МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Маркетинг и менеджмент»

МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ

Методические рекомендации к лабораторным работам для студентов специальности 1-28 01 02 «Электронный маркетинг» очной и заочной форм обучения

Могилев 2021

УДК 339.138
ББК 65.291.3
М27

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Маркетинг и менеджмент» «5» июня 2021 г., протокол № 12

Составитель канд. экон. наук, доц. А. В. Александров

Рецензент канд. экон. наук, доц. Т. В. Романькова

Методические рекомендации содержат задания для проведения лабораторных занятий по дисциплине «Маркетинговые исследования» для студентов дневной и заочной форм обучения специальности 1-28 01 02 «Электронный маркетинг».

Учебно-методическое издание

МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ

Ответственный за выпуск	А. В. Александров
Корректор	И. В. Голубцова
Компьютерная верстка	Е. В. Ковалевская

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 56 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, г. Могилев.

Содержание

Введение

Общие сведения о программном комплексе SPSS.

Статистический анализ данных, полученных в ходе маркетингового исследования, позволяет вскрыть такие закономерности и внутренние связи, которые невозможно выявить другими средствами. Подтверждение гипотез о наличии связи между переменными, оценка характера данных связей, оценка влияния частных параметров продукта на общее впечатление от него потребителей, сегментирование потребителей, прогнозирование изменений рыночной конъюнктуры – это лишь некоторые задачи, с успехом решаемые с применением статистических методов анализа.

Статистический анализ целесообразно выполнять с использованием компьютерной техники. Многие статистические расчеты можно провести в широко распространенной программе MS Excel. На новый уровень статистический анализ выводит применение специализированного программного обеспечения. Наиболее популярным в настоящее время является статистический программный комплекс SPSS.

SPSS Statistics[1] – это компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках. Этот статистический пакет развивается с 1968 г. и может быть полезен маркетологам, занимающимся как фундаментальными, так и прикладными маркетинговыми исследованиями.

Интерфейс статистического пакета напоминает программу MS Excel, данные вводятся в табличной форме. Некоторые возможности SPSS:

– ввод и хранение данных – формирование баз (файлов) данных SPSS;
– использование переменных разных типов;
– анализ данных маркетинговых исследований;
– статистическая обработка данных;
– создание иллюстраций и отчетов.

Общие требования к выполнению и защите лабораторных работ.

Выполнение лабораторных работ по дисциплине «Маркетинговые исследования» предполагает самостоятельную проработку студентом рекомендуемой литературы, использование теоретических знаний для решения практических задач под руководством преподавателя, подготовку к сдаче экзамена по дисциплине «Маркетинговые исследования».

Для выполнения каждой лабораторной работы (кроме лабораторной работы № 1) студент получает от преподавателя *файл исходных данных* (закодированные результаты опроса) и *перечень анализируемых переменных* (из числа переменных, представленных в этом файле). При выполнении лабораторной работы № 1 студент получает исходные данные в виде необработанных (незакодированных) результатов опроса.

В качестве исходных данных студент может использовать первичную

[1] Аббревиатура английских слов «Statistical Package for the Social Sciences» – «Статистический Пакет для Социальных Наук».

информацию, собранную в ходе опроса в рамках собственного маркетингового исследования, выполняемого в процессе написания курсовой работы по данной дисциплине. В этом случае перечень анализируемых переменных подлежит согласованию с преподавателем.

Перечень используемого оборудования для проведения лабораторных работ включает персональный компьютер с установленными программами IBM SPSS Statistics и MS Word.

Результаты выполнения лабораторной работы должны быть отражены в рабочих окнах программы IBM SPSS Statistics.

Отчет по лабораторной работе представляется в электронной форме в виде документа MS Word и должен содержать:

– титульный лист с указанием наименований университета и кафедры, названий учебной дисциплины и лабораторной работы, группы, фамилии, имени и отчества студента, выполнившего лабораторную работу;

– цель и задачи лабораторной работы;

– перечень использованного оборудования и программного обеспечения, исходные данные (название файла исходных данных и перечень анализируемых переменных);

– порядок выполнения работы, который включает формулировку рабочей гипотезы (при наличии таковой), изложение последовательности выполненных действий, иллюстрируемое скриншотами соответствующих диалоговых окон на заключительных этапах совершаемых операций;

– результаты анализа в виде таблиц, рисунков, графиков, полученных в ходе выполнения работы;

– анализ результатов и выводы по работе, включая оценку статистической значимости и маркетинговую интерпретацию.

Отчет по лабораторной работе составляется каждым студентом. Студент, выполнивший работу и оформивший по ней отчет, допускается к защите лабораторной работы.

Контрольные вопросы для защиты лабораторных работ[1]**.**

1 Сущность и область применения соответствующего метода статистического анализа.

2 Возможные типы анализируемых переменных.

3 Основные используемые понятия и статистики.

4 Команды меню программы, используемые при проведении анализа.

5 Критерии статистической значимости полученных результатов.

6 Маркетинговая интерпретация полученных результатов.

Защита лабораторных работ проводится по мере их выполнения в часы занятий, отведенные на выполнение лабораторных работ. Защита студентом выполненных ранее, но незащищенных лабораторных работ проводится в течение лабораторных занятий либо на консультациях в соответствии с графиком кафедры.

[1] За исключением лабораторных работ № 1 и 2, контрольные вопросы для которых приведены после их описания.

1 Формирование исходной базы данных

Цель работы: сформировать исходную базу данных в SPSS.

Задачи работы:
– изучить порядок формирования исходной базы данных в SPSS;
– изучить типы переменных в SPSS;
– провести кодирование исходных данных;
– сформировать исходную базу данных в SPSS.

Задание
На основании исходных данных сформируйте базу данных в SPSS.

Методические указания

Файл исходной базы данных для проведения статистического анализа в SPSS формируется в редакторе данных, который имеет две вкладки (два представления): *Данные*[1] и *Переменные*. Эти вкладки представляют собой таблицы, содержащие информацию о данных, собранных для проведения анализа.

Во вкладке *Переменные* представлена таблица с данными, описывающими свойства переменных. Каждая строка отображает переменную, каждый столбец – ее свойства. Переменные создаются для каждого альтернативного вопроса анкеты, а также для каждого варианта ответа поливариантного вопроса.

Во вкладке *Данные* представлена таблица с данными, описывающими значения переменных. Каждый столбец отображает переменную, каждая строка – отдельное наблюдение (объект сбора информации, каковым, как правило, является респондент).

1 Определение переменных.

В редакторе данных дважды щелкните по ячейке с надписью «*пер*» или щелкните по вкладке *Переменные* на нижнем краю таблицы. В обоих случаях Вы перейдете в режим просмотра переменных. Здесь последовательно, строка за строкой следует определить необходимые переменные и их свойства.

Имя.

Указывается выбранное имя переменной в текстовом формате (как правило, это номер вопроса в анкете или его краткое описание для идентификации). При выборе имени переменной следует соблюдать определенные правила:

– имена переменных могут содержать буквы и цифры. Кроме того, допускаются специальные символы _ (подчеркивание), . (точка), а также символы @ и #. Не разрешаются пробелы, знаки других алфавитов и специальные символы, такие как !, ?, " и *;

– имя переменной должно начинаться с буквы либо одного из символов @, # (служебная переменная) или $ (системная переменная);

[1] Здесь и далее названия рабочих окон, команд и других элементов программы IBM SPSS Statistics приведены для версии 21.

– последний символ имени не может быть точкой или знаком подчеркивания;

– длина имени переменной не должна превышать 32 символов;

– имена переменных могут состоять из любого сочетания символов в верхнем и нижнем регистрах. Регистр сохраняется при отображении имен переменных, однако имена переменных нечувствительны к регистру, т. е. прописные и строчные буквы не различаются.

Тип.

Вновь созданные в SPSS переменные по умолчанию являются численными с максимальной длиной восемь знаков, причем дробная часть состоит из двух знаков. Чтобы изменить тип переменной, следует подвести курсор в соответствующую ячейку таблицы, и при нажатии кнопки мыши на экране откроется диалоговое окно *Тип переменной*.

В SPSS существуют следующие типы переменных (таблица 1).

Таблица 1 – Типы переменных

Тип	Описание
Числовая	Переменная, значения которой являются числами. Значения отображаются в стандартном числовом формате
Запятая	Числовая переменная, значения которой отображаются с запятыми, разделяющими каждые три разряда, а для отделения дробной части используется точка (43,675.67). Если запятые опускаются при вводе, они вставляются автоматически
Точка	Числовая переменная, значения которой отображаются с точками, разделяющими каждые три разряда, а для отделения дробной части используется запятая (43.675,67). Если точки опускаются при вводе, они вставляются автоматически
Научная запись	Числовая переменная, значения которой выводятся в экспоненциальном представлении, т. е. с символом «E» и показателем степени экспоненты с основанием 10 с соответствующим знаком (4,30E+004). Редактор данных принимает в качестве таких переменных числовые значения как со степенью, так и без
Дата	Числовая переменная, значения которой отображаются в одном из нескольких форматов календарной даты или времени. Формат выбирается из списка. Разделителями могут быть слэши, дефисы, точки, запятые и пробелы
Доллар	Числовая переменная, отображаемая со значком доллара вначале ($), точками, отделяющими группы по три разряда, и точкой в качестве десятичного разделителя ($43,675.67). Если знак доллара или запятые опускаются при вводе, они вставляются автоматически
Выбираемая валюта	Числовая переменная, значения которой выводятся в одном из денежных форматов, заданных пользователем. Обозначение валюты при вводе не указывается; оно вставляется автоматически
Текстовая	Переменная, значения которой не являются числовыми. Ее невозможно использовать в вычислениях. Может содержать любые символы, однако их количество не должно превышать заданную величину. Заглавные и строчные буквы считаются разными символами
Ограниченный числовой	Переменная, значения которой ограничены неотрицательными целыми числами. Значения выводятся с предшествующими нулями, дополняющими до максимальной ширины переменной. Группировка цифр не поддерживается

При вводе и выводе данных надо учитывать следующие особенности:

– для числовых форматов десятичным разделителем может быть либо точка, либо запятая. Тип десятичного разделителя зависит от настроек Windows. Точное значение переменной хранится внутри программы, а редактор данных отображает на экране лишь заданное число десятичных разрядов. Значения, которые имеют больше десятичных разрядов, округляются. Для вычислений применяется точное значение;

– для текстовых форматов в длинных строковых переменных значения дополняются пробелами до максимальной длины;

– для форматов даты и времени в качестве разделителей между значениями дня, месяца и числа могут применяться косая черта, дефис, пробел, запятая или точка. Можно выбрать один из нескольких форматов даты (dd-mm-yyyy, dd-mmm-yy, mm/dd/yyyy и т. д.). Дата в формате dd-mmm-yy отображается с разделителем дефисом и сокращением названия месяца из трех букв. Дата в форматах dd/mm/yy и mm/dd/yy отображается с разделителем косой чертой и номером месяца вместо названия;

– всего доступно 27 различных форматов даты и времени, которые отображаются в разворачивающемся списке. В форматах времени в качестве разделителей между значениями часов, минут и секунд могут использоваться двоеточие, точка или пробел;

– форматы отображения специальной валюты CCA, CCB, CCC, CCD и CCE задаются с помощью вкладки *Валюта*, которая открывается командой меню *Правка → Параметры*.

Ширина.

Указывается число знаков, используемых для кодировки переменной.

Десятичные.

Указывается число знаков после запятой при записи значений переменной.

Метка.

Метка переменной – это название, позволяющее описать переменную более подробно, чем имя переменной. При задании меток переменных часто используются формулировки вопросов, содержащихся в анкете. Метка переменной может содержать до 256 символов. При этом различаются прописные и строчные буквы. Они отображаются в том виде, в каком были введены. Метка переменной будет отображаться во всех графиках и таблицах, представляющих результаты статистического анализа, поэтому рекомендуется использовать лаконичные метки для наглядности их представления. По умолчанию метка отсутствует.

Значения.

Здесь указываются коды и описания (метки) возможных значений переменной. Данное поле предназначено для указания вариантов ответа в альтернативных вопросах (как правило, не заполняется для поливариантных переменных). Метки значений могут быть длиной до 120 символов. По умолчанию значения не заданы.

При подведении курсора к соответствующей ячейке таблицы и нажатии клавиши мыши появляется диалоговое окно *Метки значений*. В этом окне в поле *Значение* указываются числовые коды вариантов ответа, а в поле *Метка* –

вербальные формулировки вариантов ответа. При задании меток необходимо предлагать разумные варианты ответов, учитывая, что впоследствии именно эти названия (в том же виде) будут фигурировать на графиках и в аналитических таблицах. Например, вариант ответа на вопрос о половой принадлежности респондента следует называть не «мужской» («женский»), а «мужчины» («женщины»).

Пропущенные.

Указывается, какие коды вариантов ответов следует исключить из анализа. В SPSS допускаются два вида пропущенных значений:

1) пропущенные значения, определяемые системой, – если в матрице данных есть незаполненные численные ячейки, система SPSS самостоятельно идентифицирует их как пропущенные значения. Этот факт отображается в матрице данных с помощью запятой;

2) пропущенные значения, задаваемые пользователем, – если в определенных случаях у переменных отсутствуют значения, например, если на вопрос не был дан ответ, ответ неизвестен, или существуют другие причины, пользователь может с помощью данного свойства переменной объявить эти значения как пропущенные. Пропущенные значения можно исключить из последующих вычислений.

По умолчанию пропущенные значения не заданы.

Ширина столбца.

Задает ширину, которую будет иметь данный столбец в таблице при отображении значений. Ширину столбца также можно изменить непосредственно в окне редактора данных (расширить или сузить путем перетаскивания). По умолчанию ширина столбца равна 8.

Выравнивание.

Задает вид выравнивания значений в таблице данных. Они могут быть выровнены по правому краю, по левому краю или по центру. По умолчанию числовые переменные выровнены по правому краю, а текстовые – по левому.

Шкала.

Указывается шкала, по которой измерена переменная. Может быть номинальной, порядковой или количественной. По умолчанию тип шкалы неизвестен.

Роль.

Некоторые диалоговые окна поддерживают роли, которые можно использовать для предварительного выбора переменных для анализа. При открытии такого диалогового окна переменные, которые удовлетворяют ролевым требованиям, будут автоматически отображаться в списке(ах) назначения. Доступные роли:

– входная – переменная будет использоваться в качестве входной (например, предиктор, независимая переменная);

– целевая – переменная будет использоваться в качестве выходной или целевой (например, независимая переменная);

– двойного назначения – переменная будет использоваться в качестве входной и выходной;

– нет – переменная не имеет назначения роли;

– разделения – переменная будет использоваться, чтобы разделить данные на отдельные выборки для обучения, испытания и проверки;

– расщепления – переменная будет использоваться для двустороннего совмещения с IBM SPSS Modeler.

По умолчанию всем переменным назначается роль *Входная*.

Задав свойства переменной, можно скопировать одно или несколько свойств и применить их к другим переменным. Для этого используются обычные действия по копированию и вставке.

2 Ввод данных.

Ввод данных осуществляется на вкладке *Данные*. Здесь можно настроить вывод на экран значений данных или их меток (команда меню *Вид → Метки значений*).

Вводить данные можно в любую ячейку – последовательно по каждому наблюдению или по переменным, в выбранные области или в отдельные ячейки. Если заданы метки значений и их отображение, вводить данные можно путем выбора из выпадающего списка в соответствующей ячейке. Если начать вводить значение в пустой столбец, редактор данных автоматически создаст новую переменную и присвоит ей имя по умолчанию (*VAR00001*).

При большом количестве переменных можно создавать наборы переменных, которые будут появляться в списках переменных редактора данных и диалоговых окон (команда меню *Сервис → Задать наборы переменных*). Заданные наборы сохраняются вместе с файлами данных.

3 Сохранение данных.

Файл данных в формате SPSS сохраняется с расширением «.sav». В диалоговом окне *Сохранить данные как: переменные* (команда меню *Файл → Сохранить как*) можно выбрать переменные, которые необходимо сохранить в файле данных. По умолчанию сохраняются все переменные. Кроме того, имеется возможность сохранить данные в любом из широкого диапазона форматов внешних файлов. Также можно сохранить содержимое окна вывода как файл *Viewer* («.spv»). Сохраненный документ будет включать обе панели окна вывода: схему и результаты.

Контрольные вопросы

1 Основные возможности программы SPSS.
2 Рабочие окна SPSS.
3 Основные команды меню SPSS.
4 Редактор данных SPSS.
5 Свойства переменных в SPSS.
6 Типы переменных в SPSS.
7 Особенности создания исходной базы данных в SPSS.
8 Импорт данных в SPSS.
9 Сохранение и экспорт данных в SPSS.

2 Модификация, отбор и описательный анализ данных

Цель работы: выполнить модификацию, отбор и описательный анализ данных в SPSS.

Задачи работы:

– изучить порядок отбора и модификации данных в SPSS;

– провести модификацию и отбор данных в SPSS;

– изучить описательные характеристики статистических данных, используемые в SPSS;

– провести описательный анализ данных с применением SPSS;

– интерпретировать полученные результаты.

Задание

На основании исходных данных произведите отбор, модификацию и описательный анализ данных по заданным переменным.

Методические указания

1 Отбор данных.

Выбор наблюдений по определенному условию.

Выберите в меню команду *Данные → Отобрать наблюдения*. Откроется соответствующее диалоговое окно. По умолчанию в этом диалоге выбран пункт *Все наблюдения*.

Выберите пункт *Если выполнено условие* и щелкните по кнопке *Если*. Открывшееся диалоговое окно разделено на следующие части:

– список исходных переменных (левая часть диалогового окна) – содержит переменные, содержащиеся в открытом файле данных;

– редактор условий (верхняя часть диалогового окна) – здесь записывается логическое выражение, по которому должны быть отобраны наблюдения;

– кнопка со стрелкой (между списком исходных переменных и редактором условий) – позволяет перенести переменную из списка исходных переменных в редактор условий;

– клавиатура – содержит цифры, а также арифметические, логические операторы и операторы отношения; если щелкнуть по какой-нибудь кнопке мышью, соответствующий знак будет скопирован в редактор условий;

– список функций – содержит около 140 функций. Каждую из функций можно скопировать в редактор условий двойным щелчком.

Задайте в редакторе условий требуемое условие отбора.

Щелкнув по кнопке *Продолжить* и завершив операцию при помощи щелчка по кнопке *ОК* в главном диалоговом окне, можно увидеть, что респонденты, не соответствующие заданному условию, оказались исключенными из рассмотрения (их номера перечеркнуты).

Чтобы продолжить работу только с отобранными наблюдениями, в диалоговом окне *Отобрать наблюдения* необходимо в области *Вывод* заменить

выбранный по умолчанию параметр *Отфильтровать неотобранные наблюдения* на параметр *Копировать отобранные наблюдения в новый набор данных* с указанием имени этого набора.

Можно не только временно исключить из рассмотрения респондентов, не подходящих под определенное условие, но и полностью удалить такие нерелевантные анкеты из базы данных SPSS. Для этого в диалоговом окне *Отобрать наблюдения* необходимо в области *Вывод* установить параметр *Неотобранные наблюдения удаляются*.

Извлечение случайной выборки.

Для формирования случайных выборок в диалоговом окне *Отобрать наблюдения* предусмотрен параметр *Случайная подвыборка*. Выберите его и щелкните по кнопке *Подвыборка*. Открывшееся диалоговое окно содержит два способа формирования случайной выборки: с указанием доли респондентов, которых необходимо отобрать из исходной выборки (*Примерно*); либо с указанием конкретного количества респондентов, которое необходимо отобрать (*Точно*). При этом в последнем случае необходимо также указать в поле *Из первых ... наблюдений* количество респондентов, из которого следует осуществить выбор. Для формирования случайной выборки из общего числа опрошенных в данном поле следует указать совокупный размер выборки.

2 Модификация данных.

Сортировка наблюдений.

Выберите в меню команду *Данные → Сортировать наблюдения*. В открывшемся диалоговом окне левый список содержит все доступные в текущей базе данных переменные. В область *Сортировать по* поместите переменные, по которым следует произвести сортировку. Порядок следования переменных в данной области соответствует порядку сортировки, т. е. сначала сортировка происходит по первой переменной, затем – по второй и т. д. Группа переключателей *Порядок сортировки* позволяет выбрать направление сортировки: по возрастанию или убыванию. При этом для каждой переменной можно выбрать свой тип сортировки.

Перекодирование значений переменных.

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

Перекодирование внутри одной переменной.

Выберите в меню команду *Преобразовать → Перекодировать в те же переменные*. Открывшееся диалоговое окно в левой области содержит список всех доступных переменных, а в правой, имеющей метку *Переменные*, – место для помещения перекодируемых переменных. Необходимо отметить, что за один цикл использования диалогового окна *Перекодировать в те же переменные* можно перекодировать сколько угодно переменных, но только одними и теми же кодами. Иными словами, нельзя в одной переменной нули заменить на единицы, а в другой – шестерки на строки «шесть». Для этого придется сначала перекодировать первую переменную, а затем вновь вернуться в диалоговое окно

Перекодировать в те же переменные, щелкнуть по кнопке *Сброс* и затем ввести данные для перекодировки другой переменной.

Кроме того, диалоговое окно *Перекодировать в те же переменные* так же, как и многие другие окна SPSS, содержит кнопку *Если*, позволяющую осуществить действия не для всех респондентов в выборке, а только для отдельных групп.

Поместите в поле *Переменные* переменную, значения которой требуется перекодировать. Щелкните по кнопке *Старые и новые значения*, которая открывает диалоговое окно, позволяющее задать перекодируемые значения. Это окно разделено на две части. В левой части можно указать, какие конкретно значения подлежат перекодировке, а в правой – в какие значения они будут перекодированы. Чтобы указать конкретное значение для перекодировки, введите исходное значение в левое поле *Старое значение*, а конечное значение – в правое поле *Новое значение*. Обязательно щелкните по кнопке *Добавить*, чтобы добавить указанное сочетание в список перекодировки.

Также можно задавать диапазоны значений переменной для перекодировки.

После того как были указаны все необходимые варианты перекодирования, следует закрыть окно щелчком по кнопке *Продолжить* и запустить процедуру перекодирования кнопкой *ОК*. В исходной базе данных SPSS все значения указанных переменных будут перекодированы.

Перекодирование с образованием новых переменных.

Выберите в меню команду *Преобразовать → Перекодировать в другие переменные*). Откроется соответствующее диалоговое окно, которое аналогично окну *Перекодировать в те же переменные*, только добавлена дополнительная область *Выходная переменная*, предназначенная для указания имени и метки вновь создаваемой переменной, которая будет содержать перекодированные значения.

Введите в соответствующие поля название и метку новой переменной. Откройте диалоговое окно *Старые и новые значения*, щелкнув по одноименной кнопке. В нем содержатся некоторые дополнительные полезные инструменты. По умолчанию значения исходной переменной, не указанные в списке перекодировки, не попадают в новую переменную. Изменить данное условие по умолчанию можно при помощи параметра *Скопировать старые значения*. Также есть возможность конвертации числовых значений в строковые (параметр *Новые переменные – текстовые*). При этом изменится тип всей новой переменной; следовательно, все исходные значения должны быть перекодированы как строковые. Существует и обратная возможность – конвертации строковых значений, похожих на цифры, в числовой вид (например, «5» в 5). Данная возможность реализуется при помощи параметра *Преобразовывать текст в числа*.

Автоматическое перекодирование.

Данная процедура предназначена для автоматического кодирования полей анкеты числовыми значениями типа «индекс». При ее выполнении одинаковые ответы из текстовых полей группируются, и им присваиваются соответствующие коды ответа (например, начиная с 1).

Выберите в меню команду *Преобразовать → Автоматическая перекодировка*. Перенесите интересующую строковую переменную в поле *Переменная -> Новое имя*. В текстовое поле под ним введите новое имя и щелкните по кнопке *Добавить новое имя*. В группе переключателей *Начать перекодировку с* есть два параметра, позволяющих присвоить номера вариантам ответа либо по алфавиту, начиная с самого малого значения (*Минимальное значение*), либо начиная с конца упорядоченного списка вариантов ответа (*Максимальное значение*).

После щелчка по кнопке *ОК* и выполнения указанных преобразований в базе данных будет создана новая числовая переменная с вариантами ответа согласно списку перекодировки.

Вычисление переменных.

Вычисление новых переменных.

Выберите в меню команду *Преобразовать → Вычислить переменную*. В открывшемся диалоговом окне в поле *Вычисляемая переменная* укажите имя переменной, которой присваивается вычисленное значение. В качестве выходной переменной может служить уже существующая или новая переменная. Щелкните по кнопке *Тип и Метка*, чтобы указать метку и тип новой переменной.

В поле *Числовое выражение* введите выражение, применяемое для определения значения выходной переменной. В этом выражении могут использоваться имена существующих переменных, константы, арифметические операторы и функции. Формулу можно ввести либо вручную, либо используя список переменных и клавиатуру диалогового окна.

После запуска процедуры вычисления (кнопка *ОК*) будет создана новая переменная.

Подсчет значений переменных.

В SPSS есть возможность подсчитать количество появления одного и того же значения или значений для определенной переменной (как правило, поливариантной).

Выберите в меню команду *Преобразовать → Подсчитать значения в наблюдениях*. В открывшемся диалоговом окне в полях *Вычисляемая переменная* и *Метка вычисляемой переменной* следует указать соответственно имя вновь создаваемой переменной и ее метку. В область *Переменные* поместите переменную, значения которой необходимо подсчитать.

Щелкните по кнопке *Задать значения*. Откроется диалоговое окно, которое служит для определения подсчитываемых значений. Можно задать отдельное значение, диапазон или сочетание того и другого.

3 Описательный анализ данных.

Статистические характеристики вычисляются в основном для переменных, относящихся к интервальной или относительной шкале. Для этого используются следующие четыре команды меню:

1) *Анализ → Описательные статистики → Частоты*;
2) *Анализ → Описательные статистики → Описательные*;
3) *Анализ → Описательные статистики → Разведочный анализ*;
4) *Анализ → Отчеты → Итоги по наблюдениям*.

В таблице 2 приведен обзор характеристик, рассчитываемых в SPSS.

Таблица 2 – Обзор статистических характеристик разных команд меню *Анализ*

Характеристика	Частоты	Описательные	Разведочный анализ	Итоги по наблюдениям
Среднее значение	X	X	X	X
Сумма	X	X		X
Медиана	X		X	X
Групповая медиана	X			X
Квартиль	X			
Процеитиль	X		X	
Мода	X			
Стандартное отклонение	X	X	X	X
Стандартная ошибка	X	X	X	X
Дисперсия	X	X	X	X
Минимум	X	X	X	X
Максимум	X	X	X	X
Размах	X	X	X	X
Межквартильная широта			X	
Эксцесс (вариация)	X	X	X	X
Асимметрия	X	X	X	X
Стандартная ошибка эксцесса	X	X	X	X
Стандартная ошибка асимметрии	X	X	X	X
Доверительный интервал			X	
Гармоническое среднее				X
Геометрическое среднее				X
М-оценка (Хампеля)			X	
Выброс			X	
Усеченное среднее			X	

В меню *Описательные* можно также провести стандартизацию переменных (*z*-преобразование).

Статистические характеристики, которые задаются в меню *Итоги по наблюдениям*, можно также вычислить раздельно по категориям группирующих переменных, относящихся к номинальной или порядковой шкале.

Наиболее часто при описательном анализе данных используются частотные таблицы и описательные статистики.

Частотные таблицы.

Выберите в меню команду *Анализ* → *Описательные статистики* → *Частоты*. В открывшемся диалоговом окне перенесите требуемые переменные в список выходных переменных.

С помощью кнопки *Формат* укажите, каким способом следует сортировать результаты в частотных таблицах. В открывшемся диалоговом окне в группе *Упорядочить по* можно выбрать порядок, в котором будут отображены значения в частотной таблице. Возможны следующие варианты:

– *по возрастанию значений* – данные сортируются по возрастанию значений. Это настройка по умолчанию;

– *по убыванию значений* – данные сортируются по убыванию значений;

– *по возрастанию частот* – данные сортируются по возрастанию частот;

– *по убыванию частот* – категории сортируются по убыванию частот.

Кроме того, флажок *Отключить таблицы со многими категориями* позволяет избежать вывода длинных частотных таблиц.

Подтвердите выбор кнопкой *Продолжить*.

Кнопка *Диаграммы* вызывает одноименное диалоговое окно, которое позволяет, помимо таблиц, вывести диаграммы по выбранным переменным. По умолчанию SPSS не выводит диаграммы (параметр *Нет* в группе *Тип диаграммы*). Можно построить столбиковую диаграмму, круговую или гистограмму. В группе *Значения на диаграмме* отмечается необходимость отобразить на диаграмме абсолютные значения частот или относительные значения, т. е. проценты. Подтвердите выбор кнопкой *Продолжить*.

После щелчка по кнопке *ОК* в главном диалоговом окне *Частоты* откроется окно вывода *Viewer*, в котором будут представлены частотные таблицы, а также другая информация, указанная на подготовительном этапе. Перед самой частотной таблицей выводится небольшая таблица с обзором валидных (допустимых) и пропущенных значений.

После частотных таблиц отображается диаграмма, если была указана необходимость ее построения. Дважды щелкните по области диаграммы – откроется редактор диаграмм, в котором можно придать диаграмме желаемый вид.

Описательные статистики.

Чтобы получить описательную статистику количественных переменных, нужно в диалоговом окне *Частоты* щелкнуть по кнопке *Статистики*. Откроется соответствующее диалоговое окно.

В группе *Значения процентилей* можно выбрать следующие варианты:

– *квартили* – будут показаны первый, второй и третий квартили;

– *процентили для … равных групп* – будут вычислены значения процентилей, разделяющие выборку на группы наблюдений, которые имеют одинаковую ширину, т. е. включают одно и то же количество измеренных значений. По умолчанию предлагается количество групп 10;

– *процентили* – здесь имеются в виду значения процентилей, определяемые пользователем. Введите значение процентиля в пределах от 0 до 100 и щелкните по кнопке *Добавить*. Повторите эти действия для всех желаемых значений процентилей. Значения в порядке возрастания будут показаны в списке.

В группах *Разброс*, *Расположение*, *Распределение* можно выбрать соответствующие статистические характеристики.

В диалоговом окне также есть флажок *Значения – центры групп*. Если установить его, то при вычислении медианы и остальных значений процентилей оценки этих характеристик будут определяться для концентрированных данных.

Для возврата в диалоговое окно *Частоты* щелкните по кнопке *Продолжить*. Для просмотра только таблицы с описательной статистикой деактивируйте опцию *Вывести частотные таблицы*. Щелкните по кнопке *ОК*.

Частотные таблицы для поливариантных вопросов (вопросов с множественными ответами).

Чтобы построить частотные распределения по поливариантному вопросу,

прежде всего необходимо сформировать набор множественных ответов (многовариантную переменную) для данного вопроса. Это делается при помощи команд меню *Анализ → Множественные ответы → Задать наборы множественных ответов*.

Открывшееся диалоговое окно позволяет сформировать соответствующий набор переменных (правый список) из общего списка доступных переменных (левый список). В группе *Переменные кодируются как* задается тип используемой кодировки – по умолчанию *Дихотомии* (указывает, что переменные, обозначающие варианты ответа в поливариантном вопросе, являются *дихотомическими*). В поле *Подсчитываемое значение* введите цифру-код, указывающую, что вариант ответа выбран. Заполните поля *Имя* и *Метка* для создаваемой многовариантной переменной. Затем щелкните по кнопке *Добавить*. Обратите внимание, что к имени создаваемых многовариантных переменных добавляется префикс $ (этим они отличаются от обычных одновариантных переменных). Теперь Вы можете создать еще одну или несколько многовариантных переменных, добавляя их в соответствующий список при помощи кнопки *Добавить*. Завершается процесс создания многовариантных переменных щелчком по кнопке *Закрыть*.

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

Чтобы создать частотную таблицу для полученной переменной, выберите в меню команду *Анализ → Множественные ответы → Частоты*. В открывшемся диалоговом окне в списке *Наборы множественные ответов* отображаются уже определенные наборы переменных. Перенесите интересующие наборы в список *Таблицы для* и щелкните по кнопке *ОК*.

Следует отметить, что данное меню позволяет строить только таблицы линейных распределений, но не дает возможности вывести диаграммы. Для этого применяется команда меню *Графика → Панель выбора диаграмм*. При этом многовариантная переменная должна быть создана с помощью команды меню *Данные → Задать наборы множественных ответов* (а не *Анализ → Множественные ответы*).

Контрольные вопросы

1 Отбор исходных данных в SPSS: назначение, виды, порядок выполнения.

2 Модификация исходных данных в SPSS: сортировка наблюдений.

3 Модификация исходных данных в SPSS: перекодирование значений переменных.

4 Модификация исходных данных в SPSS: вычисление переменных.

5 Получение описательных статистических характеристик с помощью различных команд меню SPSS.

6 Построение частотных таблиц в SPSS.

3 Построение таблиц сопряженности признаков и проверка гипотез

Цель работы: построить таблицы сопряженности признаков и выполнить проверку гипотез с применением SPSS.

Задачи работы:

– изучить методику построения таблиц сопряженности признаков в SPSS;
– построить таблицы сопряженности признаков с применением SPSS;
– изучить методику проверки гипотез в SPSS;
– провести проверку гипотез о связях и различиях с применением SPSS;
– оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных постройте таблицы сопряженности признаков и проведите проверку гипотез для заданных переменных. Сделайте выводы по всем таблицам.

Методические указания

1 Создание таблиц сопряженности.

Для создания таблиц сопряженности и вычисления меры связанности на их основе выберите в меню команду *Анализ → Описательные статистики → Таблицы сопряженности*.

В открывшемся диалоговом окне список исходных переменных содержит переменные открытого файла данных. Здесь можно выбрать переменные для *Строк* и *Столбцов* таблицы сопряженности. Для каждого сочетания двух переменных будет создана двухмерная таблица. При необходимости увеличения количества измерений таблицы переменные добавляются в область *Слой*. Для категории каждой из переменной слоев будет создана отдельная таблица сопряженности. Чтобы добавить новый слой, щелкните по кнопке *Следующий*. Каждый последующий уровень делит таблицу сопряженности на меньшие подгруппы. Переходить от одного слоя к другому можно при помощи кнопок *Следующий* и *Предыдущий*.

Щелкните по кнопке *Ячейки*. Открывшееся диалоговое окно предназначено для задания значений, выводимых в таблице сопряженности. По умолчанию в каждой ячейке таблицы выводится только количество респондентов (параметр *Наблюдаемые* в области *Частоты*). Область *Проценты* позволяет организовать их вывод в ячейках таблицы по *Строкам*, *Столбцам*, а также от общего числа респондентов, ответивших одновременно на все вопросы, по которым строится перекрестное распределение (*По таблице* (*слою*)).

Можно изменить порядок сортировки переменных строк в таблице сопряженности, щелкнув в диалоговом окне *Таблицы сопряженности* по кнопке

Формат. В открывшемся диалоговом окне в группе *Порядок строк* можно выбрать один из следующих вариантов сортировки значений: *По возрастанию* или *По убыванию*.

Щелкните по кнопке *ОК* в диалоговом окне *Таблицы сопряженности*, и будет создана таблица сопряженности в требуемом формате. В окне вывода *Viewer* будут показаны следующие таблицы: *Сводка обработки наблюдений* и *Таблица сопряженности <Переменная 1> * <Переменная 2> ...* . Первая таблица содержит информацию о числе самих наблюдений, вторая – это собственно таблица сопряженности.

Чтобы сделать более наглядными данные, содержащиеся в таблицах сопряженности, их можно представить визуально. Для этого установите в диалоговом окне *Таблицы сопряженности* флажок *Вывести кластеризованные столбиковые диаграммы*.

Таблицы сопряженности можно также создавать между многовариантными переменными. Для этого используется команда меню *Анализ → Множественные ответы → Таблицы сопряженности*[1].

В открывшемся диалоговом окне слева размещены два списка переменных: в верхнем – все доступные переменные из файла данных, в нижнем – сформированные многовариантные переменные. В перекрестном анализе могут принимать участие как многовариантные переменные, так и другие доступные одновариантные переменные. Можно задать несколько измерений (максимум три) при помощи введения одного дополнительного *Слоя*. Имейте в виду, что при построении перекрестных таблиц переменные, находящиеся в областях *Строки, Столбцы* и *Слой*(*и*), перекрещиваются по тройкам последовательно.

Если таблица сопряженности строится между одновариантными и многовариантными переменными, то для первых следует задать диапазон значений. Щелкните по кнопке *Задать диапазон*. В открывшемся диалоговом окне в соответствующих полях следует указать минимальное (*Минимум*) и максимальное (*Максимум*) значения, которые может принимать одновариантная переменная.

Щелкните по кнопке *Параметры*. Откроется диалоговое окно, которое позволяет указать, нужно ли выводить проценты (по *Строкам*, *Столбцам* или *Таблице* (*слою*)), а также определить, что является базой для расчета процентов: количество *Респондентов* или количество *Ответов* на вопрос. Флажок *Сопоставить переменные по наборам ответов* имеет смысл выбирать, только если таблица сопряженности строится на основе двух наборов переменных. В этом случае первая переменная из первого набора сочетается с первой переменной из второго набора и т. д.

Щелкните по кнопке *ОК* в диалоговом окне *Таблицы сопряженности для множественных ответов*, и будет создана таблица сопряженности в требуемом формате.

---

[1] Чтобы строить распределения по многовариантным переменным, сначала их нужно сформировать при помощи команды меню *Анализ → Множественные ответы → Задать наборы множественных ответов* – см. лабораторную работу № 2.

2 Проверка гипотезы о связях (распределении частот).

Исследовать существование зависимости между переменными таблицы сопряженности позволяет вычисление значений их ожидаемых частот. Чтобы определить эти значения, выберите в меню команду *Анализ → Описательные статистики → Таблицы сопряженности*.

Щелкните по кнопке *Ячейки*. В открывшемся диалоговом окне в группе *Частоты* установите флажок *Ожидаемые*, при этом флажок *Наблюдаемые* остается неизменным. Сравнение полученных в таблице наблюдаемых и ожидаемых частот позволяет подтвердить или опровергнуть предположение о взаимосвязи между изучаемыми переменными.

Более наглядную возможность выявления существования зависимости между переменными дает вычисление остатков. Эти остатки являются показателем того, насколько сильно наблюдаемые и ожидаемые частоты отклоняются друг от друга. Чтобы получить остатки частот, в диалоговом окне *Таблицы сопряженности*: *Вывод в ячейках* в группе *Остатки* установите флажок *Нестандартизированные*.

Выявить статистическую значимость зависимостей между переменными таблицы сопряженности позволяют критерий хи-квадрат (χ^2) и сопутствующие тесты.

Для того чтобы организовать наряду с таблицей сопряженности вывод соответствующих статистик, щелкните по кнопке *Статистики* в диалоговом окне *Таблицы сопряженности*. В открывшемся диалоговом окне выберите параметр *Хи-квадрат*. Это позволит впоследствии определить, имеется ли определенная связь между исследуемыми переменными.

При анализе зависимостей, кроме обнаружения наличия связи, также можно определить, насколько сильно выражена данная зависимость (установить силу связи). Сделать это позволяют релевантные статистические тесты, применяемые отдельно для каждого из трех типов переменных, участвующих в анализе. *Для номинальных* переменных следует применять один из тестов, представленных в соответствующей области. Наиболее универсальным и часто применяемым методом является *Фи и V Крамера*. Для *Порядковых* переменных следует применять один из методов, представленных в соответствующей области. Рекомендуется использовать наиболее универсальный метод *Гамма*. Теоретически этот же метод можно применять и для интервальных переменных, однако все же для них целесообразно использовать более релевантную процедуру корреляционного анализа.

Установите необходимые флажки, щелкните по кнопке *Продолжить*, а в главном диалоговом окне – по кнопке *ОК*.

В окне вывода появится таблица перекрестного распределения переменных. Кроме того, будет показана таблица с результатами теста по *Критерию хи-квадрат*. В ней для вычисления критерия χ^2 применяются три различных подхода:

1) формула Пирсона – строка *Хи-квадрат Пирсона*;
2) поправка на правдоподобие – строка *Отношение правдоподобия*;
3) тест Мантеля-Хензеля – строка *Линейно-линейная связь*.

Если таблица сопряженности имеет четыре поля и ожидаемая вероятность

менее 5, дополнительно выводится *Точный критерий Фишера*.

В последней строке приводится объем анализируемой выборки (*Кол-во валидных наблюдений*).

Необходимо отметить, что расчет всех статистических процедур производится по отдельности для каждого варианта переменной, расположенной в слоях (в случае наличия таковой).

В столбцах таблицы *Критерии хи-квадрат* представлены для соответствующих критериев *Значения*, число степеней свободы (*ст. св.*) и асимптотическая значимость двусторонняя (*Асимпт. значимость 2-стор.*).

Именно из условия статистической значимости критерия χ^2 следует статистическая значимость анализируемой зависимости между переменными. В таблице 3 представлен наиболее распространенный способ интерпретации различных уровней значимости в маркетинговых исследованиях.

Таблица 3 – Интерпретация уровней значимости

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

При этом корректность проведения теста χ^2 по формуле Пирсона определяется двумя условиями:

1) ожидаемые частоты меньше 5 должны встречаться не более чем в 20 % полей таблицы – это значение отображается в примечании «а» в первой строке после таблицы *Критерии хи-квадрат*. На практике приемлемая доля ожидаемых частот меньше 5 может отклоняться от 20 % (в пределах +5 пп.). При наличии ярко выраженной зависимости можно считать такую зависимость статистически значимой;

2) суммы по строкам и столбцам всегда должны быть больше нуля.

Альтернативой формуле Пирсона для вычисления критерия χ^2 является поправка на правдоподобие. При большом объеме выборки формула Пирсона и подправленная формула дают очень близкие результаты.

Дополнительно в таблице теста хи-квадрат выводится значение теста Мантеля-Хензеля (*Линейно-линейная связь*). Эта форма критерия χ^2 – мера линейной зависимости между строками и столбцами таблицы сопряженности. Если величина данного теста статистически значима, следовательно, между строковой и столбцовой переменными есть линейная зависимость.

В случае выбора в диалоговом окне *Таблицы сопряженности: Статистики* сопутствующих тестов далее в окне *Вывода* будут отображены таблицы *Симметричные меры*, из которых можно узнать о силе и направлении (только для порядковых и интервальных переменных) связи между анализируемыми переменными. *Значение* выбранной меры является коэффициентом корреляции.

В таблице 4 представлены словесные описания величин коэффициента корреляции.

Таблица 4 – Интерпретация величины коэффициента корреляции

Модуль значения коэффициента корреляции r	Статистическая интерпретация
$0 < r \leq 0{,}2$	Очень слабая корреляция
$0{,}2 < r \leq 0{,}5$	Слабая корреляция
$0{,}5 < r \leq 0{,}7$	Средняя корреляция
$0{,}7 < r \leq 0{,}9$	Сильная корреляция
$0{,}9 < r \leq 1$	Очень сильная корреляция

3 Проверка гипотезы о различиях (средних значениях).

При сравнении средних значений выборок предполагается, что обе выборки подчиняются нормальному распределению. Если это не так, то вычисляются медианы и для сравнения выборок используется непараметрический тест.

Для установления различий между двумя группами респондентов предназначены *t*-тесты.

Сравнение двух независимых выборок.

Под независимыми выборками понимаются бинарные категории какой-либо переменной, т. е. существуют два уровня группирующей (независимой) переменной и несколько уровней независимой переменной, на основании которых и будет выполняться различие между группами независимой переменной.

Для выполнения *t*-теста выберите в меню команду *Анализ → Сравнение средних → Т-критерий для независимых выборок*. В область *Проверяемые переменные* поместите переменную, которая будет являться зависимой. Затем в поле *Группировать по* переместите переменную, являющуюся критерием для установления различий.

Щелчком по кнопке *Задать группы* открывается окно, в котором следует ввести значения двух категорий для группирующей переменной. Обратите внимание, что если вместо дихотомии имеется группирующая переменная с интервальной шкалой, это диалоговое окно позволяет установить точку отсечения *Пороговое значение*, которая будет разделять все возможные значения данной переменной на две группы.

С помощью кнопки *Параметры* в главном диалоговом окне рассматриваемой процедуры можно установить доверительный уровень для результатов расчета *t*-теста. По умолчанию установлен уровень доверия 95 %.

После завершения процедуры расчета *t*-теста в окне просмотра *Viewer* будут отражены результаты в виде двух таблиц.

В первой таблице *Групповые статистики* содержатся количество наблюдений (*N*), *средние значения*, *стандартные отклонения* и *стандартные ошибки средних* в обеих группах.

Вторая таблица *Критерии для независимых выборок* позволяет установить статистическое различие между данными значениями. Она содержит:

– результаты *теста Ливиня* (Левена) *на равенство дисперсий*: значение (*F*) и значимость (*Знч.*);

– результаты *t*-теста: значение распределения (*t*), количество степеней свободы (*ст. св.*), вероятность ошибки *p* под обозначением «*Значимость* (*2-сторонняя*)»;

– *разность средних значений*, ее *стандартную ошибку* и *доверительный интервал*.

Анализ этой таблицы начинается с определения значимости теста Ливиня. Если он статистически незначим, то различие между двумя анализируемыми средними на основании *t*-теста определяется из строки *Предполагается равенство дисперсий*; в противном случае – из строки *Равенство дисперсий не предполагается*.

Сравнение двух зависимых выборок.

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

Для выполнения *t*-теста выберите в меню команду *Анализ → Сравнение средних → T-критерий для парных выборок*. В левом списке содержатся все доступные переменные из базы данных. Указав две переменные для анализа, перенесите их в область *Парные переменные*. Кнопка *Параметры* позволяет установить уровень доверия для производимых расчетов.

Щелкните по кнопке *ОК*, чтобы начать вычисления. В окне просмотра *Viewer* появятся три таблицы с результатами.

Первая таблица *Статистики парных выборок* содержит *средние значения*, количество наблюдений (*N*), *стандартные отклонения* и *стандартные ошибки средних* для обеих переменных.

В следующей таблице *Корреляции парных выборок* представлены коэффициент *корреляции* Пирсона между переменными и значимость (*Знч.*) его отклонения от нуля.

Третья таблица *Критерий парных выборок* содержит:

– *парные разности*: *среднее значение*, *стандартное отклонение*, *стандартная ошибка* и *доверительный интервал*;

– результаты *t*-теста: тестовая величина (*t*), количество степеней свободы (*ст. св.*), вероятность ошибки *p* под обозначением «*Значимость* (*2-сторонняя*)».

Эта таблица позволяет сделать вывод о наличии/отсутствии статистически значимого различия между тестируемыми переменными.

T-тест одной выборки.

Этот тест позволяет выяснить, отличается ли среднее значение, полученное на основе данной выборки, от предварительно заданного контрольного значения. Возможен также вариант определения, отличается ли среднее значение какого-либо параметра для определенной целевой группы респондентов от среднего значения по всей выборке.

Для выполнения *t*-теста выберите в меню команду *Анализ → Сравнение средних → Одновыборочный t-критерий*. Перенесите из левого списка всех доступных переменных в правую область *Проверяемую переменную*. В поле *Проверяемое значение* укажите значение, с которым будет сравниваться среднее тестируемой переменной. Кнопка *Параметры* позволяет указать доверительный уровень, для которого устанавливается различие.

Щелкните по кнопке *ОК*, чтобы начать вычисления. В окне просмотра *Viewer* появятся две таблицы с результатами.

В первой таблице *Статистика для одновыборочного t-критерия* отражены расчеты среднего значения исследуемой переменной: количество значений (*N*), *среднее значение*, *стандартное отклонение*, *стандартная ошибка среднего*.

Вторая таблица *Одновыборочный t-критерий* позволяет сделать вывод о статистической значимости/незначимости тестируемого различия. Она содержит значение тестовой величины (*t*), количество степеней свободы (*ст. св.*), вероятность ошибки *p* под обозначением «*Значимость* (*2-сторонняя*)», *разность* между реальным и тестируемым значениями *средних* и *доверительный интервал разности средних*.

4 Дисперсионный и ковариационный анализ данных

Цель работы: выполнить дисперсионный и ковариационный анализ данных с применением SPSS.

Задачи работы:

– изучить методику выполнения дисперсионного и ковариационного анализа в SPSS;

– провести однофакторный дисперсионный анализ данных с применением SPSS, оценить статистическую значимость и интерпретировать полученные результаты;

– провести многофакторный дисперсионный анализ данных с применением SPSS, оценить статистическую значимость и интерпретировать полученные результаты;

– провести ковариационный анализ данных с применением SPSS, оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных:

– выполните двумя способами однофакторный дисперсионный анализ для заданных переменных;

– выполните двухфакторный дисперсионный анализ для заданных переменных;

– выполните однофакторный дисперсионный анализ для заданных переменных. С помощью ковариационного анализа проверьте надежность выявленной зависимости (убедитесь, что при всех возможных ковариатах сохраняется значимость эффекта независимой переменной).

Методические указания

1 Однофакторный дисперсионный анализ.

Однофакторный одномерный дисперсионный анализ можно проводить двумя способами: посредством *Общей линейной модели GLM* или при помощи

специальной процедуры *Однофакторного дисперсионного анализа ANOVA*.

Общая линейная модель GLM.

Диалоговое окно одномерного дисперсионного анализа запускается при помощи команды меню *Анализ → Общая линейная модель → ОЛМ-одномерная*. Из левого списка всех доступных переменных переместите зависимую переменную в соответствующее поле, независимую – в поле *Фиксированные факторы*.

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

Чтобы определить это, существуют дополнительные статистические тесты, задаваемые при помощи кнопки *Апостериорные*. В соответствующем диалоговом окне перенесите из области *Фактор*(*ы*) в область *Апостериорные критерии для* те независимые переменные, которые необходимо подвергнуть тестированию на предмет установления различий между их группами (в случае однофакторного анализа такая переменная одна). Далее укажите релевантные дополнительные тесты для указанной переменной. При этом SPSS выводит различные тесты для *равных* и *неравных дисперсий*.

В общем случае неизвестно, равны ли дисперсии и, соответственно, какую группу статистических тестов следует использовать. Поэтому рекомендуется сразу вывести тесты и для равных, и для неравных дисперсий, чтобы сократить количество итераций при проведении дисперсионного анализа. SPSS предлагает много различных дополнительных тестов, помогающих определить различия между группами исследуемых переменных. Однако использовать их все нецелесообразно. Рекомендуется ограничиться наиболее популярным и универсальным тестом *Шеффе* для равных дисперсий и тестом *Т2 Тамхейна* – для неравных. Теперь можно закрыть описываемое диалоговое окно щелчком по кнопке *Продолжить*.

Параметры вывода результатов расчета можно указать в диалоговом окне *Параметры*, вызываемом одноименной кнопкой в главном диалоговом окне *ОЛМ-одномерная*. В области *Вывести* активируется необходимость проведения различных статистических тестов. Для однофакторного дисперсионного анализа можно ограничиться только одним тестом Ливиня на равенство дисперсий (параметр *Критерии однородности*). Кроме того, задайте вывод *Описательных статистик*, установив соответствующий флажок.

Следует отметить, что если исследуемая независимая переменная имеет всего две категории (дихотомия), апостериорные тесты для нее не проводятся. В таком случае установить направление различия между категориями позволяет вывод средних значений зависимой переменной в каждой из двух категорий. Для этого перенесите исследуемую независимую дихотомическую переменную из области *Факторы и их взаимодействия* в область *Вывести средние для*. Если независимая переменная имеет больше двух категорий, специально выводить для нее средние значения нет смысла (они будут выведены в таблице *Однородные подмножества*).

Остальные кнопки главного диалогового окна *ОЛМ-одномерная* предназначены для многофакторного анализа.

Теперь щелкните по кнопке *ОК*, чтобы запустить процедуру дисперсионного анализа. В окне *Вывод* будут отображены результаты расчетов.

Таблица *Межгрупповые факторы* содержит общую информацию о независимой переменной.

В таблице *Описательные статистики* в разрезе каждой группы независимой переменной представлены *средние значения*, *стандартные отклонения* и количество наблюдений (*N*).

Первой практически значимой таблицей является *Критерий Ливиня проверки равенства дисперсий* зависимой и независимых переменных. В столбце *Значимость* содержится единственное интересующее нас значение – это статистическая значимость тестовой статистики *F*. Если значение в данном столбце показывает незначимость *F*, то дисперсии равны, и в дальнейшем мы будем анализировать результаты расчета теста *Шеффе*, предполагающего равенство дисперсий. В противном случае, если *F*-статистика значима – дисперсии не равны, и при анализе различий между группами следует использовать тест *Тамхейна,* предполагающий неравенство дисперсий.

Следующая таблица – это *Оценка эффектов межгрупповых факторов*. Она является основной в выводимых результатах дисперсионного анализа и показывает наличие/отсутствие значимых различий между категориями исследуемых переменных. Таблица содержит типовую схему дисперсионного анализа, включая *Сумму квадратов*, Число степеней свободы (*ст. св.*), *Средний квадрат* для межгрупповой (независимая переменная) и внутригрупповой дисперсии (ошибка), а также значение *F*-критерия и оценку его значимости (*Знч.*).

Первое, на что следует обратить внимание при анализе описываемой таблицы, – это величина *R-квадрат* (отображается в примечании «а» после таблицы), характеризующая долю совокупной дисперсии в зависимой переменной, описываемой статистической моделью. Другими словами, это та часть вариации зависимой переменной, которую можно объяснить на основании независимой переменной. Естественно, что чем меньше независимых переменных, тем меньше величина R^2, и наоборот.

Второе, на что обращают внимание при интерпретации описываемой таблицы, – значимость различия между группами независимой переменной. Этот вывод следует из значения на пересечении строки, содержащей соответствующую независимую переменную (в случае однофакторного анализа можно рассматривать первую строку – *Скорректированная модель*), и столбца *Значимость*. Обратите внимание, что если тест Ливиня выявил факт неравенства дисперсий независимых и зависимых переменных, следует поднять порог значимости со стандартного значения 0,05 до 0,01.

Если установлено наличие статистически значимого различия между исследуемыми группами, необходимо определить, какие из имеющихся групп отличаются от остальных и каким образом (в большую или в меньшую сторону). Осуществить это можно при помощи таблицы *Множественные сравнения*. При интерпретации данной таблицы, прежде всего, вспомните результаты теста Ливиня.

Если на основании этого теста дисперсии оказались равными, в данной таблице будет рассматриваться только та ее часть, в которой приведены расчеты по методу *Шеффе*. Тест *Тамхейна* применяется, если дисперсии не равны.

В описываемой таблице представлено сравнение различий между каждой из анализируемых групп с остальными группами. На основе этих данных и определяются группы, которые значимо отличаются от других, опираясь на данные столбца *Значимость*. При этом из столбца *Разность средних* можно видеть, насколько отличается среднее значение той или иной группы от среднего значения других групп (звездочками отмечены значимые различия при 95-процентном доверительном уровне).

Наконец, в последней таблице *Однородные подмножества* представлена однозначная картина различий между группами независимой переменной. Все эти группы разделены на категории на основании различий в значениях зависимой переменной. Если бы оказалось, что статистически значимых различий не наблюдается, все группы независимой переменной были бы отнесены к одной категории (*Подмножество* было бы только 1). Иногда возникает ситуация, при которой одна и та же группа респондентов может относиться сразу к нескольким группам. В таком случае следует поднять порог значимости со стандартных 0,05, скажем, до 0,01 или любого другого значения (кнопка *Параметры* диалогового окна *ОЛМ-одномерная*).

Процедура One-way ANOVA.

Выберите в меню команду *Анализ → Сравнение средних → Однофакторный дисперсионный анализ*. В открывшемся диалоговом окне перенесите соответствующие переменные в *Список зависимых переменных* и в поле *Фактор*.

В диалоговом окне *Параметры* задайте вывод *Описательных статистик* и *Проверку однородности дисперсий*, установив соответствующие флажки.

Чтобы выполнить апостериорный тест, вернувшись в основное диалоговое окно, щелкните по кнопке *Апостериорные*. Выберите тест *Шеффе* для *равных дисперсий* и тест *Т2 Тамхейна* – для *неравных дисперсий*.

Запустите тест, щелкнув по кнопке *ОК*. В окне просмотра появятся следующие таблицы:

– *Описательные статистики*;

– *Критерий однородности дисперсий* – содержит результаты теста Ливиня на гомогенность дисперсий;

– *Однофакторный дисперсионный анализ* – содержит типовую схему дисперсионного анализа по оценке *межгрупповой* и *внутригрупповой* дисперсии;

– *Множественные сравнения*;

– *Однородные подмножества*.

Убедитесь, что значения данных в таблицах аналогичны предыдущему способу анализа.

2 Многофакторный дисперсионный анализ.

Выберите в меню команду *Анализ → Общая линейная модель → ОЛМ-одномерная*.

Переместите зависимую и независимые переменные в соответствующие об-

ласти. Щелкните по кнопке *Модель*. В открывшемся диалоговом окне можно задать исследование либо всех возможных взаимодействий между независимыми переменными (*Полная факторная*), либо только каких-то конкретных взаимодействий (*Настраиваемая*). По умолчанию установлена полнофакторная модель. В этом же диалоговом окне можно задать тип формирования *Сумм квадратов* для метода наименьших квадратов (по умолчанию – *тип III*), но для большинства задач маркетинговых исследований достаточно оставлять все эти значения неизменными. Иными словами, кнопкой *Модель* главного диалогового окна *ОЛМ-одномерная* можно не пользоваться.

То же самое касается и кнопки *Контрасты*, а также кнопки *Сохранить*, позволяющей сохранять некоторые значения. В большинстве практических случаев, встречающихся в маркетинговых исследованиях, при проведении дисперсионного анализа вам не потребуется ничего сохранять.

В диалоговом окне *Апостериорные* следует добавить к списку исследуемых переменных (*Апостериорные критерии для*) все независимые *Факторы*, кроме дихотомических. Активируйте тест *Шеффе* для равных дисперсий.

В диалоговом окне *Параметры* в поле *Показать средние значения для* перенесите переменные, для которых необходимо вывести значения среднего и стандартной ошибки. Можно также вывести эти значения для совокупной выборки (переменная *OVERALL – В ЦЕЛОМ*) и для взаимодействия переменных (обозначается их произведением с помощью знака «*»). Затем активируйте опции *Описательные статистики* и *Критерии однородности*.

После этого можно запускать процедуру дисперсионного анализа на выполнение. В окне *Вывод* будут выведены результаты расчетов. Полученные таблицы в целом аналогичны заданию 1. Отличием является присутствие в некоторых таблицах многослойного представления результатов для категорий, основанных одновременно на двух исследуемых факторах, а также расчет статистических показателей для взаимодействия факторов (если оно было указано в качестве объекта анализа). Кроме того, после таблицы *Оценка эффектов межгрупповых факторов* будут выведены таблицы *Оцененные маргинальные* (*предельные*) *средние*, содержащие описательные статистики для совокупной выборки, для отдельных слоев факторов, для взаимодействия факторов (в зависимости от выбора в диалоговом окне *Параметры*).

Проверку значимости результатов многофакторного анализа начинают с оценки значимости полного эффекта (*Скорректированная модель*). Если полный эффект статистически значимый, то на следующем этапе изучают значимость эффекта взаимодействия переменных (обозначается их произведением с помощью знака «*»). Проверка значимости эффекта для каждого отдельного фактора осуществляется только в том случае, если эффект взаимодействия является статистически незначимым.

3 Ковариационный анализ.

Ковариационный анализ является разновидностью дисперсионного и выполняется в том случае, если хотя бы одна из независимых переменных относится к интервальной шкале или к шкале отношений (метрической). Такая переменная называется ковариткой.

Процедура выполнения ковариационного анализа аналогична процедуре многофакторного дисперсионного анализа. Единственным отличием является размещение переменной – ковариаты в главном диалоговом окне дисперсионного анализа *ОЛМ-одномерная*. Эта переменная переносится в поле *Ковариаты*.

5 Корреляционно-регрессионный анализ данных

Цель работы: выполнить корреляционно-регрессионный анализ данных с применением SPSS.

Задачи работы:

– изучить методику выполнения корреляционно-регрессионного анализа в SPSS;

– провести однофакторный корреляционно-регрессионный анализ данных с применением SPSS, записать соответствующую модель, оценить статистическую значимость и интерпретировать полученные результаты;

– провести многофакторный корреляционно-регрессионный анализ данных с применением SPSS, записать соответствующую модель, оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных выполните:

– однофакторный корреляционно-регрессионный анализ для заданных переменных;

– многофакторный корреляционно-регрессионный анализ для заданных переменных.

Методические указания

1 Однофакторный корреляционно-регрессионный анализ.

Исследование линейных корреляций по Пирсону, Спирману и Кендаллу.

Выберите в меню команду *Анализ* → *Корреляции* → *Парные*. В открывшемся диалоговом окне выберите в левом списке всех доступных переменных две исследуемые и перенесите их в область тестируемых *Переменных*.

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

1) для переменных с интервальной шкалой применяется коэффициент корреляции *Пирсона* (соответствующий флажок в области *Коэффициентов корреляции*), который позволяет охарактеризовать линейную связь между двумя переменными по указанным параметрам: наличию (есть/нет), направлению (убывает/возрастает) и силе (очень слабая/слабая/умеренная/сильная). Если переменная является квазипорядковой[1], ее до начала корреляционного анализа

[1] Квазипорядковые переменные, хотя и закодированы как порядковые, по сути являются интервальными

надо перекодировать (см. лабораторную работу № 2);

2) если хотя бы одна из пары исследуемых переменных имеет порядковую или дихотомическую шкалу, используются ранговые коэффициенты корреляции *Спирмана* или *Tay-b Кендалла*. Чаще всего эти коэффициенты применяются в случаях, когда необходимо установить степень соответствия двух ранжированных списков.

Отмечать можно только один из коэффициентов. Остальные параметры в этом диалоговом окне, установленные по умолчанию, рекомендуется оставить неизменными: вывод статистической значимости коэффициентов (параметр *Двухстороняя* в области *Критерий значимости*) и *Метить значимые корреляции*. Кнопка *Параметры* не предлагает исследователю каких-либо существенных опций.

Чтобы запустить процедуру построения корреляционной таблицы, щелкните по кнопке *ОК*. В таблице *Корреляции* будут показаны коэффициенты корреляции, количество использованных пар значений переменных (*N*) и вероятность ошибки, соответствующая предположению о ненулевой корреляции (*Знч.* (*2-сторон*)). Для интерпретации полученных данных воспользуйтесь информацией таблиц 3 и 4 лабораторной работы № 3.

Частные корреляции. Выявление ложных корреляций.

На практике иногда возникают ситуации, когда в результате корреляционного анализа обнаруживаются логически необъяснимые, противоречащие объективному опыту исследователя корреляции между двумя переменными. В этом случае говорят о так называемой ложной корреляции, когда статистически значимый коэффициент корреляции является не проявлением некоторой причинной связи между двумя рассматриваемыми переменными, а в большей степени обусловлен некоторой третьей переменной. Исследовать такую ситуацию помогают частные коэффициенты корреляции.

Выберите в меню команду *Анализ → Корреляции → Частные*. В левом списке всех доступных переменных выберите переменные, между которыми обнаружена «странная» корреляция, и поместите их в область тестируемых *Переменных*. Переменную, с которой коррелируют обе исследуемые переменные, поместите в область *Исключаемые*. В этом диалоговом окне больше ничего не изменяйте – запустите программу на исполнение, щелкнув по кнопке *ОК*.

В окне вывода *Viewer* появятся результаты расчетов частных коэффициентов корреляции. В полученной таблице первая строка каждой ячейки содержит коэффициент корреляции Пирсона, а вторая – статистическую значимость данного коэффициента.

Регрессионный анализ.

Для определения формы зависимости между двумя переменными используется поле корреляции (диаграмма рассеяния).

Выберите в меню команду *Графика → Панель выбора диаграмм*. В открывшемся диалоговом окне в списке доступных переменных выделите сначала неза-

---

(например, заданные в виде интервалов значения возраста или дохода респондента).

висимую переменную, а затем, удерживая клавишу Ctrl, зависимую. Среди появившихся в правой части диалогового окна видов диаграмм выберите *Диаграмму рассеяния*. Щелкните по кнопке *ОК*.

На появившемся в окне вывода *Viewer* рисунке изучите расположение точек. Если они сгруппированы (с незначительным случайным разбросом) вокруг прямой линии, направленной вправо и вверх (вниз), следовательно, к изучаемым данным подходит линейная парная регрессионная модель.

Уравнение линейной регрессии задается формулой

$$y = b \cdot x + a, \qquad (1)$$

где b – регрессионный коэффициент;

a – смещение по оси ординат.

Для определения уравнения регрессии выберите в меню команду *Анализ → Регрессия → Линейная*. В открывшемся диалоговом окне перенесите *зависимую* и *независимую* переменные в соответствующие поля. Ничего больше не меняя, начните расчет нажатием кнопки *ОК*. Вывод основных результатов включает в себя таблицы:

– *Сводка по модели* – содержит сведения о построенной модели.

Коэффициент *R,* изменяющийся в пределах от 0 до 1, является характеристикой силы общей линейной связи между переменными в регрессионной модели. Чем он выше, тем лучше выбранная независимая переменная подходит для определения поведения зависимой переменной. Требования к коэффициенту *R* такие же, как к коэффициенту корреляции: в общем случае он должен превышать хотя бы 0,5.

Коэффициент детерминации *R-квадрат* показывает, какая доля совокупной вариации в зависимой переменной описывается выбранной независимой переменной. Величина *R-квадрат* изменяется от 0 до 1. Как правило, данный показатель должен превышать 0,5 (чем он выше, тем показательнее построенная регрессионная модель, т. е. больше степень соответствия между регрессионной моделью и исходными данными).

Величина показателя *Скорректированный R-квадрат* всегда меньше, чем нескорректированного.

Величина *Стандартной ошибки оценки* определяет качество регрессионной модели. Данный показатель варьируется в пределах от 0 до 1. Чем он меньше, тем надежнее модель (в общем случае показатель должен быть меньше 0,5);

– *Дисперсионный анализ* – отражает два источника дисперсии: дисперсию, которая описывается уравнением регрессии (*Сумма квадратов*, обусловленная *Регрессией*), и дисперсию, которая не учитывается при записи уравнения (*Остаток*). Частное от суммы квадратов, обусловленных регрессией, и общей суммы квадратов является коэффициентом *R-квадрат*.

В таблице также содержатся число степеней свободы (*Ст. св.*) и среднее значение квадрата (*Средний квадрат*).

Существование ненулевых коэффициентов регрессии проверяется посред-

ством вычисления контрольной величины *F*, к которой относится соответствующий уровень значимости (*Знч.*).

На основании таблиц *Дисперсионный анализ* и *Сводка по модели* можно судить о статистической значимости и практической пригодности построенной регрессионной модели;

– *Коэффициенты* – отражает коэффициент регрессии *b* и смещение по оси ординат *a* под именем *Константа* (в столбце *Нестандартизированные коэффициенты B*). Частное рассчитанных коэффициентов и их стандартных ошибок (*Стд. ошибка*) дают контрольную величину *t*; соответственный уровень значимости (*Знч.*) относится к существованию ненулевых коэффициентов регрессии. Значение *Стандартизированного коэффициента Бета* будет рассмотрено при изучении многофакторного корреляционно-регрессионного анализа.

После получения уравнения регрессии, можно построить регрессионную прямую. Выберите в меню команду *Анализ → Регрессия → Подгонка кривых*. В открывшемся диалоговом окне перенесите *зависимую* и *независимую* переменные в соответствующие поля. Ничего больше не меняя, нажмите кнопку *ОК*.

В окне *Вывода* появятся три информационные таблицы, таблица *Сводка модели и оценки параметров* (сравните данные в ней с предыдущими таблицами), а также диаграмма рассеяния с нанесенной линией линейной регрессии.

На диаграмму можно вывести уравнение регрессии и коэффициент детерминации. Для этого дважды щелкните по диаграмме; откроется окно *Редактора диаграмм*. Дважды щелкните по линии регрессии, в открывшемся окне установите флажок *Назначить кривой метку*. Нажмите кнопку *Применить*, а затем *Закрыть*.

В меню *Элементы* редактора диаграмм выберите *Линия аппроксимации для итога*, убедитесь, что в области *Метод аппроксимации* выбрана *Линейная регрессия*, и нажмите кнопку *Закрыть*.

После закрытия окна *Редактора диаграмм* на диаграмме в окне *Вывода* появится заданная информация.

Полученное уравнение регрессии может использоваться для прогнозирования значения зависимой переменной при различных значениях независимой переменной. Выберите в меню команду *Анализ → Регрессия → Линейная*. При необходимости перенесите заданные переменные в соответствующие поля. Нажмите кнопку *Сохранить*. В открывшемся диалоговом окне в области *Предсказанные значения* установите флажок *Нестандартизированные*. Щелкните *Продолжить*, а затем *ОК*.

В результате в файле данных появится переменная *PRE_N*, где *N* – порядковый номер регрессионной модели (если регрессионный анализ выполнялся несколько раз). Ее значения рассчитаны по уравнению регрессии для заданных значений независимой переменной. Чтобы спрогнозировать значение зависимой переменной для произвольного значения независимой переменной, добавьте в файл данных нового респондента, установите ему требуемое значение независимой переменной и повторите расчет.

2 Многофакторный корреляционно-регрессионный анализ.

Корреляционный анализ.

Процедура выполнения многофакторного корреляционного анализа (расчета коэффициентов Пирсона, Спирмана и Кендалла) аналогична процедуре однофакторного корреляционного анализа.

Регрессионный анализ.

В случае многофакторного (множественного) регрессионного анализа необходимо оценить коэффициенты уравнения

$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + \ \ldots \ + b_n \cdot x_n + a\,, \qquad (2)$$

где n – количество независимых переменных, обозначенных как x.

Выберите в меню команду *Анализ* → *Регрессия* → *Линейная*. В открывшемся диалоговом окне перенесите *зависимую* и *независимые* переменные в соответствующие поля.

Если стоит задача провести регрессионный анализ в разрезе значений некоторой переменной, выберите ее в левом списке и перенесите в область *Переменная отбора наблюдений*. Затем щелкните по кнопке *Правило*, чтобы задать конкретное значение данной переменной для регрессионного анализа. Следует отметить, что за одну итерацию можно построить регрессию только в разрезе какого-то одного значения. В дальнейшем следует повторить все этапы сначала по количеству значений этой переменной.

Если нет необходимости проводить регрессионный анализ в каком-либо разрезе, оставьте поле *Переменная отбора наблюдений* пустым.

Существует несколько *Методов* проведения регрессионного анализа (см. соответствующий список). Для множественного анализа с несколькими независимыми переменными не рекомендуется оставлять метод *Принудительного включения* всех переменных, установленный по умолчанию. Этот метод соответствует одновременной обработке всех независимых переменных, выбранных для анализа, и поэтому используется только в случае простого анализа с одной независимой переменной. Для множественного анализа следует выбрать один из пошаговых методов.

При прямом методе (*Включение*) независимые переменные, которые имеют наибольшие коэффициенты частичной корреляции с зависимой переменной, пошагово увязываются в регрессионное уравнение.

При обратном методе (*Исключение*) начинают с результата, содержащего все независимые переменные, и затем исключают независимые переменные с наименьшими частичными корреляционными коэффициентами, пока соответствующий регрессионный коэффициент не оказывается незначимым.

Пошаговый метод (*Пошаговый отбор*) устроен так же, как и прямой, однако после каждого шага переменные, используемые в данный момент, исследуются по обратному методу. При пошаговом методе могут задаваться блоки независимых переменных; в этом случае заданные блоки на одном шаге обрабатываются совместно.

Последним шагом перед запуском процедуры построения регрессионной модели является выбор пункта *Диагностики коллинеарности* в диалоговом окне, появляющемся при нажатии на кнопку *Статистики*. Установление требования

провести диагностику наличия коллинеарности между независимыми переменными позволяет избежать эффекта мультиколлинеарности, при котором несколько независимых переменных могут иметь настолько сильную корреляцию, что в регрессионной модели обозначают, в принципе, одно и то же (это неприемлемо).

Щелчок по кнопке *ОК* в главном окне *Линейная регрессия* приведет к запуску процедуры построения линейной регрессии.

Необходимо отметить, что все таблицы, представленные в отчете, содержат несколько блоков, соответствующих количеству шагов SPSS при построении модели. При интерпретации результатов регрессионного анализа следует обращать внимание только на последний блок. В примечаниях к таблицам поясняется состав влияющих переменных (*Предикторы*).

Интерпретация полученных таблиц в целом аналогична однофакторному регрессионному анализу. Пояснения здесь требуются, начиная с изменившейся таблицы *Коэффициенты*.

Прежде всего, необходимо исключить возможность возникновения ситуации мультиколлинеарности. Для этого следует в таблице *Коэффициенты* изучить значения *КРД* (коэффициент разбухания дисперсии), соответствующие каждой независимой переменной. Если величина данного показателя меньше 10 (в некоторых литературных источниках указывается в качестве критического значения 3), то эффекта мультиколлинеарности не наблюдается и регрессионная модель приемлема для дальнейшей интерпретации. Чем выше этот показатель, тем более связаны между собой переменные. Если какая-либо переменная превышает значение 10, следует пересчитать регрессию без этой независимой переменной. В данном случае автоматически уменьшится величина *R-квадрат* и возрастет величина свободного члена (константы), однако, несмотря на это, новая регрессионная модель будет более практически приемлема, чем первая.

Аналогичную роль выполняет статистика *Толерантность*. Она характеризует процент дисперсии данной переменной, который не может быть объяснен другими переменными. Значение показателя *Толерантности* более 0,3 говорит об отсутствии мультиколлинеарности.

Таблица *Диагностики коллинеарности* также позволяет оценить взаимообусловленность независимых переменных. *Собственные значения*, близкие к 0, и *Показатель обусловленности* более 15 свидетельствуют о проблемах коллинеарности.

Уравнение множественной регрессии составляется на основании таблицы *Коэффициенты*. В ее первом столбце содержатся независимые переменные, удовлетворяющие требованию статистической значимости[1]. Можно определить направление и силу влияния на зависимую переменную каждой независимой. Это позволяет сделать столбец *Бета*, содержащий *Стандартизированные* β-*коэффициенты* регрессии. Данные коэффициенты дают возможность сравнить между собой значимость независимых переменных (например, составить рейтинг их влияния). Знак («+» или «–») перед β-коэффициентом показывает

[1] *Исключенные переменные* (при использовании шаговых методов регрессионного анализа) содержатся в соответствующей таблице.

направление связи между независимой и зависимой переменными, а модуль его величины – силу влияния.

Прогнозирование по уравнению множественной регрессии аналогично однофакторному регрессионному анализу.

6 Факторный анализ данных

Цель работы: выполнить факторный анализ данных с применением SPSS.

Задачи работы:

– изучить методику выполнения факторного анализа в SPSS;

– провести факторный анализ данных с применением SPSS;

– оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных выполните факторный анализ для заданных переменных.

Методические указания

Выберите в меню *Анализ → Снижение размерности → Факторный анализ*. В открывшемся диалоговом окне перенесите переменные для анализа из левого списка в поле *Переменные*. Поле *Переменная отбора наблюдений* позволяет выбрать переменную, в разрезе которой будет проводиться анализ. Оставьте это поле пустым.

Щелкните по кнопке *Описательные*. В открывшемся диалоговом окне в области *Статистики* оставьте установленную по умолчанию опцию вывода *Начальное решение*, результаты которого включают в себя первичные относительные дисперсии простых факторов, собственные значения и процентные доли объясненной дисперсии. В области *Корреляционная матрица* выберите пункты *Коэффициенты* (задает вывод исходной корреляционной матрицы), *Воспроизведенная* (вывод вычисленной корреляционной матрицы для анализа остатков и качества модели), а также *КМО и критерий сферичности Бартлетта* (вывод соответствующих тестов, позволяющих определить, насколько имеющиеся данные пригодны для факторного анализа). Окно *Факторный анализ*: *Описательные* позволяет вывести и другие необходимые статистики. Однако в большинстве случаев маркетинговых исследований эти возможности не используются. Закройте это окно, щелкнув по кнопке *Продолжить*.

Откройте окно *Извлечение*, которое предназначено для выбора метода формирования факторной модели. В списке *Метод* выберите метод извлечения (формирования) факторов. Общая рекомендация по выбору метода состоит в следующем. Необходимо выбирать тот метод извлечения факторов, который позволяет однозначно классифицировать как можно больше переменных. Таким

образом, основные соображения здесь – число классифицированных факторов и однозначность классификации (т. е. каждая переменная должна принадлежать только одному фактору). Хорошим результатом факторного анализа является доля однозначно классифицированных переменных не менее 90 %. Выберите метод *Главные компоненты*, который является наиболее подходящим для решения большинства задач маркетинговых исследований при помощи факторного анализа.

В области *Анализ* оставьте неизменной выбранную по умолчанию *Матрицу корреляций*.

В области *Вывести* (или *Отобразить*) оставьте неизменным выбранный по умолчанию вывод *Неповернутых решений*. Активируйте также опцию *График собственных значений*, что позволит графически представить собственные значения факторов, упорядоченные по величине.

В области *Выделить* (или *Извлечь*) укажите количество образуемых факторов. По умолчанию установлен метод его определения *Основываясь на собственном значении*, который используется SPSS для установления количественного и качественного состава извлекаемых факторов. При предустановленном значении данного показателя, равном 1, количество образуемых факторов будет равно количеству переменных, значение характеристических чисел для которых больше или равно 1.

Также можно вручную указать, какое *Фиксированное количество факторов* необходимо извлекать. Эта возможность предусмотрена в SPSS для того, чтобы при слишком большом количестве переменных с характеристическим числом больше 1 вручную сократить число факторов. Большое число факторов трудно интерпретировать, поэтому, если автоматически не удается извлечь приемлемое число факторов (чем меньше, тем лучше), следует самостоятельно указать его. Необходимо отметить, что при этом иногда количество однозначно классифицированных переменных оказывается меньше, чем при методе характеристических чисел. Однако данный негативный момент нивелируется возросшей наглядностью результатов факторного анализа – ведь это позволяет освободиться от факторов, в которых нет переменных со значимым коэффициентом корреляции.

Определите количество факторов в соответствии с заданием. Подтвердите произведенные установки нажатием кнопки *Продолжить*.

Щелкните по кнопке *Вращение*. Вращение коэффициентной матрицы производится для того, чтобы максимально приблизить факторную модель к идеалу – возможности однозначно классифицировать все переменные. В поле *Метод* выберите конкретный метод вращения. В большинстве случаев наиболее приемлемым вариантом является метод *Варимакс*. Он облегчает интерпретацию факторов, минимизируя количество переменных с высокими факторными нагрузками. Наряду с выводом *Повернутого решения*, установленным по умолчанию, можно получить факторные нагрузки в графическом виде, активируя опцию *График*(*и*) *нагрузок* (в данной работе это делать необязательно). Закройте диалоговое окно, щелкнув по кнопке *Продолжить*.

Щелкните по кнопке *Значения факторов*. Это диалоговое окно служит для создания в исходном файле данных новых переменных, которые в дальнейшем

позволят отнести каждого респондента к определенной группе (фактору). Активируйте параметр *Сохранить как переменные*, а в качестве метода определения значений для этих новых переменных – метод *Регрессии*. Подтвердите произведенные установки нажатием кнопки *Продолжить*.

Последним этапом перед запуском процедуры факторного анализа является выбор некоторых дополнительных *Параметров*. Нажмите на соответствующую кнопку. В открывшемся диалоговом окне сохраните выбор по умолчанию в поле *Пропущенные значения*. В поле *Формат вывода коэффициентов* активируйте параметр *Отсортировать по величине*. Это позволит вывести переменные, входящие в каждый фактор, в порядке убывания их факторных коэффициентов (величины вклада переменной в формирование фактора).

Также можно активировать параметр *Не выводить коэффициенты с низкими значениями*, что облегчит задачу однозначной интерпретации полученных факторов. Указанное в соответствующем поле значение данного параметра отсекает переменные с факторными коэффициентами менее данного значения. Это позволяет упростить повернутую матрицу факторов, поскольку из нее исчезают незначимые переменные, входящие в каждый извлеченный фактор. Если Вы не задействуете данный параметр, для каждой переменной будет отображен факторный коэффициент по каждому фактору, что излишне перегрузит факторную модель и затруднит ее восприятие.

Данный параметр вводится, чтобы облегчить практическую интерпретацию результатов факторного анализа. Так как факторные коэффициенты в результирующей повернутой матрице коэффициентов являются коэффициентами корреляции между соответствующими переменными и факторами, в большинстве практических случаев целесообразно устанавливать начальное значение отсечения незначимых переменных на уровне 0,5. Если в результате факторного анализа окажется, что число классифицированных переменных менее приемлемого, можно пересчитать факторную модель с меньшим значением отсечения (например, 0,4). В обратной ситуации, если переменная входит в несколько факторов, можно предложить повысить уровень извлечения с 0,5 до 0,6. Это позволит устранить переменные, входящие сразу в несколько факторов, увеличив практическую пригодность результатов факторного анализа.

Для запуска процедуры факторного анализа подтвердите произведенные установки нажатием кнопки *Продолжить* и в главном диалоговом окне – кнопки *ОК*. После того как программа произведет все необходимые расчеты, откроется окно *Вывода* с результатами построения факторной модели.

Первой таблицей является *Матрица корреляций*, которая позволяет сделать вывод о наличии / отсутствии взаимосвязи между отдельными переменными.

Далее необходимо оценить пригодность имеющихся данных для факторного анализа в целом. В таблице *Мера адекватности и критерий Бартлетта* отображены соответствующие показатели: тест *адекватности выборки Кайзера – Мейера – Олкина* (КМО) и значимость теста *сферичности Бартлетта*. Результаты теста КМО позволяют сделать вывод относительно общей пригодности имеющихся данных для факторного анализа, т. е. насколько хорошо построенная факторная модель описывает структуру ответов респондентов на

анализируемые вопросы. Значение теста варьируется в интервале от 0 (факторная модель абсолютно неприменима) до 1 (факторная модель идеально описывает структуру данных). Факторный анализ следует считать пригодным, если КМО больше 0,5.

Тест Бартлетта проверяет гипотезу о том, что переменные, участвующие в факторном анализе, некоррелированы между собой. Если данный тест дает положительный результат (переменные некоррелированы), факторный анализ следует признать непригодным и использовать другие статистические методы (например, кластерный анализ). Статистикой, определяющей пригодность факторного анализа по тесту Бартлетта, является значимость (*Знч.*). При приемлемом уровне значимости (ниже 0,05) факторный анализ считается пригодным для анализа исследуемой выборочной совокупности.

Если Вы пришли к выводу, что имеющиеся данные подходят для исследования при помощи факторного анализа, то можно приступать к изучению и интепретации полученных результатов.

Таблица *Общности* отражает долю вариации (дисперсии) каждой переменной, обусловленной совокупным влиянием факторов. Общность можно сравнить с множественным коэффициентом корреляции, принимающим значение 0 в случае, если факторы не влияют на переменную, и значение 1 в случае, если дисперсия переменной целиком определяется выделяемыми факторами. Перед началом извлечения факторов величина общности, равная 1, установлена по умолчанию для всех переменных, участвующих в факторном анализе.

Исходя из данных таблицы *Полная объясненная дисперсия*, определяют количество факторов, подлежащих извлечению[1] (если только оно не задано заранее). Аналогичную функцию выполняет *График нормализованного простого стресса*, точка перегиба которого указывает на действительное число факторов.

Таблица *Матрица компонент* отражает корреляции между переменными и выделенными факторами (факторные нагрузки). Однако она редко приводит к факторам, которые можно однозначно интерпретировать, поскольку последние коррелируют со многими переменными.

Поэтому факторы подвергаются вращению, и основным итогом факторного анализа является *Матрица повернутых компонент* (таблицу *Воспроизведенные корреляции* временно пропускаем). В ней отражаются результаты классификации переменных по факторам (компонентам). Переменная относится к тому фактору, в столбце которого находится максимальное абсолютное значение факторной нагрузки. Поскольку в параметрах анализа была выбрана сортировка коэффициентов по величине, то переменные сгруппированы последовательно для каждой компоненты.

Если в данную таблицу вошли не все анализируемые переменные, можно поступить следующим образом. Необходимо просто пересчитать факторную модель, удалив в диалоговом окне *Факторный анализ*: *Параметры* ранее установленное значение отсечения. Будет построена факторная матрица, в которой вам предстоит самостоятельно определить принадлежность неклассифицированных

[1] См. пояснения для диалогового окна *Факторный анализ*: *Извлечение*.

переменных к тому или иному фактору на основании критерия наибольшего коэффициента корреляции между переменными и факторами.

Наиболее сложной задачей при проведении факторного анализа является интерпретация полученных факторов. Необходимо обнаружить смысловую связь переменных, отнесенных к каждому фактору, и дать факторам вербальное описание. Здесь не существует какого-либо универсального решения: в каждом конкретном случае аналитик использует имеющийся практический опыт для того, чтобы понять, почему факторная модель относит ту или иную переменную к данному конкретному фактору.

Иногда возникают случаи, когда переменная, отнесенная SPSS к конкретному фактору, логически никак не связана с остальными переменными, составляющими тот же фактор. Можно пересчитать факторную модель без отсечения незначимых коэффициентов и посмотреть, с каким еще фактором данная нелогичная переменная коррелирует практически с той же силой, как с фактором, к которому она была отнесена автоматически. Если же исследователь зашел в тупик и никакие средства не помогают объяснить принадлежность той или иной переменной к конкретному фактору, остается применить другую статистическую процедуру (например, кластерный анализ).

После того как все полученные факторы успешно интерпретированы, необходимо определить соответствие полученной модели исходным данным. Вернитесь к таблице *Воспроизведенные корреляции*. Если в ней содержится много (более 50 %) остатков с большими абсолютными значениями (более 0,05), то модель не обеспечивает хорошее соответствие данным и требует пересмотра.

Завершает факторный анализ *Матрица преобразования компонент*, которая содержит корреляции между выделенными факторами.

Далее рассмотрим, как можно использовать результаты факторного анализа для построения рейтингов. Вспомним о том, что факторные рейтинги (т. е. принадлежность каждого респондента к определенному фактору) были сохранены в исходном файле данных в виде новых переменных[1]. Эти переменные имеют имена типа *facX_Y*, где *X* – это номер фактора, а *Y* – порядковый номер факторной модели (если анализ выполнялся несколько раз).

Наиболее частый способ использования факторных рейтингов – это ранжирование и последующее разделение вновь созданных переменных, обозначающих извлеченные факторы, на четыре квартиля (25-процентных процентиля). Такой подход позволяет создать новые переменные с порядковой шкалой, описывающие четыре уровня каждого фактора (например: не согласен, скорее не согласен, скорее согласен, согласен).

Чтобы создать переменные, по которым далее будут группироваться респонденты, вызовите меню *Преобразовать → Ранжировать наблюдения*. В открывшемся диалоговом окне из левого списка выберите переменную, содержащую факторные рейтинги, и поместите ее в поле *Переменные*. Далее в области *Ранг 1 присвоить наблюдению* выберите пункт *С минимальным значением*.

[1] Параметр *Сохранить как переменные* в диалоговом окне *Факторный анализ*: *Значения факторов*.

Щелкните по кнопке *Типы рангов*, затем выберите *Типы*, отмените установленный по умолчанию параметр *Ранг* и вместо него выберите *N разбиение* с предустановленным числом групп, равным 4. Щелкните по кнопке *Продолжить* и затем в главном диалоговом окне – по кнопке *ОК*. Данная процедура создаст в файле данных новую переменную *nfacX_l*, распределяющую респондентов на четыре группы в зависимости от уровня фактора *X*.

Для повышения наглядности рекомендуется присвоить метки каждому из выделенных уровней; можно переименовать и саму переменную. Теперь можете проводить перекрестный анализ при помощи новой порядковой переменной, а также строить другие статистические модели, предусмотренные в SPSS.

7 Кластерный анализ данных

Цель работы: выполнить кластерный анализ данных с применением SPSS.

Задачи работы:

– изучить методику выполнения кластерного анализа в SPSS;

– провести кластерный анализ данных с применением SPSS;

– оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных:

– выполните иерархический кластерный анализ;

– получите значимые кластеры с использованием двухступенчатого кластерного анализа, самостоятельно сформировав набор переменных.

Методические указания

1 Иерархический кластерный анализ.

Выберите в меню команду *Анализ → Классификация → Иерархическая кластеризация*. В открывшемся диалоговом окне перенесите переменные, являющиеся критериями сегментирования, из левого списка в поле *Переменные*. Поле *Метить значениями* позволяет выбрать переменную для обозначения наблюдений.

Щелкните по кнопке *Статистики*. В открывшемся диалоговом окне наряду с выводом *Порядка агломерации* можно задать вывод таблицы *Принадлежности к кластерам*, в которой каждому наблюдению сопоставляется номер кластера. Если нет уверенности в *Одном решении* о количестве кластеров, следует задать *Диапазон решений*.

Однако данная таблица при достаточно большом количестве респондентов (практически во всех маркетинговых исследованиях) становится совершенно бесполезной, т. к. представляет собой длинную последовательность пар значений «номер респондента - номер кластера», в таком виде не поддающуюся

интерпретации. Технически цель кластерного анализа всегда состоит в образовании в файле данных дополнительной переменной, отражающей разделение наблюдений на целевые группы (при помощи щелчка на кнопке *Сохранить* в главном диалоговом окне кластерного анализа). Эта переменная в совокупности с номерами наблюдений и есть таблица *Принадлежность к кластерам*.

Закройте это окно, щелкнув по кнопке *Продолжить*.

Щелкните по кнопке *Графики*. Активируйте опцию вывода древовидной диаграммы – *Дендрограммы* и посредством опции *Нет* отмените вывод накопительной *Сосульчатой диаграммы*. Подтвердите произведенные установки нажатием кнопки *Продолжить*.

Щелкнув по кнопке *Метод*, следует выбрать метод образования кластеров, а также метод расчета дистанционной меры и меры преобразования. Эксперименты с данными параметрами позволяют добиться большей точности при определении оптимального числа кластеров.

Первое, что устанавливается в данном окне, – это *Метод* формирования кластеров, т. е. объединения наблюдений. Среди всех возможных вариантов статистических методик, предлагаемых SPSS, следует выбирать либо установленный по умолчанию метод *Межгрупповых связей*, либо *метод Варда*. Первый метод используется чаще ввиду его универсальности и относительной простоты статистической процедуры, на которой он основан. При использовании этого метода расстояние между кластерами вычисляется как среднее значение расстояний между всеми возможными парами наблюдений, причем в каждой итерации принимает участие одно наблюдение из одного кластера, а второе – из другого. Информация, необходимая для расчетов расстояния между наблюдениями, находится на основании всех теоретически возможных пар наблюдений. Метод Варда более сложен для понимания и используется реже. Он состоит из множества этапов и основан на усреднении значений всех переменных для каждого наблюдения с последующим суммированием квадратов расстояний от вычисленных средних до каждого наблюдения. Для решения практических задач маркетинговых исследований рекомендуется использовать метод межгрупповых связей.

После выбора статистической процедуры кластеризации следует выбрать *Меру* для вычисления расстояний между наблюдениями. Дистанционные меры зависят от типа переменной и шкалы, к которой она относится: *Интервальная* переменная; *Частоты* или номинальная переменная; *Бинарная* (дихотомическая) переменная. При этом дихотомическая шкала подразумевает только переменные, отражающие наступление/ненаступление какого-либо события (купил/не купил, да/нет и т. д.). Другие типы дихотомических переменных (например, мужчина/женщина) следует рассматривать и анализировать как номинальные. Выбор метода определения расстояний для дихотомических переменных предполагает указание конкретных значений, которые они могут принимать, в соответствующих полях *Наличие* и *Отсутствие*.

Наиболее часто используемой мерой определения расстояний для интервальных переменных является *Квадрат расстояния Евклида*, установленный по умолчанию. Именно эта мера зарекомендовала себя в маркетинговых исследованиях как наиболее точная и универсальная.

Однако для дихотомических переменных, где наблюдения представлены только двумя значениями (например, 0 и 1), данная мера не подходит. Дело в том, что она учитывает только взаимодействия между наблюдениями типа $X = 1$, $Y = 0$ и $X = 0$, $Y = 1$ (где X и Y – переменные) и не учитывает другие типы взаимодействий. Наиболее комплексной мерой расстояния, учитывающей все важные типы взаимодействий между дихотомическими переменными, является *Лямбда*.

Для переменных с номинальным типом шкалы SPSS предлагает всего две меры: *Хи-квадрат* и *Фи-квадрат*. Рекомендуется использовать первую из них как наиболее распространенную.

В диалоговом окне *Метод* есть также область *Преобразовать значения*, в которой находится поле *Стандартизация*. Данное поле применяется в том случае, когда в кластерном анализе принимают участие переменные с различным типом шкалы, разными измерителями или размерностями шкалы. Для того чтобы использовать эти переменные, следует провести стандартизацию, приводящую их к единому типу шкалы – интервальному. Самым распространенным методом стандартизации переменных является *Z-стандартизация*: все переменные приводятся к единому диапазону значений от -3 до +3 и после преобразования являются интервальными.

Так как все оптимальные методы (кластеризации и определения расстояний) установлены по умолчанию, целесообразно использовать диалоговое окно *Метод* только для указания типа анализируемых переменных, а также для указания необходимости произвести Z-стандартизацию переменных.

Щелкнув на кнопке *Продолжить*, вернитесь в главное диалоговое окно, в котором щелкните на кнопке *ОК*, чтобы запустить кластерный анализ.

В окне *Вывода* после обычной общей статистической сводки итогов по наблюдениям приводится таблица *Шаги агломерации*, из которой можно выяснить очерёдность построения кластеров, а также их оптимальное количество.

Необходимо отметить, что единого универсального метода определения оптимального числа кластеров не существует. Как правило, решающее значение имеет показатель, выводимый в графе *Коэффициенты*. Под коэффициентом подразумевается расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учетом предусмотренного преобразования значений. На том этапе, где мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить, так как в противном случае были бы объединены кластеры, находящиеся на относительно большом расстоянии друг от друга. Поэтому оптимальным считается число кластеров, равное разности количества наблюдений и количества этапов, после которого коэффициент увеличивается скачкообразно.

Далее приводится таблица *Принадлежность к кластеру* с информацией о принадлежности каждого наблюдения к определенному кластеру, если ее вывод был затребован.

В заключение приводится *Дендрограмма*, которая визуализирует процесс слияния, приведенный в таблице порядка агломерации. Она идентифицирует объединенные кластеры и значения коэффициентов на каждом шаге. При этом отображаются не исходные значения коэффициентов, а значения, приведенные

к шкале от 0 до 25. Кластеры, получающиеся в результате слияния, отображаются горизонтальными линиями. По дендрограмме можно сделать вывод об оптимальном количестве кластеров – оно определяется субъективно и равно числу горизонтальных линий справа от очередного слияния (вертикальной линии, проведенной через всю дендрограмму), после которого все величины коэффициентов (длины горизонтальных линий) становятся достаточно большими.

Определите оптимальное количество кластеров. После этого необходимо сохранить информацию о принадлежности каждого наблюдения к определенному кластеру. Для этого вновь откройте диалоговое окно *Иерархический кластерный анализ* и щелкните на кнопке *Сохранить*. Открывшееся диалоговое окно позволяет создать в исходном файле данных новую переменную, распределяющую наблюдения на полученные группы. Выберите параметр *Одно решение* и укажите в соответствующем поле необходимое количество кластеров. Запустите процедуру кластерного анализа. Новая переменная имеет имя типа *cluX_Y*, где *X* – это количество кластеров, *Y* – порядковый номер кластерной модели (если анализ выполнялся несколько раз).

Чтобы установить, насколько верно определено оптимальное число кластеров, постройте частотное распределение данной переменной с помощью команды *Анализ → Описательные статистики → Частоты*. Размер кластеров должен быть статистически значимым и практически приемлемым. При большом объеме выборки минимальный размер кластера следует установить хотя бы на уровне 10. Для небольшой выборки относительные размеры кластеров должны быть достаточно выразительными.

При необходимости снова проведите кластерный анализ, указав в диалоговом окне *Сохранить* число кластеров, состоящих из значимого количества наблюдений. Эта процедура должна повторяться до тех пор, пока не будет найдено решение, в котором все кластеры будут состоять из статистически значимого числа наблюдений[1].

Завершающий этап кластерного анализа – интерпретации кластеров. Для этого следует воспользоваться процедурой сравнения средних значений исследуемых переменных (кластерных центроидов).

Выберите в меню команду *Анализ → Сравнение средних → Средние*. В открывшемся диалоговом окне из левого списка выберите переменные, использовавшиеся в качестве критериев сегментирования, и перенесите их в *Список зависимых переменных*. Затем переменную, отражающую разделение наблюдений на кластеры, переместите из левого списка в *Список независимых переменных*.

После этого щелкните по кнопке *Параметры*. В поле *Статистики в ячейках* оставьте только вывод *Средних* значений и *Количества наблюдений*, удалив

[1] Необходимо отметить, что критерий практической и статистической значимости численности кластеров не является единственным критерием, по которому можно определить оптимальное число кластеров. Исследователь может самостоятельно, на основании имеющегося у него опыта предположить число кластеров (условие значимости должно удовлетворяться). Другим вариантом является довольно распространенная ситуация, когда в целях исследования заранее ставится условие сегментировать респондентов на заданное число целевых групп. В этом случае необходимо просто один раз провести иерархический кластерный анализ с сохранением требуемого числа кластеров и затем пытаться интерпретировать то, что получится.

другие установленные по умолчанию статистики. Закройте диалоговое окно щелчком на кнопке *Продолжить*. Из главного диалогового окна запустите процедуру сравнения средних значений (кнопка *ОК*).

В открывшемся окне *Вывода* появятся результаты работы статистической процедуры сравнения средних значений. Интерпретируйте кластеры на основании данных таблицы *Отчет*.

2 Двухступенчатый кластерный анализ.

Алгоритм, используемый данной процедурой, имеет несколько привлекательных особенностей, которые отличают его от традиционных методов кластерного анализа:

– работа одновременно с категориальными и метрическими (количественными) переменными;

– автоматический выбор числа кластеров;

– масштабируемость (формируется дерево свойств кластеров, которое является компактным представлением информации о наблюдениях).

Для проведения двухэтапного кластерного анализа выберите в меню команду *Анализ → Классификация → Двухэтапный кластерный анализ*. В открывшемся диалоговом окне перенесите переменные, являющиеся критериями сегментирования, из левого списка в поля *Категориальные переменные* и *Количественные переменные* в соответствии с их типом.

В области *Мера расстояния* задается способ вычисления сходства между кластерами. Следует иметь в виду, что *Евклидова мера* может быть использована, только когда все переменные являются количественными.

Область *Число кластеров* позволяет задать, как будет определяться количество групп. Задайте автоматическое определение, которое определит оптимальное число кластеров, используя критерий, заданный в области *Критерий кластеризации* (оставьте здесь установленный по умолчанию *Байесовский информационный критерий*). Дополнительно введите максимальное количество кластеров, которое должна рассмотреть процедура (например, 10). При необходимости можно зафиксировать число кластеров в решении (параметр *Задать*).

Группа *Количество количественных переменных* дает сводную информацию об установках, касающихся стандартизации таких переменных, заданных в диалоговом окне *Параметры*.

Щелкните по кнопке *Параметры.* Поскольку алгоритм кластеризации работает со стандартизованными количественными переменными, то все они должны быть оставлены в поле *Подлежат стандартизации*. Чтобы несколько сэкономить время и снизить вычислительные затраты, можно поместить переменные, которые уже стандартизованы (если таковые имеются), в поле *Считаются стандартизованными*. Остальные установки данного диалогового окна оставьте неизменными. Закройте окно щелчком на кнопке *Продолжить*.

Щелкните по кнопке *Вывод.* В поле *Вывод* предоставлены параметры для вывода таблиц результатов кластеризации. Оставьте установленный по умолчанию вывод *Диаграмм и таблиц в средстве просмотра моделей*. В поле *Поля оценивания* можно задать вычисление характеристик кластера для переменных, которые не использовались в создании кластера.

Задание вывода *Мобильных таблиц* целесообразно после принятия окончательного решения о составе переменных, участвующих в кластерном анализе. Одновременно следует активировать опцию *Создать переменную принадлежности к кластерам* (вида *TSC_X*, где *X* – это порядковый номер операции сохранения активного набора данных, выполненной этой процедурой в течение данного сеанса работы).

В поле *Файлы XML* можно задать возможность экспорта результатов кластерного анализа.

Подтвердите произведенные установки нажатием кнопки *Продолжить*. Из главного диалогового окна запустите процедуру анализа (кнопка *ОК*).

В открывшемся окне *Вывода* сначала (если был задан вывод мобильных таблиц) приводятся обобщенные данные процесса кластеризации, а также распределение наблюдений по кластерам и профили кластеров (центроиды для количественных переменных и частоты для категориальных), позволяющие их интерпретировать.

Однако значительно б*о*льшие возможности для интерпретации дает выводимый далее объект средства просмотра моделей. Чтобы получить информацию о кластерной модели, активизируйте его двойным щелчком.

Средство просмотра кластеров состоит из двух панелей: основной, находящейся слева, и дополнительной, находящейся справа. Имеется два основных представления (выбираются в нижней левой части окна в меню *Вид*):

1) *Сводка для модели* (по умолчанию) – включает непосредственно сводку по модели (алгоритм, входные поля и кластеры) и силуэтную меру связности и разделения кластеров с использованием затенения для индикации низкого, среднего и хорошего качества полученных результатов. Это дает возможность быстро понять, является ли качество разбиения на кластеры низким. В таком случае следует вернуться к диалоговому окну *Двухэтапный кластерный анализ*, чтобы скорректировать параметры для построения модели с целью получения более приемлемых результатов (например, исключить наименее значимые переменные);

2) *Кластеры* – содержит «сетку» кластеров по показателям, которая включает имена кластеров (метки и описания, которые можно вводить и редактировать), объемы (размеры) и профили (входные поля) каждого кластера. Общая важность показателей обозначается интенсивностью цвета фона ячейки: наиболее важный показатель является наиболее темным. По умолчанию показатели отсортированы по убыванию общей важности. Если поместить указатель мыши на ячейку показателя, то будет выведено его полное имя/метка и значение важности для этой ячейки. В зависимости от типа показателя и вида представления может быть выведена дополнительная информация. В панели представления (нижняя часть окна) можно выбрать различные способы вывода информации о кластерах.

В дополнительной панели доступны четыре вида представления (выбираются в нижней правой части окна в меню *Вид*):

1) *Важность предиктора* – показывает относительную важность каждого входного поля при оценивании модели;

2) *Размеры кластеров* (по умолчанию) – показывает круговую диаграмму, содержащую все кластеры;

3) *Распределение в ячейках* – выводит расширенную, более детальную диаграмму распределения данных для любой ячейки показателя, выбранной в таблице в представлении *Кластеры* в основной панели;

4) *Сравнение кластеров* – выводит сетку с показателями в строках и выбранными кластерами в столбцах. Этот вид представления помогает лучше понять, какие факторы формируют кластер. Он также позволяет увидеть различие между кластерами не только в сравнении со всеми данными, но и в сравнении между собой. Чтобы выбрать кластеры для вывода, щелкните по верху столбца кластера в основной панели в представлении *Кластеры*. Пользуйтесь клавишами *Ctrl* и *Shift* совместно с щелчком мышью для выбора или отмены выбора нескольких кластеров для сравнения.

Добейтесь хорошего качества модели, исключая наименее важные предикторы (переменные) или изменяя количество кластеров. При этом кластеры не должны быть сведены к простой комбинации различных категорий номинальных и порядковых переменных. Интерпретируйте полученные кластеры.

8 Дискриминантный анализ данных

Цель работы: выполнить дискриминантный анализ данных с применением SPSS.

Задачи работы:

– изучить методику выполнения дискриминантного анализа в SPSS;

– провести дискриминантный анализ данных с применением SPSS;

– оценить статистическую значимость и интерпретировать полученные результаты.

Задание

На основании исходных данных выполните дискриминантный анализ по заданной проблеме исследования.

Методические указания

Выберите в меню команду *Анализ → Классификация → Дискриминантный анализ*. В открывшемся диалоговом окне перенесите зависимую переменную, разделяющую совокупность объектов исследования на группы, из левого списка в поле *Группировать по*. После щелчка по кнопке *Задать диапазон* введите *минимальное* и *максимальное* значения этой переменной. В поле *Независимые* перенесите соответствующие переменные. Оставьте установленный по умолчанию метод *Принудительное включение*, при котором в анализе одновременно будут участвовать все независимые переменные.

Щелкнув по кнопке *Статистики*, активируйте опции *Средние*,

Однофакторный дисперсионный анализ, *Нестандартизированные коэффициенты функции*, *Матрица внутригрупповой корреляции.*

Щелкнув по кнопке *Классифицировать*, активируйте вывод *Итоговой таблицы*. Остальные установки оставьте по умолчанию.

Щелкнув по кнопке *Сохранить*, активируйте сохранение в дополнительных переменных *Предсказанной принадлежности к группе* и *Дискриминантных баллов*. Эти переменные имеют имена типа *dis_Y* и *dis1_Y*, где *Y* – порядковый номер дискриминантной модели (если анализ выполнялся несколько раз).

Начните расчёт нажатием кнопки *ОК* в главном диалоговом окне.

После вводного обзора валидных и исключенных значений (таблица *Сводка результатов обработки наблюдений*) приводятся средние значения, стандартные отклонения, количество наблюдений для каждой группы в отдельности и суммарные показатели для обеих групп (таблица *Групповые статистики*). При наличии среди независимых переменных дихотомических, закодированных при помощи 0 и 1, среднее значение указывает на долю наблюдений с кодировкой 1 (при кодировке с помощью 1 и 2 среднее значение необходимо уменьшить на единицу).

Первой важной для анализа таблицей является *Критерий равенства групповых средних*. Она показывает, насколько значимо каждая независимая переменная разделяет наблюдения на исследуемые группы (тестовой величиной служит Лямбда Уилкса).

По таблице *Объединенные внутригрупповые матрицы* можно определить, являются ли предикторы независимыми, сопоставив коэффициенты корреляции. При наличии мультиколлинеарности между предикторами (коэффициентов корреляции, превышающих 0,5) не существует однозначной меры относительной важности независимых переменных для дискриминации между группами.

Таблица *Собственные значения* позволяет оценить общее качество разделения наблюдений на заданные группы зависимой переменной. Соответствующий вывод можно сделать исходя из коэффициента канонической корреляции (характеризует взаимосвязь между рассчитанными значениями дискриминантной функции и показателем принадлежности к группе). Еще одним важным показателем в этой таблице является собственное значение дискриминантной функции. В общем случае большая его величина указывает на высокую точность подобранной дискриминантной функции.

На основании таблицы *Лямбда Уилкса* также можно оценить качество приближения дискриминантной модели. Данный тест определяет значимость различий между средними значениями дискриминантной функции в двух исследуемых группах зависимой переменной.

Следующая таблица *Нормированные коэффициенты канонической дискриминантной функции* позволяет оценить относительную важность отдельных независимых переменных. Как правило, переменные с относительно большими нормированными коэффициентами вносят больший вклад в дискриминирующую мощность функции.

Аналогичный вывод можно сделать и по таблице *Структурная матрица*, которая содержит коэффициенты корреляции между каждой из переменных и

дискриминантной функцией (переменные при этом упорядочены по их дискриминирующей силе).

Далее следуют нестандартизированные (ненормированные) *Коэффициенты канонической дискриминантной функции*, на основании которых строится дискриминантное уравнение вида

$$D = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ \ldots \ + b_m \cdot X_m, \qquad (3)$$

где D – группирующая (зависимая) переменная;

b_m – коэффициенты дискриминантной функции;

b_0 – свободный член (константа);

X_m – дискриминационные (независимые) переменные (предикторы).

На основании этого уравнения, зная характеристики объекта исследования, можно с определенной степенью уверенности определить его принадлежность к одной из исследуемых групп.

Таблица *Функции в центроидах групп* показывает средние значения дискриминантной функции в каждой анализируемой группе зависимой переменной (центроиды). По ней рассчитывается константа дискриминации (граница, разделяющая два множества) как среднее арифметическое центроидов.

Завершает вывод результатов дискриминантного анализа таблица *Результаты классификации*, в примечании к которой содержится информация о точности построенной модели. Она должна составлять не менее 50 % (в некоторых источниках – 75 %). При низкой точности необходимо исключить из модели незначимые независимые переменные и повторить дискриминантный анализ.

При выполнении дискриминантного анализа, как и для других многомерных процедур, можно применять и пошаговый метод формирования дискриминантной функции, который рекомендуется при наличии большого количества независимых переменных. Эта процедура задается в главном диалоговом окне дискриминантного анализа посредством активации опции *Шаговый отбор*. Выполните анализ с применением данного метода и сравните результаты с полученными при использовании *Принудительного включения*.

Список литературы

1 **Малхотра, Н. К.** Маркетинговые исследования. Практическое руководство : пер. с англ. / Н. К. Малхотра. – 4-е изд. – Москва: Вильямс, 2007. – 1200 с.

2 **Моосмюллер, Г.** Маркетинговые исследования с SPSS : учебное пособие / Г. Моосмюллер, Н. Н. Ребик. – 2-е изд. – Москва : ИНФРА-М, 2021. – 200 с.