая (зависимая) переменная;
b_n – коэффициенты дискриминантной функции;
a – свободный член (константа);
x_n – дискриминационные (независимые) переменные.

С помощью этой модели, зная характеристики объекта исследования, можно с определенной степенью уверенности определить его принадлежность к одной из исследованных групп. Например, требуется построить дискриминантную модель, при помощи кото-

рой на основании социально-демографических признаков (пол, возраст, образование, доход семьи) можно было бы причислить туриста к одной из двух групп: посещающих и не посещающих дискотеки (рис. 9.1).

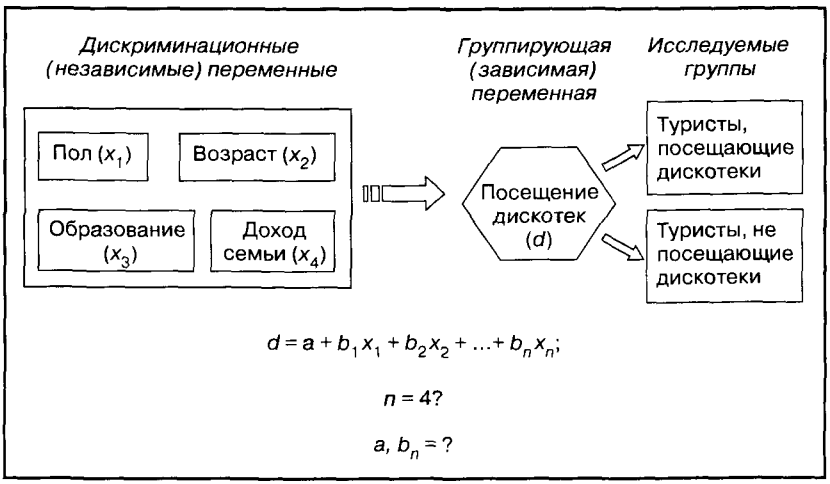


Рис. 9.1. Дискриминантный анализ: постановка цели исследования

Для того чтобы построить дискриминантную модель, следует сначала выяснить, все ли выбранные дискриминационные переменные в действительности служат отличительными признаками исследуемых групп ($n = 4$?). Только после этого можно построить дискриминантную модель (a, b_n).

В нашем примере для дискриминантного анализа используются данные, собранные в результате опроса туристов, отдыхающих в курортной зоне «Баварский лес».

Информация по группирующей переменной формируется из ответов респондентов на вопрос анкеты № 4: «Какие заведения/мероприятия Вы часто посещаете во время отдыха?». В качестве ответа на этот вопрос респондентам предлагается выбрать один или несколько вариантов из 11 предложенных ответов. В качестве ответа № 7 предлагается вариант «дискотеки».

При занесении в файл данных *SPSS* ответов на многовариантные вопросы создается несколько дихотомических переменных (см. п. 2.2 «Виды кодировки данных»). В рассматриваемом примере вопрос анкеты № 44 представлен в файле данных *SPSS* в виде семи переменных (рис. 9.2).

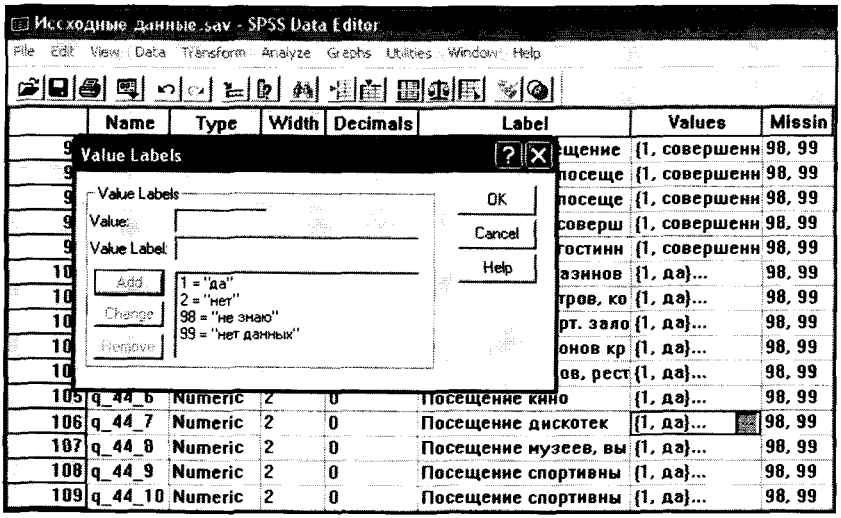


Рис. 9.2. Фрагмент вкладки «Свойства переменных» (*Data Variable*)

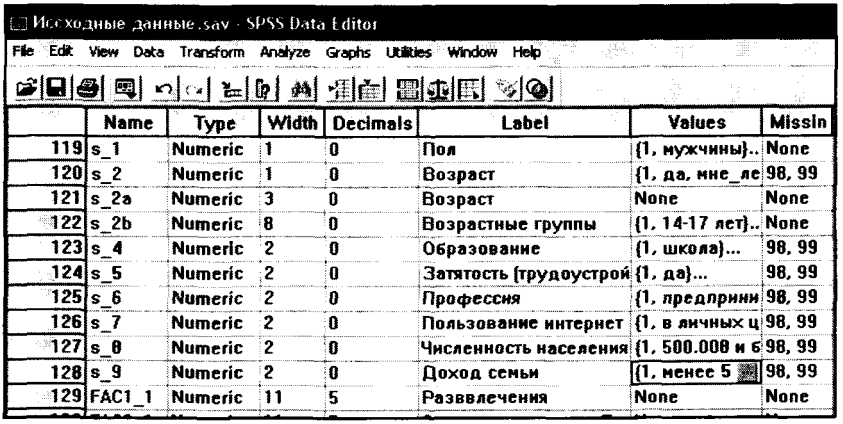
В рассматриваемом примере дискриминантного анализа в качестве группирующей переменной используется переменная с именем «q_44_7» и меткой «Посещение дискотек». Метка этой переменной имеет два значения: «1» – «да» и «2» – «нет», которые разделяют опрашиваемых туристов на две группы: посещающие и не посещающие дискотеки. Ответы респондентов, которые затруднились или не захотели отвечать на этот вопрос («98», «99»), не участвуют в исследовании, о чем есть пометка в столбце «*Missing*».

В качестве дискриминационных переменных в рассматриваемом примере используются социально-демографические признаки туристов: пол, возраст, образование и доход семьи (рис. 9.3).

Переменная с именем «s_1» и меткой «Пол» имеет всего два значения («1» – «мужчины», «2» – женщины), т.е. она является дихотомической.

Переменная с именем «s_2a» и меткой «Возраст» является метрической переменной. Ответы на соответствующий вопрос анкеты выражаются в числах, поэтому числовые коды значений метки переменной отсутствуют, о чем говорит отметка «*None*» в столбце «*Values*».

Переменная с именем «s_4» и меткой «Образование» является порядковой переменной. Значения меток этой переменной относятся к 7 категориям, соответствующим уровням иерархии системы образования в Германии.



Исходные данные.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missin
119	s_1	Numeric	1	0	Пол	{1, мужчины}..	None
120	s_2	Numeric	1	0	Возраст	{1, да, мне_ле	98, 99
121	s_2a	Numeric	3	0	Возраст	None	None
122	s_2b	Numeric	8	0	Возрастные группы	{1, 14-17 лет}..	None
123	s_4	Numeric	2	0	Образование	{1, школа}...	98, 99
124	s_5	Numeric	2	0	Затятость (трудоустрой	{1, да}...	98, 99
125	s_6	Numeric	2	0	Профессия	{1, предприни	98, 99
126	s_7	Numeric	2	0	Пользование интернет	{1, в личных ц	98, 99
127	s_8	Numeric	2	0	Численность населения	{1, 500.000 и б	98, 99
128	s_9	Numeric	2	0	Доход семьи	{1, менее 5	98, 99
129	FAC1_1	Numeric	11	5	Развлечения	None	None

Рис. 9.3. Фрагмент вкладки «Свойства переменных» (*Data Variable*)

Переменная с именем «s_9» и меткой «Доход семьи» также является порядковой. Значения этой переменной представлены 9 категориями туристов по уровню дохода семьи: «1» – «до 500 евро», «2» – «от 500 до 900 евро», «3» – «от 900 до 1250 евро», «4» – «от 1250 до 1800 евро» ... «9» – «свыше 3800 евро» в неделю.

Как отмечалось выше, не все из выбранных дискриминационных переменных в действительности могут выступать в качестве отличительных признаков исследуемых групп. Если они таковыми не являются, они должны быть исключены из дискриминантной модели. В целом при выполнении дискриминантного анализа решаются следующие задачи:

- Оценивается выбор дискриминационных переменных.
- Строится дискриминантная модель.
- Оценивается точность прогнозов на основе построенной дискриминантной модели.

9.2. КОМАНДЫ SPSS НА ВЫПОЛНЕНИЕ ДИСКРИМИНАНТНОГО АНАЛИЗА

Дискриминантный анализ, как и кластерный, относится к классификационным видам анализа. Для задания процедуры его выполнения в меню методов анализа, предлагаемых пакетом *SPSS*, следует выбрать группу методов «*Classify*», которая имеет собственное меню, включающее некоторые виды кластерного анализа и дискриминантный анализ (рис. 9.4).

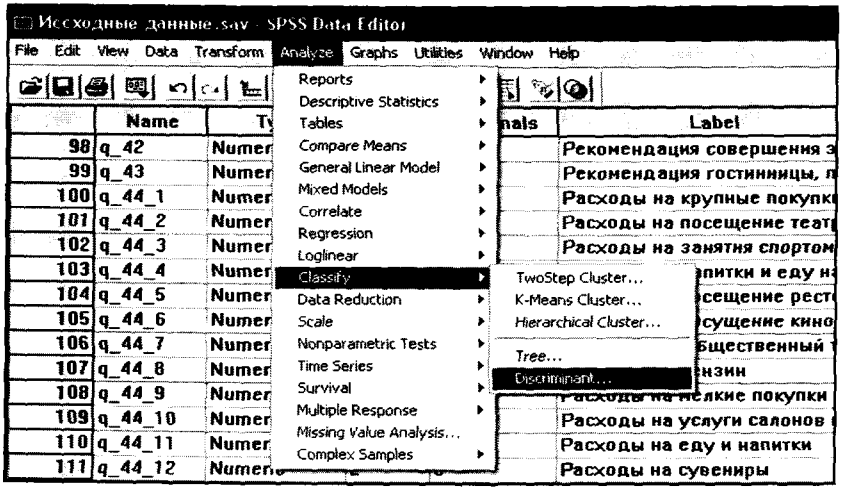


Рис. 9.4. Выбор в меню процедуры «Дискриминантный анализ»

При выборе меню «*Analyze* > *Classify* > *Discriminant*» открывается диалоговое окно «Дискриминантный анализ», в котором формируется задание на выполнение дискриминантного анализа (рис. 9.5).

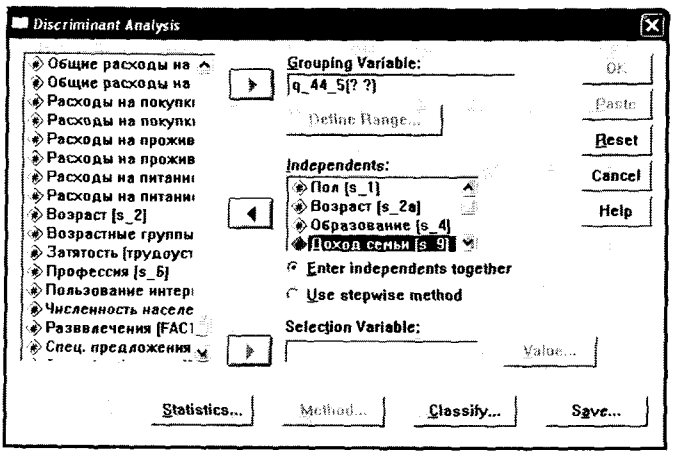


Рис. 9.5. Диалоговое окно «Дискриминантный анализ»

В левом поле диалогового окна «Дискриминантный анализ» находится список меток всех переменных, занесенных в исход-

ный файл данных. Из этого списка следует выбрать метки независимых переменных дискриминантной модели и при помощи кнопки со стрелкой поочередно перенести их в правое поле окна «*Independents*». В рассматриваемом примере это метки переменных «пол», «возраст», «образование» и «доход семьи».

Затем из списка всех меток переменных в левом поле диалогового окна «Дискриминантный анализ» следует выбрать метку группирующей переменной и при помощи кнопки со стрелкой перенести ее в правое поле «*Grouping Variable*». В рассматриваемом примере это метка переменной «Посещение дискотек».

После осуществления переноса метки группирующей переменной в поле «*Grouping Variable*» в этом поле появляется имя группирующей переменной («q_44_5») и активизируется кнопка «*Define Range*» («Определить область»). При нажатии этой кнопки открывается одноименное вспомогательное диалоговое окно (рис. 9.6).

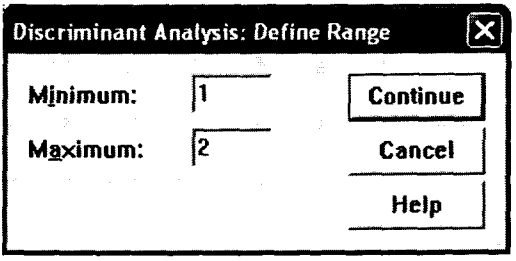


Рис. 9.6. Диалоговое окно «*Define Range*» («Определить область»)

Во вспомогательном диалоговом окне «*Define Range*» следует определить минимальное и максимальное значения числовых кодов исследуемых групп. В рассматриваемом примере исследуемых групп только две: «1» – «туристы, посещающие дискотеки» и «2» – «туристы, не посещающие дискотеки». После нажатия кнопки «*Continue*» (см. рис. 9.6) осуществляется возврат в главное диалоговое окно «Дискриминантный анализ» (см. рис. 9.5).

В главном диалоговом окне «Дискриминантный анализ» следует указать метод построения дискриминантной модели. Возможен выбор пошагового метода («*Use stepwise method*») (см. рис. 9.5), который предполагает поэтапное включение независимых переменных в дискриминантную модель. В результате применения этого метода создается несколько дискриминантных моделей по количеству независимых переменных. В рассматриваемом

примере выбран метод «*Enter independents together*» (см. рис. 9.5). Этот метод предполагает одновременное включение в дискриминантную модель всех заданных независимых переменных.

При нажатии кнопки «*Statistics*» в главном диалоговом окне «Дискриминантный анализ» открывается вспомогательное диалоговое окно «Статистические показатели» (рис. 9.7).

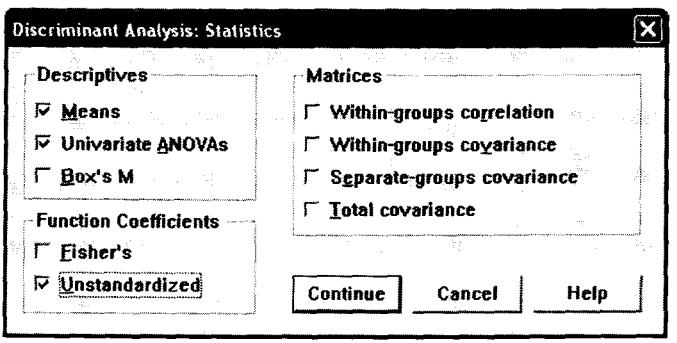


Рис. 9.7. Диалоговое окно «Статистические показатели»

В диалоговом окне «Статистические показатели» можно задать команды на расчет различных статистических показателей в процессе выполнения процедуры дискриминантного анализа. В рассматриваемом примере в поле «*Descriptives*» («Описательные статистические методы») поставлены отметки напротив команд «*Means*» и «*Univariate ANOVAs*».

В результате выполнения команды «*Means*» рассчитываются средние значения дискриминационных переменных для каждой исследуемой группы. Результаты выполнения этой команды будут представлены далее (см. табл. 9.2).

В результате выполнения команды «*Univariate ANOVAs*» («Одномерные тесты *ANOVA*») производится тест на равенство средних значений дискриминационных переменных в исследуемых группах (см. главу 5 «Сравнение средних величин в *SPSS*»). Результаты выполнения этой команды будут представлены далее (см. табл. 9.2 и 9.3).

В рассматриваемом примере в поле «*Matrices*» («Таблицы») диалогового окна «Статистические показатели» поставлена отметка напротив команды «*Within-groups correlation*» (см. рис. 9.7). В результате выполнения этой команды на экран компьютера выводится таблица «Объединенные матрицы внутри групп», содержа-

щая данные о корреляционных связях между дискриминационными переменными (см. далее табл. 9.5).

Также в рассматриваемом примере в поле «*Function Coefficients*» диалогового окна «Статистические показатели» поставлена отметка напротив команды «*Unstandardized*» (см. рис. 9.7). Это означает, что при построении дискриминантной функции будут использованы нестандартизированные коэффициенты. Значения нестандартизированных коэффициентов дискриминантной функции в рассматриваемом примере будут представлены далее (см. табл. 9.10).

При нажатии кнопки «*Continue*» в диалоговом окне «Статистические показатели» данное окно закрывается и осуществляется возврат в главное диалоговое окно «Дискриминантный анализ» (см. рис. 9.5). При нажатии кнопки «*Classify*» в главном диалоговом окне «Дискриминантный анализ» открывается вспомогательное диалоговое окно «Классификация» (рис. 9.8).

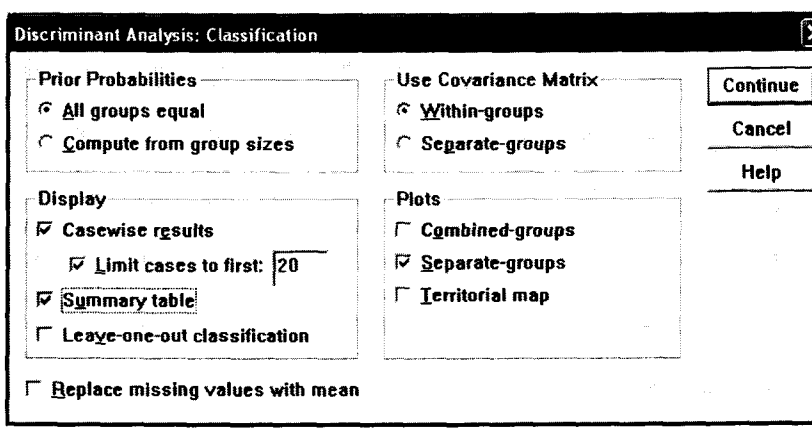


Рис. 9.8. Диалоговое окно «Классификация»

В диалоговом окне «Классификация» задаются условия и форма представления классификации объектов исследования, т.е. распределения их по группам. В рассматриваемом примере речь идет о разделении туристов на две группы: «посещающие дискотеки» и «не посещающие дискотеки».

В поле «*Plots*» («Графики») диалогового окна «Классификация» можно задать построение графиков, иллюстрирующих результаты классификации. В рассматриваемом примере поставлена отметка напротив команды «*Separate-groups*» («Разделенные группы») (см. рис. 9.8). В результате выполнения этой команды

на экран выводятся графики распределения дискриминантной функции для каждой исследуемой группы. Результаты выполнения этой команды будут представлены далее (см. табл. 9.9 и 9.10).

В поле «*Display*» диалогового окна «Классификация» задается форма представления результатов классификации. В рассматриваемом примере отмечена команда «*Casewise results*» («Результаты отдельно по каждому наблюдению»). Таким образом, на экран выводятся результаты классификации отдельно по каждому респонденту, а именно к какой группе и с какой вероятностью он может быть причислен исходя из значения дискриминантной функции.

Следует ограничить число респондентов, по которым представляются результаты классификации. Это можно сделать при помощи команды «*Limit cases to first...*» («Ограничить наблюдения по первым...»). В нашем примере задано ограничение по первым 20 респондентам (см. рис. 9.8). Результаты классификации по первым 20 респондентам будут представлены далее (см. табл. 9.12).

В поле «*Display*» диалогового окна «Классификация» также поставлена отметка напротив команды «*Summary table*» («Сводная таблица»). В результате выполнения этой команды на экран компьютера выводится сводная таблица результатов классификации (см. далее табл. 9.13).

При нажатии кнопки «*Continue*» в диалоговом окне «Классификация» данное окно закрывается и осуществляется возврат в главное диалоговое окно «Дискриминантный анализ» (см. рис. 9.5). При нажатии кнопки «*Save*» («Сохранить») в диалоговом окне «Дискриминантный анализ» открывается одноименное вспомогательное диалоговое окно, в котором можно задать команды на сохранение результатов дискриминантного анализа в виде новых переменных в исходном файле данных. В рассматриваемом примере такие операции не производятся. Запуск процедуры выполнения дискриминантного анализа осуществляется путем нажатия кнопки «*OK*» в главном диалоговом окне «Дискриминантный анализ».

9.3. ОЦЕНКА ВЫБОРА ДИСКРИМИНАЦИОННЫХ ПЕРЕМЕННЫХ

Оценка выбора дискриминационных переменных представляет собой первый этап интерпретации результатов дискриминантного анализа. Представление результатов анализа начина-

ется с обзора действительных и пропущенных значений, который выводится на экран компьютера в виде таблицы «Анализ обработанных наблюдений» (табл. 9.1).

Таблица 9.1

Анализ обработанных наблюдений

Analysis Case Processing Summary

		№	Persent
Действительные		1023	16,0
Исключенные:	Отсутствующие или находящиеся вне области кодировки данные о принадлежности к группе	4711	73,7
	По меньшей мере одна отсутствующая дискриминационная переменная	146	2,3
	Оба случая отсутствия данных (отсутствующие или находящиеся вне области кодировки данные о принадлежности к группе и по меньшей мере одна отсутствующая дискриминационная переменная)	516	8,1
	Всего	5373	84,0
Всего		6396	100,0

В нашем примере число респондентов, принявших участие в опросе (*Total*), составляет 6396; из этих данных только 1023 анкеты являются действительными (*Valid*), т.е. только эти наблюдения используются при расчетах для построения дискриминантной функции. Данные по остальным респондентам исключены из анализа (*Excluded*) ввиду отсутствия данных по ответам на нужные вопросы.

Число респондентов, не давших информации о том, посещают ли они дискотеки, составляет 4711. Как отмечалось ранее при построении дискриминантной функции, данные по этим респондентам не используются. Однако эти респонденты участвуют в классификации на основании построенной дискриминантной модели (см. далее табл. 9.12 и 9.13).

Число респондентов, не давших информацию о себе хотя бы по одному из нужных социально-демографических признаков, составляет 146. Число респондентов, не давших информации о том, посещают ли они дискотеки, и одновременно не давших информацию о себе хотя бы по одному из нужных социально-демографических признаков, составляет 516.

После обзора действительных и пропущенных значений на экран компьютера выводится таблица «Статистические показатели в группах», которая содержит данные о средних значениях

(*Mean*) дискриминационных переменных в каждой из исследуемых групп. Эти показатели дают общее представление о том, являются ли дискриминационные переменные отличительными признаками исследуемых групп (табл. 9.2).

Таблица 9.2

Статистические показатели в группах

Group Statistics

Посещение дискотек		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Да	Пол	1,42	,496	88	88,000
	Возраст	35,81	12,985	88	88,000
	Образование	6,97	13,026	88	88,000
	Доход семьи	5,99	2,307	88	88,000
Нет	Пол	1,46	,498	935	935,000
	Возраст	51,31	12,920	935	935,000
	Образование	5,56	10,346	935	935,000
	Доход семьи	6,43	1,854	935	935,000
Всего	Пол	1,45	,498	1023	1023,000
	Возраст	49,98	13,632	1023	1023,000
	Образование	5,68	10,603	1023	1023,000
	Доход семьи	6,39	1,899	1023	1023,000

N – число действительных наблюдений.

Из данных, представленных в табл. 9.2, видно, что средний возраст туристов, посещающих дискотеки, составляет около 36 лет, а средний возраст туристов, не посещающих дискотеки, составляет примерно 51 год. Средний возраст всех опрошенных респондентов составляет примерно 50 лет.

Переменная «возраст» является метрической. С точки зрения статистики только метрические переменные могут участвовать в дискриминантном анализа в качестве независимых переменных, поскольку только для них можно рассчитать среднее значение и стандартное отклонение (см. п. 2.3 «Типы шкал измерения переменных»).

Переменная «пол» является дихотомической. Расчет такого показателя, как «средний пол», является абсурдным. Однако дихотомические переменные могут рассматриваться как метрические. Если бы метки переменной «пол» имели числовые коды не

«1»(мужчины)/«2»(женщины), а «0»(мужчины)/«1»(женщины), то средние значения этой переменной для исследуемых групп были бы не 1,42/1,46, а 0,42/0,46. Это означает, что доля женщин среди туристов, посещающих дискотеки, составляет 42%, а среди туристов, не посещающих дискотеки, – 46%.

Переменные «образование» и «доход семьи» являются порядковыми, т.е. они разделяют туристов на категории по уровню образования и дохода семьи. Средние значения этих переменных не имеют никакого смысла, поскольку представляют лишь средние значения порядковых номеров категорий, указанных респондентами.

Из данных, представленных в табл. 9.2, можно сделать вывод, что уровень образования туристов, посещающих дискотеки, несколько выше уровня образования туристов, не посещающих дискотеки (6,97 > 5,56).

Что касается уровня дохода семьи, то можно сказать, что он несколько ниже у туристов, посещающих дискотеки, по сравнению с туристами, не посещающими дискотеки (5,99 < 6,43).

Неравенство средних значения заявленных дискриминационных переменных (пол, возраст, образование, доход семьи) в группах туристов, посещающих и не посещающих дискотеки, не доказывает, что данные переменные являются отличительными признаками исследуемых групп. Их можно считать отличительными признаками только в том случае, если будет доказана статистическая значимость различий их средних значений в исследуемых группах (см. главу 5 «Сравнение средних величин в *SPSS*»). Для этого проводится тест на равенство средних значений в группах (табл. 9.3).

Таблица 9.3

Тест на равенство средних значений в группах

Tests of Equality of Group Means

	Wills' Lambda*	F	df1	df2	Sig.
Пол	1,000	,451	1	1021	,502
Возраст	,898	115,751	1	1021	,000
Образование	,999	1,412	1	1021	,235
Доход семьи	,996	4,313	1	1021	,038

Для проведения теста на равенство средних значений в группах в качестве тестовой величины используется лямбда Уилкса

(*Wilks' Lambda*)[1]. Основной результат теста определяется с помощью величины «*Significance*» («Значимость»). Если значение «*Significance*» меньше 0,05, это означает, что различия между средними значениями дискриминационных переменных в исследуемых группах являются статистически значимыми (см. главу 5 «Сравнение средних величин в *SPSS*»).

В рассматриваемом примере значение «*Significance*» не превышает 0,05 только для двух заявленных дискриминационных переменных: «возраст» (0,000) и «доход семьи» (0,038). Это означает, что туристы, посещающие и не посещающие дискотеки, отличаются по возрасту и доходу семьи, поэтому переменные «возраст» и «доход семьи» могут выступать в качестве дискриминационных переменных.

Значение величины «*Significance*» для переменных «пол» (0,502) и «образование» (0,235) (см. табл. 9.3) превышает 0,05. Это означает, что между группами туристов, посещающих и не посещающих дискотеки, не существует достаточно четкого различия по половой структуре и уровню образования туристов.

Переменные «пол» и «образование» не обладают дискриминирующими (разделительными) свойствами и не могут выступать в качестве дискриминационных переменных. Они должны быть исключены из дискриминантной модели.

В случае необходимости изменения состава дискриминационных переменных следует заново сформировать задание на проведение дискриминантного анализа. Для этого следует повторить все операции, описанные в п. 9.2 «Команды *SPSS* на выполнение дискриминантного анализа», изменив лишь список независимых

[1] **Лямбда Уилкса (*Wilks' Lambda*)** – это критерий, используемый при проведении теста на предмет того, значимо ли различаются между собой средние значения дискриминантной функции в исследуемых группах.

Построенная дискриминантная модель (дискриминантная функция) должна отражать четкое разделение исследуемых групп. Для оценки четкости этого разделения проводится тест с помощью критерия Лямбда Уилкса. В качестве исходной гипотезы выступает утверждение: «Средние значения дискриминантной функции в исследуемых группах равны». Верность исходной гипотезы определяется значением показателя «*Significance*». Если значение «*Significance*» не превышает 0,05, это означает, что вероятность ошибки при отклонении нулевой гипотезы составляет менее 5% (т.е. ниже допустимого уровня при доверительном интервале 95%). В этом случае исходная гипотеза может быть отклонена, т.е. она неверна. Значение «*Significance*» менее 0,05 доказывает ошибочность исходной гипотезы и статистическую значимость различия средних значений дискриминантной функции в исследуемых группах.

переменных (*Independents*) в диалоговом окне «Дискриминантный анализ» (см. п. 9.2, рис. 9.5).

После изменения задания на выполнение дискриминантного анализа изменяются результаты. В рассматриваемом примере после исключения переменных «пол» и «образование» из состава дискриминационных переменных число действительных значений осталось без изменений – 1023 (см. табл. 9.1). Значения величины «*Significance*» при проведении теста на равенство средних значений в группах для переменных «возраст» и «доход семьи» также остались без изменений (табл. 9.3 и 9.4).

Таблица 9.4

Тест на равенство средних значений в группах

Tests of Equality of Group Means

	Wills' Lambda	F	df1	df2	Sig.
Возраст	,898	115,751	1	1021	,000
Доход семьи	,996	4,313	1	1021	,038

После того как доказано наличие дикриминирующих (разделительных) особенностей переменных «возраст» и «доход семьи», следует также доказать, что они являются действительно независимыми. Значения этих переменных не должны обусловливать друг друга. Следует доказать, что доход семьи не находится в прямой зависимости от возраста туриста. Для этого рассчитывается коэффициент корреляции, характеризующий связь между исследуемыми переменными (табл. 9.5).

Таблица 9.5

Объединенные матрицы внутри групп

Pooled Within-Groups Matrices

		Возраст	Доход семьи
Correlation	Возраст	1,000	,076
	Доход семьи	,076	1,000

В табл. 9.5 представлено осредненное значение коэффициента корреляции между независимыми переменными дискриминантной функции для обеих исследуемых групп. По данным этой таблицы коэффициент корреляции между переменными «возраст» и «доход семьи» составляет всего 0,076 (<<0,5). Это доказывает отсутствие корреляционной связи между этими переменными.

9.4. ПОСТРОЕНИЕ ДИСКРИМИНАНТНОЙ МОДЕЛИ

Построение дискриминантной модели заключается в расчете и анализе коэффициентов дискриминантной функции. Построенная дискриминантная модель должна максимально четко разделять исследуемые группы. Качество построенной дискриминантной модели в рассматриваемом примере характеризуется данными, представленными в табл. 9.5 и 9.6.

Таблица 9.6

Собственные значения

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,115[a]	100,0	100,0	,321

[a] В этом анализе используются первые канонические дискриминантные функции.

Значение коэффициента корреляции между рассчитанными значениями дискриминантной функции и реальной принадлежностью к группе «0,321» является неудовлетворительным. В табл. 9.6 также представлен такой показатель, как собственное значение дискриминантной функции (*Eigenvalue*). Высокое значение этого показателя свидетельствует о высокой точности построенной дискриминантной модели. В рассматриваемом примере этот показатель имеет весьма низкое значение 0,115, что является негативным фактором.

Показатель «Лямбда Уилкса» используется для проведения теста на значимость различий средних значений дискриминантной функции в исследуемых группах. В нашем примере значение показателя «*Significance*» составляет 0,000 (табл. 9.7), что свидетельствует о высокой значимости различий средних значений (см. главу 5 «Сравнение средних величин в *SPSS*»).

Таблица 9.7

Лямбда Уилкса

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	χ-квадрат	df	Sig.
1	,897	110,971	2	,000

В табл. 9.8 и 9.9 представлены коэффициенты, значения которых были рассчитаны в обеих группах по отдельности и затем усреднены.

Таблица 9.8

Стандартизированные канонические коэффициенты дискриминантной функции

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Возраст	,984
Доход семьи	,117

При помощи стандартизированных коэффициентов дискриминантной функции, представленных в табл. 9.8, можно оценить относительный вклад каждой дискриминационной переменной в различие двух исследуемых групп. В рассматриваемом примере возраст респондентов в 8,4 (0,984/0,117) раза больше влияет на желание туристов посещать дискотеки, чем доход их семьи.

Таблица 9.9

Структурная матрица

Structure Matrix*

	Function
	1
Возраст	,993
Доход семьи	,192

* Объединенные корреляции внутри групп между дискриминантными переменными и стандартизированными каноническими дискриминантными функциями. Переменные, расположенные в соответствии с абсолютными корреляционными величинами внутри функции.

Корреляционные коэффициенты, представленные в табл. 9.9, позволяют оценить, насколько сильна связь дискриминационных переменных со стандартизированными значениями дискриминантной функции.

В табл. 9.10 представлены нестандартизированные (канонические) коэффициенты дискриминантной функции, именно они используются для построения дискриминантной модели.

Таблица 9.10

Канонические коэффициенты дискриминантной функции

Canonical Discriminant Function Coefficients*

	Function
	1
Возраст	,076
Доход семьи	,062
(Constant)	−4,200

* Нестандартизированные коэффициенты.

В соответствии с данными, представленными в табл. 9.10, дискриминантная модель, построенная в результате проведения дискриминантного анализа, имеет следующий вид:

$$d = -4{,}2 - 0{,}076x_1 - 0{,}062x_2,$$

где x_1 – возраст;

x_2 – доход семьи.

Как отмечалось ранее, построенная дискриминантная модель должна как можно более четко разделять исследуемые группы. Четкость разделения исследуемых групп характеризуется расстоянием между средними значениями дискриминантной функции в исследуемых группах (табл. 9.11).

Таблица 9.11

Функции групповых центройдов

Structure Matrix*

	Function
Посещение дискотек	1
Да	−1,104
Нет	,104

* Нестандартизированные канонические дискриминантные функции, которые оцениваются по групповым средним значениям.

Как видно из данных, представленных в табл. 9.11, средние значение дискриминантной функции для группы туристов, посещающих дискотеки, составляет −1,104, а среднее значение дискриминантной функции для группы туристов, не посещающих дискотеки, составляет 0,104. Чем больше расстояние между средними значениями дискриминантной функции в исследуемых

группах, тем более четко прослеживается различие между исследуемыми группами.

Четкость различия между исследуемыми группами зависит также от рассеяния значений дискриминантной функции в исследуемых группах. Это рассеяние показано на графиках распределения значений дискриминантной функции в исследуемых группах (рис. 9.9 и 9.10).

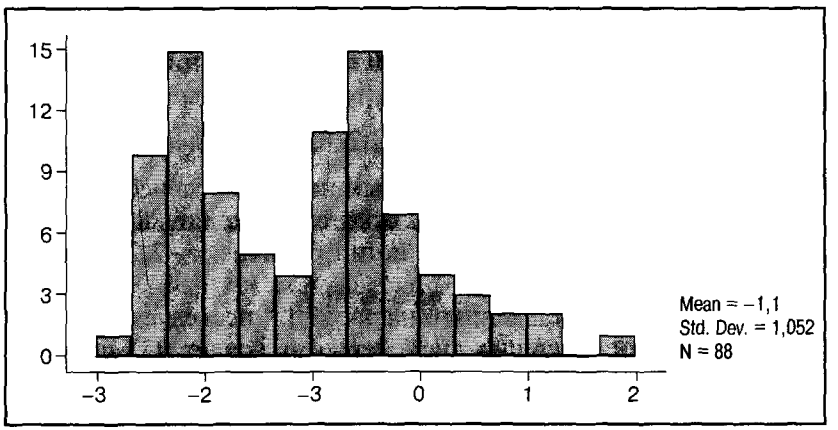


Рис. 9.9. Распределение значений дискриминантной функции для группы «туристы, посещающие дискотеки»

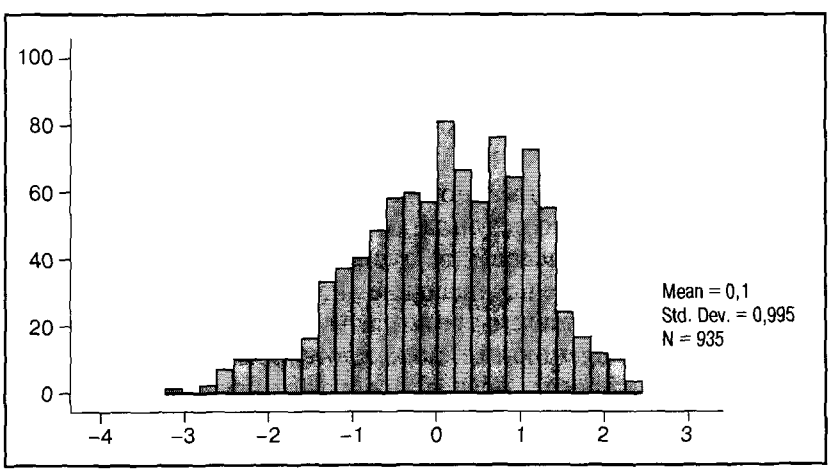


Рис. 9.10. Распределение значений дискриминантной функции для группы «туристы, не посещающие дискотеки»

Чем больше рассеяние значений дискриминантной функции в исследуемых группах, тем шире область их пересечения и слабее четкость различия между исследуемыми группами. Следовательно, чем больше такое рассеяние, тем сложнее однозначно определить принадлежность респондента к одной из исследуемых групп.

На основе построенной нами дискриминантной модели, можно сделать прогнозы посещения дискотек определенным туристом исходя из его возраста и уровня дохода семьи. Например, для туриста в возрасте 20 лет, принадлежащего по уровню дохода семьи к категории «7» (2800 – 3300 евро), значение дискриминантной функции составит

$$d = -4{,}2 - 0{,}076 \cdot 20 - 0{,}062 \cdot 7 = -2{,}246.$$

Согласно данным, представленным на рис. 9.9, в исследуемую группу «туристы, посещающие дискотеки» входят 88 туристов. Значение дискриминантной функции близкое к –2,246 имеют 15 человек.

По данным, представленным на рис. 9.10, исследуемая группа «туристы, не посещающие дискотеки» включает 935 человек. Значение дискриминантной функции, близкое к –2,246, имеют примерно 10 человек. На основании вышеизложенного можно сделать вывод, что турист в возрасте 20 лет, принадлежащий по уровню дохода семьи к категории «7» (2800–3300 евро), скорее всего, будет посещать дискотеки.

9.5. ОПРЕДЕЛЕНИЕ ТОЧНОСТИ ПРОГНОЗОВ НА ОСНОВЕ ПОСТРОЕННОЙ ДИСКРИМИНАНТНОЙ МОДЕЛИ

Точность прогнозов на основе построенной дискриминантной модели оценивается по результатам классификации, т.е. распределения объектов исследования (туристов) по исследуемым группам (посещающие и не посещающие дискотеки).

В табл. 9.12 представлены результаты классификации отдельно по каждому наблюдению, т.е. по каждому респонденту, принявшему участие в опросе и предоставившему информацию о своем возрасте и доходе семьи. Поскольку число респондентов слишком велико, в табл. 9.12 представлены только 20 наблюдений, первых по списку, – как было указано при формировании задания на проведение дискриминантного анализа (см. рис. 9.8).

В столбце «*Actual Group*» (см. табл. 9.12) указывается фактическая принадлежность респондента к одной из исследуемых групп. Так, первый по списку респондент не посещает дискотеки («2»). Остальные респонденты не ответили на вопрос, посещают ли они дискотеки, поэтому в столбце «*Actual Group*» стоит отметка «*ungrouped*» («Несгруппированное наблюдение»). Такое большое число несгруппированных наблюдений не должно удивлять. Из 6396 респондентов, принявших участие в опросе, 4717 туристов, указав свой возраст и доход семьи, не дали информации о том, посещают ли они дискотеки.

Таблица 9.12

Статистические показатели для отдельных наблюдений

Casewise Statistics

	Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df						
Original	1	2	1**	,029	1	,972	4,739	2	,028	11,800	−3,327
	2	ungrouped	2	,076	1	,954	3,141	1	,046	9,183	1,881
	3	ungrouped	2	,025	1	,974	5,052	1	,026	12,290	2,356
	4	ungrouped	2	,183	1	,922	1,774	1	,078	6,708	1,440
	6	ungrouped	2	,120	1	,940	2,414	1	,060	7,907	1,662
	8	ungrouped	2	,096	1	,947	2,765	1	,053	8,531	1,771
	9	ungrouped	2	,202	1	,917	1,627	1	,083	6,418	1,384
	10	ungrouped	2	,137	1	,935	2,211	1	,065	7,534	1,595
	11	ungrouped	2	,254	1	,903	1,303	1	,097	5,757	1,250
	12	ungrouped	2	,206	1	,915	1,600	1	,085	6,365	1,373
	13	ungrouped	2	,153	1	,930	2,046	1	,070	7,227	1,539
	15	ungrouped	2	,150	1	,931	2,076	1	,069	7,284	1,549
	16	ungrouped	2	,108	1	,943	2,580	1	,057	8,204	1,714
	18	ungrouped	2	,235	1	,908	1,410	1	,092	5,980	1,296
	19	ungrouped	2	,297	1	,891	1,088	1	,109	5,295	1,151
	20	ungrouped	2	,113	1	,942	2,513	1	,058	8,084	1,693
	21	ungrouped	2	,124	1	,939	2,371	1	,061	7,829	1,648
	22	ungrouped	2	,026	1	,973	4,958	1	,027	12,143	2,335
	23	ungrouped	2	,090	1	,949	2,873	1	,051	8,720	1,803
	24	ungrouped	2	,066	1	,957	3,383	1	,043	9,594	1,948

** Классификация не совпадает с фактической.

В столбце «*Predicted Group*» указывается прогнозируемая принадлежность респондента к одной из исследуемых групп, определяемая на основе построенной дискриминантной модели. Если прогнозируемая принадлежность к группе не совпадает с фактической, ее значение отмечается двумя звездочками (**).

В столбце «$P\ (G = g \mid D = d)$» табл. 9.12 указывается вероятность, с которой конкретный респондент может быть причислен к прогнозируемой группе. В столбце «*Discriminant Scores*» указывается значение дискриминантной функции.

Например, значение дискриминантной функции для первого респондента составляет –3,327. Согласно построенной дискриминантной модели этот респондент с вероятностью 97,2% может быть причислен к группе туристов, посещающих дискотеки, в действительности же он не посещает дискотеки.

К сожалению, из-за большого числа несгруппированных наблюдений табл. 9.12 не показывает, сколько представленных результатов классификации из 20 являются ошибочными. В результате по данным этой таблицы нельзя составить даже приблизительного представления о точности прогнозов на основе построенной дискриминантной модели.

Точность прогнозов на основе построенной дискриминантной модели определяется из данных сводной таблицы результатов классификации, т.е. причисления объектов исследования к одной из исследуемых групп (табл. 9.13).

Таблица 9.13

Результаты классификации

Classification Results[a]					
			Predicted Group Membership		
		Посещение дискотек	Да	Нет	Всего
Original	Count	Да	62	26	88
		Нет	249	686	935
		Не участвующие в группах	1280	3437	4717
	%	Да	70,5	29,5	100,0
		Нет	26,6	73,4	100,0
		Не участвующие в группах	27,1	72,9	100,0

[a] 73% первоначально сгруппированных наблюдений были классифицированы корректно.

Из данных табл. 9.13 «Результаты классификации» видно, что исследуемая группа туристов, посещающих дискотеки, состоит

фактически их 88 человек. Согласно построенной дискриминантной модели 62 туриста из 88 были корректно причислены к этой группе, а 26 – по ошибке причислены к группе туристов, не посещающих дискотеки. Итак, корректные результаты классификации составили 70,5%, а ошибочные – 29,5%.

По данным этой же таблицы исследуемая группа туристов, не посещающих дискотеки, состоит фактически их 935 человек. Согласно построенной дискриминантной модели 686 туристов из 935 были корректно причислены к этой группе, а 249 – по ошибке причислены к группе туристов, посещающих дискотеки. Итого корректные результаты классификации составили 73,4%, а ошибочные – 26,6%.

В целом корректные результаты классификации составили 73,1%, т.е. в 73,1% случаев фактическая принадлежность туриста к группе посещающих или не посещающих дискотеки совпадает с прогнозируемой, определенной на основе построенной дискриминантной модели. Это дает возможность сделать вывод, что точность прогнозов, сделанных на основе построенной дискриминантной модели составляет примерно 73%.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите цель проведения и возможности использования результатов дискриминантного анализа.
2. Как выглядит математическое описание дискриминантной модели?
3. Какие требования предъявляются к переменным, участвующим в дискриминантном анализе, относительно типов шкал измерения переменных?
4. Какие задачи решаются в ходе проведения дискриминантного анализа?
5. Каким образом и с какой целью выявляется наличие дискриминирующих свойств у переменных, выбранных в качестве независимых (дискриминационных) переменных дискриминантной модели?
6. Как можно интерпретировать результаты теста на равенство средних величин в группах, проводимого в ходе процедуры дискриминантного анализа, если значение «*Significance*» («Значимость») для определенной дискриминационной переменной составляет 0,637?
7. Что характеризует и с какой целью рассчитывается коэффициент корреляции между дискриминационными переменными? Как можно интерпретировать результаты таких расчетов, если значение коэффициента корреляции между двумя дискриминирующими переменными составляет 0,52?

8. Что характеризует и для чего рассчитывается коэффициент корреляции между расчетными значениями дискриминантной функции и реальной принадлежностью респондента к определенной группе? Как можно интерпретировать результаты, если значение этого коэффициента составляет 0,485?
9. Для чего в ходе проведения дискриминантного анализа рассчитывается показатель Лямбда Уилкса, как следует интерпретировать результаты, если значение величины «*Significance*» («Значимость») при расчете этого показателя составляет 0,02?
10. Для чего служат стандартизированные и нестандартизированные коэффициенты дискриминантной функции? Как следует интерпретировать результаты, если значения стандартизированных коэффициентов составляют: для дискриминационной переменной «1» – 0,692; для дискриминационной переменной «2» – 0,346?
11. Что характеризует расстояние между средними значениями и распределение дискриминантной функции в исследуемых группах?
12. Что представляет собой сводная таблица результатов классификации, выводимая в *SPSS* на экран компьютера среди результатов дискриминантного анализа, какие выводы можно сделать на основе данных этой таблицы?

ЗАКЛЮЧЕНИЕ

В странах с развитой рыночной экономикой, в частности в Германии, где темпы роста рынков отдельных товаров и услуг незначительны или рост вообще отсутствует, конкурирующие фирмы работают в условиях жесткой борьбы за потребителя. Для достижения успеха в этой борьбе активно используется весь спектр современных технологий маркетинговых исследований, в том числе предполагающих обработку информации при помощи статистических методов анализа.

Современные маркетологи должны владеть знаниями в области статистического анализа и навыками работы с соответствующими программными продуктами, наиболее популярным из которых в Европе является программа *SPSS*.

В Германии такие знания являются обязательными для студентов, обучающихся по специальности «Маркетинг». Все дипломные проекты и диссертационные исследования, которые защищаются по этой специальности, как правило, предполагают анализ информации с использованием программного пакета *SPSS*.

Пособие предназначено для российских студентов, обучающихся по специальности «Маркетинг», а также для всех желающих самостоятельно освоить основы статистического анализа в маркетинговых исследованиях.

Данная работа представляет лишь небольшую часть возможностей, предоставляемых пользователям *SPSS*. В частности, были рассмотрены не все методы статистического анализа, часто применяемые на практике, такие, как, например, многофакторный и многомерный дисперсионный анализ, логистическая регрессия и др. Для ознакомления с этим материалом рекомендуется использовать источники, указанные в списке литературы.

СПИСОК ЛИТЕРАТУРЫ

1. *Бююль Ахим, Цефель Петер. SPSS*: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. СПб.: ДиаСофтЮП, 2002.
2. *Гмурман В.Е.* Теория вероятностей и математическая статистика: Учеб. пособие для вузов. 9-е изд. М.: Высшая школа, 2003.
3. *Крыштановский А.О.* Анализ социологических данных с помощью пакета SPSS. М.: Изд. дом ГУ ВШЭ, 2007.
4. *Плис А.И., Сливина Н.А.* Практикум по прикладной статистике в среде *SPSS*: Учеб. пособие: В 2 ч. Ч. 1: Классические процедуры статистики. М.: Финансы и статистика, 2004.
5. *Таганов Д.Н. SPSS*: Статистический анализ в маркетинговых исследованиях. СПб.: Питер, 2005.
6. *Янкевич В.С., Безрукова Н.Л.* Маркетинг в гостиничной индустрии и туризме: российский и международный опыт / Под ред. В.С. Янкевича. М.: Финансы и статистика, 2003.
7. *Backhaus K., Erichson B., Plinke W., Weiber R.* Multivariate analysemethoden. Eine anwendungsorientierte Einführung. 9 Auflage. Berlin: Springer, 2000.
8. *Brosius F. SPSS* 11. Fundierte Einführung in *SPSS* und Statistik. Bonn: mitp-Verlag, 2002.
9. *Bühl A., Zöfel P. SPSS* 11. Einführung in die moderne Datenanalyse unter Windows. München: Pearson Studium, 2002.
10. *Jassen J., Laatz W.* Statistische Datenanalyse mit *SPSS* für Windows. Eine anwendungsorientierte Einführung in das Basissystem und das Modul exakte Tests. Berlin: Springer, 2003.
11. *Schmalen H.* Grundlagen und Probleme der Betribswirtschaf. 12 Auflage. Stuttgard: Schäffer-Poeschel Verlag, 2002.
12. *Wittenberg R.* Datenanalyse mit *SPSS* für Windows. Stuttgart: Lucius&Lucius, 2003.

ОГЛАВЛЕНИЕ

По вопросам приобретения книг обращайтесь:
Отдел продаж «ИНФРА-М» (оптовая продажа):
127282, Москва, ул. Полярная, д. 31В, стр.1
Тел. (495) 280-15-96; факс (495) 280-36-29
E-mail: books@infra-m.ru

•

Отдел «Книга–почтой»:
тел. (495) 280-15-96 (доб. 246)

Учебное издание

Гертруда Моосмюллер
Наталья Николаевна Ребик

МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ С *SPSS*

2-е издание

УЧЕБНОЕ ПОСОБИЕ

Редактор *Т.Г. Беляева*
Корректор *М.В. Литвинова*
Компьютерная верстка *Г.А. Волковой*

Подписано в печать 25.01.2015.
Формат 60×90/16. Бумага офсетная. Гарнитура Newton.
Печать офсетная. Усл. печ. л. 12,5. Уч.-изд. л. 11,68.
Доп. тираж 200 экз. Заказ № 0055

ТК 77750-460792-251110

ООО «Научно-издательский центр ИНФРА-М»
127282, Москва, ул. Полярная, д. 31В, стр. 1
Тел.: (495) 280-15-96, 280-33-86. Факс: (495) 280-36-29
E-mail: books@infra-m.ru http://www.infra-m.ru

Отпечатано в типографии ООО «Научно-издательский центр ИНФРА-М»
127282, Москва, ул. Полярная, д. 31В, стр. 1
Тел.: (495) 280-15-96, 280-33-86. Факс: (495) 280-36-29