# ОПИСАТЕЛЬНАЯ СТАТИСТИКА ДЛЯ КАТЕГОРИАЛЬНЫХ ШКАЛ
# ОПИСАТЕЛЬНАЯ СТАТИСТИКА ДЛЯ КОЛИЧЕСТВЕННЫХ ШКАЛ

# ВОПРОСЫ

1. Выберите верное утверждение:
   - ❑ Объем генеральной совокупности меньше объема выборки
   - ❑ Объем генеральной совокупности больше объема выборки
   - ❑ Объем генеральной совокупности равен объему выборки
   - ❑ Объем генеральной совокупности и объем выборки не влияют на ошибку выборки

2. Ч…ого опроса?
   - Поиск ошибок и противоречий в данных
   - Поиск недопустимых значений
   - Корректировка недопустимых значений
   - Вычисление новых переменных
   - Категоризация количественных переменных
   - Автоматическая перекодировка
   - Сортировка данных
   - Отбор наблюдений
   - Вывод значений переменных
   - Визуальная категоризация
   - Установка вывода значений и меток значений

…рафическим признакам (к

3. …

4. …
   - ❑ Затраты на рекламную кампанию
   - ❑ Район проживания
   - ❑ Время покупки
   - ❑ Оценка товара потребителем

5. Какой признак измеряется в количественной шкале?

- ❑ Цена товара
- ❑ Телефонный номер
- ❑ Должность сотрудника
- ❑ Сорт продукции

6. Какая шкала позволяет применять максимальное количество математико-статистических методов?

- ❑ Номинальная шкала
- ❑ Порядковая шкала
- ❑ Интервальная шкала
- ❑ Шкала отношений

7. Значения в какой шкале обладают только отношением равенства/неравенства?

- ❑ Шкала наименований
- ❑ Порядковая шкала
- ❑ Интервальная шкала
- ❑ Шкала отношений

8. Что такое репрезентативность?

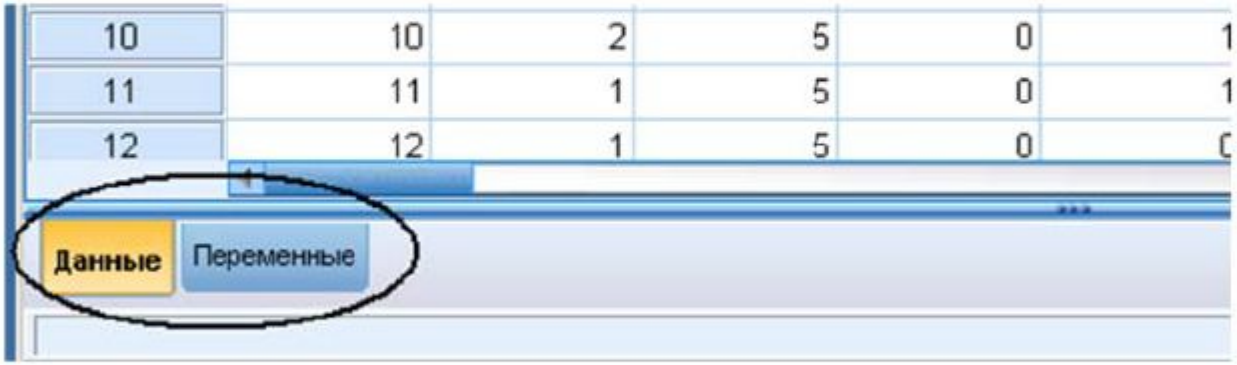
- ❑ Ошибка выборки
- ❑ Объем выборки
- ❑ Свойство выборки отражать генеральную совокупность
- ❑ Свойство генеральной совокупности отражать выборку

# Окно Редактора данных

1. Окно **Редактора данных** имеет два листа **Данные** и **Переменные**, расположенные в левом нижнем углу экрана.
2. Лист **Данные** позволяет просматривать значения данных и вносить в них изменения, переходя в нужную ячейку при помощи курсора мыши или клавиш со стрелками.
3. Строки на листе **Данные** соответствуют объектам наблюдения, а столбцы – переменным (характеристикам объектов наблюдения).
4. Лист **Переменные** позволяет изменять свойства переменных (столбцов) файла данных.
5. Строки на листе **Переменные** соответствуют переменным, а столбцы – свойствам переменных.
6. Переключиться между листами можно одним из способов:
   - нажать по ярлычку листа левой кнопкой мыши;
   - нажать комбинацию клавиш **Ctrl+T**.

# Окно Редактора данных

- В окне Редактора данных два листа:
  - ✓ Данные – для изменения значений данных
  - ✓ Переменные – для изменения свойств переменных



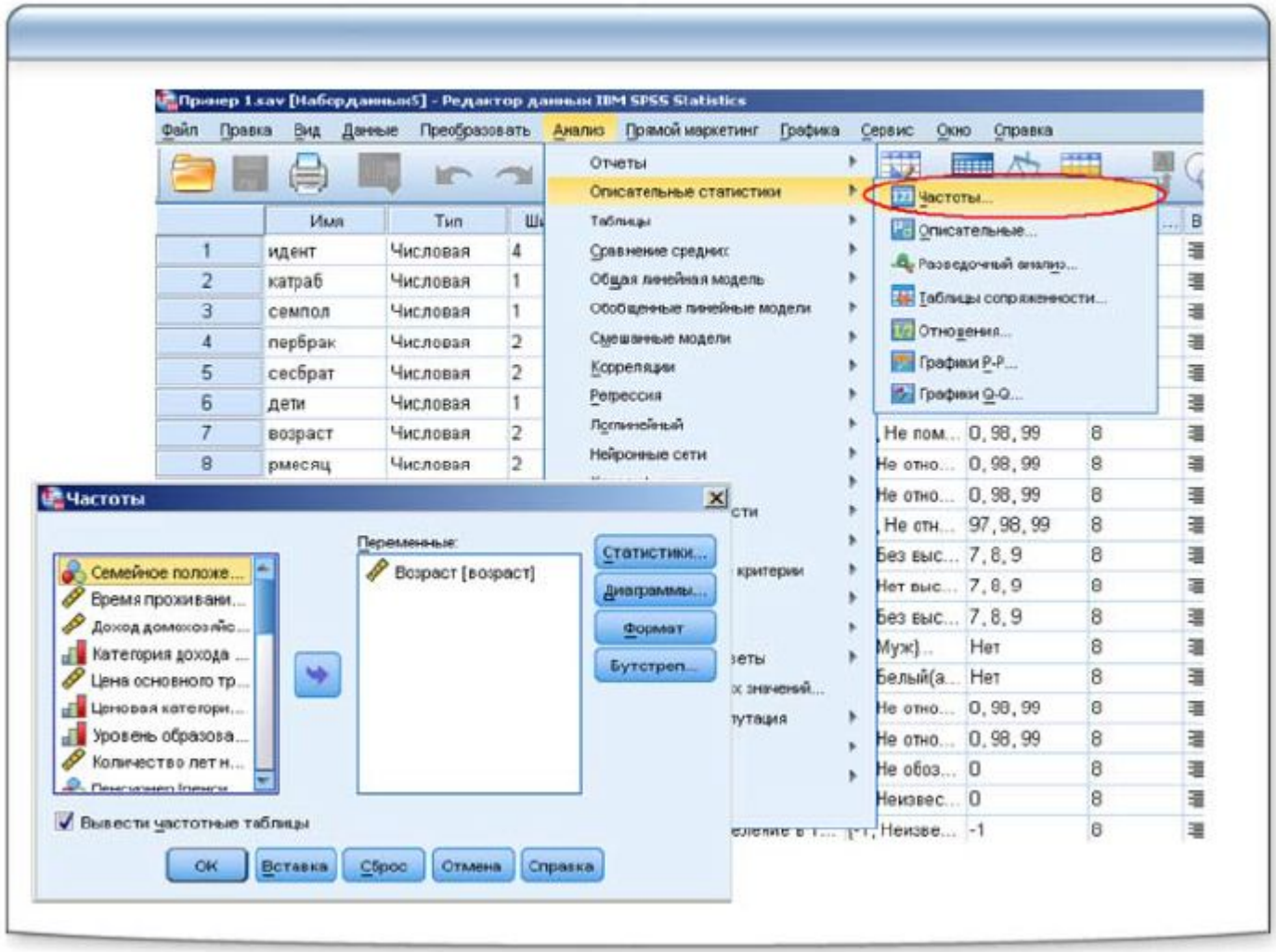
## Запуск процедур анализа

1. Все операции анализа данных осуществляются с помощью процедур в меню **Анализ**.
2. Пункт меню **Анализ** включает в себя список доступных в SPSS групп процедур анализа.
3. За названиями большинства групп следует стрелка, указывающая на наличие нескольких процедур анализа в данной группе.

*Доступный набор статистических методов зависит от того, какие модули программы SPSS были установлены. В демонстрационной триал-версии доступны все модули.*

4. Диалоговые окна статистических процедур содержат следующие компоненты:
   - список исходных данных – список всех переменных в файле данных;
   - список выбранных переменных – список, содержащий переменные файлы данных, которые отбираются для анализа;
   - командные строки – кнопки, при нажатии на которые выполняются определенные действия.

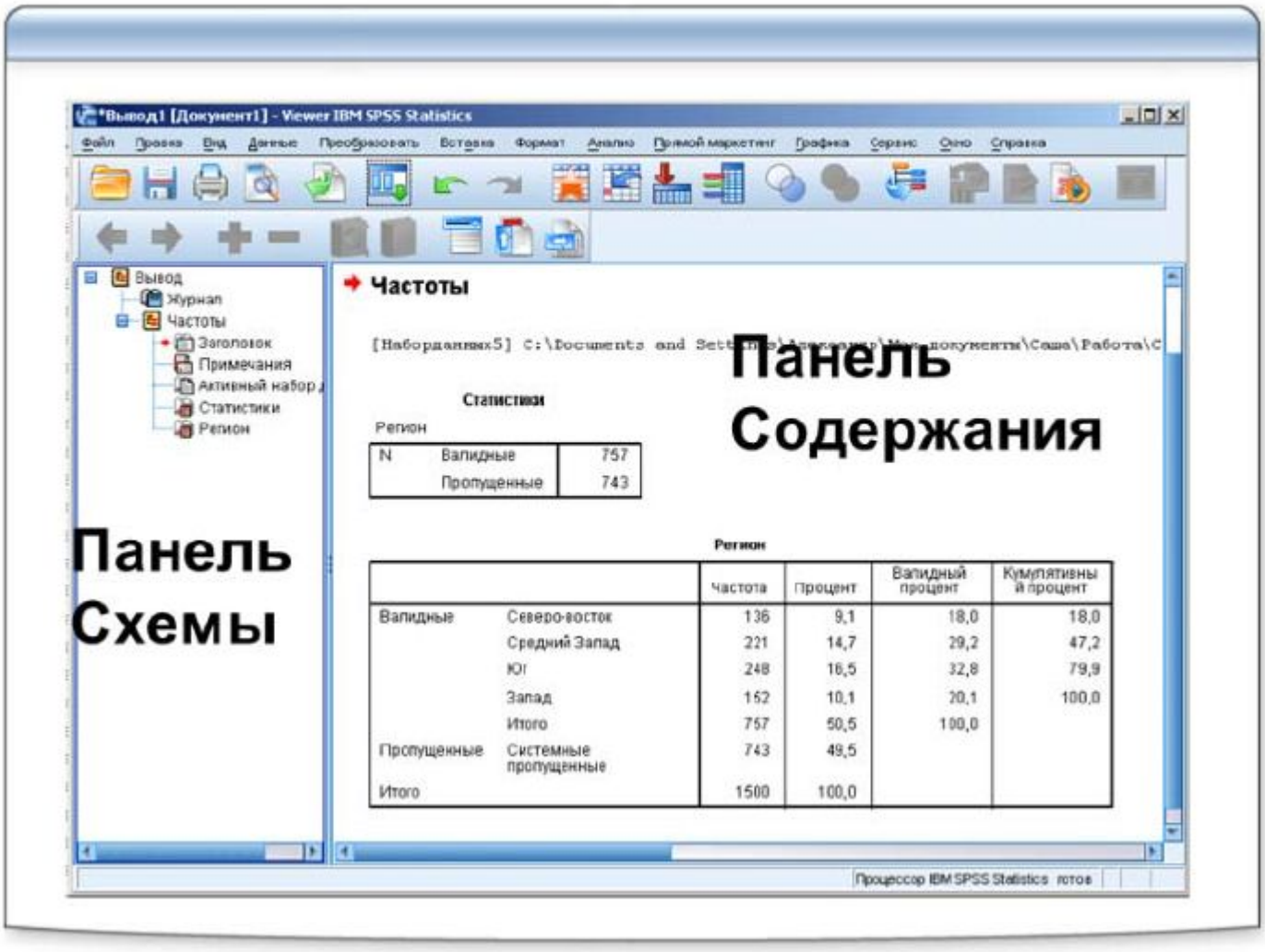
# Запуск процедур анализа



## Окно Вывода

1. Результаты статистического анализа (таблицы и диаграммы) представляются в окне **Вывода**.
2. Окно **Вывода** состоит из двух панелей:
   - панель **Схемы** (левая панель);
   - панель **Содержания** (правая панель).
3. Панель **Схемы** содержит результаты анализа в схематическом виде (пиктограммы).
4. Панель **Содержания** содержит сами статистические таблицы (диаграммы).
5. Перемещаться по панели **Содержания** можно одним из способов:
   - используя полосы вертикальной и горизонтальной прокрутки;
   - нажимая по элементу на панели **Схемы**, чтобы переходить к таблице или графику на панели **Содержания**.

# Окно Вывода


*Вывод1 [Документ1] - Viewer IBM SPSS Statistics
Файл Правка Вид Данные Преобразовать Вставка Формат Анализ Прямой маркетинг Графика Сервис Окно Справка
Вывод
Журнал
Частоты
Заголовок
Примечания
Активный набор
Статистики
Регион
Частоты
Статистики
Регион
N Валидные 757
Пропущенные 743
Регион
Валидные Северо-восток 136 9,1 18,0 18,0
Средний Запад 221 14,7 29,2 47,2
Юг 248 16,5 32,8 79,9
Запад 152 10,1 20,1 100,0
Итого 757 50,5 100,0
Пропущенные Системные пропущенные 743 49,5
Итого 1500 100,0
Панель Содержания
Панель Схемы
Процессор IBM SPSS Statistics готов

# ОРГАНИЗАЦИЯ ДАННЫХ

- Особенности организации данных
- Выбор шкал измерения переменных
- Способы кодировки данных
- Способы ввода данных
- Импорт данных из электронных таблиц
- Свойства переменных
- Изменение типа шкалы измерений
- Задание меток переменных
- Задание меток значений
- Задание пропущенных значений
- Прочие свойства переменных
- Копирование и вставка свойств переменных

## Особенности организации данных

1. Исходная информация в источниках данных не всегда представлена в виде, пригодном для обработки в SPSS, поэтому часто возникает необходимость преобразования исходной информации.
2. Каждая строка на листе **Данные** соответствуют единице анализа.

*Примеры единиц анализа:*

- *маркетинговый опрос – опрошенный респондент;*
- *научный физический эксперимент – отдельное зарегистрированное наблюдение;*
- *данные об объеме продаж – продажи за отдельно взятый месяц.*

3. Каждый столбец на листе **Данные** соответствует переменной или характеристике объекта.
4. Число столбцов может быть больше числа измеряемых характеристик по следующим причинам:
   - количество вариантов ответов в вопросе может быть больше одного (неальтернативные вопросы), и каждому варианту соответствует отдельный столбец;
   - замеры характеристики производятся в разные моменты времени, и каждому замеру соответствует отдельный столбец.

## Выбор шкал измерения переменных

1. Определение вопросов для обследования (переменных) позволяет понять, какие статистические методы можно использовать для анализа полученных данных.
2. Выбор методов производится по двум критериям:
   - какую гипотезу необходимо проверить, основываясь на ответах на данный вопрос;
   - каков уровень измерения данных, доступных при ответе на данный вопрос.
3. Ответы на большинство вопросов обследования могут быть представлены в виде чисел, которые затем можно легко анализировать.
4. Для того, чтобы полученные данные можно было обработать, прежде всего, следует создать кодировочную таблицу.
5. Кодировочная таблица устанавливает соответствие между отдельными вопросами анкеты и переменными, используемыми при компьютерной обработке данных в SPSS.
6. Кодировочная таблица также устанавливает соответствие между возможными значениями переменных и кодовыми числами.

## Способы кодировки данных

1. SPSS позволяет кодировать данные как числами, так и текстовыми кодами (или смешанными число-текстовыми кодами).
2. Числа предпочтительнее использовать для кодировки по причине того, что многие статистические процедуры не позволяют работать с текстовыми кодами.
3. SPSS позволяет присваивать каждому числовому коду соответствующий ему текстовый вариант ответа, то есть метку значения.
4. Наличие меток значений в файле данных предоставляет всю информацию, необходимую для последующего анализа.
5. При разработке схемы кодировки вопроса важно представлять все возможные варианты ответа, чтобы учесть их в формулировании закрытого вопроса (вопрос с заранее определенными возможными ответами).
6. В случае с открытыми вопросами, предполагающими свободный стиль ответа, следует использовать следующую стратегию:
   - на основе выборочного просмотра 20-30% опросных листов сформировать кодификатор;
   - проводить последующую кодировку с использованием этого кодификатора;
   - в случае появления новых уникальных вариантов ответа добавлять их в кодификатор.

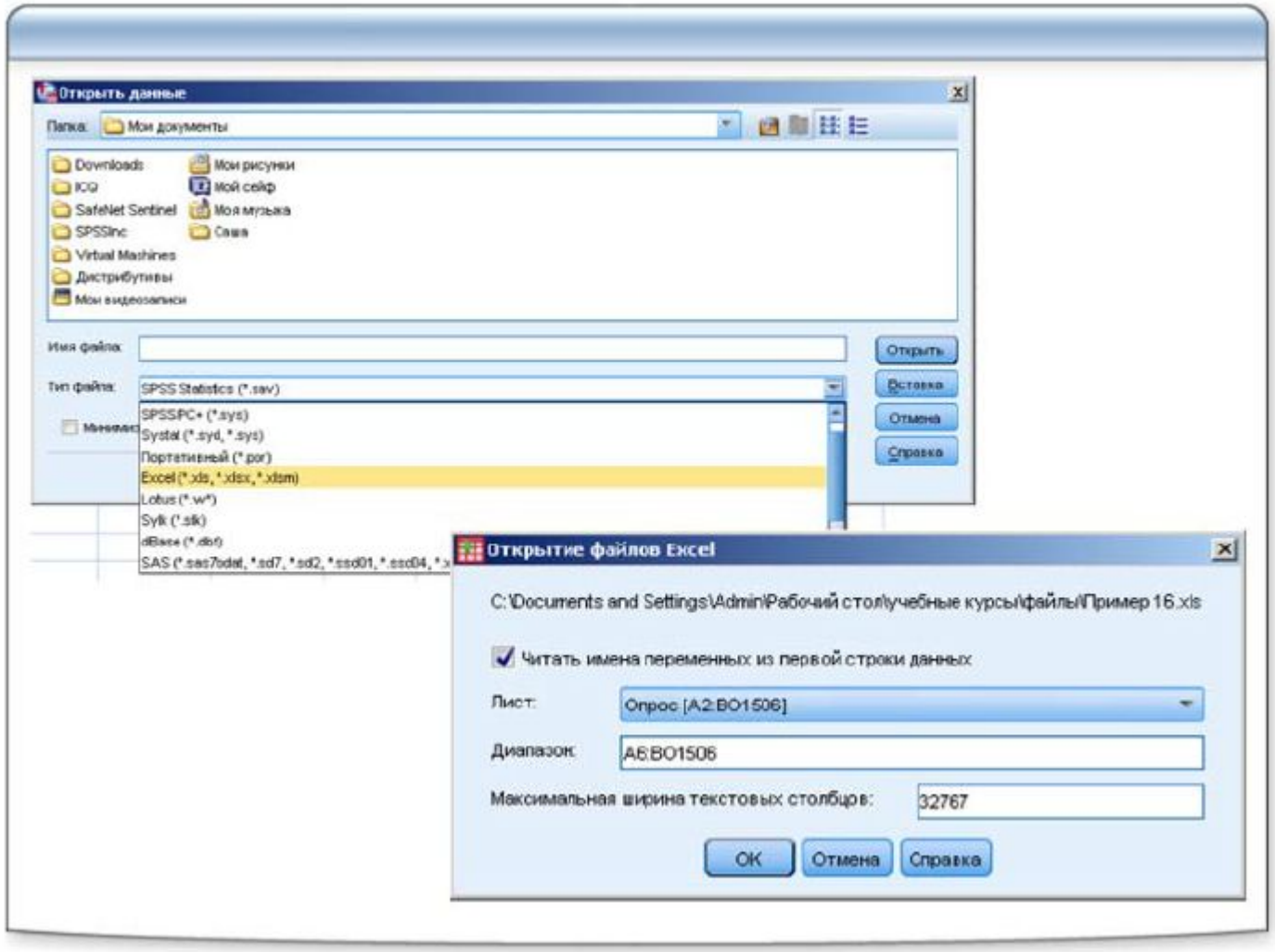
## Задание меток переменных

1. Метка переменной представляет собой расшифровку системного имени переменной, т.е. заголовка столбца на листе **Данные**.
2. В качестве метки переменной можно использовать формулировку вопроса или название регистрируемого признака.
3. Метки переменных не требуются для проведения статистического анализа данных.
4. Метки переменных рекомендуется добавлять в файл данных по следующим причинам:
   - метки переменных отображаются в таблицах и диаграммах, что полезно в том случае, если с результатами анализа должен ознакомиться пользователь, не работавший ранее с файлом данных;
   - по той же причине метки переменных полезны, когда анализ данных выполняется пользователем, не работавшим ранее с этим файлом данных.
5. Для задания меток переменных следует:
   - перейти на лист **Переменные Редактора данных**;
   - нажать левой кнопкой мыши на ячейке, находящейся на пересечении строки с именем переменной, метку которой необходимо задать, и столбца **Метка**;
   - ввести текст метки переменной.
6. В метках переменных различаются прописные и строчные буквы.

# Импорт данных из электронных таблиц

1. SPSS позволяет загружать данные в **Редактор данных** непосредственно из файла электронных таблиц, таких как Excel, SuperCalc, Lotus 1-2-3.
2. В данном модуле будет рассмотрен только вариант загрузки данных из файлов Excel.
3. Для загрузки файла электронных таблиц Excel в SPSS необходимо:
   - в меню **Файл** выбрать команду **Открыть/Данные**;
   - выбрать в выпадающем списке **Тип файла** тип **Excel (*.xls)**;
   - указать в окне открытия файлов путь к файлу электронных таблиц **Excel**;
   - дважды нажать левой кнопкой мыши по названию файла или выделить его и нажать кнопку **Открыть**;
   - в появившемся окне при помощи флажка в пункте **Читать имена переменных из первой строки данных** указать, следует ли первую строку данных листа файла Excel читать как имена переменных;
     *Чтобы корректно ответить на вопрос о чтении имен переменных, необходимо, прежде всего, открыть исходные данные в Excel и понять, присутствуют ли в первой строке данных имена переменных.*
   - при помощи выпадающего списка **Лист** указать лист файла Excel, из которого будут загружаться данные;
   - в поле **Диапазон** задать диапазон ячеек листа Excel, в котором содержатся данные;
     *Чтобы корректно задать диапазон ячеек, необходимо предварительно открыть исходные данные в Excel и уточнить, в каком диапазоне ячеек находятся данные.*
   - для загрузки данных следует нажать кнопку **ОК**.

# Импорт данных из электронных таблиц



## Упражнение

*Тема:* Чтение электронных таблиц из файлов Excel.

1. Исходная ситуация: из комментариев, прилагаемых к файлу электронных таблиц **Excel,** известно, что этот файл включает в себя только 1 лист **Опрос** и данные в этом листе располагаются в диапазоне ячеек **A6:BO1506**, причем первая строка включает в себя имена переменных.
2. В меню **Файл** выберите команду **Открыть/Данные**.
3. В выпадающем списке **Тип файла** выберите тип **Excel (*.xls)**.
4. Выберите файл *Пример.xls*.
5. В открывшемся диалоговом окне отметьте флажок в пункте **Читать имена переменных из первой строки данных**.
6. В поле **Диапазон** наберите **A6:BO1506**.
   *Имена столбцов Excel следует набирать латинскими буквами.*
7. Для открытия данных нажмите кнопку **ОК**.
8. Перейдите на лист **Данные** окна **Редактора данных**, чтобы убедиться в корректности загрузки данных.

# ИЗМЕНЕНИЕ ЗНАЧЕНИЙ ДАННЫХ

- Поиск ошибок и противоречий в данных
- Поиск недопустимых значений
- Корректировка недопустимых значений
- Вычисление новых переменных
- Категоризация количественных переменных
- Автоматическая перекодировка
- Сортировка данных
- Отбор наблюдений
- Вывод значений переменных
- Визуальная категоризация
- Установка вывода значений и меток значений

# Поиск ошибок и противоречий в данных

- Выделяют два основных вида явных ошибок в данных:
  - недопустимые значения (например, возраст респондента 325 лет)
  - ошибки логики (например, возраст респондента 30 лет и возраст первого ребенка 20 лет)
  - Латентные ошибки – скрытые неявные ошибки данных

# Поиск ошибок и противоречий в данных

1. После ввода данных в SPSS не следует сразу же приступать к их анализу.
2. На первом этапе статистического анализа данные следует подвергнуть подробному и всестороннему анализу для обнаружения ошибок.
3. Выделяют два основных вида явных ошибок в данных:
   - наличие в данных недопустимых значений – в переменной для конкретного наблюдения находится значение, которого быть не может исходя из природы самой переменной;

     *Пример: в переменной Возраст респондента находится значение 325 лет.*

   - ошибки логики – в двух (или нескольких) переменных для конкретного наблюдения находятся значения, которые не могут одновременно находиться в одном наблюдении исходя из природы этих переменных.

     *Пример: в переменной Возраст респондента находится значение 30 лет и одновременно для этого же респондента в переменной Возраст первого ребенка находится значение 20 лет.*
4. Явные ошибки в данных можно выявить и скорректировать при помощи SPSS.
5. Вместе с тем существуют еще два основных вида латентных ошибок в данных:
   - Наличие в данных допустимого, но некорректно введенного значения по сравнению с исходным материалом, на основе которого проводился ввод.

     *Пример: оператор, вводя данные, по ошибке ввел не то значение, которое было указано в анкете (причем это значение имеет допустимый характер).*

   - Наличие в данных допустимого значения, имеющего экстремальный характер.

     *Пример: респондент на вопрос о том, сколько среднем часов в день он проводит у телевизора, указал значение 20. Это значение является допустимым, поскольку в сутках 24 часа, но имеет экстремально высокий характер. Более того, такую ситуацию нельзя считать ошибкой в строгом смысле слова.*

# Категоризация количественных переменных

1. В процессе анализа может возникать необходимость определенным образом сгруппировать значения исходной переменной в категории.
   *Пример: необходимость перейти от количественной переменной возраста респондента к ограниченному количеству возрастных категорий.*
2. Для выполнения группировки данных используется процедура **Перекодировать в другие переменные**.
3. В результате работы этой процедуры в файле данных появится новая переменная, в которой вместо значений исходной переменной будут находиться значения категорий.

*Тема:* Группировка данных.

1. Откройте файл данных *Пример 4.sav*.
2. В целях анализа необходимо перекодировать количественную переменную *Возраст респондента* в новую категориальную переменную *Возраст респондента по категориям* с 5 категориями.
3. В меню **Преобразовать** выберите команду **Перекодировать в другие переменные**.
4. Перенесите переменную *Возраст респондента* из списка переменных в поле **Входная переменная –> Выходная переменная** при помощи кнопки со стрелкой.
5. В группе **Выходная переменная** в поле **Имя** наберите имя новой переменной **Возрасткатегория**, в поле **Метка** наберите метку новой переменной **Возраст респондента по категориям**.

## Упражнение (продолжение)

6. Нажмите кнопку **Изменить**.
7. Нажмите кнопку **Старые и Новые значения**.
8. В группе **Старое значение** отметьте пункт **Диапазон от наименьшего значения до указанного** и наберите в числовом поле значение **20**.
9. В группе **Новое значение** наберите в числовом поле значение **1**.
10. Нажмите кнопку **Добавить**.
11. В группе **Старое значение** отметьте пункт **Диапазон** и введите в верхнее числовое поле значение **21**, а в нижнее – **30**.
12. В группе **Новое значение** наберите в числовом поле значение **2**.
13. Нажмите кнопку **Добавить**.
14. В группе **Старое значение** отметьте пункт **Диапазон** и введите в верхнее числовое поле значение **31**, а в нижнее – **40**.
15. В группе **Новое значение** наберите в числовом поле значение **3**.
16. Нажмите кнопку **Добавить**.
17. В группе **Старое значение** отметьте пункт **Диапазон** и введите в верхнее числовое поле значение **41**, а в нижнее – **50**.
18. В группе **Новое значение** наберите в числовом поле значение **4**.
19. Нажмите кнопку Добавить.

20. В группе **Старое значение** отметьте пункт **Диапазон** от указанного значения до наибольшего и введите в числовое поле значение **51**.

21. В группе **Новое значение** наберите в числовом поле значение **5**.

22. Нажмите кнопку **Добавить**.

23. В группе **Старое значение** отметьте пункт **Системное или пользовательское пропущенное**

24. В группе **Новое значение** наберите в числовом поле значение **9**.

25. Нажмите кнопку **Добавить**.

26. Нажмите кнопку **Продолжить** для закрытия окна.

27. Нажмите кнопку **ОК**.

28. Откройте лист **Данные** окна **Редактора данных**.

29. При помощи горизонтальной полосы прокрутки перейдите к правой границе файла и убедитесь в том, что переменная *Возраскатегория* была создана.

30. Для новой переменной задайте метки значений:

- 1 – до 20 лет;
- 2 – от 21 до 30 лет;
- 3 – от 31 до 40 лет;
- 4 – от 41 до 50 лет;
- 5 – 51 год и старше.

**Обратите внимание!** Необходимо учесть, что могут быть значения 0 (пользовательские пропущенные), которые перекодируются в категорию возраста до 20 лет.

# ОПИСАТЕЛЬНАЯ СТАТИСТИКА ДЛЯ КАТЕГОРИАЛЬНЫХ ДАННЫХ

*Частотный анализ  для категориальных переменных.*
*Частотные таблицы для порядковых шкал.*
*Графическое представление категориальных переменных.*
*Таблицы сопряженности и их анализ.*
*Критерий хи-квадрат.*

# ТАБЛИЦЫ СОПРЯЖЕННОСТИ

| | $y_1$ | $y_2$ | … | $y_k$ | $z_i$ |
|---|---|---|---|---|---|
| $x_1$ | $f_{11}$ | $f_{12}$ | … | $f_{1k}$ | $z_1$ |
| $x_2$ | $f_{21}$ | $f_{22}$ | $f_{ij}$ | $f_{2k}$ | $z_2$ |
| … | … | … | … | … | … |
| $x_m$ | $f_{m1}$ | $f_{m2}$ | … | $f_{mk}$ | $z_m$ |
| $s_j$ | $s_1$ | $s_2$ | … | $s_k$ | $N$ |

**$y_j$ – категории переменной столбцов**

**$x_i$ – категории переменной строк**

**$s_j$ – суммы по столбцам**

**$z_i$ – суммы по строкам**

**$N$ – общее число наблюдений**

***Таблица сопряженности двумерного распределения*** *категориальных переменных для выявления взаимосвязи:*

*- строки таблицы задаются категориями одной переменной;*

*- столбцы таблицы задаются категориями другой переменной;*

*-на пересечении строки i и столбца j в таблице сопряженности находится количество объектов (наблюдений, записей), для которых переменная строк принимает значение i, а переменная столбцов - значение j.*

*Независимая переменная оказывает влияние на зависимую переменную.*
*Например: уровень образования респондента может оказывать влияние на категорию частоты просмотра телепередач.*

***зависимая** переменная задает строки таблицы,*
***независимая** переменная - столбцы*

# ПРОЦЕНТЫ В ЯЧЕЙКАХ ТАБЛИЦЫ СОПРЯЖЕННОСТИ

- Для более четкого понимания связи переменных между собой можно использовать не только частоты, но и проценты
- Фактические частоты: $f_{ij}$
- Проценты по строкам: $\frac{f_{ij}}{z_i} \cdot 100\%$
- Проценты по столбцам: $\frac{f_{ij}}{s_j} \cdot 100\%$

# Ожидаемые частоты и остатки в таблицах сопряженности

- Более тщательное исследовать существование зависимости между переменными позволяет вычисление значений:
  - ожидаемых частот: $e_{ij} = \frac{s_j z_i}{N}$
  - остатков: $r_{ij} = f_{ij} - e_{ij}$
  - стандартизированных остатков: $\frac{r_{ij}}{\sqrt{e_{ij}}}$
  - уточненных стандартизированных:

$$\frac{r_{ij}}{\sqrt{e_{ij}\left(1-\frac{z_i}{N}\right)\left(1-\frac{s_j}{N}\right)}}$$

# ОЖИДАЕМЫЕ ЧАСТОТЫ И ОСТАТКИ В ТАБЛИЦАХ СОПРЯЖЕННОСТИ

1. *Вычисление остатков – разница между наблюденными частотами и ожидаемыми частотами. Остатки являются показателем того, насколько сильно наблюдаемые и ожидаемые частоты отклоняются друг от друга.*
2. *Нестандартизированные – отображаются ненормированные остатки, т.е. разность между наблюдаемыми и ожидаемыми частотами;*
3. *Стандартизированные – отображаются нормированные остатки, для этого ненормированные остатки делятся на квадратный корень из ожидаемой частоты;*
4. *Скорректированные стандартизированные – нормированные остатки вычисляются с учетом сумм по строкам и столбцам.*

# ОПИСАТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

# МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ

## ЗАДАЧИ СТАТИСТИЧЕСКОГО АНАЛИЗА

*1. оценка параметров – получение точечных и интервальных оценок параметров генеральной совокупности;*

***- проверка статистических гипотез*** *– тестирование выдвинутых на этапе проектирования статистического исследования рабочих гипотез и предположений;*

*- **статистическое изучение взаимосвязи** – построение математико-статистической модели зависимости изучаемых явлений.*

*2. Выбор статистик для оценки параметров генеральной совокупности зависит от свойств выборочной совокупности.*

*3. Методы проверки статистических гипотез и изучения взаимосвязи могут быть параметрические и непараметрические, и также зависят от свойств выборочной совокупности.*

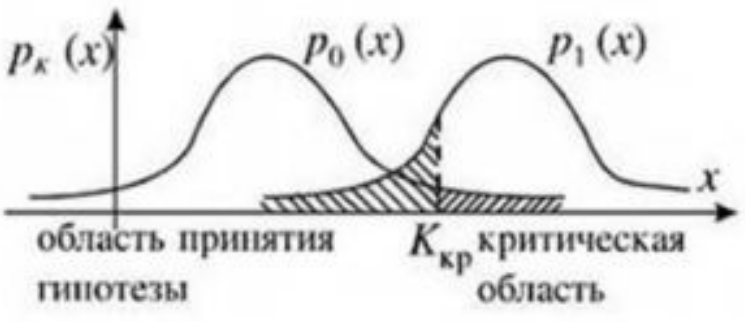
*4. **Основное мастерство статистического анализа** данных заключается в получении и интерпретации данных.*

# МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ

1. *Проверка гипотез и изучение взаимосвязи для категориальных шкал происходит, как правило, на основе таблиц сопряженности.*
2. *Таблица сопряженности – это таблица, строки которой задают категории одной переменной, а столбцы – категории другой переменной.*
3. *Проверка гипотез на основе таблиц сопряженности относят к непараметрическим методам статистики.*
4. *Для применения непараметрических методов не нужна информация о форме и параметрах распределения исследуемой переменной.*

# СТАТИСТИЧЕСКАЯ ГИПОТЕЗА

- **Статистическая гипотеза – любое предположение (утверждение) относительно неизвестного закона распределения переменной в генеральной совокупности или значениях его параметров**



# *СТАТИСТИЧЕСКАЯ ГИПОТЕЗА*

1. *Статистическая гипотеза – это любое предположение (утверждение) относительно неизвестного закона распределения переменной в генеральной совокупности или значениях его параметров.*
2. *Любое статистическое исследование направлено на определение некоторой характеристики изучаемой генеральной совокупности или выявление взаимосвязи между переменными.*

   *Статистическая достоверность – это возможность распространить результат исследования на всю генеральную совокупность.*

1. *Проверка статистической достоверности сводится к проверке статистических гипотез.*

## *НУЛЕВАЯ И АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА*

1. *При проверке статистических гипотез различают основную (выдвигаемую, нулевую) гипотезу, которую необходимо проверить, и альтернативную (конкурирующую) гипотезу.*
2. *Как правило, нулевую гипотезу формируют об отсутствии различий или отсутствии взаимосвязей, или о соответствии заданному закону распределения.*
3. *Точная формулировка гипотезы зависит от конкретного вида гипотезы.*
4. *Альтернативная гипотеза утверждает о наличии связи, различий или расхождении законов распределения.*
5. *Альтернативная гипотеза обычно является «рабочей» гипотезой исследования, которую необходимо доказать.*
6. *Альтернативная гипотеза может быть: ненаправленной – цель исследования просто опровергнуть нулевую гипотезу;*
7. *направленной – цель исследования опровергнуть нулевую гипотезу с учетом направления изменения (различий).*

# КРИТЕРИЙ ХИ-КВАДРАТ

- Проверяется гипотеза о независимости переменных в таблице сопряженности (отсутствии связи)
- Обычная формула:

$$\chi^2_P = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(f_{ij}-e_{ij}\right)^2}{e_{ij}}$$

- Формула на основе правдоподобия:

$$\chi^2_{LR} = 2\sum_{i=1}^{m}\sum_{j=1}^{k} f_{ij}\ln\left(\frac{f_{ij}}{e_{ij}}\right)$$

- Критическое значение: $\chi^2_{кр} = \chi^2\left(\alpha;\nu=(m-1)(k-1)\right)$

$f_{ij}$ – фактические частоты

$e_{ij}$ – ожидаемые частоты

$m$ и $k$ – число строк и столбцов

# КРИТЕРИЙ ХИ-КВАДРАТ

1. ***Критерий хи-квадрат*** *основан на различиях между наблюденными и ожидаемыми частотами в ячейках таблицы сопряженности:*
2. *Чем больше различия между наблюденными и ожидаемыми частотами, тем более вероятно, что между переменными существует зависимость.*
3. ***В критерии хи-квадрат проверяется гипотеза*** *об отсутствии статистической связи между переменными таблицы сопряженности. Само по себе значение критерия хи-квадрат не позволяет сделать однозначного вывода о наличии или отсутствии связи между переменными, поскольку значение этого критерия зависит еще и от количества строк и столбцов в таблице сопряженности.*
4. ***Расчетное значение критерия*** *необходимо сравнить с критическим, взятым из распределения хи-квадрат Пирсона с заданным уровнем значимости и числом степеней свободы.*
5. ***Если расчетное значение критерия больше критического****, то проверяемая нулевая гипотеза отвергается (связь есть).*
6. ***Критерий хи-квадрат*** *применим, когда ожидаемые частоты больше 5 и суммы по строкам и столбцам всегда должны быть больше 0. Обычно должно быть не более 20% ячеек с ожидаемыми частотами меньше 5.*

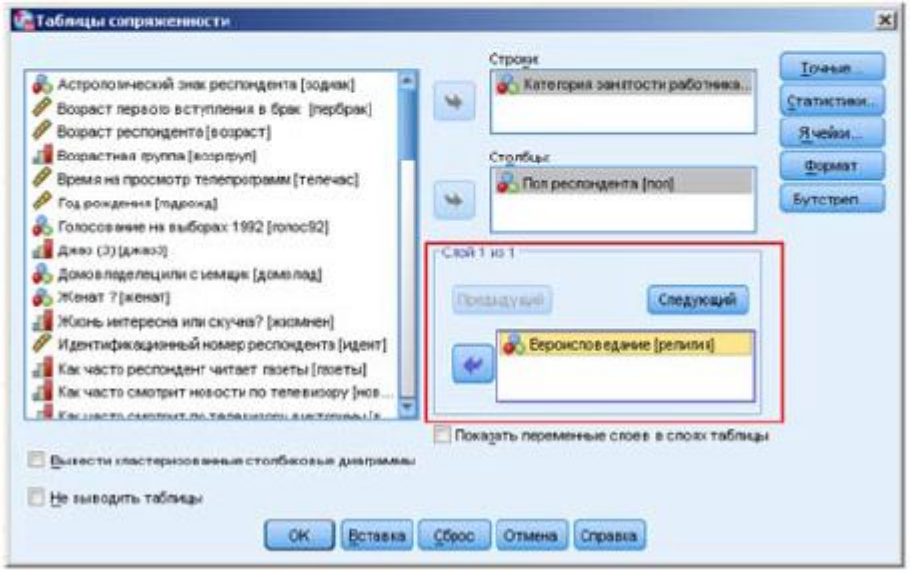
# КРИТЕРИЙ ХИ-КВАДРАТ

*7. **При вычислении критерия хи-квадрат** может использоваться альтернативная формула с поправкой на правдоподобие.*
*8. **При большом объеме** выборки формула Пирсона и подправленная формула дают очень близкие результаты.*

*При анализе основное внимание следует обращать на показатель значимости критерия хи-квадрат:*
*- **если значение в ячейке** Асимпт.значимость (2-стор.) **меньше, чем 0,05** (либо другого приемлемого уровня значимости 0,01; 0,001), то между переменными **существует статистическая связь**;*
*- если значение в этой ячейке **больше или равно 0,05**, то статистическая **связь** между переменными **отсутствует**.*

# ЗАДАНИЕ СЛОЕВ В ТАБЛИЦЕ СОПРЯЖЕННОСТИ

- Количество измерений в таблице сопряженности может превышать два измерения
- Для построения трехмерной таблицы сопряженности следует задать слой



*Процедура* ***Таблицы сопряженности также позволяет добавлять в таблице третье измерение - слои.***

*В этом случае появляется возможность ответить на вопрос, существует ли зависимость между двумя переменными, задаваемыми в строках и столбцах, для различных категорий третьей переменной, которая задается в измерении слоев.*

*Количество слоев также может быть более одного, что позволяет получать таблицы с большим количеством измерений, однако такие таблицы достаточно сложно интерпретировать.*

# ОПИСАТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

# МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ

# СТАТИСТИКИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

- Статистики центральной тенденции используются для выявления типичных значений в исследуемой совокупности
- Мода – для номинальных и порядковых переменных, для количественных переменных
- Медиана – для порядковых переменных и количественных переменных
- Среднее арифметическое – для количественных переменных

***Основные показателя оценки «типичных» значений**:*

*- среднее арифметическое – сумма всех значений числового ряда, деленная на количество значений в ряде.*

*- медиана – значение, разбивающее упорядоченный ряд на две равные части так, что 50% значений ряда имеют значение ниже медианы, а другие 50% – выше медианы.*

*- мода – наиболее часто встречающееся значение данных.*

# Медианное значение

- Не сгруппированные данные – ранжируем ряд
- Дискретный вариационный ряд – ищем накопленную частоту
- Интервальный вариационный ряд – по формуле:

$$Me = x_0 + i \cdot \frac{\frac{1}{2}\sum f_i - S_{Me-1}}{f_{Me}},$$

где $x_0$ – нижняя граница медианного интервала

$i$ – ширина медианного интервала

$f_{Me}$ – частота медианного интервала

$S_{Me-1}$ – накопленная частота интервала, предшествующего медианному

# *МЕДИАННОЕ ЗНАЧЕНИЕ*

1. *Медианное значение – альтернативная мера оценки среднего значения в генеральной совокупности вместо средней арифметической.*
2. *Медиану используют при статистически неоднородной совокупности и нарушении нормальности распределения.*
3. *Для определения медианы по не сгруппированным данным необходимо:*
   *- ранжировать наблюдения;*
   *- медианой будет центральное значение при нечетном числе наблюдений*
   *или среднее арифметическое двух центральных значений.*

# Статистики разброса

- Дисперсия – сумма квадратов разностей каждого значения переменной и среднего значения, деленная на количество наблюдений
- Стандартное отклонение – корень квадратный из дисперсии
- Максимум и минимум – наибольшее и наименьшее значения, встречающееся в данных
- Размах – разность между максимумом и минимумом

$$R = X_{\max} - X_{\min}$$

# *СТАТИСТИКИ РАЗБРОСА*

1. *Для определения, насколько сильно значения переменной отличаются друг от друга, используют статистики разброса или вариации.*
2. *Статистики разброса показывают степень разброса значений признака от средней величины.*
3. *Задача анализа статистик разброса – обобщить индивидуальные различия изучаемых единиц.*
4. *Показатели вариации при изучении многих процессов служат показателями риска.*
5. *Чем выше степень вариации (статистик разброса), тем выше степень риска (риска отклониться от среднего значения).*

***Вариация*** *– это изменение (варьирование) значений признака у каждой единицы изучаемой совокупности.*

*Используют следующие меры разброса:*

*- **дисперсия** – сумма квадратов разностей каждого значения переменной и среднего значения,*
*деленная на количество наблюдений;*

*- **стандартное отклон**ение – корень квадратный из дисперсии;*

*- максимум – наибольшее значение, встречающееся в данных;*

*-минимум – наименьшее значение, встречающееся в данных;*

*-- размах – разность между максимумом и минимумом.*

*Дисперсия позволяет дать информацию о разбросе значений в целом, но ее недостаток в том, что*
*она измеряется в возведенных в квадрат единицах исходной переменной.*

*Стандартное отклонение измеряется в тех же единицах, что исходная переменная.*

*Максимум, минимум и размах имеют второстепенный характер по сравнению с показателями*
*дисперсии и стандартного отклонения.*

# Дисперсия и стандартное отклонение

- Простая формула:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Взвешенная формула:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i - 1}$$

- Упрощенная формула: $\sigma^2 = \overline{x^2} - (\bar{x})^2$
- Стандартное отклонение: $\sigma = \sqrt{\sigma^2}$

$x_i$ – значение количественного признака

$f_i$ – частота повторения признака

$n$ – число наблюдений

# Процентили

$$P = x_0 + i \cdot \frac{P\sum_{i=1}^{n} f_i - S_{m-1}}{f_m}$$

где $x_0$ – нижняя граница интервала, в котором находится процентиль;

$P$ – процентное значение процентиля

$m$ – интервал, в котором находится процентиль

$i$ – ширина интервала

$f_m$ – частота интервала, в котором находится процентиль

$S_{m-1}$ – накопленная частота интервала класса $m$-1

# ПРОЦЕНТИЛИ

1. *Значение **процентиля** – это значение количественной переменной, которое разделяет упорядоченные данные на две группы таким образом, что определенный процент наблюдений имеет значение количественной переменной меньше или равно значений процентиля, а прочие наблюдения имеют значение больше процентиля.*

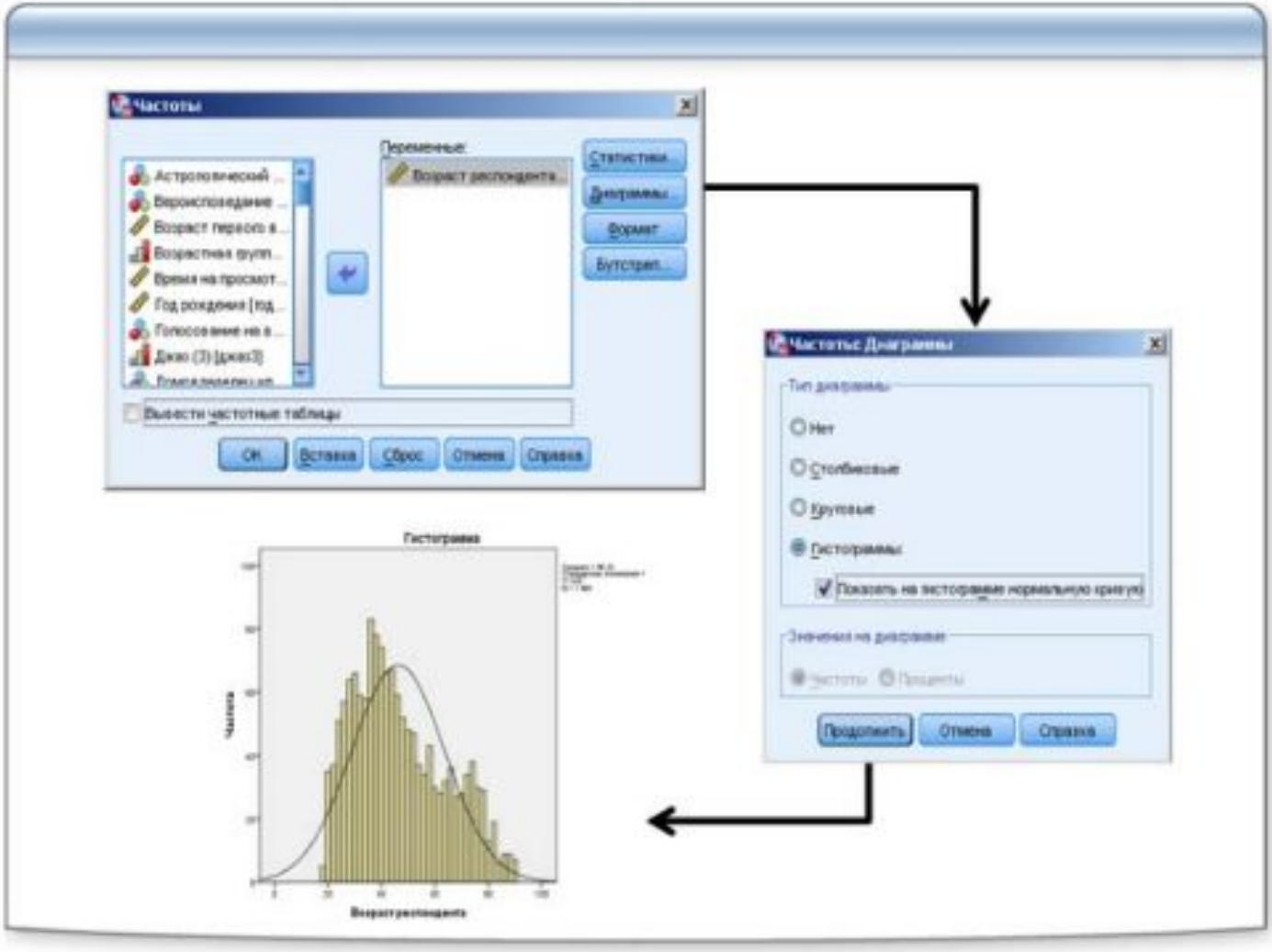
*К примеру, выражение «40% процентиль данных о доходе равен 15.000 рублей» означает, что 40% респондентов имеют доход не выше 15.000 рублей.*

2. *N-% процентиль представляет собой такое значение упорядоченного ряда, N% значений которого меньше или равно N-% процентиля.*
3. *Наиболее часто используемыми процентилями являются квартили.*
4. ***Квартили** – это значения, которые делят упорядоченный ряд на четыре равные части и отделяют, соответственно, 25%, 50% и 75% значений данных.*

*Второй квартиль, отделяющий 50% наблюдений, по определению является также и медианой.*

5. ***децили** (значения, делящие ряд на десять равных частей), квинтили (значения, деляющие ряд на пять равных частей).*
6. *Разница между 75% и 25% процентилями называется межквартильным размахом или межквартильной широтой.*

# Гистограммы



# ГИСТОГРАММЫ

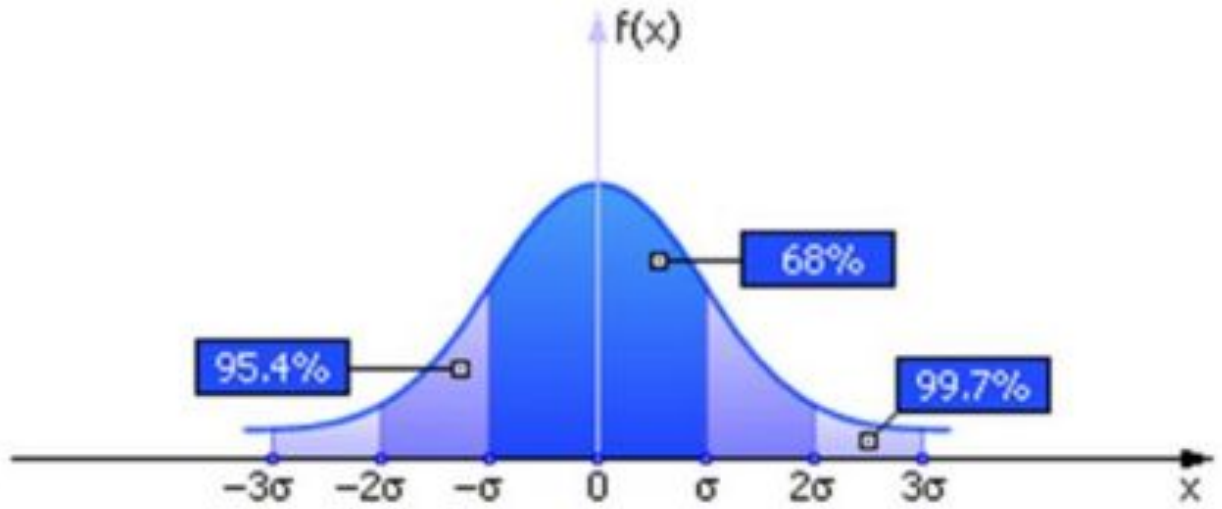
*Одним из наиболее информативных средств подытоживания и наглядного представления количественных переменных является гистограмма.*

*В отличие от столбиковой диаграммы гистограмма позволяет компактно представить в графическом виде переменные с большим количеством уникальных значений.*

*На гистограмме отображаются также стандартное отклонение, среднее значение и общее количество наблюдений, а также кривая нормального распределения.*

# Нормальный закон распределения

- Стандартизированное нормальное распределение



# НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ

*1. Предпосылкой применения большинства методов статистического анализа является нормальность распределения.*

*2. Для переменных, относящихся к интервальной шкале и подчиняющихся нормальному распределению, в качестве основной обобщающей характеристики используют среднее значение, а в качестве меры разброса – стандартное отклонение или стандартную ошибку.*

*3. Для порядковых или интервальных переменных, не подчиняющихся нормальному распределению, в качестве основной характеристики используют медиану, а в качестве меры разброса – межквартильный размах или широту (разницу между первым и третьим квартилями).*

*4. Для выборки, подчиняющейся нормальному распределению, в интервале шириной:*

*- равной удвоенному стандартному отклонению, который отложен по обе стороны от среднего значения, располагается примерно 68% всех наблюдений;*

*- равной четырем стандартным отклонениям, который отложен по обе стороны от среднего значения, располагается примерно 95% всех наблюдений;*

*- равной шести стандартным отклонениям, который отложен по обе стороны от среднего значения, располагается примерно 99,7% всех наблюдений.*

# Асимметрия и эксцесс

- Центральный момент порядка *k*:

$$\mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k}{n-1} \qquad \mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k f_i}{\sum_{i=1}^{n} f_i - 1}$$

- Коэффициент асимметрии: $A = \frac{\mu_3}{\sigma^3}$
- Коэффициент эксцесса: $E = \frac{\mu_4}{\sigma^4} - 3$

$\sigma$ – стандартное отклонение

$x_i$ – значение признака

$f_i$ – частота повторения признака

*k* – порядок момента

*n* – число наблюдений

# АСИММЕТРИЯ И ЭКСЦЕСС

1. *Асимметрия и эксцесс – это статистики, описывающие форму и симметричность распределения изучаемой переменной по сравнению с нормальным распределением.*
2. *Коэффициенты асимметрии и эксцесса (коэффициенты формы распределения) рассчитывают для оценки нормальности распределения.*
3. *У нормального распределения коэффициенты асимметрии и эксцесса равны нулю.*
4. *Если коэффициенты формы и симметричности распределения по модулю меньше 1, то распределение близко к нормальному.*
5. *Коэффициент асимметрии показывает симметричность распределения по сравнению с нормальным распределением.*
6. *Коэффициент эксцесса показывает отличие формы распределения от нормального распределения.*

# Проверка распределения на нормальность

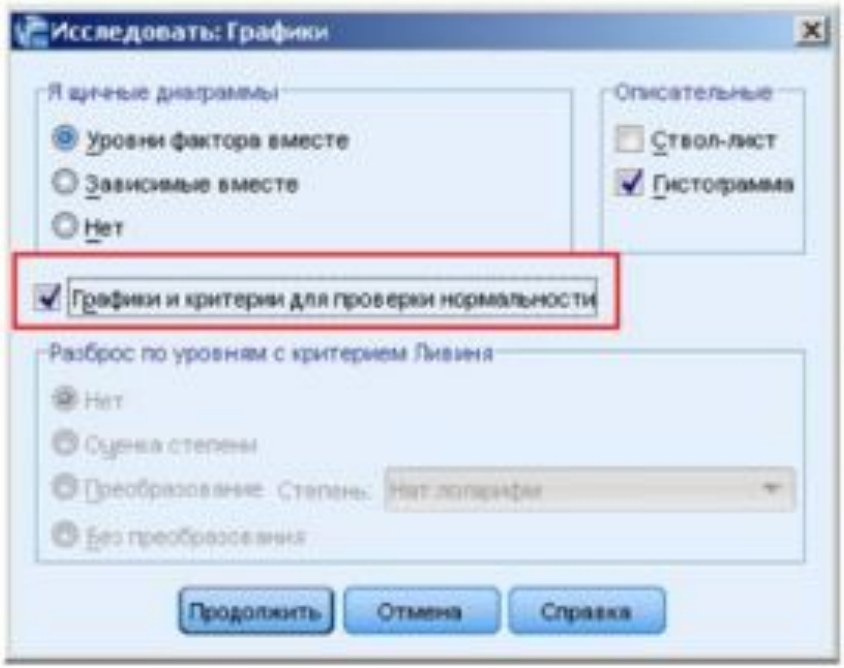
Методы проверки распределения на нормальность:

- анализ моды, медианы, среднего значения, усеченного среднего, эксцесса и асимметрии
- визуальный анализ гистограммы с наложением кривой нормального распределения
- критерии Колмогорова-Смирнова или Шапиро-Уилка

# ПРОВЕРКА РАСПРЕДЕЛЕНИЯ НА НОРМАЛЬНОСТЬ

1. *Для приблизительной оценки нормальности распределения необходимо сравнить моду, медиану и среднее значение, а также проанализировать коэффициенты асимметрии и эксцесса.*
2. *Если характеристики меры тенденции приблизительно равны между собой, а коэффициенты асимметрии и эксцесса приблизительно равны нулю, то можно признать распределение нормальным.*
3. *Для визуальной оценки нормальности распределения можно построить гистограмму с наложением кривой нормального распределения.*
4. *В качестве формального теста на нормальность можно использовать критерии КолмогороваСмирнова.*

# Тест на нормальное распределение




**Критерий нормальности**

| | Колмогоров-Смирнов[a] | | | Шапиро-Уилк | | |
|---|---|---|---|---|---|---|
| | Статистика | ст. св. | Значимость | Статистика | ст. св. | Значимость |
| Возраст респондента | ,090 | 1495 | ,000 | ,955 | 1495 | ,000 |

a. Поправка значимости Лильефорса

## *ТЕСТ НА НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ*

1. *Для проверки гипотезы о нормальности распределения применяют статистический тест Колмогорова-Смирнова или Шапиро-Уилка.*
2. *В тесте на нормальность проверяется гипотеза, что исследуемое распределение соответствует нормальному.*
3. *Для применения теста на нормальность необходимо в диалоговом окне Исследовать процедуры Разведочный анализ установить параметр Графики и критерии для проверки нормальности.*
4. *В результате выполнения теста на нормальность в окне Вывода будет получена таблица с расчетными значениями критерия Колмогорова-Смирнова и Шапиро-Уилка.*
5. *Если в таблице результатов теста на нормальность получена вероятность в столбце Значимость менее 0,05 (или другого выбранного уровня значимости), то проверяемое распределение значимо отличается от нормального.*

# ВЫЯВЛЕНИЕ СТАТИСТИЧЕСКОЙ ВЗАИМОСВЯЗИ МЕЖДУ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

# Классификация видов взаимосвязи

- Статистическая взаимосвязь – зависимость, которая проявляется не в каждом отдельном случае, а в общем, в среднем, при большом числе наблюдений
- По направлению:
  - прямая и обратная
- По тесноте:
  - слабая, умеренная и сильная
- По форме:
  - линейная и нелинейная
- По количеству признаков:
  - парная и множественная

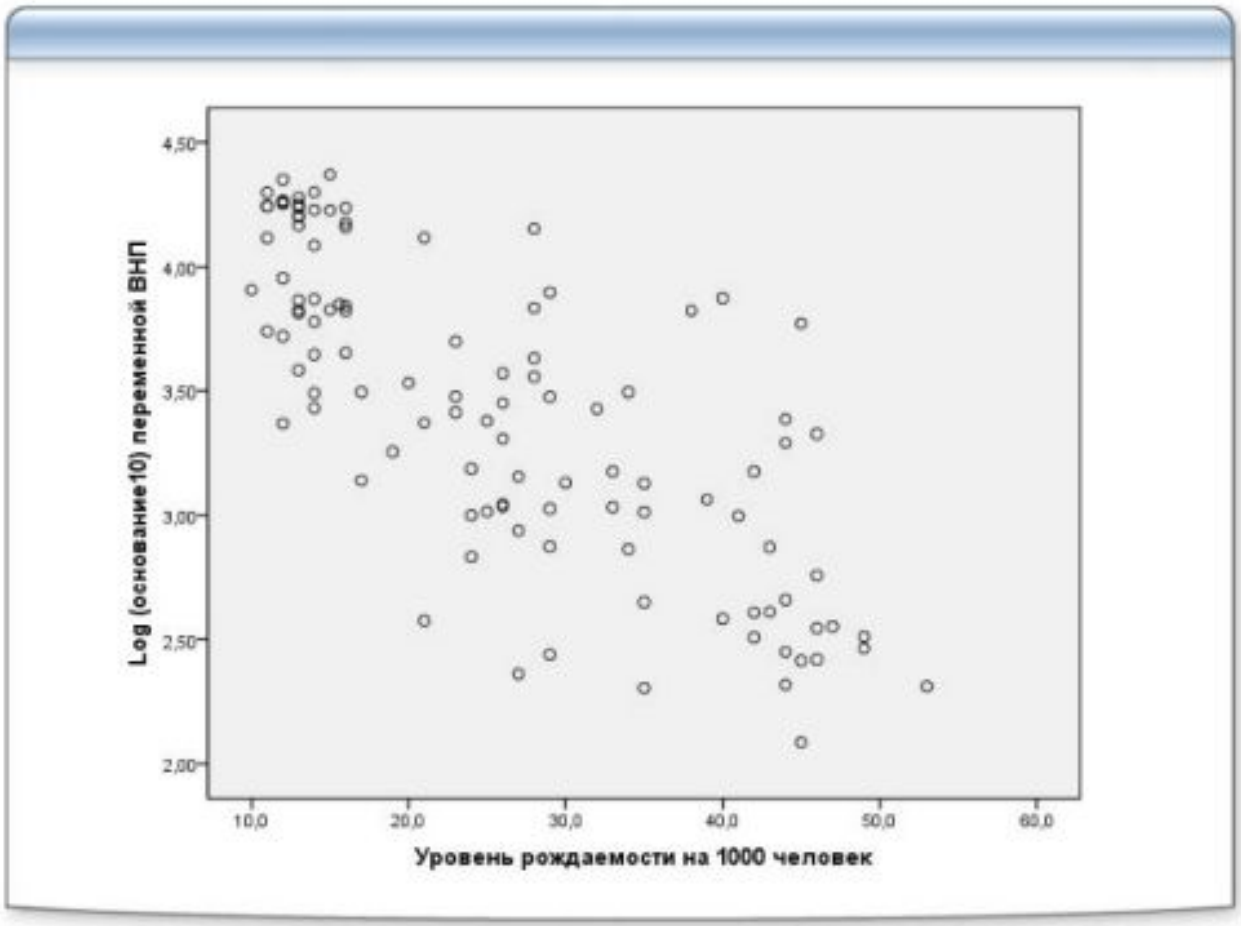
## КЛАССИФИКАЦИЯ ВИДОВ ВЗАИМОСВЯЗИ

1. *Одна из задач статистического анализа – изучение взаимосвязи между переменными.*
2. *Факторные признаки – переменные, которые обуславливают изменение других, связанных с ним переменных (результативных). Результативные признаки – переменные, значения которых формируются под воздействием факторных признаков.*
3. *Функциональная взаимосвязь – такая связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.*
4. *Статистическая взаимосвязь – зависимость, которая проявляется не в каждом отдельном случае, а в общем, в среднем, при большом числе наблюдений.*
5. *По направлению связи различают статистическую зависимость: прямую и обратную.*
6. *По тесноте связи различают статистическую зависимость: слабую, статистически незначимую; среднюю или умеренную; сильную. 8.*
7. *По форме выражения связи различают статистическую зависимость: линейную и нелинейную.*
8. *По количеству изучаемых переменных различают парные и множественные модели взаимосвязи.*

## *ДИАГРАММА РАССЕЯНИЯ*

1. *При анализе взаимосвязи двух переменных можно построить диаграмму рассеяния (поле корреляции).*
2. *Диаграмма рассеяния – это простейший графический способ изучения взаимосвязи между количественными переменными.*
3. *Диаграмма рассеяния представляет каждую единицу совокупности в пространстве двух измерений, соответствующих двум переменным.*
4. *Обычно при построении диаграммы рассеяния придерживаются следующих правил: - по оси абсцисс (горизонтали) указывают значение факторной переменной; - по оси ординат (вертикали) указывают значение результативной переменной.*
5. *При отсутствии взаимосвязи между изучаемыми переменными точки на диаграмме рассеяния будут расположены случайным образом.*
6. *Чем сильнее взаимосвязь между изучаемыми переменными, тем ближе будут группироваться точки вокруг определенной линии, выражающей форму связи.*
7. *При построении диаграммы рассеяния следует обращать внимание на: - наличие выбросов (выбросы необходимо исключить из дальнейшего анализа).*

# Диаграмма рассеяния



# Коэффициент корреляции Пирсона

- Коэффициент корреляции:

$$r = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Ковариация: $\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X - \overline{X})(Y - \overline{Y})$

$\overline{X}$ и $\overline{Y}$ – средние значения переменных

$\sigma_x$ и $\sigma_y$ – стандартные отклонения переменных

*n* – число наблюдений

- Коэффициент корреляции показывает направление и тесноту связи
- Измеряется от -1 до 1

## *КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА*

1. *Корреляционный анализ – метод определения тесноты и направления линейной взаимосвязи между двумя количественными переменными.*
2. *Теснота и направление связи выражается величиной парного коэффициента корреляции Пирсона.*
3. *Значение коэффициента корреляции безразмерная величина, принимающая значения от -1 до +1.*
4. *Чем ближе значение коэффициента корреляции по модулю к 1, тем сильнее взаимосвязь, и, наоборот, чем ближе значение к нулю – тем слабее взаимосвязь.*
5. *Знак коэффициента корреляции характеризует направление связи между изучаемыми признаками : Положительный знак говорит о прямой взаимосвязи между изучаемыми переменными, отрицательный знак – об обратной взаимосвязи.*
6. *Коэффициент корреляции можно вычислять только для количественных переменных.*
7. *При расчете коэффициента корреляции факторный и результативный признак должны иметь нормальное распределение.*
8. ***В случае отсутствия нормального распределения** можно применить нормализующее преобразование или **рассчитать альтернативный коэффициент ранговой корреляции**, рассмотренный ниже.*
9. *Значение коэффициента корреляции не изменится, если факторный и результативный признак поменять местами.*

# Коэффициент корреляции Пирсона (продолжение)

Для интерпретации силы связи можно использовать следующую таблицу сравнения абсолютной величины коэффициента корреляции Пирсона:

| Значение | Интерпретация |
|---|---|
| До 0,2 | Очень слабая |
| 0,2-0,5 | Слабая |
| 0,5-0,7 | Средняя |
| 0,7-0,9 | Высокая |
| Выше 0,9 | Очень высокая |

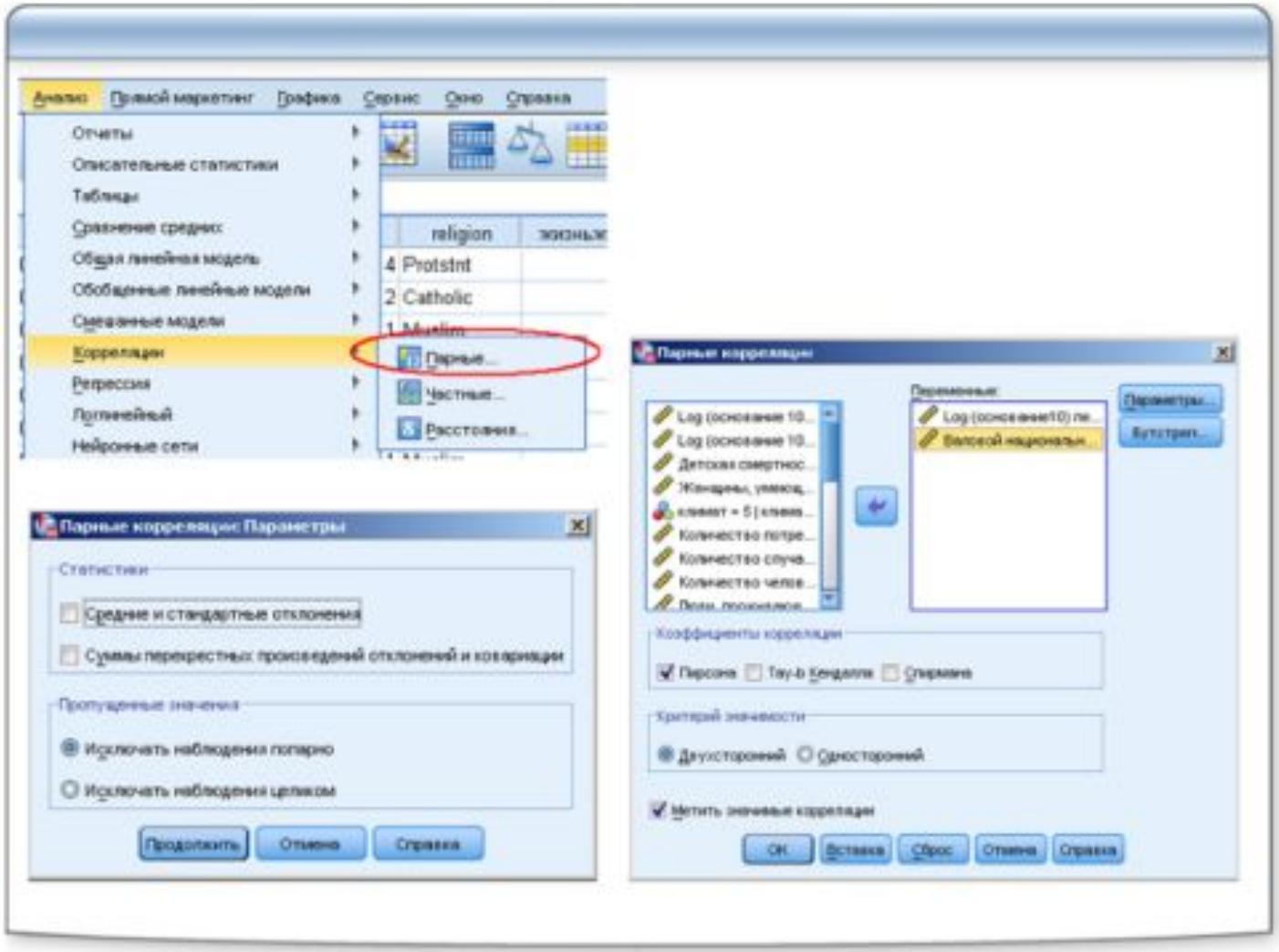
## Процедура Парные корреляции

1. Для расчета парного коэффициента корреляции необходимо вызывать процедуру **Парные корреляции** нужно:
   - в меню **Анализ** выбрать команду **Корреляции/Парные**;
   - в диалоговом окне **Парные корреляции** в списке слева выбрать переменные, для которых необходимо рассчитать коэффициенты корреляции и перенести их в поле **Переменные**;
   - в области **Коэффициенты корреляции** выбрать параметр **Пирсона**;
   - в случае необходимости задать дополнительные параметры, нажав кнопку **Параметры**:
     - в поле **Статистики** задается расчет средних величин, стандартных отклонений, суммы перекрестных произведений;
     - в поле **Пропущенные значения** параметры исключения пропущенных значений.
   - установить по желанию параметр **Метить значимые корреляции**;

*В этом случае в окне **Вывода** с результатами анализа значимые коэффициенты будут выделены двумя звездочками. О значимости коэффициентов взаимосвязи будет подробно сказано ниже.*

   - нажать кнопку **ОК**, запустив процедуру.

# Процедура Парные корреляции



# Ранговые коэффициенты корреляции

- Ранжирование – упорядочение наблюдений по возрастанию или убыванию
- Ранг – порядковый номер значения признака в ранжированной совокупности
- Непараметрический метод оценки взаимосвязи на основе рангов
- Используют в случае нарушения нормальности распределения или для порядковых переменных

# РАНГОВЫЕ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

1. *Ранжирование – это процедура упорядочения объектов изучения, которая выполняется на основе предпочтения.*
2. *Ранг – это порядковый номер значений переменной, расположенных в порядке возрастания или убывания их величин.*
3. *Принцип ранжирования значений исследуемых переменных является основой непараметрическихметодов оценки тесноты связи.*
4. *Среди непараметрическихметодов оценки тесноты взаимосвязи наибольшее распространение получили ранговые коэффициенты корреляции Спирмена (ρ) и Кендалла (τ).*
5. *При вычислении ранговых коэффициентов корреляции Спирмена и Кендалла в формуле для коэффициентов корреляции используются преобразованные значения данных: вместо значений подставляются их ранги.*
6. *Непараметрические ранговые коэффициенты корреляции применяются,* ***когда изучаемые признаки имеют разное распределение, в том числе отличное от нормального распределения****.*
7. *Ранговые коэффициенты корреляции могут рассчитываться как между количественными переменными, так и между порядковыми переменными при условии, что их значения проранжированы по возрастанию или убыванию.*

# Коэффициент корреляции Спирмена

- **Для несвязанных рангов:**

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

- **Для связанных рангов:**

$$T_{x/y} = \frac{1}{12}\sum_{j=1}^{k}\left(t_j^3 - t_j\right)$$

$$\rho = \frac{\frac{1}{6}\left(n^3-n\right) - \sum_{i=1}^{n} d_i^2 - T_x - T_y}{\sqrt{\left[\frac{1}{6}\left(n^3-n\right) - 2T_x\right]\cdot\left[\frac{1}{6}\left(n^3-n\right) - 2T_y\right]}}$$

$d_i$ – разность рангов

$n$ – число пар рангов, $T_j$ – число одинаковых пар рангов

1. Ранговые коэффициенты корреляции принимают значения от -1 до +1.
2. Обычно связь между признаками можно признать существенной, если значения ранговых коэффициентов корреляции Спирмена и Кендалла больше 0,5.
3. Коэффициент Спирмена при наличии связанных и несвязанных рангов рассчитывается по-разному.
4. Связанные ранги – это ранги с одинаковыми значениями.

# Коэффициент корреляции Кендалла

- Для несвязанных рангов:

$$\tau = \frac{2(P-Q)}{n(n-1)}$$

- Для связанных рангов: $V_{x/y} = \frac{1}{2}\sum_{j=1}^{k} t_j(t_j - 1)$

$$\tau = \frac{P-Q}{\sqrt{\left(\frac{n(n-1)}{2} - V_x\right)\left(\frac{n(n-1)}{2} - V_y\right)}}$$

P – сумма проинверсий, Q – сумма инверсий

n – число пар рангов

$V_j$ – число одинаковых пар рангов

## Коэффициент корреляции Кендалла

1. Применение коэффициента Кендалла является предпочтительным, если в исходных данных встречаются выбросы.
2. Расчет коэффициент Кендалла происходит по следующему алгоритму:
   - значения зависимой переменной ранжируют;
   - значения фактора располагают в порядке, соответствующем значениям зависимой переменной;
   - для каждого ранга фактора определяют число следующих за ним значений рангов, превышающих его величину;
   - суммируют полученные на предыдущем этапе значения и получают число соблюдения последовательностей (проинверсий);
   - для каждого ранга фактора определяют число следующих за ним значений рангов, меньших его величины;
   - суммируют полученные на предыдущем этапе значения и получают число нарушений порядка последовательностей (инверсий);
   - рассчитывают коэффициент по формуле.
3. Коэффициент корреляции Кендалла при наличии связанных и несвязанных рангов рассчитывается по-разному.
4. Коэффициент Кендалла, как правило, получается меньше коэффициента Спирмена.