SESUG Paper 172-2019

Probability Plots for Exploratory Data Analysis

Dennis J. Beal, Leidos, Oak Ridge, Tennessee

ABSTRACT

Probability plots are used in statistical analysis to check distributional assumptions, visually check for potential outliers, and see the range, median, and variability of a data set. Probability plots are an important statistical tool to use for exploratory data analysis. This paper shows SAS® code that generates normal and lognormal probability plots using the Output Delivery System (ODS) on a real environmental data set using `PROC UNIVARIATE` and interprets the results. This paper is for beginning or intermediate SAS users of Base SAS® and SAS/GRAPH®.

Key words: normal probability plots, lognormal probability plots, `PROC UNIVARIATE`

INTRODUCTION

When the data analyst receives a data set, the first step to understanding the data is to conduct an exploratory data analysis. This first step might include a `PROC CONTENTS` to list the variables that are in the data set and their data types (numeric, character, date, time, etc.). The next step might be to run `PROC FREQ` to see the values of all the variables in the data set. Another common practice is to run `PROC UNIVARIATE` to see summary statistics for all numeric variables. These steps help guide the data analyst to see whether the data needs cleaning or standardizing before conducting a statistical analysis. In addition to summary statistics, `PROC UNIVARIATE` also generates histograms and normal probability plots for numeric variables. In SAS v. 9.3 and 9.4 `PROC UNIVARIATE` generates generic histograms and normal probability plots for the numeric variables with the ODS.

In addition, `PROC GPLOT` also produces normal and lognormal probability plots. Producing graphs in `PROC GPLOT` gives the data analyst the most control of the details of the graph, such as customizing symbols, fonts, font sizes, line thickness and type, text boxes, titles, data labels, tick marks, X and Y axis scales, and colors. Multiple plots can be combined into a single plot when there are several variables to overlay. The annotation facility can also be used to further customize the graph. These options are best used in SAS v. 9.3 without ODS.

Examples using real environmental data are shown by plotting concentrations for the metal lead in two soil depth intervals: surface soil and subsurface soil. In addition to quantitative distributional tests such as the Shapiro-Wilk test to test the null hypothesis of normality or lognormality, the probability plots confirm the results of these tests visually.

This paper will show the SAS code that generates high quality normal and lognormal probability plots using `PROC UNIVARIATE`. The SAS code presented uses the SAS System for personal computers version 9.3 and 9.4 running on a Windows® 7 Professional platform.

CALCULATING PROBABILITY PLOTS

In order to understand probability plots, some calculations must first be performed if using `PROC GPLOT`. `PROC UNIVARIATE` performs these calculations automatically. For normal probability plots, the variable to be examined is plotted on a linear scale on the vertical Y axis. The X axis shows plotting positions from the inverse standard normal cumulative distribution function (cdf). For a data set with n observations, the data are sorted from smallest to largest. We denote the data as x_i for i = 1 to n, where x_1 is the smallest observation and x_n is the largest observation. Though there are several plotting position functions in the statistical literature, the most common plotting position used for probability plots is the Blom plotting position (Helsel 2005, p. 48).

For each observation x_i ($i = 1, 2, \ldots, n$), the plotting position p_i is calculated as shown in Equation 1.

$$p_i = \frac{i-0.375}{n+0.25} \quad (1)$$

Note that all p_i must be exclusively between 0 and 1 ($0 < p_i < 1$). Each p_i represents the cumulative area (percentile) under the standard normal cdf for each x_i. The x_i are then plotted on the Y axis, while the p_i are plotted on the X axis for each ordered pair (p_i, x_i) for $i = 1, 2, \ldots, n$.

A linear regression line is then fit to the data. If the data plot approximately linearly along its regression line, then the data will be approximately normally distributed on a normal probability plot (where Y has a linear scale). Any curvature, breaks in the distribution or inflection points will indicate deviations from normality. Potential outliers can also be identified on both the right and left tails of the distribution. The slope of the regression line is a measure of the standard deviation of the data. A steep regression line indicates large variability, while a flatter regression line indicates low variability. The median is the data point that is associated with p_i = 50th percentile. The range is the difference between the maximum data point (x_n) and the minimum data point (x_1).

PROBABILITY PLOTS USING PROC UNIVARIATE

Using `PROC UNIVARIATE` in SAS v. 9.4, the SAS code to generate normal probability plots is simple and generates normal probability plots, as shown in Figures 1 and 2. The `probplot` option in `PROC UNIVARIATE` creates the normal probability plot without any additional calculations programmed by the user. The example data set `metal` consists of real surface and subsurface soil concentrations for a site analyzed for total metals.

```
data metal;
 input MEDIA $1-15 Lead;
 UNITS = ‘mg/kg’;
 label lead = 'Lead (mg/kg) in Surface Soil';
datalines;
Subsurface Soil             38.5
Subsurface Soil            47.65
Subsurface Soil            48.55
Subsurface Soil             55.9
Subsurface Soil             58.1
Subsurface Soil             58.7
Subsurface Soil             59.9
Subsurface Soil             61.9
Subsurface Soil             63.8
Subsurface Soil             71.3
Subsurface Soil             74.8
Subsurface Soil             79.9
Subsurface Soil             80.2
Subsurface Soil             93.1
Subsurface Soil             99.1
Subsurface Soil              103
Subsurface Soil              106
Subsurface Soil              106
Subsurface Soil              122
Subsurface Soil              125
Subsurface Soil              140
Subsurface Soil              150
Subsurface Soil              153
Subsurface Soil              163
Subsurface Soil              174
Subsurface Soil              185
Subsurface Soil              220
Subsurface Soil              251
Subsurface Soil              462
Subsurface Soil              505
Surface Soil                36.4
Surface Soil                37.3
Surface Soil                45.7
Surface Soil                  47
Surface Soil                48.4
Surface Soil                50.8
Surface Soil                52.2
Surface Soil                55.9
Surface Soil                65.9
Surface Soil                66.1
;
run;

ods graphics on;
proc univariate data=metal;    ** for Fig. 1;
  where media = 'Surface Soil';
  var Lead;
  probplot Lead / normal(mu=est sigma=est)
    odstitle='Normal Probability Plot for Lead in Surface Soil';
  output out=statsout min=min mean=mean median=median skewness=skewness
    max=max std=std cv=cv probn=probn; run; quit;
ods graphics off;

proc print data=statsout; run;

data metal2;
  set metal;
  label lead = 'Lead (mg/kg) in Subsurface Soil';
run;

ods graphics on;
proc univariate data=metal2;   ** for Fig. 2;
  where media = 'Subsurface Soil';
  var Lead;
  probplot Lead / normal(mu=est sigma=est)
    odstitle='Normal Probability Plot for Lead in Subsurface Soil';
  run; quit;
ods graphics off;
```

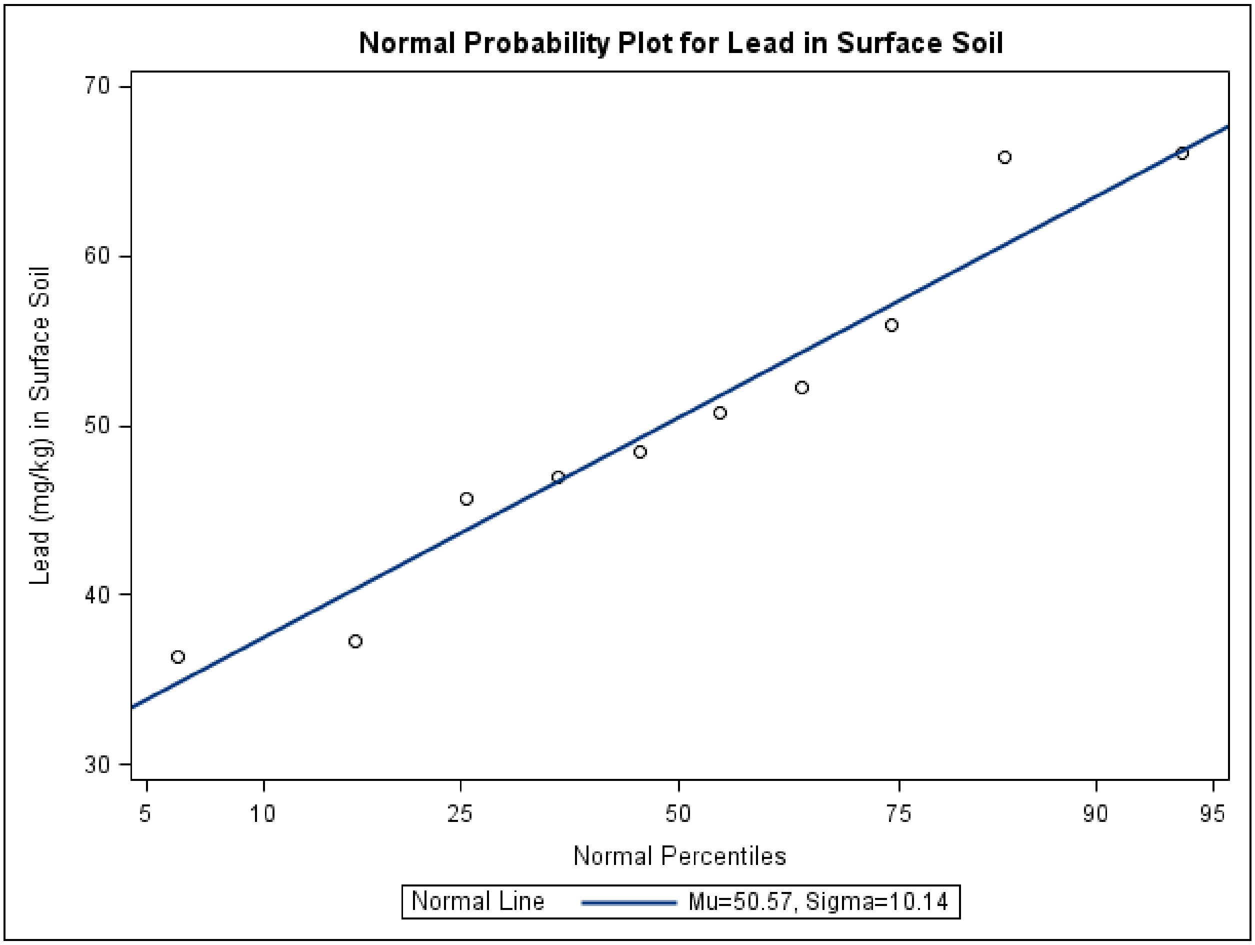

Figure 1. Normal Probability Plot for Lead in Surface Soil

Figure 1 shows 10 lead concentrations in surface soil. The data plot approximately linearly along its linear regression line, which indicates the data are approximately normally distributed. The regression line is generated using the `normal(mu=est sigma=est)` option, which estimates the two parameters mean (`mu`) and standard deviation (`sigma`) from the data for an assumed normal distribution. Summary statistics such as the minimum, mean, median, maximum, standard deviation, skewness and coefficient of variation were also produced by `PROC UNIVARIATE` and output into the SAS data set `statsout`. Figure 1 shows the median concentration is 49.6 mg/kg at the 50th percentile. The Shapiro-Wilk (SW) test *p*-value calculated using the `probn` option within `PROC UNIVARIATE` confirms the surface soil is normally distributed with a *p*-value of 0.4689.

Figure 2 shows 30 lead concentrations in subsurface soil. The distribution has significant curvature compared to its linear regression line and a break in the distribution between the second and third highest concentrations. The two highest concentrations appear as possible outliers given that they stand out from the rest of the distribution. The plot indicates a statistical outlier test should be conducted for the two highest concentrations to determine if these concentrations are statistical outliers at a specified significance level. Figure 2 shows the median concentration is 101.5 mg/kg at the 50th percentile. The SW *p*-value for subsurface soil is <0.0001, which indicates the data are not normally distributed.

Because the subsurface soil data in Figure 2 are not normally distributed due to significant curvature, a break in the distribution and two possible outliers in the right tail, the data should be tested to see if it is lognormally distributed. One way to do that is to calculate the natural logarithm (ln) of each data point and create a normal probability plot on the ln transformed concentrations. If the ln transformed concentrations plot approximately linearly along its linear regression line, then the data are approximately lognormally distributed.

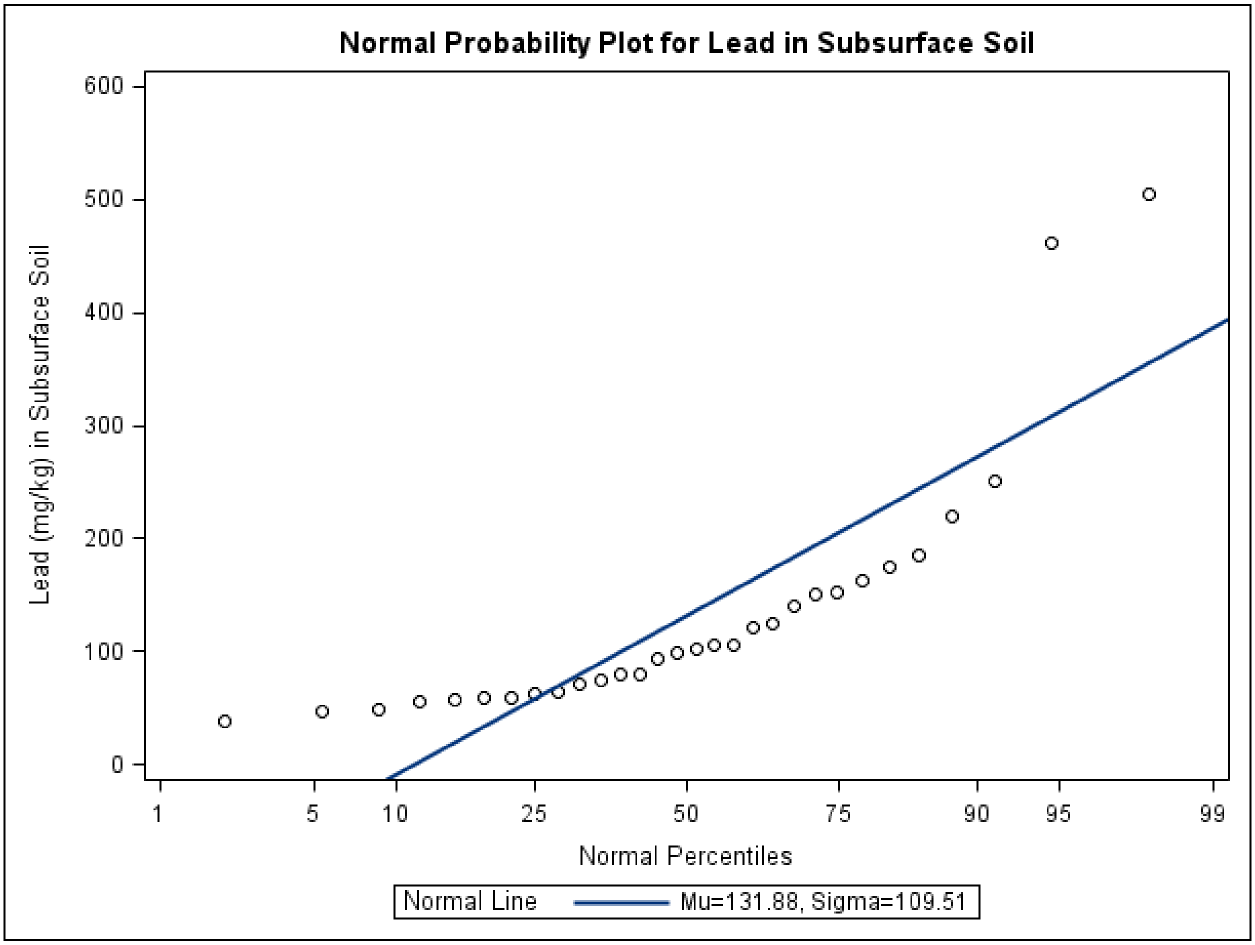

Figure 2. Normal Probability Plot for Lead in Subsurface Soil

The SAS code to transform the data and produce a normal probability plot on the ln transformed data is shown next. The SAS function `log` calculates the natural logarithm on positive concentrations. A normal probability plot on the ln transformed concentrations is shown in Figure 3.

```
data metal3;
  set metal;
    where media = 'Subsurface Soil';
  if lead > 0 then Ln_Lead = log(lead);
  label Ln_Lead='Ln(Lead) in Subsurface Soil [ln(mg/kg)]';
run;

ods graphics on ;
proc univariate data=metal3;      ** for Fig. 3;
  var Ln_Lead;
  probplot Ln_Lead / normal(mu=est sigma=est)
    odstitle='Normal Probability Plot for Ln(Lead) in Subsurface Soil';
  output out=statsout min=min mean=mean median=median skewness=skew
    max=max std=std cv=cv probn=probn;
  run; quit;
  ods graphics off;
  proc print data=statsout; run;
```

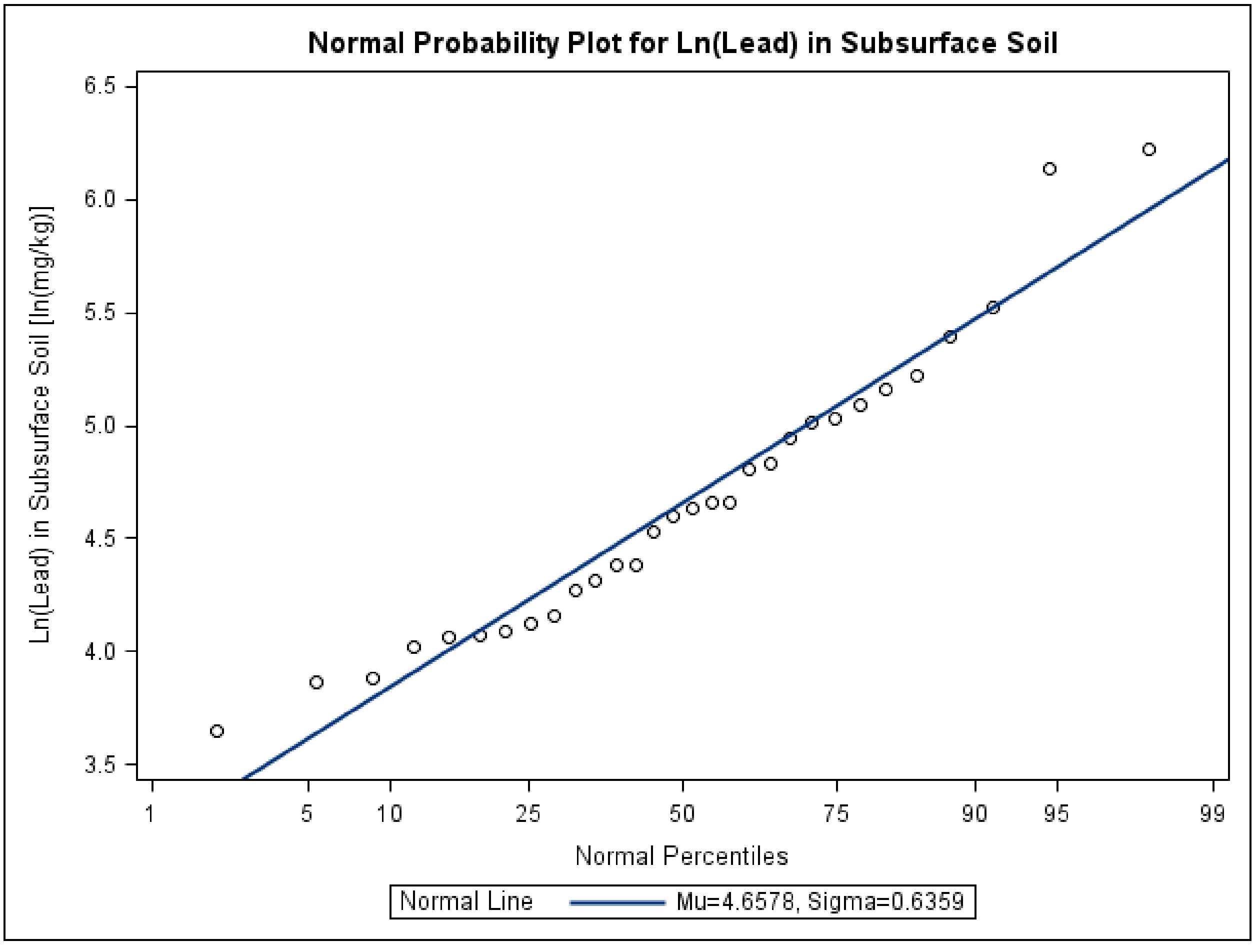

Figure 3. Normal Probability Plot for Ln (Lead) in Subsurface Soil

Figure 3 shows the ln transformed concentrations plot approximately linearly along its linear regression line, which indicates the data are lognormally distributed. The SW test *p*-value is 0.1555, which confirms the data are lognormally distributed at the 0.05 significance level. The Y axis in Figure 3 is in ln transformed units and is on a linear scale. The two highest concentrations still appear to be possible outliers. The median transformed concentration is 4.62 ln(mg/kg), so the median untransformed concentration is exp(4.62) = 101.5 mg/kg. When the X and Y axes are on a linear scale, the probability plot is a normal probability plot. When the X or Y axis is on a $\log_{10}$ scale using untransformed concentrations, the probability plot is a lognormal probability plot.

Another way to test if the data are lognormally distributed without first transforming the concentrations is to use the `lognormal` option within `PROC UNIVARIATE`. However, a lognormal distribution has more parameters than a normal distribution. The three parameters that must be estimated in order to generate a linear regression line through the data are scale, sigma, and theta. Sigma is the shape parameter, while theta is the threshold parameter. The following SAS code produces the lognormal probability plot shown in Figure 4. The `odstitle` option can be used to change the title of the graph.

```
ods graphics on ;
proc univariate data=metal2;
  where media = 'Subsurface Soil';   ** for Fig. 4;
  var Lead;
  probplot Lead / lognormal(scale=est sigma=est theta=0)
    odstitle='Lognormal Probability Plot for Lead in Subsurface Soil';
   run; quit;
ods graphics off;
```

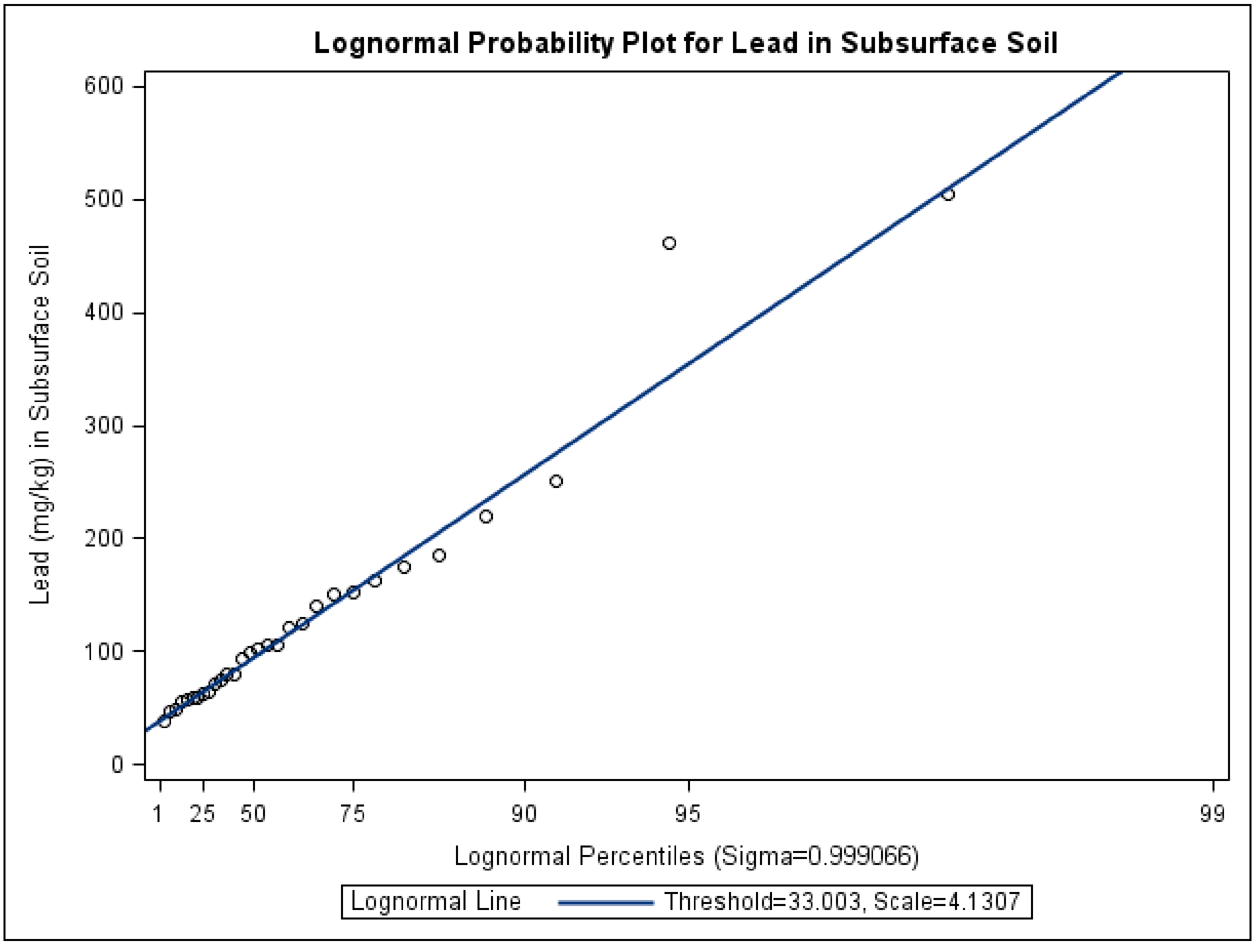

Figure 4. Lognormal Probability Plot for Lead in Subsurface Soil

Figure 4 shows the data plot very nicely along the regression line except the second highest concentration. The parameter estimates from the data are for the threshold parameter theta (33.003), the scale parameter (4.1307), and shape parameter sigma (0.999066). Figure 4 is a lognormal probability plot because the scale on the X axis assumes percentiles from a lognormal distribution.

Probability plots for many other continuous distributions (such as beta, exponential, Gumbel, Pareto, power, Rayleigh, and Weibull) can also be generated for testing the distribution of data simply by specifying these distributions and their appropriate parameters as options in the `probplot` statement. The advantage of generating normal and lognormal probability plots using `PROC UNIVARIATE` is the simplicity of the code.

CONCLUSION

Probability plots are a useful exploratory data analysis tool that data analysts can use for exploring data. Probability plots reveal whether the data follow a parametric distribution such as normal or lognormal, identify possible outliers, show variability in the data and central tendency estimates. SAS code was presented to easily produce normal or lognormal probability plots using the `probplot` option within `PROC UNIVARIATE` using ODS.

REFERENCES

Helsel, Dennis R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Hoboken, NJ: John Wiley & Sons.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dennis J. Beal, Ph.D.
Senior Statistician/Risk Scientist
Leidos
301 Laboratory Road
P.O. Box 2502
Oak Ridge, Tennessee 37831
e-mail: beald@leidos.com