Introduction

Statistical methods are based on various assumptions that uphold the methods. One of them is normality, which is commonly assumed. Thus, statistical models often require checking the normality of variables. Otherwise, interpretations and inferences based on the models are not reliable, if not valid.

There are two ways of checking normality. Graphical methods visualize differences between empirical data distribution and theoretical distribution like a normal distribution. Numerical methods conduct statistical tests on the null hypothesis that the variable is normally distributed.

Graphical Methods (Visualization)

Graphical methods visualize the distribution using graphs, such as stem-and-leaf-plot, (skeletal) box plot, dot plot, and histogram. When N is small, a stem-and-leaf plot or dot plot is useful to summarize data; the histogram is appropriate for large N samples. A stem-and-leaf plot assumes continuous variables, while a dot plot works for categorical variables.

A box plot presents 25 percentile, 50 percentile (median), 75 percentile, and mean in a box. If a variable is normally distributed, its histogram looks bell-shaped. Thus, 25 and 75 percentile become symmetry; median and mean are located at the same point, exactly the middle point.

However, the P-P plot and Q-Q plot are commonly used to check normality. These methods provide visual ways of analyzing distributions of variables. The probability-probability plot (P-P plot or percent plot) compares the empirical cumulative distribution function of a variable with a specific theoretical cumulative distribution function (e.g., the standard normal distribution function).

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 form a linear pattern that passes through the origin and has a unit slope. So, the P-P plot and the Q-Q plot are used to see how well a theoretical distribution models the empirical data. STATA has a nice feature of producing P-P and Q-Q plot.

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

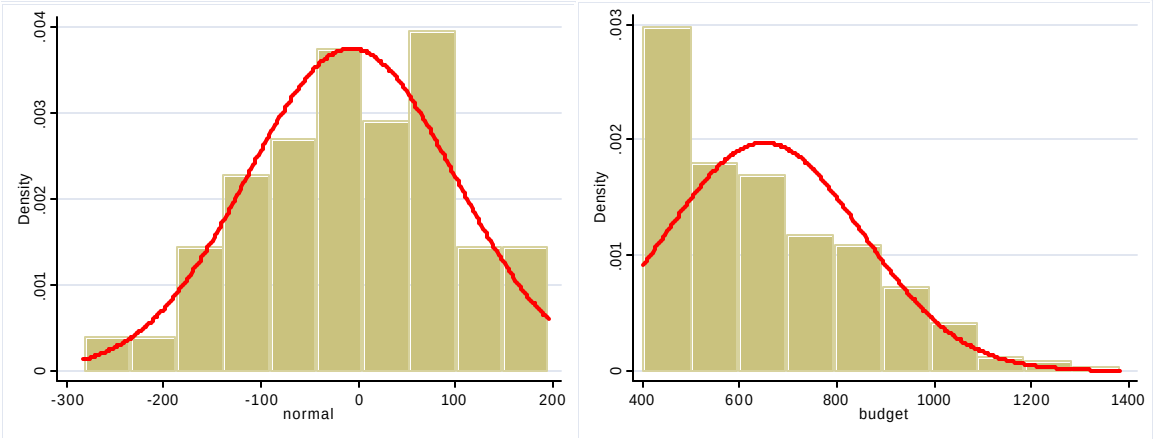
Although visually appealing, these graphical methods do not provide objective criteria to determine the normality of variables. Interpretations are matter of judgments. Here is necessity of numerical methods.

Numerical Methods (Statistical Test)

Skewness (the third central moment) and Kurtosis (the fourth moment) are commonly used descriptive statistics. They show how the distribution of a variable deviates from symmetry and how thick the tails of the distribution are.[1]

$$\text{Skewness}=\frac{E[(x-m)^3]}{[Var(x)]^{3/2}}=\frac{\sum(x-\bar{x})^3}{\left(\sum(x-\bar{x})^2\right)^{3/2}}$$

$$\text{Kurtosis}=\frac{E[(x-m)^4]}{[Var(x)]^2}$$



If a variable is normally distributed, its Skewness and Kurtosis are zero and three, respectively (see the left histogram above). If Skewness is greater than zero, the distribution is skewed to the right, having more observations on the left (see the right histogram above). If Kurtosis is less than three (or if Kurtosis -3 is less than zero), the distribution has thicker tails compared to the normal distribution. Like graphical methods, however, Skewness and Kurtosis do not provide a conclusive way of interpretations.

Table 1. Numerical Methods of Testing Normality

Test	Stat.	N	Dist.	SAS	STATA	SPSS
Jarque-Bera (S-K) test	c^2	-	$c^2(2)$	Manually	. sktest	Manually
Shapiro-Wilk	W	7=N= 2,000	-	YES	. swilk	YES
Shapiro-Francia	W	5=N= 5,000	-	-	. sfrancia	-
Kolmogorov-Smirnov	D	> 2,000	EDF	YES	*	YES
Cramer-vol Mises	W^2	> 2,000	EDF	YES	-	-
Anderson-Darlling	A^2	> 2,000	EDF	YES	-	-

* STATA . ksmirnov command is used for the one or two samples Kolmogorov-Smirnov test.

Numerical methods of testing normality include the Kolmogorov-Smirnov (K-S) test (Lilliefors test), Shapiro-Wilk' test, Anderson-Darling test, and the Cramer-von

[1] SAS and SPSS produce Kurtosis -3, while STATA gives us the Kurtosis. SAS uses its weighted kurtosis formula. So, if N is small, it may reports a different Kurtosis.

Mises test (SAS Institute 1995).[2] The K-S D test and Shapiro-Wilk' W test are commonly used. 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).

The Shapiro-Wilk statistic (1965) is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance.[3] The statistic is positive and less than or equal to one; being close to one indicate normality of the variable. The W statistic requires that the sample size need to greater than or equal to seven and less than or equal to 2,000.

$$W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$$, where a'=$(a_1, a_2, \ldots, a_n) = m'V^{-1}[m'V^{-1}V^{-1}m]^{-1/2}$, 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 test is an approximate test that modified the Shapro-Wilk test. The S-F statistic uses b'=$(b_1, b_2, \ldots, b_n) = m'(m'm)^{-1/2}$ 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 three to 5,000. However, SAS and SPSS do not provide this statistic.

The K-S D statistic, a supremum class of empirical distribution function statistics, is based on the largest vertical difference between $F(x)$ and $F_n(x)$.[4] The K-S D statistic is computed when the sample size is greater than 2000. Anderson-Darling A^2 and Cramer-von Misers W^2 are based on the squared difference $\left(F_n(x) - F(x)\right)^2$ The SAS UNIVARIATE and CAPABILITY procedures use the modified Kolmogorov-Smirnov D statistic to test the data against a normal distribution with mean and variance equal to the sample mean and variance (SAS Institute 1995).

However, these numerical methods tend to reject the null hypothesis when N becomes large (see the experiment result in the next section). Given a large number of observations, the Jarque-Bera test and STATA version of the Skewness and Kurtosis test will be promising alternatives of testing normality. The Jarque-Bera statistic is known to follow the chi-squares distribution with two degrees of freedom. Under the null hypothesis of normality, the expected value of the statistic is two. Note that $6/n$ and $24/n$ are respectively variances of Skewness and Kurtosis.

---

[2] The UNIVARIATE procedure has the NORMAL option to produce four statistics, while the CAPABILITY procedure has the NORMAL or NORMALTEST option.

[3] 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 original W statistic (1965) 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 seven and 2000.

[4] $F_n(x)$ is a step function that takes a step of height 1/n at each observation (SAS Institute 1995).

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

The above formula gives a penalty of increasing the number of observations. According to the Central Limit Theorem, the Skewness and Kurtosis-3 approach zero as N goes infinity. These support a good asymptotic property of the Jarque-Bera test.

STATA provides the .sktest command to compute the statistic, which is based on D'Agostino, Belanger, and D'Agostino, Jr. (1990) and Royston(1991). In SAS and SPSS, researchers need to manually compute or write a program to get the Jarque-Bera statistic.

Table 2. Comparison of Graphical Methods and Numerical Methods

	Graphical Methods	Numerical Methods
Common Methods	P-P plot Q-Q plot	Shapiro-Wilk, Shapiro- Francia test Kolmogorov-Smirnov test (Lillefors test) Anderson-Darling/Cramer-von Mises tests Jarque-Bera (Skewness and Kurtosis) test
Descriptive	Leaf-stem-plot, (skeletal) boxplot, dot plot, and histogram	Skewness Kurtosis
Pros and Cons	Easy to read (intuitive) Not conclusive (subjective evaluation)	Providing objective criteria

The above table summarizes graphical and numerical methods for testing normality. Note that graphical methods' visualization makes it easy to read, while numerical methods provide objective criteria for evaluating normality. Computation difficulty in the numerical methods is no longer a major problem these days.

Table 3. Comparison of Procedures and Commands Available

	SAS 8.2	STATA 8.0 SE	SPSS 11.0
Descriptive statistics (Skewness/Kurtosis)	UNIVARIATE	.summarize .tabstat	Descriptives, Frequencies Examine
Stem-leaf-plot	UNIVARIATE*	.stem	Examine
Histogram, dot plot	UNIVARIATE CHART, PLOT	.histogram*** .dotplot	Graph, Igraph, Examine, Frequencies
Box plot	UNIVARIATE*		Examine, Igraph
P-P plot	CAPABILITY**	.pnorm	Pplot
Q-Q plot	UNIVARIATE	.qnorm	Examine, Pplot
Detrended Q-Q/P-P plot			Pplot, Examine
Jarque-Bera (S-K) test		.sktest	
Shapiro-Wilk	UNIVARIATE	.swilk	Examine
Shapiro-Francia	UNIVARIATE	.sfrancia	
Kolmogorov-Smirnov	UNIVARIATE		Examine
Cramer-vol Mises	UNIVARIATE		
Anderson-Darlling	UNIVARIATE		

* Only the UNIVARIATE procedure can provide the graph.
** Only the CAPABILITY procedure can provide the graph.
*** The command is newly added in version 8.0. It is equivalent to ".graph *var*, normal" in 7.0.

Testing Normality in SAS

SAS provides the UNIVARIATE and CAPABILITY procedures to draw graphs and conduct statistical tests on normality. Two procedures have the almost similar features, while the UNIVARIATE is included in the SAS/BASE module and the CAPABILITY in the SAS/QC.

Suppose we have one variable with 100 observations, which were randomly generated from a normal distribution. Note that the seed 1234567 is used for replication.

```
DATA normal;
DO i=1 to 100;
Normal=INT(NORMAL(1234567)*100); OUTPUT;
END;
```

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

```
SYMBOL V=SQUARE COLOR=GREEN H=.5;
PROC UNIVARIATE NORMAL PLOT;
VAR normal;
QQPLOT normal /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
INSET MEAN STD /CFILL=BLANK FORMAT=5.2 ;
RUN;
```

The NORMAL option is specified to conduct normality testing. The PLOT option draws a leaf-and-stem plot and a box plot. The QQPLOT statement is used to draw a Q-Q plot. The CAPABILITY procedure also produces the similar result except a leaf-and-stem plot, a box plot, and a normal probability plot. Unlike the UNIVARIATE, the procedure has PPPLOT option to draw a P-P plot.

Following is an example of the CAPABILITY procedure. Note that the procedure does not provide a stem plot, a box plot, and a normal probability plot.

```
PROC CAPABILITY NORMALTEST;
VAR Normal;
PPPLOT Normal /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
QQPLOT Normal /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
INSET MEAN STD /CFILL=BLANK FORMAT=5.2 ;
RUN;
```

Moments

N	100	Sum Weights	100
Mean	-7.07	Sum Observations	-707
Std Deviation	106.577639	Variance	11358.793
Skewness	-0.2280012	Kurtosis	-0.503857
Uncorrected SS	1129519	Corrected SS	1124520.51
Coeff Variation	-1507.4631	Std Error Mean	10.6577639

Basic Statistical Measures

Location		Variability	
Mean	-7.070	Std Deviation	106.57764
Median	-5.500	Variance	11359
Mode	-147.000	Range	479.00000
		Interquartile Range	160.50000

NOTE: The mode displayed is the smallest of 9 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-0.66337	Pr > |t|	0.5086
Sign	M	-2	Pr >= |M|	0.7644
Signed Rank	S	-120	Pr >= |S|	0.6820

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.984225	Pr < W	0.2789
Kolmogorov-Smirnov	D	0.068943	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.077217	Pr > W-Sq	0.2288
Anderson-Darling	A-Sq	0.459524	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Quantile	Estimate
100% Max	196.0
99%	195.5
95%	166.5
90%	117.5
75% Q3	74.5
50% Median	-5.5
25% Q1	-86.0
10%	-143.5
5%	-180.0
1%	-265.0
0% Min	-283.0

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
-283	29	171	25
-247	73	177	4
-218	19	192	56
-213	46	195	43
-181	77	196	16

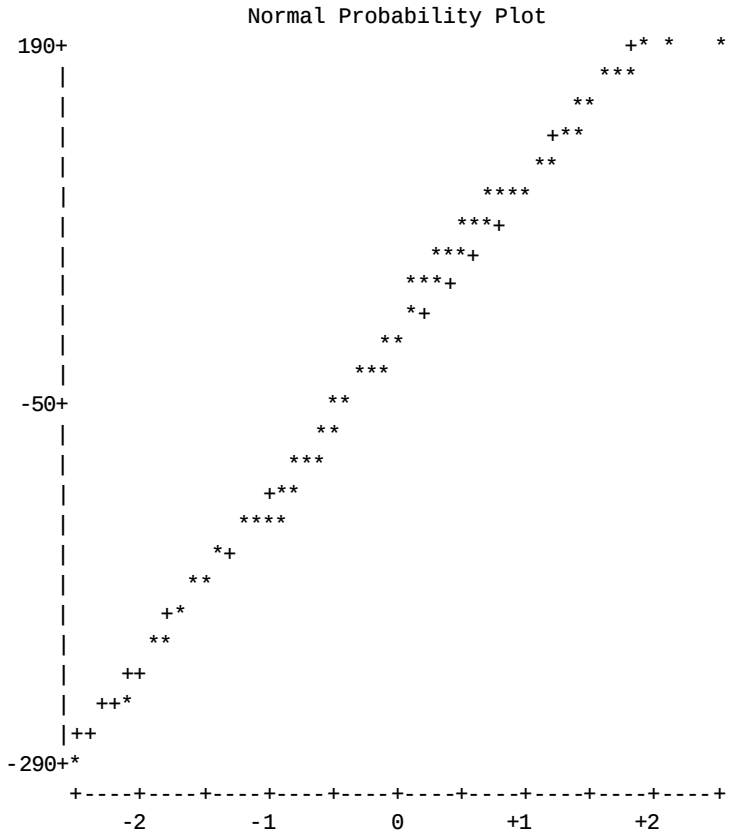
Skewness and Kurtosis-3 are respectively -.2280 and -.5039, indicating a symmetric distribution with thicker tails. However, these statistics do not provide conclusive information for normality. Note that when N is small (e.g., 10), the weighted Kurtosis formula of SAS produces a quite different value (see the experiment result in page nine).

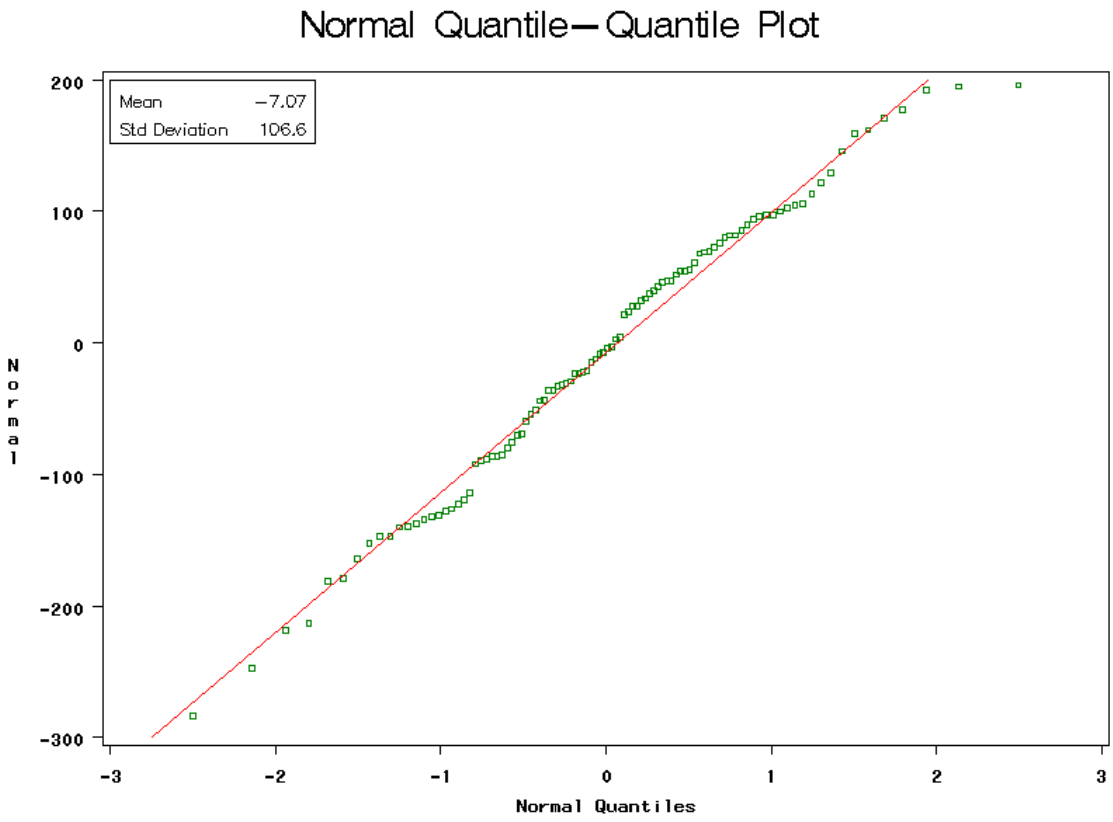
SAS provides four different statistics for testing normality. Since the number of observations is less than 2,000, we have to take a look at the Shapiro-Wilk W statistic and it p value. W and its p are respectively .9842 and .2789, providing solid evidence not to reject the null hypothesis that the variable is normally distributed. Although rest three statistics do not reject the null hypothesis, it is not relevant to interpret them. The Jarque-Bera test also indicates normality of the variable at the five percent level (p value of .38).

$$100\left[\frac{-.2280^2}{6}+\frac{-.5039^2}{24}\right]\sim 1.9244(2)$$

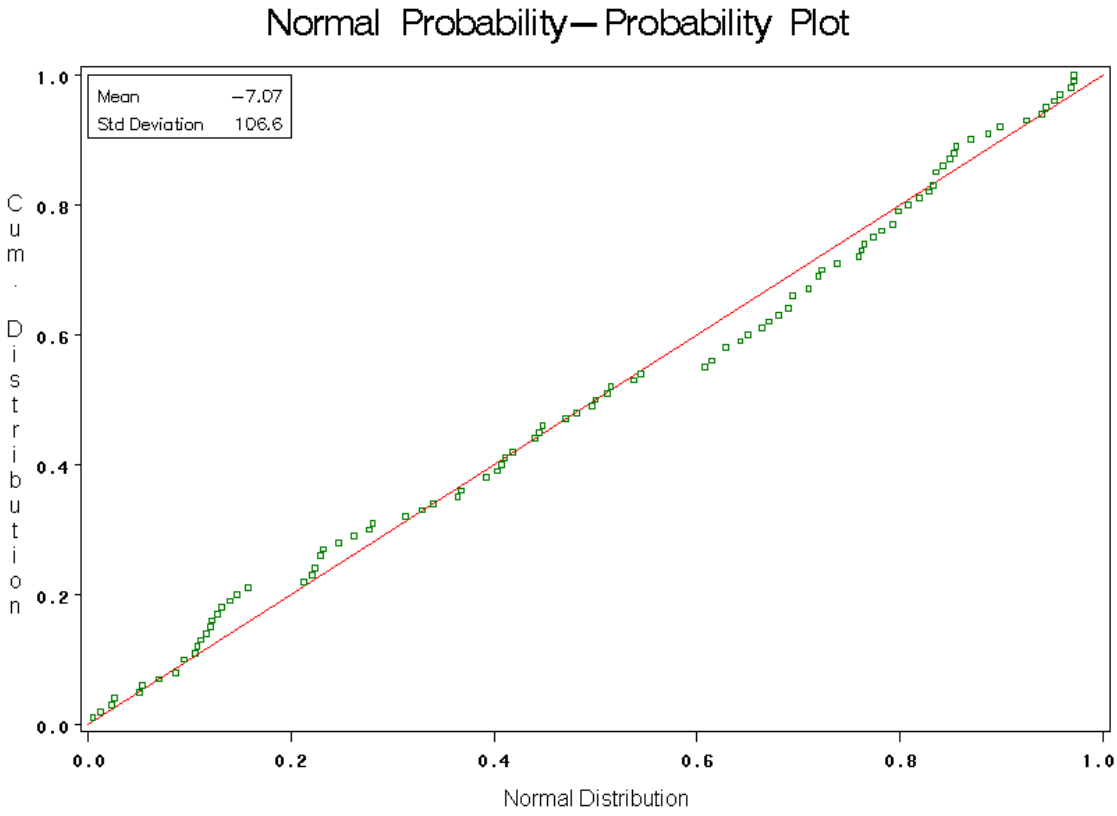
The stem-and-leaf plot and box plot illustrate that the variable is normally distributed. The locations of first quantile, mean, median, and third quintile indicate a bell-shaped distribution. Note that the mean -7.07 and median -5.5 are very close. The normal probability plot shows a straight line, implying normality of the variable.

```
Stem Leaf                     #  Boxplot
  18 256                      3     |
  16 217                      3     |
  14 69                       2     |
  12 29                       2     |
  10 03563                    5     |
   8 022604677                9     |
   6 189036                   6  +-----+
   4 036772556                9  |     |
   2 2488248                  7  |     |
   0 35                       2  |     |
  -0 528743                   6  *--+--*
  -2 6632193321              10  |     |
  -4 94143                    5  |     |
  -6 509                      3  |     |
  -8 2986650                  7  +-----+
 -10 94                       2     |
 -12 97421862                 8     |
 -14 2770                     4     |
 -16 94                       2     |
 -18 1                        1     |
 -20 83                       2     |
 -22                                |
 -24 7                        1     |
 -26                                |
 -28 3                        1     |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```


Normal Probability Plot
190+
-50+
-290+
-2
-1
0
+1
+2


Normal Quantile–Quantile Plot
Mean -7.07
Std Deviation 106.6
200
100
0
-100
-200
-300
Normal
-3
-2
-1
0
1
2
3
Normal Quantiles

The P-P and Q-Q plots show that the data points are not seriously deviated from the straight line. They consistently indicate that the variable is normally distributed.



The following table compares the experiment results when N increases. The data were generated from the normal random generator with the same seed of 1234567. As N grows, the mean and median approach zero, while standard deviation, 1st and 3rd quantiles approach certain values. Of course, Skweness and Kurtosis become zero as the number of observation beyond 5,000.

N	10	100	500	1,000	5,000	10,000	100,000
Mean	52.0000	-7.0700	-9.4660	-.9540	-1.5204	-1.9052	-.1096
Std	95.1210	106.5776	99.9414	100.5120	100.6815	100.2597	99.8342
Median	63.5000	-5.5000	-11.5000	-3.0000	-2.0000	-2.0000	.0000
1st quantile	-23.0000	-86.0000	-80.0000	-70.5000	-70.0000	-71.0000	-67.0000
3rd quantile	146.0000	74.5000	61.0000	70.0000	66.0000	64.0000	67.0000
Skewness	-.1602	-.2280	-.0225	.0100	-.0386	-.0391	-.0028
Kurtosis	-1.4471	-.5039	-.3860	-.2494	.0094	-.0044	-.0160
Jarque-Bera	.9153 (.63)	1.9244 (.38)	3.1463 (.21)	2.6083 (.27)	1.2600 (.53)	2.5561 (.28)	1.1973 (.55)
S-W W	.9366 (.52)	.9842 (.28)	.9956 (.17)	.9980 (.30)	-	-	-
K-S D	.1377 (.15)	.0689 (.15)	.0289 (.15)	.0188 (.15)	.0113 (.12)	**.0109 (.01)**	**.0058 (.01)**
C-M W^2	.0340 (.25)	.0772 (.23)	.0866 (.18)	.0592 (.25)	.0520 (.25)	.0932 (.14)	**.0325 (.01)**
A-D A^2	.2491 (.25)	.4595 (.25)	.5502 (.16)	.4199 (.25)	.2864 (.25)	.5171 (.20)	**1.464 (.01)**

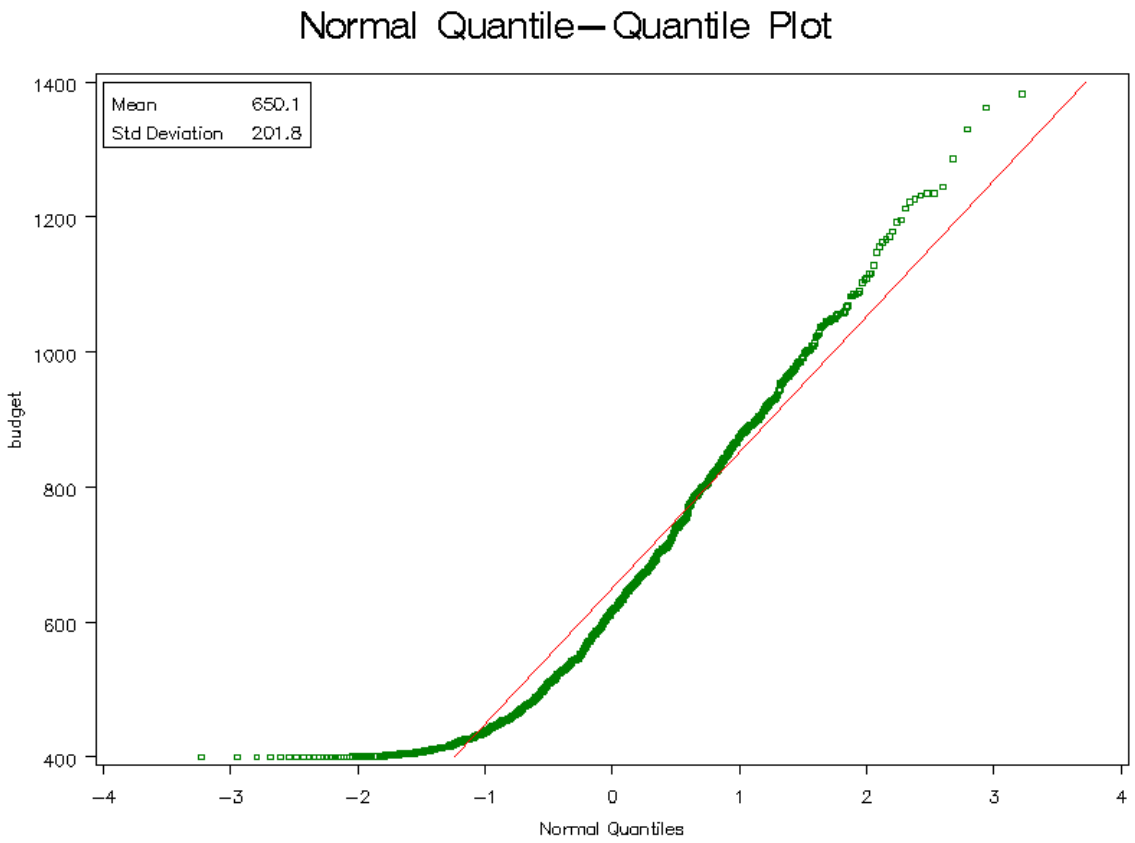
It is notable that the S-W W statistic is not reported when N is greater than 2,000. All four statistics do not reject the null hypothesis as long as N is not 10,000 or 100,000, where K-S D test rejects the null hypothesis at the 1 percent level. This result illustrates

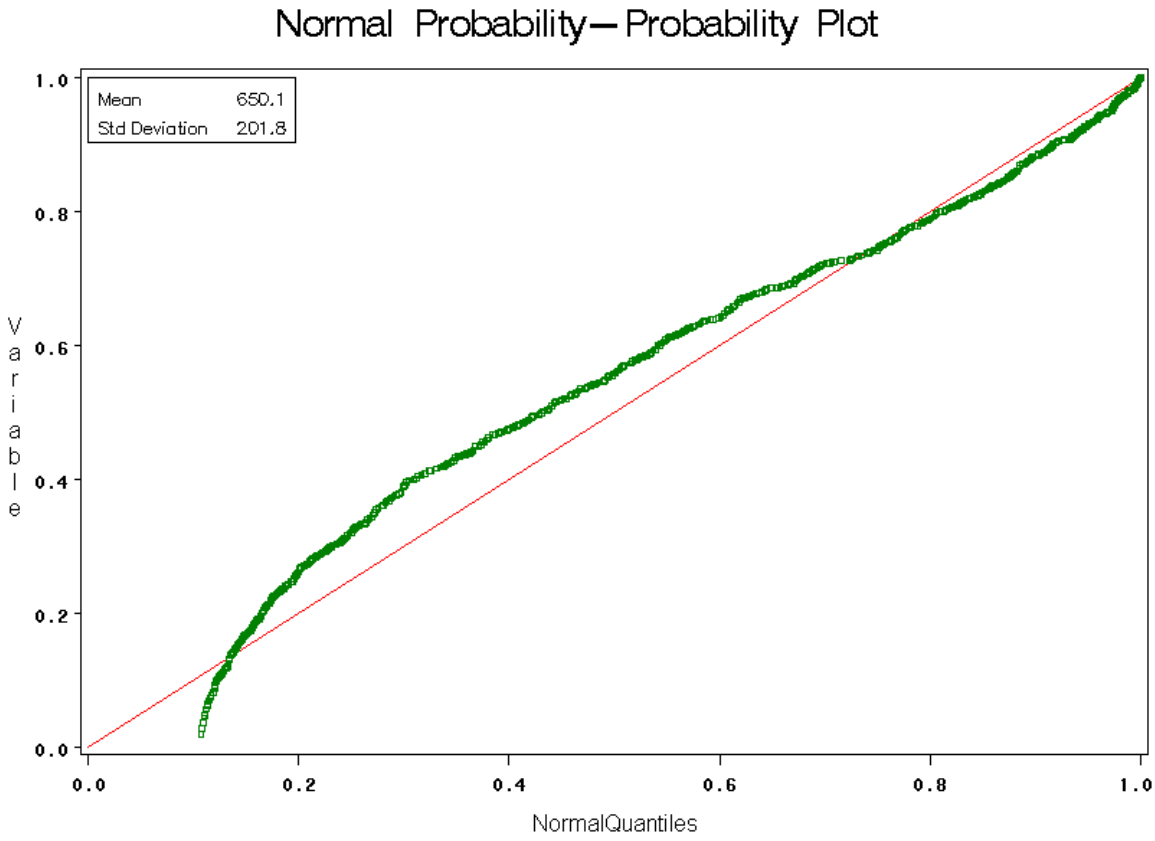
that the statistical tests are likely to reject the null hypothesis of normality as N grows large. Interestingly, the Jarque-Bera test show consistent results regardless of the number of observations. It seems that the asymptotic property makes the statistic robust.

Let us take an example of not being normally distributed. Following are selective statistics of the UNIVARIATE output. The large Skewness of .77 and the box plot indicate that the variable is seriously skewed to the top. Not that the histogram replaces the stem-leaf-plot since the number or observations is large.

```
                          Moments

N                        1000    Sum Weights               1000
Mean                   650.126    Sum Observations        650126
Std Deviation      201.844215    Variance            40741.0872
Skewness           0.77211688    Kurtosis            -0.0038015
Uncorrected SS      463364162    Corrected SS         40700346.1
Coeff Variation    31.0469379    Std Error Mean      6.38287453

                  Tests for Normality

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

Shapiro-Wilk          W     0.931252    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.107635    Pr > D     <0.0100
Cramer-von Mises      W-Sq  2.668104    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  17.88223    Pr > A-Sq  <0.0050

                         Histogram                        #  Boxplot
   1375+*                                                 2     0
        .*                                                1     0
        .*                                                1     0
        .**                                               7     |
        .**                                               7     |
        .**                                               7     |
        .*****                                           17     |
        .******                                          22     |
        .********                                        29     |
        .*********                                       35     |
        .**************                                  53     |
        .***************                                 58     |
        .*************                                   50  +-----+
        .******************                              72  |     |
        .**********************                          84  |  +  |
        .********************                            80  *-----*
        .******************                              76  |     |
        .***************************                     105  |     |
        .***************************                     108  +-----+
    425+*********************************************   186     |
        ----+----+----+----+----+----+----+----+----+--
        * may represent up to 4 counts
```

The Shapiro-Wilk test rejected the null hypothesis. The Jarque-Bera test (99.3621, p<.0000) also indicates that the variable is not normally distributed. Q-Q and P-P plots also support our conclusion of non-normality.


Normal Quantile—Quantile Plot
Mean 650.1
Std Deviation 201.8
budget
1400
1200
1000
800
600
400
−4
−3
−2
−1
0
1
2
3
4
Normal Quantiles


Normal Probability—Probability Plot
Mean 650.1
Std Deviation 201.8
Variable
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
NormalQuantiles

Testing Normality in STATA

In STATA, researchers have to use individual commands to get adequate statistics. The .summarize and .tabstat commands produce descriptive statistics. The .stem command generates a stem-and leaf plot, while the .histogram draws a histogram. The .pnorm and .qnorm respectively produce standardized normal P-P and Q-Q plots.

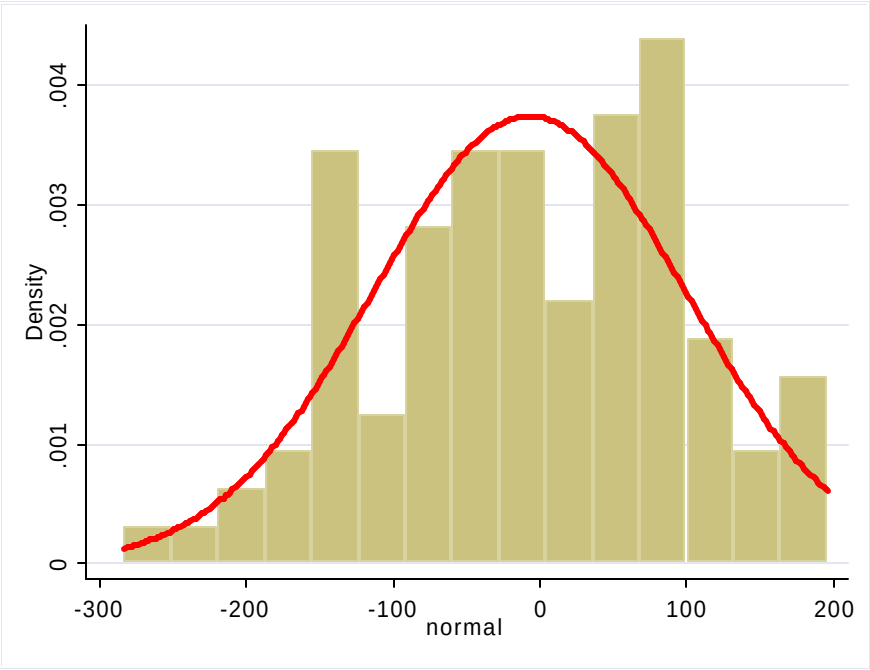
```
. summarize normal, detail

                        variable
-------------------------------------------------------------
      Percentiles      Smallest
 1%         -265           -283
 5%         -180           -247
10%       -143.5           -218       Obs                 100
25%          -86           -213       Sum of Wgt.         100
50%         -5.5                      Mean              -7.07

                       Largest       Std. Dev.      106.5776
75%         74.5            177
90%        117.5            192       Variance       11358.79
95%        166.5            195       Skewness      -.2245668
99%        195.5            196       Kurtosis       2.46158
```

```
. tabstat normal, 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	variable
N	100
mean	-7.07
sum	-707
max	196
min	-283
range	479
sd	106.5776
variance	11358.79
se(mean)	10.65776
skewness	-.2245668
kurtosis	2.46158
p50	-5.5
p1	-265
p5	-180
p10	-143.5
p25	-86
p50	-5.5
p75	74.5
p90	117.5
p95	166.5
p99	195.5
iqr	160.5
p25	-86
p50	-5.5
p75	74.5

It is notable that the Skewness and Kurtosis are different from those of SAS and SPSS. STATA gives us Skewness of -.2245668, which is slightly smaller than -.2280012 in SAS. STATA's Kurtosis of 2.46158 is slightly greater than SAS's 2.4961(=-.503857+3). It is due to the difference in formula used in SAS and STATA.

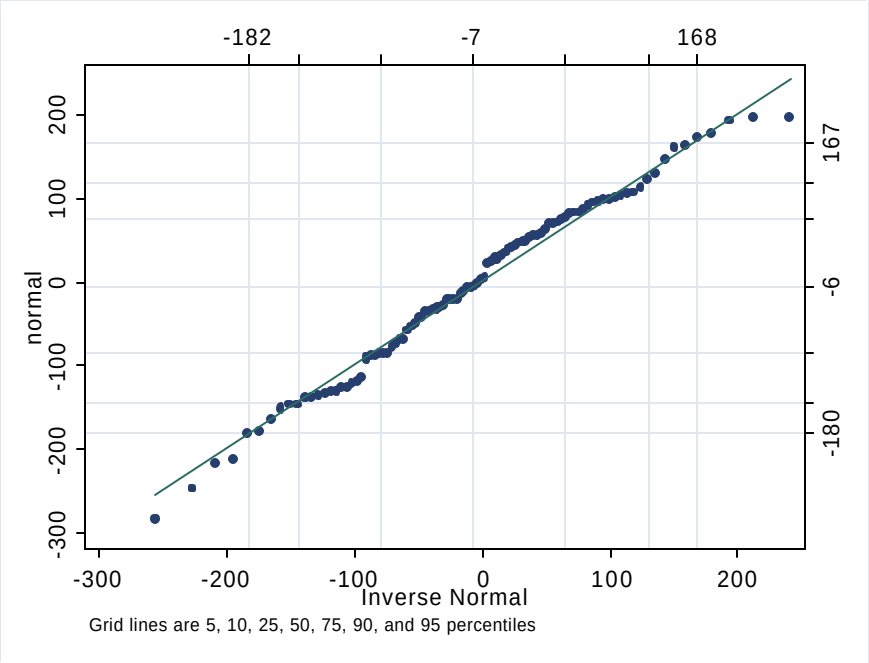
```
. stem normal

Stem-and-leaf plot for normal
 -2** | 83
 -2** |
 -2** | 47
 -2** |
 -2** | 18,13
 -1** | 81
 -1** | 79,64
 -1** | 52,47,47,40
 -1** | 39,37,34,32,31,28,26,22
 -1** | 19,14
 -0** | 92,89,88,86,86,85,80
 -0** | 75,70,69
 -0** | 59,54,51,44,43
 -0** | 36,36,33,32,31,29,23,23,22,21
 -0** | 15,12,08,07,04,03
  0** | 03,05
  0** | 22,24,28,28,32,34,38
  0** | 40,43,46,47,47,52,55,55,56
  0** | 61,68,69,70,73,76
  0** | 80,82,82,86,90,94,96,97,97
  1** | 00,03,05,06,13
  1** | 22,29
  1** | 46,59
  1** | 62,71,77
  1** | 92,95,96
```



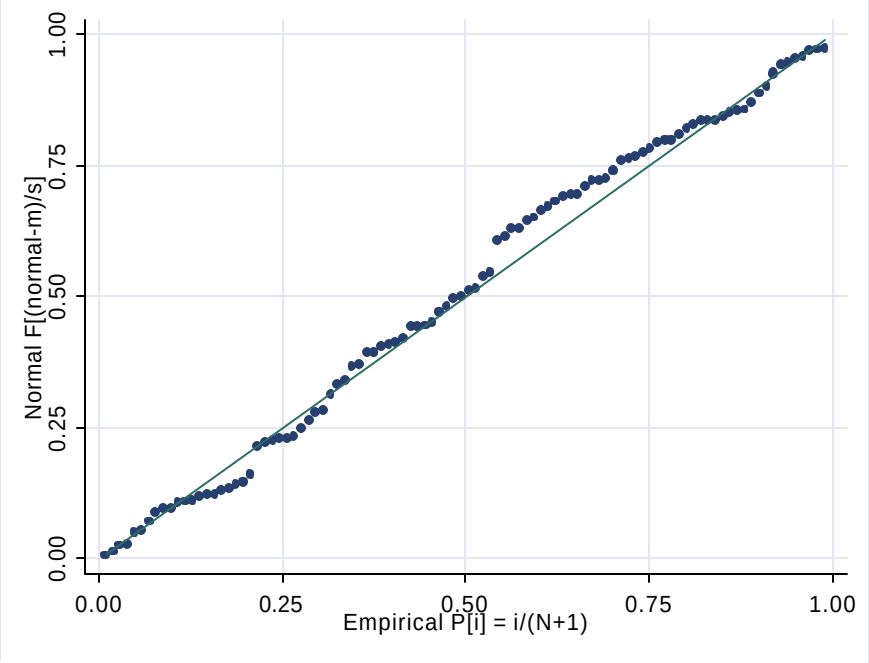
```
. histogram normal, bin(15) normal normopts( clcolor(red) clwidth(thick) )
graphregion(fcolor(white) lcolor(none))
. dotplot normal, median graphregion(fcolor(white) lcolor(none))
```

The above .histogram command draws a histogram. The .dotplot and ".graph box" commands respectively produce a dot plot and a box plot of the variable. Now, let us draw a Q-Q plot and a P-P plot using the .qnorm and .pnorm commands below.

```
. qnorm normal, grid graphregion(fcolor(white) lcolor(none))
```




```
. pnorm normal, grid graphregion(fcolor(white) lcolor(none))
```



STATA is able to conduct the Skewness-Kurtosis test, Shapiro-Wik test, and Shapiro-Francia test using the .sktest, .swilk, and .sfrancia command, respectively. All the three tests consistently show that the variable is normally distributed. Note that the "noadjust" option in .sktest suppresses the empirical adjustment made by Royston (1991).

```
. sktest normal

                  Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      normal |      0.334         0.220            2.50       0.2864

. sktest normal, noadjust

                  Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      normal |      0.334         0.220            2.44       0.2957
```

The STATA Skewness-Kurtosis test produces Chi squares of 2.44 (p<.30), which is larger than Jarque-Bera statistic of 1.9244 (p < .38) in SAS. The Jarque-Bera in STATA is calculated as 2.0484045= 100*[(-.2245668)^2/6+(2.46158-3)^2/24] (p<.3591). It is because STATA produces different Skewness and Kurtosis.

```
. swilk normal

                   Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
      normal |    100    0.98428      1.298      0.579  0.28140

. sfrancia normal

                  Shapiro-Francia W' test for normal data
    Variable |    Obs        W'         V'         z     Prob>z
-------------+-------------------------------------------------
      normal |    100    0.98755      1.125      0.240  0.40498
```

The .swilk test produces the same statistic as that of SAS, although the slight difference under decimal point.

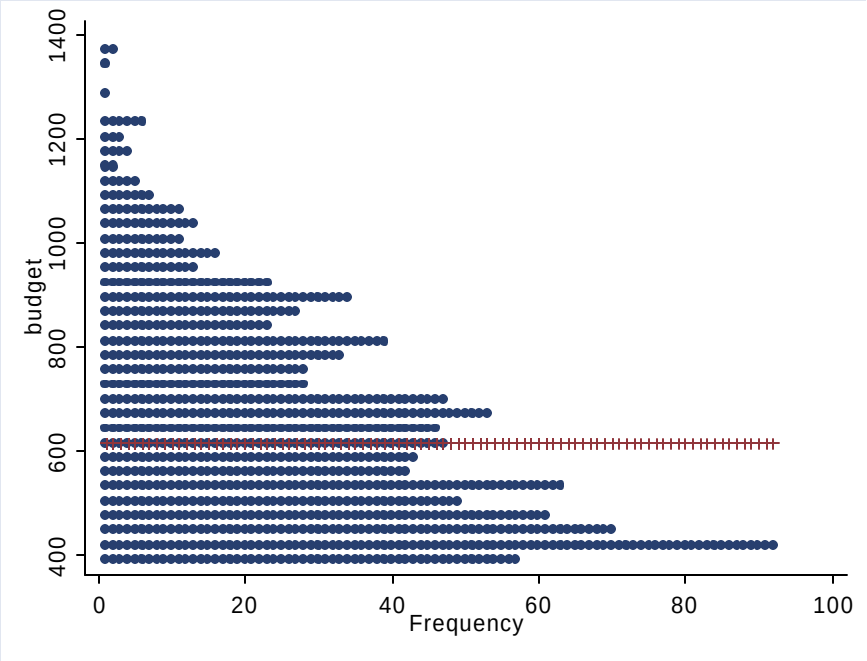
Now, consider the other variable that is not likely to be normally distributed. Note that the variable has large Skewness of .7710 and large difference between its median and mean (616 versus 650).

```
. summarize budget, detail

                          budget
-------------------------------------------------------------
      Percentiles      Smallest
```

```
 1%          400             400
 5%          405             400
10%          416             400       Obs                 1000
25%          477             400       Sum of Wgt.         1000

50%          616                       Mean            650.126
                        Largest       Std. Dev.       201.8442
75%          790            1286
90%          930            1330       Variance        40741.09
95%       1037.5            1362       Skewness        .7709582
99%         1218            1383       Kurtosis        2.990223
```

The following dot plot shows the distribution is seriously skewed to the top. The red line indicates the median of the variable.

```
. dotplot budget, median graphregion(fcolor(white) lcolor(none))
```

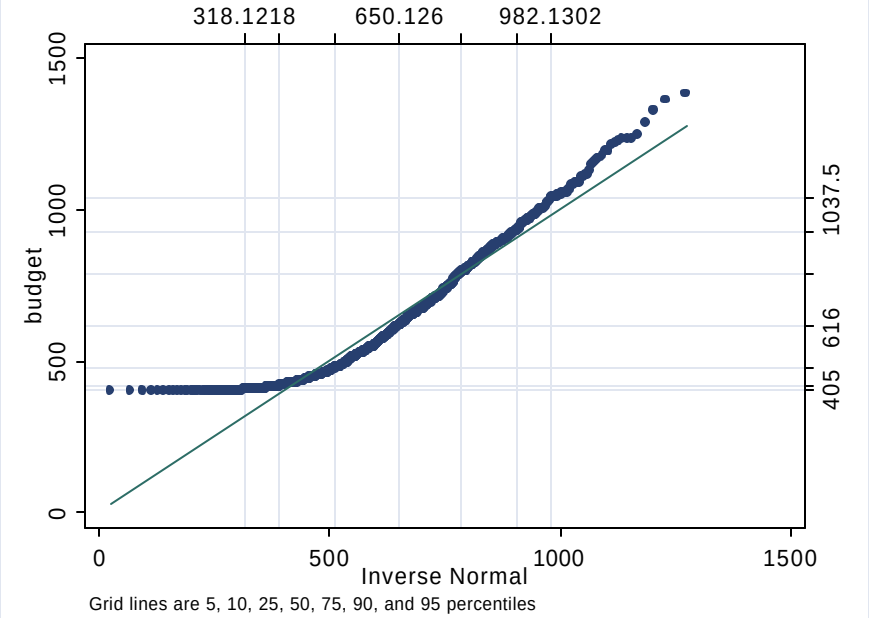


The Skewness-Kutosis test and Shapiro-Wilk test allow us to reject the null hypothesis that the variable is normally distributed.

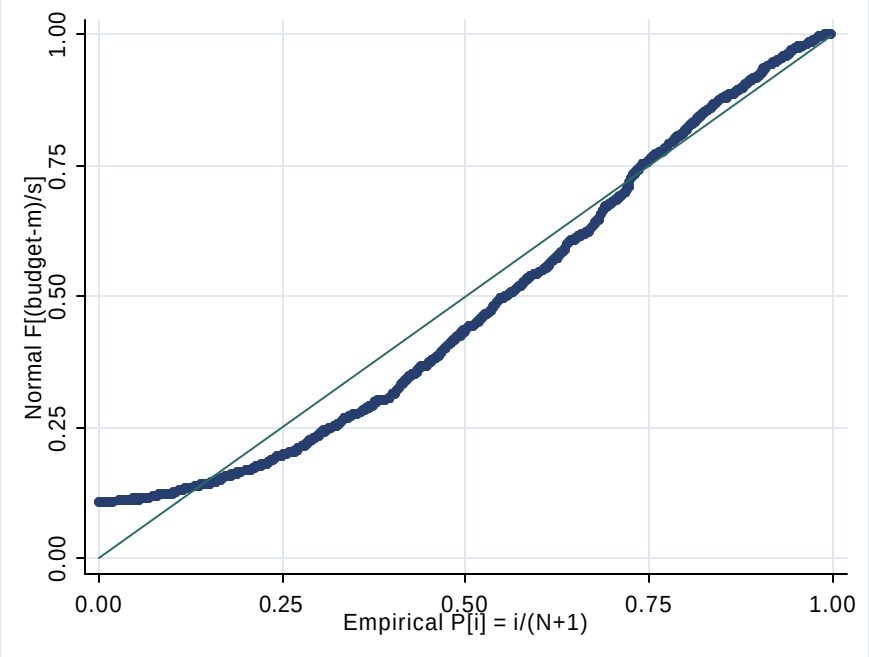
```
. sktest budget, noadjust

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      budget |      0.000         0.960           80.15       0.0000

. swilk budget

                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
      budget |   1000    0.93243     42.617      9.292  0.00000
```

The .qnorm and .pnorm commands respectively generate the Q-Q and P-P plots. Both plots show that data points are systemically deviated from the straight line. Thus, we can conclude that the variable is not likely to be normally distributed. Compare the plots with those of the normally distributed variable.

```
. qnorm budget, grid graphregion(fcolor(white) lcolor(none))
```




```
. pnorm budget, grid graphregion(fcolor(white) lcolor(none))
```



Testing Normality in SPSS

SPSS has the DESCRIPTIVES and FREQUENCIES commands to produce descriptive statistics. GRAPH and IGRAPH commands draw a histogram and box plot. PPLOT command produces P-P and Q-Q plots. Like the UNIVARIATE procedure of SAS, the EXAMINE command of SPSS can produce both descriptive statistics and various plots, such as a stem-leaf-plot, histogram, box plot, and Q-Q plot. Unlike SAS and STATA, however, SPSS is able to draw the detrended Q-Q plot easily using the EXAMINE command. The EXAMINE command also conducts Kolmogorov-Smirnov test and Shapiro-Wilk test for normality.

The DESCRIPTIVES command is usually applied to continuous variables and the FREQUENCIES command to categorical variables. But they both produce similar descriptive statistics including Skewness and Kurtosis. The statistics are specified in the /STATISTICS subcommand using corresponding keywords. The output of the DESCRIPTIVES command is skipped in this paper, since the command gives us a long single row of statistics. [5]

```
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.
```

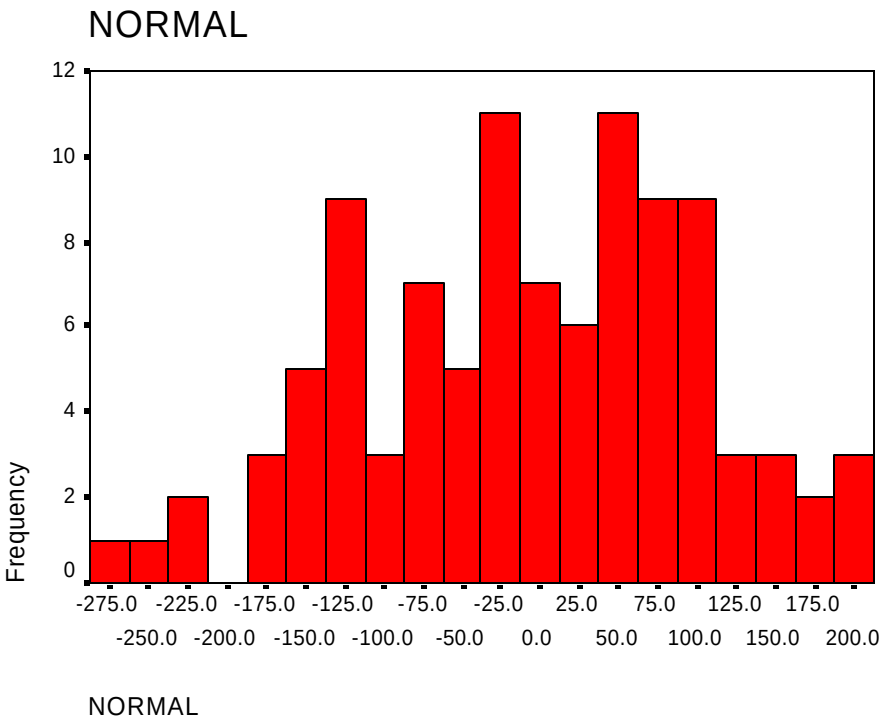
N	Valid	100
	Missing	0
Mean		-7.07
Std. Error of Mean		10.658
Median		-5.50
Mode		-147(a)
Std. Deviation		106.578
Variance		11358.793
Skewness		-.228
Std. Error of Skewness		.241
Kurtosis		-.504
Std. Error of Kurtosis		.478
Range		479
Minimum		-283
Maximum		196
Sum		-707
Percentiles	25	-86.00
	50	-5.50
	75	75.25

a Multiple modes exist. The smallest value is shown

[5] You may click Analyze→Descriptive Statistics on the menu to use the commands.

Note that the /HISTOGRAM subcommand of the FRIQUENCIES command ask SPSS to draw a histogram of the variable. You can get the identical result using the following GRAPH command or the EXAMINE command. The following IGRAPH command also produces the similar histogram.

```
GRAPH /HISTOGRAM=normal.

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.
```

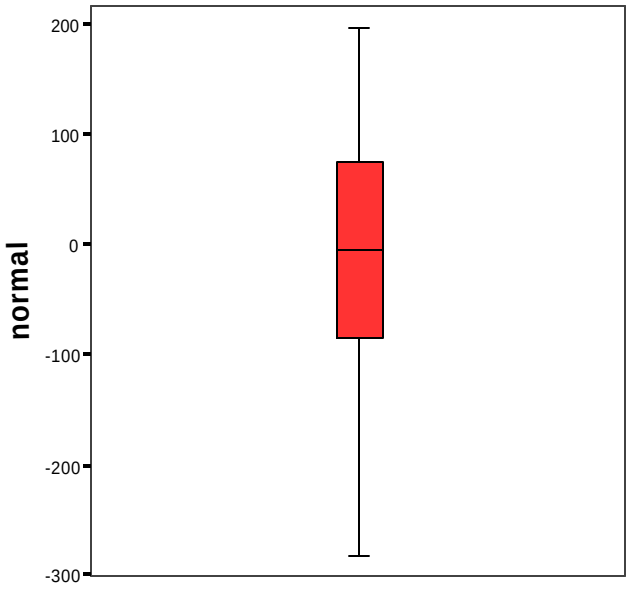


The EXAMINE command can draw a stem-and-leaf plot using the /PLOT subcommand with the STEMLEAF option.

```
VARIABLE Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00       -2 .  8
     3.00       -2 .  114
     4.00       -1 .  5678
    13.00       -1 .  1122233333444
    13.00       -0 .  5556778888889
    18.00       -0 .  000011222223333344
    14.00        0 .  00222233344444
    19.00        0 .  5555666777888899999
     8.00        1 .  00001224
     7.00        1 .  5677999

 Stem width:       100
 Each leaf:       1 case(s)
```

The following IGRAPH command draws a box plot, which is also produced by the EXAMINE command with its /PLOT BOXPLOT subcommand.[6]

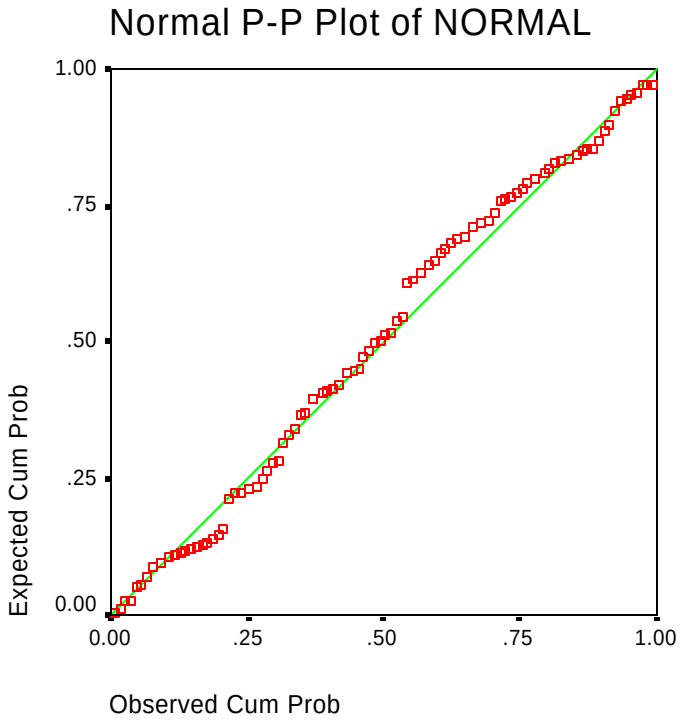
```
IGRAPH /VIEWNAME='Boxplot' /Y = VAR(normal) TYPE = SCALE
/COORDINATE = VERTICAL /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0 /CHARTLOOK='NONE'
/BOX OUTLIERS = ON EXTREME = ON MEDIAN = ON WHISKER = T.
```

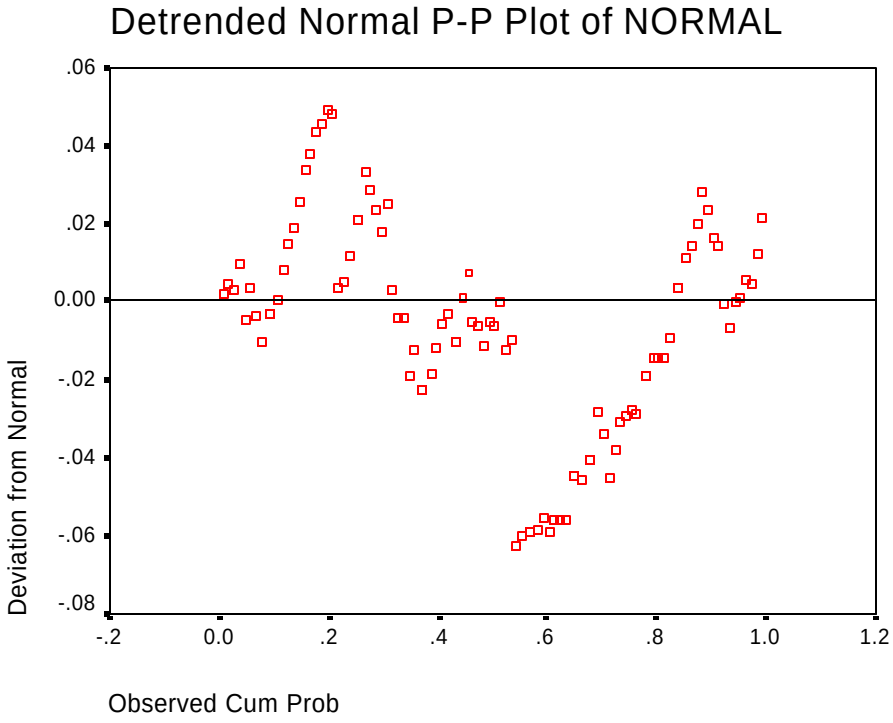


In order to get both P-P and Q-Q plots, we need to use the PPLOT command. The Q-Q plot of the PPLOT command is identical to that of the EXAMINE command. Consider the following PPLOT command to draw a P-P plot. Note that the /TYPE and /DIST respectively specify a P-P plot and the standard normal distribution.

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

MODEL:  MOD_1.
Distribution tested: Normal
Proportion estimation formula used: Tukey's
Rank assigned to ties: Mean

For variable NORMAL ...
Normal distribution parameters estimated: location = -7.07 and scale = 106.57764
```

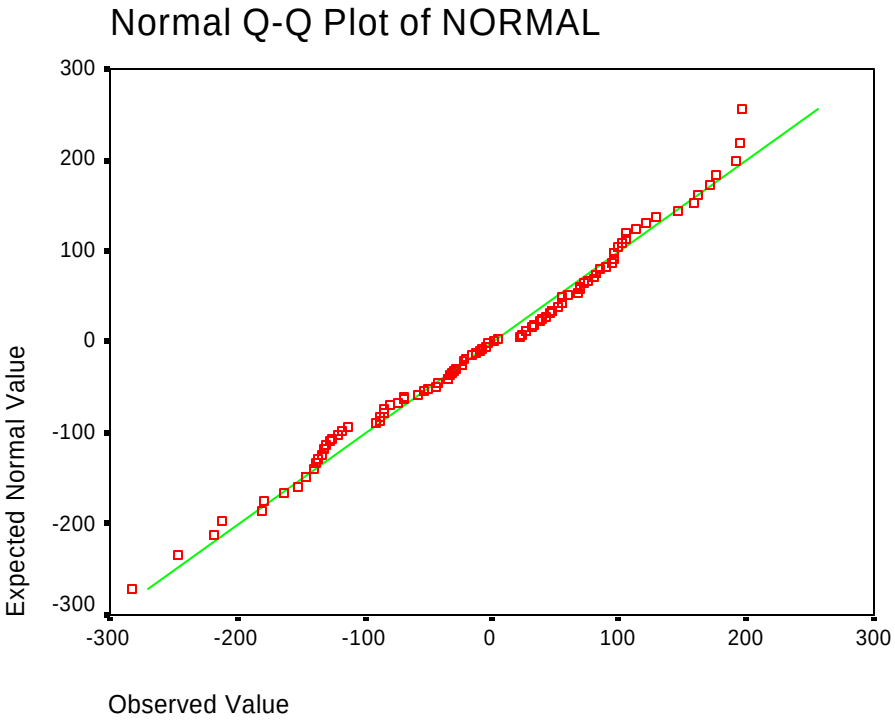
[6] To use the menu for plotting, click Graphs→Histogram, Q-Q, or P-P. For interactive graphs, sequentially click Graphs→Interactive→Histogram or Boxplot.

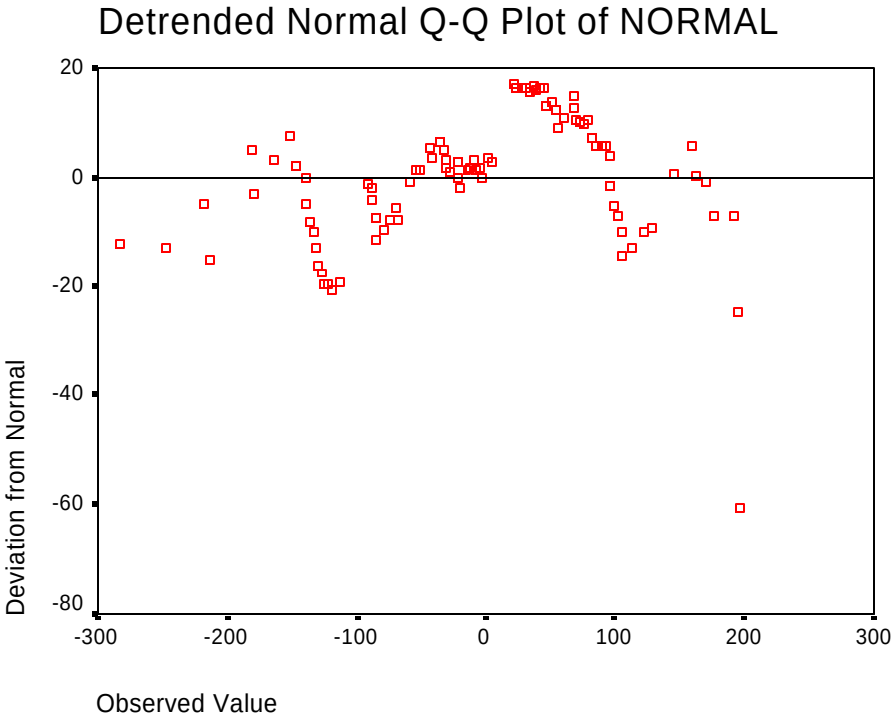




Following PPLOT command draws a Q-Q plot of the variable.

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





The P-P and Q-Q plots are identical to corresponding plots in SAS and STATA: no significant deviation from the line. Note that the Q-Q plot and detrended Q-Q plot has observed quantiles on the X axes and normal quantiles on the Y axes. In SAS, by contrast, X axes lists the normal quantiles; the position is switched.

Now, we move on to the numerical ways of testing normality of a variable. SPSS has the EXAMINE command to do the job. The EXAMINE command gives us the Kolmogorov-Smirnov and Shapiro-Wilk statistics.

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

The EXAMINE command produces descriptive statistics (/STATISTICS DESCRIPTIVES), a box plot (/PLOT BOXPLOT), a stem-and-leaf plot (/PLOT STEMLEAF), and a normal P-P plot (/PLOT NPPLOT). Then, it conducts a normality test. [7]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
NORMAL	100	100.0%	0	.0%	100	100.0%

Descriptives

			Statistic	Std. Error
NORMAL	Mean		-7.07	10.658
	95% Confidence Interval for Mean	Lower Bound	-28.22	
		Upper Bound	14.08	
	5% Trimmed Mean		-5.51	
	Median		-5.50	
	Variance		11358.793	
	Std. Deviation		106.578	
	Minimum		-283	
	Maximum		196	
	Range		479	
	Interquartile Range		161.25	
	Skewness		-.228	.241
	Kurtosis		-.504	.478

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
NORMAL	.069	100	.200(*)	.984	100	.279

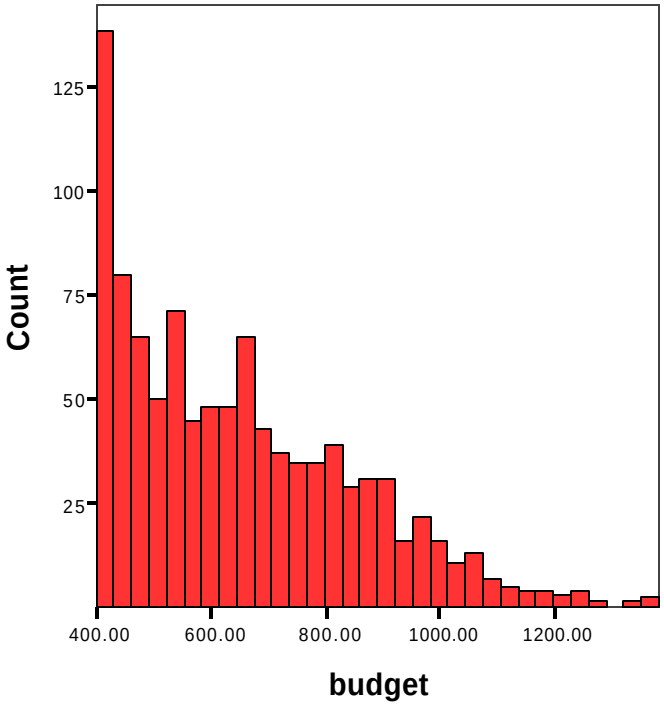
* This is a lower bound of the true significance.
a Lilliefors Significance Correction

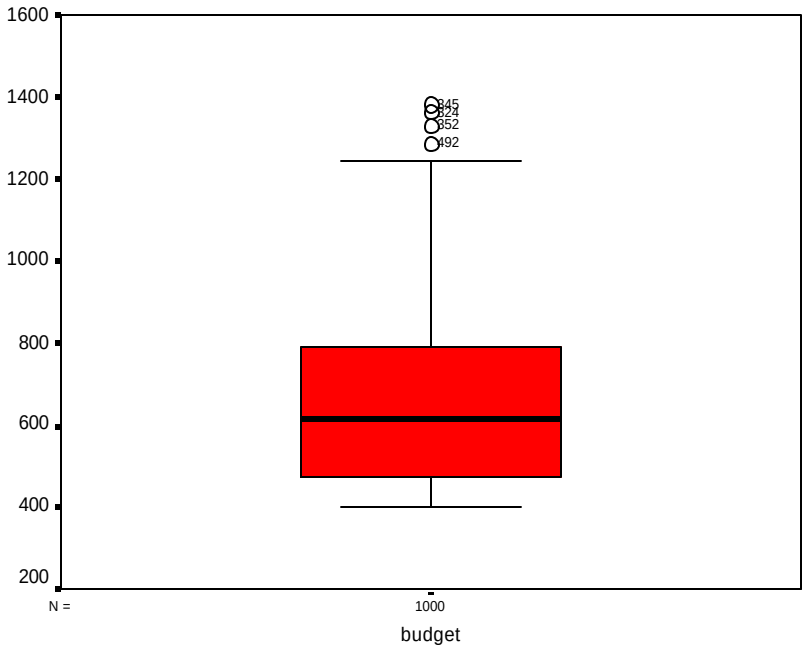
Note that SPSS reports the lower bound of significance .2, which is larger than the p value for K-S .15. Since N is less than 2,000, we have to use the Shapiro-Wilk statistic.

[7] To run the EXAMINE command using menu, sequentially click Analyze→Descriptive Statistics→Explore, then, include the variable you want to examine.

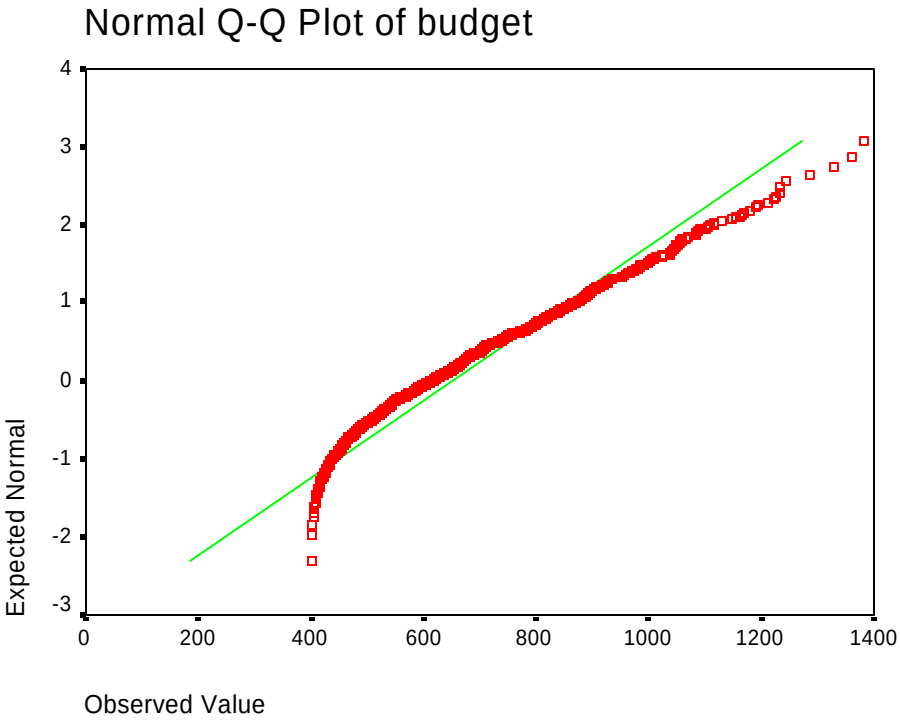
Let us consider a variable that is not normally distributed. Descriptive statistics and S-W test are skipped. Following histogram is the output of the IGRAPH command.

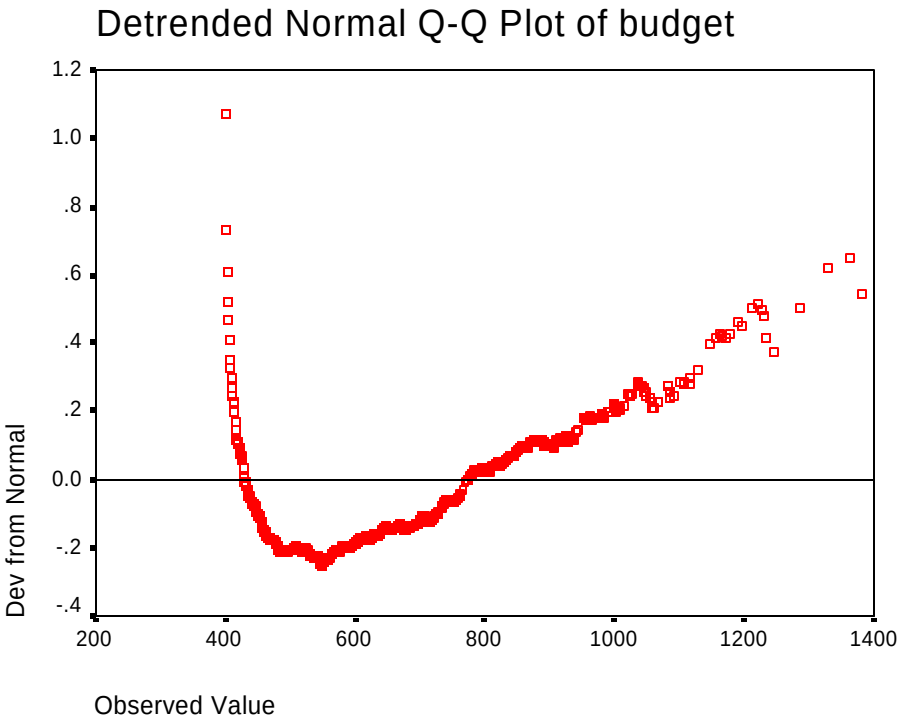
```
IGRAPH /VIEWNAME='Histogram' /X1 = VAR(budget) TYPE = SCALE /Y = $count
 /COORDINATE = VERTICAL  /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0
 /CHARTLOOK='NONE' /Histogram SHAPE = HISTOGRAM CURVE = OFF X1INTERVAL AUTO X1START = 0.
```





The above box plot, below Q-Q and detrended Q-Q plots are the output of the EXAMINE command with the same options as the previous one. The box plot shows that the distribution is heavily skewed to the top. The Q-Q plot illustrates that data points in the two extremes are significantly deviated from the straight line. We can observe an obvious pattern in the detrended Q-Q plot.





References

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

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.

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

SAS Institute. 1999. *SAS Procedures Gudide, Version 8*. 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. 2003. *STATA Reference Manual Release 8*. College Station, TX: STATA Press.