

УДК 519.222

**ПЕРВИЧНЫЕ НАВЫКИ РАБОТЫ В ПРОГРАММЕ STATSOFT
STATISTICA – ПРОЦЕДУРА DESCRIPTIVE STATISTICS****Г. А. Качалова***Старший преподаватель,
Московский государственный гуманитарный
университет, им. М. А. Шолохова,
г. Москва, Россия***PRIMARY SKILLS OF WORK WITH THE PROGRAM STATSOFT
STATISTICA – PROCEDURE DESCRIPTIVE STATISTICS****G. A. Kachalova***Senior teacher,
Moscow State Humanitarian
University named after M. A. Sholokhov,
Moscow, Russia*

Summary. This article describes the procedure Descriptive statistics of the statistical programme StatSoft STATISTICA for primary processing of experimental data. Shows the graphics capabilities of the Basic Statistics and Tables. The analysis of the data at a specific example.

Keywords: the program StatSoft STATISTICA; module Basic Statistics and Tables; the procedure Descriptive statistics.

Статистика – отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных; изучение количественной стороны массовых общественных явлений в числовой форме.

Статистический анализ данных используется практически во всех сферах человеческой деятельности, таких как экономика, бизнес, маркетинг, промышленность, медицина, научные исследования и др. Одной из наиболее популярных статистических программ для анализа и визуализации данных, поиска закономерностей, прогнозирования экономических ситуаций является система *STATISTICA*, разработанная компанией *StatSoft*. В данной системе существует возможность проводить классические и новейшие методы анализа данных: дисперсионный, корреляционный, факторный и кластерный анализ; линейную и нелинейную регрессии; нейронные сети и др. Визуализация исходных, промежуточных, выходных данных может быть осуществлена выбором из большого числа различных графиков, пиктографиков и диаграмм. Применение программы *STATISTICA* позволяет эффективно решать сложные проблемы и осуществлять аналитическую поддержку принятия решений.

Система *STATISTICA* представляет собой интегрированную систему статисти-

ческого анализа и обработки данных. Она состоит из 5 компонентов:

1) электронных таблиц для ввода и задания исходных данных, а также специальных таблиц для вывода результатов статистического анализа;

2) графической системы визуализации данных и результатов статистического анализа;

3) набора статистических модулей, в которых собраны группы логически связанных между собой статистических процедур;

4) специального инструментария для подготовки отчётов;

5) встроенных языков программирования, позволяющих расширить стандартные возможности системы.

В любом конкретном модуле можно выполнить определённый способ статистической обработки, не обращаясь к процедурам других модулей. Переключаться между модулями можно как между обычными *Windows*-приложениями, выбирая их на панели переключателей модулей щелчком мыши.

Графические возможности доступны в любом статистическом модуле на любом шаге анализа.

В данной статье покажем на конкретном примере, как можно без труда рассчитать основные описательные статистики и визуализировать данные с помощью модуля *Basic Statistics and Tables* в системе *STATISTICA 10*.

Расчёт описательных статистик производится при помощи модуля *Basic Statistic and Tables* (основные статистики и таблицы). В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры.

В стартовой панели модуля приводится перечень статистических процедур этого модуля (рис. 1).

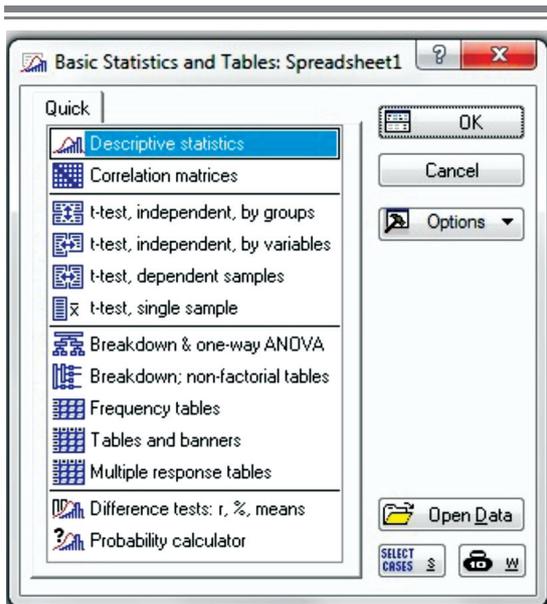


Рис. 1. Стартовое окно модуля с перечнем статистических процедур

Descriptive statistics – описательные статистики;

Correlation matrices – корреляционные матрицы (парные и частные корреляции);

t-test, independent, by groups – t-критерий для независимых выборок;

t-test, independent, by variables – t-критерий для независимых переменных;

t-test, dependent samples – t-критерий для зависимых выборок;

t-test, single sample – одновыборочный t-критерий;

Breakdown & one-way ANOVA – группировка и однофакторный дисперсионный анализ и др.

Воспользуемся процедурой *Descriptive statistics*.

Рассмотрим возможности этой процедуры на примере. Результаты испытаний представлены в таблице. Имеется выборка объёмом 23 измерения, содержащая данные по 6 переменным (рис. 2).

После выбора процедуры *Descriptive statistics* на экране появится одноименное диалоговое окно (рис. 3).

В этом окне при помощи кнопки *Variables* следует выбрать переменные для анализа (рис 3).

На первом этапе обработки данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака.

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6
1	117.0	22.0	96.0	8.0	7.5	32.8
2	126.0	27.2	113.0	12.0	7.5	30.1
3	120.0	31.3	72.0	4.0	10.1	32.5
4	110.5	19.5	92.0	16.0	7.0	29.7
5	118.0	23.7	96.0	15.0	6.5	26.6
6	124.0	22.9	103.0	15.0	6.1	26.3
7	120.0	21.9	122.0	14.0	7.9	33.3
8	124.5	25.3	86.0	14.0	7.5	32.4
9	128.5	26.7	89.0	8.0	11.5	37.7
10	114.0	18.8	93.0	15.0	7.0	29.7
11	113.5	22.6	86.0	12.0	7.6	32.9
12	123.0	26.7	139.0	15.0	7.1	32.9
13	121.5	25.0	89.0	15.0	7.3	30.9
14	115.5	21.0	47.0	13.0	12.7	38.0
15	120.5	20.3	105.0	13.0	7.5	31.0
16	123.0	22.4	131.0	17.0	6.6	26.4
17	118.5	21.9	86.0	12.0	12.0	40.1
18	121.0	23.0	132.0	18.0	6.9	26.6
19	122.0	22.3	138.0	12.0	7.3	31.9
20	127.5	27.9	123.0	13.0	6.7	26.6
21	115.5	22.7	105.0	8.0	9.1	34.1
22	113.0	21.1	108.0	12.0	7.6	28.8
23	119.0	26.7	104.0	13.0	7.8	32.4

Рис. 2. Окно файла данных

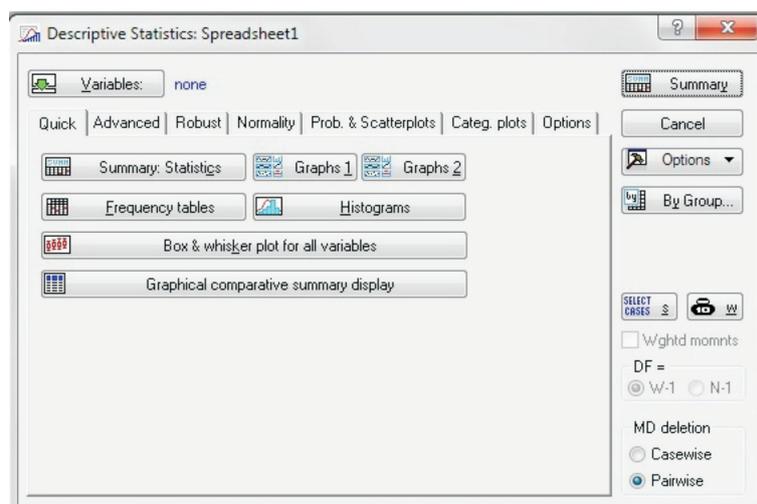


Рис. 3. Диалоговое окно *Descriptive statistics*

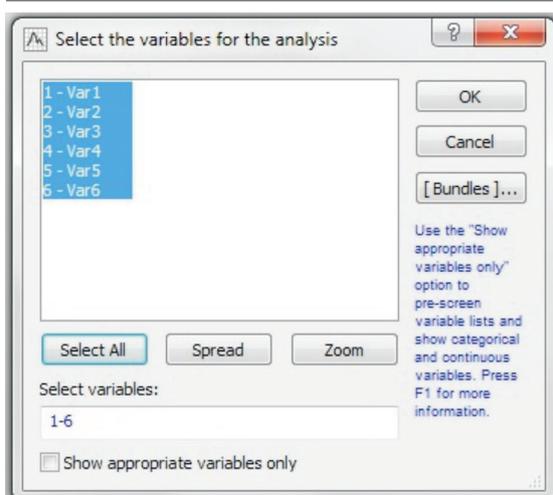


Рис. 4. Окно выбора переменных

В данном примере сгруппируем данные по 5 классам (интервалам) (рис. 5).

Число классов (интервалов) группировки данных устанавливается при помощи счётчика переключателя *Number of intervals* окна *Descriptive statistics*. Внизу от кнопок *Distribution* находятся две опции *Categorization* (группировка), позволяющие задать число интервалов группировки или установить величину интервала равную целому числу. Если заактивировать переключатель *Integer intervals (categories)*, то классы (интер-

валы) группировки будут представлять из себя целые числа.

Результаты группировки переменной *Var1* представлены на рис. 6.

Для построения гистограмм и таблиц частот используется группа кнопок *Distribution* окна *Descriptive statistics*. Представим распределение переменных на гистограммах. Для этого предназначена кнопка *Histograms* окна *Descriptive statistics*.

На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова-Смирнова, Лиллиефорса; вычислить статистику Шапиро-Уилкса. Для этого в группе опций *Distribution* необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

На рис. 7 в качестве примера приводится гистограмма распределения сгруппированной переменной *Var1*.

На гистограмме показана кривая плотности нормального распределения, а также критерий Колмогорова-Смирнова (*d*). Статистика Колмогорова-Смирнова оказалась равной 0,079. Чем меньше величина этой статистики, тем ближе распределение случайной величины к нормальному виду. Вероятность нулевой гипотезы (*p*) более 0,20.

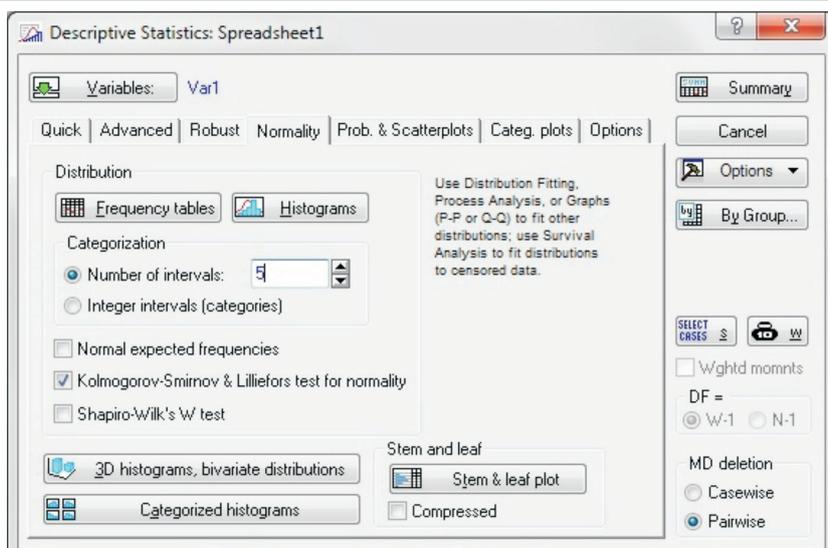


Рис. 5. Окно группировки переменной Var1 на классы (интервалы)

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
105,0000<x<=110,0000	0	0	0,00000	0,0000	0,00000	0,0000
110,0000<x<=115,0000	4	4	17,39130	17,3913	17,39130	17,3913
115,0000<x<=120,0000	8	12	34,78261	52,1739	34,78261	52,1739
120,0000<x<=125,0000	8	20	34,78261	86,9565	34,78261	86,9565
125,0000<x<=130,0000	3	23	13,04348	100,0000	13,04348	100,0000
Missing	0	23	0,00000		0,00000	100,0000

Рис. 6. Таблица группировки переменной Var1 на классы (интервалы)

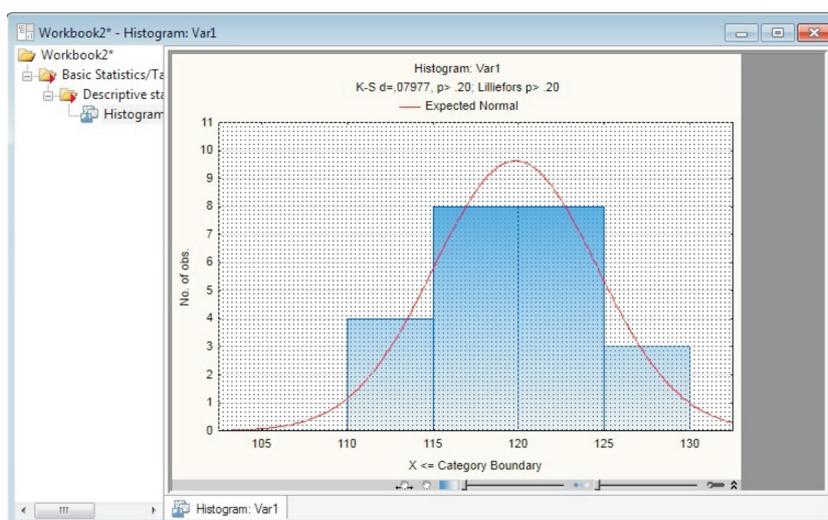


Рис. 7. Гистограмма распределения сгруппированной переменной Var1

О нормальности распределения можно судить по графику на нормальной вероятностной бумаге. Его легко построить при помощи опции *Normal probability plots* окна *Descriptive statistics* (рис. 3). Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию (рис. 8). Этот метод оценки является фактически «глазомерным». В сомни-

тельных случаях проверку на нормальность можно продолжить с использованием специальных статистических критериев (Колмогорова-Смирнова, Омега-квадрат (w^2)). Однако детальная проверка гипотезы о нормальности выборки требует довольно значительных объёмов выборки (по мнению некоторых авторов не менее 100 наблюдений).

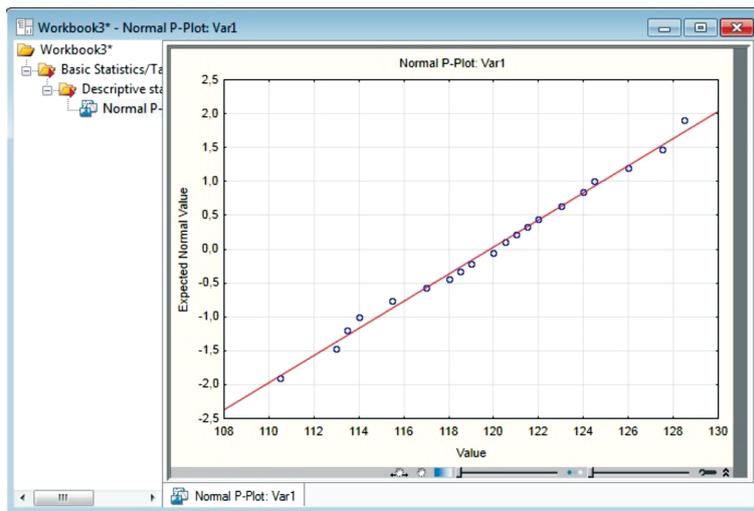


Рис. 8. График на нормальной вероятностной бумаге для переменной Var1

Чтобы выбрать статистики, подлежащие вычислению, удобнее всего

воспользоваться кнопкой *Advanced* (рис. 9).

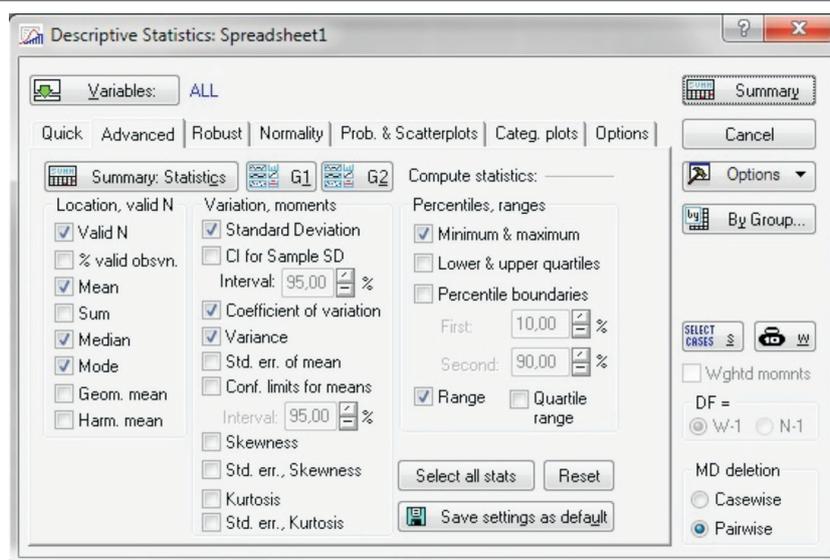


Рис. 9. Окно выбора статистик

Valid N – объём выборки;
Mean – средняя арифметическая;
Sum – сумма;
Median – медиана;
Mode – мода;
Geom. mean – геометрическая средняя;
Harm. mean – гармоническая средняя;
Standard Deviation – стандартное отклонение;
Variance – дисперсия;
Standard error of mean – стандартная ошибка среднего;
95% confidence limits of mean – 95%-й доверительный интервал для среднего;
Skewness – асимметрия;
Standard error of Skewness – стандартная ошибка асимметрии;

Kurtosis – эксцесс;
Standard error of Kurtosis – стандартная ошибка эксцесса;
Minimum, maximum – минимальное и максимальное значения;
Lower, upper quartiles – нижний и верхний квартили;
Range – размах;
Quartile range – интерквартильная широта.

Напротив статистик, подлежащих вычислению (рис. 9), следует поставить флажок.

После нажатия на кнопку ОК окна *Descriptive statistics* на экране появится таблица с результатами расчётов описательных статистик (рис. 10).

Variable	Valid N	Mean	Confidence -95.000%	Confidence 95.000%	Median	Mode	Frequency of Mode	Sum	Minimum	Maximum	Range	Variance	Std. Dev.	Confidence SD -95.000%	Confidence SD +95.000%	Coef. Var.
Var1	23	119.8261	117.7649	121.8872	120.0000	Multiple	2	2756.000	110.5000	128.5000	18.00000	22.7184	4.76638	3.68629	6.74610	3.97775
Var2	23	23.6043	22.2930	24.9157	22.7000	26.70000	3	542.900	18.8000	31.3000	12.50000	9.1959	3.03247	2.34530	4.29201	12.84709
Var3	23	102.3913	92.8513	111.9314	103.0000	86.00000	3	2355.000	47.0000	139.0000	92.00000	486.7036	22.06136	17.06214	31.22458	21.54613
Var4	23	12.7826	11.3786	14.1866	13.0000	Multiple	5	294.000	4.0000	18.0000	14.00000	10.5415	3.24677	2.51103	4.59532	25.39989
Var5	23	8.0348	7.2535	8.8161	7.5000	7.500000	4	184.800	6.1000	12.7000	6.60000	3.2642	1.80671	1.39730	2.55713	22.48607
Var6	23	31.4652	29.8304	33.1000	31.9000	26.60000	3	723.700	26.3000	40.1000	13.80000	14.2924	3.78053	2.92384	5.35077	12.01493

Рис. 10. Окно с результатами расчёта описательных статистик

Для визуализации описательных статистик можно построить статистические графики типа «коробок» (или «поле усов участка для всех переменных»). Это легко можно сделать при

помощи кнопки *Box & Whisker plot for all variable* окна *Descriptive statistics* (рис. 3). На графике можно отобразить 3 статистики, установив переключатель в одно из 4-х положений (рис. 11).

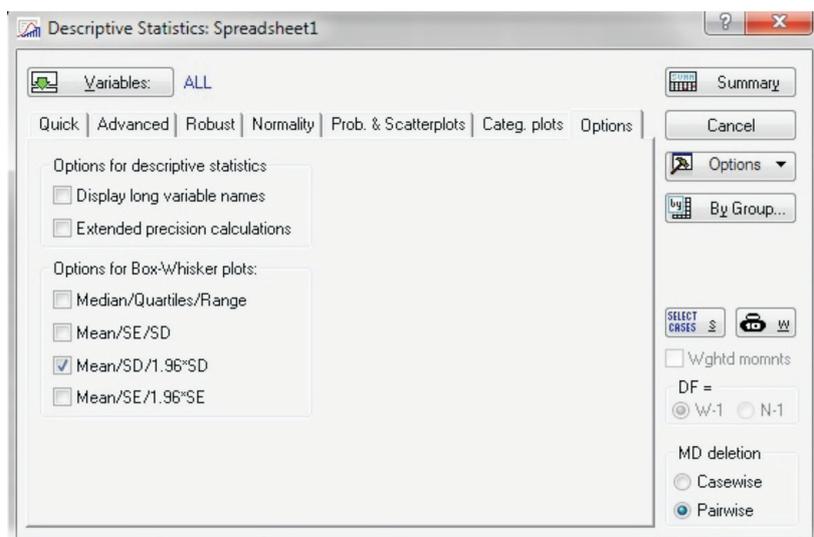


Рис. 11. Окно выбора статистик для графика коробок

Median/Quartiles/Range – медиана / квантили / размах;

Mean/SE/SD – среднее / ошибка среднего / стандартное отклонение

*Mean/SD/1.96*SD* – среднее / стандартное отклонение / интервал $1,96 \cdot$ стандартного отклонения;

*Mean/SE/1.96*SE* – среднее / ошибка среднего / интервал $1,96 \cdot$ ошибки среднего.

Визуализация описательных статистик переменных VAR1–VAR6 рассматриваемого примера при помощи графика коробок представлена на рис. 12.

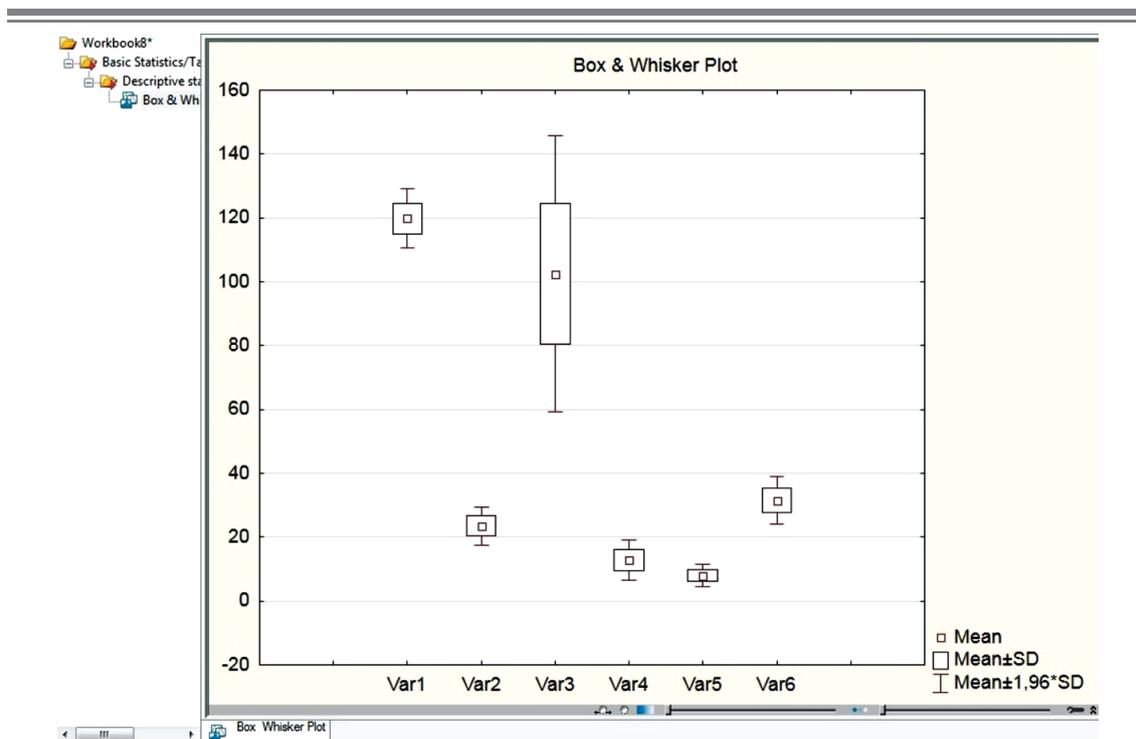


Рис. 12. Описательные статистики в графическом виде

Итак, на примере данной статьи мы показали, как без труда, используя программу *StatSoft STATISTICA*, можно произвести первичный анализ статистических данных, вычислить описательные статистики, используя процедуру *Descriptive statistics*. Навыки работы в программе *StatSoft STATISTICA* полезны при подготовке бакалавров любого направления к их будущей профессиональной деятельности.

Библиографический список

1. Боровиков В. П. STATISTICA: Искусство анализа данных на компьютере. – СПб. : Питер, 2003. – 688 с.
2. Вуколов Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использовани-

ем пакетов Statistica и Excel. – М. : Форум, 2010. – 464 с.

3. Качалова Г. А., Власов Д. А. Технологии Wolframalpha при изучении элементов прикладной математики студентами бакалавриата // Молодой учёный. – 2013. – № 6. – С. 683–691.

Библиографический список

1. Borovikov V. P. STATISTICA: Искусство анализа данных на компьютере. – SPB. : Piter, 2003. – 688 с.
2. Vukolov E. A. Osnovyi statisticheskogo analiza. Praktikum po statisticheskim metodam i issledovaniyu operatsiy s ispolzovaniem paketov Statistica i Excel. – М. : Forum, 2010. – 464 с.
3. Kachalova G. A., Vlasov D. A. Tehnologii Wolframalpha pri izuchenii elementov prikladnoy matematiki studentami bakalavriata // Molodoy uchyonyiy. – 2013. – № 6. – S. 683–691.

© Качалова Г. А., 2014.