Paired Sample Comparison

Summary

The **Paired Sample Comparison** procedure is designed to compare data in 2 numeric columns where the values in each row are paired, i.e., correspond to the same subject or experimental unit. The primary reason for such a comparison is typically to determine whether or not the factor that differentiates the columns has a significant effect on the data.

Note: the *Paired Sample Comparison* procedure parallels the *One Variable Analysis* procedure in terms of output and options, since it operates on a single numeric column of paired differences. A more detailed discussion of the mathematical basis of the calculations performed may be found in the *One Variable Analysis* documentation.

Sample StatFolio: *paired.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	MPG City
Acura	Integra	31	25
Acura	Legend	25	18
Audi	90	26	20
Audi	100	26	19
BMW	535i	30	22
Buick	Century	31	22
Buick	LeSabre	28	19
Buick	Roadmaster	25	16
Buick	Riviera	27	19
Cadillac	DeVille	25	16
Cadillac	Seville	25	16
Chevrolet	Cavalier	36	25

In this file, the experimental unit is a specific automobile model. It is of interest to compare the miles per gallon achieved by a particular model during highway driving with its performance during city driving.

Data Input

The data to be analyzed consist of two numeric columns of paired observations.
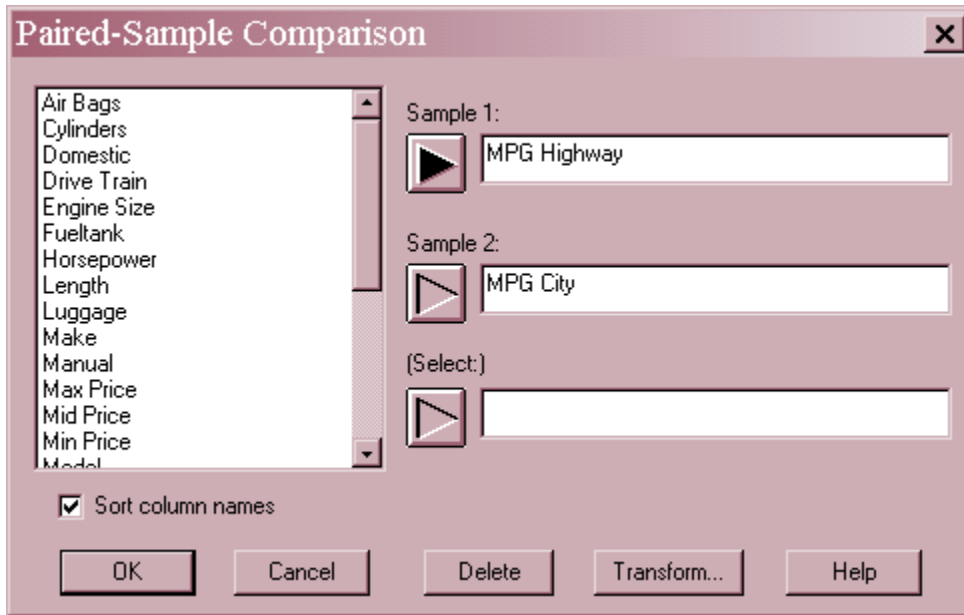
- **Sample 1:** numeric column containing the first value for each experimental unit.

- **Sample 2:** numeric column containing the second value for each experimental unit.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows the number of paired observations in the data column.

Paired Samples - MPG Highway & MPG City
Data variable: MPG Highway-MPG City
93 values ranging from 2.0 to 11.0

In the analysis that follows, primary attention is paid to the paired differences

$$d_i = x_{i,1} - x_{i,2} \tag{1}$$

which represent the differences between each pair of observations. In the sample data, d_i represents the additional mileage achieved by model i in highway driving versus that achieved in city driving.

Scatterplot

The *Scatterplot* plots each paired difference.

Scatterplot
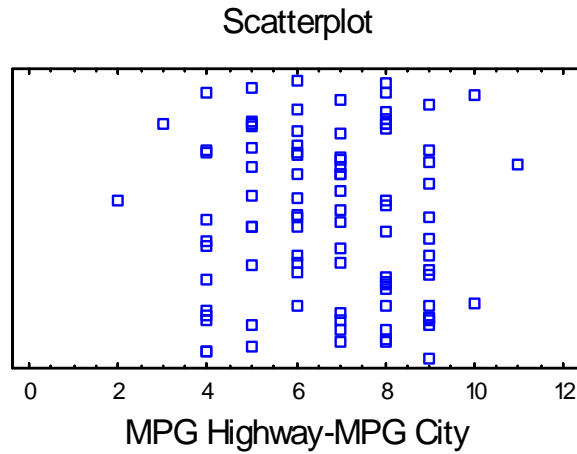
The differences are plotted along the horizontal axis. Along the vertical axis, the points are "jittered", i.e., they are offset randomly up or down. This is done to prevent identical points from overplotting each other. The amount of jitter is controlled by the *Jittering* button on the analysis toolbar:
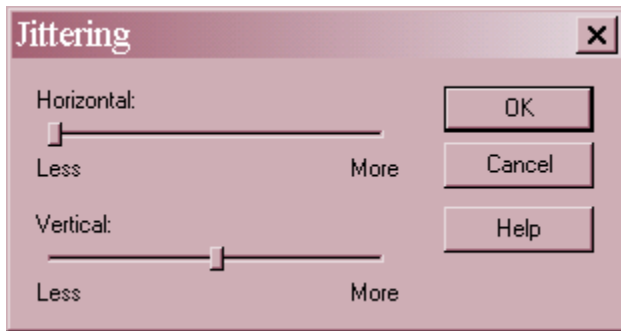
Reducing the amount of *Vertical* jitter will reduce the amount of random offset:

Scatterplot
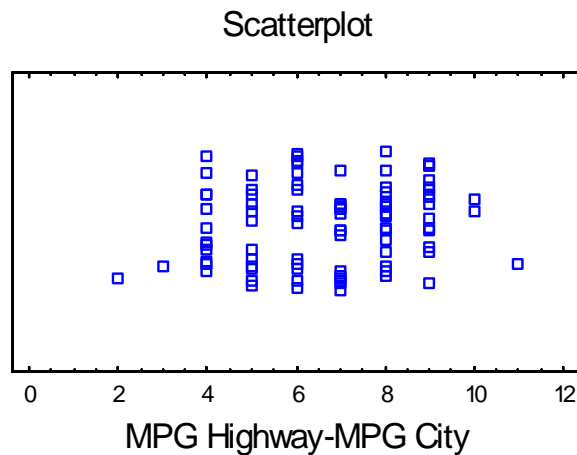
Summary Statistics

The *Summary Statistics* pane calculates a number of different statistics for the *n* paired differences:

Summary Statistics for MPG Highway-MPG City	
Count	93
Average	6.72043
Median	7.0
Variance	3.44273
Standard deviation	1.85546
Coeff. of variation	27.6092%
Minimum	2.0
Maximum	11.0
Range	9.0
Stnd. skewness	-0.686851
Stnd. kurtosis	-1.33445

Most of the statistics fall into one of three categories:

1. measures of *central tendency* – statistics that characterize the "center" of the data.
2. measure of *dispersion* – statistics that measure the spread of the data.
3. measures of *shape* – statistics that measure the shape of the data relative to a normal distribution.

The statistics included in the table by default are controlled by the settings on the *Stats* pane of the *Preferences* dialog box. Within the procedure, the selection may be changed using *Pane Options*. The meaning of each statistic is described in the documentation for the *One Variable Analysis* procedure.

Of particular interest for the current example is the average or median, which show that automobiles tend to average almost 7 miles per gallon more while driving on the highway than while driving in the city. The range of that difference was 2 to 11. The standardized skewness and standardized kurtosis are both between -2 and +2, indicating no significant deviation in shape from a normal distribution.
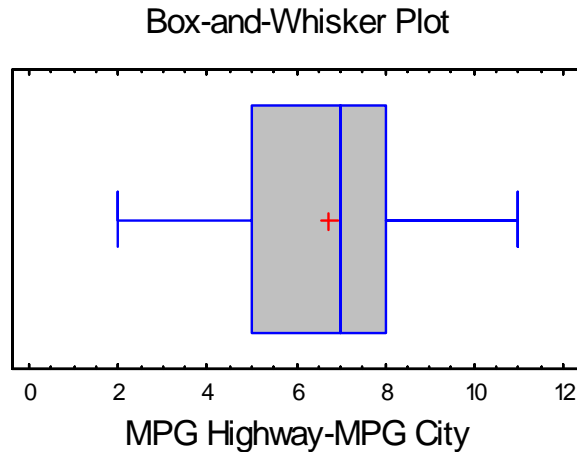
Pane Options

Select the desired statistics.

Box-and-Whisker Plot

This pane displays a box-and-whisker plot of the paired differences.

Box-and-Whisker Plot

The 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 differences 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 difference.

- Whiskers are drawn from the edges of the box to the largest and smallest differences, 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 paired differences is very symmetric. Although the vertical line for the median lies somewhat to the right of the plus sign for the mean, this is probably due to the fact that miles per gallon is only measured to the nearest whole number. There are no outside points.

Pane Options
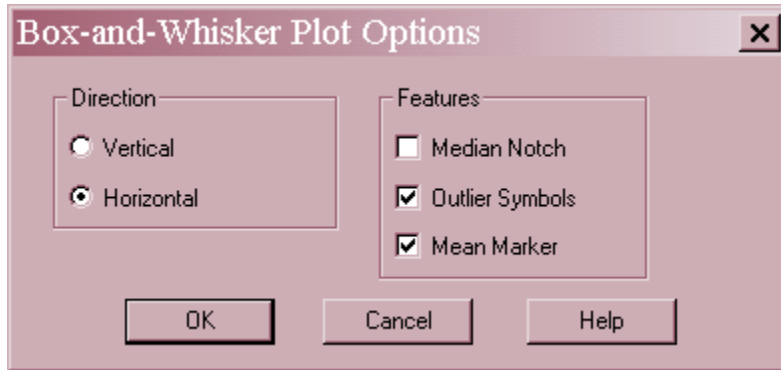
- **Direction**: the orientation of the plot, corresponding to the direction of the whiskers.
- **Median Notch**: if selected, a notch will be added to the plot showing an approximate 100(1-α)% confidence interval for the median difference at the default system confidence level (set on the *General* 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.

Example – Notched Box-and-Whisker Plot
The following plot shows the addition of a median notch at the 95% confidence level.
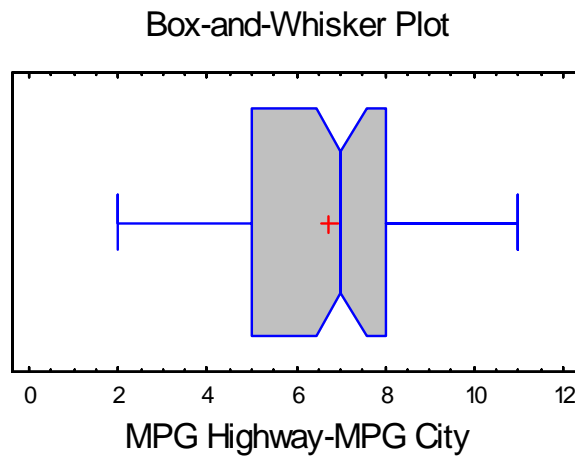
The notch, which ranges from approximately 6.5 to 7.5, provides an indication of the potential sampling error in the median, assuming that the differences are a random sample from a normal population.

Frequency Tabulation

A common method of summarizing quantitative data is to construct *k* intervals covering the range of the data and then calculate the number of observations falling within each of the intervals. STATGRAPHICS displays this type of tabulation in the *Frequency Tabulation* pane:

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
	at or below	0.5		0	0.0000	0	0.0000
1	0.5	1.5	1.0	0	0.0000	0	0.0000
2	1.5	2.5	2.0	1	0.0108	1	0.0108
3	2.5	3.5	3.0	1	0.0108	2	0.0215
4	3.5	4.5	4.0	12	0.1290	14	0.1505
5	4.5	5.5	5.0	12	0.1290	26	0.2796
6	5.5	6.5	6.0	15	0.1613	41	0.4409
7	6.5	7.5	7.0	16	0.1720	57	0.6129
8	7.5	8.5	8.0	18	0.1935	75	0.8065
9	8.5	9.5	9.0	15	0.1613	90	0.9677
10	9.5	10.5	10.0	2	0.0215	92	0.9892
11	10.5	11.5	11.0	1	0.0108	93	1.0000
12	11.5	12.5	12.0	0	0.0000	93	1.0000
	above	12.5		0	0.0000	93	1.0000

Frequency Tabulation for MPG Highway-MPG City

Mean = 6.72043   Standard deviation = 1.85546

This table is linked to the *Frequency Histogram* and displays the following information for each interval or "class":

- **Lower Limit** - the lower limit of the class.

- **Upper Limit** - the upper limit of the class.

- **Midpoint** - the class midpoint (halfway between the low and the high).

- **Frequency** - the number of observations that are greater than the lower limit of the class and less than or equal to the upper limit.

- **Relative Frequency** - the proportion of observations that lie in each class.

- **Cumulative Frequency** - the number of observations lying in the current or previous classes.

- **Cumulative Relative Frequency** - the proportion of observations lying in the current or previous classes.
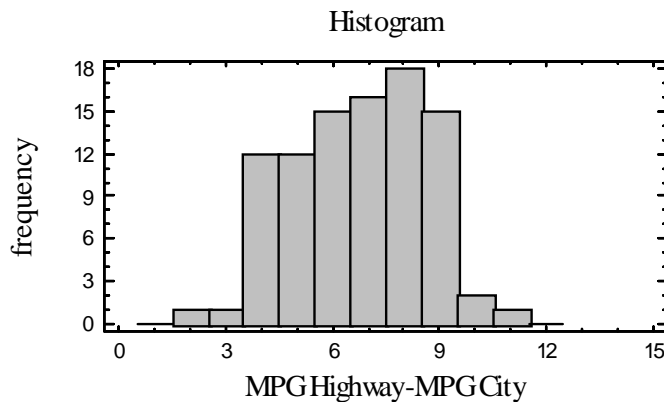
Pane Options
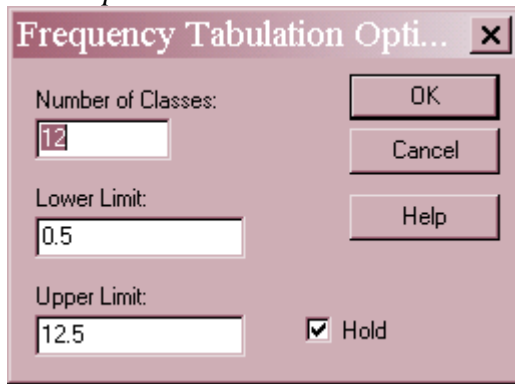
Frequency Tabulation Opti... ✕

Number of Classes:
12

Lower Limit:
0.5

Upper Limit:
12.5

OK

Cancel

Help

☑ Hold

- **Number of classes**: the number of intervals into which the data will be divided. Intervals are adjacent to each other and of equal width.

- **Lower Limit**: lower limit of the first interval.

- **Upper Limit**: upper limit of the last interval.

- **Hold**: maintains the selected number of intervals and limits even if the source data changes. By default, the number of classes and the limits are recalculated whenever the data changes. This is necessary so that all observations are displayed even if some of the updated data fall beyond the original limits.

The number of intervals into which the data are grouped by default is set by the rule specified on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu. For heavily rounded data such as the current example, care should be taken to insure that the midpoints of the classes line up with the rounded values. Otherwise, a distorted picture could be presented.

Frequency Histogram

The *Frequency Histogram* pane displays the results of the tabulation in the form of a barchart or lineplot, depending on the *Pane Options* setting.

Histogram

The height of each bar in the plot above represents the number of differences equal to each whole number.

Pane Options

- **Number of classes**: the number of intervals into which the data will be divided. Intervals are adjacent to each other and of equal width.

- **Lower Limit**: lower limit of the first interval.

- **Upper Limit**: upper limit of the last interval.

- **Hold**: maintains the selected number of intervals and limits even if the source data changes. By default, the number of classes and the limits are recalculated whenever the data changes. This is necessary so that all observations are displayed even if some of the updated data fall beyond the original limits.

- **Counts**: if *Relative*, the height of the bars represents the observations in a single interval. If *Cumulative*, the height represents the observations in the indicated interval and all intervals to its left.

- **Plot Type**: if *Histogram*, the class frequencies are displayed as a barchart. If *Polygon*, the frequencies are displayed using a connected line chart.

Example – Cumulative Frequency Polygon

Setting *Plot Type* to *Polygon* and checking the *Cumulative* and *Relative* boxes gives a display of the cumulative distribution of the paired differences:

Histogram
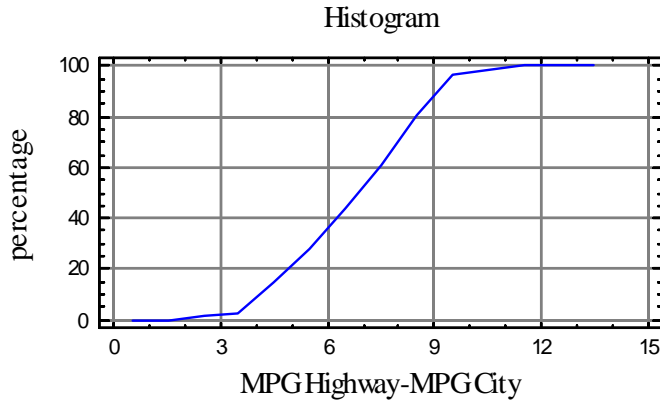
The above plot shows an almost uniform increase between 3.5 and 9.5.

Stem-and-Leaf Display

The stem-and-leaf display also presents a tabulation of the differences.

```
Stem-and-Leaf Display for MPG: unit = 0.1   1|2 represents 1.2

       1        2|0
       2        3|0
      14        4|000000000000
      26        5|000000000000
      41        6|000000000000000
     (16)       7|0000000000000000
      36        8|000000000000000000
      18        9|000000000000000
       3       10|00
       1       11|0
```

This display, due to John Tukey (1977), takes each data value and divides it into a *stem* and a *leaf*. For example, the difference between the highway and city miles per gallon for the automobile in the first row equals 6.0. Let the first digit ("6") be called the stem, and the second digit ("0") be called the leaf. Each row of the stem-and-leaf display corresponds to values with the same stem, shown to the left of the vertical line. To the right of the vertical line, a single digit is shown displaying the leaf for each data value. For example, the row showing

```
        7|0000000000000000
```

indicates that 16 automobiles did 7.0 miles per gallon better on the highway than in the city. Outside points, defined in the same manner as for the box-and-whisker plot, are plotted on special HI and LO stems.

The numbers in the far left-hand column, called *depths*, give a cumulative count of the observations inward from the top and bottom of the display. At the row containing the median, the number of observations in that row is shown instead and placed in parentheses.

Although similar to a histogram turned on its side, Tukey thought the stem-and-leaf plot was preferable to a barchart since the data values could be recovered from the display. He used the depths to help locate the median and quartiles when tabulating data by hand.

Pane Options

Flag Outliers: if checked, outside points will be placed on separate HI and LO stems. Otherwise, they will be included in the main part of the plot.

Percentiles

The *p-th percentile* of a continuous probability distribution is defined as that value of X for which the probability of being less than or equal to X equals p/100. For example, the 90-th percentile is that value below which lies 90% of the population. The *Percentiles* pane displays a table of selected percentiles based on the sample differences.

Percentiles for MPG Highway-MPG City

	Percentiles
1.0%	2.0
5.0%	4.0
10.0%	4.0
25.0%	5.0
50.0%	7.0
75.0%	8.0
90.0%	9.0
95.0%	9.0
99.0%	11.0

For example, the 90th percentile of the differences equals 9.0, implying that 90% of all automobiles had differences of 9 miles per gallon or less. If requested using *Pane Options*, lower and upper confidence limits or one-sided confidence bounds may also included, assuming that the differences follow a normal distribution.

Pane Options

- **Percentiles**: the percentages at which percentiles should be calculated. Set to 0 to suppress the calculation.

- **Include normal limits**: check to include confidence limits or bounds based on the assumption that the differences are random samples from a normal distribution.

- **Confidence level**: level for the limits or bounds.

- **Type**: select *Two-Sided* for a confidence interval or a one-sided bound to calculate a lower or upper bound for the percentile.

Quantile Plot

This pane plots the quantiles (percentiles) of the differences.
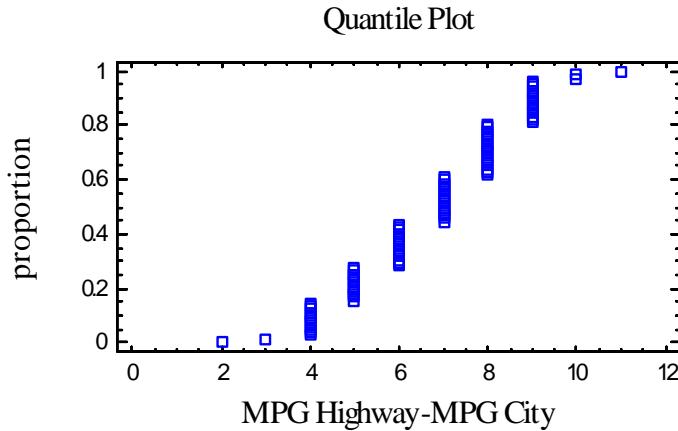

In this plot, the differences are sorted from smallest to largest and plotted at the coordinates

$$\left(x_{(j)}, \frac{j - 0.5}{n} \right) \tag{2}$$

In the example above, the heavy rounding of the data causes the differences to form vertical stacks at integer values.

Normal Probability Plot

Like the *Quantile Plot*, the *Normal Probability Plot* displays the paired differences from smallest to largest. However, it does so in a manner that makes it possible to judge whether or not the differences come from a normal distribution.
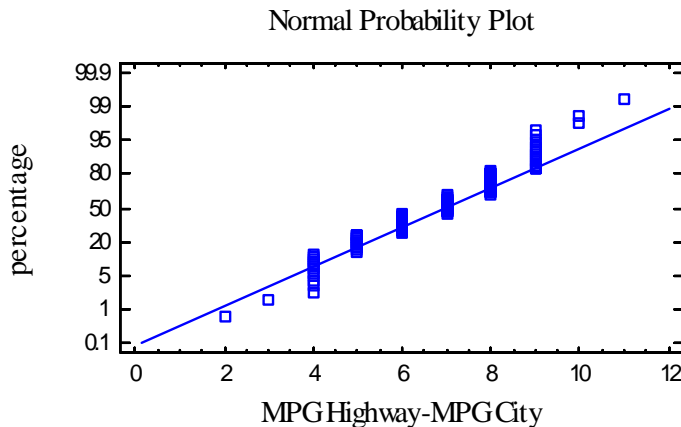

The vertical axis is scaled in such a way that, if the differences come from a normal distribution, the points should lie approximately along a straight line. In order to help determine how closely

the points correspond to a straight line, a reference line is superimposed on the plot corresponding to a normal distribution with mean μ and standard deviation σ. Note: set the default method for fitting lines on normal probability plots using the *EDA* pane on the *Preferences* dialog box, accessible from the *Edit* menu. A more detailed discussion may be found in the documentation for the *Normal Probability Plot* procedure.

In the current example, the points deviate from the line at the upper end, not extending out as much as would be expected for data from a normal distribution. The short upper tail is also evident in the *Frequency Histogram* illustrated earlier.

Pane Options
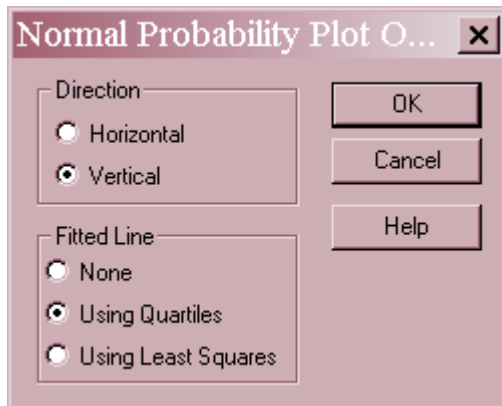
- **Direction**: the orientation of the plot. If vertical, the *Percentage* is displayed on the vertical axis. If *Horizontal*, *Percentage* is displayed on the horizontal axis.

- **Fitted Line**: the method used to fit the reference line to the differences. If *Using Quartiles*, the line passes through the median when *Percentage* equals 50 with a slope determined from the interquartile range. If *Using Least Squares*, the line is fit by least squares regression of the normal quantiles on the observed order statistics. The former method based on quartiles puts more weight on the shape of the data near the center and is often enable to show deviations from normality in the tails that would not be evident using the least squares method.
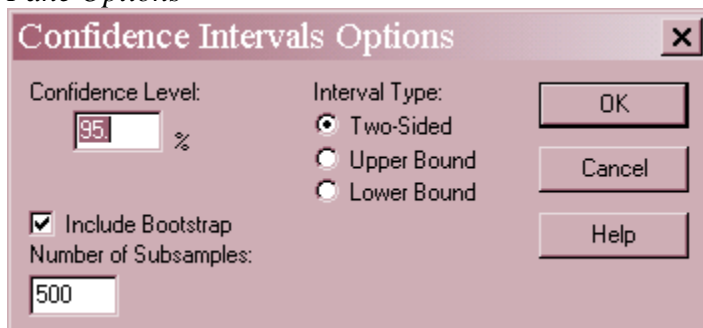
Confidence Intervals

The *Confidence Intervals* pane displays confidence intervals for the mean and standard deviation of the paired differences. It also includes bootstrap intervals for the mean, median, and standard deviation if requested.

Confidence Intervals for MPG Highway-MPG City
95.0% confidence interval for mean: 6.72043 +/- 0.382128   [6.3383,7.10256]
95.0% confidence interval for standard deviation: [1.62175,2.1685]

Bootstrap Intervals
Mean: [6.34409,7.11828]
Standard deviation: [1.63228,2.05965]
Median: [6.0,7.0]

95% confidence intervals are constructed in such a way that, in repeated sampling, 95% of such intervals will contain the true value of the parameter being estimated. You can also view a confidence interval as specifying the "margin of error" in the same manner as stated when taking an opinion poll. In the above example, although the mean difference in the sample was 6.72, the mean difference in the population from which the data were sampled may well differ from that estimate by 0.38 miles per gallon in either direction.

Confidence intervals for the mean and standard deviation rely on the assumption that the data come from a normal distribution. If this is not tenable, then an alternative is to construct intervals using the *bootstrap* method. In this method, q subsamples are formed by randomly selecting m observations from the original sample with replacement (i.e., the same observation may be selected more than once). For each of the q subsamples, the mean, median, and standard deviation are computed. Confidence intervals or bounds are then obtained using percentiles of the observed distribution of the subsample statistics. If the data do not come from a normal distribution, the bootstrap intervals may differ considerably from those obtained analytically. Also, because of the random nature of this procedure, different results will be obtained each time the bootstrap method is performed.

Pane Options

- **Confidence Level**: level of confidence for the interval or bound.
- **Type**: select two-sided for a confidence interval or a one-sided for a confidence bound.
- **Include Bootstrap**: include bootstrap intervals in the display.
- **Number of Subsamples**: the number of subsamples q on which the interval or bound will be based. Note: each subsample will contain $m = n$ observations, sampled with replacement.

Hypothesis Tests

Circumstances frequently arise when it is necessary to determine whether a sample comes from a distribution with a particular mean or standard deviation. For example, it is often of interest to test whether the mean difference could be equal to zero. To determine whether or not this is a reasonable statement given the data that has been collected, two approaches are possible:

1. Construct a *confidence interval* for the mean difference and determine whether or not 0 is within the confidence interval.

2. Perform a formal statistical *hypothesis test*.

The *Hypothesis Tests* pane supports the latter approach.

<u>t Test for the Mean</u>

The top section of the output is shown below:

```
Hypothesis Tests for MPG Highway-MPG City
Sample mean = 6.72043
Sample median = 7.0


t-test
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = 34.9291
P-Value = 0.0
Reject the null hypothesis for alpha = 0.05.
```

To run a hypothesis test, two competing hypotheses are formulated:

- **Null hypothesis**: a hypothesis such as $\mu = 0$ that will be given the benefit of the doubt. The value specified by the null hypothesis is labeled μ_0.

- **Alternative hypothesis**: a hypothesis such as $\mu \neq 0$ that will lead to rejection of the null hypothesis if there is sufficient evidence against the null.

The standard statistical approach to this problem is to construct a t-test using:

$$t = \frac{\bar{d} - \mu_0}{s/\sqrt{n}} \tag{3}$$

and compare it to Student's t distribution with $\nu = n$ - 1 degrees of freedom.

The above table shows the results of this test:

- *Computed t statistic* – the calculated value $t = 34.9291$

- *P-Value* – a value that may to used to reject the null hypothesis if it is small enough. At the $\alpha = 5\%$ significance level, the null hypothesis would be rejected if $P < 0.05$.

In this case, there is extremely strong evidence that the data do not come from a population in which the mean difference equals 0.

Tests for the Median

If the distribution from which the differences come is not normal, it may be of more interest to test a hypothesis concerning the population median rather than the mean. STATGRAPHICS performs two such tests: a sign test and a signed rank test.

---

sign test
Null hypothesis: median = 0.0
Alternative: not equal

Number of values below hypothesized median: 0
Number of values above hypothesized median: 93

Large sample test statistic = 9.53996 (continuity correction applied)
P-Value = 0.0
Reject the null hypothesis for alpha = 0.05.

signed rank test
Null hypothesis: median = 0.0
Alternative: not equal

Average rank of values below hypothesized median: 0.0
Average rank of values above hypothesized median: 47.0

Large sample test statistic = 8.39782 (continuity correction applied)
P-Value = 0.0
Reject the null hypothesis for alpha = 0.05.

The *Sign Test* is based on comparing the number of differences below the hypothesized median to the number of differences above the median. A large discrepancy leads to rejection of the null hypothesis. The signed rank test ranks the absolute differences between the paired differences and the hypothesized median from smallest to largest and compares the average rank of the paired differences below the hypothesized median to the average rank of those above.

Of primary importance in the above table are the P-Values. Small values (below 0.05 if operating at the 5% significance level) lead to rejection of the null hypothesis. In the current example, both tests reject the idea that the median difference equals 0.

Pane Options

- **t Test, Sign Test, Signed Rank Test, Chi-Squared Test**: specify the tests to be performed.

- **Mean/Median**: μ_0, the value of the mean or median specified by the null hypothesis.

- **Standard Deviation**: σ_0, the value of the standard deviation specified by the null hypothesis.

- **Alpha**: the significance level of the test, usually set to 0.01, 0.05, or 0.10. This equals the probability of rejecting the null hypothesis if it is true. It does not affect the P-Value, only the conclusion stated immediately below the P-Value.

- **Alt. Hypothesis:** the alternative hypothesis may be two-sided ("Not Equal") or one-sided (such as μ < 0 if "Less Than" is specified.).

Density Trace

The *Density Trace* provides a nonparametric estimate of the probability density function of the population from which the differences were sampled. It is created by counting the number of differences that fall within a window of fixed width moved across the range of the data.

Density Trace
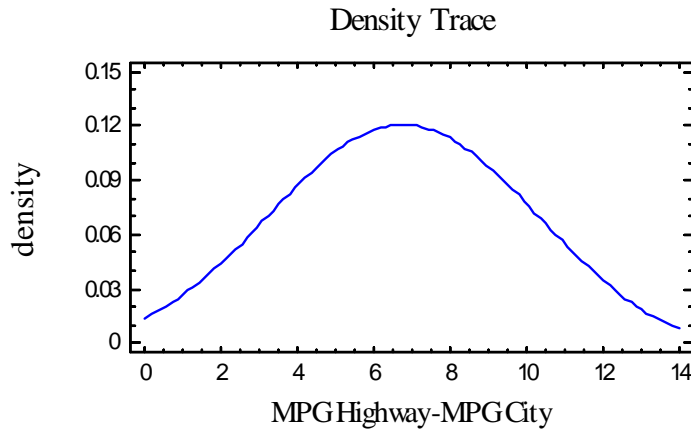
Pane Options
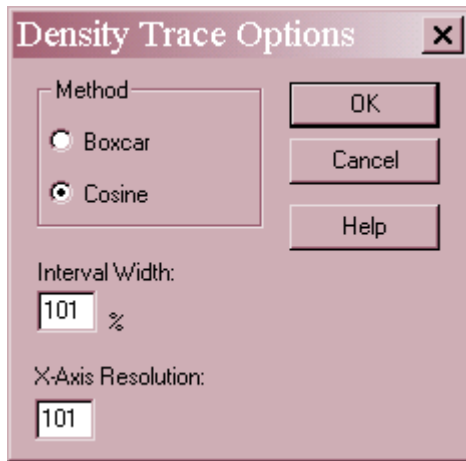
- **Method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog 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.

- **X-Axis Resolution**: the number of points at which the density is estimated.

Symmetry Plot

The Symmetry Plot is used to help judge whether the differences come from a symmetric distribution, i.e., a distribution that has a density function with the same shape on either side of the median.

Symmetry Plot
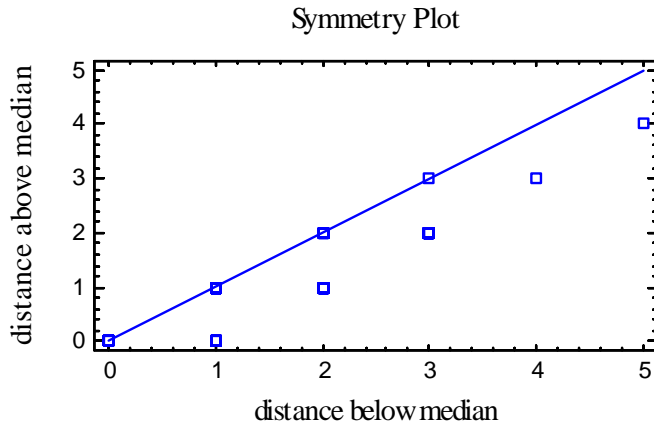
To create this plot, the differences are sorted and then paired based on their location with respect to the median. For example, with 93 differences, the sorted points are paired as:

$$(d_{(46)}, d_{(48)}), \ (d_{(45)}, d_{(49)}), \ (d_{(44)}, d_{(50)}), \ \dots, (d_{(1)}, d_{(93)})$$

The distance of each pair below and above the median is plotted. If the differences come from a symmetric distribution, the points should lie close to a 45-degree line. If not, the points will deviate from the line in a particular direction. The plot is not very useful for a sample with heavily rounded data.

Save Results

You may save the following results to the datasheet:

1. **Summary Statistics** – the values of the statistics displayed on the *Summary Statistics* pane.
2. **Statistic Labels** – the labels for the statistics displayed on the *Summary Statistics* pane.
3. **Percentiles** – the values of the percentiles displayed on the *Percentiles* pane.
4. **Frequencies** – the class frequencies displayed on the *Frequency Tabulation* pane.
5. **Cumulative Frequencies** – the class cumulative frequencies displayed on the *Frequency Tabulation* pane.
6. **Relative Frequencies**– the class relative frequencies displayed on the *Frequency Tabulation* pane.
7. **Cum. Rel. Frequencies** – the class cumulative relative frequencies displayed on the *Frequency Tabulation* pane.

Calculations

Consult the documentation for the *One Variable Analysis* procedure for details on the calculations performed.