Multiple Violin Plot Statlet

statgraphics 18®
centurion

Revised: 10/11/2017

Summary

The *Multiple Violin Plot Statlet* displays data for 2 or more quantitative samples using a combination of a box-and-whisker plot and a nonparametric density estimator. It is very useful for visualizing the shape of the probability density function for the populations from which the data came.

Sample StatFolio: *multiple violinplot.sgp*

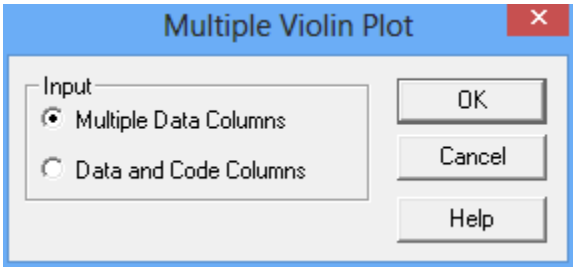
Sample Data

The file *iris.sgd* contains a famous set of data from Fisher (1936). The data consist of a total of $n = 150$ irises, 50 from each of $g = 3$ different species: *setosa*, *versicolor*, and *virginica*. Measurements were made on $p = 4$ variables, describing the length and width of the sepal and petal. The table below shows a partial list of the data in that file:

Sample	*Sepal length*	*Sepal width*	*Petal length*	*Petal width*	*Species*
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
…	…	…	…	…	…

Data Input

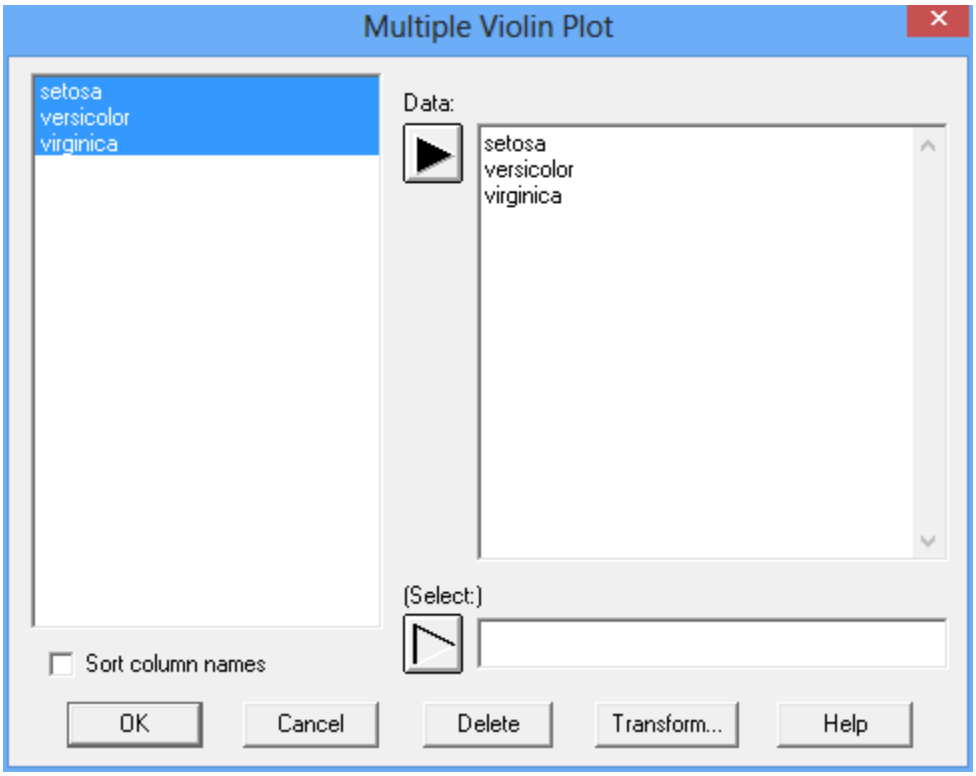
When the procedure is selected from the main menu, the first dialog box displayed asks you to specify the format in which the data have been entered:



- **Multiple Data Columns**: indicates that each sample has been placed into a separate column.
- **Data and Code Columns**: indicates that all observations have been placed into a single column, with a second column indicating which sample each observation belongs to.

Multiple Data Columns

If the data have been placed in separate columns for each sample, the column names must be entered on the second dialog box:

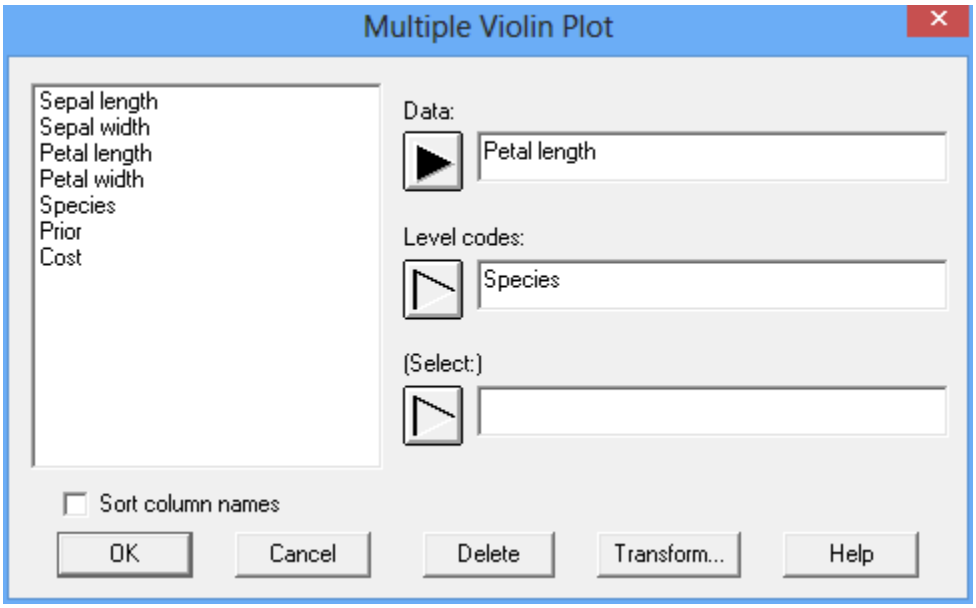


- **Data:** two or more numeric columns containing the observations, one column for each sample.

- **Select:** subset selection.

Data and Code Columns

If the data from all samples have been placed into a single column, then enter the name of that column and the column containing the group identifiers:



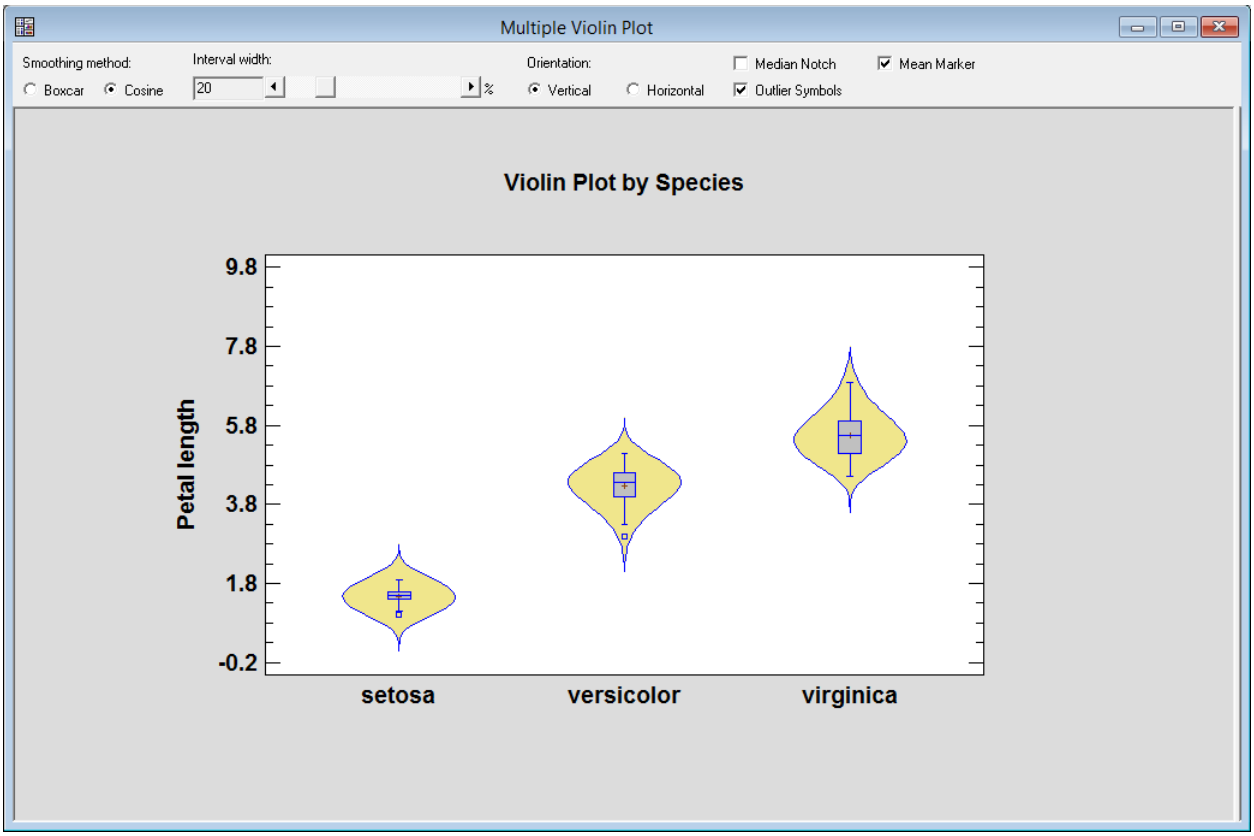
- **Data:** numeric column containing the observations from all samples.

- **Level codes:** numeric or non-numeric column containing an identifier for the sample corresponding to each data value.

- **Select:** subset selection.

Statlet

The output of this procedure is displayed in a dynamic Statlet window:



The difference between the 3 samples is clearly visible.

The first 2 controls on the toolbar determine how the density estimate is calculated. Basically, the estimate is created by counting the number of observations that fall within a window of fixed width moved across the range of the data.

- **Smoothing method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog box accessible from the *Edit* menu.

- **Interval Width:** the width of the window *h* within which observations affect the estimated density, as a percentage of the range covered by the x-axis. *h* = 60% is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.

The next several controls specify options for the box-and-whisker plot:

- **Orientation**: the orientation of the plot, corresponding to the direction of the whiskers.

- **Median Notch**: if selected, a notch will be added to the plot showing an approximate 100(1-α)% confidence interval for the median at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu).

- **Outlier Symbols**: if selected, indicates the location of outside points.

- **Mean Marker**: if selected, shows the location of the sample mean as well as the median.

Calculations

For information on how the nonparametric density estimate is calculated, refer to the PDF file titled *Distribution Fitting (Uncensored Data).* For information on the options for the box-and-whisker plot, refer to the PDF file titled *Box-and-Whisker Plot*.

References

Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems." Ann. Eugenics 7, Pt. II, 179-188.