# INTRODUCTION TO STATISTICA

STRUCTURE OF *STATISTICA*

STATISTICA consists of modules, each containing a group of related statistical procedures. Use the **Statistics** menu to select the various analyses available in your particular version of *STATISTICA*.
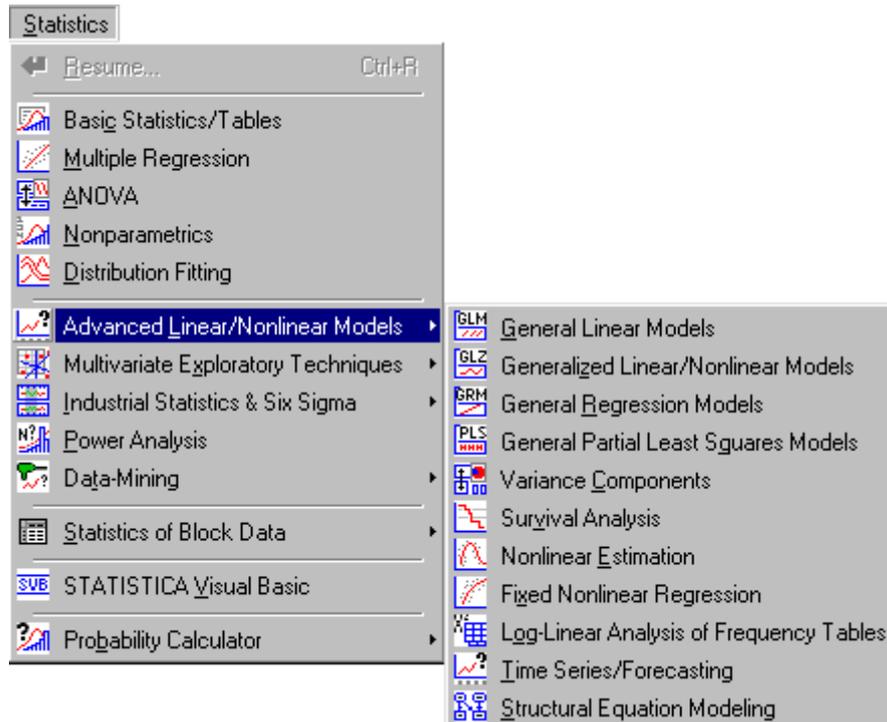
You can have several copies of *STATISTICA* open at the same time. Each of them can run similar or different types of analyses. Moreover, in one *STATISTICA* application, multiple analyses can be open simultaneously. They can be of the same or a different type (e.g., three **Multiple Regressions** and two **ANOVAs**), and each of them can be performed on the same or a different input data file (multiple input data files can be open simultaneously).

All "general-purpose" facilities (such as the data spreadsheet, general graphing procedures, and automation facilities) are available in every module and at every point of the analysis.

OUTPUT

There are three basic channels to which you can direct all output: workbooks, reports, and stand-alone windows. One can adjust the program to display output in the form of their choice by selecting **Options** from the **Tools** menu, and clicking on the **Output Manager** tab.

WORKBOOKS

Workbooks are the default way of managing output. Each output document is stored as a tab in the workbook. Documents can be organized into hierarchies of folders and documents using a tree view where individual documents and folders or entire branches of the tree can be flexibly managed.

Dependnt Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual
GENDER	0.319861	0.102311	-0.099669	1.2605	9	0.14005	11.0595
ADVERT	0.462205	0.213634	0.036702	2.6533	9	0.29481	9.7667
MEASURE01	0.447795	0.200520	0.020638	55.0429	9	6.11587	219.4571
MEASURE02	0.473719	0.224410	0.049902	91.6533	9	10.18370	316.7667
MEASURE03	0.277088	0.076778	-0.130947	27.9486	9	3.10540	336.0714
MEASURE04	0.411576	0.169395	-0.017491	58.6919	9	6.52132	287.7881
MEASURE05	0.672721	0.452553	0.329378	153.8319	9	17.09243	186.0881
MEASURE06	0.398296	0.158640	-0.030666	70.8676	9	7.87418	375.8524
MEASURE07	0.543868	0.295793	0.137346	90.2819	9	10.03132	214.9381
MEASURE08	0.511253	0.261379	0.095190	100.9133	9	11.21259	285.1667

Test of SS Whole Model vs. SS Residual (Adstudy.sta)

For example, selections of documents can be extracted (drag-copied or drag-moved) to the report window or to the application where they will be displayed in stand-alone windows. Even entire branches can be placed into other workbooks to build a specific folder organization.

REPORTS

Reports in *STATISTICA* offer a more traditional way of handling output, where each object is displayed sequentially in a word processor style document. The advantages of this format are the ability to insert notes and comments as well as its support for the more traditional way of scrolling through and reviewing the output.
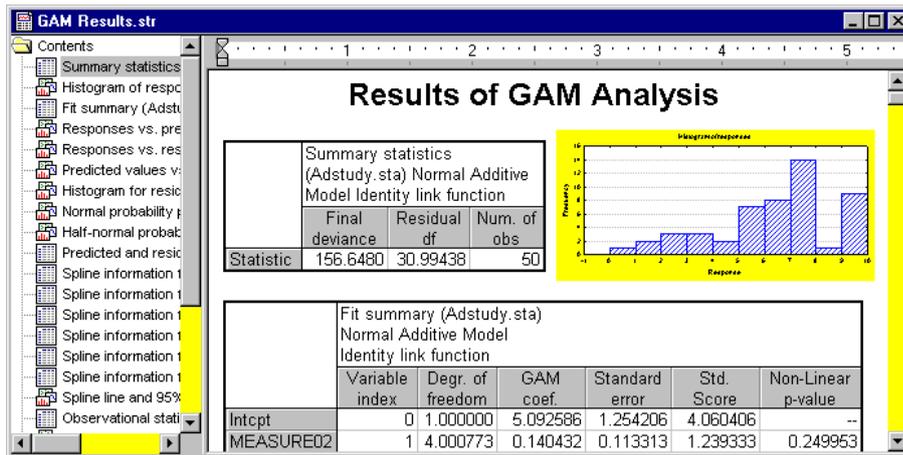
The report output includes and preserves a record of the supplementary information, which contains a detailed log of the options specified for the analyses.

STAND-ALONE WINDOWS

Finally, *STATISTICA* output documents can also be directed to a queue of stand-alone windows that can be easily custom arranged within the *STATISTICA* application workspace. This is useful for creating reference documents to compare to new output.
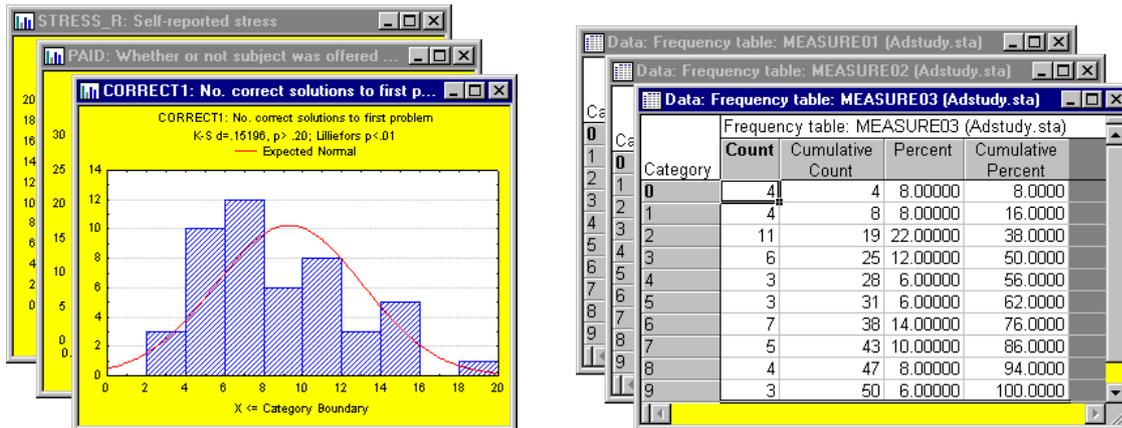
DATA FILES

The name of the data file is displayed in the title bar, along with the number of variables and cases in the file. Below, the data file title is *Adstudy.sta,* and it contains 25 variables and 50 cases.

Immediately below the title bar is the one-line file header, an optional short description of the data file. Double-click on the header to type in the field.

Additional information about the data can be entered in the **Info Box** in the upper-left corner of the data file. The **Info Box** can be accessed by double-clicking in the box. In the data file above, the **Info Box** contains the data *August 10th*.

STATISTICA data files are organized into variables and cases. The columns are the variables (equivalent to fields in data base programs), and the rows, or observations, are the cases (equivalent to records in data base programs).

The variable names are listed across the top of the data file, and the case names, optionally, are displayed down the left side of the data file. In the data file above, the first variable's name is *GENDER* and the first case's name is *R. Rafuse*.

VARIABLE AND CASE SPECIFICATIONS

Variables and cases can be modified in several ways: add variables, move cases, recalculate variables, etc. Most options for modifying variables and cases can be accessed from the

Vars [Vars ▾] and **Cases** [Cases ▾] toolbar button menus.

VARIABLES

Each variable has a set of properties or specifications associated with it. Click on a variable and select **Specs…** from the **Vars** toolbar button menu to display the **Variable** specification dialog.

Variable 1

Name: GENDER Type: Double OK

MD code: -9999 Length: 8 Cancel

<< >>

Display format

General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom

Decimal places: 0

1000; -1000
1,000; -1,000
1000; (1000)
1,000; (1,000)

All Specs...

Text Labels...

Values/Stats...

Long name (label or formula with [Functions]): □ Function guide

Gender of the subjects (May 15, 1996).

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

The **Name** box contains the variable name, which is displayed in the column header in the spreadsheet.

The **Type** box contains the variable data type.

The **MD code** box is used to specify a missing data code for blank cells or specific values that you intend to ignore in calculations.

The **Length** box, which is available only if you have selected **Text** as the data type for the variable, is used to specify the maximum number of characters allowed for the variable.

Under **Display format**, you can select a format for the variable. When certain display formats are chosen, a box to the right lists additional formats that are compatible with the selected display format.

The **Decimal places** box (which is only available when **Number**, **Scientific**, **Currency**, or **Percentage** is chosen as the display format) is used to specify the number of decimal places to be displayed in the spreadsheet.

The **Long name** box is used to provide a longer description of the variable, which can optionally be printed out with statistical results, or can be used to define a spreadsheet formula (with help from the **Function guide**).

VARIABLE SPECIFICATIONS EDITOR

Specifications for all variables can be reviewed or edited in the **Variable Specifications Editor**, a spreadsheet-like editor accessible by selecting **All Specs** from the **Vars** toolbar button menu.

|	Name	Type	MD code	Length	Long Name (label or formula)
1	GENDER	Double	-9999		Gender of the subjects (May 15, 1996).
2	ADVERT	Double	-9999		Ad shown to the subjects (May 15, 1996
3	MEASURE01	Double	-9999		
4	MEASURE02	Double	-9999		
5	MEASURE03	Double	-9999		
6	MEASURE04	Double	-9999		
7	MEASURE05	Double	-9999		
8	MEASURE06	Double	-9999		
9	MEASURE07	Double	-9999		
10	MEASURE08	Double	-9999		
11	MEASURE09	Double	-9999		
12	MEASURE10	Double	-9999		
13	MEASURE11	Double	-9999		
14	MEASURE12	Double	-9999		
15	MEASURE13	Double	-9999		

It is convenient for comparison of or editing specifications of several variables, especially when you need to copy and paste between variables or extend a format definition or missing data code from one variable to subsequent variables.

Right-click on the **Variable Specifications Editor** to access a shortcut menu containing the following commands: **Add Vars**, **Delete Vars**, **Cut**, **Copy**, **Paste**, and **Fill/Copy Down**.

VARIABLE OPERATIONS

The **Vars** toolbar button menu contains access to the most common data management operations. Each of the commands on this menu by default operates on the currently selected variable(s) in the *STATISTICA* spreadsheet.

These operations include simple tasks such as adding, deleting, copying, and moving selected groups of variables in the data file. Other operations include transformations of date values, recalculation of existing spreadsheet formulas, lagging one or more variables against the rest of the data file, converting raw data values into their relative ranks within the variable, recoding a variable using logical selection conditions based on other variables in the data file, and creating a subset of data.
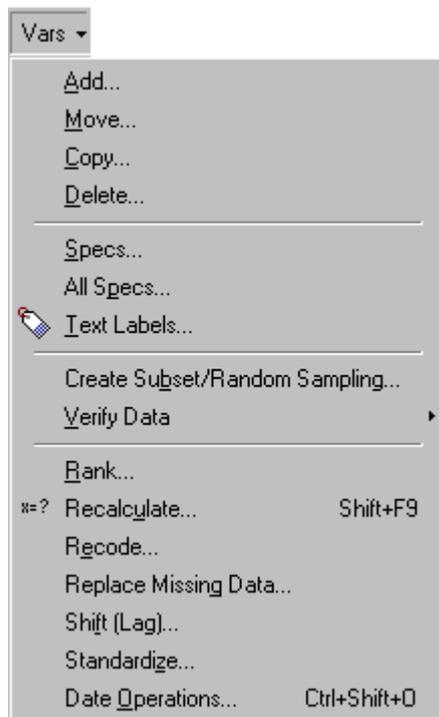
Note: Choose **Copy** from the **Vars** toolbar button menu to copy several variables and insert them in another location in the data file. This command produces different results from the **Copy** command available from the **Edit** menu, which copies the highlighted block of data to the Clipboard. The former (global) copy performs operations on entire variables (as units); the latter (Clipboard-based) copy only operates on blocks of data within variables. The same logic applies to the **Delete** command in this menu.

CASE OPERATIONS

The commands on the **Cases** toolbar button menu are used to perform operations on selected groups of cases in the data file. Again, the **Copy** and **Delete** commands here are much different from the commands of the same name on the **Edit** menu.
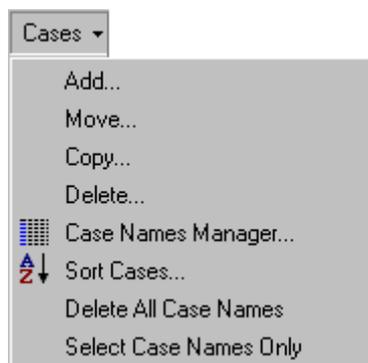

CASE NAMES

Case names can be used as long, unique identifiers for the observations in the spreadsheet. They are also used by default as labels for many graphs.

To enter case names into the spreadsheet, double-click on the gray header of any case, and type the desired name into the field. Press the *Enter* key on your keyboard to move to the next case name.

CASE NAMES MANAGER

The maximum number of characters in a case name and the case header width can be adjusted in the **Case Names Manager** dialog, accessible by selecting **Case Names Manager** from the **Cases** toolbar button menu.
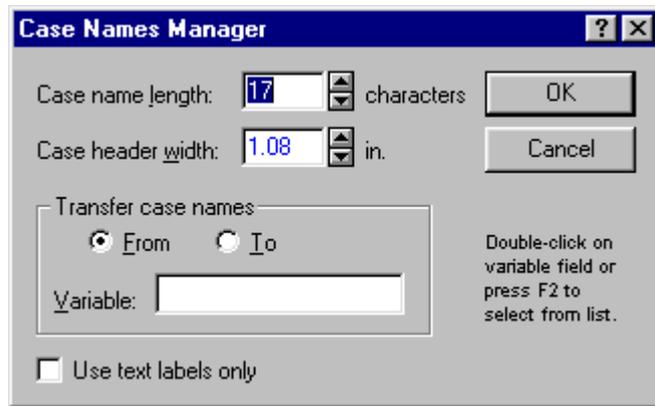
Also, case names can be transferred from a particular variable to the case name headers in the spreadsheet.

VARIABLE TYPES

You can specify the data type of each variable in the variable specifications dialog (accessible by selecting **Variable Specs** from the **Data** menu). *STATISTICA* Spreadsheet data files support the four basic data types listed below:

Double is the default format for storing numeric values in *STATISTICA*. Each numeric value can have a unique text label attached. When your data type is **Double**, each cell takes up 8 bytes of storage (plus the optional text label).

Integer is the data type to select for whole number values. You cannot enter numeric values containing decimals into a variable of this type. Each numeric value can have a unique text label attached. When your data type is **Integer**, each cell takes up 4 bytes of storage (plus the optional text label). Hence, this data type offers a more economical way of storing numbers and is recommended for storing integer data in large data files.

Byte is the data type for integers between and including 0 through 255. You cannot enter numeric values containing decimals into a variable of this type. Each byte value can have a unique text label attached. The advantage of specifying **Byte** as your data type is that it offers the most economical storage for values that are small integers, as each cell takes up only 1 byte of storage (plus the optional text label).

Text is optimized for storing sequences of any characters of long length. The length of a field reserved for text variable type is not constant and can be adjusted.

TEXT LABELS

For many statistical data analysis applications, it is useful to use text labels that can aid in the interpretation of their respective numeric values. For example, in the input spreadsheet, you could enter the values *1* and *2* in the variable *GENDER* to refer to males and females, respectively. Then, using the **Text Labels Editor** (accessible by selecting **Text Labels Editor** from the **Data** menu), you can assign *MALE* to the value of *1* and *FEMALE* to the value of *2*.

When you click the **OK** button, all the *1*'s in the column of the *GENDER* variable will automatically change to *MALE*, and all the *2*'s will change to *FEMALE*.

EXAMPLE 1

Let's begin by creating a hypothetical data file. We will enter information about 18 people. The spreadsheet will contain the gender, eye color, hair color, height, weight, and age of each person.

1. **Create a new data file with 6 variables and 18 cases.** To do this, select **New** from the **File** menu to display the **Create New Document** dialog. On the **Spreadsheet** tab, enter 6 in the **Number of variables** box and 18 in the **Number of cases** box. For this example, select the **As a stand-alone window** option button, and then click the **OK** button.
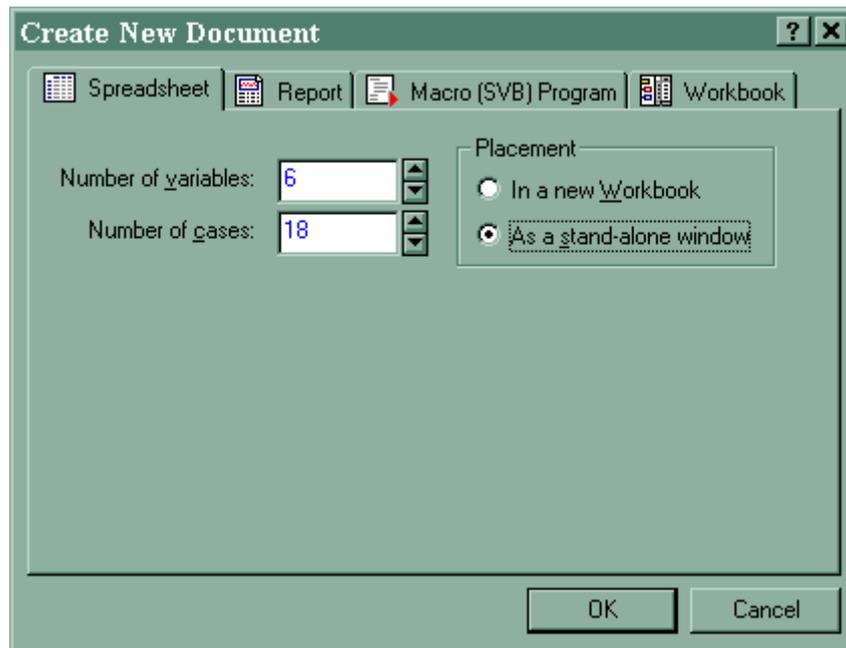
2. **Save the spreadsheet.** Select **Save As** from the **File** menu to display a standard **Save As** dialog, and name this empty spreadsheet *Information.sta*. Click the **Save** button.

3. **Give the 6 variables the appropriate names as listed above.** You can do this several different ways, but the easiest may be to select **All Specs** from the **Vars** toolbar button menu to display the **Variable Specifications Editor**. In the **Name** column, type in the 6 characteristics. To go from one variable name to the next, use the down arrow key on your keyboard.

4. **Change the variable type.** Since we know that *Gender* and *Eye Color* will contain only text characters, we may want to change the type of these two variables to **Text**. You can do this in the **Variable Specifications Editor** as well. Click the arrow button ▾ adjacent to *Gender*, and select **Text** from the drop-down menu; then repeat for *Eye Color*. Now, click the **OK** button.

Variable Specifications Editor

A | Arial ▼ | 10 ▼ | **B** *I* U | **A** ▾ | ✂ 📋 📋 | Vars ▾

	Name	Type	MD code	Length	Long Name (label or formula)
1	Gender	Text ▼		8	
2	Eye Color	Text ▼		8	
3	Hair Color	Double ▼	-9999		
4	Height	Double ▼	-9999		
5	Weight	Double ▼	-9999		
6	Age	Double ▼	-9999		

OK | Cancel

5. **Enter the data.** For this example, lets say the first 9 subjects are female and the last 9 subjects are male. You can enter this information easily by using the **Fill** option. To do this, type *Female* in the first cell of the variable *Gender*. Then select the first 9 rows of this variable including the word *Female*. Right-click in this highlighted area, and select **Fill/ Standardize Block - Fill/Copy Down** from the shortcut menu. Do the same for the males by entering *Male* into the tenth cell, etc.
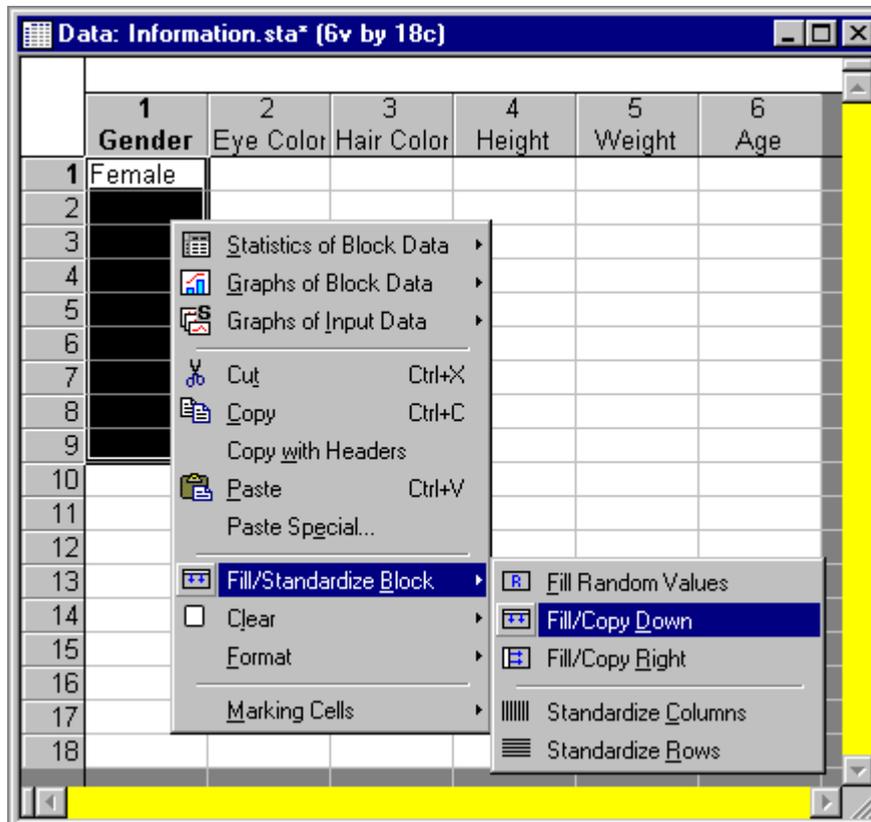
6. **Enter the values for *Eye Color*.** Say we observed blue eyes, brown eyes, and green eyes. Enter *Blue*, *Brown*, and *Green* into the first three cells under *Eye Color*. Then, select those three values and use the Excel-style drag-and-drop feature to fill in the remainder of the cells replicating this pattern: Move the cursor to the lower-right corner of the highlighted area until it changes to a black plus ✚. Then click and drag down the column to the 18th row.

	1 Gender	2 Eye Color	3 Hair Color	4 Height	5 Weight	6 Age
1	Female	Blue				
2	Female	Brown				
3	Female	Green				
4	Female	Blue				
5	Female	Brown				
6	Female	Green				
7	Female	Blue				
8	Female	Brown				
9	Female	Green				
10	Male	Blue				
11	Male	Brown				
12	Male	Green				
13	Male	Blue				
14	Male	Brown				
15	Male	Green				
16	Male	Blue				
17	Male	Brown				
18	Male	Green				

Data: Information.sta* (6v by 18c)

7. **Color the cells under *Gender*.** Specify blue for the cells containing *Female* and yellow for the cells containing *Male*. To do this, select the cells that contain *Female*, click the arrow on the **Fill Color** toolbar button, and select blue from the color palette. Or you can right-click on the highlighted cells and select **Format – Cell**s from the resulting shortcut menus to display the **Format Cells** dialog. On the **Font** tab, click the arrow under **Background Color** to display a color palette, select blue, and click the **OK** button. Repeat the process for the cells containing *Male*, selecting the color yellow.

8. **Color the text in the cells under *Eye Color*.** Specify green for the text in the cells that contain *Green*. To do this, click on one of these cells, and then click the arrow on the **Font Color** toolbar button. Or you can right-click on a cell and select **Format – Cells** from the resulting shortcut menus to display the **Format Cells** dialog. On the **Font** tab, click the arrow under **Text Color** to display a color palette, select green, and click the **OK** button. Repeat the process for the other cells that list *Green* for the variable *Eye Color*. Alternatively, you can hold down the CTRL key on your keyboard while you click on each cell that contains *Green*, and then apply the green text to all of them at once.

9. **Save your changes.** Even though we have not yet completed entering the data, save your changes to this data file by selecting **Save** from the **File** menu.

DESCRIPTIVE STATISTICS

Several types of basic statistical analyses are available from the **Basic Statistics and Tables** dialog, accessible by selecting **Basic Statistics/Tables** from the **Statistics** menu. These procedures give you more depth and control over the output than statistics of input data.
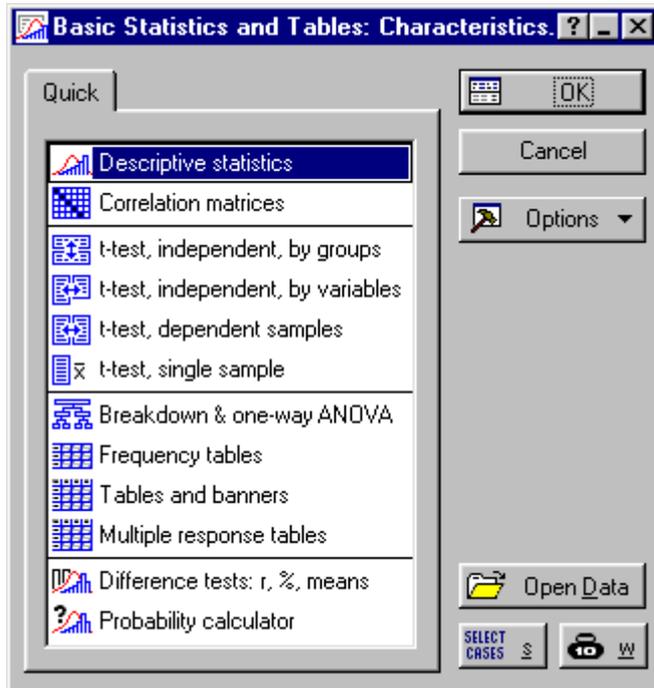
For the following demonstrations, the *Characteristics.sta* data file will be used. The output will be displayed in a workbook. If you would like your output to be displayed in the workbook automatically, select **Options** from the **Tools** menu to display the **Options** dialog. Select the **Output Manager** tab, and select the **Workbook** and the **Single Workbook (common for all Analyses/graphs)** option buttons. Then click **OK** to close the dialog.

Select **Descriptive Statistics** to display the **Descriptive Statistics** dialog.
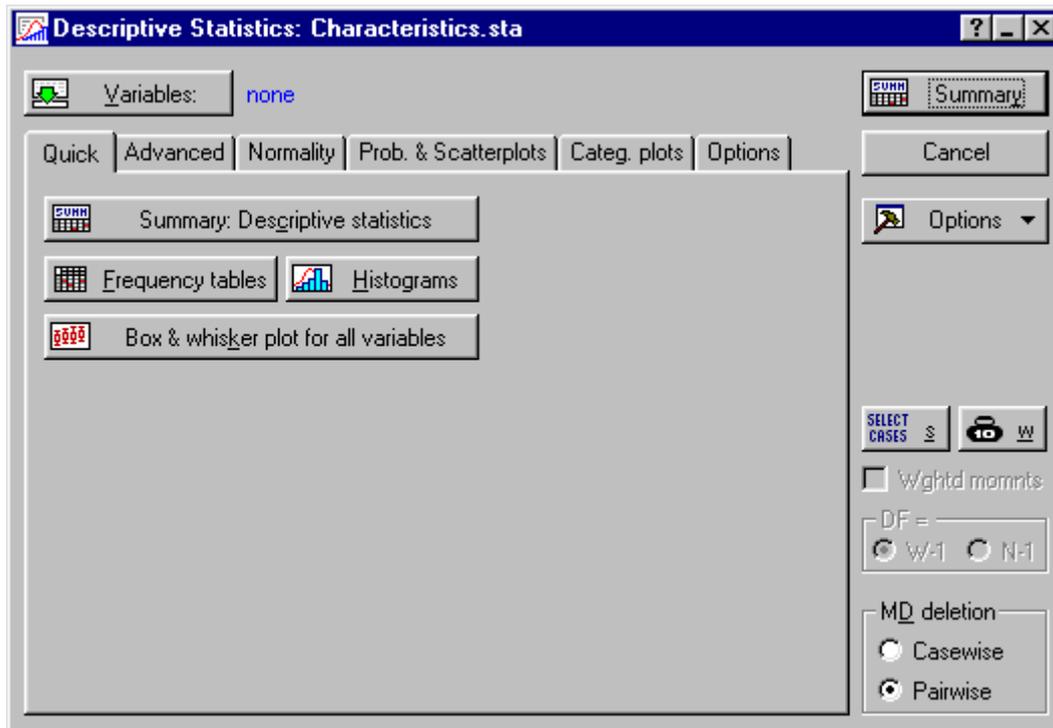
The analysis dialog is structured so that similar features are conveniently grouped together on tabs for easier navigation and selection of commonly used procedures and graphs to describe your data.

SELECTING VARIABLES AND STATISTICS

Click the *Variables* button to display a variable selection dialog, which contains a list of all available variables in the data file. Here, you select the variables to be analyzed. You can select consecutive variables by selecting the first variable with the mouse pointer and then dragging it down the to the last variable you wish to select. You can also select a discontinuous list of variables by holding down the CTRL key and clicking with the mouse at the same time. A third technique for selecting variables is typing in the variable numbers into the *Select Variables* field.
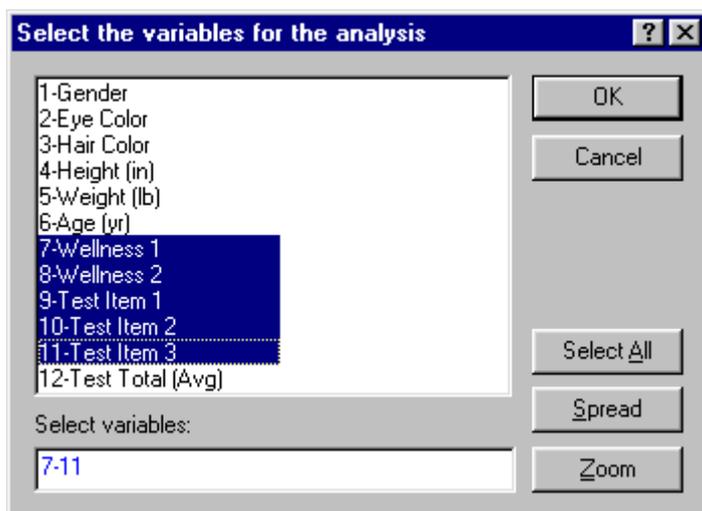
After selecting the desired variables, click the *Summary* button to produce the results spreadsheet with the default selection of statistics.

By default, the valid N, mean, minimum, maximum, and standard deviation are displayed (for definitions of these statistics, see the glossary). A more comprehensive selection of descriptive statistics to compute is available on the **Advanced** tab. On this tab, you can precisely control the statistics that are computed and displayed in the results spreadsheet.

NORMALITY

Many basic tests of statistical significance are based upon assumptions of normality regarding the data used for the test. The **Normality** tab of the **Descriptive Statistics** dialog contains many of the most common tools for checking normality assumptions. These tools include frequency tables, histograms with a normal fit, and statistical tests for normality.

EXAMPLE 3

1. Select **Open** from the **File** menu. Browse to the *C:\Program Files\StatSoft\STATISTICA 6\Examples\Datasets* directory and select *Characteristics.sta*.

2. Select **Basic Statistics/Tables** from the **Statistics** menu. Select **Descriptive statistics** and click **OK**.

3. Click the **Variables** button and highlight variables *4-8*. Click **OK**.

Select the variables for the analysis

```
1-Gender
2-Eye Color
3-Hair Color
4-Height (in)
5-Weight (lb)
6-Age (yr)
7-Wellness 1
8-Wellness 2
9-Test Item 1
10-Test Item 2
11-Test Item 3
12-Test Total (Avg)
```

OK
Cancel
Select All
Spread
Zoom

Select variables:

`4-8`
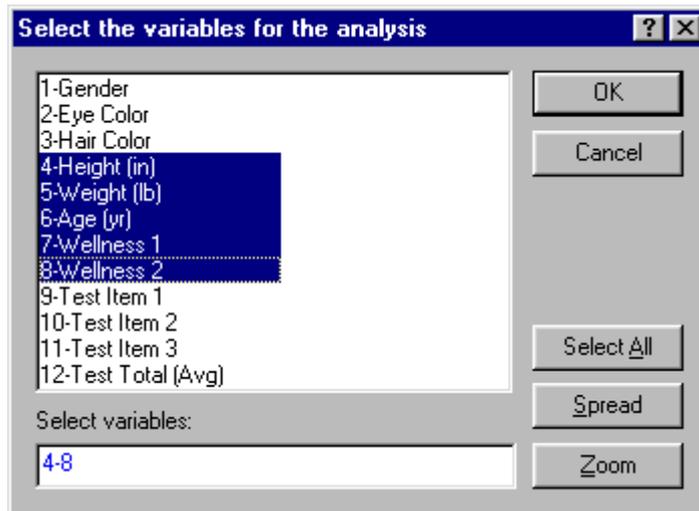
4. Click the **Summary** button on the **Descriptive Statistics** dialog to display the default descriptive statistics for the selected variables.

Workbook1.stw* - Descriptive Statistics (Characteristics.sta)

Workbook1.stw*
Basic Statistics/Tab
Descriptive stat
Descriptive

Variable	Descriptive Statistics (Characteristics.sta)				
	Valid N	Mean	Minimum	Maximum	Std. Dev.
Height (in)	100	67.8900	57.00000	75.0000	3.57289
Weight (lb)	100	184.9800	80.00000	282.0000	36.16711
Age (yr)	100	44.0800	20.00000	79.0000	15.97579
Wellness 1	100	49.3707	19.55478	79.6092	11.84380
Wellness 2	100	60.7563	35.02135	87.3602	11.85713

Descriptive Statistics (Characteristics.sta)

5. Resume the analysis by clicking the **Descriptive Statistics** button on the *Analysis* bar in the lower-left corner of the screen. Click the **Histograms** button to produce histograms with a normal fit for each selected variable.
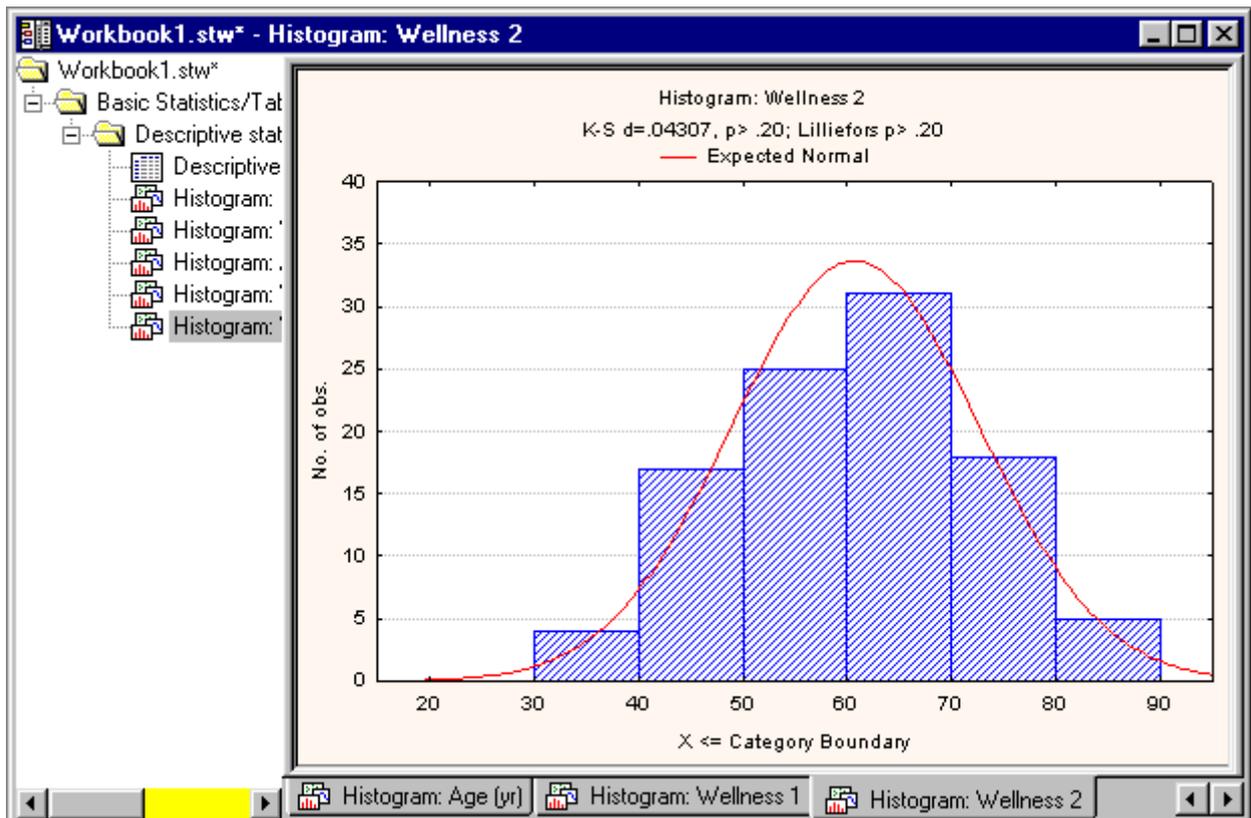
6. Resume the analysis and click on the **Advanced** tab. Select only the following statistics: Under **Location, valid N**, select **Mean**; under **Variation, moments**, select **Skewness** and **Kurtosis**; under **Percentiles, ranges**, select **Minimum & maximum** and **Range**. Also, click the **Variables** button to reselect only the variables *Height (in)*, *Weight (lb)*, and *Age (yr)*.

7. Click the **Summary: Descriptive statistics** button to view the selected computations.
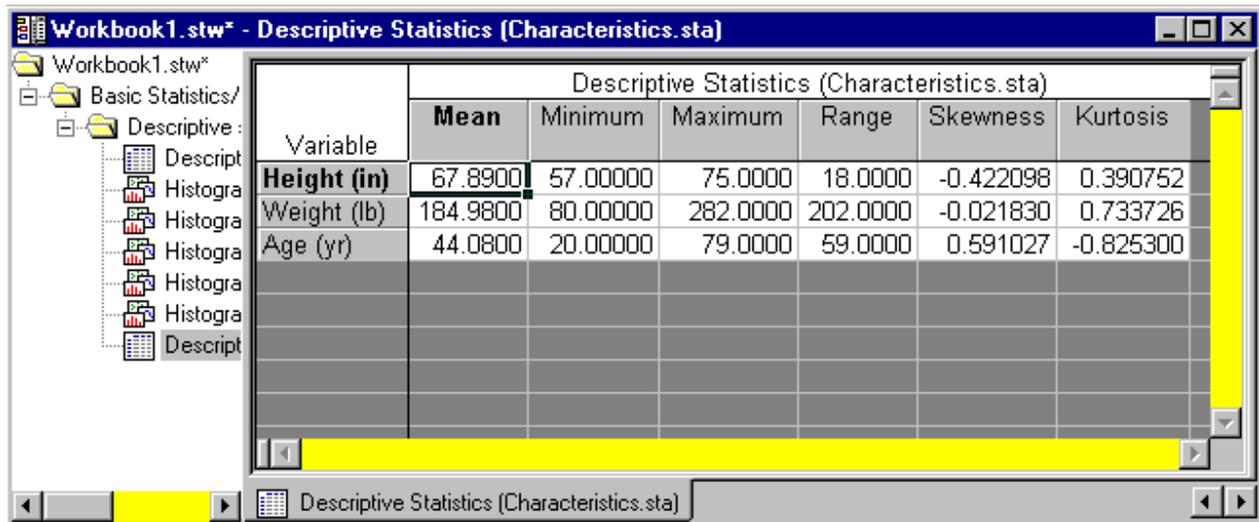
Notice the positive skewness (*0.059*) for the variable *Age (yr)*. Compare this with the histogram for *Age (yr)* you made previously in this example. Positive skewness is indicated by the data being skewed to the right.

8. To further investigate the distribution of a variable, you can view a normal probability plot. Create this type of plot for the three variables we have chosen by resuming the analysis and clicking on the **Prob.& Scatterplots** tab. Click the **Normal probability plot** button.
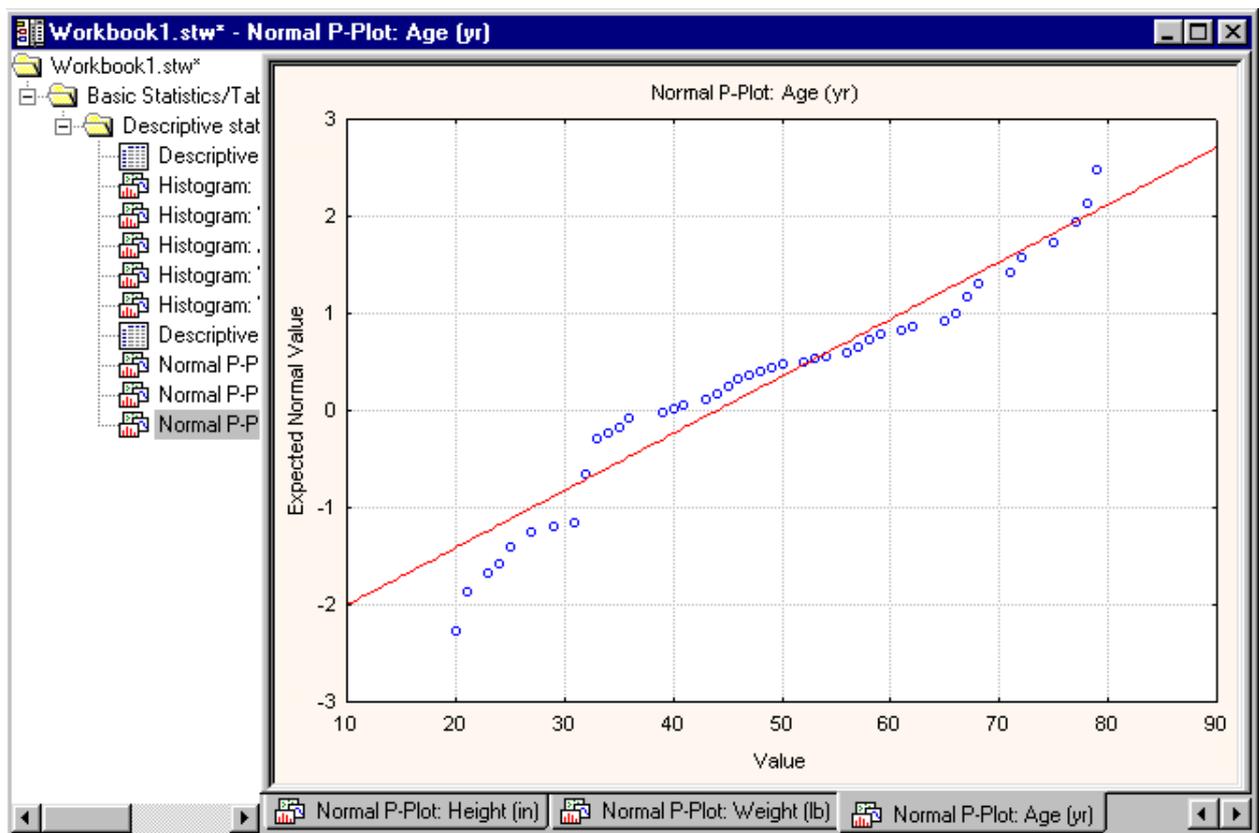
Three plots will be placed into the workbook, one for each selected variable. Notice that *Age (yr)* is the variable that deviates from the Normal distribution the most when compared to the *Height (in)* and *Weight (lb)* variables.

The t-test is the most commonly used method to evaluate differences in means between two groups. The test assumes that the data follow a normal distribution within each group, and that the variance within each group is the same across the groups. Alpha levels, p-values, variances, standard deviations and other terms are important when discussing t-tests. You may want to refer to the glossary or any elementary statistics book to understand the concepts of these terms. On the **Basic Statistics and Tables** (Startup Panel), there are four different t-tests available. This section will explain when each should be used in an analysis.

When one variable contains codes for two groups and the second variable contains measurements or values of a dependent variable, one should use a t-test by groups to compare the group means.

EXAMPLE 5

1. Open the *Characteristics.sta* data file. Select **Basic Statistics/Tables** from the **Statistics** menu to display the **Basic Statistics and Tables** dialog. Then, select **t-test, independent, by groups** and click **OK** to display the **T-Test for Independent Samples by Groups** dialog.

2. Click the **Variables** button, select *Height (in)* as the dependent variable, and select *Gender* as the grouping variable. Click **OK**.
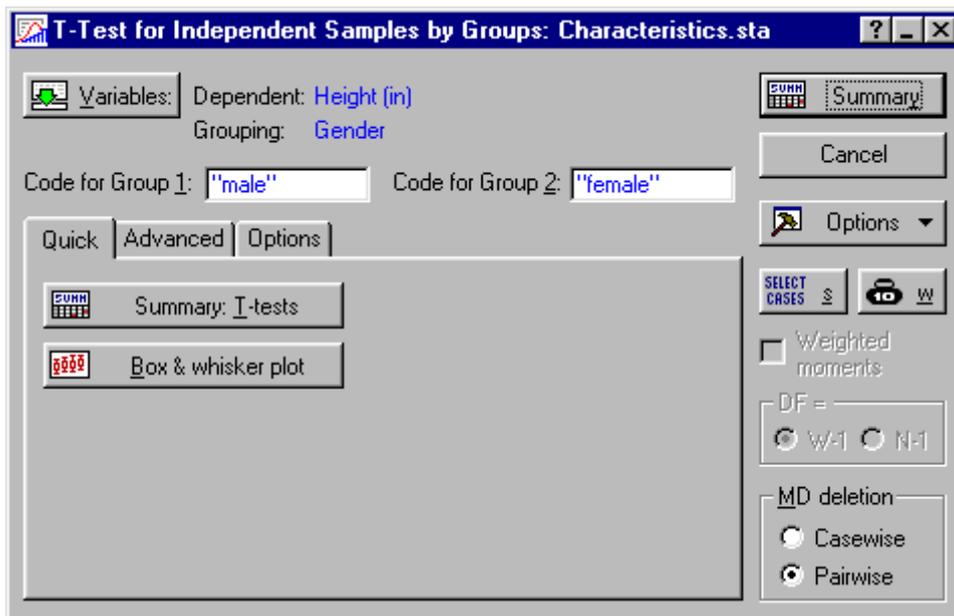
3. Click the **Summary** button to display the results of the t-test.

| | T-tests; Grouping: Gender (Characteristics.sta) Group 1: male Group 2: female | | | | | | | |
Variable	Mean male	Mean female	t-value	df	p	Valid N male	Valid N female	Std.Dev. male
Height (in)	67.78846	68.00000	-0.294428	98	0.769054	52	48	3.488561

Notice that several statistics are given in this spreadsheet. Scroll to the right in the spreadsheet to view all of the results. The t-test compared the sample mean for female (*x-bar* = 68) to the sample mean for male (*x-bar* = 67.78). It computed a t-value, along with its corresponding p-value, in order for the user to evaluate and decide whether the means are significantly different from each other. In this example, with a large p-value of *0.769*, we can conclude that the means are not different from each other. The female group is similar to the male group with regard to the *Height (in)* variable.

t-TEST, INDEPENDENT, BY VARIABLES

Instead of the data file having one variable that holds the group codes (like *Gender* in the previous example), maybe the data file has two variables of data, one for each group. For example, in the previous example, the data file would have one variable containing the *male* data and one variable containing the *female* data. Of course, independent comparison methods would still apply. When the two groups to be compared reside in separate variables, it is more appropriate to choose a t-test by variables.

EXAMPLE 6

Suppose we are still interested in the height measurement. We want to know if the average male height significantly differs from the average female height. The data resides in the *CharacteristicsHeight.sta* data file. Open this file and notice the format. The data is the same regarding the height measurements and the gender.

1. Select **Basic Statistics/Tables** from the **Statistics** menu. Select **t-test, independent, by variables** and click **OK**.

2. Click the **Variables (groups)** button and select *Male Height* as the first variable and *Female Height* as the second variable. Click **OK**.

3. Click the **Summary** button to display the results of the t-test.


Notice that the same result is computed when compared to EXAMPLE 5. The difference is only in the structure of the data file. We can conclude that the means are not different from each other.

t-TEST, DEPENDENT SAMPLES

The t-test for dependent samples helps you take advantage of one specific type of design in which an important source of within-group variation can easily be explained. If the two groups being compared were measured twice on the same variable, then a considerable portion of the within-group variation can be attributed to the individual differences between measurements on the same subjects. To illustrate this, follow the next example.

EXAMPLE 7

1. Again, open the *Characteristics.sta* data file. Select **Basic Statistics/Tables** from the **Statistics** menu. Select **t-test, dependent samples** and click **OK**.

2. Click on the **Variables** button and select *Test Item 1* as the first variable and *Test Item 2* as the second variable. Click **OK**.
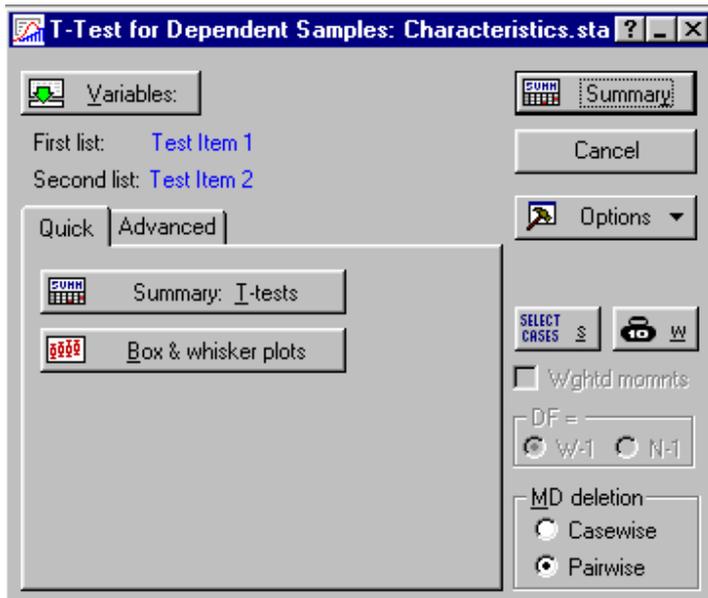
3. Click the **Summary** button to see the results of the t-test.

We are assuming that *Test Item 1* was measured on all 100 subjects at a certain time and then *Test Item 2* was measured on the same 100 subjects, under the same conditions, but at a different time. These are then dependent variables. Is there a difference between the two variables? Yes. With a p-value of 0, you have sufficient evidence to conclude that *Test Item 1* is different from *Test Item 2*.

4. Resume the analysis. Click the **Box & whisker plots** button to graphically display the difference you found between the two variables.
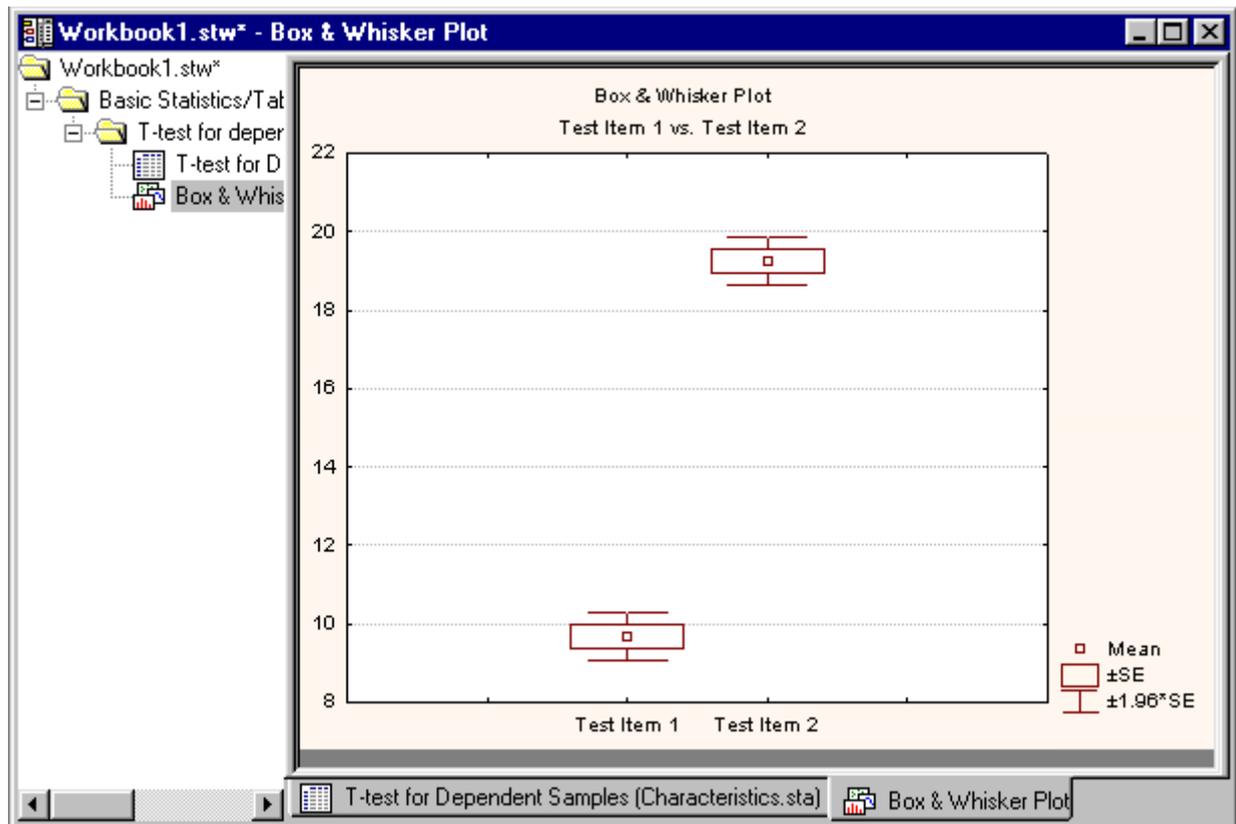
Using the single sample t-test, you can compare the mean of a particular variable to a specified value. Let us demonstrate again with an example.

EXAMPLE 8

1. With the *Characteristics.sta* data file open, select **Basic Statistics/Tables** from the **Statistics** menu. Select **t-test, single sample** and click **OK**.

2. Click the **Variables** button and select *Weight (lb)* as the variable for the analysis. Click **OK**.

3. Select the **Test all means against** option button and enter *200* into the adjacent box. You are testing to see if the mean of *Weight (lb)* significantly differs from 200 pounds.
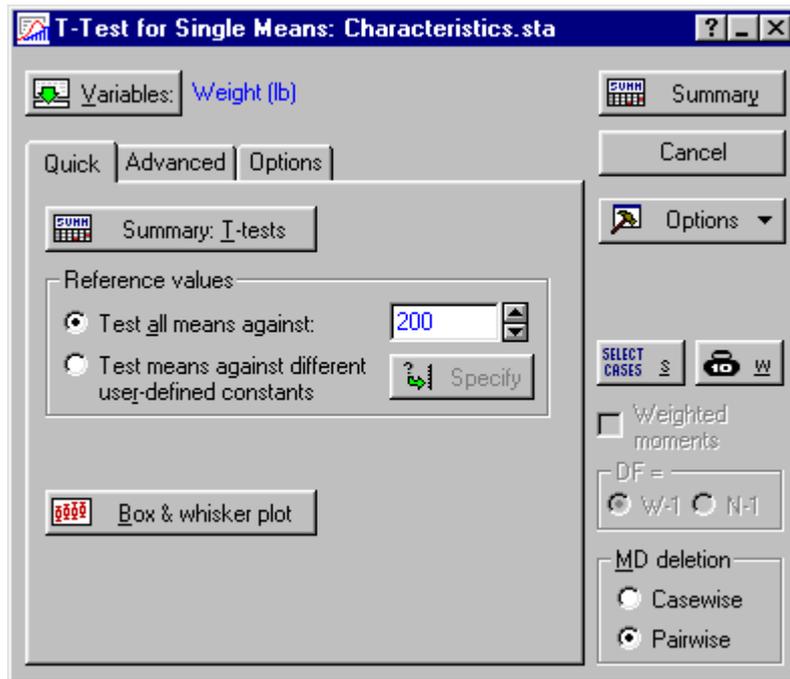
4. Click the **Summary** button to view the results of the t-test.

With a small p-value of almost 0, you can conclude that the mean of *Weight (lb),* in which its sample mean is 184.98, is indeed, significantly different from 200.

5. Resume the analysis. Click the **Box & whisker plot** button. Select **Mean/SE/1.96*SE** in the **Box-whisker Type** dialog box. Click **OK** to create the plot.
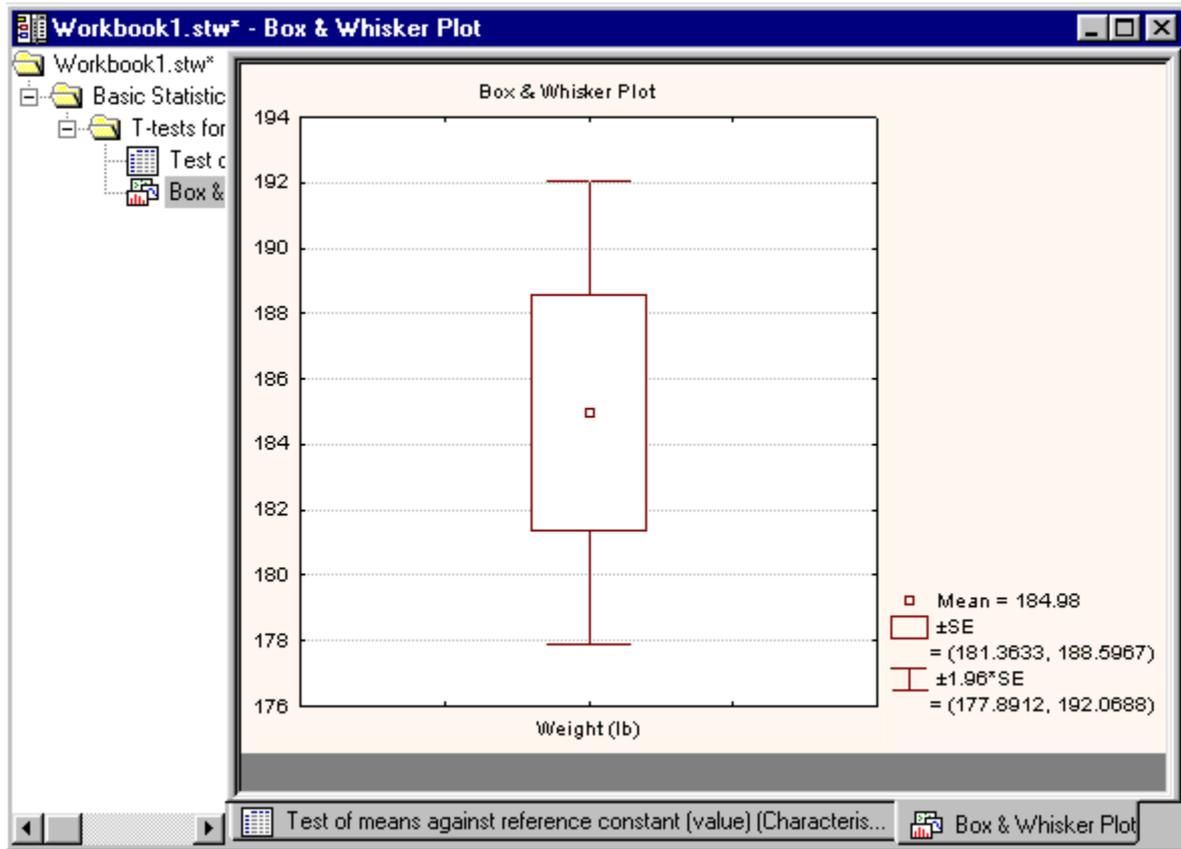
The whiskers in this type of Box & Whisker plot, display the 95% confidence interval about the sample mean. Notice that 200 does not lie within the whiskers (or interval).