

**Электронное методическое пособие по курсу
«Методы обработки данных в политологии»
Бакалавриат факультета прикладной политологии
2 курс, 3-4 модуль**

Пакет статистических программ SPSS

Содержание

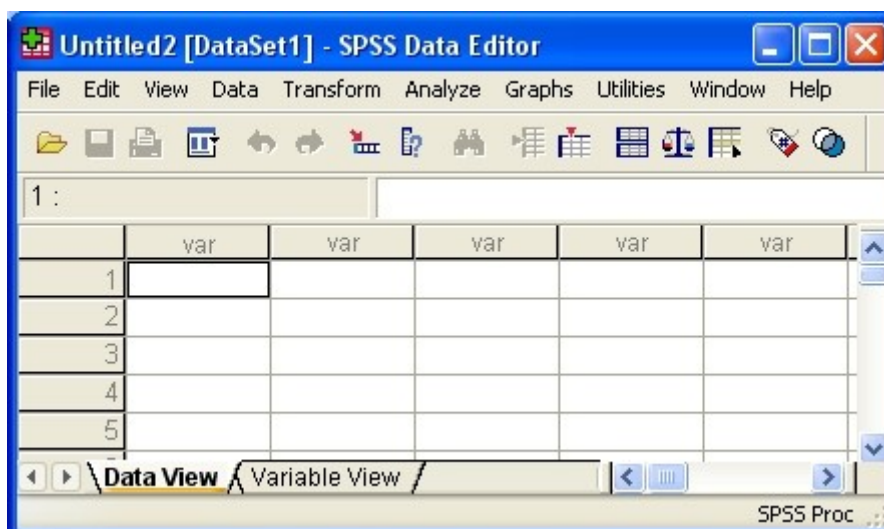
<u>Запуск SPSS, ввод и загрузка данных.....</u>	<u>2</u>
<u>Окно процедуры обработки.....</u>	<u>4</u>
<u>Выбор объектов для анализа.....</u>	<u>11</u>
<u>Редактирование графиков.....</u>	<u>12</u>
<u>Сравнение двух выборок с неизвестным распределением.....</u>	<u>19</u>
<u>Сравнение двух независимых выборок (Mann-Whitney U, Wilcoxon W).....</u>	<u>19</u>
<u>Сравнение двух связанных выборок (Sign Test, Wilcoxon Signed Ranks Test).....</u>	<u>21</u>
<u>Сравнение двух выборок с известным распределением.....</u>	<u>24</u>
<u>Сравнение двух независимых выборок (Independent-Samples T Test).....</u>	<u>24</u>
<u>Сравнение двух связанных выборок (Related-Samples T Test).....</u>	<u>26</u>
<u>Сравнение нескольких независимых выборок.....</u>	<u>28</u>
<u>Критерий Краскела-Уоллиса (Kruskal-Wallis H).....</u>	<u>28</u>
<u>Однофакторный дисперсионный анализ (One-Way ANOVA).....</u>	<u>31</u>
<u>Корреляционный анализ.....</u>	<u>34</u>
<u>Регрессионный анализ</u>	<u>37</u>

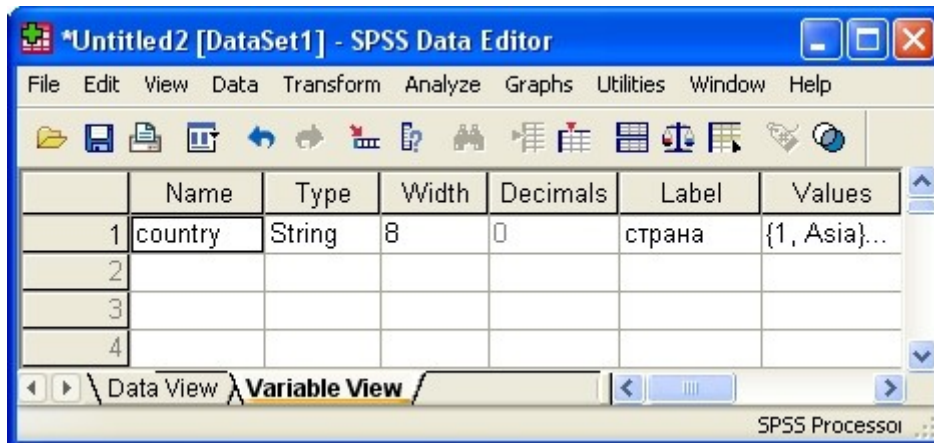
Запуск SPSS, ввод и загрузка данных

Запустите пакет **SPSS** с помощью значка главного меню Windows. Для того чтобы ввести собственные данные, установите переключатель **Type in data**. Если Вы уже имеете массив данных в формате **.sav**, то установите переключатель **Open an existing data source**.

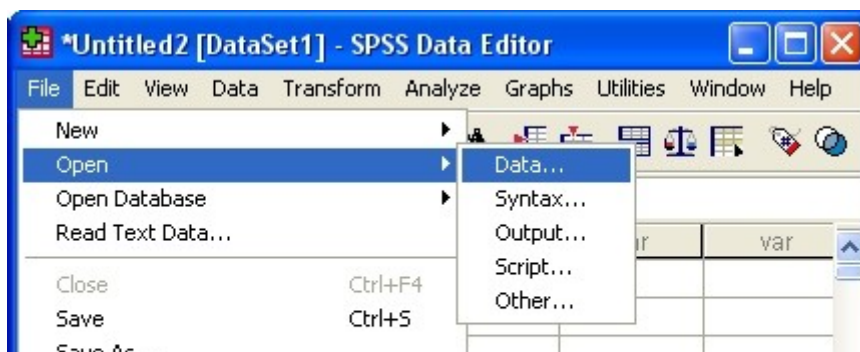


Ввод данных осуществляется через окно редактора данных в закладке **Data View** (Просмотр данных). Вторая закладка **Variable View** (Просмотр переменных) позволяет создавать новые переменные: задавать их названия (**Name**), задавать их тип (**Type**), ширину (**Width**), количество знаков после запятой (**Decimals**), описывать переменные (**Label**), задавать значения категоризованных переменных (**Values**). Обратим внимание, что названия переменных не должны содержать пробелов, предпочтительнее заменять их на нижнее подчеркивание.

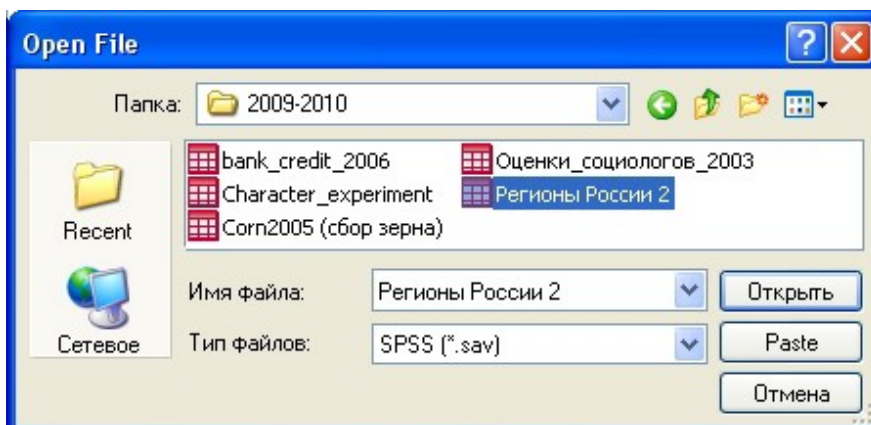




Для того чтобы воспользоваться уже имеющимся массивом данных, можно открыть его через команду **Open** в строке панели управления с помощью выпадающего меню **File**. Для открытия файлов расширения **.sav** выберите команду **Data.**, для файлов **.spo** – команду **Output**.



В появившемся диалоговом окне **Open File** выберите файл данных **Регионы_России_2.sav**. Нажмите на кнопку **Открыть**.



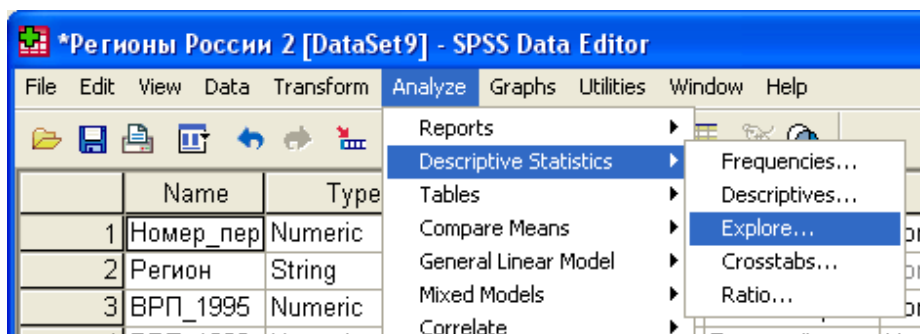
Увидеть весь массив данных можно через закладку **Data View**. Переменные располагаются по горизонтали, наблюдения – по вертикали.


	Номер_переменной	Регион	ВРП_1995	ВРП_1996	ВРП_1997	ВРП_1998
1	69	Белгородская область	8613,40	10153,10	11406,30	12938,10
2	83	Брянская область	5314,30	7720,40	7697,60	8206,30
3	75	Владимирская область	6563,40	7928,70	9137,20	9882,40
4	67	Воронежская область	6633,10	8103,20	9442,20	9769,10

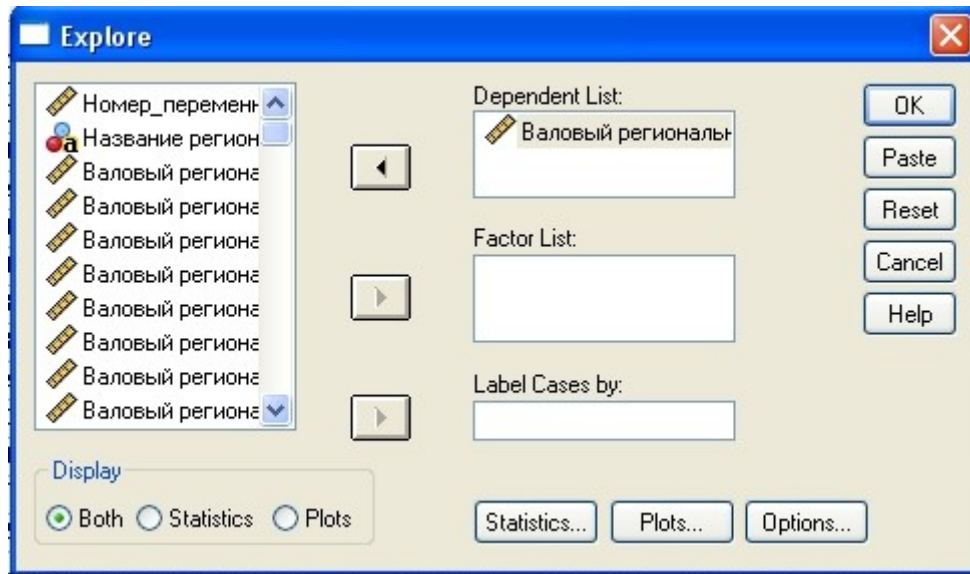
Окно процедуры обработки

Почти все окна процедуры обработки имеют сходное устройство, продемонстрируем это на примере процедуры **Explore**.

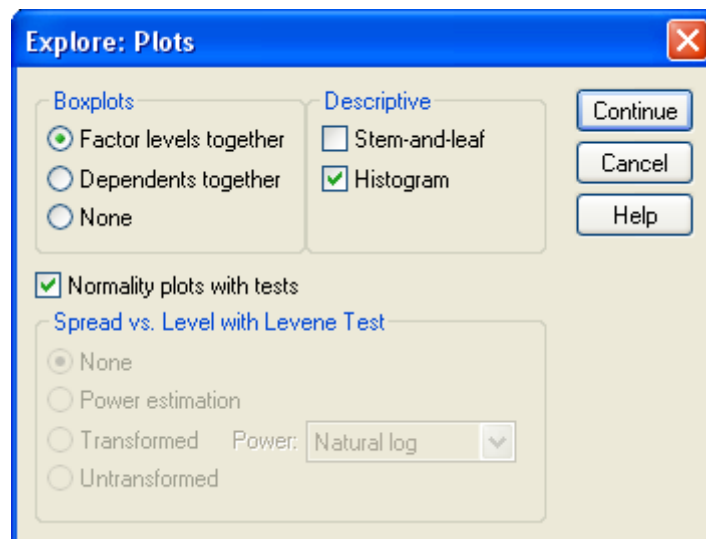
Процедура **Explore** позволяет получить описание выборки и проверяет нормальность ее распределения. Для того чтобы ею воспользоваться, нажмите на команду **Explore** в меню **Analyze** (Анализ), **Descriptive Statistics** (Описательные статистики).



Слева в диалоговом окне располагаются все переменные, доступные для анализа. Справа находятся поля, которые необходимо заполнить переменными, например, **Dependent List** (Список зависимых переменных) или **Factor List** (список *факторов*). С помощью кнопки  помещаем анализируемую переменную интервального типа **ВРП_1995** в поле **Dependent List**.



В нижней части окна нажмите кнопку **Plots** (Графики) и в появившемся окне снимите флажок **Stem-and-leaf** и установите флажки **Histogram** (Гистограмма) для того, чтобы представить распределение визуально, и **Normality plots with tests** (Тест на нормальность распределения), чтобы пакет вычислил значение статистики Z Колмогорова-Смирнова. Нажмите **Continue** (Продолжить).



Нажмите **OK** в окне **Explore**, чтобы получить результаты в отдельном файле **Output** (окно выдачи). Интерпретация: количество валидных наблюдений $N = 79$ (89,8%), количество пропущенных наблюдений $N = 9$ (10,2%). В таблице **Descriptives** находятся значения среднего, доверительного интервала для среднего, медианы, дисперсии, стандартного отклонения, значениями максимума и минимума. Значение статистики Колмогорова-Смирнова находится из таблицы **Tests of Normality** равно 0,136. Минимальный

уровень значимости равен 0,001, поэтому нулевую гипотезу о нормальности распределения можно отвергнуть.

Descriptives

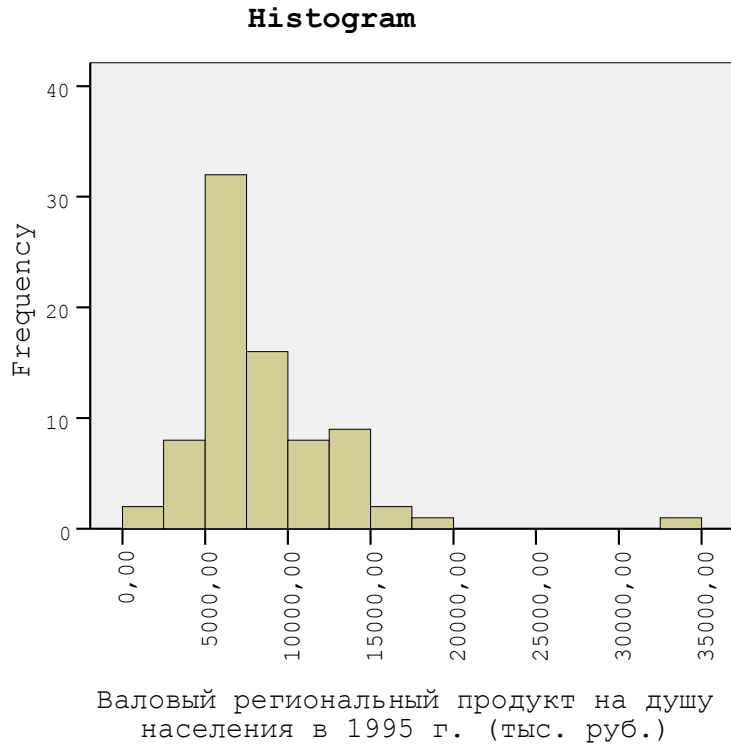
			Statistic	Std. Error
Валовый региональный продукт на душу населения в 1995 г. (тыс. руб.)	Mean		8583,18	518,58
	95% Confidence Interval for Mean	Lower Bound	7550,76	
		Upper Bound	9615,60	
	Median		7466,30	
	Variance		21245279,70	
	Std. Deviation		4609,26	
	Minimum		1877,70	
	Maximum		34335,60	

Tests of Normality

	Kolmogorov-Smirnov(a)		
	Statistic	df	Sig.
Валовый региональный продукт на душу населения в 1995 г. (тыс. руб.)	,136	79	,001

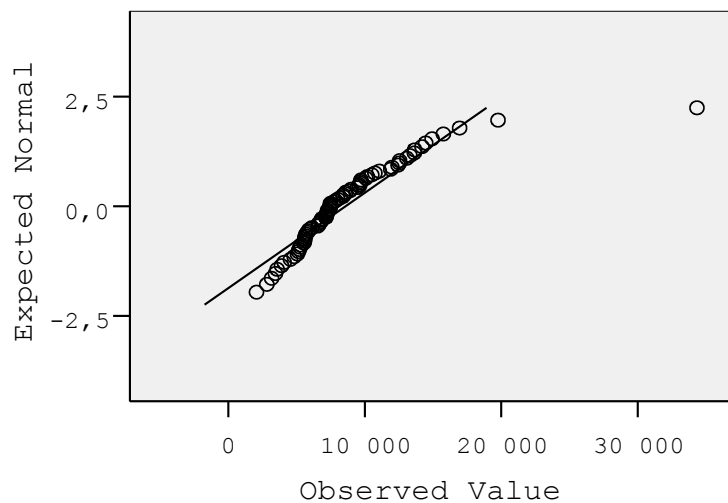
Помимо таблиц окно выдачи SPSS содержит несколько графиков, рассмотрим их подробнее.

На гистограмме (**Histogram**) по оси ОХ откладываются значения анализируемой переменной «Валовый региональный продукт на душу населения в 1995 г. (тыс. руб.)», по оси ОУ – частота (**Frequency**). Отредактировать гистограмму, в том числе изменить цену деления шкалы и получить столбики другой ширины, можно через редактирование гистограмм (см. подробнее пункт **Редактирование графиков**).



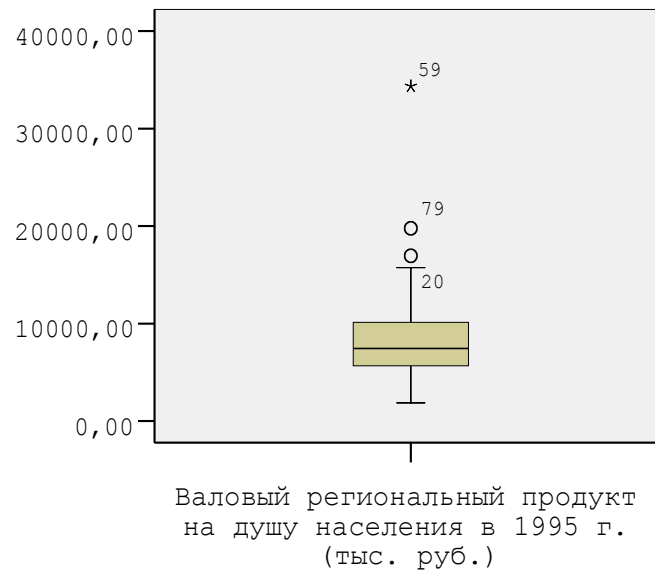
Нормальная вероятностная бумага (**Normal Q-Q Plot**) тоже визуально проверяет нормальность распределения. Проведённая прямая линия – график функции нормального распределения, и если наблюдения располагаются прямо по линии, то можно предположить, что наша выборка нормальна.

Normal Q-Q Plot of Валовый региональный продукт на душу населения в 1995 г. (тыс. руб.)



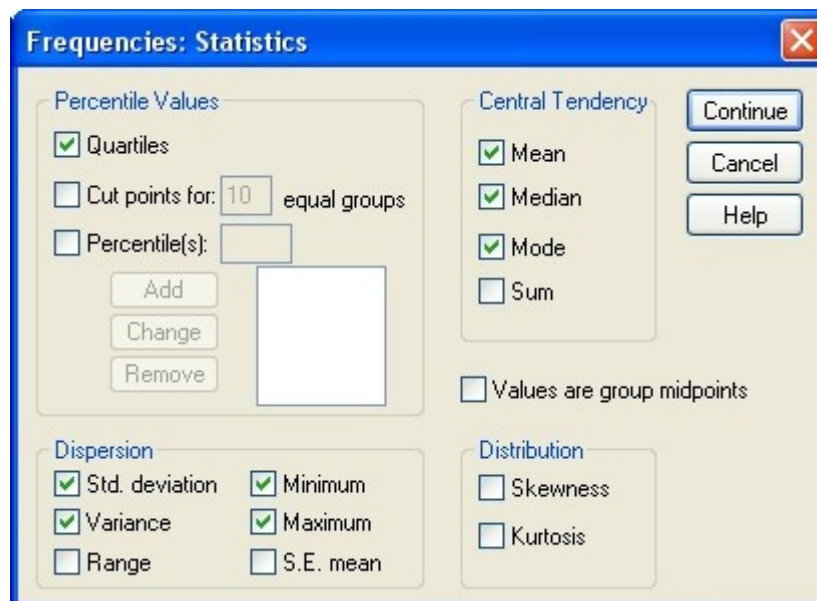
На графике «Ящик с усами» (**Boxplot**) обозначены медиана ($Q_{0,5}$), верхний ($Q_{0,75}$) и нижний квартиль ($Q_{0,25}$), межквартильный размах ($Q_{0,75}-Q_{0,25}$), максимальное и минимальное

значение, потенциальные выбросы (**suspected outliers**) и выбросы (**outliers**). Последние отмечены на рисунке звездочкой. По оси ОУ откладываются значения переменной «Валовый региональный продукт...».

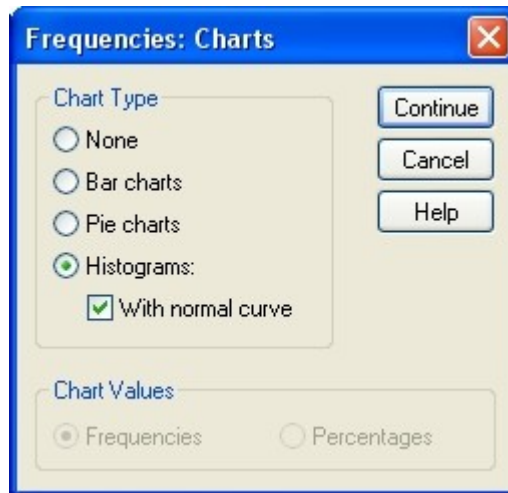


Другие способы получить описание переменных в массиве данных тоже находятся в меню **Descriptive Statistics** (Описательные статистики). Процедуры **Frequencies** (Частоты) и **Descriptives** (Описание) устроены стандартно. С их помощью можно:

- получить числовые характеристики распределения вероятностей установив флажки в соответствующих полях и диалоговых окнах (например, **Mean**, **Median**, **Variance** и т.д.);



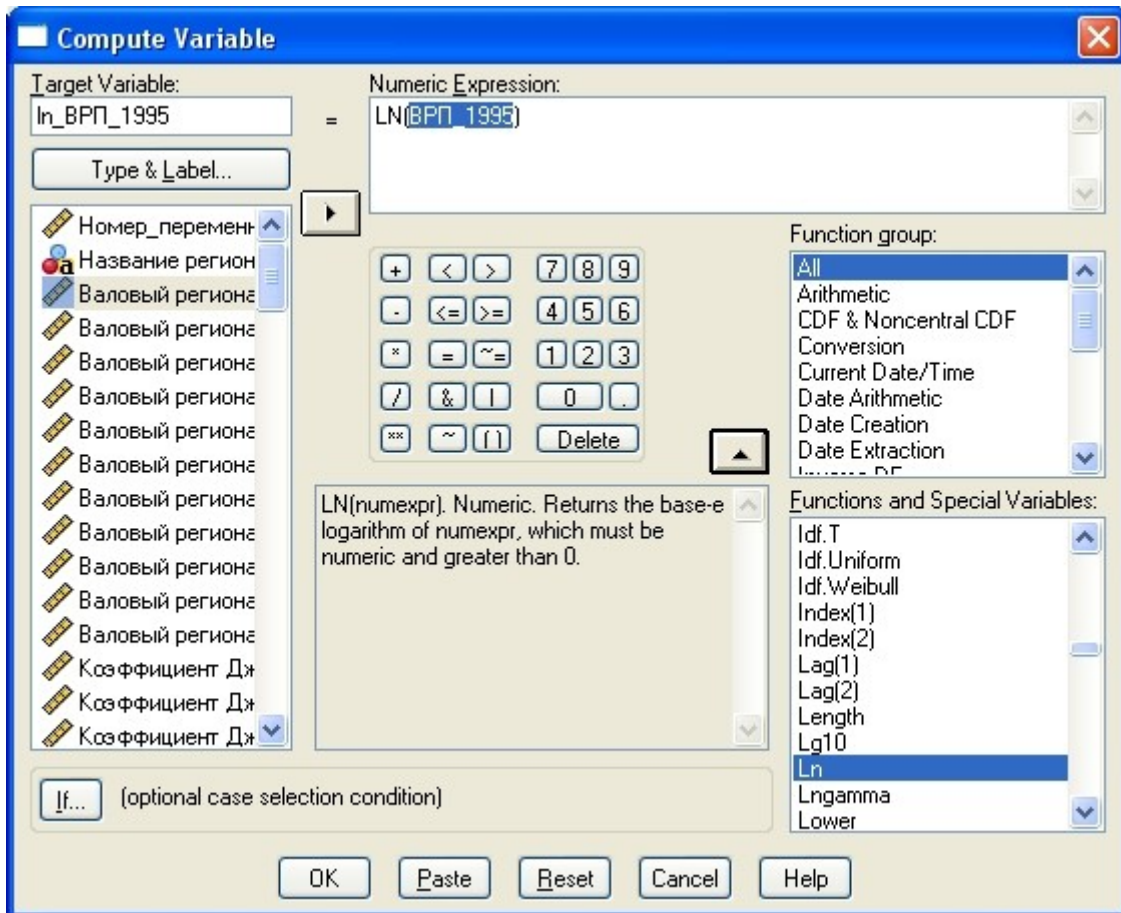
- получить графическое представление данных (флажки **Histograms** и **With normal curve**);




- сохранить стандартизованные значения в качестве новых переменных (**Save standardized values as variables**).

Вернемся к содержательной задаче.

ВРП на душу населения измеряется в тысячах рублей. Попробуем преобразовать данные и разбить выборку на подгруппы. Сначала прологарифмируем переменную **ВРП_1995** через команду **Compute** выпадающего меню **Transform** на панели управления.



Вкратце опишем устройство диалогового окна **Compute Variable**. Как уже было сказано, слева находится список всех доступных переменных, над ним – поле **Target Variable**, в которое необходимо вписать название новой, преобразованной, переменной. Задать функцию преобразования в поле **Numeric Expression** можно с помощью панели калькулятора. Второй способ – выбрать из доступных функций и специальных переменных (**Functions and Special Variables**) в области справа. Для того чтобы функция появилась в поле **Numeric Expression**, необходимо сначала выбрать или из группы функций (**Function Group**) или из их полного списка, выбрав строку **All**. Затем необходимо выбрать собственно функцию, их описание появится внизу под панелью калькулятора. Нам требуется натуральный логарифм Ln, поэтому с помощью кнопки  перенесем его наверх. После чего вместо знака вопроса нужно выбрать переменную для логарифмирования LN(BPP_1995). Нажмите **OK**. В закладке **Variable View** появится описание новой переменной ln_BPP_1995, а в закладке **Data View** – ее значения.

Вновь запустим процедуру **Explore** и будем анализировать уже логарифм ВРП. Для того чтобы условно разбить выборку по некоторому категоризованному признаку, добавим в поле **Factor List** переменную **Федеральный округ**.

В окне выдачи описательные статистики (таблица **Descriptives**) и значения статистики Колмогорова-Смирнова (таблица **Tests of Normality**) будут посчитаны для каждого из федеральных округов в отдельности.

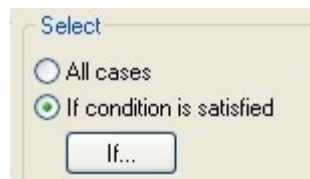
Выбор объектов для анализа

Для того чтобы отобрать только те наблюдения, которые нам необходимы, воспользуемся процедурой **Select cases** в окне выпадающего меню **Data**.

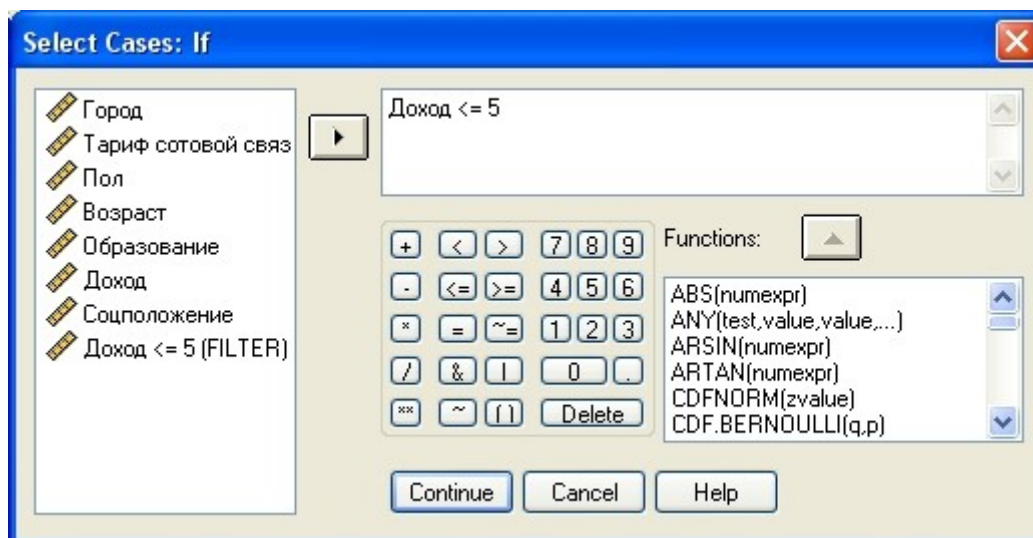
К примеру, очистим массив данных от ошибок ввода данных. Для этого откройте файл «Рег.предпочтение тарифов» (его описание представлено в [...]). Переменная «доход» разбита на 5 категорий и «отказ отвечать». Для того чтобы выяснить есть ли в массиве ошибки ввода данных, построим столбиковую диаграмму (**Bar** в меню **Graphs**).

На диаграмме в окне выдачи **Output** помимо ожидаемых столбцов 1, 2,...6 появятся столбцы 7, 8, 9 и 14. Они позволяют определить наличие ошибок ввода данных. Для того чтобы очистить от них выборку, воспользуемся процедурой **Select cases**.

В поле поставьте флажок **If condition is satisfied** и нажмите кнопку **If...**

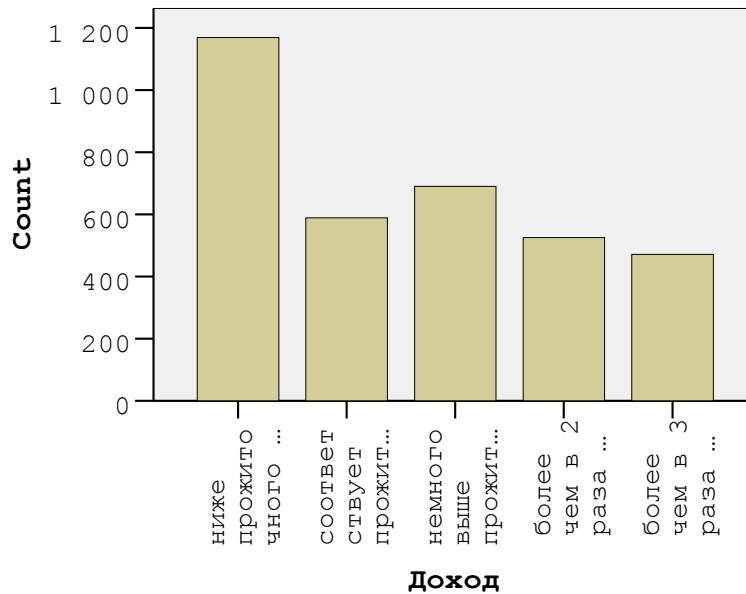


Появится новое окно, в котором нужно задать условие отбора наблюдений. Это можно сделать с панели калькулятора или используя список функций. Например, условие может выглядеть так: **Доход <= 5**, т.е. отберем только те случаи, которые могут предоставить информацию о категории дохода респондента.



Чтобы продолжить нажмите **Continue**, затем **OK**. В результате, в закладке **Variable View** некоторые объекты будут зачеркнуты.

Если заново воспроизвести процедуру **Graphs, Bar**, то на столбиковой диаграмме будут отражены только 5 категорий дохода, как это показано на рисунке ниже.

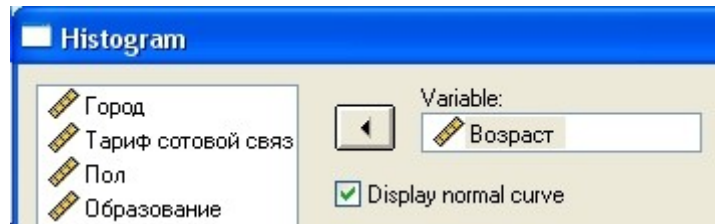


Редактирование графиков

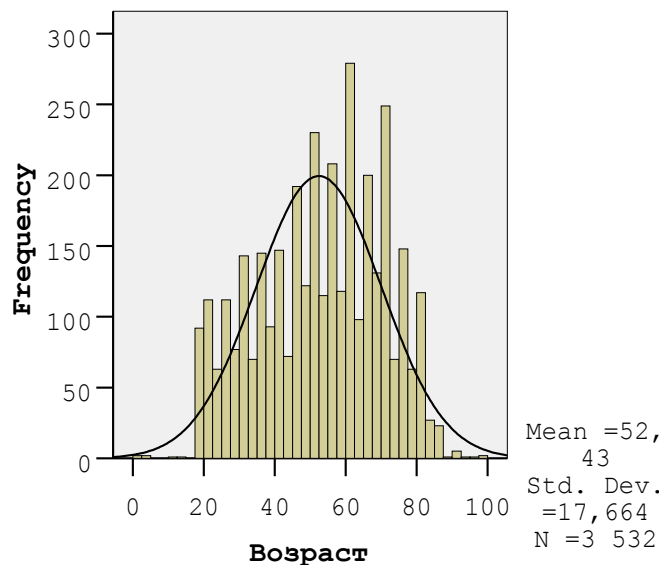
Окно графического редактора похоже на окно любого текстового редактора. Оно содержит панель управления, несколько тематических панелей инструментов, рабочее поле и возможность вызвать диалоговое окно кликом мыши на объекте графика или рабочем поле. Покажем работу графического редактора на примере построения графика распределения некой переменной.

Откроем файл «Рег. предпочтение тарифов». Файл содержит два вида переменных (их описание дано в [...]). Для количественных переменных используется процедура **Histogram**, для категоризованных – процедура **Bar** (столбиковые диаграммы). Все процедуры, позволяющие визуально представить данные в том или ином виде, находятся в меню **Graphs**. В качестве примера посмотрим, как распределен возраст в выборке.

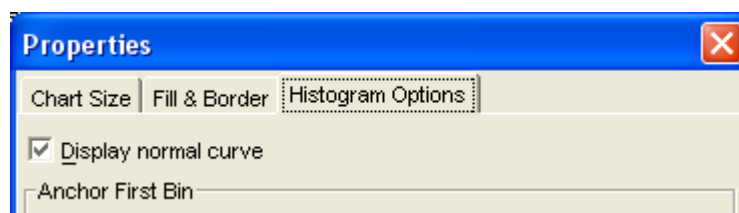
Для того чтобы построить график распределения количественной переменной «возраст», выберите процедуру **Histogram**. В диалоговом окне нужно перенести переменную в поле **Variable**, как это показано на рисунке ниже. Поставьте флажок **Display normal curve**.



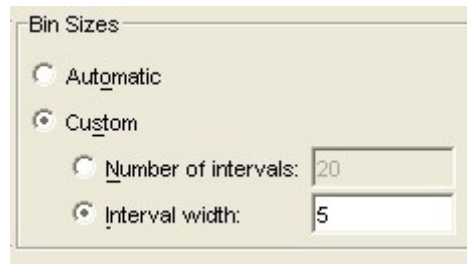
В окно выдачи **Output** появится график, на котором по оси ОХ показан возраст, а по ОУ – частота, с которой встречается тот или иной возраст. Гистограмму можно редактировать, в том числе, изменить цену шкалы деления, т.е. возрастной интервал, образующий столбик.



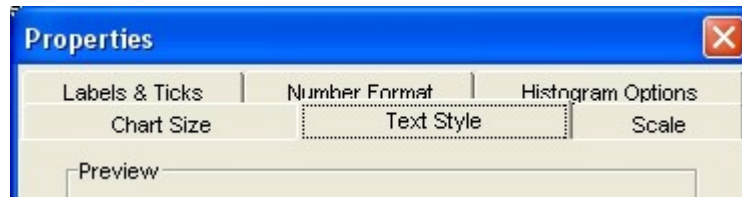
Для того чтобы появилось диалоговое окно **Properties**, необходимо дважды нажать на график. В закладке **Chart size** можно задать размер рисунка, в закладке **Fill & Border** можно изменить оформление рисунка: цвета, стили, шрифты. Изменить интервал, задающий шкалу деления, можно в закладке **Histogram options**.




Поменяем ширину столбца в поле **Bin sizes**, установив переключатель **Custom**, а затем **Interval width** и поставим, к примеру, 5. Нажмите **Apply**.



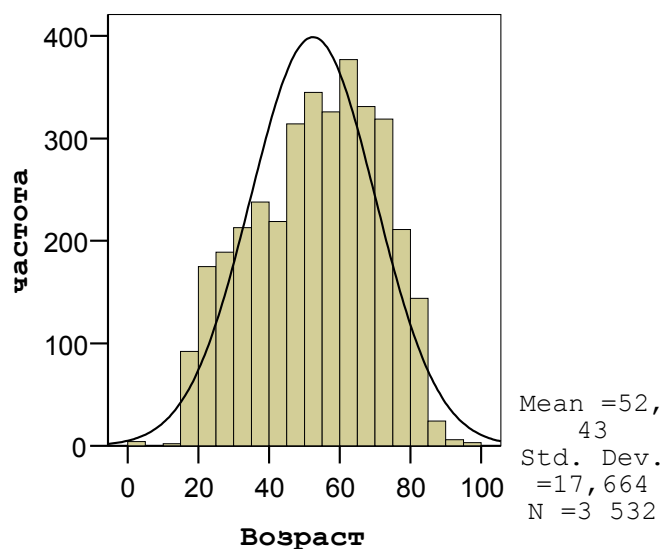
Кликнув на текстовую часть рисунка в закладке **Text style** можно изменить размер, стиль, цвет шрифта и многое другое. Аналогичным образом устроены прочие закладки.



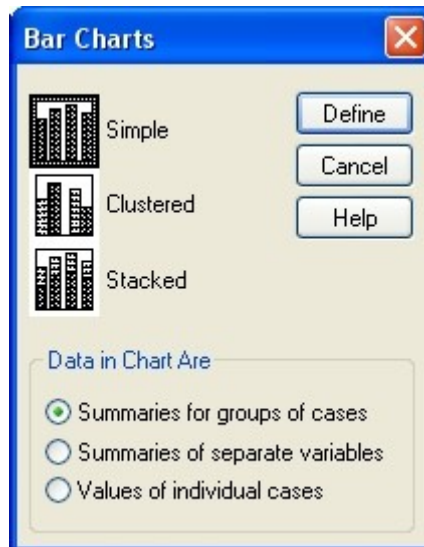
Окно графического редактора имеет свою панель инструментов. Воспользуемся кнопкой **Insert a title** , чтобы сделать для гистограммы заголовок.

Результат некоторых преобразований, сделанных нами, представлен на рисунке.

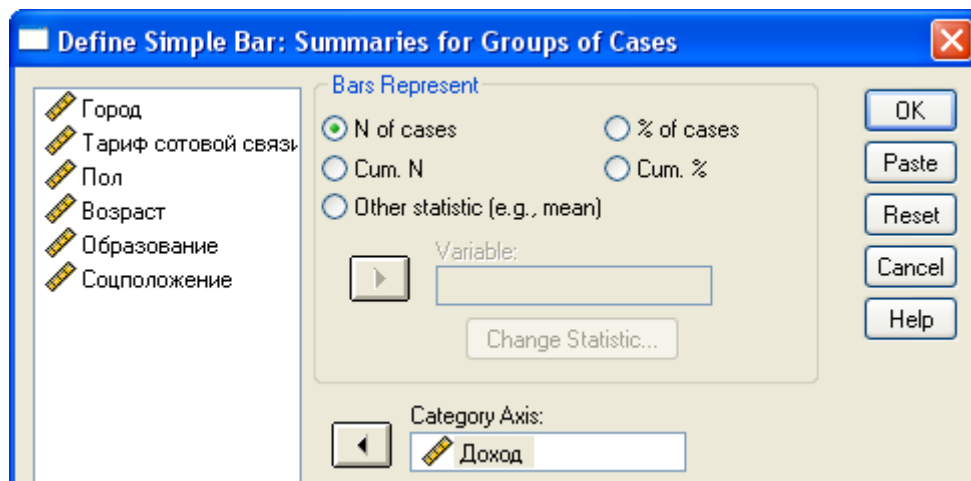
распределение возрастной структуры



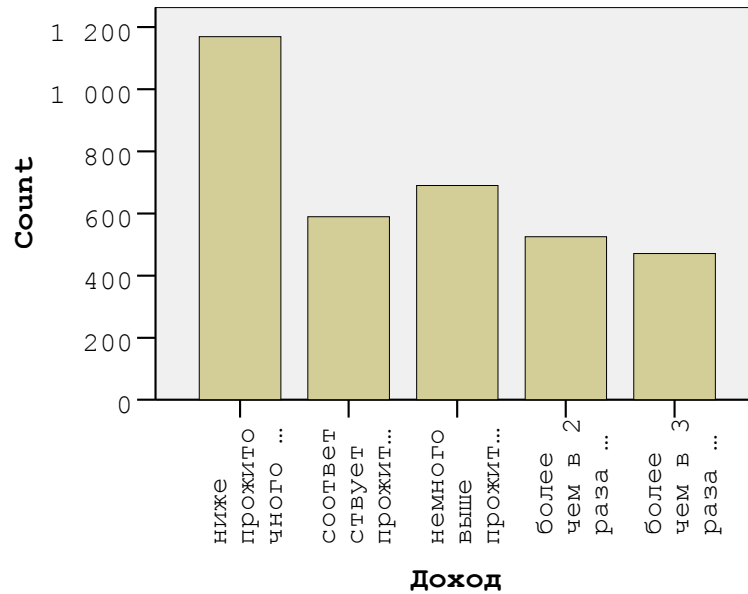
Предположим, теперь мы хотим посмотреть то, как распределен уровень образования респондентов. В меню **Graphs** нужно выбрать команду **Bar**. В окне **Bar Charts** нажмите на тип **Simple**, а в поле **Data in Chart Are** поставьте переключатель **Summaries for groups of cases**.



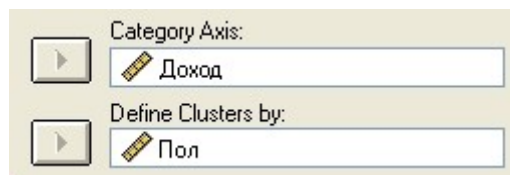
В новом окне перенесите порядковую переменную «образование» в поле **Category Axis**. В поле **Bar Represents** выберите один из вариантов представления данных на столбиковой диаграмме, например, число наблюдений (**N of cases**), и поставьте соответствующий переключатель. Затем нажмите **ОК**.



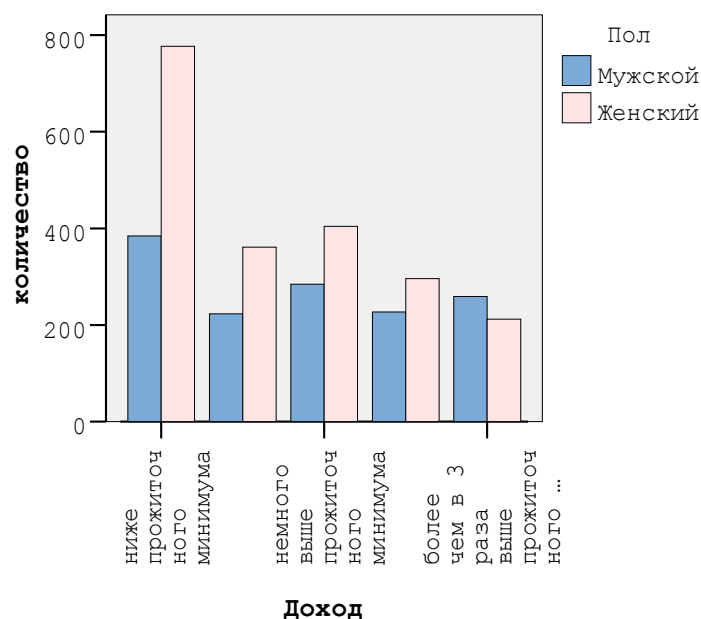
До того, как представлять данные графически, необходимо очистить массив данных от ошибок ввода данных (процедура **Select cases** в меню **Data**). Тогда столбиковая диаграмма будет выглядеть как на рисунке ниже.



Если в окне **Bar Charts** вместо **Simple** выбрать тип **Clustered**, то в диалоговом окне **Define Clustered Bar** станет возможным разбить выборку по некоторому признаку. В новое поле **Define Clusters by** необходимо поместить фактор, например, по пол.



Как следует из диаграммы, почти вдвое больше опрошенных женщин имеют доход ниже прожиточного минимума, и большее количество мужчин имеет доход более чем в 3 раза выше прожиточного минимума. В оставшихся трех категориях по количеству преобладают женщины.



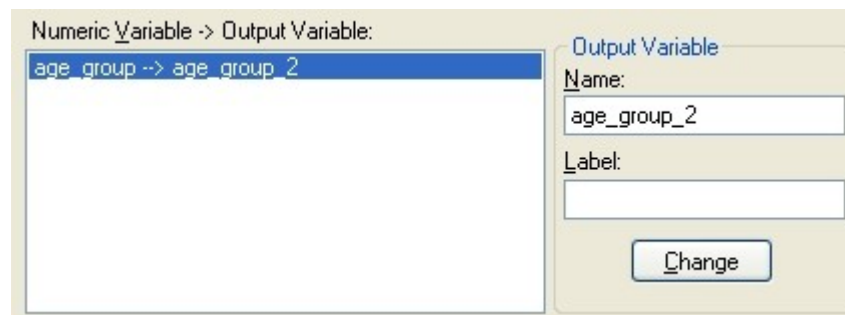
Редактировать можно любые графики и рисунки: гистограммы (**Histogram**), столбиковые диаграммы (**Bar**), линейные диаграммы (**Line**), «ящики с усами» (**Boxplot**), нормальную вероятностную бумагу (**Normal Q-Q Plot**), диаграммы рассеяния (**Scatter/ Dot**).

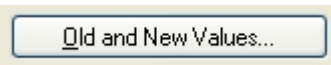
Перекодирование данных

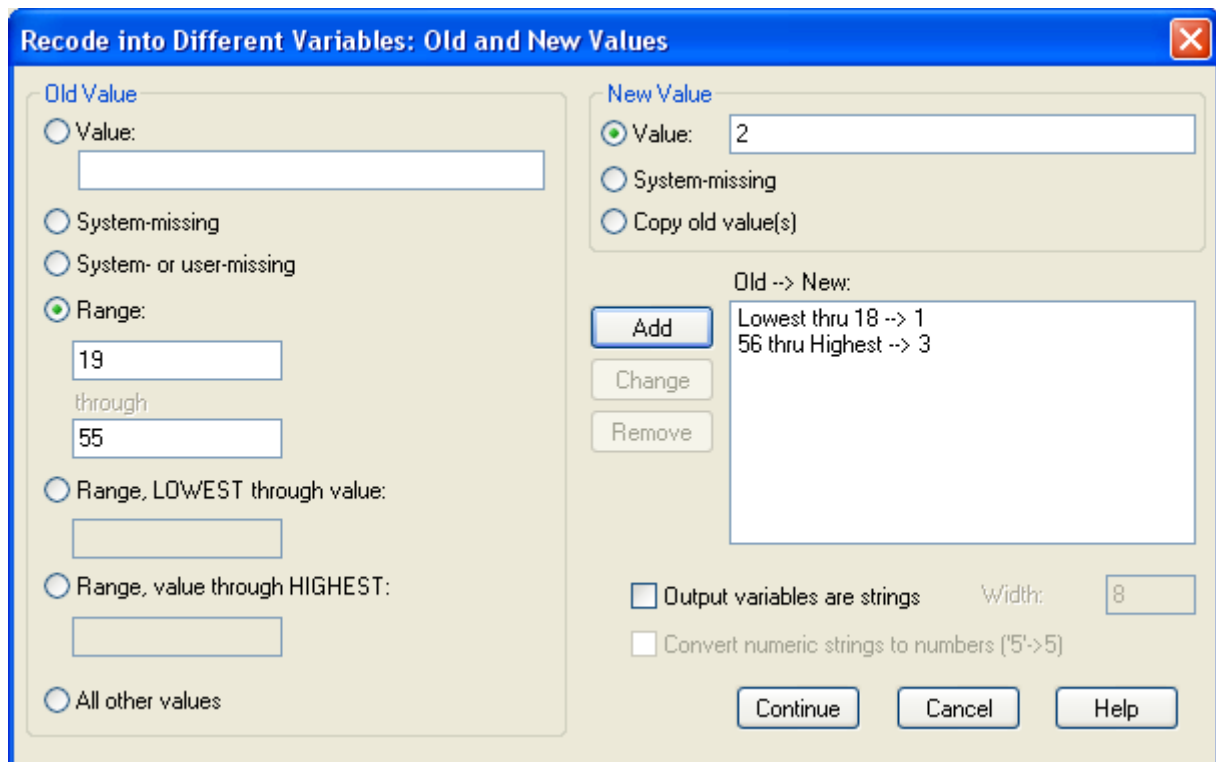
Откройте файл «Moscow_demography_2006». В списке переменных есть две, указывающие на возраст: «age» и «age_group». Первая – количественная, вторая – порядковая с возрастными группами по пять лет. Допустим, что нам нужны возрастные группы, но составленные по другому принципу: до 18 лет, с 19 до 55, от 56 и старше.

На панели управления нужно выбрать выпадающее меню **Transform**, а затем команду **Recode, Into Different Variables**.

В диалоговом окне **Recode into Different Variables** количественной переменной для перекодирования будет «age_group», ее нужно поместить в поле **Numeric Variable - Output Variable**. В поле **Output Variable** введите имя новой переменной, например «age_group_2», и нажмите кнопку **Change**. После чего, оно появится в поле **Numeric Variable - Output Variable**, как это показано на рисунке.



Затем с помощью кнопки  в новом диалоговом окне необходимо задать правило, согласно которому будет перекодирована переменная «age_group».



Правило может выглядеть следующим образом:

Возраст до 18 лет – первая группа, с 19 до 55 лет – вторая, от 56 лет и старше – третья.

Можно поочередно присваивать каждой группе (их всего 21) новую категорию, и воспользоваться для этого переключателем **Value** в поле **Old Value**. В определенных случаях это целесообразно, но в этой задаче удобнее воспользоваться другим способом.

В поле **Old Value** установите переключатель **Range, LOWEST through value** и введите значение 18. В поле **New Value** установите переключатель **Value** и введите значение 1. нажмите **Add**, и в поле **Old-New** ниже появится первое условие. Затем установите переключатель **Range value through HIGHEST** и значение 56. В поле **New Value** введите значение 3 и нажмите **Add**. Для второй категории возрастной категории установите переключатель **Range through** со значениями 19 и 55, а затем значение 2 в поле **New Value**. Еще раз нажмите **Add** и запустите процедуру с помощью кнопок **Continue** и **OK**.

В закладке Variable View появилась новая переменная «age_group_2», принимающая всего 3 значения.

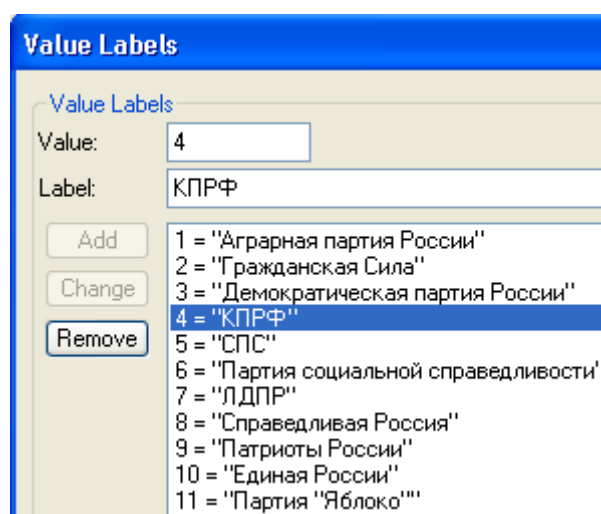
Сравнение двух выборок с неизвестным распределением

Сравнение двух независимых выборок (Mann-Whitney U, Wilcoxon W)

В политологическом анализе возникает необходимость сравнить возрастную структуру электората двух политических партий. Это необходимо, например, для того, чтобы определить круг инструментов, которые будут задействованы в предвыборной кампании. Могут ли они конкурировать за одну и ту же группу избирателей? Читает ли электорат КПРФ и ЛДПР одни и те же печатные издания?

Когда распределение, которым описывается массив данных неизвестно, для того, чтобы сравнить средние значения какого-либо признака в двух независимых выборках, используется непараметрический критерий Вилкоксона. Проверяется нулевая гипотеза о том, распределения двух выборок однородны, против альтернативной гипотезы о том, что они разные. Для ее проверки в статистическом пакете SPSS есть процедура «**2 Independent Samples**», которая находится в выпадающем меню **Analyze, Nonparametric Tests**.

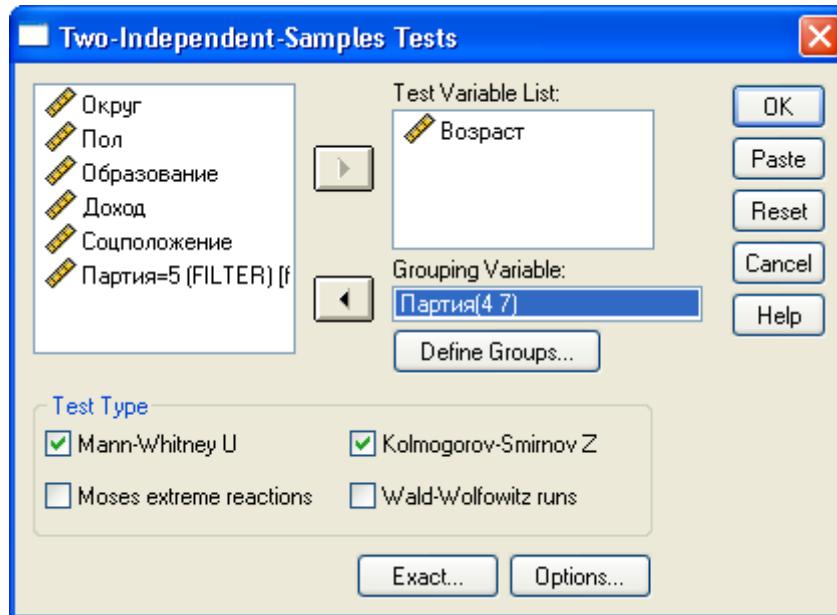
Загрузите файл данных «Данные по выборам 2007». Допустим, анализируемой переменной будет «Возраст», а переменной, которая выделит из всего массива данных две независимые выборки, будет «Партия». Для того чтобы определить выборки с помощью кнопки **Define Groups**, необходимо заранее выяснить, под какими номерами закодированы интересующие нас партии. В окне «**Variable View**» в столбце **Values** перечислены все партии и те номера, под которыми они фигурировали в избирательных списках на парламентских выборах в 2007. Номер КПРФ – «4», ЛДПР – «7». Соответственно, в поля «**Group 1**» и «**Group 2**» окна **Define Groups** нужно поместить цифры 4 и 7. После чего нажмите **OK**.



Критерии, которые будут применены к анализу данных, задаются флажками в поле «**Test Type**». Критерий Манна-Уитни дает возможность установить различия в степени

выраженности переменной в двух выборках Mann-Whitney U. Одновременно со статистикой U, процедура осуществляет подсчет статистики Вилкоксона W.

Диалоговое окно **Two-Independent-Samples Tests** должно выглядеть, как это показано на рисунке. Запустите процедуру.



Первая таблица «**Ranks**» в окне выдачи показывает число наблюдений в каждой выборке, которое может быть неодинаковым, средний ранг по выборке и сумму рангов. Важно отметить, что для подсчета рангов наблюдения были объединены в один вариационный ряд. Таким образом, уже на этом этапе по средним рангам видно, что КПРФ имеет намного более взрослый электорат, чем ЛДПР.

Ranks

Партия	N	Mean Rank	Sum of Ranks
Возраст КПРФ	451	409,65	184753,50
ЛДПР	231	208,44	48149,50
Total	682		

Test Statistics(a)

	Возраст
Mann-Whitney U	21353,500
Wilcoxon W	48149,500
Z	-12,626
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: Партия

В таблице «**Test Statistics**», посчитаны значения статистик Манна-Уитни и Вилкоксона, а также стандартизованное значение статистики Z. На уровне значимости 0,000 гипотезу о равенстве средних можно отвергнуть.

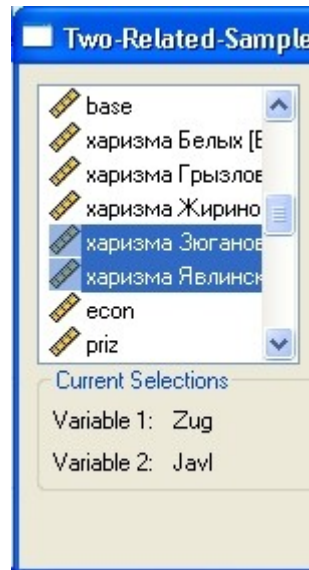
Для того чтобы сравнить распределения трех и более независимых выборок, используется непараметрический критерий Краскела-Уоллиса.

Сравнение двух связанных выборок (Sign Test, Wilcoxon Signed Ranks Test)

В случае, если выборки не являются независимыми (например, повторные наблюдения на одних и тех же объектах до и после воздействия), используется критерий знаков или критерий знаковых ранговых сумм Вилкоксона. Таким способом можно сравнить два измерения переменной, полученных на одной выборке. Проверяется нулевая гипотеза об однородности наблюдений внутри каждой пары (иначе, гипотеза об отсутствии эффекта обработки). Критерий знаков сопоставляет количество положительных и отрицательных разностей значений, затем высчитывается стандартизованное Z значение. Критерий знаковых ранговых сумм Вилкоксона учитывает, насколько велика разность между первой и второй выборкой.

Загрузите файл данных «orgos_05_07», в котором представлены результаты опроса студентов по ряду вопросов: предпочтения операторов мобильной связи, зарубежных стран для проведения отпуска или каникул, восприятие политиков и их хобби.

Сравним, по-разному ли воспринимают Г. Зюганова и Г. Явлинского одни и те же люди. В массиве данных восприятие политиков отражают переменные «харизма Зюганова» «харизма Явлинского». Гипотезу об отсутствии эффекта обработки проверяет процедура «**2 Related Samples**» в выпадающем меню **Analyze, Nonparametric Tests**. В диалоговом окне слева необходимо выбрать две сравниваемые переменные, они появятся ниже в поле «**Current Selections**», и потом перенести их в поле для анализа.



Для того чтобы выбрать критерии, их нужно отметить флажками в поле **Test Type**. Переменная «харизма» измерена в порядковой шкале, поэтому поставим два флажка «**Wilcoxon**» и «**Sign**». Запустите процедуру.



В таблице «**Ranks**» для критерия знаковых ранговых сумм Вилкоксона посчитаны положительные и отрицательные разности, средние ранги для выборок и суммы рангов. Negative Ranks = 121 означает, что 121 раз харизма Явлинского была оценена ниже, чем харизма Зюганова.

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
харизма Явлинского - харизма Зюганова	Negative Ranks	121(d)	105,55	12772,00
	Positive Ranks	77(e)	89,99	6929,00
	Ties	1(f)		
	Total	199		

d харизма Явлинского < харизма Зюганова

e харизма Явлинского > харизма Зюганова

f харизма Явлинского = харизма Зюганова

В таблице «**Test Statistics**» вычислено стандартизованное *Z* значение и двусторонний уровень значимости **Asymp. Sig. (2-tailed)**. Гипотеза об отсутствии эффекта отработки отвергается, т.е. разница в восприятии двух политиков есть.

Test Statistics(b)

	харизма Явлинского - харизма Зюганова
Z	-3,687(a) ,000

a Based on positive ranks.

b Wilcoxon Signed Ranks Test

Таблицы выдачи для критерия знаков устроены аналогичным образом, с той лишь разницей, что в таблице «**Frequencies**» нет информации о рангах и показано только количество положительных и отрицательных разностей.

Sign Test

Frequencies

		N
харизма Явлинского - харизма Зюганова	Negative Differences(b)	121
	Positive Differences(d)	77
	Ties(f)	1
	Total	199

b харизма Явлинского < харизма Зюганова

d харизма Явлинского > харизма Зюганова

f харизма Явлинского = харизма Зюганова

Test Statistics(a)

	харизма Явлинского - харизма Зюганова
Z	-3,056
Asymp. Sig. (2-tailed)	,002

a Sign Test

Сравнение двух выборок с известным распределением

Сравнение двух независимых выборок (Independent-Samples T Test)

Для того чтобы определить принадлежат ли две выборки одной генеральной совокупности, используется t-критерий Стьюдента для двух независимых выборок. Он проверяет две гипотезы: о равенстве дисперсий и равенстве средних значений.

Откроем файл данных «Регионы России». Сравним валовой региональный продукт в 2005 году по двум федеральным округам: Центральному и Южному.

Перед тем, как применить t-критерий, необходимо проверить выборку на нормальность с помощью критерия Колмогорова-Смирнова (процедура «**Explore**»). Значения статистики критерия и уровня значимости позволяют сказать о том, что данные не описываются нормальным распределением. Попробуем преобразовать их и далее работать с натуральным логарифмом ВРП, а не исходными значениями. На уровне значимости 0,200 распределение является нормальным.

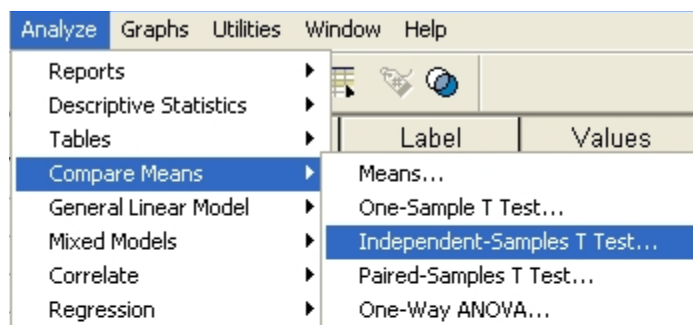
Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Ln_ВРП_2005	,064	79	,200(*)	,965	79	,027

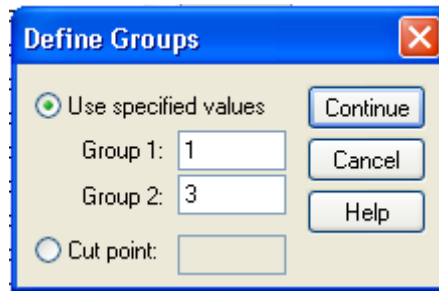
* This is a lower bound of the true significance.

a Lilliefors Significance Correction

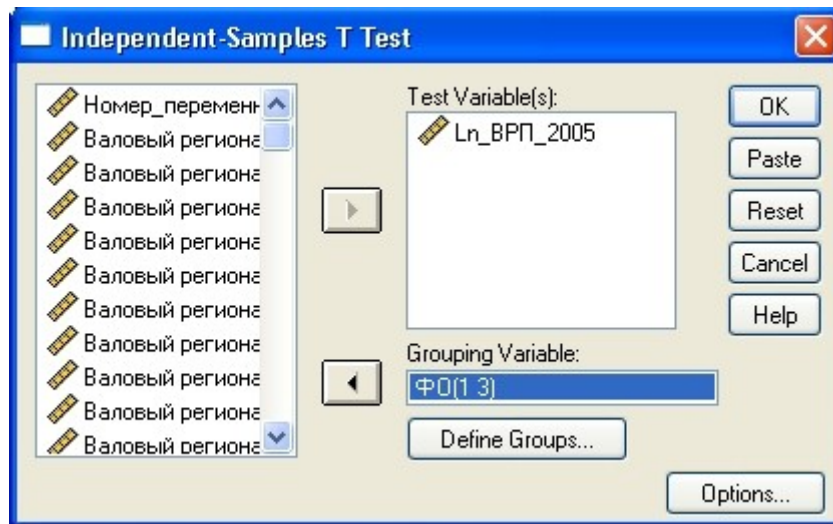
Процедура «**Independent-Samples T Test**» находится в меню **Analyze, Compare Means**.



В появившемся диалоговом окне необходимо выбрать две переменные: «Ln_ВРП_2005», которая будет проанализирована, и качественная переменная «Федеральный округ», по которой будет разделена выборка. Первую необходимо перенести в поле «Test Variable(s)», вторую – в поле «Grouping Variable». Конкретные федеральные округа, определяются через кнопку **Define Groups**. ЦФО соответствует значению 1 переменной «Федеральный округ» и ЮФО – значение 3.



Окно «Independent-Samples T Test» должно выглядеть, как это показано на рисунке. Нажмите **ОК**.



В первой таблице «**Group Statistics**» окна выдачи представлены некоторые описательные статистики.

T-Test

Group Statistics

Федеральный округ		N	Mean	Std. Deviation	Std. Error Mean
Ln_ВРП_2005	ЦФО	18	11,2574	,48172	,11354
	ЮФО	12	10,7172	,44526	,12853

Следующую таблицу «**Independent Samples Test**» можно разбить на две части: проверка равенства дисперсий и проверка равенства средних для двух случаев.

Критерий Ливиня вычисляет значение статистики с F-распределением Фишера. Согласно расчетам, гипотезу о равенстве дисперсий на уровне значимости 0,931 отвергнуть нельзя.

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
Ln_ВРП_2005	Equal variances assumed	,008	,931
	Equal variances not assumed		

Далее, во второй части таблицы, нас будет интересовать строка, которой соответствует допущение о равенстве дисперсий. На основании значения t-статистики и уровня значимости принимается решение отвергнуть нулевую гипотезу. Таблица содержит оценку средней разницы в ВРП регионов двух округов, а так же верхнее и нижнее значения доверительного интервала этой оценки.

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Ln_ВРП_2005	Equal variances assumed	3,099	28	,004	,54017	,17431	,18310	,89723
	Equal variances not assumed	3,150	25,011	,004	,54017	,17150	,18696	,89338

Сравнение двух связанных выборок (Related-Samples T Test)

Сравнить объекты до и после некоторого воздействия (события или просто по истечении периода времени) можно с помощью t-критерий Стьюдента для двух связанных выборок. Этот метод применяется в том случае, если данные описываются известным распределением. Нулевая гипотеза в этом случае звучит так: различия между наблюдениями в паре отсутствуют.

Процедура «**Related-Samples T Test**» находится в меню **Analyze, Compare Means**. В появившемся диалоговом окне необходимо выбрать две переменные, по которым будет проведено сравнение. Пусть это будут прологарифмированные значения ВРП по регионам России в 1995 и 2005 годах. Они отразились в поле «Current Selections»



Выбранные переменные следует перенести в поле «Paired Variables».



В данном случае наблюдения очевидно являются парными, т.к. показатель характеризует те же самые объекты, но спустя десять лет. Запустите процедуру.

T-Test

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Ln_ВРП_2005	11,3269	79	,57025	,06416
Ln_ВРП_1995	8,9406	79	,48796	,05490

Таблица «**Paired Samples Correlations**» позволяет узнать, что наблюдения коррелируют на уровне 0,912, и что эта корреляция значима. Необходимо учитывать, что при недостаточно большом количестве наблюдений, даже столь высокие значения коэффициента корреляция могут оказаться незначимы.

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 Ln_ВРП_2005 & Ln_ВРП_1995	79	,912	,000

В таблице «**Paired Samples Test**» нас интересует значение наблюдаемой t-статистики и уровня значимости. Нулевая гипотеза о том, что различия между наблюдениями в паре отсутствуют, отвергается. Положительное значение показателя **Mean** в столбце **Paired Differences** говорит о том, что логарифм ВРП в 2005 году выше, чем в 1995.

Paired Samples Test

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Pair 1 Ln_ВРП_2005 - Ln_ВРП_1995	2,38624	,23623	89,781	78	,000

Сравнение нескольких независимых выборок

Часто в сравнительном анализе возникает задача выяснить, насколько значимым является тот или иной фактор при сравнении стран по некоторому признаку. Иными словами, можно ли объяснить изменчивость признака разницей в том, что на выборки оказали воздействие факторы разного уровня? Здесь встает задача проверки гипотезы о том, что выборки принадлежат одному и тому же распределению.

Задача может иметь *параметрическую* и *непараметрическую* постановку в зависимости от того, будет ли зависимая переменная (отклик) иметь нормальное распределение.

Загрузите файл данных «country_compar_24-02-2010». Устройство файла таково, что каждая страна за определенный временной период (год) выступает как отдельное наблюдение и потому несколько раз повторяется в столбце (такие данные называются *панельными*). Необходимо отобрать только те наблюдения, которые относятся к одному году, например, к 2005, с помощью процедуры **Select Cases**.

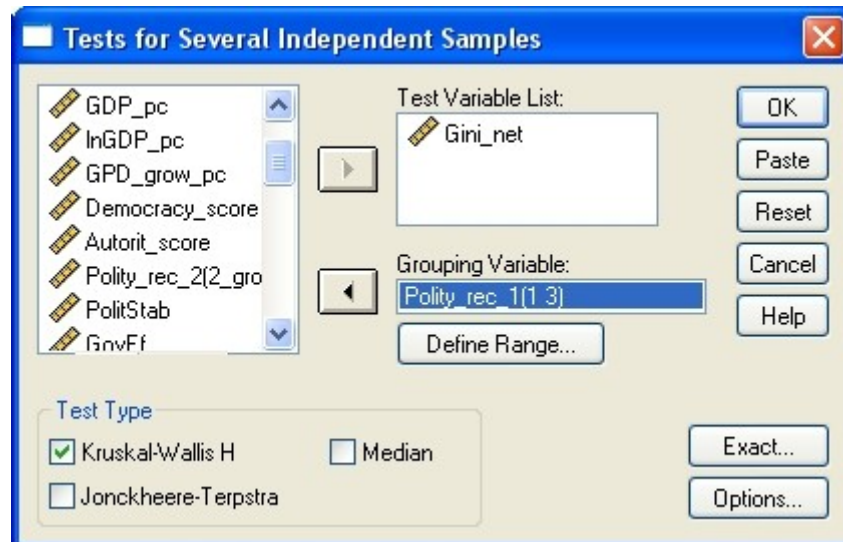
Все страны можно разбить на нескольких независимых выборок и сравнить их по выбранным для анализа показателям. В качестве независимой переменной выберем значение индекса POLITY IV. Она является номинальной и имеет три градации: демократические, переходные и авторитарные режимы. (описание файла дано в [...]).

Критерий Краскела-Уоллиса (Kruskal-Wallis H)

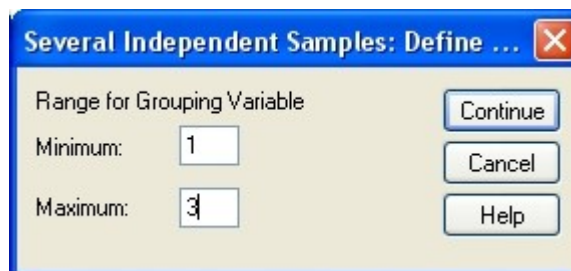
Критерий Краскела-Уоллиса (**Kruskal-Wallis H**) относится к непараметрическим методам и применяется в случаях, когда распределение отклика неизвестно, ошибки являются независимыми одинаково распределенными случайными величинами.

Зависимой переменной будет значение коэффициента Джини в 2005 году («Gini_net»), фактором, влияющим на отклик, – политический режим (переменная «Polity_rec_1»).

Процедура **K independent samples** для вычисления статистики H Краскела-Уоллиса находится в меню **Analyze, Nonparametric tests**.



Минимальное и максимальное значения фактора задаются через нажатие кнопки **Define Ranges**. В нашем случае, в окне **Several independent samples: Define Ranges** нужно поставить 1 и 3 так, как показано на рисунке ниже, предварительно определив, какие значения принимает переменная и какие мы будем анализировать. В поле Test Type необходимо поставить флажок **Kruskal-Wallis H**.



В окне выдачи представлено посчитанное значение H-статистики, имеющей распределение хи-квадрат, оно равно 1,775. Учитывая, что уровень значимости Asymp. Sig. равен 0,412, нулевую гипотезу о том, что степень неравенства в доходах в обществе не зависит от политического режима, отвергнуть нельзя.

Test Statistics(a,b)

	Gini_net
Chi-Square	1,775
df	2
Asymp. Sig.	,412

a Kruskal Wallis Test

b Grouping Variable: Polity_rec_1(3_groups)

Ranks

Polity rec 1(3 groups)		N	Mean Rank
Gini_net	1	7	31,57
	2	3	50,33
	3	61	35,80
Total		71	

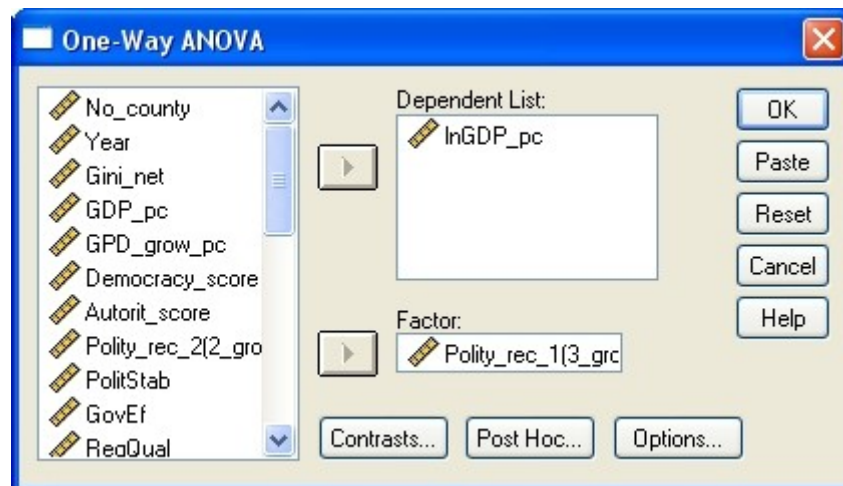
В таблице «Ranks» посчитаны средние ранги коэффициента Джини по каждой выборке.

Однофакторный дисперсионный анализ (One-Way ANOVA)

Для аддитивной модели факторного анализа $x_{ij} = a_j + e_{ij}$, где $a_j = \mu + \tau_j$, x_{ij} – изучаемый признак, а μ – среднее значение по выборке и об ошибках e_{ij} можно сказать, что они независимы, одинаково распределены, имеют нулевое среднее значение и постоянную дисперсию, т.е. описываются нормальным распределением.

Проверим, являются ли статистически значимыми различия в размере ВВП на душу населения у стран с разными политическими режимами. В качестве зависимой переменной используем натуральный логарифм ВВП, имеющий нормальное распределение.

Вновь отберем показатели 2005 года. Затем в выпадающем меню **Analyze** необходимо выбрать процедуру **Compare Means, One-Way**, и перенести переменную «lnGDP_pc» в поле **Dependent List**, а «Polity_rec_1» – в поле **Factor**. Запустим процедуру.



Значения статистики Фишера и уровня значимости позволяют нам отвергнуть нулевую гипотезу о том, у стран с разными политическими режимами нет различий в подушевом ВВП.

ANOVA

lnGDP_pc

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27,571	2	13,785	10,941	,000
Within Groups	133,560	106	1,260		
Total	161,131	108			

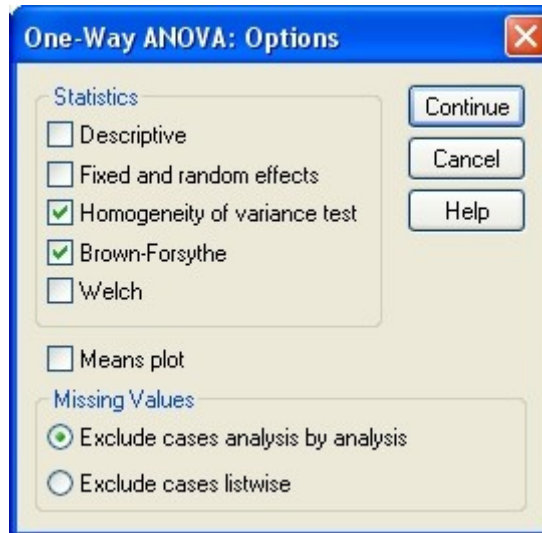
В окне **Options** где поставим флажки **Homogeneity of variance test** и **Brown-Forsythe**. Тем самым мы можем проверить выборки на равенство дисперсий, т.к. в таблице «Test of Homogeneity of Variances» будет посчитана статистика Левина. Этот критерий не требует нормальности распределения данных. Уровень значимости высок, 0,145, поэтому отвергнуть гипотезу о равенстве дисперсий мы не можем.

Test of Homogeneity of Variances

lnGDP_pc

Levene Statistic	df1	df2	Sig.
1,967	2	106	,145

Для проверки равенства средних между группами можно применить критерий Brown-Forsythe. В случаях, когда дисперсии выборок не являются равными, статистику Brown-Forsythe предпочтительнее, чем F-статистика.



Robust Tests of Equality of Means

lnGDP_pc

	Statistic(a)	df1	df2	Sig.
Brown-Forsythe	12,668	2	24,405	,000

a. Asymptotically F distributed.

Табличное значение статистики Brown-Forsythe и уровня значимости позволяют отвергнуть гипотезу о равенстве средних.

После того, как была установлена статистически значимая разница в подушевом ВВП между группами, нас может заинтересовать, между какими именно группами существует разница. Установим дополнительные настройки. Для этого нажмем кнопку **Post Hoc** (Постфактум). В окне **One-Way ANOVA: Post Hoc Multiple Comparisons** предложено множество критериев для процедуры проведения попарных сравнений в случаях с равной и разной дисперсией выборок.

Отметим флажком тест **Scheffe** в поле **Equal Variances Assumed**. Этот критерий основан на сравнении возможных комбинаций средних значений и использует F-распределение Фишера.



В таблице «Multiple Comparisons» представлены попарные сравнения средних для трех выборок. Значения уровней значимости из столбца «Sig.» позволяют сделать вывод о том, что разницы в подушевом ВВП между выборками 1 и 2 нет. Средние значения логарифма подушевого ВВП из таблицы «lnGDP_pc» подтверждают этот вывод: 7,7104 и 7,9883 для первой и второй выборок, и 9,0101 для третьей. Размеры выборок не одинаковы, поэтому было посчитано гармоническое среднее.

Multiple Comparisons

Dependent Variable: lnGDP_pc
Scheffe

(I) Polity_rec_1(3_grou ps)	(J) Polity_rec_1(3_groups)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	,27794	,42959	,811	-,7886	1,3445
	3	-1,02177(*)	,29283	,003	-1,7488	-,2947
2	1	-,27794	,42959	,811	-1,3445	,7886
	3	-1,29971(*)	,36096	,002	-2,1959	-,4035
3	1	1,02177(*)	,29283	,003	,2947	1,7488
	2	1,29971(*)	,36096	,002	,4035	2,1959

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

lnGDP_pc

Scheffe

Polity_rec_1(3_groups)	N	Subset for alpha = .05	
		1	2
2	11	7,7104	
1	18	7,9883	
3	80		9,0101
Sig.		,749	1,000

Means for groups in homogeneous subsets are displayed.

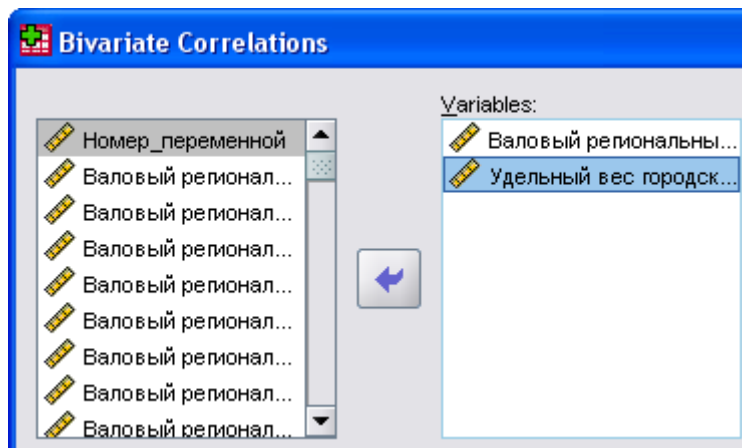
a Uses Harmonic Mean Sample Size = 18,872.

b The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

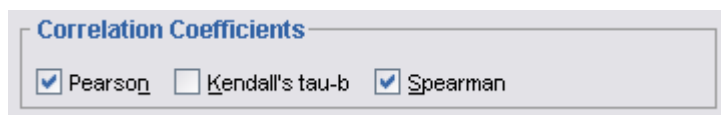
Корреляционный анализ

При анализе данных может возникнуть задача выяснить насколько тесно взаимосвязаны исследуемые признаки. Для проверки того, существует ли статистически значимая связь между переменными, можно посчитать коэффициент корреляции. Если Ваши переменные измерены в интервальной (количественной) шкале, то рассчитывается значение коэффициента корреляции К. Пирсона, а если Ваши переменные измерены в порядковой (ординальной) шкале, то рассчитывается значение коэффициента ранговой корреляции Ч.Э. Спирмена.

Проверим, существовала ли статистически значимая связь между размером валового регионального продукта (ВРП) на душу населения и удельным весом городского населения в общей численности населения региона (в %) в 2005 году. Для этого откройте файл данных «Регионы России 2». В меню **Analyze** выберите команду **Correlate, Bivariate**. В диалоговом окне Bivariate Correlations поместите исследуемые переменные в поле Variables (переменных может быть более двух, если перед Вами стоит задача установления попарной связи между тремя и более переменными).



Т.к. обе переменные измерены в интервальной шкале, то в поле Correlation Coefficients был бы достаточно оставить флажок Pearson. При этом нужно учитывать то, что коэффициента корреляции Пирсона менее робастен, чем коэффициента корреляции Спирмена, поэтому отметим еще и Spearman, как это показано на рисунке ниже.



Для случаев, когда исследуемых переменных несколько и поиск статистически значимых коэффициентов корреляции в большой таблице окна выдачи становится

затруднительным, облегчить задачу можно поставив флажок Flag significant correlations



В таблице **Correlations** окна выдачи показаны значение коэффициента корреляции, уровень значимости и число наблюдений. Таким образом, между размером ВРП на душу населения и удельным весом городского населения существует статистически значимая связь, и в социальных науках значение коэффициента корреляции, равное 0,427, может считаться неплохим показателем.

		ВРП на душу в 2005 г. (руб.)	Удельный вес городского населения
ВРП на душу в 2005 г. (руб.)	Pearson Correlation	1	,427**
	Sig. (2-tailed)		,000
	N	79	79
Удельный вес городского населения	Pearson Correlation	,427**	1
	Sig. (2-tailed)	,000	
	N	79	86

** . Correlation is significant at the 0.01 level (2-tailed).

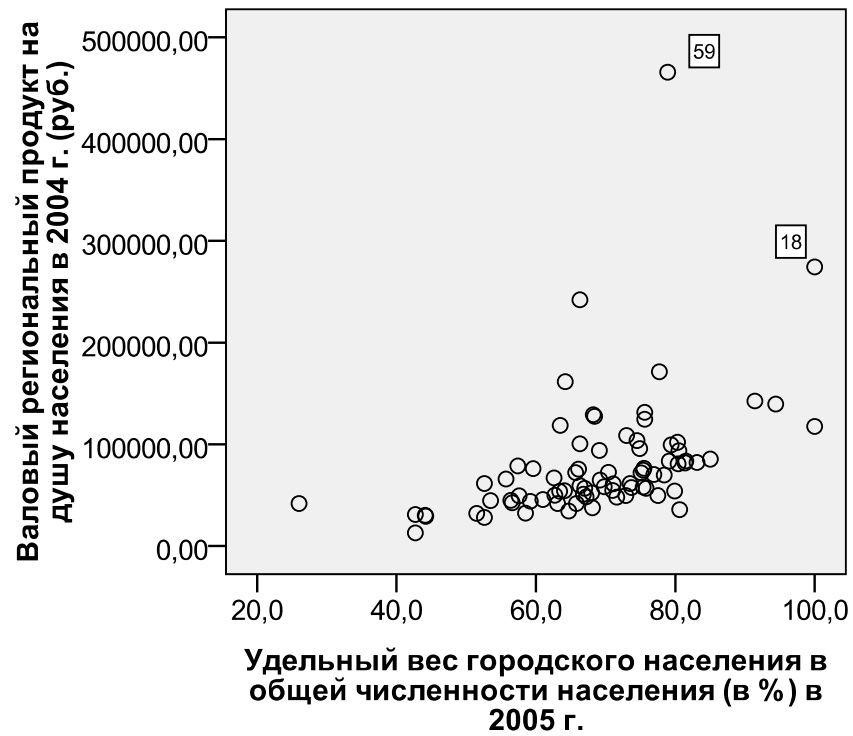
Вторая таблица содержит аналогичную информацию. Коэффициент ранговой корреляции Спирмена (Spearman's rho) имеет значение 0,616.

Nonparametric Correlations

		ВРП на душу в 2005 г. (руб.)	Удельный вес городского населения
Spearman's rho ВРП на душу в 2005 г. (руб.)	Correlation Coefficient	1,000	,616**
	Sig. (2-tailed)		,000
	N	79	79
Удельный вес городского населения	Correlation Coefficient	,616**	1,000
	Sig. (2-tailed)	,000	
	N	79	86

** . Correlation is significant at the 0.01 level (2-tailed).

Большая разница между двумя значениями коэффициентов корреляции обусловлена тем, что есть наблюдение (№ 59), которое сильно отличается от остальных. Это Тюменский регион, который имел большой душевой размер ВРП ввиду сверхдоходов от экспорта нефти. Второй выброс в рассматриваемом массиве – наблюдение № 18, Москва (что можно установить из описательных статистик для рассматриваемой переменной).



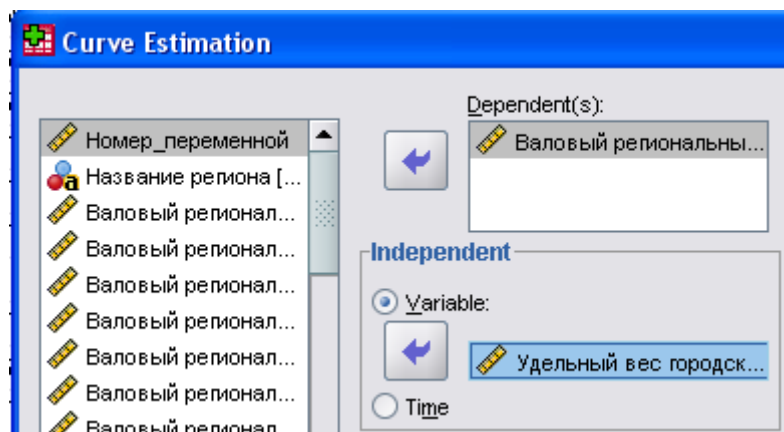
Регрессионный анализ

В случае, когда перед Вами стоит задача установления не просто взаимосвязи, а причинно-следственной связи, то для ее решения можно оценить регрессионную модель

$$y_i = \alpha + \beta X_i + \varepsilon_i.$$

в которой y – зависимая переменная (отклик), а X_i – независимая переменная (предиктор, фактор, регрессор), α – константа (интерсепт), β – коэффициент перед независимой переменной, а ε – случайный член.

Откройте файл данных «Регионы России 2». Чтобы оценить вид зависимости, в меню **Analyze** выберите команду **Regression, Curve Estimation**. В диалоговом окне Curve Estimation необходимо перенести зависимую переменную в поле **Dependent(s)**, а независимую – в поле **Independent**, установив флажок **Variable**. Флажок **Time** используется, когда в качестве независимой переменной выступает время.



В оцениваемой модели мы хотим посмотреть влияет ли Удельный вес городского населения в общей численности населения (в %) в 2005 г. (независимая переменная *Гор_нас_2005*) на Валовый региональный продукт на душу населения в 2005 г. (руб.) (зависимая переменная *ВРП_2005*).

В поле **Models** необходимо установить несколько флажков помимо «Linear», если ожидаемый тип связи не является линейным.

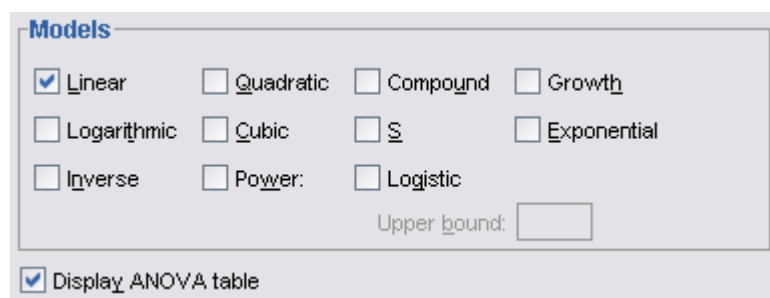
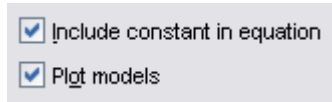
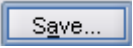
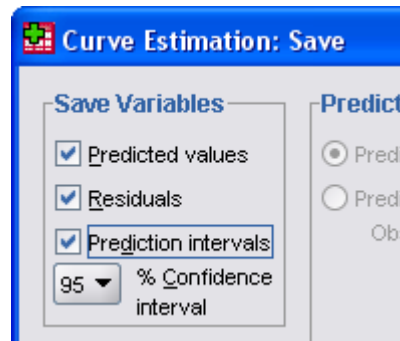


Таблица дисперсионного анализа (ANOVA) позволяет получить значение F-статистики для оцениваемой модели, поэтому рекомендуется поставить флажок Display ANOVA table.

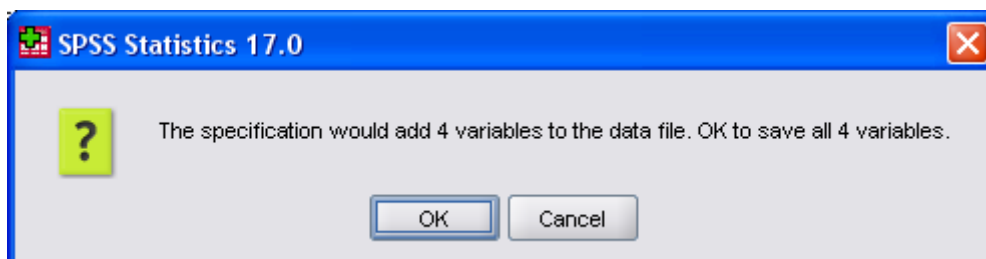
Для того чтобы включить в уравнение регрессии константу и получить диаграмму рассеяния, необходимо соответственно поставить флажки Include constant in equation и Plot models.



Существует возможность сохранить оцененные значения зависимой переменной, остатки, а также границы 95 %-ного доверительного интервала для зависимой переменной с помощью кнопки . В появившемся диалоговом окне **Curve Estimation: Save** поставьте флажки напротив тех переменных, которые необходимы Вам для дальнейшего анализа (рекомендуется как минимум сохранить остатки (*Residuals*)).



Запустите процедуру, нажав ОК. Перед тем как выдать результаты регрессионного анализа программа предупреждает о том, что в массив данных будут добавлены новые переменные, которые были запрошены Вами.



В окне выдачи с содержательной стороны интересны следующие таблицы:

Model Summary, в которой показаны коэффициент корреляции (R), коэффициент детерминации (R Square) и скорректированный коэффициент детерминации (Adjusted R Square). Оцененная модель может быть охарактеризована как плохая, т.к. значение коэффициента детерминации относительно низкое.

Linear

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,427	,182	,172	77761,912

The independent variable is Удельный вес городского населения в общей численности населения (в %) в 2005 г..

В таблице дисперсионного анализа (**ANOVA**) представлены объясненная регрессией сумма квадратов (Regression), сумма квадратов остатков (Residual), общая сумма квадратов (Total), соответствующие им степени свободы, средний квадрат, значение F-статистики и уровень значимости. F-тест проверяет на значимость одновременно все коэффициенты в модели, и нулевая гипотеза формулируется так:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0, \text{ где } k - \text{число независимых переменных.}$$

Таким образом, на значение F-статистики целесообразно смотреть, когда оценивается модель множественной регрессии. В нашем случае нулевая гипотеза может быть отвергнута на уровне значимости 0,000.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1,037E11	1	1,037E11	17,154	,000
Residual	4,656E11	77	6,047E9		
Total	5,693E11	78			

The independent variable is Удельный вес городского населения в общей численности населения (в %) в 2005 г..

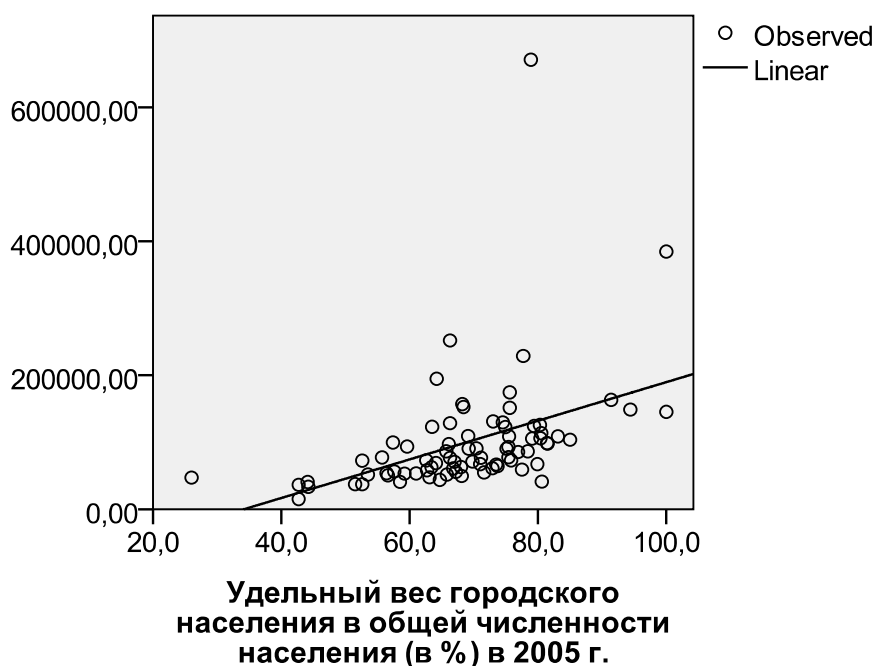
Таблица **Coefficients** содержит значения коэффициентов, стандартные ошибки коэффициентов, стандартизованные значения коэффициентов, t-статистики и уровни значимости. t-тест проверяет гипотезу о равенстве истинного значения коэффициента нулю. Из таблицы видно, что на уровне значимости 0,05 статистически значимы все коэффициенты: константа и коэффициент β перед переменной «Удельный вес городского населения».

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Удельный вес город. населения в общей численности населения (в %) в 2005 г.	2879,858	695,333	,427	4,142	,000
(Constant)	-98143,989	48673,046		-2,016	,047

На диаграмме рассеяния показана регрессионная линия. Два отдельно стоящих наблюдения, выброса, – Тюменская область и Москва. Они являются выбросами, что можно увидеть в описательных статистиках переменной ВРП_2005.

Валовый региональный продукт на душу населения в 2005 г. (руб.)



Если не исключить два наблюдения из анализа, то значения коэффициента детерминации и коэффициента перед независимой переменной изменятся.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,502	,252	,242	39100,077

The independent variable is Удельный вес городского населения в общей численности населения (в %) в 2005 г..

ANOVA

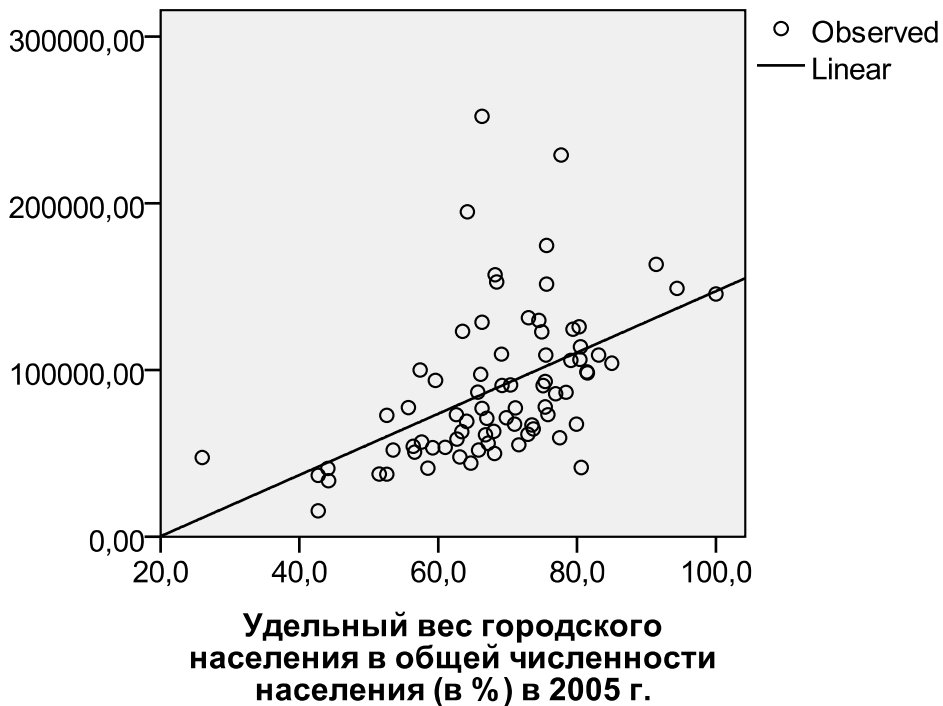
	Sum of Squares	df	Mean Square	F	Sig.
Regression	3,854E10	1	3,854E10	25,206	,000
Residual	1,147E11	75	1,529E9		
Total	1,532E11	76			

The independent variable is Удельный вес городского населения в общей численности населения (в %) в 2005 г..

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Удельный вес городского населения в общей численности населения (в %) в 2005 г.	1837,406	365,977	,502	5,021	,000
(Constant)	-36484,692	25399,140		-1,436	,155

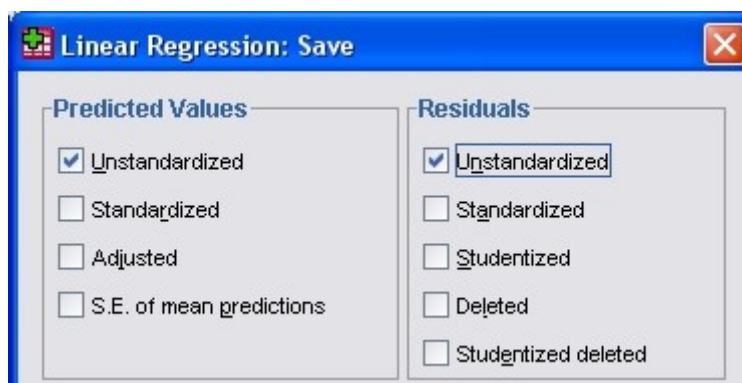
Валовый региональный продукт на душу населения в 2005 г. (руб.)



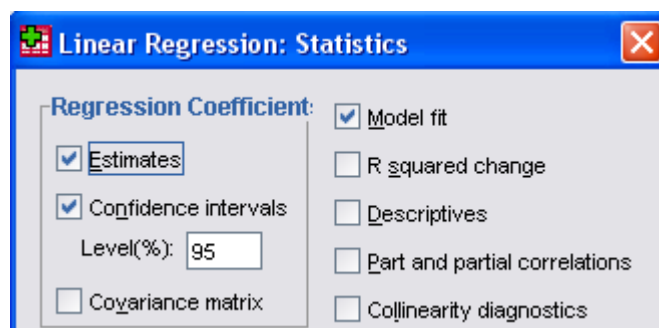
Все возможности для оценивания *модели линейной регрессии* доступны в меню **Analyze, Regression, Linear**. Они в значительной степени уже были описаны при оценивании вида зависимости в модели, поэтому мы сосредоточимся на дополнительных возможностях команды **Linear Regression**.

Определив переменные для модели аналогично тому, как Вы это уже делали, и оставьте метод Enter . Его можно заменить на Stepwise, если Вы оцениваете модель множественной регрессии и хотели бы, чтобы статистический пакет самостоятельно исключил из модели статистически незначимые объясняющие переменные. Но стоит иметь в виду, что зачастую присутствие некоторых переменных в уравнении является обязательным с точки зрения теории.

Нажмите кнопку и сохраните нестандартизованные оцененные значения зависимой переменной (т.н. *y-hat* или *y* с крышкой) и остатки.



Нажав на кнопку Statistics можно поставить флажок Confidence intervals и получить доверительные интервалы для оценок коэффициентов регрессии. Результаты будут показаны в таблице **Coefficients^a** в столбце **95,0% Confidence Interval for B**.

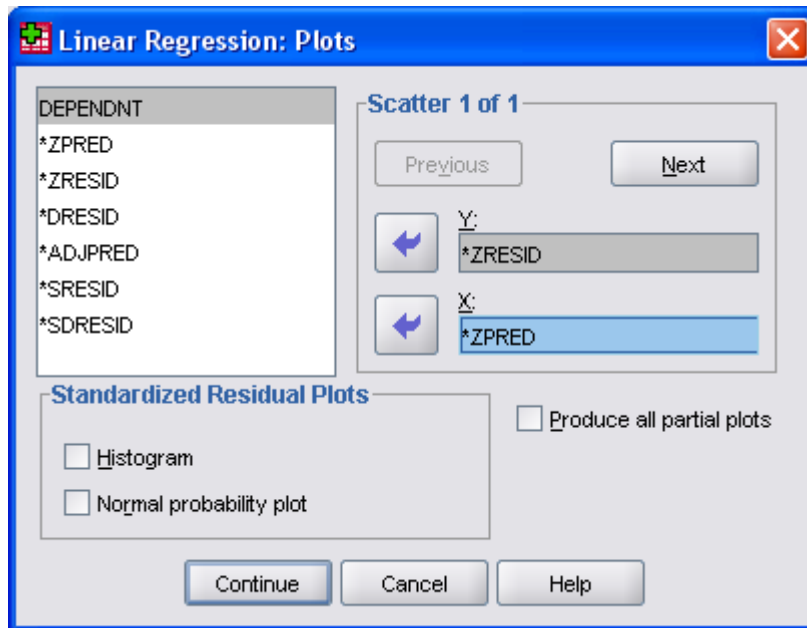


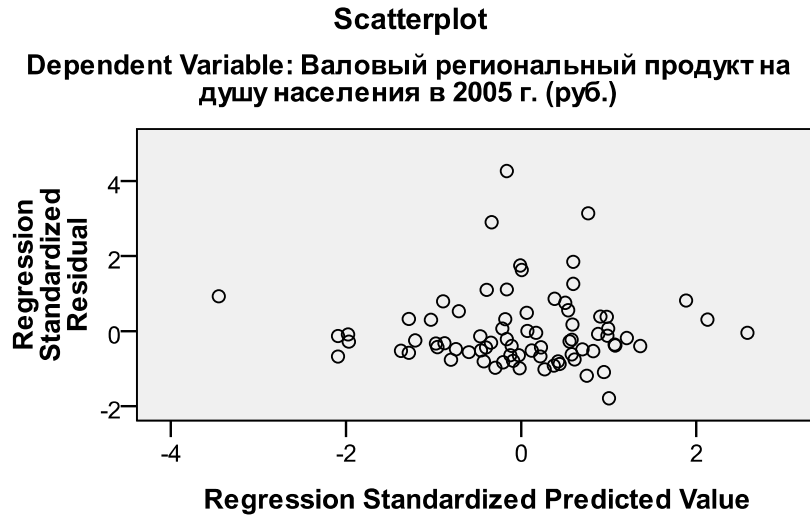
Coefficients^a

Model		Unstandardized Coefficients		95,0% Confidence Interval for B	
		B	Std. Error	Lower Bound	Upper Bound
1	(Constant)	-36484,692	25399,140	-87082,374	14112,989
	Удельный вес город. населения в общей численности населения (в %) в 2005 г.	1837,406	365,977	1108,343	2566,469

a. Dependent Variable: Валовой региональный продукт на душу населения в 2005 г. (руб.)

Нажав на кнопку Plots можно получить диаграммы рассеяния остатков против оцененных значений зависимой переменной. Для этого необходимо перенести переменную ZRESID в поле Y (т.е. остатки будут отложены по оси Y), а переменную ZPRED в поле X, и нажать Continue. Тогда в окне выдачи появится диаграмма рассеяния, как это показано на рисунке.





Дополнительно в окне выдачи появляется таблица с описательными статистиками для остатков **Residuals Statistics^a**.

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	11287,8613	147255,8906	89055,4688	22517,62745	77
Residual	-70088,62500	1,66762E5	,00000	38841,98761	77
Std. Predicted Value	-3,454	2,585	,000	1,000	77
Std. Residual	-1,793	4,265	,000	,993	77

a. Dependent Variable: Валовый региональный продукт на душу населения в 2005 г. (руб.)