StatSoft

STATISTICA

data analysis • data mining • quality control • web-based analytics

Introduction to *STATISTICA*

2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

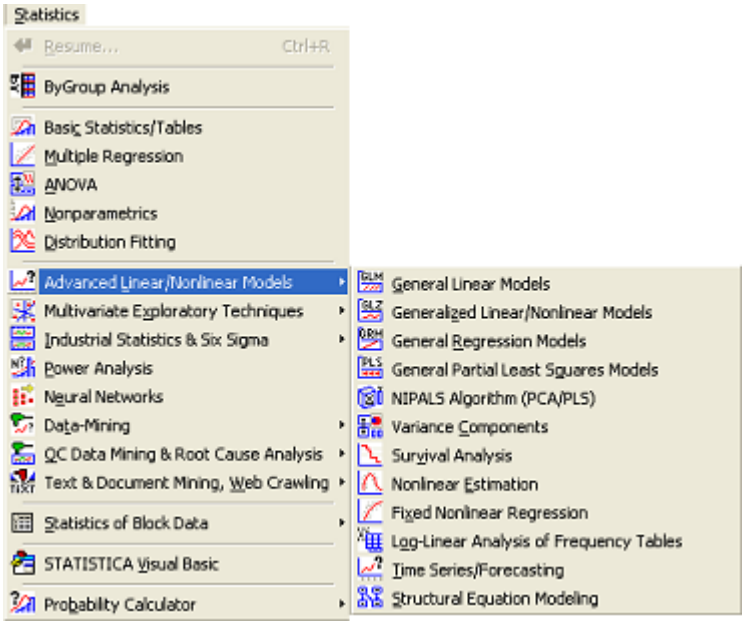
Australia: StatSoft Pacific Pty Ltd. | Germany: StatSoft GmbH | Japan: StatSoft Japan Inc. | Portugal: StatSoft Iberica Ltda. | Spain: StatSoft Espana
Brazil: StatSoft Brazil Ltda. | Hungary: StatSoft Hungary Ltd. | Korea: StatSoft Korea | Russia: StatSoft Russia | Sweden: StatSoft Scandinavia AB
Czech Republic: StatSoft Czech Rep. s.r.o. | Israel: StatSoft Israel Ltd. | Netherlands: StatSoft Benelux BV | Singapore: StatSoft Singapore | Taiwan: StatSoft Taiwan
France: StatSoft France | Italy: StatSoft Italia srl | Poland: StatSoft Polska Sp. z o. o. | S. Africa: StatSoft S. Africa (Pty) Ltd. | UK: StatSoft Ltd.

Outline

After today's training session you will:

- Know the three basic output channels of STATISTICA and how to change the setting to your liking
- Be able to manipulate STATISTICA spreadsheets (e.g. edit variables and cases)
- Perform basic data management operations(recoding data, using spreadsheet formulas, setting case selection conditions, etc.)
- Import data from external data sources such as excel files, databases (as well as create random samples directly from the database)
- Review basic descriptive statistics such as the mean, median, standard deviation, and compute correlations.

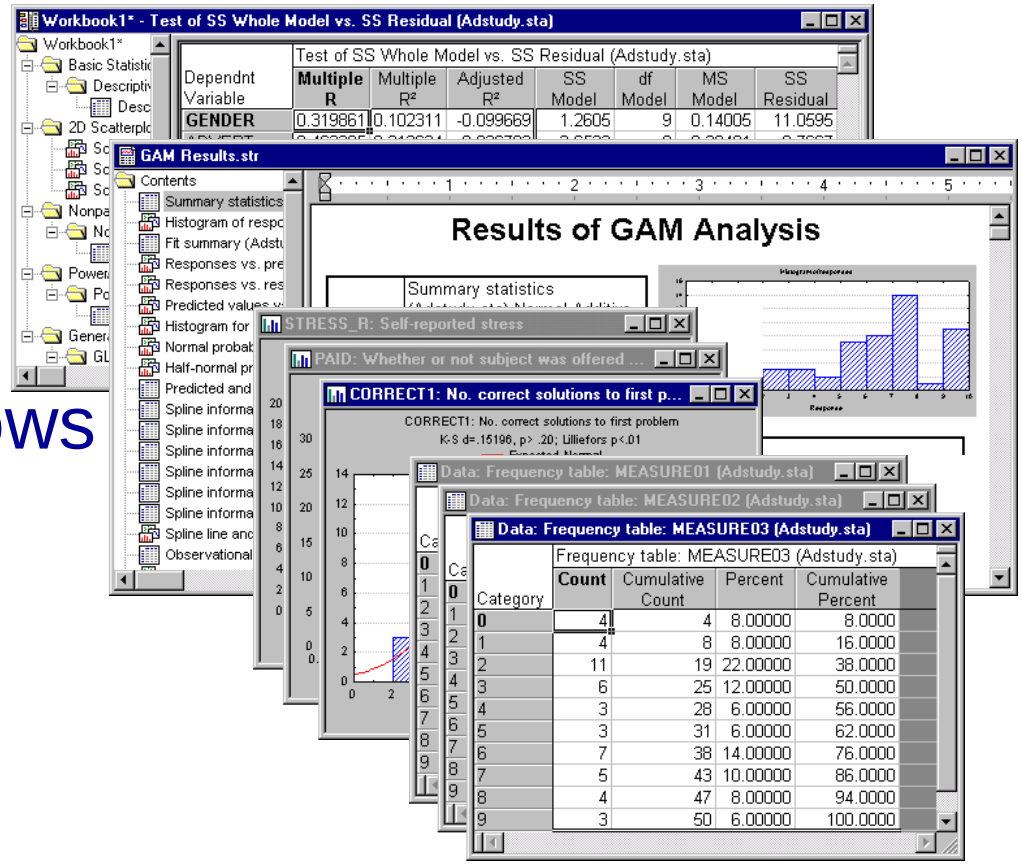
Structure of *STATISTICA* 10



- Use the ***Statistics*** menu to access various analyses
- Multiple analyses can be open simultaneously
- All general purpose facilities are available in every module

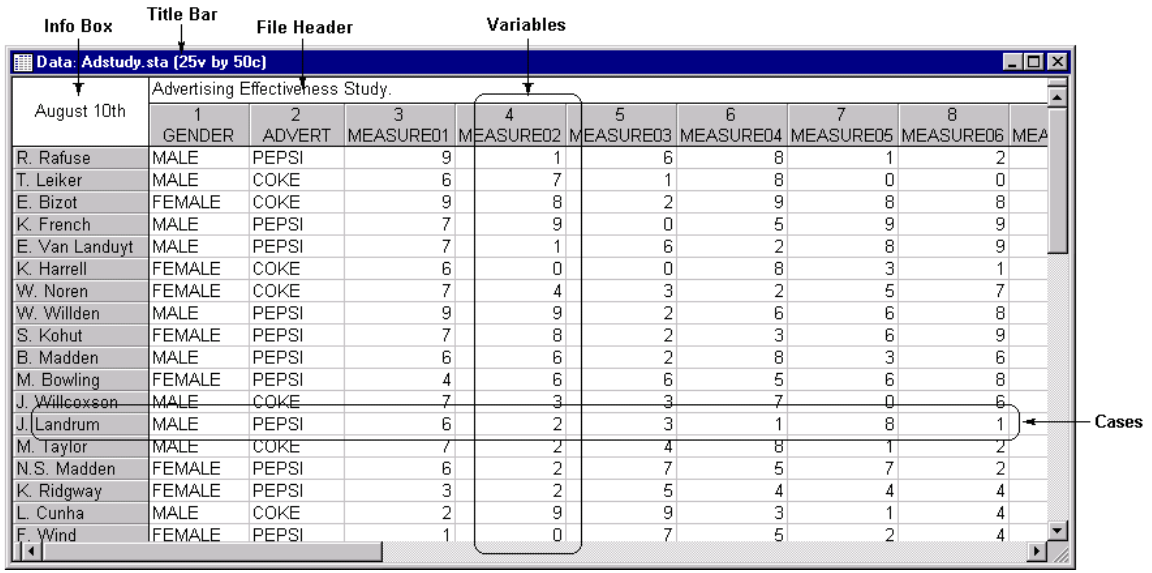
3 Basic Channels for Output

- Workbooks
- Reports
- Stand-alone Windows



Data Files

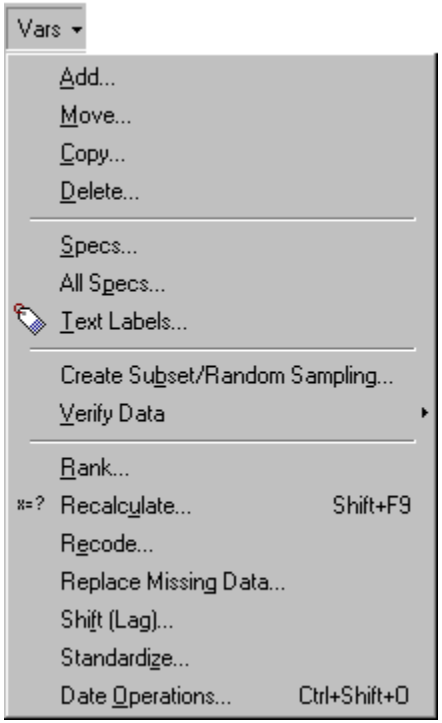
- Info Box
- Title Bar
- File Header
- Variables
- Cases



August 10th	1 GENDER	2 ADVERT	3 MEASURE01	4 MEASURE02	5 MEASURE03	6 MEASURE04	7 MEASURE05	8 MEASURE06	MEA
R. Rafuse	MALE	PEPSI	9	1	6	8	1	2	
T. Leiker	MALE	COKE	6	7	1	8	0	0	
E. Bizot	FEMALE	COKE	9	8	2	9	8	8	
K. French	MALE	PEPSI	7	9	0	5	9	9	
E. Van Landuyt	MALE	PEPSI	7	1	6	2	8	9	
K. Harrell	FEMALE	COKE	6	0	0	8	3	1	
W. Noren	FEMALE	COKE	7	4	3	2	5	7	
W. Willden	MALE	PEPSI	9	9	2	6	6	8	
S. Kohut	FEMALE	PEPSI	7	8	2	3	6	9	
B. Madden	MALE	PEPSI	6	6	2	8	3	6	
M. Bowling	FEMALE	PEPSI	4	6	6	5	6	8	
J. Willcoxson	MALE	COKE	7	3	3	7	0	6	
J. Landrum	MALE	PEPSI	6	2	3	1	8	1	
M. Taylor	MALE	COKE	7	2	4	8	1	2	
N.S. Madden	FEMALE	PEPSI	6	2	7	5	7	2	
K. Ridgway	FEMALE	PEPSI	3	2	5	4	4	4	
L. Cunha	MALE	COKE	2	9	9	3	1	4	
F. Wind	FEMALE	PEPSI	1	0	7	5	2	4	

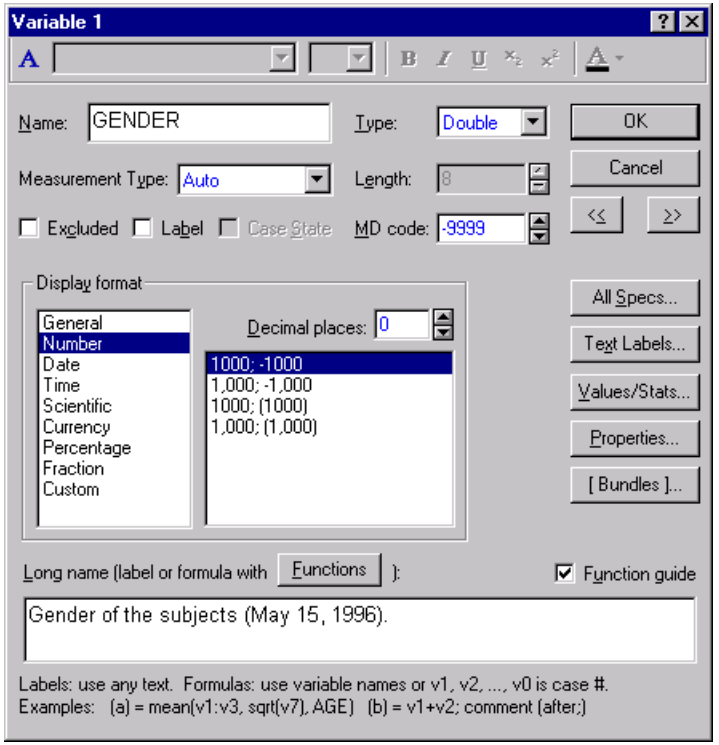
Variable Operations (The Vars button)

Access to the most common data management operations



Variable Dialog

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.



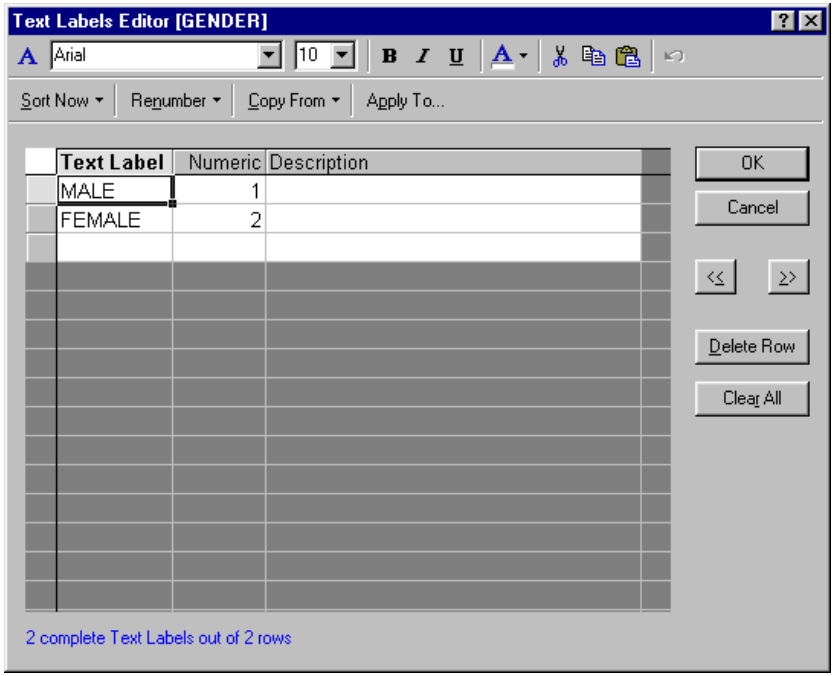
Variable Types

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.
- ***Integer*** is the data type to select for whole number values.
- ***Byte*** is the data type for integers between and including 0 through 255.
- ***Text*** is optimized for storing sequences of any characters of long length.

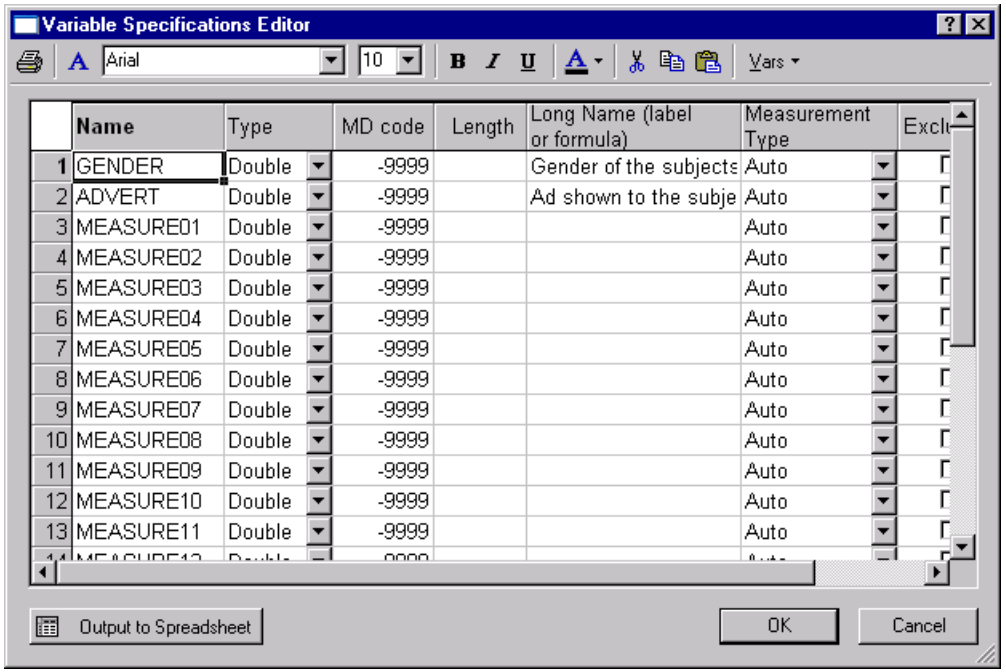
Text Labels

Text labels can aid in the interpretation of their respective numeric values.



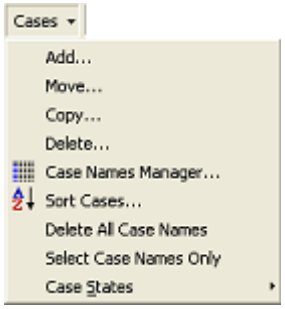
Variables Specifications Editor

All variables can be reviewed or edited in the ***Variable Specifications Editor***



Case Operations (The *Cases* button)

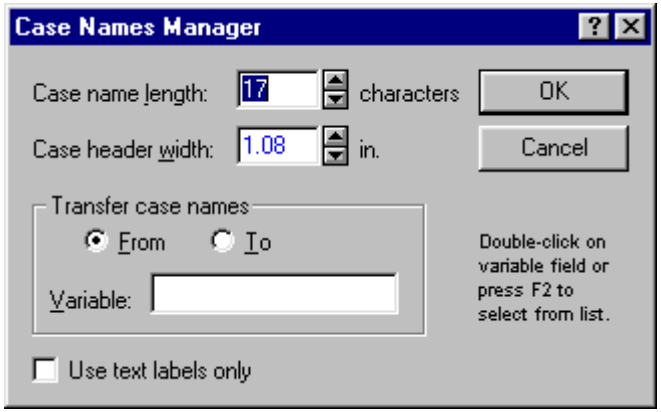
Used to perform operations on selected groups of cases in the data file



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.

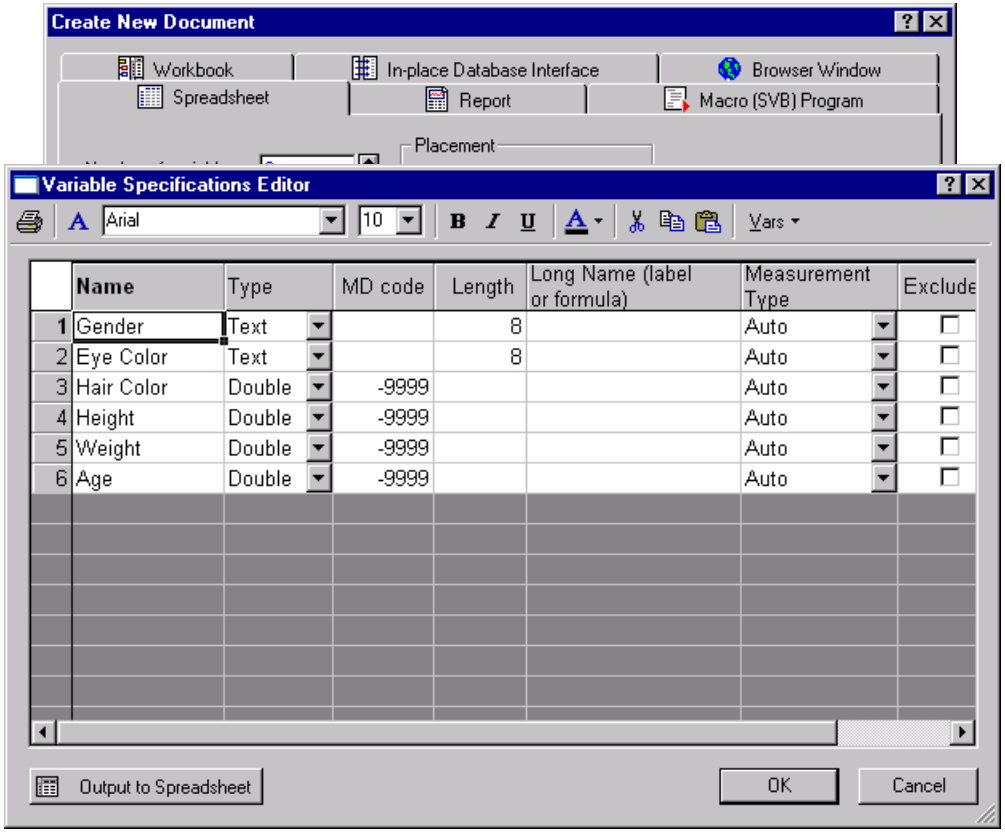
Use the Case Names Manager to adjust case name length and width and to transfer case names to or from a variable



Example 1: Creating a hypothetical data file

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.

Example Steps



1. Create a new data file with 6 variables and 18 cases.
2. Save the spreadsheet.
3. Give the 6 variables these names *Gender*, *Eye Color*, *Hair Color*, *Height*, *Weight*, and *Age* .
4. Change the variable *Type*.

Example Steps

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

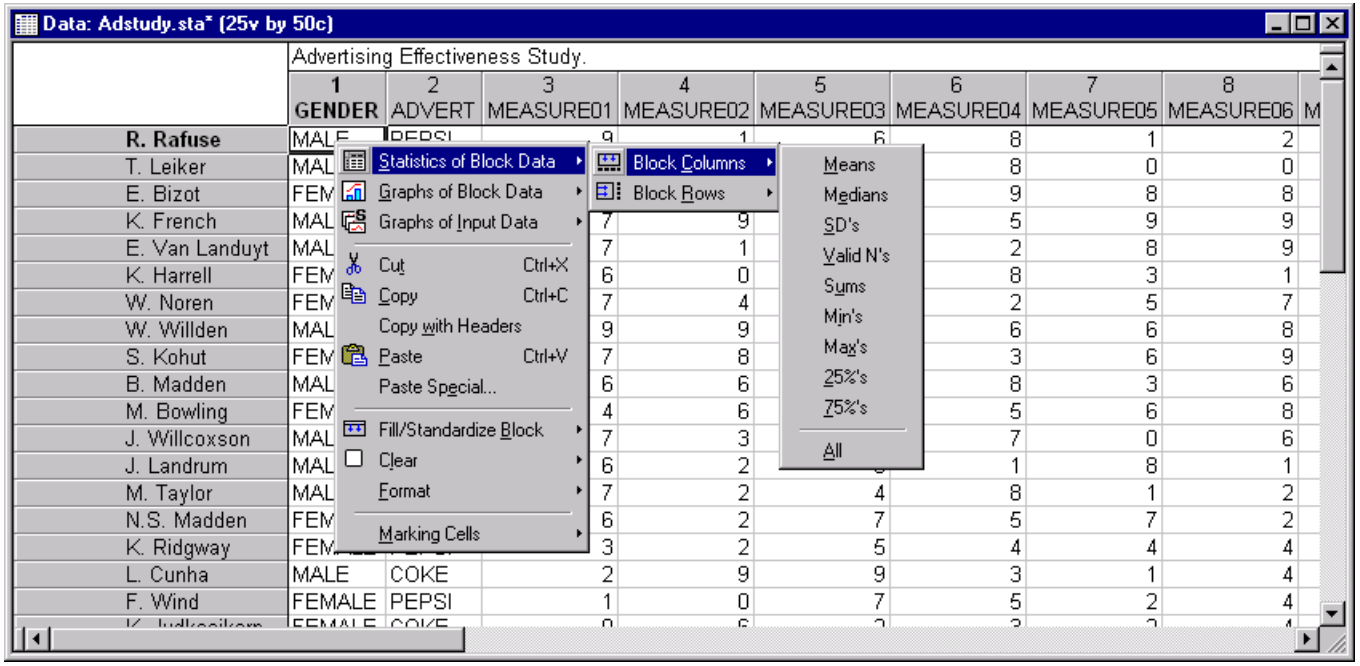
Data: Information.sta* (6v by 18c)

	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				

5. Enter the data.
6. Enter the values for *Eye Color*.
7. Color the cells under *Gender*.
8. Color the text in the cells under *Eye Color*.
9. Save your changes.

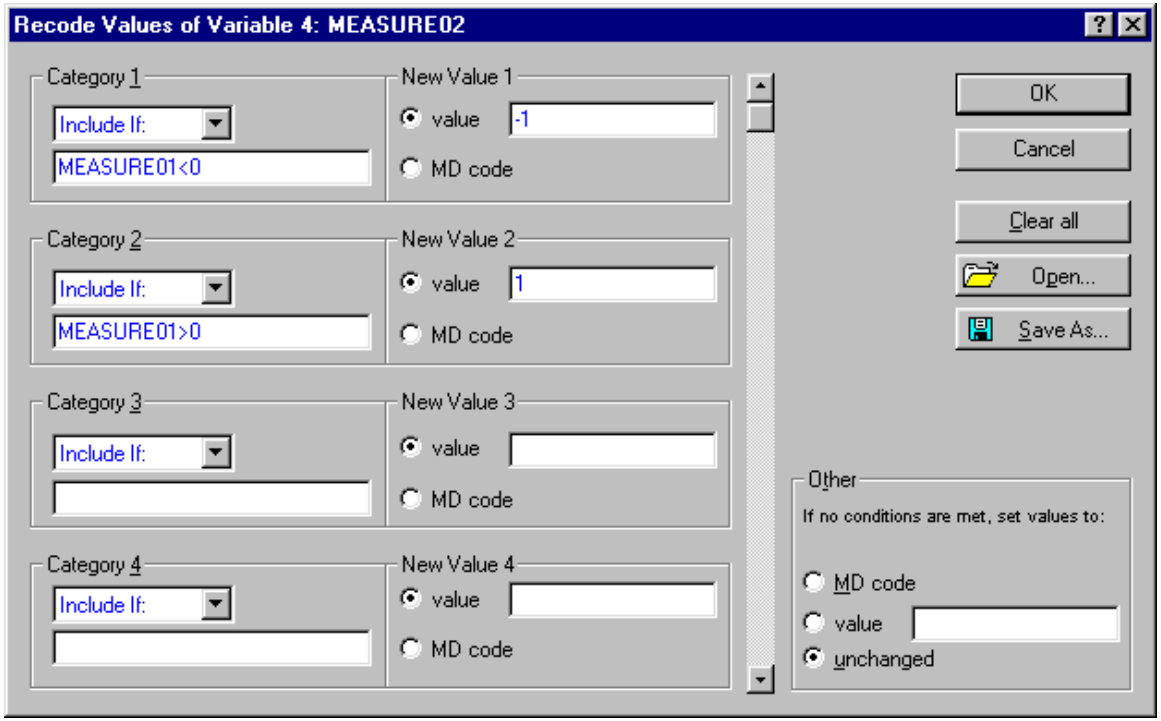
Block Stats

Statistics for each row or column in a selected block can be computed and added to the spreadsheet by selecting the desired ***Statistics of Block Data*** from the shortcut menu.



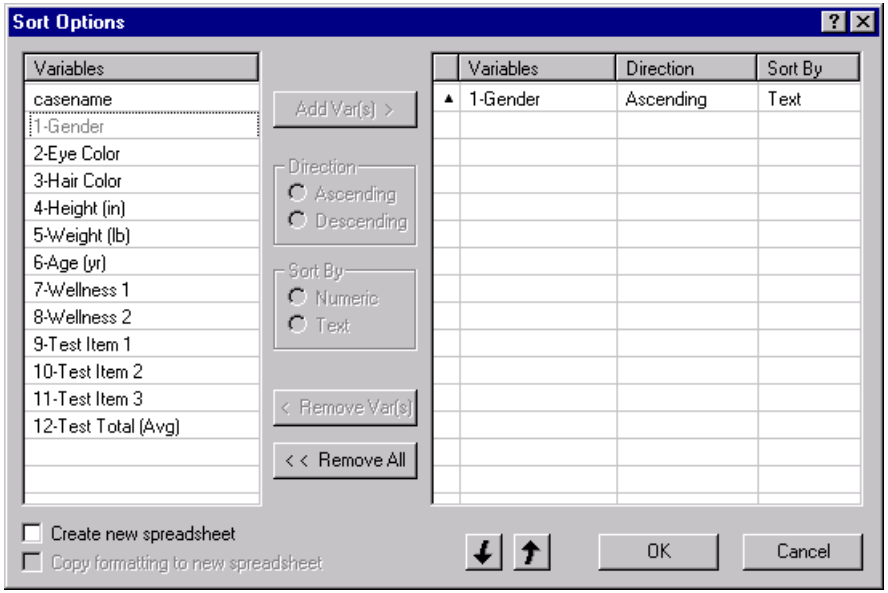
Recode

The Recode dialog is used to define new values of the selected variable depending on the specific conditions that you define.



Sorting

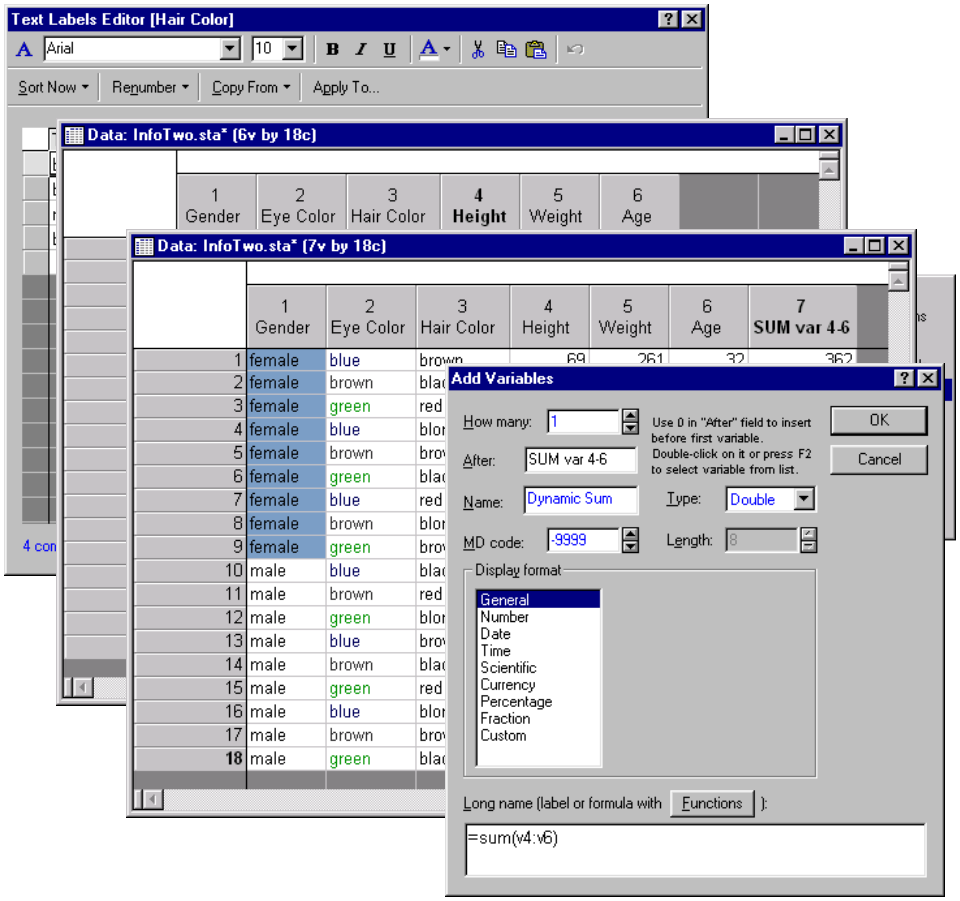
Sorting can be performed on both text and numeric values, as well as on case names, in either descending or ascending order.



Example 2: Working with Variables

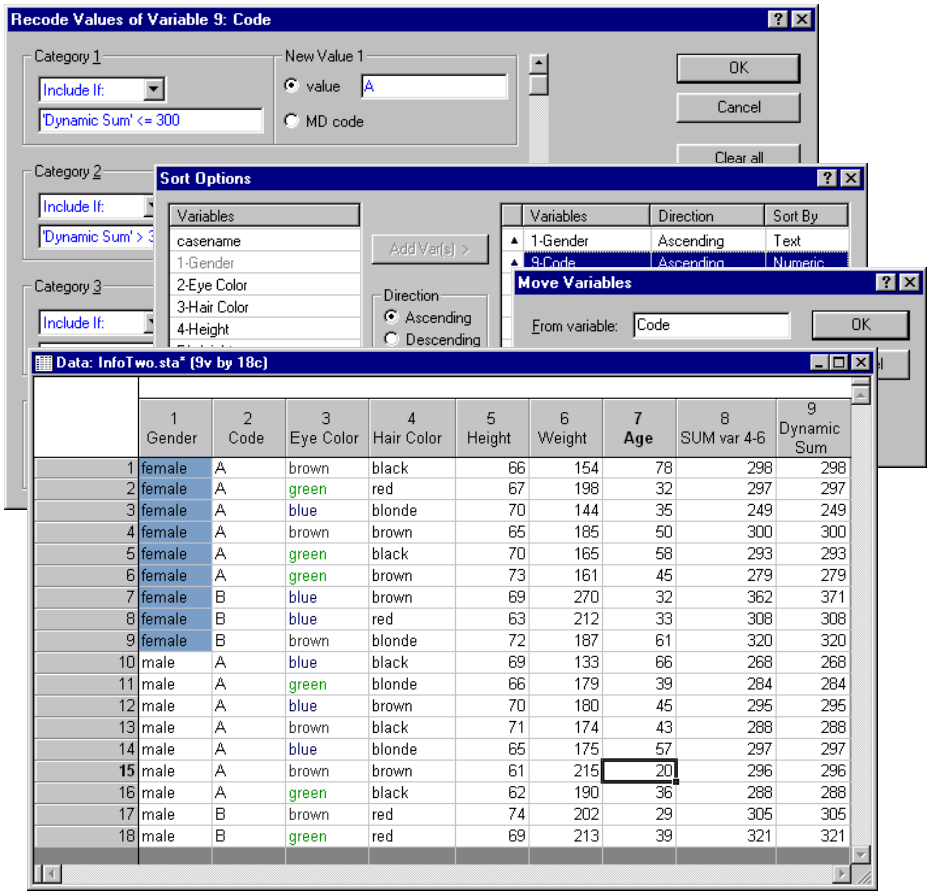
1. **Open the *InfoTwo.sta* data file.** Notice that this data file is similar to the one you created earlier. Data has been entered for the variables *Height*, *Weight*, and *Age*.

Example Steps



2. Enter values for the *Hair Color* variable.
3. Find the sum of the *Height*, *Weight*, and *Age* variables for each case.
4. Create a dynamic variable that automatically updates as the data change.

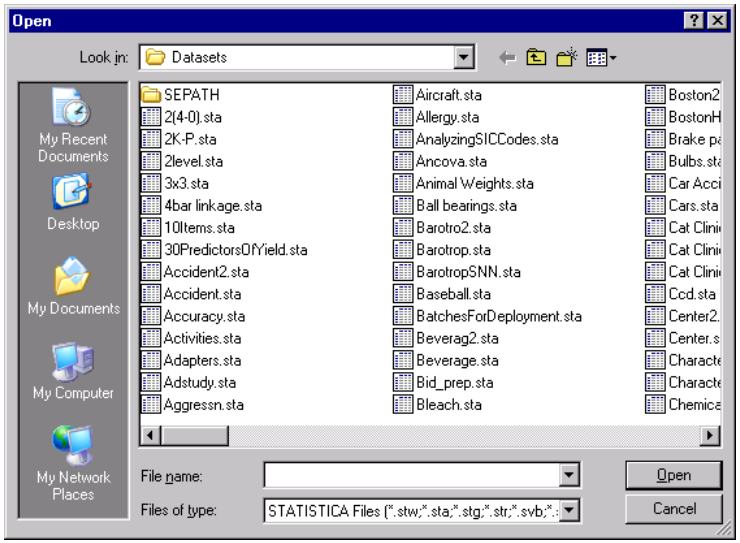
Example Steps

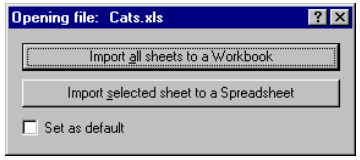


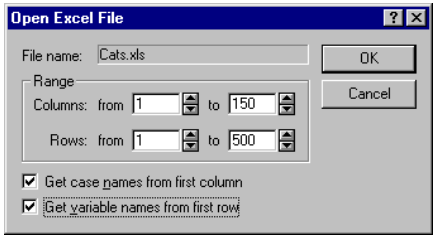
	1 Gender	2 Code	3 Eye Color	4 Hair Color	5 Height	6 Weight	7 Age	8 SUM var 4-6	9 Dynamic Sum
1	female	A	brown	black	66	154	78	298	298
2	female	A	green	red	67	198	32	297	297
3	female	A	blue	blonde	70	144	35	249	249
4	female	A	brown	brown	65	185	50	300	300
5	female	A	green	black	70	165	58	293	293
6	female	A	green	brown	73	161	45	279	279
7	female	B	blue	brown	69	270	32	362	371
8	female	B	blue	red	63	212	33	308	308
9	female	B	brown	blonde	72	187	61	320	320
10	male	A	blue	black	69	133	66	268	268
11	male	A	green	blonde	66	179	39	284	284
12	male	A	blue	brown	70	180	45	295	295
13	male	A	brown	black	71	174	43	288	288
14	male	A	blue	blonde	65	175	57	297	297
15	male	A	brown	brown	61	215	20	296	296
16	male	A	green	black	62	190	36	288	288
17	male	B	brown	red	74	202	29	305	305
18	male	B	green	red	69	213	39	321	321

5. Create another new variable that will contain codes.
6. Sort the data file.
7. Rearrange the order of the variables.
8. Rank variables.

Importing Data



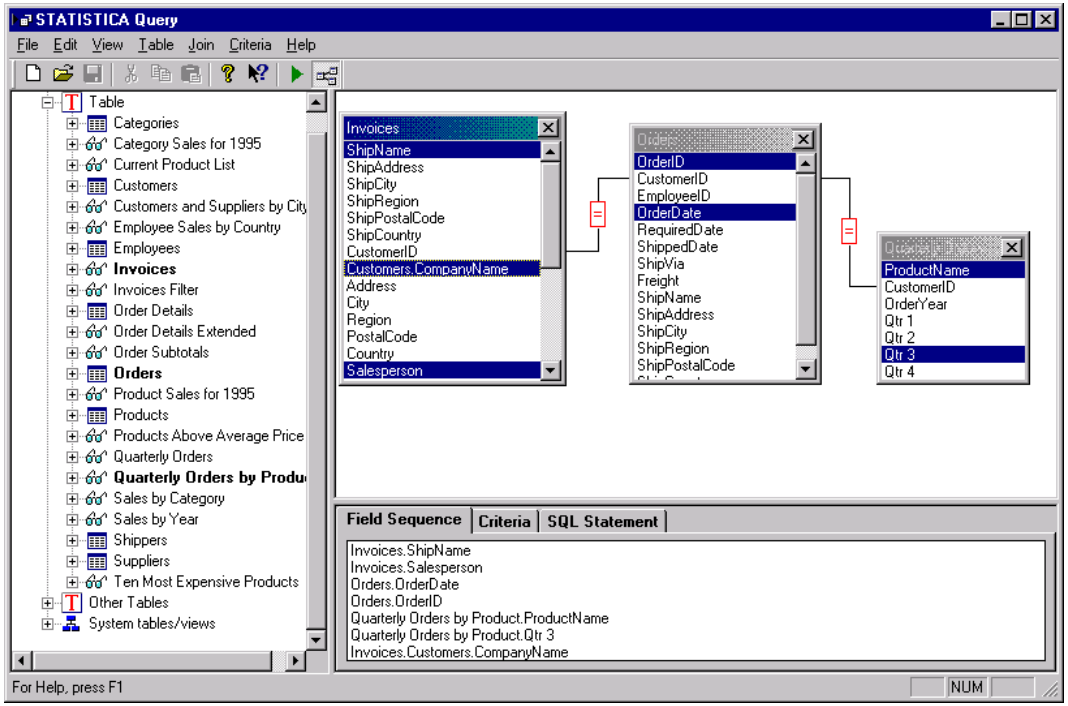




- Use the Clipboard
- *File – Open*

STATISTICA Query

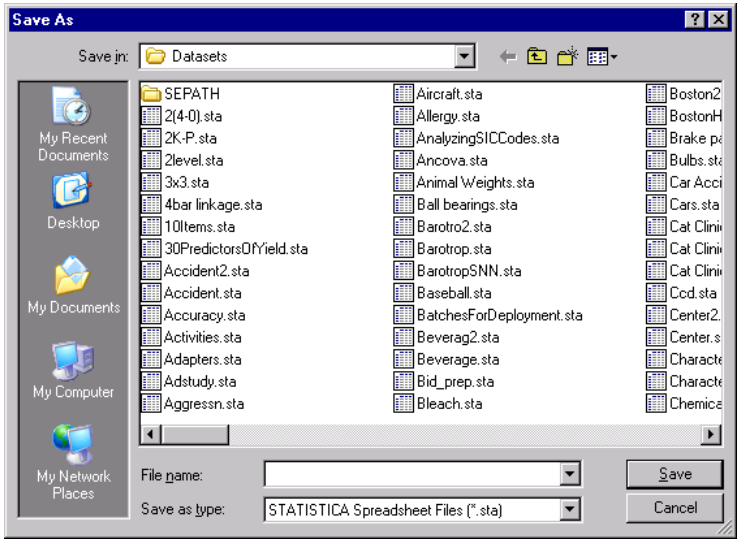
STATISTICA provides access to most databases (including many large system databases such as Oracle, Sybase, etc.) via *STATISTICA* Query



Exporting Data

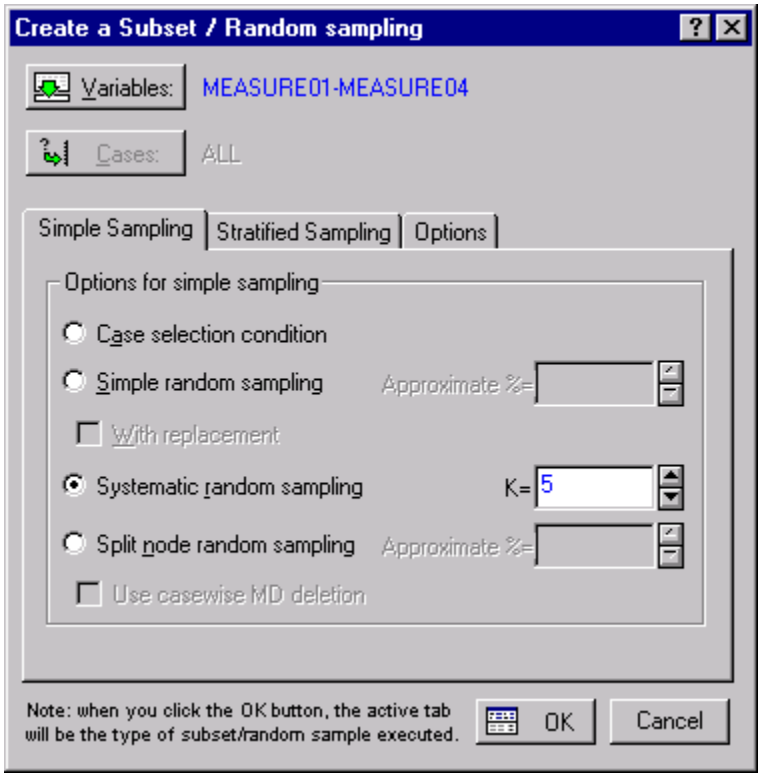
To export, select ***Save As*** from the ***File*** menu to display the ***Save As*** dialog.

A wide variety of files is available in the ***Save as type*** drop-down list (*STATISTICA* spreadsheet, Excel, SPSS (Data and Portable files), SAS (Data and Transport files), JMP, Minitab, dBASE, Text, HTML, Lotus Worksheets, Quattro Pro/DOS, and PDF).



Random Subsets

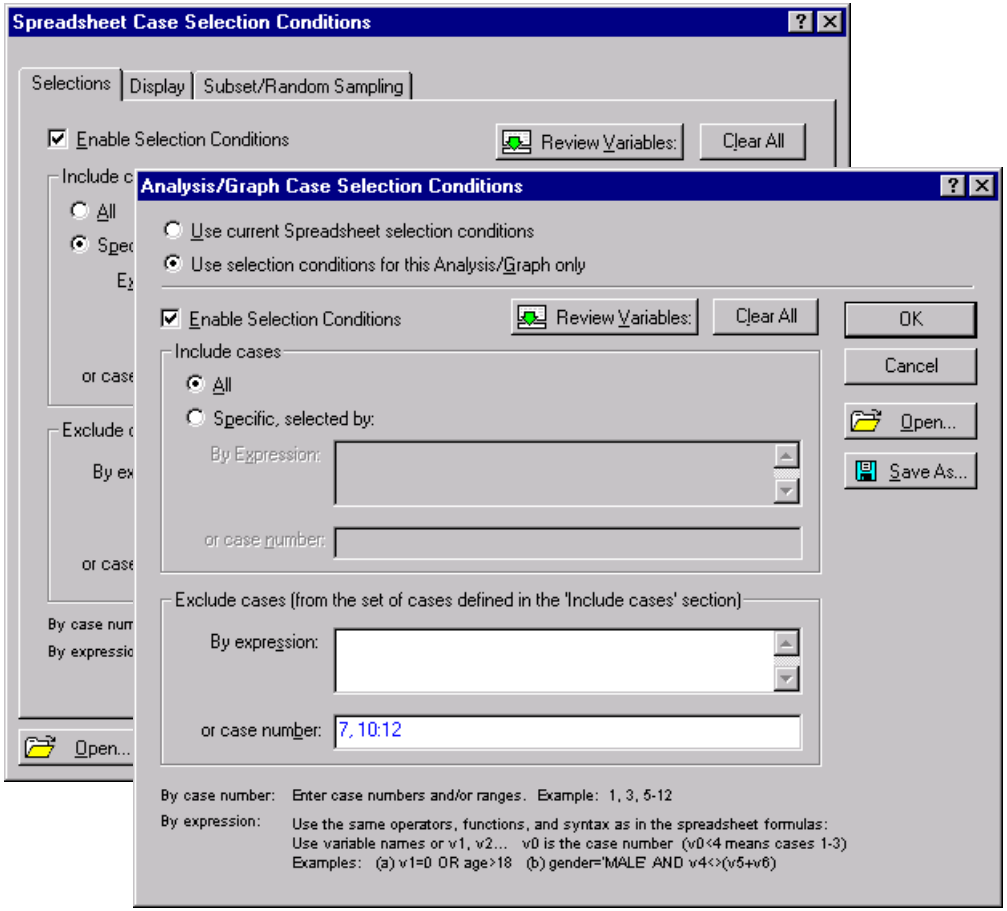
STATISTICA includes a facility that allows the user to randomly sample any dataset in order to create a subset dataset for analysis.



Case Selection Conditions

If you only want to analyze a specific subset of your data, you can use case selection conditions.

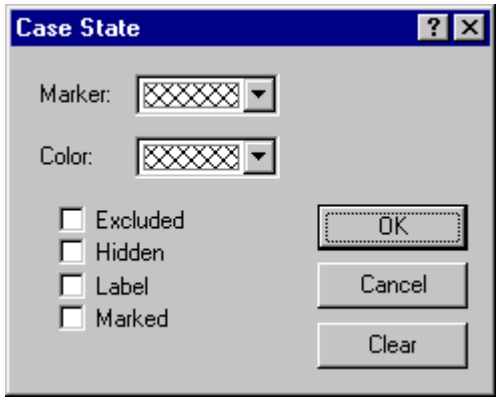
Cases can be selected in the spreadsheet or for a specific graph or analysis.



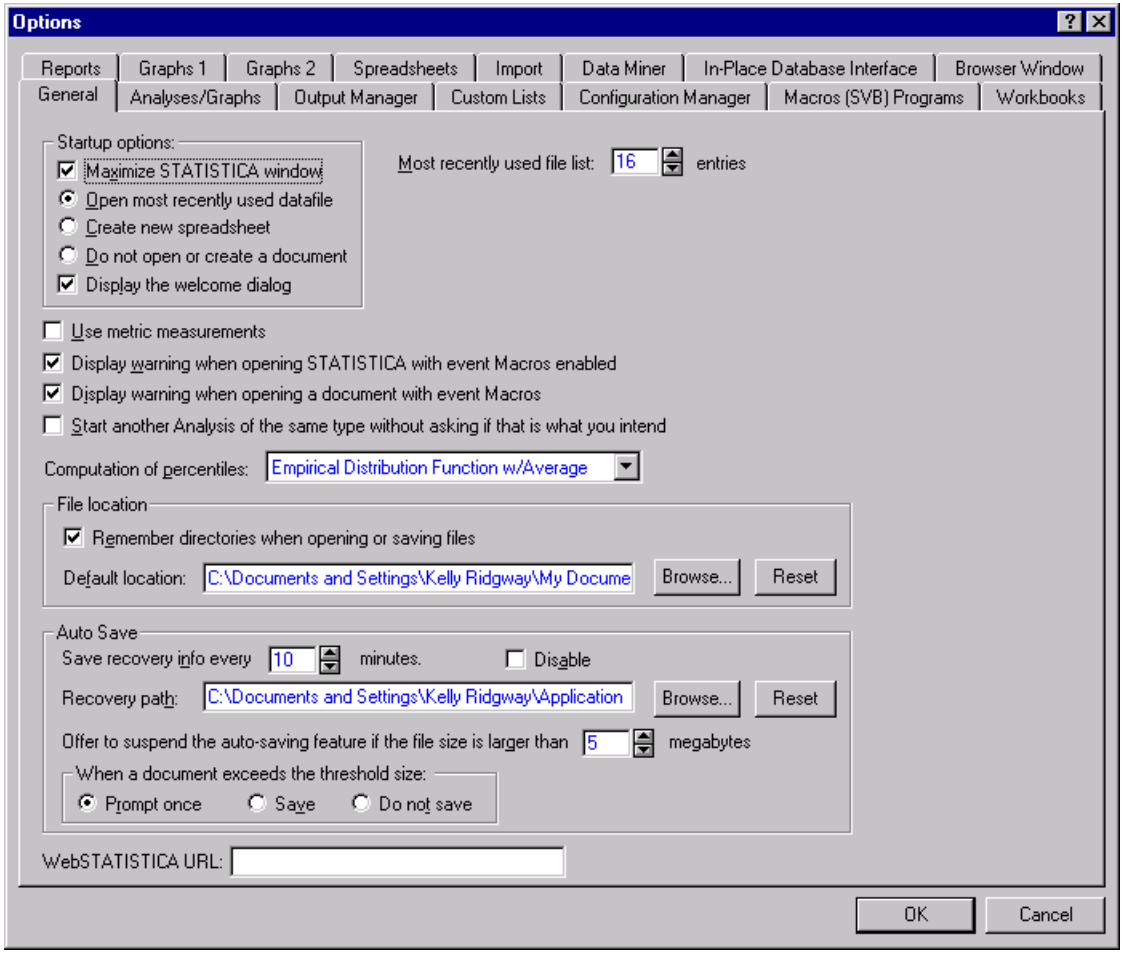
Case States

Assign case states to cases in order to customize the appearance of points in graphical displays

Assign unique point markers, or specify cases as *Excluded*, *Hidden*, *Labeled* or *Marked*.



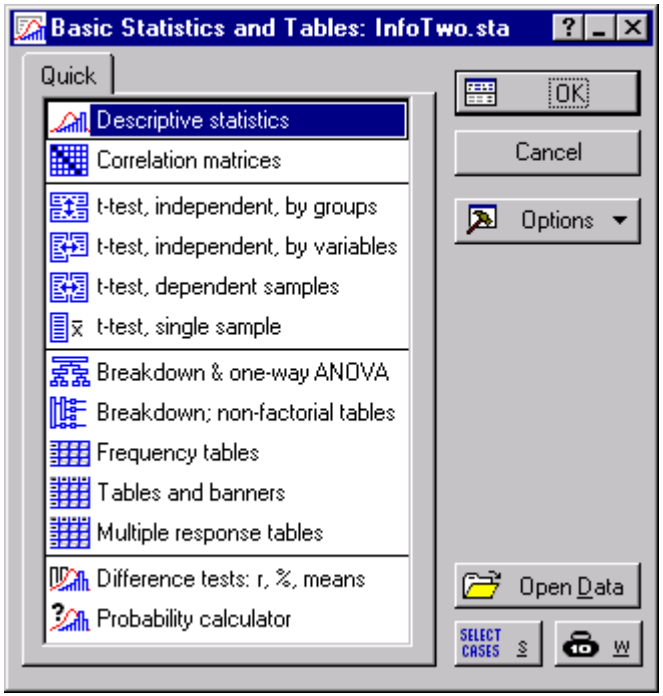
Options



Descriptive Statistics

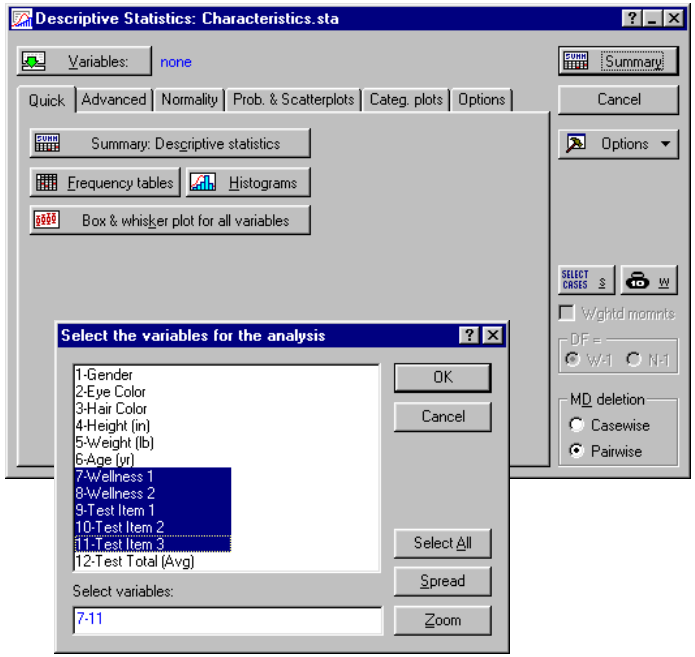
The *Basic Statistics and Tables* module gives you more depth and control over the output than statistics of block data.

Select ***Basic Statistics/Tables*** from the *Statistics* menu.



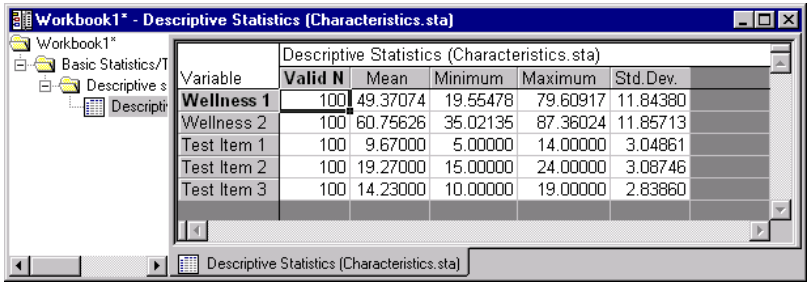
Specifying an Analysis

1. Use *Characteristics.sta*.
2. Select *Descriptive Statistics*
3. Click *Variables* and select the variables for the analysis

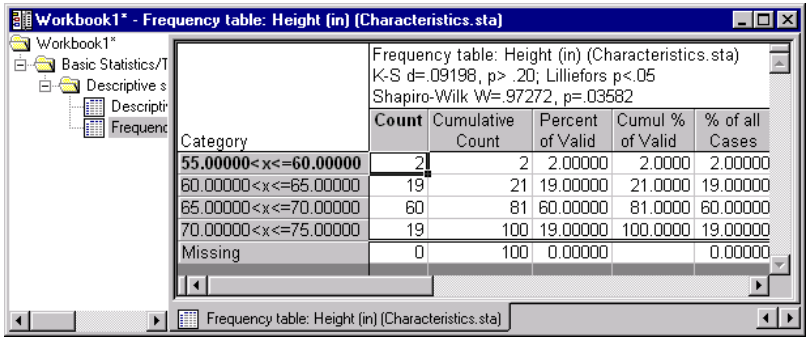


Reviewing Results

4. Click the ***Summary*** button to produce the results spreadsheet with the default selection of statistics



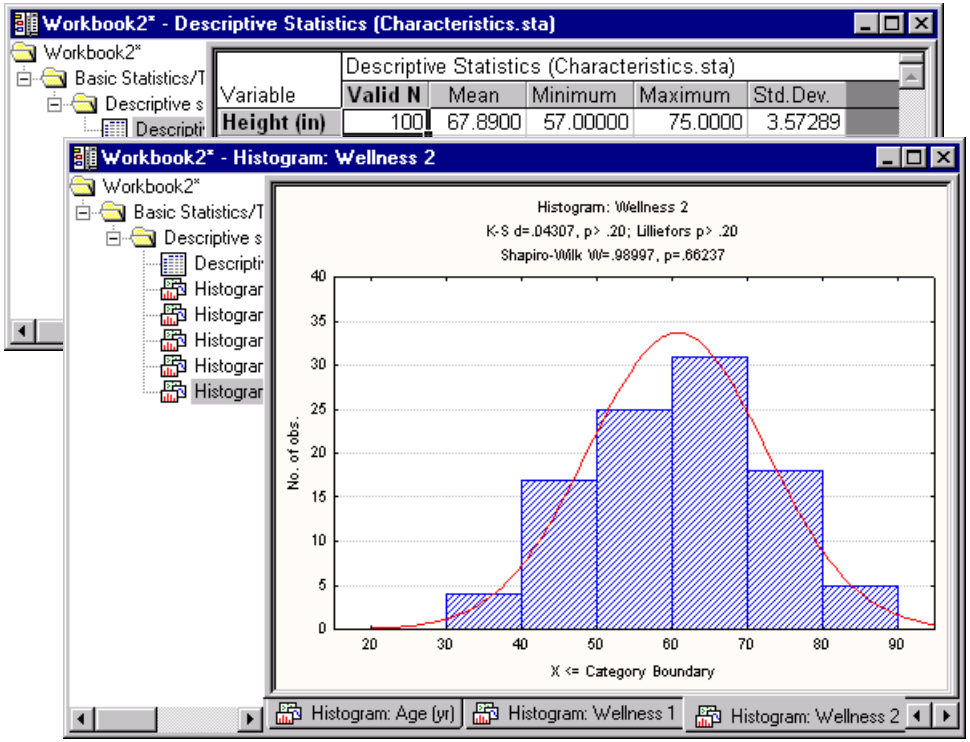
5. The ***Normality*** tab of the ***Descriptive Statistics*** dialog contains many of the most common tools for checking normality assumptions



Example 3: Descriptive Stats

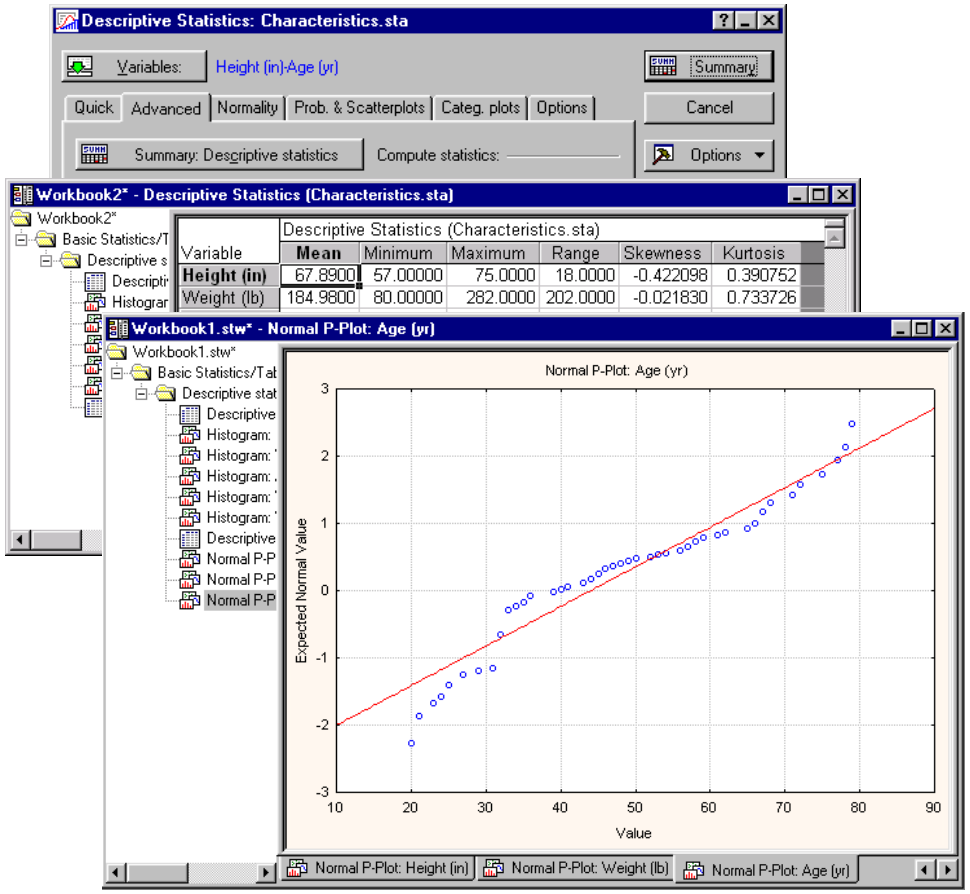
1. Open the *Characteristics.sta* data file.
2. From the *Statistics* menu, select *Basic Statistics/Tables*, then select *Descriptive Statistics*.

Example Steps



3. Click the ***Variables*** button and highlight variables *4-8.*
4. Click the ***Summary*** button.
5. Resume the analysis by clicking the ***Descriptive Statistics*** button on the *Analysis* bar, then click the ***Histograms*** button.

Example Steps



6. Resume the analysis and click on the ***Advanced*** tab. Select only the statistics shown at left.
7. Click the ***Summary: Descriptive statistics*** button
8. Resume the analysis and on the ***Prob.& Scatterplots*** tab, click the ***Normal probability plot*** button.

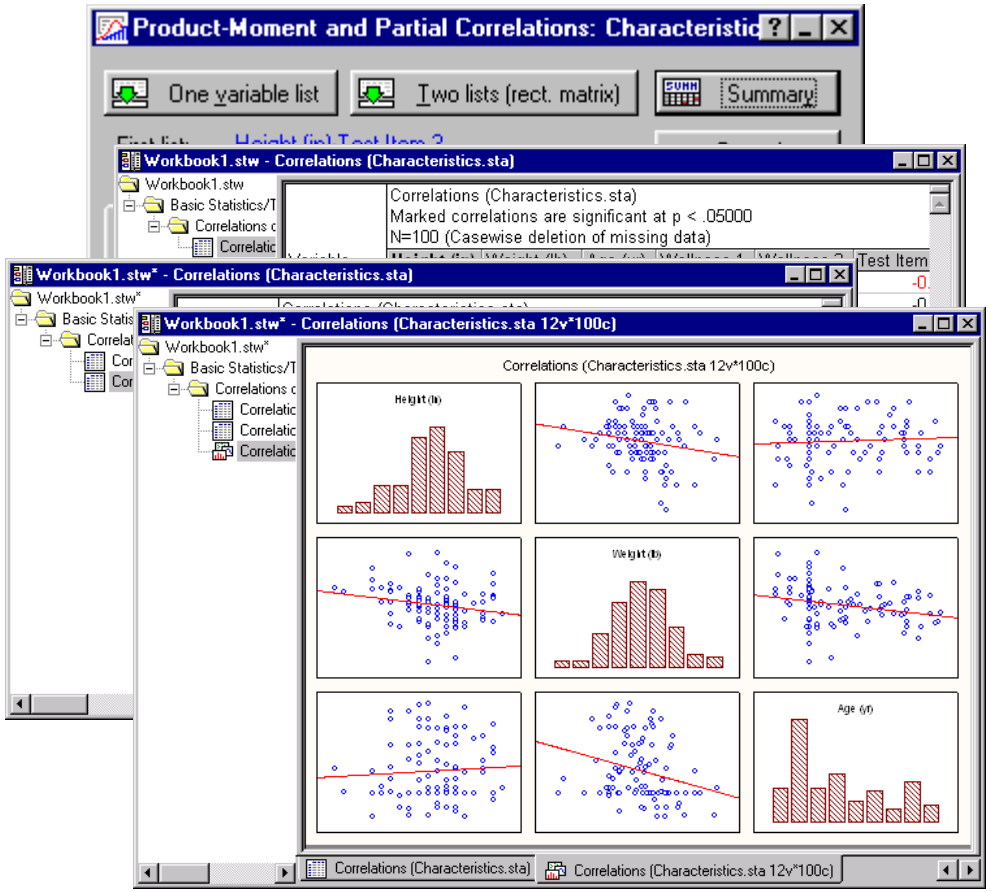
Correlations

Correlation is a measure of the relationship between two or more variables. The measurement scales used should be at least interval scales. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.

Example 4: Correlations

1. Continue using the *Characteristics.sta* data file. Select ***Basic Statistics/Tables*** from ***Statistics*** menu, then select ***Correlation matrices*** and click the ***OK*** button to display the ***Product-Moment and Partial Correlations*** dialog.

Example Steps



2. Click the ***One variable list*** button and select variables 4-11. Click ***OK.***
3. Click the ***Summary*** button.
4. Resume analysis. Click the ***Options*** tab. Select the ***Display r, p-levels, and N's*** option button. Click ***Summary***
5. Resume analysis. On the ***Quick*** tab, click ***Scatterplot matrix for variables***. Select variables *4-6* and click ***OK***.

t-tests

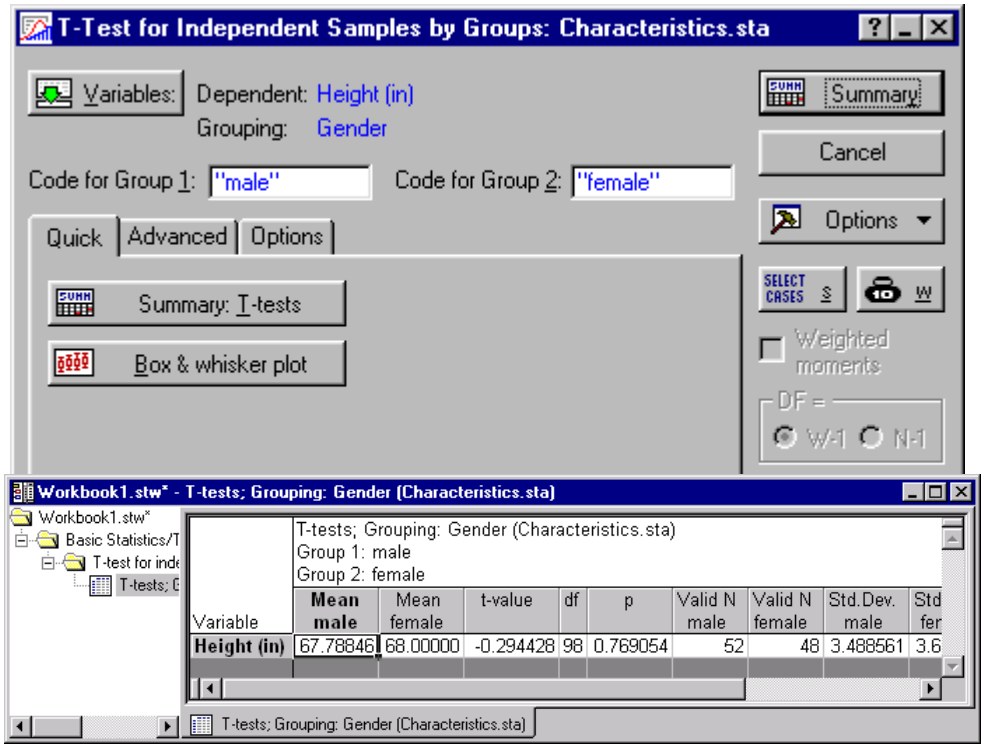
There are four different types of t-tests in the Basic Statistics and Tables module. The following examples illustrate when to use each test.

1. Independent, by groups
2. Independent, by variables
3. Dependent samples
4. Single sample

Example 5: t-test Independent, By Groups

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 Steps

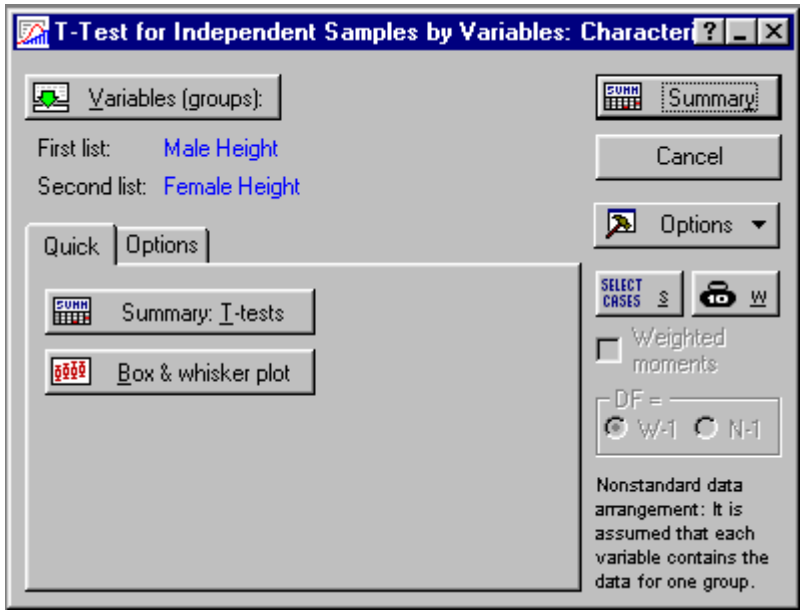


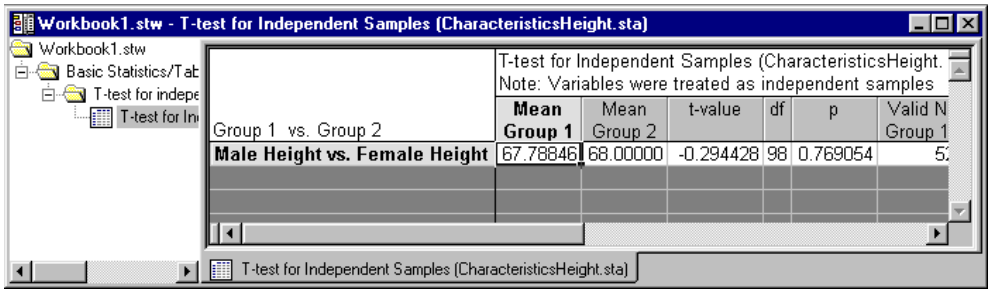
1. Use *Characteristics.sta*. Select ***Statistics-Basic Statistics/Tables,*** then, select ***t-test, independent, by groups*** and click ***OK***.
2. Click ***Variables***, select *Height (in)* as the dependent variable, and *Gender* as the grouping variable. Click ***OK***.
3. Click ***Summary***.

Example 6: t-test, Independent, By Variables

When the two groups to be compared reside in separate variables, it is more appropriate to choose a t-test by variables.

Example Steps



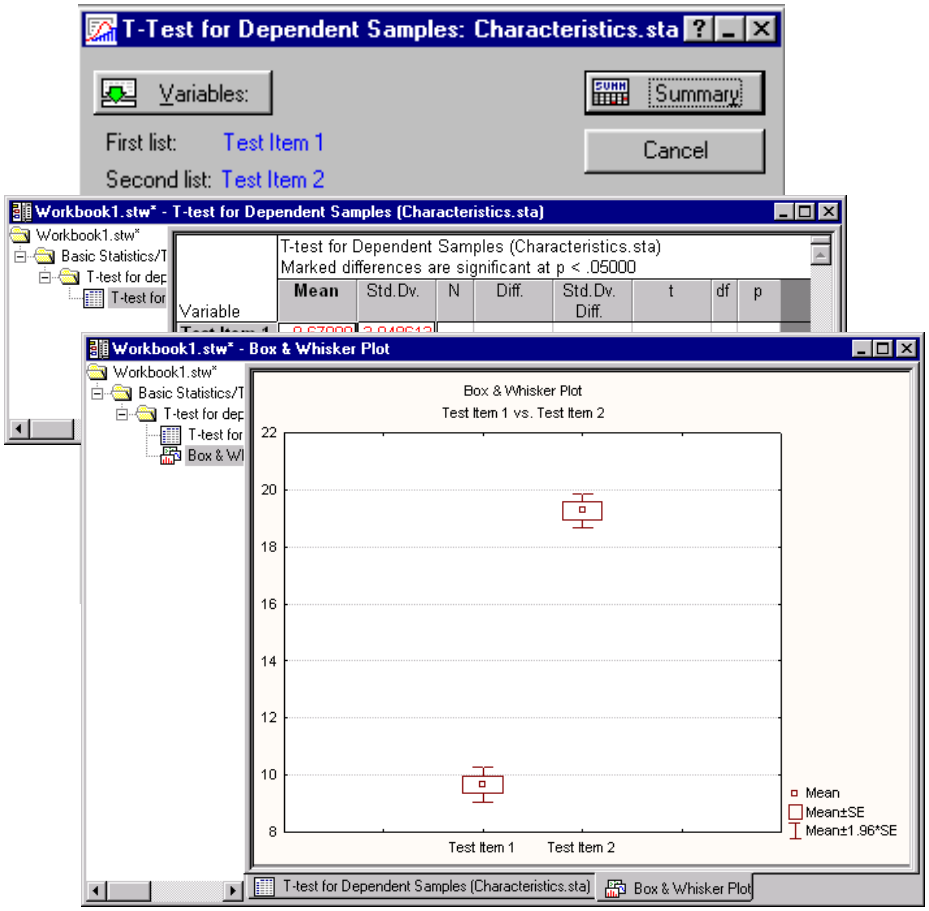


1. Use *CharacteristicsHeight.sta*.
2. On the ***Basic Statistics and Tables*** dialog, select ***t-test, independent, by variables*** and click ***OK***.
3. Select *Male Height* as the first variable and *Female Height* as the second variable.
4. Click ***Summary***.

Example 7: t-test, Dependent Samples

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.

Example Steps

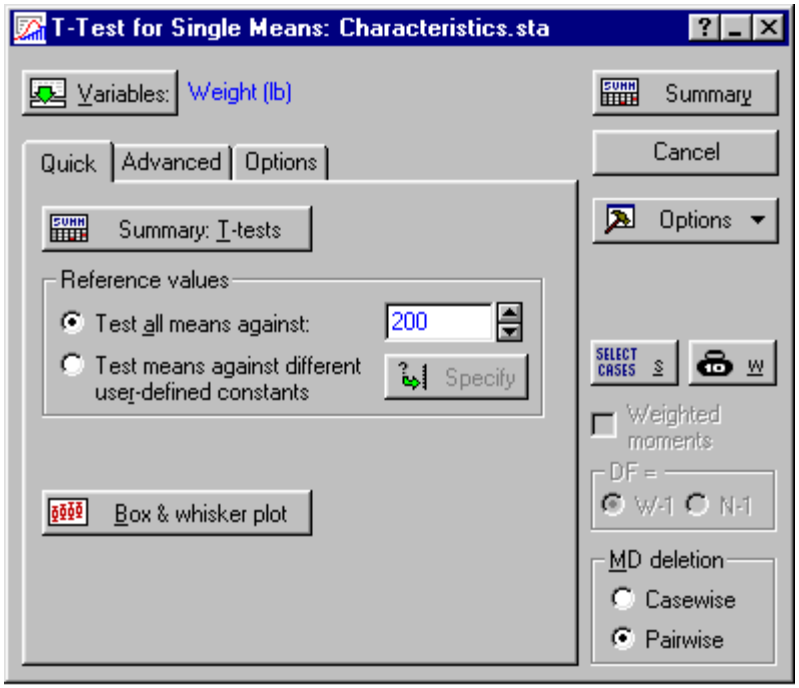


1. Use *Characteristics.sta*. On the ***Basic Statistics and Tables*** dialog, select ***t-test, dependent samples.***
2. Select *Test Item 1* as the first variable and *Test Item 2* as the second variable.
3. Click the ***Summary*** button.
4. Resume the analysis. Click the ***Box & whisker plots*** button.

Example 8: t-test, Single Sample

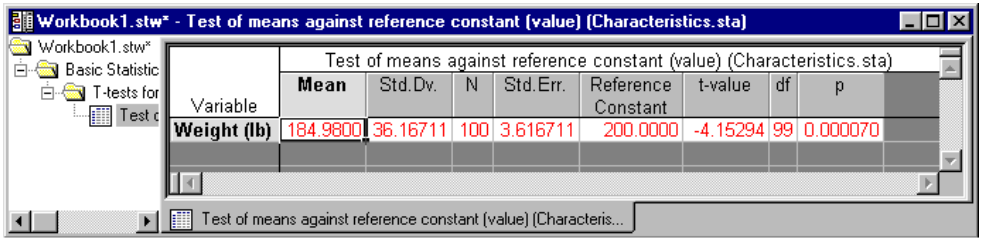
Using the single sample t-test, you can compare the mean of a particular variable to a specified value.

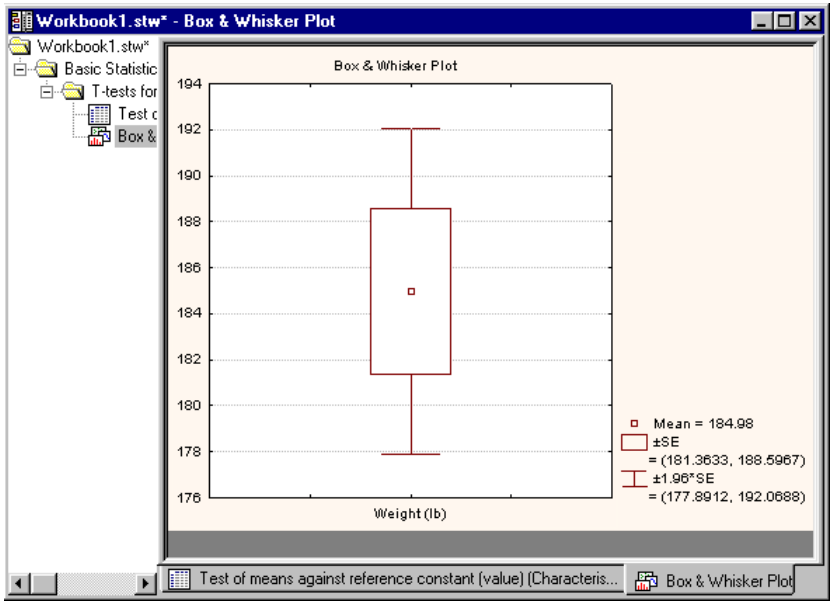
Example Steps



1. Use *Characteristics.sta*. On the ***Basic Statistics and Tables*** dialog, Select ***t-test, single sample.***
2. Use *Weight (lb)* as variable.
3. Select the ***Test all means against*** option button and enter *200* into the adjacent box.

Example Steps





4. Click the ***Summary*** button.
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***.

Breakdown and One-way ANOVA

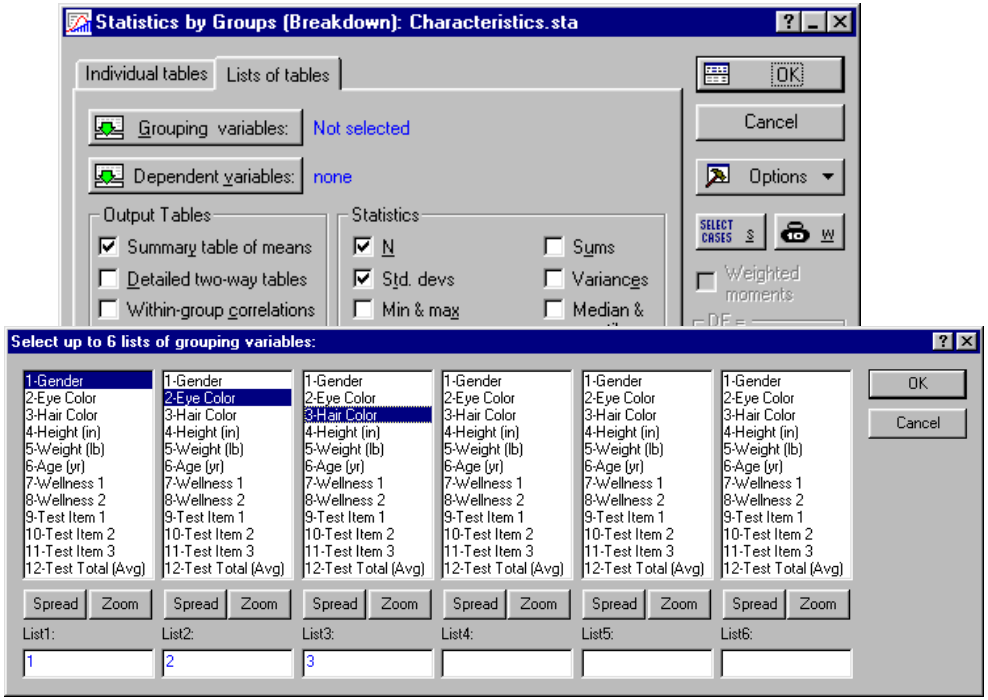
Typically used as an EDA Technique, breakdowns answer the question, “Are the groups different regarding the dependent variable?”

- Examine group means with one-way ANOVA.
- Investigate variation with homogeneity of variance tests
- Conduct a variety of Post-hoc tests

Example 9: Breakdowns and One-way ANOVA

1. Continue using the *Characteristics.sta* data file. Select ***Breakdown & One-way ANOVA*** from the ***Basic Statistics and Tables*** (Startup Panel), and click ***OK***.

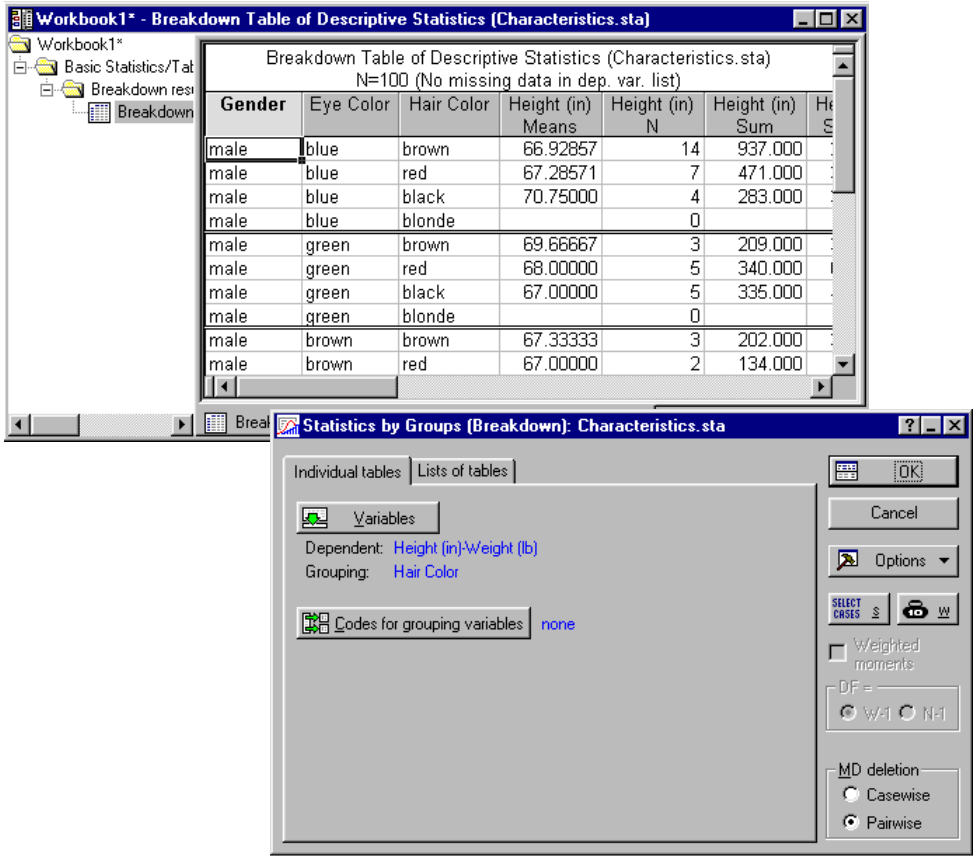
Example Steps



2. On the ***Statistics by Groups (Breakdown)*** dialog, select ***List of tables*** tab, and click ***Grouping variables***. For ***List1*** highlight *Gender*, for ***List2*** highlight *Eye Color*, and for ***List3*** highlight *Hair Color*.

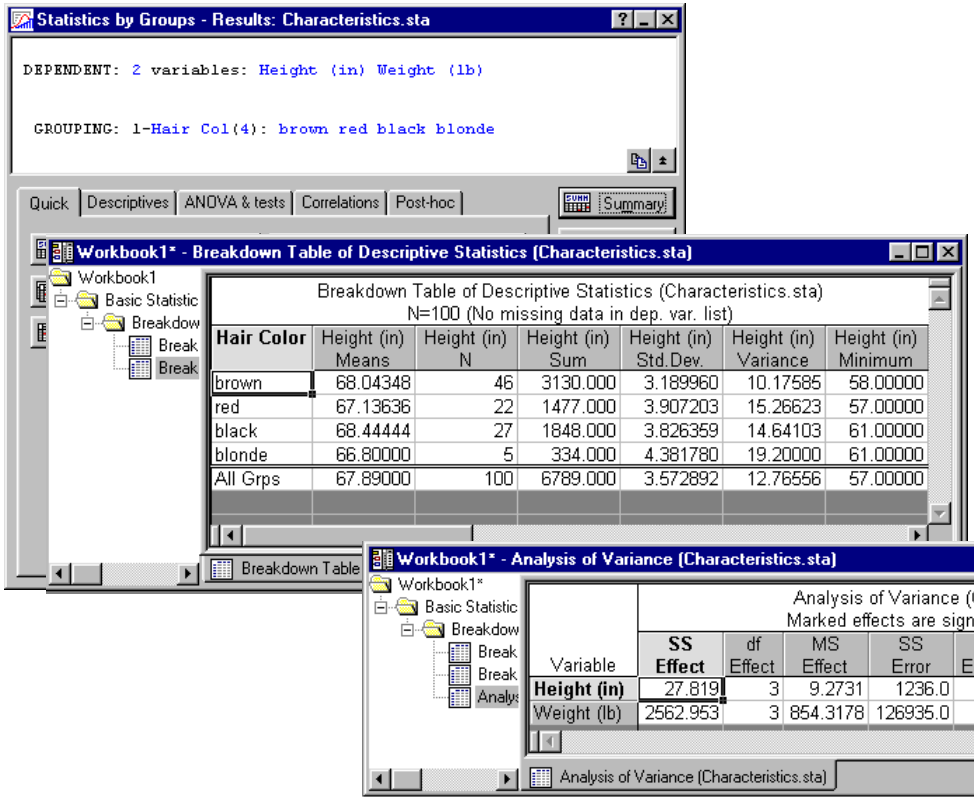
3. Click ***OK***. Click ***Dependent variables***. Highlight the *Height (in)* and *Weight (lb)* variables and click ***OK***.

Example Steps



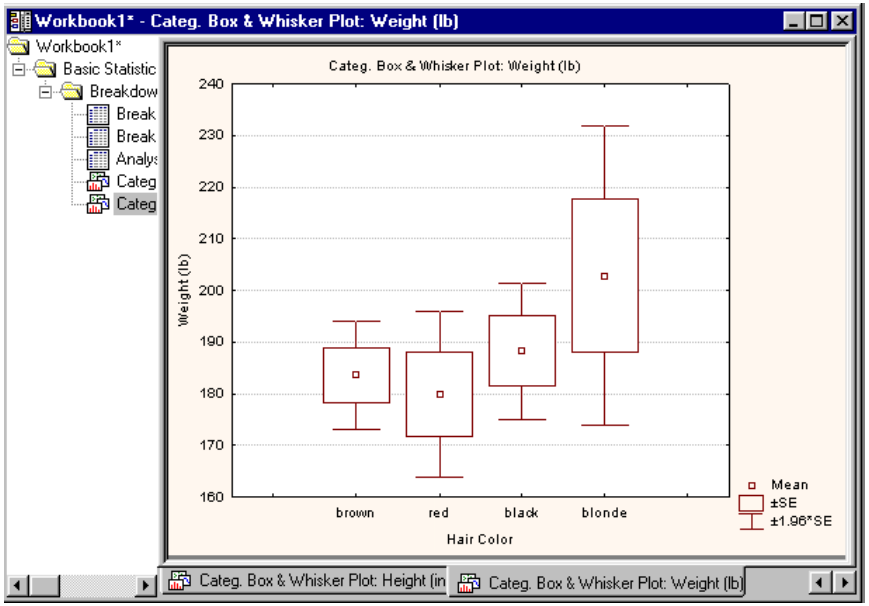
4. On the ***List of tables*** tab, select ***Sums***, ***Variances***, ***Std. devs***, ***Min & max*** and ***N*** in the ***Statistics*** section. Under ***Output*** tables, select ***Summary table of means***. Click ***OK***.
5. Resume analysis. Specify the ***Individual tables*** tab as shown. Click ***OK***.

Example Steps



6. Click ***OK*** to display results dialog.
7. Click the ***Summary*** button
8. Press CTRL+R on your keyboard, then click the ***Analysis of Variance*** button

Example Steps



9. Resume the analysis. Click the ***Categorized box & whisker plot*** button. Select both dependent variables and click ***OK***.

The ***ANOVA & tests*** tab and the ***Post-hoc*** tab are also available if you would like to further test for equal variances or to see which levels differ from each other (via post-hocs).

Nonparametric Methods

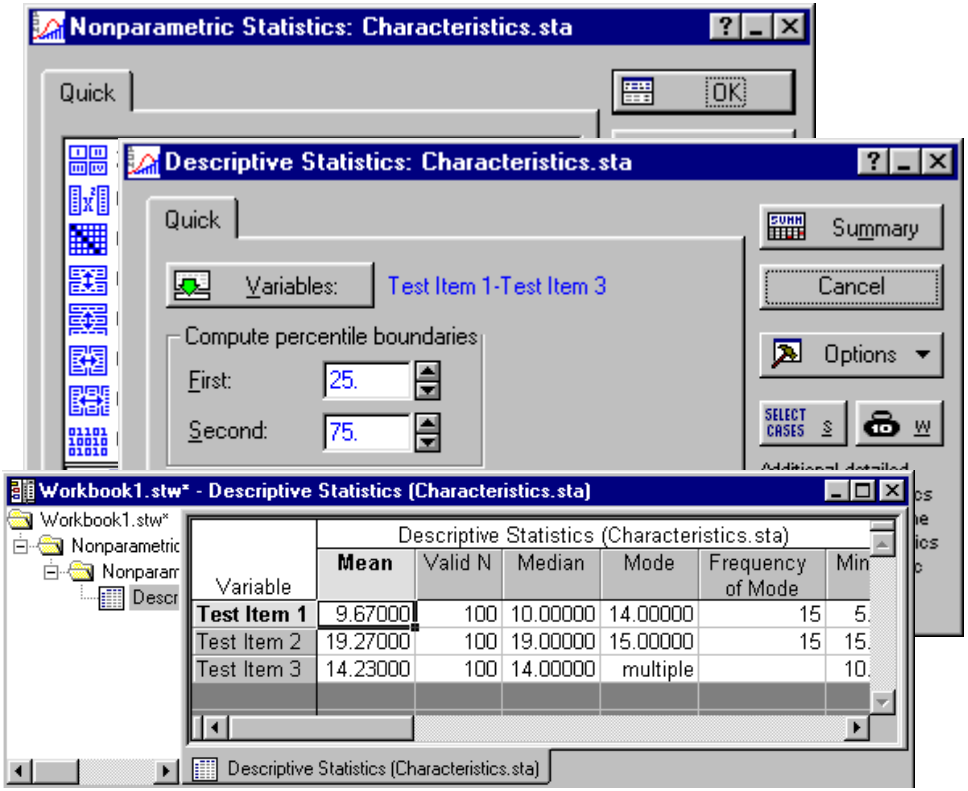
- Used when the researcher knows nothing about the parameters of the variable of interest in the population or in cases where the usual parametric assumptions do not hold
- Do not rely on the estimation of parameters

Example 10: Descriptives & Percentiles

Ordinal descriptive statistics computes a wide variety of measures of location (mean, median, mode, etc.) and dispersion (variance, average deviation, quartile range, etc.) to provide a more "complete picture" of your data

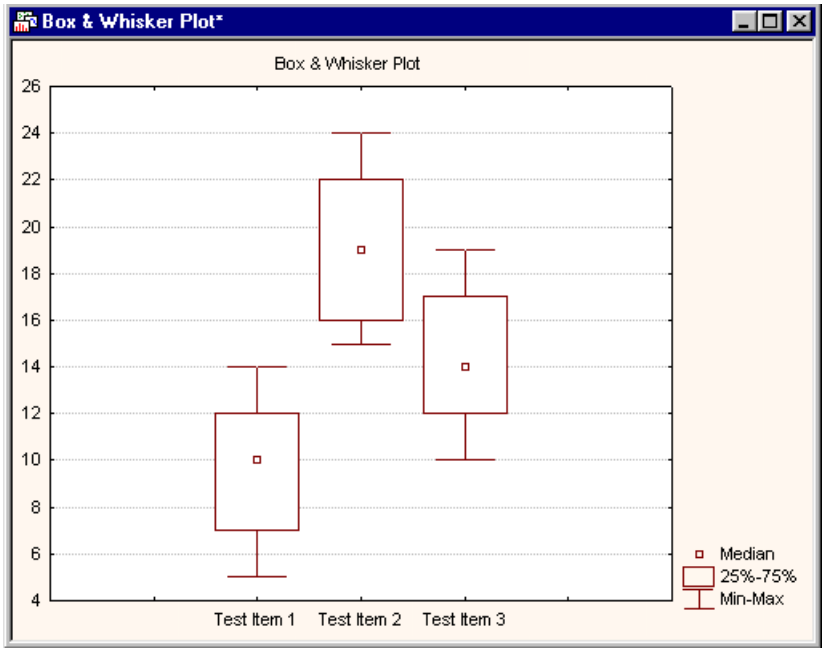
For this example, use *Characteristics.sta*.

Example Steps



1. From the ***Statistics*** menu, select ***Nonparametrics*** to display the ***Nonparametric Statistics*** (Startup Panel).
2. Choose ***Ordinal descriptive statistics (median, mode, …)*** and click ***OK***. On the ***Descriptive Statistics*** dialog, click ***Variables*** and select *Test Item 1*, *Test Item 2*, and *Test Item 3*. Click ***OK***.
3. Click the ***Summary*** button.

Example Steps



4. Press CTRL+R and click the ***Box & whisker plot for all variables*** button. On the ***Box-Whisker Type*** dialog, select the ***Median/Quart. /Range*** option button and click ***OK*** .

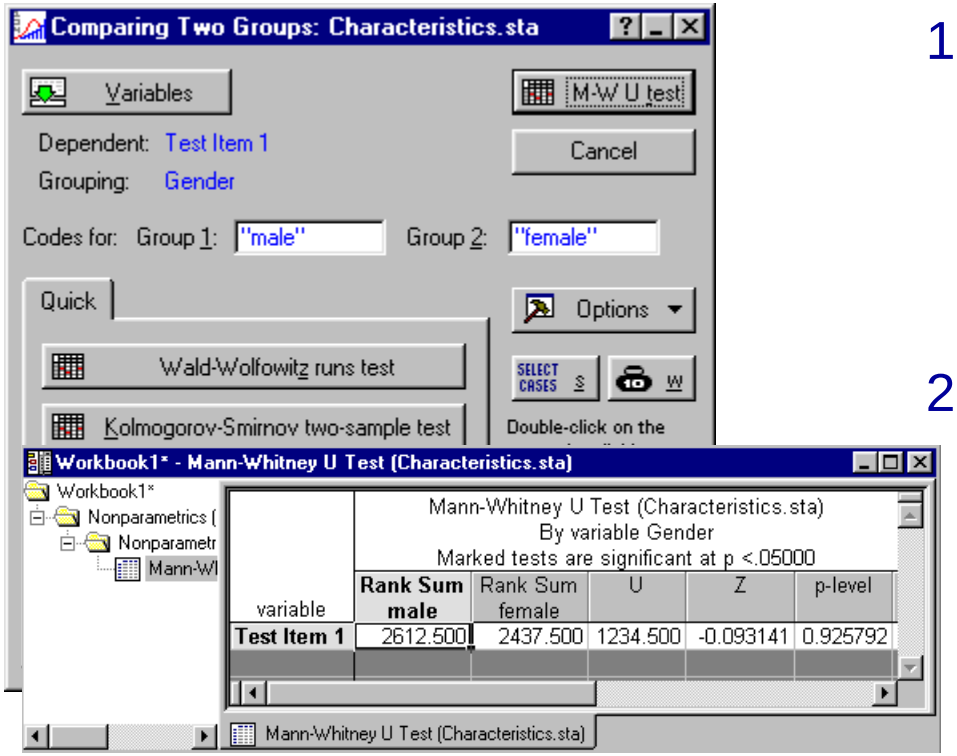
5. Resume the analysis. Enter *10* and *90* into the ***First*** and ***Second*** boxes, respectively, under ***Compute percentile boundaries***. Click ***Summary***.

Example 11: Comparing 2 Independent Samples (Groups)

Nonparametric alternatives for the t-test for independent samples available in *STATISTICA* are the Wald-Wolfowitz runs test, the Mann-Whitney U test, and the Kolmogorov-Smirnov two sample test.

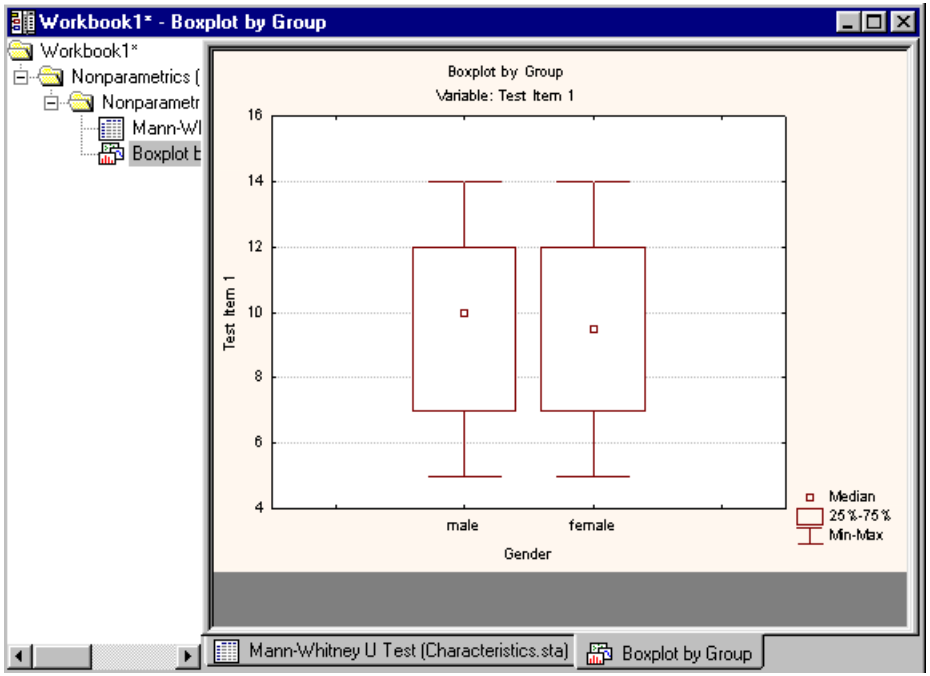
For this example, use *Characteristics.sta*.

Example Steps



1. From the ***Statistics*** menu, select ***Nonparametrics***, then choose ***Comparing two independent samples (groups)*** and click ***OK***.
2. Click the ***Variables*** button and highlight *Test Item 1* in the ***Dependent variable list*** box and ***Gender*** in the ***Indep. (grouping) variable*** box. Click ***OK***.
3. Click the ***M-W U test*** button.

Example Steps



4. Now, look at a box plot to visualize the results. To do this, resume the analysis. Click the ***Box & Whisker plot by group*** button on the ***Comparing Two Groups*** dialog.

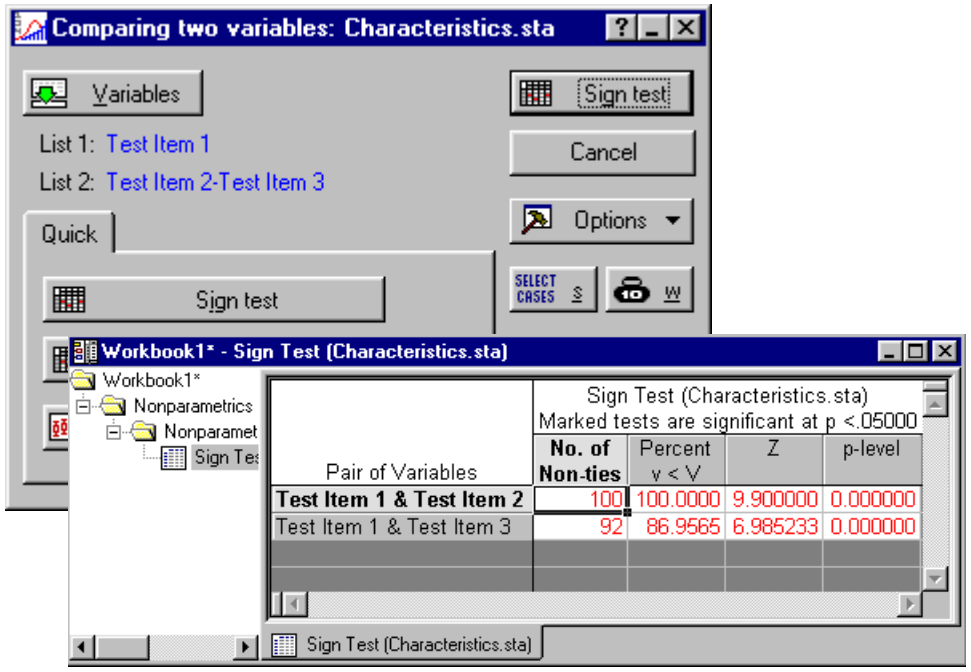
The plot confirms our conclusion that there is little difference between males and females in respect to *Test Item 1*.

Example 12: Comparing Two Dependent Samples

When the data are nonparametric, use the Sign test or Wilcoxon's matched pairs test to compare two dependent samples.

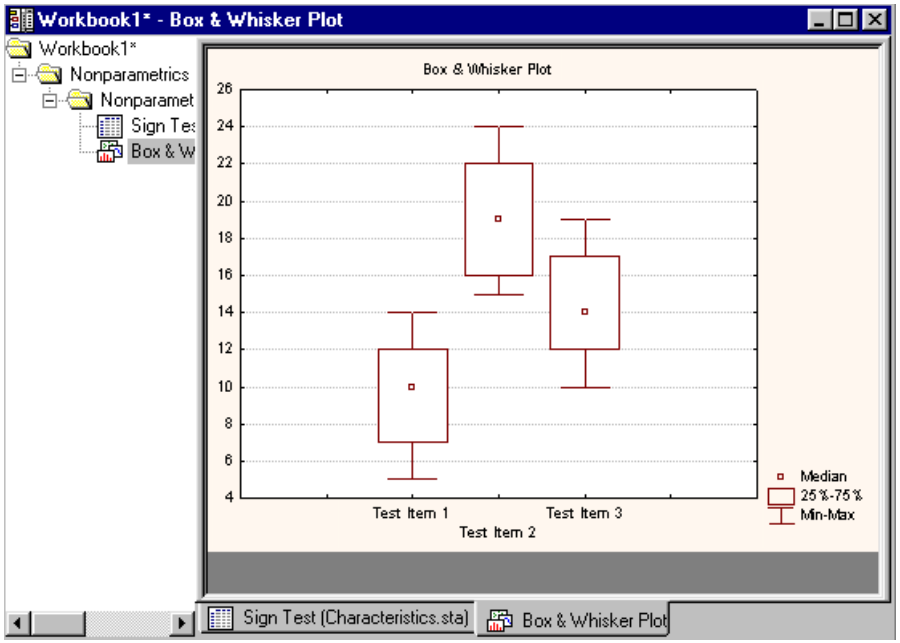
For this example, use *Characteristics.sta*.

Example Steps



1. Select ***Nonparametrics*** from the ***Statistics*** menu.
2. Choose ***Comparing two dependent samples (variables)*** and click ***OK***.
3. Select the variable *Test Item 1* in the ***First variable list*** box and select both *Test Item 2* and *Test Item 3* in the ***Second variable list*** box.
4. Click the ***Sign test*** button.

Example Steps



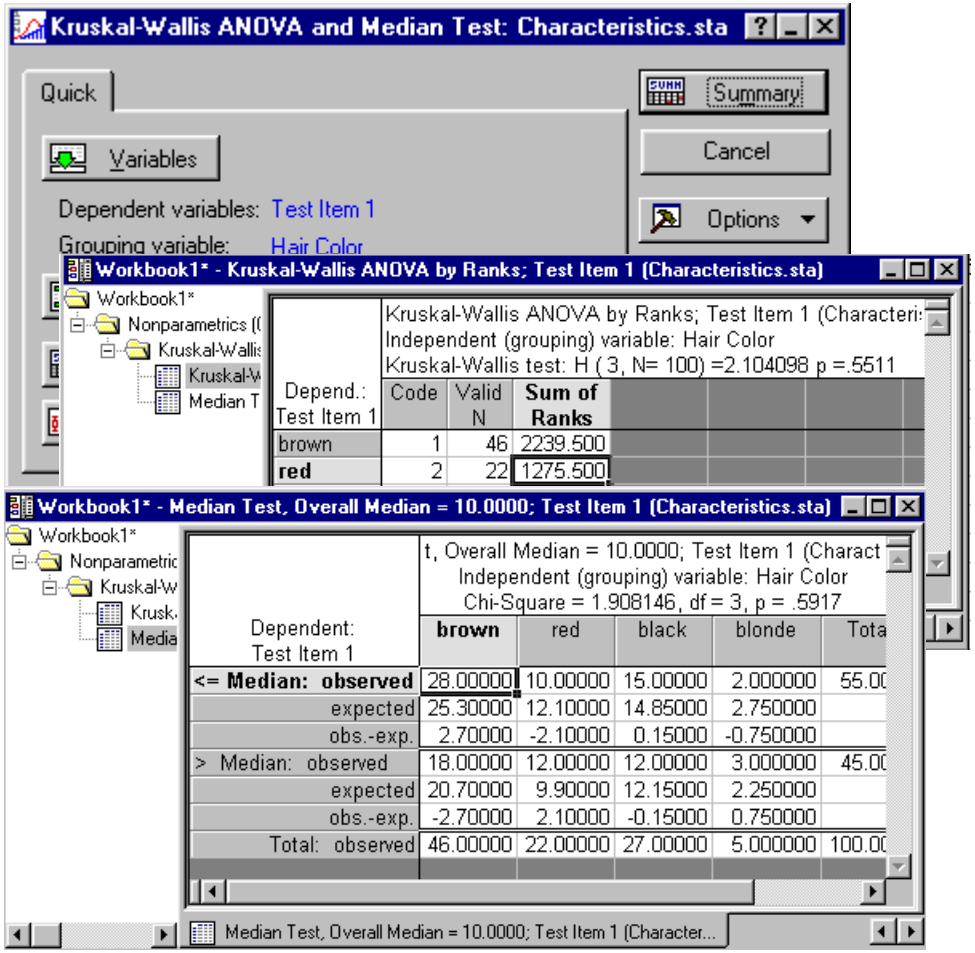
5. Resume the analysis and create a box plot to graphically visualize the results. Click the ***Box & whisker plots for all variables*** button. Select all three variables and click ***OK***. Select the ***Median/Quart./ Range*** option button on the ***Box-Whisker Type*** dialog, and click ***OK*** to produce the graph.

Example 13: Comparing Multiple Independent Samples

The nonparametric equivalents to the one-way ANOVA method are the Kruskal-Wallis analysis of ranks and the Median test.

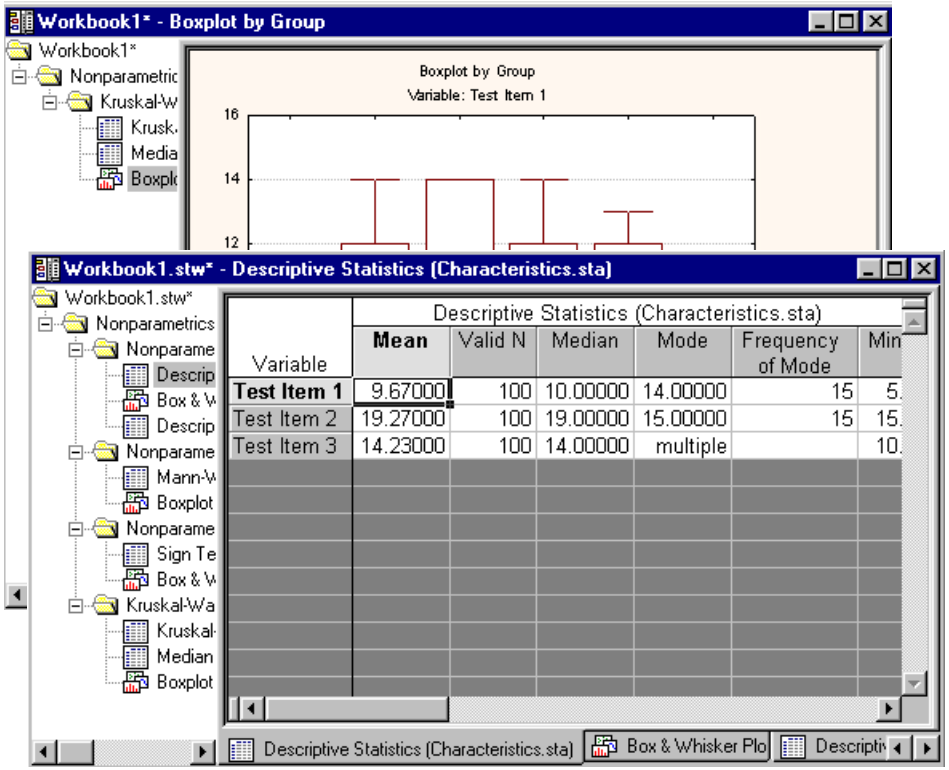
For this example, use *Characteristics.sta*.

Example Steps



1. Select ***Nonparametrics*** from the ***Statistics*** menu.
2. Choose ***Comparing multiple indep. samples (groups)***, then click the ***Variables*** button and select *Test Item 1* as the ***Dependent variable list***, and *Hair Color* as the ***Indep. (grouping) variable***. Click ***OK***.
3. Click ***Summary***. The ***Median test*** results are displayed next in the workbook

Example Steps



4. Now, look at a box plot to graphically visualize the results. This graph is available via the ***Box & whisker*** button on the ***Kruskal-Wallis ANOVA and Median Test*** dialog.

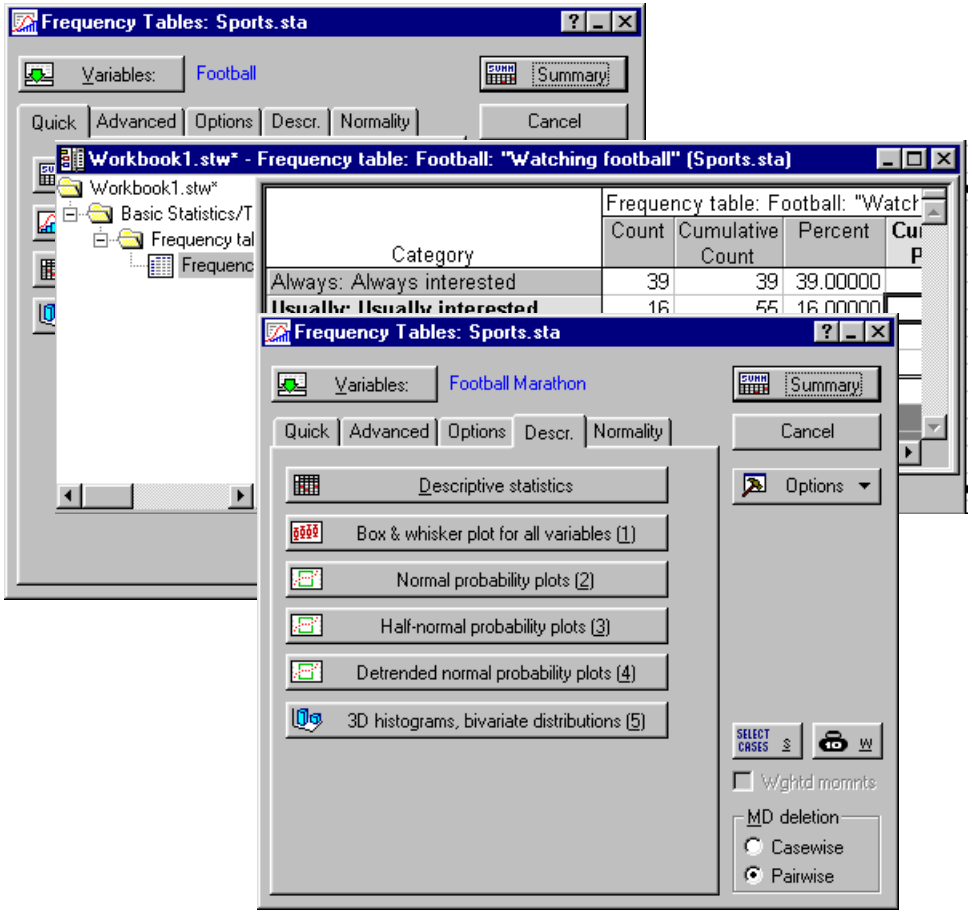
Notice that each analysis in this workbook was a new analysis. If you placed all of the documents in the same workbook, your workbook could look similar to this.

Example 14: Frequency Tables

Frequency or one-way tables represent the simplest method for analyzing categorical (nominal) data. They are often used as an exploratory procedure to review how different categories of values are distributed in the sample.

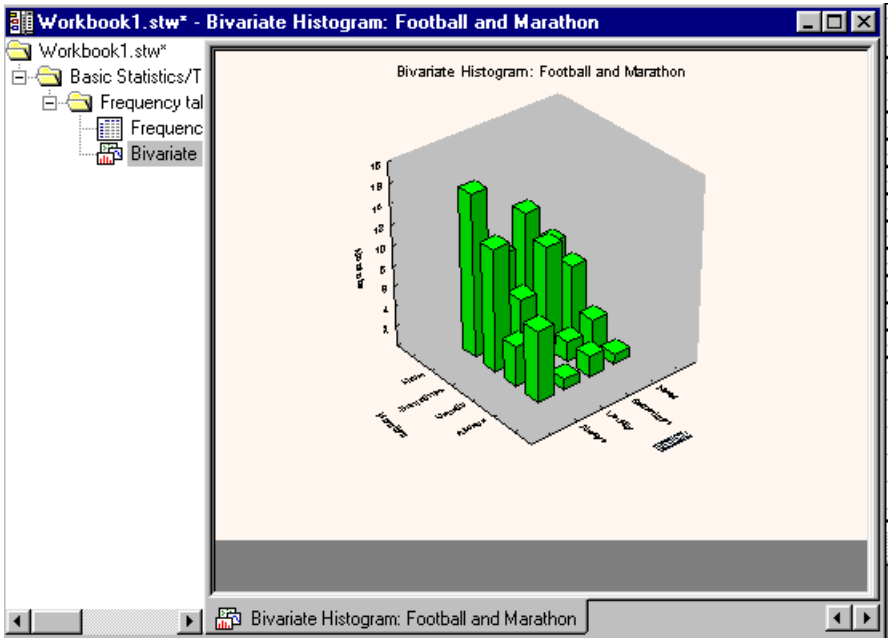
For this example, use *Sports.sta*.

Example Steps



1. Choose ***Frequency tables*** from the ***Basic Statistics and Tables*** (Startup Panel). Select *Football* as the ***Variable***, and click ***OK***.
2. Click the ***Summary*** button
3. Press CTRL+R and select Football and Marathon.
4. Return to the ***Frequency Tables*** dialog and select the ***Descr.*** tab.

Example Steps



5. Click the ***3D histograms, bivariate distributions*** button. Select *Football* for the first variable list and *Marathon* for the second variable list.

   Click ***OK***.

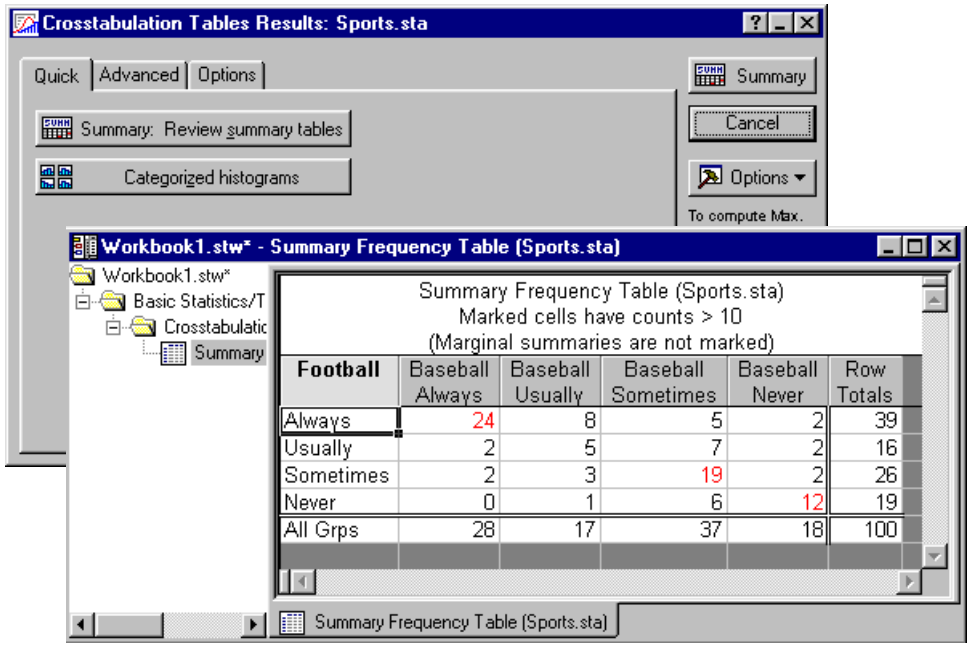
From this graph, you can see that regarding the two sports selected, the largest group of men said that they never watch marathons on television and always watch football.

Example 15: Crosstabulation Tables

Crosstabulation allows us to examine frequencies of observations that belong to specific categories on more than one variable. By examining these frequencies, we can identify relationships between variables.

For this example, use *Sports.sta*.

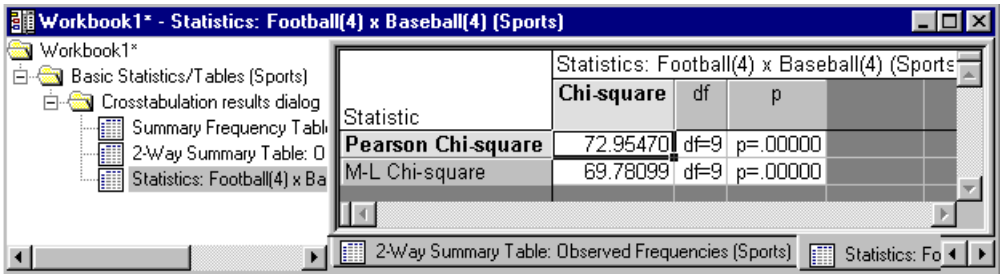
Example Steps

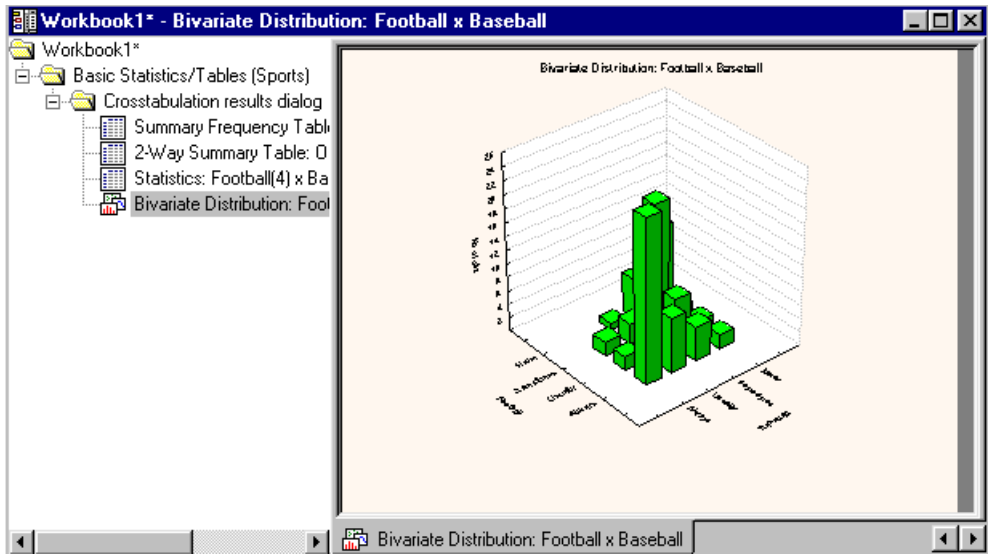


Football	Baseball Always	Baseball Usually	Baseball Sometimes	Baseball Never	Row Totals
Always	24	8	5	2	39
Usually	2	5	7	2	16
Sometimes	2	3	19	2	26
Never	0	1	6	12	19
All Grps	28	17	37	18	100

1. Choose ***Tables and banners*** from the ***Startup Panel*** and click ***OK***.
2. On the ***Crosstabulation*** tab, click the ***Specify tables (select variables)***, and select *Football* in ***List1*** and *Baseball* in ***List2***. Click ***OK***. Click ***OK*** on the ***Crosstabulation Tables*** dialog.
3. On the ***Crosstabultaion Tables Results*** dialog, click ***Summary***.

Example Steps

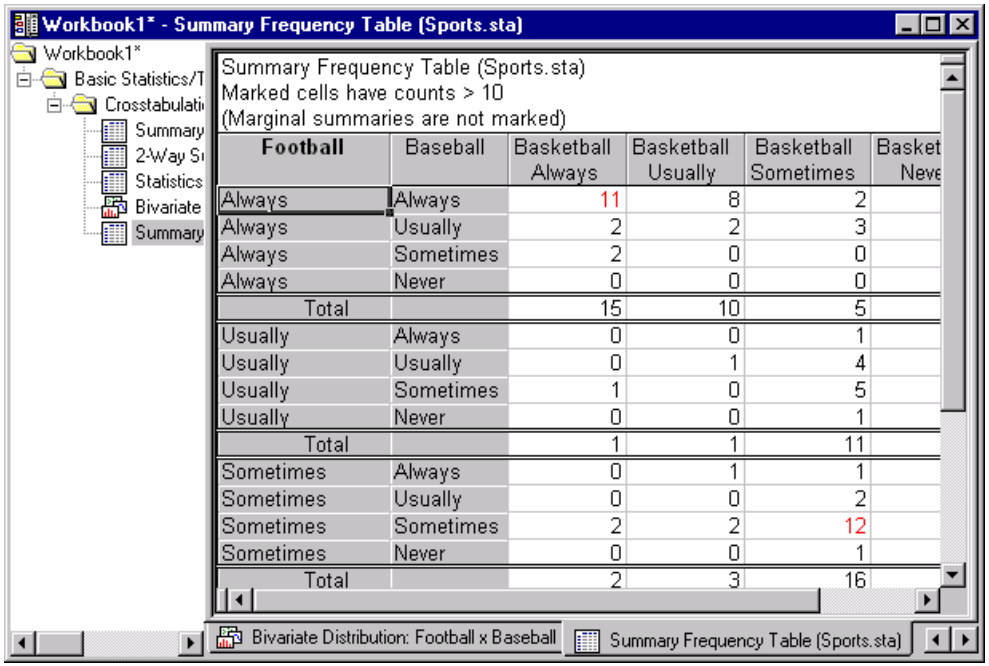




4. On the ***Options*** tab, select the ***Pearson & M-L Chi-square*** check box. On the ***Advanced*** tab, click ***Detailed two-way tables***.

5. Press CTRL+R, then click ***3D histograms*** on the ***Advanced*** tab.

Example Steps



Football	Baseball	Basketball Always	Basketball Usually	Basketball Sometimes	Basket Neve
Always	Always	11	8	2	
Always	Usually	2	2	3	
Always	Sometimes	2	0	0	
Always	Never	0	0	0	
Total		15	10	5	
Usually	Always	0	0	1	
Usually	Usually	0	1	4	
Usually	Sometimes	1	0	5	
Usually	Never	0	0	1	
Total		1	1	11	
Sometimes	Always	0	1	1	
Sometimes	Usually	0	0	2	
Sometimes	Sometimes	2	2	12	
Sometimes	Never	0	0	1	
Total		2	3	16	

6. Press CTRL+R, then cancel to return to ***Crosstabulation Tables*** dialog. Specify new variables: *Football* on ***List1***, *Baseball* on ***List2***, and *Basketball* on ***List3***.

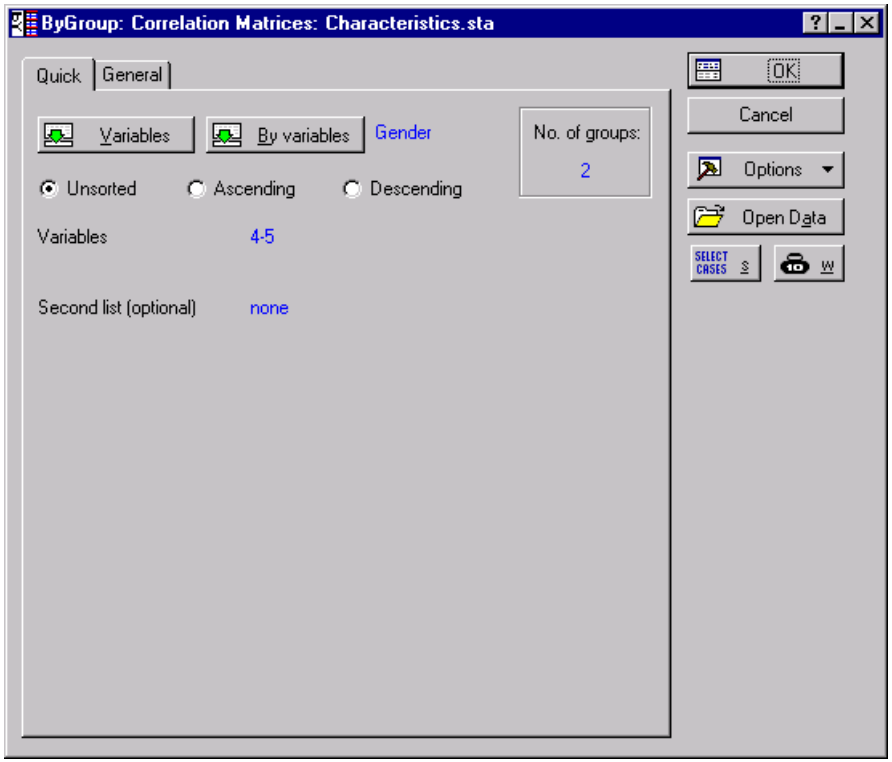
7. Click ***OK*** to run the analysis. Click ***Summary***. In this survey, 12 men said they watch baseball sometimes, football sometimes, and basketball sometimes.

Example 16: ByGroup Analysis

Use ***ByGroup*** analysis to do the wide variety of analyses and graphs in *STATISTICA* while taking a grouping variable into consideration.

For this example, use *Characteristics.sta*.

Example Steps

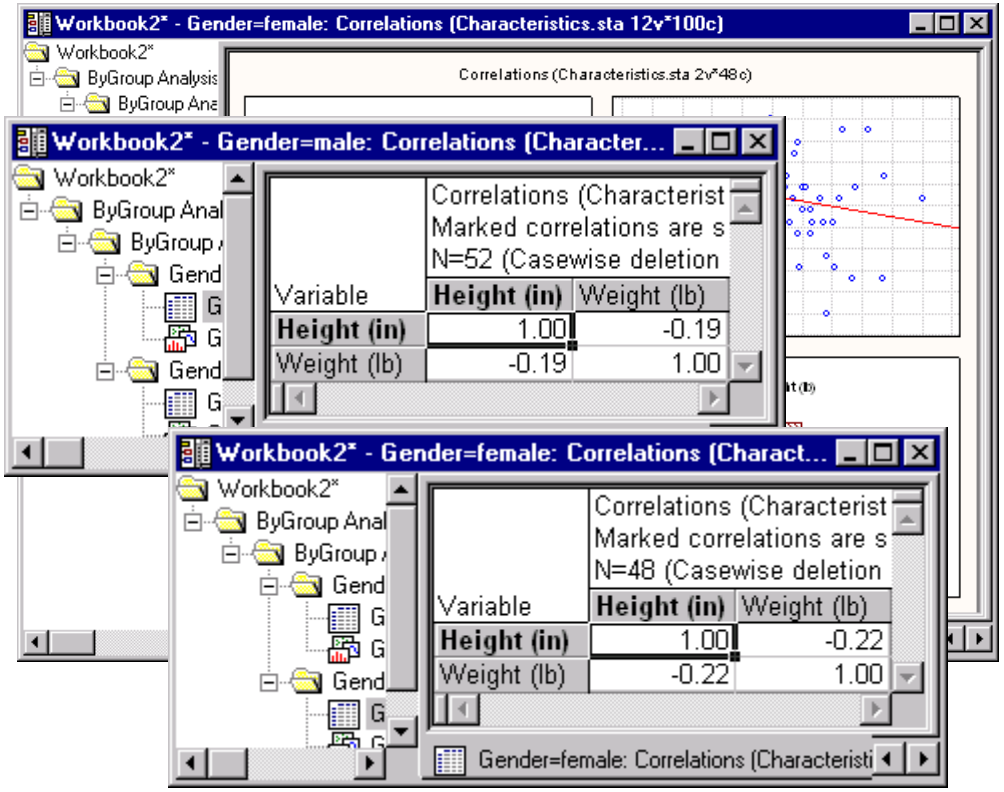


1. Select ***Statistics - ByGroup Analysis***. In the ***ByGroup Statistics Browser*** select *Basic Statistics and Tables* in the left pane and *Correlation Matrices* in the right pane.

   Click ***OK***.

2. On the ***Quick*** tab, click ***Variables*** and select *Height(in)*, *Weight(lb)*, and *Age(yr)* in the first list; click ***OK***. Click ***By Variables*** and select *Gender*. Click ***OK***.

Example Steps



3. Select the ***General*** tab and select ***All results*** for the ***Detail of computed results reported***.
4. Click ***OK*** to display the results.

When you compare the results for *Males* and *Females*, you can see that only the correlation between *Weight* and *Age* for *Males* was significant.