One Variable Analysis

Summary

The **One-Variable Analysis** procedure is one of the primary procedures for analyzing a single column of numeric data. It calculates summary statistics, performs hypothesis tests, and creates a variety of graphical displays. The graphs include a scatterplot, histogram, box-and-whisker plot, quantile plot, normal probability plot, density trace, and symmetry plot. The tables include percentiles and a stem-and-leaf display.

Sample StatFolio: *onevar.sgp*

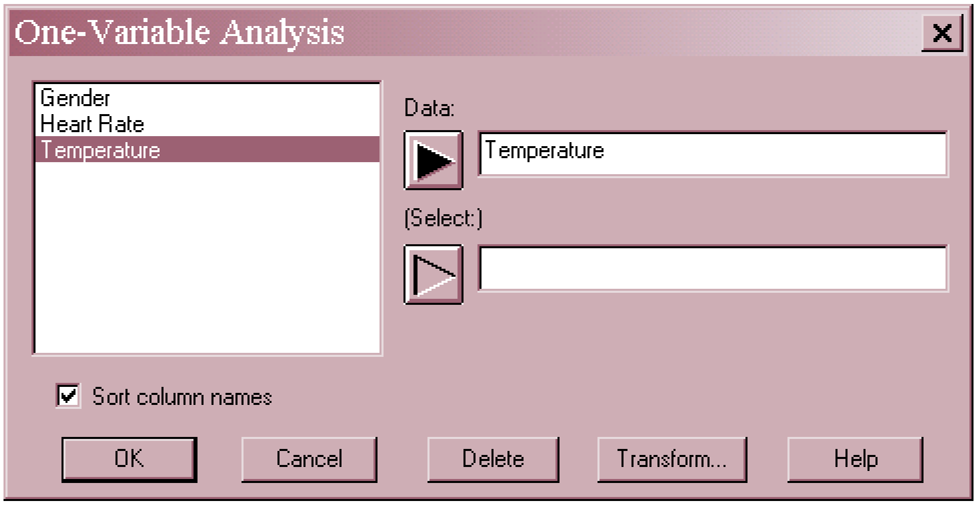
Sample Data:

The file *bodytemp.sgd* file contains data describing the body temperature of a sample of *n* = 130 people. It was obtained from the Journal of Statistical Education Data Archive (www.amstat.org/publications/jse/jse_data_archive.html) and originally appeared in the Journal of the American Medical Association. The first 20 rows of the file are shown below.

Temperature	*Gender*	*Heart Rate*
98.4	Male	84
98.4	Male	82
98.2	Female	65
97.8	Female	71
98	Male	78
97.9	Male	72
99	Female	79
98.5	Male	68
98.8	Female	64
98	Male	67
97.4	Male	78
98.8	Male	78
99.5	Male	75
98	Female	73
100.8	Female	77
97.1	Male	75
98	Male	71
98.7	Female	72
98.9	Male	80
99	Male	75

Data Input

The data to be analyzed consist of a single numeric column containing *n* = 2 or more observations.



Data: Temperature

(Select:)

☑ Sort column names

- **Data :** numeric column containing the data to be summarized.
- **Select:** subset selection.

Analysis Summary

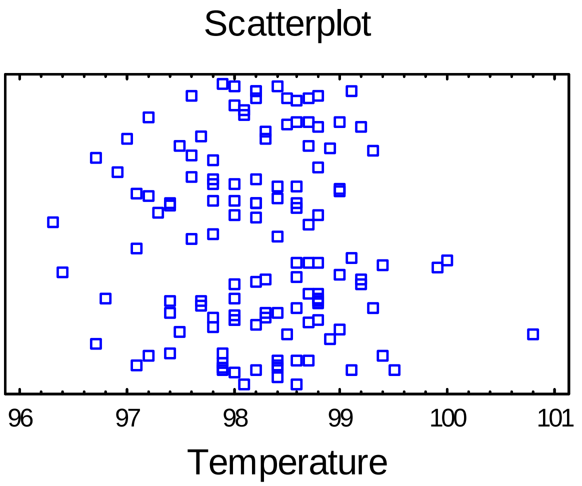
The Analysis Summary shows the number of observations in the data column.

One-Variable Analysis - Temperature
Data variable: Temperature
130 values ranging from 96.3 to 100.8

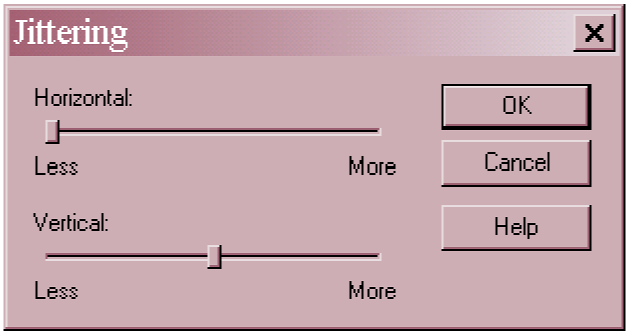
Also displayed are the largest and smallest values.

Scatterplot

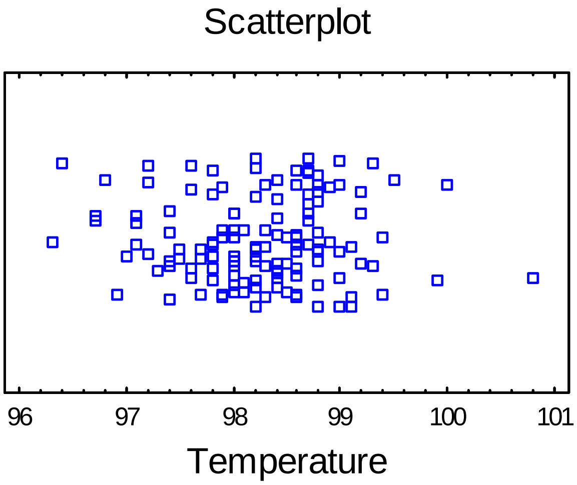
The *Scatterplot* plots each data value.



The data values 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:



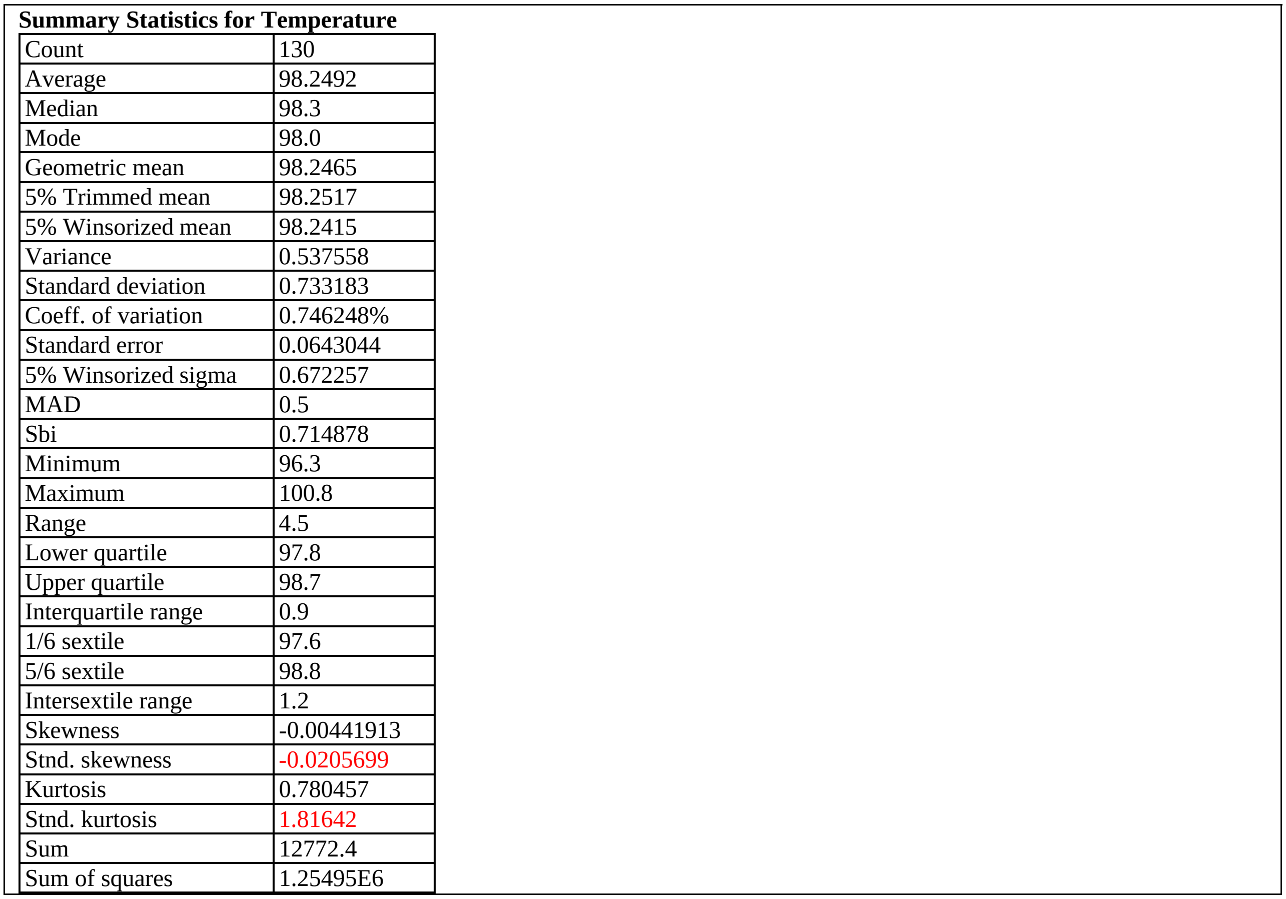
Reducing the amount of *Vertical* jitter will reduce the amount of random offset:



Notice that the points are most dense near the middle range of temperature and become less dense at higher or lower values. There is also one point at 100.8° that looks somewhat unusual. If you click on that point, you will see that it corresponds to row #15 of the file.

Summary Statistics

The *Summary Statistics* pane calculates a number of different statistics that are commonly used to summarize a sample of *n* observations:



Summary Statistics for Temperature

Count	130
Average	98.2492
Median	98.3
Mode	98.0
Geometric mean	98.2465
5% Trimmed mean	98.2517
5% Winsorized mean	98.2415
Variance	0.537558
Standard deviation	0.733183
Coeff. of variation	0.746248%
Standard error	0.0643044
5% Winsorized sigma	0.672257
MAD	0.5
Sbi	0.714878
Minimum	96.3
Maximum	100.8
Range	4.5
Lower quartile	97.8
Upper quartile	98.7
Interquartile range	0.9
1/6 sextile	97.6
5/6 sextile	98.8
Intersextile range	1.2
Skewness	-0.00441913
Stnd. skewness	-0.0205699
Kurtosis	0.780457
Stnd. kurtosis	1.81642
Sum	12772.4
Sum of squares	1.25495E6

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 shown below.

- **Count** - the sample size *n*, the number of non-missing entries in the column.

- **Average** or arithmetic **mean** (measure of central tendency) - the center of mass of the data, given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

- **Median** (measure of central tendency) - the middle value when the data are sorted from smallest to largest. If n is odd, the sample median equals $x_{(0.5+n/2)}$, where $x_{(i)}$ represents the *i-th* smallest observation If n is even, the sample median is the average of the two middle values:

$$\frac{x_{(n/2)} + x_{(1+n/2)}}{2} \tag{2}$$

- **Mode** (measure of central tendency) - the most frequently occurring data value (if any). If no single value occurs more often than any other, this statistic is not calculated.

- **Geometric Mean** (measure of central tendency) - estimates the center of the data according to

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \tag{3}$$

  This statistic is often used for data that are positively skewed, since it will be closer to the peak of the distribution than the arithmetic mean. Note: this statistic is only defined for a sample of data in which all values are greater than 0. The program calculates the statistic by averaging the natural logarithms of the data values and taking the inverse logarithm of the result.

- **α% Trimmed Mean** (measure of central tendency) – the mean of the sample after removing a fraction α each of the smallest and largest data values:

$$T(\alpha) = \frac{1}{n(1-2\alpha)}\left[k\left(x_{(r+1)} + x_{(n-r)}\right) + \sum_{i=r+2}^{n-r-1} x_{(i)}\right] \tag{4}$$

  where $r = \lfloor \alpha n \rfloor$ and $k = 1 - (\alpha n - r)$. By default, STATGRAPHICS trims 15% from each end, although that value may be changed using *Pane Options*.

- **Winsorized mean** (measure of central tendency) – a resistant measure obtained by calculating the sample mean after copies of $x_{(r+1)}$ and $x_{(n-r)}$ have replaced the data values which would be trimmed away by a trimmed mean:

$$T_W = \frac{1}{n}\left\{\sum_{i=r+1}^{n-r} x_{(i)} + r\left[x_{(r+1)} + x_{(n-r)}\right]\right\} \tag{5}$$

  Both the trimmed mean and the Winsorized mean are less affected by outliers than the arithmetic mean.

- **Variance** (measure of dispersion) - a measure of the average squared deviation around the sample mean:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{6}$$

- **Standard deviation** (measure of dispersion) - the square root of the sample variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{7}$$

- **Coefficient of variation** or **relative standard deviation** (measure of dispersion) - measures the magnitude of the standard deviation as a percentage of the sample mean according to:

$$CV = 100 \frac{s}{\bar{x}} \% \tag{8}$$

It is only defined if $\bar{x} > 0$.

- **Standard error** (measure of dispersion) - the standard error of the mean:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{9}$$

- α**% Winsorized sigma** (measure of dispersion) – a Winsorized estimate of variability around the Winsorized mean:

$$S_W = \sqrt{\frac{n\left\{\sum_{i=r+1}^{n-r} \left(x_{(i)} - T_W\right)^2 + r\left[\left(x_{(r+1)} - T_W\right)^2 + \left(x_{(n-r)} - T_W\right)^2\right]\right\}}{(n-2r)(n-2r-1)}} \tag{10}$$

- **MAD** – the median absolute deviation:

$$MAD = median_i \left\{\left|x_i - \tilde{x}\right|\right\} \tag{11}$$

- **Sbi** (measure of dispersion) - an estimate based on a weighted sum of squares around the sample median:

$$S_{bi} = \frac{\sqrt{n\sum_{i=1}^{n} \left(x_i - \tilde{x}\right)^2 \left(1 - u_i^2\right)^4}}{\left|\sum_{i=1}^{n} \left(1 - u_i^2\right)\left(1 - 5u_i^2\right)\right|} \tag{12}$$

where

$$u_i = \frac{x_i - \tilde{x}}{9MAD} \tag{13}$$

- **Minimum** - the smallest data value $x_{(1)}$.

- **Maximum** - the largest data value $x_{(n)}$.

- **Range** (measure of dispersion) - the maximum minus the minimum:

$$R = x_{(n)} - x_{(1)} \tag{14}$$

- **Lower quartile** - the 25-th percentile. Approximately 25% of the data values will lie below this value.

- **Upper quartile** - the 75-th percentile. Approximately 75% of the data values will lie below this value.

- **Interquartile range** (measure of dispersion) - the distance between the quartiles:

$$\text{IQR} = \text{upper quartile - lower quartile} \tag{15}$$

- **1/6 sextile** - the 16.67-th percentile.

- **5/6 sextile** - the 83.33-th percentile.

- **Intersextile range** (measure of dispersion) - the distance between the sextiles:

$$\text{ISR} = \text{upper sextile - lower sextile} \tag{16}$$

- **Skewness** (measure of shape) - a measure of asymmetry calculated according to:

$$g_1 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} \tag{17}$$

A value close to 0 would correspond to a nearly symmetric data sample. Positive skewness indicates a longer upper tail than lower, while negative skewness indicates a longer lower tail.

- **Standardized skewness** (measure of shape) - converts the skewness statistic computed above to a value that has approximately a standard normal distribution in large samples:

$$z_1 = \frac{g_1}{\sqrt{6/n}} \tag{18}$$

At the 5% significance level, significant skewness could be asserted if z_1 fell outside the interval (-2, +2).

- **Kurtosis** (measure of shape) - a measure of relative peakedness or flatness compared to a bell-shaped curve:

$$g_2 = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{19}$$

  A value close to 0 would correspond to a nearly bell-shaped normal distribution. Positive kurtosis indicates a distribution that is more peaked in the center and has longer tails than the normal. Negative kurtosis indicates a distribution that is flatter than the normal with shorter tails. This measure is usually only relevant for characterizing symmetric data samples.

- **Standardized kurtosis** (measure of shape) - converts the kurtosis statistic computed above to a value which has approximately a standard normal distribution in large samples:

$$z_2 = \frac{g_2}{\sqrt{24/n}} \tag{20}$$

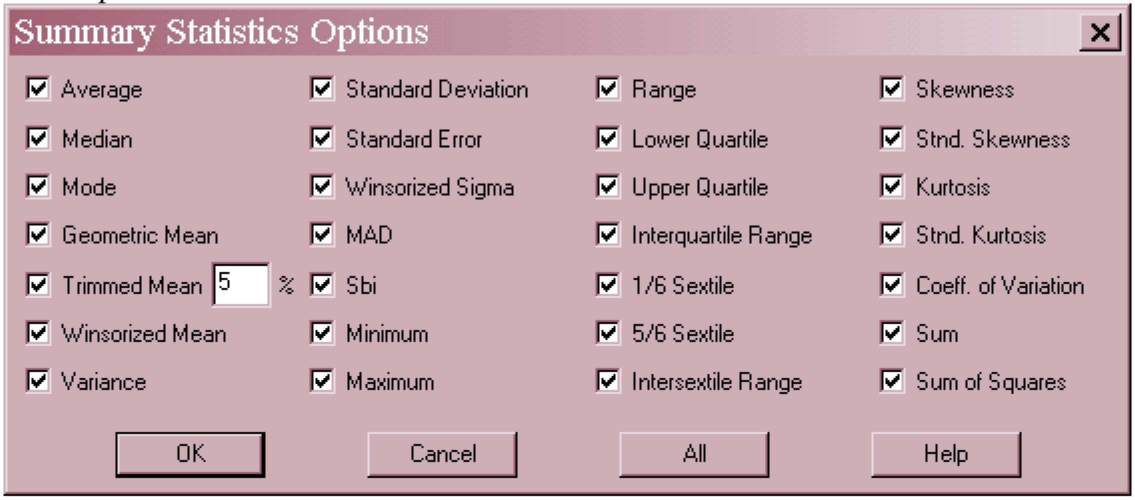
  At the 5% significance level, significant kurtosis could be asserted if z_2 fell outside the interval (-2,+2).

- **Sum** - the sum of the data values.

- **Sum** of squares - the sum of the data values squared.

For the data on body temperatures, all of the measures of central tendency are very similar, as they should be if body temperatures follow a symmetric distribution such as the normal. The standardized skewness and standardized kurtosis are also both between -2 and +2, indicating no significant deviation in shape from a normal distribution.

Pane Options



Summary Statistics Options

☑ Average
☑ Median
☑ Mode
☑ Geometric Mean
☑ Trimmed Mean 5 %
☑ Winsorized Mean
☑ Variance

☑ Standard Deviation
☑ Standard Error
☑ Winsorized Sigma
☑ MAD
☑ Sbi
☑ Minimum
☑ Maximum

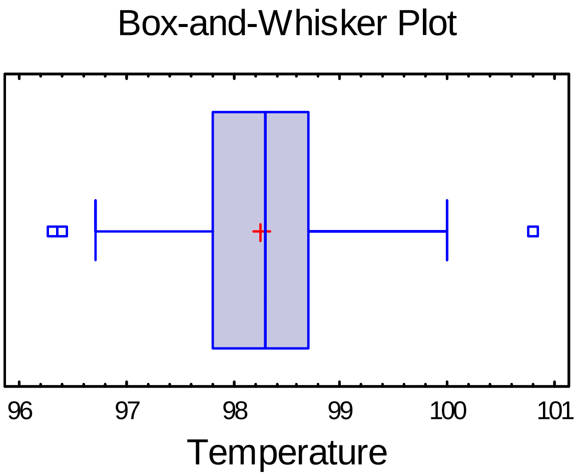
☑ Range
☑ Lower Quartile
☑ Upper Quartile
☑ Interquartile Range
☑ 1/6 Sextile
☑ 5/6 Sextile
☑ Intersextile Range

☑ Skewness
☑ Stnd. Skewness
☑ Kurtosis
☑ Stnd. Kurtosis
☑ Coeff. of Variation
☑ Sum
☑ Sum of Squares

Select the desired statistics.

Box-and-Whisker Plot

This pane displays the 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 data values when sorted from smallest to largest.

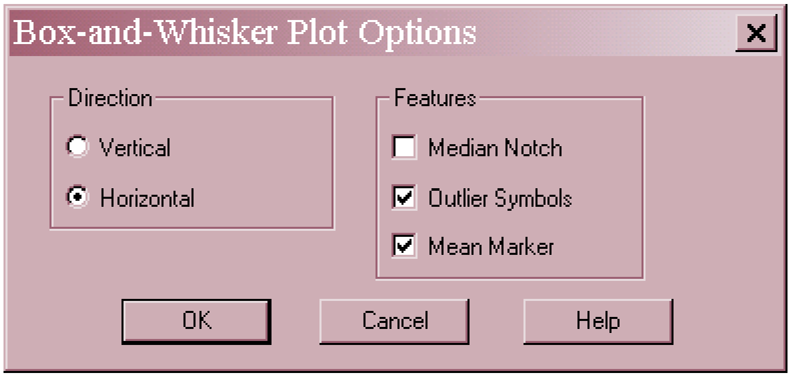
- A vertical line is drawn at the *median* (the middle value).

- 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 body temperature data is very symmetric. The plus sign for the mean lies very close to the line for the median, while the whiskers are of approximately equal length. There are 3 outside points. When sampling 130 observations from a normal distribution, outside points can be expected to occur just by chance about half the time, but usually only one or two. Far outside points, of which there is none, occur extremely rarely.

Pane Options



Box-and-Whisker Plot Options

Direction

○ Vertical

● Horizontal

Features

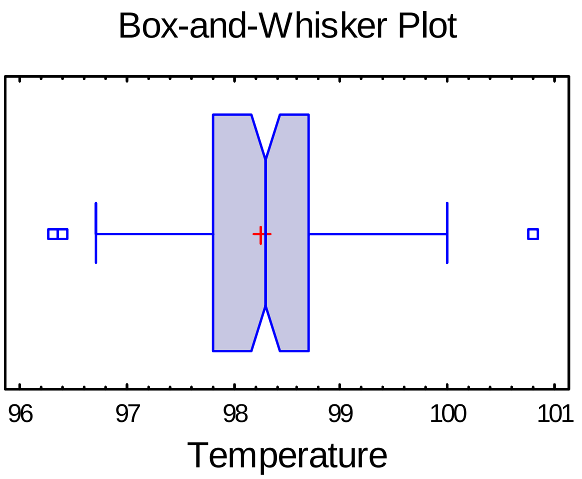
□ Median Notch

☑ 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 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.

Example – 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

$$\textit{sample median} \pm z_{\alpha/2}\frac{1.25(IQR)}{1.35\sqrt{n}} \tag{21}$$

where *IQR* is the sample interquartile range, *n* is the sample size, and $z_{\alpha/2}$ is the upper (α/2)% critical value of a standard normal distribution. The notch, which ranges from approximately 98.16 to 98.44, provides an indication of the potential sampling error in the median, assuming that the data are a random sample from a normal population. Note that this interval does not contain the usually quoted value for average human body temperature of 98.6°.

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:

Frequency Tabulation for Temperature

	Lower	*Upper*			*Relative*	*Cumulative*	*Cum. Rel.*
Class	*Limit*	*Limit*	*Midpoint*	*Frequency*	*Frequency*	*Frequency*	*Frequency*
	at or below	96.0		0	0.0000	0	0.0000
1	96.0	96.25	96.125	0	0.0000	0	0.0000
2	96.25	96.5	96.375	2	0.0154	2	0.0154
3	96.5	96.75	96.625	2	0.0154	4	0.0308
4	96.75	97.0	96.875	3	0.0231	7	0.0538
5	97.0	97.25	97.125	6	0.0462	13	0.1000
6	97.25	97.5	97.375	8	0.0615	21	0.1615
7	97.5	97.75	97.625	7	0.0538	28	0.2154
8	97.75	98.0	97.875	23	0.1769	51	0.3923
9	98.0	98.25	98.125	13	0.1000	64	0.4923
10	98.25	98.5	98.375	17	0.1308	81	0.6231
11	98.5	98.75	98.625	18	0.1385	99	0.7615
12	98.75	99.0	98.875	17	0.1308	116	0.8923
13	99.0	99.25	99.125	6	0.0462	122	0.9385
14	99.25	99.5	99.375	5	0.0385	127	0.9769
15	99.5	99.75	99.625	0	0.0000	127	0.9769
16	99.75	100.0	99.875	2	0.0154	129	0.9923
17	100.0	100.25	100.125	0	0.0000	129	0.9923
18	100.25	100.5	100.375	0	0.0000	129	0.9923
19	100.5	100.75	100.625	0	0.0000	129	0.9923
20	100.75	101.0	100.875	1	0.0077	130	1.0000
21	101.0	101.25	101.125	0	0.0000	130	1.0000
22	101.25	101.5	101.375	0	0.0000	130	1.0000
23	101.5	101.75	101.625	0	0.0000	130	1.0000
24	101.75	102.0	101.875	0	0.0000	130	1.0000
	above	102.0		0	0.0000	130	1.0000

Mean = 98.2492   Standard deviation = 0.733183

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 f_j 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, given by f_j/n.

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

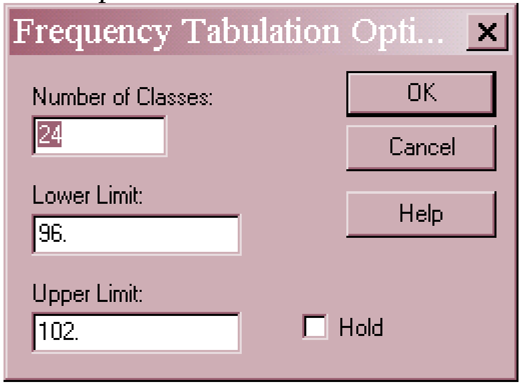
$$\sum_{i=1}^{j} f_i \tag{22}$$

- **Cumulative Relative Frequency** - the proportion of observations lying in the current or previous classes:

$$\frac{\sum_{i=1}^{j} f_i}{n} \tag{23}$$

The rightmost column is of considerable interest, since it corresponds to the *cumulative distribution* of the observations. For example, 62.31% of the data are less than or equal to 98.5°.

Pane Options



Number of Classes: 24

Lower Limit: 96.

Upper Limit: 102.

□ 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 is grouped by default is set by the rule specified on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu. Each rule determines the number of intervals *m* as a function of the sample size *n*. The rules are:

Sturges' rule: $m = ceiling(1 + 3.322 \log(n))$ (24)

10 log10(n): $m = ceiling(10 \log(n))$ (25)

Scott's rule: $m = ceiling[\,(max-min) / (3.5\, s / n^{1/3})\,]$ (26)

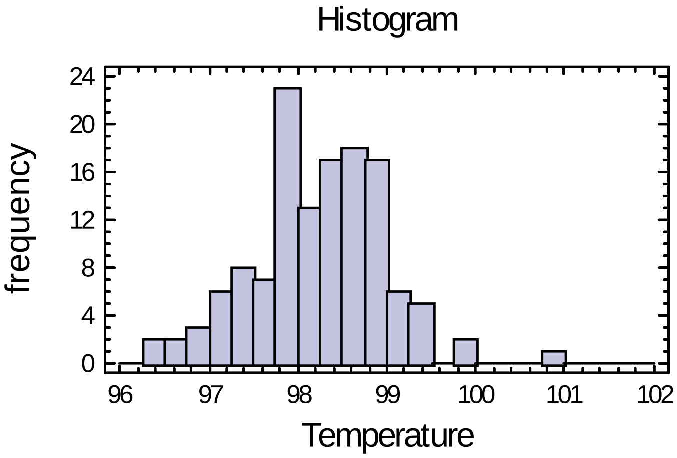
Freedman-Diaconis rule: $m = ceiling[\ (max\text{-}min) /(2.0\ IQR/\ n^{1/3})\ ]$ (27)

Fixed number: *m = pre-specified number* (28)

where *min* equals the smallest data value in the sample, *max* equals the largest data value, *s* equals the sample standard deviation, *IQR* equals the sample interquartile range, and the *ceiling* function finds the smallest integer greater than or equal to its argument. You can experiment with different rules to determine which rule gives a good number of intervals for your most common type of data.

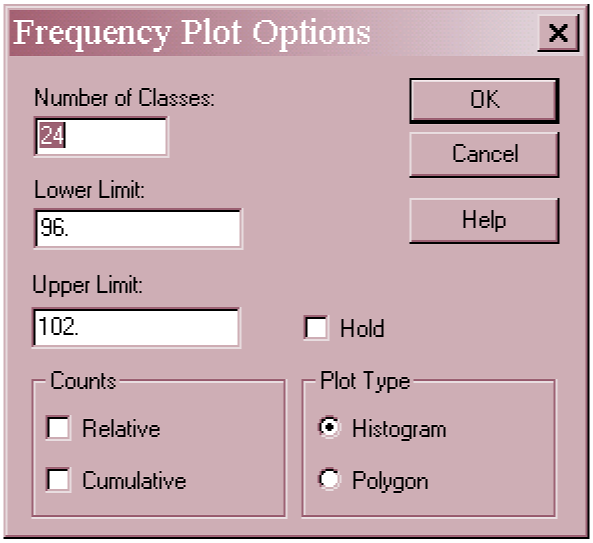
Frequency Histogram

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



The height of each bar in the plot above represents the number of observations in each class.

Pane Options



Number of Classes: 24

Lower Limit: 96.

Upper Limit: 102.

□ Hold

Counts

□ Relative

□ Cumulative

Plot Type

● Histogram

○ Polygon

- **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.

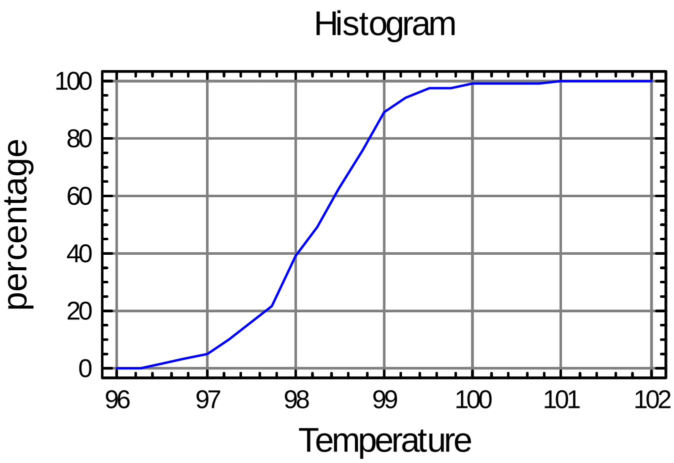
- **Hold**: maintains the selected number of intervals and limits even if the source data changes. By default, the number of classes and the limits are recalculated whenever the data changes. This is necessary so that all observations are displayed even if some of the updated data fall beyond the original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

<u>Example</u> – Cumulative Frequency Polygon

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



The above plot shows the percentage of observations at or below the upper limit of each interval into which the data were grouped. It can be seen that about 50% of the data fall below 98.3°.

Stem-and-Leaf Display

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

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

```
          LO|96.3 96.4

  2       96|
  6       96|7789
 19       97|0111222344444
 40       97|556666777888888899999
(38)      98|00000000000111222222222233333444444444
 52       98|555666666666677777777888888888899
 19       99|000001112223344
  4       99|59
  2      100|0

          HI|100.8
```

This display, due to John Tukey (1977), takes each data value and divides it into a *stem* and a *leaf*. For example, the temperature of the first subject in the data sample had a body temperature of 98.4°. Let the first two digits ("98") be called the stem, and the third digit ("4") 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

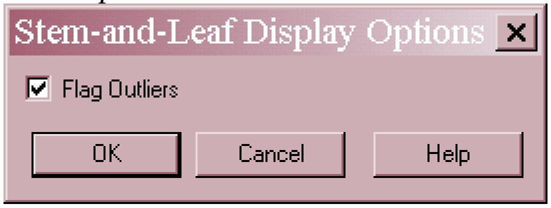
```
98|00000000000111222222222233333444444444
```

indicates that there were11 subjects with temperatures of 90.0°, 3 subjects with temperatures of 98.1°, 10 with 98.2°, 5 with 98.3°, and 9 with a 98.4°. 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

- **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 data.

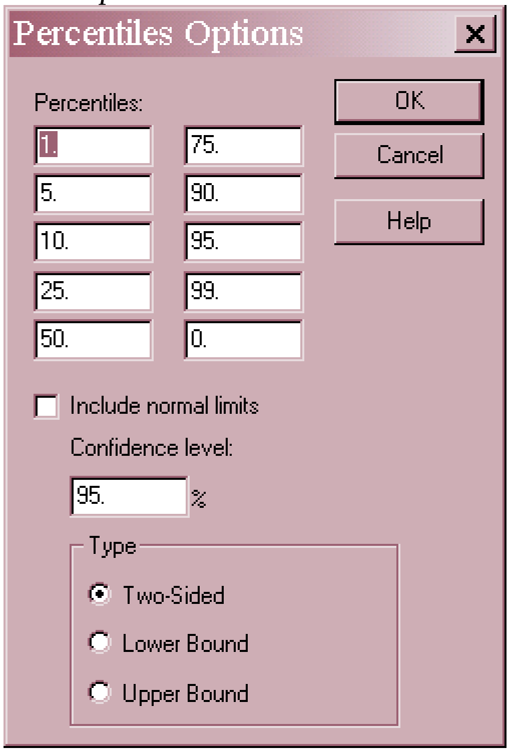
Percentiles for Temperature

	Percentiles	*Lower Limit*	*Upper Limit*
1.0%	96.4	96.2713	96.7643
5.0%	97.0	96.829	97.2211
10.0%	97.25	97.1232	97.4677
25.0%	97.8	97.6062	97.8882
50.0%	98.3	98.1222	98.3762
75.0%	98.7	98.6102	98.8922
90.0%	99.1	99.0308	99.3753
95.0%	99.3	99.2774	99.6695
99.0%	100.0	99.7342	100.227

Output includes 95.0% normal confidence limits.

For example, the 90th percentile of the body temperature data equals 99.1°, implying that 90% of all subjects had temperatures of 99.1° or lower. If requested using *Pane Options*, lower and upper confidence limits or one-sided confidence bounds may also included, assuming that the data are samples from a normal distribution. The 95% confidence interval for the temperature at or below which one would find 90% of all individuals similar to those in the study ranges from 99.03° to 99.38°.

Pane Options



Percentiles Options

Percentiles:

1.
5.
10.
25.
50.
75.
90.
95.
99.
0.

☐ Include normal limits

Confidence level: 95. %

Type

● Two-Sided
○ Lower Bound
○ Upper Bound

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

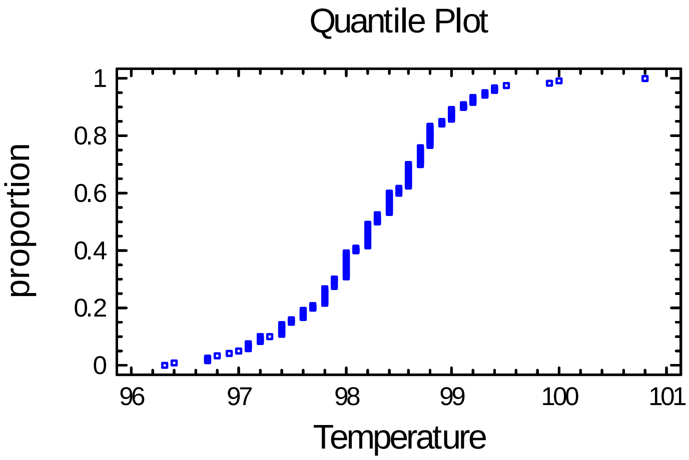
- **Include normal limits**: check to include confidence limits or bounds based on the assumption that the data are random samples from a normal distribution.

- **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 data.



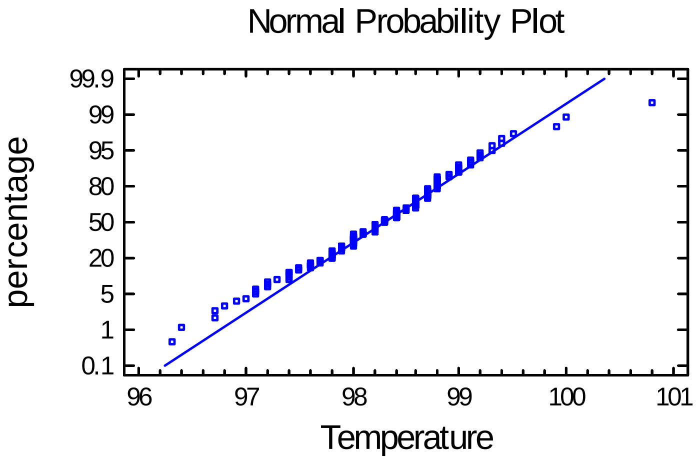
In this plot, the data are sorted from smallest to largest and plotted at the coordinates

$$\left(x_{(j)}, \frac{j-0.5}{n}\right) \tag{29}$$

The S-shape shown above is typical of data from a bell-shaped normal distribution.

Normal Probability Plot

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



The vertical axis is scaled in such a way that, if the data come from a normal distribution, the points should lie approximately along a straight line. In constructing the plot, the points are plotted at coordinates equal to

$$\left(x_{(j)}, \Phi^{-1}\left(\frac{j-0.375}{n+0.25} \right) \right) \tag{30}$$

where $\Phi^{-1}(u)$ represents the inverse standard normal distribution evaluated at *u*. The labels along the vertical axis equal 100*u*%, for values of *u* ranging between 0.001 and 0.999.

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. There are two options for fitting the line:

1. Using the median and the sample quartiles:

$$\hat{\mu} = \text{sample median} \tag{31}$$

$$\hat{\sigma} = \text{interquartile range / 1.35} \tag{32}$$

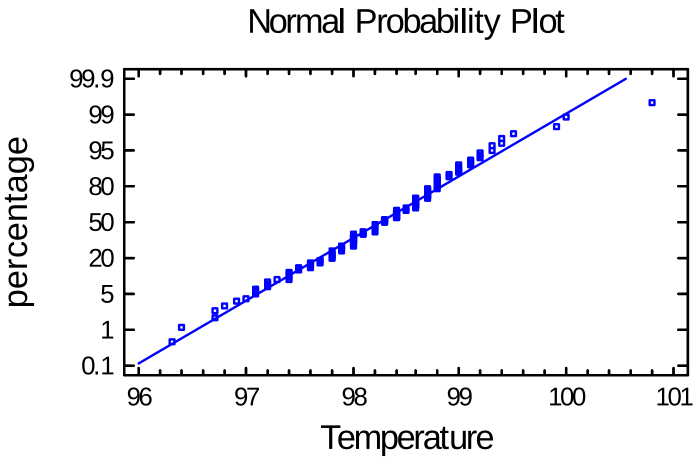
2. Fitting a least squares regression of the normal quantiles on the sorted data values.

$$\hat{\mu} = \text{- intercept / slope} \tag{33}$$

$$\hat{\sigma} = \text{1 / slope} \tag{34}$$

The first method is more robust to deviations from normality in the tails of the distribution, since it essentially relies only on the middle half. Outliers or long tails will have a greater influence on the fit using the least squares method.

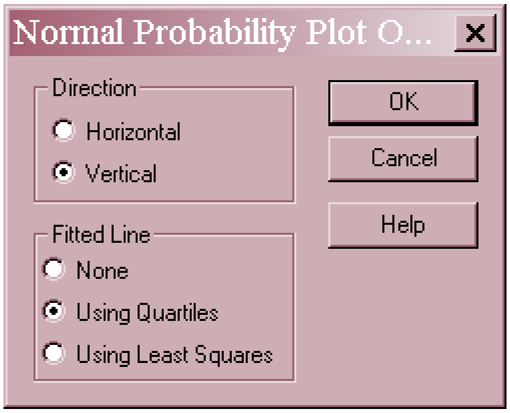
As is quite often the case, the least squares option shows a much closer fit to the temperature data:



Except for one data value, the other points are very close to the line.

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.

Pane Options



Direction
○ Horizontal
● Vertical

Fitted Line
○ None
● Using Quartiles
○ Using Least Squares

- **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 data. 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. It also includes bootstrap intervals for the mean, median, and standard deviation if requested.

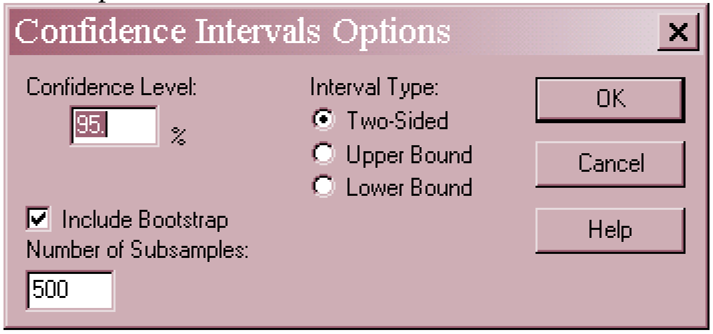
```
Confidence Intervals for Temperature
95.0% confidence interval for mean: 98.2492 +/- 0.127228   [98.122,98.3765]
95.0% confidence interval for standard deviation: [0.653586,0.835043]

Bootstrap Intervals
Mean: [98.1277,98.3808]
Standard deviation: [0.619503,0.832856]
Median: [98.15,98.4]
```

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 temperature in the sample was 98.25°, the mean in the population from which the data were sampled may well differ from that estimate by 0.13° 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: 95 %

Interval Type:
● Two-Sided
○ Upper Bound
○ Lower Bound

☑ Include Bootstrap

Number of Subsamples: 500

- **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 the sample data come from a distribution with a particular mean or standard deviation. For example, it is commonly assumed that the mean temperature of human beings is 98.6°. 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 and determine whether or not 98.6° is within the confidence interval.

2. Perform a formal statistical *hypothesis test*.

The *Hypothesis Tests* pane supports the latter approach.

t Test for the Mean

The top section of the output is shown below:

```
Hypothesis Tests for Temperature
Sample mean = 98.2492
Sample median = 98.3
Sample standard deviation = 0.733183

t-test
Null hypothesis: mean = 98.6
Alternative: not equal

Computed t statistic = -5.45482
P-Value = 4.37123E-7
Reject the null hypothesis for alpha = 0.05.
```

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

- **Null hypothesis**: a hypothesis such as $\mu = 98.6°$ 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 $\mu \neq 98.6°$ 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:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{35}$$

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 $t = -5.455$

- *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 very strong evidence that the data do not come from a population in which the mean equals 98.6°.

Tests for the Median

If the distribution from which the data 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 = 98.6
Alternative: not equal

Number of values below hypothesized median: 81
Number of values above hypothesized median: 39

Large sample test statistic = 3.74277 (continuity correction applied)
P-Value = 0.000182057
Reject the null hypothesis for alpha = 0.05.

signed rank test
Null hypothesis: median = 98.6
Alternative: not equal

Average rank of values below hypothesized median: 67.7222
Average rank of values above hypothesized median: 45.5

Large sample test statistic = 4.86 (continuity correction applied)
P-Value = 0.00000117545
Reject the null hypothesis for alpha = 0.05.
```

The *Sign Test* is based on comparing the number of observations below the hypothesized median to the number of observations above the median. A large discrepancy leads to rejection of the null hypothesis. The signed rank test ranks the absolute differences between the data and the hypothesized median from smallest to largest and compares the average rank of the observations 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 body temperature equals 98.6°.

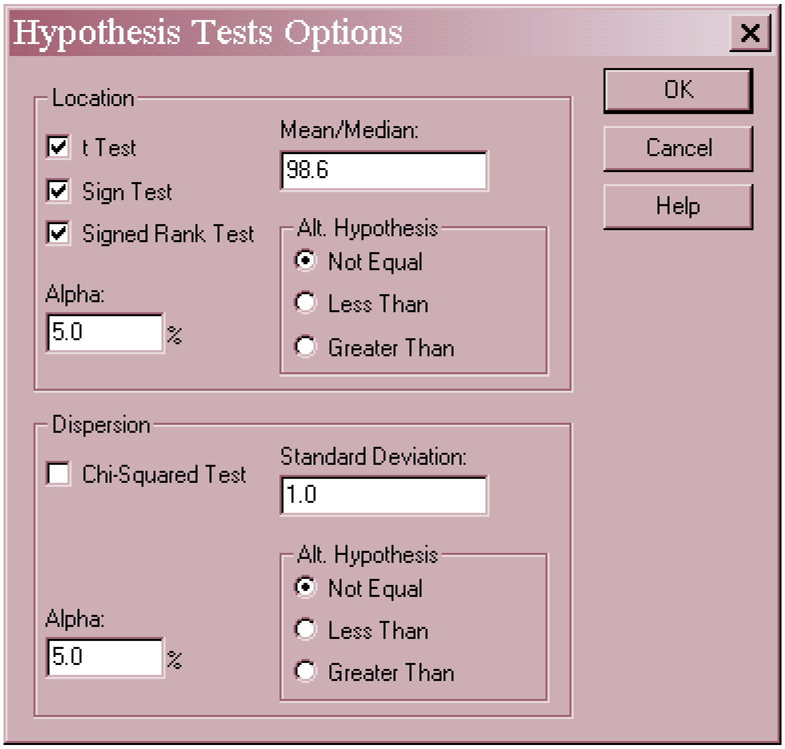
Test for the Standard Deviation

It is also possible to test hypotheses about the population standard deviation. The test statistic is

$$\mathrm{X}^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{36}$$

which is compared to a chi-squared distribution with $\nu = n - 1$ degrees of freedom. Small P-values lead to rejection of the standard deviation value σ_0 specified by the null hypothesis.

Pane Options



Hypothesis Tests Options

Location

☑ t Test

☑ Sign Test

☑ Signed Rank Test

Mean/Median: 98.6

Alpha: 5.0 %

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Dispersion

☐ Chi-Squared Test

Standard Deviation: 1.0

Alpha: 5.0 %

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

- **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.

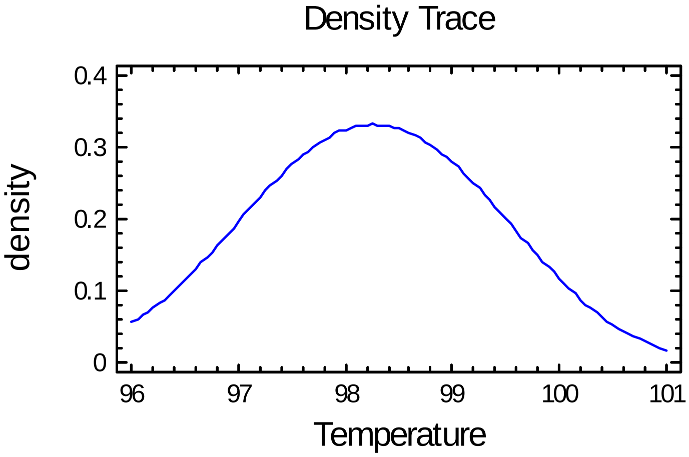
- **Standard Deviation**: σ_0, the value of the standard deviation specified by the null hypothesis.

- **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 $\mu < 98.6$ 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 data were sampled. It is created by counting the number of observations which fall within a window of fixed width moved across the range of the data.



The estimated density function is given by:

$$f(x) = \frac{1}{hn}\sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right) \tag{37}$$

where *h* is the width of the window in units of *X* and *W(u)* is a weighting function determined by the selection on the *Pane Options* dialog box. Two forms of weighting function are offered:

Boxcar Function

$$W(u) = \begin{cases} 1 & if\,|u| \le 1/2 \\ 0 & otherwise \end{cases} \tag{38}$$

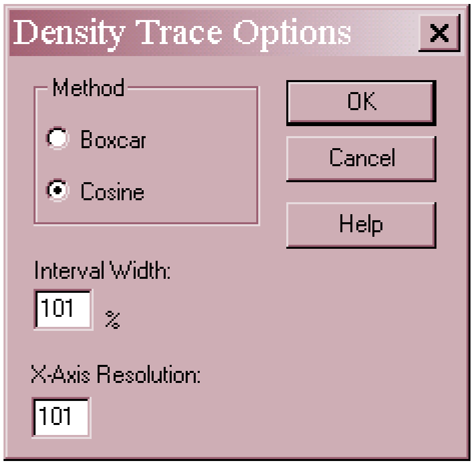
Cosine Function

$$W(u) = \begin{cases} 1 + \cos(2\pi u) & if\,|u| < 1/2 \\ 0 & otherwise \end{cases} \tag{39}$$

The latter selection usually gives a smoother result, with the desirable value of *h* depending on the size of the data sample.

For the sample data, the density trace closely resembles a normal distribution.

Pane Options



Density Trace Options

Method

○ Boxcar

● Cosine

Interval Width: 101 %

X-Axis Resolution: 101

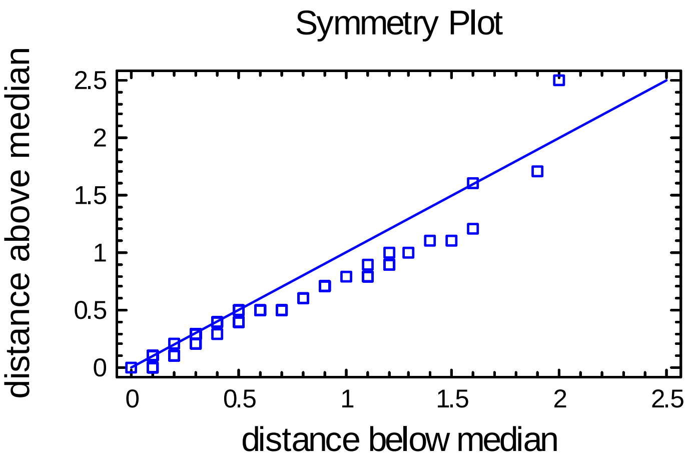
- **Method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog 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 data come from a symmetric distribution, i.e., a distribution that has a density function with the same shape on either side of the median.



To create this plot, the data values are sorted and then paired based on their location with respect to the median. For example, with 130 observations, the sorted points are paired as:

$(x_{(65)},x_{(66)})$, $(x_{(64)},x_{(67)})$, $(x_{(63)},x_{(68)})$, …, $(x_{(1)},x_{(100)})$

The distance of each pair below and above the median is plotted. If the data 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 above tends to deviate below the diagonal line over much of the range of X, which would indicate a longer lower tail than upper. A few unusual values at the end, however, disrupt that pattern.

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

Percentiles

1. Calculate the order statistics $x_{(j)}$ = *j-th* smallest data value.
2. For p-th percentile, let $q=p/100$. (40)
3. If *nq* is an integer, let

$$j_1=nq \tag{41}$$

$$j_2 = 1+nq \tag{42}$$

4. Else if *nq* is not an integer, let

$$j_1 = j_2 = floor(1+nq) \tag{43}$$

where the *floor* function returns the largest integer less than or equal to its argument.

5. The *p*-th percentile is then given by

$$\frac{x_{(j_1)}+x_{(j_2)}}{2} \tag{44}$$

Confidence Interval for Mean

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{45}$$

Confidence Interval for Standard Deviation

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}},\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}\right] \tag{46}$$

Sign Test

Given a hypothesized median θ_0, let

$$n_- = \text{number of } x_i < \theta_0 \tag{47}$$

$$n_+ = \text{number of } x_i > \theta_0 \tag{48}$$

Then

$$z = \frac{\max(n_-, n_+) - 0.5 - \frac{(n_- + n_+)}{2}}{\sqrt{\frac{n_- + n_+}{4}}} \tag{49}$$

is compared to a standard normal distribution.

Signed Rank Test

Given a hypothesized median θ_0, rank the deviations from the hypothesized median $|x_i - \theta_0|$. Let

$$T^- = \text{sum of ranks for all } x_i < \theta_0 \tag{50}$$

$$T^+ = \text{sum of ranks for all } x_i > \theta_0 \tag{51}$$

Then

$$z^- = \frac{T^- - 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{S}{48}}} \tag{52}$$

$$z^+ = \frac{T^+ - 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{S}{48}}} \tag{53}$$

where $n = n_- + n_+$ and S=0 unless there are tied observations. If there are g groups of tied observations, and t_j equals the size of the *j-th* tied group, then

$$S = \sum_{j=1}^{g} t_j (t_j - 1)(t_j + 1) \tag{54}$$

For a two-sided test, the larger of the two Z statistics is then compared to a standard normal distribution. For a one-sided test, only the statistic corresponding to the direction of the alternative hypothesis is used.