Multiple Box-and-Whisker Plot

statgraphics 18®
centurion

Revised: 10/11/2017

Summary

The **Multiple Box-and-Whisker Plot** procedure creates a plot designed to illustrate important features of a numeric data column when grouped according to the value of a second variable. The box-and-whisker plot was first described by John Tukey (1977) in his box Exploratory Data Analysis. Box-and-whisker plots summarize data samples through 5 statistics:

1. minimum
2. lower quartile
3. median
4. upper quartile
5. maximum

They can also indicate the presence of outliers.

Sample StatFolio: *boxplot.sgp*

Sample Data

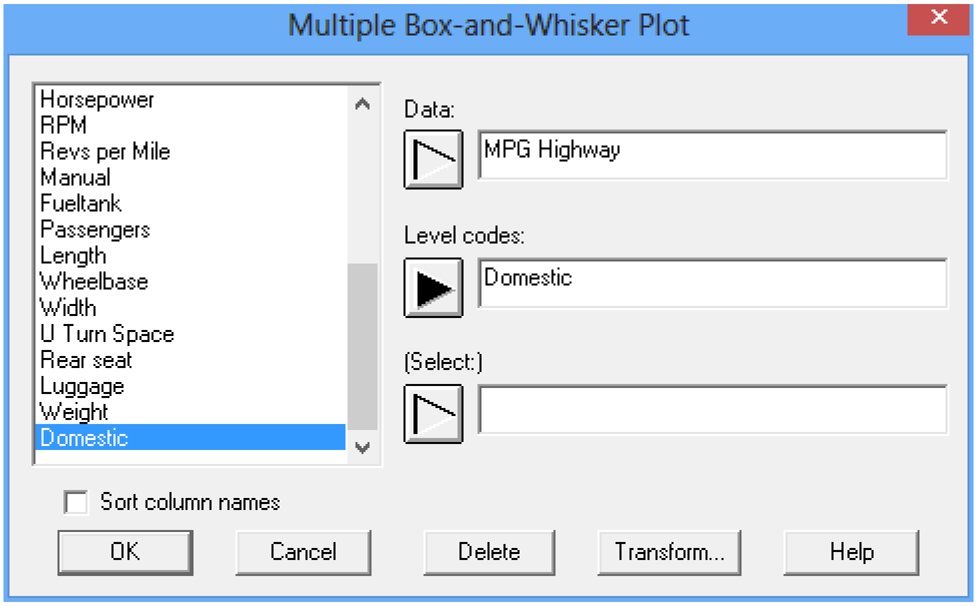
The file *93cars.sgd* contains information on 26 variables for *n* = 93 makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of 4 columns from that file:

Make	*Model*	*MPG Highway*	*Domestic*
Acura	Integra	31	0
Acura	Legend	25	0
Audi	90	26	0
Audi	100	26	0
BMW	535i	30	0
Buick	Century	31	1
Buick	LeSabre	28	1
Buick	Roadmaster	25	1
Buick	Riviera	27	1
Cadillac	DeVille	25	1
Cadillac	Seville	25	1
Chevrolet	Cavalier	36	1

The *Domestic* column contains a 1 for a car manufactured in the United States and a 0 otherwise.

Data Input

The data to be analyzed consist of a numeric column containing *n* = 2 or more observations and a second column (numeric or character) containing identifiers by which to group the data.



Data: MPG Highway

Level codes: Domestic

(Select:)

- **Data:** numeric column containing the data to be summarized.
- **Level codes:** numeric or non-numeric column containing group identifiers.
- **Select:** subset selection.

Analysis Summary

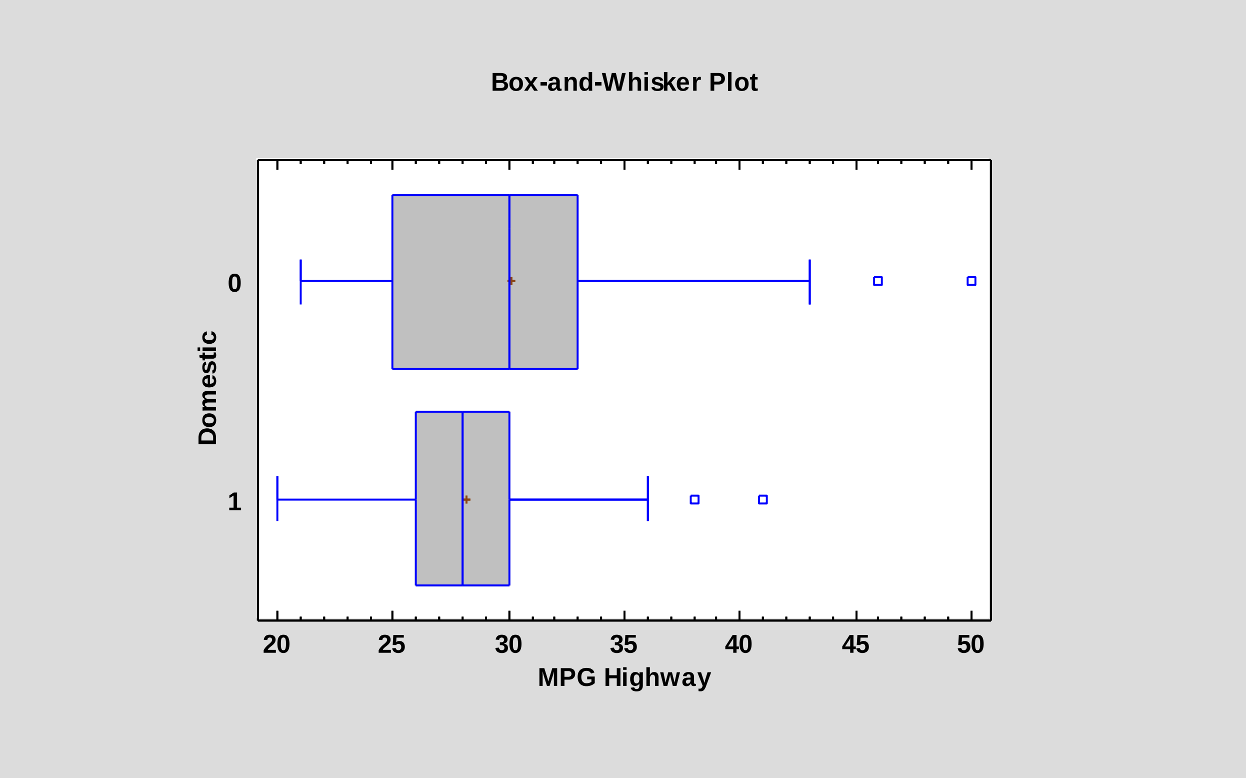
The *Analysis Summary* shows the number of observations in the data column and the number of levels (groups) into which the data has been divided.

```
Multiple Box-and-Whisker Plot
Dependent variable: MPG Highway
Factor: Domestic

Number of observations: 93
Number of levels: 2
```

Box-and-Whisker Plot

This pane displays a box-and-whisker plot for the data corresponding to each level code.



For each level code, a plot is constructed in the following manner:

- A box is drawn extending from the *lower quartile* of the sample to the *upper quartile*. This is the interval covered by the middle 50% of the data values when sorted from smallest to largest.

- A vertical line is drawn at the *median* (the middle value).

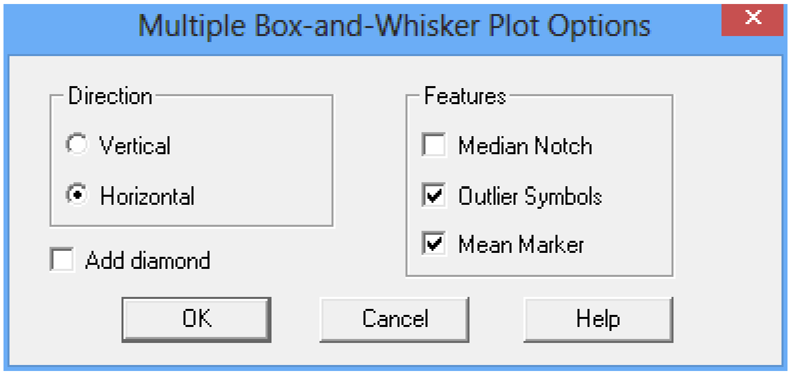
- If requested, a plus sign is placed at the location of the sample mean.

- Whiskers are drawn from the edges of the box to the largest and smallest data values, unless there are values unusually far away from the box (which Tukey calls *outside points*). Outside points, which are points more than 1.5 times the interquartile range (box width) above or below the box, are indicated by point symbols. Any points more than 3 times the interquartile range above or below the box are called *far outside points*, and are indicated by point symbols with plus signs superimposed on top of them. If outside points are present, the whiskers are drawn to the largest and smallest data values which are not outside points.

The above plot for the data on highway miles per gallon shows the data grouped by whether or not it is domestic (manufactured in the United States). The upper plot displays information for non-domestic automobiles, while the lower plot displays information for domestic automobiles.

The greater width of the upper plot indicates that there is more variability amongst cars made outside the U.S. The difference between the median lines also shows that non-domestic automobiles have greater miles per gallon on average than domestic automobiles. 4 outside points can also been observed.

Pane Options



Multiple Box-and-Whisker Plot Options

Direction

○ Vertical

● Horizontal

□ Add diamond

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

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

- **Median Notch**: if selected, a notch will be added to each plot to allow a comparison between the sample medians at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu). The notches are constructed such that if 2 notches do not overlap, there is a statistically significant difference between the medians of those 2 groups.

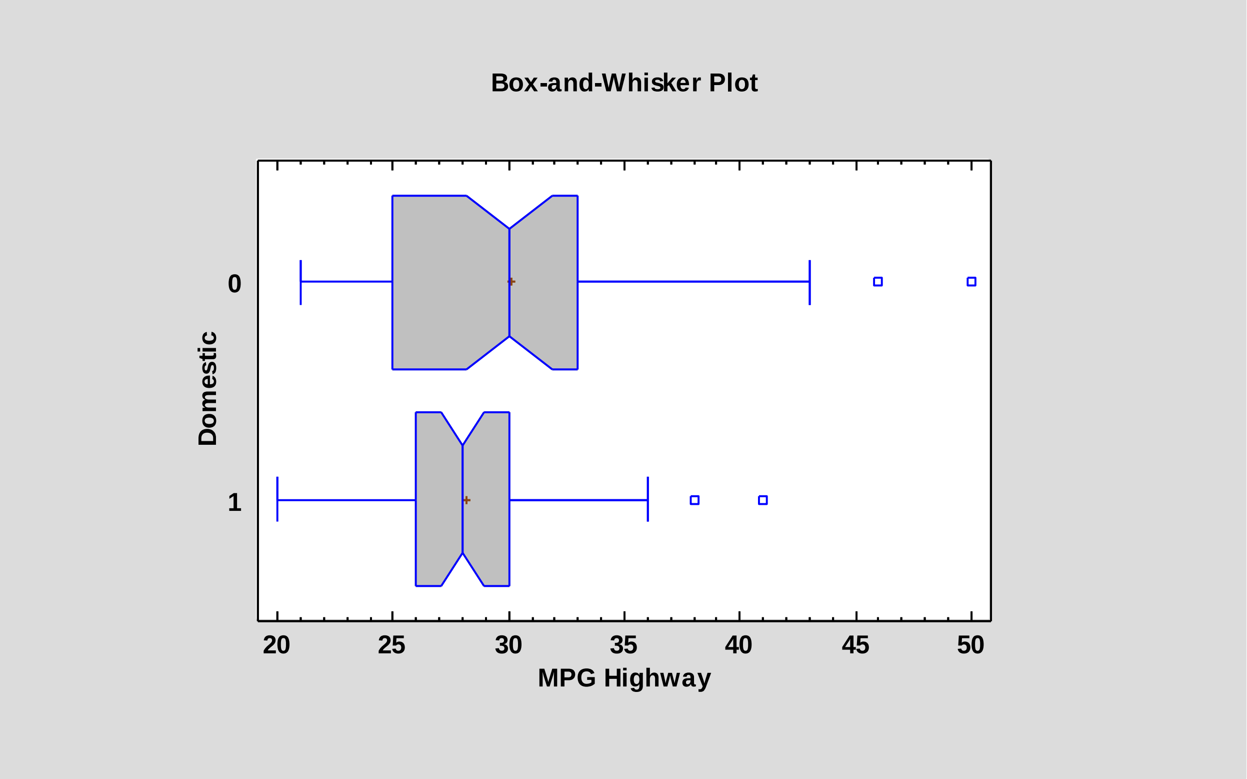
- **Outlier Symbols**: if selected, indicates the location of outside points.

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

- **Add diamond:** if selected, a diamond will be added to the plot showing a 100(1-α)% confidence interval for the mean at the default system confidence level.

<u>Example</u> – Notched Box-and-Whisker Plot

The following plot shows the addition of a median notch at the 95% confidence level.



The notch covers the interval

$$\tilde{x}_j \pm \frac{z_{\alpha/2}}{2} \frac{1.25(IQR_j)}{1.35\sqrt{n_j}} \left(1 + \frac{1}{\sqrt{2}}\right) \qquad (1)$$

where $\tilde{x}_j$ is the median of the data in group *j*, IQR_j is the sample interquartile range, n_j is the sample size, and $z_{\alpha/2}$ is the upper (α/2)% critical value of a standard normal distribution. In the above plot, the 2 notches overlap horizontally, indicating that the medians of the two groups of data are **not** significantly different at the 5% significance level.