МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение
высшего образования
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ
УНИВЕРСИТЕТ» (РГГУ)

Институт социально-экономических наук

Факультет маркетинга и рекламы

Ю. Ю. Шитова

АНАЛИЗ МАРКЕТИНГОВОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ ПРОГРАММЫ SPSS

Учебное пособие

Часть вторая

Научное электронное издание

Чебоксары
Издательский дом «Среда»
2025

УДК 339.138
ББК 65.291.21я73
Ш64

Рецензенты:
д-р экон. наук, профессор
ФГАОУ ВО «Южный федеральный университет»
Радина Оксана Ивановна;
канд. экон. наук, доцент
ФГБОУ ВО «Чувашский государственный
университет им. И.Н. Ульянова»
Толстова Мария Леонидовна
Автор:
Шитова Ю.Ю. – д-р экон. наук, профессор кафедры интегрированных коммуникаций и рекламы Института социально-экономических наук РГГУ.

Шитова Ю. Ю.

Ш64 **Анализ маркетинговой информации с помощью программы SPSS :** учебное пособие / Ю. Ю. Шитова; Российский государственный гуманитарный университет. – Ч. 2. – Чебоксары: Среда, 2025. – 183 с. – 1 CD-ROM. – Загл. с титул. экрана. – Текст : электронный.

ISBN 978-5-907965-71-3

Настоящее пособие предназначено для освоения базовых методик анализа информации при помощи программного пакета Software Package for Social Science (SPSS) – популярного программного продукта, ориентированного на решение задач в данной области. Пособие может представлять интерес для студентов старших курсов направления реклама и связи с общественностью, менеджмент, экономика, социология. А также для всех интересующихся вопросами обработки социально-экономической информации.

Минимальные системные требования: PC с процессором Intel 1,3 ГГц и выше ; 256 Мб (RAM) ; Microsoft Windows, MacOS; дисковод CD-ROM ; Adobe Reader

УДК 339.138
ББК 65.291.21я73

ISBN 978-5-907965-71-3
DOI 10.31483/a-10753

ОГЛАВЛЕНИЕ

Предисловие

Умение анализировать массивы экономических и социальных данных при помощи программных пакетов обработки превратилось для современных маркетологов и специалистов в области рекламы в насущную необходимость. Без этих навыков им все сложнее конкурировать на рынке труда, соответствовать возрастающим требованиям со стороны работодателей.

Настоящее пособие предназначено для освоения базовых методик анализа социально-экономической информации при помощи популярного программного пакета Software Package for Social Science (SPSS). Выбор SPSS связан с его специализацией именно в области социальных наук.

Пособие построено на основе справочной информации и документации пакета SPSS 17.0 и представляет собой пошаговые инструкции к решению определенных социально-экономических задач, исходной информацией для которых служат учебные файлы данных, входящие в дистрибутив SPSS 17.0. Поэтому читатель имеет возможность (и ему рекомендуется это сделать в целях обучения) решить задачу самостоятельно, последовательно выполняя действия, подробно описанные в пособии.

Решая задачи из разных разделов, обучающийся знакомится с различными статистическими методиками решения задач. Для большинства методик приводится их общее описание, сфера решаемых задач и критерии применимости, параметры, трактовка результатов тестов. По ходу решения задач читателю предлагаются вопросы, нацеленные на повторение пройденного материала, самостоятельный анализ результатов расчетов, с последующей самопроверкой. Сознательная проработка вопросов повышает эффективность обучения.

1. Первое знакомство с пакетом SPSS.

В настоящей главе обсуждаются основы SPSS: представление и вывод данных на примере частотного анализа.

1.1. Представление данных.

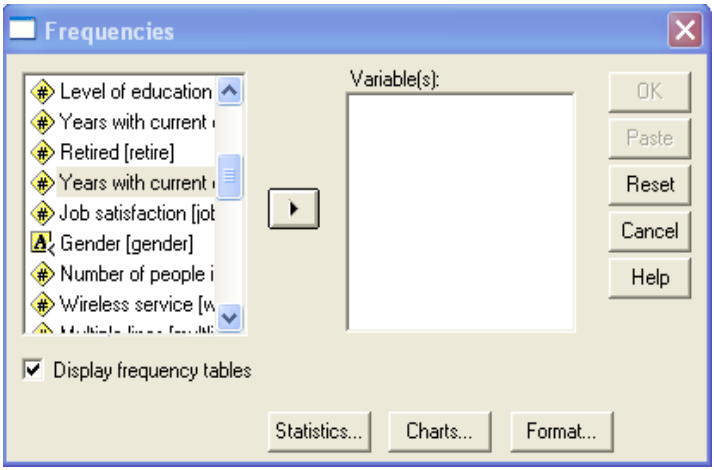
Используемый файл данных: **tutorial\sample_files\demo.sav**

Открытие данных: меню ***file->open->data***, кнопка ***Open***  , открывается в редакторе данных. Разные способы представления данных

- Метки – меню ***View->Value labels***, кнопка ***Value labels*** 
- Подробное описание переменных ***View->Variables***, комбинация клавиш ***Cntrl+T***, кнопка ***Variables*** 

1.2. Частотный анализ в один шаг.

Используемый файл данных: **tutorial\sample_files\demo.sav**

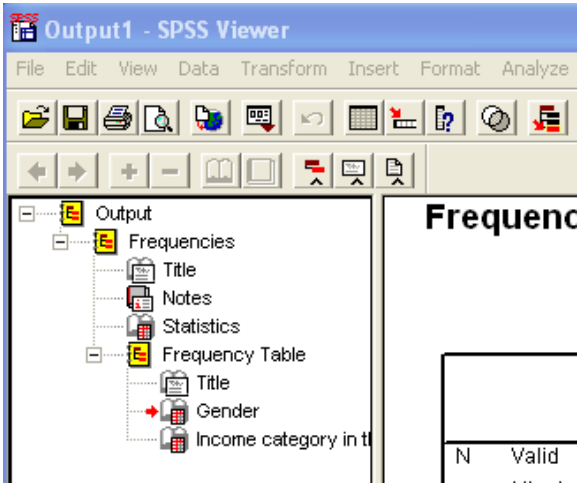


Открываем файл, выбираем меню ***Analyze->Descriptive Statistics->Frequencies***, получаем окно справа. Обращаем внимание на дополнительную информацию о переменных (список в левом окне): значком  обозначены цифровые переменные,  - строковые (string, lphanumeric). Нажатие правой кнопки мыши над переменной и выбор опции ***Variable Information*** выдает дополнительную информацию.

Выбираем переменные ***Gender [gender]*** и Income ***Category [inccat]*** (кнопка ) и нажимаем ***OK***.

1.3. Вывод результатов в виде таблиц.

Действие предыдущего параграфа выдает окно выдачи информации (***SPSS Viewer***)



Основные моменты:

- Все объекты активны, щелчок правой кнопки мыши по объекту показывает дополнительные возможности. Наиболее полезные опции:
  - объяснение результатов (***Results Coach***)
  - объяснение методик (***Case studies***)
- редактирование
- экспорт в другие форматы (кнопка )

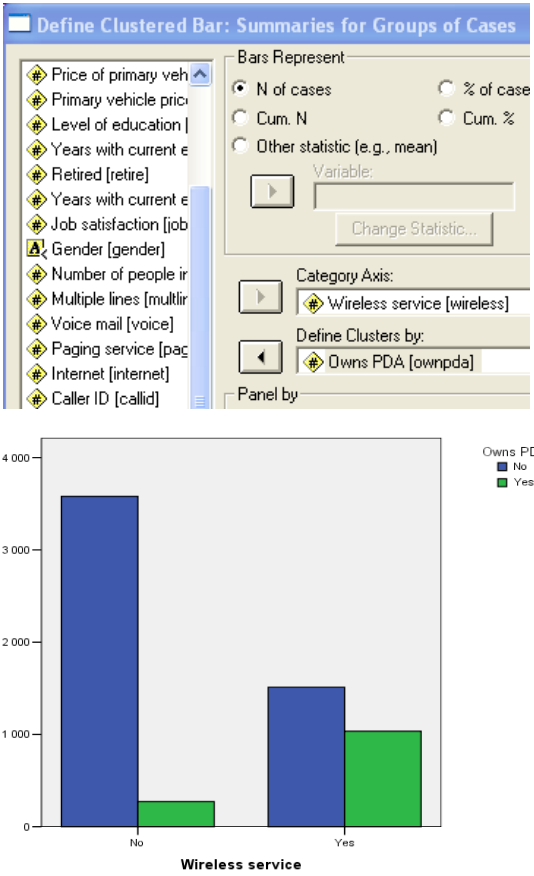
1.4. Печать результатов в виде графиков.

Используем файл данных: **tutorial\sample_files\demo.sav**

Задача: построить график, показывающий взаимосвязь подключения к беспроводной связи (***wireless telephone***) и наличия карманного компьютера ***PDA (personal digital assistant***).

Строим сгруппированный график. Меню ***Graphs->Bar***, выбираем тип ***Clustered*** (***Define***).

Выбираем переменную***Wireless service [wireless]*** в качестве объясняемой (***Category Axis variable***), а переменную ***Owns PDA [ownpda]*** в качестве группирующей(***Define Clusters By variable***) и получаем график:



1.5. Помощь.

Стандартный набор (информация по категориям, индекс и поиск). Существуют дополнительные возможности:

- Мастера статистики (***Statistics Coach***), в режиме диалога подсказывают необходимую статистику для решения задачи пользователя.
- Мастер результатов (***Results Coach***) пошаговым образом объясняет смысл использования результатов.

1.6. Подробней о частотном анализе (Frequency analysis).

Применяется для:

- определения типичного (среднего) значения величины, ее диапазона значений.
- Оценки распределения величины, его адекватности дальнейшему анализу (например, многие стат.методы требуют близкого к нормальному (гауссовому) распределения величины)
- Оценка качества данных: наличие потерянных, отсутствующих в ответах данных.
- В количественных переменных (***Scale variable***) существует ряд важных показателей:
- Среднее значение (***Mean***) и медиана, (взвешенное среднее, ***Median***) определяют главную тенденцию показателя. Сильное отличие медианы от среднего значения показывает наличие выбросов, которые влияют на среднее значение, но не на медиану.

Стандартная ошибка (***корень дисперсии, Std. deviation***) показывают разброс значений величины вокруг средней и имеют смысл для нормально распределенных величин.

Асимметрия (***skewness***) показывает меру отличия распределения величины от нормальной. Для нормального распределения она равна нулю, для длинного правого хвоста она больше нуля, и меньше нуля для левого хвоста. Распределение существенно асимметрично, если |Асимметрия|>1.

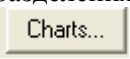
Иногда асимметричное распределение может быть сведено к нормальному путем преобразований. Чаще всего используется логарифмирование.

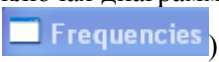
Если не удается свести распределение величины к нормальному, необходимо использовать непараметрические тесты, не требующие нормального распределения измеряемых величин.

Используемый файл данных: **tutorial\sample_files\contacts.sav**

Вы менеджер фирмы, продающей компьютерные комплектующие в фирмы, занимающиеся разработкой программного обеспечения. Статистика контактов в указанном файле.

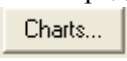
Задача: при помощи частотного анализа исследовать показатели.

1. Строим частотную таблицу по подразделениям (***Department***), включая диаграмму (***Pie chart***) (кнопка 

в меню-таблице 

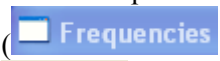
) Обратите внимание на количество недостающей информации (11.4%).

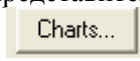
2. Для определения моды и распределения постройте линейчатый график в обратном порядке. Для этого возвращаемся к меню 

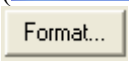
в 

выбираем ***Bar chart***, а в 

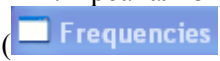
выбираем ***Descending count***.

3. Постройте линейчатый график представителей компаний (

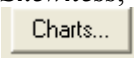
для Company rank, 

-> ***Bar chart***, 

->***Descending values***)

4. Проанализировать статистику последних покупок компаний (

для ***sale***). Снимите флаг с ***Display frequency table*** (возникающее предупреждение нормально, поскольку не выбрано никакого расчета-вывода результатов). Нажмите кнопку 

и в открывшемся меню обозначьте для вывода ***Quartiles***, ***Std.Deviation***, ***Minimum***, ***Maximum***, ***Mean***, ***Median***, ***Skewness***, ***Kurtosis***). Потом добавьте графическое построение 

-> ***Histograms with normal curve***. Проанализируйте результат

a. Половина распределения между 12 и 52.8 (***Percentiles***)

b. Полный диапазон между мин-макс

c. Медиана существенно отличается от средней – первое свидетельство асимметрии распределения.

d. Большая положительная величина коэффициента асимметрии (***Skewness***) – распределение имеет большой правый хвост.

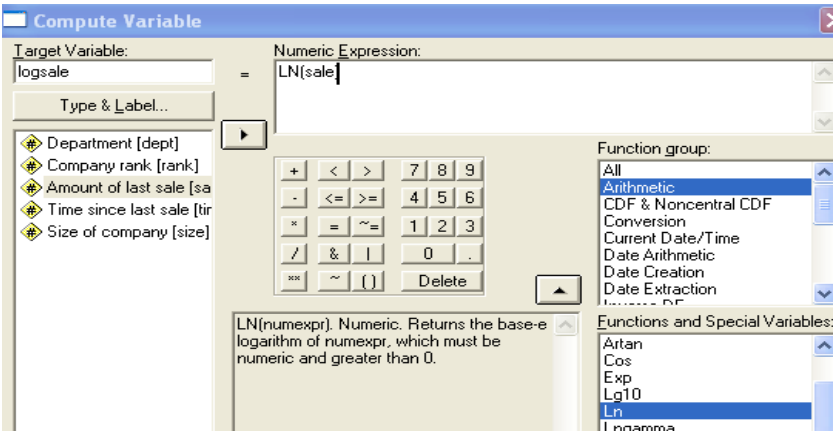
e. Большая асимметрия плюс п.с -> стандартная ошибка завышена и не является полезным показателем.

f. Большой коэффициент выброса (***kurtosis***) показывает, что распределение имеет более сильный пик и большие хвосты по отношению к нормальному.

g. Гистограмма показывает визуальное представление вышесказанного.

5. Попробуем преобразовать величину ***sale*** к нормальному виду путем логарифмирования (она по определению > 0, поэтому может быть логарифмирована). Для этого вызываем меню ***Transform->Compute*** Даем имя новой переменной ***logscale*** (в окошке ***Target variable***), а ее величину, считаем как логарифм от продаж ***LN(sale)***:

6. Повторяем п. 4 для переменной ***logsale***, комментируем результат.



7. Сохраняем файл результатов расчетов.

1.7. Вопросы для самоконтроля.

1. Что означает аббревиатура SPSS?

1. Для каких задач можно применять SPSS?

2. Какие типы переменных поддерживает SPSS, как посмотреть тип и описание переменных в окне данных?

3. В каком виде осуществляется вывод данных в SPSS, активны ли объекты вывода и что подразумевается под термином «активный»?

4. Какие возможности имеет система помощи в SPSS, какие особенности отличают ее от систем помощи в других программных пакетах?

5. Для чего применяется частотный анализ (ЧА)?

6. Как вызвать мастер ЧА в SPSS?

7. Какой из терминов идентичен ЧА: распределение процентных показателей, матрица рассеяния?

8. Какие количественные параметры можно рассчитать и вывести в ЧА? Как это сделать?

9. Как анализировать данные, не распределенные по нормальному закону? Каким стандартным преобразованием иногда можно свести данные к нормальному распределению? Как это сделать в SPSS?

10. Что показывает параметр асимметрии? При каком значении асимметрии распределение считается симметричным?

11. Потренируйте частотный анализ на результатах переписи населения США (файл **1991 U.S. General Social Survey.sav**), попробуйте получить нетривиальные результаты.

2. Работа с данными.

В данной главе описаны основные операции работы с исходными данными, преобразование и создание вторичных переменных.

2.1. Графический редактор (Chart Editor) на примере корреляционной матрицы (матрицы рассеяния, scatter plot).

Используемый файл данных: **tutorial\sample_files\car_sales.sav.**

Задача: исследовать взаимосвязь между расходом горючего (***mpg*** – mills per gallon миль на галлон) и весом машины (***curb_weight***).

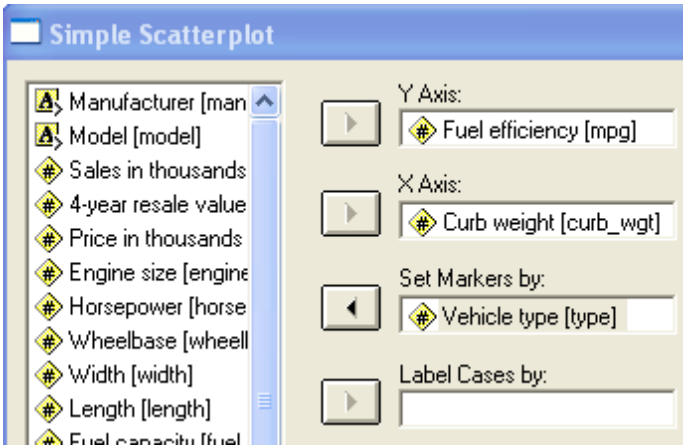
Вопрос (повторение пройденного): можно ли использовать корреляционную матрицу для исследования переменных ***mpg*** и ***curb_weight***? (Нужно построить частотные гистограммы и посмотреть статистические показатели и убедиться, что они не сильно отличаются от нормальных).

Выбираем меню ***Graphs->Scatter/Dot…->Simple Scatter***.

Дополнительный вопрос: почему для переменной ***Manufacturer*** нельзя построить корреляционную матрицу (внести в качестве переменной по осям)?

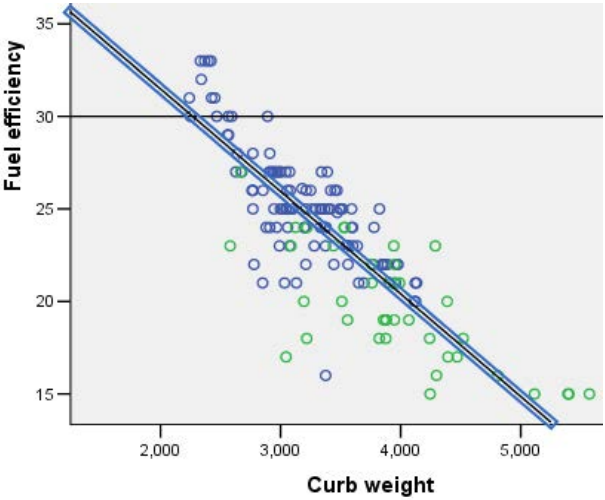
Ответ:Для строковых переменных нельзя построить корреляционную матрицу.

Берем ***Fuel efficiency*** и ***Curb weight*** как переменные по осям у/х соответственно, а ***Vehicle type*** в качестве группирующей (***Set Markers by***):



В окне результатов (***Viewer***) двойным кликом по построенной матрице вызывается редактор изображений (***Chart Editor***) в котором можно осуществлять редактирование полученной картинки и ряд расчетов.

- Можно активировать все данные, отдельные группы и отдельные значения, совершая над ними определенные действия (например, закрашивая или выделяя определенные точки контрастным цветом с помощью линейки )
- Можно редактировать картинку, добавляя к ней графические объекты (см. кнопки или содержание меню)
- Фитировать данные линией. Выбираем меню ***Elements->Fit line at total*** или кнопка  и строим линейный ***(linear***) плот:



Важный показатель – ***коэффициент детерминации*** $\mathbf{R^2}$ изменяется от 0 (корреляции нет) до 1 (полная линейная зависимость) и показывает, какой процент корреляции объясняет прямая. Чем он больше, тем лучше, и в данном случае R^2=0.67 (линейный фит объясняет 67% данных).

- Отобразить числовые величины данных на графике.

Выбираем меню ***Elements->Show Data Labels*** или кнопка  . По умолчанию видим порядковый номер точки в данных (***Case Number***), но в появляющемся меню опций (***Properties***) можно выбрать другие варианты. Повторное нажатие на ту же кнопку или меню ***Elements->Hide Data Labels*** убирает числовые метки.

- Редактирование отдельных элементов на графике. Уберите числовые метки, если они активированы (см. предыдущий пункт) и выберите меню ***Elements->Data Label Mode*** (кнопка ). Курсор примет вид, аналогичный кнопке. Кликнув им на точке данных, вы активируете, тем самым, ее числовое значение. Дезактивация этой моды происходит повторным выбором ***Elements->Data Label Mode*** или нажатием кнопки  .

Аналогичным образом в Редакторе графики (***Chart Editor***) можно редактировать любые иные графики, полученные из других методик обработки данных

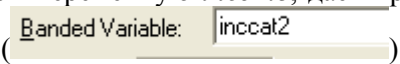
2.2. Преобразование данных.

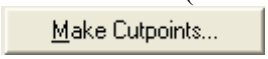
При анализе данных часто требуются различные преобразования исходных переменных (мы уже использовали логарифмирование для приведения частотного распределения к нормальному) и последующее использование вторичных переменных. Рассмотрим этот вопрос подробней:

2.2.1. Создание порядковой (категориальной) переменной из непрерывной.

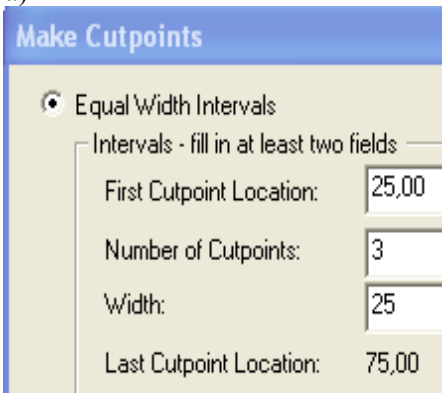
Используемый файл данных: **tutorial\sample_files\demo.sav**.

В данном файле уже есть такая ***inccat***, сделанная из непрерывной переменной ***income***. Задача: создать идентичную переменную ***inccat*** средствами программы.

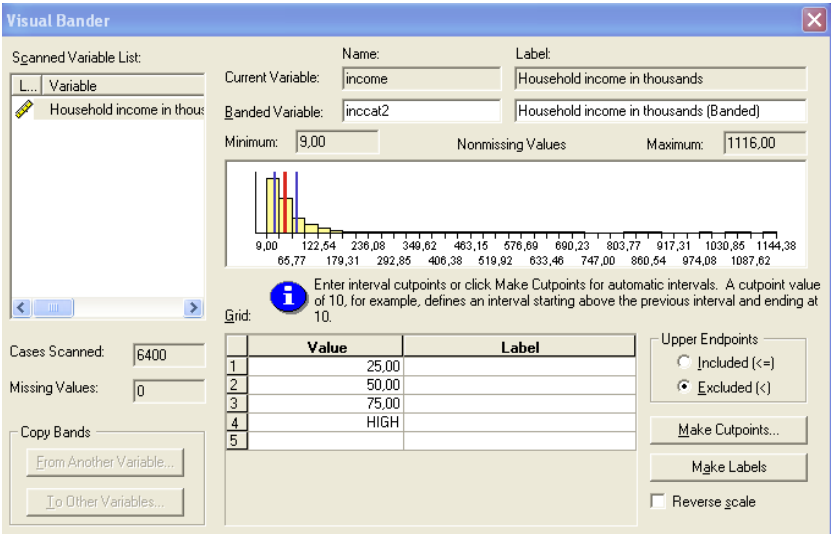
Вызываем меню ***Transform->Visual Bander*** и выбираем для группировки (***banding***) переменную ***income***. Нажимаем Continue и в окне связей выделяем переменную ***income***, даем групповой переменной имя ***inccat2*** (

) и нажимаем кнопку 

и выбираем параметры группировки (одинаковые диапазоны, начальную точку, количество делений и ширину диапазона)



и нажимаем ***Apply***:



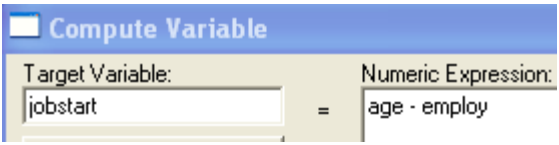
Для последнего диапазона  выбираем  и нажимаем  , что дает автоматические обозначения (метки, ***Label***) для отдельных значений переменных. Правило хорошего тона – давать значениям групповой переменной внятные имена для понимания ситуации. Метки можно отредактировать по своему усмотрению. Жмем ***OK*** и получаем новую переменную ***inccat2***.

2.2.2. Создание расчетной переменной.

Задача: построить переменную «Возраст выхода на последнюю работу».

Для этого из текущего возраста нужно вычесть количество лет работы на последнем рабочем месте.

Выбираем меню ***Transform->Compute*** и собираем соответствующее выражение:



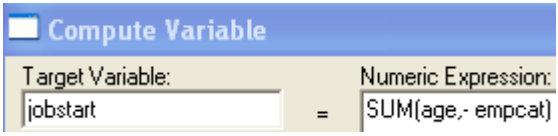
После нажатия ***OK*** в таблице данных появляется переменная ***jobstart***

2.2.3. Использование функций для создания вторичной переменной.

При создании новых переменных можно использовать:

- Арифметические функции
- Логические
- Статистические
- Распределения
- Агрегирующие и выделяющие функции от времени и даты
- Функции для отсутствующих значений (***missing values***)
- Перекрестные (***cross-case***) функции
- Строковые функции

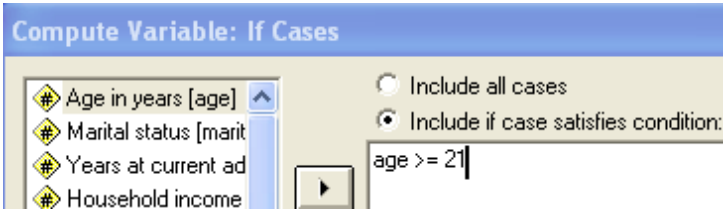
Например, предыдущая задача может быть скомпонована следующим образом:

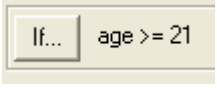


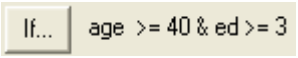
2.2.4. Использование условных выражений.

При расчетах переменных можно накладывать ограничения на используемые выражения. Для этого в окне вычисления новой переменной (***Compute variable***) используем кнопку 

. В частности, чтобы учитывать только совершеннолетних респондентов, накладываем следующее условие:



Нажимаем ***Continue*** и видим условие в окне расчета новой переменной: 

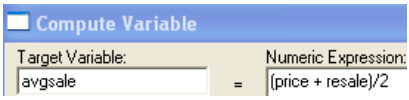
Возможны комбинированные условия: 

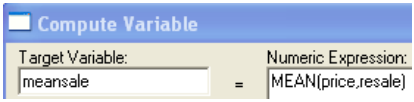
2.2.5. Работа с пропущенными значениями.

Используемый файл данных: **tutorial\sample_files\car_sales.sav.**

Пропущенные значения в записях обрабатываются разными функциями различным образом. На этот момент нужно обращать внимание.

Задача: рассчитать среднюю величину стоимости машины при первой продаже и после 4 лет эксплуатации. Рассчитаем двумя способами:





Обратите внимание на результаты расчетов в таблице данных, объяснить (первый вариант расчетов не возвращает ответ, если нет одного из значений, вторая функция (**mean**) считает результат, отбрасывая отсутствующую переменную):

sales	resale
16,919	16,360
39,384	19,875
14,114	18,225
8,588	29,725
20,397	22,255
18,780	23,555
1,380	39,000
19,747	.

avgsale	meansale
16,64	16,64
29,63	29,63
16,17	16,17
19,16	19,16
21,33	21,33
21,17	21,17
20,19	20,19
.	19,75

Решение вопроса – указание через точку после функции минимального количества непустых переменных в функции. В предыдущем варианте для аналогичного ***avgsale*** результата необходимо ввести:

что дает ⇒

meansale2
16,64
29,63
16,17
19,16
21,33
21,17
20,19
.

2.2.6. Функции даты и времени.

Используем специальный Мастер дат и времен (***Date and Time Wizard***) который позволяет:

- Создавать переменные типа даты-времени из строковых, включающих дату-время
- Составлять переменную даты-времени из переменных, содержащих по отдельности параметры даты-времени
- Добавлять-прибавлять величины к дате-времени, включая две переменных типа дата-время
- Извлекать отдельные показатели (например, год, минуты и т.д.) из переменных дата-время.

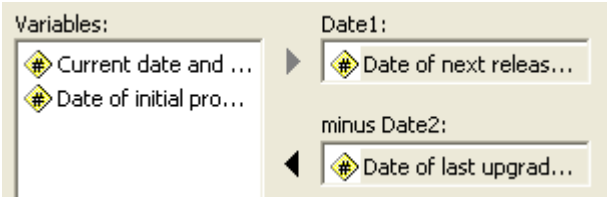
Вычитание дат

Используемый файл данных: **tutorial\sample_files\upgrade.sav**.

Задача: найти время между выходом новой версии продукта (***NextRel***) и датой последнего апгрейда (***LastUp***).

Выбираем меню ***Transform->Date/Time***. В получившемся мастере выберете ***Calculate with dates and times*** (при желании можно получить краткую справку о переменных типа время-дата через опцию ***Learn how dates and times are represented in SPSS***) и выберите опцию ***Далее>*** Выберите ***Calculate the number of time units***

between two dates и еще раз ***Далее>*** и выберите необходимые переменные и нажмите ***Далее>***:



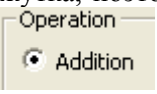
Введите название новой переменной ***YearsLastUp*** и ***Years since last upgrade*** в качестве метки. Поставьте флажок ***Create the variable now*** и нажмите ***Готово***.

Прибавление величин к дате

Задача: подсчитать срок окончания технической поддержки. Для этого необходимо прибавить к дате покупки число лет технической поддержки.

В мастере дат-времен выберем ***Calculate with dates and times-> Add or subtract a duration from a date*** Выберите ***Date of initial product license*** в качестве ***Date***, а ***Years of tech support*** как переменную срока (***Duration Variable***). Поскольку последняя цифровая, надо указать цену промежутка, поэтому ставим

чае, наконец, операция – сложение



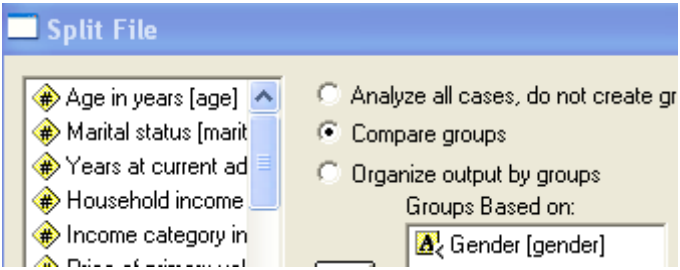
2.3. Сортировка данных.

В некоторых видах анализа бывает полезно или необходимо отсортировать исходные данные по одной или нескольким переменным. Мастер вызывается из меню ***Data->Sort Cases…***

2.4. Обработка данных по отдельным группам.

Используемый файл данных: **tutorial\sample_files\demo.sav**.

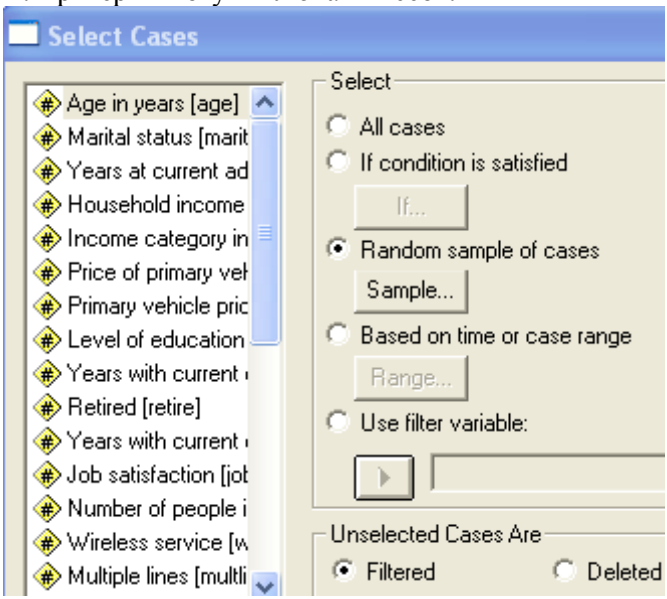
Провести анализ доходов по половому признаку. Для разделения данных на группы для отдельного анализа воспользуемся меню ***Data->Split File…***



После чего остальные операции будут делаться с учетом группировки. Строим частотные таблицы для переменной ***inccat***. Для снятия разделения выбираем меню ***Analyze all cases***.

2.5. Выделение и обработка выборок (subsets) из данных.

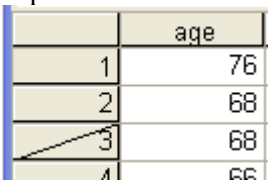
Можно ограничить свой анализ на основании выборок, отобранных на основании значений переменных и комплексных выражений. Критерии могут включать в себя:



- Значения переменных и их диапазоны, включая диапазоны времен и дат.

- Номера записей данных.
- Арифметические, логические и функциональные выражения.

Воспользуемся меню ***Data->Select Cases***. Как видно из мастера, можно ввести выражение (***if condition…***), взять случайную выборку, условие на диапазон, фильтрацию по переменной. Невыбранные записи могут просто отфильтровываться, либо удаляться. Если выбрано второе, то при сохранении данных удаленные записи уже нельзя восстановить! Отфильтрованные записи обозначены диагональным перечеркиванием.



Если объем данных велик, то удаление отфильтрованных данных существенно ускоряет процесс обработки.

2.6. Вопросы для самоконтроля.

1. Что такое корреляция? Что такое коэффициент детерминации? Для каких типов переменных он определен? Приведите диапазоны значений коэффициента детерминации для сильной и слабой корреляции.
2. Приведите синонимы термина «корреляционная матрица»?
3. Продемонстрируйте способы изменения графических параметров точки и группы точек на графике в SPSS.
4. Что показывает фит корреляционной матрицы?
5. Для чего нужны преобразования переменных при анализе данных?
6. Как преобразовать переменную в SPSS?
7. Для чего нужны расчетные (вторичные) переменные? Какие функции для их создания позволяет применять SPSS?
8. Для чего применяются условные выражения при создании переменных? Как это делается в SPSS?

9. В чем тонкости обработки пропущенных значений при анализе данных? Продемонстрировать различия обработки пропущенных значений в SPSS разными функциями.

10. Можно ли использовать функции даты в SPSS и каким образом?

11. Как осуществляется сортировка и отбор данных (создание выборок) в SPSS? В каких случаях это необходимо и полезно при анализе данных?

3. Базовые статистики и методы анализа.

В раздел вошли основополагающие (отсюда термин «базовые»), наиболее часто использующиеся (особенно, на первичном этапе работы с первичной информацией) методики работы с данными: описательные статистики и основы работы с первичной информацией, сравнение перекрестных таблиц категориальных переменных, различные методики сравнения средних величин по выборкам данных, а также односторонний вариационный анализ.

3.1. Использование описательных статистик (Descriptive procedure).

Описательные статистики используются для общего сравнения нормально распределенных шкалируемых переменных и определения выбросов расчетом стандартизованных отклонений. Последняя величина именуется также Z-величиной (***Z-score***), рассчитывающейся как разница значений переменной и среднего, поделенная на стандартное отклонение. Z-величины используются также для сравнения переменных в разных шкалах.

Используемый файл данных: **tutorial\sample_files\telco.sav**

Задача: телекоммуникационная компания поддерживает базу пользователей, в которой отражена статистика пользования дальней связью, бесплатными услугами (***toll-free***), арендой оборудования, карточками и беспроводной связью. Определить наиболее доходные виды услуг.

Решение:

1. Выбираем меню ***Analyze->Descriptive Statistics->Descriptives*** и берем в качестве переменных все необходимые переменные: ***Long distance last month***, ***Toll free last month***, ***Equipment last***

month, ***Calling card last month***, and ***Wireless last month***. Подтверждая выбор кнопкой ***OK*** получим результат:

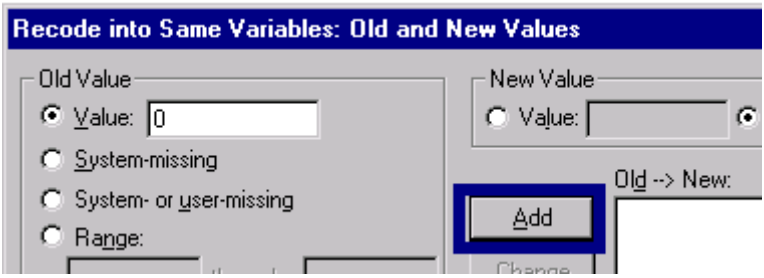
	N	Minimum	Maximum	Mean	Std. Devia
Long distance last month	1000	.90	99.95	11.7231	10.3E
Toll free last month	1000	.00	173.00	13.2740	16.9C
Equipment last month	1000	.00	77.70	14.2198	19.0E
Calling card last month	1000	.00	109.25	13.7810	14.0E
Wireless last month	1000	.00	111.95	11.5839	19.71
Valid N (listwise)	1000				

Вопрос: как вы проанализируете полученные табличные результаты?

2. **Ответ:** на основании полученной таблицы трудно сделать выводы – средние значения не сильно отличаются, основная проблема, много пользователей не пользуются каким-то видом связи, что показывают нули в столбце ***Minimum***. Эти нули, будучи учитываемыми в расчетах, дают дисбаланс частотных распределений (для проверки построить, к примеру, частотную гистограмму для ***Calling card last month***). Другой фактор, показывающий это – большое стандартное отклонение. Решение проблемы – учет в расчетах только ненулевых значений переменных. Иными словами, в SPSS необходимо указать, что нулевые значения переменных должны трактоваться как отсутствующие (***missing***) данные.

Для этого используем меню ***Transform->Recode->Into Same Variables…*** Выбираем те же переменные и нажимаем кнопку 

. В появившемся мастере перекодировки устанавливаем преобразование нулевых значений старой переменной (***Old value***) в отсутствующие (***System-missing***) величины нового параметра (***New Value***) при помощи кнопки ***Add***:



Повторите п.1 с одним изменением: после ввода описуемых переменных нажмите кнопку ***Options***, дезактивируйте вывод на экран показателей ***Minimum/Maximum***, добавив вместо них ***Skewness/Kurtosis***.

	N	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Long distance last month	1000	11.7231	10.36349	2.966	.077	14.052	.155
Toll free last month	475	27.9453	13.82910	3.465	.112	26.735	.224
Equipment last month	386	36.8389	10.39568	.756	.124	.641	.248
Calling card last month	678	20.3260	12.62916	2.150	.094	7.572	.187
Wireless last month	296	39.1348	15.32916	1.359	.142	3.079	.282
Valid N (listwise)	131						

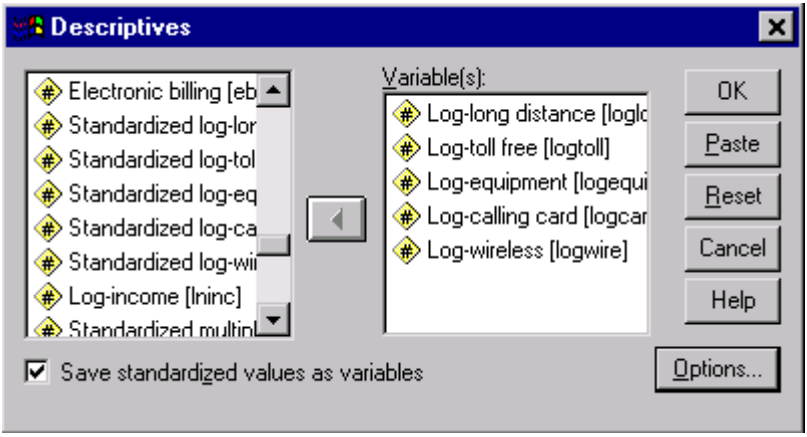
Задача: проанализировать результаты.

Ответ: видны кардинальные изменения. Выявились сервисы, приносящие наибольший доход, и сервисы, нуждающиеся в большем продвижении. При этом аренда оборудования имеет наименьшее стандартное отклонение. Коэффициенты асимметрии и эксцесса велики, что говорит о существенном отклонении распределения данных от нормальных. Помочь может логарифмирование, которое уже было сделано в данных этого файла.

3. Для поиска пользователей, тратящих значительно больше или меньше других, используем расчет Z-величин. Однако их можно использовать для величин, распределение которых близко к нормальному. В данном случае, подсчитаем z-отклонения для логарифмированных показателей объема услуг, оказанных пользователей.

Повторяем п.1 с логарифмированными переменными:

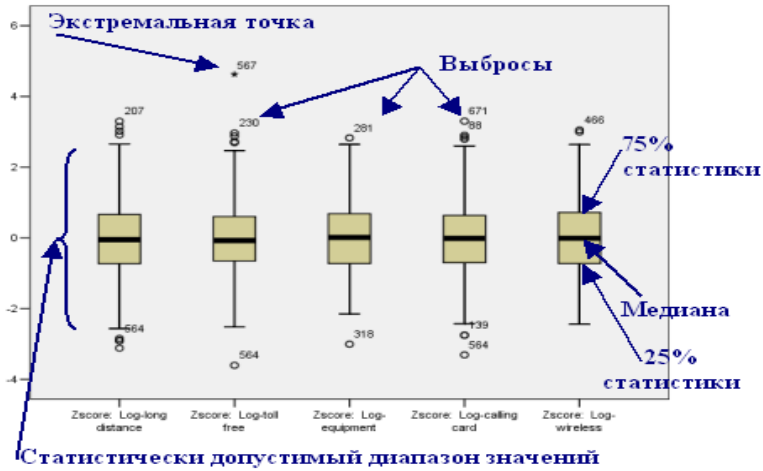
Активируем флажок ***Save standardized values as variables***, что позволяет сохранить статистику (переменные ***Zxxx*** в данных) для последующего анализа.



За исключением свободных-сервисов, ситуация нормализовалась в прямом и переносном смысле.

	N	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Log-long distance	1000	2.1821	.73455	.166	.077	-.001	.155
Log-toll free	475	3.2397	.41381	.304	.112	1.107	.224
Log-equipment	386	3.5681	.27756	.057	.124	-.344	.248
Log-calling card	678	2.8542	.55729	.081	.094	.109	.187
Log-wireless	296	3.5983	.36729	.200	.142	-.168	.282
Valid N (listwise)	131						

4. Для выявления пользователей свободных услуг, «выпавших из толпы», построим блочный график (***Boxplot***) Меню ***Graphs->Boxplot…->Simple*** с флагом ***Summaries of separate variables***. Выберем для изображения на графике (окно ***Boxes Represent***:) все переменные ***Zscore…*** (они подсчитаны автоматически на предыдущем шаге). Нажимаем кнопку ***Options*** и выбираем флаг ***Exclude cases variable by variable***. На результирующем графике виден пользователь (запись 567), использующий бесплатный сервис существенно больше других. Он и отвечает за искажение нормального спектра переменной.



Выводы: установлено, что аренда и беспроводной сервис приносят больший доход, хотя последний имеет больший разброс. Найден «ненормальный» пользователь бесплатных услуг. Дальнейшее исследование необходимо, случаен ли этот эффект, или это тенденция.

3.1.1. Блочный график Boxplot.

Блочный график (см. рис. предыдущего параграфа) представляет собой удобное визуальное представление статистических показателей по нескольким переменным или группировкам. Толстая черная линия внутри каждой метки – медиана, нижние и верхние края прямоугольника проводятся на уровнях 25 и 75% распределения переменной (тем самым, прямоугольник покрывает 50% всех данных). Усы показывают диапазон в пределах которого лежат статистически распределенные данные. Соответственно, данные лежащие вне усов, являются выбросами (обозначены кружком) и экстремальными значениями (обозначенными звездочкой). Рядом с выбросами и экстремальными точками указаны номера соответствующих записей в данных.

3.1.2. Вопросы для самоконтроля.

1. Для чего применяются описательные статистики? Какие типы переменных анализируются с помощью них?
2. Для чего используется преобразование некоторых значений переменных в отсутствующие?
3. Для чего в анализе первичные данные часто подвергают логарифмированию?
4. Для чего используется блочный график? Какие основные статистики он демонстрирует?

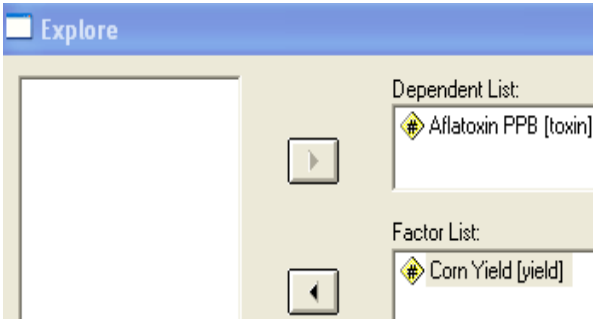
3.2. Первичное исследование и сортировка данных.

Первичное исследование данных необходимо для выяснения адекватности применения тех или иных статистических методов. В SPSS существует большое количество визуальных и численных методик первичной оценки данных, объединенных термином Explore procedure, как для всех, так и для группированных данных. Explore procedure позволяет:

- Сканировать данные
- Определять выбросы
- Проверять предположения
- Оценивать различия в группах данных.

Используемый файл данных: **tutorial\sample_files\aflatoxin.sav.**

Задача: зерна кукурузы должны проверяться на афлатоксин, яд, концентрация которого широко варьируется в разных партиях зерна. На элеватор привезли 8 партий зерен, однако распределение афлатоксина в миллиардных долях (**in parts per billion** (**PPB**)) обязательно необходимо проверить перед переработкой. Файл данных содержит анализ 16 образцов из каждой из 8 партий зерна.



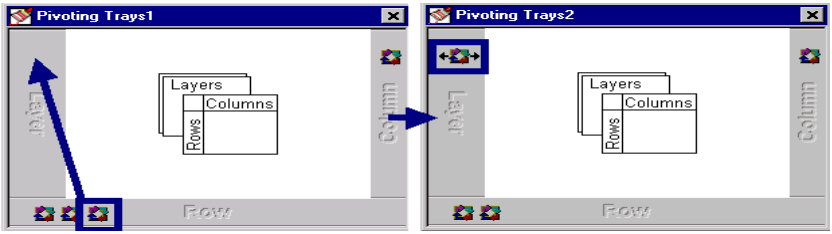
Решение: выбираем меню ***Analyze->Descriptive Statistics->Explore…*** Выбираем ***Aflatoxin PPB*** в качестве зависимой переменной (***Dependent List***), а ***Corn Yield*** – независимой (***Factor List***):

После нажатия ***OK*** в окне вывода представлены результаты расчетов. Для того чтобы посмотреть статистику в компактном виде, щелкните двойным кликом по таблице описаний (***Discriptives***)

Descriptives

	Corn Yield			Statistic	Std. Error
Aflatoxin PPB	1	Mean		20,2500	1,07819
		95% Confidence Interval for Mean	Lower Bound	17,9519	
			Upper Bound	22,5481	

И выберите меню ***Pivot->Pivoting Trays…*** показ статистики в окне прокрутки (***Pivoting Trays1***). Перетащите кнопку статистики в колонку слоя, как показано на рисунке:



После ввода, описательная таблица примет вид:

Descriptives

Mean

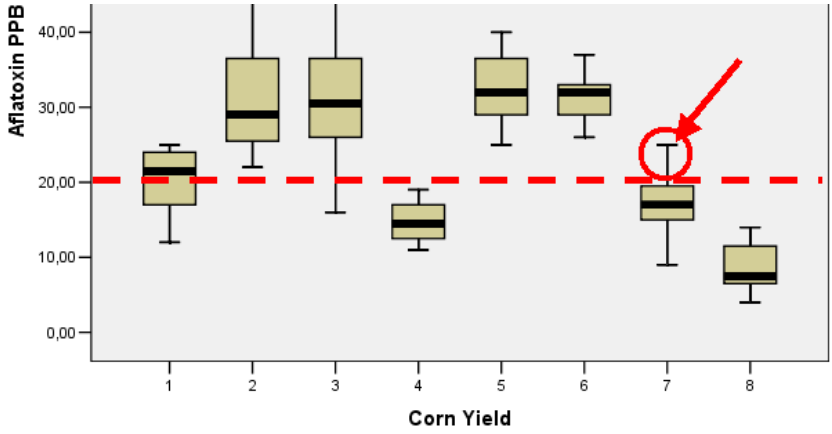
	Corn Yield	Statistic	Std. Error
Aflatoxin PPB	1	20,2500	1,07819
	2	33,0625	3,04339
	3	32,6875	2,57669
	4	14,6875	,66281
	5	33,0000	1,55724
	6	31,3750	,71224
	7	17,0625	1,04670
	8	8,4375	,76903

Согласно американским законам, содержание афлатоксина в пищевой кукурузе не должно превышать 20 PPB. Вопрос: какие партии пригодны для употребления?

Ответ: 4-я,7-я,8-я.

Вопрос: почему 7-я партия входит в данную группу ответов?

Ответ: посмотреть Boxplot (нарисован предыдущей процедурой и находится после таблиц), на нем видно, что верхний статистический ус блока больше 20, а значит, есть вероятность, что измеренная нами величина уровня отравы на самом деле больше 20 PPB.

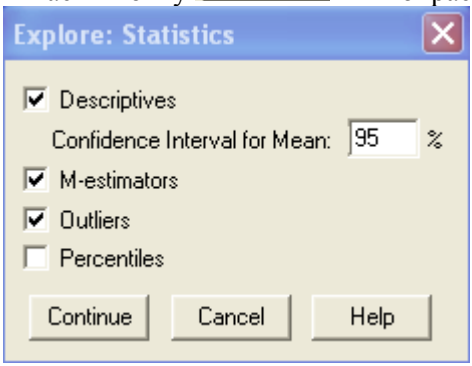


Используемый файл **tutorial\sample_files\ceramics.sav**.

Задача: фабрика использует нитрид серебра для производства керамических подшипников, которые должны выдерживать температуру 1500 ОС. Известно, что тепловое сопротивление стан-

дартного сплава нормально распределено. Однако впервые тестируется новый сплав, распределение которого неизвестно.

Решение: выбираем меню ***Analyze->Descriptive Statistics->Explore…*** Выбираем ***Degrees Centigrade*** в качестве зависимой переменной (***Dependent List***), ***Alloy*** – независимой (***Factor List***), ***labrunid*** в качестве маркировочной переменной (***Label cases by***). Поскольку нам необходимы робастные (надежные) оценки общей тенденции, нажимаем кнопку Statistics... и выбираем расчет:



Кроме того, для проверки нормальности распределения нажимаем кнопку Plots... и отмечаем флаг Normality plots with tests

В описывающей таблице видим, что для стандартного сплава (***Alloy standard***) распределение нормально: среднее, 5% усеченное среднее (***trimmed Mean***, при расчете n% усеченного среднего отбрасываем n процентов статистики снизу и сверху. Такая оценка менее чувствительна к выбросам (которые в отброшенной статистике), поэтому лучше выявляет среднее значение) и медиана близки друг другу, а ***skewness/kutosis*** близки к нулю.

Новый сплав (***alloy premium***) – другое дело. Среднее тут выше усеченного среднего и медианы, а величины ***skewness/kutosis*** существенно больше нуля – распределение отличается от нормального.

Робастные оценки среднего (***M-estimators,*** М-оценки) близки к медиане, отличаясь от среднего, что говорит о ненормальности распределения:

M-Estimators

	Alloy	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
Degrees Centigrade	Premium	1540,0953	1539,5658	1540,2052	1539,5506
	Standard	1514,6413	1514,6925	1514,6828	1514,6955

a. The weighting constant is 1,339.

b. The weighting constant is 4,685.

c. The weighting constants are 1,700, 3,400, and 8,500

d. The weighting constant is 1,340*pi.

Таблица пяти максимальных, пяти минимальных значений выглядит, как показано на следующем рисунке. Из нее видно, что, с одной стороны, разброс верхних значений для нового сплава существенно больше стандартного (разброс среди нижних меньше), с другой стороны, температуру новый сплав держит существенно лучше стандартного, никогда не опускаясь ниже 1530,44.ОС

Extreme Values

	Alloy			Case Number	labrunid	Value
Degrees Centigrade	Premium	Highest	1	211	d421	1591,04
			2	417	g837	1574,62
			3	17	a 17	1571,77
			4	437	h917	1568,10
			5	357	f657	1567,07
		Lowest	1	139	c289	1530,44
			2	475	h955	1530,73
			3	199	d379	1530,75
			4	373	g733	1530,76
			5	207	d387	1530,79
	Standard	Highest	1	408	g828	1537,99
			2	198	d378	1534,29
			3	20	a 20	1534,06
			4	168	c318	1533,43
			5	184	d364	1533,35
		Lowest	1	396	g816	1488,30
			2	100	b190	1488,36
			3	80	b170	1494,09
			4	154	c304	1494,64
			5	240	d450	1495,15

Тест Колмогорова–Смирнова показывает качество вписывания нормальной кривой в данные. Если тест существенный (**significant**, обычно существенным считается значение, большее 0,05, хотя в данном примере, согласно сноске * это 0,2). Если количество записей меньше 50, то показывается также и тест **Шапиро–Вилка**.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Alloy	Statistic	df	Sig.	Statistic	df	Sig.
Degrees Centigrade	Premium	,123	240	,000	,888	240	,000
	Standard	,027	240	,200*	,995	240	,602

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

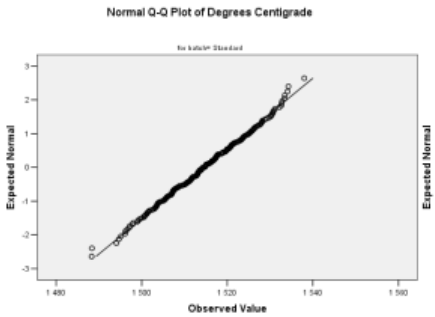
График «ствол-листья» – оригинальный алфавитно-цифровой способ (очевидно, использовался на старых машинах в «дографическую эру») изображения распределения переменной (фактически, это повернутая на бок гистограмма):

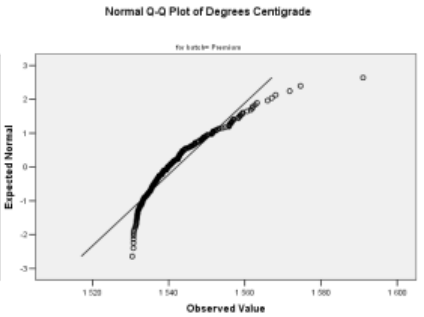
Stem-and-Leaf Plots

```
Degrees Centigrade Stem-and-Leaf Plot for
batch= Premium

 Frequency     Stem &  Leaf

    24,00        153 .  000000011111111111111111
    22,00        153 .  22222222223333333333
    26,00        153 .  44444444445555555555555555
    26,00        153 .  66666666666666777777777777
    24,00        153 .  88888888888888999999999999
    19,00        154 .  0000000000111111111
    25,00        154 .  22222222222222223333333333
    10,00        154 .  4444455555
    12,00        154 .  666666667777
    10,00        154 .  8888999999
     8,00        155 .  00111111
     4,00        155 .  2223
     6,00        155 .  445555
     6,00        155 .  666667
     6,00        155 .  888899
     3,00        156 .  011
     3,00        156 .  223
     6,00 Extremes     (>=1566)

 Stem width:      10,00
 Each leaf:        1 case(s)
```

Из графика видно, что в диапазоне нижних температур распределение близко к равномерному - в диапазоне 1530-1543 ОС, а в высоких температурах очевиден существенный разброс показателей.

Наконец, **Q-Q график** также показывает отклонение наблюдаемой величины от ожидаемой нормальной. Из графика следует, что функциональная зависимость для нового сплава отличается от нормальной, особенно в области высоких энергий.


Normal Q-Q Plot of Degrees Centigrade
Expected Normal
Observed Value


Normal Q-Q Plot of Degrees Centigrade
Expected Normal
Observed Value

3.2.1. Вопросы для самоконтроля.

1. Для чего необходим первичный анализ данных?
2. Каковы основные показатели, приводимые описательной таблице SPSS (Descriptives)?
3. Какие основные статистики используются в описательном анализе?
4. Что показывает отличие среднего от медианы? Для чего применяют усеченное среднее?
5. Как определить асимметрию и не нормальное распределение переменной?
6. Что такое М-оценка (M-estimator)?
7. Что показывает тест Колмогорова–Смирнова? Признак значимости этого теста?
8. Каковы критерии применимости тестов Колмогорова–Смирнова и Шапиро–Вилка?
9. Что такое графики «ствол-листья», Q-Q? Постройте их в SPSS.

3.3. Перекрестные таблицы (crosstabs).

Исследование перекрестных таблиц – классическая техника анализа категориальных переменных, имеющих ограниченный набор значений (номинальных или ординарных переменных), возможно, при контроле третьей переменной слоя (***layering variable***, если первые две анализируемые переменные дают строки и колонки таблицы, третья переменная слоя дает глубину).

Перекрестные таблицы используются для проверки независимости, ассоциативности и согласия переменных. Дополнительно можно оценить относительный риск события, состоящего в наличии или отсутствии определенной характеристики.

3.3.1. Перекрестный анализ номинальных переменных.

Используемый файл **tutorial\sample_files\satisf.sav.**

Задача: Для определения уровня удовлетворенности покупателей компания-ритейлер провела опрос 582 клиентов в 4 супермаркетах. Из исследования вы определили, что качество сервиса

было наиболее важным фактором общей удовлетворенности покупателей. Имея эту информацию, мы хотим проверить, предоставляет ли каждый из супермаркетов одинаковый и адекватный уровень обслуживания покупателей.

Решение: с помощью перекрестных таблиц проверяем гипотезу, что уровень удовлетворенности клиентов одинаков по всем магазинам.

1. Выбираем меню ***Analyze->Descriptive Statistics->Crosstabs...*** затем ***Store*** по строкам и ***Service Satisfaction*** по столбцам. Нажимаем кнопку 

и выделяем ряд показателей ***Chi-square***, ***Contingency Coefficient***, ***Phi and Cramer's V***, ***Lambda***, and ***Uncertainty coefficient***.

В ячейках **перекрестной таблицы** отображены частотные показатели опроса. Например, количество крайне недовольных покупателей в третьем магазине оказалось 15.

Store * Service satisfaction Crosstabulation

Count

		Service satisfaction					
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	Total
Store	Store 1	25	20	38	30	33	146
	Store 2	26	30	34	27	19	136
	Store 3	15	20	41	33	29	138
	Store 4	27	35	44	22	34	162
Total		93	105	157	112	115	582

Задание: проанализировать таблицу.

Ответ: покупатели менее довольны магазином 2 и более довольны магазином 3.

Однако из перекрестной таблицы нельзя определить, являются ли данные различия значимыми, или это игра статистики (статистическая флуктуация). Для этого нужны другие способы.

3.3.2. Хи-квадрат тест (Chi-square test).

Хи-квадрат тест показывает различия между наблюдаемыми данными и теми, что получились бы при независимых столбцах и колонках. Малая (обычно в качестве порога берут 0.05) асимптотическая значимость (***Asymptotic Significance***, асимптотическая означает использование известного статистического распределе-

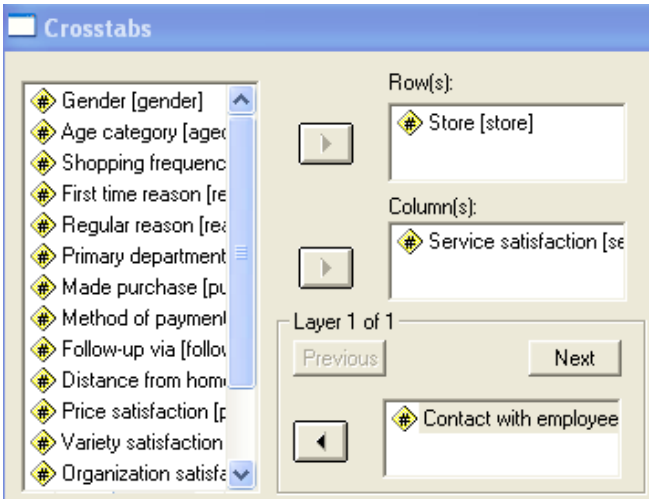
ния, к которому стремится измеренное при устремлении количества измерений к бесконечности) сигнализирует о взаимосвязи переменных, высокая – об их независимости. В нашем случае:

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.293[a]	12	.178
Likelihood Ratio	17.012	12	.148
Linear-by-Linear Association	.084	1	.772
N of Valid Cases	582		

асимптотическая значимость больше 0,1, поэтому можно утверждать, что различия связаны с игрой статистики.

Однако не все респонденты контактировали с представителями магазинов, поэтому не все могут объективно оценить уровень сервиса. Попробуем проанализировать анкеты только тех, кто сталкивался с работниками магазинов.

2. Повторяем п.1 и устанавливаем переменную слоя ***Contact with employee***:



После этого перекрестная таблица разбивается с учетом переменной слоя:

Store * Service satisfaction * Contact with employee Crosstabulation

Count

Contact with employee			Service satisfaction					Total
			Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	
No	Store	Store 1	16	9	18	17	19	79
		Store 2	2	15	16	13	12	58
		Store 3	9	14	23	22	14	82
		Store 4	17	14	19	10	10	70
	Total		44	52	76	62	55	289
Yes	Store	Store 1	9	11	20	13	14	67
		Store 2	24	15	18	14	7	78
		Store 3	6	6	18	11	15	56
		Store 4	10	21	25	12	24	92
	Total		49	53	81	50	60	293

Видно, что среди респондентов, общавшихся с работниками магазина, недовольство магазином 2 более ярко выражено. Хи-квадрат тест показывает следующую статистику:

Chi-Square Tests

Contact with employee		Value	df	Asymp. Sig. (2-sided)
No	Pearson Chi-Square	20,898[a]	12	,052
	Likelihood Ratio	22,937	12	,028
	Linear-by-Linear Association	3,514	1	,061
	N of Valid Cases	289		
Yes	Pearson Chi-Square	25,726[b]	12	,012

Для респондентов, не пересекавшихся с работниками магазина, значимость теста на пороге, что позволяет предположить зависимость переменных, но подтвердить ее можно только в других тестах.

Для респондентов, вкусивших прелесть общения с работниками магазина, налицо взаимосвязь переменных, и можно утверждать, что выявленное различие в перекрестной таблице – реальное, а не игра статистики.

Важно отметить, что критерий хи-квадрат показывает только наличие связи, а не направление и силу. Для последних показателей используем другие методы.

3.3.3. Перекрестный анализ порядковых переменных.

Используемый файл: **tutorial\sample_files\satisf.sav.**

Задача: Компания-ритейлер из того же исследования заинтересована определить взаимосвязь между частотой покупок и об-

щим уровнем удовлетворенности покупателей. С учетом того, что обе переменные порядковые, необходимо определить силу и знак (направление) взаимосвязи.

Решение: выбираем меню ***Analyze->Descriptive Statistics->Crosstabs...*** затем ***Shopping frequency*** по строкам и ***Overall satisfaction*** по столбцам. Нажимаем кнопку 

и выделяем статистики ***Gamma***, ***Somer's d***, ***Kendall's tau-b***, and ***Kendall's tau-c***. Тем самым, мы получим перекрестные таблицы и измерения связи между порядковыми (***ordinal***) переменными.

Перекрестная таблица не дает четкой зависимости. Проверим гипотезу, что посетители, делающие покупки чаще, более удовлетворены.

Shopping frequency * Overall satisfaction Crosstabulation

Count

		Overall satisfaction					
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	Total
Shopping frequency	First time	5	13	15	14	5	52
	< 1/month	26	38	39	34	16	153
	1/month	27	43	46	55	30	201
	1/week	7	36	33	40	26	142
	> 1/week	1	8	10	5	10	34
Total		66	138	143	148	87	582

3.3.3.1. Парные измерения (pairwise measures).

Парные (симметричные, направленные) измерения базируются на идее подсчета согласования против антисогласования при анализе каждой пары записей в данных. Они показывают не только наличие связи, но ее силу и направленность. Пара записей в данных считается:

- **сочетаемой** (***concordant***), если большему значению переменной в ряду соответствует большее значение переменной в столбце. Согласование подразумевает положительную ассоциацию переменных в ряду и столбцах.
- **несочетаемой** (***discordant***) в противоположном случае
- **связанной по переменной** (***tied on variable***), если при одинаковом значении одной переменной другая принимает различные значения.
- **связанной** (***tied***) при равенстве обеих записей.

Измерения отличаются друг от друга способом подсчета каждой сравниваемой пары.

Gamma – разница между вероятностями (кол-во (не)сочетаемых пар/общее количество пар) сочетаемых и не сочетаемых пар без учета связанных. К сожалению, часто эта статистика исключает большое число связанных пар.

Somer's d – модификация ***Gamma***, в котором учитываются пары, связанные по переменной. Gamma уменьшается по мере роста количества пар, связанных по переменной.

Kendall's tau-b так же учитывает связанные по переменной пары, но эта связь не имеет простой интерпретации. ***Tau-b*** так же уменьшается по мере роста количества пар, связанных по переменной, как и ***Sommer's d***.

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.107	.033	3.267	.001
	Kendall's tau-c	.102	.031	3.267	.001
	Gamma	.140	.043	3.267	.001
N of Valid Cases		582			

			Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	.107	.033	3.267	.001
		Shopping frequency Dependent	.104	.032	3.267	.001
		Satisfied with overall experience Dependent	.110	.034	3.267	.001

Kendall's tau-c сходна с ***tau-b***, но может быть предпочтительней, если число категорий в переменных разное.

Правила оценки

1. Если значимость тестов < 0.05 (как в данном случае 0.001), то между переменными (в данном случае ***Shopping frequency*** и ***Overall satisfaction***) существует статистически значимая связь.

2. Величина статистик изменяется от –1 (полная отрицательная ассоциация) до + 1 (полная положительная связь). В нашем случае величины менее 0.15, то есть связь слабая положительная.

Таким образом, используя перекрестные измерения порядковых переменных, мы нашли статистическую, однако слабую, по-

зитивную связь между переменными ***Shopping frequency*** и ***Overall satisfaction***.

3.3.4. Перекрестная оценка риска события.

Используемый файл: **tutorial\sample_files\demo.sav.**

Задача: компания, продающая подписку на журналы, обычно ежемесячно делает рассылку по покупаемой базе адресов. Частота ответов низка, поэтому необходимо найти способ лучшего отбора клиентов. Одно из предложений – рассылать письма подписчикам газет в предположении, что читатели с большей долей вероятности выпишут журнал.

Решение: для проверки этого предположения оценим риск, что подписчик газеты ответит на письмо компании. Строим перекрестные таблицы для переменных ***Newspaper subscription*** (ряд) и ***Response*** (таблица), как это было описано в двух предыдущих параграфах, только в меню Statistics... отмечаем пункт ***Risk***, а нажав кнопку Cells... выберем ☑ Row.

Задание: проанализировать кросс-таблицу результатов:

Newspaper subscription * Response Crosstabulation

			Response		Total
			Yes	No	
Newspaper subscription	Yes	Count	380	2388	2768
		% within Newspaper subscription	13,7%	86,3%	100,0%
	No	Count	299	3333	3632
		% within Newspaper subscription	8,2%	91,8%	100,0%
Total		Count	679	5721	6400
		% within Newspaper subscription	10,6%	89,4%	100,0%

Ответ: с одной стороны, отклик небольшой, с другой стороны, подписчики газет откликаются чаще.

Относительный риск – это отношение вероятностей событий. В данном случае риск ответа – это отношение вероятности ответа подписчика к вероятности ответа не подписчика. Аналогично вычисляется риск не ответа (не ответ подписчика/не ответ не подписчика).

Вывод: подписчик газет в 1,6 раза выгоден для получения ответа, чем неподписчик, и в 0,94 раза выгодней как неответчик по сравнению с неподписчиком.

Отношение шансов (***odds ratio***) – отношение вероятностей шансов события. **Шанс события** (***odds of event***)– отношение вероятности свершения события к вероятности его не свершения. Из перекрестной таблицы (предыдущей таблицы) следует, что шанс ответа подписчиков 13.7%/86.3% = 0.158, шанс ответа неподписчиков 8.2%/91.8% = 0.089, соответствующее отношение шансов 0.158/0.089 = 1.775 (отношение относительного риска ответа к риску неответа 1.668/0.940 = 1.775).

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Newspaper [illegible]	1.774	1.511	2.082
For cohort Response = Yes	1.668	1.445	1.924
For cohort Response = No	.940	.924	.957
[illegible]	[illegible]		

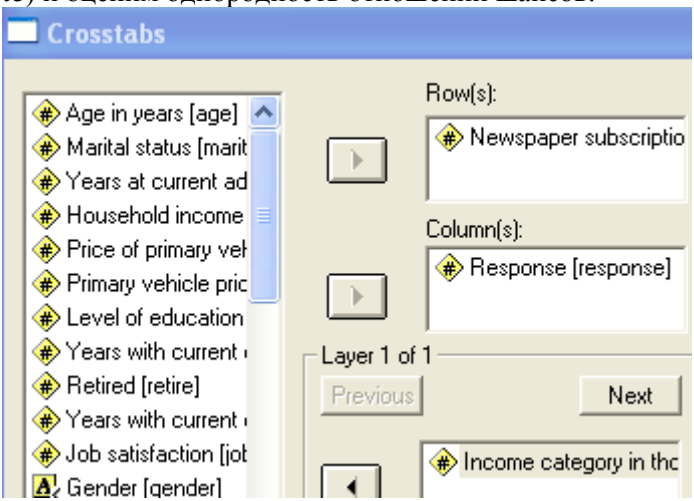
Отношение шансов как отношение отношений трудно интерпретируется. Относительный риск легче понять, тем самым отношение шансов не очень помогает. Однако есть несколько часто встречающихся ситуаций, когда оценка относительного риска не очень хороша, а вместо нее можно использовать отношение шансов:

- вероятность интересующего события мала (<0.1)
- методика анализа включает контроль и отброс некоторых событий. В этом случае относительный риск – плохая (смещенная) оценка.

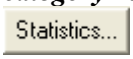
В нашем случае первое условие выполнено, а второе – нет, поэтому более надежно использовать показатель относительного риска 1.668, нежели отношение шансов 1.775.

Таким образом, отправляя рассылку только подписчикам газет, мы повысим в 1.668 раза количество откликов.

Задача: можем ли мы далее повысить эффективность рассылок? Добавим ***Income category*** как переменную слоя (***layer variable***) и оценим однородность отношений шансов.



Решение: выбираем меню ***Analyze->Descriptive Statistics->Crosstabs...*** затем ***Newspaper subscription*** по строкам и ***Response*** по столбцам и ***Income category*** как слой.

Нажимаем кнопку 

и выделяем статистики Risk и Cochran's and Mantel-Haenszel statistics.

Risk Estimate

Income category in thousands		Value	95% Confidence Interval	
			Lower	Upper
Under $25	Odds Ratio for Newspaper subscription (Yes / No)	2,513	1,830	3,451
	For cohort Response = Yes	2,146	1,652	2,789
	For cohort Response = No	,854	,805	,907
	N of Valid Cases	1174		
$25 - $49	Odds Ratio for Newspaper subscription (Yes / No)	1,829	1,418	2,357
	For cohort Response = Yes	1,701	1,361	2,125
	For cohort Response = No	,930	,900	,961
	N of Valid Cases	2388		
$50 - $74	Odds Ratio for Newspaper subscription (Yes / No)	2,182	1,387	3,434
	For cohort Response = Yes	2,056	1,351	3,128
	For cohort Response = No	,942	,909	,976
	N of Valid Cases	1120		
$75+	Odds Ratio for Newspaper subscription (Yes / No)	1,594	1,097	2,315
	For cohort Response = Yes	1,539	1,088	2,177
	For cohort Response = No	,966	,940	,992

Задание: проанализировать таблицу.

Ответ: Можно сделать предположение, что риск вроде бы уменьшается с возрастанием дохода. Проверим это предположение при помощи тестов однородности отношений шансов.

3.3.4.1. Breslow-Day and Tarone's statistics.

Проверяет однородность отношений шансов относительно переменной слоя. Иными словами, 0-гипотеза – однородность отношений шансов. Низкая значимость (обычно < 0.05 или 0.10) означает, что отношения шансов изменяются относительно переменной слоя.

В данном случае значимость большая, то есть риски однородны, поэтому наше предположение в данной статистике не подтверждается.

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	4,030	3	,258
Tarone's	4,026	3	,259

Однако попробуем применить другую оценку.

3.3.4.2. Cochran's and Mantel-Haenszel statistics.

Тестирует независимость бинарных факторных переменных (в нашем случае ***Newspaper subscription*** и ***Response***). Статистики рассчитываются на основании ковариантных матриц исследуемых переменных, построенных в разрезе одной или более переменных слоя (в нашем случае ***Income category***). 0-гипотеза тут также независимость переменных в разрезе переменной слоя. Низкая значимость показывает наличие связи, как в нашем случае:

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	68,916	1	,000
Mantel-Haenszel	68,178	1	,000

Однако, силу и направление связи данные статистики не указывают. Оценка отношения шансов есть **t-test** (статистический тест, сравнивающий средние двух групп. Если тест значим, то средние групп отличны друг от друга) общего отношения шансов. В нашем случае это полезный тест, поскольку раннее при помощи ***Breslow-Day and Tarone's statistics*** нами была установлена однородность отношений шансов. С учетом распределения по переменной Income category мы получили фактор1.976 , что лучше, чем чем фактор 1.668, полученный нами ранее из таблицы рисков. Нулевая значительность гипотезы означает, что первичный фактор (1.0 по умолчанию) несостоятелен.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1,976
ln(Estimate)			,681
Std. Error of ln(Estimate)			,083
Asymp. Sig. (2-sided)			,000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1,678
		Upper Bound	2,327
	ln(Common Odds Ratio)	Lower Bound	,518
		Upper Bound	,845

В итоге выяснили, что число откликов можно увеличить в 1,6 раза, если рассылать рекламу подписчикам газет, и до 1,9 раза, если осуществлять рассылку подписчикам с низким доходом.

3.3.5. Перекрестные оценки согласованности (association measures).

Используемый файл: **tutorial\sample_files\site.sav.**

Задача: компания планирует расширение бизнеса и ей необходимо выбрать среди 20 возможных участков для застройки. Наняты два эксперта-консультанта, которые раздельно оценивают эти участки. В итоге оба представили отчет с оценкой всех участков (хороший, средний, плохой). Необходимо определить, в какой участок инвестировать, и, по возможности, какого из экспертов оставить для дальнейшей работы.

Решение: выбираем меню ***Analyze->Descriptive Statistics->Crosstabs...*** затем ***Rating from consultant 1*** по строкам и ***Rating from consultant 2*** по столбцам. Нажимаем кнопку Statistics... и выделяем статистики ***Kappa***.

Величина ***Kappa*** статистически отлична от нуля (Value 0.429, Approx.Sig.<0.05), что означает, что рейтинги экспертов в целом совпадают за некоторыми исключениями.

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Measure of Agreement	Kappa	.429	.131	3.333	.001
N of Valid Cases		20			

Перекрестная таблица рейтингов дает обширную информацию.

Вопрос: проанализировать таблицу.

Rating from consultant 1 * Rating from consultant 2 Crosstabulation

Count

		Rating from consultant 2			
		Poor	Fair	Good	Total
Rating from consultant 1	Poor	6 (C1)	0	0	6
	Fair	5	2 (C3)	2 (C4)	9
	Good	1 (C2)	0	4	5
Total		12	2	6	20

Судя по всему, следует отказаться от участков  (С1), сосредоточить внимание на  (С4), участки  (С3) можно брать во внимание, а можно и нет, в зависимости от временных и денежных ограничений. На участке  (С2) следует остановиться особо, необходимо проанализировать подробные отчеты, почему эксперты дали столь противоречивые оценки? Возможно, обнаруженные разногласия покажут уровень экспертов (один заметил то, что не заметил другой).

3.3.6. Вопросы для самоконтроля.

1. Для каких типов задач применяется анализ перекрестных таблиц? Какие типы переменных при этом используются?
2. Для чего применяется тест хи-квадрат? Признаки значимости этого теста?
3. Какова максимальная размерность перекрестных таблиц в SPSS? Что такое переменная слоя и для чего она нужна?
4. Чем отличается перекрестный анализ порядковых и номинальных переменных?
5. Что такое парные измерения? Для чего они применяются и их сущность и используемые статистики, критерии их значимости?
6. В чем сущность метода перекрестной оценки риска? Для каких типов задач он применяется? Что такое относительный риск, отношение шансов?
7. Какие статистики применяются для оценки рисков?
8. Для решения каких задач применяется перекрестная оценка согласованности? В чем суть метода?

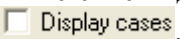
3.4. Суммирующие процедуры (Summarize procedures).

Используемый файл: **tutorial\sample_files\marketvalues.sav.**

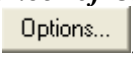
Суммирующие процедуры показывают таблицы различных статистик и/или индивидуальные таблицы записей одной или более шкалированных переменных. Целый ряд статистик показывают центральную тенденцию (значение распределения, вокруг которого группируются данные. Главные показатели – среднее, медиана, мода) и дисперсию (характеристику частотного распределения, показывающую разброс данных. Показатели – вариация,

стандартное отклонение, межквартильное расстояние). Суммарные показатели могут быть подсчитаны по всем данным, или по группам, или по отобранной выборке.

Задача: Агентство недвижимости помогает клиенту продать квартиру. Собрав информацию о продажах за год, оно готовит отчет.

Решение: выбираем меню ***Analyze->Reports->Case Summaries...*** затем ***Purchase Price*** как переменную, и ***House street*** в качестве группирующей переменной (***grouping variable***). Поскольку у нас групповой рейтинг, отдельные списки данных нам не нужны,и мы деактивируем их показ 

. Нажимаем кнопку 

и выделяем для показа статистики ***Mean***, ***Median***, ***Minimum***, и ***Maximum***. Обратите внимание, что переменная ***Number of Cases*** появляется в списке автоматически. Нажав кнопку 

, вводим заголовок и подпись к суммарным статистикам. Допустим, клиент живет на ***Fairway View Drive***.

Home Sale Statistics

Purchase Price

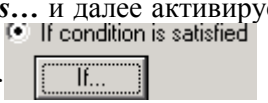
House Street	N	Mean	Median	Minimum	Maximum
Bunker Hill Dr	28	$279,821.43	$281,000.00	$200,000	$416,000
Dawson Ln	23	$131,391.30	$132,000.00	$118,000	$140,000
Fairway View Dr	7	$283,000.00	$271,000.00	$243,000	$343,000
Lakeview Dr	8	$304,000.00	$300,500.00	$289,000	$334,000
Par Dr	7	$303,714.29	$305,000.00	$271,000	$349,000
Persimmon Dr	9	$312,111.11	$300,000.00	$281,000	$351,000
Wintergreen Te	12	$322,833.33	$317,500.00	$273,000	$403,000
Total	94	$256,159.57	$281,000.00	$118,000	$416,000

Grouped by Street

Задание: проанализировать данные.

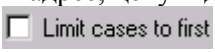
Ответ: среднее сдвинуто относительно медианы (которая лучше, как оценка центра распределения), причина – как минимум, один существенно более дорогой дом.

Задача: клиент желает продать квартиру не менее, чем за $335000. Он просит распечатать список домов в районе дороже данной суммы и время, потребовавшееся для продажи.

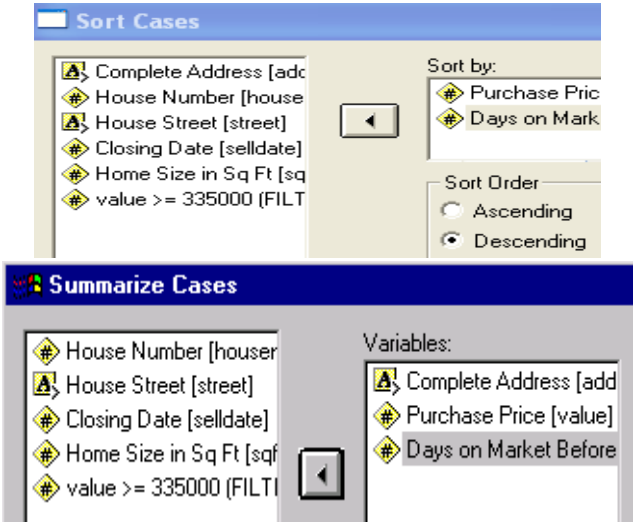
Решение: 1. Для выбора домов выбираем меню ***Data->Select Cases…*** и далее активируем отбор по условию и нажимаем кнопку ***If*** 

2. Сортируем данные по цене и дням на рынке в обратном порядке (***Descending***) через ***Data->Sort Cases…***(рис. слева).

3. Создаем отчет через ***Analyze->Reports->Case Summaries...***Используем кнопку 

для сброса предыдущих данных, выводим адрес, цену и дни продажи, дезактивируйте ограничения вывода 

(рис. справа):



Нажав 

, выбираем среднее и медиану, в 

название и подпись, наконец, ***OK*** выводит конечную таблицу.

3.4.1. Вопросы для самоконтроля.

1. Для чего нужны суммирующие процедуры?
2. Перечислите главные показатели.
3. Приведите методику оценки симметричности распределения переменной.
4. Приведите методику оценки нормальности распределения переменной.
5. Каким образом вывести и распечатать список определенных данных в SPSS?

3.5. Расчет средних (Means procedure).

Расчет средних полезен как для анализа, так и описания шкалируемых переменных. Используя описательные свойства, можно использовать множество статистик для получения центральной тенденции и дисперсии. Любое количество переменных могут быть использованы для группировки, или определенные ячейки данных могут быть отобраны для формирования групп. Односторонний (***one-way***) ***ANOVA*** может быть использован для оценки разности между группами. Вместе с тестами линейности и ассоциативными измерениями, ***ANOVA*** позволяет проверять структуру и силу связи между переменными и их группами.

Используемый файл: **tutorial\sample_files\hourlywagedata.sav.**

Задача: Для анализа заработной платы медсестер журнал собрал информацию о часовых заработной платах медсестр на разных позициях и с разным уровнем опыта. Проанализировать связь уровня зарплат с уровнем опыта и типом позиции.

Решение: выбираем меню ***Analyze->Compare Means->Means...*** затем ***Hourly Salary*** и ***Years Experience*** в качестве зависимой (***Dependent***) и независимой (***Independent***) переменных соответственно.

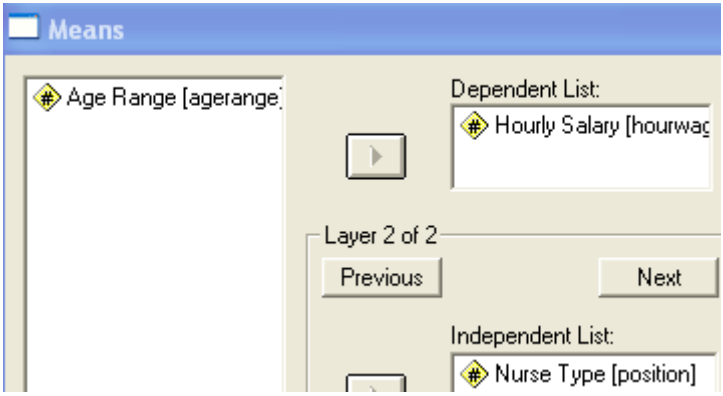
Вопрос: проанализировать таблицу.

Hourly Salary

Years Experience	Mean	N	Std. Deviation
5 or less	18.0416	221	3.86667
6-10	18.9169	460	3.77816
11-15	19.6616	752	3.90528
16-20	20.2876	729	3.82786
21-35	21.2594	539	4.08669
36 or more	21.6342	210	3.61826
Total	20.0159	2911	4.00309

Ответ: заработная плата растет с опытом.

Вместе с тем заработок медсестер, работающих в офисе поликлиники (office) и в стационаре (hospital), возможно, отличаются из-за неодинаковых начальных заработных плат и их временных изменений. Для проверки воспользуемся таким типом позиции как переменная слоя. Возвращаясь в мастер среднего, нажимаем кнопку Next и вводим вторую, независимую переменную ***Nurse Type***.



Задание: проанализировать таблицы результатов.

Hourly Salary

Years Experience	Nurse Type	Mean	N	Std. Deviation
5 or less	Hospital	19,0753	147	3,37129
	Office	15,9882	74	3,98762
	Total	18,0416	221	3,86667
6-10	Hospital	19,4846	313	3,35218
	Office	17,7082	147	4,32447
	Total	18,9169	460	3,77816
11-15	Hospital	20,2412	518	3,41065
	Office	18,3784	234	4,57662
	Total	19,6616	752	3,90528
16-20	Hospital	21,1369	471	3,29487
	Office	18,7373	258	4,23293
	Total	20,2876	729	3,82786
21-35	Hospital	21,8601	350	3,48989
	Office	20,1471	189	4,82372
	Total	21,2594	539	4,08669
36 or more	Hospital	22,0641	146	3,14466
	Office	20,6534	64	4,38931
	Total	21,6342	210	3,61826

Report

Hourly Salary

Years Experience	Nurse Type	Mean	N	Std. Deviation
5 or less	Hospital	19,0753	147	3,37129
	Office	15,9882	74	3,98762
	Total	18,0416	221	3,86667
6-10	Hospital	19,4846	313	3,35218
	Office	17,7082	147	4,32447
	Total	18,9169	460	3,77816
11-15	Hospital	20,2412	518	3,41065
	Office	18,3784	234	4,57662
	Total	19,6616	752	3,90528
16-20	Hospital	21,1369	471	3,29487
	Office	18,7373	258	4,23293
	Total	20,2876	729	3,82786
21-35	Hospital	21,8601	350	3,48989
	Office	20,1471	189	4,82372
	Total	21,2594	539	4,08669
36 or more	Hospital	22,0641	146	3,14466
	Office	20,6534	64	4,38931
	Total	21,6342	210	3,61826

Ответ: видно, что офисные медсестры зарабатывают в среднем меньше, однако с опытом разница уменьшается. Эффект, возможно, связан с набором требований к медсестрам. От офисных медсестер требуется стабильный набор умений общего свойства. От

сестер в стационаре требуются более динамичные и специализированные умения.

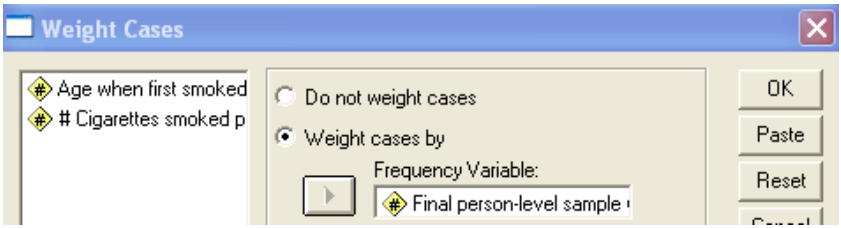
Используемый файл: **tutorial\sample_files\smokers.sav**.

Задача: исследуется курение среди молодежи. Гипотеза исследования – более тяжелые курильщики начинают курить в более раннем возрасте, причем зависимость линейная.

В исходном файле представлены данные национального исследования потребления наркотиков, являющейся вероятностной выборкой в отношении домохозяйств.

Решение: 1. Шаг – взвесим данные таким образом, чтобы они отражали распределение населения.

Выбираем меню ***Data->Weight Case…*** затем ***Final person-level sample weight*** в качестве весовой частотной переменной:



2. Выбираем меню ***Analyze->Compare Means->Means...*** затем ***Age when first smoked a cigarette*** и ***# Cigarettes smoked per day past 30 days*** в качестве зависимой (***Dependent***) и независимой (***Independent***) переменных соответственно. Нажимаем Options... ***Anova table and eta*** и ***Test for linearity***.

Вопрос: проанализировать таблицу-отчет средних.

Report

Age when first smoked a cigarette

# Cigarettes smoked per	Mean	N	Std. Deviation
1 to 5 cigarettes each day	15,81	1119	4,452
6 to 15 cigarettes (about 1/2 pack) each	15,89	1594	4,820
16 to 25 cigarettes (about 1 pack) each	15,63	1604	5,450
26 to 35 cigarettes (about 1 1/2 pk) eac	14,18	622	4,066
35 or more cigarettes (about 2 packs) ea	14,45	461	4,376
Total	15,48	5400	4,866

Ответ: действительно, те, кто курит меньше, начали курить позже.

3.5.1. Таблица ANOVA.

Проверяет разницу между средними разных групп. Полная вариация разделяется на две компоненты:

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	Between Groups (C1)	(Combined)	1974,095	4	493,524	21,2	,000 (C3)
	C4	Linearity	1321,500	1	1321,5	56,7	,000 (C5)
		Deviation from Linearity	652,595	3	217,532	9,33	,000
	Within Groups		125841	5395	23,326		
	Total (C2)		127815	5399			

Межгрупповая (***Between***) (C1) описывает вариацию средних групп относительно общего среднего.

Внутригрупповая (***Within Groups***) (C2) описывает вариацию индивидуальных показателей относительного среднего группы.

Наличие связи определяется показателем значимости теста (C3). Если Sig.<0.05, это показывает наличие взаимосвязи, как в нашем случае. При наличии взаимосвязи в таблице появляется групповой раздел (***Groups***) (C4), в котором оцениваются два типа зависимости – линейная (***Linearity***) и нелинейная (***Deviation from Linearity***)[1]. Наличие зависимостей показывают значимости

[1] Для правильности расчетов независимая переменная должна быть отсортирована по категориям!

(C5), в нашем случае существует и линейная и нелинейная компоненты связи переменных.

3.5.2. Ассоциации переменных (association).

Четыре показателя взаимосвязи переменных предлагаются в таблице ассоциаций:

Measures of Association

	R	R Squared	Eta	Eta Squared
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	-,102	,010	,124	,015

R, $\mathbf{R^2}$ – коэффициенты детерминации используются в случае линейной зависимости переменных. R^2 показывает процент вариации, описываемый линейной зависимостью между перемеными. В данном случае линейная зависимость составляет всего 1% вариации.

$\boldsymbol{\eta}$, $\boldsymbol{\eta^2}$ – коэффициенты детерминации без предположения о линейной зависимости переменных. η^2 показывает процент вариации, объясняемый разницей между группами, 1,5% в данном случае. Данный показатель необходимо использовать в случае существования заметной нелинейной зависимости между переменными, когда R^2 и η^2 существенно отличаются друг от друга. В данном случае, оба показателя близки друг другу и малы.

Таким образом, взаимосвязь была обнаружена, однако в силу ее слабости (1–1,5%) практическая значимость ее не очевидна.

3.5.3. Вопросы для самоконтроля.

1. Для чего необходим расчет средних?
2. Для чего применяется взвешивание данных при анализе средних?
3. В чем различие линейной и нелинейной взаимосвязи между переменными? Приведите графически переменные с линейной и нелинейной корреляцией.
4. Что такое межгрупповая и внутригрупповая вариации?

5. Как оценивается ассоциация переменных. Смысл и границы применимости показателей η^2 и R^2?

3.6. OLAP-куб анализ (OLAP Cube procedure).

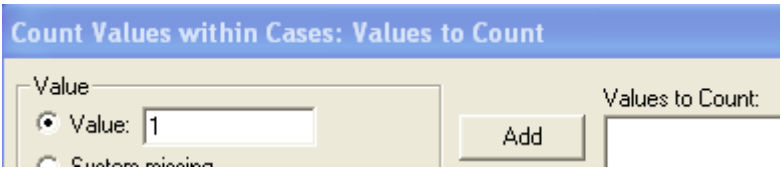
OLAP-куб анализ (OnLine Analytical Processing) – еще один способ расчета суммарных статистик для переменных, разбитых по категориям одной или более группирующих переменных. Широкие и гибкие возможности представления и расчетов в разрезе групп делают удобным применение OLAP именно для межгруппового анализа показателей.

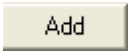
Задача: Телекоммуникационная фирма хочет уменьшить «текучку» (churn) – процент пользователей, перешедших к другому провайдеру. Проанализировать ситуацию с пользователями, сменившими и не сменившими провайдеров в течение последнего месяца в трех районах деятельности компании.

Используемый файл: **tutorial\sample_files\telco.sav.**

Решение: 1. Для начала необходимо создать переменную, показывающую количество сервисов, используемых пользователем. Выбираем меню ***Transform->Count…*** и создаем переменную services в качестве целевой переменной (***Target Variable***) с описанием (***Target Label***) ***# of services***. В качестве переменных (***Numeric Variables***) выбираем переменные от Multiple lines до Electronic billing и нажимаем кнопку 

. В открывшемся окне введите 1 как значение

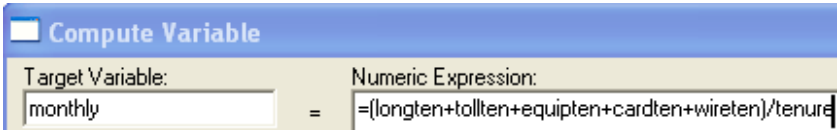


и нажимаем кнопку 

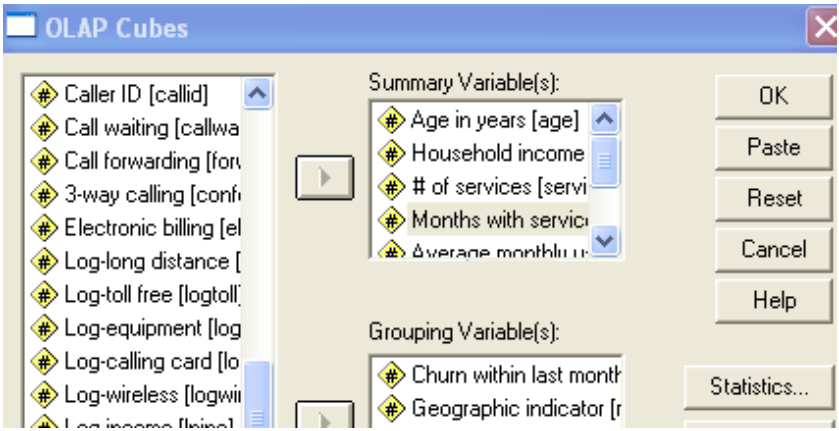
потом ***OK*** – переменная services появляется в данных.

2. Другой важный показатель активности пользователя – среднемесячная стоимость используемых сервисов. Вводим соответствующую новую переменную ***monthly*** (под именем ***Average monthly expense***) через ***Transform->Compute…*** как среднее пере-

менных от ***Long distance over tenure*** до ***Wireless over tenure*** по переменной ***tenure***.



3. Для OLAP-анализа выбираем ***Analyze->Reports->OLAP Cubes....***, где в качестве суммарных переменных (***summary variables***) ***Months with service***, ***Age in years***, ***Household income in thousands***, ***# of services***, и ***Average monthly expense***. А в качестве групповых переменных ***Churn within last month*** и ***Geographic indicator***:



В меню ***Statistics*** дезактивируйте ***Sum***, ***Percent of Total Sum***, и ***Percent of Total N*** и выберите ***Mean***. С помощью 

установите название и подпись. Нажав, ***OK*** в итоге получаем таблицу:

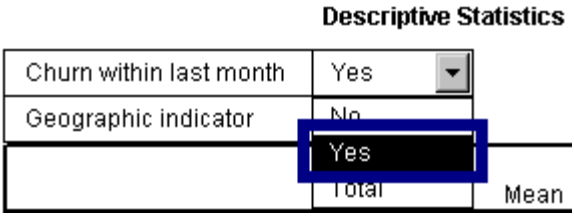
Descriptive Statistics

Churn within last month: Total
Geographic indicator: Total

	N	Mean	Std. Deviation	Median
Months with service	1000	35.53	21.360	34.00
Age in years	1000	41.68	12.559	40.00
Household income in thousands	1000	77.5350	107.0442	47.0000
# available services	1000	3.7400	2.60138	3.0000
Average monthly expense	1000	62.3230	50.04173	48.8956

By Customer Churn and Geographic Region

Суммарная статистика, это полезно, но ее можно получить (и мы ранее получали) другими способами. Специфика OLAP-анализа – возможность посмотреть в той же таблице данные по разным группам. Активация таблицы происходит двойным щелчком, при этом по групповым переменным можно осуществить выбор необходимых значений. Например, для обзора всех пользователей, сменивших провайдера в последнем месяце, выбираем ***Churn within last month – Yes***.



Задание: проанализировать полученную таблицу.

Churn within last month: Yes

Geographic indicator: Total

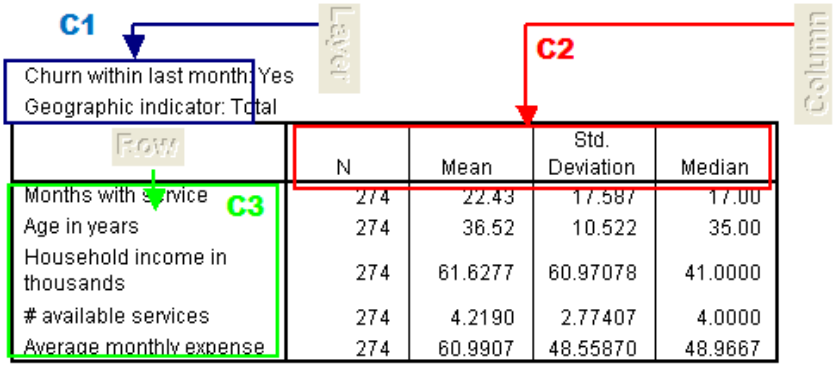
	N	Mean	Std. Deviation	Median
Months with service	274	22.43	17.587	17.00
Age in years	274	36.52	10.522	35.00
Household income in thousands	274	61.6277	60.97078	41.0000
# available services	274	4.2190	2.77407	4.0000
Average monthly expense	274	60.9907	48.55870	48.9667

By Customer Churn and Geographic Region

Ответ: 274 человека сменили провайдера. Разница медианы и среднего переменных Household income in thousands и Average monthly expense сигнализирует о наличии правых хвостов.

3.6.1. Скроллинг таблиц (pivoting table).

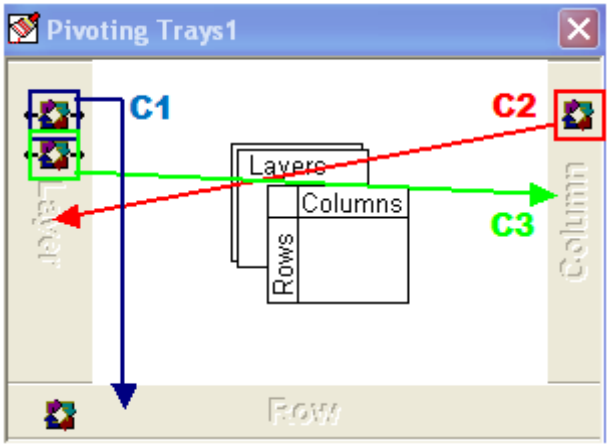
Во многих таблицах результатов можно манипулировать переменными в рядах (***Row***) (С3), столбцах (***Column***) (С2) и слоях (***Layer***) (С1):



	N	Mean	Std. Deviation	Median
Months with service	274	22.43	17.587	17.00
Age in years	274	36.52	10.522	35.00
Household income in thousands	274	61.6277	60.97078	41.0000
# available services	274	4.2190	2.77407	4.0000
Average monthly expense	274	60.9907	48.55870	48.9667

By Customer Churn and Geographic Region

В настоящем случае, хотя в OLAP-таблице можно посмотреть таблицы любых комбинаций групповых переменных, однако удобно просмотреть все детали прокруткой (скроллингом, **pivoting out**) групповых переменных в рядах или столбцах. Для этого активируем двойным кликом таблицу и из меню обозревателя выбираем ***Pivot->Pivot Trays***. Перетаскиваем значок переменной ***Churn within last month*** в область Ряда (***Row***) (C1); статистику из области Колонок (***Column***), в раздел Слой (***Layer***) (C2); географический фактор – из раздела Слой в раздел Колонок (C3), из Статистики (переменная слоя теперь) выбираем показ среднего (***Mean***):



Задание: проанализировать полученную таблицу.

Mean

	Churn within last month	Geographic indicator			
		Zone 1	Zone 2	Zone 3	Total
Age in years	No	42,89	43,88	44,08	43,63
	Yes	37,59	36,32	35,66	36,52
	Total	41,41	41,75	41,88	41,68
Household income in thousands	No	77,4569	85,8375	86,9213	83,5386
	Yes	63,0889	62,3191	59,4444	61,6277
	Total	73,4410	79,2186	79,7326	77,5350
# of services	No	3,5733	3,5000	3,6024	3,5592
	Yes	4,2111	4,2234	4,2222	4,2190
	Total	3,7516	3,7036	3,7645	3,7400
Months with service	No	39,85	40,25	41,24	40,47
	Yes	23,00	23,05	21,21	22,43
	Total	35,14	35,41	36,00	35,53
Average monthly user payment	No	62,3930	63,1913	62,8760	62,8259
	Yes	61,4338	60,2713	61,2990	60,9907
	Total	62,1249	62,3695	62,4634	62,3230

C2 — Household income in thousands, Yes

C1 — Months with service, Yes

C3 — Average monthly user payment, Yes

Ответ:

1. Сменившие провайдера пользовались сервисом почти в два раза меньше (C1).

2. При примерно одинаковой интенсивности пользования услугами (C3) это люди в среднем с меньшими доходами (C2).

Перейдем от изучения таблицы средних к таблице медиан. Для этого активируем OLAP-таблицу (двойной щелчок) и выведем на экран данные по медиане (***Median***):

OLAP Cubes

Median

	Churn within last month	Geographic indicator			
		Zone 1	Zone 2	Zone 3	Total
Age in years	No	42,00	43,00	43,00	43,00
	Yes	35,00	35,00	32,50	35,00
	Total	C2 40,00	41,00	40,00	40,00
Household income in thousands C1	No	47,0000	52,0000	49,0000	49,0000
	Yes	51,0000	40,0000 C3	37,5000	41,0000
	Total	47,5000	48,0000	46,0000	47,0000
# of services	No	3,0000	3,0000	3,0000	3,0000
	Yes	4,0000	3,0000	4,0000	4,0000
	Total	3,0000	3,0000	4,0000	3,0000
Months with service	No	40,50	42,00	42,00	41,50
	Yes	20,50	17,50	12,50	17,00
C1	Total	33,00	34,50	35,00	34,00
Average monthly user payment	No	49,6790	50,8417	46,8825	48,8956
	Yes	49,2440	50,5370	46,3302	48,9667
	Total	49,5103	50,7997	46,5078	48,8956

Задание: проанализировать полученную таблицу.

Ответ:

1. Доходы и траты пользователей существенно ниже, чем в предыдущей таблице средних (C1, правые хвосты распределений), очевидно, что данные оценки средних ближе к истине, чем средние показатели.

2. Заметна аномалия в зоне 1, где доходы отключившихся больше доходов оставшихся клиентов (C2), что против средней тенденции.

3. Заметна географическая корреляция, где доходы отключившихся заметно спадают от зоны 1 к зоне 3 (C3).

Таким образом, пользуясь OLAP-таблицами, мы быстро просмотрели срезы данных в разных проекциях, фокусируясь на отключившихся пользователях. Полученные результаты помогут в дальнейшем исследовании проблемы.

3.6.2. Вопросы для самоконтроля.

1. Как расшифровывается аббревиатура OLAP?
2. Для каких целей применяется OLAP-анализ?
3. Для чего и каким образом осуществляется скроллинг таблиц?

3.7. T-тесты (t-tests).

T-тест – статистическая методика, нацеленная на сравнительный анализ групповых показателей по различным выборкам, а также сравнение экспериментальных данных с теоретическими предсказаниями и моделями.

3.7.1. T-тест на единичной выборке (one sample t-test).

T-тест на единичной выборке позволяет:

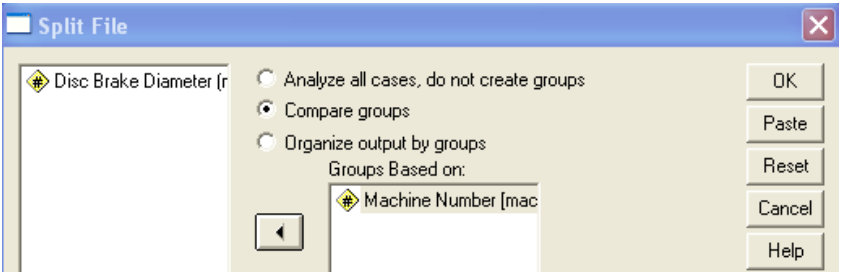
- Проверить разницу между средним выборки известной гипотетической величиной.
- Позволяет варьировать уровень достоверности (***level of confidence***)
- Выдает таблицы описательных статистик для каждой тестируемой переменной.

Используемый файл: **tutorial\sample_files\brakes.sav.**

Задача: автомобильная компания производит тормозные диски (***disc brakes***) которые должны быть 322 мм в диаметре. При контроле качества случайным образом отбирается 16 дисков, сделанных каждым из 8 станков, для измерения диаметра. Определить, правильно ли настроены автоматы.

Решение: используем т-тест для проверки, значимо ли отклоняются средние каждой из 8 выборок от требуемого диаметра 322 мм.

1. В файле данных присутствует номинальная переменная ***Machine number***. Поскольку данные по каждой машине тестируются отдельно, разделим файл данных на группы данной переменной через вызов ***Data->Split File…*** (см. Раздел 2.4).



2. Выбираем ***Analyze->Compare Means->One-Sample T Test...***

Выбираем меню Disc Brake Diameter (mm) в качестве тестируемой переменной (***Test Variable***), и выставляем тестируемое значение Test Value: 322 В Options... выбираем 90% доверительный интервал Confidence Interval: 90 % затем Continue и OK.

В то время как описательная таблица (***One-Sample Statistics***) показывает уже виденные нами неоднократно среднее, отклонение и ошибку, т-таблица (***One-Sample Test***) показывает результаты статистики. Ее содержание:

One-Sample Statistics

Machine Number		N	Mean	Std. Deviation	Std. Error Mean
1	Disc Brake Diameter (mm)	16	321,9985	,0111568	,0027892

C5

$t = (Mean - Test\ Value)/Std.Deviation$

One-Sample Test

Machine Number		Test Value = 322				C6
		C1	C2	C3	C4	90% Confidence Interval of the Difference
		t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
1	Disc Brake Diameter (mm)	-,533	15	,602	-,00149	-,006375	,00340
2	Disc Brake Diameter (mm)	5,336	15	,000	,014263	,009577	,01895
3	Disc Brake Diameter (mm)	-,655	15	,522	-,00172	-,006311	,00288
4	Disc Brake Diameter (mm)	-2,61	15	,020	-,00456	-,007628	-,0015
5	Disc Brake Diameter (mm)	1,847	15	,085	,004249	,000216	,00828
6	Disc Brake Diameter (mm)	1,134	15	,274	,002452	-,001337	,00624
7	Disc Brake Diameter (mm)	2,650	15	,018	,006181	,002092	,01027
8	Disc Brake Diameter (mm)	-1,71	15	,107	-,00330	-,006680	,00008

- ***t***-колонка (C1) содержит т-статистику, которая есть отношение разницы среднего по выборке и тестируемого значения к величине стандартной ошибки (C5);
- ***df***-колонка (C2) показывает число степеней свободы (degree of freedom), которое есть количество записей в выборке минус единица N-1;
- значимость (Sig. 2-tailed) (C3) показывает вероятность получения значения, равного или больше полученной t-статистики с df степенями свободы, в случае чисто случайного

разброса между средним по выборке Mean и тестируемым значением Test value. Чем меньше это значение, тем более значимо отличие среднего выборки и тестируемого значения.

- Среднее отклонение (***Mean Difference***) (C4) показывает разницу между средним по выборке и тестируемым значением.
- Доверительный интервал разницы (C6) показывает допустимый диапазон, в котором лежит с заданной вероятностью (90% в данном случае) истинное среднее отклонение для каждой машины. Отсутствие 0 в допустимом диапазоне показывает значительность отклонения.

Задание: проанализировать т-таблицу

Ответ: Из доверительных интервалов (C6) видно, что станки 2-й,5-й,7-й делают диски большего диаметра, а станок 4-й – меньшего, чем необходимо.

3.7.2. T-тест по парным выборкам (Pair-sample t-test).

В экспериментальных исследованиях очень часто используется метод «до-и-после» (**pre-post design**). В этом случае измерения проводят до и после некоторого воздействия (применения лекарства, осуществления некоторого события и т.д.). 0-гипотеза исследования – это отсутствие эффекта и равенство средних по обоим измерениям. С другой стороны, наличие эффекта от воздействия (намеренного или непреднамеренного) дает ненулевую значимую разницу средних и отрицает 0-гипотезу.

Т-тест на парных выборках предназначен для проверки 0-гипотезы. Данные могут представлять собой пару измерений одного объекта, или одно измерение, разбитое на парные выборки. Дополнительно считаются:

- Описательные статистики для каждой анализируемой переменной.
- Корреляция Пирсона и ее значимость.
- Доверительный интервал с изменяемой по вашему усмотрению достоверностью (95% по умолчанию)

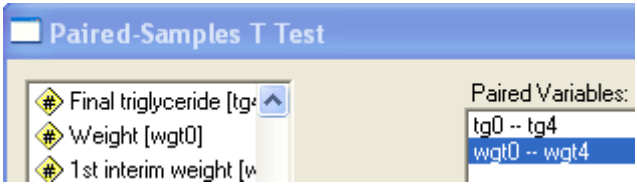
Используемый файл: **tutorial\sample_files\dietstudy.sav.**

Задача: физиолог оценивает влияние новой диеты на своих пациентов с наследственной предрасположенностью к сердечным заболеваниям. Для оценки диеты 16 пациентов сидели на диете в течение 6 месяцев. Их веса и уровень триглицерида (***triglyceride***) в

крови были измерены до и после диеты. Проанализировать результаты.

Решение: используем т-тест по парным выборкам для оценки значимости отличий весов и уровня триглецирида.

Берем ***Analyze->Compare Means->Paired-Samples T Test...*** Выбираем ***Triglyceride*** и ***Final Triglyceride*** в качестве первой пары переменных (***Paired Variables***), а ***Weight*** и ***Final Weight*** – в качестве второй:



После нажатия ***OK*** получаем результаты.

Задание: проанализировать статистику парных выборок (***Paired Samples Statistics***).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Triglyceride	138,44	16	29,040	7,260
	Final triglyceride	124,38	16	29,412	7,353
Pair 2	Weight	198,38	16	33,472	8,368
	Final weight	190,31	16	33,508	8,377

Ответ: сравнение средних показывает, что и уровень триглицерина и веса упал на 14–15 и 8 пунктов соответственно. При этом разброс показателей в относительном выражении (***Std. Deviation/Mean***) больше по триглицерину.

Задание: какие выводы можно сделать из корреляций по парным выборкам (***Paired Samples Correlations***).

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Triglyceride & Final triglyceride	16	-,286	,283
Pair 2	Weight & Final weight	16	,996	,000

Ответ: несмотря на то, что в целом уровень триглецирина ниже после диеты, корреляция здесь статистически незначима (Sig.>0.05). Это означает, что наряду с больными, у которых понизился уровень триглицерина, есть пациенты, у которых уровень не изменился, либо увеличился[2]. С другой стороны, корреляция (Пирсона) по весу практически близка с 100 процентам и значима. То есть все пациенты теряли вес.

Задание: проанализировать таблицу с результатами т-теста:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Triglyceride - Final triglyceride	14,063	46,875	11,719	-10,915	39,040	1,200	15	,249
		C1	C2	C3	C4		C5		
Pair 2	Weight - Final weight	8,063	2,886	,722	6,525	9,600	11,17	15	,000

Ответ: в т-таблице по парным выборкам указывается:

- средняя разница между выборками (C1);
- ее разброс (C2);
- ожидаемый показатель среднего при повторении аналогичных исследований (опросов по 16 пациентов) (C3);
- диапазон значений, в котором с вероятностью 95% будет лежать средняя разница выборок при повторении исследований (C4);
- показатели т-статистики (C5), аналогичные описанным в предыдущем параграфе.

Т-тест показывает нулевую значимость для веса, что означает, что полученную разницу в весе нельзя объяснить случайным разбросом, а можно приписать диете. Для уровня триглицерина зна-

[2] При отсутствии значимой корреляции (малое значение ***Correlation*** и большое ***Sig.***), как в этом случае, можно попробовать т-тест независимых выборок.

чимость теста высокая, тем самым диета не является существенным фактором снижения уровня триглицерина.

3.7.3. Т-тест независимых выборок (Independent-sample t-test).

Проверяет значимость разницы между средними независимых выборок. Также изображаются:

- описательные статистики переменных;
- тест идентичности вариаций;
- интервалы достоверности

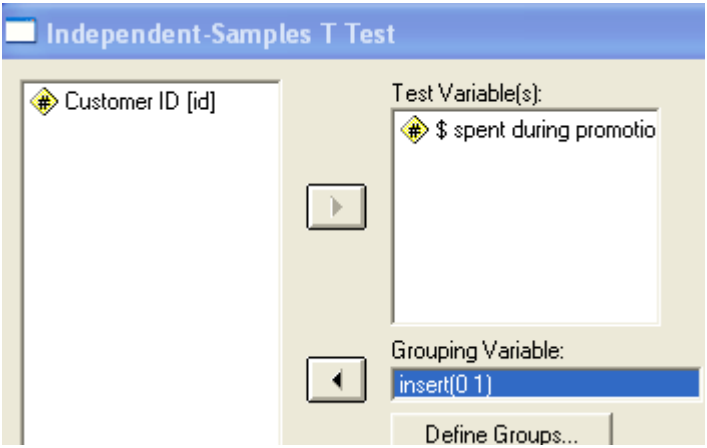
Обычно, группирующая переменная для т-тестов фиксирована и имеет одно значение для каждой группы (например, мужчины и женщины). Однако, иногда группировка делается на базе шкалируемой переменной, и тогда формирование групп зависит от выбора порогового значения (***cut point***). Программа делит данные согласно заданному порогу и делает т-тест. Достоинство метода в том, что можно изменять пороговое значение без необходимости каждый раз пересоздавать групповую переменную для нового порога.

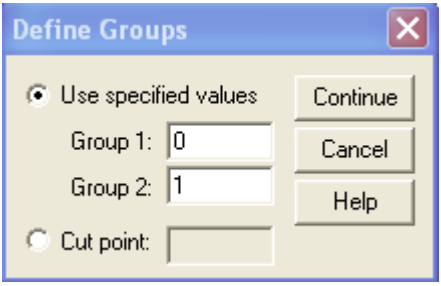
Используемый файл: **tutorial\sample_files\creditpromo.sav**

Задача: аналитик отдела продаж хочет оценить результаты недавней рекламной компании кредитных карт. Для этого выбрано случайным образом 500 держателей карт. Половина из них – владельцы карт с рекламируемым дисконтом на покупки в течение следующих трех месяцев, другая половина – владельцы обычных карт. Проанализировать поведение отобранных групп.

Решение: используем т-тест независимых выборок для оценки значимости отличий в поведении групп.

Берем ***Analyze->Compare Means->Independent-Samples T Test...*** Выбираем ***$ spent during promotional period*** в качестве тестируемой переменной (***Test Variable(s)***), а ***Type of mail insert received*** в качестве группируемой переменной (***Grouping variable***)





В меню Define Groups... выбираем значения 0 и 1 для первой и второй групп соответственно. Нажимаем Continue и ***OK*** соответственно

Задание: проанализировать результаты групповой статистики (***Group Statistics***)

Group Statistics

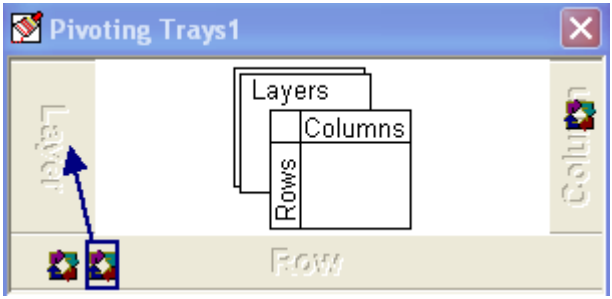
	Type of mail insert received	N	Mean	Std. Deviation	Std. Error Mean
$ spent during promotional period	Standard	250	1566,3890	346,67305	21,92553
	New Promotion	250	1637,5000	356,70317	22,55989

Ответ: расходы владельцев новых карт возросли примерно на 70 пунктов, при этом разброс расходов (***Std.Deviation***) также вырос на 10 пунктов.

Задание: проанализировать результаты теста (***Independent Sample Test***)

Ответ: вначале необходимо установить, эквивалентен или нет разброс (вариация, variation) внутри сравниваемых групп. Проверкой этого служит тест Левене (***Levene's Test***), 0-гипотеза которого есть равенство вариаций сравниваемых средних. Поскольку в нашем случае значимость теста высока (Sig.>0.10), это означает, что обе группы пользователей имеют одинаковый разброс и не рассматривать второй вариант (разный разброс у сравниваемых групп). При помощи прокрутки переменных таблицы (pivoting tables, см. Раздел 3.6.1) меняем вид таблицы, перемещая переменную ***Assumptions*** в разряд переменных слоя (см. картинку).

		Levene's Test for Equality of Variances	
		F	Sig.
$ spent during promotional	Equal variances assumed	1,190	,276
	Equal variances		



в итоге получаем таблицу:

Independent Samples Test

Equal variances assumed

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
$ spent during promotional period	1,190	,276	-2,26	498	,024	-71,1	31,46	-132,9	-9,302

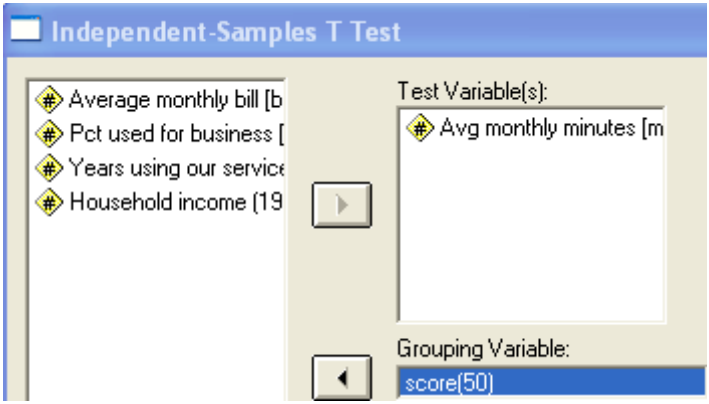
из которой видно, что разница $71 не является игрой случая (Sig<0.05, допустимый диапазон не содержит 0, подробней параметры т-статистики описаны в Разделе 3.7.1)

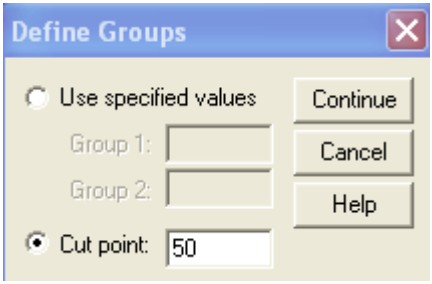
Используем файл **tutorial\sample_files\cellular.sav.**

Задача: в сотовой компании анализируется склонность пользователей к уходу. Рейтинг колеблется от 0 до 100, значение 50 и выше рассматривается как возможность ухода клиента. Необходимо сравнить 50 пользователей с рангом выше порогового с 200 пользователями ниже порога по количеству минут, используемых в месяц.

Решение: используем т-тест независимых выборок для оценки значимости отличий в поведении групп.

Берем ***Analyze->Compare Means->Independent-Samples T Test...*** Выбираем ***Avg monthly minutes*** в качестве тестируемой переменной (***Test Variable(s)***), а ***Propensity to leave*** в качестве группируемой переменной (***Grouping variable***)





В меню Define Groups... выбираем пороговое значение (***Cut point***) 50. Нажимаем Continue и ***OK*** соответственно.

Задание: проанализировать результаты

Ответ: по аналогии с предыдущей задачей видим, что

- описательная статистика показывает, что группа склонных к смене провайдера пользуется связью более интенсивно.
- Вариации средних по группам одинаковы согласно тесту Левене, с помощью прокрутки (Pivoting) отображаем в таблице только результаты теста в предположении одинаковых вариаций средних.
- Т-таблица показывает не случайность разницы по группам (78 минут), причем эта разница с 95% вероятностью не ниже 67 минут.

Вывод: компании необходимо изыскивать методы удержания этих клиентов, возможное решение – бонусы для многоговорящих абонентов.

Equal variances assumed

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Avg monthly minutes	,966	,327	14,35	248	,000	78,256	5,45394	67,514	88,998

3.7.4. Вопросы для самоконтроля.

1. Для чего применяются т-тесты? Какие три основные вида т-тестов используются при решении задач?

2. Каковы основные параметры результирующей таблицы т-теста на единичной выборке. Их трактовка?

3. Для каких типов задач используется т-тест по парным выборкам? Какова 0-гипотеза этого теста и какие дополнительные показатели можно рассчитать в SPSS?

4. Для каких задач применяется т-тест независимых выборок? Каковы критерии его значимости? Какие статистики считаются дополнительно в SPSS?

5. Для чего и для каких типов переменных можно использовать пороговое значение переменной в т-тесте независимых выборок?

6. В чем сущность теста Левене (0-гипотеза, критерии значимости)?

3.8. Односторонний вариационный анализ (One-way ANOVA).

Односторонний (однонаправленный, ***One-way***) вариационный анализ ANOVA (***ANalysis Of VAriance***) используется для проверки гипотезы, что средние двух или более групп не являются статистически значимыми. Кроме того, ANOVA позволяет:

- Оценить статистику зависимой переменной в разрезе групп.
- Провести тест вариационной идентичности
- Построить график групповых средних
- Провести ранговые тесты (***Range test***), попарные множественные сравнения и контраст-тесты для определения природы различий в группах.

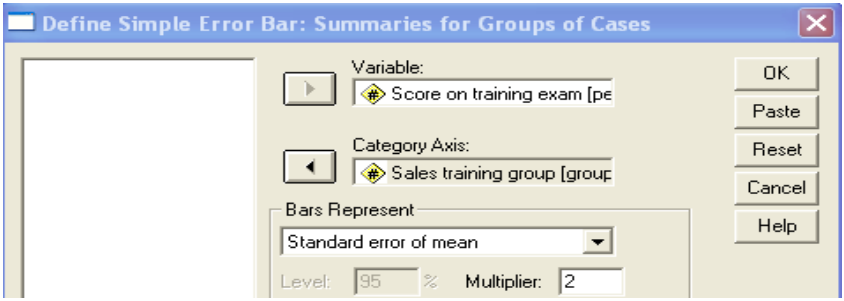
3.8.1. Эквивалентность вариаций групп.

Важным первым шагом анализа является проверка априорных условий, при которых можно применять ANOVA. Одно предположение ANOVA о том, что вариации групп эквивалентны.

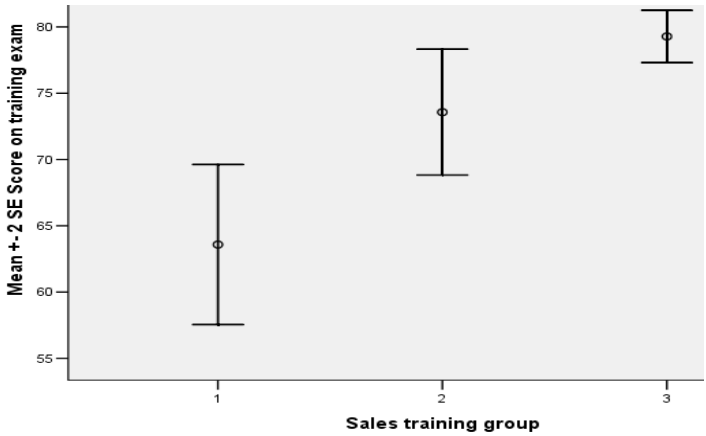
Используемый файл: **tutorial\sample_files\salesperformance.sav.**

Задача: менеджер по продажам желает определить оптимальное количество дней тренинга для новых работников. Он имеет характеристики труда работников с одним, двумя и тремя днями обучения.

Решение: прежде чем применить ANOVA, построим график средних и ошибок. Выбираем ***Graphs->Error Bar...***затем ***Simple***. В качестве анализируемой переменной (***Variable***) выбираем ***Score on training exam*** а ***Sales training group*** в качестве категориальной переменной (***Category Axis***). Выбираем меню ***Standard error of mean*** из выпадающего меню ***Bars Represent***.



Задание: проанализируйте полученный график.



Ответ: с одной стороны видно, что показатели работы растут у тех, кто учился дольше. С другой стороны, вариация (ее показывает ошибка) внутри групп уменьшается по мере увеличения обучения. Поэтому предположение об эквивалентности вариаций внутри групп, требуемое ANOVA, может здесь не выполняться.

Для проверки эквивалентности вариаций внутри групп запустим ANOVA ***Analyze->Compare Means->One-Way ANOVA...*** В качестве зависимой переменной (***Dependent List***) выбираем ***Score on training exam***, а ***Sales training group*** в качестве факториальной переменной (***Factor***). В разделе Options... выбираем ***Descriptive*** и ***Homogeneity of variance test***.

Задание: проанализировать таблицы.

Descriptives

Score on training exam

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	20	63,580	13,50858	3,02061	57,2576	69,9020	32,68	86,66
2	20	73,568	10,60901	2,37225	68,6025	78,5328	47,56	89,65
3	20	79,279	4,40754	,98556	77,2165	81,3420	71,77	89,69
Total	60	72,142	12,00312	1,54960	69,0415	75,2430	32,68	89,69

Test of Homogeneity of Variances

Score on training exam

Levene Statistic	df1	df2	Sig.
4,637	2	57	,014

Ответ: подтверждается снижение вариаций по мере увеличения времени обучения и визуально ☐, и математически через тест Левене, показавшим низкую значимость (Sig.<0.05) гипотезы, что вариации групп одинаковы ☐.

Таким образом, вариации по исследуемым группам не идентичны. В этом случае ANOVA может быть применен для анализа, если размер выборок групп одинаков или почти одинаков. Однако, можно применить преобразование переменных или непараметрические тесты, свободные от данного предположения.

3.8.2. Выполнение анализа.

Используемый файл: **tutorial\sample_files\dvdplayer.sav**.

Задача: В ответ на пожелания пользователей фирма электроники разрабатывает новый DVD-плеер. Прототип был предложен фокусной группе, ее оценки были собраны отделом маркетинга. Необходимо оценить, различаются ли мнения пользователей в зависимости от возраста.

Решение: используем ANOVA ***Analyze->Compare Means->One-Way ANOVA...*** В качестве зависимой переменной (***Dependent List***) выбираем ***Total DVD assessment***, а ***Age Group*** в качестве факториальной переменной (***Factor***). В разделе Options... выбираем ***Means plot***. Нажимаем ***Continue*** и ***OK***.

Вопрос: проанализировать полученную таблицу.

ANOVA

Total DVD assessment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	733,274	5	146,655	4,601	,001
Within Groups	1976,417	62	31,878		
Total	2709,691	67			

Ответ: Величина F-теста и ее близкая к нулевой значимость однозначно показывают значимость различий между мнениями групп. График подтверждает, что пользователи среднего возраста оценили плеер выше других возрастных групп. Если необходим дополнительный более детальный анализ, можно воспользоваться парными сравнениями (см. раздел 3.3.3.1), ранговыми тестами или контраст-опцией ANOVA, обсуждаемой в следующем параграфе.

3.8.3. Контраст-анализ средних (Contrasts between means).

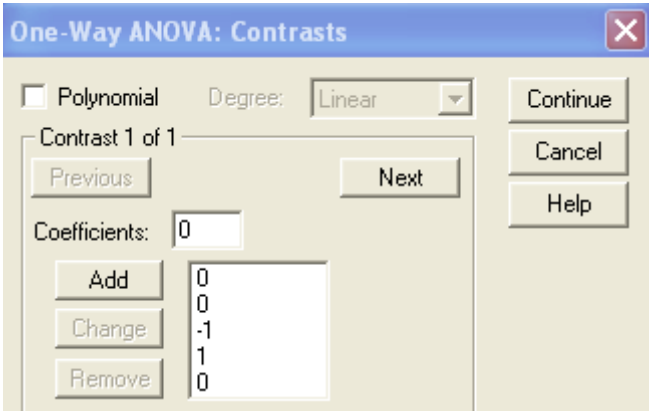
В общем случае F-статистика устанавливает наличие или отсутствие связи, а графики средних показывают, где может лежать разница. Контраст-анализ в однонаправленном ANOVA позволяет точно определить величину отклонений, и исследовать их параметры.

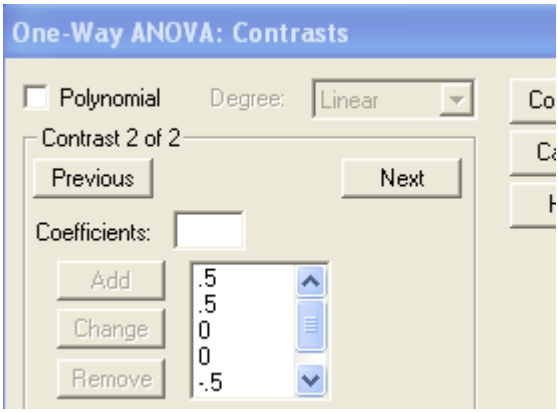
Задача: анализируя DVD-данные задачи предыдущего параграфа, аналитик задается вопросами:

- Две группы между 32 и 45 годами действительно отличаются?
- Можно ли считать группы возрастов менее 32 и более 45 лет статистически эквивалентными?

Решение: используем ANOVA ***Analyze->Compare Means->One-Way ANOVA.*** В качестве зависимой переменной (***Dependent List***) выбираем ***Total DVD assessment***, а ***Age Group*** в качестве факториальной переменной (***Factor***). Нажимаем кнопку 

В первом контрасте анализируем только группы 3 (вес минус 1) и 4 (вес 1), остальные (две первых, и две последних) не учитываем, присваивая им нулевые веса при помощи кнопки **Add**:





После нажатия **Next** вводим второй контраст, маскируя группы 3 и 4 (вес 0), первым двум группам присваиваем веса 0.5 и последним двум – минус 0.5. Нулевыми коэффициентами мы отсекаем группы, не участвующие в анализе, а задавая противоположные контрастные коэффициенты, мы проверяем, является ли значимой разница средних групп, участвующих в анализе. Сумма контрастных коэффициентов равна 0, а таблица контраст-коэффициентов позволяет проверить правильность их ввода:

Contrast Coefficients

Contrast	Age Group					
	18-24	25-31	32-38	39-45	46-52	53-59
1	0	0	-1	1	0	0
2	,5	,5	0	0	-,5	-,5

Задание: проанализировать полученную таблицу теста

Contrast Tests

			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Total DVD assessment	Assume equal variances	1	2,20	2,525	,871	62	,387
		2	2,61	1,633	1,596	62	,116
	Does not assume equal variances	1	2,20	2,678	,821	17,209	,423
		2	2,61	1,594	1,635	42,280	,110

Ответ: Результаты теста приводятся отдельно для предположений, что вариации средних по группам одинаковы и различны. В этом случае мы следуем первому предположению (**Дополнительное задание:** проверить правильность этого предположения! **Ответ:** провести анализ согласно Разделу 3.8.1). Результаты обоих контраст-тестов значимы (Sig>0.1), что говорит о статистической идентичности двух сформированных групп пользователей: 1) возраст 32–45, 2) возраст до 32 и после 45.

Контраст-метод очень полезен, когда в анализе вы сравниваете четко определенные группы и присваиваете необходимые вам веса. В ряде случаев у вас нет возможности или необходимости сравнения таких групп. ANOVA позволяет вам сравнивать любую группу относительно любой другой при помощи многочисленных парных сравнений, описываемых в следующем параграфе.

3.8.4. Множественные парные сравнения[3].

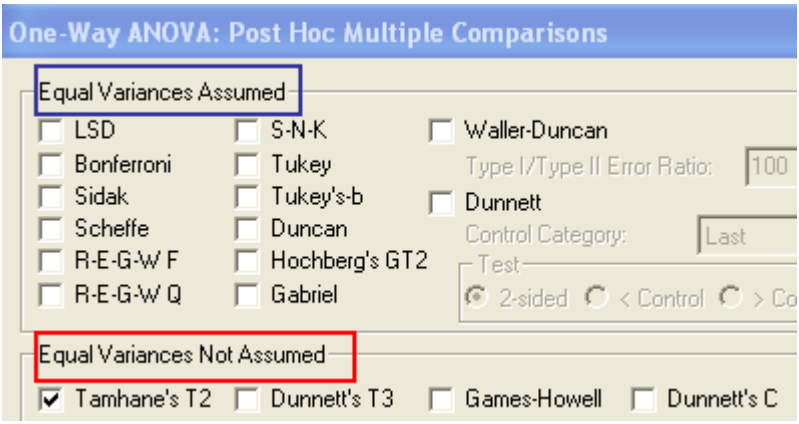
Используемый файл: **tutorial\sample_files\salesperformance.sav.**

Задача: менеджер по продажам продолжает анализ данных по обучению персонала (см. задачу в разделе 3.8.1). Несмотря на то, что отличия в группах были признаны значимыми, нет никакого априорного предположения о природе различий. Поэтому решено сравнить каждую группу с каждой другой.

Решение: используем ANOVA ***Analyze->Compare Means->One-Way ANOVA...*** В качестве зависимой переменной (***Dependent List***) выбираем ***Score on training exam***, а ***Sales training group*** в качестве факториальной переменной (***Factor***), и нажимаем

[3] Pairwise multiple comparisons.

кнопку Post Hoc..., статистики в котором выбираются, исходя из равенства ☐ или различия ☐ вариаций внутри исследуемых групп. Поскольку в нашем случае вариации групп различны (**Вопрос:** откуда это следует? **Ответ:** из теста Левене (***Levene's Test***), описанного в разделе 3.7.3 и уже проводившегося в разделе 3.8.1), выбираем статистику ***Tamhane's T2*** из второй группы ☐:



Задание: проанализируйте полученную таблицу:

Multiple Comparisons

Dependent Variable: Score on training exam

Tamhane

(I) Sales training group	(J) Sales training group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-9,98789*	3,84079	,040	-19,6053	-,3705
	3	-15,69947*	3,17733	,000	-23,8792	-7,5198
2	1	9,98789*	3,84079	,040	,3705	19,6053
	3	-5,71158	2,56883	,102	-12,2771	,8539
3	1	15,69947*	3,17733	,000	7,5198	23,8792
	2	5,71158	2,56883	,102	-,8539	12,2771

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

Решение: первая группа значимо отличается от двух других ☐ (Sig <0.1), в то же время, разброс между второй и третьей ☐ на грани значимости. Таким образом, выбор менеджера фокусируется вокруг вопроса, отменять или нет третий день обучения. В

пользу сохранения третьего дня говорит существенное уменьшение вариации в третьей группе по сравнению с первой (см. график в разделе 3.8.1).

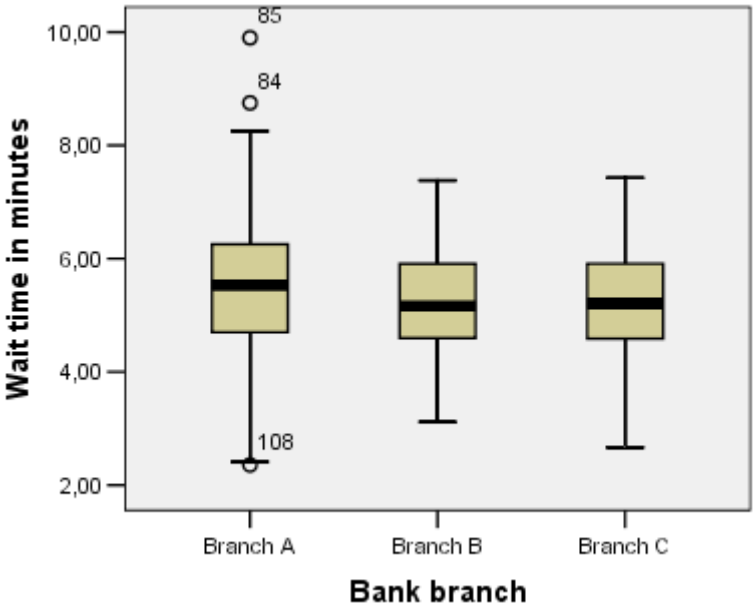
3.8.5. Робастные оценки вариаций

Статистики множественных парных сравнений надежны в той же степени, в которой стандартная F-статистика устойчива к нарушениям допущений анализа. В частности, F-статистика устойчива для групп с разными вариациями, если размеры групп равны или близки. Если же отличны и размеры выборок, и вариации, в этом случае F-статистика теряет силу и может выдавать некорректные результаты. В данном разделе показан ANOVA-анализ, работающий в данных случаях.

Используемый файл: **tutorial\sample_files\waittimes.sav.**

Задача: менеджмент банка получает жалобы клиентов о задержках обслуживания в одном из филиалов. Для оценки ситуации были проведены серии измерений в этом филиале и двух случайно выбранных для сравнения. Оценить значимость различий.

На блочном графике (см. раздел 3.1.1) видны выбросы и больший разброс показаний в банке А по отношению к двум другим. Отличаются и размеры выборки.



Решение: применим ANOVA-анализ для получения робастной оценки F-статистики. ***Analyze->Compare Means->One-Way ANOVA...*** В качестве зависимой переменной (***Dependent List***) выбираем ***Wait time in minutes***, а ***Bank branch*** в качестве факториальной переменной (***Factor***). В разделе Options... выбираем ***Homogeneity of variance test***, ***Brown-Forsythe***, и ***Welch***. Нажимаем ***Continue*** и ***OK***.

Задание: проанализировать полученные результаты.

Test of Homogeneity of Variances

Wait time in minutes

Levene Statistic	df1	df2	Sig.
4,062	2	312	,018

ANOVA

Wait time in minutes

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8,017	2	4,009	2,985	,052
Within Groups	418,983	312	1,343		
Total	427,000	314			

Ответ: тест Левене подтверждает предположение, что вариации групп различны. Значимость ANOVA-теста близка к 0,05, однако из-за разброса вариаций и размера групп этот результат не является надежным.

Robust Tests of Equality of Means

Wait time in minutes

	Statistic[a]	df1	df2	Sig.
Welch	2,622	2	206,327	,075
Brown-Forsythe	3,185	2	307,396	,043

a. Asymptotically F distributed.

В частности, тест Брауна–Форсайта оказывается существенно меньше 0,05, а тест Велша – больше 0.05. Тесты Велша и стандартный (F-статистика) более значимы (мощны) при равенстве вариаций и размеров групп. В данном случае более надежен тест Брауна-Форсайта, который показывает значимое отличие групп (Sig < 0.05). Необходимо заметить, что при увеличении размеров групп статистики Велша и Брауна–Форсайта асимптотически стремятся к стандартной F-статистике.

Таким образом, ANOVA тестирует гипотезу, что средние групп одинаковы. При помощи ANOVA-анализа вы можете:

- Проверить предположение о равенстве вариаций для разных групп;

- Оценить визуально поведение групповых средних показателей;
- Сделать контраст-тесты согласно вашим гипотезам;
- Сравнить попарно средние групп в предположении равенства вариаций или нет;
- Сделать робастные оценки вариаций групп.

Если условия и данные задачи не удовлетворяют предположениям ANOVA, вы можете использовать анализ среднего и непараметрические тесты.

3.8.6. Вопросы для самоконтроля.

1. Что значит аббревиатура ANOVA?
2. Для каких типов задач применяется ANOVA?
3. Для чего необходимо оценивать эквивалентность средних внутри групп и как это сделать в SPSS табличным и графическим образом?
4. При помощи какого теста можно оценить значимость расхождений групповых вариаций?
5. В чем суть ANOVA контраст-анализа, и как он осуществляется в SPSS?
6. Каковы границы применимости контраст-метода?
7. При каких предположениях справедлива стандартная F-статистика в ANOVA?
8. Может ли оставаться корректной F-статистика, если вариации групп различны? Если да, то при каких условиях?
9. В каких случаях применяется парные сравнения в ANOVA?
10. Какие статистики следует применять случае равных и неравных групповых вариаций в парных сравнениях ANOVA?
11. Как ведут себя статистики Велша и Брауна–Форсайта при увеличении размера анализируемых групп?

4. Регрессионный анализ.

Регрессионный анализ, пожалуй, самый широко используемый статистический метод анализа данных, предназначенный для выявления зависимостей между различными переменными согласно определенной гипотезе.

4.1. Линейная регрессия.

Линейная регрессия, является простейшей и наиболее часто используемой регрессионной моделью, в которой предполагается, что поведение зависимой (определяемой, ***dependent***) переменной y_i моделируется линейной комбинацией одной или нескольких описывающих (определяющих) переменных x_i (предикторов, ***predictors***):

$$y_i=b_0+b_1x_{i1}+\ldots+b_px_{ip}+e_i \qquad 4.1$$

где:

y_i –значение зависимой переменной в i-той строке данных, p – число определяющих переменных-предикторов, $\underline{b_j}$ –коэффициент j-того предиктора (поиск которых и является основной целью анализа), x_{ij} – значение j-того предиктора в i-той строке данных, e_i – остаточная ошибка (***residual***) наблюдения для i-строки данных.

Отметим также, что b_0 – это сечение[4] (***intercept***), предсказываемое регрессионной моделью значение y при нулевых значениях предикторов.

Кроме этого, для корректного применения линейной регрессии необходимо выполнение нескольких условий в отношении остаточной ошибки e_i:

- Она нормально распределена вокруг нуля.
- Вариация остаточной ошибки постоянно по всем данным и не зависит от предикторов. Остаточная ошибка с непостоянной вариацией является гетероскедастичной (***heteroscedastic***).

[4] Иногда его называют средним значением, что не совсем верно.

- Величина остаточной ошибки для конкретной строки данных независима от величин предикторов модели а также от значений остаточных ошибок в других строках данных

Проверка выполнения этих условий является ***необходимым шагом*** *регрессионного анализа, прямо определяющим его применимость!* Потому что довольно часто пользователи ограничиваются лишь механическим расчетом корреляции. А при наличии серьезных проблем, которые как раз и можно обнаружить при анализе остаточной ошибки, полученные результаты будут неверными.

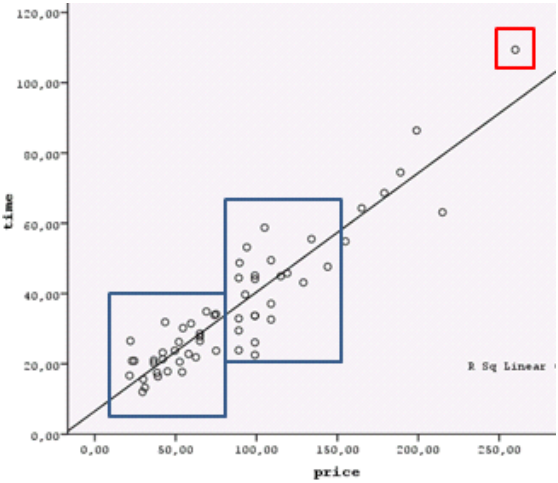
4.2. *Однопараметрическая линейная регрессия.*

Задача РА1: Компания «Nambe Mills» имеет линию по выпуску металлической столовой посуды, которая в процессе изготовления проходит стадию полировки. Для временного планирования процесса для 59 образцов продукции разных типов и относительных (к радиусу детали) размеров было измерено и записано время полировки. При помощи линейной регрессии определить, можно ли предсказать время полировки в зависимости от размера продукта? Перед использованием регрессии необходимо исследовать корреляционную матрицу (***scatterplot***) время полировки/размер продукта на предмет возможности применения линейной модели в отношении этих переменных.

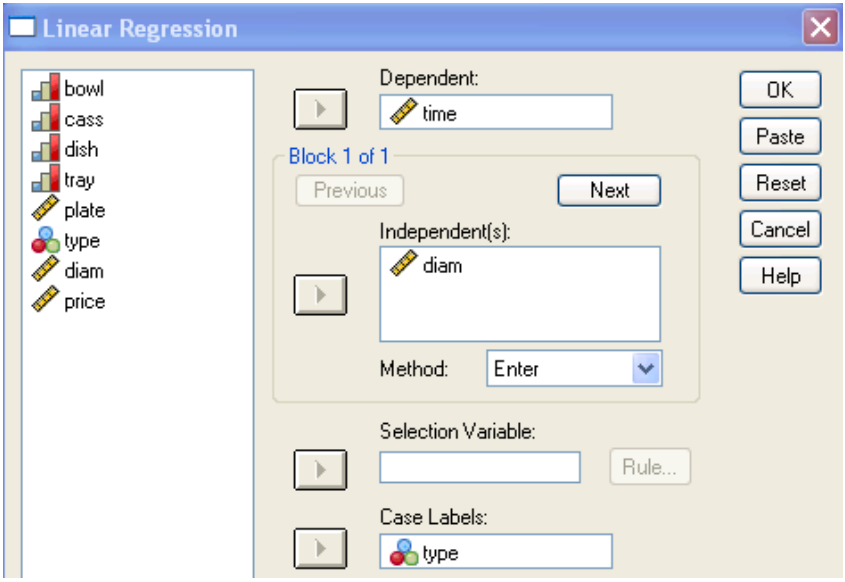
Используемый файл данных: **tutorial\sample_files\polishing.sav**.

Корреляционную матрицу строим через мастер графиков ***Graphs->Chart Builder->Scatter/Dot->Simple Scatter***, фитируем ее прямой линией (см. 2.1, как это сделать). На результирующем графике отметим две важные особенности (в дальнейшем мы вернемся к этому вопросу):

- разброс точек увеличивается с диаметром изделий.
- одна удаленная точка может оказывать неоправданно большое влияние на результаты.

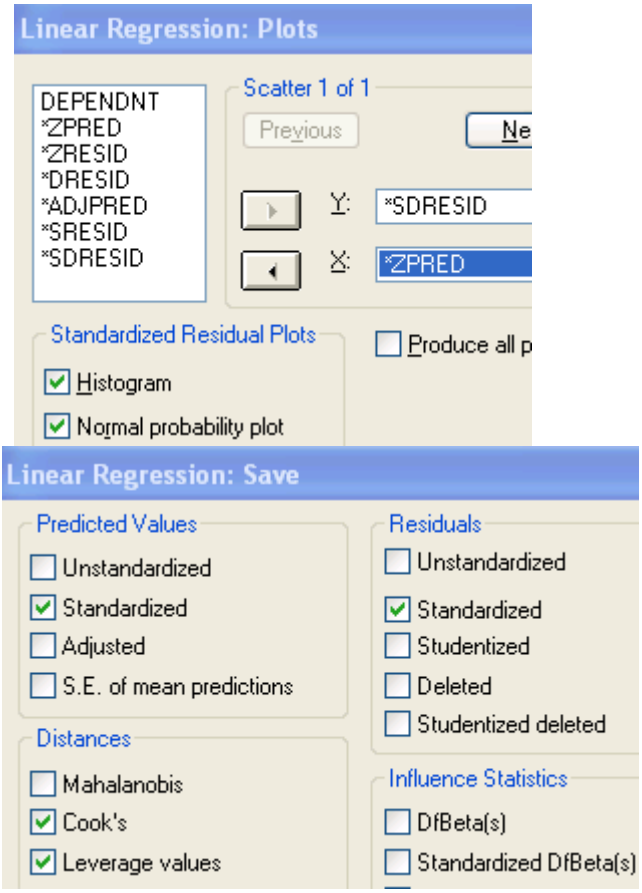


Регрессию осуществляем при помощи ***Analyze->Regression->Linear***, выбирая время (***time***) в качестве зависимой переменной, а диаметр изделия (***diam***) – предикатором, переменную тип (***type***) используем для надписей на точках данных (***Case Labels***):



В качестве изображаемых графиков (кнопка ***Plots...***) выбираем ****SDRESID/*ZPRED*** для переменных Y/X а также гистограммы (***Histogram***) и вероятностного графика (***Normal probability plot***).

В разделе сохраняемых величин (кнопка ***Save...***) сохраняем стандартные показатели предсказанных значений и остаточной ошибки, а также дистанционные показатели (***Predicted Values->Standardized***, ***Residuals->Standardized***, ***Distances->Cook's, Leverage values***) – все это необходимо для последующей диагностики качества регрессии.



Сохраняемые величины рассчитываются для каждой строки и заносятся в лист данных в качестве обычных переменных и могут быть использованы в дальнейшем таким же образом, как исходные данные, для построения графиков и расчетов. Каждая из расчетных величин имеет формат ***NAME_X***, где ***NAME*** – уникальное имя, предустановленное в программе (например, ***PRED*** – не стандартизованное прогнозируемое значение, ***RES*** – не стандартизованная ошибка и т.д.), ***X*** – номер, который автоматически увеличивается на единицу при запуске очередного расчета регрессии. Для очистки листа данных можно выделить и удалить соответствующие расчетные переменные

После запуска регрессии в окне выходных данных появляется ряд таблиц и графиков. Необходимые нам коэффициенты b_i модели (4.1) (таблица ***Coefficients***) дают результат:

Время обработки (***time***) =

3.457*Диаметр изделия (***diam***) – 1.955 4.2

Однако отметим при этом, что величина константы оказалась не значимой (вероятность, что это случайно полученное значение, равна 71.9% (***Sig.***) , это же видно по величине стандартной ошибки, которая ~2.5 раза больше самого значения)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1,955	5,402		-,362	,719
	diam	3,457	,467	,700	7,407	,000

a. Dependent Variable: time

Вариационный анализ (***ANalysis Of Variance, ANOVA***) позволяет оценить приемлемость модели со статистической точки зрения. В данном случае видно, что учтенная моделью вариация значима (вероятность случайности расчета (***F-statistics***, ***Sig.***) нулевая (<0.05[5])) составляет (***Sum of Squares***) 10287 против 10687 неучтенной . Таким образом, линейная модель описывает чуть менее половины вариации времени обработки деталей. Вариаци-

[5] 5% часто используется в качестве границы значимости/не значимости нулевой гипотезы.

онный анализ является важным статистическим инструментом, показывая, может ли модель адекватно описывать вариации исследуемой переменной. Однако, он не может напрямую показать силу этой зависимости (***strength of relationship***).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10287,173	1	10287,173	54,865	,000[a]
	Residual	10687,511	57	187,500		
	Total	20974,684	58			

a. Predictors: (Constant), diam

В суммарной таблице (***Model Summary***) множественный коэффициент корреляции (МКК, ***multiple correlation coefficient***) показывает корреляцию между наблюдаемым и предсказываемым моделью поведением описываемой переменной. Чем ближе он к 1, тем лучше модель[6]. Коэффициент детерминации является квадратом МКК (***R square***) и показывает долю корреляции, описываемой моделью (что соответствует предыдущему результату ANOVA). Наконец, еще одна оценка силы модели – это сравнение ошибок оценки модели (***Std. Error of the Estimate***) и стандартного отклонения от среднего (***Std. Deviation***) для исследуемой переменной (построение описательных статистик (***Descriptive Statistics***) см. в разделе 3.1). Фактически, описательная статистика есть не что иное, как тривиальная линейная регрессия без предикатов с одной лишь константой, которая в этом случае превращается в среднее значение описываемой величины. Из таблиц видно, что линейная модель с учетом диаметра заготовки более точна (дает меньшую ошибку), чем модель простого усреднения

[6] Стоит отметить, что в реальных условиях корреляция всегда меньше 1 (к примеру, точное равенство получается для математически точной функциональной зависимости). Поэтому при получении значений, близких к единице, нужно тщательно проверять модель и данные.

Model Summary[b]

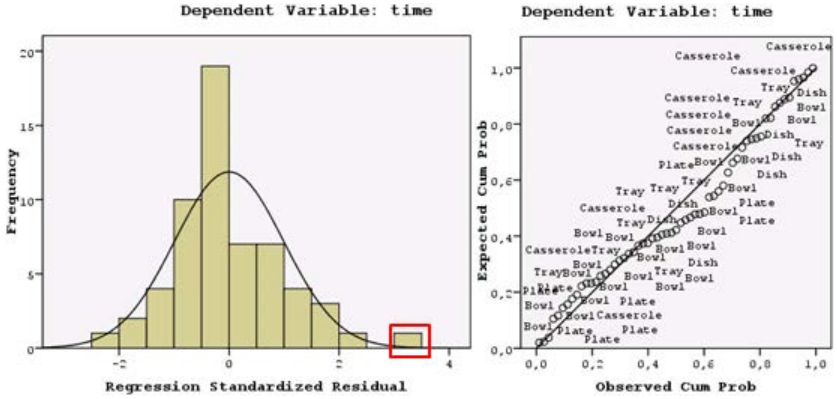
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,482	13,69307

a. Predictors: (Constant), diam

Descriptive Statistics

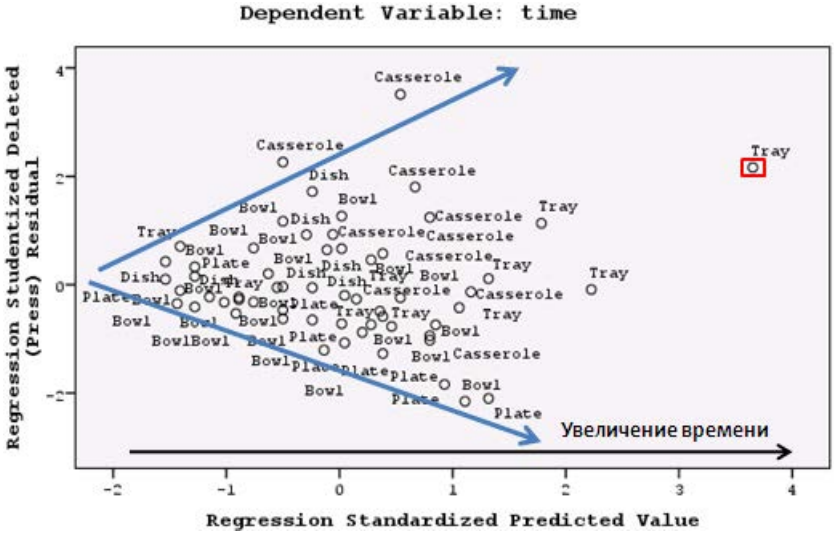
	N	Mean	Std. Deviation
time	59	35,8171	19,01664
Valid N (listwise)	59		

Следующий **необходимый** этап – анализ выполнения рассматривавшихся выше требований к остаточной ошибке модели, складывающейся из остатков (***residuals***) – разностей между наблюдаемым и модельным значением для каждой строки данных. По частотной гистограмме (слева) проверяется нормальное распределение остатков вокруг нуля (выбросы □ изучаются отдельно). Другой способ – график наблюдаемой и модельной накопленной вероятностей (***P-P plot***), который должен следовать линии с наклоном 45% между точками 0,0 и 1,1. Из рисунков видно удовлетворительное соответствие.

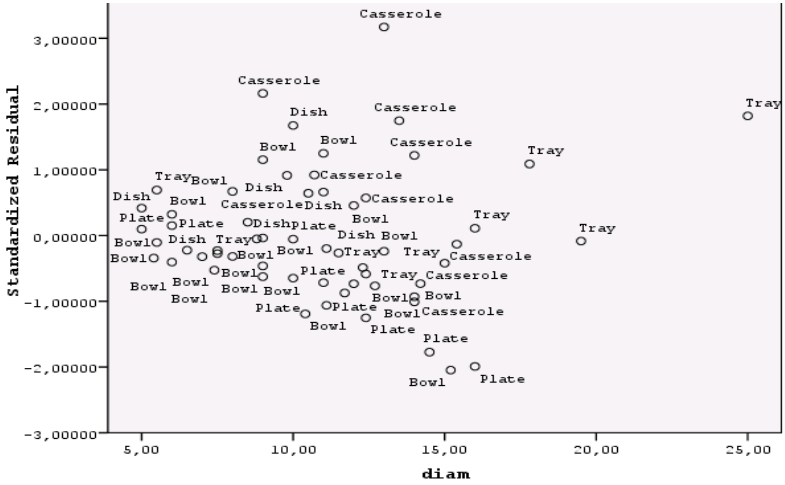


Корреляция остатков как функции описываемой величины (в идеальном случае должно наблюдаться равномерное рассеяние вокруг 0) показывает, что разброс ошибок растет с увеличением

времени обработки (что было замечено еще при анализе корреляционной матрицы). Вместе с тем, рассеяние ошибок хорошее.

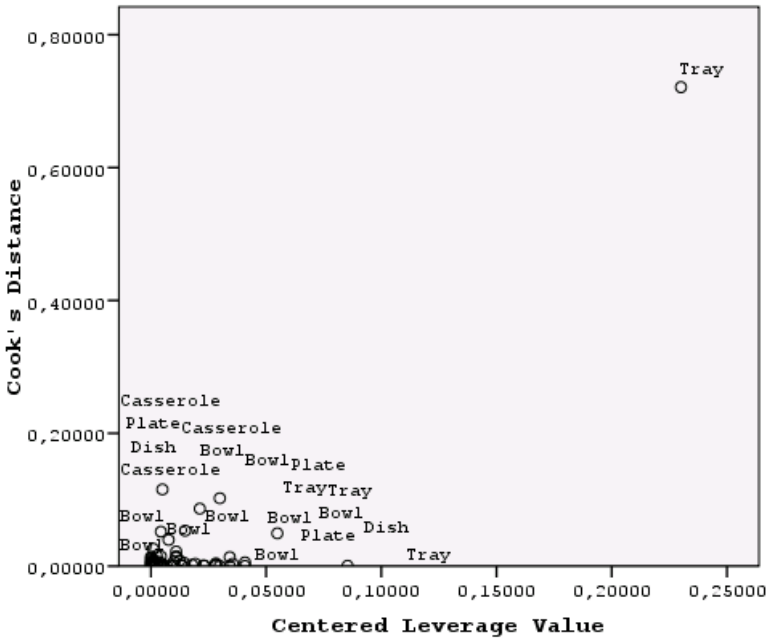


Теперь построим корреляцию остатков в зависимости от предиктора (диаметра), снова воспользовавшись описанным выше мастером графиков (***Graphs->Chart Builder->Scatter/Dot->Simple Scatter***) между переменными ***diam*** и ***Standard Residual*** (которая по нашей просьбе была сохранена вместе с данных при расчете регрессии наряду с другими запрошенными для сохранения переменными). Получили аналогичный эффект, который и называется **гетероскедастичностью**.



Идентичности графиков следовало ожидать для однопараметрической регрессии, где разброс описываемой переменной полностью определяется разбросом одного предиктора. В многопараметрической регрессии, где поведение зависимой переменной определяется суммарным воздействием нескольких предикторами, корреляционные матрицы остатков по определяемой переменной и индивидуальным предикторам могут различаться. В случае обнаружения эффектов, к примеру, гедероскедастичности остатков, задачей пользователя становится выявление конкретных предикторов, вызывающих те или иные эффекты.

Возвращаясь к примеру, отметим, что для борьбы с проблемой гедероскедастичности можно использовать весовые коэффициенты $w=f(\boldsymbol{diam})$ при расчете регрессии. Используемая весовая функция $f(\boldsymbol{diam})$ не всегда очевидна, искусство ее выбора определяется задачей, опытом обработчика, а иногда методическим подбором. В данном случае можно использовать простейший вариант весовой функции $w=1/\boldsymbol{diam}^2$ (эта задача будет предложена для самостоятельного решения).



Наконец, финальной стадией проверки является исследование точек с аномальным отклонением от общего поведения (выбросов). Как мы уже видели на корреляционной матрице, гистограмме остатков, у нас есть в наличии одна такая точка. Для наглядной оценки ситуации необходимо построить (например, через мастер графиков) матрицу ***Leverage value*** (переменная ***LEV***) против ***Cook's Distance*** (***COO***).

Первый показатель ***Leverage value*** является фактором «рычага» (***S***)[7] в том смысле, что точки, имеющие большое значение этого показателя, заметно воздействуют на наклон (коэффициент b_i) регрессионной линии (действуя на нее своим «рычагом», откуда и название). Второй показатель ***Cook's Distance*** является фактором «силы» (***F***): чем больше значение этого показателя, тем больший вес имеет точка в расчетах. В целом, данный график можно назвать матрицей моментов следуя аналогии с физической фор-

[7] Leverage – действие рычага, система рычагов (англ.).

мулой ***M=F*S***[8]. Таким образом, в наших данных присутствует точка с очень большим моментом, влияющим на расчеты регрессии. Уменьшение влияния этой точки осуществляется либо снижением ее индивидуального веса в расчетах, либо полным удалением из расчетов. В реальной ситуации, где это возможно, практикуется повторение измерений, в которых наблюдались выбросы. Это позволяет выяснить, не являются ли выбросы результатом грубой ошибки.

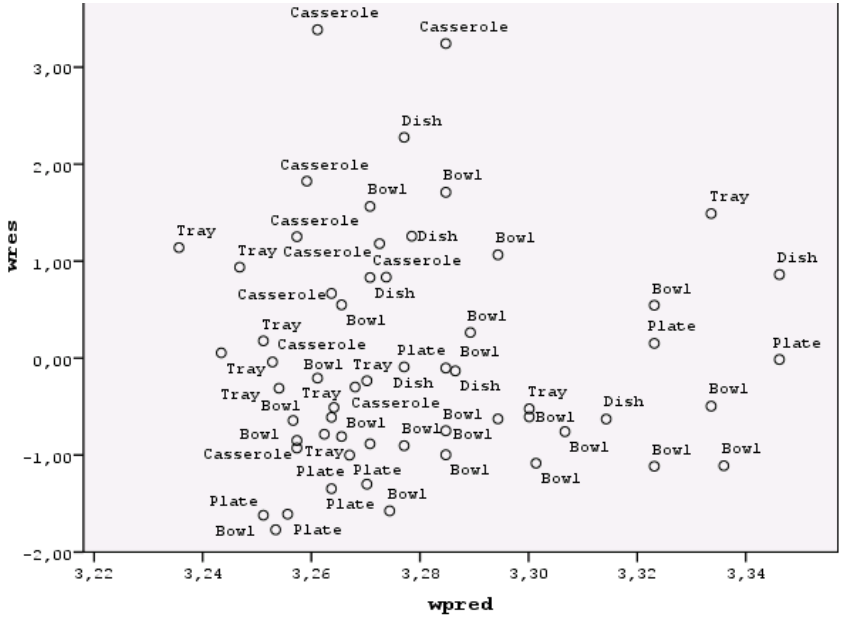
Подводя итоги, можно сказать, что, используя линейную регрессию, мы определили время полировки посуды как функцию ее размера, а также убедились, что применение метода обосновано с определенными оговорками, а именно, наличием проблем выбросов и гетероскедастичности. Решение этих проблем (одна из которых предлагается далее в качестве самостоятельного задания) позволит уточнить полученные результаты. В итоге данная информация может быть использована фирмой «Nambe Mills» для разработки или обновления графика работ в производстве.

4.3. Задания для самостоятельной работы.

1. Для рассмотренной задачи проведите линейный регрессионный анализ с учетом веса по диаметру образцов (чем больше диаметр, тем меньше вес). Для этого необходимо:

Добавить расчетную переменную ***wdiam***=1/***diam***^{2} (***Transform->Compute Variable…***, см. 2.2.2), и использовать ее в качестве веса (***WLS Weight***). При этом в расчетах сохранить обычные (***Unstandardized***) предсказанные значения и остатки, появляющиеся после расчетов в листе данных в виде переменных ***PRE_N*** и ***RES_N***. Для построения корреляции остатков взвешенной регрессии используем мастер графиков с еще двумя расчетными переменными ***WPRED=SQRT(WDIAM)*PRE_N*** и ***WRES=SQRT(WDIAM)*RES_N***, чтобы получить:

[8] Отметим, что в регрессионных формулах расчеты моментов присутствуют в явном виде.



Исследуйте качество новой модели аналогично рассмотренному примеру, проведите сравнительный анализ моделей и сделайте вывод, какая из них лучше и почему?

2. В общем случае весовые коэффициенты можно описывать силовой функцией $1/VAR^N$, где N может принимать любые значения, в том числе дробные и отрицательные. Найдите оптимальное значение N для рассматриваемой нами задачи, воспользовавшись инструментом **Analyze->Regression->Weight Estimation…**. После нахождения оптимальной величины N сделайте перерасчет по п.1 и получите наиболее оптимальные коэффициенты для формулы (4.2).

4.4. Многопараметрическая линейная регрессия.

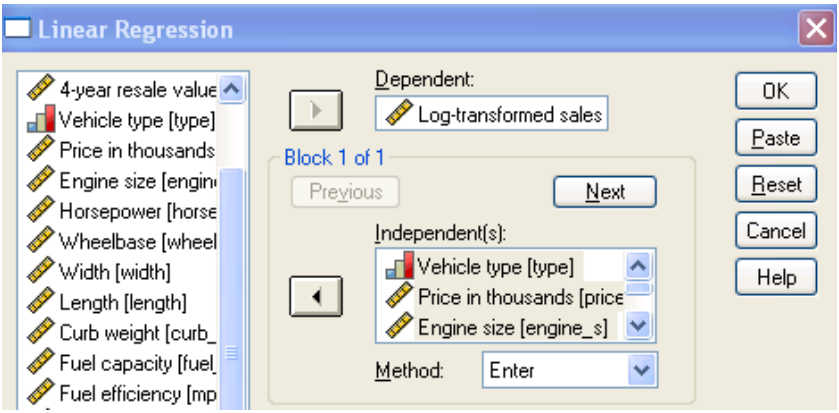
Однопараметрическая модель линейной регрессии используется для решения простых задач и незаменима для изучения основ регрессионного анализа. Однако задачи из реальной жизни, как правило, связаны с обработкой многомерных массивов данных с большим количеством переменных, где широко применяется многопараметрическая линейная регрессия, модели которой имеют

более одного предиктора. В этом случае возникает ряд специфических вопросов и задач, таких как взаимовлияние данных, вклад индивидуальных переменных и др., которые рассматриваются в настоящем параграфе на примере предлагаемой задачи.

Задача РА2: компания, торгующая автомобилями, ведет учет продаж своей продукции. В поисках оптимальной стратегии необходимо определить, какие характеристики машин в большей степени влияют на объем продаж, выявить недооцененные и переоцененные модели. Используя линейную регрессию, нужно установить наиболее плохо продающиеся модели.

Используемый файл данных: **tutorial\sample_files\car_sales.sav**.

В мастере линейной регрессии (***Analyze->Regression->Linear***) выбираем логарифм объема продаж (***Log-transformed sales***) в качестве зависимой переменной, а все характеристики машин от типа (***Vehicle type***) до потребления горючего (***Fuel efficiency***) – в качестве предикторов. Логарифмирование переменной (случайной величины) приближает ее распределение к нормальному, что улучшает поведение большинства статистических методов анализа, включая регрессионный.



Нажав кнопку ***Statistics***, выбираем взаимные корреляции (***Part and partial correlations***) и диагностику коллинеарности (***Collinearity diagnostics***) и запускаем расчет регрессии.

Согласно результатам вариационного анализа (***ANOVA***) получаем значимые результаты, а из суммарной таблицы (***Model***

Summary) определяем, что модель описывает примерно половину вариации продаж.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130,300	10	13,030	13,305	,000[a]
	Residual	138,082	141	,979		
	Total	268,383	151			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,697[a]	,486	,449	,98960

И хотя в целом результат удовлетворительный, из таблицы коэффициентов видно, что большинство переменных не значимы (т.е. вносят малый вклад в модель) (***Sig.>0.05***).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3,017	2,741		-1,101	,273
	Vehicle type	,883	,331	,293	2,670	,008
	Price in thousands	-,046	,013	-,502	-3,596	,000
	Engine size	,356	,190	,281	1,871	,063
	Horsepower	-,002	,004	-,092	-,509	,611
	Wheelbase	,042	,023	,241	1,785	,076
	Width	-,028	,042	-,073	-,676	,500
	Length	,015	,014	,148	1,032	,304
	Curb weight	,156	,350	,075	,447	,655
	Fuel capacity	-,057	,047	-,167	-1,203	,231
	Fuel efficiency	,081	,040	,262	2,023	,045

Вторая часть таблицы позволяет получить ряд ответов на вопросы о взаимовлиянии предикторов. В частности, для большинства предикторов заметно существенное падение парциальных (***Partial***) и частичных (***Part***) корреляций по отношению к нулевым (***Zero-order***). Это означает, что большая доля вариации, объясняемая предиктором (ценой в данном случае), также объясняется другими переменными.

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Vehicle type	.274	.219	.161	.304	3.293
	Price in thousands	-.552	-.290	-.217	.187	5.337
	Engine size	-.135	.156	.113	.162	6.159
	Horsepower	-.389	-.043	-.031	.112	8.896
	Wheelbase	.292	.149	.108	.200	4.997
	Width	.037	-.057	-.041	.313	3.193
	Length	.215	.087	.062	.178	5.605
	Curb weight	-.041	.038	.027	.131	7.644
	Fuel capacity	-.016	-.101	-.073	.189	5.303
	Fuel efficiency	.121	.168	.122	.217	4.604

Толерантность (***tolerance***) показывает долю вариации предиктора, которая не может быть объяснена другими переменными. Малые величины означают, что 70-90% вариации предикторов не могут быть объяснены другими переменными. Близкая к 0 толерантность предиктора свидетельствует о сильной мультиколлинеарности, и стандартная ошибка может быть завышена (***inflated***). Обычно проблематичным считается величина инфляционного коэффициента вариации (***variance inflation factor***, ***VIF***) большая 2, что в нашем случае касается всех переменных.

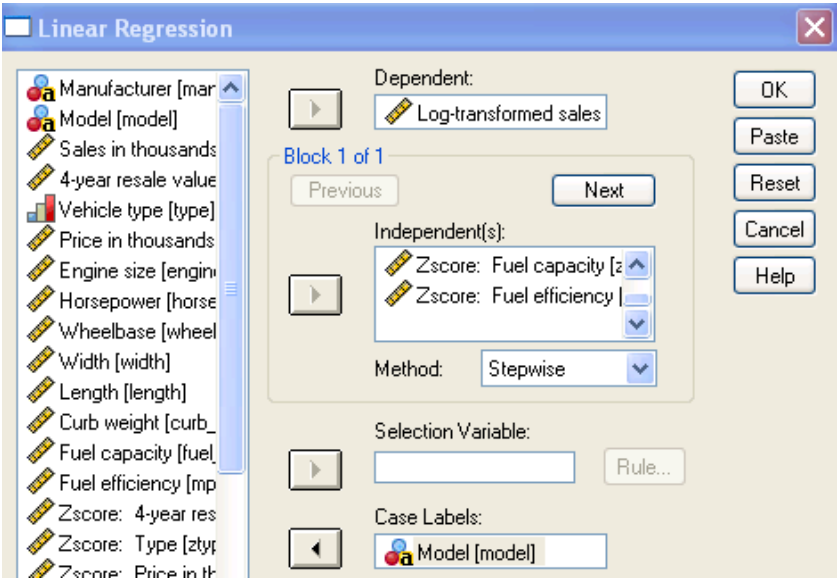
Model	Dimension	Eigenvalue	Condition Index
1	1	9,920	1,000
	2	,733	3,678
	3	,259	6,193
	4	,050	14,051
	5	,019	22,589
	6	,008	35,942
	7	,005	44,275
	8	,003	58,480
	9	,002	76,175
	10	,001	130,747
	11	,000	148,267

a. Dependent Variable: Log-transformed sales

Диагностика коллинеарности (таблица ***Collinearity Diagnostics***) подтверждает наличие серьезных проблем с мультиколлинеарностью. Во-первых, близость к нулю собственных значений (***Eigenvalue***) для большинства предикторов ☐ показывает сильную внутреннюю корреляцию между ними. Это означает, даже небольшие изменения в данных могут привести к существенным колебаниям коэффициента корреляции для данного предиктора. Во-вторых, наличие затруднений демонстрирует условный индекс (***Condition index***) (= квадратный корень из отношения наибольшего собственного значения к собственному значению текущего предиктора). Проблемы с коллинеарностью возникают при величинах, большие 15, а величины, больших 30, свидетельствуют о серьезных проблемах. В нашем случае ☐ имеются 6 переменных с условным индексом большим 30, что говорит о серьезных проблемах коллинеарности предикторов модели.

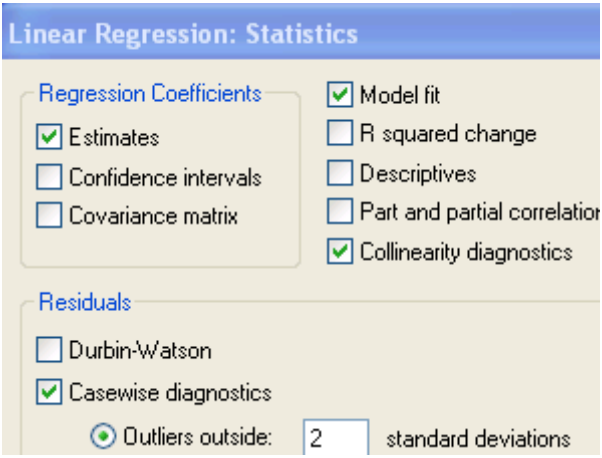
Попробуем решить проблему коллинеарности, повторив регрессию для стандартизованных величин переменных (***z-score***, рассмотрены в параграфе 3.1), используя последовательную методику (***stepwise method***) включения наиболее значимых переменных в регрессионную модель.

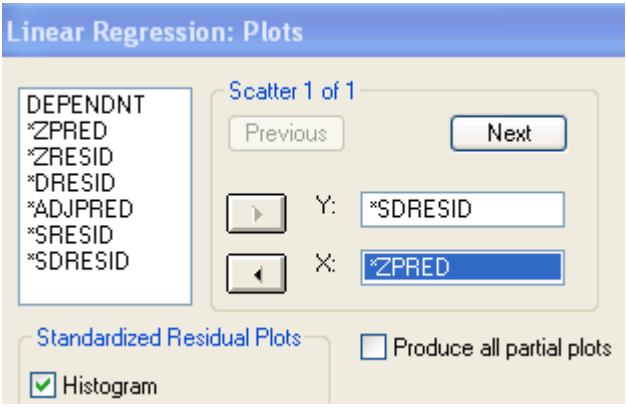
Повторяем процедуру выбора регрессионной модели, описанную в начале параграфа (***Analyze->Regression->Linear***), с той лишь разницей, что в качестве предиктора выбираем стандартизованные переменные, от типа (***Zscore: Vehicle type***) до потребления горючего (***Zscore: Fuel efficiency***) – в качестве предикторов. В качестве метода выбираем пошаговый (***stepwise***), а для надписей (***Case Labels***) используем переменную ***Model***. Выбор пошагового метода означает, что при расчетах предикторы будут выбираться из расчета убывания по весомости до тех пор, пока не будут выбраны все значимые.



В разделе статистики (кнопка ***Statistics…***) контролируем коллинеарность (***Collinearity diagnostics***) и выбросы данных более двух стандартных ошибок (***Casewise diagnostics: Outliers outside 2 standard deviations***). Последнее очень полезно для исследования аномалий. В качестве сохраняемых графиков рисуем стандартизованные остатки (****SDRESID***) как функцию предсказываемой описываемой величины (****ZPRED***).

Наконец, для сохранения (кнопка ***Save…***) выбираем стандартизованные предсказываемые величины (***Predicted values: standardized***) и остатки (***Residuals: standardized***), вес и расстояние (***Distances->Cook's, Leverage values***) аналогично тому, как это было сделано в предыдущем параграфе для однопараметрической прогрессии.





Согласно результатам теста собственные значения и условный индекс теперь указывают на неколлинеарность предикторов. В ходе последовательной методики было пройдено два шага, и, соответственно, построено две модели:

1. ***Const*** + ***Zscore: Price*** (однопараметрическая);
2. ***Const*** + **Zscore: Price** + ***Zscore: Wheelbase*** (колесная база (англ.) – расстояние между передней задней осями автомобиля, фактически, базовый параметр величины автомобиля).

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Zscore: 4-year resale value	Zscore: Wheelbase
1	1	1,004	1,000	,50	,50	
	2	,996	1,004	,50	,50	
2	1	1,056	1,000	,04	,42	,48
	2	1,002	1,026	,87	,12	,00
	3	,942	1,059	,08	,46	,52

a. Dependent Variable: Log-transformed sales

Сравнение результатов (Таб. справа) с моделью первого шага (Таб. слева) показывает, что учет всего лишь двух предикторов описывает 43% вариации, в то время как полная модель дает 49%. А с учетом коэффициента дискриминации, скорректированного с учетом сложности модели (такое сравнение более справедливо) разница еще меньше.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,697[a]	,486	,449	,98960

a. Predictors: (Constant), Fuel efficiency, Length, Price in thousands, Vehicle type, Width, Engine size, Fuel capacity, Wheelbase, Curb weight, Horsepower

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,552[a]	,304	,300	1,11553
2	,655[b]	,430	,422	1,01357

a. Predictors: (Constant), Zscore: Price in thousands

b. Predictors: (Constant), Zscore: Price in thousands, Zscore: Wheelbase

c. Dependent Variable: Log-transformed sales

Из коэффициентов следует отрицательная корреляция с ценой, и положительная – с размером автомашины (колесной базой), т.е. наибольшим спросом пользуются недорогие большие автомобили.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	3,286	,090		36,316	,000		
	Zscore: Price in thousands	-,732	,090	-,552	-8,104	,000	1,000	1,000
2	(Constant)	3,290	,082		40,020	,000		
	Zscore: Price in thousands	-,783	,083	-,590	-9,487	,000	,988	1,012
	Zscore: Wheelbase	,470	,082	,356	5,718	,000	,988	1,012

a. Dependent Variable: Log-transformed sales

Таблица исключенных предикторов позволяет проанализировать процедуру. Цена (***Price***) была выбрана первой по причине наибольшей корреляции с объясняемой переменной. Оставшиеся анализируются с целью определить лучший предикторов, если таковой вообще найдется, для дополнения модели. Из таблицы видно, что все коэффициенты (***Beta in*** показывает коэффициент предиктора, если его включить в модель) ненулевые и значимы (***Sig.***<=0.05), то есть любой из них может быть включен в модель. Для выбора оптимального обратим внимание на парциальную корреляцию (***Partial correlation***), которая показывает линейную корреляцию между предиктором и описываемой переменной после вычитания эффекта текущей модели. По этому показателю и была выбрана колесная база ***Wheelbase***. Наконец, после добавления в модель данного предиктора, больше не осталось ни одного параметра , который вносил бы значимый вклад в уже построенную модель. Такое резкое изменение за один шаг (все значимые – все незначимые) означает, что все остававшиеся предикторы описывали вариацию объясняемой переменной практически одинаковым образом, и выбор переменной ***Wheelbase*** был обоснован лишь небольшим ее преимуществом перед остальными.

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Coll Tolerance
1	Zscore: Type	,251[a]	3,854	,000	,301	,998
	Zscore: Engine size	,342[a]	4,128	,000	,320	,611
	Zscore: Horsepower	,257[a]	2,062	,041	,167	,293
	Zscore: Wheelbase	,356[a]	5,718	,000	,424	,988
	Zscore: Width	,244[a]	3,517	,001	,277	,892
	Zscore: Length	,308[a]	4,790	,000	,365	,976
	Zscore: Curb weight	,346[a]	4,600	,000	,353	,722
	Zscore: Fuel capacity	,266[a]	3,687	,000	,289	,820
	Zscore: Fuel efficiency	-,198[a]	-2,584	,011	-,207	,758
2	Zscore: Type	,129[b]	1,928	,056	,157	,835
	Zscore: Engine size	,145[b]	1,576	,117	,128	,445
	Zscore: Horsepower	,028[b]	,229	,819	,019	,256
	Zscore: Width	-,025[b]	-,275	,784	-,023	,470
	Zscore: Length	,027[b]	,237	,813	,020	,290
	Zscore: Curb weight	,105[b]	1,028	,306	,084	,365
	Zscore: Fuel capacity	,002[b]	,024	,981	,002	,443
	Zscore: Fuel efficiency	,014[b]	,164	,870	,014	,559

a. Predictors in the Model: (Constant), Zscore: Price in thousands

b. Predictors in the Model: (Constant), Zscore: Price in thousands, Zscore: Wheelbase

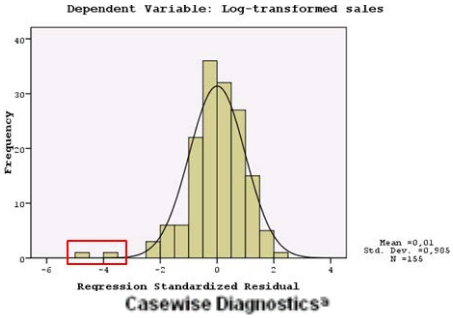
c. Dependent Variable: Log-transformed sales

Далее заметим, что параметр Тип машины (***Vehicle type***) находится на грани значимости (0.056 близко к 5%-ной отсечке) и имеет высокую толерантность (***Tolerance***), поэтому можно попытаться его использовать. Следующая по качеству переменная Размер двигателя (***Engine size***) проигрывает по значимости, но самое главное, имеет в два раза меньший коэффициент толерантности, что означает большую коррелированность с уже входящими в модель переменными.

Гистограмма остатков (рис. слева) хорошо вписывается в нормальное распределение, однако есть выбросы, требующие внимания.

Для этого обратимся к таблице диагностики данных (***Case Diagnostics***, рис. справа), в которой по нашей просьбе выведена информация точках с отклонением более, чем в 2 стандартных ошибки. Из нее можно сделать вывод, что наиболее слабые показатели на рынке у *Cutlass* и *3000GT* (отрицательное стандартное

отклонение), в меньше степени, у *Breeze*, *Prowler* и *SW*, в то же время, Explorer демонстрирует показатели выше ожиданий.

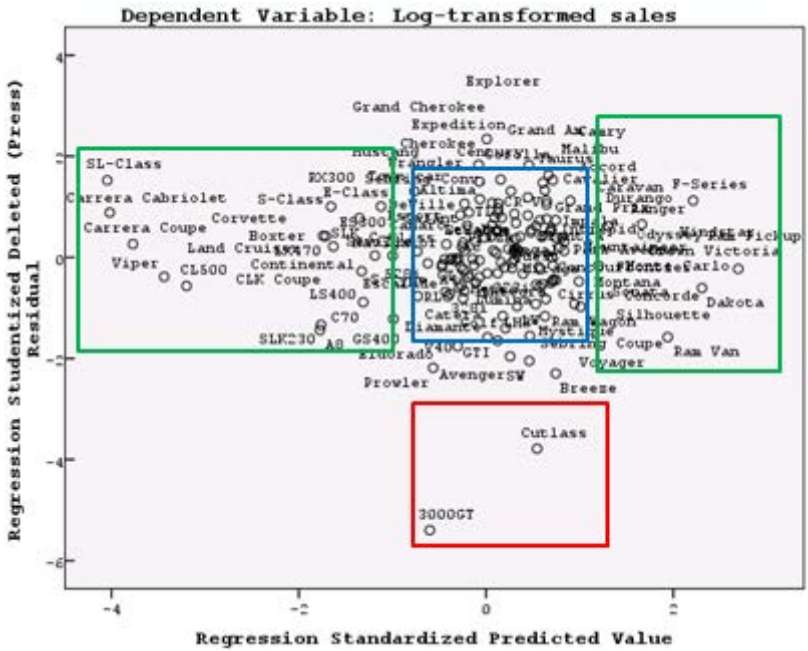


Casewise Diagnostics[a]

Case Number	Model	Std. Residual	Log-transformed sales	Predicted Value	Residual
53	Explorer	2,297	5,62	3,2953	2,32778
84	3000GT	-4,905	-2,21	2,7638	-4,97111
109	Cutlass	-3,610	,11	3,7651	-3,65892
116	Breeze	-2,252	1,66	3,9393	-2,28296
118	Prowler	-2,139	,63	2,7955	-2,16849
132	SW	-2,012	1,65	3,6927	-2,03967

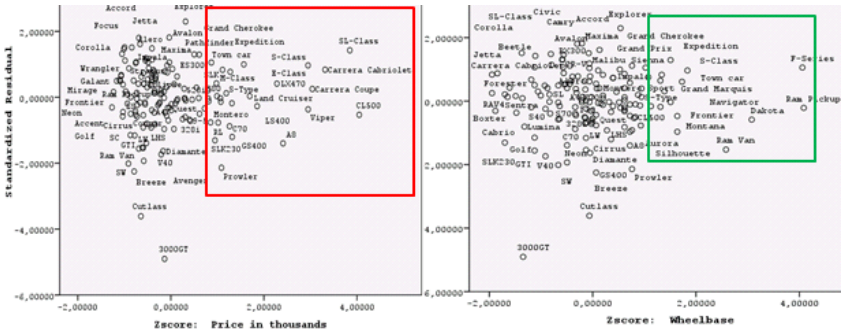
a. Dependent Variable: Log-transformed sales

Выводы в отношении ***Cutlass*** и *3000GT* еще более наглядны на матрице остатков, а вот остальные модели не сильно выделяются на фоне основного кластера, и их отклонения больше похоже на игру случая (статистики). Вместе с тем, озабоченность вызывает наличие группировок справа, и, особенно, слева от основного кластера. Эти точки могут давать чрезмерный вес в регрессионные расчеты.

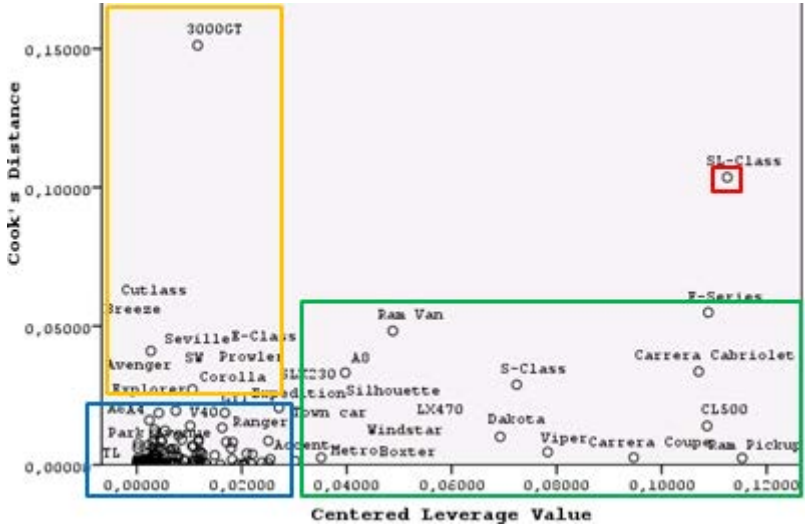


Для дальнейшего исследования построим матрицу остатков (Standardized Residual) как функции цены (Zscore: price in thousands) мастером графиков (***Graphs->Chart Builder->Scatter/Dot->Simple Scatter***), использовав переменную ***Model*** для подписей (через ***Groups/Point ID tab ->Point ID Label***). Из матрицы (Рис. слева) становится очевидно, что отклоняющиеся точки в остатках описываемой величины – дорогие машины. Очевидно, в распределении цен существует правая асимметрия. Поэтому модель с лог-нормированными ценами, возможно, даст лучшие результаты.

Строим аналогичную матрицу в отношении второй переменной ***Wheelbase*** (Рис. справа). Из нее видно, что модельный состав правого кластера ☐ такой же, что и правого кластера на матрице остатков как функции предсказываемой переменной (показан тем же цветом на предыдущем Рис.). Так же, как с случае цен, можно попробовать логарифмировать эту переменную.



В заключение построим «матрицу моментов» (описана в конце предыдущего раздела) из переменных рычага ***Leverage value*** (переменная ***LEV***) и силы (веса) ***Cook's Distance*** (***COO***).



Как уже обсуждалось ранее (в конце предыдущего раздела), опасность для расчетов представляют точки, имеющие значимые и вес, и рычаг, дающие заметный момент ***LEV*COO***, к которым относятся *SL-class* ☐. Лежащие вне основного кластера ☐ точки имеют либо большой рычаг, но малый вес ☐, либо большой вес, но малый рычаг ☐. В результате в обоих случаях получается малый результирующий момент. Поэтому данные кластеры не вносят заметные искажения в результаты, в том числе, марка 3000GT,

казавшаяся проблематичной. Зато выявилась другая модель SL-class, имеющая большую цену и колесную базу, и являющаяся причиной проблем в матрице остатков.

Итоги. При помощи множественной линейной регрессии мы построили «наилучшую» модель, предсказывающую объем продаж. Выяснилось, что из множества данных значимые лишь два – цена и колесная база, были найдены переоцененные и недооцененные модели.

Было обнаружено, что на результаты модели могут влиять асимметрия исходных данных и выбросы, предложены варианты решения проблемы (логарифмирование цен и колесной базы).

В теоретическом плане изучена методика множественной регрессии, показаны способы построения и оценки качества модели.

4.5. Задания для самостоятельной работы.

1. Построить три расширенные регрессионные модели рассмотренной задачи, включив в первую из них переменную «Тип Машины» (***Vehicle type***), в другую – «Размер Двигателя» (***Engine size***), а в третью – оба этих показателя. Провести сравнительный анализ качества моделей.

2. Построить модель с логарифмированными ценами и колесной базой, выбросить переоцененную точку, сделать сравнительный анализ результатов этой модели с представленными выше.

3. Построить три расширенные регрессионные модели рассмотренной задачи, включив в первую из них переменную «Тип Машины» (***Vehicle type***), в другую – «Размер Двигателя» (***Engine size***), а в третью – оба этих показателя. Провести сравнительный анализ качества моделей.

4. Построить модель с логарифмированными ценами и колесной базой, выбросить переоцененную точку, сделать сравнительный анализ результатов этой модели с представленными выше.

5. Факторный анализ.

Факторный анализ применяется в многопараметрических моделях для **уменьшения размерности** (**data reduction**) за счет поиска оптимальных факторизованных переменных. При этом устраняются проблемы, связанные с коррелированными переменными.

Кроме того, при помощи него осуществляется **поиск значимых структур и взаимосвязей в многомерных данных** (**structure detection**). Выявление такой изначально скрытой (латентной) информации является задачей популярного современного направления анализа данных исследований (**data mining**), призванного находить полезную информацию в огромных массивах данных, генерируемых фирмами в ходе своей коммерческой деятельности.

Существует несколько механизмов поиска решений (**extraction methods**) методом факторного анализа:

Уменьшение размерности. Базовый метод начинается с поиска первой **компоненты (фактора)**, являющейся линейной комбинации исходных переменных, максимальным образом описывающей вариацию исходных данных. Затем осуществляется поиск второй компоненты, не коррелированной с первой компонентой и максимально описывающей остаток вариации переменных, не описанный первой компонентой. Далее процесс продолжается до тех пор, пока количество найденных компонент не станет равным количеству исходных переменных. В большинстве случаев несколько первых найденных компонент описывают большую часть вариации данных, и в этом случае найденные компоненты могут заменить набор исходных данных в анализе системы (объекта).

Поиск структур. Далее делаем шаг вперед и выдвигаем предположение, что часть вариации данных не может быть описана компонентами (факторами). Тем самым, вариация, объясняемая факторными переменными, меньше, чем исходными, однако добавление структуры найденных факторов в факторную модель дает идеальную возможность анализа взаимосвязей между переменными.

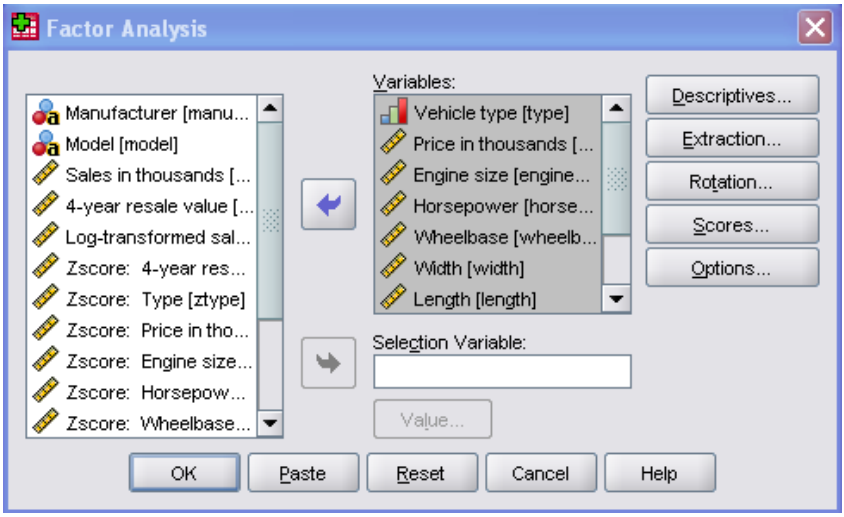
Для любого метода поиска возникают два главных вопроса: «Какое количество компонент нужно использовать для анализа?», «Какой смысл несут новые факторные переменные?»

5.1. Прогноз продаж автомашин.

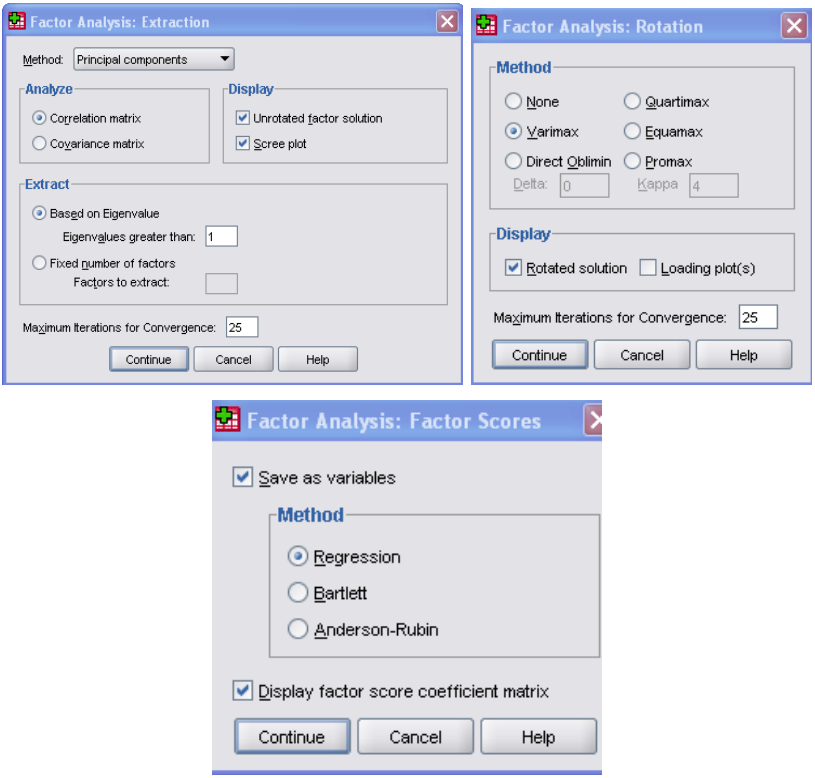
Задача ФА1: Индустриальный аналитик хочет спрогнозировать продажи автомашин, исходя из анализа ряда исходных данных (предикаторов). Однако, исходные данные коррелированны и могут исказить результаты. Необходимо сформировать ряд надежных показателей при помощи факторного анализа.

Используемый файл данных (директория Samples в SPSS): **car_sales.sav**.

Открываем файл и вызываем факторный анализ ***Analyze->Dimension Reduction->Factor…*** Выбираем переменные от ***Vehicle type*** до ***Fuel efficiency*** (для выделения группы переменных используйте мышку в комбинации с клавишей *Shift*).



Для установки опций метода нажимаем кнопку ***Extraction***, где выбираем ***Scree plot***, затем нажимаем ***Rotation*** и выбираем ***Varimax*** в разделе методов. Наконец, нажав по кнопке ***Score***, активируем опции ***Save as variables*** и ***Display factor score coefficient matrix***:



При нажатии ***OK*** запускается **метод главных компонент** (***principal components extraction***), которые потом подвергаются вращению с целью облегчения интерпретации. Компоненты с собственными значениями (***eigenvalues***) >1 сохраняются в рабочем файле для того, чтобы их можно было бы использовать в дальнейшем анализе.

Общности

	Начальные	Извлеченные
Vehicle type	1,000	,930
Price in thousands	1,000	,876
Engine size	1,000	,843
Horsepower	1,000	,933
Wheelbase	1,000	,881
Width	1,000	,776
Length	1,000	,919
Curb weight	1,000	,891
Fuel capacity	1,000	,861
Fuel efficiency	1,000	,860

Метод выделения: Анализ главных компонент.

Таблица общностей (***communalities***) демонстрирует качество описания вариаций исходных переменных найденными компонентами. Если изначально ☐ они все равнялись 1 (для метода главных компонент это всегда так), то колонка «Извлеченные» ⌐ показывает, какую степень вариации каждой исходной переменной описывает совокупность найденных компонент. В данном примере видно, что описание хорошее (не худшем случае 78% для переменной ***Width***). Если общности малы, то может потребоваться дополнительный расчет компонент.

Таблица «Полная объясненная дисперсия» (Variance explained) показывает долю объясненных вариаций в разбивке по найденным компонентам. В первой секции таблицы показаны результаты по собственным значениям. В первой колонке «Итого» ☐ показаны собственные значения, а именно, доля вариации начальных переменных, описанная каждой компонентой. К примеру, первая компонента описывает почти 6 исходных показателей (напомним, что по определению, каждая последующая найденная компонента слабее предыдущей). Во второй и третьей колонках «% Диспер-

сии» и «Кумулятивный %» отражена доля вариации всех переменных, описываемая каждой компонентой по отдельности и нарастающим (кумулятивным) итогом соответственно.

Компонента	Начальные собственные значения		
	Итого	% Дисперсии	Кумулятивный %
1	5,994	59,938	59,938
2	1,654	16,545	76,482
3	1,123	11,227	87,709
4	,339	3,389	91,098
5	,254	2,541	93,640
6	,199	1,994	95,633
7	,155	1,547	97,181
8	,130	1,299	98,480
9	,091	,905	99,385
10	,061	,615	100,000

В начальном решении число компонент равно числу переменных и в корреляционном анализе сумма собственных значений равна числу компонент. В опциях мы выбрали сохранение только компонент с собственными значениями >1, то есть трех первых компонент данного решения.

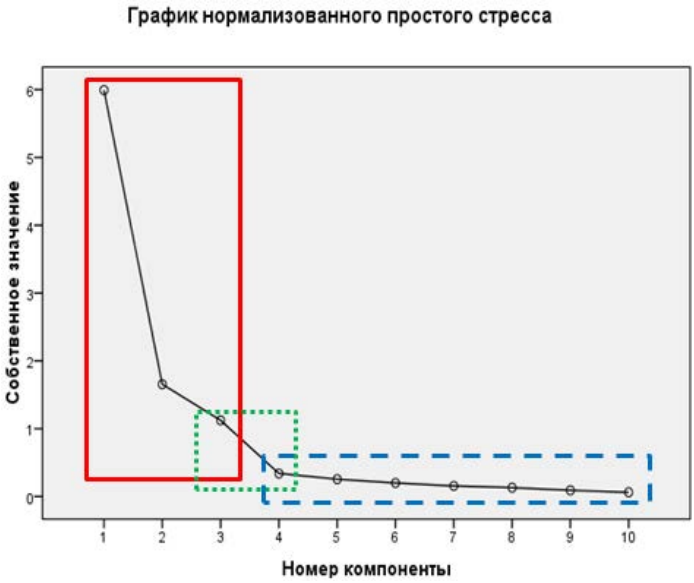
Вторая секция «Суммы квадратов нагрузок извлечения» (***Extraction Sums of Squared Loadings***) показывает отдельно только сохраненные компоненты. Как видно из таблицы, три компоненты описывают почти 88% исходной вариации, поэтому можно существенно уменьшить сложность анализа, перейдя от анализа 10 компонент к 3 с потерей всего лишь 12% информации. Последняя третья секция «Суммы квадратов нагрузок вращения» (***Rotation Sums of Squared Loadings***) показывает показатели компонент после применения к ним специального преобразования (вращение матрицы, отсюда и названия), задача которого – максимально равномерное распределение собственных значений, что видно из сопоставления колонок «Итого» во 2–3 секциях.

Суммы квадратов нагрузок извлечения		
Итого	% Дисперсии	Кумулятивный %
5,994	59,938	59,938
1,654	16,545	76,482
1,123	11,227	87,709

Суммы квадратов нагрузок вращения		
Итого	% Дисперсии	Кумулятивный %
3,220	32,199	32,199
3,134	31,344	63,543
2,417	24,166	87,709

Это делается из тех соображений, что, чем больше собственное значение компоненты, тем проще ее интерпретация в аспекте исходных переменных. Поэтому компоненты после примененной процедуры вращения легче интерпретировать, чем до него.

Стрессовый график (***Scree plot***) дает удобное визуальное представление о поведении собственных значений компонент. Из него видно, что три компоненты начальной ступеньки дают максимальный вклад в описание исходной вариации, а компоненты на пологом подножии практически не добавляют полезной информации. В данном примере разрыв между 3 и 4 компонентами (разрыв между ступенькой и подложкой) очевиден и выбор количества используемых компонент очевиден. Гораздо менее очевиден выбор при более пологом спадании стрессового графика.



Матрица повернутых компонент (***rotated component matrix***) помогает установить характер связи между компонентами и исходными переменными.

Первая компонента ☐ наиболее сильно связана с ценой (***Price in thousands***) и мощностью (***Horsepower***) машины. Причем цена – лучший представитель, поскольку она меньше коррелированна с другими двумя выбранными компонентами. Вторая компонента ⌐ наиболее сильно связана с длиной машины (***Length***). Наконец, третья компонента ⋯ наиболее сильно связана с типом машины (***Vehicle type***).

Таким образом, у нас есть два способа действия.

1. Первый и наиболее простой: сфокусироваться на трех основных переменных (цена, длина, тип).

2. Второй и лучший – воспользоваться значениями сохраненных компонент (***component scores***), рассчитанными из начальных переменных для каждой строки данных с учетом коэффициентов, представленных в матрице повернутых компонент. Как уже отмечалось, использование этих компонент гарантирует всего 12% потери исходной информации.

Второй способ также предпочтительней тем, что компоненты представляют информацию обо всех 10 исходных переменных, и линейно не коррелированы друг с другом. При этом по-прежнему необходимо проверять наличие выбросов и нелинейных корреляций между параметрами. Для этого вызываем мастер ***Graphs->Chart Builder->Scatter/Dot->Scatterplot Matrix*** и перетаскиваем все три переменные ***REGR factor score 1-3*** для построения матрицы рассеяния. В клетке i-той строки и j-того столбца изображаются матрицы рассеяния i-той компоненты (по Y) и j-того столбца (по X). Из графиков видно, что первая компонента имеет ассиметричное (skewed) распределение, и причиной этому – ассиметрия исходной переменной цена (***Price***). Поэтому факторный анализ с использованием логарифмированной цены может дать лучший результат. Разделение на графиках с третьей компонентой связано с тем, что определяющая ее компонента – тип машины – является бинарной переменной (0 – автомобиль, 1 – грузовик). Между первой и третьей компонентой видна связь – все дорогие машины – автомобили. Возможно, эта проблема уменьшится при логарифмировании цены. Если же это не поможет, можно рассмотреть вопрос о раздельной обработке данных по автомобилям и грузовикам.

Таким образом, количество исследуемых переменных может быть уменьшено с 10 до 3 базовых компонент. Отметим, что, хотя интерпретация результатов немного усложняется, однако достоинства снижения размерности и использование некоррелированных предикаторов перевешивают потери.

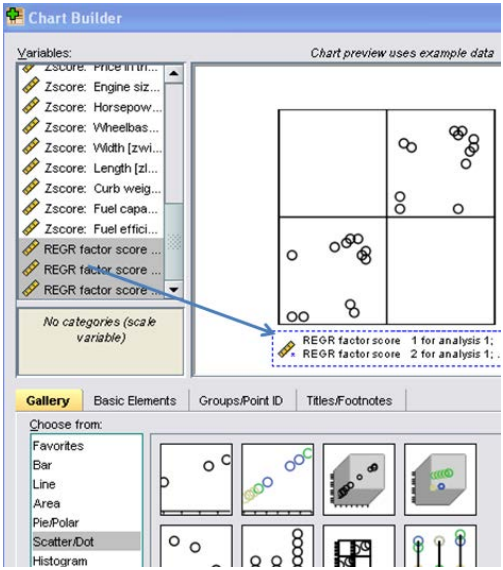
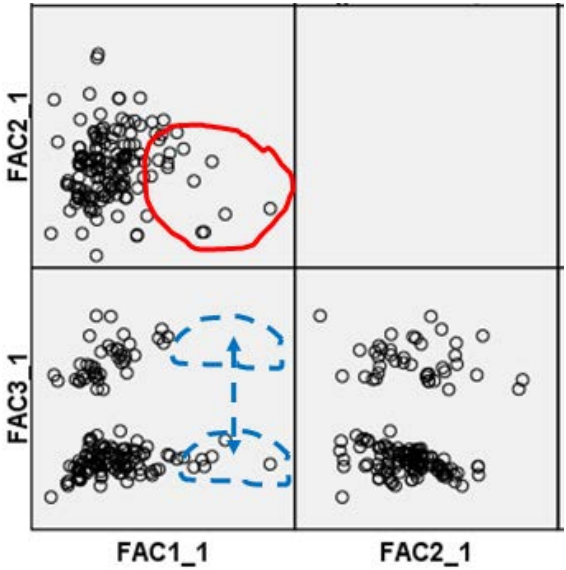


Chart Builder
Variables:
Chart preview uses example data
Zscore: Engine siz...
Zscore: Horsepow...
Zscore: Wheelbas...
Zscore: Width [zwi...
Zscore: Length [zl...
Zscore: Curb weig...
Zscore: Fuel capa...
Zscore: Fuel effici...
REGR factor score ...
REGR factor score ...
REGR factor score ...
No categories (scale variable)
REGR factor score 1 for analysis 1;
REGR factor score 2 for analysis 1;
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram

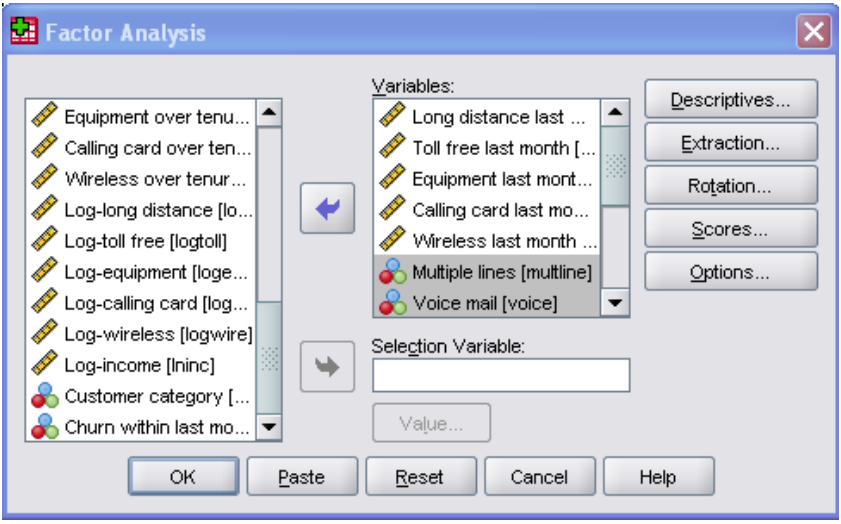

FAC2_1
FAC3_1
FAC1_1
FAC2_1

5.2. Структурирование сервисов провайдера.

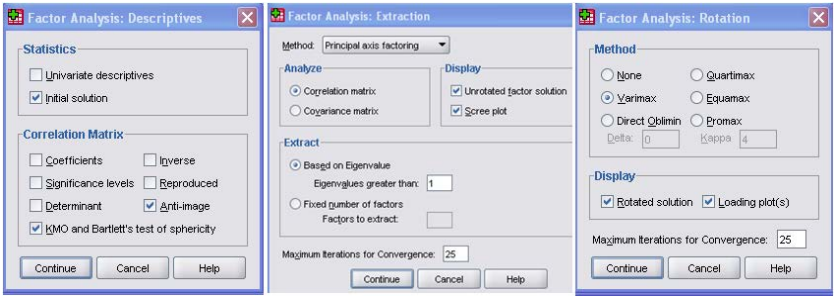
Задача ФА2. Провайдер телекоммуникаций хочет лучше понять поведение пользователей (выделить структуру, паттерны их поведения) на основании своей клиентской базы. Если удастся осуществить кластеризацию используемых сервисов, компания может предложить более выгодные пакеты своим пользователей. Используя методы факторного анализа, выделите структуру использования сервисов компании пользователями.

Используемый файл данных (директория Samples в SPSS): **telco.sav**.

Открываем файл и вызываем факторный анализ ***Analyze->Dimension Reduction->Factor…*** Выбираем анализируемые переменные от дальней связи ***Long distance last month*** до безпроводной ***Wireless last month***, и от ***Multiple lines*** до ***Electronic billing***.



Для установки опций метода нажимаем кнопку ***Descriptives***, выбираем ***Anti-image*** и ***KMO and Bartlett's test of sphericity***. На кнопке ***Extraction*** выбираем метод ***Principal axis factoring*** и ***Scree plot***, затем нажимаем ***Rotation*** и выбираем ***Varimax*** в разделе методов и ***Select Loading plot(s)*** в ***Display group***:



При нажатии ***OK*** активируется факторный анализ по методу принципиальных осевых компонент с последующим Varimax-вращением.

В первой таблице «Мера адекватности и критерий Бартлетта» представлены результаты двух тестов на адекватность использования факторного анализа для исходных данных:

1) КМО-тест адекватности выборки (***The Kaiser-Meyer-Olkin Measure of Sampling Adequacy***) – индикатор, показывающий долю вариации исходных данных, за которыми могут стоять скрытые факторные переменные (***underlying factors***) . Чем ближе величина к 1 (как в данном случае 0.88), тем лучше могут быть результаты факторного анализа. При величине КМО-теста < 0.5 факторный анализ данных малосодержателен.

2) Тест Бартлетта на сферичность проверяет гипотезу того, что корреляционная матрица является матрицей идентичности (***identity matrix***), что означало бы, что исходные переменные не связаны друг с другом и потому бесполезно искать заложенную в них структуру. Малые значения значимости (Знч.) <0.05 означают, что факторный анализ может быть полезен .

Мера адекватности и критерий Бартлетта

Мера выборочной адекватности Кайзера-Мейера-Олкина.		,888
Критерий сферичности Бартлетта	Прибл. хи-квадрат	6230,901
	ст.св.	91
	Знч.	,000

Начальные значения общностей (***communalities***) в корреляционном анализе являются долями вариации данной переменной, описываемыми всеми остальными переменными.

В колонке «Извлеченные» приводится тот же показатель для найденного факторного решения.

Общности

	Начальные	Извлеченные
Long distance last month	,297	,748
Toll free last month	,510	,564
Equipment last month	,579	,697
Calling card last month	,266	,307
Wireless last month	,660	,708
Multiple lines	,276	,340
Voice mail	,471	,501
Paging service	,527	,541
Internet	,455	,525
Caller ID	,552	,623
Call waiting	,545	,610
Call forwarding	,532	,596
3-way calling	,506	,561
Electronic billing	,416	,488

Метод выделения: Факторизация главных осей.

Переменные с малыми значениями (<0.5) плохо вписываются в найденное решение и, возможно, должны быть исключены из анализа. В данном случае результаты приемлемы почти для всех переменных, за исключением ***Multiple lines*** и ***Calling card***, вписавшихся в решение хуже других.

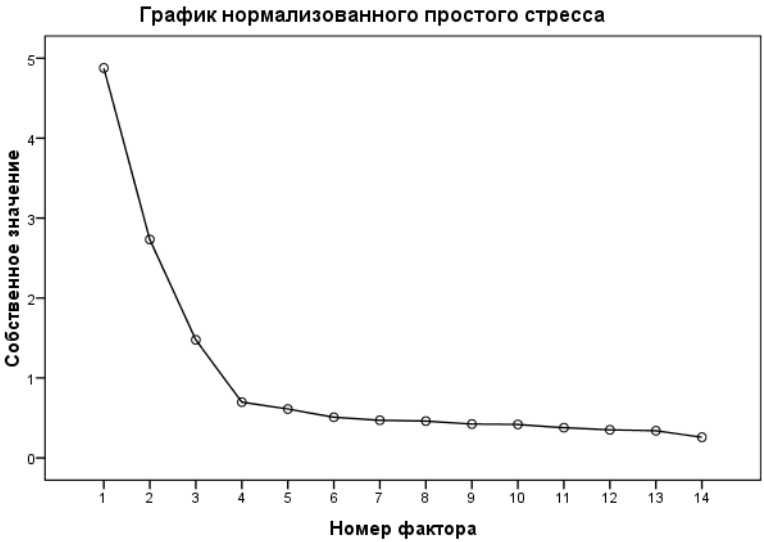
Основные показатели компонент представлены в таблица «Полная объясненная дисперсия» (***Variance explained***). В первой секции показаны доли вариации, объясненной начальным решением. Всего три компоненты имеют собственные значения >1, объясняя в обще совокупности 65% вариации всех исходных данных. Тем самым, можно предположить, что эти три латентных фактора ассоциированных с использованием сервисов, однако

существенная доля вариации (33%) остается необъясненной. Во второй секции таблицы «Суммы квадратов нагрузок извлечения» (***Extraction Sums of Squared Loadings***) показаны доли вариаций значимых компонент. Общая доля описанной вариации составила 56% по отношению к 65% начального решения. Разница 9% возникла за счет латентных факторов оригинальных переменных, а также вариации, которая не может быть объяснена факторной моделью. Наконец, в третьей секции (на Рис. справа внизу) «Суммы квадратов нагрузок вращения» (***Rotation Sums of Squared Loadings***) представлена решения после Varimax-вращения с целью выравнивания собственных значений компонент. Из таблицы видно , что выравниваются 1–2 компоненты, тогда как значение третьей практически не изменяется.

Фактор	Начальные собственные значения			Суммы квадратов нагрузок извлечения			Суммы квадратов нагрузок вращения		
	Итого	% Дисперсии	Кумулятивный %	Итого	% Дисперсии	Кумулятивный %	Итого	% Дисперсии	Кумулятивный %
1	4,877	34,838	34,838	4,461	31,864	31,864	3,747	26,763	26,763
2	2,733	19,518	54,357	2,297	16,409	48,273	2,872	20,514	47,277
3	1,476	10,546	64,903	1,053	7,519	55,792	1,192	8,515	55,792
4	,697	4,979	69,883						
5	,612	4,368	74,251						
6	,508	3,628	77,879						
7	,470	3,358	81,237						
8	,461	3,289	84,526						
9	,423	3,021	87,548						
10	,418	2,988	90,536						
11	,377	2,695	93,231						
12	,350	2,503	95,733						
13	,339	2,424	98,158						
14	,258	1,842	100,000						

Метод выделения: Факторизация главных осей.

Стрессовый график (***Scree plot***) подтверждает выбор трех значимых компонент.



Содержательность компонент более-менее ясна. В третьем факторе доминирует ***Long distance last month***. Во втором - ***Equipment last month***, ***Internet***, и ***Electronic billing***. Первая компонента имеет целый набор: ***Toll free last month***, ***Wireless last month***, ***Voice mail***, ***Paging service***, ***Caller ID***, ***Call waiting***, ***Call forwarding***, и ***3-way calling***.

Матрица факторов[a]

	Фактор		
	1	2	3
Long distance last month	,146	-,254	,814
Toll free last month	,652	-,373	,020
Equipment last month	,494	,671	,054
Calling card last month	,364	-,243	,339
Wireless last month	,799	,261	,037
Multiple lines	,257	,280	,442
Voice mail	,669	,228	-,038
Paging service	,692	,246	-,050
Internet	,323	,648	-,014
Caller ID	,689	-,345	-,172
Call waiting	,678	-,366	-,126
Call forwarding	,684	-,336	-,128
3-way calling	,662	-,338	-,093
Electronic billing	,250	,652	-,035

Однако некоторые факторы «первого порядка» одной компоненты связаны (позитивно и негативно) с факторами «второго порядка» в других компонентах. В целом, большое количество исходных данных (сервисов) имеют большие весовые коэффициенты (>0.2), что смазывает картину. Вращение матрицы должно поправить ситуацию. Матрица преобразования факторов показывает параметры Varimax-вращения оригинальной матрицы компонент в повернутую матрицу. Меньшие □/большие □значения недиагональных элементов соответствуют меньшим/большим вращениям.

Матрица преобразования факторов

Фактор	1	2	3
1	,837	,516	,184
2	-,497	,856	-,140
3	-,230	,026	,973

Метод выделения: Факторизация главных осей
Метод вращения: Варимакс с нормализацией Кайзера.

В матрице повернутых компонент (***rotated component matrix***) ситуация с третьей компонентой практически не изменилась (а с ней все было ясно и в оригинальной матрице), а вот первые две компоненты интерпретировать стало легче. Первая наибольшим образом связана с ***Toll free last month***, ***Caller ID***, ***Call waiting***, ***Call forwarding***, и ***3-way calling***, которые практически не коррелированы с другими компонентами. Второй фактор наиболее коррелирован с ***Equipment last month***, ***Internet***, и ***Electronic billing***. Тем самым, сформировались три основные группы сервисов, назовем их условно «Экстра», «Техника» и «Дальняя связь» по специфике наиболее коррелированных исходных данных. После данной группировки нужно проанализировать оставшиеся сервисы, по которым можно сделать следующие выводы:

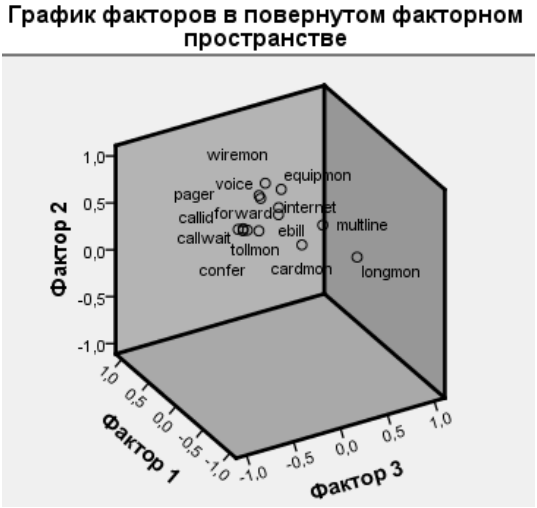
Матрица повернутых факторов[a]

	Фактор		
	1	2	3
Long distance last month	,062	-,121	,854
Toll free last month	,726	,018	,191
Equipment last month	,067	,831	,049
Calling card last month	,348	-,012	,431
Wireless last month	,530	,637	,146
Multiple lines	-,025	,384	,438
Voice mail	,455	,539	,054
Paging service	,468	,566	,044
Internet	-,049	,722	-,045
Caller ID	,787	,056	,008
Call waiting	,779	,033	,054
Call forwarding	,768	,062	,048
3-way calling	,743	,050	,078
Electronic billing	-,107	,686	-,080

- Из-за существенных корреляций с 1–2 компонентами сервисы ***Wireless last month***, ***Voice mail***, и ***Paging service*** тяготеют одновременно к группам «Экстра» и «Техника».
- Сервис ***Calling card last month*** умеренно коррелирован с первым и третим факторами, то есть находится на стыке групп «Экстра» и «Дальняя связь».
- По тем же причинам ***Multiple lines*** коррелирован с 2–3 компонентами и лежит на стыке групп «Техника» и «Дальняя связь».

Эти зависимости подсказывают комбинированные продажи. К примеру, пользователи экстра сервисов могут быть более предрасположены к покупке специальных предложений по беспроводной связи, чем других сервисов.

График факторных нагрузок (***Factor loading plot***) дает визуальное представление о матрице вращения. В сложных случаях он может помочь интерпретировать взаимосвязи факторов.



Резюме: используя метод поиска принципиальных осевых компонент, мы обнаружили три латентных фактора, описывающие соотношения между переменными. Анализ этих компонент продемонстрировал паттерны использования сервисов, которые можно использовать для увеличения комбинированных продаж.

Факторный анализ пытается обнаружить внутренние переменные (компоненты), определяющие паттерны корреляций, стоящих за исходными данными. Эта процедура наиболее часто используется для уменьшения размерности исходных данных, но также может применяться для поиска скрытой (латентной) структуры, присутствующей в исходных данных.

При подозрениях на нелинейные корреляции между переменными попробуйте применить более подходящий для этого случая анализ бивариационных корреляций (***Bivariate Correlations***)

Для не размерных переменных можно попробовать воспользоваться иерархическим кластерным анализом (***Hierarchical Cluster Analysis***), как альтернативой факторного анализа для поиска структур в данных.

5.3. Вопросы для самоконтроля.

1. В каких моделях и с какой целью применяется факторный анализ?
2. Каковы преимущества и каковы недостатки факторного анализа?
3. Для чего требуется «уменьшение размерности» данных, каковы плюсы и минусы этого подхода?
4. Что такое «метод главных компонент»
5. Что такое стрессовый график, и для чего он применяется.
6. Как определить значимые компоненты факторного анализа?
7. Какие тесты применимости факторного анализа к конкретной выборке вы знаете?
8. Для чего применяются КМО-тест и тест Бартлетта?
9. Как определяется качество факторного анализа (объясняемая вариация)?
10. Для чего используется график факторных нагрузок?

6. Кластерный анализ.

Под кластерным анализом подразумевают технику группировки (классификации) начальных переменных по близости свойств. Существует разные методики такой группировки, из которых мы рассмотрим наиболее популярные.

6.1. Двушаговый кластерный анализ (TwoStep Cluster Analysis).

Двушаговый кластерный анализ (ДКА) относится к исследовательским методикам, использующимся для обнаружения изначально не очевидных группировок (кластеров) внутри исходных данных. Данный алгоритм имеет несколько полезных свойств, отличающих его от других методик:

- Возможность поиска кластера, как по категориальным, так и по непрерывным переменным.
- Автоматическое определение количества кластеров.
- Возможность эффективного анализа больших объемов данных.

Для совместной обработки категориальных и непрерывных переменных процедура ДКА использует вероятностную дистанционную метрику, предполагающую независимость переменных кластерной модели. Кроме того, предполагается, что каждая непрерывная переменная имеет нормальное (Гаусово) распределение, а каждая категориальная – мультиноминальное. Эмпирические внутренние проверки показывают, что процедура достаточно устойчива к нарушениям обеих предположений, однако все равно необходимо следить за тем, насколько хорошо выполняются указанные предположения. Методика ДКА в целом выглядит следующим образом:

Шаг первый. Сначала строится дерево кластерных свойств (***Cluster Feature Tree***, ***CFT***). Строительство дерева начинается с помещения первой записи в лист дерева, содержащий информацию о данной записи. Последующие записи либо добавляются в существующий лист, либо формируют новые листья в зависимости от «степени похожести» листьев, рассчитывающейся на основании дистанционной метрики как критерия похожести. Узел, содержащий множество листьев, содержит расчетную суммарную информацию от всех входящих в него членов. Тем самым, CFT обеспечивает капсулированную интегральную информацию о данных.

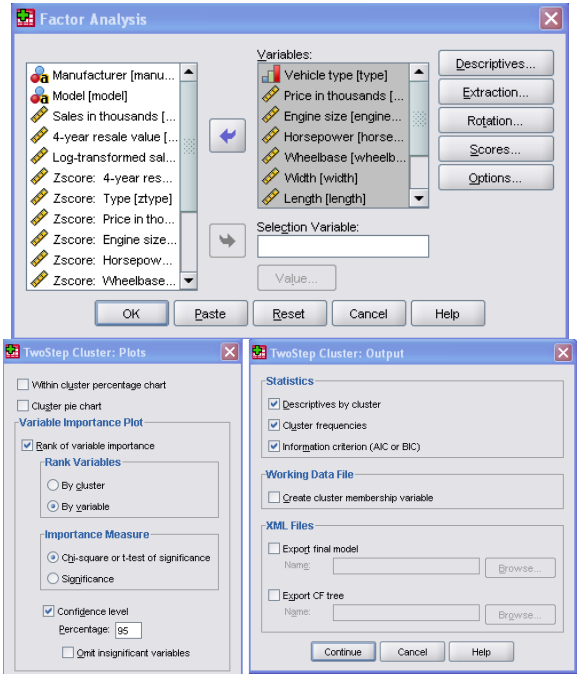
Шаг второй. Производится группировка листьев CFT на основе агломерационного кластерного алгоритма, предлагающего ряд решений. Для определения лучшего решения производится сравнение каждого из них по кластеризационным критериям: Шварцевскому Баейсовскому критерию (***Schwarz's Bayesian Criterion***, ***BIC***) или информационному критерию Акейка (***Akaike Information Criterion***, ***AIC***).

Задача КА1. Производителям автомобилей необходимо оценить рынок с целью определить наилучшую нишу для своих автомобилей. Если машины могут быть сгруппированы по имеющимся данным, эта задача может быть в большей части автоматизирована при использовании кластерного анализа. Необходимо сгруппировать машины по цене и физическим свойствам, используя ДКА.

Используемый файл данных (директория Samples в SPSS): **car_sales.sav**.

Открываем файл и вызываем факторный анализ ***Analyze->Classify->Two Step Cluster…*** Выбираем категориальную переменную ***Vehicle type*** и непрерывные от ***Price in thousands*** до ***Fuel efficiency*** (для выделения группы переменных используйте мышку в комбинации с клавишей *Shift*).

Для установки опций метода нажимаем кнопку Графики (***Plots***), где выбираем изображение графика рейтинга важности переменных (***Rank of variable importance***), выбор по переменным (***By variable*** в ***Rank Variables group***). Кроме того, выбираем построение конфиденциального интервала (***Confidence level***). По кнопке Вывод (***Output***) выбираем информационный критерий ***Information criterion (AIC or BIC)*** в группе статистики.



При нажатии ***OK*** запускается ДКА. В таблице «Автоматическая кластеризация» отражен процесс формирования кластеров по их количеству. Кластеризационный критерий (в данном случае - BIC) рассчитан для каждого потенциального числа кластеров .

Меньшие значения BIC соответствуют лучшим моделям, тем самым, лучшее кластерное решение имеет минимальный BIC.

Автоматическая кластеризация

Число кластеров	Байесовский информационный критерий Шварца (BIC)	Изменение BIC[a]	Отношение изменений BIC[b]	Отношение мер расстояния[c]
1	1214,377			
2	974,051	-240,326	1,000	1,829
3	885,924	-88,128	,367	2,190
4	897,559	11,635	-,048	1,368
5	931,760	34,201	-,142	1,036
6	968,073	36,313	-,151	1,576
7	1026,000	57,927	-,241	1,083
8	1086,815	60,815	-,253	1,687
9	1161,740	74,926	-,312	1,020
10	1237,063	75,323	-,313	1,239
11	1316,271	79,207	-,330	1,046
12	1396,192	79,921	-,333	1,075
13	1477,199	81,008	-,337	1,076
14	1559,230	82,030	-,341	1,301
15	1644,366	85,136	-,354	1,044

Однако существуют случаи, когда BIC продолжает снижаться при увеличении количества кластеров, хотя улучшение кластерного результата согласно измерению BIC (***BIC Change***) не стоит увеличения сложности модели с большим количеством кластеров. В таких случаях для определения лучшего решения используют показатели изменения в BIC и метрике расстояний. Хорошее решение будет иметь относительно большие значения Отношения изменения BIC (***Ratio of BIC Changes***) и Отношения мер расстояния (***Ratio of Distance Measures***).

Распределение по кластерам

		N	% объединенных	% от итога
Кластер	1	62	40,8%	39,5%
	2	39	25,7%	24,8%
	3	51	33,6%	32,5%
	Объединенный	152	100,0%	96,8%
Исключенные наблюдения		5		3,2%
Итог		157		100,0%

Таблица «Распределение по кластерам» показывает размерность каждого кластера. В данном случае 5 записей было исключено из анализа из-за отсутствующих значений переменных.

Таблица «Центроиды» (***Centroids***) показывает среднее и стандартное отклонение по кластерам. Поскольку по умолчанию она изображается слишком широкой таблицей, для удобства просмотра переформатируем ее (***pivot***) в более удобный формат. Для этого в окне выходных данных активируем таблицу двойным щелчком мыши и выберем меню ***Pivot->Pivoting Trays***. Перетаскиваем Номер кластера (***Cluster ID***) в заголовки колонки, а остальные переменные Статистика (***Statistics***) и Непрерывные переменные (***Continuous Variables***) – в заголовки рядов и закрываем таблицу форматирования.

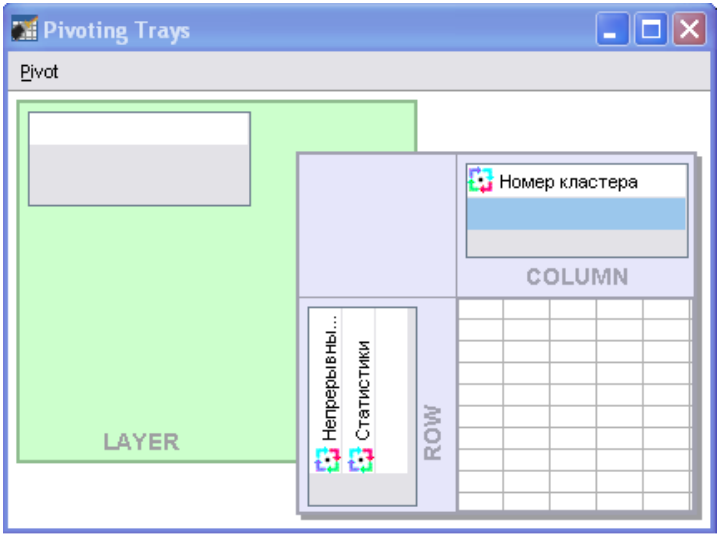


Таблица показывает, что три кластера хорошо разделяются по исходным непрерывным переменным:

1) Машины в первом кластере недорогие, небольшие и экономичны по потреблению топлива) .

2) Машины второго кластера средней цены, тяжелые, с большим баком, возможно, для компенсации их большого расхода горючего .

3) Машины третьего кластера дороги, большие со средним расходом топлива .

Центроиды

		Кластер			
		1	2	3	Объединенный
Price in thousands	Среднее	19,61671	26,56182	37,29980	27,33182
	Стд. отклонение	7,644070	10,185175	17,381187	14,418669
Engine size	Среднее	2,194	3,559	3,700	3,049
	Стд. отклонение	,4238	,9358	,9493	1,0498
Horsepower	Среднее	143,24	187,92	232,96	184,81
	Стд. отклонение	30,259	39,049	54,408	56,823
Wheelbase	Среднее	102,595	112,972	109,022	107,414
	Стд. отклонение	4,0799	9,6537	5,7644	7,7178
Width	Среднее	68,539	72,744	72,924	71,089
	Стд. отклонение	1,9366	4,1781	2,1855	3,4647
Length	Среднее	178,235	191,110	194,688	187,059
	Стд. отклонение	9,6534	14,4415	10,3512	13,4712
Curb weight	Среднее	2,83742	3,96759	3,57890	3,37618
	Стд. отклонение	,310867	,671766	,297204	,636593
Fuel capacity	Среднее	14,979	22,064	18,443	17,959
	Стд. отклонение	1,8699	4,2894	2,0445	3,9376
Fuel efficiency	Среднее	27,24	19,51	23,02	23,84
	Стд. отклонение	3,578	2,910	2,060	4,305

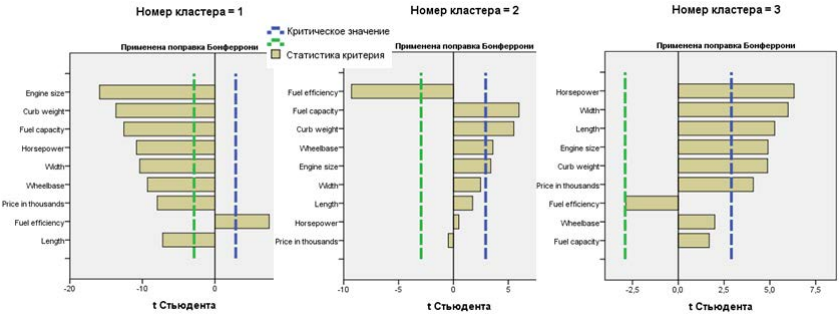
Частотная таблица по типам автомобилей (***Vehicle type***) подтверждает эти выводы, при этом вторую группу составили грузовики.

Vehicle type

		Automobile		Truck	
		Частота	Процент	Частота	Процент
Кластер	1	61	54,5%	1	2,5%
	2	0	,0%	39	97,5%
	3	51	45,5%	0	,0%
	Объединенный	112	100,0%	40	100,0%

Комбинированный график важности переменных (***"by variable" importance chart***) объединяет графики отдельных переменных, и построен в убывающем порядке их важности. Вертикальные пунктирные линии показывают критические значения, определяющие значимость каждой переменной. Для значимых переменных величина t-статистики должна превышать критическое значение любого знака. Для первого кластера все переменные оказываются значимыми. Отрицательная t-статистика означает, что переменная в целом имеет меньшее значение, чем средняя величина по кла-

стеру, и наоборот. То есть для первого кластера экономичность по кластеру больше, чем средняя по данным, а все остальные переменные – меньше, что подтверждает выводы анализа таблицы «Центроиды».



Результаты для второго кластера (грузовики) показывают не значимость переменных ***Width***, ***Length***, ***Horsepower***, и ***Price in thousands*** для формирования кластера. В третьем кластере не важны ***Wheelbase*** и ***Fuel capacity***, в то время как параметр экономичности ***Fuel efficiency*** на гране значимости.

Резюме. Используя ДКА, мы разделили машины на три достаточно широких категории. Дальнейшая детализация внутри групп возможно при наличии дополнительной информации.

ДКА работает с непрерывными и категориальными переменными и удобен для натуральной группировки данных, включая выборки больших размеров.

Для данных небольших размеров, выбирая между несколькими методами, способами трансформации переменных и желая измерять различия между кластерами, можно попытаться использовать **иерархический кластерный анализ** (ИКА), о котором речь пойдет далее. Отличительным достоинством ИКА является возможность кластеризации не только данных, но и исходных переменных.

6.2. Иерархический кластерный анализ (Hierarchical Cluster Analysis).

Иерархический кластерный анализ (ИКА) – поисковый метод, призванный обнаруживать групповые взаимосвязи, не очевидные

при обычном рассмотрении данных. Он особенно эффективен для данных небольшого размера (менее нескольких сотен).

Объектами ИКА по выбору пользователя могут быть данные, либо переменные.

На начальной стадии ИКА каждый объект сам по себе является кластером. На каждом шаге анализа происходит ослабление **критерия**, разделяющего кластеры до той степени, чтобы объединить два наиболее схожих кластера в один. Далее процедура повторяется до полного выстраивания дерева классификации.

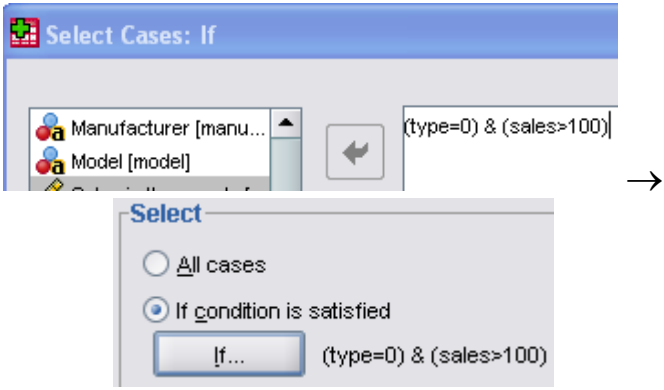
Базовым критерием любой кластеризации является расстояние. Близкие объекты должны принадлежать одному кластеру, а удаленные – к разным. Кластеризация выборки зависит от того, как определить следующие параметры:

- **Кластерный метод** (***cluster method***) определяет правила формирования кластеров. К примеру, для определения расстояния между кластерами можно использовать пару самых близких кластеров, или пару самых удаленных, или компромисс между этими крайними случаями.
- **Метрика** (***measure***) определяет формулу расчета расстояния. Например, Евклидова метрика определяет расстояние как «прямую линию» между двумя кластерами. Интервальная метрика предполагает, что переменные шкалируемы; в счетной метрике переменные являются дискретными числами, а в бинарной метрике переменные имеют два значения.
- **Стандартизация или нормировка** (***Standardization***) позволяет уравнивать переменные разного типа.

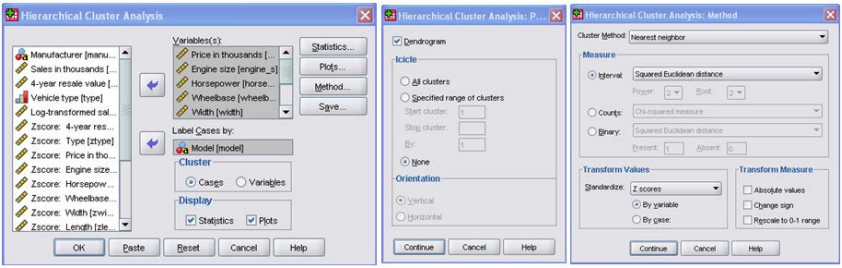
Задача КА2. Решить **Задачу КА1** методом ИКА, сосредоточившись на наиболее продаваемых марках машин: автомобилях, проданных в количестве более 100 тыс. единиц.

Используемый файл данных (директория Samples в SPSS): **car_sales.sav**.

Для создания анализируемой выборки воспользуемся отбором записей по условия. Выбираем меню ***Data->Select cases***, указываем опцию ***If condition is satisfied*** и нажимаем кнопку ***if*** (см. Рис.). В поле условия компонуем необходимое условие, используя встроенный редактор: ***(type=0) & (sales>100)***:



Вызываем мастер ИКА через ***Analyze->Classify->Hierarchical Cluster…*** Выбираем переменные от ***Price in thousands*** до ***Fuel efficiency*** (для выделения группы переменных используйте мышку в комбинации с клавишей *Shift*) и указываем ***Model*** в качестве маркировочной переменной. В опциях ***Plots*** выбираем ***Dendrogram*** и ***None*** в группе ***Icicle***. В опции ***Method*** выбираем ***Nearest neighbor*** и ***Z scores*** в качестве метода стандартизации в разделе ***Transform Values***.



При нажатии ***OK*** запускается ИКА. Дендрограмма – графическое представление найденного кластерного решения. По вертикальной оси перечислены метки исходных данных. По горизонтальной оси показано расстояние между кластерами на момент объединения. Анализ количества кластеров – субъективный процесс. Обычно смотрят на «щели (зазоры)» между моментами объединения кластеров (вертикальными объединительными линиями дерева), начиная справа (конечной стадии). В частности, заметен зазор между 20-25, разделяющий автомобили на два кла-

стера . Другая щель 9-14 в середине дерева указывает на существование 6 кластеров .

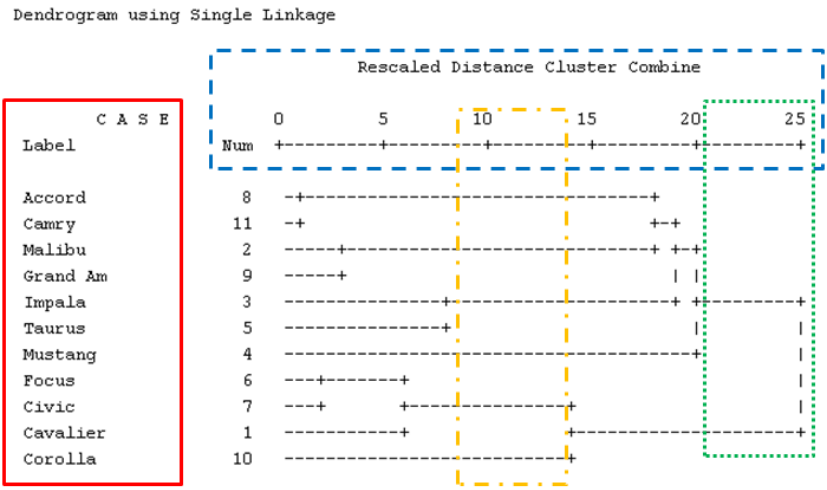
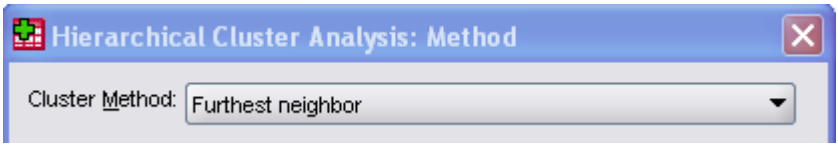


Таблица «Шаги агломерации» (***Agglomeration schedule***) представляет численный отчет о кластеризации. На первом этапе были объединены наиболее близкие записи 8 и 11, их совместный кластер в дальнейшем будет объединен на седьмом шаге . На седьмом шаге объединяются кластеры, созданные на шагах 1 и 3, и образовавшийся кластер далее появляется на шаге 8 . При большом количестве данных таблица может быть весьма длинной, однако часто ее все равно легче сканировать, чем дендограмму. При хорошем кластерном решении ожидается наибольший прыжок (зазор) между коэффициентами (расстояния). В данном случае наибольший зазор наблюдается между шагами 5 и 6, сигнализируя о хорошем 6-кластерном решении, и двукластерное хорошее решение наблюдается на шаге 9-10 . То есть аналогично тому, что мы видели на дендрограмме.

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	8	11	1,260	0	0	7
2	6	7	1,579	0	0	4
3	2	9	1,625	0	0	7
4	1	6	2,318	0	2	6
5	3	5	2,619	0	0	8
6	1	10	3,670	4	0	10
7	2	8	4,420	3	1	8
8	2	3	4,505	7	5	9
9	2	4	4,774	8	0	10
10	1	2	5,718	6	9	0

Полученные решения не совсем удовлетворительные, поскольку не наблюдается явно выраженной кластеризации. Попробуем другой кластеризационный метод полной линковки (***complete linkage***) или «дальнего соседа» ***Furthest neighbor*** при выборе опции в разделе ***Method***.



Первые несколько шагов данного метода аналогичны предыдущему. Но на конечных шагах (начиная с 5-6) разница весьма существенна, поскольку метод полной линковки дает более обособленные кластеры, особенно на последних двух шагах .

Шаги агломерации

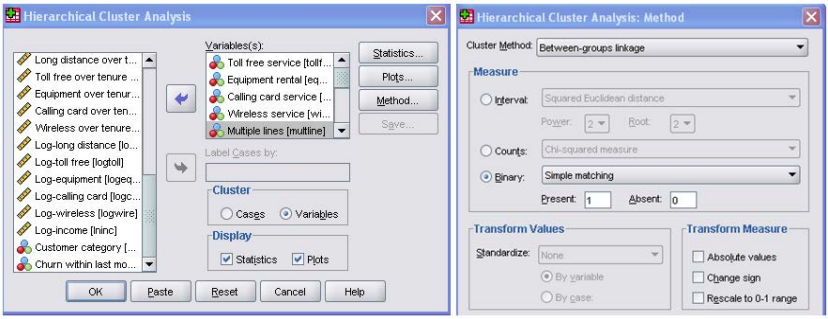
Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	8	11	1,260	0	0	7
2	6	7	1,579	0	0	5
3	2	9	1,625	0	0	6
4	3	5	2,619	0	0	6
5	6	10	4,012	2	0	9
6	2	3	7,333	3	4	8
7	1	8	9,183	0	1	9
8	2	4	12,440	6	0	10
9	1	6	25,486	7	5	10
10	1	2	54,607	9	8	0

Исключительность решения подтверждает дендрограмма, демонстрирующая формирование двух кластеров. Верхний кластер содержит меньшие машины. Нижний – большие. Кластер малых машин можно разделить далее на малые и экономичные машины. *Civic* и *Corolla* меньше и дешевле своих оппонентов *Accord* и *Camry* соответственно.

Таким образом, метод полной линковки (дальнего соседа) оказывается удовлетворительным, поскольку приводит к более явной кластеризации, чем метод ближнего соседа. Используя метод полной линковки, можно оценивать конкуренцию среди автомобилей на стадии разработки, задавая их спецификации в новых записях и прогоняя анализ по новой.

Задача КА3. Провайдер телекоммуникаций хочет лучше понять поведение пользователей (выделить структуру, паттерны их поведения) на основании своей клиентской базы. Если удастся осуществить кластеризацию используемых сервисов, компания может предложить более выгодные пакеты своим пользователей. Используя методы ИКА, исследуйте взаимосвязи между сервисами провайдера. **Примечание.** Ранее эта задача была решена методами факторного анализа (**Задача ФА2**).

Используемый файл данных (директория Samples в SPSS): **telco.sav**..



Вызываем мастер ИКА через ***Analyze->Classify->Hierarchical Cluster…*** Выбираем переменные от ***Toll free service*** до ***Wireless service*** и от ***Multiple lines*** до ***Electronic billing*** (для выделения группы переменных используйте мышку в комбинации с клавишей *Shift*). Поскольку кластеризировать будет переменные, а не данные, необходимо выбирать ***Variables*** в ***Cluster group***. В опциях ***Plots*** выбираем ***Dendrogram*** и ***None*** в группе ***Icicle*** (см. аналогичный Рис. к **Задаче КА2**). В опции ***Method*** оставляем метод по умолчанию ***Between-groups-linkage*** и выбираем бинарную метрику ***Binary*** в группе ***Measure*** методом простых совпадений (***Simple Matching***) и запускаем ИКА кнопкой ***OK***.

Использование бинарной метрики необходимо из-за наличия бинарных переменных, метод простых совпадений или метрика Жаккарда (***Jaccard measures***) – хорошие отправные методики для анализа.

Для бинарной метрики коэффициенты таблицы «Шаги агломерации» (***Agglomeration schedule***) содержат оценки похожести (***similarity***), которые убывают по ходу анализа ❑.

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	4	7	,861	0	0	2
2	4	6	,825	1	0	9
3	1	10	,823	0	0	5
4	2	8	,812	0	0	8
5	1	9	,812	3	0	6
6	1	11	,806	5	0	7
7	1	12	,795	6	0	10
8	2	13	,791	4	0	9
9	2	4	,707	8	2	11
10	1	3	,666	7	0	12
11	2	5	,623	9	0	12
12	1	2	,562	10	11	0

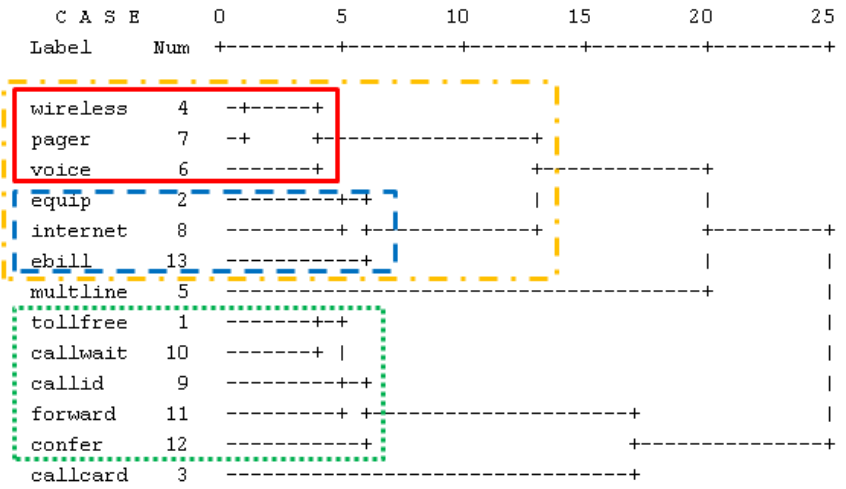
Полученные данные трудно интерпретировать, поэтому обратимся к дендрограмме. Прежде всего отметим, что сервисы ***MULTLINE*** и ***CALLCARD*** имеют особый статус, отличаясь от всех остальных. Далее, наблюдается формирование трех кластеров:

1) ***WIRELESS***, ***PAGER*** и ***VOICE***.

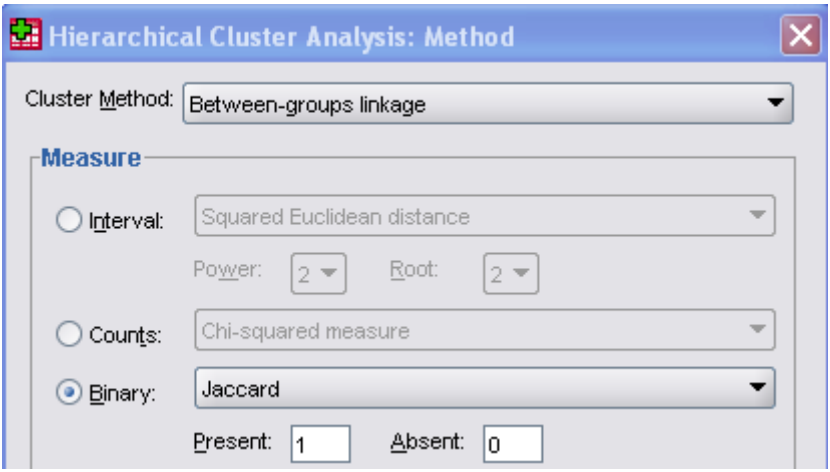
2) ***EQUIP***, ***INTERNET*** и ***EBILL***.

3) ***TOLLFREE***, ***CALLWAIT***, ***CALLID***, ***FORWARD*** и ***CONFER***.

Кроме того, первые два кластера ближе друг другу, чем сервисы из третьего кластера, в частности ***wireless*** и ***internet*** ближе друг другу, чем ***callwait***.

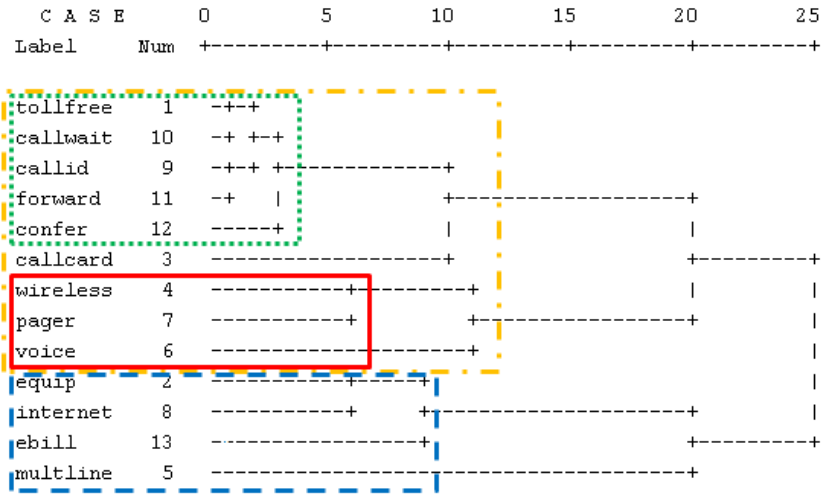


Перезапустим ИКА, используя метрику Жаккарда (***Jaccard measures***):



Из дендрограммы видно, что количество и состав кластеров остались, однако расстояния между кластерами изменились: теперь первый и третий кластеры (в первоначальной нумерации, приведенной выше) ближе друг другу, чем второй. В частности ***wireless*** и ***internet*** теперь дальше друг от друга, чем ***callwait***.

Разница между методом простых совпадений и метрикой Жаккарда в том, что последняя считает похожими два сервиса, только если пользователь подписан на оба. А методика простых совпадений вдобавок считает похожими сервисы, если пользователь не подписан ни на один из них. Отсюда разница в кластерных решениях между метриками, поскольку в выборке есть много пользователей, не подписанных ни на wireless, ни на internet. Поэтому содержащие их кластеры ближе друг другу в методике простых совпадений, чем в методике Жаккарда. Таким образом, используемая метрика зависит от того, какое из определений «похожести» лучше подходит для конкретной задачи.



Таким образом, мы рассмотрели двушаговый и иерархический кластерные анализы, еще один популярный метод – **кластерный анализ по К-среднему** (ККА), который хотя и ограничен шкалируемыми переменными, однако позволяет работать с большими объемами данных, сохранять расстояния между кластерами.

6.3. Кластерный анализ по К-среднему (K-mean Cluster Analysis).

Кластерный анализ по К-среднему (ККА) – инструмент, предназначенный для сортировки исходных данных к фиксированному, заранее заданному количеству кластеров с заранее неизвестными характеристиками. Он наиболее полезен в случае большой статистики (тысячи записей). Хороший КА:

- Эффективен (***efficient***) в смысле формирования минимально требуемого количества кластеров.
- Эффективен (***effective***) в аспекте вычленения всех статистически и коммерчески значимых кластеров. К примеру, кластер из 5 пользователей может значимо отличаться от остальной, при этом он может быть не интересен в плане коммерческой выгоды.

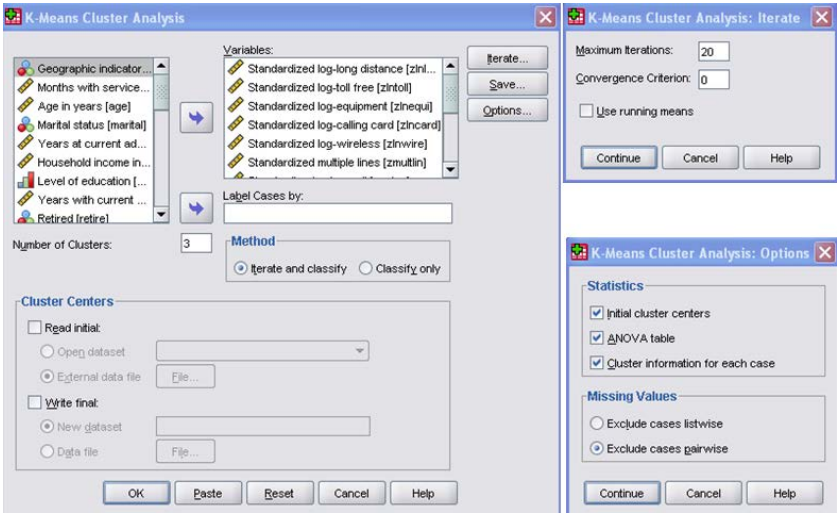
ККА начинается с выбора k начальных кластерных центров. Это делается либо вручную по выбору пользователя, либо по определенной процедуре, к примеру, выбору исходных k записей, хорошо удаленных друг от друга. После выбора начальных кластеров происходит:

- Присвоение записей исходных данных определенным центрам, исходя из расстояния от них до центров.
- Перерасчет центров кластеров, являющих собой усреднение по входящим в кластер исходным записям.

Описанная процедура повторяется до тех пор, пока переназначение записей между кластерами улучшает кластеризацию (делает кластеры внутренне более вариативными или внешне более похожими).

Задача КА4. Решить **задачу КА3** методом ККА для поиска «похожих» пользователей, используя стандартизованные переменные.

Используемый файл данных (директория Samples в SPSS):**telco_extra.sav**.



Вызываем мастер ККА через ***Analyze->Classify->K-mean Cluster…*** Выбираем переменные от ***Standardized log-long distance*** до ***Standardized log-wireless*** и от ***Standardized multiple lines*** до ***Standardized electronic billing*** (для выделения группы переменных используйте мышку в комбинации с клавишей *Shift*). Количество кластеров (***Number of clusters***) возьмите равным 3. В опциях ***Iterate*** выбираем 20 итераций максимально, а в опции Options выбираем ***ANOVA table*** и ***Cluster information for each group*** в группе ***Statistics***, а также ***Exclude cases pairwise*** в группе ***Missing Values***. Последнее связано с тем, что большинство пользователей не подписаны на все сервисы, поэтому попарное исключение позволяет максимизировать статистику исследований. Платой за это является возможный систематический сдвиг (***bias***) результата.

При нажатии ***OK*** запускается ККА. Данные установки генерят следующий командный синтаксис (в принципе, может быть задан через редактор, минуя графический интерфейс):

```
ERASE  FILE='C:\DOCUME~1\PURGEN~1\LOCALS~1\Temp\spss4024\spssclus.tmp'.
QUICK CLUSTER zlnlong zlntoll zlnequi zlncard zlnwire zmultlin zvoice zpager zinterne zcall
id zcallwai zforward zconfer zebill
  /MISSING=PAIRWISE
  /CRITERIA=CLUSTER(3) MXITER(20) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /PRINT INITIAL ANOVA CLUSTER DISTAN.
```

- Процедура *QUICK CLUSTER* запускается по переменным от ***zlnlong*** до ***zebill***.
- Подкоманда *MISSING* указывает метод *PAIRWISE*, то есть запись учитывается в кластере, если все переменные кластеризации ненулевые.
- Подкоманда *CRITERIA* определяет три формируемых кластера, формируемых за 20 максимальных итераций.
- Подкоманда *PRINT* указывает вывести начальные центры, конечные расстояния между кластера и таблицу ANOVA с описательными F-тестами.
- Остальные не указанные опции равные своим значениям, заданным по умолчанию.

В таблице «Начальные центры кластеров» (***Initial cluster centers***) приведены исходные данные по k хорошо-разделенным переменным. Таблица «История итераций» (***Iteration history***) показывает прогресс кластеризации, выражаемый в движении центров кластеров, на каждом шаге. Из нее видно, что вначале изменения центров кластеров были достаточно существенны (шаги 1-12), а в конце (в данном случае, начиная с 14 итерации) стали незначительными. Если алгоритм останавливается из-за превышения максимального количества допустимых итераций, эту цифру можно попробовать увеличить, поскольку полученное решение может быть неустойчивым. Таблица ANOVA демонстрирует значимость переменных для кластеризации. Переменные с большими F-величинами дают большее разделение между кластерами.

История итераций[a]

Итерация	Изменения центров кластеров		
	1	2	3
1	3,298	3,590	3,491
2	1,016	,427	,931
3	,577	,320	,420
4	,240	,180	,195
5	,119	,125	,108
6	,093	,083	,027
7	,069	,094	,032
8	,059	,051	,018
9	,035	,085	,063
10	,025	,359	,333
11	,068	,439	,287
12	,079	,368	,177
13	,125	,139	,078
14	,077	,096	,020
15	,041	,047	,015
16	,014	,027	,000
17	,019	,038	,000
18	,000	,000	,000

ANOVA

	Кластер		Ошибка		
	Средний квадрат	ст.св.	Средний квадрат	ст.св.	F
Standardized log-long distance	13,063	2	,976	997	13,387
Standardized log-toll free	43,418	2	,820	472	52,932
Standardized log-equipment	99,056	2	,488	383	202,999
Standardized log-calling card	6,301	2	,984	675	6,402
Standardized log-wireless	52,879	2	,646	293	81,873
Standardized multiple lines	38,032	2	,926	997	41,084
Standardized voice mail	236,301	2	,528	997	447,554
Standardized paging	298,992	2	,402	997	743,348
Standardized internet	123,447	2	,754	997	163,642
Standardized caller id	308,104	2	,384	997	802,474
Standardized call waiting	294,674	2	,411	997	717,172
Standardized call forwarding	288,343	2	,424	997	680,718
Standardized 3-way calling	262,397	2	,476	997	551,678
Standardized electronic billing	112,782	2	,776	997	145,381

Конечные центры кластеров являются средним значением по каждой переменной, рассчитанной внутри каждого финального кластера. Финальные центры можно интерпретировать как типичные (средние) значения для записи, принадлежащей каждому кластеру. Они представлены в таблице «Конечные центры кластеров» (***The final cluster centers***).

Конечные центры кластеров

	Кластер		
	1	2	3
Standardized log-long distance	,05	,22	-,16
Standardized log-toll free	,24	,12	-1,05
Standardized log-equipment	,81	-,19	-,69
Standardized log-calling card	,17	,02	-,17
Standardized log-wireless	,42	-,75	-1,00
Standardized multiple lines	,48	-,29	-,05
Standardized voice mail	1,26	-,24	-,44
Standardized paging	1,43	-,38	-,44
Standardized internet	,81	-,59	-,02
Standardized caller id	,82	,71	-,81
Standardized call waiting	,76	,72	-,80
Standardized call forwarding	,78	,69	-,79
Standardized 3-way calling	,74	,67	-,75
Standardized electronic billing	,70	-,63	,05

Расстояния между конечными центрами кластеров

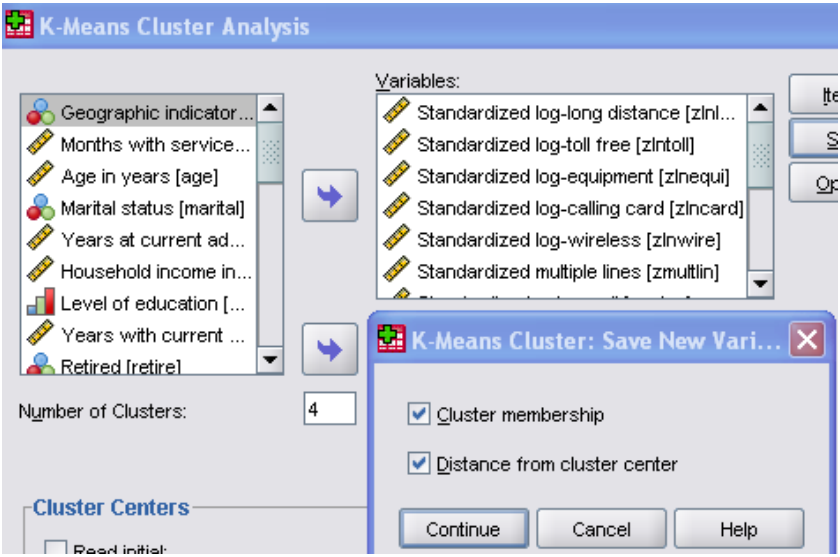
Кластер	1	2	3
1		3,500	4,863
2	3,500		3,396
3	4,863	3,396	

Число наблюдений в каждом кластере

Кластер	1	226,000
	2	292,000
	3	482,000
Валидные		1000,000
Пропущенные значения		,000

Из нее видно, что пользователи первого кластера ☐ тратят много и подписаны на множество сервисов. Пользователи второго

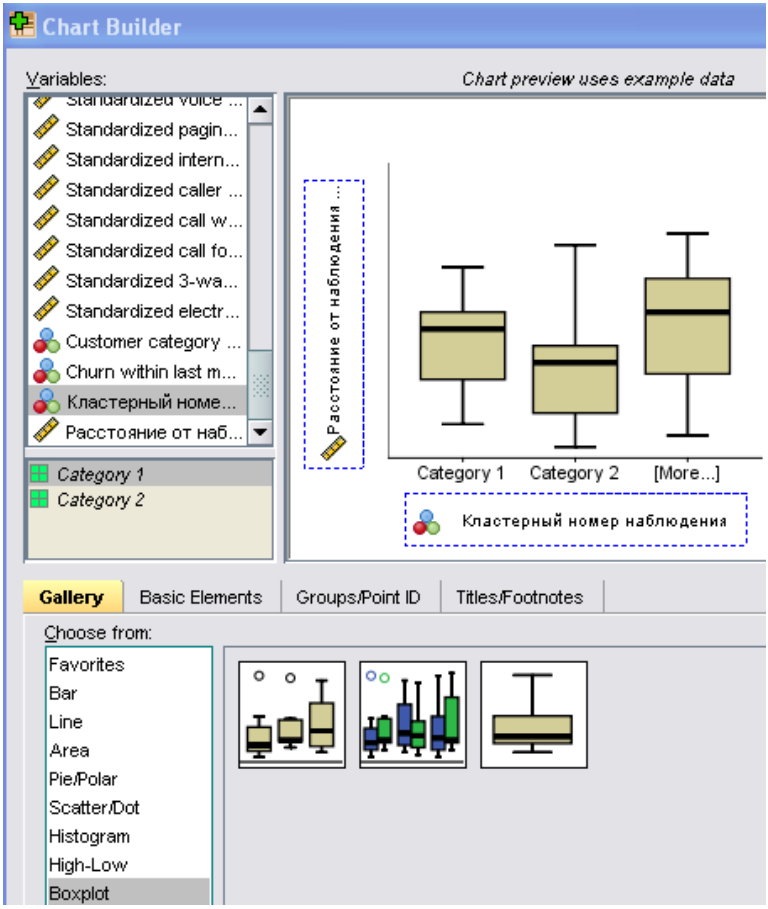
класса являются умеренными потребителями, тратясь в основном на call-услуги  . Потребители третьего класса тратят очень мало и не покупают много сервисов  . Таблица «Расстояния между конечными центрами кластеров» (***Distances between Final Cluster Centers***) показывает Евклидово расстояние между кластерами. Чем больше расстояние, тем больше разница между кластерами. В данном случае наибольшие различия наблюдаются между пользователями первого и третьего кластеров  , а пользователи второго кластера примерно одинаково удалены от пользователей первого и третьего кластеров  . Эти же выводы можно сделать из анализа таблицы финальных центров. Однако применение такого интуитивного подхода становится более затруднительными при увеличении числа кластеров. Наконец, размер кластеров представлен в Таблице «Число наблюдений в каждом кластере» (***Number of Cases in Each Cluster***). Из нее видно, что наименее выгодная группа пользователей третьего кластера, к сожалению, наиболее многочисленна  . Возможно, из этой группы «базовых пользователей» можно выделить еще один, четвертый кластер более выгодных клиентов. Для этого повторяем ККА через ***Analyze->Classify->K-mean Cluster…***, выбирая количество кластеров равным четырем. В опции ***Save*** активируем ***Cluster membership*** и ***Distance from cluster center***.



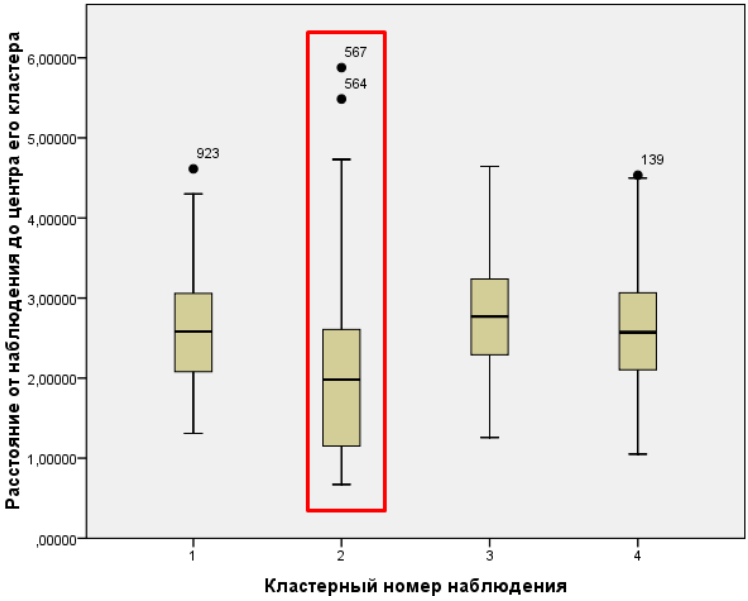
После запуска ККА командный синтаксис в окне вывода:

```
QUICK CLUSTER zlnlong zlntoll zlnequi zlncard zlnwire zmultlin zvoice zpager zinterne zcall
id zcallwai zforward zconfer zebill
  /MISSING=PAIRWISE
  /CRITERIA=CLUSTER(4) MXITER(20) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER DISTANCE
  /PRINT INITIAL ANOVA CLUSTER DISTAN.
```

показывает, что производится поиск четырех-кластерного решения, а расстояние до центра кластера для каждой записи сохраняется в виде переменной в исходном файле (подкоманда *SAVE*). Эти переменные можно визуализировать, например, при помощи ***boxplot***. Выбираем меню ***Graphs→Chart Builder...***, в вкладке типов графиков ***Gallery*** выбираем ***Boxplot*** и перетаскиваем иконку ***Simple Boxplot*** в верхнюю канву. Перетаскиваем переменную ***Расстояние от наблюдения до центра его кластера*** (Distance of Case from its Classification Cluster Center) в качестве оси Y, а переменную ***Кластерный номер наблюдения*** (Cluster Number of Case) – в качестве оси X.



Полученный график позволяет определять выбросы (записи, сильно отличающиеся от других) а также степень вариативности данных внутри кластеров, определяемая по высоте столбика.



В данном случае видно, что наиболее вариативным является второй кластер, однако изменения – в пределах разумного. Из таблицы конечных кластеров следует, что пользователи первых двух кластеров в большинстве своем сформированы из малобюджетного третьего кластера в предыдущем трех-кластерном решении. Однако, при этом представители первого кластера. Однако, пользователи первого класса являются при этом потребителями Internet-сервисов, что выделяет явную, и, возможно, выгодную группу пользователей. Третий и четвертый кластеры соответствуют первому и второму кластерам предыдущего трехкластерного решения. Из таблицы «Расстояния между конечными центрами кластеров» видно, что наиболее сходны между собой представители первого и второго кластеров. Это логично, поскольку они были выделены из одного кластера трехкластерного решения. Наиболее сильны различия между кластерами 2 и 3, являющимися антагонистами по интенсивности потребления сервисов. Кластер 4 по-прежнему равноудален от всех остальных. В новой группе оказывается почти 25% поль-

зователей «Е-сервисов» (таблица «Число наблюдений в каждом кластере») , что весьма существенно для доходов компании.

Конечные центры кластеров

	Кластер			
	1	2	3	4
Standardized log-long distance	,23	-,48	,05	,24
Standardized log-toll free	-,75	-1,10	,26	,11
Standardized log-equipment	-,36	-1,28	,86	-,20
Standardized log-calling card	-,06	-,37	,18	,06
Standardized log-wireless	-,84	-1,54	,45	-,71
Standardized multiple lines	,75	-,74	,45	-,26
Standardized voice mail	-,21	-,59	1,30	-,23
Standardized paging	-,32	-,51	1,50	-,37
Standardized internet	-,57	-,52	,77	-,56
Standardized caller id	-,78	-,73	,86	,72
Standardized call waiting	-,76	-,73	,80	,74
Standardized call forwarding	-,71	-,82	,83	,76
Standardized 3-way calling	-,69	-,79	,84	,72
Standardized electronic billing	,58	-,38	,67	-,63

Расстояния между конечными центрами кластеров

Кластер	1	2	3	4
1		2,568	4,454	3,631
2	2,568		5,746	3,675
3	4,454	5,746		3,515
4	3,631	3,675	3,515	

Число наблюдений в каждом кластере

Кластер	1	236,000
	2	272,000
	3	212,000
	4	280,000
Валидные		1000,000
Пропущенные значения		,000

Таким образом, используя ККА, мы сгруппировали пользователей по трем кластерам. Однако, найденное решение оказалось не вполне удовлетворительным, и был произведен перерасчет ККА с четырьмя кластерами. Результат стал лучше, в дополнительном кластере была обнаружена достаточно большая и потенциально выгодная группа интернет-пользователей, упущенная в первоначальном трех-кластерном анализе.

Следующим шагом для компании является попытка создания модели, классифицирующих пользователей на основе их индивидуальных демографических данных. С помощью этой модели компания может предложить более точные предложения, адресованные определенным группам потребителей. Осуществить подобный анализ можно, к примеру, при помощи дискриминантного анализа.

6.4. Задания для самостоятельной работы.

1. Сравнить результаты решений одних и тех же задач методами факторного анализа (предыдущий раздел) и кластерного анализа. Сформулировать достоинства и недостатки каждого метода.

2. Сформулируйте преимущества и недостатки по отношению друг к другу изученных методик кластерного анализа.

6.5. Вопросы для самоконтроля.

1. Что такое кластерный анализ и для решения каких задач он применяется?
2. Какие два основных вида кластерного анализа применяются на практике?
3. Что такое двушаговый кластерный анализ (ДКА), каковы его полезные свойства, и какова логика схемы применения?
4. Как оценивается важность исходных переменных в ДКА?
5. Каков принцип иерархического кластерного анализа (ИКА) и для каких задач он применяется?
6. Что такое метрика в КА, и какую роль она играет? Какие используемые метрики вы знаете?
7. Что такое дендрограмма, и как она строится?
8. В чем основная суть анализа по K-среднему (K-mean), для какого рода задач он применяется?
9. Как оценивается качество кластерного анализа?

7. Дискриминантный анализ.

Дискриминантный анализ (ДА) используется для моделирования зависимой категориальной переменной, основанного на ее зависимости от одного или более предикторов (***predictors***).

7.1. Теория ДА.

На основании ряда независимых переменных ДА пытается найти линейные комбинации этих переменных, наилучшим образом группирующие исходные данные. Эти комбинации называются ***дискриминантными функциями*** (***ДФ***, ***discriminant functions***), имеющими форму:

$$d_{ik}=b_{0k}+b_{1k}x_{i1}+\ldots+b_{pk}x_{ip} \qquad 7.1$$

где

d_{ik} величина k-той ДФ для i-той записи

p Число предикаторов

b_{jk} величина j-того коэффициента k-той ДФ

x_{ij} величина j-того предикатора i-той записи

Количество ДФ k равно меньшему из чисел: количество категорий (групп) зависимой переменной – 1, число независимых переменных-предикторов.

Процедура ДА автоматически выбирает первую ДФ, разделяющую группы максимально сильно. Затем выбирается вторая ДФ, которая: а) не коррелирует с первой ДФ, б) обеспечивает дальнейшее максимальное разделение групп. Процедура продолжается *k* раз.

Применение ДА накладывает следующие допущения на исходные данные:

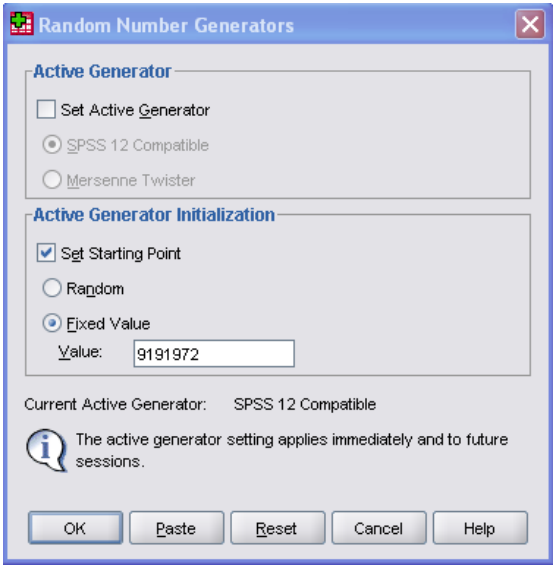
1. Предикторы не сильно коррелированы друг с другом.
2. Среднее значение и вариация предиктора не коррелированы.
3. Корреляции между двумя предикторами постоянны по всем категориям (группам) зависимой переменной.
4. Предикторы распределены по нормальному закону.

7.2. Расчет кредитных рисков.

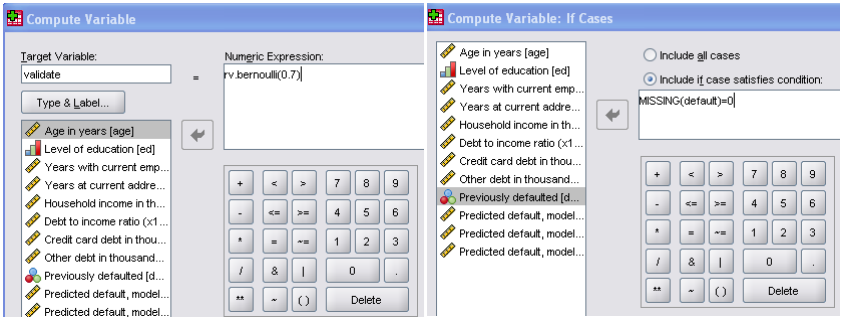
Задача ДА1. Кредитному менеджеру необходимо идентифицировать параметры, по которым можно выявлять пользователей с высоким риском невозврата кредитов. В дальнейшем эти параметры будут использованы для расчетов хороших/плохих кредитных рисков. Предположим, что информация о 850 прошлых и нынешних пользователей доступна в исходном файле данных **bankloan.sav**. Первые 700 записей соответствуют заемщикам, бравшим деньги в прошлом. Построить ДА-модель на основании случайной выборки из этих пользователей, произвести ее проверку на оставшихся данных, а затем использовать для классификации оставшихся 150 будущих заемщиков с точки зрения хороших/плохих рисков.

Используемый файл данных (директория Samples в SPSS): **bankloan.sav**.

Для начала установим фиксированное начальное значение (***random seed***) генератора случайных чисел. Это позволяет повторять одну и ту же случайную выборку при повторных прогонах анализа. Вызываем меню ***Transform->Random numbers…*** и активируем установку исходной точки (***Set Starting Point***) фиксированным значением (***Fixed Value***) величиной ***9191972***.



Теперь создадим переменную-селектор для проверки вызовом меню ***Transform->Compute Variables…***, в текстовом окне ***Target Variable*** печатаем имя переменной ***validate***, а в окне ***Numeric Expression*** вводим формулу (или выбираем из списка ***Random Numbers*** окна ***Function Groups***) ***rv.bernoulli(0.7)***, устанавливающую генерацию распределения Бернулли с параметром вероятности, равным 0.7. Переменная ***validate*** используется только для записей, на основании которых будет создаваться ДА-модель, то есть для бывших клиентов. Оставшиеся в исходном файле 150 пользователей с отсутствующими значениями переменной ***default*** являются текущими клиентами. Для проведения расчетов только на выборке бывших клиентов нажимаем кнопку **if**, выбираем опцию ***Include if case satisfies condition*** и вводим формулу ***MISSING(default) = 0*** в качестве условия. Это гарантирует, что переменная ***validate*** будет рассчитана только для бывших заемщиков, у которых задана переменная ***default***.

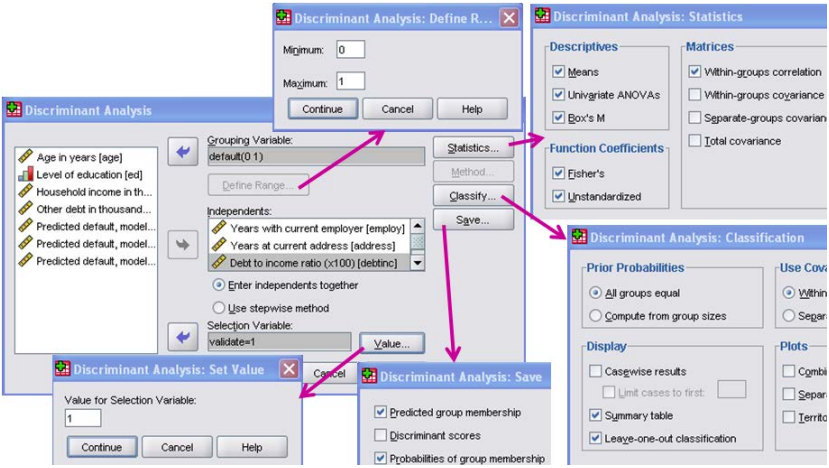


Исходя из заданной формулы, переменная ***validate*** будет иметь значение 1 примерно у 70% записей. Эти пользователи будут использованы для создания модели. Остальные будут использоваться для проверки модели.

Теперь вызываем основной мастер ДА анализа из меню ***Analyze→Classify →Discriminant...*** В качестве группирующей (категориальной) переменной используем предыдущие дефолты (***Previously defaulted***), а в качестве независимых переменных выбираем ***Years with current employer***, ***Years at current address***, ***Debt to income ratio (x100)***, и ***Credit card debt in thousands***. Выбираем ***validate*** в качестве переменной-селектора (***selection variable***). Диапазон группирующей переменной устанавливаем через кнопку ***Define Range…***, задавая 0/1 в качестве минимума/максимума, и значение 1 для переменной-селектора – через кнопку ***Value...***

Следующий этап – установка опций самого ДА. Под кнопкой ***Statistics*** выбираем ***Select Means***, ***Univariate ANOVAs***, и ***Box's M*** в группе ***Descriptives***, ***Fisher's*** и ***Unstandardized*** в группе ***Function Coefficients***, и активируем ***Within-groups correlation*** во вкладке ***Matrices***.

В опции ***Classify*** выбираем ***Summary table*** и ***Leave-one-out classification*** в группе ***Display***. Наконец, под кнопкой ***Save*** выбираем ***Select Predicted group membership*** и ***Probabilities of group membership***.



По нажатию кнопки ***OK*** *SPSS* строит ДА-модель на базе четырех заданных независимых переменных.

Классификация пользователей с высокими и низкими кредитными рисками. Таблица «Результаты классифицирующих функций» показывает результаты присвоения исходных значений по группам.

Коэффициенты классифицирующей функции

	Previously defaulted	
	No	Yes
Years with current employer	,277	,109
Years at current address	,145	,085
Debt to income ratio (x100)	,291	,386
Credit card debt in thousands	-,734	-,303
(Константа)	-3,485	-3,676

Для каждой группы сформирована своя ДФ. Для каждой записи производится расчет классификационного значения (***classification score***) по обеим ДФ. ДА-модель относит запись к той группе, ДФ которой дает большее классификационное значение. Коэффициенты для ***Years with current employer*** и ***Years at current address*** меньше для ДФ ***Yes***, чем для ДФ ***No*** . Это озна-

чает, что клиенты, живущие по одному адресу и работающие в одной компании много лет, допускают дефолты по кредиту с меньше вероятностью. По той же причины вероятность дефолта выше для пользователей, имеющих долги .

Для примера сравним записи 701 и 703. В первом случае клиент, проработал 16 лет у одного работодателя, 13 лет прожил в одном и том же доме, имеет долги в 10,9% дохода, из которых 540$ по кредитной карте. Модель определяет вероятность дефолта по данному кредиту на уровне 8.5%, что можно отнести к хорошему риску . Во втором случае клиент проработал и прожил на одном месте меньшее число лет и имеет худшие долги, и вероятность дефолта по нему предсказывается на уровне 81%, то есть это плохой риск .

716 : preddef1 | 0,3598649771399876

	age	ed	employ	address	debtinc	creddebt	Dis2_1
701	36	1	16	13	10,90	0,54	0,08452
702	50	1	6	27	12,90	1,32	0,26429
703	40	1	9	9	17,00	4,88	0,81455

Оценка коллинеарности предикторов. Корреляции между предикторами показаны в таблице «Объединенные внутригрупповые матрицы» (***The within-groups correlation matrix***). Наибольшие корреляции с другими переменными возникают у предиктора ***Credit card debt in thousands*** .

Объединенные внутригрупповые матрицы

		Years with current employer	Years at current address	Debt to income ratio (x100)	Credit card debt in thousands
Корреляция	Years with current employer	1,000	,286	,104	,508
	Years at current address	,286	1,000	,140	,290
	Debt to income ratio (x100)	,104	,140	1,000	,508
	Credit card debt in thousands	,508	,290	,508	1,000

Однако сейчас трудно оценить, велики ли они настолько, чтобы вызывать опасения. Для уверенного вывода необходимо посмотреть на разницу между структурной матрицей (***structure matrix***) и коэффициентами ДФ (***discriminant function coefficients***), что будет проделано далее.

Корреляции среднего и его вариации. Из таблицы групповых статистик видны потенциально еще более серьезные проблемы – для всех четырех предикторов большее среднее значение переменных соответствует большему отклонению (что противоречит описанному в предыдущем параграфе предположению 2 для ДА) . Особенно это заметно для долговых переменных . Поэтому для более точного анализа следует рассмотреть вопрос о преобразовании исходных переменных с целью устранения данной проблемы.

Групповые статистики

Previously defaulted		Среднее	Стд. отклонение
No	Years with current employer	9,5840	6,67766
	Years at current address	8,8800	6,94239
	Debt to income ratio (x100)	8,8179	5,69545
	Credit card debt in thousands	1,2554	1,41769
Yes	Years with current employer	5,1855	5,72737
	Years at current address	6,3548	6,27836
	Debt to income ratio (x100)	14,4468	7,97554
	Credit card debt in thousands	2,3656	3,36732

Проверка однородности ковариационной матрицы. Логарифм определителя (***log determinants***) ковариационной матрицы – мера изменчивости группы. Чем больше значение, тем большая вариативность наблюдается внутри группы. Большая разница в величинах для разных групп является индикатором различия их ковариационных матриц.

Лог. определители

Previously defaulted	Ранг	Лог. определитель
No	4	11,185
Yes	4	12,253
объединенные внутри групп	4	11,957

Напечатаны ранги и натуральные логарифмы определителей групповых ковариационных матриц.

Результаты теста

М Бокса		252,117
F	Приблизительно	24,893
	ст.св1	10
	ст.св2	245917,239
	Знч.	,000

Проверка нулевой гипотезы о равенстве ковариационных матриц.

Box's M тест проверяет предположение о равенстве ковариаций по группам. Поскольку тест оказывается значимым , необходим проводимый нами в дальнейшем анализ отдельных групповых матриц на предмет радикальных различий в результатах классификации.

Оценка вклада отдельных предикторов. Для оценки индивидуального вклада независимых переменных в модель строятся несколько таблиц.

Тесты эквивалентности групповых средних представлены в таблице «Критерий равенства групповых средних» и показываю т потенциал каждой независимой переменной перед созданием модели. Для каждого предиктора проводится односторонний ***ANOVA***-тест с использованием группирующей переменной в качестве фактора, результаты которого и представлены в таблице . Если уровень значимости более, чем 0.10, переменная, вероятно, не дает вклад в ДА-модель. В данном случае все переменные являются значимыми . Показатель Лямбда-Уилкса (***Wilks' lambda***) – еще одна метрика потенциала . Чем меньше эта величина, тем лучше переменная для дискриминации по группам. В данном случае наилучшей переменной в этом смысле является ***Debt to income ratio (x100)***.

Критерий равенства групповых средних

	Лямбда Уилкса	F	ст.св1	ст.св2	Знч.
Years with current employer	,920	43,262	1	497	,000
Years at current address	,975	12,911	1	497	,000
Debt to income ratio (x100)	,871	73,534	1	497	,000
Credit card debt in thousands	,949	26,597	1	497	,000

Нормированные коэффициенты канонической дискриминантной функции

	Функция
	1
Years with current employer	-,784
Years at current address	-,295
Debt to income ratio (x100)	,437
Credit card debt in thousands	,649

Структурная матрица

	Функция
	1
Debt to income ratio (x100)	,644
Years with current employer	-,494
Credit card debt in thousands	,387
Years at current address	-,270

Нормированные коэффициенты канонической ДФ (***Standardized Canonical Discriminant Function Coefficients***) позволяют сравнивать переменные в разных шкалах. Коэффициенты с большей абсолютной величиной соответствуют переменным с большей дискриминационной силой . В данном случае критерий понижает значение переменной ***Debt to income ratio (x100)***, при этом порядок значимости остальных переменных не изменился.

Структурная матрица показывает корреляцию каждого предиктора с ДФ, упорядоченного по убыванию величины (значимости) . Получившийся порядок аналогичен результату теста групповых средних и отличается от группировки в таблице нормированных коэффициентов ДФ. Наиболее вероятная причина разногласий – коллинеарность между переменными ***Years with current employer*** и ***Credit card debt in thousands***, отмеченная выше в корреляционной матрице. Поскольку коллинеарность не влияет на структурную матрицу, будет справедливым считать, что данная коллинеарность послужила причинной завышения значимости переменных ***Years with current employer*** и ***Credit card debt in thousands*** в таблице нормированных коэффициентов ДФ. Тем са-

мым, переменная ***Debt to income ratio (x100)*** наилучшим образом дискриминирует клиентов по дефолту.

Анализ качества модели. В дополнение к анализу вклада предикторов в модель, процедуры ДА позволяют оценить, насколько хорошо в целом модель описывает исходные данные.

Собственные значения, представленные в одноименной таблице, показывают относительную эффективность каждой ДФ. В случае двух групп наиболее важной метрикой является показатель канонической корреляции (***canonical correlation***) , эквивалентный корреляции Пирсона[9] между коэффициентами дискриминации по группам.

Коэффициент лямбда Уилкса (***Wilks' lambda***) показывает насколько хорошо ДФ разделяет данные по группам . Он равен доле полной вариации в расчетах, не объясненных разницей между группами. Меньшие значения показателя показывают большую разделительную способность ДФ. Сопутствующий -тест проверяет гипотезу, равенства средних по группам, сформированных при помощи ДФ . Малая значимость говорит о том, что ДФ работает лучше, чем случайное разделение на группы.

Собственные значения

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция
1	,357[a]	100,0	100,0	,513

a. В анализе использовались первые 1 канонические дискриминантные функции.

Лямбда Уилкса

Проверка функций(й)	Лямбда Уилкса	Хи-квадрат	ст.св.	Знч.
1	,737	151,007	4	,000

[9] Описывает степень линейной корреляции между двумя параметрами.

Результаты классификации[b,c,d]

				Предсказанная принадлежность к группе		
			Previously defaulted	No	Yes	Итого
Выбранные наблюдения	Исходные	Частота	No	281	94	375
			Yes	30	94	124
		%	No	74,9	25,1	100,0
			Yes	24,2	75,8	100,0
	Кросс-проверенные[a]	Частота	No	278	97	375
			Yes	31	93	124
		%	No	74,1	25,9	100,0
			Yes	25,0	75,0	100,0
Не выбранные наблюдения	Исходные	Частота	No	106	36	142
			Yes	10	49	59
			Несгруппированные наблюдения	95	55	150
		%	No	74,6	25,4	100,0
			Yes	16,9	83,1	100,0
			Несгруппированные наблюдения	63,3	36,7	100,0

a. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.

b. 75,2% выбранных исходных сгруппированных наблюдений классифицировано правильно.

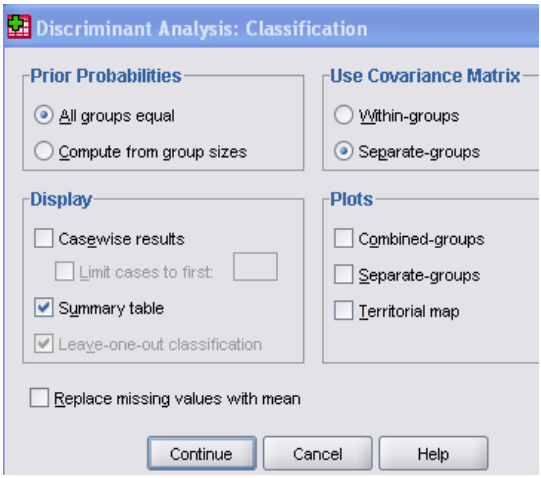
c. 77,1% невыбранных исходных сгруппированных наблюдений классифицировано правильно

d. 74,3% выбранных кросс-проверяемых сгруппированных наблюдений классифицировано правильно.

Проверка модели. Таблица «Результаты классификации» демонстрирует практические результаты работы ДА-модели. Правильно были предсказаны 281 и 375 заемщиков, не допустивших дефолта, и 94 из 124 заемщиков, оказавшихся банкротами ., соответственно, доля правильных предсказаний составила 75.2% . Классификация на основе всех записей является слишком оптимистичной в смысле возможного завышения доли правильных предсказаний. Для коррекции этого используется кросс-проверка, состоящая в исключении из расчетов того случая, который прове-

ряется . Однако, и данный метод может давать (и обычно дает) завышенный результат по сравнению с проверкой по под-выборке (***subset validation***), осуществляемой по тем пользователям, которые не использовались при расчете модели. Эти результаты показаны в секции «Не выбранные наблюдения (***Cases Not Selected***) таблицы . Согласно ему 77.1% случаев корректно классифицируются моделью . Что в целом означает, что полученная модель корректно предсказывает около трех случаев из четырех. Наконец, 150 не вошедших в группу – это текущие пользователи, и полученные результаты дают предсказания построенной модели в отношении этих заемщиков .

Поскольку ***Box's M*** тест оказался значительным, стоит прогнать ДА еще раз, используя ковариационные матрицы отдельных групп, выбираемые в опции. Для этого запускаем еще раз мастер ДА через ***Analyze→Classify →Discriminant...*** и в опции ***Classify*** выбираем ***Separate-groups*** в группе ***Use Covariance Matrix***. Отметим, что в этом режиме недоступен расчет ***Leave-one-out classification*** в группе ***Display***.



Результаты перерасчета мало отличаются от первого результата. Это, скорей всего, означает, что не стоит использовать групповые ковариационные матрицы. ***Box's M*** тест может быть избыточ-

но чувствительным в случае большого объема исходных данных, что имеет место в нашем случае.

Результаты классификации[a,b]

				Предсказанная принадлежность к группе		
			Previously defaulted	No	Yes	Итого
Выбранные наблюдения	Исходные	Частота	No	287	88	375
			Yes	31	93	124
		%	No	76,5	23,5	100,0
			Yes	25,0	75,0	100,0
Не выбранные наблюдения	Исходные	Частота	No	107	35	142
			Yes	10	49	59
			Несгруппированные наблюдения	96	54	150
		%	No	75,4	24,6	100,0
			Yes	16,9	83,1	100,0
			Несгруппированные наблюдения	64,0	36,0	100,0

a. 76,2% выбранных исходных сгруппированных наблюдений классифицировано правильно.

b. 77,6% невыбранных исходных сгруппированных наблюдений классифицировано правильно

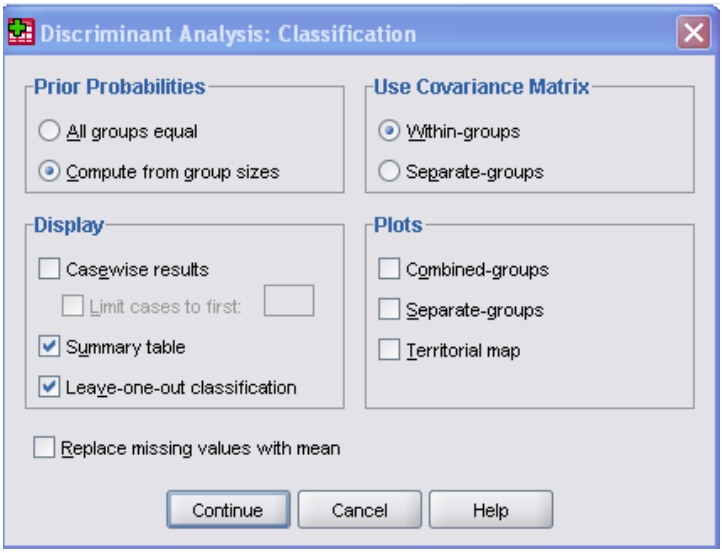
Настройка первичных (априорных) вероятностей. Результаты модели зависят от начального предположения о распределении по категориям исследуемой переменной. В данном случае, доли предполагаемых будущих банкротов из общего числа заемщиков. Используемые априорные оценки показаны в таблице «Априорные вероятности для групп». Априорная вероятность, это вероятность принадлежности записи к определенной группе (категории) при отсутствии дополнительной информации. Пока не предположено другого, по умолчанию подразумеваются равные вероятности причисления пользователя к группе исследуемой переменной. В данном случае при построении ДФ предполагается, что клиент с вероятностью фифти-фифти может быть банкротом

. При расчетах делается поправка (взвешивание) с учетом размера групп .

Априорные вероятности для групп

Previously defaulted	Априорные	Наблюдения, использованные в анализе	
		Невзвешенные	Взвешенные
No	,500	375	375,000
Yes	,500	124	124,000
Итого	1,000	499	499,000

Для получения классификации с использованием неравномерных априорных предположений о принадлежности клиентов к группе исследуемых переменных необходимо прогнать ДА еще раз через ***Analyze→Classify →Discriminant...*** и в опции ***Classify*** выбрать ***Compute from group sizes*** в группе ***Prior probabilities***, вернув опцию ***Within-groups*** в группе ***Use Covariance Matrix***.



Априорные вероятности теперь базируются на статистике бывших клиентов. Не допустивших дефолт в выборке 75.2%, по-

этому при расчетах ДФ более тяготеет в сторону признания клиента хорошим заемщиком, который не допустит дефолта.

Априорные вероятности для групп

Previously defaulted	Априорные	Наблюдения, использованные в анализе	
		Невзвешенные	Взвешенные
No	,752	375	375,000
Yes	,248	124	124,000
Итого	1,000	499	499,000

После перерасчета, вероятность правильного предсказания в ДА-модели поднялась до 80% . К сожалению, платой за это стало существенное (более чем в два раза) увеличение доли будущих банкротов, идентифицированных, как благонадежные заемщики .

Результаты классификации[b,c,d]

			Previously defaulted	Предсказанная принадлежность к группе: No	Предсказанная принадлежность к группе: Yes	Итого
Выбранные наблюдения	Исходные	Частота	No	356	19	375
			Yes	75	49	124
		%	No	94,9	5,1	100,0
			Yes	60,5	39,5	100,0
	Кросс-проверенные[a]	Частота	No	355	20	375
			Yes	77	47	124
		%	No	94,7	5,3	100,0
			Yes	62,1	37,9	100,0
Не выбранные наблюдения	Исходные	Частота	No	137	5	142
			Yes	31	28	59
			Несгруппированные наблюдения	130	20	150
		%	No	96,5	3,5	100,0
			Yes	52,5	47,5	100,0
			Несгруппированные наблюдения	86,7	13,3	100,0

a. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.

b. 81,2% выбранных исходных сгруппированных наблюдений классифицировано правильно.

c. 82,1% невыбранных исходных сгруппированных наблюдений классифицировано правильно

d. 80,6% выбранных кросс-проверяемых сгруппированных наблюдений классифицировано правильно.

Выбор той или иной стратегии зависит от политики банка. В консервативном варианте вашей целью является лучшая идентификация будущих банкротов, в этом случае лучше использовать равные априорные предположения. В случае агрессивной тактики банка в отношении займов можно прибегнуть к неэквивалентным априорным предположениям.

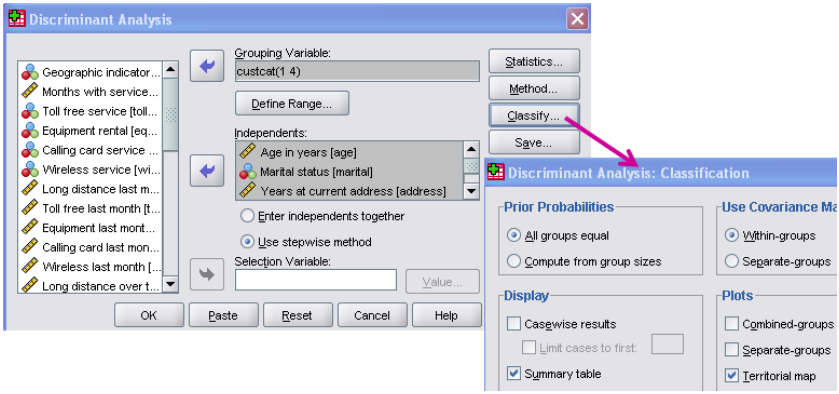
Резюме. На основе ДА-анализа была создана модель классификации пользователе по степени кредитного риска. Обнаруженные возможные проблемы с ковариационной матрицей (***Box's M*** тест) объясняются большим объемом данных. Использование неравных априорных предположений дает выигрыш в увеличении доли правильных предсказаний по всем данным за счет большего пропуска будущих банкротов.

7.3. Классификация пользователей провайдера.

Задача ДА2. Фирма-провайдер разделила своих пользователей на четыре группы на основании паттернов использования сервисов[10]. Если можно было бы априори определять принадлежность пользователя к той или иной группе, исходя из его персональных демографических данных, фирма могла бы более точно формировать пакеты услуг, нацеливая их на определенные группы будущих пользователей. Постройте такую модель классификации пользователей, используя методы ДА.

Используемый файл данных (директория Samples в SPSS): **telco.sav**.

Открываем файл данных и вызываем меню ***Analyze→Classify →Discriminant...*** Нажимаем кнопку ***Reset*** для сброса всех параметров в состояние по умолчанию. В качестве группирующей (категориальной) переменной используем переменную ***Customer category*** (***custcut***), а в качестве независимых переменных выбираем переменные от ***Age in Years*** до ***Number of people in household***. Выбираем пошаговый метод решения, активируя опцию ***Use stepwise method***. Диапазон группирующей переменной устанавливаем через кнопку ***Define Range…***, задавая 1/4 в качестве минимума/максимума. В опции ***Classify*** выбираем ***Summary table*** в группе ***Display*** и ***Territorial map*** в группе ***Plots***.



[10] Это было осуществлено методами кластерного анализа по К-среднему, рассмотренного нами ранее.

При нажатии ***OK*** запускается ДА с пошаговой методикой выбора переменных. При большом количестве исходных предикторов пошаговый метод оказывается полезным для отбора лучших переменных для использования в модели.

Таблица «Переменные, не включенные в анализ» (***Variables Not in the Analysis***) показывает статистику пошагового ДА. На нулевом шаге ни одна переменная не входит в модель. На очередном шаге в модель включается предиктор, имеющий наибольшее значение величины ***F включения*** (***F to Enter***), при этом оно превышающее определенный порог (3.84 по умолчанию). В данном случае это переменная ***Level of education***. После включения переменной в модель он исключается из не используемых предикторов, производится перерасчет показателей и на следующем первом шаге в модель включается следующий предиктор ***Years with current employer***.

Процедура повторяется до тех пор, пока в списке не останутся предикторы, включение которых в модель нецелесообразно, поскольку не значимо.

Шаг		F включения
0	Age in years	7,521
	Marital status	3,500
	Years at current address	8,433
	Household income in thousands	6,689
	Level of education	61,454
	Years with current employer	16,976
	Retired	3,005
	Gender	,373
	Number of people in household	3,976

Шаг		F включения
1	Age in years	6,125
	Marital status	3,803
	Years at current address	8,487
	Household income in thousands	6,022
	Years with current employer	14,933
	Retired	1,432
	Gender	,358
	Number of people in household	3,967

Шаг		F включения
2	Age in years	,252
	Marital status	1,507
	Years at current address	3,514
	Household income in thousands	,687
	Retired	,353
	Gender	,395

Таблица «Переменные в анализе» показывает статистику ряда параметров включения предикторов в модель. ***Толерантность*** – пропорция вариации переменной, не учитываемой любой другой независимой переменной, включенной в модель. Переменная с низкой толерантностью вносит мало информации в модель и может вызвать проблемы при расчетах. Параметр ***F исключения*** (***F to Remove***) полезен для оценки того, что произойдет при исклю-

чении переменной из модели . Аналогичен параметру ***F включения*** (***F to Enter***) из предыдущей таблицы.

Переменные в анализе

Шаг		Толерантность	F исключения	Лямбда Уилкса
1	Level of education	1,000	61,454	
2	Level of education	,953	59,108	,951
	Years with current employer	,953	14,933	,844
3	Level of education	,951	60,046	,940
	Years with current employer	,934	15,824	,834
	Number of people in household	,979	4,841	,807

Предостережение! Пошаговый ДА – удобный метод, но имеет свои ограничения. Следует всегда иметь в виду, что данная методика отбора основана исключительно на статистической значимости, поэтому в ходе ДА в качестве значимых переменных могут быть выбраны параметры, не имеющие практической значимости. Поэтому, при наличии опыта и ожиданий в отношении значимости предикторов, необходимо учитывать эти знания и делать поправки в отношении выводов пошагового ДА. Однако, если же у вас много предикторов и никаких идей в отношении их значимости, тогда запуск пошагового ДА и отладка полученной модели лучше, чем полное отсутствие модели.

Согласно данным таблицы «Собственные значения», практически вся вариация (99,6%), описанная моделью, осуществляется за счет двух ДФ . Хотя автоматически фитируются три функции, из-за мизерного собственного значения можно спокойно игнорировать третью ДФ. Тест лямбда Уилкса (***Wilks' lambda***) подтверждает значимость только первых двух функций. Для каждого набора функций этот тест проверяет гипотезу равенства средних ДФ по группам исследуемой переменной. Тест для третьей функции дает значимость теста, большую 10% (0.10) , что означает, что функция практически не дает вклада в модель, поскольку привносимые ей различия по группам не значимы.

Собственные значения

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция
1	,198[a]	80,2	80,2	,407
2	,048[a]	19,4	99,6	,214
3	,001[a]	,4	100,0	,031

a. В анализе использовались первые 3 канонические дискриминантные функции.

Лямбда Уилкса

Проверка функции(й)	Лямбда Уилкса	Хи-квадрат	ст.св.	Знч.
от 1 до 3	,796	227,345	9	,000
от 2 до 3	,953	47,486	4	,000
3	,999	,929	1	,335

В таблице «Структурная матрица» показана взаимосвязь между предикторами и ДФ.

Структурная матрица

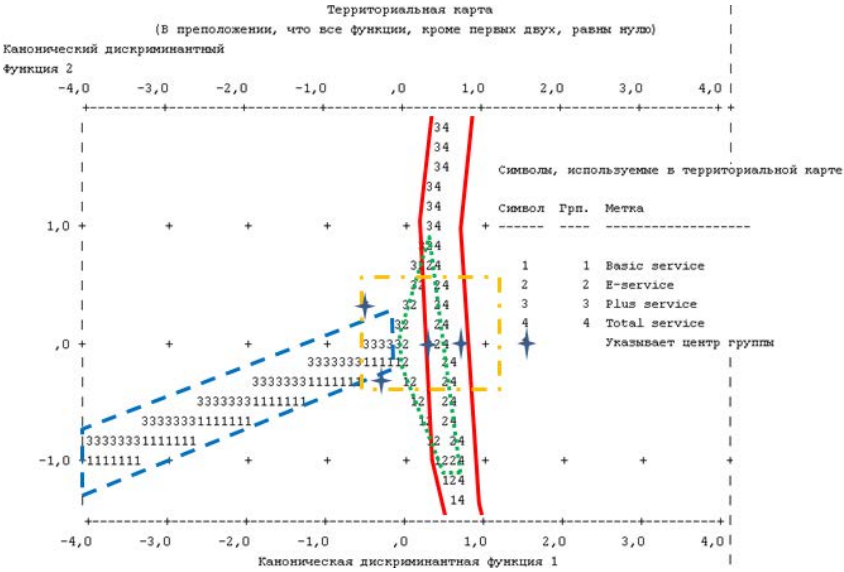
	Функция		
	1	2	3
Level of education	,966*	-,090	-,244
Years with current employer	-,182	,964*	-,193
Age in years[a]	-,162	,598*	-,285
Household income in thousands[a]	,109	,514*	-,190
Years at current address[a]	-,151	,394*	-,214
Retired[a]	-,108	,230*	-,137
Gender[a]	,008	,054*	,009
Number of people in household	,232	,097	,968*
Marital status[a]	,132	,134	,600*

Объединенные внутригрупповые корреляции между дискриминантными переменными и нормированными каноническими дискриминантными функциями.
Переменные упорядочены по абсолютной величине корреляций внутри функции.

*. Максимальная по абсолютной величине корреляция между переменными и дискриминантными функциями.

a. Эта переменная не используется в анализе.

При количестве ДФ>1 звездочкой (*) отмечаются наиболее коррелирующие по абсолютной величине с каждой ДФ переменные. При наличии таких переменных для одной ДФ, они сортируются в порядке убывания корреляции. В данном случае первая ДФ коррелирует с одним единственным предиктором Level of education на уровне 96.7% . Фактически между первой ДФ и этой переменной можно поставить знак равенства. Вторая ДФ связана с предикторами ***Years with current employer***, ***Age in years***, ***Household income in thousands***, ***Years at current address***, ***Retired***, и ***Gender***, причем два последних коррелированы в меньшей степени. На основе совокупности наиболее коррелируемых предикторов вторую ДФ можно назвать «функцией стабильности». Третья ДФ практически полностью связана с переменной ***Number of people in household*** и совсем чуть-чуть с ***Marital status*** . Однако выше мы уже выяснили бесполезность данной функции, и, тем самым, предикторов, ее определяющих.



Двумерный график «Территориальная карта» помогает исследовать отношения между группами и ДФ, визуализируя взаимосвязь между группами и предикторами. Первая функция, отложенная по горизонтальной оси, отделяет четвертую группу ***Total***

service customers от всех остальных . Поскольку основу этой ДФ составляет переменная уровня образования ***Level of education***, это означает, что пользователи большого количества сервисов наиболее образованны. Вторая функция разделяет группы 1 (***Basic service***) и 3 (***Plus service***). Последние, как правило, постарше и дольше работают на текущем месте работы по сравнению с базовыми пользователями. Интернет-пользователи (***E-service***) не сильно выделяются, но по карте можно предположить, что они более образованы и имеют умеренный опыт работы . В целом, близость центров групп, обозначенных значками в области , демонстрирует, что разделение между группами не очень сильное.

Изображены только первые две функции, однако, поскольку ранее выяснилось, что третья функция малозначима, данной карты достаточно для полного охвата текущей дискриминантной модели.

Результаты классификации. Тест лямбда Уилкса выше показал, что модель работает лучше, чем случайное предположение, однако необходимо определить, насколько лучше.

Эти оценки можно сделать на основании таблицы «Результаты классификации». При отсутствии модели – «нулевой гипотезе» мы априори записывали бы всех пользователей в группу ***Plus service***, как самую многочисленную из четырех . При этом точность такого предсказания 281/1000 (общее число пользователей по всем группам) = 28.1%. Построенная модель дает средний результат на 11.4% лучше, или 39.5% . При этом наилучшие предсказания дает идентификация пользователей ***Total service*** – более 60% . К сожалению, точность предсказаний ***E-service*** пользователей очень низкая – всего 6.9% , в такой ситуации, возможно, требуется поиск нового, более значимого предиктора для выделение данных пользователей, если таковые доступны из имеющейся статистики.

Результаты классификации[a]

		Customer category	Предсказанная принадлежность к группе				Итого
			Basic service	E-service	Plus service	Total service	
Исходные	Частота	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47,0	4,1	22,9	25,9	100,0
		E-service	22,6	6,9	26,7	43,8	100,0
		Plus service	36,3	5,0	39,9	18,9	100,0
		Total service	16,9	6,8	15,7	60,6	100,0

a. 39,5% исходных сгруппированных наблюдений классифицировано правильно.

Резюме. Была создана ДА-модель, классифицирующая пользователей, исходя из их персональных данных, по предопределенным группам по уровню подписанных сервисов. Используя структурную матрицу и территориальную карту, удалось определить переменные, дающие наилучшую сегментацию групп. Из классификационных результатов выяснилось низкое качество предсказаний пользователей ***E-service***. Требуются дополнительные исследования с целью поиска предикторов, которые позволили бы выделить данную категорию пользователей. Несмотря на это, модель может быть адекватной для предсказания принадлежности пользователей другим группам. Например, если основной доход фирма-провайдер получает по категориям ***Plus service*** и ***Total service***, модель дает требуемую информацию.

В целом ДА применяется для классификации (моделирования) категорий зависимой переменной от одного или более шкалируемых предикторов. Если зависимой является шкалируемая переменная, можно попробовать использовать регрессионный анализ (линейная регрессия, GLM-унивариатный подход).

В случае мультиколлинеарности переменных и/или необходимости уменьшения их числа необходимо использовать факторный анализ, рассмотренный нами ранее.

7.4. Вопросы для самоконтроля.

1. Что такое дискриминационный анализ и для решения каких задач он применяется?
2. Что такое дискриминационные функции?
3. В чем похожесть дискриминационного и кластерного анализов?
4. Каковы требования ДА к качеству исходных переменных?

5. Чем опасны корреляции исходных переменных для ДА, как их определить, и как от них избавиться?

6. Как оценить значимость дискриминационных функций?

7. Как определить качество ДА (коэффициент лямбда Уилкса, проверка моделей)?

8. Для чего и как применяется пошаговый ДА, каковы его преимущества и недостатки?

9. Для чего применяется и как строится график «Дорожная карта»?

Заключение

Первая часть пособия (главы 1–3) посвящена основным элементам работы в пакете SPSS, освоение которых соответствует базовому уровню подготовки. Как показывает практика, чем проще статистическая методика, тем чаще она применяется для решения задач, возникающих в повседневной жизни. Несмотря на свою простоту, описанные методики (описательные статистики, перекрестные таблицы, т-тесты и ANOVA анализ) охватывают широкий класс задач, который предстоит решать в своей профессиональной деятельности специалистам экономико-социальных направлений.

Вторая часть пособия (главы 4-7) посвящена классифицирующему анализу в пакете SPSS. Изложенные методики регрессионного, факторного, кластерного, дискриминационного анализа лежат в основе решение подавляющего числа прикладных и научных задач в самых разных областях, включая маркетинг и рекламу, поэтому их знание необходимо в профессиональной деятельности специалистов социально-экономических направлений деятельности самого широкого спектра.

Автор надеется, что освоение настоящего пособия стимулирует стремление читателя совершенствовать свои навыки работы в SPSS. Он может это сделать самостоятельно с помощью документации, учебников и руководств по SPSS, в числе которых будут и следующие части настоящего пособия.

Список литературы

1. Бессокирная, Г. П. Анализ социологических данных с помощью SPSS: обзор учебной литературы / Г. П. Бессокирная // Социология 4М. – 2008. – №26. – С. 168–175.

2. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цефель. – Санкт-Петербург: ДиасофтЮП, 2005. – 608 с.

3. Крыштановский, А. О. Анализ социологических данных с помощью пакета SPSS: учеб. пособие для вузов / А. О. Крыштановский. – Москва: Изд. дом ГУ ВШЭ, 2006. – 281 с.

4. Малхорта, Нэреш К. Маркетинговые исследования. Практическое руководство / Н. К. Малхорта; пер. с англ. – 3-е изд. – Москва: Вильямс, 2002. – 960 с.

5. Моосмюллер, Г. Маркетинговые исследования с SPSS: учебное пособие / Г. Моосмюллер, Н. Ребик. – 2-е изд. – Москва: Ozon, 2011. – 200 с.

6. Наследов, А. Компьютерный анализ данных в психологии и социальных науках / А. Наследов. – Санкт-Петербург: Питер, 2005. – 418 с.

7. Пациорковский, В. В. SPSS для социологов: учебное пособие / В. В. Пациорковский, В. В. Пациорковская. – Москва: ИСЭПН РАН, 2005. – 433 с.

8. Таганов Д. Статистический анализ в маркетинговых исследованиях / Д. Таганов. – Санкт-Петербург: Питер, 2005. – 192 с.

9. Field A. Discovering Statistics Using SPSS. 2-nd edition. SAGE publication Ltd. London, 2005. 780 p.

10. Janssens W., De Pelsmacker P., Wijnen K., Van Kenhove P. Marketing Research with SPSS. Pearson Education, 2008, UK. 441 p.

11. Joaquim P. Markes de Sá Applied statistics using SPSS, MATHLAB and R. 2-nd edition. Springer-Verlag, Berlin-Heidelberg, 2007, 520 p.

Онлайн-ресурсы

1. Иллюстрированный самоучитель по SPSS. – URL: http://www.datuapstrade.lv/rus/spss/ (дата обращения: 18.06.2025).

2. Хомич А. Программа SPSS (версий 10–11) / А. Хомич. – URL: http://khomich.narod.ru/metodichka/SPSS/IV.htm (дата обращения: 18.06.2025).

3. Анализ социологических данных с применением статистического пакета SPSS / Новосибирский государственный университет. – URL http://www.webstarstudio.com/marketing/theor.htm (дата обращения: 18.06.2025).

4. IBM SPSS Statistics V27: краткое руководство. – URL: https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/ru/IBM_SPSS_Statistics_Brief_Guide.pdf (дата обращения: 18.06.2025).

5. Курсы по теме «IBM SPSS Statistics». – URL: https://www.udemy.com/ru/topic/spss (дата обращения: 18.06.2025).

6. Литература по SPSS. – URL: https://www.statmethods.ru/stati/literatura-po-spss/ (дата обращения: 18.06.2025).

7. Raynald's SPSS Tools: коллекция синтаксиса, макросов, скриптов и советов, созданная в помощь решению задач управления данными и анализа данных в IBM SPSS Statistics. – URL: https://www.spsstools.net/ru/ (дата обращения: 18.06.2025).

8. Основы работы в SPSS (для психологов). – URL: https://handbook.mathpsy.com/?page_id=988 (дата обращения: 18.06.2025).

9. Инструментарий аналитика: SPSS: подборка ресурсов от Филиппа Управителева. – URL: https://thinkcognitive.org/ru/blog/instrumentarij-analiika-spss (дата обращения: 18.06.2025).

Научное электронное издание

Шитова Юлия Юрьевна

АНАЛИЗ МАРКЕТИНГОВОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ ПРОГРАММЫ SPSS

Учебное пособие

Чебоксары, 2025 г.

Компьютерная верстка *Е. В. Кузнецова*

Подписано к использованию 29.07.2025 г.
Объем 3,87 Мб. Тираж 20 экз.
Уч. изд. л. 4.46.

Издательский дом «Среда»
428023, Чебоксары, Гражданская, 75, офис 12
+7 (8352) 655-731
info@phsreda.com
https://phsreda.com