I n d i a n a U n i v e r s i t y

University Information Technology Services

Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*

Hun Myoung Park, Ph.D.

Last modified on November 2008

University Information Technology Services
Center for Statistical and Mathematical Computing
Indiana University
410 North Park Avenue Bloomington, IN 47408
(812) 855-4724 (317) 278-4740
http://www.indiana.edu/~statmath

* The citation of this document should read: "Park, Hun Myoung. 2008. *Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University." http://www.indiana.edu/~statmath/stat/all/normality/index.html

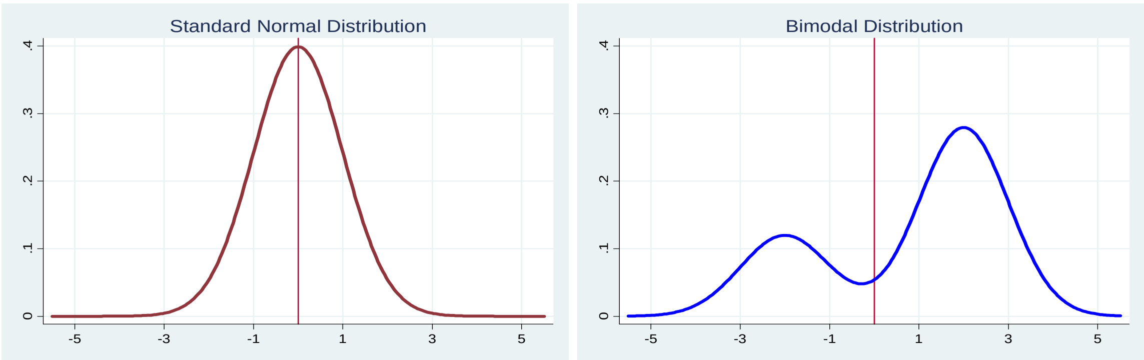
This document summarizes graphical and numerical methods for univariate analysis and normality test, and illustrates how to do using SAS 9.1, Stata 10 special edition, and SPSS 16.0.

1. Introduction

Descriptive statistics provide important information about variables to be analyzed. Mean, median, and mode measure central tendency of a variable. Measures of dispersion include variance, standard deviation, range, and interquantile range (IQR). Researchers may draw a histogram, stem-and-leaf plot, or 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



The t-test and ANOVA (Analysis of Variance) compare group means, assuming a variable of interest follows a normal probability distribution. 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?

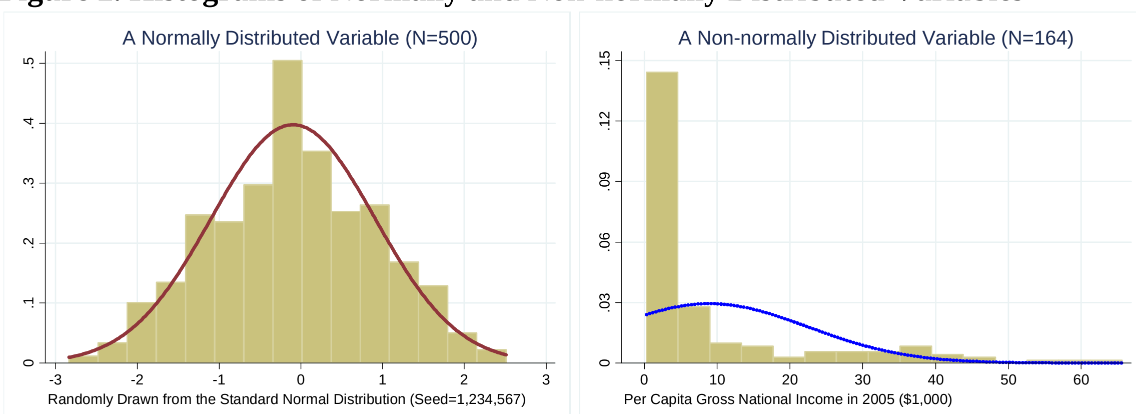
There are two ways of testing normality (Table 1). Graphical methods visualize the distributions of random variables or differences between an empirical distribution and a theoretical distribution (e.g., the standard normal distribution). Numerical methods present summary statistics such as skewness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

Table 1. Graphical Methods versus Numerical Methods

	Graphical Methods	Numerical Methods
Descriptive	Stem-and-leaf plot, (skeletal) box plot, dot 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. A 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 normality test, 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 mean 0 and variance 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.

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 and dot plot are useful to summarize continuous or event count data. Figure 3 and 4 respectively present a stem-and-leaf plot and a dot plot of the unemployment rate of three states.

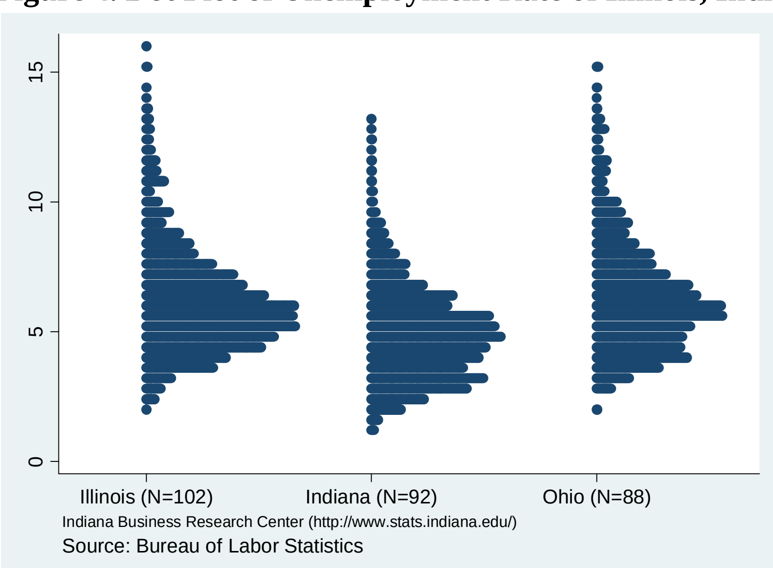
Figure 3. Stem-and-Leaf Plot of Unemployment Rate of Illinois, Indiana, Ohio

```
-> state = IL

Stem-and-leaf plot for rate(Rate)

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

  3. | 7889
  4* | 011122344
  4. | 5566666666677778888999
  5* | 00111222223333333344444
  5. | 5555667777777888999
  6* | 000011222333444
  6. | 555579
  7* | 0033
  7. |
  8* | 0
  8. | 8
```

```
-> state = IN

Stem-and-leaf plot for rate(Rate)

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

  3* | 1
  3. | 89
  4* | 012234
  4. | 566666778889999
  5* | 00000111222222233344
  5. | 555666666777889
  6* | 002222233344
  6. | 5666677889
  7* | 1113344
  7. | 67
  8* | 14
```

```
-> state = OH

Stem-and-leaf plot for rate (Rate)

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

  3* | 8
  4* | 014577899
  5* | 01223333445556667778888888999
  6* | 001111122222233444446678899
  7* | 01223335677
  8* | 1223338
  9* | 99
 10* | 1
 11* |
 12* |
 13* | 3
```

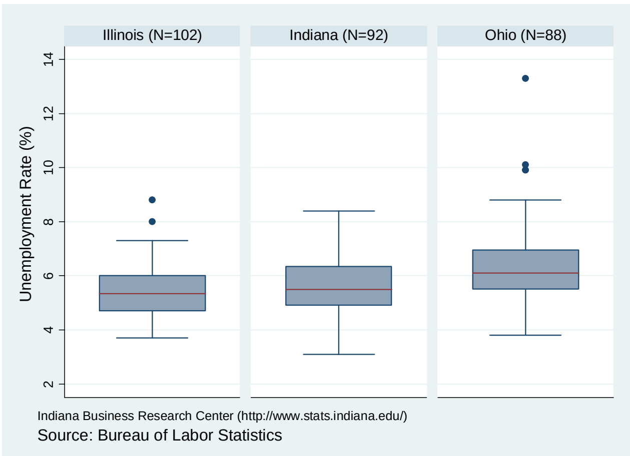
Figure 4. Dot Plot of Unemployment Rate of Illinois, Indiana, Ohio



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]

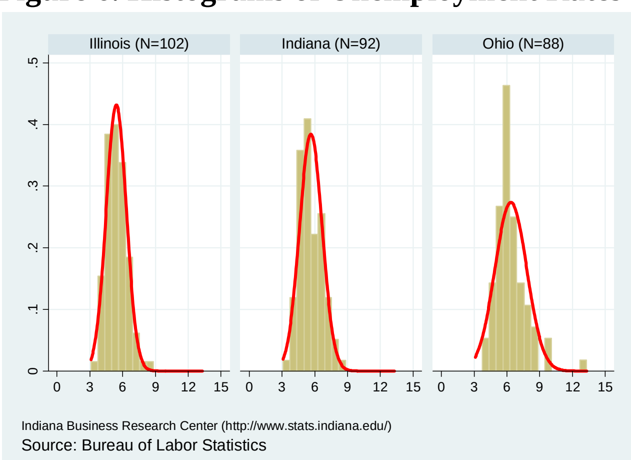
In Figure 5, you should see outliers in Illinois and Ohio that affect the shapes of corresponding boxes. By contrast, the Indiana unemployment rate does not have outliers, and its symmetric box implies that the rate appears to be normally distributed.

Figure 5. Box Plots of Unemployment Rates of Illinois, Indiana, and Ohio



The histogram graphically shows how each category (interval) accounts for the proportion of total observations and is appropriate when N is large (Figure 6).

Figure 6. Histograms of Unemployment Rates of Illinois, Indiana and Ohio



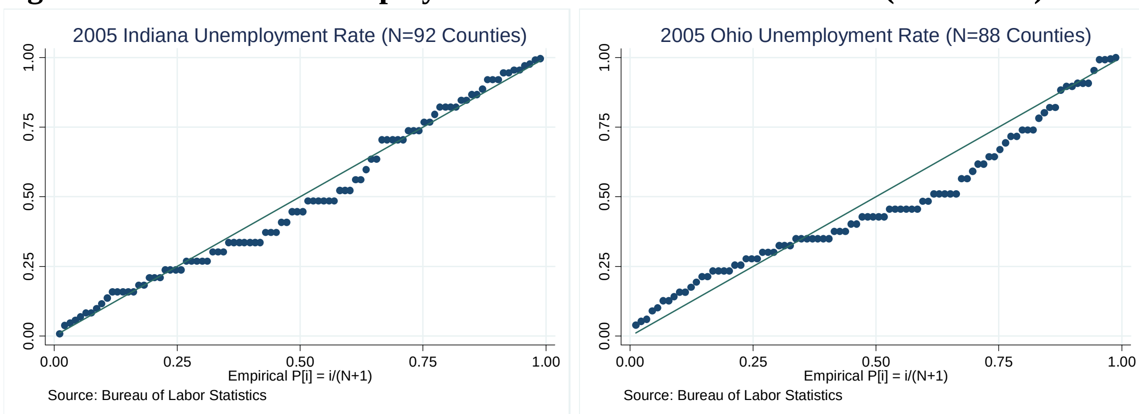
[1] The first quartile cuts off lowest 25 percent of data; the second quartile, median, cuts data set in half; and the third quartile cuts off lowest 75 percent or highest 25 percent of data. See http://en.wikipedia.org/wiki/Quartile

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

2.2 Theory-driven Plots

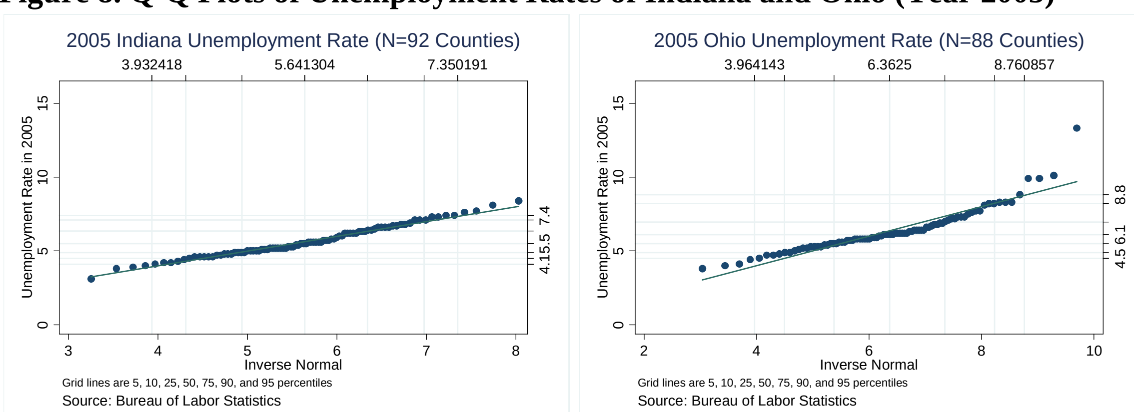
P-P and Q-Q plots are considered here. The probability-probability plot (P-P plot or percent plot) compares an empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function (e.g., the standard normal distribution function). In Figure 7, Ohio appears to deviate more from the fitted line than Indiana.

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



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 8, Indiana appears to have a smaller variation in its unemployment rate than Ohio. By contrast, Ohio appears to have a wider range of outliers in the upper extreme.

Figure 8. Q-Q Plots of Unemployment Rates of Indiana and Ohio (Year 2005)



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.

3. Numerical Methods

Graphical methods, although visually appealing, do not provide objective criteria to determine normality of variables. Interpretations are thus a matter of judgments. 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 their mean. The second central moment is

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

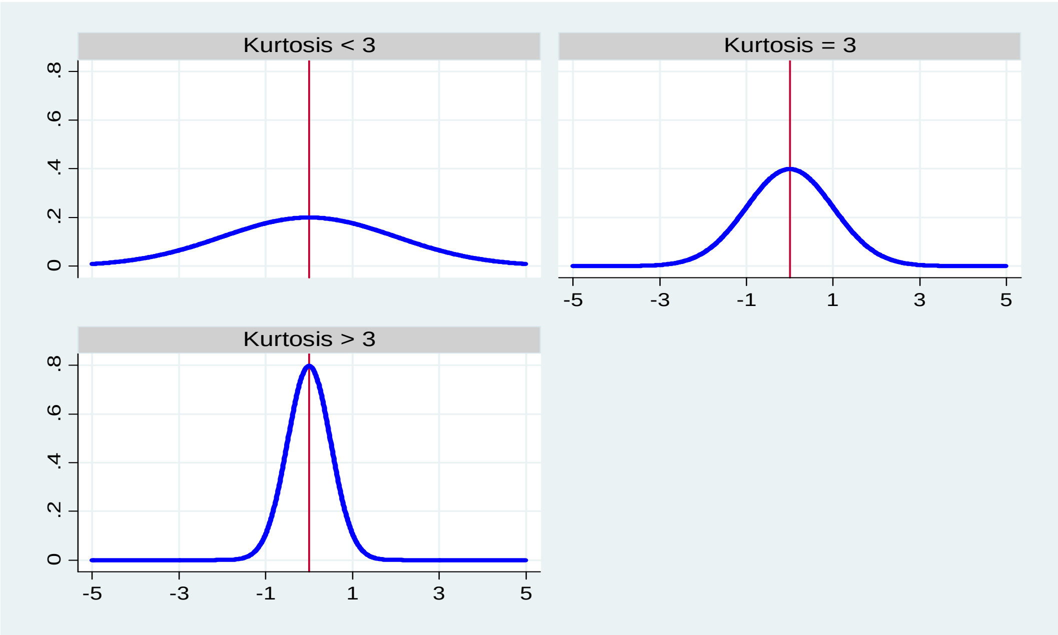
Skewness is a third standardized moment that measures the degree of symmetry of a probability distribution. If skewness is greater than zero, the distribution is skewed to the right, having more observations on the left.

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

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

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

Figure 9. Probability Distributions with Different Kurtosis



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 9).[3] By contrast, kurtosis larger than 3 indicates a higher peak and thin tails (last plot). A normally distributed random variable should have skewness and kurtosis near zero and three, respectively (second plot in Figure 9).

```
 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
----------------------------------------------------------------------------------------
```

In short, skewness and kurtosis show how the distribution of a variable deviates from a normal distribution. These statistics are based on the empirical data.

3.2 Theory-driven Statistics

The numerical methods of normality test 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 of Testing Normality

Test	Statistic	N Range	Dist.	SAS	Stata	SPSS
Jarque-Bera	χ^2		$\chi^2(2)$	-	-	-
Skewness-Kurtosis	χ^2	$9 \le N$	$\chi^2(2)$	-	`.sktest`	-
Shapiro-Wilk	W	$7 \le N \le 2,000$	-	YES	`.swilk`	YES
Shapiro-Francia	W’	$5 \le N \le 5,000$	-	-	`.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.

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.

---

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

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 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. Since N is not large, you need to read Shapiro-Wilk, Shapiro-Francia, Jarque-Bera, and Skewness-Kurtosis statistics.

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 alternative ways of normality test.

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.

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

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

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

For Illinois: `12.292825 = 102*(0.66685022^2/6 + 1.0553068^2/24)`
For Indiana: `1.9458304 = 92*(0.34732004^2/6 + (-0.1583764)^2/24)`
For Ohio: `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
1st quantile	-.2372	-.8674	-.8052	-.7099	-.7034	-.7121
Median	.6411	-.0625	-.1196	-.0309	-.0224	-.0219
3rd quantile	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 sas	.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 parentheses

Table 4 presents results of normality tests for random variables with different numbers of observations. 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, and the standard deviation gets close to 1. The Kolmogorov-Smirnov D, Anderson-

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

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.

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.

By 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. Since SPSS has changed graph-related features over time, you need to check menus, syntaxes, and reported bugs.

Table 5 summarizes SAS procedures and Stata/SPSS commands that are used to test normality of random variables.

Table 5. Comparison of Procedures and Commands Available

	SAS	Stata	SPSS
Descriptive statistics (Skewness/Kurtosis)	UNIVARIATE	`.summarize` `.tabstat`	Descriptives, Frequencies Examine
Histogram, dot plot	UNIVARIATE CHART, PLOT	`.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	`.qnorm`	Pplot, Examine
Detrended Q-Q/P-P plot			Pplot, Examine
Jarque-Bera (S-K) test		`.sktest`	
Shapiro-Wilk W	UNIVARIATE	`.swilk`	Examine
Shapiro-Francia W'		`.sfrancia`	
Kolmogorov-Smirnov	UNIVARIATE		Examine
Cramer-vol Mises	UNIVARIATE		
Anderson-Darling	UNIVARIATE		

* The UNIVARIATE procedure can provide the plot.
** The CAPABILITY procedure can provide the plot.

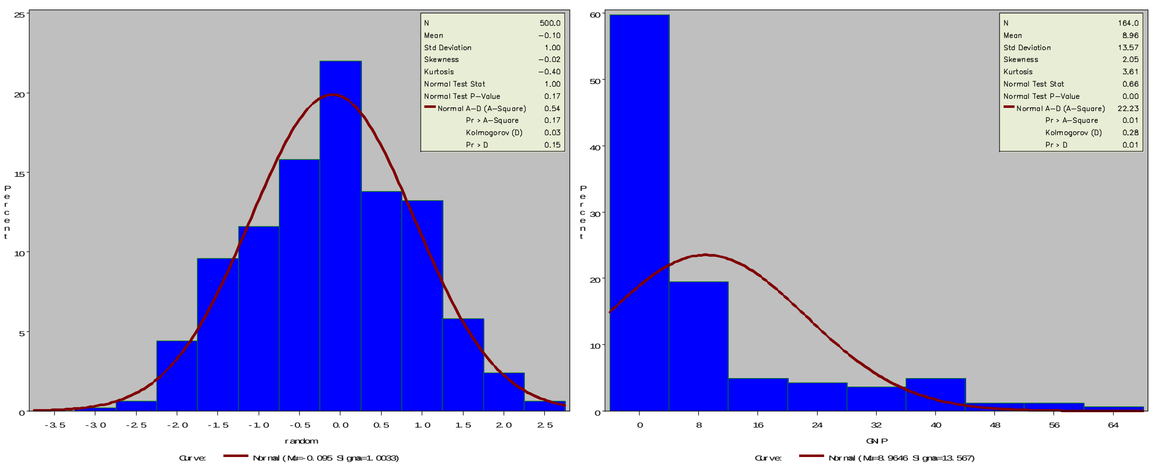
[8] MINITAB also performs the Kolmogorov-Smirnov and Anderson-Darling tests.

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 a stem-and-leaf plot, box plot, and normal probability plot, while CAPABILITY provides P-P plot and CDP plot 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 (see Figure 10).

Figure 10. Histogram of Normally and Non-normally Distributed Variables



4.1 A Normally Distributed Variable

The UNIVARIATE procedure provides a variety of descriptive statistics, Q-Q plot, leaf-and-stem-plot, and box plot. This procedure also conducts Kolmogorov-Smirnov test, Shapiro-Wilk' test, Anderson-Darling, and Cramer-von Misers tests.

Let us take a look at an example of the UNIVARIATE procedure. The NORMAL option conducts normality testing; PLOT draws a leaf-and-stem plot and a box plot; 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 a leaf-and-stem plot, a box plot, and a normal probability plot; this procedure does not have the PLOT option available in UNIVARIATE.

4.1.1 SAS Output of Descriptive Statistics

The following is an example of the CAPABILITY procedure. QQPLOT, PPPLOT, and HISTOGRAM statements respectively draw a Q-Q plot, P-P plot, and histogram. Note that the INSET statement adds summary statistics to graphs such as histogram and Q-Q plot.

```
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
               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 plot and box plot, produced by the UNIVARIATE produre, illustrate that the variable is normally distributed (Figure 11). 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 11. Stem-and-Leaf Plot and Box Plot 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
```

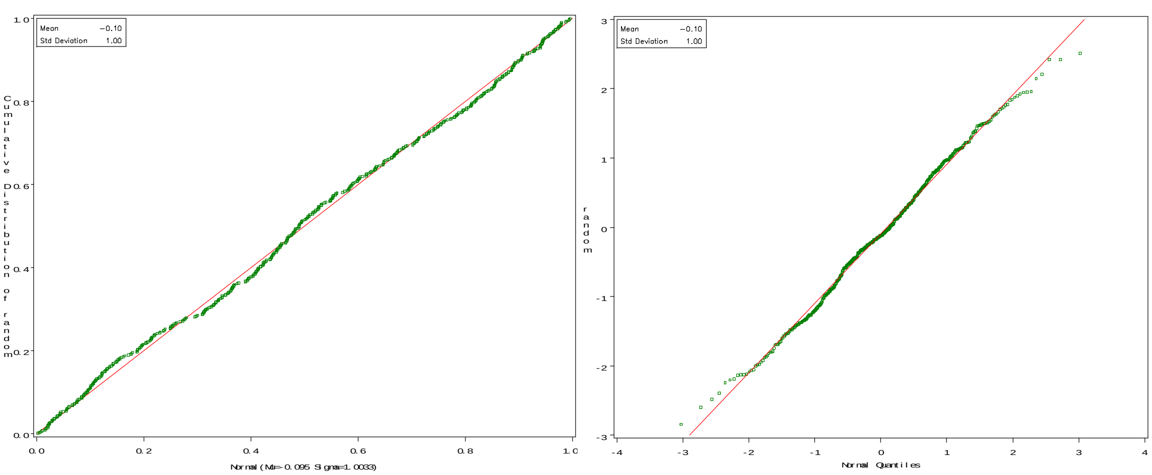
The normal probability plot available in UNIVARIATE shows a straight line, implying the normality of the randomly drawn variable (Figure 12).

Figure 12. 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 below show that the data points are not seriously deviated from the fitted line. They consistently indicate that the variable is normally distributed.

Figure 13. P-P plot and Q-Q Plot 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 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 (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, however, Shapiro-Wilk W test will be appropriate for this case.

The Jarque-Bera test also indicates the normality of the randomly drawn variable (p=.1875). Note that -.3988 is kurtosis -3.

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

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

4.2 A Non-normally Distributed Variable

Let us examine the per capita gross national income as an example of non-normally distributed variables. 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. The variable has mean 8.9646 and median 2.0495, where are substantially different. Variance 184.0577 is extremely large.

```
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 plot, box plot, and normal probability plots all indicate that the variable is not normally distributed (Figure 14). Most observations are highly concentrated on the left side of the distribution. See the stem-and-leaf plot and box plot in Figure 14.

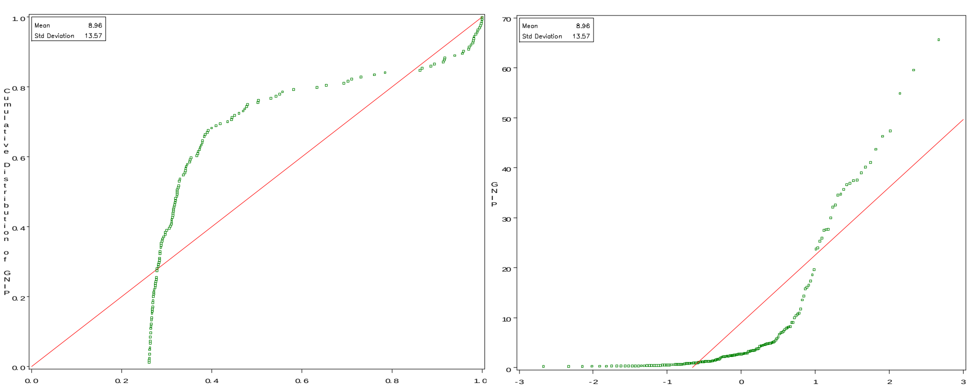
Figure 14. Stem-and-Leaf Plot, Box Plot, and Normally Probability Plot

```
                  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
```

The following P-P and Q-Q plots show that the data points are seriously deviated from the fitted line (Figure 15).

Figure 15. P-P plot and Q-Q Plot of a Non-normally Distributed Variable



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 rejected 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 normally distributed and non-normally distributed variables using graphical and numerical methods.

5.1 Graphical Methods

A histogram is the most widely used graphical method. The histograms of normally and non-normally distributed variables are presented in the introduction. 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
```

Let us draw a stem-and-leaf plot using the `.stem` command. The stem-and-leaf plot of the randomly drawn `normal` shows a bell-shaped distribution (Figure 16).

```
. stem normal
```

Figure 16. 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* | 9999887665555444333221111100
  -1* | 88887777766555444333222221110
  -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
```

By contrast, per capita gross national income is highly skewed to the right, having most observations within $10,000 (Figure 17).

```
. stem gnip
```

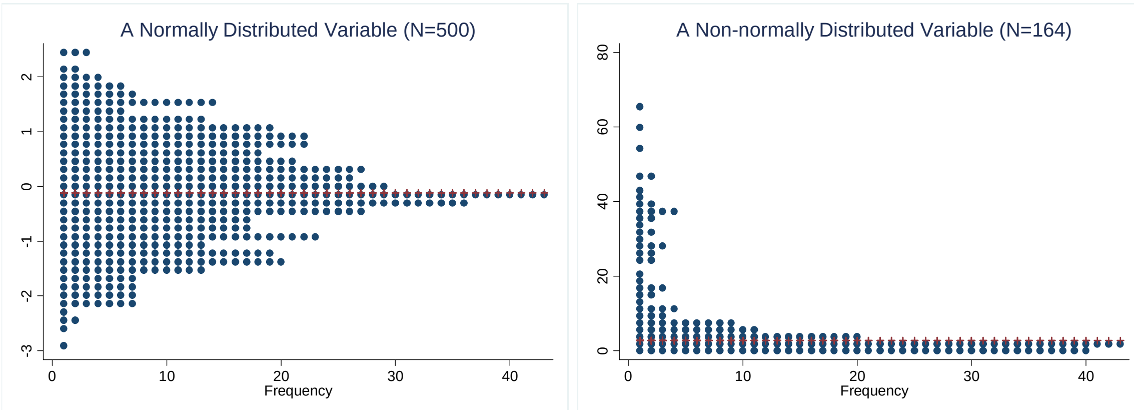
Figure 17. Stem-and-Leaf Plot of a Non-normally Distributed Variable

```
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 dot plot, very similar to the stem-and leaf plot, in a descending order (Figure 18).

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

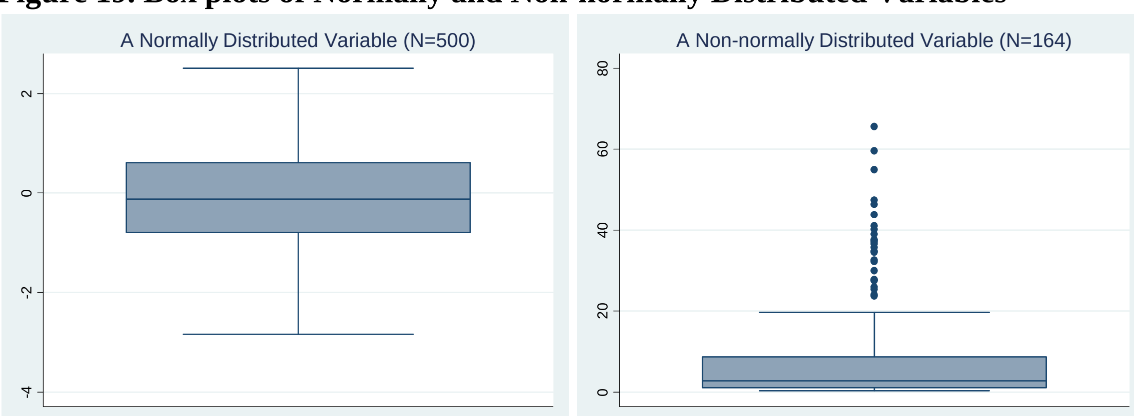
Figure 18. Dotplots of Normally and Non-normally Distributed Variables



The `.graph box` command draws a box plot. In the left plot of Figure 19, the shaded box represents the 25th percentile, median, and 75th percentile, which are symmetrically arranged. The right plot has an asymmetric box with many outliers beyond the adjacent maximum line.

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

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

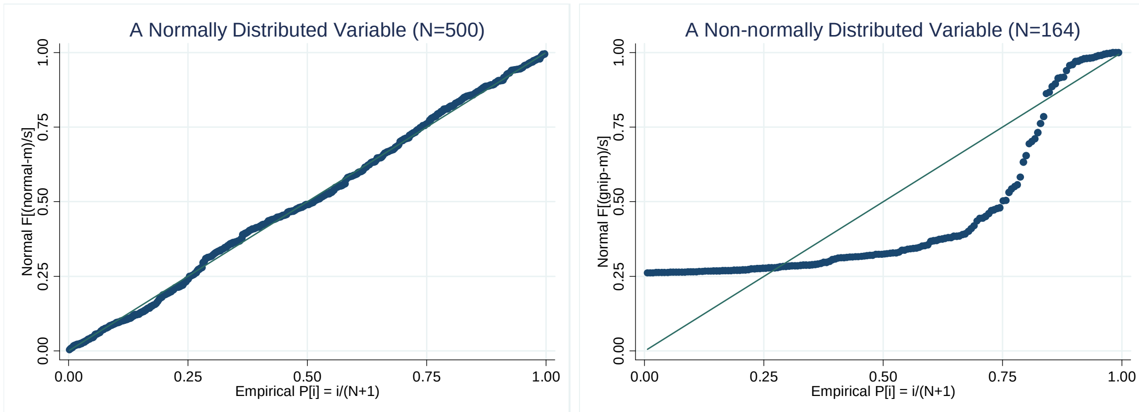


The `.pnorm` command produces standardized normal P-P plot. The left plot shows almost no deviation from the line, while the right depicts an s-shaped curve that is largely deviated from the fitted line. In Stata, a P-P plot has the cumulative distribution of an empirical variable on the x axis and the theoretical normal distribution on the y axis.[9]

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

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

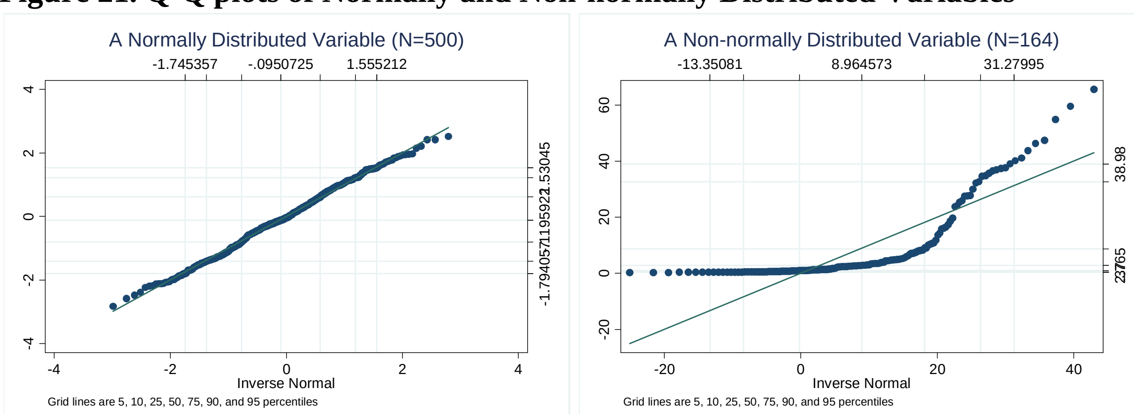
Figure 20. P-P plots of Normally and Non-normally Distributed Variables



The `.qnorm` command produces a standardized normal Q-Q plot. The following Q-Q plots show a similar pattern that P-P plots do (Figure 21). In the right plot, data points are systematically deviated from the straight fitted line.

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

Figure 21. 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 including mean, standard deviation, minimum, and maximum. Skewness and kurtosis 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
```

The `.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 of 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 conducts 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 but surprisingly reject the null hypothesis at the .1 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, both S-K tests below reject the null hypothesis that per capita gross national income is normally distributed at the .01 significance level.

```
. 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 `normal` 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 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). The Jarque-Bera test appears to be more reliable than the Stata S-K test (see Table 4).

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

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 in addition to frequency tables. 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, (detrended) P-P plot, and (detrended) Q-Q plot. EXAMINE also performs the Kolmogorov-Smirnov and Shapiro-Wilk tests for normality.

6.1 A Normally Distributed Variable

DESCRIPTIVES summarizes interval or continuous variables and FREQUENCIES reports frequency tables of discrete variables and summary statistics. The /STATISTICS subcommand in both commands specify statistics to be produced.

The following DESCRIPTIVES command reports the number of observations, sum, mean, variance, standard deviation of `normal`.[10] The mean of -.10 and standard deviation 1 implies that the variable is normally distributed.

```
DESCRIPTIVES VARIABLES=normal
        /STATISTICS=MEAN SUM STDDEV VARIANCE.
```

Descriptive Statistics

	N	Sum	Mean	Std. Deviation	Variance
normal	500	-47.54	-.0951	1.00330	1.007
Valid N (listwise)	500				

The following FREQUENCIES produces various statistics of `normal`, a frequency table, and a histogram.[11] Since `normal` is continuous, its frequency table is long and thus skipped here. The /HISTOGRAM subcommand draws a histogram, which is the same as what the GRAPH command in the next page produces.

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

Statistics

[10] In order to execute this command, open a syntax window , copy and paste the syntax into the window, and then click Run menu. Alternatively, click Analysis→ Descriptive Statistics→Descriptives and provide a variable of interest.
[11] Click Analysis→ Descriptive Statistics→Frequencies and then specify statistics using the **Statistics** option.

normal

N	Valid	500.000
	Missing	.000
Mean		-.095
Std. Error of Mean		.045
Median		-.120
Mode		-2.837[a]
Std. Deviation		1.003
Variance		1.007
Skewness		-.020
Std. Error of Skewness		.109
Kurtosis		-.399
Std. Error of Kurtosis		.218
Range		5.349
Minimum		-2.837
Maximum		2.512
Sum		-47.536
Percentiles	25	-.807
	50	-.120
	75	.613

a. Multiple modes exist. The smallest value is shown

The variable has a mean -.10 and a unit variance. The median -.120 is very close to the mean. The kurtosis-3 is -.399 and skewness is -.020.

6.1.1 Graphical Methods

Like the /HISTOGRAM subcommand of FREQUENCIES, the GRAPH command draws a histogram of the variable `normal` (left plot in Figure 22).[12]

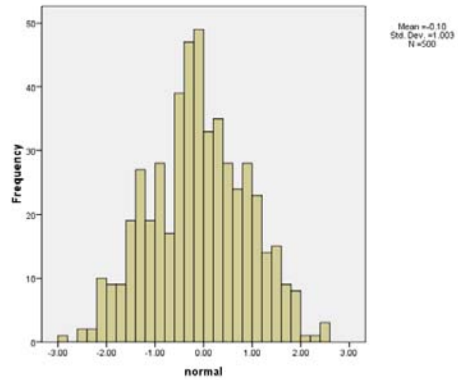
```
GRAPH /HISTOGRAM=normal.
```

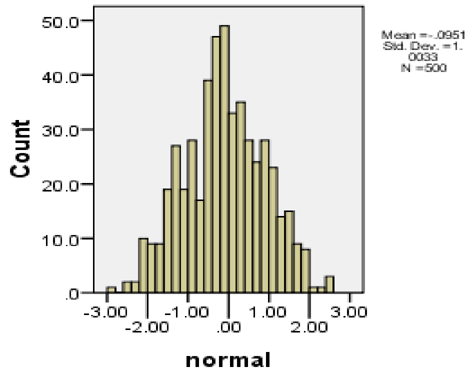
The IGRAPH command can produce a similar histogram (right plot in Figure 22) but its syntax appears to be messy.[13] Two histograms report mean -.1 and standard deviation 1 on the right top corner and suggest that the variable is normally distributed.

```
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 22. Histogram of a Normally Distributed Variable

[12] Click Graphs→Legacy Dialogs→Histogram.
[13] Click Graphs→Legacy Dialogs→Interactive→Histogram.





The EXAMINE command can produce descriptive statistics as well as a stem-and-leaf plot and a box plot (Figure 23 and 24).[14] The /PLOT subcommand with STEMLEAF and BOXPLOT draws two plots that is very similar to the histogram in Figure 22.

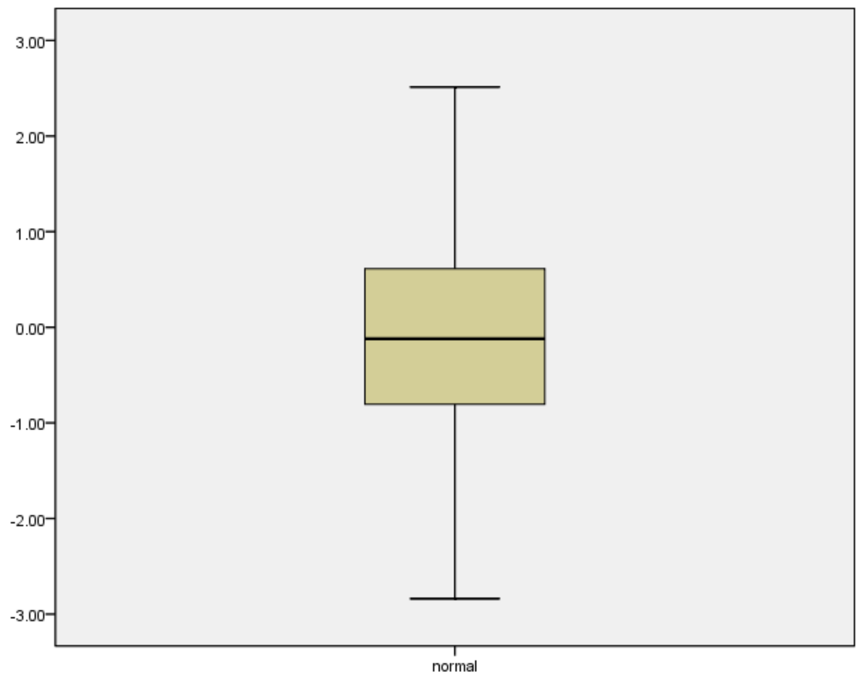
```
EXAMINE VARIABLES=normal
       /PLOT BOXPLOT STEMLEAF
       /COMPARE GROUP
       /STATISTICS DESCRIPTIVES
       /CINTERVAL 95
       /MISSING LISTWISE
       /NOTOTAL.
```

Figure 23. Stem-and-Leaf Plot of a Normally Distributed Variable

```
normal Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00       -2 .  &
    13.00       -2 .  00111&
    27.00       -1 .  555566678899
    56.00       -1 .  0001111112222223333333344444
    64.00       -0 .  5555555556667777788888888999999
   116.00       -0 .  0000000000111111111111112222222222222223333333333334444444444
    80.00        0 .  0000000111111111222222222233333333444444
    68.00        0 .  5555555566666777777788888899999999
    46.00        1 .  00000111111222334444
    23.00        1 .  55566778899
     4.00        2 .  4&
     1.00        2 .  &

 Stem width:   1.00
 Each leaf:       2 case(s)


 & denotes fractional leaves.
```

Figure 24. Box Plot of a Normally Distributed Variable

[14] Click Analyze→Descriptive Statistics→Explore, and then include the variable you want to examine.

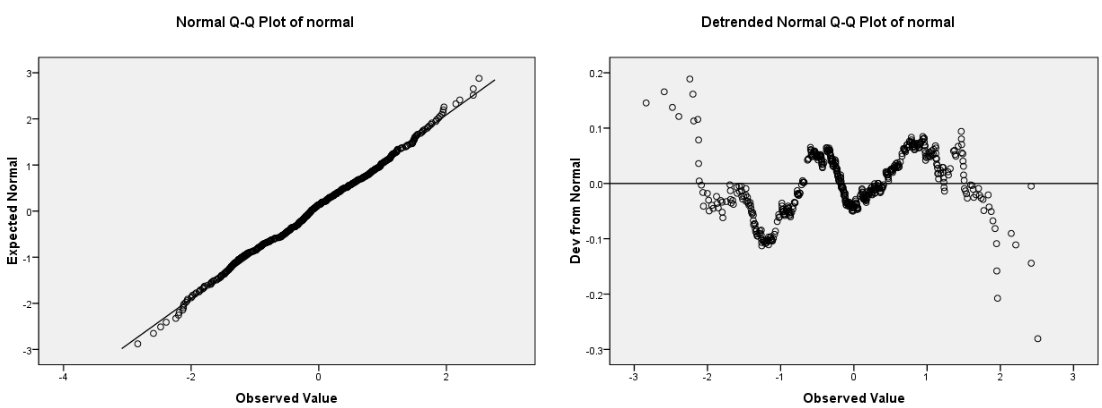


The both extremes (i.e., minimum and maximum), the 25th, 50th, and 75th percentiles are symmetrically arranged in the box plot.

EXAMINE also produces a histogram and normal Q-Q plot and detrended normal Q-Q plot using HISTOGRAM and NPPLOT option (Figure 25).[15] NPPLOT conducts normality test and draw the two Q-Q plots.

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

Figure 25. Q-Q and Detrended Q-Q Plots of a Normally Distributed Variable

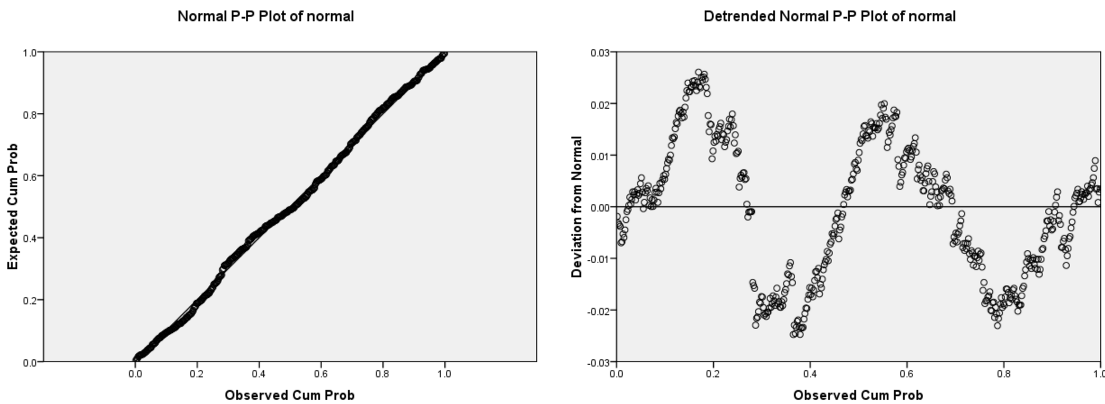


[15] In the **Explore** dialog box, choose **Plots** option and then check **Normality plots with tests** option.

The PPLOT command produces P-P and Q-Q plots as well. [16] The /TYPE subcommand chooses either P-P or Q-Q plot and /DIST specifies a probability distribution (e.g., the standard normal distribution). The following PPLOT command draws normal P-P and detrended normal P-P plots (Figure 26); the output of other descriptive statistics is skipped here.

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

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



The following PPLOT command draws normal Q-Q and detrended normal Q-Q plots of the variable (see Figure 25).

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

Both P-P and Q-Q plots show no significant deviation from the fitted line. As in Stata, the normal Q-Q plot and detrended Q-Q plot has observed quantiles on the X axis and normal quantiles on the Y axis.

6.1.2 Numerical Methods

EXAMINE has the /PLOT NPPLOT subcommand to test normality of a variable. This command produces descriptive statistics (/STATISTICS DESCRIPTIVES), outliers (EXTREME), draws a normal Q-Q plot (/PLOT NPPLOT), and performs the Kolmogorov-Smirnov and Shapiro-Wilk tests.

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

Case Processing Summary

[16] In SPSS 16.0, you may not see P-P and Q-Q under the Graphs menu, which were available in previous versions.

	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

Since N is less than 2,000, we have to read the Shapiro-Wilk statistic and do not reject the null hypothesis of normality (p<.168). Like SAS, 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.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Normal	.027	500	.200*	.996	500	.168

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

6.2 A Non-normally Distributed Variable

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

6.2.1 Graphical Methods

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

```
EXAMINE VARIABLES=gnip
       /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
       /STATISTICS DESCRIPTIVES EXTREME
       /CINTERVAL 95 /MISSING LISTWISE /NOTOTAL.
```

Figure 27 illustrates that the distribution is heavily skewed to the right and there exist many outliers beyond the extreme line in the box plot (right plot). The median and the 25th percentile are close to each other.

Figure 27. Histogram and Box Plot a Non-normally Distributed Variable

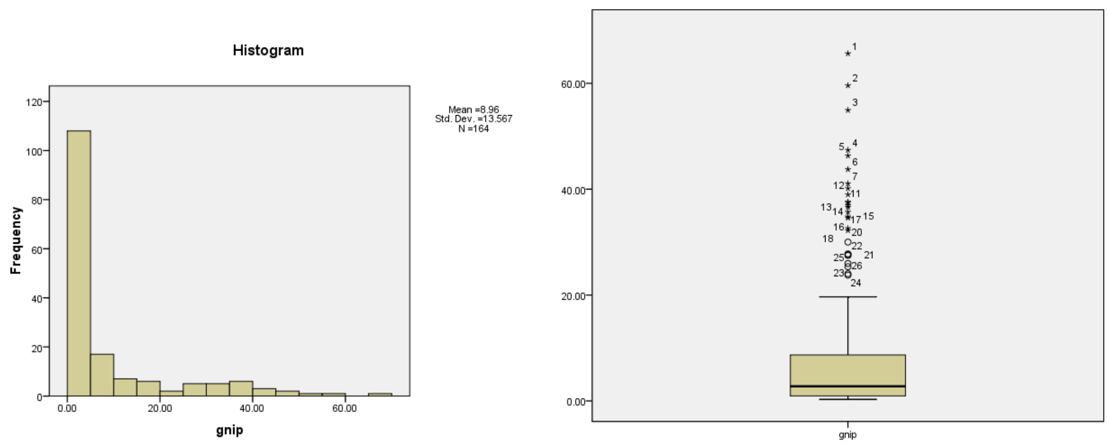
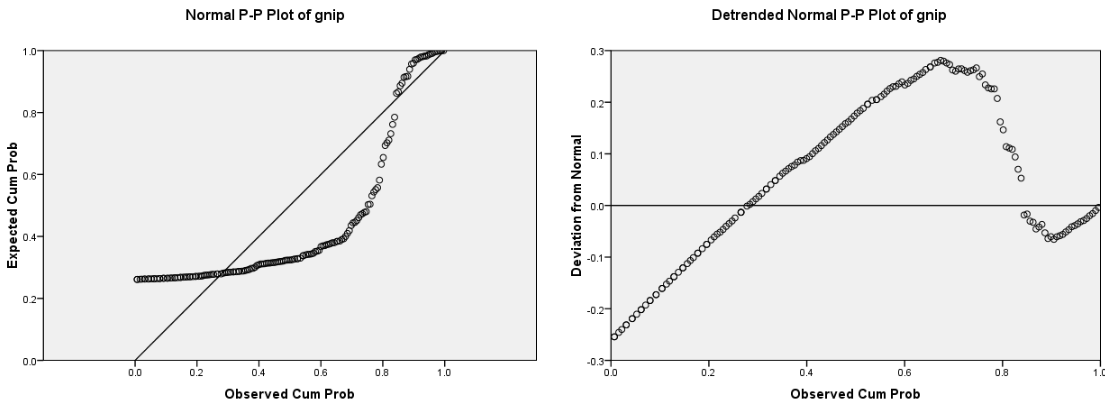


Figure 28 presents the P-P and detrended P-P plots where data points are significantly deviated from the straight fitted line.

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

```
      /DIST=NORMAL.
```

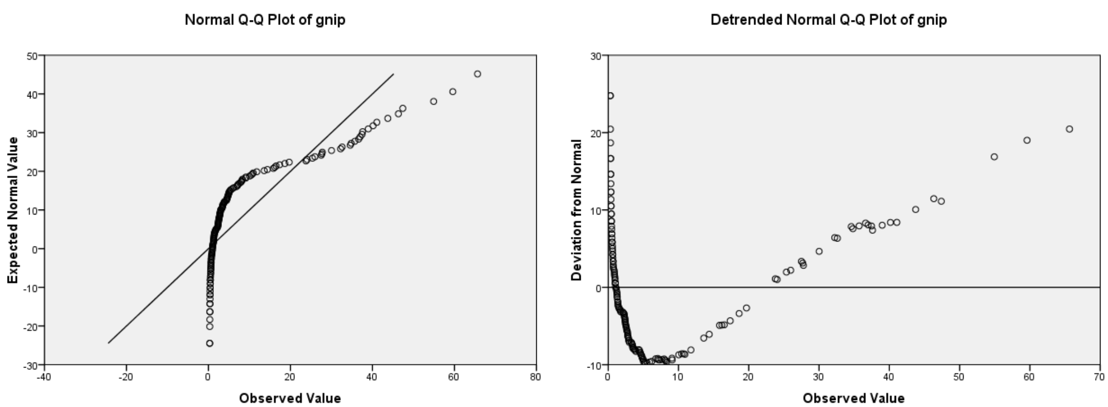
Figure 28. P-P and Detrended P-P Plots of a Non-normally Distributed Variable



The Q-Q and detrended Q-Q plots also show a significant deviation from the fitted line (Figure 26).

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

Figure 29. 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. The following tables are the output of the above EXAMINE command.

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

Extreme Values

			Case Number	Value
gnip	Highest	1	1	65.63
		2	2	59.59
		3	3	54.93
		4	4	47.39
		5	5	46.32
	Lowest	1	164	.29
		2	163	.29
		3	162	.31
		4	161	.33
		5	160	.34[a]

a. Only a partial list of cases with the value .34 are shown in the table of lower extremes.

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 appropriate analysis methods to be employed.

Normality is commonly assumed in many statistical and economic methods, although often conveniently assumed in reality 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). 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 varialbe 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 9).

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 packages, is a consistent method of 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 to conduct 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 Bureau of Labor Statistics. Actual data were downloaded from http://www.stats.indiana.edu/, 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 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.

SPSS Inc. 2007. *SPSS 16.0 Command Syntax Reference*. Chicago, IL: SPSS Inc.

Stata Press. 2007. *Stata Base Reference Manual Release 10*. College Station, TX: Stata Press.

Stata Press. 2007. *Stata Graphics Reference Manual Release 10*. 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. 11 Revision with new data.
- 2008. 11 Revision with new versions of software packages.