

ЛАБОРАТОРНАЯ РАБОТА № 1

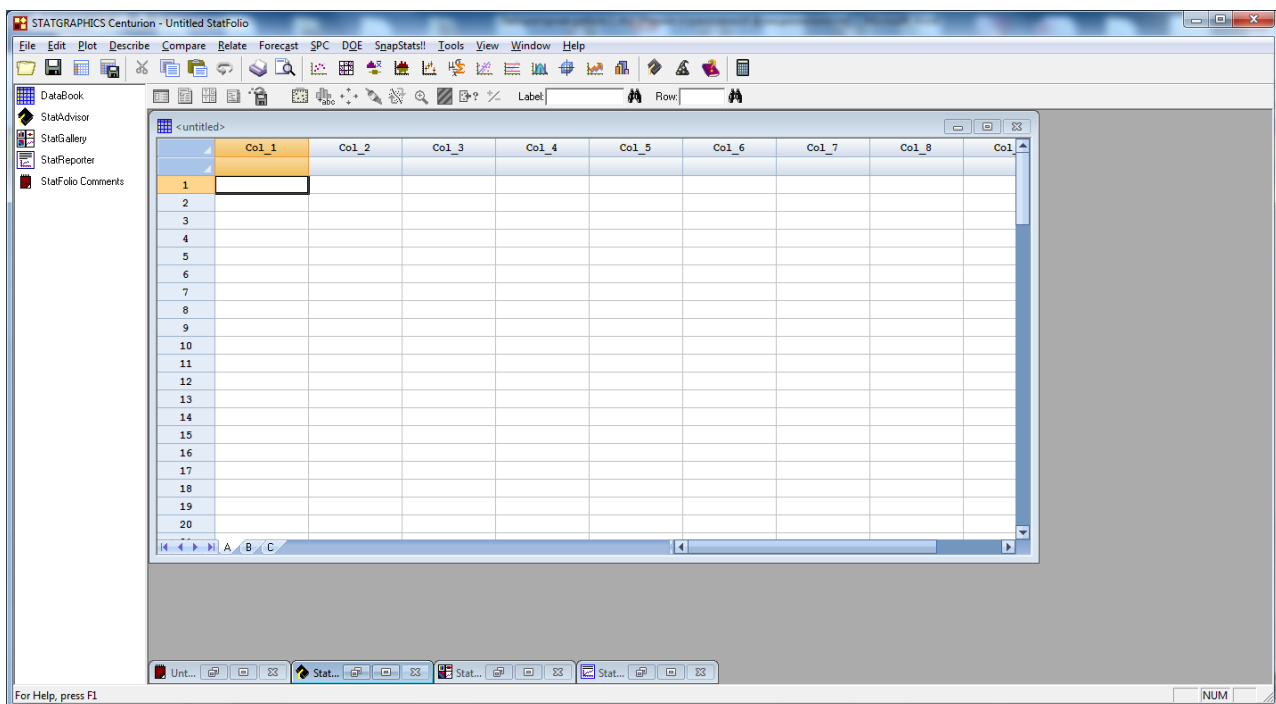
ВВОД, ПЕРВИЧНАЯ ОБРАБОТКА И ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ ДАННЫХ

Пусть мы имеем выборку из нескольких чисел. Возникает простейшая задача: ввести данные, сгруппировать их (построить ряды частот, частостей, накопленных частот и накопленных частостей). Затем необходимо вычислить *выборочные числовые характеристики*: среднее, моду, медиану, размах, дисперсию, стандартное отклонение, асимметрию, эксцесс. Все это достаточно легко сделать с помощью пакета Statgraphics Centurion.

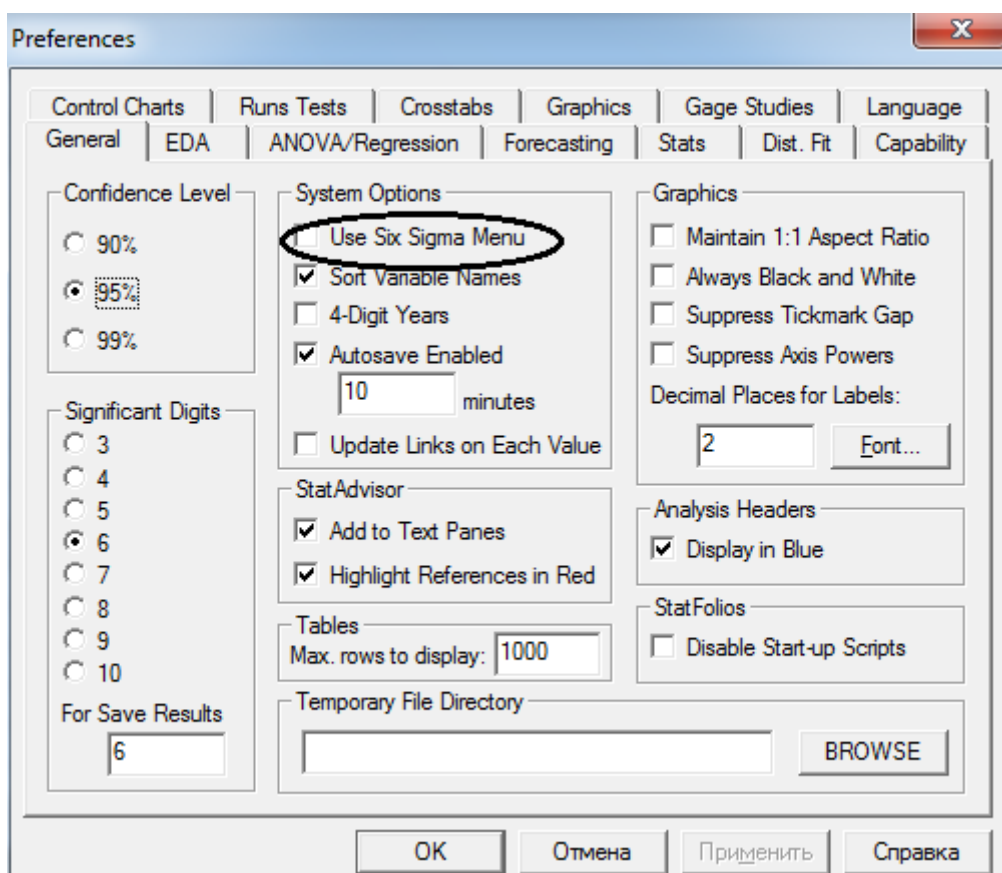
ЗАДАЧА. Администрацию универсама интересует оптимальный уровень запасов продуктов в торговом зале, а также среднемесячный объем покупок товаров, не являющихся предметом ежедневного потребления в семье (таких, например, как сода). Для выяснения этого вопроса менеджер универсама в течение января регистрировал частоту покупок стограммовых пакетиков с содой и собрал следующие данные (x_i):

4 4 9 3 3 1 7 0 4 2 3 5 7 10 6 5 7 3 2 9 8 1 4 2 1 0 8 9

Войдем в *Statgraphics*. Экран должен иметь следующий вид:

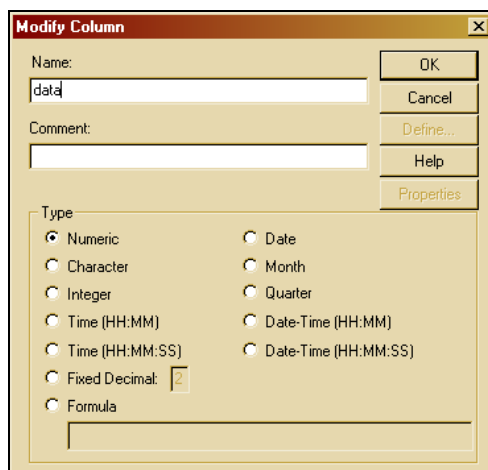


В пакете есть выбор из двух меню, поэтому, если у вас меню не такое, как здесь, надо его перенастроить. Для этого в меню выберите **Edit, Preferences**, откроется окно, в котором надо *отключить* **Use Six Sigma Menu**.



Введем данные в первый столбец окна **Data**. Щелкните там мышкой и вводите числа, по одному в каждой ячейке. Переход в следующую ячейку по нажатию клавиши Enter.

Если вдруг вам захочется изменить название столбца или тип данных, для этого щелкните мышкой по нынешнему названию, столбец выделится. Щелкните правой кнопкой по заголовку, появится контекстное меню. Выберите в нем **Modify Column**, затем щелкните левой кнопкой, на экране появится диалоговое окно, в котором можно менять имя и тип.



Название должно быть написано латинскими буквами. (Поле **Name** содержит от 1 до 32 символов. Не должно быть следующих 19 символов: ‘ “ . > < ~ ! & , ; + - * / ^ = | () Имя не может начинаться с цифры. Пробелы допустимы. Имя не чувствительно к регистру. Поле **Comment** содержит от 0 до 64 символов. Оно служит для того чтобы ввести дополнительную информацию о содержимом столбца, может содержать русские буквы, но лучше его написать на английском.)

В нижней части окна можно выбрать тип данных. Поскольку вы в этом примере работаете с целыми числами, выберите **Integer**.

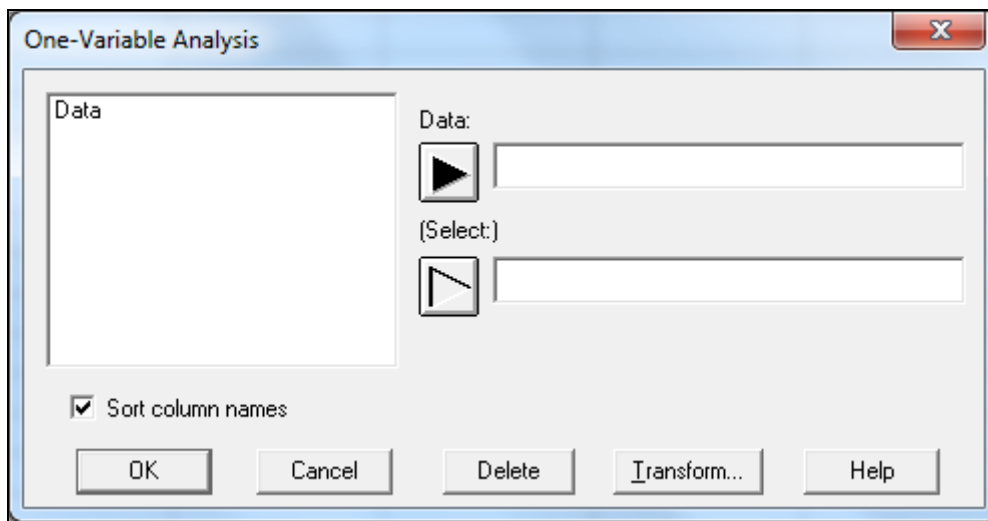
Введите название **Data**, тип оставьте **Numeric** и нажмите OK.


После заполнения таблицы сохраним файл данных. *Обратите внимание*: мы сейчас сохраняем только данные, введенные нами (они будут иметь расширение *sf6*.) Также пакет может работать с данными предыдущих версий, у них расширение *sf*. В пакете *StatGraphics* есть возможность сохранить результаты анализа данных (это называется *StatFolio*), этот файл хранится отдельно от файла данных и имеет расширение *.sgp*.

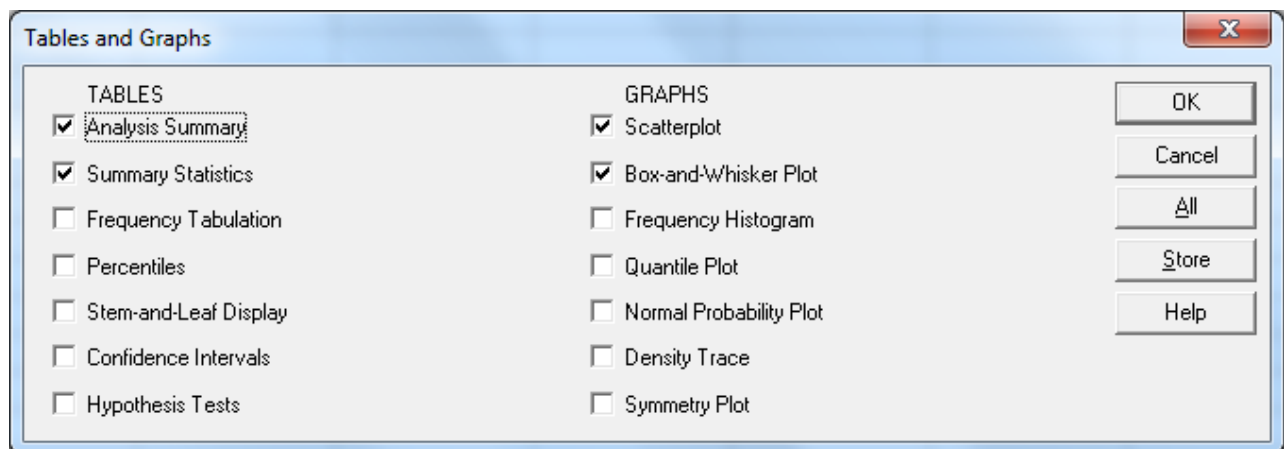
Для сохранения данных в строке меню выберите **Save As**, затем в раскрывшемся меню выберите **Save Data File As**, потом выберите папку, в которой хотите сохранить файл, напишите желаемое имя файла, нажмите на кнопку **Сохранить**. То же самое можно сделать, нажав клавишу **F12**.

Займемся описательным статистическим анализом введенных данных, найдем все интересующие нас числовые характеристики выборки. В строке ме-

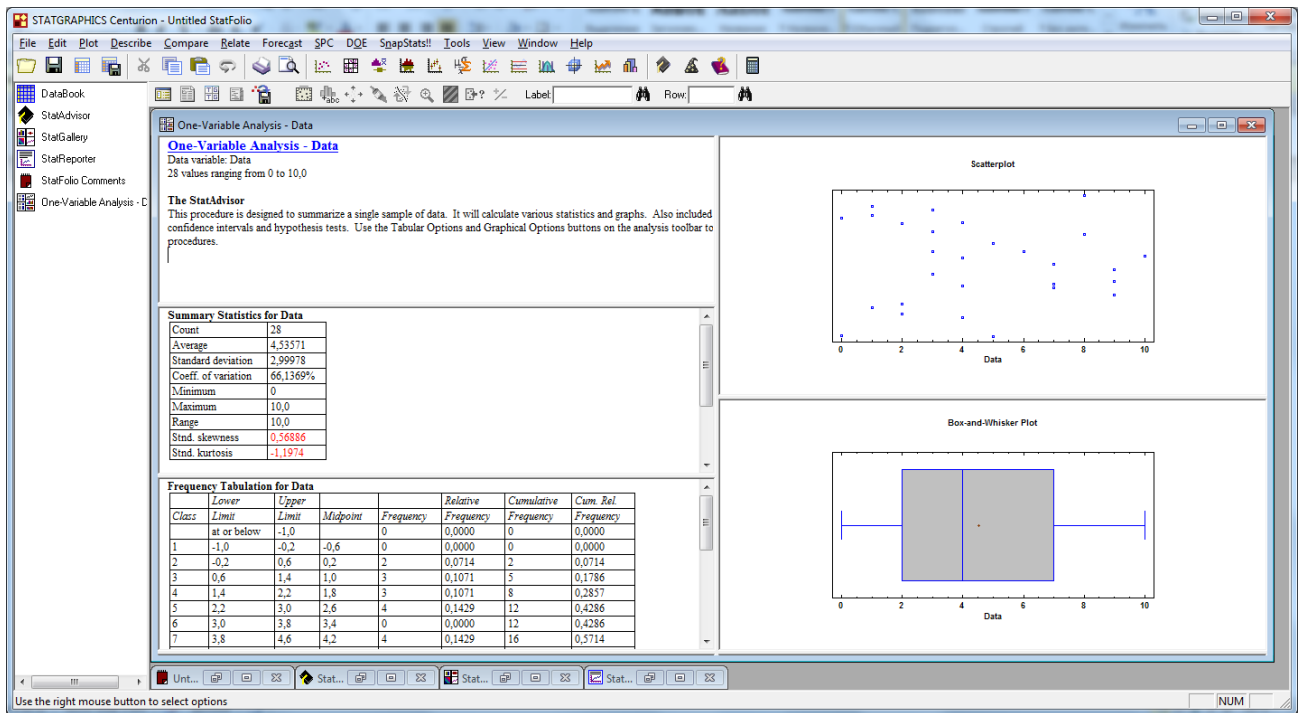
ню выберите **Describe**, в раскрывшемся меню выберите **Numeric Data**, а затем **One-Variable Analysis**. Раскроется диалоговое окно, в левой части которого надо выделить столбец с данными, которые мы хотим проанализировать,



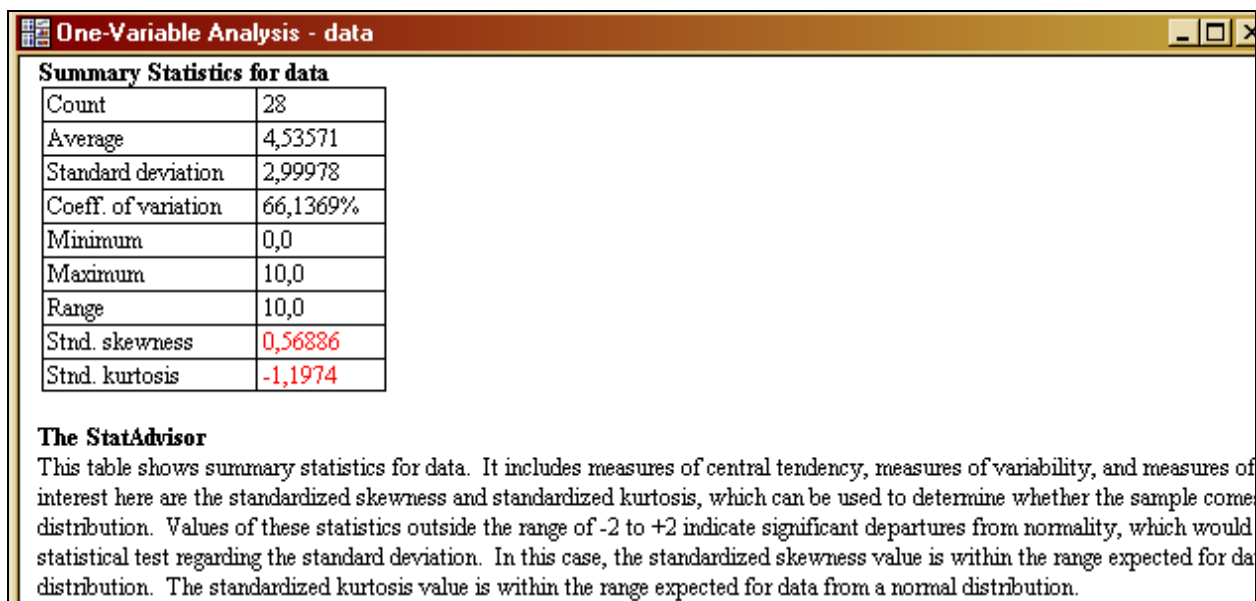
Затем нажмите на кнопку , название выбранного нами столбца перейдет в правую часть экрана. Нажмите на кнопку ОК. Перед вами раскроется следующее окно,



Нажмите ОК. раскроется окно, разделенное на несколько частей

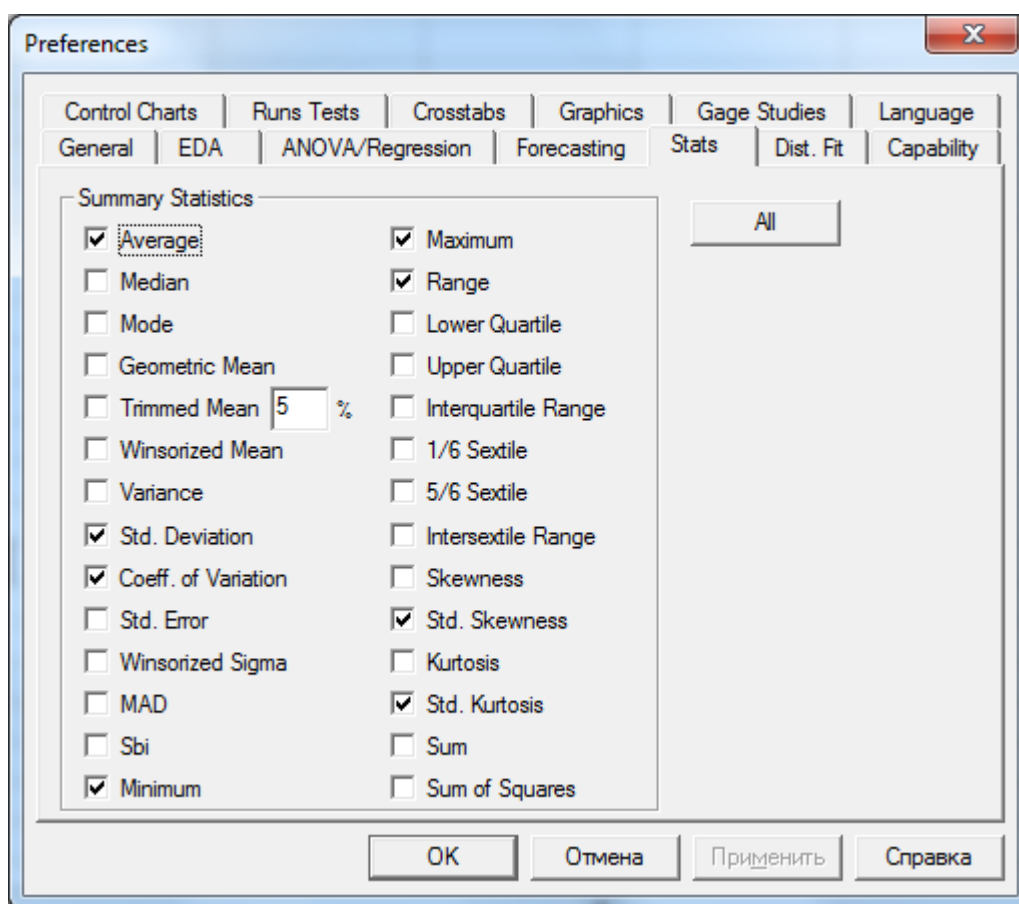


Щелкните дважды по окну *Summary Statistics*. Раскроется следующее окно:

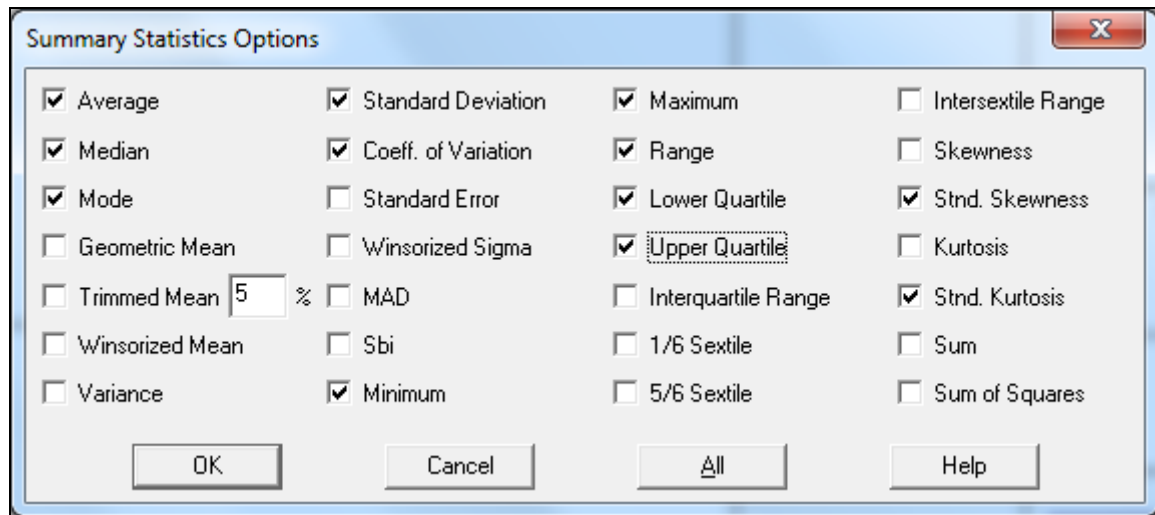


Среди вычисленных по умолчанию числовых характеристик нашей выборки есть те, которые нас не интересуют, и нет некоторых из нужных нам характеристик. Чтобы выбрать те, что мы желаем получить, в раскрывшемся окне *Summary Statistics* щелкните **правой** кнопкой, в появившемся контекстном меню выберите *Pane Options*, щелкните левой кнопкой. Раскроется диалоговое окно *Summary Statistics Options*, в этом окне выберите *Average* (среднее), *Median* (медиана), *Mode* (мода), *Variance* (дисперсия), *Standard Deviation* (стандартное отклонение), *Minimum* (минимальное значение выборки), *Maximum* (макси-

мальное значение), *Range* (размах выборки), *Std.Skewness* (стандартизованная асимметрия), *Std.Kurtosis* (стандартизованный эксцесс), *Sum* (сумма), *Coeff.of Variation* (коэффициент вариации), Lower Quartile (нижний квартиль), Upper Quartile (верхний квартиль). Все остальные числовые характеристики можно убрать. Если вам постоянно нужны другие характеристики, можно изменить настройки. В меню выберите **Edit, Preferences**, затем выберите вкладку **Stats**



здесь нужно выбрать нужные характеристики.



Нажмите кнопку ОК.

Summary Statistics for Col_1

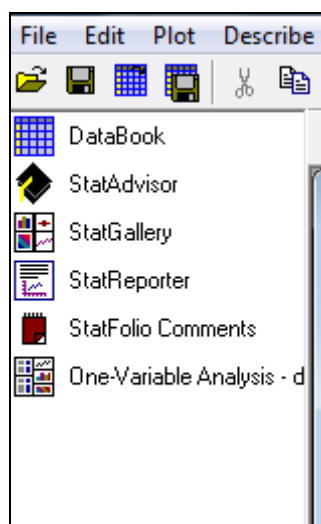
Count	28
Average	4,53571
Median	4,0
Mode	
Standard deviation	2,99978
Coeff. of variation	66,1369%
Minimum	0,0
Maximum	10,0
Range	10,0
Lower quartile	2,0
Upper quartile	7,0
Skewness	0,263331
Kurtosis	-1,10858

The StatAdvisor


This table shows summary statistics for Col_1. It includes measures of central tendency, measures of variability, and measures of shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate any statistical test regarding the standard deviation. In this case, the standardized skewness value is within the range expected for data from a normal distribution. The standardized kurtosis value is within the range expected for data from a normal distribution.

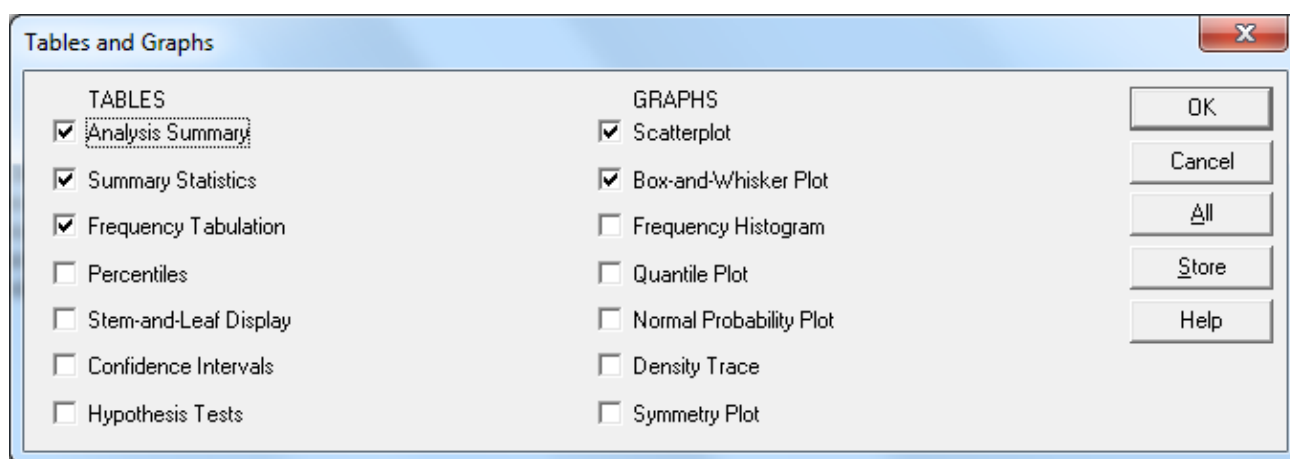
Обсудим полученные результаты Count. – количество. Всего 28 элементов в выборке. Average – выборочное среднее. Оно равно 4,5. Медиана равна 4,0. Медиана не сильно отличается от среднего, это говорит о том, что выборка достаточно однородная, нет резких выбросов. Обратите внимание на моду. Она отсутствует. Это происходит оттого, что одинаково часто встречаются несколько значений. Тогда пакет оставляет пустое место вместо моды. Range – это разность между максимальным и минимальным значением.

В левой части окна есть панель перехода

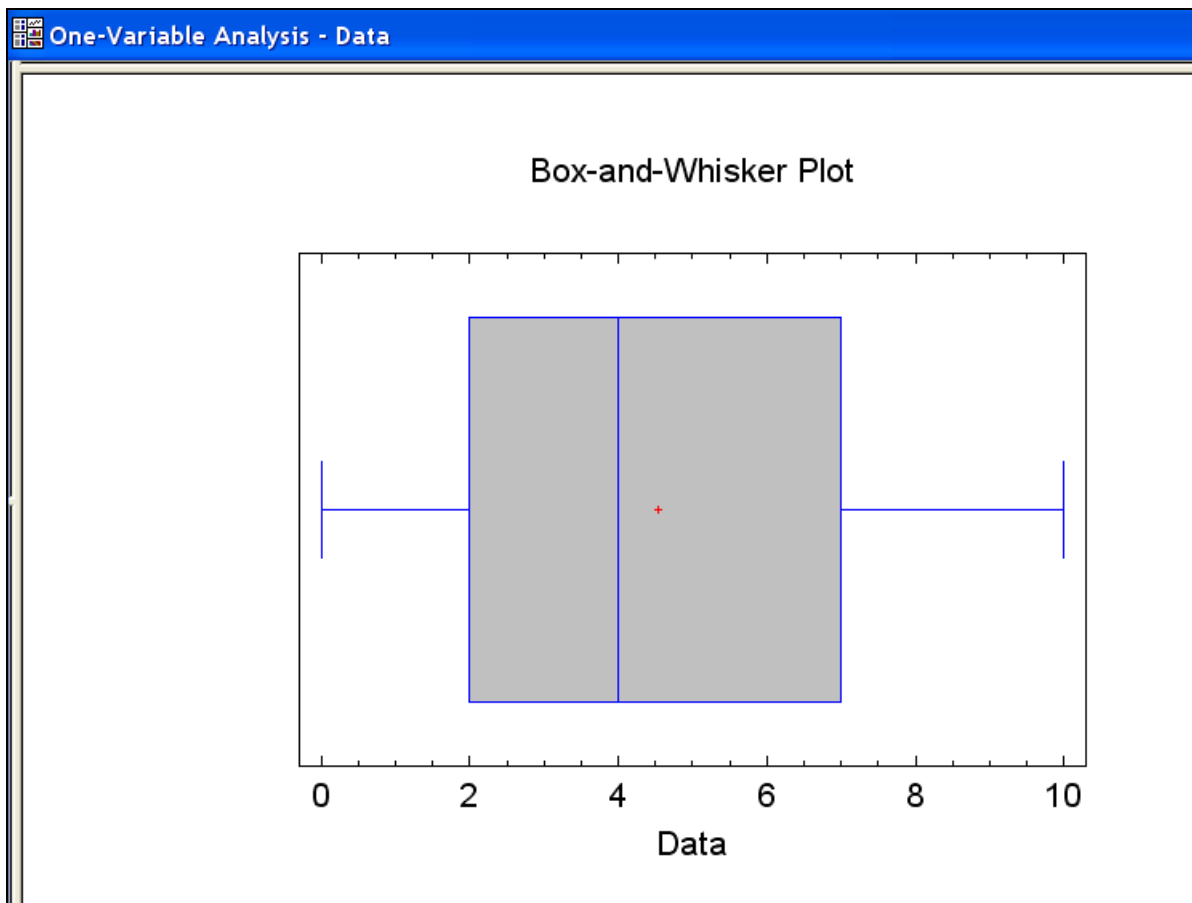


Сейчас мы хотим снова войти в таблицу с данными. Щелкните по DataBook, увеличьте первый элемент выборки в четыре раза. Щелкнув по ***One-Variable Analysis*** – Data на панели перехода, перейдете к окну, в котором находится анализ данных. Посмотрите, как изменится среднее, мода и медиана. Щелкните дважды по окну с суммарными статистиками, чтобы уменьшить его.

Посмотрим сейчас еще на один вариант графического представления данных. В первоначальном окне анализа есть график Box-and-Whisker Plot. (Если его нет, нажмите кнопку  ***Tables and graphs*** и выберите этот анализ в раскрывшемся окне).

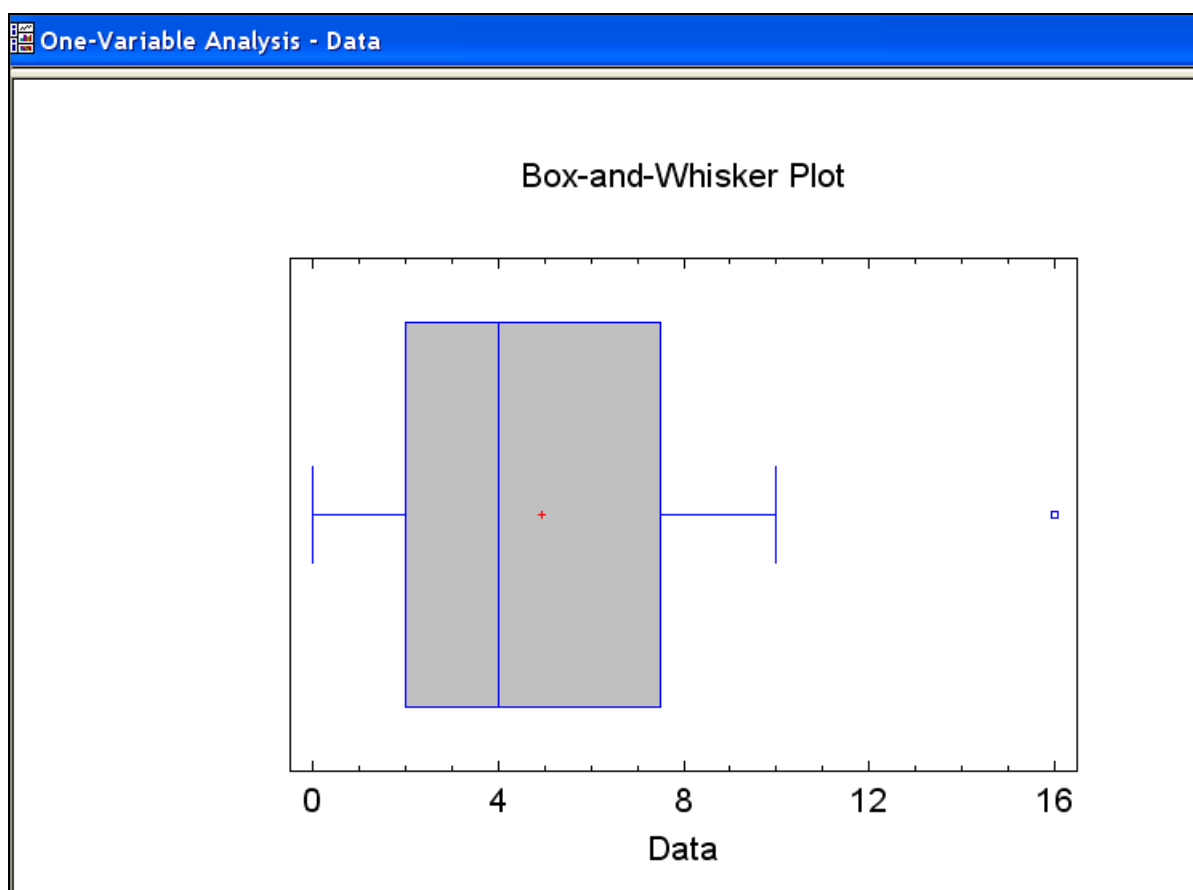


Щелкните дважды, раскроется окно



(Box-and-Whisker Plot в переводе встречается как ящиковая диаграмма, коробчатая диаграмма и даже ящик с усами). Этот прямоугольник представляет пространство между первым и третьим квартилями, т.е. от 25% до 75% вариационного ряда попадают в него. Линия внутри прямоугольника соответствует медиане. Плюс в коробке обозначает выборочное среднее. В нашем случае выборочное среднее не лежит на медиане, о чем это говорит?

Могут быть выбросы, они обозначаются отдельными точками. Увеличим снова первое значение в 4 раза, получится график



Появится точка справа, если по ней щелкнуть, то в поле Row появится цифра 1




Это – номер ряда, в котором был выброс. Обратите внимание, выборочное среднее еще дальше отошло от медианы, коробка стала еще менее симметричной. Точка становится выбросом, если она отстоит на расстояние более, чем в полтора раза превышающее ширину коробки.

Снова вернемся к таблице с данными, увеличим значение в первом ряду в 10 раз. Щелкните по графику, обратите внимание, коробка стала еще несимметричнее, а внутри выброса появился красный крестик. Это – обозначение так называемого «далекого» выброса (или экстремального выброса). Так обозначаются точки, если они отстоят более, чем на 3 расстояния, равного ширине коробки.

Графическое представление Box-and-Whisker Plot используется для первоначального анализа данных, например, чтобы проверить, не было ли при вводе

данных опечаток: все экстремальные выбросы надо проверить. В следующих работах мы вернемся к рассмотрению выбросов, а пока пойдем дальше.

Если данных много, обычно их **группируют**, т.е. строят интервальный ряд частот. Найдите окно ***Frequency Tabulation for data*** (частотная группировка).

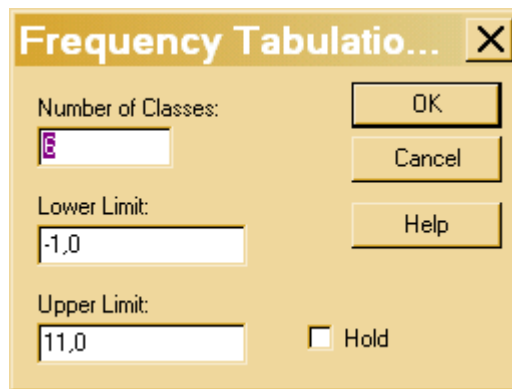
Если его на экране нет, щелкните по кнопке  ***Tables and graphs*** и выберите ***Frequency Tabulation***. Для того чтобы увеличить его, щелкните по нему дважды. Так можно увеличить любое окно, в том числе и графики. Для того чтобы вернуть все в исходное состояние, нужно снова дважды щелкнуть по окну. Пусть сейчас окно раскрыто полностью.

Frequency Tabulation for Col_1

	<i>Lower</i>	<i>Upper</i>			<i>Relative</i>	<i>Cumulativ</i>	<i>Cum. Rel.</i>
<i>Class</i>	<i>Limit</i>	<i>Limit</i>	<i>Midpoi</i>	<i>Frequenc</i>	<i>Frequency</i>	<i>Frequenc</i>	<i>Frequency</i>
	at or below	-1,0		0	0,0000	0	0,0000
1	-1,0	1,0	0,0	5	0,1786	5	0,1786
2	1,0	3,0	2,0	7	0,2500	12	0,4286
3	3,0	5,0	4,0	6	0,2143	18	0,6429
4	5,0	7,0	6,0	4	0,1429	22	0,7857
5	7,0	9,0	8,0	5	0,1786	27	0,9643
6	9,0	11,0	10,0	1	0,0357	28	1,0000
	above	11,0		0	0,0000	28	1,0000

Mean = 4,53571 Standard deviation = 2,99978

Здесь пакет разбивает данные на 6 групп. В таблице приводятся данные: номер группы (***Class***), нижняя граница (***Lower Limit***), верхняя граница (***Upper Limit***), средняя точка (***Midpoint***), частота (***Frequency***) (сколько элементов выборки попало в данный интервал), относительная частота (***Relative Frequency***) (частота, деленная на количество элементов в выборке). Она же называется частотность. ***Cumulative Frequency*** — накопленная частота, ***Cum.Rel. Frequency*** — накопленная частотность. Можно изменить количество классов. Щелкните правой кнопкой по окну, выберите **Pane Options**, раскроется окно

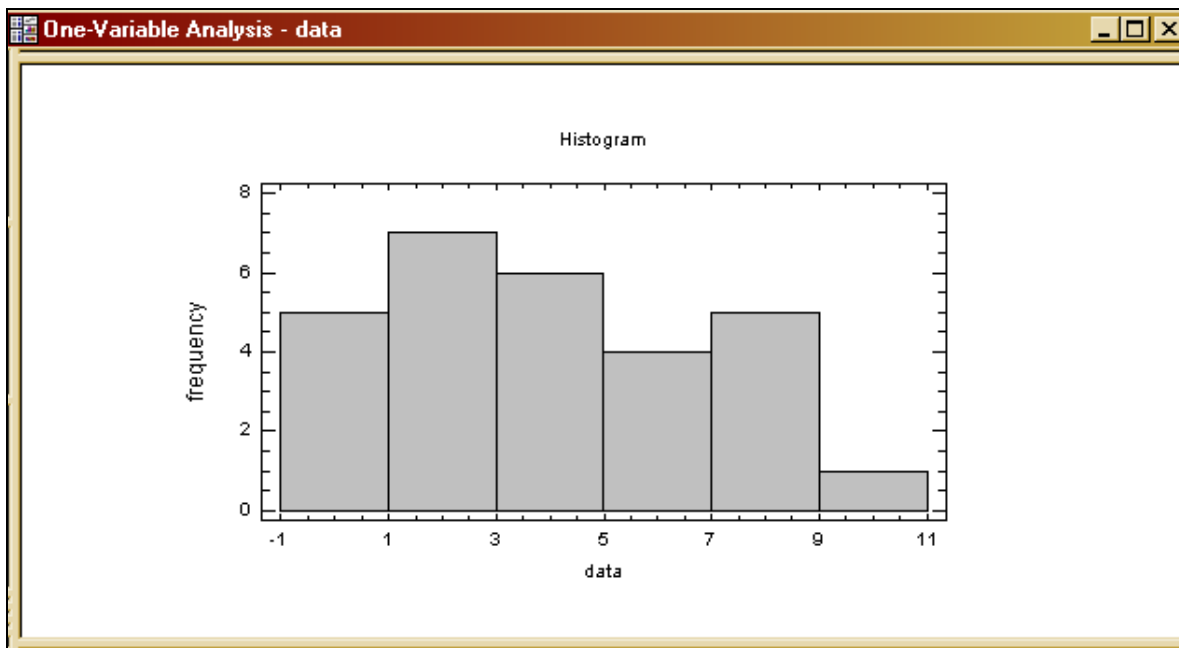


В этом окне можно менять нижний и верхний пределы, а также количество классов. Попробуйте сделать это. Затем верните исходные данные. Дело в том, что количество классов зависит от объема выборки, увеличение или уменьшение их количества не принесут, как правило, улучшения результатов.

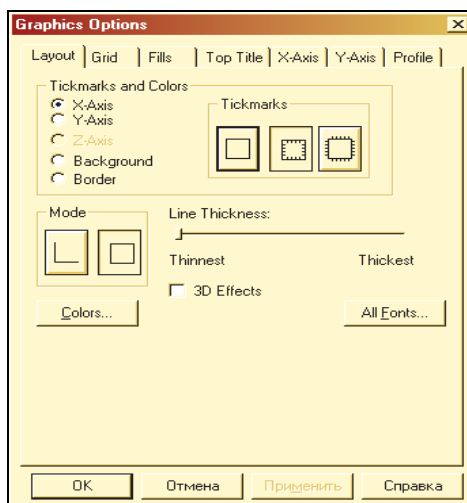
Посмотрим теперь на окно с гистограммой. Для этого щелкните по кнопке



Tables and graphs и выберите **Frequency Histogram**. Раскроется следующее окно:

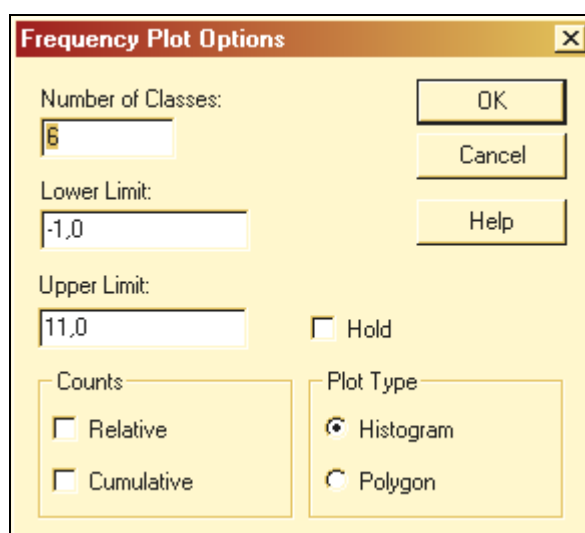


С этой гистограммой можно производить некоторые преобразования. Например, изменить цвет ряда данных или фона. Для этого надо щелкнуть по окну с гистограммой правой кнопкой, выбрать **Graphics Options**, раскроется окно, в котором в первой вкладке **Layout** можно выбрать цвет заполнения фона.



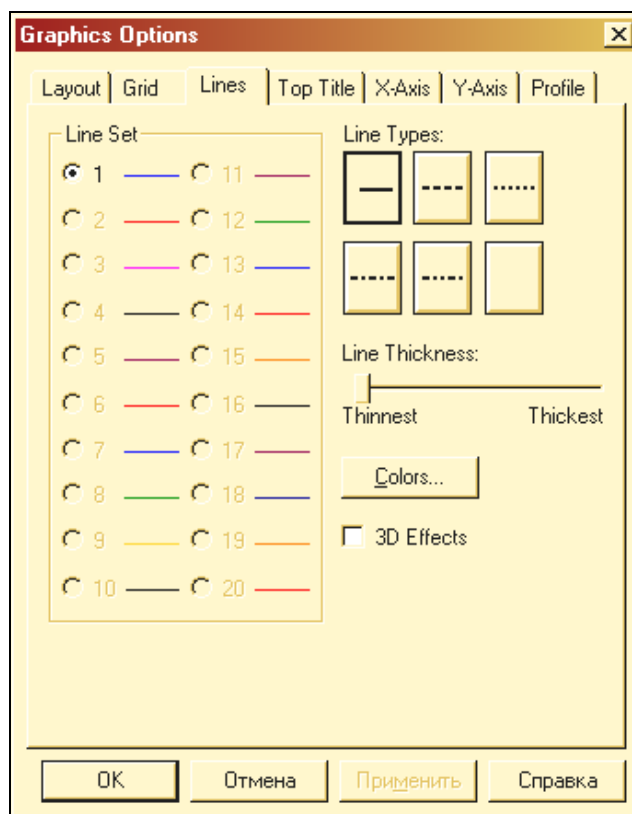
Для этого надо выбрать **Background**, затем надо нажать на кнопку **Colors**, в раскрывшемся окне выбрать нужный цвет, нажать кнопку ОК. Во вкладке **Fills** можно поработать с видом гистограммы. Попробуйте разные варианты. Во вкладке **Top Title** можно переименовать гистограмму. Также в соответствующих вкладках можно поработать с осями координат. Во вкладке **Profile** можно настраивать общий вид гистограммы.

Превратим теперь гистограмму в **полигон**. Для этого щелкните по окну с гистограммой правой кнопкой, в раскрывшемся контекстном меню выберите **Pane Options**, перед вами раскроется диалоговое окно:

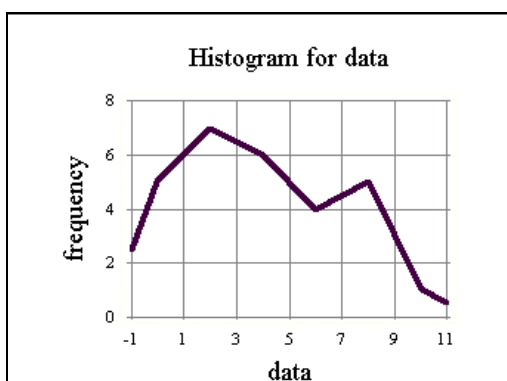


В этом окне выберите **Polygon** и щелкните ОК. На экране появится полигон, у которого вы можете изменить толщину и цвет линии. Для этого надо увеличить размер экрана, выделить линию, щелкнув по ней левой кнопкой, а затем щелк-

нуть по линии **правой** кнопкой мыши и выбрать в контекстном меню **Graphics Options**. Появится следующее диалоговое окно:



В этом диалоговом окне нужно выбрать нужные параметры — тип линии, ее цвет и толщину и нажать на кнопку ОК. Должно получиться примерно следующее:

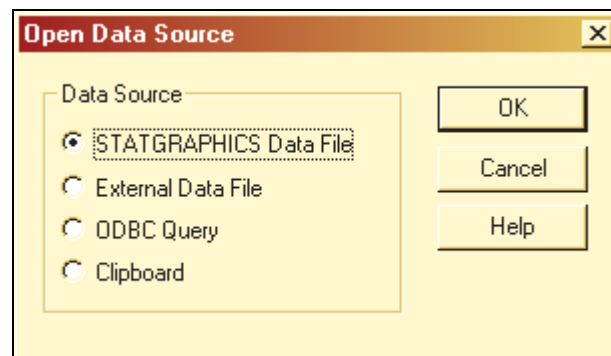


Кстати, в том же окне, где полигон и гистограмма, можно выбрать **кумулятивную гистограмму** (эмпирическую функцию распределения), если выбрать **Cumulative Histogram**, или кумуляту, если выбрать **Cumulative polygon**. Если же выбрать **Relative**, то на графике по оси Y будут относительные частоты, т.е. ча-

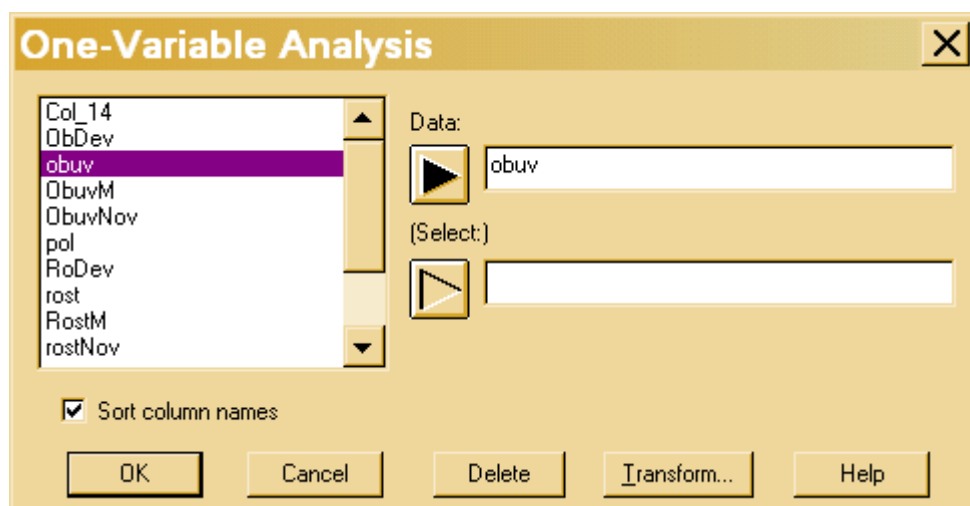
стости. Попробуйте самостоятельно сделать это. *Результаты покажите преподавателю.*

Для того чтобы сохранить весь проведенный анализ, не прилагая никаких новых усилий по заданию табличных и графических опций, нужно сохранить анализ в виде файла *StatFolio*. В строке меню выберите **File**, потом в раскрывшемся меню выберите *Save As, Save StatFolio As*, задайте имя, выберите папку, нажмите кнопку **Сохранить**. Закройте теперь анализ, выбрав в меню **File, Close StatFolio**. Для того чтобы показать анализ на другом компьютере, нужно иметь оба файла – и StatFolio, и файл с данными.

Проведем теперь анализ данных, полученных в результате анонимного опроса 40 студентов-математиков. Имеются следующие данные: рост, размер обуви, вес, пол. Данные находятся в файле Rost_Razmer.sf. Для того чтобы открыть файл, выберите в строке меню **File**, затем в раскрывшемся меню выберите **Open Data Source** (обратите внимание, не *Open StatFolio*) и перед вами раскроется диалоговое окно:

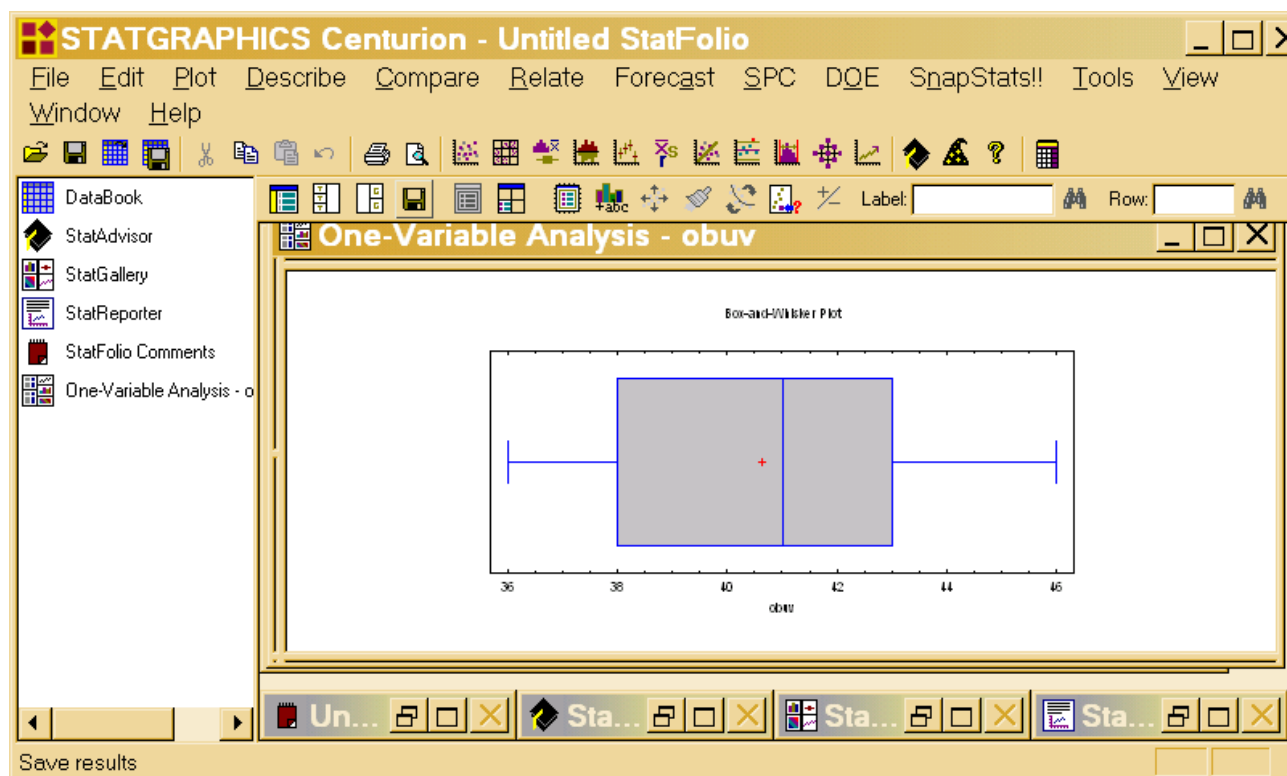


в нем надо выбрать *STATGRAPHICS Data File*, нажмите OK. Выберите нужный файл. Интересно проанализировать данные по размеру обуви. В строке меню выберите **Describe, Numeric Data, One-Variable Analyses**, раскроется окно, в нем выберите Obuv.



Нажмите кнопку ОК.

По данным ящичковой диаграммы (Box-and-Whisker Plot) видно, что резких выбросов нет.




Посмотрите выборочные характеристики, выведите на экран моду, медиану, выборочное среднее, верхний и нижний квартили, коэффициент вариации. Проанализируем полученные данные.

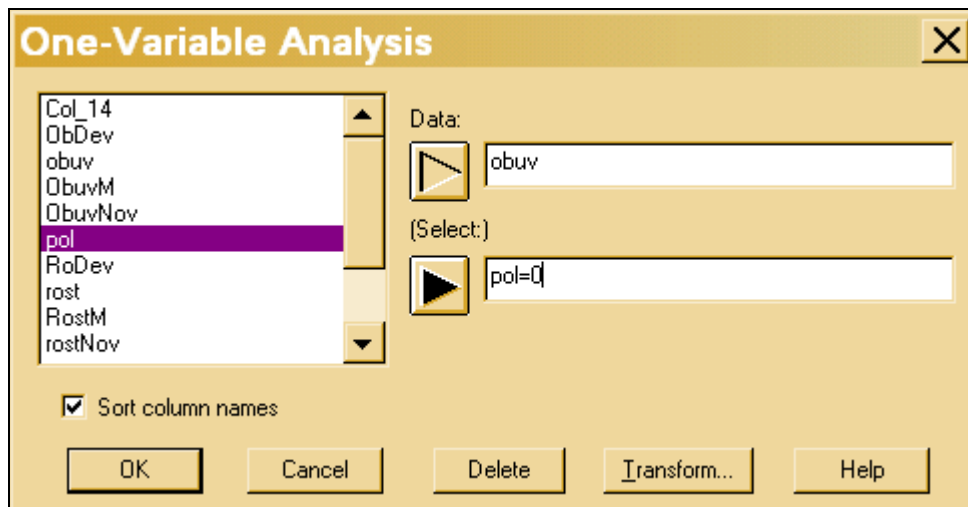
Summary Statistics for obuv

Count	40
Average	40,625
Median	41,0
Mode	43,0
Standard deviation	2,82559
Coeff. of variation	6,95531%
Minimum	36,0

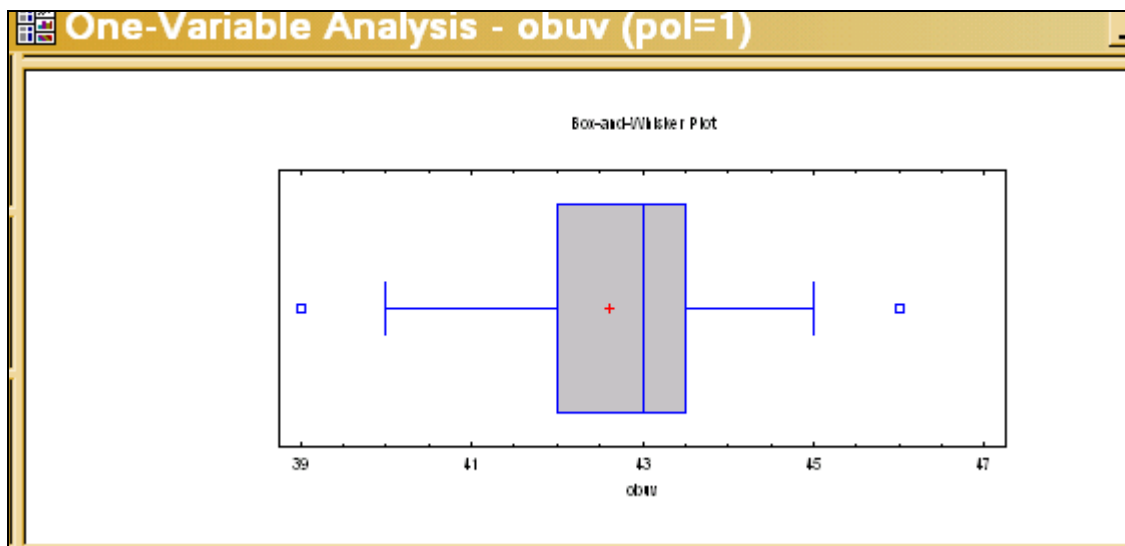
Maximum	46,0
Range	10,0
Lower quartile	38,0
Upper quartile	43,0
Std. skewness	-0,199539
Std. kurtosis	-1,66471

Выборочное среднее 40, 625 почти не отличается от медианы (41). Это говорит о том, что данные достаточно однородные. Мода равна 43, это – наиболее часто встречающийся размер обуви. Нижний квартиль (Lower quartile) равен 38, верхний (Upper quartile) – 43. Следовательно, между 38 и 43 размером находится половина всех размеров обуви. Коэффициент стандартизованной асимметрии (Std.Skewness) близок к нулю, это говорит о возможности нормального распределения (хотя коэффициент стандартизованного эксцесса (Std.Kurtosis) от нуля далек).

Проанализируем теперь данные для девочек. Щелкните по кнопке  **Input Dialog**, раскроется окно, в котором в поле **Select** надо выбрать **pol** и ввести с клавиатуры =0.

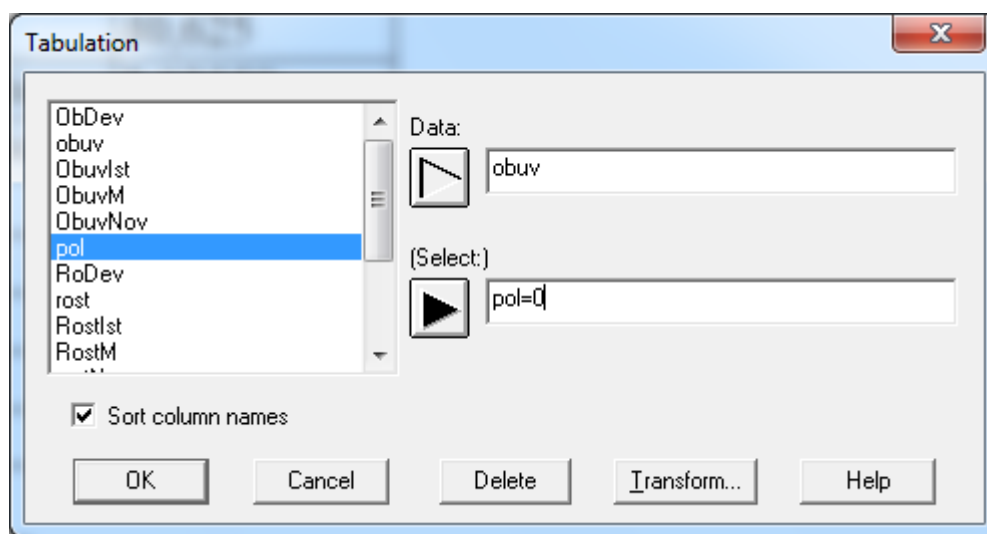


Нажмите кнопку ОК. В раскрывшемся окне все характеристики пересчитаются для девочек. Проанализируйте самостоятельно полученный результат. Теперь выведем анализ для мальчиков (pol=1). Видно, что у мальчиков есть выбросы, это наглядно показывает ящичковая диаграмма

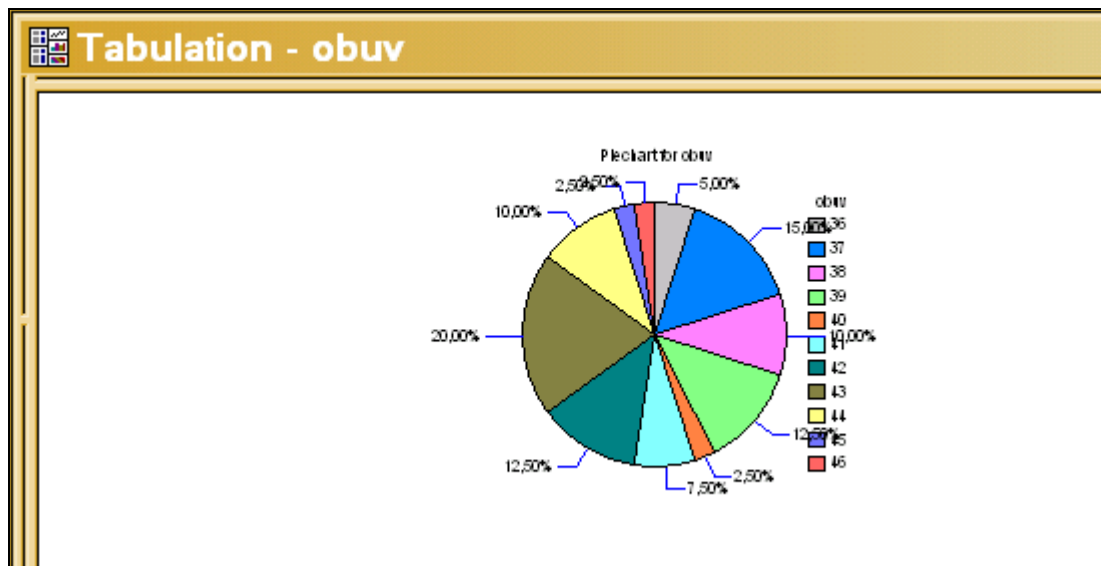


Как вы думаете, у кого разброс в размерах больше: у девочек или у мальчиков? На этот вопрос ответ дает **коэффициент вариации** (Coeff. of variation). У кого он больше? Покажите преподавателю самостоятельно сделанный анализ данных размеров обуви у девочек и у мальчиков.

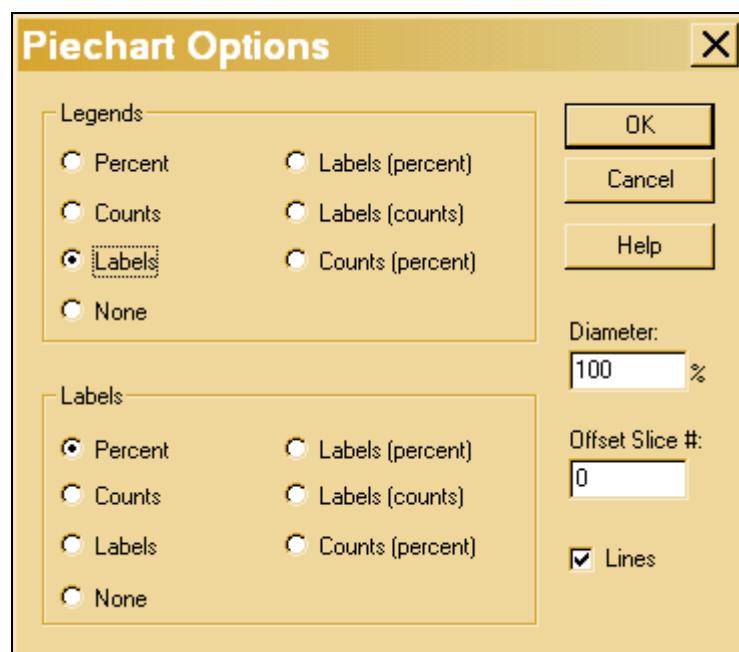
Можно также представить наглядное представление о распределении размеров обуви. В строке меню выберите **Describe, Categorical Data, Tabulation**.



Выберите **obuv**, в поле **Select** выберите **pol**, с клавиатуры введите **=0**, нажмите ОК.



На представленной диаграмме хорошо видна доля каждого размера обуви (в процентах), а также сразу видна мода – самый большой «кусок». Щелкните по этому окну правой кнопкой, выберите **Pane Options**. Раскроется окно



В этом окне можно выбрать различные варианты настройки диаграммы, например, не в процентах, а по количеству (выберите в Labels Counts). Сейчас в поле Offset Slice поставьте 1. Что получилось? Постройте самостоятельно круговые диаграммы для девочек; для мальчиков.

В пакете *Statgraphics* имеется некоторое количество встроенных примеров. Далее будем работать с ними и для того, чтобы случайно не испортить какой-нибудь из них, скопируйте эти примеры к себе в папку. Они находятся в папке

Data, которая, скорее всего, находится в папке *Program Files, Statgraphics Centurion*.

Сейчас откройте один из файлов.

В своей папке **Data** выберите *93Cars.sf*. В этом файле представлены характеристики подержанных автомашин различных марок. Проанализируйте переменную **horsepower** — это мощность автомобиля в лошадиных силах. От вас требуется:

- 1) просмотреть выборку на экране;
- 2) произвести интервальную группировку данных (построить интервальные вариационные ряды частот, частостей, накопленных частот и накопленных частостей);
- 3) построить гистограмму, полигон и кумуляту частостей;
- 4) вычислить выборочные числовые характеристики: среднее, моду, медиану, дисперсию, стандартное отклонение, стандартизованную асимметрию, стандартизованный эксцесс и коэффициент вариации.
- 5) Построить круговую диаграмму распределения количества машин по количеству пассажиров (passengers).
- 6) результаты своего анализа покажите преподавателю.

В пакете *Statgraphics* есть возможность получить характеристики нескольких величин одновременно. Найдём и сравним выборочные характеристики мощностей (**horsepower**) и максимальных цен (**Max price**). В строке меню выберите **Describe**, затем **Numeric Data, Multiple Variable Analyses**. Откроется диалоговое окно, в нём надо выбрать в левой части окна оба столбца данных (**horsepower, Max price**), переведите их в правую часть. В открывшемся диалоговом окне выберите **Summary Statistics**. С помощью **Pane Options** (из контекстного меню) выведите на экран все нужные выборочные характеристики. Ответьте на вопрос: где больше разброс — в ценах на автомобили или в мощностях? Как вы думаете, почему? Результаты покажите преподавателю.

Можно работать с файлом, сохранённым в каком-либо другом формате. Имеется в виду Excel files (*.xls), Text files (*.txt; *.csv; *.dat), XML (*.xml).

ЗАДАНИЕ

Проанализируйте данные о росте студентов из файла `Rost_razmer.sf`. От вас требуется

- 1) Просмотреть выборку *Rost* на экране;
- 2) произвести интервальную группировку данных (построить интервальные вариационные ряды частот, частостей, накопленных частот и накопленных частостей); построить гистограмму, полигон и кумуляту частостей; можно, ли глядя на гистограмму, предположить, что рост имеет нормальное распределение? Постройте гистограммы для размера обуви и веса, что можно сказать об их нормальности? Какова может быть причина отсутствия нормальности в наших данных? Как исправить ситуацию, учитывая пол?
- 3) вычислить выборочные числовые характеристики: среднее, моду, медиану, размах, дисперсию, стандартное отклонение, стандартную асимметрию, стандартный эксцесс. Каковы должны быть асимметрия и эксцесс, если рост имеет нормальное распределение?
- 4) вывести на экран все выборочные характеристики о росте, размере обуви и весе одновременно. Где больше разброс: в росте весе или размере? Почему?
- 5) Прodelать этот же анализ для девочек и мальчиков по отдельности.

ВОПРОСЫ

1. Как записывается выборочное среднее для не сгруппированных данных?
2. Как записывается несмещенная выборочная дисперсия для не сгруппированных данных?
3. Что такое (выборочная) мода (можно на примере).
4. Что такое (выборочная) медиана (можно на примере).
5. Что характеризуют асимметрия и эксцесс?
6. Для чего используется коэффициент вариации?
7. Что вы понимаете под репрезентативностью выборки?

8. Что такое гистограмма частостей, статистическим аналогом чего она является?
9. Что такое кумулята частостей, статистическим аналогом чего она является?
10. Как записывается выборочное среднее для сгруппированных данных?