

## Checking for normality

(Adapted from Univariate Analysis and Normality Test Using SAS, STATA and SPSS by Hun Myoung Park)

(<http://www.indiana.edu/~statmath/stat/all/normality/normality.pdf>)

Descriptive statistics provide important information about variables. Mean, median, and mode measure the central tendency of a variable. Measures of dispersion include variance, standard deviation, range, and interquartile range (IQR). Researchers may draw a histogram, a stem-and-leaf plot, or a box plot to see how a variable is distributed.

Statistical methods are based on various underlying assumptions. One common assumption is that a random variable is normally distributed. In many statistical analyses, normality is often conveniently assumed without any empirical evidence or test. But normality is critical in many statistical methods. When this assumption is violated, interpretation and inference may not be reliable or valid.

There are two ways of testing normality (Table 1). Graphical methods display the distributions of random variables or differences between an empirical distribution and a theoretical distribution (e.g., the standard normal distribution). Numerical methods present summary statistics such as skewness and kurtosis, or conduct statistical tests of normality. Graphical methods are intuitive and easy to interpret, while numerical methods provide more objective ways of examining normality.

Table 1	Graphical Methods	Numerical Methods
Descriptive	Stem-and-leaf plot, (skeletal) box plot Histogram	Skewness Kurtosis
Theory-driven	P-P plot Q-Q plot	Shapiro-Wilk, Shapiro- Francia test Kolmogorov-Smirnov test (Lillefors test) Anderson- Darling/Cramer-von Mises tests Jarque-Bera test, Skewness- Kurtosis test

Graphical and numerical methods are either descriptive or theory-driven. The dot plot and histogram, for instance, are descriptive graphical methods, while skewness and kurtosis are descriptive numerical methods. The P-P and Q-Q plots are theory-driven graphical methods for testing normality, whereas the Shapiro-Wilk  $W$  and Jarque-Bera tests are theory-driven numerical methods.

## Results from the Summarize

The variable has a mean close to zero and a unit variance. The kurtosis is close to 3 and skewness approach zero.

```
. summarize x, detail
```

```

      x
-----
Percentiles      Smallest
1%      -2.338288    -3.534935
5%      -1.683915    -3.058428
10%     -1.315623    -2.958493
25%     -.6848071    -2.844061

50%      .0094563

75%      .7548931
90%      1.339141
95%      1.718761
99%      2.416142

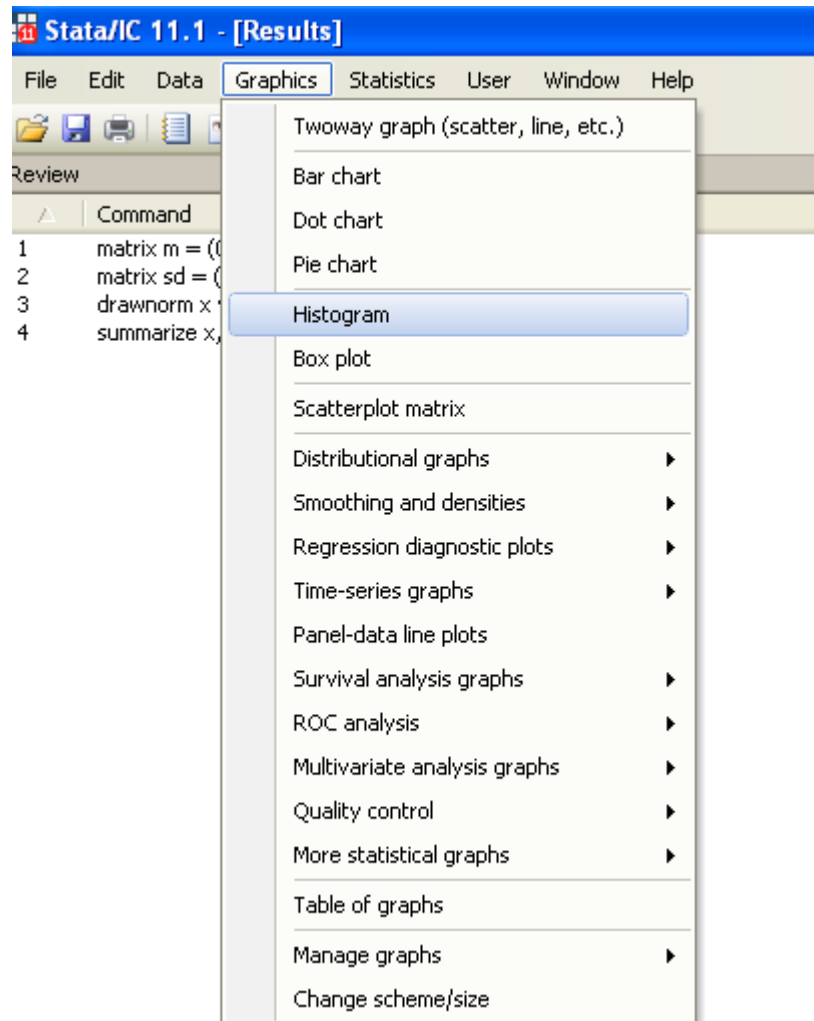
      Largest
      2.9914

      Obs          2000
      Sum of wgt.  2000

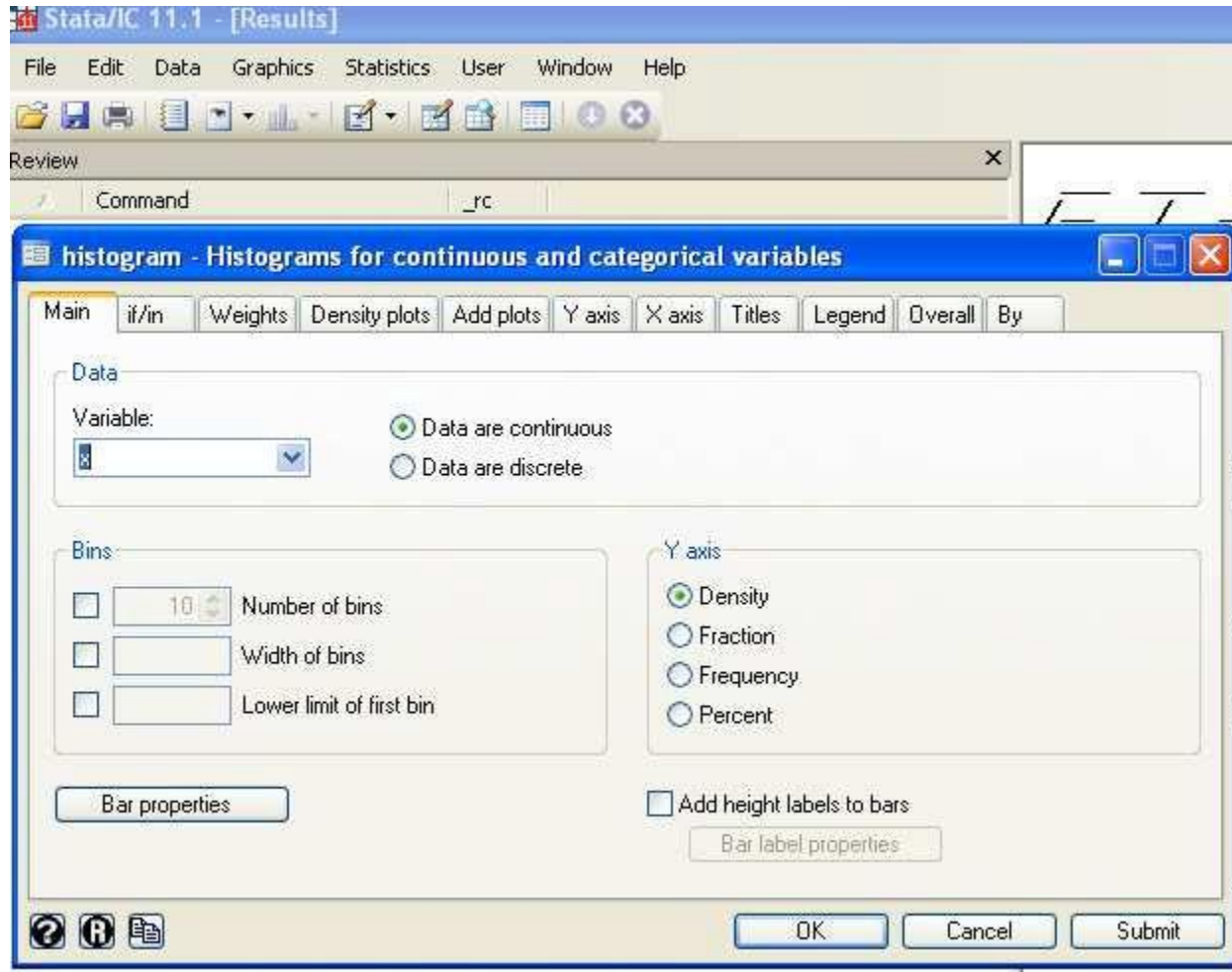
      Mean          .0270618
      Std. Dev.     1.028202

      Variance      1.057198
      Skewness      .0043508
      Kurtosis      2.88071
```

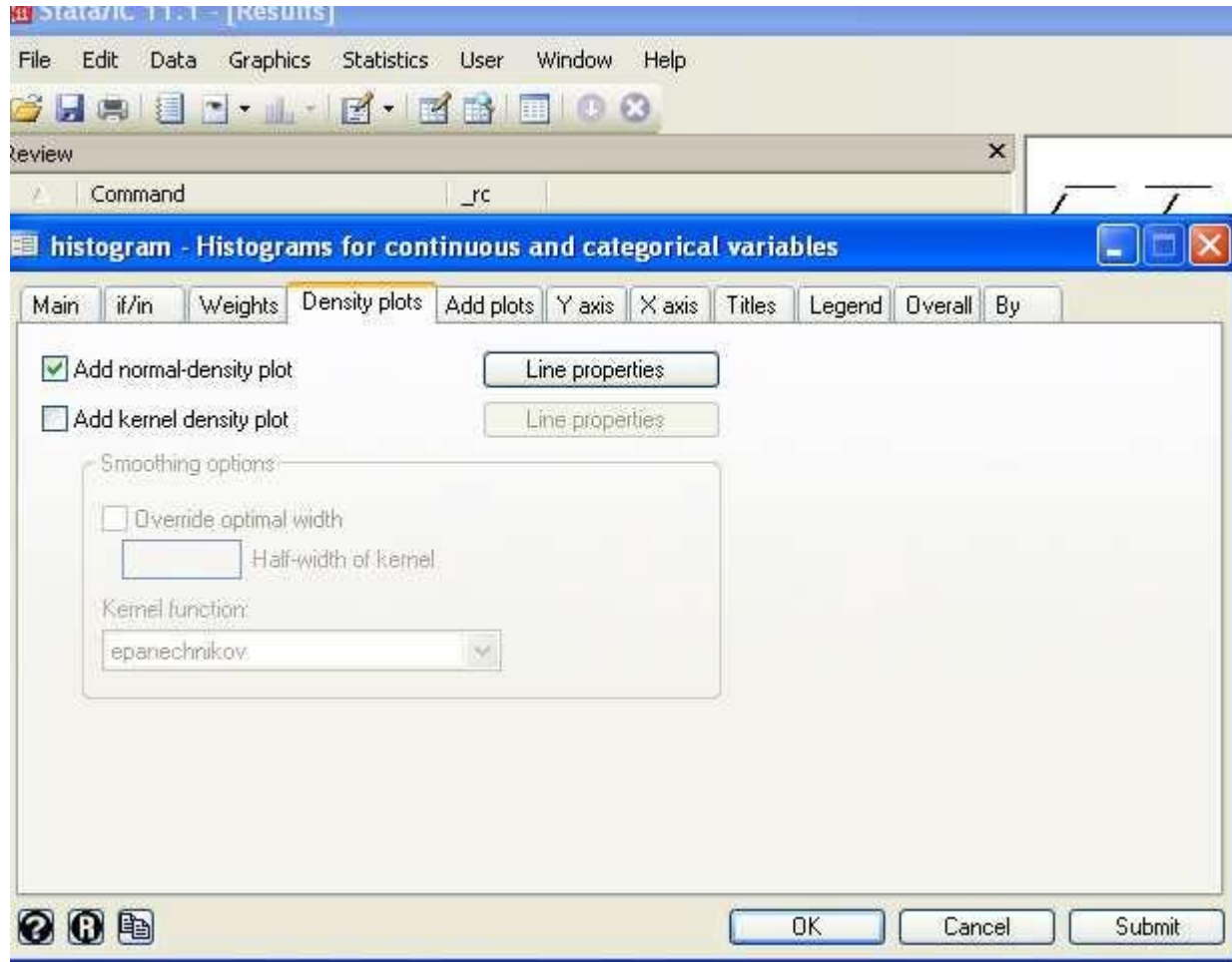
## Graphical Methods



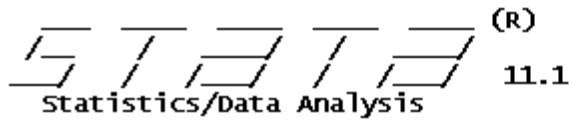
A histogram is the most widely used graphical method. A histogram can be generated using the Graphics menu (Graphics → Histogram)



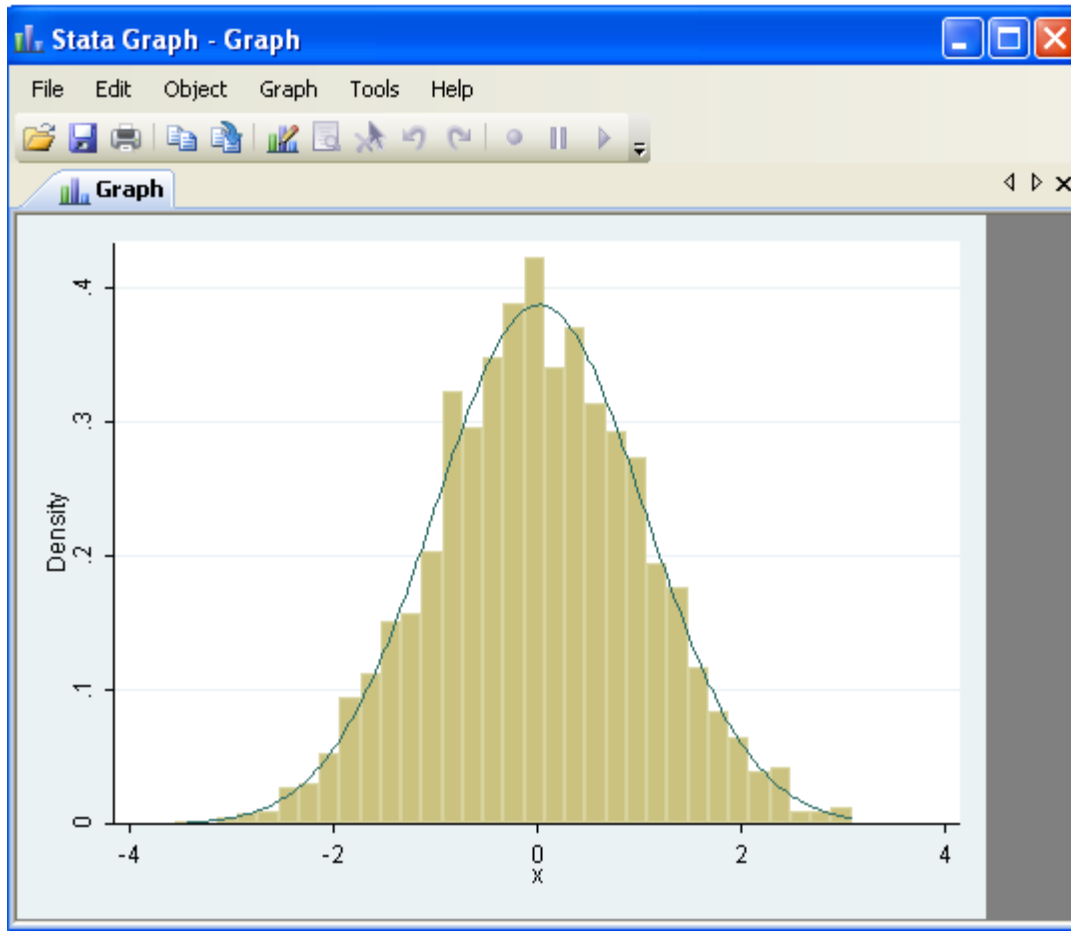
Select variable from the variable list. This also allows you to add a normal density curve to the histogram. By clicking on the Density plots you will be able to add a normal density curve on the histogram.

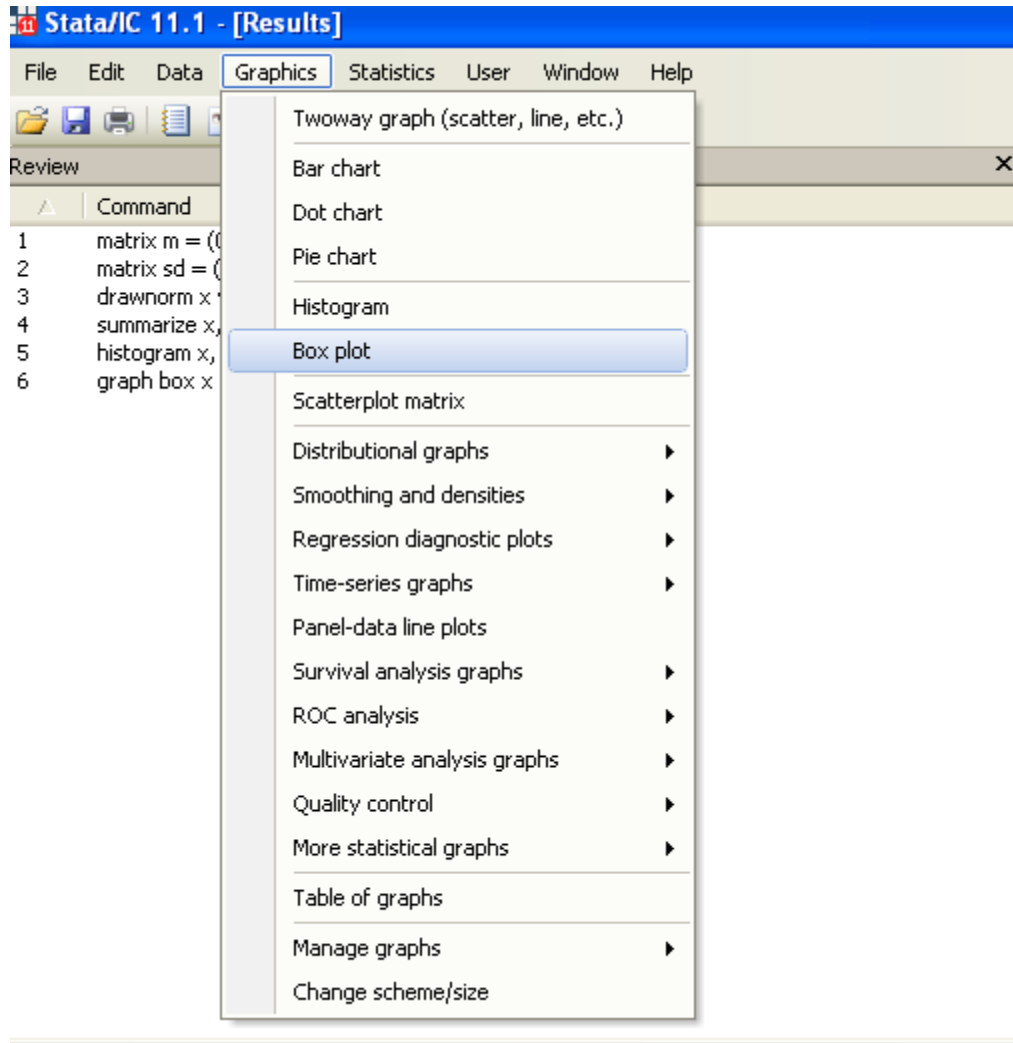


Check Add normal-density plot  
Click Ok




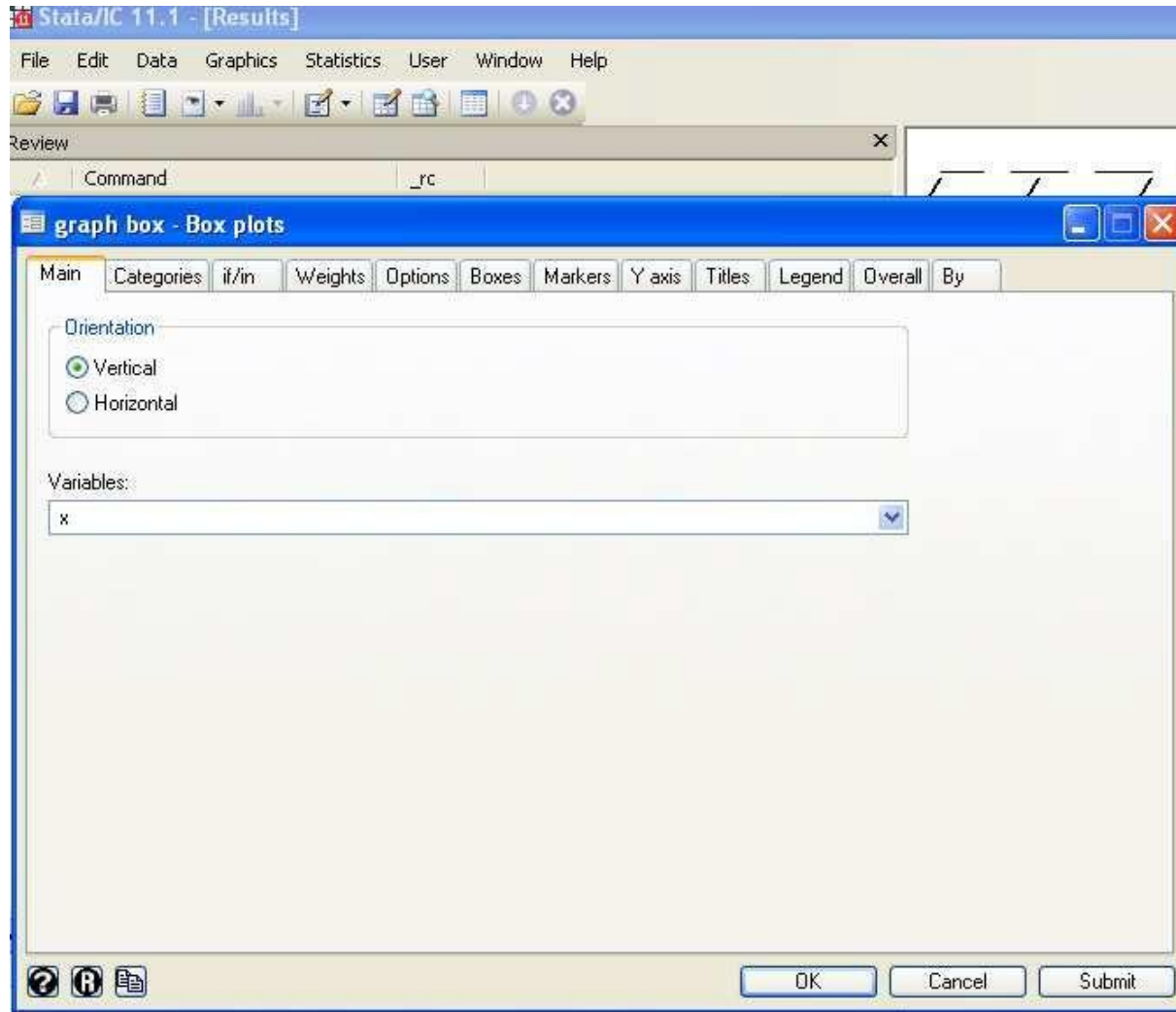
Copyright 2009 StataCorp LP  
StataCorp  
4905 Lakeway Drive  
College Station, Texas 77845 U.S.A.  
800-STATA-PC <http://www.stata.com>  
979-696-4600 [stata@stata.com](mailto:stata@stata.com)  
979-696-4601 (fax)





A box plot presents the minimum, 25<sup>th</sup> percentile (1st quartile), 50<sup>th</sup> percentile (median), 75<sup>th</sup> percentile (3rd quartile), and maximum in a box and lines.<sup>1</sup> Outliers, if any, appear at the outsides of (adjacent) minimum and maximum lines. As such, a box plot effectively summarizes these major percentiles using a box and lines. If a variable is normally distributed, its 25th and 75th percentile are symmetric, and its median and mean are located at the same point exactly in the center of the box.

To generate a box plot, click on Graphics  Box plot



Select variable from the variable list.

Click Ok

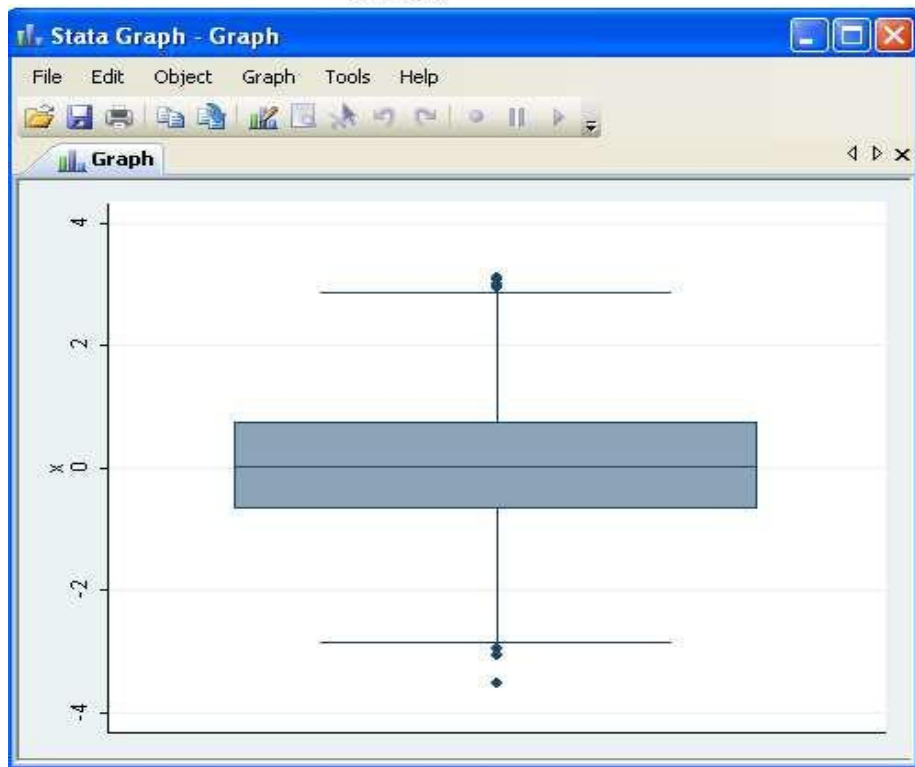


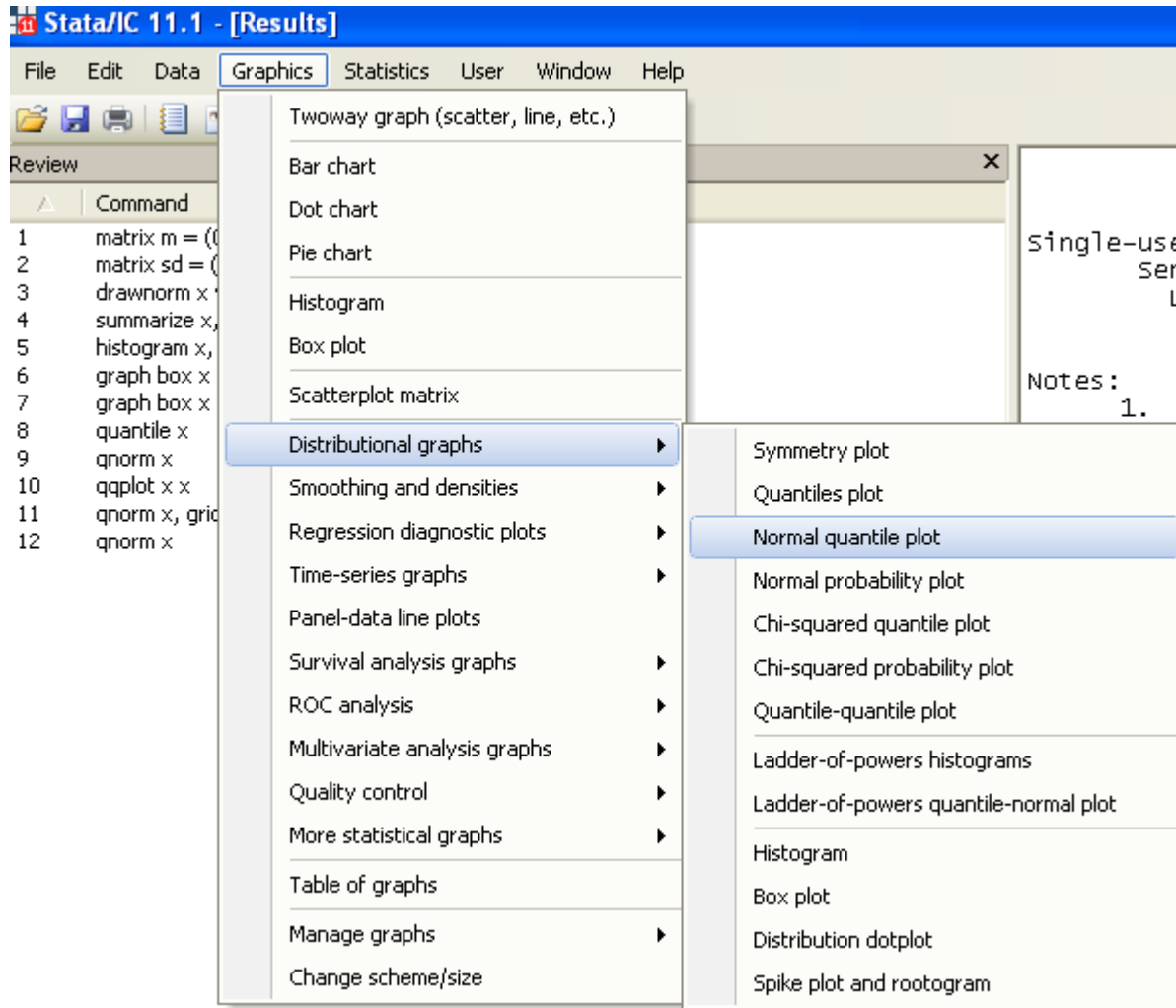


Copyright 2009 stataCorp  
StataCorp  
4905 Lakeway Drive  
College Station, Texas 77845  
800-STATA-PC http://www.stata.com  
979-696-4600 stat@stata.com  
979-696-4601 (fax)

The both extremes (i.e., minimum and maximum), the 25th, 50th, and 75th percentiles are symmetrically arranged in the box plot.

single-user stata perpetual license:  
Serial number: 30110536678  
Licensed to: Michelle Edwards  
DRC ug

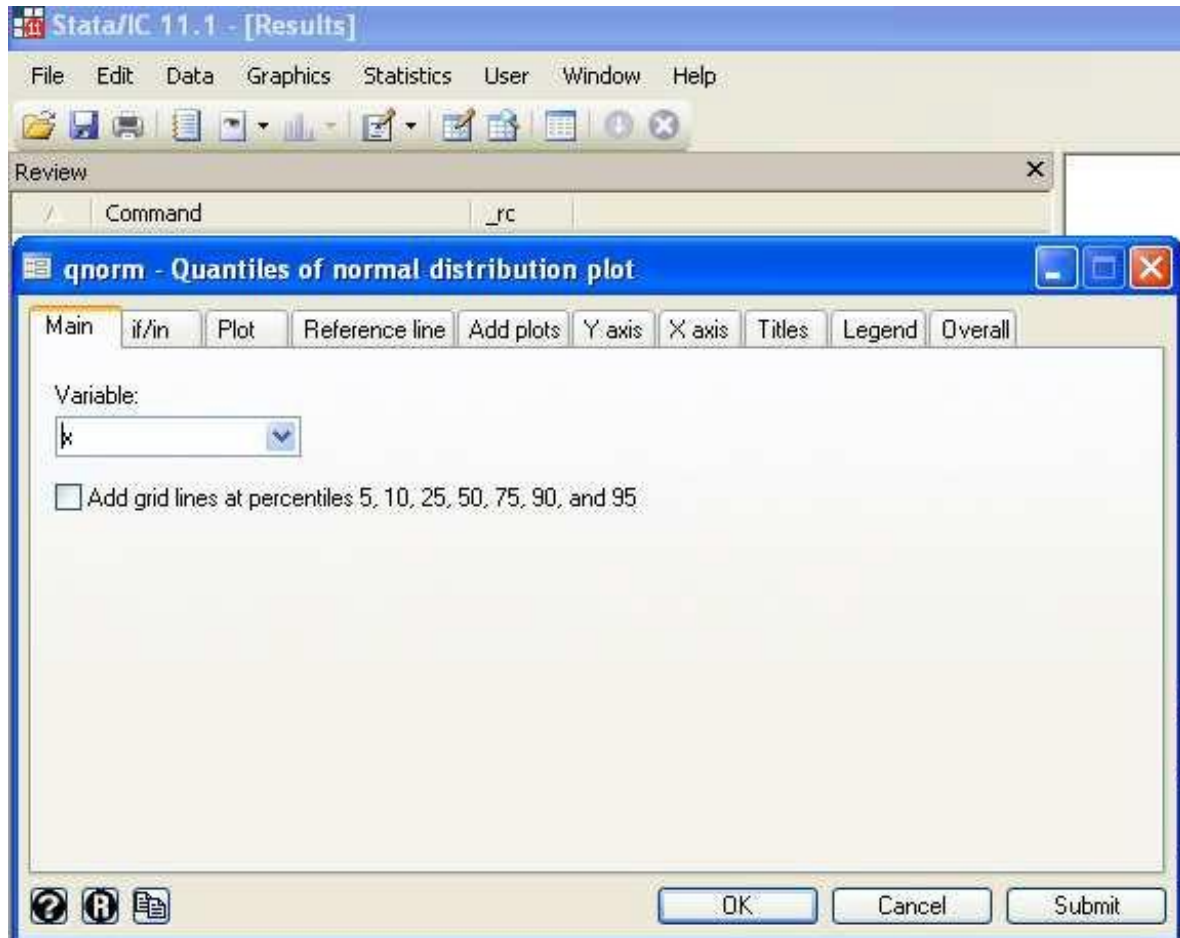




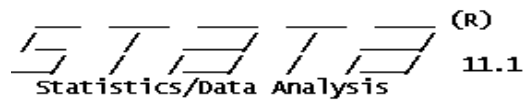
The Quantile plot (Q-Q plot) compares ordered values of a variable with quantiles of a specific theoretical distribution (i.e., the normal distribution). If two distributions match, the points on the plot will form a linear pattern passing through the origin with a unit slope. P-P and Q-Q plots are used to see how well a theoretical distribution models the empirical data.

To generate a box plot

Click on Graphics → Distributional graph → Normal quantile plot



Select variable from the variable list.  
Click Ok

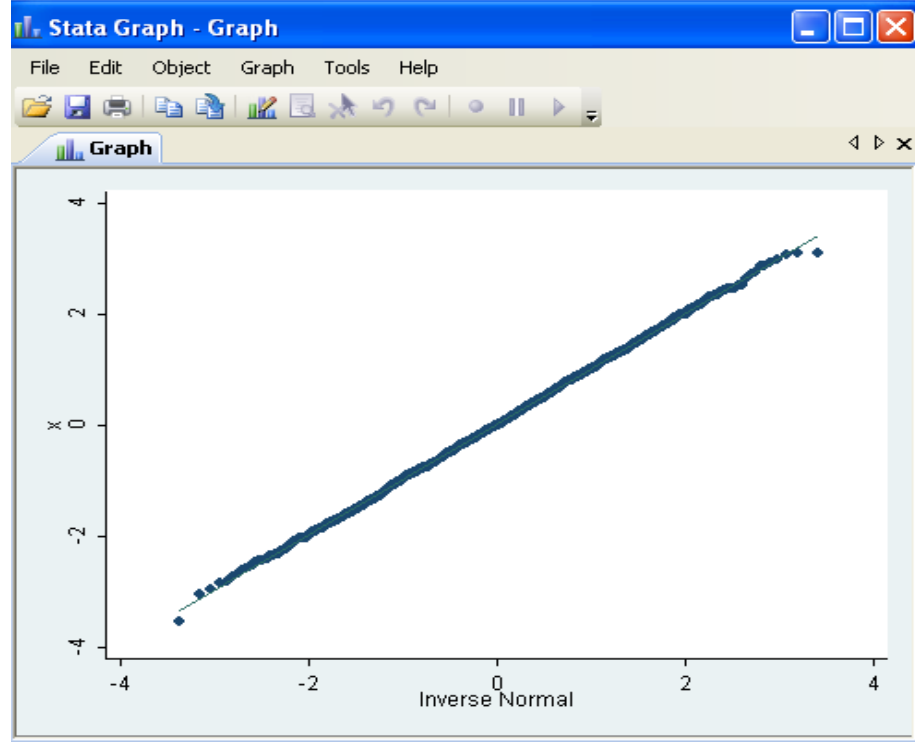


Copyright 2009 StataCorp  
StataCorp  
4905 Lakeway Drive  
College Station, Texas 77  
800-STATA-PC [http:  
stata](http://www.stata.com)  
979-696-4600  
979-696-4601 (fax)

single-user stata perpetual license:  
serial number: 30110536678  
Licensed to: Michelle Edwards  
DRC ug


## notes:

1. (/m# option or -set memory-) 10.00 MB allocated to da
2. New update available; type `-update all-`



The Q-Q plot indicate no significant deviation from the fitted line.

**Theory Driven Statistics**

Skewness and kurtosis are based on the empirical data. The numerical methods for testing normality compare empirical data with a theoretical distribution. Widely used methods include the Kolmogorov-Smirnov (K-S) D test (Lilliefors test), Shapiro-Wilk test, Anderson-Darling test, and Cramer-von Mises test (SAS Institute 1995).<sup>4</sup> The K-S D test and Shapiro-Wilk W test are commonly used. To generate Tests of Normality, click on **Statistics**  **Distributional plots and tests.**

shapiro-wilk w test for normal data

variable	obs	w	v	z	Prob>z
x	2000	0.99946	0.642	-1.127	0.87014

. sfrancia x

Shapiro-Francia w' test for normal data

variable	obs	w'	v'	z	Prob>z
x	2000	0.99961	0.417	-2.444	0.99273

In both Shapiro-Wilk and Shapiro-Francia statistic fail to reject the null hypothesis of normality.