

ДЕПАРТАМЕНТ ОБРАЗОВАНИЯ И НАУКИ  
ХАНТЫ-МАНСКИЙСКОГО АВТОНОМНОГО ОКРУГА

---

ГОУ ВПО «СУРГУТСКИЙ  
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ХАНТЫ-МАНСКИЙСКОГО АВТОНОМНОГО ОКРУГА – ЮГРЫ»

---

Кафедра политологии

Е.В. Дорогонько

**ОБРАБОТКА И АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ  
С ПОМОЩЬЮ ПАКЕТА SPSS**

Учебно-методическое пособие

Сургут  
Издательский центр СурГУ  
2010

## СОДЕРЖАНИЕ

Введение .....	3
<b>1. Обработка данных на компьютере. Подготовительный этап.....</b>	<b>5</b>
1.1. Определение структуры данных .....	5
1.2. Запуск SPSS. Окна программы .....	6
<b>2. Создание и редактирование файлов данных.....</b>	<b>10</b>
2.1. Ввод данных .....	14
<b>3. Управление данными.....</b>	<b>16</b>
3.1. Выбор объектов для анализа.....	16
3.2. Перекодировка в новую переменную.....	18
<b>4. Одномерный описательный анализ социологических данных. Построение частотных (линейных) распределений.....</b>	<b>21</b>
4.1. Частоты.....	21
4.2. Описательные статистики .....	24
<b>5. Взаимосвязь переменных.....</b>	<b>29</b>
5.1. Двумерный анализ социологических данных. Парные распределения.....	29
5.2. Коэффициенты корреляции.....	33
<b>6. Анализ множественных ответов.....</b>	<b>35</b>
6.1. Анализ множественных ответов с применением категориального метода.....	35
6.2. Таблицы сопряженности (парные распределения) вопросов с множественными ответами.....	36
<b>7. Анализ взаимосвязей качественных и количественных переменных. Средние значения.....</b>	<b>39</b>
7.1. Команда T- test для сравнения двух независимых выборок.....	39
7.2. Однофакторный дисперсионный анализ.....	40
<b>8. Регрессионный анализ .....</b>	<b>42</b>
8.1. Парный регрессионный анализ.....	42
8.2. Множественный регрессионный анализ.....	44
<b>9. Факторный анализ.....</b>	<b>47</b>
9.1. Исследование структуры данных.....	47
9.2. Значения факторов .....	51
<b>10. Кластерный анализ.....</b>	<b>53</b>
10.1. Иерархический кластер-анализ.....	54
10.2. Кластерный анализ при большом количестве наблюдений (Кластерный анализ методом к-средних).....	55
<b>Заключение.....</b>	<b>59</b>
<b>11. Словарь основных терминов, используемых в процедурах прикладного социологического исследования.....</b>	<b>60</b>

В пособие рассматриваются статистические методы, применяемые в социологии и политологии с помощью компьютерной программы SPSS. Пособие содержит подробные пошаговые инструкции по выполнению команд, необходимых для получения статистической информации.

Данное пособие окажет помощь студентам специальностей «политология» и «связи с общественностью» при работе со SPSS: в учете и организации исходных данных, в выборе наиболее адекватного метода исследования, в вычислении статистических показателей, в проведении более глубокого анализа данных и интерпретации результатов исследований.

## ВВЕДЕНИЕ

Данное пособие представляет собой практическое руководство по анализу данных с помощью широко используемой в исследовательской практике программы статистической обработки информации – SPSS. Компьютерные методы социологических и политологических исследований изучаются в рамках общепрофессиональных дисциплин «Социология массовых коммуникаций», «Политический анализ и прогнозирование», «Политическая социология», входящих в федеральный компонент государственного образовательного стандарта специальностей «Связи с общественностью» и «Политология».

На сегодня SPSS является самой распространённой программой для обработки статистической информации. В научной литературе в последнее время появилось достаточно много научных работ и практических руководств по работе с программой SPSS<sup>1</sup>. Наиболее полное описание того, как можно анализировать статистические данные с помощью пакета SPSS содержится в учебниках Наследова А. «SPSS 15 профессиональный статистический анализ данных», Бююля А., Цёфеля П. «SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей». Ценным в этих учебниках является то, что описание большинства процедур, предназначенных для анализа данных, приведено в виде подробных и пошаговых инструкций по выполнению команд, необходимых для получения статистической информации.

Такой же принцип соблюден в данном пособии, отличительная черта которого – в использовании практических примеров на материале социологических исследований, проведенных на территории Ханты-Мансийского автономного округа, лабораторией социологических исследований СурГУ. В пособии рассматриваются методы, используемые социологами и политологами в практических исследованиях: построение и анализ одномерных (линейных) и двумерных (парных) частотных распределений и таблиц, анализ взаимосвязи качественных и количественных переменных с помощью различных корреляционных коэффициентов, анализ средних значений, регрессионных (парный и множественный) анализ, факторный анализ, способы табличного и графического представления данных. Подробно описывается, каким образом эти методы могут применяться с помощью пакета SPSS. В пособии обращается внимание на особенности интерпретации результатов анализа социологических данных.

После каждой темы предлагается выполнение небольшого задания для самостоятельной работы, часть из заданий основана на базах данных практических исследований, содержащих большой массив единиц анализа, соответствующий репрезентативной выборке. Поэтому данное пособие сопровождается диском SD-ROM, содержащим файлы с базами данных проведенных исследований.

Данное пособие окажет помощь студентам специальностей «политология» и «связи с общественностью» при работе со SPSS: в учете и организации исходных данных, в выборе наиболее адекватного метода исследования, в вычислении статистических показателей, в проведении более глубокого анализа данных и интерпретации результатов исследований.

Предполагается, что студенты, приступающие к изучению статистического пакета SPSS, имеют знания по базовым курсам математики, информатики, социологии, политической социологии и методов социологического исследования. Достаточно много полезной информации по статистическим методам измерения социологической информации содержится в учебниках В.А. Ядова, И.Ф. Девятко, Е.М. Бабосова.<sup>2</sup> Технологии и методы анализа политической жизни подробно рассматриваются в учебниках Г.П. Артемова и А.С. Ахрименко<sup>3</sup>. Практические методы исследований массовых коммуникаций приводятся в книгах Л.Н. Федотовой, Т.В. Науменко, М.М. Назарова<sup>4</sup>.

Базовые знания технологий и методов исследования социологической и политической информации в сочетании с умениями и навыками работы с компьютерной программой SPSS по статистической обработке и анализу данных помогут студентам в проведении практических исследований, являющихся важной частью работы в области связей с общественностью, прикладного и теоретического анализа политики.

SPSS – это аббревиатура от Statistical Package of the Social Science (статистический пакет для социальных наук). Как следует из названия, SPSS представляет собой набор различных программ обработки данных. Эти программы облегчают процесс ввода информации, позволяют гибко менять структуру данных, использовать самые современные методы обработки и получать результаты в удобной и наглядной форме. В 2009 г. вышла новая версия статистического пакета SPSS 18.0 (теперь PASW Statistics). С информацией о лицензионных версиях SPSS можно познакомиться в Интернете по адресу: <http://www.spss.ru>.

<sup>1</sup> Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. М.:DiaSoft, 2002; Наследов А. SPSS 15 профессиональный статистический анализ данных. СПб: Питер, 2008; Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере/ под ред. В.Э. Фигурнова. М.:ИНФРА-М,2002; Крыштановский А.О. Анализ социологических данных. М: Изд.дом ГУ ВШЭ, 2007.

<sup>2</sup> Ядов В.А. Стратегия социологического исследования. М.: Добросвет, 2003; Девятко И.Ф. Методы социологического исследования. М.:КДУ, 2003, Бабосов Е.М Прикладная социология. Минск:ТетраСистемс, 2000.

<sup>3</sup> Артемов Г.П. Политическая социология. М.:Логос, 2002; Ахрименко А.С. Политический анализ и прогнозирование. М.: Гардарики, 2006.

<sup>4</sup> Федотова Л.Н. Социология массовых коммуникаций. М., 2002; Науменко Т.В. Социология массовой коммуникации. СПб., 2005; Назаров М.М. Массовая коммуникация в современном мире: методология анализа и практика исследований. М., 2003.

Отечественным пользователям удобнее использовать для работы версии, начиная с 12.0, где стала возможной русификация окон ввода и обработки результатов. Скачать демонстрационную версию программы SPSS, срок действия которой 14 дней можно по адресу: <http://www.spss.ru/products/spss/index.htm>.

## 1. Обработка данных на компьютере. Подготовительный этап

Основу программы SPSS составляет SPSS Base (базовый модуль), предоставляющий разнообразные возможности доступа к данным и управления данными. Он содержит методы анализа, которые применяются чаще всего. SPSS Base включает все процедуры ввода, отбора и корректировки данных, а также большинство предлагаемых в SPSS статистических методов. Наряду с простыми методиками статистического анализа, такими как частотный анализ, расчет статистических характеристик, таблиц сопряженности, корреляций, построения графиков, этот модуль включает t-тесты и большое количество других непараметрических тестов, а также усложненные методы, такие как многомерный линейный регрессионный анализ, дискриминантный анализ, факторный анализ, кластерный анализ, дисперсионный анализ, анализ пригодности (анализ надежности) и многомерное шкалирование.

В данном пособии содержится базовая информация об основных методах компьютерной обработки данных.

Анализ данных с применением компьютера включает в себя несколько этапов.

1. Подготовительный этап: определение структуры данных.
2. Ввод данных в компьютер в соответствии с их структурой и требованиями программы.
3. Задание метода обработки данных в соответствии с задачами исследования.
4. Интерпретация результатов обработки.

### 1.1. Определение структуры данных

При работе с социологическими данными используются два основополагающих понятия: единица анализа и переменная. *Единица анализа* – это элементарная, единичная часть объекта исследования. Единица анализа чаще всего совпадает с единицей наблюдения, в социологии, как правило, этой единицей является отдельный респондент. Следовательно, единицей анализа, становится информация, содержащаяся в анкете, чаще всего заполняемой одним респондентом.

*Переменная* - элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа. Простейшие переменные – вопросы анкеты, к примеру, пол и возраст респондента. Значения переменных – варианты анкеты, выбранные респондентами в качестве ответа.

Например, необходимо провести опрос, и выяснить электоральные предпочтения избирателей в отношении политических партий. Анкета может выглядеть следующим образом:

Если проведен опрос 30 респондентов, то единицами анализа в данном случае будут все данные опроса –

Анкета: номер анкеты (заполняется интервьюером) _____	
<b>1. За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?</b>	
1. Единая Россия	
2. Справедливая Россия	
3. КППРФ	
4. Правое дело	
5. ЛДПР	
6. затрудняюсь ответить	
<b>2. Ваш пол</b>	
1. мужской	
2. женский	
<b>3. Ваш возраст (напишите)_____</b>	
<b>4. В каком населенном пункте Вы проживаете (напишите)</b> _____	
<b>var2</b>	<b>2.Ваш пол</b> 1. мужской 2. женский
<b>age</b>	<b>3. Ваш возраст (напишите)_____</b>
<b>terr</b>	<b>4. В каком населенном пункте Вы проживаете (напишите)</b> _____

30 анкет, заполненных респондентами. Одна анкета – одна единица анализа. Переменными каждой единицы анализа будут вопросы анкеты, значения переменных – ответы респондентов – отмеченные варианты вопроса анкеты (если применяется номинальное шкалирование), числовые значения (например, возраст респондента), или буквенные значения (текстовая информация, например, населенный пункт).

#### Кодирование и кодировочная таблица

Для того чтобы полученные данные можно было обработать, прежде всего, следует создать кодировочную таблицу. Кодировочная таблица устанавливает соответствие между отдельными вопросам анкеты и переменными, используемыми при компьютерной обработке данных.

Например, вопросу анкеты «За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?» может

соответствовать имя переменной var1, или party. В версии 13.0 SPSS имя переменной нужно задавать латинскими буквами, цифрами, без пробелов, до 8 символов, причем первым символом имени должна быть буква. Переменные могут принимать различные значения. Например, переменная «пол» может иметь два возможных значения: "женский" и "мужской". Кодировочная таблица определяет кодовые числа, соответствующие отдельным значениям переменных; например, значению "женский" может соответствовать цифра "1", а значению "мужской" — "2". Для нашей анкеты мы можем составить следующую кодировочную таблицу. Она приводится в самой анкете.

### Матрица данных

Предположим, что 10 анкет были заполнены следующим образом:

number	Var1	Var2	age	terr
1	Единая Россия	женский	45	Сургут
2	Единая Россия	мужской	22	Нефтеюганск
3	КПРФ	мужской	19	Нефтеюганск
4	Единая Россия	женский	42	Нефтеюганск
5	Правое дело	мужской	34	Нижневартовск
6	КПРФ	женский	72	Нижневартовск
7	Справедливая Россия	мужской	38	Сургут
8	Справедливая Россия	женский	56	Сургут
9	Справедливая Россия	мужской	61	Сургут
10	Единая Россия	женский	77	Сургут

Приведенная выше таблица называется матрицей данных. Данные, предназначенные для обработки в SPSS для Windows, должны быть представлены в виде такой матрицы. Матрица данных состоит из определенного числа строк и столбцов. Строки и столбцы образуют прямоугольную таблицу. При этом каждая строка соответствует одной анкете, а каждый столбец — одной переменной. Так как в нашем небольшом опросе участвовало 10 респондентов, матрица содержит 10 строк. Каждая строка включает четыре столбца для переменных number (номер анкеты), var1 (первый вопрос анкеты «За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?»), var2 (пол), age (возраст) и terr (населенный пункт).

**Задание.** 1. Разработать анкету для проведения социологического опроса. 2. Подготовить анкету к компьютерной обработке данных, закодировав переменные.

### 1.2. Запуск SPSS. Окна программы

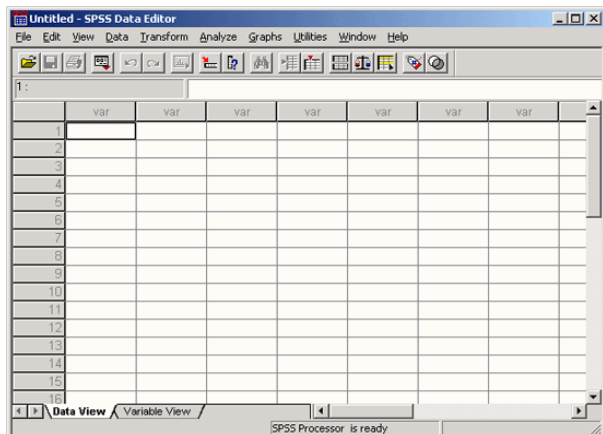
Начнем с ввода данных для небольшого примера анализа.

- Запустите SPSS для Windows, дважды щелкнув левой кнопкой мыши на значке SPSS.

Откроется редактор данных SPSS.

Редактор данных (Data Editor) — это одно из многих окон SPSS. Здесь можно вводить новые данные или загружать существующие из файлов данных с помощью команд меню File (Файл) Open... (Открыть...)

Так как при запуске SPSS ни один файл данных еще не загружен, в заголовке редактора данных стоит "Untitled" (Без имени). Над изображением таблицы в редакторе данных имеются строка меню и панель символов.



Редактор данных это приложение, напоминающее электронную таблицу. Под электронной таблицей подразумевается рабочий лист, разделенный на строки и столбцы, который позволяет про сто и эффективно вводить данные. Отдельные строки таблицы соответствуют отдельным единицам анализа. Например, при обработке данных опроса одна строка содержит данные одного респондента. Отдельные столбцы соответствуют отдельным переменным. При обработке данных наблюдений анкеты в одной переменной хранятся ответы на отдельный вопрос. Отдельные ячейки таблицы содержат значения переменных для каждого отдельного наблюдения; в каждой ячейке хранится одно значение переменной.

Строка меню содержит команды для выполнения почти всех операций, предусмотренных в программе SPSS. Как правило, выполнение команды начинается с появления диалогового окна, в котором пользователю полагается установить значения параметров.

Краткое описание основных меню.

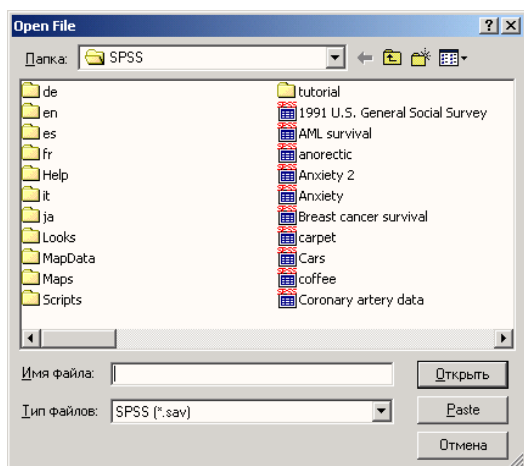
File (Файл)	Команды, предназначенные для открытия, чтения и сохранения файлов, а также команду выхода из программы SPSS
Edit (Редактирование)	Команды редактирования, такие как команды копирования, вставки, замены, поиска и т.п.
View (Просмотр)	Набор команд, влияющих на представление информации на экране. Например, команды Value Labels (метки значений), Fonts (шрифты)
Data (Данные)	Команды для управления вводом и представлением данных
Transform (Преобразование)	Команды, модифицирующие введенные данные, а также создающие новые данные на основе существующих
Analyze (Анализ)	С этого меню начинаются все процедуры анализа данных
Graphs (Графики)	Команды, создающие различные диаграммы
Utilities (Утилиты)	Команды служат для упрощения сложных операций над данными, предназначены для опытных пользователей
Window (Окно)	С помощью этого меню можно управлять взаимным расположением и статусом открытых окон программы SPSS
Help (Помощь)	Меню предназначено для доступа к справочной информации

### Диалоговое окно открытия файла

Диалоговое окно Open File (Открыть файл) является стандартным окном операционной системы и позволяет открыть ранее созданные файлы данных.

■ Для того, чтобы вызвать это окно, выберите в меню File команду Open – Data, либо щелкните мышью на кнопке Open File панели инструментов.

Обратите внимание, что все файлы, созданные с помощью редактора данных SPSS, имеют расширение .sav.

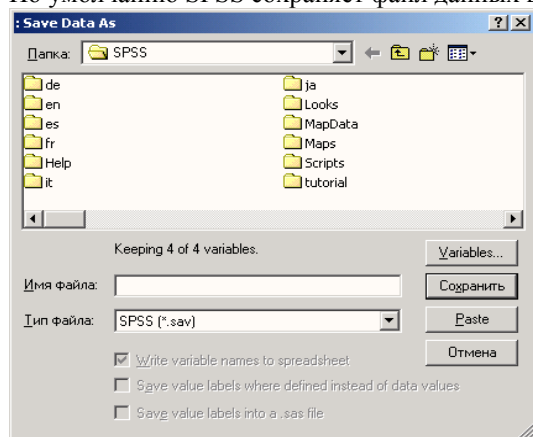


### Сохранение файла данных

Для того, что бы сохранить созданный файл данных поступите следующим образом:

- Выберите в меню команды File (Файл) Save as... (Сохранить как...) Откроется диалоговое окно Save Data as (Сохранить данные как).

По умолчанию SPSS сохраняет файл данных в текущем каталоге с расширением .sav.



Диалоговые окна статистических процедур содержат следующие компоненты:

1. Список исходных переменных — список переменных в файле данных. Например, на рисунке исходных переменных присутствуют следующие переменные: номер анкеты (nom), пол (var1), возраст (var2), партия (var3). Перед именем переменной стоит значок; по которому можно определить, является ли эта переменная численной или текстовой.

2. Список выбранных переменных — список, содержащий переменные файла данных, которые были выбраны для анализа. Список выбранных переменных также называют целевым списком или списком тестируемых переменных. Этот список имеет заголовок Variable(s) (Переменная(ые)). Так как мы еще не выбрали ни одной переменной, этот список пуст.

3. Командные кнопки — кнопки, при щелчке на которые выполняются определенные действия. В этом диалоговом окне расположены кнопки ОК, Paste (Вставить), Reset (Сброс или Отклонить), Cancel (Отмена) и Help (Справка), а также кнопки, открывающие вспомогательные диалоговые окна: Statistics... (Статистика), Charts... (Диаграммы или Графики) и Format... (Формат). Кнопки вспомогательных диалоговых окон отличаются троеточием (...) после названия.

Пять стандартных командных кнопок в главном диалоговом окне имеют следующее назначение:

- ОК — кнопка ОК запускает соответствующую процедуру. Одновременно она закрывает диалоговое окно.
- Paste — эта кнопка переносит выбранный в диалоговом окне синтаксис команды в редактор синтаксиса. Здесь можно отредактировать синтаксис команды и дополнить его другими опциями, недоступными в данном диалоговом окне.
- Reset — эта кнопка отменяет перенос выделенной переменной в целевой список переменных.
- Cancel — эта кнопка отменяет все изменения, сделанные с момента последнего открытия диалогового окна, и закрывает его.
- Help — эта кнопка выводит контекстно-чувствительную справку. При щелчке на ней открывается окно справки, содержащее сведения о текущем диалоговом окне.

### Выбор переменных

Сначала мы построим частотное распределение для переменной var3 (партия). Выполните следующие действия:

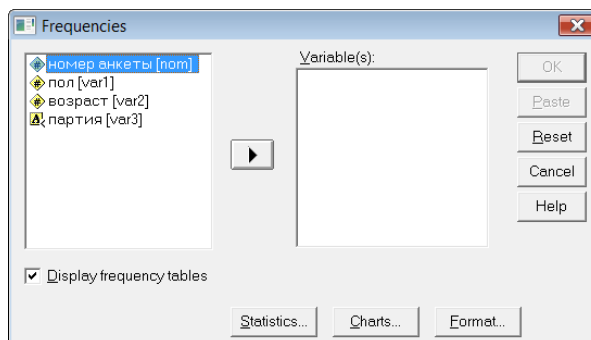
- Выделите переменную var3 (партия) в списке исходных переменных.
- Щелкните на кнопке, которая находится рядом со списком выбранных переменных. Переменная «партия» будет перенесена из списка исходных переменных в список выбранных переменных. Можно также дважды щелкнуть на нужной переменной, и она будет перенесена в список выбранных переменных.
- Подтвердите операцию, щелкнув на кнопке ОК. Результаты будут отображены в окне просмотра (Viewer).

- Задайте имя файла, соответствующее соглашению об именах в DOS. Для рассматриваемого примера мы предлагаем имя файла "opros.sav". Расширение .sav SPSS присваивает файлам данных по умолчанию. Поэтому расширение .sav вводить не обязательно.

### Диалоговое окно процедуры обработки

Каждая процедура обработки имеет собственное диалоговое окно. Несмотря на это, практически все диалоговые окна построены по одному и тому же принципу.

Приведем пример диалогового окна процедуры Frequencies (Частоты).



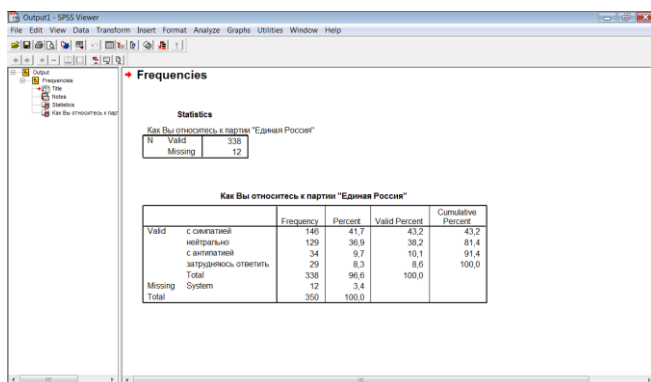
цедур

всех пе-  
сунке в  
следую-  
(var1),  
каждой  
но опре-  
ной или



## Окно просмотра (вывода данных)

Результаты анализа данных отображены в окне Output - SPSS Viewer. Окно просмотра разделено на две части. В левой отображается структура вывода, а в правой — собственно выводимые данные. В разделе вывода отображаются как таблицы, так и графики.



Statistics

N	Valid	Missing
	338	12

Как Вы относитесь к партии "Единая Россия"

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid с сомнением	146	41,7	43,2	43,2
нейтрально	129	38,9	38,2	81,4
с антипатией	34	9,7	10,1	91,4
затрудняюсь ответить	29	8,3	8,6	100,0
Total	338	96,6	100,0	
Missing System	12	3,4		
Total	350	100,0		

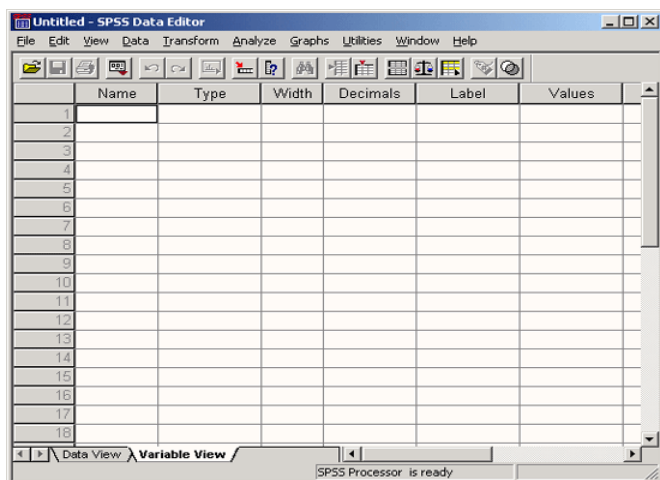
В отличие от файла данных (с расширением .sav) файл вывода данных имеет расширение .spo. Для сохранения файла вывода данных в меню File (Файл) выберите команду Save (Сохранить), в появившемся диалоговом окне задайте имя файла и щелкните по кнопке Save.

Другой способ сохранения окна вывода результатов – копирование в буфер обмена (при помощи правой кнопки мыши) элементов окна вывода данных и последующая вставка в открытый документ Word. При этом доступно два варианта копирования и вставки. При выборе команды Copy Object (Копирование объекта) выбранная таблица будет вставлена в документ Word как рисунок, недоступный для дальнейшего редактирования. Перенос таблицы из SPSS в Word как рисунка (объекта) гарантирует сохранность формата таблицы. А при выборе команды Copy (Копировать) таблица будет вставлена как обычная таблица, доступная для редактирования, но такой перенос может нарушить структуру таблицы.

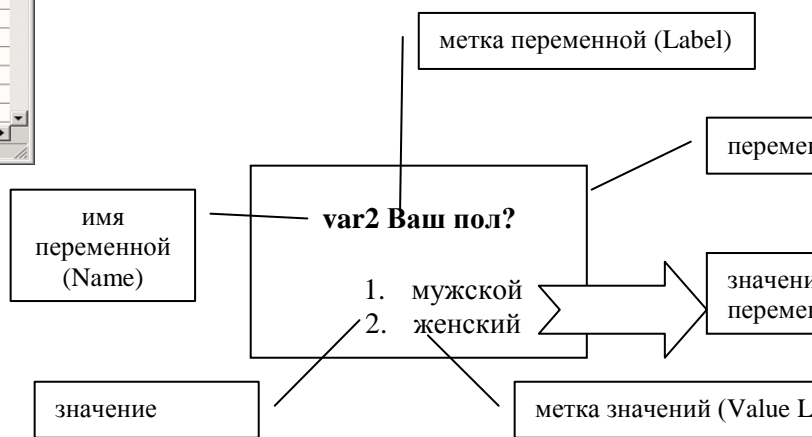
## 2. Создание и редактирование файлов данных

После запуска программы SPSS в открывшемся диалоговом окне редактора данных Data Editor нужно перейти на вкладку Variable View (Просмотр переменных), щелкнув на ее ярлычке мышью (или в редакторе данных дважды щелкните на ячейке с надписью var).

Вкладка Data View (Просмотр данных), которая отображается сразу после запуска редактора, предназначена для ввода значений в создаваемый файл данных.



Вкладка Variable View (Просмотр переменных) позволяет задать структуру файла данных (создать макет данных), то есть определить имена, метки и структуры переменных. Заголовки столбцов представляют собой параметры каждой из переменных: Name (Имя), Type (Тип), Width (Ширина), Decimals (Дробная часть), Label (Метка), Values (Значения), Missing (Пропуски), Columns (Столбцы), Align (Выравнивание), Measure (Измерение).



Имя переменной – Name

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	M...
1	number	Numeric	8	0	номер анкеты	None	None	8	Right	Scale
2	var1	Numeric	8	0	За какую партию Вы бы проголосовали?	None	None	8	Right	Scale
3	var2	Numeric	8	0	Ваш пол	None	None	8	Right	Scale
4	age	Numeric	8	0	Ваш возраст	None	None	8	Right	Scale
5	terr	String	13	0	место проживания	None	None	13	Left	Nomi
6										
7										
8										
9										
10										
11										

Чтобы задать имя переменной, нужно поступить следующим образом:

- Введите в текстовом поле Name (Имя) выбранное имя переменной. В нашем примере мы сначала определим переменную number. Для этого введите в поле Name текст "number". При выборе имени переменной сле-

дует соблюдать определенные правила:

- Имена переменных могут содержать буквы латинского алфавита и цифры. Кроме того, допускаются специальные символы \_ (подчеркивание), . (точка), а также символы @ и #. Не разрешаются, например, пробелы, знаки других алфавитов и специальные символы, такие как !, ?, " и \*.
- Имя переменной должно начинаться с буквы.
- Последний символ имени не может быть точкой или знаком подчеркивания (\_).
- Длина имени переменной не должна превышать восьми символов.
- Имена переменных нечувствительны к регистру, то есть прописные и строчные буквы не различаются.

Примеры допустимых имен переменных:

budget09, gender, zarplata, party, quest\_13, q13, var3\_1\_2, var1.

Чтобы задать имя первой переменной, просто введите его с клавиатуры в текущую ячейку. Имя второй переменной вводится в том же столбце под именем первой, то есть во второй строке, имя третьей переменной – в третьей строке и т.д.

В нашем примере (см. анкету на стр.7) во второй строке будет содержаться имя первой переменной анкеты - var1 для вопроса «За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?», в третьей строке – var2 – «ваш пол», в четвертой строке – age – «ваш возраст», в пятой – terr – «место проживания».

Тип переменной – Variable Type

Параметр Type определяет тип переменной. Текущим типом переменной является тип Numeric (Численный). В большинстве случаев при вводе социологических данных используется числовой тип. В тех редких случаях, когда значения переменных представляют собой буквы или буквосочетания, необходимо установить переключатель String (Строка).

Numeric (Численный)	К допустимым значениям относятся цифры, перед которыми стоит знак плюс или минус и десятичный разделитель. Знак плюс перед числом, в отличие от минуса, не отображается. В текстовом поле Width (Ширина) задается максимальное количество знаков, включая позицию для десятичного разделителя. В текстовом поле Decimals (Десятичные разряды) вводится количество отображаемых знаков дробной части.
String (Строка)	Строка символов. К допустимым значениям относятся: буквы, цифры и специальные символы. Различаются короткие и длинные строковые переменные. Короткие строковые переменные могут содержать не более восьми знаков. В большинстве процедур SPSS применение длинных строковых переменных ограничивается или вообще не допускается.

Как правило, строковые переменные не подлежат обработке. Поэтому их следует избегать, за исключением редких случаев, когда данная переменная содержит имена людей или названия городов.

- Если требуется изменить тип переменной, щелкните в ячейке на кнопке с тремя точками: ...  
Откроется диалоговое окно Define Variable Type (Определение типа переменной).

The screenshot shows the 'Variable Type' dialog box. The 'Numeric' radio button is selected. The 'Width' field contains the value 8, and the 'Decimal Places' field contains the value 2. The 'OK', 'Cancel', and 'Help' buttons are on the right side of the dialog.

The screenshot shows the 'Variable Type' dialog box. The 'String' radio button is selected. The 'Characters' field contains the value 8. The 'OK', 'Cancel', and 'Help' buttons are on the right side of the dialog.

Пример №1: вопрос: «Ваш пол», имя переменной - var2, тип переменной числовой – Numeric, так как коды значений переменной – целые числа – 1,2 (1- мужской, 2 – женский), то в текстовом поле Decimal Places устанавливается значение – 0 (вместо 2 по умолчанию). В текстовом поле Width (Ширина) оставляется максимальное значение – 8 по умолчанию.


Пример №2: вопрос: «В каком населенном пункте Вы проживаете?», имя переменной terr, тип переменной строковый тип - String, так как респонденты словами записывали название населенного пункта. Длину – в текстовом поле Characters можно выставить 13 символов, так как самое длинное слово, по примеру нашей анкеты – «Нижевартовск» - содержит 13 букв.

С такими переменными нельзя выполнять никаких вычислительных операций, но можно проводить, например, подсчеты повторяемости.

- Нажмите клавишу <Tab>, чтобы перейти к установке формата столбца.


## Формат столбца, ширина (Width)

Параметр Width (ширина) позволяет задать максимальное количество знаков, которое может иметь значение переменной, включая дробную часть. По умолчанию задана ширина – 8 знаков, в большинстве случаев нет необходимости менять заданную ширину переменной.

- В нашем примере для переменной var1 («За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?») можно задать число позиций в столбце, равное "1". Так как коды значений (1 – для варианта «Единая Россия», 2 – для варианта «Справедливая Россия» и т.д.) – ширина переменной не более 1 знака. Если число вариантов ответов в вопросе было  $\geq 10$ , то ширина переменной можно было бы определить в 2 знака. Но для экономии времени можно оставить ширину «8», заданную по умолчанию.
  - Для переменной «terr» нужно задать ширину  $\geq 13$ , для того, что бы поместились названия населенных пунктов, в том числе самое длинное название города «Нижевартовск».
  - Чтобы изменить этот формат представления переменной, перенесенный из диалога Define Variable Type, щелкните на кнопке лифта: 
  - В этом случае выбранное значение ширины подтверждается клавишей <Tab>.
- Но обычно, для экономии времени, величину ширины переменной по умолчанию («8») не меняют.

## Десятичные разряды, дробная часть (Decimals)

Параметр Decimals (Дробная часть) предназначен для задания числа десятичных знаков после запятой в случае, если тип переменной допускает использование дробных чисел. Для строковых переменных значение в ячейке Decimals (Дробная часть) автоматически устанавливается равным нулю, а для числовых переменных – равным 2.

- Например, так как переменная «terr» - «место проживания» является строковой, для нее задано количество десятичных разрядов "0". Увеличение или уменьшение этого значения, определенного настройкой в диалоге Define Variable Type, также производится при помощи кнопки лифта:  Подтвердите значение "0", нажав клавишу <Tab>.

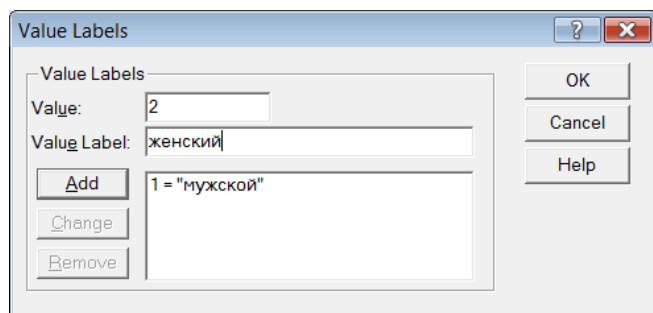
## Метка переменной (Label)

Метка переменной — это название, позволяющая описать переменную более подробно. Метка переменной может содержать до 256 символов. В метках переменных различаются прописные и строчные буквы. Они отображаются в том виде, в каком были введены.

В нашем примере для переменной number введите в качестве метки в поле Variable label текст "Номер анкеты". Для переменной var1 - метка переменной будет: «За какую партию Вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?».

## Метки значений (Values)

Метки значений — это название, позволяющее более подробно описать возможные значения переменной. Так, например, в случае переменной var2 – «пол» - можно задать метку "мужской" для значения "1" и метку "женский" для значения "2". Подтвердите настройку по умолчанию None (Нет) клавишей <Tab>. Впрочем, ввод данных также можно подтвердить клавишей <Enter>.



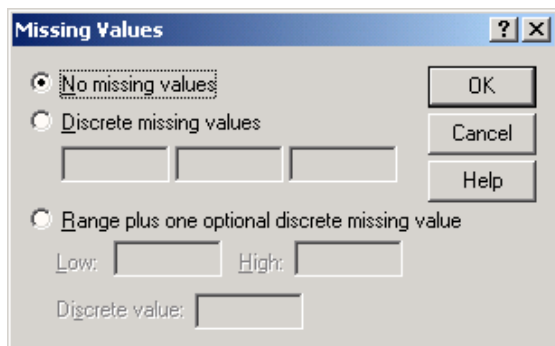
Метки значений определяются следующим образом:


- Вначале введите в поле Value (Значение) число "1". Нажмите клавишу <Tab>.
  - Введите в поле Value label (Метка значения) текст "мужской".
  - Щелкните на кнопке Add (Добавить). Метка значения будет добавлена в список. Для этой цели можно также нажать комбинацию клавиш <Alt>+<h>.
  - Повторите эти действия для значений "2" — "женский".
- Максимально допустимая длина метки значения составляет 60 знаков.
- Подтвердите введенные данные кнопкой ОК, а затем — клавишей <Tab>.

## Пропущенные значения (Missing values)

Параметр Missing values (Пропущенные значения) используется очень редко, поскольку программа и так позволяет учитывать пропуски в данных. Необходимость в этом параметре возникает, когда требуется различать причины пропусков значений. Например, пропуск в данных может быть обусловлен тем, что респондент еще не был опрошен, а может быть, он отказался отвечать на данный вопрос.

Поэтому можно для еще не опрошенных оставлять пустую ячейку, при вводе данных, а для не определившихся можно обозначить кодом «9». Если ввести значение «0» в столбец Missing values, то оно не будет использоваться в дальнейшем при обработке наряду с пустыми ячейками.



- Чтобы задать пропущенные значения, щелкните в поле Missing на кнопке с тремя точками . Откроется диалоговое окно Define Missing Values (Определение пропущенных значений).
- По умолчанию предлагается вариант No missing values (Нет пропущенных значений), то есть все значения в настоящее время рассматриваются как допустимые. Подтвердите настройку по умолчанию None (Нет) клавишей <Enter>.
- Щелкните на пункте Discrete missing values (Отдельные пропущенные значения). Для одной переменной нужно задать до трех пользовательских пропущенных значений. Введите значение "9".

Если в матрице данных есть незаполненные численные ячейки, система SPSS самостоятельно идентифицирует их как пропущенные значения. Этот факт отображается в матрице данных с помощью запятой (,).

## Столбцы (Columns)

Поле Columns определяет ширину, которую будет иметь в таблице данный столбец при отображении значений. Ширину столбца также можно изменить непосредственно в окне редактора данных. Для этого поместите указатель мыши на разделитель между двумя заголовками столбцов с именами переменных. Вид указателя изменится. Появившаяся двойная стрелка указывает, что соответствующий столбец можно расширить или сузить путем перетаскивания.

- Подтвердите настройку по умолчанию "8" клавишей <Enter>.

## Выравнивание (Alignment)

Здесь можно задать вид выравнивания значений, т.е. определить, как они будут отображаться в таблице. Возможные виды выравнивания — "Right" (по правому краю), "Left" (по левому краю) и "Center" (по центру).

Чтобы задать вид выравнивания, щелкните на кнопке .

- Подтвердите настройку по умолчанию Right клавишей <Enter>.

## Шкала измерения (Measure)

Здесь можно задать шкалу переменной, которая может быть номинальной (шкала наименований), порядковой или метрической. По умолчанию принимается метрическая шкала измерения. Правда, это различие имеет значение только при создании интерактивных графиков, где номинальная и порядковая шкала измерений объединяются в "категориальный" тип.

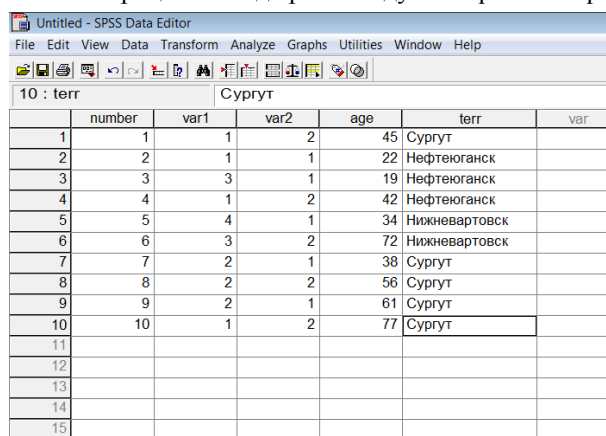
Если вы загружаете файлы, созданные в предыдущих версиях SPSS, или шкала измерений не определяется явно, SPSS вначале автоматически предполагает метрическую шкалу. Однако если соответствующая переменная имеет метки значений или принимает менее 24 различных значений, то задается порядковая шкала.

- Подтвердите настройку по умолчанию Nominal (шкала наименований) клавишей <Tab>. Затем снова поместите курсор в поле Name, чтобы начать объявление следующей переменной.

**Задание.** Создать макет для разработанной анкеты, определив основные параметры переменных. Сохранить файл.

## 2.1. Ввод данных

В процессе ввода рекомендуется время от времени производить сохранение файла во избежание случай-



	number	var1	var2	age	terr	var
1	1	1	2	45	Сургут	
2	2	1	1	22	Нефтеюганск	
3	3	3	1	19	Нефтеюганск	
4	4	1	2	42	Нефтеюганск	
5	5	4	1	34	Нижневартовск	
6	6	3	2	72	Нижневартовск	
7	7	2	1	38	Сургут	
8	8	2	2	56	Сургут	
9	9	2	1	61	Сургут	
10	10	1	2	77	Сургут	
11						
12						
13						
14						
15						

ной порчи или утери введенных данных. Перед вводом данных следует перейти на выкладку Data View (просмотр, редактор данных).

Данные можно вводить по отдельным наблюдениям (строкам) или по отдельным переменным (столбцам). Действуйте следующим образом:

- Щелкните на ячейке в левом верхнем углу. Это будет переменная с именем: number – номер анкеты. Вокруг ячейки появится рамка. Таким образом, эта ячейка обозначается как активная.
- Введите значение, в нашем примере это "1" (анкета №1) Это значение отобразится в редакторе ячеек в верхней части окна редактора данных.
- Нажмите клавишу <Tab>. Значение из редактора ячеек отобразится в ячейке.

В следующих таблицах показано, каким клавишам в редакторе данных соответствует какая функция. Здесь, как и далее, мы предполагаем, что активизирована таблица просмотра данных.

### Позиционирование

Клавиша	Функция
<Tab> или <стрелка вправо>	Перемещает курсор на ячейку вправо.
<Enter> или <стрелка вниз>	Перемещает курсор на ячейку вниз.
<стрелка вверх>	Перемещает курсор на ячейку вверх.
<Shift> <Tab> или <стрелка влево>	Перемещает курсор на ячейку влево, т.е. в предыдущее поле.
<Home>	Перемещает курсор в первую ячейку строки или случая.
<End>	Перемещает курсор в последнюю ячейку случая.
<Ctrl> <стрелка вверх>	Перемещает курсор в первый случай столбца.
<Ctrl> <стрелка вниз>	Перемещает курсор в последний случай столбца.
<Ctrl> <Home>	Перемещает курсор в первую ячейку первого случая.
<Ctrl> <End>	Перемещает курсор в последнюю ячейку последнего случая.
<Page Up>	Прокручивает таблицу на одну страницу вверх.
<Page Down>	Прокручивает таблицу на одну страницу вниз.

### Выделение

<Shift> <пробел>	Выделяет всю строку.
<Ctrl> <пробел>	Выделяет весь столбец.
<Shift> <клавиши со стрелками>	Выделение области случаев и переменных. Также можно щелкнуть мышью и перетянуть ее из верхнего левого угла области в нижний правый угол.

### Редактирование

F2	Переключает в режим редактирования. Следующее нажатие <F2> отключает режим редактирования.
<стрелка вправо>	Переместить позицию редактирования в ячейке вправо на один знак.
<стрелка влево>	Переместить позицию редактирования в ячейке влево на один знак.
<Home>	Перейти в начало значения ячейки.
<End>	Перейти в конец значения ячейки.

**Вставка нового объекта.** Если необходимо вставить новый объект (строку) между двумя соседними строками, щелкните сначала на нижней из них, а затем на кнопке Insert Cases (Вставка объектов) панели инструментов. В результате будет создана пустая строка, а номера строк, находящиеся ниже, увеличатся на единицу.

**Вставка новой переменной.** Чтобы вставить новую переменную между двумя соседними, щелкните сначала на правой из них, а затем – на кнопке *Inset Variable* (вставка переменной). Будет создан пустой столбец, а все переменные, находящиеся справа, окажутся сдвинутыми на один столбец.

**Поиск данных.** Очень удобным вспомогательным средством работы с данными является функция поиска. В меню *Edit* (Редактирование) выберите команду *Find* (поиск), или щелкните на кнопке *Find* (Поиск) на панели инструментов. На экране появится диалоговое окно *Find Data in Variable <name>* (поиск данных в переменной), с помощью которой можно найти заданное слово или значение. Таким образом, можно в больших файлах данных обнаружить недопустимые или неверные значения какой-либо из переменных.

**Задание.** Ввести данные анкет (единиц анализа) в созданный макет файла. Сохранить файл.

### 3. Управление данными

После того, как создана матрица данных, практически всегда существует необходимость в предварительной подготовке и преобразовании исходных данных. В процессе работы могут понадобиться агрегированные данные, то есть данные являющиеся результатом некоторых действий над исходными данными файла. Иногда желательно упорядочить данные файла по какому-либо признаку, например, по результатам выполнения какого-либо задания. Нередко возникает необходимость обработки не всех данных, а лишь их подмножества, выделяемого по определенными критериям.

Основные команды управления данными:

- В меню File - команда Display Data File Information (Показать информацию о файле) позволяет получить сведения о переменных как открытого, так и любого внешнего файла данных SPSS: имена, метки имен и значений;
- В меню Analyze – команда Reports – Case Summaries (Отчеты – Сводка по данным) предназначена для проверки состава и качества данных;
- Команда Transform – Replace Missing Values (Преобразование – Заменить пропущенные значения) работает с отсутствующими значениями переменных;
- Команда Transform – Compute (Преобразование – Вычислить) позволяет путем вычислений создавать новые переменные на основе существующих;
- Команда Transform – Rank Cases (Преобразование – Ранжировать объекты) создать новую переменную путем ранжирования значений существующей переменной;
- Команды подменю Recode (Перекодирование) меню Transform (Преобразование) предназначены для изменения способа кодирования переменных, например, для уменьшения числа возможных значений;
- С помощью команды Data – Select Cases (Данные – Выбор объектов) можно выбрать подмножество объектов для дальнейшего анализа;
- Команда Data – Sort Cases (Данные – Сортировка объектов) позволяет упорядочивать объекты в соответствии с назначенными критериями;
- Команды подменю Merge Files (Слияние файлов) меню Data используются для добавления в файл новых переменных или объектов из другого файла.

Более подробно с этими командами управления данными можно познакомиться в работах: Бююль А., Цёфель П. «SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей», Наследов А. «SPSS 15 профессиональный статистический анализ данных».

Мы же рассмотрим наиболее необходимые при анализе социологической информации команды управления данными.

#### 3.1. Выбор объектов для анализа

Команда Select Cases (Выбор объектов) позволяет пользователю выбирать для обработки не все, а часть данных, удовлетворяющих заданным условиям.

Пример. В анкете содержатся два вопроса:

**var1. Принимали ли вы участие в голосовании на прошлых выборах в городскую думу?**

1. участвовал
2. не участвовал
3. не помню
4. затрудняюсь ответить

**var2. Будете ли вы участвовать в голосовании на будущих выборах в городскую думу?**

1. да, обязательно
2. скорее всего, да
3. скорее всего, нет
4. нет
5. еще не решил
6. затрудняюсь ответить

В результате проведенного анализа, были получены следующие частотные распределения:

**участие в прошлых выборах**



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	участвовал	316	60,7	60,8	60,8
	не участвовал	162	31,1	31,2	91,9
	не помню	33	6,3	6,3	98,3
	затрудняюсь	9	1,7	1,7	100,0
	Total	520	99,8	100,0	
Missing	System	1	,2		
Total		521	100,0		

В прошлых выборах из всех опрошенных респондентов (521 человек) участвовали в выборах 60,7% , или 316 человек, не участвовали 162 опрошенных респондента или 31,2% от общего числа.

#### участие в будущих выборах

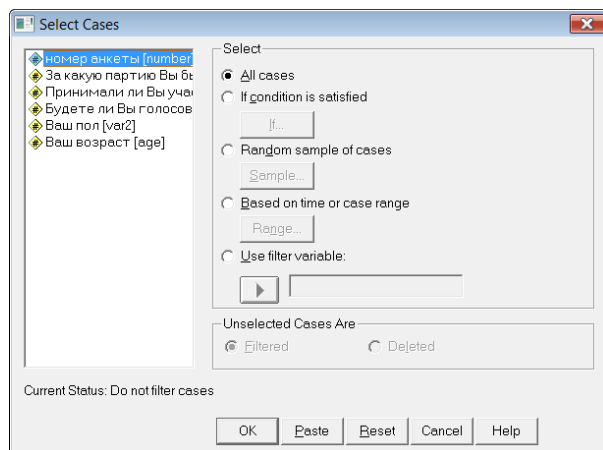
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	обязательно да	219	42,0	42,2	42,2
	скорее всего да	142	27,3	27,4	69,6
	скорее всего нет	19	3,6	3,7	73,2
	нет	44	8,4	8,5	81,7
	не решил	78	15,0	15,0	96,7
	затрудняюсь	17	3,3	3,3	100,0
	Total	519	99,6	100,0	
Missing	System	2	,4		
Total		521	100,0		

Из числа всех опрошенных респондентов (521 человек) обязательно пойдут на будущие выборы 42,0%, еще не решили – 15,0%.

Необходимо выяснить - сколько человек, не участвовавших в прошлых выборах, будут голосовать на будущих выборах.

Для решения этой задачи, необходимо выделить (отфильтровать) данные тех респондентов, кто на вопрос «Участвовали ли Вы в прошлых выборах?» выбрал вариант 2: «не участвовал».

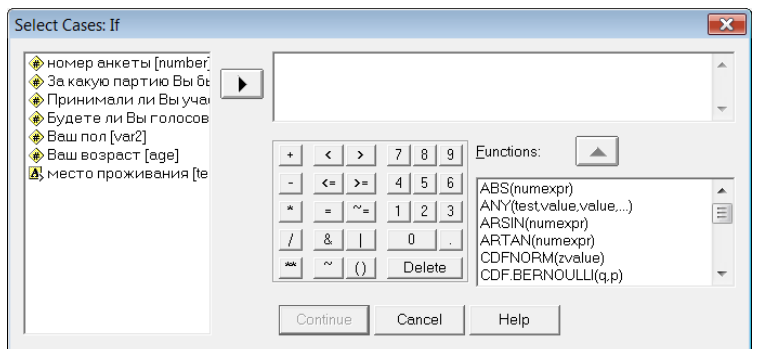
Нужно поступить следующим образом:



- Выбрать в меню команды Data (Данные) Select Cases... (Выбрать наблюдения)

По умолчанию в этом диалоге выбран пункт All cases (Все наблюдения).

- Выбрать пункт If condition is satisfied (Если выполняется условие) и щелкнуть на кнопке If... (Если). Откроется диалоговое окно Select Cases: If



Это диалоговое окно разделено на следующие части:

- Список исходных переменных: Содержит переменные, содержащиеся в открытом файле данных. В нашем случае должны быть переменные «участие в прошлых выборах» (var1.a) и «участие в будущих выборах» (var1.b).
- Редактор условий: Здесь записывается логическое выражение, по которому должны быть отобраны наблюдений. В данный момент редактор условий пока пуст.
- Кнопка с треугольником: Эта кнопка позволяет перенести переменную из списка исходных переменных в редактор условий.
- Клавиатура: Содержит цифры, а также арифметические, логические операторы и операторы отношения; с ней можно работать как с обыкновенным калькулятором. Если щелкнуть на какой-нибудь кнопке мышью, соответствующий знак, например, +, \*, 7, будет скопирован в редактор условий.

- Список функций: Содержит около 140 функций. Каждую из функции можно скопировать в редактор условий двойным щелчком.

Для того, что бы отобрать тех респондентов, кто не участвовал в прошлых выборах, нужно в диалоговом окне Select Cases: If выделить переменную var1.a, с помощью кнопки с треугольником переместить ее в редактор условий (пустое поле справа) и задать условия с помощью клавиатуры: var1.a = 2 (где, 2 – значение варианта «не участвовал»). После создания условий щелкнуть на кнопке Continue (Продолжение), что бы закрыть первое окно и на кнопке ОК, чтобы закрыть второе диалоговое окно и вернуться в окно редактора данных. При этом в окне редактора данных появится новый столбец filter\_\$, где отобразятся отфильтрованные единицы анализа (значение 1).

После выполнения этого шага при любой обработке будут учитываться только данные для респондентов, которые не участвовали в прошлых выборах. Чтобы сделать доступными все данные, достаточно в окне Select Cases установить переключатель All cases (Все объекты).

После создания фильтра (отбора нужных данных) можно вычислить частотные распределения (подробнее об этом в разделе 4 данного пособия).

В нашем примере, не участвовали в прошлых выборах 162 респондента. Из них обязательно будут голосовать на будущих выборах 20,4%, еще не решили – 21,0%.

**участие в будущих выборах**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	обязательно да	33	20,4	20,4	20,4
	скорее всего да	42	25,9	25,9	46,3
	скорее всего нет	12	7,4	7,4	53,7
	нет	32	19,8	19,8	73,5
	не решил	34	21,0	21,0	94,4
	затрудняюсь	9	5,6	5,6	100,0
Total		162	100,0	100,0	

### 3.2 Перекодировка в новую переменную

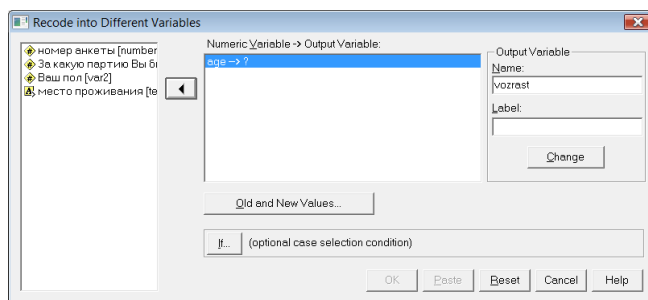
Команда Transform – Recode – Into Different Variables (Преобразование – Перекодировка – в другие переменные) создает новую переменную, ее значения определяются на основе замены множества значений существующей переменной небольшим числом категорий.

Например, на вопрос «Ваш возраст», респонденты выбирали не варианты ответа, а указывали свой возраст в цифровом (натуральном) значении. При вычислении частотных распределений переменной «возраст» получается множество значений. Для того, чтобы последующий анализ сделать более удобным необходимо перекодировать переменную «возраст» в новую переменную, где значения будут сгруппированы в три категории:

1. от 18 до 29 лет,
2. от 30 до 50 лет,
- 3 старше 50 лет.

возраст					
	Frequency	Percent	Valid Percent	Cumulative Percent	
Valid	18	28	5,4	5,4	5,4
	19	10	1,9	1,9	7,3
	20	8	1,5	1,5	8,8
	21	5	1,0	1,0	9,8
	22	5	1,0	1,0	10,8
	23	6	1,2	1,2	11,9
	24	6	1,2	1,2	13,1
	25	14	2,7	2,7	15,8
	26	6	1,2	1,2	16,9
	27	5	1,0	1,0	17,9
	28	11	2,1	2,1	20,0
	29	19	3,6	3,7	23,7
	30	14	2,7	2,7	26,3
	31	10	1,9	1,9	28,3
	32	15	2,9	2,9	31,2
	33	19	3,6	3,7	34,8
	34	12	2,3	2,3	37,1
	35	14	2,7	2,7	39,8
	36	12	2,3	2,3	42,1
	37	19	3,6	3,7	45,8
	38	9	1,7	1,7	47,5
	39	24	4,6	4,6	52,1
	40	30	5,8	5,8	57,9
	41	19	3,6	3,7	61,5
	42	21	4,0	4,0	65,6
	43	12	2,3	2,3	67,9
	44	11	2,1	2,1	70,0
	45	12	2,3	2,3	72,3
	46	7	1,3	1,3	73,7
	47	10	1,9	1,9	75,6
	48	11	2,1	2,1	77,7
	49	21	4,0	4,0	81,7
	50	15	2,9	2,9	84,6
	51	5	1,0	1,0	85,6
	52	15	2,9	2,9	88,5
	53	10	1,9	1,9	90,4
	54	3	,6	,6	91,0
	55	2	,4	,4	91,3
	56	5	1,0	1,0	92,3
	57	3	,6	,6	92,9
	58	7	1,3	1,3	94,2
	59	5	1,0	1,0	95,2
	60	7	1,3	1,3	96,5
	61	4	,8	,8	97,3
	62	2	,4	,4	97,7
	63	2	,4	,4	98,1
	64	1	,2	,2	98,3
	65	4	,8	,8	99,0
	67	1	,2	,2	99,2
	68	2	,4	,4	99,6
	69	1	,2	,2	99,8
	70	1	,2	,2	100,0
Total	520	99,8	100,0		
Missing System	1	,2			
Total	521	100,0			

Команда перекодировки выполняется с помощью диалогового окна:



Нужно выделить «старую» переменную – «age» и с помощью кнопки с треугольником перенести в правое пустое поле – список Input Variable – Output Variable (Входная переменная – Выходная переменная). Затем в поле Name в области Output Variable ввести имя новой переменной (в данном случае vozrast). Щелчок на Change приведет к появлению переменной vozrast в предыдущем списке: его содержимое будет иметь вид age-vozrast.

(Старое  
вой  
ство-  
мен-  
ные  
хо-  
го  
сти

Щелчок на Old and New Values (Старые и новые величины) вызовет диалоговое окно:

В нем можно задать градации новой переменной, которые будут соответствовать диапазонам уровней старой переменной.

Правое подокно - Old Values (Старые величины) содержит следующие необходимые нам переключатели:

- Value (Значение) при установке этого переключателя нужно в поле рядом ввести значение. Например, если нам нужно задать отдельную категорию для респондентов в возрасте 18 лет, то в поле нужно поставить – 18.
- Range (Диапазон) этому переключателю соответствуют два окна, позволяющие задать верхнюю и нижнюю границы диапазона значений. Например, для возрастной категории – от 19 до 29 лет, в первое поле нужно поставить цифру 19, во второе – 29.
- Range lowest through (Диапазон от наименьшего до заданного значения). Например, для возрастной категории – моложе 18 лет, в поле нужно поставить цифру 18, и тогда в новой категории переменной будут учитываться респонденты, чей возраст не превышает 18 лет.
- Range through highest (Диапазон от наибольшего до заданного значения). В нашем примере, для учета респондентов старше 50 лет, в поле ставится цифра 50.
- All other values (Все другие значения) Позволяет присвоить новое значение всем остальным величинам исходной переменной.

В левом подокне – New Value, устанавливаются новые значения новой переменной.

Последовательность действий в нашем примере будет следующая. Возрасту от 18 до 29 лет, установленному в поле Old Value – Range, в диапазоне – от 18 до 29 лет, в окне New Value, в поле Value будет соответствовать цифра 1. Для перекодировки, нужно щелкнуть на кнопке Add (Добавить) и новое перекодированное значение появится в окне Old – New в виде строки:

18 thru 29 → 1

Возрасту от 30 лет до 50 лет соответствует новое значение – 2, в окне Old – New:

30 thru 50 → 2

Возрасту старше 50 лет, установленному в поле Range through highest, соответствует новое значение – 3, в окне Old – New:

50 thru Highest → 3.

После создания новых значений переменных нужно щелкнуть на кнопке Continue (Продолжение), затем в окне Recode Into Different Variables – на кнопке OK.

В окне редактора данных появится новая переменная с перекодированными данными старой переменной, для нее можно назначить метки, вручную изменять значения. С новой переменной можно производить статистические процедуры, например, вычислять частотные распределения.

В нашем примере, можно получить данные по возрасту опрошенных респондентов, сгруппированных в 3 возрастных категории.

		vozrast			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	от 18 до 29 лет	123	23,6	23,7	23,7
	от 30 до 50 лет	302	58,0	58,1	81,7
	старше 50 лет	95	18,2	18,3	100,0
	Total	520	99,8	100,0	
Missing	System	1	,2		
Total		521	100,0		

- Задание.** 1. Выяснить, как оценивают молодые респонденты, негативно относящиеся к политической деятельности, возможности молодежи повлиять на власть. Сравнить эти оценки с оценками тех респондентов, кого интересует политика и кто намерен ею заниматься. По массиву данных файла `orgos.sav` работать с переменными «Как Вы относитесь к политической деятельности?» (`var11`), «Согласны ли вы с тем, что у молодых людей нет возможности повлиять на власть» (`var13`). Проанализировать полученные данные.
2. По массиву данных файла `orgos.sav` перекодировать переменную «Ваш доход в месяц» (`var45`) в переменную «доход» со сгруппированными значениями (интервальной шкалой). Такую же операцию совершить с переменной «Сколько Вам лет» (`var62`), создав переменную «`vozrast`».
3. Исходя из задач собственного исследования, определить категории (часть данных), удовлетворяющим условиям задачи, и провести вычисления по выделенной (отфильтрованной) части данных.
4. Исходя из задач собственного исследования, определить переменную, сгруппировать значения переменной, перекодировав в новую переменную.

#### 4. Одномерный описательный анализ социологических данных. Построение частотных (линейных) распределений

Анализ частотных распределений результатов количественного социологического исследования – это первый шаг при обработке социологической информации. Первый шаг одномерного описательного анализа для объяснения какого-либо явления – его описание. Результаты любого массового опроса содержат ответы большого числа респондентов на широкий круг анкетных вопросов. Даже в рамках одного вопроса анкеты объем исходной информации достаточно велик. В матрице данных ответы представлены в виде числовых кодов. Поскольку полностью матрица содержит множество ответов респондентов, а объем выборки достаточно часто превышает 1500 и 2000 респондентов, просто просмотр ответов всех опрошенных либо на экране компьютера, либо в распечатанном виде не дает возможности осмыслить такой массив информации.

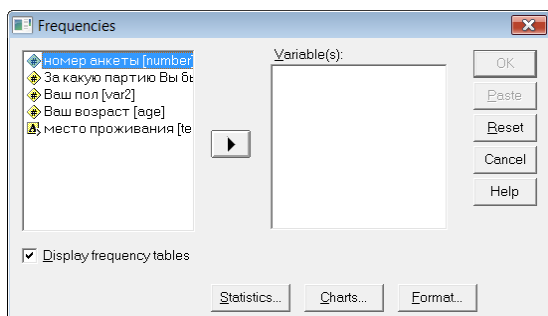
В этом случае методы одномерного описательного анализа решают задачу сжатия исходной информации, ее компактного представления. Как правило, в процессе исследования бывает важно получить совокупные характеристики отдельных предметов через призму какого-либо конкретного свойства. Вместо большого числа отдельных показателей нам требуется одно значение, которое было бы типичным (репрезентативным) для всей совокупности объектов. Принадлежность к какой социальной или возрастной группе наиболее типична для членов определенной партии? Сколько раз в среднем в месяц студенты смотрят общественно-политические передачи? Ответы на эти вопросы дает анализ одномерных (частотных) распределений, в частности подсчет средних величин для разных уровней измерения. Анализ одномерных распределений позволяет заодно установить, насколько типичное значение в действительности типично, репрезентативно по отношению к совокупности данных.

В одномерном описательном анализе используются методы:

- Построения частотных распределений;
- Графического представления поведения анализируемой переменной;
- Получения статистических характеристик распределения анализируемой переменной.

##### 4.1. Частоты

Команда Frequencies (Частоты) являются одной из самых простых и часто используемых команд SPSS. Действие команды сводится к подсчету количества объектов в каждой категории переменной. Это и называется распределением частот по категориям переменной.



Для создания частотных распределений в меню Analyze (Анализ) нужно выбрать команду Descriptive Statistic (Описательные статистики), затем Frequencies (Частоты). Появится диалоговое окно.

В левой части окна расположен список всех доступных переменных. В нем необходимо выбрать те переменные, для которых необходимо вычислить распределение частот. Для этого щелчком выделяется нужная переменная и с помощью кнопки с треугольником перемещается в целевой список Variable(s) (Переменные).

Если необходимо удалить переменную из целевого списка, достаточно выделить ее в нем, затем воспользоваться кнопкой с направленной влево стрелкой, переменная вновь переместится в исходный список. Чтобы полностью очистить целевой список, можно щелкнуть на кнопке Reset (Сброс).

После создания целевого списка, для получения частотных распределений, нужно щелкнуть на кнопке OK. Программа SPSS сформирует окно вывода с результатами выполнения команды.

Пример частотного распределения вопроса: «За какую партию Вы голосовали бы в ближайшее воскресенье?» (опрос проводился в 2006 г.)

**За какую из партий вы проголосовали в ближайшее воскресенье?**

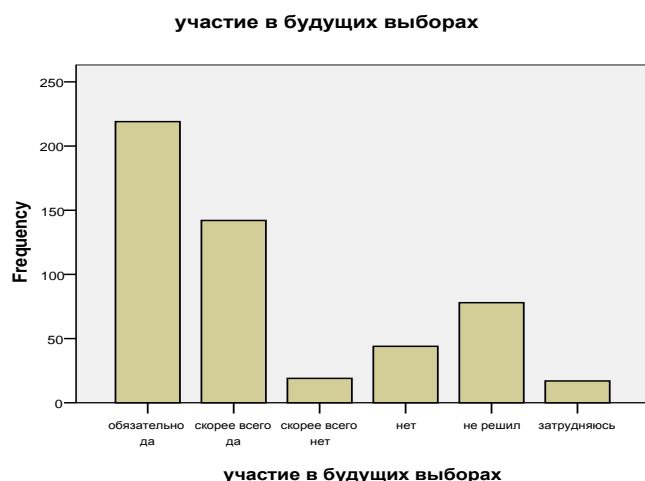
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Союз правых сил	5	1,5	1,6	1,6
	Яблоко	5	1,5	1,6	3,2
	Родина	8	2,5	2,5	5,7
	ЛДПР	37	11,5	11,7	17,4
	Единая Россия	118	36,5	37,3	54,7
	КПРФ	15	4,6	4,7	59,5
	против всех	38	11,8	12,0	71,5
	не стал бы участвовать в выборах	45	13,9	14,2	85,8
	затрудняюсь ответить	45	13,9	14,2	100,0
	Total	316	97,8	100,0	
Missing	System	7	2,2		
Total		323	100,0		

Интерпретация данных таблицы частотных распределений по вопросу: «За какую партию Вы проголосовали бы в ближайшее воскресенье?» В опросе принял участие 316 респондентов (по строке Total), из них 7 респондентов или 2,2% из общего числа не ответили на поставленный вопрос. Из тех респондентов, кто ответил на вопрос анкеты, большинство – 37,3% опрошенных - проголосовали бы за «Единую Россию», на втором месте – респонденты с протестным голосованием – «против всех» проголосовали бы 12,0%, на третьем – приверженцы партии ЛДПР – 11,7%. Достаточно много респондентов – 14,2% - заявили, что они не стали бы участвовать в выборах, и столько же затруднились с ответом.

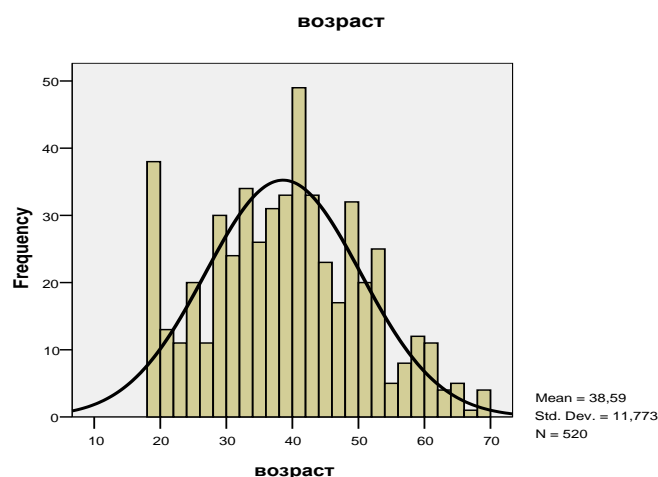
Ниже дана трактовка терминов, используемых программой в окне вывода данных.

- Frequency (Частота) – число объектов, соответствующих каждой категории (градации) переменной (число респондентов, выбравших соответствующий вариант ответа)
- Percent (Процент) – процент от общей численности (с учетом пропусков). Если в файле есть пропущенные значения, то их процент указан в предпоследней строке Missing System.
- Valid percent (Валидный процент) – процент значений для каждой категории за вычетом пропущенных значений.
- Cumulative percent (Кумулятивный процент) – накопленный процент величины Valid percent.
- Valid (Валидные значения) – список градаций (значений) переменной.
- Total (Итого) – итоговые значения.

**Столбиковые диаграммы.** Для того, чтобы создать столбиковую диаграмму для дискретных данных (например, распределение респондентов по полу, предпочтений в выборе партий) необходимо в диалоговом окне Frequencies (Частоты) щелкнуть на кнопке Charts (Диаграммы) и выбрать тип диаграммы с помощью переключателей Bar charts (Столбиковая), Pie charts (Круговая), Histograms (Гистограмма). В зависимости от величины, которую нужно использовать для отображения частот, в группе Chart Values (Значения в диаграмме) устанавливается переключатель Frequencies (Частоты) Percentages (Проценты). Для закрытия диалогового окна нужно щелкнуть на кнопке Continue (Продолжить). Для завершения операции в диалоговом окне Frequencies щелкнуть на кнопке ОК. После этого программа сгенерирует диаграмму, соответствующую выбранной переменной. Созданные диаграммы можно просмотреть в окне вывода, просмотра данных.



**Гистограммы.** Используются для отображения распределения частот непрерывных переменных (например, переменная возраста, или переменные отражающие среднюю отметку учащегося и т.д.). Для построения гистограммы в диалоговом окне Frequencies (Частоты) щелкнуть на кнопке Charts (Диаграммы), выбрать тип диаграммы - Histograms (Гистограмма). Если необходимо установить флажок With normal curve (с нормальной кривой), щелкнуть на кнопке Continue (Продолжить), вернуться в окно Frequencies. Затем сбросить флажок Display frequencies tables (показывать таблицы частот) и щелкнуть на кнопке ОК. Справа от гистограммы помещены вычисленные параметры: среднее значение (Mean), стандартное отклонение распределения (Std.Dev), а также общее число объектов (N).



**Задание. 1.** По массиву данных opros.sav вычислить частотные (линейные) распределения вопросов: «Играет ли молодежь заметную роль в общественной жизни города?», «Как вы относитесь к политической деятельности?», «Удовлетворены ли вы уровнем своего образования?». Построить диаграммы. Проанализировать полученные данные.

**2.** Исходя из задач собственного исследования, создать линейные распределения для переменных анкеты. Построить диаграммы. Проанализировать полученные данные.

## 4.2 Описательные статистики

*Описательные статистики* – это различные вычисляемые показатели, характеризующие распределение значений переменной. Их можно разбить на несколько групп<sup>5</sup>. Первая группа – меры центральной тенденции, вокруг которых «группируются» данные: среднее значение, медиана, мода. Вторая группа характеризует изменчивость значений переменной относительно среднего: среднее отклонение и дисперсия. Диапазон изменчивости характеризуется минимумом, максимумом и размахом. Ассиметрия и эксцесс представляют меру отклонения формы распределения от нормального вида. При помощи команды Descriptives (Описательные статистики) можно вычислить любую из указанных величин.

**Меры центральной тенденции** – характеристики, предназначенные для описания центра распределения.

- Среднее арифметическое значение (mean) равно сумме всех значений распределения, деленной на их количество. Для распределения [3 5 7 5 6 8 9] среднее значение равно  $(3+5+7+5+6+8+9)/7=6,14$
- Медиана (median) определяется как значение, находящееся в середине распределения, полученного из исходного путем упорядочивания по возрастанию. Для распределения [3 5 7 5 6 8 9] медиана равна 6, поскольку значение, равное 6 находится в центре последовательности [3 5 5 6 7 8 9].
- Мода (mode) равна наиболее часто встречающемуся значению. В распределении [3 5 7 5 6 8 9] мода равна 5, поскольку число 5 встречается в нем дважды.

**Меры изменчивости** – показывают как далеко, в среднем, отдельные значения разбросаны по отношению к среднему арифметическому значению.

- Дисперсия (variance) равна сумме квадратов отклонений каждого значения от среднего, деленной на  $N-1$ , где  $N$  - число значений в распределении. Для распределения [3 5 7 5 6 8 9] дисперсия равна  $((3 - 6,14)^2 + (5 - 6,14)^2 + (7 - 6,14)^2 + (5 - 6,14)^2 + (6 - 6,14)^2 + (8 - 6,14)^2 + (9 - 6,14)^2)/6 = 4,1429$
- Стандартное отклонение (standard deviation) равно квадратному корню из дисперсии. Для распределения [3 5 7 5 6 8 9] стандартное отклонение равно 2,0354.

**Характеристики диапазона распределения.**

- Минимум (minimum) равен наименьшему из значений распределения. Для распределения [3 5 7 5 6 8 9] минимум равен 3.
- Максимум (maximum) равен наибольшему из значений распределения. Для распределения [3 5 7 5 6 8 9] максимум равен 9.
- Размах (range) составляет разность между максимумом и минимумом распределения. Для распределения [3 5 7 5 6 8 9] размах равен  $9 - 3 = 6$ .
- Сумма (sum) равна сумме всех значений распределения. Для распределения [3 5 7 5 6 8 9] сумма равна  $3+5+7+5+6+8+9 = 43$ .

**Характеристики формы распределения.**

Используются для отражения близости формы распределения к нормальному виду.

- Эксцесс (kurtosis) – мера «сглаженности» («остро» и «плосковершинности») распределения. Если значение эксцесса близко к 0, это означает, что форма распределения близка к нормальному виду. Положительный эксцесс указывает на «плосковершинное» распределение, у которого максимум вероятности выражен не столь ярко, как у нормального. Значения эксцесса, превышающие 5,0, говорят о том, что по краям распределения находится больше значений, чем вокруг среднего. Отрицательный эксцесс характеризует «островершинное» распределение, график которого более вытянут по вертикальной оси, чем график нормального распределения. Считается, что распределение с эксцессом от  $-1$  до  $+1$  примерно соответствует нормальному виду.
- Асимметрия (skewness) показывает, в какую сторону относительно среднего сдвинуто большинство значений распределения. Нулевое значение асимметрии означает симметричность распределения относительно среднего значения, положительная асимметрия указывает на сдвиг распределения в сторону меньших значений, а отрицательная – в сторону больших значений. В большинстве случаев за нормальное распределение принимается распределение с асимметрией в пределах  $-1$  до  $+1$ .

**Стандартная ошибка (standard error)** – характеристика точности, или стабильности, величины, для которой она вычисляется. Чем меньше значение стандартной ошибки, тем выше стабильность величины, для которой она вычисляется.

Для вычисления описательных статистик в меню Analyze нужно выбрать команду Descriptive Statistics – Descriptives. В диалоговом окне необходимо задать переменные, для которых будут вычислены описательные статистики, перенести их в целевой список. По умолчанию в программе можно получить данные, включающие среднее значение (mean), стандартное отклонение (standard deviation), максимум (maximum), минимум (minimum). Для этого в окне Descriptives при заданном целевом списке нужно щелкнуть на кнопке ОК.

Чтобы вычислить дополнительные характеристики – размах (range), сумму (sum), дисперсию (variance), эксцесс (kurtosis), асимметрию (skewness) нужно перед щелчком на кнопке ОК щелкнуть на кнопке Options

<sup>5</sup> Наследов А. SPSS 15 профессиональный статистический анализ данных. СПб: Питер, 2008. – С.115-116



(Параметры). Откроется диалоговое окно Descriptives: Options, в котором с помощью флажков можно задать дополнительные характеристики, за исключением двоих: медианы (median) и моды (mode).

В зависимости от того, какие уровни измерения используются для статистического анализа, применяются разные методы вычисления описательных статистик для переменных.

Выделяется три основных уровня измерения переменных: номинальный, порядковый, интервальный.

Наиболее полную информацию дают интервальные измерения. Они позволяют численно выражать и сравнивать различия между объектами измерения. Например, переменная «возраст» может быть измерена с помощью интервальной шкалы, иногда бывает достаточно трех значений: молодежь в возрасте 18 до 35 лет, средний возраст – 36-55 лет, старший возраст – более 55 лет. Или может быть измерена в натуральных числах – годах с момента рождения человека. Объяснение свойства интервальных измерений численно выражать различия между объектами заложено в их названии: измерение осуществляется с помощью неизменного интервала, который выступает эталоном меры. Такими эталонами являются, например, градус, метр, килограмм, минута, процент или рубль. На интервальном уровне измерения осуществимы все операции с натуральными числами. Это имеет большое практическое значение, так как позволяет применять к интервальным переменным статистические методы любой сложности. Методику перевода переменной с натуральными числами в новую с интервальной шкалой мы приводили в разделе «Перекодировка в новую переменную».

На порядковом уровне измерения присутствует упорядочивание категорий с точки зрения возрастания/убывания интенсивности признака. С помощью порядковых (ранговых) шкал измеряют интенсивность оценок каких-то свойств, суждений, событий, степени согласия или несогласия с предложенными утверждениями.

Построение порядковой шкалы можно проиллюстрировать на примере переменной «политическое участие гражданина»<sup>6</sup>, использованием измерения, позволяющего ранжировать граждан по классам, различающимся количеством данного свойства, а именно:

- 1) отсутствие политического участия;
- 2) эпизодическое или регулярное участие в выборах в качестве избирателя;
- 3) регулярное участие в выборах, членство в политической партии;
- 4) регулярное участие в различных политических компаниях, акциях и т.д.
- 5) участие в выборах в качестве кандидата;
- 6) повседневное участие в принятии политических решений.

В приведенном примере интенсивность политического участия возрастает от первого класса к шестому. Можно утверждать, что в классе 2 (участие в выборах в качестве избирателя) признак «политическое участие» выражен больше, чем в классе 1 (отсутствие участия), но меньше, чем в классе 5 (участие в выборах в качестве кандидата). Относя изучаемых нами граждан к определенным классам политического участия, мы тем самым ранжируем их по данному признаку. Но такое ранжирование по классам не дает точных показателей, как фиксированный интервал, «эталон меры» политического участия. Поэтому по сравнению с интервальными шкалами возможности математических операций со значениями порядковых переменных ограничены.

Порядковые измерения имеют широкое применение в социологических исследованиях. Например, такие распространенные характеристики, как социальный статус или уровень образования измеряются по порядковой шкале. Порядковыми по своей природе являются такие переменные, как «политическая активность», «интерес к политике», «степень доверия к правительству», «отношение к той или иной политической партии».

Наименее полную информацию дают номинальные измерения (шкала наименований). Номинальная шкала устанавливает отношения равенства между явлениями, которые включены в один класс. Каждый элемент шкалы существует как бы сам по себе, и в целом шкала не упорядочена. Единственное условие состоит в том, что все элементы должны иметь единое основание для выделения. Номинальные переменные отражают сугубо качественные признаки, такие как «политическая ориентация», «членство в партии», «тип политического режима». При помощи номинальных переменных также измеряются преимущественно объективные признаки респондентов (пол, возраст, партийность, семейное положение, род занятий и др.). Соответственно, числовые значения на номинальном уровне не отражают каких-либо свойств объектов, а служат своего рода «ярлыками», «опознавательными кодами» классов.

Для номинальных и порядковых переменных с небольшим количеством категорий существует общее название: категориальные, или *неметрические*. Соответственно, интервальные и порядковые переменные с большим количеством категорий называют *метрическими*.

Для номинальных переменных наиболее важными вычислениями являются частотные распределения, процентные соотношения, мода и стандартное отклонение.

Пример описательных статистик номинальной переменной - «За какую партию вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?».

---

<sup>6</sup> Пример порядковой переменной приведен из учебного пособия Ахрименко А.С. Политический анализ и прогнозирование. – М.: Гардарики, 2006. – С.39

Descriptive Statistics											
	N	Range	Minimum	Maximum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
За какую из партий вы проголосовали бы в ближайшее воскресенье?	316	8	1	9	6,01	1,924	3,702	,007	,137	-,628	,273
Valid N (listwise)	316										

Число респондентов – 316 человек (пропущенные значения не учитываются). Стандартное отклонение – 1,924. Ассиметрия положительная +007, нормальная - в пределах +1, стабильная – стандартная ошибка 0,137. Эксцесс отрицательный -0,628, график «островершинный», соответствует нормальному виду – в пределах -1, величина стабильная – стандартная ошибка – 0,273.

Statistics		
За какую из партий вы проголосовали бы в ближайшее воскресенье?		
N	Valid	316
	Missing	7
Mean		6,01
Median		5,00
Mode		5
Std. Deviation		1,924
Variance		3,702
Skewness		,007
Std. Error of Skewness		,137
Kurtosis		-,628
Std. Error of Kurtosis		,273
Range		8
Minimum		1
Maximum		9

Мода – 5, соответствует варианту «Единая Россия», следовательно, в опрошенной группе наиболее распространены приверженцы партии «Единая Россия». Необходимо выяснить, насколько эта средняя (мода) в действительности отражает характер распределения, то есть насколько предпочтения партии «Единая Россия» типичны (репрезентативны) для группы в целом. Стандартное отклонение (Std. Deviation) показывает насколько существенен разброс значений вокруг средней. Стандартное отклонение – 1,924.

Для порядковых переменных основной средней величиной для порядковых переменных является медиана (median). Медиана представляет собой середину ранжирования числового ряда. В случае, когда число элементов является четным и возникают как бы две середины числового ряда, медиана – их среднее арифметическое.

Распространенный способ измерить разброс значений вокруг средней на порядковом уровне является вычисление *квартилей* - четвертей ранжированного ряда. Квартиль является естественным развитием медианы, с той разницей, что квартильное разбиение делит всех респондентов не на 2, а на 4 части. Первый квартиль – это такая точка на шкале, значения меньше (либо равные) которой отметили 25% опрошенных. Второй квартиль – точка, меньше которой отметили 50% опрошенных (следовательно, второй квартиль совпадает с медианой). Наконец, третий квартиль – точка, градации меньше которой отметили 75% опрошенных.

Квартильное отклонение – это разница между третьим и первым квартилями.

Вычисление квартилей, как и моды (и)или медианы возможно через команду Frequencies (Частоты). В диалоговом окне щелкнуть на кнопке Statistics (статистические показатели), в окне Percentile Values с помощью флажка задать квартили (Quartiles). При этом можно снять флажок Display Frequencies tables и не показывать на экране таблицы с частотными распределениями.

**Frequencies: Statistics**

Percentile Values

☒ Quartiles

☐ Cut points for 10 equal groups

☐ Percentile(s):

Add Change Remove

Central Tendency

☐ Mean

☐ Median

☐ Mode

☐ Sum

☐ Values are group midpoints

Dispersion

☐ Std. deviation

☐ Variance

☐ Range

☐ Minimum

☐ Maximum

☐ S.E. mean

Distribution

☐ Skewness

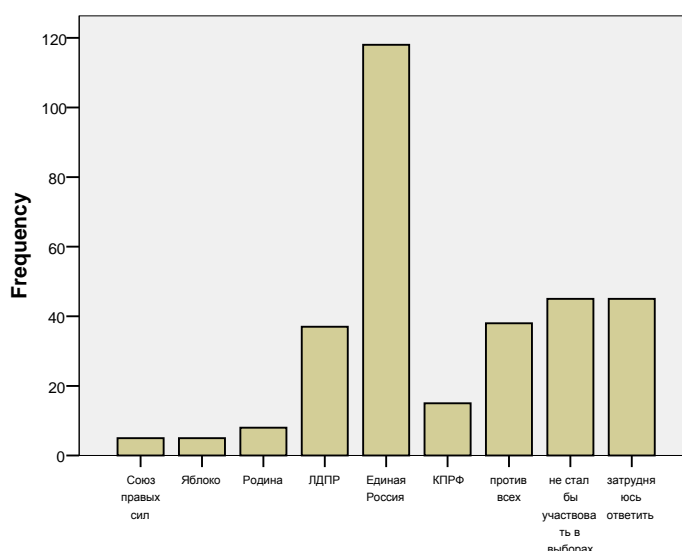
☐ Kurtosis

Continue Cancel Help

Вычисление моды и медианы возможно через команду Frequencies (Частоты). В диалоговом окне щелкнуть на кнопке Statistics (статистические показатели), с помощью флажков задать моду (mode) и медиану (median), а также здесь можно задать все остальные описательные характеристики.

В примере с вопросом «За какую партию вы проголосовали бы, если бы выборы состоялись в ближайшее воскресенье?» это будет выглядеть следующим образом. Самое часто встречающееся значение – мода – 5, медиана – 6,01.

За какую из партий вы проголосовали бы в ближайшее воскресенье?



Например, для порядковой переменной «удовлетворенность своим образованием» медианой является 2 значение, однако по

процентному соотношению нельзя сказать, насколько точно модель средней тенденции (медиана) отражает поведение переменной. Из таблицы видно, что достаточно большое количество респондентов имеют значение переменной – 3.

**В какой мере вы удовлетворены уровнем вашего образования?**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	полностью удовлетворен	196	15,3	15,5	15,5
	в основном удовлетворен	479	37,5	37,9	53,4
	не совсем удовлетворен	470	36,8	37,2	90,6
	совсем не удовлетворен	88	6,9	7,0	97,5
	затрудняюсь ответить	31	2,4	2,5	100,0
	Total	1264	98,9	100,0	
Missing	System	14	1,1		
Total		1278	100,0		

И только по квартильному разбиению можно сказать, что значение переменной 2 - «в основном удовлетворен» является важной средней характеристикой для всей выборочной совокупности.

Квантильное разбиение для переменной «удовлетворенность уровнем образования» будет выглядеть следующим образом:

#### Statistics

В какой мере вы удовлетворены уровнем вашего образования?

N	Valid	1264
	Missing	14
Percentiles	25	2,00
	50	2,00
	75	3,00

переменных является децильное

В этом случае в окне Frequencies Statistics в Percentile Values с помощью флажка нужно отметить умолчанию (в окошке появится 4 части (квартили), а на 10 рав-

Например, с помощью децильного разбиения можно изучить насколько высока неоднородность доходов, получаемых респондентами в месяц.

Децильное разбиение для переменной «доход» выглядит следующим образом:

Данные таблицы говорят о том, что доход до 5000 рублей в месяц получают 10% респондентов (граница первого дециля), а также о том, что для 10% опрошенных доход в месяц составляет 25000 руб. и выше (граница десятого дециля). Децильное отношение – это отношение десятого дециля к первому. Этот показатель демонстрирует, во сколько раз больше получают 10% наиболее высокооплачиваемых респондентов по сравнению с 10% наименее оплачиваемых. В нашем примере децильное отношение составляет 5,00, что показывает степень неоднородности доходов респондентов.

На интервальном уровне измерения, предполагающим не только упорядочение категорий по признаку «больше - меньше», но и установление фиксированного интервала измерения. Поэтому можно осуществлять все операции с натуральными числами.

Наиболее распространенной средней величиной для интервальных вычислений является среднее арифметическое (mean). Характерной особенностью среднего арифметического является высокая чувствительность к кренам в распределении, связанным с наличием в совокупности одного или нескольких предельных значений.

Традиционной мерой разброса значений вокруг средней на интервальном уровне выступает стандартное отклонение.

По методике вычисления описательных статистик проведем одномерный анализ интервальной переменной «доход в месяц». Ход вычислений: Analyze – Descriptive Statistics – Frequencies – Statistics – флажки на Mean, Median, Mode, Std. Deviation, Variance, Range, Minimum, Maximum.

Среднее арифметическое для переменной «доход в месяц» составляет 13,737 (средний доход в месяц составляет 13 тыс. руб.), стандартное отклонение – 12,2770 (достаточно большое значение, показывающее на разброс значений, минимальное значение – 1000 руб. в месяц, максимальное – 149 тыс. руб.).

**Задание:** 1. Определить какие шкалы

#### Statistics

Доход в месяц

N	Valid	643
	Missing	635
Percentiles	10	5,000
	20	6,000
	30	8,000
	40	10,000
	50	10,000
	60	13,000
	70	15,000
	80	20,000
	90	25,000

Полезным и нередко используемым показателем при анализе количественных отношений.

Percentile Values с помощью флажка нужно отметить Cut points for equal group. При этом по цифре 10) все респонденты делятся на 10 равных частей.

Децильное разбиение можно использовать для изучения насколько высока неоднородность доходов, получаемых респондентами в месяц.

переменной «доход» выглядит следующим образом:

#### Statistics

Доход в месяц

N	Valid	643
	Missing	635
Mean		13,737
Median		10,000
Mode		10,0
Std. Deviation		12,2770
Variance		150,725
Range		149,0
Minimum		1,0
Maximum		150,0

необходимы для измерения

менных «Какие городские проблемы вызывают у Вас сейчас наибольшую тревогу?», «Как Вы оцениваете эффективность работы городской администрации в решении существующих в городе проблем?», «Как часто вы смотрите передачи на политические темы по телевидению?». Сформулировать значения переменных (варианты ответов).

2. (по массиву данных файла `opros.sav`). На основании вычисления описательных статистик (моды, стандартного отклонения, асимметрии и эксцесса), а так же частоты и процентных соотношений определить характер распределения респондентов по категориям отношения к политической деятельности – переменная «интерес к политической жизни» (`var11`). Выяснить какая категория (значение переменной) типична для выборочной группы. Построить столбиковую диаграмму.

3. Создать описательные статистики, выбранных двух-трех переменных собственного исследования.

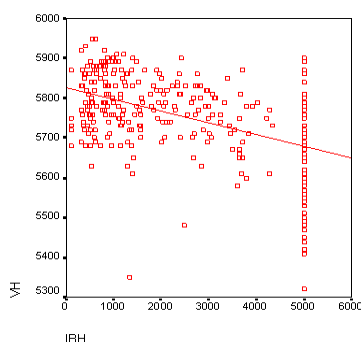
## 5. Взаимосвязь переменных.

### 5.1. Двумерный анализ социологических данных. Парные распределения.

Обработка социологических данных с помощью одномерных частотных распределений, как правило, является исходным этапом анализа собранной информации. Вместе с тем наиболее интересные для социологов вопросы связаны с одновременным анализом значений более одной переменной.

Процесс анализа собранных данных предполагает формирование гипотез типа: «социальные группы с разным уровнем образования (дохода, должностью, местом жительства и т.д.) отличаются по электоральным предпочтениям (степенью удовлетворенности жизнью и т.д.)». Другими словами, допускается, что существует переменная (такая как «принадлежность к определенной социальной группе»), которая объясняет поведение других переменных. Таким образом, есть объясняющие переменные, которые называются *независимыми*, и объясняемые переменные – *зависимые*.

Корреляционный анализ основан на расчете отклонения значений изучаемого признака от линии регрессии (от лат. regression – возврат, в данном случае – возврат к средней) – условной линии, к которой эти значения тяготеют. Чем меньше разброс значений, тем сильнее связи.



Корреляция (от лат. correlatio - соотношение) – это статистическая взаимосвязь между признаками изучаемого явления. Корреляционный анализ представляет собой математическую процедуру, с помощью которой изучается эта взаимосвязь.

Наиболее частыми инструментами изучения взаимосвязи двух переменных являются двумерные методы анализа *таблицы сопряженности*.

При анализе зависимостей двух переменных важнейшим является вопрос о том, какую из переменных считать зависимой, то есть подверженной влиянию, а какую – независимой, то есть влияющей.

Например, примем переменную «возраст» как независимую переменную, а переменную «электоральная активность» как зависимую. По гипотезе исследования возраст респондента оказывает влияние на готовность прийти на выборы. В таблице сопряженности (парном распределении) данные будут выглядеть следующим образом.

**Возрастная категория \* Собираетесь ли участвовать в выборах? Crosstabulation**

% within Возрастная категория		Собираетесь ли участвовать в выборах?				Total
		да	нет	не решил	затрудняюсь	
Возрастная категория	18-30 лет	55,3%	24,3%	13,2%	7,2%	100,0%
	31-40 лет	64,3%	13,2%	20,2%	2,3%	100,0%
	41-50 лет	64,3%	14,3%	21,4%		100,0%
	старше 50 лет	74,1%	10,3%	10,3%	5,2%	100,0%
Total		62,4%	17,0%	16,5%	4,0%	100,0%

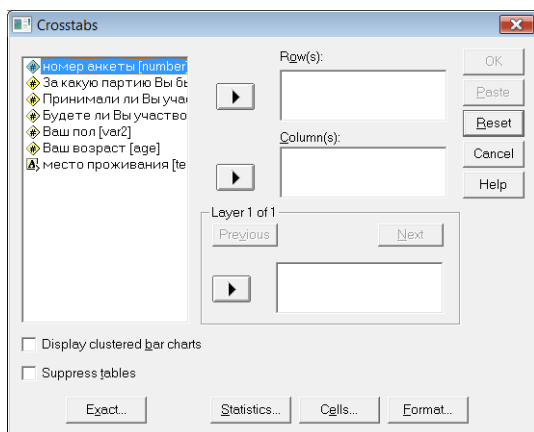
По данным в таблице можно увидеть, что действительно есть прямая зависимость возраста респондента и его электоральной активности. Среди респондентов старше 50 лет подавляющее большинство – 74,1% - готово голосовать на выборах, что свидетельствует о высокой электоральной активности людей старшей возрастной категории. Среди молодых респондентов в возрасте до 30 лет готовность голосовать на выборах продемонстрировали всего лишь 55,3% респондентов, почти четверть из них – 24,3% - заявили, что не будут участвовать в голосовании. Таким образом, чем старше возраст респондентов, тем выше их электоральная активность.

Если же принять переменную «электоральная активность» за независимую, а переменную «возраст» за зависимую, то можно получить несколько другие данные таблицы, где нормирование можно провести не от сумм по строкам, а от сумм по колонкам.

**Возрастная категория \* Собираетесь ли участвовать в выборах? Crosstabulation**

% within Собираетесь ли участвовать в выборах?		Собираетесь ли участвовать в выборах?				Total
		да	нет	не решил	затрудняюсь	
Возрастная категория	18-30 лет	31,8%	51,4%	28,6%	64,7%	35,9%
	31-40 лет	31,4%	23,6%	37,1%	17,6%	30,5%
	41-50 лет	20,5%	16,7%	25,7%		19,9%
	старше 50 лет	16,3%	8,3%	8,6%	17,6%	13,7%
Total		100,0%	100,0%	100,0%	100,0%	100,0%

В этом случае распределения необходимо сравнивать по разным колонкам таблицы, а не по строкам. Из тех респондентов, кто не собирается голосовать на выборах, большинство составляет молодежь в возрасте до 30 лет (51,4%), респондентов в возрасте 50 лет среди них всего 8,3%. Таким образом, низкая электоральная активность в большей степени характерна для молодых людей, чем для старшего поколения.



Для работы с таблицами сопряженности в программе SPSS используется команды Analyze – Descriptive Statistics - Crosstabs (Таблицы сопряженности). Например, нам нужно выяснить есть ли зависимость готовности голосовать на выборах от возраста респондентов.

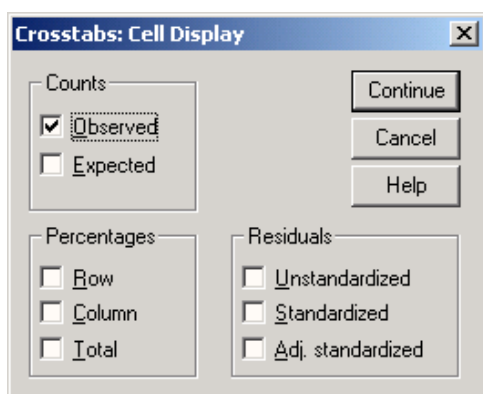
Исследуем эту зависимость чуть более детально; для этого нам понадобится точно ответить на следующие вопросы:

- Существует ли зависимость вообще?
- Что можно сказать об интенсивности этой зависимости?
- Что можно сказать о направлении и характере этой зависимости?

Для создания таблицы с переменными «возраст» и «готовность голосовать», нужно сначала выделить переменную «возраст» и с помощью кнопки с треугольником переместить в список Row(s) (Строки), а переменную «готовность голосовать» в список Column(s) (Столбцы).

Раздел Layer 1 of 1 диалогового окна позволяет построить таблицу сопряженности для трех и более переменных.

Для получения данных в процентах нужно щелкнуть на кнопке Cells (Ячейки), открыть диалоговое окно Crosstabs: Cells Display.



Например, нужно установить, существует ли на самом деле статистическая зависимость двух переменных – «возраст» и «готовность голосовать на выборах».

По умолчанию установлен флажок Observed (Наблюдаемые) в группе Counts (Значения), так как наблюдаемые частоты являются главной вычисляемой величиной. При установке флажка Expected (Ожидаемые) в группе Counts (Значения) отображается значение ожидаемой частоты для каждой ячейки. Ожидаемая частота – количество респондентов, которые должны быть в ячейках таблицы в случае независимости переменных. Сопоставляя эти ожидаемые частоты с наблюдаемыми частотами мы можем судить о том, действительно ли два номинальных признака независимы. Чем больше расхождение наблюдаемых и ожидаемых частот, тем эти два признака сильнее связаны друг с другом. При установке флажка Unstandardized (Нестандартизированные) в группе Residuals (Остатки) отображается разность между наблюдаемой и ожидаемой частотами.

Возрастная категория * Собираетесь ли участвовать в выборах? Crosstabulation							
			Собираетесь ли участвовать в выборах?				Total
			да	нет	не решил	затрудняюсь	
Возрастная категория	18-30 лет	Count	84	37	20	11	152
		Expected Count	94,9	25,9	25,2	6,1	152,0
		Residual	-10,9	11,1	-5,2	4,9	
	31-40 лет	Count	83	17	26	3	129
		Expected Count	80,5	22,0	21,3	5,2	129,0
		Residual	2,5	-5,0	4,7	-2,2	
	41-50 лет	Count	54	12	18	0	84
		Expected Count	52,4	14,3	13,9	3,4	84,0
		Residual	1,6	-2,3	4,1	-3,4	
	старше 50 лет	Count	43	6	6	3	58
		Expected Count	36,2	9,9	9,6	2,3	58,0
		Residual	6,8	-3,9	-3,6	,7	
Total	Count	264	72	70	17	423	
	Expected Count	264,0	72,0	70,0	17,0	423,0	

Как показывают данные в таблице реальные частоты Count и ожидаемые частоты Expected Count разные в большинстве ячеек таблицы. Следовательно, можно сделать вывод о том, что независимость переменных не подтверждается.

Установление соответствия между наблюдаемыми и ожидаемыми значениями возможно при применении критерия независимости  $\chi^2$  (хи-квадрат), величина которого определяется, как сумма отношений суммы квадратов отклонений наблюдаемой величины  $f_o$  от ожидаемой величины  $f_e$  к ожидаемой величине в каждой ячейке.

Для того, чтобы провести тест хи-квадрат с помощью SPSS, нужно выполнить следующие действия:

- выбрать в меню команды Analyze (Анализ) Descriptive Statistics (Дескриптивные статистики) Crosstabs... (Таблицы сопряженности)
- кнопкой Reset (Сброс) удалите возможные настройки.
- перенести переменную «возраст» в список строк, а переменную «готовность голосовать» — в список столбцов.
- щелкнуть на кнопке Cells... (Ячейки). В диалоговом окне установить, кроме предлагаемого по умолчанию флажка Observed, еще флажки Expected и Standardized. Подтвердить выбор кнопкой Continue.
- щелкнуть на кнопке Statistics... (Статистика).

Откроется описанное выше диалоговое окно Crosstabs: Statistics.

- установить флажок Chi-square (Хи-квадрат). Щелкнуть на кнопке Continue, а в главном диалоговом окне — на ОК.

Получится следующая таблица сопряженности.

			Собираетесь ли участвовать в выборах?				Total
			да	нет	не решил	затрудняюсь	
Возрастная категория	18-30 лет	Count	84	37	20	11	152
		Std. Residual	-1,1	2,2	-1,0	2,0	
	31-40 лет	Count	83	17	26	3	129
		Std. Residual	,3	-1,1	1,0	-1,0	
	41-50 лет	Count	54	12	18	0	84
		Std. Residual	,2	-,6	1,1	-1,8	
	старше 50 лет	Count	43	6	6	3	58
		Std. Residual	1,1	-1,2	-1,2	,4	
Total		Count	264	72	70	17	423

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	23,472 <sup>a</sup>	9	,005
Likelihood Ratio	26,133	9	,002
Linear-by-Linear Association	3,826	1	,050
N of Valid Cases	423		

a. 2 cells (12,5%) have expected count less than 5. The minimum expected count is 2,33.

(2 ячейки (12,5%) имеют ожидаемую величину менее 5. Минимальная ожидаемая величина 2,33.)

Принимаются во внимание абсолютные значения остатков, превышающие 1,65. Это служит индикатором существования значимой статистической зависимости между изучаемыми признаками. Знак «плюс» в стандартизованных остатках свидетельствует о том, что реальное количество наблюдений больше ожидаемого, знак «минус» - о том, что оно меньше ожидаемого. Следует учитывать, что величина стандартизованных остатков указывает лишь на вероятность наличия линейной зависимости между изучаемыми переменными, но не на направление и интенсивность этой зависимости.

Для вычисления критерия хи-квадрат применяются три различных подхода: формула Пирсона (Pearson Chi-Square), поправка на правдоподобие (Likelihood Ratio) и тест «линейно-линейная связь» (Linear-by-Linear Association). Если таблица сопряженности имеет четыре поля и ожидаемая вероятность менее 5, дополнительно выполняется точный тест Фишера (Fishers Exact Test).

Df (Ст.св.) – степени свободы, произведение количеств градаций переменных, уменьшенных на 1. Это количество ячеек таблицы, которые могут быть заполнены числами, прежде чем содержание всех остальных ячеек станет постоянным.

Asymp.Sig. (Асимпт. значимость) – вероятность случайности связи или  $p$ -уровень значимости. Чем меньше эта величина, тем выше статистическая значимость (достоверность) связи. При  $p$ -уровне значимости  $p > 0,05$  считается, что различия между наблюдаемыми и ожидаемыми значениями незначительны.

### Критерий хи-квадрат по Пирсону

Обычно для вычисления критерия хи-квадрат используется формула Пирсона:

Здесь вычисляется сумма квадратов стандартизованных остатков по всем полям таблицы сопряженности.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Поэтому поля с более высоким стандартизованным остатком вносят более весомый вклад в численное значение критерия хи-квадрат и, следовательно, — в значимый результат. Стандартизованный остаток (Std. Residual) 2 или более указывает на значимое расхождение между наблюдаемой и ожидаемой частотами.



В рассматриваемом нами примере формула Пирсона дает максимально значимую величину критерия хи-квадрат ( $p < 0,001$ ). Если рассмотреть стандартизованные остатки в отдельных полях таблицы сопряженности, то на основе вышеприведенного правила можно сделать вывод, что эта значимость в основном определяется полями, в которых переменная «готовность голосовать» имеет значение "нет". У молодых людей до 30 лет это значение повышено (2,2).

Корректность проведения теста хи-квадрат определяется двумя условиями: во-первых, ожидаемые частоты  $< 5$  должны встречаться не более чем в 20% полей таблицы; во-вторых, суммы по строкам и столбцам всегда должны быть больше нуля.

В рассматриваемом примере это условие выполняется полностью. Как указывает примечание после таблицы теста хи-квадрат, только 12,5% полей имеют ожидаемую частоту менее 5.

### Критерий хи-квадрат с поправкой на правдоподобие

Альтернативой формуле Пирсона для вычисления критерия хи-квадрат является поправка на правдоподобие. При большом объеме выборки формула Пирсона и подправленная формула дают очень близкие результаты. В нашем примере критерий хи-квадрат с поправкой на правдоподобие составляет 26,133.

### Тест «линейно-линейная связь» (Linear-by-Linear Association)

Дополнительно в таблице сопряженности под обозначением linear-by-linear ("линейный-по-линейному") выводится значение теста Мантеля-Хэнзеля (3,826). Эта еще одна мера линейной зависимости между строками и столбцами таблицы сопряженности. Она определяется как произведение коэффициента корреляции Пирсона на количество наблюдений, уменьшенное на единицу:

$$\chi^2 = r^2(n-1)$$

Полученный таким образом критерий имеет одну степень свободы. Метод Мантеля-Хэнзеля использует всегда, когда в диалоговом окне Crosstabs: Statistics установлен флажок Chi-square. Однако для данных, относящихся к номинальной шкале, этот критерий неприменим.

Таблицы сопряженности, пример которых мы рассмотрели выше, имеют тот недостаток, что в них приводятся только абсолютные значения. Чтобы узнать, насколько эти значения важны по отношению к общему количеству, надо определить их процентную долю, для вычисления процентных значений нужно выполнить следующие действия:

- выбрать в меню команды Analyze (Анализ) Descriptive Statistics (Дескриптивные статистики) Crosstabs... (Таблицы сопряженности)
- Не изменяя прежних настроек, щелкнуть на кнопке Cells... Откроется диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Отображение ячеек). В группе Percentages (Проценты) можно выбрать один или более из нижеследующих вариантов отображения:
- Row (По строкам): Вычисляются процентные значения по строкам: количество наблюдений в каждой ячейке, отнесенное к сумме по строке.
- Column (По столбцам): Вычисляются процентные значения по столбцам: количество наблюдений в каждой ячейке в отношении к сумме столбца.
- Total (Полные): Вычисляются полные процентные значения: количество наблюдений в каждой ячейке, отнесенное к общей сумме наблюдений.

Таким образом, можно получить данные в двумерной таблице по строкам и столбцам и интерпретировать их в зависимости от заданной задачи. Возможно создание общей таблицы, где представлены проценты по строкам и колонкам таблицы, а так же частоты.



Возрастная категория \* Собираетесь ли участвовать в выборах? Crosstabulation

			Собираетесь ли участвовать в выборах?				Total
			да	нет	не решил	затрудняюсь	
Возрастная категория	18-30 лет	Count	84	37	20	11	152
		% within Возрастная категория	55,3%	24,3%	13,2%	7,2%	100,0%
		% within Собираетесь ли участвовать в выборах?	31,8%	51,4%	28,6%	64,7%	35,9%
	31-40 лет	Count	83	17	26	3	129
		% within Возрастная категория	64,3%	13,2%	20,2%	2,3%	100,0%
		% within Собираетесь ли участвовать в выборах?	31,4%	23,6%	37,1%	17,6%	30,5%
	41-50 лет	Count	54	12	18	0	84
		% within Возрастная категория	64,3%	14,3%	21,4%	,0%	100,0%
		% within Собираетесь ли участвовать в выборах?	20,5%	16,7%	25,7%	,0%	19,9%
	старше 50 лет	Count	43	6	6	3	58
		% within Возрастная категория	74,1%	10,3%	10,3%	5,2%	100,0%
		% within Собираетесь ли участвовать в выборах?	16,3%	8,3%	8,6%	17,6%	13,7%
Total		Count	264	72	70	17	423
		% within Возрастная категория	62,4%	17,0%	16,5%	4,0%	100,0%
		% within Собираетесь ли участвовать в выборах?	100,0%	100,0%	100,0%	100,0%	100,0%

По данным таблицы можно сказать, что среди молодых респондентов в возрасте до 30 лет готовность прийти на выборы гораздо ниже, чем у респондентов других возрастных категорий. Только 55,3% молодых респондентов готовы прийти и проголосовать на выборах. В категории респондентов старше 50 лет тех, кто придет голосовать значительно больше – 74,1%.

С другой стороны, из числа тех, кто не собирается голосовать на выборах, большинство составляют молодые респонденты – 51,4%, в возрасте от 45 до 50 лет таких респондентов в три раза меньше – 16,7%, среди пожилых – всего 8,3%. В категории тех, кто еще не решил голосовать ли ему не выборах, больше всего респондентов в возрасте от 31 до 40 лет – 37,1%.

**Задание. 1.** (по массиву данных файла opros.sav). Построить таблицу сопряженности двух переменных «Как вы относитесь к политической деятельности» и «vozrast» (с использованием интервальной шкалы). Проанализировать процентные соотношения.

2. Провести тест хи-квадрат переменных «Как вы относитесь к политической деятельности» и переменной «vozrast», выявить корреляционную зависимость/независимость этих переменных, по стандартизированному остатку, критерию хи-квадрата проанализировать связи переменных.

3. Исходя из задач и гипотез собственного исследования, выбрать переменные, удовлетворяющие условиям зависимости. Определить зависимые и независимые переменные. Построить таблицы сопряженности переменных собственного исследования, проанализировать данные на наличие зависимости переменных. Выяснить интенсивность зависимости переменных с помощью теста хи-квадрат.

## 5.2. Коэффициенты корреляции

До сих пор мы выясняли лишь сам факт существования статистической зависимости между двумя признаками. Далее мы попробуем выяснить, какие заключения можно сделать о силе или слабости этой зависимости, а также о ее виде и направленности. Критерии количественной оценки зависимости между переменными называются *коэффициентами корреляции* или *мерами связанности*. Значение коэффициента служит показателем интенсивности связи.

Следует отметить, что коэффициенты корреляции выражают не причинную (обусловленность одного признака другим), а функциональную (взаимная согласованность изменения признаков) зависимость между признаками. Различают парную (между двумя признаками) и множественную (между несколькими признаками) корреляции.

Две переменные коррелируют между собой *положительно*, если между ними существует прямое, однонаправленное соотношение. Положительная корреляция соответствует значениям  $0 < r < 1$ . Положительную корреляцию следует интерпретировать следующим образом: если значения одной переменной возрастают, то значения другой имеют тенденцию к возрастанию. Чем коэффициент корреляции ближе к 1, тем сильнее эта тенденция, и, наоборот, с приближением коэффициента корреляции к 0 тенденция ослабевает.

Для словесного описания величин коэффициента корреляции применяется следующая таблица:

Значение коэффициента корреляции $r$	Интерпретация
$0 < r \leq 0,2$	Очень слабая корреляция
$0,2 < r \leq 0,5$	Слабая корреляция
$0,5 < r \leq 0,7$	Средняя корреляция
$0,7 < r \leq 0,9$	Сильная корреляция
$0,9 < r \leq 1$	Очень сильная корреляция

Пример сильной положительной корреляции служит зависимость между ростом и весом человека. (если,  $r = 0,83$ )

Отсутствие корреляции определяется значением  $r = 0$ . Нулевой коэффициент корреляции говорит о том, что значения переменных никак не связаны друг с другом. Примером пары величин с нулевой корреляцией является рост человека и результат его IQ-теста.

Две переменные коррелируют между собой *отрицательно*, если между ними существует обратное, разнонаправленное соотношение. Отрицательная корреляция соответствует значениям  $-1 < r < 0$ . Если значения одной переменной возрастают, то значения другой имеют тенденцию к убыванию. Чем коэффициент корреляции ближе к  $-1$ , тем сильнее эта тенденция, и, наоборот, с приближением к 0 тенденция ослабевает.

Для изучения взаимосвязи признаков, измеренных с помощью различных типов шкал, используются разные коэффициенты корреляции. В качестве коэффициента корреляции между переменными, принадлежащими порядковой шкале применяется коэффициент Спирмена, а для переменных, принадлежащих к интервальной шкале — коэффициент корреляции Пирсона (момент произведений). При этом следует учесть, что каждую дихотомическую переменную, то есть переменную, принадлежащую к номинальной шкале и имеющую две категории, можно рассматривать как порядковую. Коэффициент Спирмена равен  $+1$ , когда два ряда проранжированы строго в одном порядке,  $-1$ , когда два ряда проранжированы в строго обратном порядке, и равен нулю при полном взаимном беспорядочном расположении рангов. Коэффициент корреляции Пирсона равен  $+1$  при строгой (полной) прямой взаимозависимости двух признаков, равен  $-1$  при строгой (полной) обратной взаимозависимости.

Для начала мы проверим существует ли корреляция между переменными «возраст» и «готовность голосовать на выборах». Нужно выполнить следующие действия:

- выбрать в меню команды Analyze (Анализ) Descriptive Statistics (Дескриптивные статистики) Crosstabs... (Таблицы сопряженности)
- перенести переменную «возраст» в список строк, а переменную «готовность голосовать» — в список столбцов.
- щелкнуть на кнопке Statistics... (Статистика). В диалоге Crosstabs: Statistics установить флажок Correlations (Корреляции). Подтвердить выбор кнопкой Continue.
- В диалоге Crosstabs нужно отказаться от вывода таблиц, установив флажок Suppress tables (Подавлять таблицы). Щелкнуть на кнопке ОК.

Будут вычислены коэффициенты корреляции Спирмена и Пирсона, а также проведена проверка их значимости:

Symmetric Measures				
	Value	Asymp. Std. Error <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
Interval by Interval Pearson's R	-,095	,048	-1,963	,050 <sup>c</sup>
Ordinal by Ordinal Spearman Correlation	-,107	,048	-2,203	,028 <sup>c</sup>
N of Valid Cases	423			

a. Not assuming the null hypothesis.

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

c. Based on normal approximation.

Так как здесь порядковая переменная, мы рассмотрим коэффициент корреляции Пирсона. Он составляет  $-0,095$ .  $p$ -уровень  $-0,050$

Исходя из данных таблицы, можно сделать следующие заключения: Между переменными «возраст» и «готовность голосовать на выборах» существует слабая корреляция (заключение о силе зависимости), переменные коррелируют отрицательно (заключение о направлении зависимости).

Следовательно, разнонаправленность соотношения можно интерпретировать следующим образом: чем моложе респонденты, тем ниже их готовность прийти на выборы, и наоборот, чем старше респонденты, тем чаще они готовы голосовать на выборах. Таким образом, электоральная активность респондентов в некоторой степени зависит от возраста респондентов.

**Задание. 1.** (по массиву данных *opros.sav*) с помощью коэффициентов корреляции определить направленность, характер и интенсивность связи между переменными «Как вы относитесь к политической деятельности?» и «Согласны ли Вы с утверждением - «политических деятелей не заботит что думают такие люди как я».

2. Исходя из задач и гипотез собственного исследования, выбрать переменные, удовлетворяющие условиям зависимости. Проанализировать данные на наличие зависимости переменных с помощью коэффициентов корреляции. Выяснить интенсивность, характер и направленность зависимости переменных.

## 6. Анализ множественных ответов

В данном разделе мы рассмотрим особенности кодирования и анализа множественных ответов. Вопросы, на которые можно дать несколько ответов одновременно (это и есть множественные ответы), имеются во многих анкетных исследованиях. Для кодировки анализа таких множественных ответов SPSS представляет два различных метода: метод множественной дихотомии и категориальный метод. Наиболее удобным и часто используемым является категориальный метод, который и будет рассмотрен более подробно.

### 6.1. Анализ множественных ответов с применением категориального метода

В анкетных опросах достаточно часто встречаются вопросы, на которые можно дать несколько ответов одновременно. Возьмем в качестве примера вопрос о симпатиях респондентов различным политическим силам. Например, в анкете содержится вопрос под номером 27:

**«Каким политическим силам Вы симпатизируете? (возможно любое число вариантов ответа)»**

1. сторонникам коммунистической идеологии
2. сторонникам социалистических идей, другим левым силам
3. сторонникам социально-ориентированного государства
4. тех, кто выступает против наплыва мигрантов
5. сторонникам возрождения в стране православия
6. демократам и правозащитникам
7. сторонникам нынешнего политического курса
8. сторонникам радикальных рыночных реформ
9. иное
10. никому
11. затрудняюсь ответить

В макете данных Variable View создаются несколько одинаковых вариантов переменных var27 - «**Каким политическим силам Вы симпатизируете?**». Число вариантов переменных зависит от максимального количества вариантов, выбранных одним респондентом (или заранее заданными исследователями, вместо «*возможно любое число вариантов ответа*» в вопросе поставить: «*возможно отметить не более 3 вариантов ответа*»). Допустим, максимальное число выбранных вариантов – 3, поэтому нужно создать три переменных с одинаковыми метками переменной Label и метками значений Values – var27a, var27b, var27c.

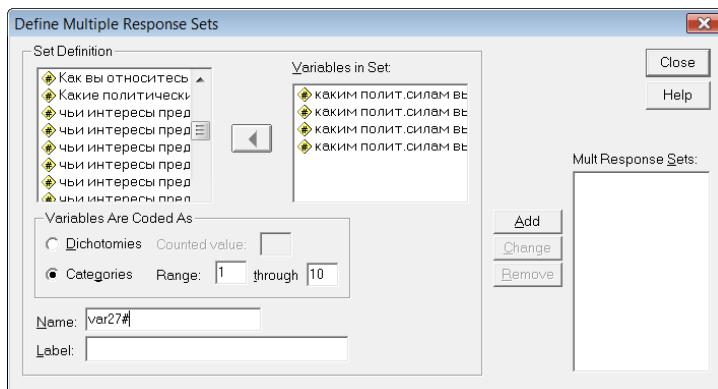
Таким образом, эти три переменных должны принадлежать к одному «набору переменных» - вопросу var27. Для этого нужно выбрать в меню команды Analyze (Анализ) - Multiple Response (Множественные ответы) - Define Sets... (Определить наборы)

Откроется диалоговое окно Define Multiple Response Sets (Определение наборов ответов).

- нужно выделить в списке исходных переменных переменные var27a, var27b, var27c и перенесите их в список Variables in Set (Переменные в наборе).
- задать кодировку переменных (опция Categories в группе Variables Are Coded As). В поле Range ввести «1», в поле through ввести «11» (по общему числу вариантов в вопросе – диапазону переменной)
- присвоить набору имя Name «var27#» и метку «симпатии политическим силам».
- щелкнуть на кнопке Add (Добавить), и созданный набор будет внесен в список наборов множественных ответов (Mult Response Sets).

SPSS начинает имена наборов переменных со знака \$; следовательно, вновь созданный набор получит имя \$var27#.

- щелкнуть на кнопке Close (Заккрыть), чтобы закончить процесс определения набора.



Частотные таблицы (линейные распределения) для вопросов с множественными ответами.

- Чтобы создать частотную таблицу, выберите команды меню Analyze (Анализ) - Multiple Response (Множественные отве-

ты) - Frequencies... (Частоты).

Откроется диалоговое окно Multiple Response - Frequencies (Частоты множественных ответов).

В списке Mult Response Sets этого диалога отображаются уже определенные наборы переменных; в нашем примере это набор \$var27#.

- Перенесите набор \$var27# в список Table(s) for (Таблицы для).
- Щелкните на кнопке ОК.

В окне просмотра появятся следующие результаты:

<b>\$var27# Frequencies</b>				
		Responses		Percent of Cases
		N	Percent	
\$var27# <sup>a</sup>	сторонникам коммунистической идеологии	17	4,0%	5,3%
	сторонникам соц.идей, другим левым силам	4	,9%	1,2%
	сторонникам социально-ориентиров. гос-ва	81	19,2%	25,1%
	тех, кто против наплыва мигрантов	71	16,8%	22,0%
	сторонникам возрождения православия	48	11,4%	14,9%
	демократам и правозащитникам	57	13,5%	17,6%
	сторонникам нынешнего политич. курса	40	9,5%	12,4%
	сторонникам радикальных рыночных реформ	9	2,1%	2,8%
	иное	6	1,4%	1,9%
	никому	53	12,6%	16,4%
	затрудняюсь ответить	36	8,5%	11,1%
Total		422	100,0%	130,7%

a. Group

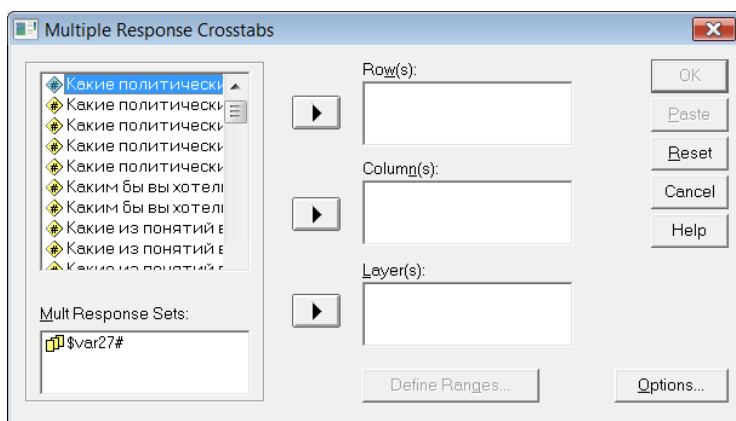
Столбец Percent содержит данные ответов респондентов в процентном отношении. Для анализа данных таблицы нам необходим столбец Percent of Cases, то есть процент от наблюдаемых случаев (процент от числа всех выборов вариантов ответа). Из таблицы видно, что четверть опрошенных (25,1%) симпатизируют идее социально-ориентированного государства, 22,0% респондентов на стороне тех, кто выступает против наплыва в страну мигрантов, сторонниками демократов являются 17,6% опрошенных, а нынешний политический курс одобряют 12,4%.

**Задание. 1.** (по массиву файла opros.sav) Создать линейное распределение переменных с множественными ответами «Какие из проблем представляют опасность для молодежи в округе?», «Что важно для достижения успеха в жизни?». Проанализировать полученные данные.

## 6.2.Таблицы сопряженности (парные распределения) вопросов с множественными ответами

Таблицы сопряженности можно создавать между двумя наборами переменных, а также между набором и "обычной" переменной. К примеру, нам необходимо в одной таблице сопряженности отобразить соотношение между набором \$var27# и переменной vozrast, которая характеризует возраст респондентов и содержит 4 варианта возрастных категорий (диапазон переменной): 1 – от 18 до 30 лет, 2 – от 31 до 40 лет, 3 – от 41 до 50 лет, 4 – старше 50 лет.

- Нужно выбрать в меню команды Analyze (Анализ) - Multiple Response (Множественные ответы) - Crosstabs... (Таблицы сопряженности). Появится диалоговое окно Multiple Response Crosstabs.



В списке наборов множественных ответов должен быть показан ранее определенный набор - \$var27#.

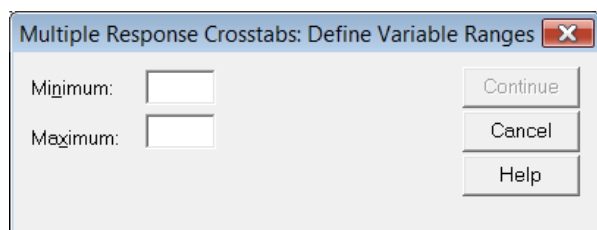
- Нужно перенести в список переменных строк Row(s) набор \$var27#, а в список переменных столбцов Column(s) — переменную возраст. Эта переменная появится в списке столбцов с двумя вопросительными знаками, заключенными в скобки. Если таблица сопряженности строится между элементарными переменными (не являющимися наборами) и

наборами, то для первых следует задать диапазон значений.

- щелкнуть на кнопке Define Ranges... (Определить диапазоны).

Откроется диалоговое окно Multiple Response Crosstabs: Define Variable Range (Таблицы сопряженности для множественных ответов: Определить диапазон переменной).

- нужно задать минимальное значение (Minimum) "1", а максимальное (Maximum) — "4".



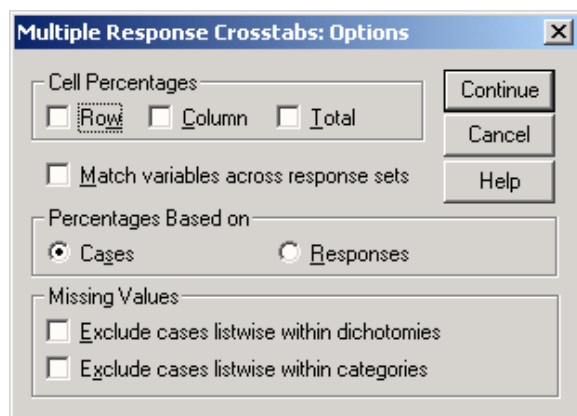
просительные знаки заменены значениями "1" и "4".

- щелкнуть на кнопке Options... (Параметры).

Откроется диалоговое окно Multiple Response Crosstabs: Options.

Абсолютные частоты в ячейках выводятся всегда. Дополнительно в группе Cell Percentages (Проценты в ячейках) можно выбрать одну или несколько характеристик:

- Row (По строкам): Отображаются проценты для строки.



- Column (По столбцам). Отображаются проценты для столбца.
- Total (Полные): Отображаются общие проценты для таблицы.

В группе Percentages based on (Проценты вычисляются на основе) можно выбрать одну из следующих опций:

- Cases (Наблюдения): Это настройка по умолчанию. Основанием для расчёта процентных показателей в ячейках является число наблюдений, соответствующие количеству опрошенных респондентов.
- Responses (ответы): Основой расчета процентного отношения в ячейке является количество ответов. Для множественных наборов количество ответов равно частоте учитываемого значения во всех наблюдениях.

Флажок Match variables across response sets (Учитывать переменные из наборов попарно) имеет смысл, только если таблица сопряженности строится на основе двух наборов переменных. В этом случае первая переменная из первого набора сочетается с первой переменной из второго набора, и т.д.

- Если в группе Percentages based on сохранить настройку по умолчанию Cases.
- В группе Cell Percentages установить флажки Row и Column.
- Подтвердить ввод кнопкой Continue, а затем — ОК. В окне просмотра будет показана следующая таблица.

\$var27#\*vozrast Crosstabulation

			возраст				Total
			от 18 до 30 лет	от 31 до 40 лет	от 41 до 50 лет	старше 50 лет	
\$var27# коммунистам	Count		4	1	6	5	16
	% within \$var27#		25,0%	6,3%	37,5%	31,3%	
	% within v ozrast		2,1%	,8%	10,7%	12,8%	
социалистич. силам	Count		1	1	0	2	4
	% within \$var27#		25,0%	25,0%	,0%	50,0%	
	% within v ozrast		,5%	,8%	,0%	5,1%	
социально-ориен. гос-	Count		34	24	11	11	80
	% within \$var27#		42,5%	30,0%	13,8%	13,8%	
	% within v ozrast		17,6%	18,8%	19,6%	28,2%	
против мигрантов	Count		30	23	8	8	69
	% within \$var27#		43,5%	33,3%	11,6%	11,6%	
	% within v ozrast		15,5%	18,0%	14,3%	20,5%	
возрожд. православия	Count		21	19	4	3	47
	% within \$var27#		44,7%	40,4%	8,5%	6,4%	
	% within v ozrast		10,9%	14,8%	7,1%	7,7%	
демократам, правозащ.	Count		30	19	6	2	57
	% within \$var27#		52,6%	33,3%	10,5%	3,5%	
	% within v ozrast		15,5%	14,8%	10,7%	5,1%	
нынеш. полит. курсу	Count		12	18	7	2	39
	% within \$var27#		30,8%	46,2%	17,9%	5,1%	
	% within v ozrast		6,2%	14,1%	12,5%	5,1%	
радикальным реформам	Count		5	3	1	0	9
	% within \$var27#		55,6%	33,3%	11,1%	,0%	
	% within v ozrast		2,6%	2,3%	1,8%	,0%	
иное	Count		4	1	0	1	6
	% within \$var27#		66,7%	16,7%	,0%	16,7%	
	% within v ozrast		2,1%	,8%	,0%	2,6%	
никому	Count		28	11	9	5	53
	% within \$var27#		52,8%	20,8%	17,0%	9,4%	
	% within v ozrast		14,5%	8,6%	16,1%	12,8%	
затрудняюсь ответить	Count		24	8	4	0	36
	% within \$var27#		66,7%	22,2%	11,1%	,0%	
	% within v ozrast		12,4%	6,3%	7,1%	,0%	
Total	Count		193	128	56	39	416

Percentages and totals are based on responses.

a. Group

Если сравнить данные по возрастным категориям респондентов, то можно увидеть, что молодые респонденты в возрасте до 30 лет в большей степени симпатизируют демократам и правозащитникам (18,9%), чем например, респонденты старше 50 лет. С другой стороны, среди пожилых респондентов значительно больше тех, кто придерживается коммунистической идеологии (17,2%). Среди респондентов среднего возраста от 31 до 40 лет больше сторонников нынешнего политического курса (20,5%), по сравнению с другими возрастными категориями.

Сторону тех, кто выступает против наплыва в страну мигрантов, в большей степени поддерживают молодые респонденты (43,5%), за возрождение в стране православных традиций выступают люди молодого и среднего возраста – 44,7% и 40,4% соответственно.

Полученные проценты соответствуют отношению частот к числу допустимых наблюдений. К сожалению, длина меток переменных ограничивается лишь двадцатью символами, поэтому варианты ответа приводятся в сокращенном виде.

Для множественных ответов SPSS не проводит проверку значимости с помощью критерия хи-квадрат.

**Задание.** 1. (по массиву файла orgos.sav) Выяснить мнение о проблемах молодежи и наиболее важных ценностях в зависимости от пола респондентов. Создать парное распределение переменных с множественными ответами «Какие из проблем представляют опасность для молодежи в округе?», «Что важно для достижения успеха в жизни?» и переменной «пол». Проанализировать полученные данные.

## 7. Анализ взаимосвязей качественных и количественных переменных. Средние значения

Достаточно распространенная задача – демонстрация средних значений каких-то количественных показателей в социальных, демографических или каких-то иных группах. Например, необходимо сопоставить величину средней заработной платы в группах респондентов, опрошенных в разных типах населенных пунктов, либо сравнить средний возраст людей, проголосовавших за разных кандидатов на выборах, и т.п.

Построение статистических таблиц в рамках пакета программ SPSS выполняется с помощью команды Means (Средние) в рамках блока команд Compare Means.

В главном меню команды Means видно, что необходимо задать два типа переменных. Первый тип переменных – Dependent List (зависимые переменные) – это переменные, средние значения которых необходимо вычислять. Например, переменная «доход в месяц». Второй тип переменных – Independent List (независимые переменные) – это те переменные, которые определяют разделение всей совокупности опрошенных на определенные группы. Например, переменная «место жительства». Самый большой в среднем доход демонстрируют респонденты г. Сургута – около 17 тыс. рублей в месяц, самый низкий – опрошенные из Березовского района (8 тыс. рублей).

### Report

Доход в месяц			
Место жительства	Mean	N	Std. Deviation
г. Сургут	16,995	233	14,7374
г. Нижнеартовск	13,537	145	13,6886
г. Ханты-Мансийск	11,163	46	10,8234
г. Урай	14,458	24	11,4245
г. Мегион	9,221	52	4,1117
г. Пыть-Ях	12,103	29	7,6221
Сургутский р-н	11,337	75	6,2629
Октябрьский р-н	10,427	24	4,7557
Березовский р-н	7,900	15	2,2377
Total	13,737	643	12,2770

Исходя из данных таблицы, мы можем визуально убедиться в наличии различий в средних доходах респондентов, проживающих в разных территориях ХМАО. Но, например, различия в средних возрастах респондентов разных территорий визуально могут быть неочевидны. Для этого требуются математические доказательства.

Наличие либо отсутствие различий средних значений можно вычислить с помощью команды T-test и One-Way ANOVA (дисперсионный анализ).

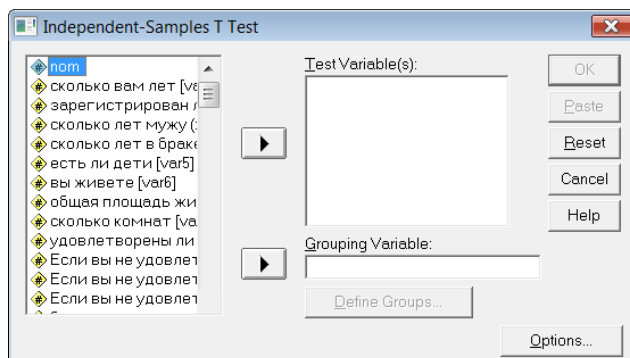
Команда T-test (тест Стьюдента) решает задачу доказательства наличия различий средних значений количественной переменной в усеченном виде, а именно в случае, когда имеются лишь две сравниваемые группы.

### 7.1. Команда T-test (тест Стьюдента) для сравнения двух независимых выборок

Пример независимых выборок – разные населенные пункты, пол респондента.

Нужно выбрать в подменю команду Independent-Samples T Test... (t-тест для независимых выборок). Откроется диалоговое окно Independent-Samples T Test.

- В списке исходных переменных щелкнуть на переменной «доход» и щелчком на кнопке с треугольником перенести ее в список тестируемых переменных (Test Variable(s)).
- Таким же способом перенести переменную «место жительства» -terr- в поле Grouping Variable (Группирующая переменная).
- Щелчком на кнопке Define Groups... (Определить группы) открывается окно, в котором можно ввести значения двух категорий для группирующей переменной «место жительства». Мы будем сравнивать две группы, удовлетворяющие условиям соответственно terr = 1 и terr = 9. Поэтому внесите в поле Group 1 (Группа 1) значение 1, а в поле Group2 — значение 9.



- Щелчком на кнопке Continue вернуться в основное диалоговое окно.
- Теперь следует выяснить, какие параметры установлены по умолчанию. Щелкнуть для этого на кнопке Options... (Параметры). Не изменяя настроек, щелкнуть на кнопке Continue и вернуться в основное диалоговое окно. Запустить t-тест, щелкнув на ОК. В окне просмотра появятся следующие результаты:

Group Statistics

Место жительства	N	Mean	Std. Deviation	Std. Error Mean
Доход в месяц г. Сургут	233	16,995	14,7374	,9655
Березовский р-н	15	7,900	2,2377	,5778

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Доход в месяц	Equal variances assumed	5,320	,022	2,384	246	,018	9,0948	3,8151	1,5805	16,6092
	Equal variances not assumed			8,083	136,926	,000	9,0948	1,1251	6,8699	11,3198

Первая часть таблицы – статистический тест Ливина (Levene's Test for Equality of Variances) – тест проверки равенства дисперсий. F-статистика этого теста равна 5,320, а значимость этой статистики Sig. – 0,022. Значимость меньше 0,05. Дисперсии двух распределений статистически значимо различаются. Вторая часть таблицы (t-test for Equality of Means) – проверка равенства средних. Включает две строки – Equal variances assumed - соответствует равным дисперсиям и Equal variances not assumed – соответствует различным дисперсиям.

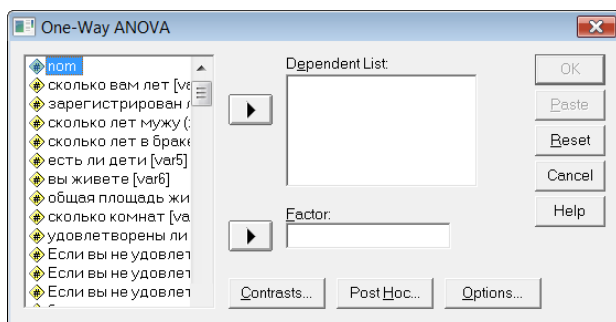
Полученные результаты говорят о различиях в средних доходах по двум территориям – г. Сургут и Березовский район. Различия статистически достоверны на высоком уровне значимости ( $p=0,000$ ).

## 7.2. Однофакторный дисперсионный анализ

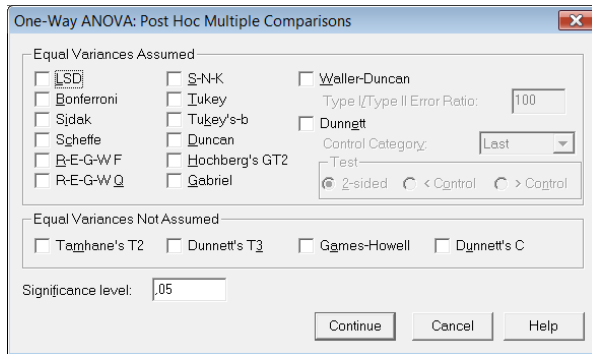
T-тест позволяет сопоставить только две градации. Проанализировать средние значения всех переменных можно с помощью метода однофакторного дисперсионного анализа One-Way ANOVA. Задача может быть сформулирована следующим образом: оказывает ли значимое влияние на значение некоторой количественной переменной интересующая нас переменная, которая измерена на номинальном или порядковом уровне?

Та переменная, которая должна оказывать влияние на конечный результат называется *фактором*. Например, в модели, объясняющей различия в готовности респондентов голосовать на выборах их возрастом, переменная «Собираетесь ли вы голосовать на выборах» будет выступать фактором. Конкретное значение фактора (например, готовность голосовать) называют *уровнем фактора*. Значение измеряемого признака (в нашем случае – возраст) называют *откликом*.

- Выберите в подменю команду One-Way ANOVA... (Однофакторный дисперсионный анализ) Подобная возможность есть и в первом пункте подменю (Means...), но она дает значительно более ограниченные возможности для анализа, и поэтому мы ее не рассматриваем. Появится диалоговое окно One-Way ANOVA.
  - Перенесите переменную «возраст» в список зависимых переменных (Dependent List), а переменную «собираетесь ли вы голосовать на выборах» — в поле Factor (Фактор).
  - Посмотрите, какие параметры можно задать для этого теста (кнопка Options...). Задайте вывод описательной статистики (флажок Descriptive) и проверку на гомогенность дисперсий (флажок Homogeneity-of-variance).
  - С помощью флажка Means plot (График средних) можно построить диаграмму, на которой будут изображены средние значения для каждой выборки.
  - Запустите тест, щелкнув на ОК.
- Выведенные результаты будут содержать:
- результаты теста Ливина на гомогенность дисперсий,
  - типовую схему дисперсионного анализа, включая вероятность ошибки  $p$  (значимость) для оценки общей значимости,
  - график средних.







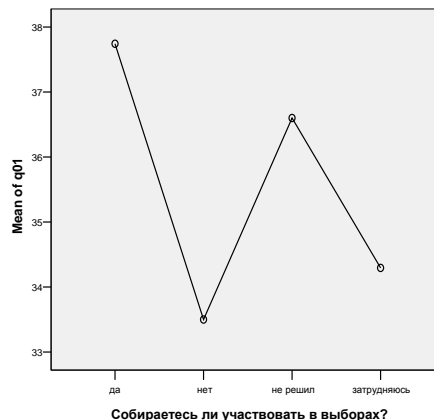
#### Descriptives

Возраст								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
да	264	37,74	11,895	,732	36,30	39,18	18	78
нет	72	33,50	13,694	1,614	30,28	36,72	18	82
не решил	70	36,60	10,661	1,274	34,06	39,14	18	68
затрудняюсь	17	34,29	13,004	3,154	27,61	40,98	20	64
Total	423	36,69	12,140	,590	35,53	37,85	18	82

Установка флажка Descriptives показывает: количество наблюдений, средние значения, стандартные отклонения и стандартные ошибки средних, 95 % доверительные интервалы, минимумы и максимумы для всех слоев фактора.

#### Test of Homogeneity of Variances

Возраст			
Levene Statistic	df 1	df 2	Sig.
1,051	3	419	,370



Критерий однородности дисперсий (Test of Homogeneity of Variances) позволяет вывести информацию о степени пригодности данных к дисперсионному анализу. Значимость критерия однородности дисперсии Ливина – 0,370 (больше 0,05) показывает, что дисперсии для каждой из групп статистически достоверно не различаются. Следовательно, результаты ANOVA могут быть признаны корректными.

#### ANOVA

Возраст					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1123,233	3	374,411	2,569	,054
Within Groups	61072,81	419	145,759		
Total	62196,05	422			

Таблица однофакторного дисперсионного анализа. Самым важным в этой таблице является уровень значимости Sig.  $p = 0,054$ . Он показывает, что разность между средними значениями переменной «готовность голосовать» для разных возрастов статистически незначительна.

#### Задание.

- По массиву данных файла opros.sav найти средние значения доходов респондентов в месяц. С помощью t-теста для независимых выборок провести анализ различий средних значений доходов от места жительства респондентов.
- По массиву данных файла opros.sav выяснить насколько доход в месяц (var42) различен для респондентов с разным уровнем образования (var60). Провести однофакторный дисперсионный анализ. Проанализировать полученные результаты.

## 8. Регрессионный анализ

Целью регрессионного анализа является измерение связи между зависимой переменной и одной (парный регрессионный анализ) или несколькими (множественный) независимыми переменными.

Независимые переменные называют также факторными, объясняющими, определяющими, регрессорами и предикторами. Зависимую переменную иногда называют определяемой, объясняемой, «откликом». Регрессионный анализ это не только удобный инструмент тестирования гипотез, но и эффективный метод моделирования и прогнозирования.

Первые действия при использовании регрессионного анализа будут практически идентичны вычислениям коэффициента корреляции. На первом этапе строятся диаграммы рассеяния, проводится статистический описательный анализ переменных и вычисляется линия регрессии. Линии регрессии строятся методом наименьших квадратов.

Например, нам нужно выяснить существует ли корреляционная связь между переменными «общая площадь жилья» и «удовлетворенность жилищными условиями». Р-уровень со значением 0,000 (меньше 0,05) и коэффициент Пирсона (для порядковых переменных) со значением - 0,368 говорят о достаточно значимой, отрицательной статистической связи между переменными.

Symmetric Measures

	Value	Asy mp. Std. Error <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
Interval by Interval Pearson's R	-,368	,050	-7,154	,000 <sup>c</sup>
Ordinal by Ordinal Spearman Correlation	-,393	,050	-7,712	,000 <sup>c</sup>
N of Valid Cases	328			

a. Not assuming the null hypothesis.

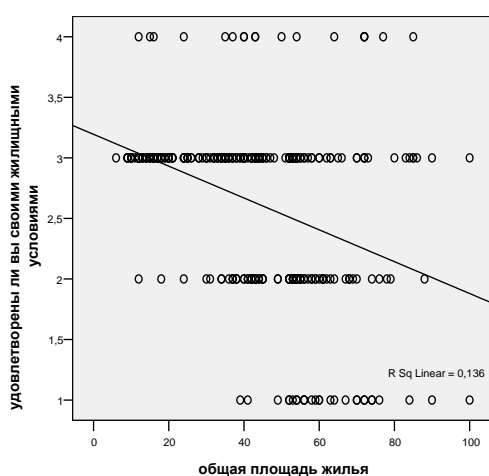
b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Построим диаграмму рассеяния.

### Диаграмма рассеяния

Для графического представления подобной связи можно использовать прямоугольную систему координат с осями, которые соответствуют обеим переменным. Каждая пара значений маркируется при помощи определенного символа.



Такой график, называемый «диаграммой рассеяния» для двух зависимых переменных можно построить путём вызова меню Graphs... (Графики) Scatter plots... (Диаграммы рассеяния).

Образовавшееся скопление точек показывает, что чем меньше площадь жилья, тем не удовлетвореннее молодежь своими жилищными условиями. Это, конечно же, не является неожиданностью; данный пример был выбран, чтобы продемонстрировать наличие явной связи.

Статистика говорит о корреляции между двумя переменными и указывает силу связи при помощи некоторого критерия взаимосвязи, который получил название *коэффициента корреляции*. Этот коэффициент, всегда обозначаемый латинской буквой *r*, может принимать значения между -1 и +1, причём если значение находится ближе к 1, то это означает наличие сильной связи, а если

ближе к 0, то слабой.

Переменная «удовлетворенность жилищными условиями» включает следующие значения: 1- «полностью удовлетворены», 2 - «отчасти удовлетворены», 3 - «нет», 4 - «трудно сказать».

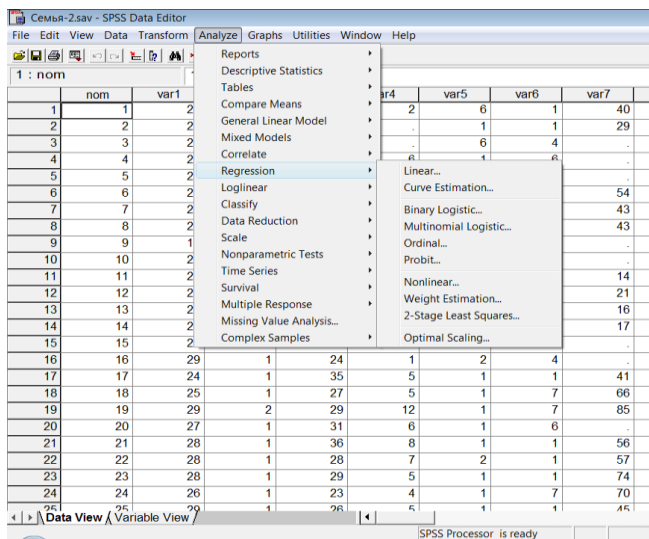
### 8.1. Парный регрессионный анализ

- Чтобы вызвать регрессионный анализ в SPSS, выберите в меню Analyze... (Анализ) Regression... (Регрессия) Откроется соответствующее подменю.

Разделы этой главы соответствуют опциям вспомогательного меню. Причём при изучении линейного регрессионного анализа снова будут проведено различие между простым анализом (одна независимая переменная) и множественным анализом (несколько независимых переменных). Собственно говоря, никаких принципиальных отличий между этими видами регрессии нет, однако простая линейная регрессия является простейшей и применяется чаще всех остальных видов.

СИОННО-  
ИМЕТЬ  
ТО ЖЕ  
ВЫЯВЛЯ-  
НОЙ ОТ  
ЛЮБОЙ

ЛИВЫ И  
МЕННАЯ  
ДВУХ  
БУДЕТ  
СИЯ.  
СЯ ПО-  
ПОЛЗУ-  
СЯТСЯ К



анализировать и нелинейные связи между переменными, которые относятся к интервальной шкале. Для этого предназначен метод нелинейной регрессии.

Принципиальная идея регрессионного анализа состоит в том, что, имея общую тенденцию для переменных – в виде линии регрессии – можно предсказать значение зависимой переменной, имея значения независимой.

Этот вид регрессии лучше всего подходит для того, чтобы продемонстрировать основополагающие принципы регрессионного анализа. Рассмотрим для этого диаграмму рассеяния, которая иллюстрирует зависимость показателя удовлетворенности жильем от общей площади жилья. Можно легко заметить очевидную связь: обе переменные развиваются в одном направлении и множество точек, соответствующих наблюдаемым значениям показателей, явно концентрируется (за некоторыми исключениями) вблизи прямой (прямой регрессии). В таком случае говорят о линейной связи.

$$y = b \cdot x + a$$

где  $b$  — регрессионные коэффициенты,  $a$  — константа, задающая смещение по оси ординат.

Смещение по оси ординат соответствует точке на оси  $y$  (вертикальной оси), где прямая регрессии пересекает эту ось. Коэффициент регрессии  $b$  через соотношение

$$b = \tan(a)$$

указывает на угол наклона прямой.

При проведении простой линейной регрессии основной задачей является определение параметров  $b$  и  $a$ . Оптимальным решением этой задачи является такая прямая, для которой сумма квадратов вертикальных расстояний до отдельных точек данных является минимальной.

Если мы рассмотрим показатель «удовлетворенность жильем» ( $var1$ ) как зависимую переменную ( $y$ ), а исходную величину «общая площадь жилья»  $var2$  как независимую переменную ( $x$ ), то тогда для проведения регрессионного анализа нужно будет определить параметры соотношения

$$var1 = b \cdot var2 + a$$

После определения этих параметров, зная исходный показатель общей площади жилья, можно спрогнозировать показатели удовлетворенности жильем.

В нашем примере, простой регрессионный анализ позволяет получить следующие таблицы.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,368 <sup>a</sup>	,136	,133	,669

a. Predictors: (Constant), общая площадь жилья

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22,881	1	22,881	51,182	,000 <sup>a</sup>
	Residual	145,741	326	,447		
	Total	168,622	327			

a. Predictors: (Constant), общая площадь жилья

b. Dependent Variable: удовлетворены ли вы своими жилищными условиями

Для проведения линейного регрессивного анализа зависимая переменная должна интервальную (или порядковую) шкалу. В время, бинарная логистическая регрессия ет зависимость дихотомической переменной от другой переменной, относящейся к шкале.

Те же условия применения справедливы для пробит-анализа. Если зависимая переменная является категориальной, но имеет более категорий, то здесь подходящим методом мультиномиальная логистическая регрессия. Новшеством уже в 10 версии SPSS является порядковая регрессия, которую можно использовать, когда зависимые переменные относятся к порядковой шкале. И, наконец, можно

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,194	,087		36,567	,000
	общая площадь жилья	-,013	,002	-,368	-7,154	,000

a. Dependent Variable: удовлетворены ли вы своими жилищными условиями

Уравнение регрессии будет выглядеть таким образом:  $Var1 = -0,013 \cdot var2 + 3,194$

Можно вычислить какова будет удовлетворенность жилищными условиями, если общая площадь жилья составит 70 кв.м.  $-0,013 \cdot 70 + 3,194 = 2,28$ , таким образом, удовлетворенность будет иметь значение 2 – «отчасти».

Одним из главных показателей регрессионного анализа является множественный коэффициент корреляции R – коэффициент корреляции между исходными и предсказанными значениями зависимой переменной. В парном регрессионном анализе он равен обычному коэффициенту корреляции Пирсона между зависимой и независимой переменной, в нашем случае – 0,368. Чтобы содержательно интерпретировать множественный R, его необходимо преобразовать в коэффициент детерминации. Это делается так же, как и в корреляционном анализе – возведением в квадрат. Коэффициент детерминации R-квадрат ( $R^2$ ) показывает долю вариации зависимой переменной, объяснимую независимой (независимыми) переменными.

В нашем случае,  $R^2 = 0,136$ . Чем больше величина коэффициента детерминации, тем выше качество модели.

Другим показателем качества модели является стандартная ошибка оценки (Std.Error of Estimate). Это показатель того насколько точки «разбросаны» вокруг линии регрессии. Мерой разброса для интервальных переменных является стандартное отклонение. Чем выше его значение, тем сильнее разброс, тем хуже модель. В нашем случае, стандартная ошибка составляет 0,669. Именно на эту величину наша модель будет «ошибаться в среднем» при прогнозировании значения переменной «удовлетворенность жильем».

Регрессионная статистика включает в себя также дисперсионный анализ (ANOVA). С его помощью выясняем: 1) какая доля вариации (дисперсии) зависимой переменной объясняется независимой переменной; 2) какая доля дисперсии зависимой переменной приходится на остатки (необъясненная часть), 3) каково отношение этих двух величин (F-отношение). Дисперсионная статистика очень важна. Для выборочных исследований она показывает, насколько вероятно наличие связи между независимой и зависимой переменными в генеральной совокупности, для сплошных исследований – проверяют «не случайность» выявленной статистической закономерности.

В нашем случае F-отношение 51,182 значимо на уровне 0,000. Соответственно, мы можем с уверенностью отвергнуть нулевую гипотезу (что обнаруженная связь носит случайный характер).

## 8.2. Множественный регрессионный анализ

В общем случае в регрессионный анализ вовлекаются несколько независимых переменных. Это, конечно же, наносит ущерб наглядности получаемых результатов, так как подобные множественные связи в конце концов становится невозможно представить графически.

В случае множественного регрессионного анализа речь идёт необходимо оценить коэффициенты уравнения

$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_n \cdot x_n + a,$$

где n — количество независимых переменных, обозначенных как  $x_1$  и  $x_n$ , а — некоторая константа.

Переменные, объявленные независимыми, могут сами коррелировать между собой; этот факт необходимо обязательно учитывать при определении коэффициентов уравнения регрессии для того, чтобы избежать ложных корреляций.

При работе с множественной регрессией, в отличие от парной, необходимо определять алгоритм анализа. Стандартный алгоритм включает в итоговую регрессионную модель все имеющие предикторы. Пошаговый алгоритм предполагает последовательное включение (исключение) независимых переменных, исходя из объяснительного «веса». Пошаговый метод хорош, когда имеется много независимых переменных; он «очищает» модель от откровенно слабых предикторов, делая ее более компактной и лаконичной.

Дополнительным условием корректности множественной регрессии (наряду с интервальностью, нормальностью, линейностью) является отсутствие мультиколлинеарности – наличия сильных корреляционных связей между независимыми переменными.

Проведем множественный регрессионный анализ зависимой переменной «желание взять ипотечный кредит» (var1) и независимыми переменными «общая площадь жилья» (S), «возможность кредита при условии его погашения при рождении детей» (A), «доход» (D).

- Выберите в меню Analyze... (Анализ) Regression...(Регрессия) Linear... (Линейная)

Поместите переменную var1 в поле для зависимых переменных, объявите переменные: ««общая площадь жилья», «согласие на кредит, при условии погашения его при рождении детей», «доход» независимыми. В меню Method установлен по умолчанию – Enter (Включение), соответствующий стандартному алгоритму. Этот

метод соответствует одновременной обработке всех независимых переменных, выбранных для анализа, и поэтому он может рекомендоваться для использования только в случае простого анализа с одной независимой переменной.

Для множественного анализа следует выбрать один из пошаговых методов. При выборе пошагового алгоритма в списке Method – Forward (Прямой) – пошаговое включение переменных с проверкой на значимость их частной корреляции с критерием. В результате в уравнение включаются все переменные, имеющие значимую частную корреляцию с переменной-критерием. Включение производится в порядке возрастания  $r$ -уровня.

При выборе Backward (Обратный) – пошаговый метод, сначала включающий в уравнение регрессии все независимые переменные, а затем поочередно удаляющий все переменные, чья корреляция с критерием имеет уровень значимости выше заданного порогового значения. Как правило, пороговым значением является  $p=0,1$ .

При выборе Stepwise (По шагам) – комбинация пошаговых методов Forward (Прямой) и Backward (Обратный). Основная идея – изменение доли влияния независимой переменной на критерий при появлении в уравнении других независимых переменных. Если влияние какой-либо из включенных переменных становится слишком слабым, то она исключается из уравнения. Подобный метод используется в регрессионном анализе наиболее часто.

Применим его к нашему случаю.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,185 <sup>a</sup>	,034	,032	,848
2	,214 <sup>b</sup>	,046	,041	,844

a. Predictors: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?

b. Predictors: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?, ваш доход

**ANOVA<sup>c</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10,033	1	10,033	13,965	,000 <sup>a</sup>
	Residual	282,356	393	,718		
	Total	292,390	394			
2	Regression	13,329	2	6,664	9,362	,000 <sup>b</sup>
	Residual	279,061	392	,712		
	Total	292,390	394			

a. Predictors: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?

b. Predictors: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?, ваш доход

c. Dependent Variable: Хотели бы вы взять ипотечный кредит?

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,500	,090		16,745	,000
	Согласились бы вы взять кредит по условию погашения его при рождении детей?	,192	,051	,185	3,737	,000
2	(Constant)	1,712	,133		12,893	,000
	Согласились бы вы взять кредит по условию погашения его при рождении детей?	,184	,051	,178	3,588	,000
	ваш доход	-,078	,036	-,106	-2,152	,032

a. Dependent Variable: Хотели бы вы взять ипотечный кредит?

**Excluded Variables<sup>c</sup>**

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	общая площадь жилья	-,090 <sup>a</sup>	-1,811	,071	-,091	,987
	ваш доход	-,106 <sup>a</sup>	-2,152	,032	-,108	,995
2	общая площадь жилья	-,086 <sup>b</sup>	-1,742	,082	-,088	,985

a. Predictors in the Model: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?

b. Predictors in the Model: (Constant), Согласились бы вы взять кредит по условию погашения его при рождении детей?, ваш доход

c. Dependent Variable: Хотели бы вы взять ипотечный кредит?

Как видно из таблиц, переменная «общая площадь жилья» исключается из анализа. Значимыми переменными остаются «доход» и «согласие взять кредит при условии погашения его при рождении детей». Перемен-

ная «уровень дохода» отрицательно влияет на желание взять ипотечный кредит, возможности взять ипотечный кредит в большей степени рассматривают респонденты с небольшим доходом.

Уравнение регрессии для прогнозирования значения var1 (возможность взять ипотечный кредит) выглядит следующим образом:

$$\text{Var1} = 0,184 * A - 0,78 * D + 1,712$$

Важным моментом является анализ остатков, то есть отклонений наблюдаемых значений от теоретически ожидаемых. Остатки должны появляться случайно (то есть не систематически) и подчиняться нормальному распределению. Это можно проверить, если с помощью кнопки Charts... (Диаграммы) построить гистограмму остатков.

Проверка на наличие систематических связей между остатками соседних случаев (что, однако, является уместным только при наличии так называемых данных с продольным сечением), может быть произведена при помощи теста Дарбина-Ватсона (Durbin-Watson) на автокорреляцию. Этот тест вычисляет коэффициент, лежащий в диапазоне от 0 до 4. Если значение этого коэффициента находится вблизи 2, то это означает, что автокорреляция отсутствует. Тест Дарбина-Ватсона можно активировать через кнопку Statistics (Статистические характеристики).

Ещё одной дополнительной возможностью является задание переменной отбора в диалоговом окне Linear Regression (Линейная регрессия). Здесь, с помощью кнопки Rule... (Правило) в диалоговом окне Linear Regression: Define Selection Rule (Линейная регрессия: ввод условия отбора), Вы получаете возможность при помощи избирательного признака сформулировать условие, которое будет ограничивать количество случаев, вовлеченных в анализ.

## 9. Факторный анализ

### 9.1 Исследование структуры данных

Собирая данные, исследователь руководствуется определенными гипотезами. Полученная в ходе исследования информация относится к избранному предмету и теме исследования, но нередко она представляет собой сырой материал, в котором можно изучить структуру показателей, характеризующих объекты, а также выявить однородные группы объектов. Информацию лучше представить в геометрическом пространстве, лаконично отразить ее особенности в классификации объектов и переменных. Такая работа создает предпосылки к выявлению типологий объектов и формулированию «социального пространства», в котором обозначены расстояния между объектами наблюдения, позволяет наглядно представить свойства объектов.

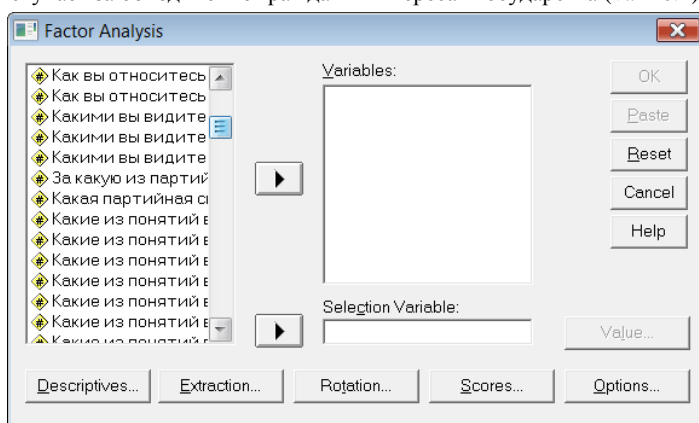
Факторный анализ является одним из наиболее мощных статистических средств анализа данных. В его основе лежит процедура объединения групп коррелирующих друг с другом переменных («корреляционных узлов») в несколько факторов.

Цель факторного анализа – сконцентрировать исходную информацию, выражая большое число рассматриваемых признаков через меньшее число более емких внутренних характеристики, которые, однако, не поддаются непосредственному измерению (являются латентными).

Факторный метод будет изложен на примере опроса, проведенного с целью выяснения политических ориентаций жителей города. В ходе опроса респондентам предложили выбрать высказывания, соответствующие их мнению, и отдать свой голос в поддержку тех, кто:

1. согласен с нынешним политическим курсом (var 21.1)
2. выступает с критикой нынешнего политического курса (var 21.2)
3. выступает за вхождение России в западную цивилизацию (var 22.1)
4. против сближения России с Западом (var 22.2)
5. выступает за неведение жесткого порядка (var 23.1)
6. считает главным демократию, политические и личные свободы граждан (var 23.2)
7. выступает за усиление влияния Церкви на государство (var 24.1)
8. считает, что государство и Церковь не должны вмешиваться в жизнь граждан (var 24.2)
9. считает, что государство не должно вмешиваться в свободную рыночную экономику (var 25.1)
10. выступает за государственный контроль бизнеса (var 25.2)
11. выступает за объединение граждан в интересах государства (var 26.1)

12. считает, что граждане должны добиваться успеха сами (var 26.2)



Оценки ставились по двухбалльной шкале: 1) поддерживаю, 2) не поддерживаю.

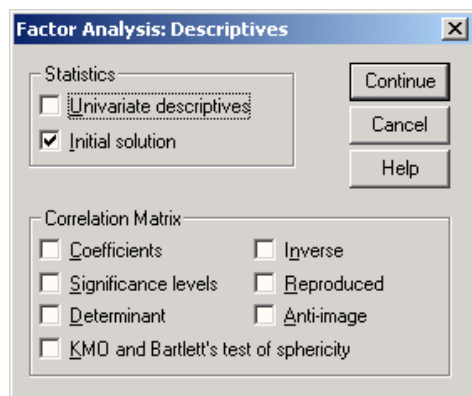
Для факторного анализа:

- Выберите в меню Analyze (Анализ) Data Reduction (Сокращение объема данных) Factor... (Факторный анализ)

Откроется диалоговое окно Factor Analysis (Факторный анализ)

Переменные var21-...var26 поместите в поле тестируемых переменных и ознакомьтесь с возможностями, предлагаемыми различными кнопками этого диалогового меню.

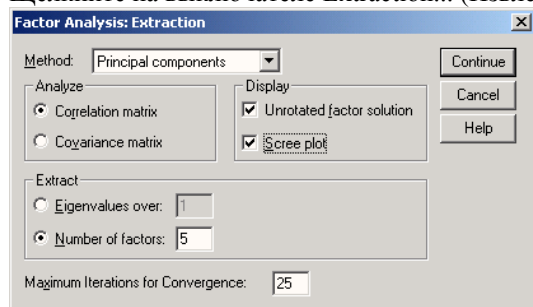
После щелчка по кнопке Descriptive Statistics (Дескриптивные статистики) оставьте вывод первичных результа-



тов, которые включают в себя первичные относительные дисперсии простых факторов, собственные значения и процентные доли объясненной дисперсии. Довольно часто бывает необходим также вывод одномерных статистик и корреляционных коэффициентов. В группе Correlation Matrix (Корреляционная матрица) целесообразно отметить флажком KMO and Bartlett test of sphericity (Критерии КМО и сферичности Бартлетта), вычисляется два критерия – на многомерную нормальность (Бартлетта) и адекватность выборки (КМО определяет применимость факторного анализа к выбранным переменным).

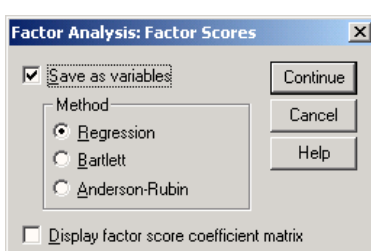
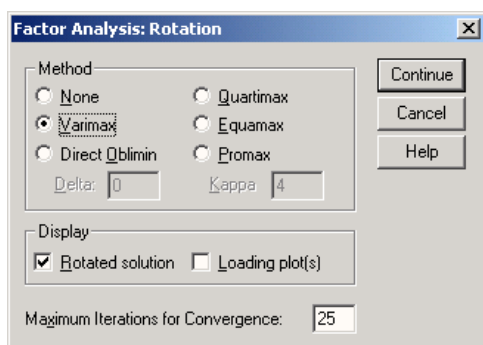
С помощью кнопки Extraction... (Отбор) можно выбрать метод отбора. Если оставить здесь анализ главных компонентов, установленный по умолчанию, то количество отобранных в этом случае факторов приравняется к числу собственных значений, превосходящих единицу. Также есть возможность собственноручно указать это количество.

Щёлкните на выключателе Extraction... (Извлечение), оставьте установку Principal components (Анализ главных



компонентов). В нашем примере количество факторов сознательно ограничим тремя. Если бы мы не сделали такого ограничения, то в соответствии с начальными установками было бы создано двенадцать факторов, количество, которое очень тяжело поддается обзору.

Можно построить график собственных значений или диаграмму каменистой осыпи, установив флажок на Scree plot. Точками показаны соответствующие собственные значения, в пространстве двух координат. Этот тип диаграммы обычно используется при определении достаточного числа факторов перед вращением. При этом руководствуются следующим правилом: оставлять нужно лишь те факторы, которым соответствуют первые точки на графике до того, как кривая станет более пологой.

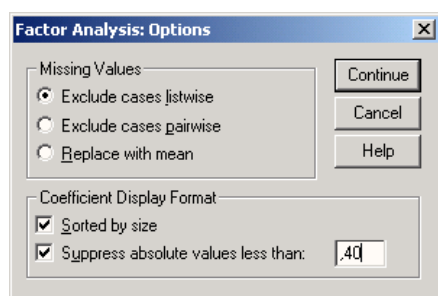


Выключатель Rotation... (Вращение) позволяет выбрать метод вращения. Вращение требуется потому, что изначально структура факторов, будучи математически корректной, как правило, трудна для интерпретации. Целью вращения является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам.

Факторные нагрузки можно представить как коэффициенты корреляции каждой переменной с каждым из выявленных факторов. Чем теснее связь переменной с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую, а отрицательный знак – на обратную связь переменной с фактором.

Активируйте метод варимакса (Varimax) и оставьте активированным вывод повернутой матрицы факторов. Далее вы можете организовать вывод факторных нагрузок в графическом виде, в котором первые три фактора будут представлены в трёхмерном пространстве; в случае наличия только двух факторов в слое приводится только одно изображение. При этом установите флажок на Loading plot(s).

Если Вы хотите найти значения факторов и сохранить их в виде дополнительных переменных задействуйте выключатель Scores... (Значения) и отметьте Save as variables (Сохранить как переменные). По умолчанию установлен регрессионный метод.



Выключатель Options... (Опции) предназначен для обработки пропущенных значений. Здесь обеспечивается возможность заменить пропущенные значения средними значениями соответствующих переменных.

При факторном анализе постоянно появляются сообщения об ошибках, например 2,56E-02 и т.п. Действительно такой формат вывода в глазах непосвященного пользователя очень портит картину всей таблицы. Это, так называемый, Е-формат, знакомый всем программистам

по языку Фортран (Fortran), где буква E соответствует 10 в некоторой степени; для числа 2,5E-02 можно было бы записать и 0,0256.

Можно запретить вывод малых факторных нагрузок и для этого установим граничное значение выводимых нагрузок равным 0,4. Достоинство этого шага состоит в том, что устраняется непривлекательное отображение малых значений в Е-формате. Для этого активируйте опцию Suppress absolute values less than: (Не выводить абсолютные значения меньшие, чем:) и введите предельное значение 0,4.

- Для проведения расчётов щёлкните на ОК.
- В окне обзора появятся результаты. Сначала приводятся первичные статистики: Критерий сферичности Бартлетта показывает статистически достоверный результат ( $p < 0,05$ ), данные вполне приемлемы для факторного анализа.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,461
Bartlett's Test of Sphericity	Approx. Chi-Square	1252,661
	df	66
	Sig.	,000



Communalities

	Initial	Extraction
поддержка нынешнего политического курса	1,000	,581
поддержка жесткой критики политического курса	1,000	,492
за наведение жесткого порядка	1,000	,369
за демократические свободы	1,000	,653
за свободную рыночную экономику	1,000	,366
за государственный контроль бизнеса	1,000	,494
граждане должны добиваться успеха сами	1,000	,515
за объединение граждан в интересах государства	1,000	,731
за усиление влияния Церкви	1,000	,351
за невмешательство государство и церкви в жизнь граждан	1,000	,534
за вхождение России в западную цивилизацию	1,000	,639
против сближения России с Западом	1,000	,337

Extraction Method: Principal Component Analysis.

В таблице Communalities перечислены переменные и общности. Столбцы второй таблицы Total Variance Explained содержат характеристики выделенных факторов: их порядковые номера, суммы квадратов нагрузок, процент общей дисперсии, обусловленной фактором, и соответствующий кумулятивный (накопленный) процент (до и после вращения). Чем больше процент дисперсии, обусловленный фактором, тем больший вес имеет данный фактор. А чем больше кумулятивный процент, накопленный к последнему фактору, тем более состоятельным является факторное решение. Если он составляет менее 50%, следует либо сократить количество переменных, либо увеличить количество факторов.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,590	21,584	21,584	2,590	21,584	21,584	2,202	18,353	18,353
2	1,777	14,808	36,392	1,777	14,808	36,392	2,007	16,727	35,079
3	1,696	14,131	50,522	1,696	14,131	50,522	1,853	15,443	50,522
4	1,323	11,025	61,547						
5	1,156	9,636	71,183						
6	1,110	9,248	80,431						
7	,983	8,195	88,626						
8	,374	3,119	91,745						
9	,334	2,784	94,529						
10	,265	2,210	96,739						
11	,206	1,719	98,458						
12	,185	1,542	100,000						

Extraction Method: Principal Component Analysis.

В данном примере насчитывается шесть собственных значений, превосходящих единицу, что означало бы отбор шести факторов, если бы мы не изменили установку по умолчанию Eigenvalues over: 1 (Собственные значения, превосходящие единицу) и не ограничили бы количество рассматриваемых факторов тремя.

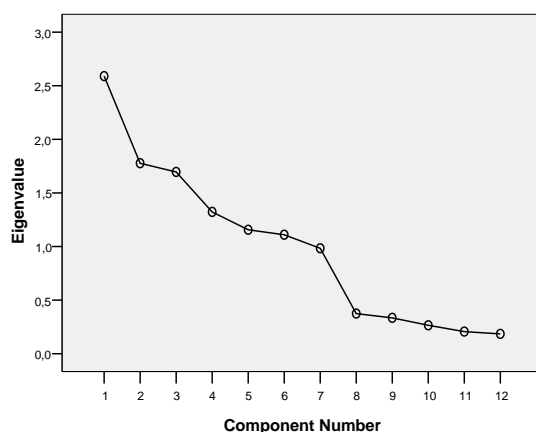
В качестве вспомогательного средства для определения задаваемого числа факторов может послужить специальная точечная диаграмма. Слово Screeplot, употребляемое для обозначения этой диаграммы состоит из

двух частей: английского слова scree, что означает щебень и слова plot, что в английском соответствует графическому представлению. Такая диаграмма служит для того, чтобы маловажные факторы — щебень — можно было отделить от самых значимых факторов. Эти значимые факторы на графике образуют своего рода склон, то есть ту часть линии, которая характеризуется крутым подъёмом. В приведенной диаграмме такой крутой подъём наблюдается в области первых восьми факторов.

Если посмотреть на график, то можно заметить что склон, то есть область значимых факторов, наблюдается выше восьмого фактора (восьмой, седьмой, шестой, пятый ...), а ниже восьмого фактора (девятый, десятый, одиннадцатый, двенадцатый...) расположился щебень, область незначимых факторов. Можно самостоятельно провести расчет с использованием модели, включающей различное число факторов; в рассмотренном примере было бы уместным произвести сравнение моделей

исходную структуру данных в большинстве более важна таблица вернутых компонент). часть факторного отобранных факторы. торной матрицы нужно имеет наибольшее аб-

Scree Plot



с учётом восьми, семи и шести факторов.

Программа SPSS включает в вывод факторных нагрузок (до вращения). Эти случаев не представляют интереса, для нас Rotated Component Matrix (Матрица пере-

Здесь начинается самая интересная лиза: мы должны попытаться объяснить. Для этого в каждой строке повернутой факторной нагрузки, которая

Component Matrix<sup>a</sup>

	Component		
	1	2	3
поддержка нынешнего политического курса	,458	-,348	,500
поддержка жесткой критики политического курса	-,427	,542	-,128
за наведение жесткого порядка	-,592	-,061	-,121
за демократические свободы	,579	,258	,502
за свободную рыночную экономику	,588	-,124	,065
за государственный контроль бизнеса	-,541	,327	,308
граждане должны добиваться успеха сами	,369	,587	-,183
за объединение граждан в интересах государства	-,404	-,416	,628
за усиление влияния Церкви	-,328	-,468	,159
за невмешательство государство и церкви в жизнь граждан	,343	,491	,418
за вхождение России в западную цивилизацию	-,342	,327	,645
против сближения России с Западом	,482	-,291	-,141

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

солютное значение.

Эти факторные нагрузки следует понимать как корреляционные коэффициенты между переменными и факторами. Так переменная var21.1 сильнее всего коррелирует с фактором 2, а именно, величина корреляции составляет 0,549, переменная var21.2 сильнее всего коррелирует с фактором 1 (0,589), переменная же var22.1

Rotated Component Matrix<sup>a</sup>

	Component		
	1	2	3
поддержка нынешнего политического курса	-,371	,549	,378
поддержка жесткой критики политического курса	,589	-,192	-,329
за наведение жесткого порядка	,356	-,455	,187
за демократические свободы	-,100	,793	-,121
за свободную рыночную экономику	-,481	,358	-,078
за государственный контроль бизнеса	,693	,011	,115
граждане должны добиваться успеха сами	,015	,244	-,675
за объединение граждан в интересах государства	,263	,123	,804
за усиление влияния Церкви	,017	-,204	,556
за невмешательство государства и церкви в жизнь граждан	,182	,658	-,260
за вхождение России в западную цивилизацию	,662	,384	,231
против сближения России с Западом	-,572	,092	-,028

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 10 iterations.

коррелирует сильнее всего с фактором 1 (0,356) и т.д. В большинстве случаев включение отдельной переменной в один фактор, осуществляемое на основе коэффициентов корреляции, является однозначным. В исключительных случаях, переменная может относиться к двум факторам одновременно. Могут быть также и переменные, которыми нельзя нагрузить ни один из отобранных факторов.

Варианты мнений, указанные вначале рассмотрения примера, можно отнести в следующем порядке к двум факторам:

*Фактор 1:*

1. поддержка жесткой критики политического курса (var21.2);
2. за наведение жесткого порядка (var 23.1);
3. за государственный контроль бизнеса (var25.2);
4. за вхождение России в западную цивилизацию (var 22.2).

*Фактор 2:*

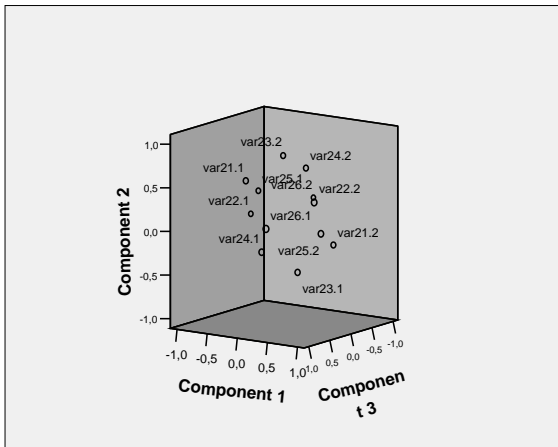
1. поддержка нынешнего политического курса (var21.1);
2. за демократические свободы (var 23.2);
3. за свободную рыночную экономику (var 25.1);
4. граждане должны добиваться успеха сами (var 26.2);
5. против сближения России с Западом (var21.2);
6. за невмешательство государства и церкви в жизнь граждан (var 24.2);

*Фактор 3*

1. за объединение граждан в интересах государства (var 26.1);
2. за усиление влияния Церкви (var 24.1).

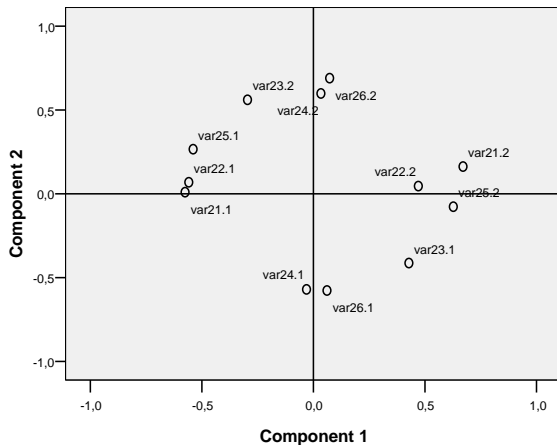
Ниже расположены диаграммы, где представлены факторные нагрузки трех и двух факторов.

Component Plot in Rotated Space



Для интерпретации факторов было бы оптимально, если бы точки лежали ближе к осям и подальше от точки начала отсчёта; тогда каждая переменная имела бы значительную нагрузку для одного фактора и незначительную для другого.

Component Plot in Rotated Space



В соответствии с порядком изложения наши три сгруппированных фактора можно кратко охарактеризовать при помощи следующих выражений: «правые государственники», «либералы», «консерваторы». Однако столь явно, как в приведенном примере факторы удаётся объяснить не всегда. Если нет возможности провести вербальное объяснение факторов, то факторный анализ можно считать неудавшимся.

## 9.2 Значения факторов

Поскольку мы пожелали произвести расчёт значений факторов, то в соответствии с тремя отобранными факторам были сгенерированы три новые переменные, названные fac1\_1, fac2\_1 и fac3\_1, которые содержат

вычисленные значения факторов. Если просмотреть текущий файл после поведения факторного анализа, то можно увидеть имеющиеся нормализованные значения факторов.

По каждому из отобранных фактору для каждого опрошенного было рассчитано специальное факторное значение. Факторное значение, как правило, лежит в пределах —3 до +3.

Рассмотрим факторную переменную `fac1_1`. Она включает следующие элементарные переменные: `var21.2`, `var 23.1`, `var25.2`, `var 22.2`. В качестве метки для этого фактора мы выбрали выражение: "авторитарные государственники". Большое положительное значение фактора означает одобрение элементарных переменных, то есть положений, входящих в этот фактор. Одобрение элементарных переменных, относящихся к первому фактору, тождественно ярко выраженными взглядам, характеризующимися ориентацией на усиление государственного влияния на экономику, установление жесткого государственного порядка, критику нынешнего политического курса.

Рассмотрим факторную переменную `fac2_1`. К ней относятся элементарные переменные: `var21.1`, `var23.2`, `var25.1`, `var26.2`, `var21.2`, `var24.2`. В качестве метки для этого фактора мы выбрали выражение: "либералы". Большое положительное значение фактора означает полное согласие. Полное согласие соответствует мнению о свободной рыночной экономики, поддержки нынешнему политическому курсу, приверженности демократическим принципам.

В заключение рассмотрим факторную переменную `fac3_1`. К ней относятся элементарные переменные `var 26.1`, `var 24.1`. В качестве метки для этого фактора мы выбрали выражение: "консерваторы". Большое положительное значение фактора означает одобрение элементарных переменных. Одобрение элементарных переменных тождественно ярко выраженным консервативным взглядам, соответствующим консолидации граждан в интересах государства, идейное влияние консервативно-традиционных национальных взглядов.

В файле находятся ещё несколько дополнительных переменных, а именно:

• <code>var28</code>	За какую политическую партию Вы проголосовали, если бы выборы состоялись в ближайшее воскресенье?
• <code>vozrast</code>	возраст

Эти переменные можно использовать для того, чтобы устанавливать связи для факторных значений. Самым распространённым методом для этого является разбиение факторных значений на четыре группы процентов. Покажем это на примере первого факторного значения (переменная `fac1_1`).

- Выберите в меню Transform (Трансформировать) Rank Cases... (Создать иерархию наблюдений)

Откроется диалоговое окно Rank Cases (Создать иерархию наблюдений).

- Переменную `fac1_1` перенесите в список тестируемых переменных.

- Щёлкните на выключателе Rank Types... (Типы иерархии), деактивируйте установленную по умолчанию опцию Rank (Ранг) и активируйте опцию Fractional rank as % (Дробный ранг как проценты). Оставьте установленное по умолчанию количество групп равное 4.

- Подтвердите свой выбор нажатием на Continue (Далее) и затем на ОК.

Будет создана переменная `nfac1_1`, которая содержит значения 1 до 4 с примерно равномерной частотой.

- Перейдите в редактор данных и измените имя переменной `nfac1_1` на более удобное имя `avtorit`, в поле метки наберите «правые государственники» и значения присвойте следующие метки: 1 = отсутствует, 2 = слабое, 3 = сильное и 4 = очень сильное. Теперь создадим таблицу сопряженности для новой переменной и переменной `var28` (голосование за политическую партию).

- Выберите в меню Analyze (Анализ) Descriptive Statistics (Дескриптивные статистики) Crosstabs... (Таблицы сопряженности)

- В диалоговом окне Crosstabs (Таблицы сопряженности) переменную `stellung` поместите в поле строк, а переменную `avtorit` в поле столбцов и через выключатель Cells... (Ячейки) сделайте дополнительно запрос на вывод процентных значений по строкам.

В окне просмотра появится следующая таблица сопряженности.

**За какую из партий вы проголосовали бы в ближайшее воскресенье? \* авторитарные  
государственники Crosstabulation**

% within За какую из партий вы проголосовали бы в ближайшее воскресенье?

		авторитарные государственники				Total
		отсутствует	слабое	сильное	очень сильное	
За какую из партий вы проголосовали бы в ближайшее воскресенье?	Союз правых сил		25,0%	25,0%	50,0%	100,0%
	Яблоко	40,0%	60,0%			100,0%
	Родина	28,6%		28,6%	42,9%	100,0%
	ЛДПР	5,7%	22,9%	31,4%	40,0%	100,0%
	Единая Россия	38,3%	28,7%	17,4%	15,7%	100,0%
	КПРФ	7,1%	14,3%	50,0%	28,6%	100,0%
	против всех	15,8%	23,7%	31,6%	28,9%	100,0%
	не стал бы участвовать в выборах	23,3%	20,9%	27,9%	27,9%	100,0%
затрудняюсь ответить		34,9%	18,6%	30,2%	16,3%	100,0%
Total		27,0%	24,0%	25,7%	23,4%	100,0%

Далее, можно создать таблицы сопряженности с переменными «либералы», «консерваторы» и «голосование за политические партии».

**За какую из партий вы проголосовали бы в ближайшее воскресенье? \* либералы Crosstabulation**

% within За какую из партий вы проголосовали бы в ближайшее воскресенье?

		либералы				Total
		отсутствует	слабое	сильное	очень сильное	
За какую из партий вы проголосовали бы в ближайшее воскресенье?	Союз правых сил		75,0%		25,0%	100,0%
	Яблоко	20,0%		60,0%	20,0%	100,0%
	Родина		42,9%	14,3%	42,9%	100,0%
	ЛДПР	8,6%	42,9%	28,6%	20,0%	100,0%
	Единая Россия	24,3%	20,0%	27,8%	27,8%	100,0%
	КПРФ	21,4%	57,1%	7,1%	14,3%	100,0%
	против всех	34,2%	21,1%	26,3%	18,4%	100,0%
	не стал бы участвовать в выборах	23,3%	27,9%	14,0%	34,9%	100,0%
затрудняюсь ответить		44,2%	14,0%	23,3%	18,6%	100,0%
Total		25,3%	25,7%	24,0%	25,0%	100,0%

**За какую из партий вы проголосовали бы в ближайшее воскресенье? \* консерваторы  
Crosstabulation**

% within За какую из партий вы проголосовали бы в ближайшее воскресенье?

		консерваторы				Total
		отсутствует	слабое	сильное	очень сильное	
За какую из партий вы проголосовали бы в ближайшее воскресенье?	Союз правых сил	75,0%		25,0%		100,0%
	Яблоко	40,0%		40,0%	20,0%	100,0%
	Родина	14,3%		57,1%	28,6%	100,0%
	ЛДПР	17,1%	11,4%	31,4%	40,0%	100,0%
	Единая Россия	24,3%	32,2%	20,9%	22,6%	100,0%
	КПРФ	21,4%	21,4%	35,7%	21,4%	100,0%
	против всех	18,4%	23,7%	28,9%	28,9%	100,0%
	не стал бы участвовать в выборах	34,9%	23,3%	16,3%	25,6%	100,0%
затрудняюсь ответить		20,9%	34,9%	25,6%	18,6%	100,0%
Total		24,3%	25,7%	25,0%	25,0%	100,0%

Проанализировав данные трех таблиц, можно прийти к выводу, что например, в среде приверженцев партии «Союз правых сил» более всего распространены идеи авторитарного государства, но в меньшей степени национально-консервативные идеи, а, следовательно, в большей степени для сторонников «СПС» важна ориентация развития России по западному пути. Для сторонников КПРФ важны как идея «сильной государственной власти», так и традиционно-консервативные ценности. Сторонники «Яблока» в большей степени поддерживают либеральные идеи. Среди приверженцев «Единой России» ярко выраженные идейные позиции не проявляются, либеральные и консервативные идеи разделяет примерно половина сторонников «ЕР», «правых государственников» меньше - примерно треть из них.

## 10. Кластерный анализ.

Кластерный анализ (от англ. cluster – группа, пучок) – это процедура, позволяющая классифицировать различные объекты. С его помощью можно разбить респондентов на группы, сходные по ряду признаков.

Цель кластерного анализа — классификация объектов на относительно однородные (однородные) группы исходя из рассматриваемого набора переменных. Объекты в группе относительно схожи между собой и отличаются от объектов в других группах. Если кластерный анализ использовать именно таким образом, то он становится составной частью факторного анализа, так как снижает количество объектов, а не количество переменных, группируя их в меньшее количество кластеров.

С кластерным анализом связаны следующие статистики и понятия.

**План агломерации, объединения (agglomeration schedule).** Дает информацию об объектах (событиях, случаях), которые должны быть объединены на каждой стадии процесса иерархической кластеризации.

**Кластерный центроид (cluster centroid).** Среднее значение переменных для всех случаев или объектов в конкретном кластере.

**Кластерные центры (cluster centers).** Исходные начальные точки в неиерархической кластеризации. Кластеры строят вокруг этих центров, или зерен кластеризации.

**Принадлежность кластеру (cluster membership).** Указывает кластер, к которому принадлежит каждый случай или объект.

**Древовидная диаграмма (дендрограмма) (dendrogram).** Ее также называют древовидный граф — графическое средство для показа результатов кластеризации. Вертикальные линии представляют объединяемые кластеры. Положение вертикальной линии на шкале расстояния (горизонтальная ось) показывает расстояния, при которых объединяли кластеры. Древовидную диаграмму читают слева направо.

**Расстояния между кластерными центрами (distances between cluster centres).** Указывают, насколько разнесены отдельные пары кластеров. Кластеры, которые разнесены широко, ясно выражены и поэтому желательны.

**Сосульчатая диаграмма (icicle diagram).** Это графическое отображение результатов кластеризации. Она названа так потому, что имеет сходство с рядом сосул, свисающих с крыши дома. Сосульчатую диаграмму читают сверху вниз.

**Матрица сходства, или матрица расстояний между объединяемыми объектами (similarity/distance coefficient matrix).** Матрица сходства (расстояний) — это нижняя треугольная матрица, содержащая значения расстояния между парами объектов или случаев.

Программа SPSS реализует три метода кластерного анализа: 2-этапный (Two-step), К-средних (K-means) и иерархический (Hierarchical).

*2-этапный кластерный анализ* позволяет выявить группы (кластеры) объектов по заданным переменным, если эти группы действительно существуют. При этом программа автоматически определяет количество существующих кластеров. Если невозможно определить количество кластеров, все объекты помещаются в один.

Наиболее часто в анализе социологической информации используется иерархический кластер-анализ и метод К-средних.

### 10.1 Иерархический кластер-анализ

Смысл *иерархического кластерного анализа* заключается в следующем. Перед началом кластеризации все объекты считаются отдельными кластерами, которые в ходе алгоритма объединяются. Вначале берется N объектов и между ними попарно вычисляются расстояния. Далее выбирается пара объектов, которые расположены наиболее близко друг от друга, и эти объекты объединяются в один кластер. В результате количество кластеров становится равным N-1. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования определяют способы вычисления расстояния между объектами и определения близости между кластерами.

Выделяют несколько этапов кластерного анализа.

1. выбор переменных-критериев для кластеризации. Например, с целью изучения мотивации электорального выбора выбираются переменные: *персональные электоральные предпочтения респондентов на выборах* (кандидат 1, кандидат 2, кандидат 3); *мотивация этих предпочтений*: (1.1 «он мне нравится», 1.2. «не хочу перемен», 2.1 «меня устраивает его программа», 2.2 «он знает, как решить проблемы страны», 3.1 «я ему доверяю», 3.2 «ему нет достойной замены»); *мотивация голосования за списки политических партий на парламентских выборах* (4.1 «они заставят правительство думать о народе», 4.2 «они смогут решить проблемы страны», 4.3 «устраивает программа партии»).
2. выбор способа измерения расстояния между объектами или кластерами. Для определения расстояния между парой кластеров могут использоваться разные подходы. По умолчанию используется квадрат Евклидова расстояния, согласно которому расстояние между объектами равно сумме квадратов разностей между значениями одноименных переменных объектов.
3. формирование кластеров. Существует два основных метода формирования кластеров метод слияния и метод дробления. В первом случае исходные кластеры увеличиваются путем объединения до тех пор, пока не будет сформирован единственный кластер, содержащий все данные. Метод дробления основан на обратной операции: сначала все данные объединяются в один кластер, который затем делится на ча-

сти до тех пор, пока не будет достигнут желаемый результат. По умолчанию программой SPSS используется метод слияния. Иерархический кластерный анализ организует данные в наглядные «древовидные структуры, или дендрограммы».

Желаемое число кластеров и оценка результатов анализа зависит от целей исследования. В нашем гипотетическом примере наиболее предпочтительными числом кластеров может быть – 3. Мотивы электоральных выборов можно разделить на три группы: первая группа – выбор кандидата 1, мотивы – «1.2. «не хочу перемен», «3.2 «ему нет достойной замены», вторая группа – выбор кандидата 2, мотивы - 2.2 «он знает, как решить проблемы страны», 3.1 «я ему доверяю», 4.1 «они заставят правительство думать о народе», третья группа – выбор кандидата 3, мотивы 1.1 «он мне нравится», 2.1 «меня устраивает его программа», 4.3 «устраивает программа партии».

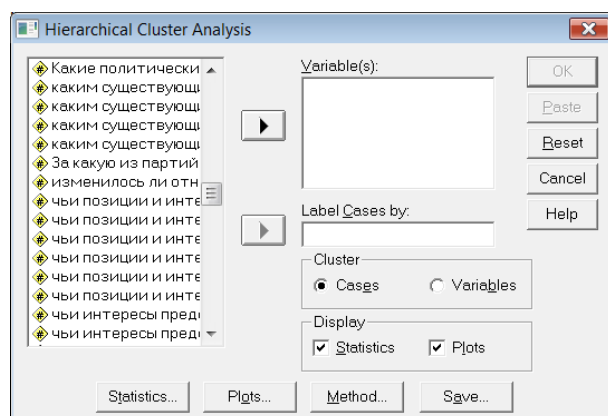
Пошаговый алгоритм иерархического кластерного анализа.

Соберём мотивы электоральных выборов в кластеры при помощи параметров «кандидаты» и «мотивации выбора кандидата» и «предпочтение политпартий».

- Выберите в меню Analyze (Анализ) Classify (Классифицировать) Hierarchical Cluster... (Иерархический кластерный анализ)

Появится диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ)

Переменные «мотивации выбора кандидата» и «предпочтение политпартий» поместите в поле тестируемых переменных, а текстовую переменную «кандидаты» в поле с именем Label cases by: (Наименования (метки) наблюдений:).



- Щелчком по выключателю Statistics... (Статистики) откройте диалоговое окно Hierarchical Cluster Analysis: Statistics (Иерархический кластерный анализ: Статистики) и наряду с выводом последовательности слияния (Agglomeration schedule) активируйте вывод показателя принадлежности к кластеру для каждого наблюдения

щёлкните по выключателю Plots... (Диаграммы). Активируйте опцию вывода древовидной диаграммы (Dendrogram) и посредством опции None (Нет) отмените вывод накопительной диаграммы.

- С помощью кнопки Method... (Метод) можно выбрать метод образования кластеров, а также метод расчета дистанционной меры и меры подобия соответственно.

SPSS предлагает, в общей сложности, семь различных методов объединения. Метод Between-groups linkage (Связь между группами) устанавливается по умолчанию.

Дистанционные меры и меры подобия зависят от вида переменных, участвующих в анализе, то есть выбор меры зависит от типа переменной и шкалы, к которой она относится: интервальная переменная, частоты или бинарные (дихотомические) данные. Для данных, относящихся к интервальной шкале по умолчанию в качестве дистанционной меры устанавливается квадрат евклидова расстояния (Squared Euclidean distance). Оставьте предварительные установки и в поле Transform Values (Преобразовывать значения) установите z-преобразование (стандартизацию) значений. Вернутся назад в главное диалоговое окно и начать расчёт нажатием ОК.

## 10.2 Кластерный анализ при большом количестве наблюдений (Кластерный анализ методом k-средних)

Процедура иерархического кластерного анализа эффективна для малого числа объектов. Ее преимущественно в том, что каждый объект можно рассмотреть в отдельности. Но эта процедура не годится для массивов большого объема

Поэтому при наличии большого количества наблюдений применяют другие методы. В такой ситуации наиболее приемлем алгоритм, носящий название «k-средних». Он реализуется в пакете командой меню K-means. Алгоритм заключается в следующем: выбирается заданное число k точек и на первом шаге эти точки рассматриваются как «центры» кластеров. Каждому кластеру соответствует один центр. Объекты распределяются по кластерам по принципу: каждый объект относится к кластеру с ближайшим к этому объекту центром. Таким образом, все объекты распределились по k кластерам.

Затем заново вычисляют центры этих кластеров, которыми после этого момента считаются координатные средние кластеров. После этого опять распределяют объекты. Вычисление центров и перераспределение объектов происходит до тех пор, пока центры не стабилизируются.

В качестве примера расчёта по этому алгоритму, рассмотрим выборку из результатов опроса 1200 молодых респондентов, в котором задавался вопрос относительно их жизненных стратегий – «что важно для достижения успеха в жизни» с вариантами ответов<sup>7</sup>:

59. Происходить из материально обеспеченной семьи
60. Иметь хорошее образование
61. Иметь амбиции для продвижения по жизни
62. Иметь высокопоставленных родителей
63. Иметь связи в криминальном мире
64. Иметь везение, счастливый случай
65. Иметь природные задатки
66. Много работать
67. Иметь необходимые знакомства, связи
68. Иметь нравственные убеждения
69. Проживать в определенном регионе
70. Важно, каков твой пол

Ответы на эти вопросы хранятся в переменных v59-v70 в файле opros.sav. В этом файле также находятся и другие переменные, использовавшиеся при исследовании (пол, возраст, место жительства, профессия). На основании вопросов о жизненных стратегиях молодежи попытаемся определить группы (кластеры) респондентов. Для начала рекомендуется сократить количество переменных при помощи факторного анализа.

Откройте файл opros.sav.

Выберите в меню Analyze (Анализ) Data Reduction (Преобразование данных) Factor... (Факторный анализ)

- Переменные v59-v70 внесите в список целевых переменных.
- Через выключатель Extraction... (Отбор) деактивируйте вывод неповёрнутого факторного решения.
- Через выключатель Rotation... (Вращение) для осуществления вращения активируйте метод варимакса.
- Минув выключатель Options... (Опции) в разделе Coefficient Display Format (Формат отображения коэффициентов) (подразумеваются факторные нагрузки) активируйте Sorted by Size (Отсортированные по размеру). Затем активируйте опцию Suppress absolute values less than: (Не выводить абсолютные значения меньше чем:) и введите значение ,40.
- В заключение щёлкните по выключателю Scores... (Значения), чтобы значения факторов сохранить в виде новых переменных.

В результате расчёта было отобрано три фактора и добавлено в файл три переменные от (fac1\_1 до fac3\_1), которые и отображают эти три фактора. Среди результатов присутствует повёрнутая факторная матрица (см. сле-

**Rotated Component Matrix<sup>a</sup>**

	Component		
	1	2	3
высокопоставленные родители	,840		
обеспеченная семья	,732		
знакомства	,685		
связи с криминалом	,613		
много работать		,659	
хорошее образование		,629	
нравственные убеждения		,587	
задатки		,582	
амбиции		,505	
везение, случай			
пол			,805
регион			,788

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

дующую таблицу).

Факторная матрица красноречиво демонстрирует, что отобранные факторы могут быть расположены в следующей смысловой последовательности:

- группа «пассивных», для которых достижение успеха связано со статусом родителей, материальной обеспеченностью семьи, знакомствами с нужными людьми.

<sup>7</sup> Кодировка переменных представлена также как в анкете опроса.

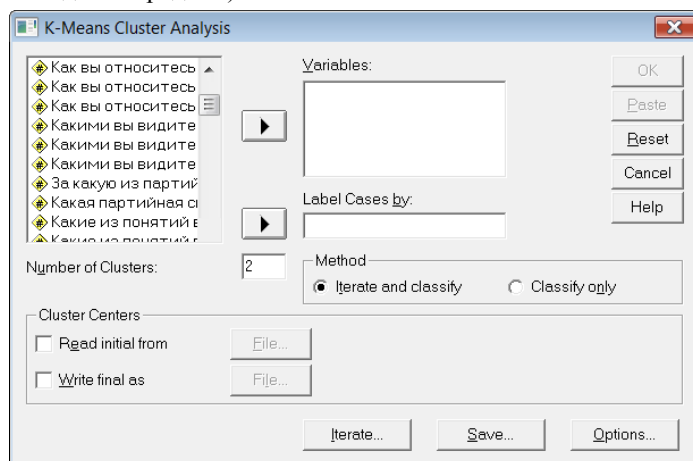


- группа «активных, самостоятельных», для которых важно много работать, иметь хорошее образование, нравственные убеждения, задатки и амбиции.
- группа «ориентированных на случай или на природные задатки»

Теперь используем сохранённые нами значения этих трех факторов для проведения кластерного анализа для респондентов. Так как количество наблюдений равно 1085 слишком велико для иерархического кластерного анализа, выберем метод анализа кластерных центров.

- Присвойте переменным fac1\_1-fac3\_1 метки: "пассивные", "активные", "ориентированные на случай" соответственно.
- Выберите в меню Analyze (Анализ) Classify (Классифицировать) K-Means Cluster... (Кластерный анализ методом к-средних)

От-  
сред-



кроется диалоговое окно K-Means Cluster Analysis (Кластерный анализ методом к-средних).

- Переменные от fac1\_1 до fac3\_1 поместите в поле тестируемых переменных. Теперь нужно указать количество кластеров. Подходящим вариантом было бы сначала провести иерархический кластерный анализ для произвольно выбранных наблюдений и получившееся количество кластеров принять за оптимальное. Но можно провести и несколько опытных, пробных расчётов с различным количеством кластеров и после этого определиться с подходящим вариантом решения.
- Мы остановимся на трех кластерах; введите это значение в поле Number of Clusters (Количество кластеров).
- Через выключатель Iterate... (Итерации) укажите число итераций равно 99; установленное по умолчанию количество итераций равно 10, оказалось бы недостаточным.
- Щёлкните по выключателю Save... (Сохранить), чтобы при помощи дополнительных переменных зафиксировать принадлежность наблюдений к кластеру.
- Щёлкните на ОК, чтобы начать расчёт.

Сначала приводятся первичные кластерные центры и обобщённые данные итерационного процесса (30 итераций); затем выводятся окончательные кластерные центры и информация о количестве наблюдений.

**Final Cluster Centers**

	Cluster		
	1	2	3
пассивные	,07957	-,47909	,56511
самостоятельные	-,63520	,15716	1,27435
ориентированные на случай	-,42038	,97864	-,53644

При оценке кластерных центров следует в первую очередь обратить внимание на то, что здесь речь идёт о средних значениях факторов, которые находятся в пределах примерно от -3 до +3. К тому же, надо помнить, что в соответствии с кодировкой ответов (1 - очень важно, 5 - не важно) большое отрицательное значение фактора означает его большую степень его проявления, то есть сигнализирует о высокой компетентности, и наоборот, большое положительное значение фактора подразумевает низкую степень его проявления.

Если учесть всё вышесказанное, то наши три кластера можно интерпретировать следующим образом:

Кластер1: самостоятельные респонденты

Кластер2: пассивные респонденты

Кластер3: ориентированные на случай

В заключение выводятся показатели количества наблюдений, относящихся к каждому из кластеров. Группа пользователей (кластер 1) наиболее многочисленна.

#### Number of Cases in each Cluster

Cluster	1	278,000
	2	183,000
	3	116,000
Valid		577,000
Missing		320,000

К исходному файлу была добавлена переменная qcl\_1, отражающая принадлежность к определённому кластеру. Эту переменную можно использовать для обнаружения возможных связей между кластерной принадлежностью и полом, возрастом, профессией или отношением к политической деятельности (исходя из задач исследования).

Cluster Number of Case \* отношение к политической деятельности Crosstabulation

			отношение к политической деятельности					Total
			я уже занимаюсь политикой	меня интересует полит. деятельность	политика важна, но заниматься ею не буду	политика бесполезна, приносит неприятности	политика мне не интересна	
Cluster Number of Case	пассивные	Count	1	27	117	23	63	231
		% within отношение к политической деятельности	25,0%	60,0%	53,7%	48,9%	41,7%	49,7%
	самостоятельные и активные	Count	2	16	61	7	53	139
		% within отношение к политической деятельности	50,0%	35,6%	28,0%	14,9%	35,1%	29,9%
	ориентированы на случай и задатки	Count	1	2	40	17	35	95
		% within отношение к политической деятельности	25,0%	4,4%	18,3%	36,2%	23,2%	20,4%
Total		Count	4	45	218	47	151	465
		% within отношение к политической деятельности	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Как видно из таблицы, среди тех, тех, кто считает политику бесполезной и относится к ней негативно большинство составляют респонденты с пассивной жизненной установкой (48,9%), не многим меньше (36,2%) респонденты ориентированные на случай. Такого же мнения придерживаются всего лишь 14,9% из числа активной молодежи.

### **Заключение.**

Цель данного пособия - познакомить студентов с базовыми техниками и методиками программы SPSS, наиболее часто применяемыми в практической исследовательской работе. Кроме описанных в данном пособии статистических методов обработки данных, программа SPSS позволяет проводить кластерный анализ, дискриминантный анализ, многомерное шкалирование, логлинейный метод и метод логистической регрессии. Подробнее с этими методами можно познакомиться в специальной литературе (см. стр.3).

В заключении можно сказать, что программа SPSS, как и широко распространенные программы Excel и Statistica, является эффективным инструментом для практической работы в области социологического и политического анализа.

## 11. Словарь основных терминов, используемых в процедурах прикладного социологического исследования, в работе с компьютерной программой SPSS.

**Валидность** – мера пригодности применяемых в прикладной социологии методик решения исследовательских задач, степень соответствия переменных и индикаторов эмпирическим данным, позволяющая получать надежные, репрезентативные и достоверные результаты исследования.

**Диаграмма рассеяния** – график совместного распределения двух количественных переменных.

**Дисперсия** – мера разброса данных, разброс данных относительно среднего арифметического. Дисперсия (variance) равна сумме квадратов отклонений каждого значения от среднего, деленной на N-1, где N - число значений в распределении.

**Дисперсионный анализ** служит для проверки гипотезы о статистической значимости различий между средними величинами в нескольких группах наблюдений.

**Единица анализа** – это элементарная, единичная часть объекта исследования. Единица анализа чаще всего совпадает с единицей наблюдения, в социологии, как правило, этой единицей является отдельный респондент. Следовательно, единицей анализа, становится информация, содержащаяся в анкете, чаще всего заполняемой одним респондентом.

**Интервальная шкала** – измерительная шкала, пункты которой расположены на одинаковом расстоянии друг от друга.

**Каузальность** – причинность, причинный характер связи между явлениями, процессами, событиями.

**Коэффициент вариации** – мера разброса данных, вычисляется по формуле

$$V = \frac{\sigma}{\bar{X}} \cdot 100\%$$

измеряет среднее квадратическое отклонение в процентах от среднего арифметического.

**Корреляционный анализ** – измерение статистической взаимозависимости между двумя и более переменными.

**Кластерный анализ** представляет собой группу алгоритмов многомерной классификации объектов, под которой понимается упорядочение в наглядные структуры или группы сходства/различия объектов, обладающих множеством характеристик.

**Медиана (median)** – мера центральной тенденции, представляет собой значение признака, соответствующее 50% накопленной частоте.

**Меры связи** – коэффициенты, предназначенные для измерения тесноты связи.

**Меры изменчивости (меры разброса данных)** – показывают как далеко, в среднем, отдельные значения разбросаны по отношению к среднему арифметическому значению (дисперсия, среднее квадратическое отклонение).

**Меры центральной тенденции** – характеристики, предназначенные для описания центра распределения (мода, медиана, среднее арифметическое).

**Мода (mode)** – мера центральной тенденции, значение обладающее максимальной частотой. Периодическая смена образцов культуры и массового поведения.

**Номинальная шкала** - измерительная шкала, предназначенная для классификации объектов, градации шкалы не упорядочены.

**Переменная** - элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа. Простейшие переменные – вопросы анкеты, к примеру, пол и возраст респондента.

**Порядковая шкала** – измерительная шкала, упорядочивающая объекты по некоторому критерию.

**Распределение частот** – способ представления обобщенных данных исследования, совокупность значений признаков и их частот (относительных, абсолютных, накопленных).

**Регрессионный анализ** – измерение связи между зависимой переменной и одной (парный регрессионный анализ) или несколькими (множественный) независимыми переменными.

**Среднее арифметическое значение (mean)** мера центральной тенденции, равная сумме всех значений распределения, деленной на их количество.

**Стандартное отклонение (standard deviation)**, среднее квадратическое отклонение, равно квадратному корню из дисперсии.

**Таблица сопряженности** – средство предоставления совместного распределения двух признаков, таблица, строки которой предназначены для значений одной переменной, столбцы – для значений другой переменной, на пересечении строки и столбца указывается частота совместного появления значений двух переменных.

**Уравнение линейной регрессии** – уравнение, описывающее линейную связь между двумя переменными:  $y = bx + a$ .

**Факторный анализ** предназначен для концентрации исходной информации, представления большого числа рассматриваемых признаков через меньшее число более емких внутренних латентных характеристик, которые не поддаются непосредственному измерению.

**Частота абсолютная** – количество объектов, обладающих данным значением признака.

**Частота накопленная** – сумма частот значений, не превосходящих данное значение.

**Частота относительная** – доля или процент объектов, обладающих данным значением признака, по отношению к объему выборки.