Do Normal Probability Plots Tell the Truth When Sample Size is Small?

Ananda Jayawardhana

Pittsburg State University

Pittsburg, KS

Abstract:

Most of the modern introductory statistics books discuss the use of normal probability plots for informal assessment of the normality of a set of data. Textbook authors describe that if the population which the data came from is normal, then the pattern of the points should be reasonably close to a straight line and the pattern should not have a symmetric pattern that is not a linear pattern. Recognizing the pattern is a subjective decision. It is well-known that normal probability plots should be interpreted loosely for small sample sizes but usually strictly for large sample sizes. How large is large? Peck and Devore (2012) discuss using the correlation test to check for normality. In this paper, author will discuss an assignment he has given to the elementary statistics students and the lessons learned. Also the author will discuss the results of a simulation study using correlation test for normality for small sample sizes when the population is not normal.

Key words: Normal Probability Plots

1. Introduction:

Probability plots are used as the first and informal check for testing whether a set of data belongs to a particular distribution. Plotting the cumulative distribution function against ordered data produces an S shaped increasing curve for continuous distributions. By transforming the vertical scale appropriately, one can transform the S shaped curve to a straight line. Probability papers for different distributions have already done the vertical transformation. Plotting the cumulative distribution function against ordered data on a probability paper should provide a plot which is close to a straight line pattern if the data came from the same distribution the probability paper was made for. Free probability papers are available on the internet and (http://www.weibull.com/GPaper/) is one such Universal Resource Locator to find probability papers. Ryan and Joiner (1976) states that "One problem confronting persons inexperienced with probability plots is that considerable practice is necessary before one can learn to judge them with any degree of confidence." When the sample sizes are large, it is not difficult to make a conclusion but when the sample sizes are small, there is a high possibility of making both Type I and Type II errors, that is to reject a set of data from a normal distribution as non-normal and not to reject a set of data from a non-normal distribution as non-normal, respectively. Peck and Devore (2012) state that "A strong linear pattern in a normal probability plot suggests that population normality is plausible. On the other hand, systematic departures from a straight-line pattern (such as curvature in the plot) indicate that it is not reasonable to assume that the population distribution is normal."

2. An Example

Tabor, J. (2003) provides a data set on teacher's salary in thousands of dollars as 39.9, 47.6, 49.3, 51.6, 47, 46.2, 48.4, 51.7, 58.1, 56.1, and 63.7.

Definition of the Empirical CDF is given by

$$F(x)=\begin{cases}0 & \text{if } x < x_{(1)} \\ \dfrac{i}{n} & \text{if } x_{(i)} \le x < x_{(i+1)} \\ 1 & \text{if } x \ge x_{(n)}\end{cases}$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are ordered data. Figure 1 and Figure 2 represent normal probability plots for Tabor's (2003) data created using Minitab and R respectively.

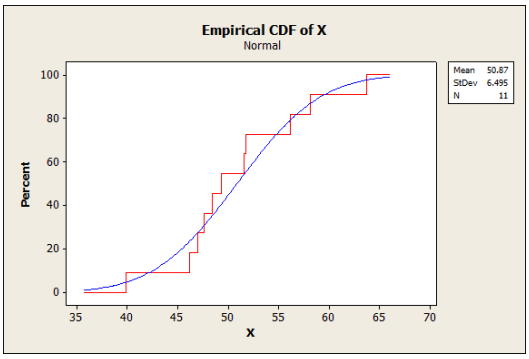

Figure 1

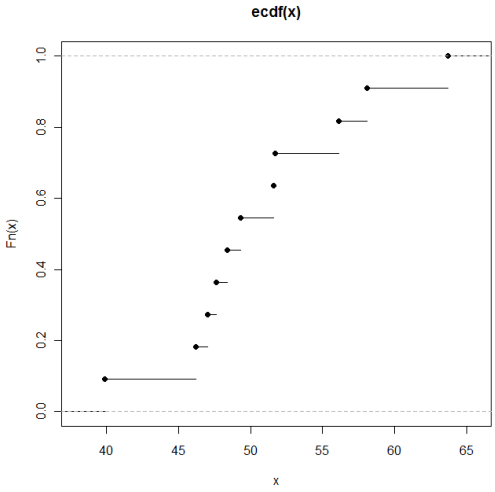

Figure 2

2.1 Modified Empirical CDF

Blom (1958) proposed a class of estimates of the empirical cumulative distribution function as $p_i = \frac{i-\alpha}{n-2\alpha+1}$; where i is the ordered rank of the data point and $0 \le \alpha < 1$. For example $\alpha = 0.375$ provides $p_i = \frac{i-0.375}{n+0.25}$.

The choice of quantity $\alpha = 0.375$ in the definition provides the following properties:

a. $p_n < 1$

b. $p_{\left(\frac{n+1}{2}\right)} = \frac{\left(\frac{n+1}{2}\right)-0.375}{n+0.25} = 0.5$

c. $1-p_i = 1-\frac{i-0.375}{n+0.25} = \frac{\left(n-\left(i-1\right)\right)-0.375}{n+0.25} = p_{(n-(i-1))}$

2.1.1 CDF of a Standard Normal Distribution

Cumulative distribution function of a standard normal distribution is defined as $\Phi(y) = \int_{-\infty}^{y} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy$.

Let $p = \Phi(y)$. Plotting the points $\left(X_{(i)}, \Phi^{-1}(p_i)\right)$ will create a normal probability plot. One can use the standard normal table to find $Y_i = \Phi^{-1}(p_i)$ but Shapiro et al. (1958) provide the approximation $\Phi^{-1}(p_i) \approx 4.91\left[p_i^{0.14} - (1-p_i)^{0.14}\right]$, which is much easier to calculate

Ordered data, ordered ranks, percentile points, and inverted percentile points for Tabor (2003) data are given in Table 1. One can plot the ordered data against the corresponding $Y_i = \Phi^{-1}(p_i)$ on a regular graph paper or ordered data against the p_i on a normal probability paper to create a normal probability plot.

Table 1

Ordered Data	39.9	46.2	47	47.6	48.4	49.3	51.6	51.7	56.1	58.1	63.7
Ordered Rank	1	2	3	4	5	6	7	8	9	10	11
p_i	0.055	0.144	0.233	0.322	0.411	0.500	588	0.678	0.767	0.856	0.945
$\Phi^{-1}(p_i)$	-1.59	-1.06	-0.73	-0.46	-0.22	0	0.22	0.46	0.73	1.06	1.59

2.2 Using Minitab to Create Probability Plots

For introductory statistics courses, it will be sufficient to show students how to use a statistical package to create a normal probability plot and how to interpret it. Minitab has several other distributions to choose from if one wants to check whether data came from a different distribution. Minitab plots ordered data on the X -axis and 100 p_i on the Y -axis. Normal probability plot created using Minitab for Tabor (2003) data is given in Figure 3.

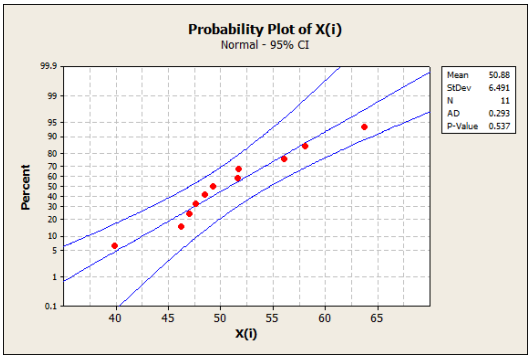

Figure 3

Minitab also produces the results of Anderson-Darling test for normality. The null hypothesis is that the data came from a normal distribution. For Tabor (2003) data, a p-value equal to 0.537 is reported and such a high p-value supports that the null hypothesis that data came from a normal distribution.

2.3 Using R to Generate Normal Probability Plots

R is getting popular among faculty and students due to its strengths as well as its free availability. R plots $\Upsilon_i = \Phi^{-1}(p_i)$ on the X -axis and ordered data on the Y -axis. Normal probability plot created using R for Tabor (2003) data is given in Figure 4.

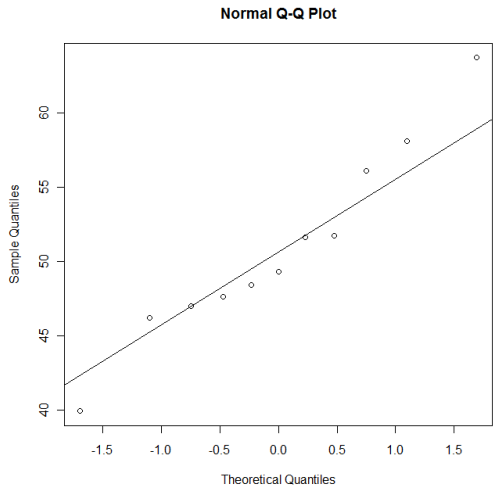

Figure 4

Shapiro-Wilk normality test

W = 0.9616, p-value = 0.7918

R can perform different tests for normality and a p-value of 0.7918 for the Shapiro-Wilk test also supports that data came from a normal distribution.

2.4 Using Excel to Create Probability Plots

Excel does not have a direct way to create normal probability plots but one can create a normal probability plot using the Excel NORMINV function. One can either enter the data in increasing order or order the data using Rank and Percentile option under data analysis. Once a column of ordered data and a column of corresponding ranks are available, one can create another column of normal percentile points of the empirical distribution using the equation NORMINV((B1-0.375)/(n+0.25, 0, 1). For example NORMINV((1-0.375)/11.25, 0, 1)= -1.59322. Using the Chart Wizard and plotting data verses corresponding percentiles, one can create a normal probability plot. Using a shift of data one can get a better plot. Normal probability plot created using Tabor (2003) data is given in Figure 5.

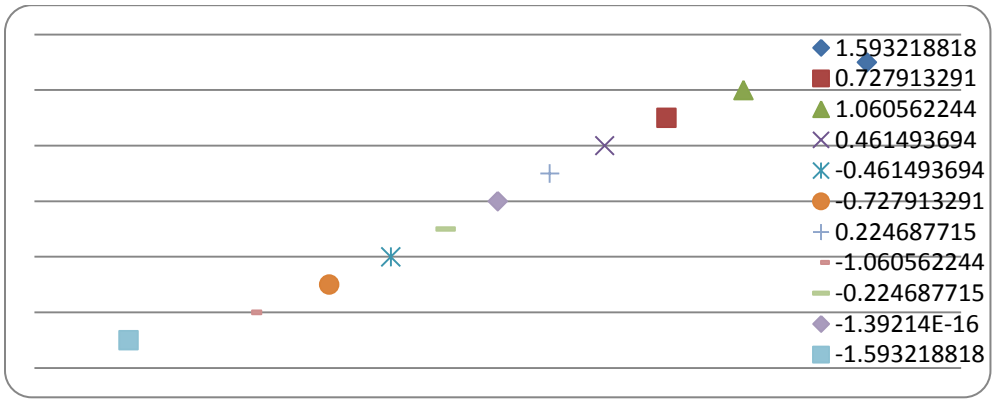

Figure 5

3. Class Assignments

One method to demonstrate the reliability of the normal probability plots is to generate a lot of normal probability plots using random data from normal distributions and other distributions. It is easy to generate hundreds of columns of random data using Minitab. Also Minitab allows input of many columns to create separate normal probability plots. Students can observe the plots and delete them one at a time. Generating random data, creating probability plots, and observing the plots for 100 data sets could be done in 5 to 10 minutes. In R one can use the up arrow key to call a previous statement. Repeating this will create a new random data set and its normal probability plot every time.

Data from a normal distribution could create a normal probability plot which does not have a linear pattern. Figure 6 represents a normal probability plot of 10 random numbers from a normal

distribution. This plot was selected from probability plots of 100 random normal data sets of size 10. This particular plot suggests that data came from a non-normal distribution.

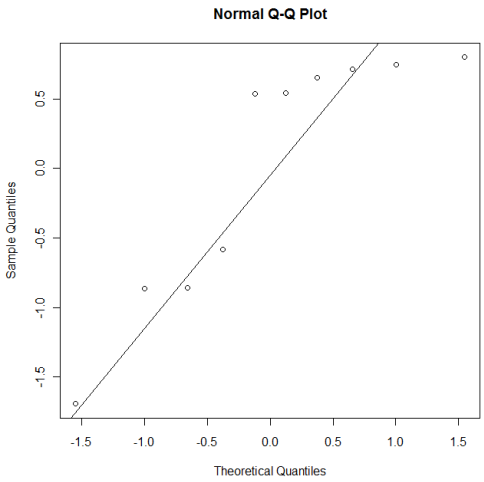

Figure 6

Shapiro-Wilk normality test

W = 0.8101, p-value = 0.01923

R command Shapiro.text created a p-value of 0.01923, which supports that data did not come from a normal distribution.

Student Assignment 1

Generate 100 samples of size 10 from a normal distribution and create normal probability plots. What percentage of these normal probability plots look like that data came from a normal distribution? Repeat the experiment with sample sizes 20 and 30.

This exercise demonstrates that even data from a normal distribution can produce a normal probability plot which suggests that data did not come from a normal distribution.

Student Assignment 2

Generate 100 samples of size 10 from an exponential distribution and create normal probability plots. What percentage of these normal probability plots look like that data came from a normal distribution? Repeat the experiment with sample sizes 20 and 30.

Plots in Figures 7 and 8 were selected among normal probability plots of 100 random data sets of size 10 from an exponential distribution. Figure 7 represents a case where the normal probability plot agrees that data did not come from a normal distribution while Figure 8 represents a case where the normal probability plot has a linear pattern indicating that the data came from a normal distribution though that data came from an exponential distribution.

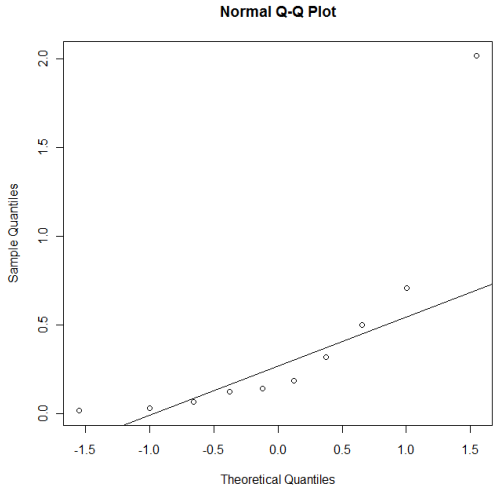

Figure 7

Shapiro-Wilk normality test

W = 0.6712, p-value = 0.0003953

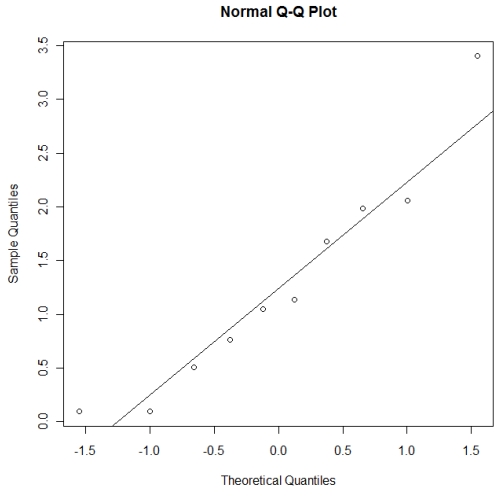

Figure 8

Shapiro-Wilk normality test

W = 0.9312, p-value = 0.4599

Figures 9 is a normal probability plot of a set of random data of size 30 from an exponential distribution. The normal probability plot displays a symmetric non-linear pattern suggesting that the data came from a non-normal distribution.

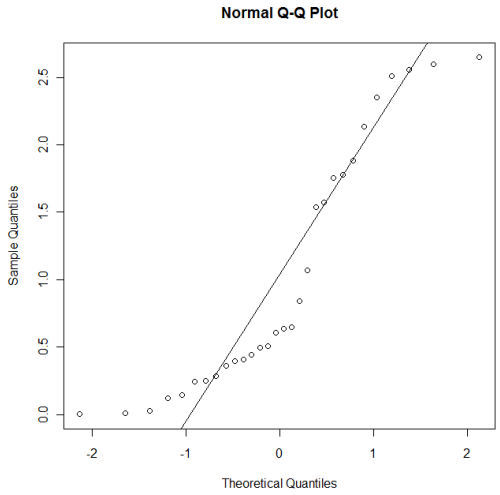

Figure 9

Shapiro-Wilk normality test

W = 0.8607, p-value = 0.001045

This exercise demonstrates to the students that when sample sizes are small, even if the data came from a non-normal distribution, there is a high probability that the graph will not suggest the data came from a non-normal distribution. But, as sample size increases this probability gets smaller. If the sample size is bigger than a certain number n one can make a conclusion with certain probability that the data came from a normal distribution. This number n changes with the distribution and the shape of the distribution.

4. Ryan and Joiner Test

Ryan and Joiner (1976) proposed a correlation like test statistic to test for normality. At elementary level, students understand the concept of correlation. Instructors usually do not talk about the requirement of the two variables having a bivariate normal distribution at that level. Actually one of the variables in this case is not even random. This method calculates a correlation like coefficient. If the correlation between the ordered data and the corresponding percentage points from a standard normal distribution is very high, one can conclude that the data came from a normal distribution.

Ryan and Joiner defined a correlation like coefficient between $X_{(i)}$ and $Y_i = \Phi^{-1}(p_i)$ as follows:

$$R_p = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} \text{ and } \sum y = \sum \Phi^{-1}(p\ ) = 0.$$

If the plot between $X_{(i)}$ and $\Phi^{-1}(p_i)$ fall nearly on a straight line, then the R_p will be near 1.

4.1 Rejection Rule for Ryan and Joiner Test (1976)

With α level of significance, reject the hypothesis that the data came from a normal distribution, if $R_P < CV(n)$, where

$$CV(n) = 1.0071 - \frac{0.1331}{\sqrt{n}} - \frac{0.3682}{n} + \frac{0.7780}{n^2} \text{ for } \alpha = 0.10,$$

$$CV(n) = 1.0063 - \frac{0.1288}{\sqrt{n}} - \frac{0.6118}{n} + \frac{1.3505}{n^2} \text{ for } \alpha = 0.05,$$

and

$$CV(n) = 0.9963 - \frac{0.0211}{\sqrt{n}} - \frac{1.4106}{n} + \frac{3.0791}{n^2} \text{ for } \alpha = 0.01.$$

Example

For Tabor, J. (2003) data: $R_p = 0.975$,

$$CV(11) = 1.0071 - \frac{0.1331}{\sqrt{11}} - \frac{0.3682}{11} + \frac{0.7780}{11^2} = 0.9399 \text{ for } \alpha = 0.10,$$

$$CV(11) = 1.0063 - \frac{0.1288}{\sqrt{11}} - \frac{0.6118}{11} + \frac{1.3505}{11^2} = 0.9230 \text{ for } \alpha = 0.05,$$

$$CV(11) = 0.9963 - \frac{0.0211}{\sqrt{11}} - \frac{1.4106}{11} + \frac{3.0791}{11^2} = 0.8880 \text{ for } \alpha = 0.01,$$

and $R_p > CV(11)$ for α= 0.01, 0.05, 0.10.

Since $R_p > CV(11)$ for α= 0.01, 0.05, 0.10 one can't reject the hypothesis that data came from a normal distribution even at 10% level of significance. Ryan-Joiner test is not commonly available in most elementary level statistics texts but Peck and Devore (2012) discusses this method.

5. Simulations Study

Since it is not practical to generate normal probability plots and decide what proportion would be accepted as plots of data from a normal distribution, a simulation study was conducted using the Ryan-Joiner test. Ryan and Joiner (1976) states that "The notion of using familiar correlation coefficient as a means of judging the straightness of a normal probability plot is intuitively appealing. This test has the virtues of being simple, easily remembered, and powerful. It encourages the use and comparison of a visual test (the probability plot) with an objective measure R_p." The assumption is that higher linearity of the points in a normal probability plot will generate a higher value for R_p for the same data. For sample sizes 5(5)50 and 100 random data were generated using one of the distributions from Exponential(1), Weibull(shape=2, scale=1), Weibull(shape=0.5, scale=1), Uniform(0,1), t-distribution with degrees of freedom of 3, Cauchy(location=0, scale=1), Scale Contaminated Normal (0.95N(0,1)+0.05N(0,5)), Location Contaminated Normal (.95N(0,1)+0.05N(2,1)), and lognormal. For each random data set R_p was calculated and checked whether it was smaller than the critical values $CV(n)$ for the specified level of significance of 0.05 or 0.10. The empirical power is reported in Tables 2 and 3.

6. Conclusion

Ryan-Joiner test has higher power against skewed distributions like Exponential, Weibull(Shape=0.5, Scale=1), and lognormal. Power of the test is low against symmetric bell shaped distributions such as Weibull (Shape=2, Scale=1) and t-distribution. This test has higher power against distribution with thicker tails such as Cauchy distribution and less power for distributions without tails such as uniform distribution. Though power against contaminated normal distributions is low, testing against location contaminated normal data had relatively higher power than that for scale contaminated data.

Unless a distribution is skewed and or has thicker tails Ryan-Joiner test is not reliable for sample sizes less than 20. Operating assumption of this paper is that degree of straightness of the normal probability plot is directly proportional to the value of the R_p statistic. The conclusion of this study is that when a distribution is skewed and or has thicker tails, normal probability plot are fairly reliable for sample sizes more than 20. For some distributions like Weibull (shape=2, Scale=1) and contaminated normal the power of the test is relatively very low.

Table 2

Empirical Power of the Correlations Test for Selected Alternative Distribution when $\alpha = 0.05$											
Distribution	Power										
	Sample size										
	5	10	15	20	25	30	35	40	45	50	100
$Exponential(1)$	0.18	0.42	0.64	0.79	0.89	0.95	0.98	0.98	0.99	1.00	1.00
$Weibull(Shape = 2, Scale = 1)$	0.06	0.07	0.09	0.13	0.16	0.19	0.25	0.27	0.29	0.36	0.75
$Weibull(Shape = .5, Scale = 1)$	0.46	0.88	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$Uniform(0,1)$	0.05	0.05	0.05	0.08	0.11	0.15	0.22	0.29	0.39	0.45	0.99
$t-distribution\ (d.f. = 3)$	0.09	0.21	0.32	0.38	0.44	0.50	0.56	0.60	0.67	0.69	0.90
$Cauchy(Loc. = 0, Scale = 1)$	0.28	0.60	0.80	0.90	0.93	0.97	0.98	0.99	1.00	1.00	1.00
$0.95N(0,1) + 0.05N(0,5)$	0.10	0.23	0.30	0.37	0.45	0.49	0.54	0.61	0.63	0.66	0.88
$0.95N(0,1) + 0.05N(2,1)$	0.06	0.07	0.08	0.08	0.09	0.11	0.11	0.13	0.12	0.12	0.21
$Lognormal$	0.24	0.57	0.80	0.91	0.97	0.99	1.00	1.00	1.00	1.00	1.00

Table 3

Empirical Power of the Correlations Test for Selected Alternative Distribution when $\alpha = 0.10$											
Distribution	Power										
	Sample size										
	5	10	15	20	25	30	35	40	45	50	100
$Exponential(1)$	0.26	0.53	0.77	0.86	0.94	0.99	0.97	0.99	1.00	1.00	1.00
$Weibull(Shape = 2, Scale = 1)$	0.12	0.17	0.19	0.19	0.26	0.32	0.37	0.39	0.47	0.46	0.83
$Weibull(Shape = .5, Scale = 1)$	0.58	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$Uniform(0,1)$	0.12	0.13	0.13	0.20	0.23	0.33	0.42	0.47	0.59	0.67	0.99
$t-distribution\ (d.f. = 3)$	0.16	0.26	0.39	0.46	0.50	0.58	0.66	0.65	0.73	0.75	0.93
$Cauchy(Loc. = 0, Scale = 1)$	0.38	0.69	0.84	0.93	0.97	0.98	0.99	0.99	1.00	1.00	1.00
$0.95N(0,1) + 0.05N(0,5)$	0.17	0.27	0.38	0.44	0.49	0.55	0.60	0.64	0.69	0.72	0.91
$0.95N(0,1) + 0.05N(2,1)$	0.11	0.12	0.13	0.14	0.16	0.17	0.18	0.19	0.20	0.20	0.31
$Lognormal$	0.35	0.69	0.88	0.96	0.99	0.99	1.00	1.00	1.00	1.00	1.00

7. References

1. Blom, G. (1958) Statistical Estimates and Transformed Beta Variables, John Wiley and Sons, New York.
2. Peck, R. and Devore, J. L. (2012) Statistics the Explorations & Analysis of Data, 7th edition, Brooks/Cole, Carnage Learning, Boston.
3. Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968) "A Comparative Study of Various Tests for Normality." JASA, No. 63, pp 1343-1372.
4. Tabor, J. (2003) Understanding Regression Output, STATS, 36, pp 24-27.

Appendix

R and Minitab Commands and Codes

The following R-codes were used to produce Figure 2 for the same data:

```
x<-c(39.9, 47.6,49.3, 51.6, 47, 46.2, 48.4, 51.7, 58.1, 56.1, 63.7)

plot.ecdf(x)
```

The following Minitab commands were used to generate Figure 3.

Using Minitab, one can follow the commands, Graph> Probability Plot> Single> OK> Choose the variable> OK. The default distribution in Minitab is the normal distribution.

R Commands to generate a normal probability plot (Figure 4) and to do the Shapiro-Wilk test for normality are as follows:

```
x<-c(39.9, 47.6,49.3, 51.6, 47, 46.2, 48.4, 51.7, 58.1, 56.1, 63.7)

qqnorm(x); qqline(x)

shapiro.test(x)
```

Using SAS to Create Probability Plots

Procedure capability produces a nice normal probability plot. If one does not specify the "noprint" option, capability procedure prints a lot of information beyond the plot. For the name, the name of the SAS data set and for the variable, name of the variable should be used.

```
proc capability data=Name noprint;
  probplot variable;
```

Figure 6 was created by the following R commands.

```
x<-rnorm(10,0,1)

qqnorm(x);qqline(x)

shapiro.test(x)
```

Figures 7 and 8 and corresponding Shapiro-Wilk tests were created by the following R commands.

```
x<-rexp(10,1)

qqnorm(x);qqline(x)

shapiro.test(x)
```

R codes for simulation study:

```
#For Alpha=0.10

size = 5;

output = c()

while(size <= 100)

{

  result=0; index=0;

 while(index<1000)

 {

x=rexp(size,1) # Exponential random data


 x_sort = sort(x)

  i = 0; y = c();

 while(i<size)

  {

  i=i+1;

  y[i] = (i-0.375)/(size+.25);

  }

y = qnorm(y)

  if(cor(x_sort,y) > (1.0071 - 0.1371/sqrt(size) - 0.3682/size + 0.7780/size^2)) {result=result+1}

  index = index+1;

 }

  output[size] = 1-(result/1000)

  size = size + 1

}
```

Output