PROBABILITY PLOTTING IN SAS

Daniel M. Chilko, West Virginia University
Gerry Hobbs, West Virginia University
E. James Harner, West Virginia University

Introduction

Bar charts or histograms are the simplest and most frequently used graphical representation of data. They reveal many properties of the data - the range of data values, the number of modes, whether the distribution is symmetric or skewed, the existence of outliers. Although bar charts reveal the general shape of the distribution, it is sometimes difficult to determine whether or not the data can be viewed as a sample from some hypothesized distribution. Probability plots are a graphical representation of data that focus on the distributional aspects of data. Bar charts are easy to produce in SAS using PROC CHART and PROC GCHART. Probability plots are also easy to produce using SAS.

Probability plots

A random variable, x, is characterized by its distribution function

$$F(x) = Pr(X <= x)$$

For example, the graph of F(x) for a random variable with a standard normal distribution is shown in Figure 1. This graph can be turned into a straight line by transforming the x axis to F(x) (both axes would be probabilities) or by transforming the F(x) axis to x (both axes would be quantities).

A sample distribution function can be produced in SAS using PROC RANK with the PERCENT option and PROC PLOT. See Figure 2. For data from symmetric distributions, this function is characteristically S-shaped and its point of inflection makes it difficult to work with. If the probability axis is transformed, the plot is now a probability plot. That is, a probability plot scales the probability axis of a sample distribution function according to some probability distribution such that, if we chose the correct distribution, the resulting plot is more or less a straight line.

Normal Distribution Function

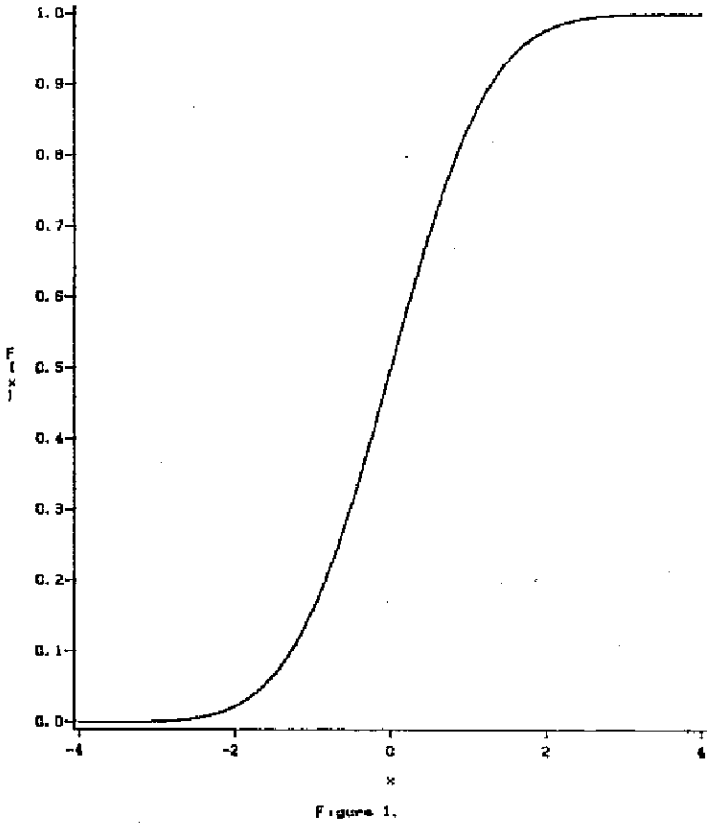

Figure 1.

Sample Cumulative Distribution Function

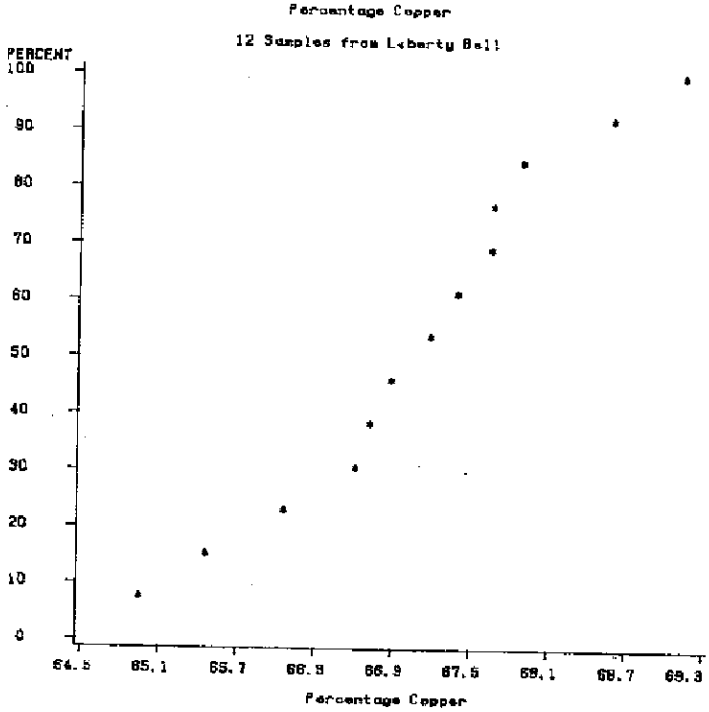

Figure 2.

If $x_1, x_2, \ldots, x_n$ are the ordered observations from a sample of size n, the scaling of the axis is achieved by finding a set of values, $y_1, y_2, \ldots, y_n$ (say) such that

$$F(y_i) = p_i$$

where p_i's denote appropriate chosen fractions of the distribution corresponding to the x_i's. Plotting the pairs $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ should result in a straight line if the x_i's are a sample from a distribution having a distribution function F(x). Since it seems reasonable to consider the data as dependent upon the distribution function, probability plots are usually constructed with the variable of interest as the vertical axis and the probability distribution scalei values as the horizontal axis.

Kimball (1963) investigated the question of how to chose the values of p , for a given size n, for use in probability plots. Some commonly used values are

I $\quad p_i = i/(n+1)$

II $\quad p_i = (i-.5)/n$

III $\quad p_i = (i-.375)/(n+.25)$

For illustration, we chose I.

The problem of scaling the probability axis now becomes the problem of finding the inverse of the distribution function. That is,

$$y_i = F^{-1}(i/(n+1))$$

Furthermore the scaling is independent of any scale and location parameters, so that the scaling reduces to finding the inverse of the distribution function of a "standardized" random variable. For example, if x has a probability distribution with location parameter A and a scale parameter B, a plot of the points $(x_1,z_1), (x_2,z_2), \ldots, (x_n,z_n)$, where $z_i = (y_i-A)/B$, would still result in a straight line.

Normal probability plots

Many statistical procedures assume the data to have a normal distribution. It is true that random variables having normal or near normal distributions occur quite often in nature, perhaps because the normal distribution is the limiting distribution of a random variable which represents the sum of a series of independent and identically distributed random variables.

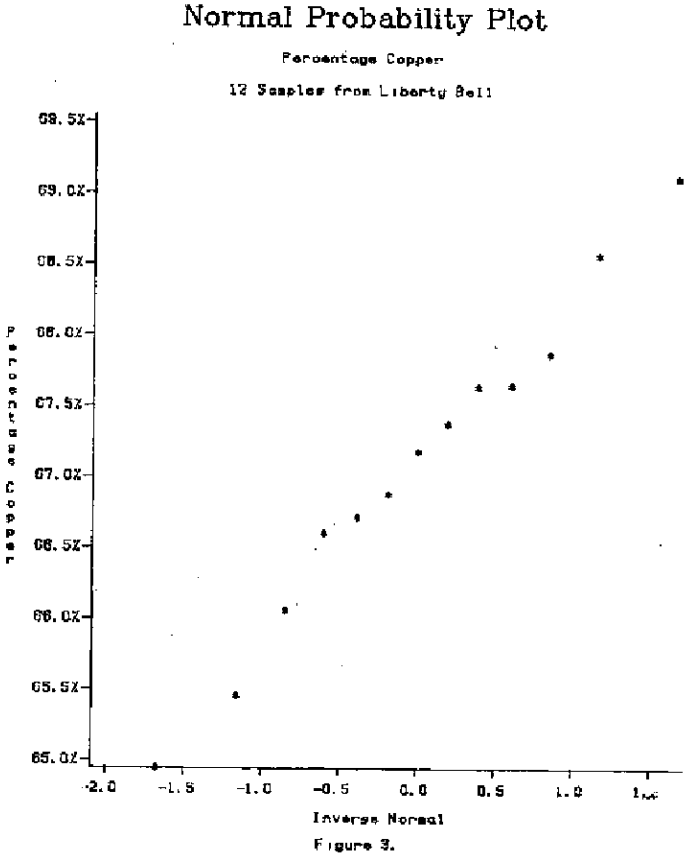

Figure 3.

Normal probability plots provide an informal test of the hypothesis that a sample comes from a normal distribution and can be produced in SAS by using PROC RANK with the NORMAL option and PROC PLOT. Figure 3 is a normal probability plot for the percentage of copper in 12 samples from the Liberty Bell.

Since construction of probability plots is independent of the estimation of scale and location parameters, one potential use is the estimation of these parameters from the plot itself (Ferrell, 1958). Interest may instead focus on the estimation of a particular percentile of the distribution. Estimation of parameters from a probability plot requires fitting a straight line to the plot. Harner et. al. (1981) described the estimation of the 99th percentile of a distribution using various regression techniques to fit the straight line using SAS.

Interpretation of non-linear plots

Quite often when data is plotted on a particular probability plot, the plot does not appear too straight. Abbot (1960) and King (1965) have investigated non-linear plots. Their studies show that many such plots have a simple and straightforward explanation.

Figure 4 shows a good fit in the middle of the plot but that the plot tends to flatten out at each end. A scarcity of values at the high end usually indicates a inspecition and selection process that removes unacceptable values. A scarcity of values at the low end may indicate selection to a minimum specification or measuring equipment which may not have resolution below some particular value.

A plot characterized by two fairly straight portions connected by a S-shaped connection indicates a bimodal distribution. See Figure 5. The detection of two sources for the data when only one is expected can be an important benefit.

A convex plot usually indicates a left-skewed distribution. A concave plot indicates a right-skewed distribution. See Figure 6. A log-normal probability plot is a good next step for this pattern. See Figure 7.

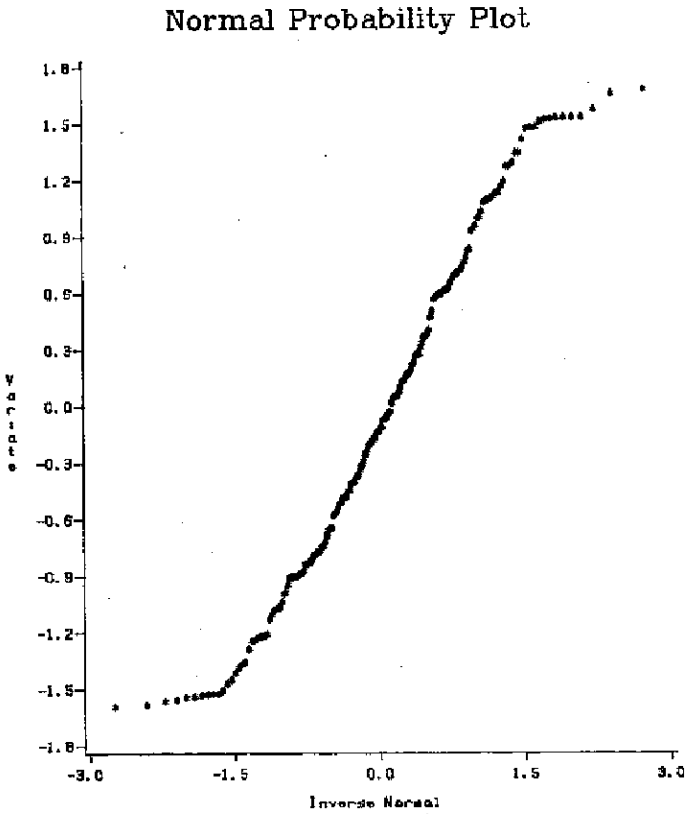

Figure 4.

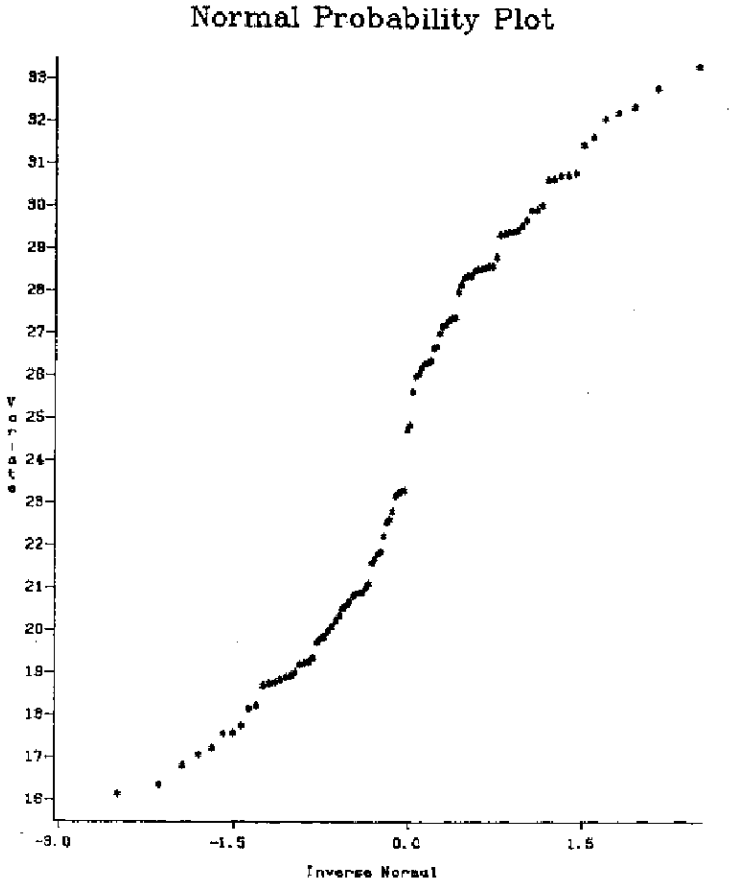

Figure 5.

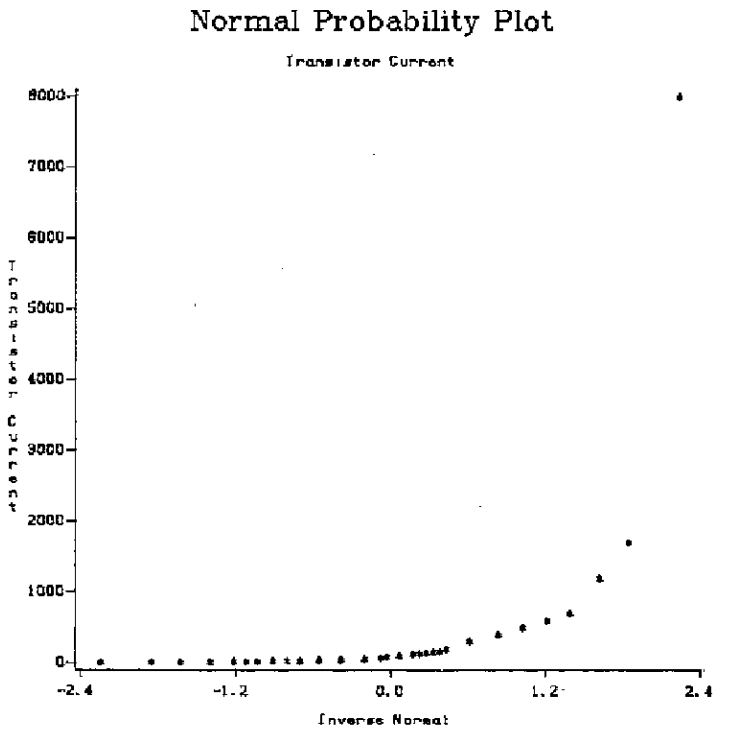

Figure 6.

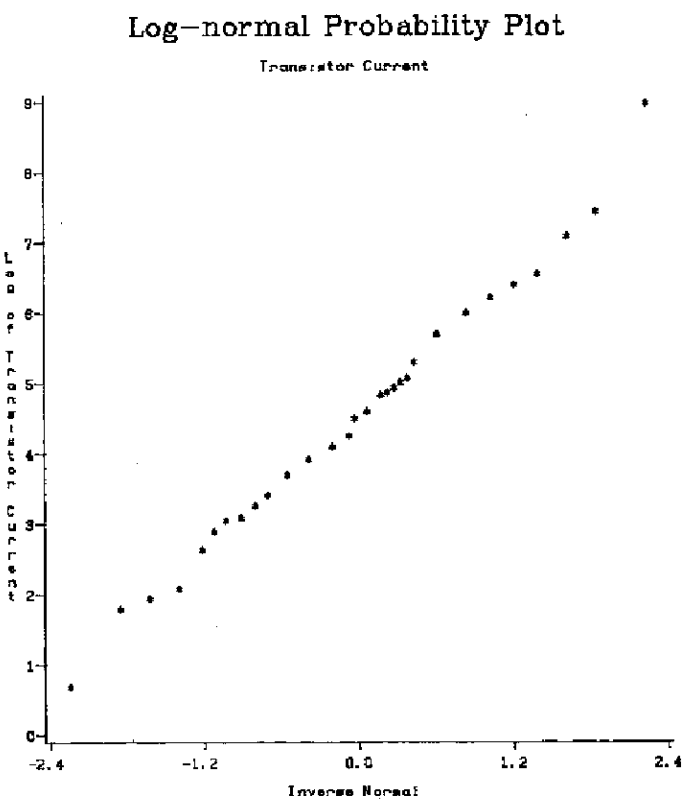

Figure 7.

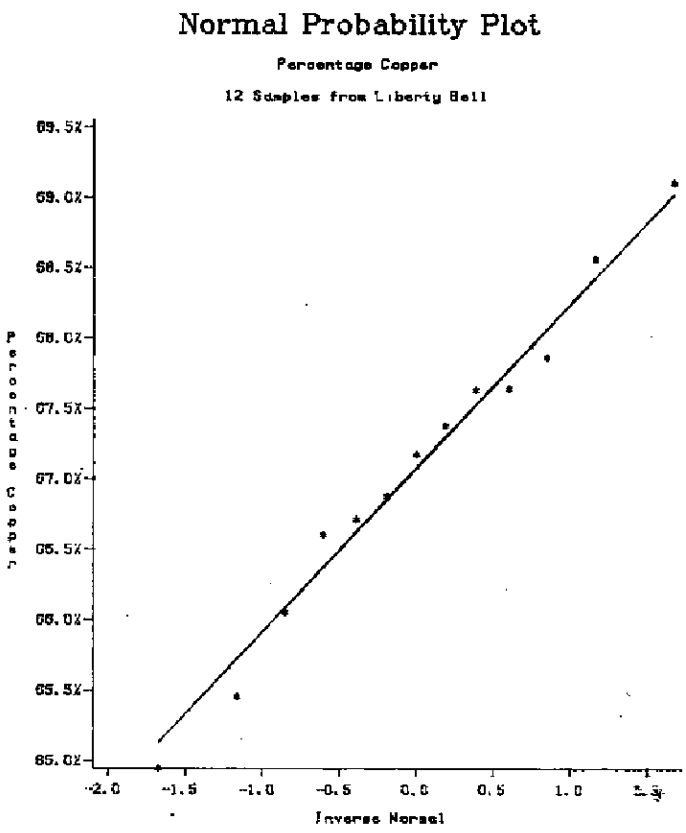

Figure 8.

Reference lines

An additional aid to the interpretation of a normal probability plot is a reference line which corresponds to a normal distribution with a specified mean and variances. PROC MEANS can be used to produce a data set containing the usual moment estimates. A short DATA step that processes this data set makes it easy to add reference lines to probability plots in SAS. See Figure 8.

The sample mean and variance are not robust and estimates of scale and location parameters based on order statistics are more useful when outliers are present in the data. Hillyer (1978) investigated the use of moment and quantile estimators in a context similar to probability plotting. The present authors duplicated his results using SAS. PROC SORT, for example, produces order statistics.

Half-normal probability plots

When a random variable has a normal distribution with mean zero, the absolute value of this random variable is said to have a half-normal distribution. In the linear regression framework

$$Y = XB + E$$

E is usually assumed to be a vector of independently and identically distributed random variables, each normally distributed with mean zero and constant variance.

In a regression analysis, the differences between observed and predicted values are called residuals. That is,

$$r = Y - \hat{Y}$$

where $\hat{Y} = Xb$ and b are the least squares estimates. If the underlying model assumptions are true, then the r's have normal distributions, each with mean zero. They do not, in general, have the same variance nor are they independently distrbuted.

Half-normal probability plots provide an informal test of the normality of the residuals in a regression analysis. Half-normal plots show more sensitivity to kurtosis at the expense of not revealing skewness. A detailed discussion of producing half-normal probability plots in SAS was given by Sall (1978). A useful exposition on the interpretation of half-normal plots was given by Daniel and Wood (1971).

Gamma probability plots

While the normal distribution is of singular importance in statistics, the gamma distribution is also encountered frequently. The general gamma distribution depends on a location, scale, and shape parameter. The general gamma distribution can be transformed to a standardized distribution with only a shape parameter. The chi-square and exponential distributions are special cases of the gamma distribution. A chi-square random variable with degrees of freedom d is a gamma random variable with shape parameter equal to d/2; an exponential random variable is a gamma distribution with shape parameter equal to 1. Wilk et. al. (1962) described the construction and interpretation of gamma probability plots. The SAS function GAMINV can be used to produce gamma probability plots.

Exponential probability plot

The exponential distribution is often used to characterize failure or waiting time distributions. Figure 9 is an exponential probability plot produced by SAS for the waiting times between major train wrecks in the U. S. during the period from 1900 to 1960.

Chi-square probability plot

Sample variances or mean squares from a normal population have a chi-square distribution. A chi-square probability plot can be used to provide an informal test of the homogeniety of sample variances. Figure 10 is a chi-square probability plot for the sample variances of the amount of nitrogen in 5 red clover plants innoculated with 6 different bacteria strains.

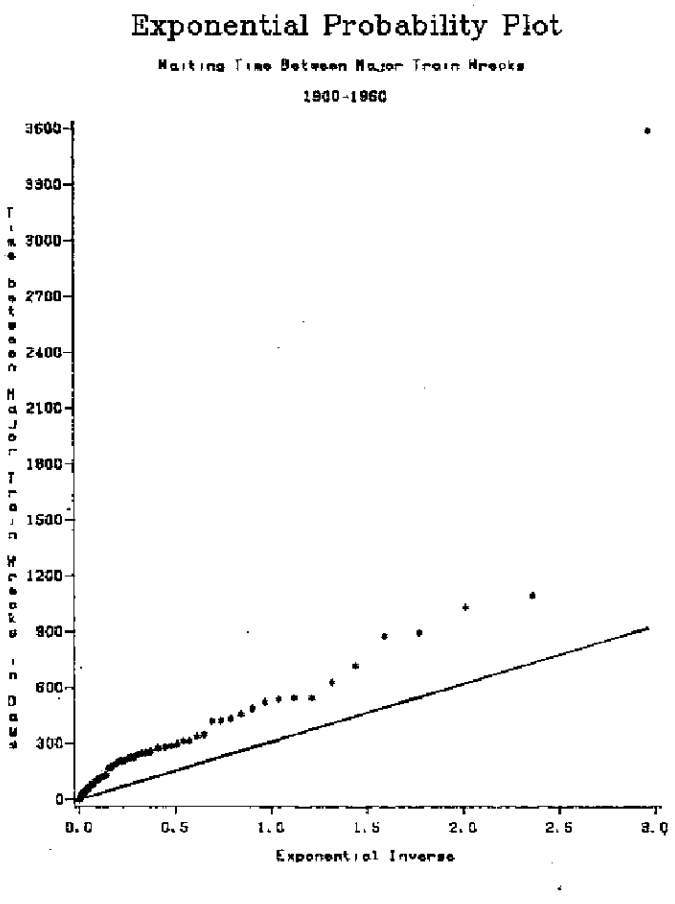

Figure 9.

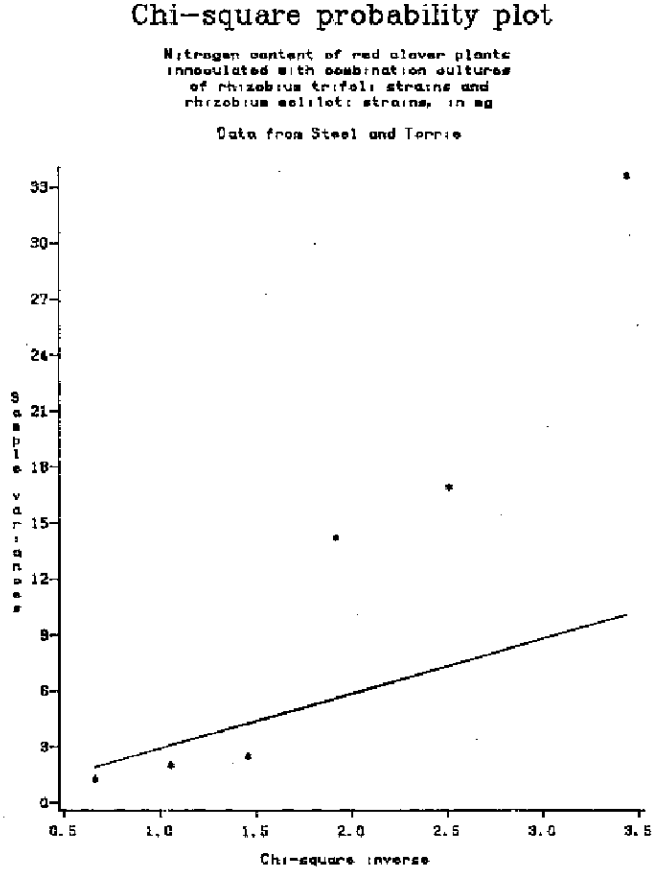

Figure 10.

References

Abbot, W.H. (1960), Probability Charts, Private Publication, St. Petersburg, FA.

Daniel, C. and F. Wood (1971), Fitting Equations to Data, John Wiley and Sons, New York, NY.

Ferrell, E.B. (1958), Plotting Experimental Data on Normal or Log-normal Probability Paper, Industrial Quality Control, 15, pp. 12-15.

Hanson, V.F., J.H. Carlson, K.M. Papauchado, and N.A. Nielson, (1976), The Liberty Bell: Composition of the Famous Failure, American Scientist, 64, pp. 614-619.

Harner, E.J., G.R. Hobbs, E.C. Keller Jr., A.G. Everett, and D.M. Chilko (1981), Assessing Estimates of the 99th Percentile of a Distribution, Environmetrics Procedings, (to appear).

Hillyer, M.J. (1978), Evaluation of the Effect of Distributional Assumptions on Statistical Forms of the Photochemical Oxidant Standard, Systems Applications, Inc., San Rafael, CA.

Kimball, B.F. (1960), On the Choice of Plotting Positions on Probability Paper, Journal of American Statistical Association, 55, pp. 546-560.

King, J.R. (1965), Graphical Data Analysis with Probability Papers, Technical and Engineering Aids for Management, Lowell, MA.

Sall, J.P. (1978), SAS Regression Applications, SAS Technical Report A-102, SAS Institute, Inc., Cary, NC.

Wilk, M.B., R. Gnanadeskan, and M.J. Huyet (1962), Probability Plotting for the Gamma Distribution, Technometrics, 4, pp. 1-20.