1 Normal Probability Plots

Abstract

A graphical procedure is presented to determine if a set of measurement values follows the normal distribution. Hand prepared plots on special normal probability paper are described and the Minitab commands for generating normal plots are given. Several examples are presented and discussed.

Introduction

The normal distribution is one of the most important and commonly occurring probability distributions. It is used to characterize measurement data and the sampling distribution of many statistics. It is also used to approximate many other probability distributions under certain conditions. It is one of the primary tools of statistical and quality control work.

A technique is required to determine when data follow the normal distribution. There are many quantitative tests for normality but the graphical method presented here, called a normal probability plot, can provide extra insight into the behavior of data. The normal plot as a test for normality is much more powerful than using a histogram. Normal plots are also very useful for analyzing nonnormal data.

Where the Technique is Used

Normal probability plots are used to:

- Test data for normality before setting specifications.
- Test data for normality before performing t, F, χ^2, or any other hypothesis test that requires normal data.
- Test data for normality before calculating process capability statistics like c_p and c_{pk}.
- Check data for skewness or kurtosis.
- Detect the presence of a tail or an outlier in a data set.
- Estimate the fraction defective of a process (normal or nonnormal data).
- Test model residuals for normality when performing regression or ANOVA.
- Identify the significant regression coefficients in a large multiple regression model.

Data

Data must be variables data (measurement data) or attribute data for which a normal distribution model is appropriate. The sample should contain 20 and preferably more values although the technique is still useful for smaller data sets.

Assumptions

- The data consist of a random sample from a single population.
- The measurement scale is accurate and precise.

Hypotheses Tested

H_0 : The data come from a single normally distributed population.
H_A : The data do not come from a single normally distributed population.

Procedure

1. Collect a random sample of n parts from the population of interest.
2. Measure the parts with the appropriate accuracy and precision.
3. Order the measurements from smallest to largest.
4. Determine the midband percentile p_i for each ordered data point from:

$$p_i = \left(\frac{i - \frac{1}{2}}{n} \right) 100\% \qquad ((1))$$

   where $i = 1$ for the smallest value, $i = 2$ for the next value, ..., and $i = n$ for the largest value.
5. On normal probability graph paper plot the midband percentiles on the probability axis vs. the measurement values on the linear axis. A blank copy of normal probability graph paper for photocopying is shown in Figure 1.
6. Draw a straight line through the data. Fit the line closely to the bulk of the data and pay less attention to any outliers near the ends.
7. Determine the degree of fit between the plotted data points and the line by observation. If the agreement is good, accept H_0 and conclude that the population is normal. If there is evidence of significant curvature in the form of a hook up or down, or an S-shaped curve, or a break or jump in the pattern of points, accept H_A and conclude that the population is not normal.

Example: *Check the following data set for normality using a normal probability plot: 425, 410, 370, 421, 391, 370, 413, 416, 399, 432, 397, 415, 356, 404, 382, 382, 392, 397, 428, 416, 401, 420, 408, 397, 378.*

Solution: *There are $n = 25$ values in the data set. The data ordered from smallest to largest are: 356, 370, 370, 378, 382, 382, 391, 392, 397, 397, 397, 399, 401, 404, 408, 410, 413, 415, 416, 416, 420, 421, 425, 428, 432. The corresponding midband percentiles are:*

$$p_1 = \frac{\left(1 - \frac{1}{2}\right)}{25} 100\% = 2\%$$

$$p_2 = \frac{\left(2 - \frac{1}{2}\right)}{25} 100\% = 6\%$$

$$p_3 = \frac{\left(3 - \frac{1}{2}\right)}{25} 100\% = 10\%$$

$$\vdots$$

$$p_{25} = \frac{\left(25 - \frac{1}{2}\right)}{25} 100\% = 98\%$$

The ordered data values, their midband percentiles, and the normal probability plot are shown in Figure 2.

Discussion

The Midband Percentiles

The method used to determine midband percentiles is not necessarily obvious. Suppose that a sample data set contains $n = 10$ data points. Then each point in the sample represents 10% of the population being sampled. If the points are ordered from smallest to largest then the smallest value represents the smallest 10% of the values in the population. Since these values span the 0 to $10th$ percentile of the population it makes sense to assign the smallest value to the $5th$ percentile, at the middle of the range of percentiles that it represents. The second smallest value represents the $10th$ to $20th$ percentile of the population so its plotting percentile should be 15%. The other midband percentiles are determined the same way. The largest data point represents the $90th$ to $100th$ percentile of the population so it plots at 95%.

The midband percentiles are the easiest percentiles to use but they are not the only ones or the most accurate. The probability plotting positions given by:

$$p_i = \left(\frac{i - 3/8}{n + 1/4}\right) 100\% \qquad ((2))$$

are more accurate than the midband percentiles and are the most common ones implemented in software. For large samples the difference between the two methods is minimal but you might notice small differences, especially in the tails of the distribution, for smaller sample sizes. If you're hand plotting points and don't have a table of the plotting positions given by Equation 2 definitely use those from Equation 1.

Normal Probability Plots with Minitab

Normal plots are used so frequently that all statistical software packages and most spreadsheet programs can make them. To construct a normal plot from Minitab the data must be loaded into a single column. It's not necessary to sort the data or calculate the midband percentiles. Minitab does all of that for you.

Suppose that sample data has been loaded into column c2 of a Minitab worksheet. Construct the normal plot of the data by typing one of the following commands at the Minitab prompt:

mtb> %normplot c2

mtb> %qqplot c2

Mouse users can perform the same operation by selecting **Stat> Basic Stats> Normality Test** or **Graph> Probability Plot** from the pull down menus and selecting c2 in the Variables window.

Example: *Use Minitab to construct the normal plot of the data from the example problem.*
Solution: *The data were entered into column c1 of the Minitab worksheet. The %normplot command was used to generate the normal plot of the data in Figure 3.*

Deviations from Normality

Step 7 in the Procedure described some of the patterns that indicate that data are not normally distributed. If you look closely enough at any data set you will find systematic deviations of the data from the line that indicates normality. These are to be expected so don't read too much into such patterns. A good technique to use is to complete the normal plot and look away from it. Then give yourself a quick glance, one full second is about right, and look away again. If no significant patterns strike you in that one second then conclude that the data are normal or at least approximately normal.

Normal plots of some data sets which significantly deviate from normality are shown in Figure 4. Hooks, S-shaped curves, and jumps or breaks indicate that the data are not normal. Even when data are not normal important conclusions can be drawn from their normal plot.

Example: *Evaluate the data in Figure 5 for normality. Use the normal plot to estimate the fraction of the population that falls outside the specification of 0.240/0.220.*
Solution: *There is a tremendous amount of curvature in the normal plot so the data are very probably not from a normal population. The specification limits are drawn on the Figure and from where they intersect the data it appears that less than 1% of the population falls below the lower spec limit of 0.220 and about 12% of the population falls above the upper spec limit of 0.240. Although there are more units falling out of spec on the high end than on the low end of the specification a small shift in the population mean to a lower value would be disasterous.*

Detecting Outliers

Normal plots are a powerful tool for identifying outliers in data sets. When a data set is well behaved, even if it is not normal, the data will create a nice smooth pattern on the normal plot. Since outliers, by definition, are either very much smaller or very much larger than the rest of the data they will fall well away from that pattern. Be careful about calling a point an outlier though. Even if a point does fall well away from the rest of the data look for other points near it that are following the same pattern. If several points taken together create their own pattern which ends in a potential outlier, the outlier is probably not a true outlier. The collection of points is indicating that another effect or pattern is present in the data.

Do not just throw a suspected outlier out of a data set. It is necessary to determine the cause that makes the suspected outlier different from the rest of the data. If the cause that distinguishes the outlier from the rest of the data can be found, then the outlier can be removed.

References

Freund and Simon, *Modern Elementary Statistics*, 9th Ed., Prentice-Hall, 1997.
Freund and Walpole, *Mathematical Statistics*, 3rd Ed., Prentice-Hall, 1980.

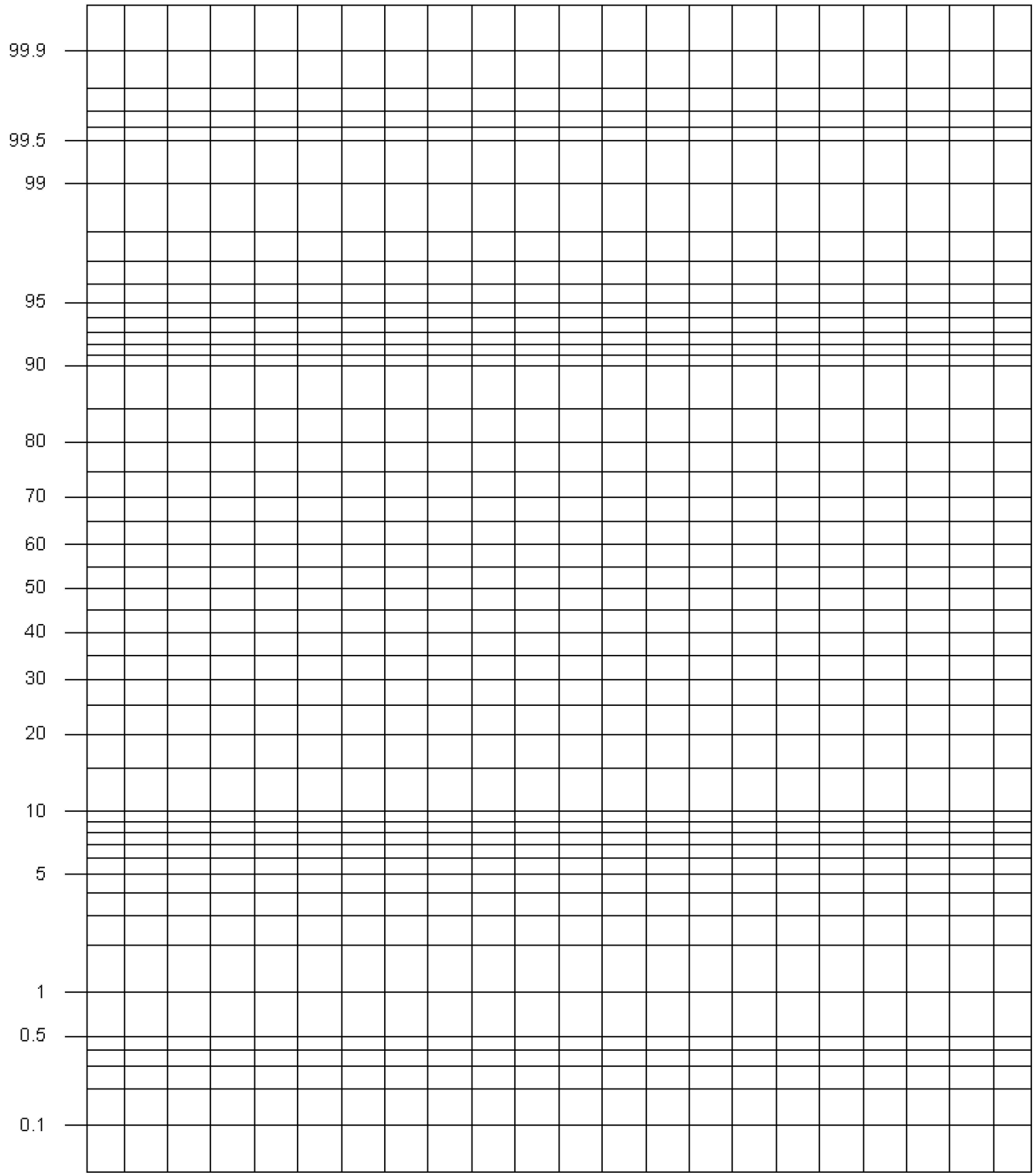

Figure 1: Normal Probability Paper

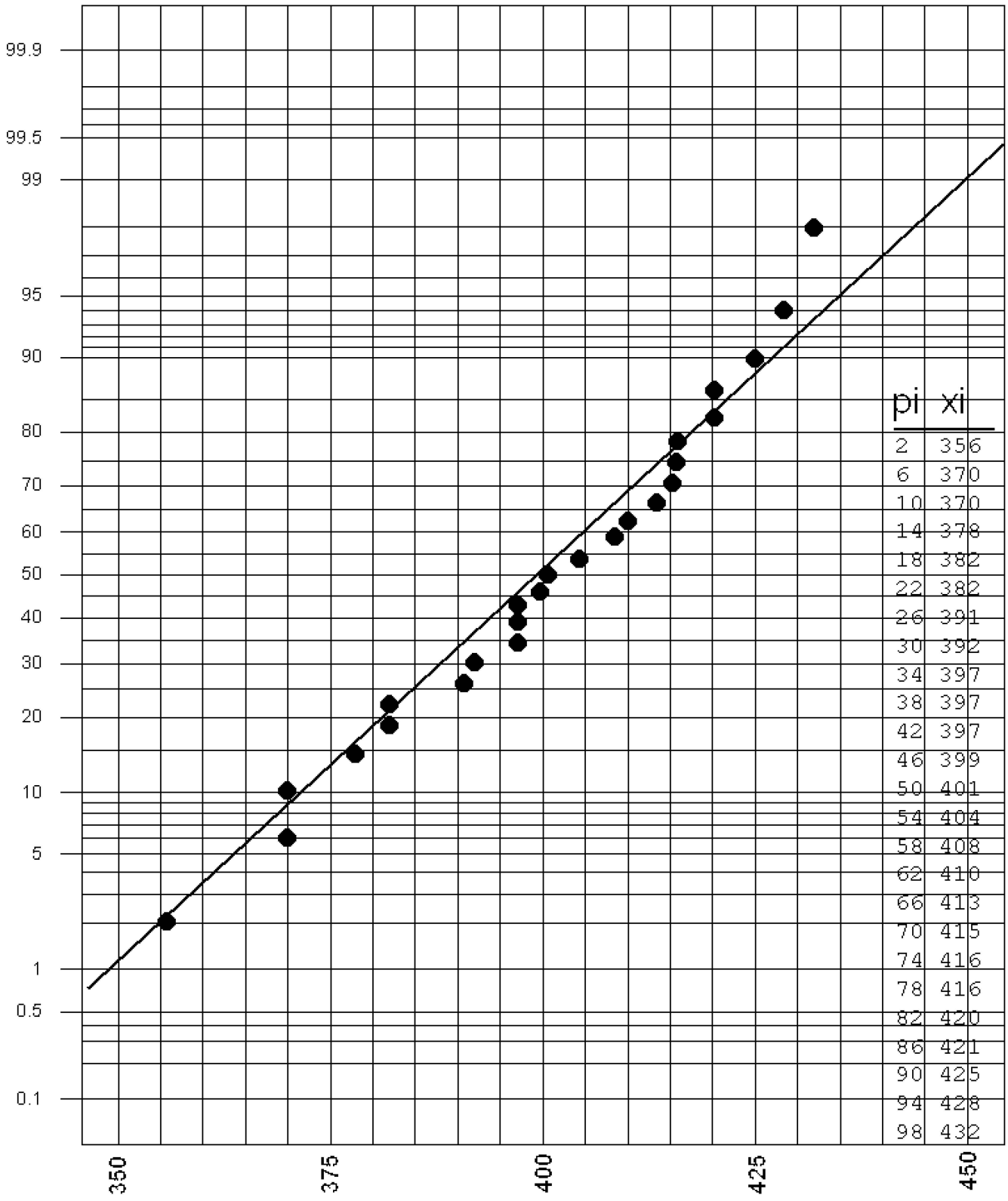

Figure 2: Manually Created Normal Plot of Example Data

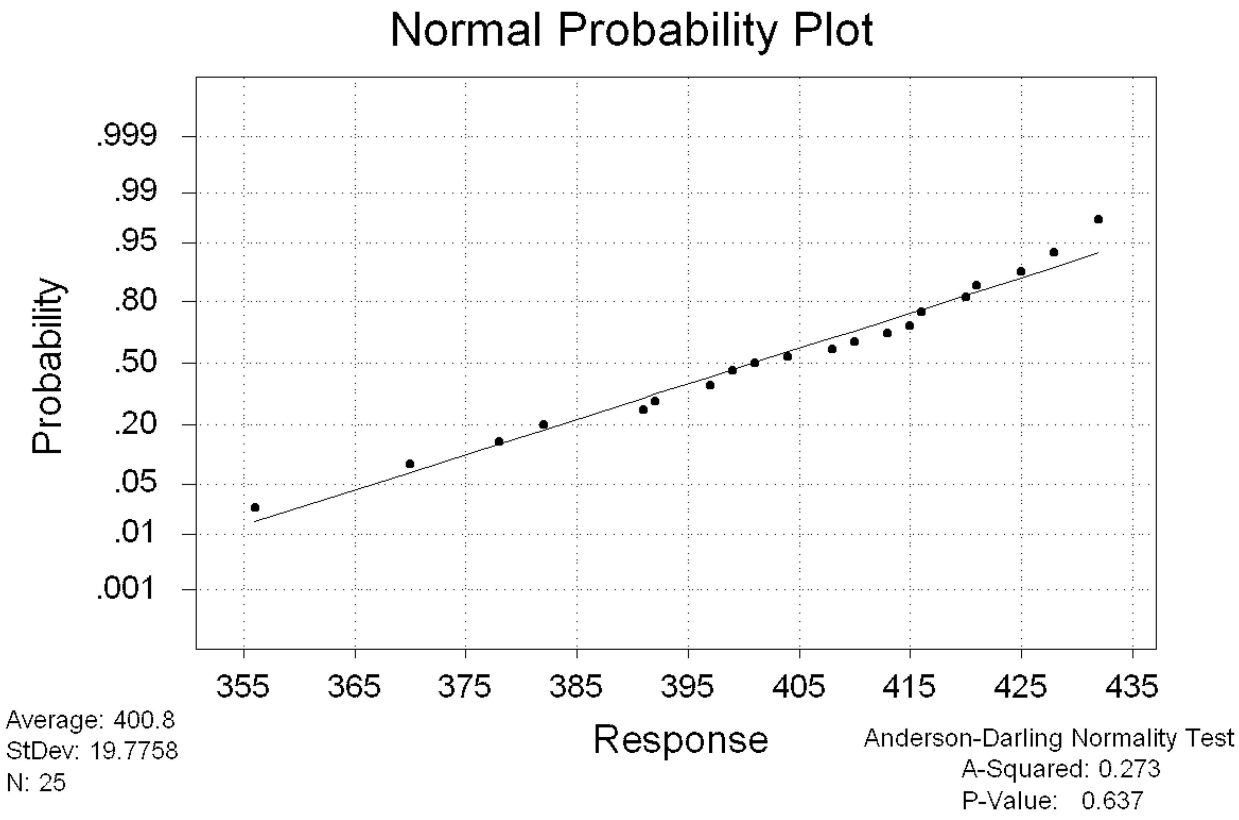

Figure 3: Normal Plot Example Using Minitab

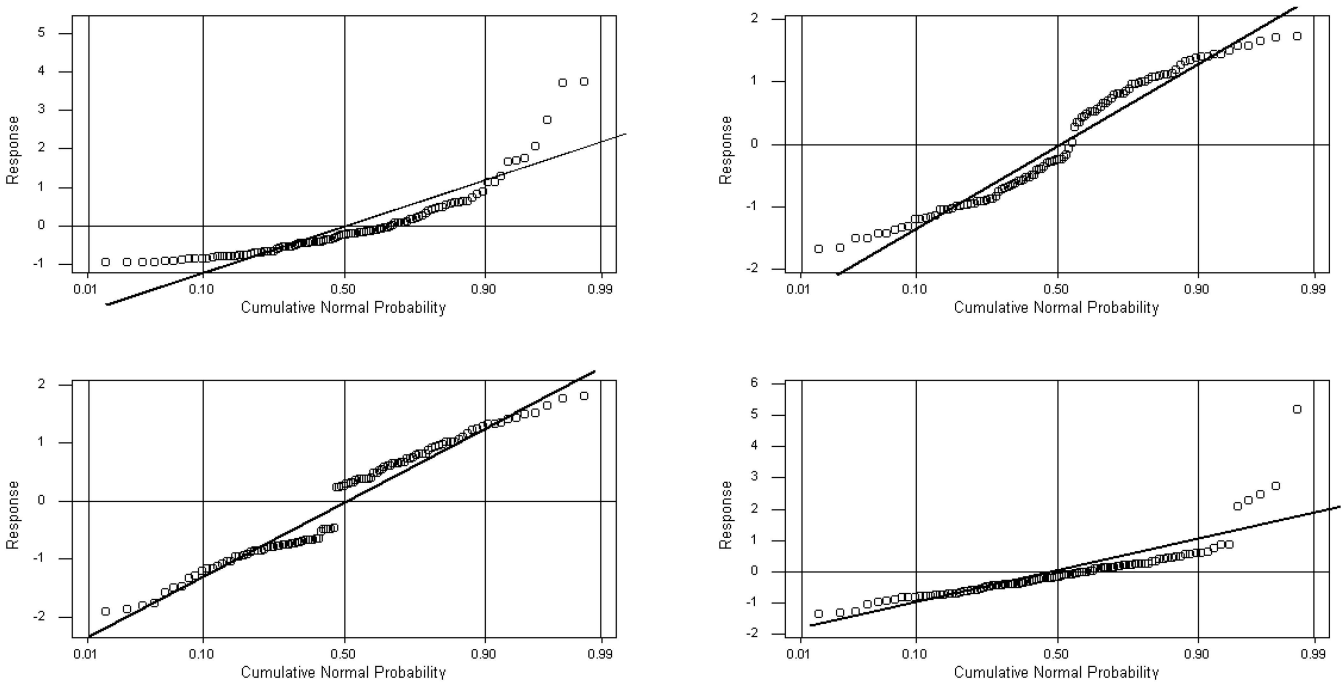

Figure 4: Normal Plots of Some Non-normal Data

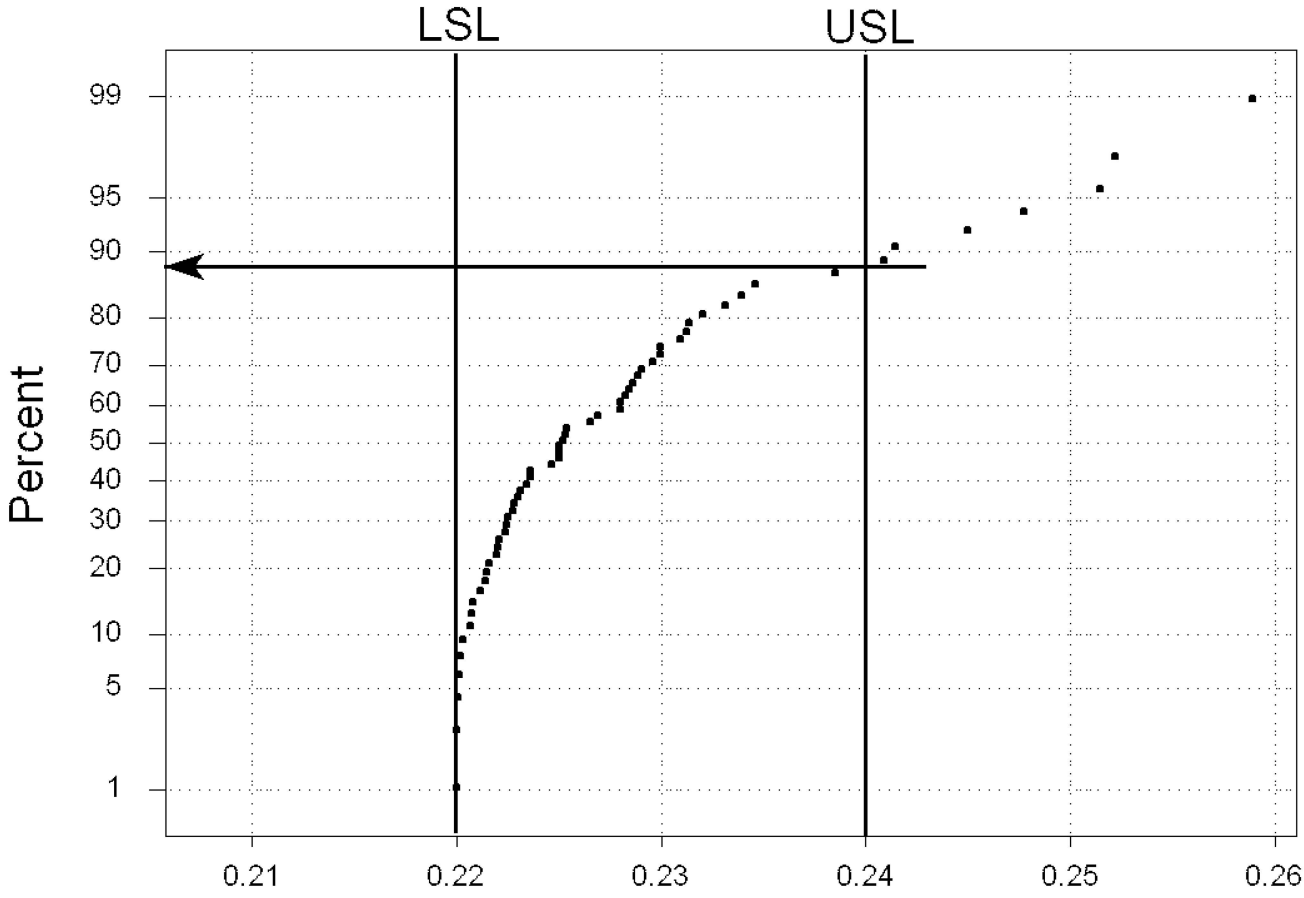

Figure 5: Using a Normal Plot to Analyze Non-normal Data