Univariate Analysis and Normality Test Using SAS, STATA, and SPSS

Hun Myoung Park

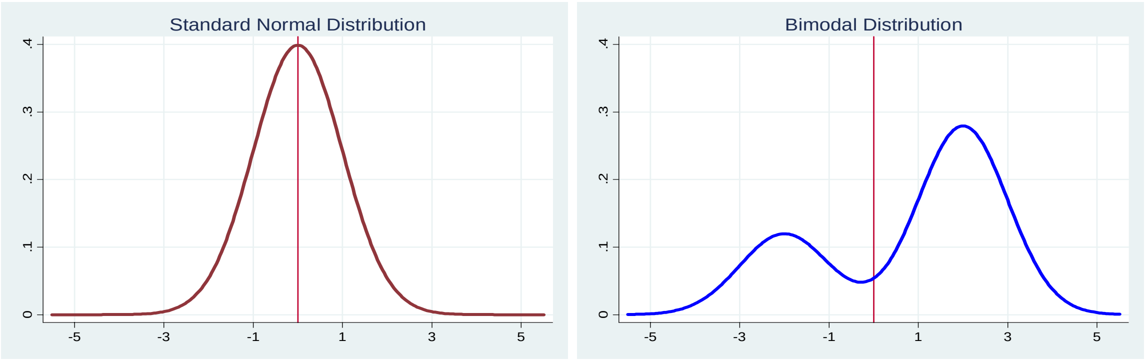
This document summarizes graphical and numerical methods for univariate analysis and normality test, and illustrates how to test normality using SAS 9.1, STATA 9.2 SE, and SPSS 14.0.

1. Introduction

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 interquantile 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.

Figure 1. Comparing the Standard Normal and a Bimodal Probability Distributions



T-test and ANOVA (Analysis of Variance) compare group means, assuming variables follow normal probability distributions. Otherwise, these methods do not make much sense. Figure 1 illustrates the standard normal probability distribution and a bimodal distribution. How can you compare means of these two random variables?

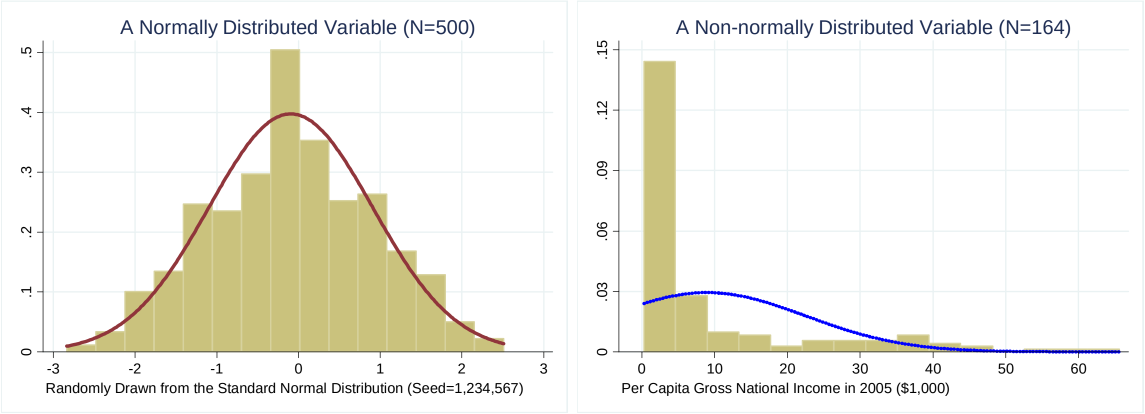
There are two ways of testing normality (Table 1). Graphical methods display the distributions of random variables or differences between an empirical distribution and a theoretical distribution (e.g., the standard normal distribution). Numerical methods present summary statistics such as skewness and 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 versus Numerical Methods

	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.

Figure 2. Histograms of Normally and Non-normally Distributed Variables



Three variables are employed here. The first variable is unemployment rate of Illinois, Indiana, and Ohio in 2005. The second variable includes 500 observations that were randomly drawn from the standard normal distribution. This variable is supposed to be normally distributed with zero mean and a variance of 1 (left plot in Figure 2). An example of a non-normal distribution is per capita gross national income (GNI) in 2005 of 164 countries in the world. GNIP is severely skewed to the right and is least likely to be normally distributed (right plot in Figure 2). See the Appendix for details about these variables.

2. Graphical Methods

Graphical methods visualize the distribution of a random variable and compare the distribution to a theoretical one using plots. These methods are either descriptive or theory-driven. The former method is based on the empirical data, whereas the latter considers both empirical and theoretical distributions.

2.1 Descriptive Plots

Among frequently used descriptive plots are the stem-and-leaf-plot, dot plot, (skeletal) box plot, and histogram. When N is small, a stem-and-leaf plot or dot plot is useful to summarize data. A stem-and-leaf plot and dot plot work well for continuous or event count variables. Figure 3 presents the stem-and-leaf plots for unemployment rates of three states.

Figure 3. Stem-and-Leaf Plot of Unemployment Rate of Illinois, Indiana, Ohio

```
-> state = IL                        -> state = IN                       -> state = OH

Stem-and-leaf plot for rate(Rate)    Stem-and-leaf plot for rate(Rate)   Stem-and-leaf plot for rate (Rate)

rate rounded to nearest multiple     rate rounded to nearest multiple    rate rounded to nearest multiple
of .1                                of .1                               of .1
plot in units of .1                  plot in units of .1                 plot in units of .1

  3. | 7889                            3* | 1                               3* | 8
  4* | 011122344                       3. | 89                              4* | 014577899
  4. | 5566666666677778888999          4* | 012234                          5* | 0122333344555666777888888888999
  5* | 0011122222333333344444          4. | 566666778889999                 6* | 001111122222233444446678899
  5. | 5555667777777888999             5* | 000001112222222233344           7* | 01223335677
  6* | 000011222333444                 5. | 555666666777889                 8* | 1223338
  6. | 555579                          6* | 002222233344                    9* | 99
  7* | 0033                            6. | 5666677889                     10* | 1
  7. |                                 7* | 1113344                        11* |
  8* | 0                               7. | 67                             12* |
  8. | 8                               8* | 14                             13* | 3
```

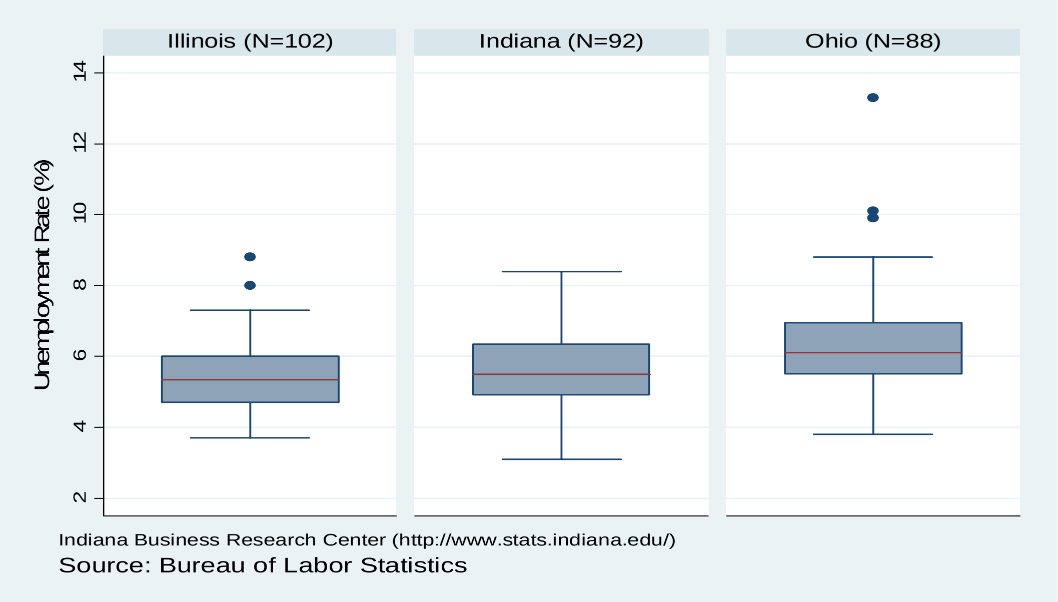
A box plot presents the minimum, 25th percentile (1st quartile), 50th percentile (median), 75th percentile (3rd quartile), and maximum in a box and lines.[1] Outliers, if any, appear at the outsides of (adjacent) minimum and maximum lines. As such, a box plot effectively summarizes these major percentiles 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.[2]

In Figure 4, you should see outliers in Illinois and Ohio that affect the shapes of corresponding boxes. In contrast, the Indiana unemployment rate does not have outliers, and its symmetric box implies that the rate appears to be normally distributed.

[1] The first quartile cuts off lowest 25 percent of data; the second quartile cuts data set in half; and the third quartile cuts off lowest 75 percent or highest 25 percent of data. See http://en.wikipedia.org/wiki/Quartile.

[2] SAS reports a mean as "+" between (adjacent) minimum and maximum lines.

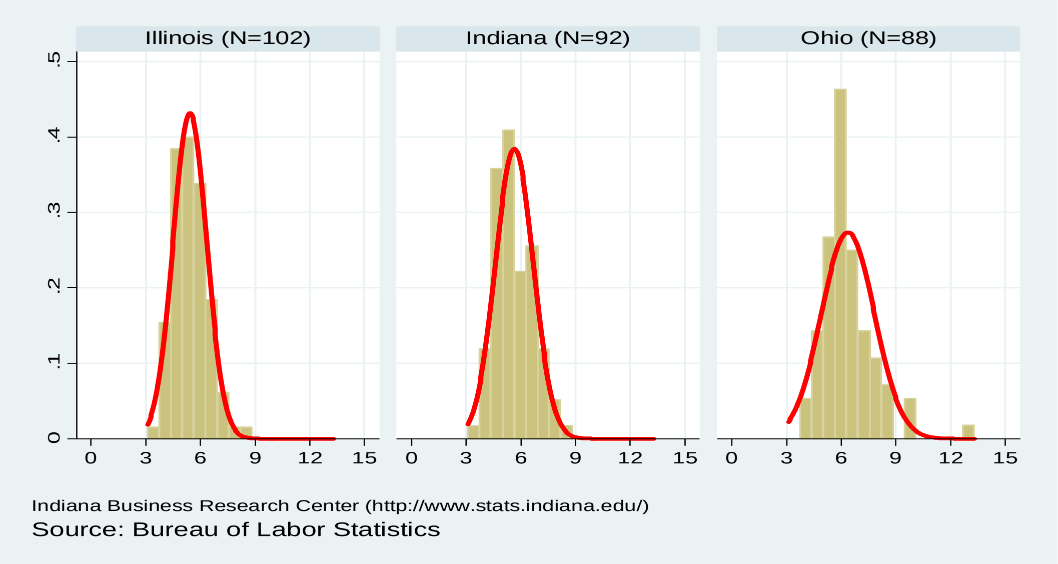
Figure 4. Box Plots of Unemployment Rates of Illinois, Indiana, and Ohio



Indiana Business Research Center (http://www.stats.indiana.edu/)
Source: Bureau of Labor Statistics

The histogram graphically shows how each category (interval) accounts for the proportion of total observations and is more appropriate for large N samples (Figure 5).

Figure 5. Histograms of Unemployment Rates of Illinois, Indiana and Ohio

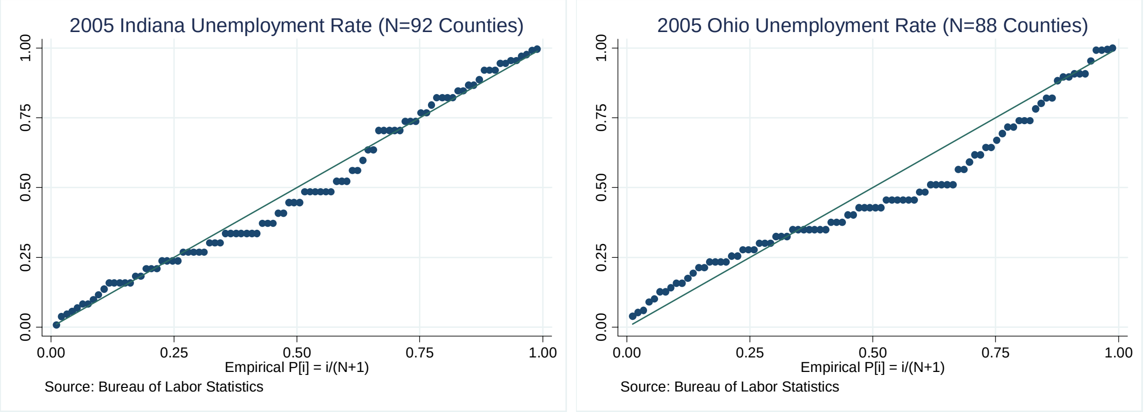


Indiana Business Research Center (http://www.stats.indiana.edu/)
Source: Bureau of Labor Statistics

2.2 Theory-driven Plots

The P-P and Q-Q plots are considered here. The probability-probability plot (P-P plot or percent plot) compares an empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function (e.g., the standard normal distribution function). In Figure 6, Ohio appears to deviate more from the fitted line than Indiana.

Figure 6. P-P Plots of Unemployment Rates of Indiana and Ohio (Year 2005)



2005 Indiana Unemployment Rate (N=92 Counties)

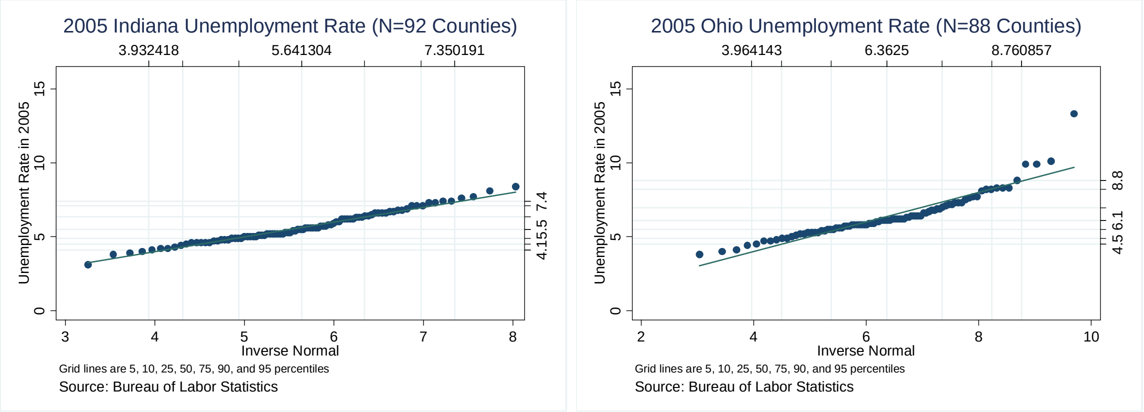
2005 Ohio Unemployment Rate (N=88 Counties)

Source: Bureau of Labor Statistics

Source: Bureau of Labor Statistics

Similarly, the quantile-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. In Figure 7, Indiana appears to have a smaller variation in its unemployment rate than Ohio. In contrast, Ohio appears to have a wider range of outliers in the upper extreme.

Figure 7. Q-Q Plots of Unemployment Rates of Indiana and Ohio (Year 2005)



2005 Indiana Unemployment Rate (N=92 Counties)

Grid lines are 5, 10, 25, 50, 75, 90, and 95 percentiles

Source: Bureau of Labor Statistics

2005 Ohio Unemployment Rate (N=88 Counties)

Grid lines are 5, 10, 25, 50, 75, 90, and 95 percentiles

Source: Bureau of Labor Statistics

Detrended normal P-P and Q-Q plots depict the actual deviations of data points from the straight horizontal line at zero. No specific pattern in a detrended plot indicates normality of the variable. SPSS can generate detrended P-P and Q-Q plots.

Although visually appealing, these graphical methods do not provide objective criteria to determine normality of variables. Interpretations are thus a matter of judgments.

3. Numerical Methods

Numerical methods use descriptive statistics and statistical tests to examine normality.

3.1 Descriptive Statistics

Measures of dispersion such as variance reveal how observations of a random variable deviate from the mean. The (sample) variance of a variable is computed from the second central moment.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Skewness is based pm the third standardized moment that measures the degree of symmetry of a probability distribution. If skewness is greater than zero, the distribution is skewed to the right, having more observations on the left.

$$\frac{E[(x-\mu)^3]}{\sigma^3} = \frac{\sum (x_i - \bar{x})^3}{s^3(n-1)} = \frac{\sqrt{n-1}\sum (x_i - \bar{x})^3}{[\sum (x_i - \bar{x})^2]^{3/2}}$$

The following is a list of descriptive statistics for unemployment rates of three states. Indiana has the smallest skewness of .3416 that is close to zero. In contrast, Ohio has a large skewness of 1.6653 indicating many observations on the left of the probability distribution.

```
 state |        N      mean    median       max       min  variance  skewness  kurtosis
-------+--------------------------------------------------------------------------------
    IL |      102  5.421569      5.35       8.8       3.7  .8541837  .6570033  3.946029
    IN |       92  5.641304       5.5       8.4       3.1  1.079374  .3416314  2.785585
    OH |       88  6.3625         6.1      13.3       3.8  2.126049  1.665322  8.043097
-------+--------------------------------------------------------------------------------
 Total |      282  5.786879      5.65      13.3       3.1  1.473955   1.44809  8.383285
----------------------------------------------------------------------------------------
```

Kurtosis, based on the fourth central moment, measures the thinness of tails or "peakedness" of a probability distribution.

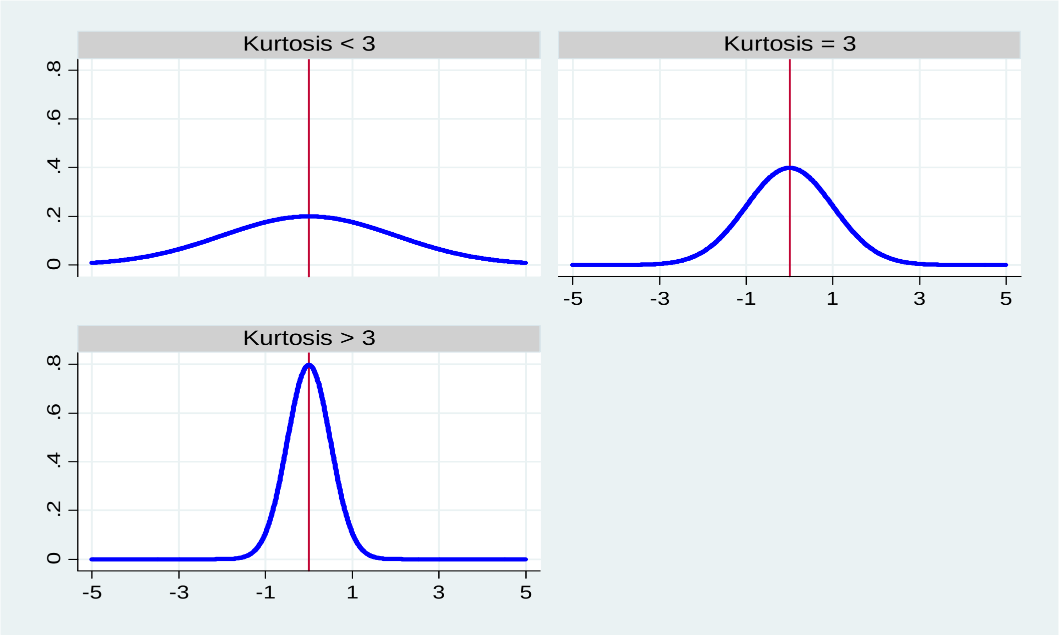
$$\frac{E[(x-\mu)^4]}{\sigma^4} = \frac{\sum (x_i - \bar{x})^4}{s^4(n-1)} = \frac{(n-1)\sum (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2}$$

Like skewness, kurtosis also shows how the distribution of a variable deviates from a normal distribution. If kurtosis of a random variable is less than three (or if kurtosis-3 is less than zero), the distribution has thicker tails and a lower peak compared to a normal distribution (first plot in Figure 8).[3] See the kutosis of 2.7856 of Indiana. In contrast,

[3] SAS and SPSS produce (kurtosis -3), while STATA returns the kurtosis. SAS uses its weighted kurtosis formula with the degree of freedom adjusted. So, if N is small, SAS, STATA, and SPSS may report different kurtosis.

kurtosis larger than 3 indicates a higher peak and thin tails (last plot). Note that Ohio has a large kurtosis of 8.0431. A normally distributed random variable should have skewness and kurtosis near zero and three, respectively (second plot in Figure 8).

Figure 8. Probability Distributions with Different Kurtosis



3.2 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).[4] The K-S D test and Shapiro-Wilk W test are commonly used. The K-S, Anderson-Darling, and Cramer-von Misers tests are based on the empirical distribution function (EDF), which is defined as a set of N independent observations x_1, x_2, …x_n with a common distribution function F(x) (SAS 2004).

Table 2. Numerical Methods for Testing Normality

Test	Stat.	N	Dist.	SAS	STATA	SPSS
Jarque-Bera	χ^2		$\chi^2(2)$	-	-	-
Skewness-Kurtosis	χ^2	9≤N	$\chi^2(2)$	-	`.sktest`	-
Shapiro-Wilk	W	7≤N≤ 2,000	-	YES	`.swilk`	YES
Shapiro-Francia	W'	5≤N≤ 5,000	-	-	`.sfrancia`	-
Kolmogorov-Smirnov	D		EDF	YES	*	YES
Cramer-vol Mises	W^2		EDF	YES	-	-
Anderson-Darling	A^2		EDF	YES	-	-

* STATA `.ksmirnov` command is not used for testing normality.

[4] The UNIVARIATE and CAPABILITY procedures have the NORMAL option to produce four statistics.

The Shapiro-Wilk W is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk 1965).[5] The statistic is positive and less than or equal to one. Being close to one indicates normality. The W statistic requires that the sample size is greater than or equal to 7 and less than or equal to 2,000 (Shapiro and Wilk 1965).[6]

$$W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$$

where a'=(a_1, a_2, …, a_n) $= m'V^{-1}[m'V^{-1}V^{-1}m]^{-1/2}$, m'=($m_1$, m_2, …, m_n) is the vector of expected values of standard normal order statistics, V is the n by n covariance matrix, x'=(x_1, x_2, …, x_n) is a random sample, and $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

The Shapiro-Francia (S-F) W' test is an approximate test that modifies the Shapro-Wilk W. The S-F statistic uses b'=(b_1, b_2, …, b_n) $= m'(m'm)^{-1/2}$ instead of a'. The statistic was developed by Shapiro and Francia (1972) and Royston (1983). The recommended sample sizes for the STATA `.sfrancia` command range from 5 to 5,000 (STATA 2005). SAS and SPSS do not support this statistic. Table 3 summarizes test statistics for 2005 unemployment rates of Illinois, Indiana, and Ohio.

Table 3. Normality Test for 2005 Unemployment Rates of Illinois, Indiana, and Ohio

State	Illinois		Indiana		Ohio	
	Test	P-value	Test	P-value	Test	P-value
Shapiro-Wilk ^{sas}	.9714	.0260	.9841	.3266	.8858	.0001
Shapiro-Wilk ^{stata}	.9728	.0336	.9855	.4005	.8869	.0000
Shapiro-Francia ^{stata}	.9719	.0292	.9858	.3545	.8787	.0000
Kolmogorov-Smirnov ^{sas}	.0583	.1500	.0919	.0539	.1602	.0100
Cramer-von Misers ^{sas}	.0606	.2500	.1217	.0582	.4104	.0050
Anderson-Darling ^{sas}	.4534	.2500	.6332	.0969	2.2815	.0050
Jarque-Bera	12.2928	.0021	1.9458	.3380	149.5495	.0000
Skewness-Kurtosis ^{stata}	10.59	.0050	1.99	.3705	43.75	.0000

The SAS UNIVARIATE and CAPABILITY procedures perform the Kolmogorov-Smirnov D, Anderson-Darling A^2, and Cramer-von Misers W^2 tests, which are useful especially when N is larger than 2,000.

3.3 Jarque-Bera (Skewness-Kurtosis) Test

The test statistics mentioned in the previous section tend to reject the null hypothesis when N becomes large. Given a large number of observations, the Jarque-Bera test and Skewness-Kurtosis test will be the alternatives for testing normality.

[5] The W statistic was constructed by considering the regression of ordered sample values on corresponding expected normal order statistics, which for a sample from a normally distributed population is linear (Royston 1982). Shapiro and Wilk's (1965) original W statistic is valid for the sample sizes between 3 and 50, but Royston extended the test by developing a transformation of the null distribution of W to approximate normality throughout the range between 7 and 2000.

[6] STATA `.swilk` command, based on Shapiro and Wilk (1965) and Royston (1992), can be used with from 4 to 2000 observations (STATA 2005).

The Jarque-Bera test, a type of Lagrange multiplier test, was developed to test normality, heteroscedasticy, and serial correlation (autocorrelation) of regression residuals (Jarque and Bera 1980). The Jarque-Bera statistic is computed from skewness and kurtosis and asymptotically follows the chi-squared distribution with two degrees of freedom.

$n\left[\frac{skewness^2}{6}+\frac{(kurtosis-3)^2}{24}\right]\sim\chi^2(2)$, where n is the number of observations.

The above formula gives a penalty for increasing the number of observations that implies a good asymptotic property of the Jarque-Bera test. The computation for 2005 unemployment rates is as follows.[7]

```
12.292825 = 102*(0.66685022^2/6 + 1.0553068^2/24)
1.9458304 = 92*(0.34732004^2/6 + (-0.1583764)^2/24)
149.54945 = 88*(1.69434105^2/6 + 5.4132289^2/24)
```

The STATA Skewness-Kurtosis test is based on D'Agostino, Belanger, and D'Agostino, Jr. (1990) and Royston (1991) (STATA 2005). Note that in Ohio the Jarque-Bera statistic of 150 is quite different from the S-K statistic of 44 (see Table 3).

Table 4. Comparison of Methods for Testing Normality

N	10	100	500	1,000	5,000	10,000
Mean	.5240	-.0711	-.0951	-.0097	-.0153	-.0192
Standard deviation	.9554	1.0701	1.0033	1.0090	1.0107	1.0065
Minimum	-.8659	-2.8374	-2.8374	-2.8374	-3.5387	-3.9838
1stquantile	-.2372	-.8674	-.8052	-.7099	-.7034	-.7121
Median	.6411	-.0625	-.1196	-.0309	-.0224	-.0219
3rdquantile	1.4673	.7507	.6125	.7027	.6623	.6479
Maximum	1.7739	1.9620	2.5117	3.1631	3.5498	4.3140
Skewness ^{sas}	-.1620	-.2272	-.0204	.0100	.0388	.0391
Kurtosis-3 ^{sas}	-1.4559	-.5133	-.3988	-.2633	-.0067	-.0203
Jarque-Bera	.9269	1.9580	3.3483	2.9051	1.2618	2.7171
	(.6291)	(.3757)	(.1875)	(.2340)	(.5321)	(.2570)
Skewness ^{stata}	-.1366	-.2238	-.0203	.0100	.0388	.0391
Kurtosis ^{stata}	1.6310	2.4526	2.5932	2.7320	2.9921	2.9791
S-K ^{stata}	1.52	2.52	4.93	3.64	1.26	2.70
	(.4030)	(.2843)	(.0850)	(.1620)	(.5330)	(.2589)
Shapiro-Wilk W	.9359	.9840	.9956	.9980	*.9998*	*.9999*
	(.5087)	(.2666)	(.1680)	(.2797)	*(.8727)*	*(.8049)*
Shapiro-F W'^{stata}	.9591	.9873	.9965	.9983	.9998	*.9998*
	(.7256)	(.3877)	(.2941)	(.4009)	(.1000)	*(.1000)*
Kolmogorov-S D^{sas}	.1382	.0708	.0269	.0180	.0076	.0073
	(.1500)	(.1500)	(.1500)	(.1500)	(.1500)	(.1500)
Cramer-M W$^{2\ sas}$	.0348	.0793	.0834	.0607	.0304	.0652
	(.2500)	(.2167)	(.1945)	(.2500)	(.2500)	(.2500)
Anderson-D A$^{2\ sas}$	.2526	.4695	.5409	.4313	.1920	.4020
	(.2500)	(.2466)	(.1712)	(.2500)	(.2500)	(.2500)

* P-value in parenthesis

[7] Skewness and kurtosis are computed using the SAS UNIVARIATE and CAPABILITY procedures that report kurtosis minus 3.

Table 4 presents results of normality tests for random variables with different values for N. The data were randomly generated from the standard normal distribution with a seed of 1,234,567 in SAS. As N grows, the mean, median, skewness, and (kurtosis-3) approach zero, while the standard deviation gets close to 1. The Kolmogorov-Smirnov D, Anderson-Darling A^2, Cramer-von Mises W^2 are computed in SAS, while The Skewness-Kurtosis and Shapiro-Francia W' are computed in STATA.

All four statistics do not reject the null hypothesis of normality regardless of the number of observations (Table 4). Note that the Shapiro-Wilk W is not reliable when N is larger than 2,000 and S-F W' is valid up to 5,000 observations. The Jarque-Bera and Skewness-Kurtosis tests show consistent results when N is large.

3.4 Software Issues

The UNIVARIATE procedure of SAS/BASE and CAPABILITY of SAS/QC compute various statistics and produce P-P and Q-Q plots. These procedures provide many numerical methods including Cramer-vol Mises and Anderson-Darling.[8] The P-P plot is generated only in CAPABILITY.

Table 5. Comparison of Procedures and Commands Available

	SAS 9.1	STATA 9.2 SE	SPSS 14.0
Descriptive statistics	UNIVARIATE	`.summarize`	Descriptives, Frequencies
(Skewness/Kurtosis)	CAPABILITY	`.tabstat`	Examine

Histogram, dot plot	UNIVARIATE
	CAPABILITY
		`.histogram`
		`.dotplot`
			Graph, Igraph, Examine, Frequencies

Stem-leaf-plot	UNIVARIATE	`.stem`	Examine
Box plot	UNIVARIATE	`.graph box`	Examine, Igraph
P-P plot	CAPABILITY	`.pnorm`	Pplot
Q-Q plot	UNIVARIATE CAPABILITY	`.qnorm`	Pplot, Examine
Detrended Q-Q/P-P plot			Pplot, Examine
Jarque-Bera (S-K) test		`.sktest`	
Shapiro-Wilk W	UNIVARIATE CAPABILITY	`.swilk`	Examine
Shapiro-Francia W'		`.sfrancia`	
Kolmogorov-Smirnov	UNIVARIATE CAPABILITY		Examine
Cramer-vol Mises	UNIVARIATE CAPABILITY		
Anderson-Darling	UNIVARIATE CAPABILITY		

In contrast, STATA has many individual commands to examine normality. In particular, STATA provides `.sktest` and `.sfrancia` to conduct Skewness-Kurtosis and Shapiro-Francia W' tests, respectively. SPSS EXAMINE provides numerical and graphical methods for normality test. The detrended P-P and Q-Q plots can be generated in SPSS. Table 5 summarizes SAS procedures and STATA/SPSS commands that are used to test normality of random variables.

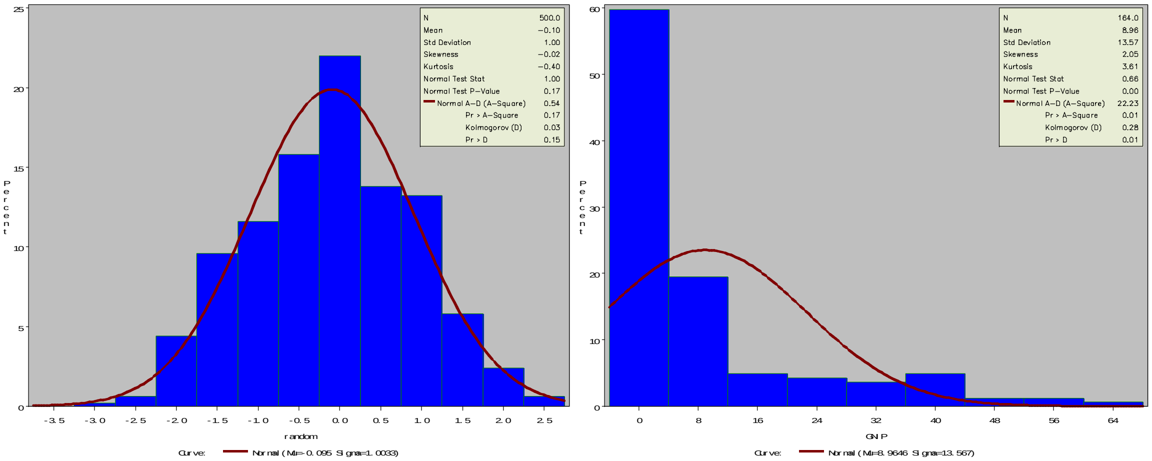
[8] MINITAB also performs the Kolmogorov-Smirnov and Anderson-Darling tests.

4. Testing Normality in SAS

SAS has the UNIVARIATE and CAPABILITY procedures to compute descriptive statistics, draw various graphs, and conduct statistical tests for normality. Two procedures have similar usage and produce similar statistics in the same format. However, UNIVARIATE produces stem-and-leaf, box, and normal probability plots, while CAPABILITY provides P-P and CDP plots that UNIVARIATE does not.

This section illustrates how to summarize normally and non-normally distributed variables and conduct normality tests of these variables using the two procedures. Figure 9 presents histograms of these variables.

Figure 9. Histograms of Normally and Non-normally Distributed Variables



4.1 A Normally Distributed Variable

The UNIVARIATE procedure provides a variety of descriptive statistics, and draws Q-Q , stem-and-leaf, normal probability, and box plots. This procedure also conducts Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Misers tests. The Shapiro-Wilk W will be reported only if N is not larger than 2000.

Let us take a look at an example of the UNIVARIATE procedure. The NORMAL option performs normality tests; the PLOT option draws a stem-and-leaf and a box plots; finally, the QQPLOT statement draws a Q-Q plot.

```
PROC UNIVARIATE DATA=masil.normality NORMAL PLOT;
       VAR random;
       QQPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
RUN;
```

Like UNIVARIATE, the CAPABILITY procedure also produces various descriptive statistics and plots. CAPABILITY can draw a P-P plot using the PPPLOT option but does not support stem-and-leaf, box, and normal probability plots.

4.1.1 SAS Output of Descriptive Statistics

In the following CAPABILITY procedure, QQPLOT, PPPLOT, and HISTOGRAM statements respectively draw a Q-Q plot, a P-P plot, and a histogram. Note that the INSET statement adds summary statistics to graphs such as a histogram and a Q-Q plot.[9]

```
PROC CAPABILITY DATA=masil.normality NORMAL;
        VAR random;
        QQPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
        PPPLOT random /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
        HISTOGRAM /NORMAL(COLOR=MAROON W=4) CFILL  = BLUE CFRAME = LIGR;
        INSET MEAN STD /CFILL=BLANK FORMAT=5.2 ;
RUN;
```

```
                                   The CAPABILITY Procedure
                                      Variable:  random

                                           Moments

                N                         500    Sum Weights                500
                Mean               -0.0950725    Sum Observations    -47.536241
                Std Deviation      1.00330171    Variance            1.00661432
                Skewness           -0.0203721    Kurtosis            -0.3988198
                Uncorrected SS     506.819932    Corrected SS        502.300544
                Coeff Variation    -1055.3019    Std Error Mean      0.04486902


                                  Basic Statistical Measures

                         Location                    Variability

                     Mean     -0.09507     Std Deviation            1.00330
                     Median   -0.11959     Variance                 1.00661
                     Mode       .          Range                    5.34911
                                           Interquartile Range      1.41773


                                  Tests for Location: Mu0=0

                         Test           -Statistic-    -----p Value------

                         Student's t    t  -2.11889    Pr > |t|    0.0346
                         Sign           M       -28    Pr >= |M|   0.0138
                         Signed Rank    S     -6523    Pr >= |S|   0.0435


                                     Tests for Normality

                    Test                  --Statistic---    -----p Value-----

                    Shapiro-Wilk          W      0.995564    Pr < W      0.168
                    Kolmogorov-Smirnov    D      0.026891    Pr > D     >0.150
                    Cramer-von Mises      W-Sq   0.083351    Pr > W-Sq   0.195
```

[9] Unlike the UNIVARIATE statement, the CAPABILITY statement does not have the PLOT option that draws stem-and-leaf, box, and normal probability plots.

```
                 Anderson-Darling      A-Sq  0.540894    Pr > A-Sq   0.171


                                Quantiles (Definition 5)

                               Quantile          Estimate

                               100% Max       2.511694336
                               99%            2.055464409
                               95%            1.530450397
                               90%            1.215210586
                               75% Q3         0.612538495
                               50% Median    -0.119592165
                               25% Q1        -0.805191028
                               10%           -1.413548051
                               5%            -1.794057126
                               1%            -2.219479314
                               0% Min        -2.837417522


                                  Extreme Observations

                      -------Lowest-------        -------Highest------

                            Value      Obs             Value      Obs

                      -2.83741752       29        2.14897641      119
                      -2.59039285      204        2.21109349      340
                      -2.47829639       73        2.42113892      325
                      -2.39126554      391        2.42171307      139
                      -2.24047386      393        2.51169434      332
```

4.1.2 Graphical Methods

The stem-and-leaf and box plots, produced by the UNIVARIATE procedure, illustrate that the variable is normally distributed (Figure 10). The locations of first quantile, mean, median, and third quintile indicate a bell-shaped distribution. Note that the mean -.0951 and median -.1196 are very close.

Figure 10. Stem-and-Leaf and Box Plots of a Normally Distributed Variable

```
                     Histogram                        #             Boxplot
      2.75+*                                          1                |
          .**                                         4                |
          .********                                  23                |
          .****************                          46                |
          .***********************                   68             +-----+
          .***************************               80             |     |
          .***************************************  116             *--+--*
          .**********************                    64             +-----+
          .*******************                       56                |
          .*********                                 27                |
          .*****                                     13                |
     -2.75+*                                          2                |
           ----+----+----+----+----+----+----+----
           * may represent up to 3 counts
```

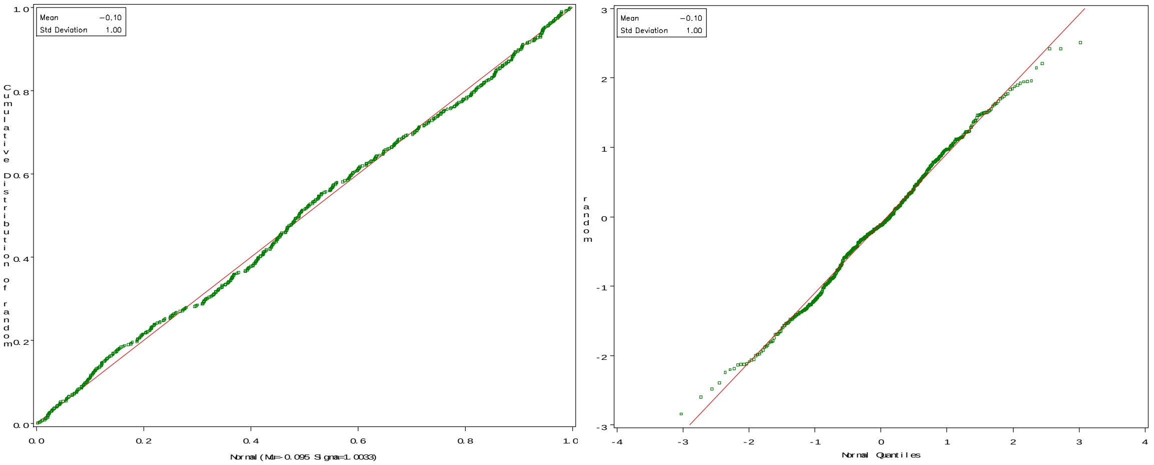
The normal probability plot available in UNIVARIATE shows a straight line, implying normality of the randomly drawn variable (Figure 11).

Figure 11. Normal Probability Plot of a Normally Distributed Variable

```
                    Normal Probability Plot
 2.75+                                                  *
     |                                              +++**
     |                                         ********
     |                                   *******
     |                               *****
     |                          ******
     |                    *******
     |                *****
     |           ******
     |      ******
     |*******
-2.75+*+
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1         0        +1        +2
```

The P-P and Q-Q plots show that the data points do not seriously deviates from the fitted line (Figure 12). They consistently indicate that the variable is normally distributed.

Figure 12. P-P and Q-Q Plots of a Normally Distributed Variable



4.1.3 Numerical Methods

The mean of -.0951 is very close to 0 and variance is almost 1. The skewness and kurtosis-3 are respectively -.0204 and -.3988, indicating an almost perfect normal distribution. However, these descriptive statistics do not provide conclusive information about normality.

SAS provides four different statistics for testing normality. Shapiro-Wilk W of .9956 does not reject the null hypothesis that the variable is normally distributed at the .05 level (p<.168). Similarly, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests do not reject the null hypothesis. Since the number of observations is less than 2,000, Shapiro-Wilk W test will be appropriate for this case.

The Jarque-Bera test also indicates the normality of the randomly drawn variable at the .05 level (p=.1875). Computation is

$$500\left[\frac{-0.0203721^2}{6}+\frac{-0.3988198^2}{24}\right]\sim 3.3482776(2)$$

Consequently, we can safely confirm that the randomly drawn variable is normally distributed.

4.2 A Non-normally Distributed Variable

Let us take the per capita gross national income as an example of non-normally distributed variable. See the appendix for details about this variable.

4.2.1 SAS Output of Descriptive Statistics

This section employs the UNIVARIATE procedure to compute descriptive statistics and perform normality tests.

```
PROC UNIVARIATE DATA=masil.gnip NORMAL PLOT;
       VAR gnip;
       QQPLOT gnip /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
       HISTOGRAM / NORMAL(COLOR=MAROON W=4) CFILL  = BLUE CFRAME = LIGR;
RUN;
```

```
                    The UNIVARIATE Procedure
                        Variable:  GNIP

                            Moments

N                         164    Sum Weights                164
Mean                8.9645732    Sum Observations    1470.19001
Std Deviation      13.5667877    Variance            184.057728
Skewness           2.04947469    Kurtosis            3.60816725
Uncorrected SS     43181.0356    Corrected SS        30001.4096
Coeff Variation    151.337798    Std Error Mean      1.05938813


                   Basic Statistical Measures

         Location                    Variability

     Mean     8.964573     Std Deviation            13.56679
     Median   2.765000     Variance                184.05773
     Mode     1.010000     Range                    65.34000
                           Interquartile Range       7.72500
```

```
                      Tests for Location: Mu0=0

           Test           -Statistic-    -----p Value------

           Student's t    t  8.462029    Pr > |t|    <.0001
           Sign           M        82    Pr >= |M|   <.0001
           Signed Rank    S      6765    Pr >= |S|   <.0001


                         Tests for Normality

      Test                  --Statistic---    -----p Value------

      Shapiro-Wilk          W     0.663114    Pr < W     <0.0001
      Kolmogorov-Smirnov    D     0.284426    Pr > D     <0.0100
      Cramer-von Mises      W-Sq  4.346966    Pr > W-Sq  <0.0050
      Anderson-Darling      A-Sq  22.23115    Pr > A-Sq  <0.0050


                        Quantiles (Definition 5)

                         Quantile      Estimate

                         100% Max        65.630
                         99%             59.590
                         95%             38.980
                         90%             32.600
                         75% Q3           8.680
                         50% Median       2.765
                         25% Q1           0.955
                         10%              0.450
                         5%               0.370
                         1%               0.290
                         0% Min           0.290


                          Extreme Observations

                ----Lowest----        ----Highest----

                Value      Obs         Value      Obs

                 0.29      164         46.32        5
                 0.29      163         47.39        4
                 0.31      162         54.93        3
                 0.33      161         59.59        2
                 0.34      160         65.63        1
```

4.2.2 Graphical Methods

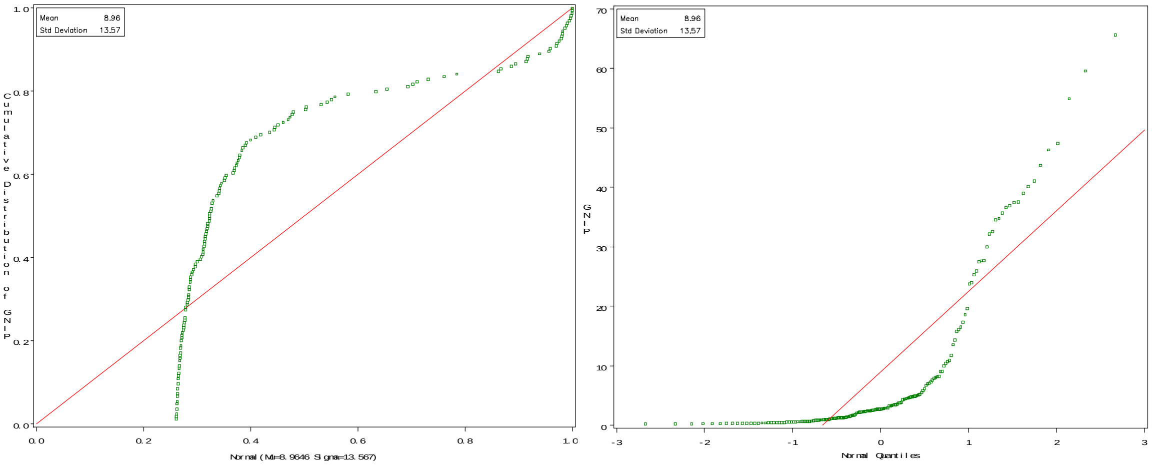
The stem-and-leaf, box, and normal probability plots all indicate that the variable is not normally distributed (Figure 13). Most observations are highly concentrated on the left side of the distribution.

Figure 13. Stem-and-Leaf, Box, and Normally Probability Plots

```
                 Histogram                    #             Boxplot
67.5+*                                        1                *
    .
    .*                                        1                *
    .*                                        1                *
    .*                                        2                *
    .*                                        3                *
    .**                                       6                *
    .**                                       5                0
    .**                                       5                0
    .*                                        2                0
    .**                                       6                |
    .***                                      7                |
    .******                                  17             +--+--+
 2.5+************************************   108             *-----*
     ----+----+----+----+----+----+----+-
     * may represent up to 3 counts
```

```
                         Normal Probability Plot
      67.5+                                                  *
          |
          |                                                *
          |                                              *
          |                                            **
          |                                          **    +++
          |                                       ***   +++
          |                                     **  ++++
          |                                    **+++
          |                                  +*+
          |                              ++++**
          |                          ++++  **
          |                       +++   ****
       2.5+* * *** **********************
           +----+----+----+----+----+----+----+----+----+----+
               -2        -1         0        +1        +2
```

Figure 14. P-P and Q-Q Plots of a Non-normally Distributed Variable



The P-P and Q-Q plots in Figure 14 show that the data points seriously deviate from the fitted line.

4.2.3 Numerical Methods

Per capita gross national income has a mean of 8.9646 and a large variance of 184.0557. Its skewness and kurtosis-3 are 2.0495 and 3.6082, respectively, indicating that the variable is highly skewed to the right with a high peak and thin tails.

It is not surprising that the Shapiro-Wilk test rejects the null hypothesis; W is .6631 and p-value is less than .0001. Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests also report similar results.

Finally, the Jarque-Bera test returns 203.7717, which rejects the null hypothesis of normality at the .05 level (p<.0000).

$$164\left[\frac{2.04947469^2}{6}+\frac{3.60816725^2}{24}\right]\sim 203.77176(2)$$

To sum, we can conclude that the per capita gross national income is not normally distributed.

5. Testing Normality Using STATA

In STATA, you have to use individual commands to get specific statistics or draw various plots. This section contrasts a normally distributed and a non-normally distributed variable using graphical and numerical methods.

5.1 Graphical Methods

A histogram is the most widely used graphical method. Histograms of normally and non-normally distributed variables are presented in introduction (Figure 2). The STATA `.histogram` command is followed by a variable name and options. The `normal` option adds a normal density curve to the histogram.

```
. histogram normal, normal
. histogram gnip, normal
```

Now let us draw a stem-and-leaf plot using the `.stem` command. The stem-and-leaf plot of the randomly drawn variable shows a bell-shaped distribution (Figure 15).

```
. stem normal
```

Figure 15. Stem-and-Leaf Plot of a Normally Distributed Variable

```
Stem-and-leaf plot for normal

normal rounded to nearest multiple of .01
plot in units of .01

 -28* | 4
 -27* |
 -26* |
 -25* | 9
 -24* | 8
 -23* | 9
 -22* | 40
 -21* | 93221
 -20* | 8650
 -19* | 8842
 -18* | 875200
 -17* | 94
 -16* | 9987550
 -15* | 97643320
 -14* | 87755432110
 -13* | 98777655433210
 -12* | 8866666433210
 -11* | 987774332210
 -10* | 875322
  -9* | 88887665542210
  -8* | 99988777533110
  -7* | 77766544100
  -6* | 998332
  -5* | 99988877654433221110
  -4* | 9998766655444433321
  -3* | 88766654433322221100
  -2* | 9999887665555444333221111 00
  -1* | 88887777766555444332222221110
  -0* | 99887776655433333111
   0* | 01233344445669
   1* | 0111222333445666778
   2* | 0001234444556889999
   3* | 1133444556667899
   4* | 014455667777
   5* | 00112334556888
   6* | 0001123668899
   7* | 00233466799999
   8* | 1122334667889
   9* | 012445666778889
  10* | 1133457799
  11* | 1222334445689
  12* | 122233489
  13* | 26889
  14* | 2777799
  15* | 00112459
  16* | 1347
  17* | 02467
  18* | 358
  19* | 03556
  20* |
  21* | 5
  22* | 1
  23* |
  24* | 22
  25* | 1
```

In contrast, per capita gross national income is highly skewed to the right, having most observations within $10,000 (Figure 16).

```
. stem gnip
```

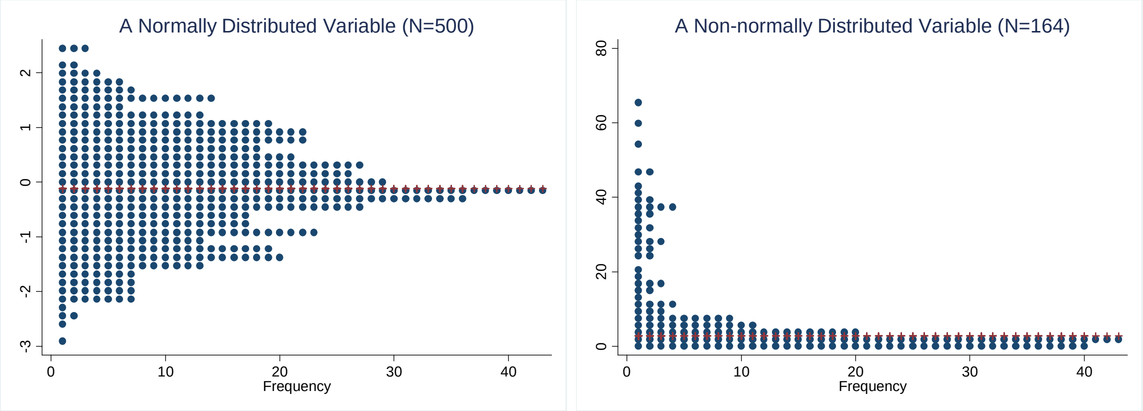
Figure 16. Stem-and-Leaf Plot of a Non-normally Distributed Variable

```
Stem-and-leaf plot for gnip

gnip rounded to nearest multiple of .1
plot in units of .1

  0** | 03,03,03,03,03,03,03,03,04,04,04,04,04,04,04,04,04,04,05,05, ... (64)
  0** | 21,22,23,23,23,24,24,24,24,25,25,25,26,26,26,27,28,28,28,28, ... (34)
  0** | 44,45,45,46,46,47,48,48,50,50,50,52,53,55,59
  0** | 62,68,71,71,73,76,79
  0** | 81,82,83,91,91
  1** | 00,04,07,09,18
  1** | 36
  1** | 44,58
  1** | 62,65,74
  1** | 86,97
  2** |
  2** | 38
  2** | 40,54
  2** | 60,75,77,78
  2** |
  3** | 00
  3** | 22,26
  3** | 46,48,57
  3** | 66,70,75,76
  3** | 90
  4** | 02,11
  4** | 37
  4** |
  4** | 63,74
  4** |
  5** |
  5** |
  5** | 49
  5** |
  5** | 96
  6** |
  6** |
  6** | 56
```

The `.dotplot` command generates a dotplot, very similar to the stem-and leaf plot, in a descending order (Figure 17).

```
. dotplot normal
. dotplot gnip
```

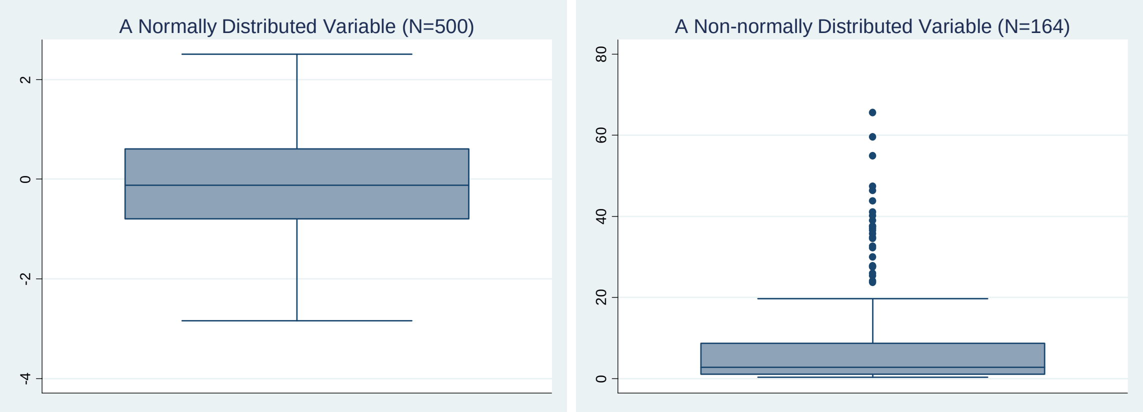
Figure 17. Dotplots of Normally and Non-normally Distributed Variables



The `.graph box` command draws a box plot. In the left plot of Figure 18, the shaded box represents the 25th percentile, median, and 75th percentile. The right plot, in contrast, has an asymmetric box with many outliers beyond the adjacent maximum line.

```
. graph box normal
. graph box gnip
```

Figure 18. Box plots of Normally and Non-normally Distributed Variables

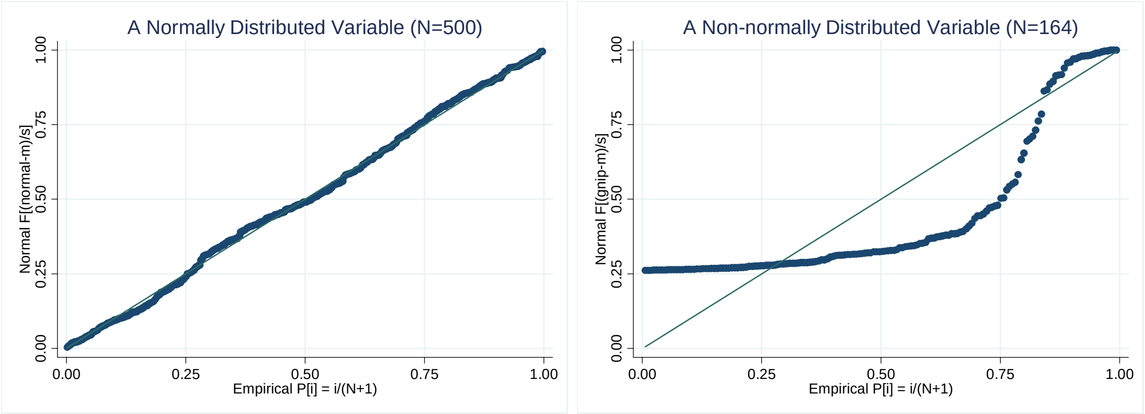


The `.pnorm` command produces a standardized normal P-P plot. The left plot shows almost no deviation from the fitted line, while the right depicts an s-shaped curve that largely deviates from the line (Figure 19). In STATA, a P-P plot has the cumulative distribution of an empirical variable on the X axis and the theoretical normal distribution on the Y axis.[10]

```
.pnorm normal
.pnorm gnip
```

[10] In SAS, these distributions are located reversely.

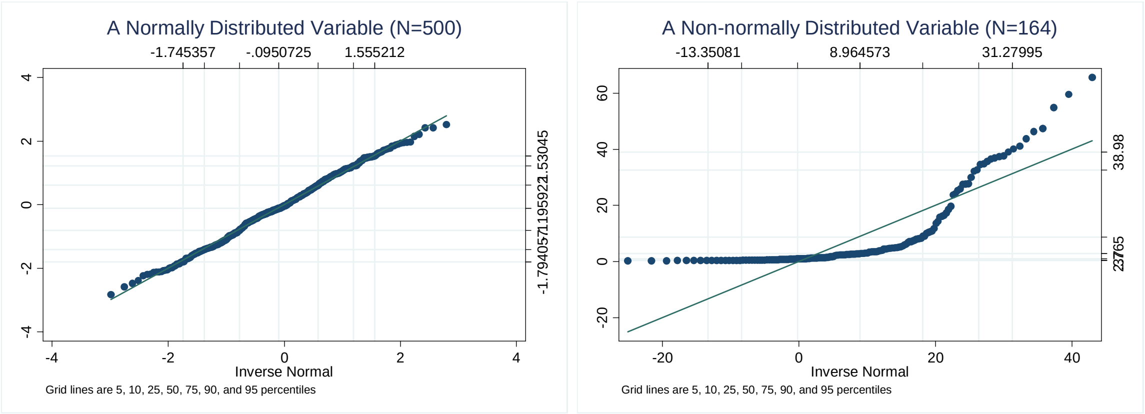
Figure 19. P-P plots of Normally and Non-normally Distributed Variables



The `.qnorm` command produces a standardized normal Q-Q plot. The Q-Q plots in Figure 20 show a similar pattern to P-P plots. In the right plot, data points systematically deviate from the straight fitted line.

```
.qnorm normal
.qnorm gnip
```

Figure 20. Q-Q plots of Normally and Non-normally Distributed Variables



5.2 Numerical Methods

Let us first get summary statistics using the `.summarize` command. The `detail` option lists various statistics in addition to the mean, standard deviation, minimum, and maximum. Skewness and kurtoshsis of a randomly drawn variable are respectively close to 0 and 3, implying normality. Per capital gross national income has large skewness of 2.03 and kurtosis of 6.46, being skewed to the right with a high peak and flat tails.

```
. summarize normal, detail

                           normal
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -2.219479      -2.837418
 5%    -1.794057      -2.590393
10%    -1.413548      -2.478296       Obs                 500
```

```
25%     -.805191      -2.391266       Sum of Wgt.         500

50%    -.1195922                      Mean          -.0950725
                        Largest       Std. Dev.      1.003302
75%     .6125385       2.211093
90%     1.215211       2.421139       Variance       1.006614
95%      1.53045       2.421713       Skewness      -.0203109
99%     2.055464       2.511694       Kurtosis       2.593181

. sum gnip, detail

                            gnip
-------------------------------------------------------------
      Percentiles      Smallest
 1%          .29            .29
 5%          .37            .29
10%          .45            .31       Obs                 164
25%         .955            .33       Sum of Wgt.         164

50%        2.765                      Mean           8.964573
                        Largest       Std. Dev.      13.56679
75%         8.68          47.39
90%         32.6          54.93       Variance       184.0577
95%        38.98          59.59       Skewness       2.030682
99%        59.59          65.63       Kurtosis       6.462734
```

STATA `.tabstat` command is vary useful to produce descriptive statistics in a table form. The `column(variable)`option lists statistics vertically (in table rows). The command for the variable `normal` is skipped.

```
. tabstat gnip, stats(n mean sum max min range sd var semean skewness kurtosis ///
  median p1 p5 p10 p25 p50 p75 p90 p95 p99 iqr q) column(variable)

    stats |    normal                              stats |      gnip
---------+----------                           ---------+----------
        N |       500                                  N |       164
     mean | -.0950725                               mean |  8.964573
      sum | -47.53624                                sum |   1470.19
      max |  2.511694                                max |     65.63
      min | -2.837418                                min |       .29
    range |  5.349112                              range |     65.34
       sd |  1.003302                                 sd |  13.56679
 variance |  1.006614                           variance |  184.0577
 se(mean) |   .044869                           se(mean) |  1.059388
 skewness | -.0203109                           skewness |  2.030682
 kurtosis |  2.593181                           kurtosis |  6.462734
      p50 | -.1195922                                p50 |     2.765
       p1 | -2.219479                                 p1 |       .29
       p5 | -1.794057                                 p5 |       .37
      p10 | -1.413548                                p10 |       .45
      p25 |  -.805191                                p25 |      .955
      p50 | -.1195922                                p50 |     2.765
      p75 |  .6125385                                p75 |      8.68
      p90 |  1.215211                                p90 |      32.6
      p95 |   1.53045                                p95 |     38.98
      p99 |  2.055464                                p99 |     59.59
      iqr |   1.41773                                iqr |     7.725
      p25 |  -.805191                                p25 |      .955
      p50 | -.1195922                                p50 |     2.765
      p75 |  .6125385                                p75 |      8.68
--------------------                           --------------------
```

Now let us conduct statistical tests for normality. STATA provide three testing methods: Shapiro-Wilk test, Shapiro-Francia test, and Skewness-Kurtosis test. The `.swilk` and `.sfrancia` commands respectively conduct the Shapiro-Wilk and Shapiro-Francia tests. Both tests do not reject normality of the randomly drawn variable and reject normality of per capita gross national income.

```
. swilk normal

                   Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
      normal |    500    0.99556      1.492      0.962  0.16804

. sfrancia normal

                  Shapiro-Francia W' test for normal data
    Variable |    Obs        W'         V'         z     Prob>z
-------------+-------------------------------------------------
      normal |    500    0.99645      1.273      0.541  0.29412

. swilk gnip

                   Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
        gnip |    164    0.66322     42.309      8.530  0.00000

. sfrancia gnip

                  Shapiro-Francia W' test for normal data
    Variable |    Obs        W'         V'         z     Prob>z
-------------+-------------------------------------------------
        gnip |    164    0.66365     45.790      7.413  0.00001
```

STATA's `.sktest` command performs the Skewness-Kurtosis test that is conceptually similar to the Jarque-Bera test. The `noadjust` option suppresses the empirical adjustment made by Royston (1991). The following S-K tests do not reject normality of a randomly drawn variable at the .05 level.

```
. sktest normal

                   Skewness/Kurtosis tests for Normality
                                                 ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      normal |      0.851         0.027            4.93       0.0850

. sktest normal, noadjust

                   Skewness/Kurtosis tests for Normality
                                                 ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      normal |      0.851         0.027            4.93       0.0850
```

Like the Shapiro-Wilk and Shapiro-Francia tests, the S-K test rejects normality of per capita gross national income. Surprisingly, the S-K test rejects normality of the variable randomly drawn at the .1 level. The Jarque-Bera statistic is 3.4823 = 500*(-.0203109^2/6+(2.593181-3)^2/24), which is not large enough to reject the null hypothesis (p<.1753). The Jarque-Bera test appears more reliable than the STATA S-K test (see Table 4).

```
. sktest gnip

                   Skewness/Kurtosis tests for Normality
                                                 ------- joint ------
```

```
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
        gnip |      0.000         0.000           55.33       0.0000
```

```
. sktest gnip, noadjust
```

```
                   Skewness/Kurtosis tests for Normality
                                                 ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
        gnip |      0.000         0.000           75.39       0.0000
```

The Jarque-Bera statistic of the per capita gross national income is 194.6489 = 164*(2.030682^2/6+(6.462734-3)^2/24). This large chi-squared rejects the null hypothesis (p<.0000).

In conclusion, the graphical methods and numerical methods provide sufficient evidence that the randomly drawn variable is normally distributed and per capita gross national income is not.

6. Testing Normality Using SPSS

SPSS has the DESCRIPTIVES and FREQUENCIES commands to produce descriptive statistics. DESCRIPTIVES is usually applied to continuous variables, but FREQUENCIES is also able to produce various descriptive statistics including skewness and kurtosis. The IGRAPH command draws histogram and box plots. The PPLOT command produces (detrended) P-P and Q-Q plots.

The EXAMINE command can produce both descriptive statistics and various plots, such as a stem-leaf-plot, histogram, box plot, P-P plot, and Q-Q plot. This command is able to draw the detrended Q-Q plot that SAS and STATA do not support. EXAMINE also performs the Kolmogorov-Smirnov and Shapiro-Wilk tests for normality.

6.1 A Normally Distributed Variable

Let us get summary statistics using DESCRIPTIVES and FREQUENCIES. The statistics are specified in the /STATISTICS subcommand. The output of the following DESCRIPTIVES command is skipped here.[11]

```
DESCRIPTIVES VARIABLES=normal
       /STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS SKEWNESS.

FREQUENCIES VARIABLES=normal /NTILES= 4
       /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
       SUM SKEWNESS SESKEW KURTOSIS SEKURT
       /HISTOGRAM
       /ORDER= ANALYSIS.
```

Statistics

normal

N	Valid	500
	Missing	0
Mean		-.0951
Std. Error of Mean		.04487
Median		-.1196
Mode		-2.84(a)
Std. Deviation		1.00330
Variance		1.007
Skewness		-.020
Std. Error of Skewness		.109
Kurtosis		-.399
Std. Error of Kurtosis		.218
Range		5.35
Minimum		-2.84
Maximum		2.51
Sum		-47.54

[11] In order to execute this command, open a syntax window and paste it into the window. Or click Analysis→ Descriptive Statistics→Descriptives and provide a variable of interest.

Percentiles	25	-.8066
	50	-.1196
	75	.6132

a Multiple modes exist. The smallest value is shown

The variable has a mean zero and a unit variance. The median is very close to the mean. The kurtosis-3 and skewness approach zero. Note that SPSS, like SAS, reports kurtosis-3.

6.1.1 Graphical Methods

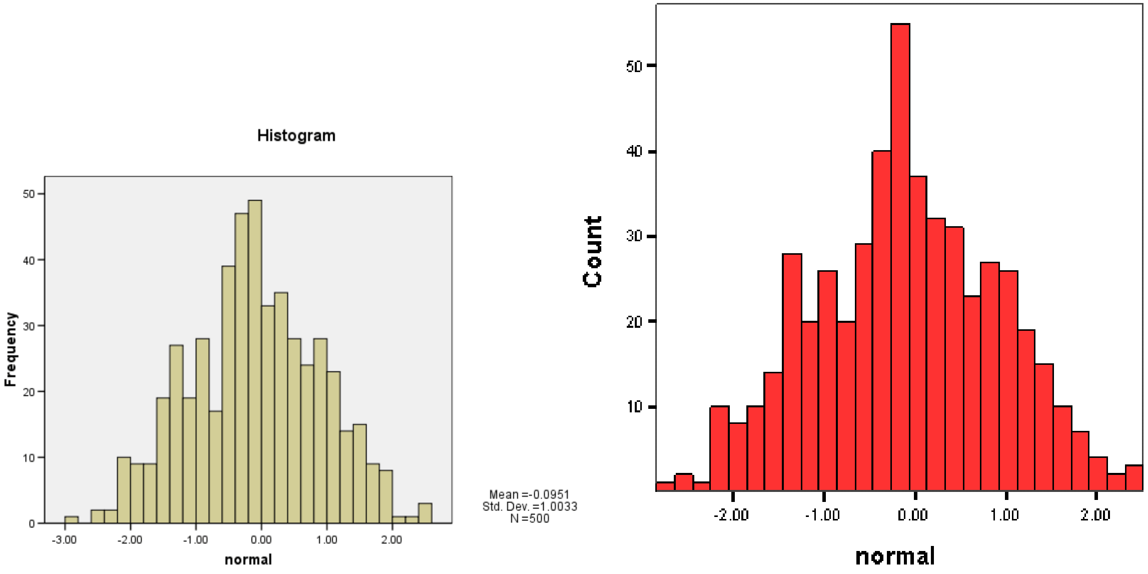
The /HISTOGRAM subcommand of FREQUENCIES tells SPSS to draw a histogram of the variable (left plot in Figure 20). You can get the identical plot using EXAMINE or the following GRAPH command.

```
GRAPH /HISTOGRAM=normal.
```

The IGRAPH command can produce a better histogram (right plot in Figure 21).[12] The histogram suggests that the variable is normally distributed.

```
IGRAPH /VIEWNAME='Histogram'
       /X1 = VAR(normal) TYPE = SCALE /Y = $count /COORDINATE = VERTICAL
       /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0 /CHARTLOOK='NONE'
       /Histogram SHAPE = HISTOGRAM CURVE = OFF X1INTERVAL AUTO X1START = 0.
```

Figure 21. Histograms of a Normally Distributed Variable



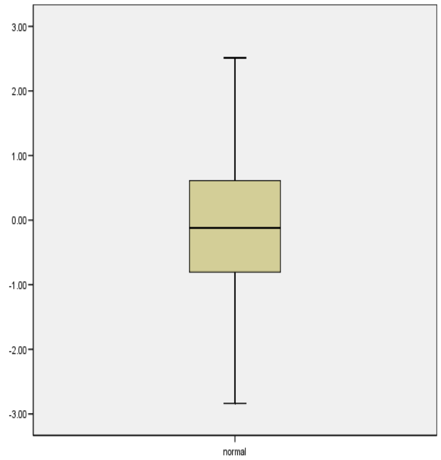
EXAMINE can produce a stem-and-leaf plot and a box plot using the /PLOT subcommand with the STEMLEAF and BOXPLOT option (Figure 22).[13] The stem-and-left plot is very similar to the histogram in Figure 20.

[12] To run this command from the menu, click Graphs→Interactive→Histogram and then specify a variable.
[13] From the menu, click Analyze→Descriptive Statistics→Explore, and then include the variable you want to examine.

```
EXAMINE VARIABLES=normal
        /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT.
        /MISSING LISTWISE /NOTOTAL.
```

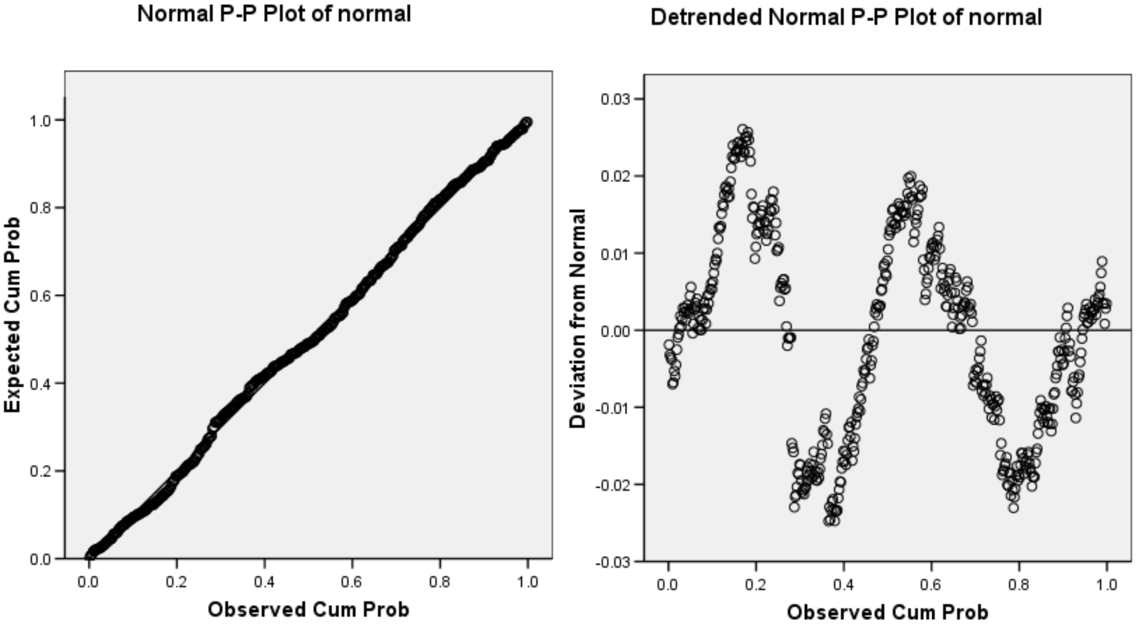
Figure 22. Stem-and-Leaf Plot and Box Plot of a Normally Distributed Variable

```
normal Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00       -2 .  &
    13.00       -2 .  011&
    27.00       -1 .  555667889
    56.00       -1 .  001111222233333444
    64.00       -0 .  55555566777888889999
   116.00       -0 .  000000011111111122222222233333334444444
    80.00        0 .  000001111112222223333334444
    68.00        0 .  5555556667777888899999
    46.00        1 .  000111122233444
    23.00        1 .  5567789
     4.00        2 .  4&
     1.00        2 .  &

 Stem width:   1.00
 Each leaf:       3 case(s)

 & denotes fractional leaves.
```



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

Figure 23. P-P and Detrended P-P Plots of a Normally Distributed Variable



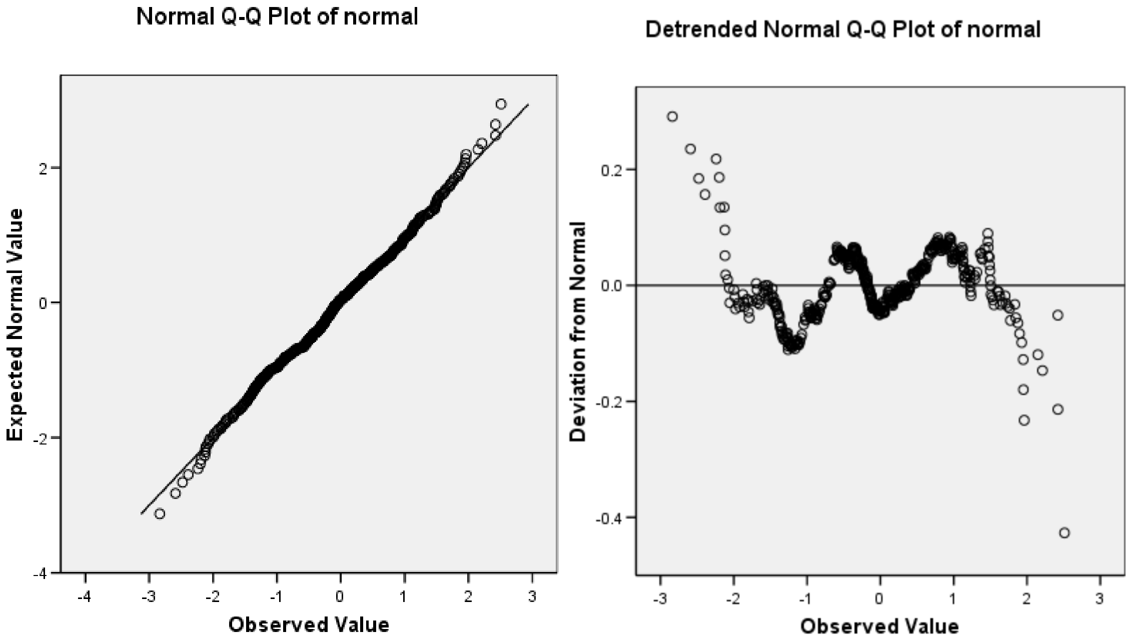
Now let us draw P-P and Q-Q plots using the PPLOT command.[14] This command automatically produces detrended P-P and Q-Q plots as well. The /TYPE chooses either P-P or Q-Q plot and /DIST specifies a probability distribution (e.g., the standard normal distribution). The variable does not deviate far away from the fitted line (Figure 23).

```
PPLOT /VARIABLES=normal
  /NOLOG /NOSTANDARDIZE
  /TYPE=P-P /FRACTION=BLOM /TIES=MEAN
  /DIST=NORMAL.
```

The following PPLOT command draws a Q-Q and detrended Q-Q plots of the variable (Figure 24).[15] The /PLOT NPPLOT subcommand of EXAMINE can also produce these plots.

```
PPLOT /VARIABLES=normal
  /NOLOG /NOSTANDARDIZE
  /TYPE=Q-Q /FRACTION=BLOM /TIES=MEAN
  /DIST=NORMAL.
```

Figure 24. Q-Q and Detrended Q-Q Plots of a Normally Distributed Variable



The P-P and Q-Q plots indicate no significant deviation from the fitted line. As in STATA, the Q-Q plot and detrended Q-Q plot has observed quantiles on the X axes and normal quantiles on the Y axes.

[14] From the menu, click Graphs→P-P and then specify a variable of interest. Probably due to bugs, the latest SPSS 14.02 (EXAMINE and PPLOT) does not produce color P-P and Q-Q plots, which are available in SPSS 13.xx and 12.xx.

[15] Click Graphs→Q-Q and then specify a variable

6.1.2 Numerical Methods

EXAMINE has the /PLOT NPPLOT subcommand to test normality of a variable. This command performs the Kolmogorov-Smirnov and Shapiro-Wilk tests and draws a normal (detrended) Q-Q plot as well.

```
EXAMINE  VARIABLES=normal
        /PLOT NPPLOT
        /STATISTICS DESCRIPTIVES EXTREME
        /CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

The above EXAMINE command first produces descriptive statistics (/STATISTICS DESCRIPTIVES), draws a normal Q-Q plot (/PLOT NPPLOT), and then conducts normality tests.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
normal	500	100.0%	0	.0%	500	100.0%

Descriptives

			Statistic	Std. Error
normal	Mean		-.0951	.04487
	95% Confidence Interval for Mean	Lower Bound	-.1832	
		Upper Bound	-.0069	
	5% Trimmed Mean		-.0933	
	Median		-.1196	
	Variance		1.007	
	Std. Deviation		1.00330	
	Minimum		-2.84	
	Maximum		2.51	
	Range		5.35	
	Interquartile Range		1.42	
	Skewness		-.020	.109
	Kurtosis		-.399	.218

Extreme Values

			Case Number	Value
Normal	Highest	1	332	2.51
		2	139	2.42
		3	325	2.42
		4	340	2.21
		5	119	2.15
	Lowest	1	29	-2.84

		2	204	-2.59
		3	73	-2.48
		4	391	-2.39
		5	393	-2.24

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Normal	.027	500	.200(*)	.996	500	.168

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Since N is less than 2,000, we have to read the Shapiro-Wilk statistic that does not reject the null hypothesis of normality (p<.168). SPSS reports the same kolmogorov-smirnov statistic of .027, but it provides an adjusted p-value of .200, a bit larger than the .150 that SAS reports.

6.2 A Non-normally Distributed Variable

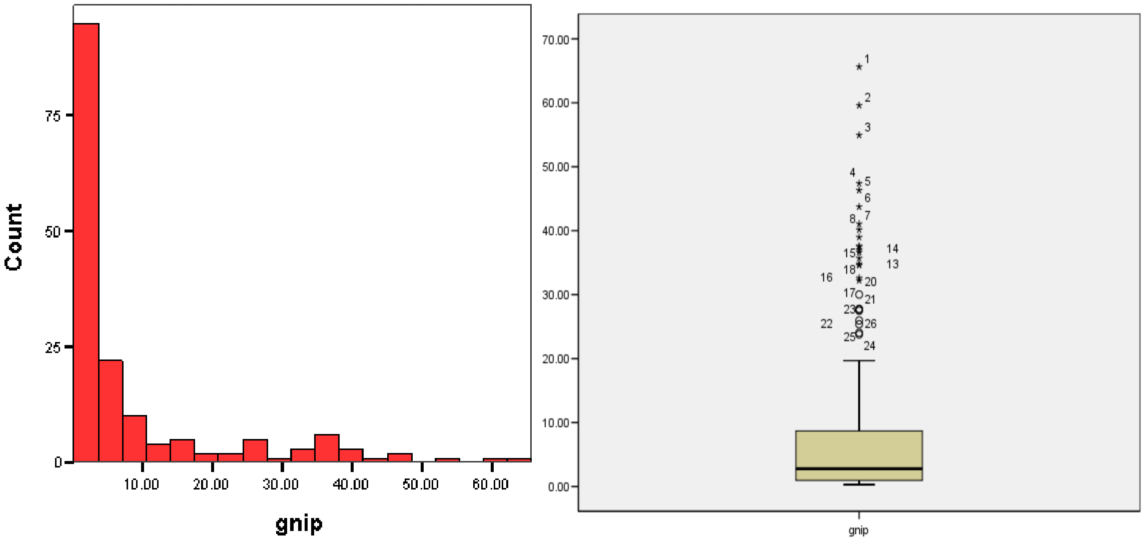
Let us consider a variable of per capita national gross income that is not normally distributed.

6.2.1 Graphical Methods

The following IGRAPH and EXAMINE command produce the histogram, stem-and-leaf plot, and box plot of a non-normally distributed variable `gnip`.

Figure 25 illustrate that the distribution is heavily skewed to the right. There are many outliers beyond the extreme line in the box plot (left plot of Figure 25). The median and the 25th percentile are close to each other. The stem-and-leaf plot is skipped here.

Figure 25. Histogram and Box Plot a Non-normally Distributed Variable

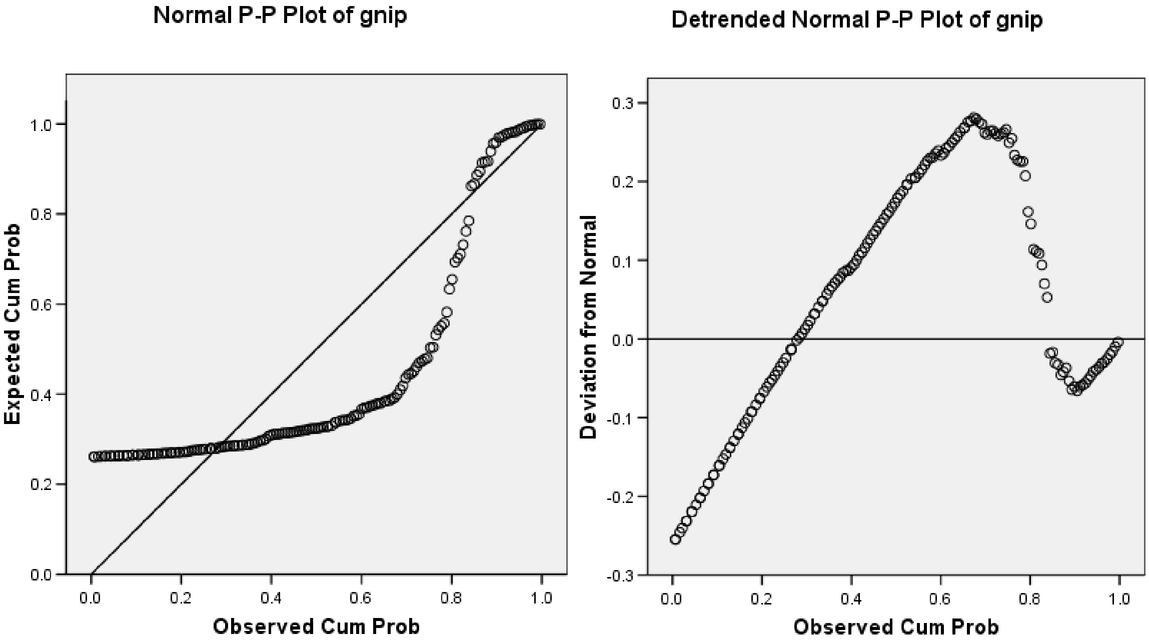


```
IGRAPH  /VIEWNAME='Histogram'
        /X1 = VAR(gnip) TYPE = SCALE /Y = $count /COORDINATE = VERTICAL
        /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0 /CHARTLOOK='NONE'
        /Histogram SHAPE = HISTOGRAM CURVE = OFF X1INTERVAL AUTO X1START = 0.

EXAMINE VARIABLES=gnip
        /PLOT BOXPLOT STEMLEAF HISTOGRAM
        /MISSING LISTWISE /NOTOTAL.
```

Figure 26 presents the P-P and detrended P-P plots where data points significantly deviate from the straight fitted line. The detrended P-P plot shows a systematic deviation of data points.

```
PPLOT /VARIABLES=gnip
        /NOLOG /NOSTANDARDIZE
        /TYPE=P-P /FRACTION=BLOM
        /TIES=MEAN
        /DIST=NORMAL.
```

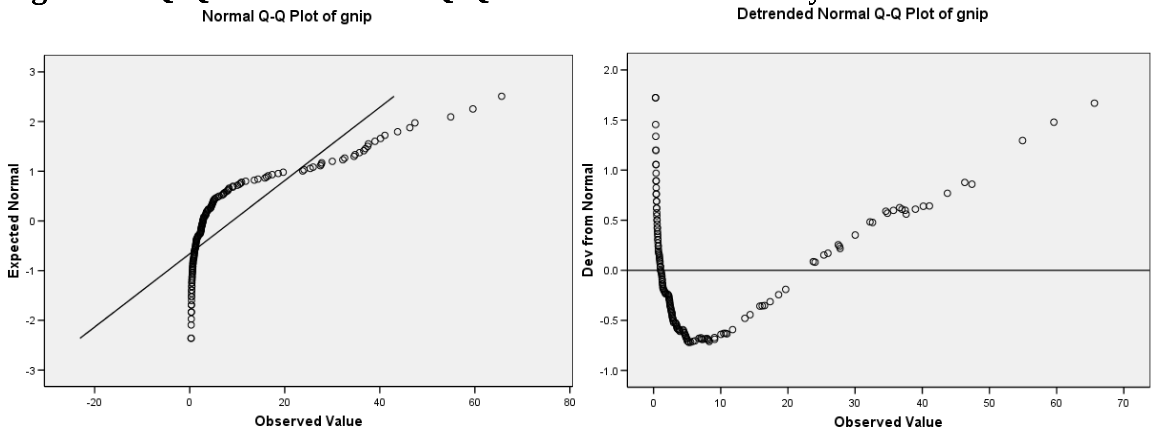
Figure 26. P-P and Detrended P-P Plots of a Non-normally Distributed Variable



The Q-Q and detrended Q-Q plots also show a significant deviation from the fitted line (Figure 27). The detrended Q-Q plot also presents a systematic pattern of deviation indicating non-normality of the variable.

```
PPLOT /VARIABLES=gnip
        /NOLOG /NOSTANDARDIZE
        /TYPE=Q-Q /FRACTION=BLOM
        /TIES=MEAN
        /DIST=NORMAL.
```

Figure 27. Q-Q and Detrended Q-Q Plots of a Non-normally Distributed Variable



6.2.2 Numerical Methods

The descriptive statistics of `gnip` indicates that the variable is not normally distributed. There is a large gap between the mean of 8.9646 and the median of 2.7650. The skewness and kurtosis -3 are 2.049 and 3.608, respectively. The variable appears severely skewed to the right with a higher peak and flat tails.

```
EXAMINE VARIABLES=gnip
       /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
       /MISSING LISTWISE /NOTOTAL.
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gnip	164	100.0%	0	.0%	164	100.0%

Descriptives

			Statistic	Std. Error
gnip	Mean		8.9646	1.05939
	95% Confidence Interval for Mean	Lower Bound	6.8727	
		Upper Bound	11.0565	
	5% Trimmed Mean		7.1877	
	Median		2.7650	
	Variance		184.058	
	Std. Deviation		13.56679	
	Minimum		.29	
	Maximum		65.63	
	Range		65.34	
	Interquartile Range		7.92	
	Skewness		2.049	.190

	Kurtosis	3.608	.377

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
gnip	.284	164	.000	.663	164	.000

a Lilliefors Significance Correction

The Shapiro-Wilk test rejects the null hypothesis of normality at the .05 level. The Jarque-Bera test also rejects the null hypothesis with a large statistic of 204. Its computation is skipped (see section 4.2.3). Based on a consistent result from both graphical and numerical methods, we can conclude the variable `gnip` is not normally distributed.

7. Conclusion

Univariate analysis is the first step of data analysis once a data set is ready. Various descriptive statistics provide valuable basic information about variables that is used to determine what method of analysis should be employed.

Normality is commonly assumed in many statistical and economic methods without any empirical test. Violation of this assumption will result in unreliable inferences and misleading interpretations.

There are graphical and numerical methods for conducting univariate analysis and normality tests (Table 1). The graphical methods produce various plots such as a stem-and-leaf plot, histogram, and a P-P plot that are intuitive and easy to interpret. Some are descriptive and others are theory-driven.

Numerical methods compute a variety of measures of central tendency and dispersion such as mean, median, quantile, variance, and standard deviation. Skewness and kurtosis provide clues to the normality of a variable. If skewness and kurtosis-3 are close to zero, the variable may be normally distributed. Keep in mind that SAS and SPSS report kurtosis-3, while STATA returns kurtosis itself.

If the skewness of a variable is larger than 0, the variable is skewed to the right with many observations on the left of the distribution; a negative skewness indicates many observations on the right. If kurtosis-3 is greater than 0 (or kurtosis is greater than 3), the distribution has a high peak and flat tails (third plot in Figure 8). If kurtosis is smaller than 3, the variable has a low peak and thick tails (first plot in Figure 8).

In addition to these descriptive statistics, there are formal ways to perform normality tests. The Shapiro-Wilk and Shapiro-Francia tests are proper when N is less than 2,000 and 5,000, respectively. The Kolmogorov-Smirnov, Cramer-vol Mises, and Anderson-Darling tests are recommended when N is large. The Jarque-Bera test, although not supported by most statistical software, is a good alternative for normality testing.

The SAS UNIVARIATE and CONTENTS procedures provide a variety of descriptive statistics and normality testing methods including Kolmogorov-Smirnov, Cramer-vol Mises, and Anderson-Darling tests (Table 5). These procedures produce stem-and-leaf, box plot, histogram, P-P plot, and Q-Q plot as well. STATA has various commands for univariate analysis and graphics. In particular, STATA supports the Shapiro-Francia test, a modification of the Shapiro-Wilk test, and the skewness-kurtosis test. But there is no command for the Kolmogorov-Smirnov test for normality in STATA. SPSS can produce detrended P-P and Q-Q plots, and perform the Shapiro-Wilk and Kolmogorov-Smirnov tests with Lilliefors significance correction.

Appendix A: Data Sets

This document uses the following three variables.

1. Unemployment rate of Illinois, Indiana, and Ohio in 2005

This unemployment rate is provided by the Bureau of Labor Statistics. Actual data were downloaded from http://www.stats.indiana.edu/, the Indiana Business Research Center of the Kelley School of Business, Indiana University.

```
. tabstat rate, stat(mean sd p25 median p75 skewness kurtosis) by(state)

Summary for variables: rate
     by categories of: state

 state |      mean        sd       p25       p50       p75  skewness  kurtosis
-------+----------------------------------------------------------------------
    IL |  5.421569  .9242206       4.7      5.35         6  .6570033  3.946029
    IN |  5.641304  1.038929       4.9       5.5      6.35  .3416314  2.785585
    OH |    6.3625  1.458098       5.5       6.1      6.95  1.665322  8.043097
-------+----------------------------------------------------------------------
 Total |  5.786879  1.214066         5      5.65       6.4   1.44809  8.383285
------------------------------------------------------------------------------
```

2. A Randomly Drawn Variable

This variable includes 500 observations that were randomly drawn from the standard normal distribution with a seed of 1,234,567. The RANNOR() of SAS was used as a random number generator.

```
%LET n=500; %LET dataset=n500;

DATA masil.&dataset;
seed=1234567;
DO i=1 TO &n;
       normal=RANNOR(seed); OUTPUT;
END;
RUN;
```

```
. tabstat normal, stat(mean sd p25 median p75 skewness kurtosis)

    variable |      mean        sd       p25       p50       p75  skewness  kurtosis
-------------+--------------------------------------------------------------------
      normal | -.0950725  1.003302  -.805191 -.1195922  .6125385 -.0203109  2.593181
----------------------------------------------------------------------------------
```

3. Per Capita Gross National Income in 2005.

This data set includes per capita gross national incomes of 164 countries in the world that are provided by the World Bank (http://web.worldbank.org/).

```
. tabstat gnip, stat(mean sd p25 median p75 skewness kurtosis)

    variable |      mean        sd       p25       p50       p75  skewness  kurtosis
-------------+--------------------------------------------------------------------
        gnip |  8.964573  13.56679      .955     2.765      8.68  2.030682  6.462734
----------------------------------------------------------------------------------
```

References

Bera, Anil. K., and Carlos. M. Jarque. 1981. "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence." *Economics Letters*, 7(4):313-318.

D'Agostino, Ralph B., Albert Belanger, and Ralph B. D'Agostino, Jr. 1990. "A Suggestion for Using Powerful and Informative Tests of Normality." *American Statistician*, 44(4): 316-321.

Jarque, Carlos M., and Anil K. Bera. 1980. "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals." *Economics Letters*, 6(3):255-259.

Jarque, Carlos M., and Anil K. Bera. 1987. "A Test for Normality of Observations and Regression Residuals." *International Statistical Review*, 55(2):163-172.

Mitchell, Michael N. 2004. *A Visual Guide to STATA Graphics*. College Station, TX: Stata Press.

Royston, J. P. 1982. "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples." *Applied Statistics*, 31(2): 115-124.

Royston, J. P. 1983. "A Simple Method for Evaluating the Shapiro-Francia W' Test of Non-Normality." *Statistician*, 32(3) (September): 297-300.

Royston, J.P. 1991. "Comment on sg3.4 and an Improved D'Agostino test." *Stata Technical Bulletin*, 3: 13-24.

Royston, P.J. 1992. "Approximating the Shapiro-Wilk W-Test for Non-normality." *Statistics and Computing*, 2:117-119.

SAS Institute. 1995. *SAS/QC Software: Usage and Reference I and II*. Cary, NC: SAS Institute.

SAS Institute. 2004. *SAS 9.1.3 Procedures Guide Volume 4*. Cary, NC: SAS Institute.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, 52(3/4) (December).:591-611.

Shapiro, S. S., and R. S. Francia. 1972. "An Approximate Analysis of Variance Test for Normality." *Journal of the American Statistical Association*, 67 (337) (March): 215-216.

STATA Press. 2003. *STATA Graphics Reference Manual Release 8*. College Station, TX: STATA Press.

STATA Press. 2005. *STATA Reference Manual Release 9*. College Station, TX: STATA Press.

Acknowledgements

I am grateful to Jeremy Albright and Kevin Wilhite at the UITS Center for Statistical and Mathematical Computing for comments and suggestions.

Revision History

- 2002 First draft
- 2006 Revision with new data