Paper 239-29

DATA: "Getting to Know You; Getting to Know All About You"

Larry Douglass, University of Maryland, College Park, Maryland
Sue Douglass, Independent Consultant, College Park, Maryland

ABSTRACT

Before data can be appropriately analyzed, researchers and/or analysts, must get to know their data. Getting to know your data involves both descriptive statistics and graphics techniques. PROC UNIVARIATE provides many descriptive statistics, graphics tools, and some simple inferential statistics to help you examine your data. Starting in Version 8.2, a number of high-quality graphics options were added to PROC UNIVARIATE to help you visualize your data and compare it to a number of possible theoretical distributions (normal, exponential, lognormal, etc.). Goodness-of-fit tests are also provided. A side benefit of a thorough examination of your data is that you will also identify possible errors and outliers in your data. PROC UNIVARIATE should be the tool of choice for getting to know your data.

This tutorial is a beginner's approach to using PROC UNIVARIATE. It will concentrate on understanding default output along with a few simple options that will easily allow SAS® users at any level to quickly examine their data.

INTRODUCTION

The purpose of this beginning tutorial is to introduce SAS users to PROC UNIVARIATE as tool for quickly examining and visualizing your data. The Univariate Procedure produces a large number of descriptive statistics and graphical presentations. The minimum syntax required to use the Univariate Procedure is very simple and produces many descriptive statistics. Although several statements and many options are available, the knowledge of just a few statements and options will allow you to conduct a very through examination of your data, including graphics output. You will be able to produce low density stem-and-leaf plots, box-and-whisker plots and normal probability plots. Additional statements can be use to produce high density graphic plots, including histograms overlaid with various distributions, normal probability plots and Q-Q plots.

We have selected two data sets to illustrate the Univariate Procedure. The first data set is referred to as the "Belize Height Data" and contains growth variables on 3 to 5 year-old male and female children from rural villages in Belize. The second data set is labeled the "Dairy Cattle Neospara Data" and contains sera titer values on animals of different genders and ages housed at four locations. This infectious disease causes abortion in dairy cattle and can be a serious problem in some herds.

GETTING STARTED WITH PROC UNIVARIATE

The "Belize Height Data" used in this manuscript contains the child's ID, gender, age, height (HT) and the child's height adjusted for age (HAZ) expressed as a Z score. A few of the 41 lines are printed below.

Belize Height Data

Obs	ID	SEX	AGE	HT	HAZ
1	1704	M	5	99.06	-2.77
2	1712	F	5	96.52	-2.58
3	1715	F	5	100.30	-1.61
4	1718	M	5	85.09	-5.22
5	1722	M	5	97.79	-2.21
6	1730	M	4	90.17	-3.58
7	1736	M	4	95.25	-1.94
8	1738	F	4	90.17	-2.96
9	1756	M	3	93.98	-0.78
.					
.					
.					
41	1884	M	3	83.82	-3.22

The following syntax is the simplest that could be used to produce output from the Univariate Procedure.

```
PROC UNIVARIATE;
```

This would generate a large number of descriptive statistics on all numeric variables in the currently active SAS data set.

For our first example, we added a couple of options and one statement to the PROC. Before proceeding, just a note about the convention used throughout this paper to present SAS syntax. To help you understand where the syntax originated, uppercase words are SAS specified syntax and lowercase words are user supplied. Of course, you do not need to follow this convention in your own program.

```
PROC UNIVARIATE DATA=height NORMAL PLOT;
   VAR ht haz;
```

The DATA= option specifies which of the two data sets the Univariate Procedure will process. The NORMAL option requests that Univariate generate test statistics for the null hypothesis that the sampled population is normally distributed. The PLOT option requests that the output include a stem-and-leaf plot, a box-and-whisker plot and a normal probability plot. Now, let's examine the output in detail.

MOMENTS

You probably recognize a number of the statistics. N, Mean, Std Deviation, Variance, Coeff Variation, SS (sums of squares) and Std Error Mean should be familiar to you. Of the remaining statistics, Skewness is a measure of asymmetry while Kurtosis is a measure of flatness or peakness of the sample distribution relative to a normal distribution. For both statistics, a normal distribution will have a value of zero. However, for samples, these values would seldom be exactly zero, even if drawn from a normal distribution. The question is, for a given sample size, how close to zero would be acceptable? Tables of acceptable ranges (or critical values) of skewness and kurtosis values are available in many introductory statistical textbooks. For skewness or kurtosis values outside the acceptable range, negative values are skewed left or flat, and positive values are skewed right or peaked.

Belize Height Data
The UNIVARIATE Procedure
Variable: HT

Moments

N	41	Sum Weights	41
Mean	91.8834146	Sum Observations	3767.22
Std Deviation	7.42419141	Variance	55.118618
Skewness	-0.2048877	Kurtosis	-1.0874654
Uncorrected SS	348349.782	Corrected SS	2204.74472
Coeff Variation	8.08001252	Std Error Mean	1.1594639

BASIC STATISTICAL MEASURES

The range is the largest minus the small observed value in the sample. The interquartile range is the 75th percentile minus 25th percentile value. It is used in the construction of a box and whisker plot and represents the range of values that covers the center 50% of the data. It is also used to identify potential outliers in the data. Be sure and look for the note regarding the Mode. If the sample is multimodal, Univariate lists only the smallest of the modes.

Basic Statistical Measures

Location		Variability	
Mean	91.88341	Std Deviation	7.42419
Median	93.98000	Variance	55.11862
Mode	82.55000	Range	26.64000
		Interquartile Range	13.97000

NOTE: The mode displayed is the smallest of 2 modes with a count of 5.

TESTS OF LOCATION

I don't generally find these default tests of significance very useful. These are three simple tests of the null hypothesis that the mean (t test) or the median (sign or signed rank test) of the population is 0. In this case it would be nonsense to test the hypothesis that height equals zero. However, for HAZ, the researcher might find these hypotheses of interest, because a value of zero represents mean normal height.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's	t	79.24646	Pr > |t|	<.0001
Sign	M	20.5	Pr >= |M|	<.0001
Signed Rank	S	430.5	Pr >= |S|	<.0001

TESTS OF NORMALITY

There are four tests of normality generated by the NORMAL option on the PROC UNIVARIATE statement. I generally use the Shapiro-Wilks test of normality and will describe it in more detail. The idea of the W statistic can be understood if one thinks of W as a measure of correlation, where W=1 represents perfect correlation between the sampled distribution and a truly normal distribution with the sample mean and standard deviation. The magnitude of W may be more relevant than the significance of W. Often the researcher wishes to show that the sampled distribution is normal. In effect, this is a test of equivalence. This often makes interpretation of the test statistic probability (Pr < W) misleading. With a small sample size, if the population distribution is non normal, the power of the test is likely to be insufficient to reject the null hypothesis. On the other hand with large sample size, it is likely that the population could differ negligibly from normal, yet given the high sensitivity the null hypothesis is likely to be rejected.

If normality of the residuals is being examined as a requirement for a parametric test of mean differences (t-tests or anova), my experience is that Ws between 0.95 and 1.0 certainly shows adequate normality. Ws between 0.90 and 0.95 are of concern with few replicates per treatment. Ws below 0.9 are of serious concern unless the number of replicates per treatment is large. When examining residuals you are looking at the worst case. That is because the central limit theorem has not had an opportunity to exercise its influence. The normality of the distribution of treatment means will only get better as you average across more and more replicates, and parametric tests of hypotheses about treatment means are valid when the distributions of treatment means are essentially normal.

Tests for Normality

Test	--Statistic --		-----p Value------	
Shapiro-Wilk	W	0.947421	Pr < W	0.0568
Kolmogorov-Smirnov	D	0,129575	Pr > D	0.0825
Cramer-von Mises	W-Sq	0.124833	Pr > W-Sq	0.0496
Anderson-Darling	A-Sq	0.778628	Pr > A-Sq	0.0413

PERCENTILES

Quantiles provides estimates of the percentiles of the sample distribution. Be careful when sample size is small. Note that even with 41 observations, the sample size is too small to distinguish between some of the percentiles. For example, the value for 90% is the same as the 95% and the 99% is the same as the maximum, while on the other end of the distribution the values for 5% are the same as the 10% and 1% is the same as the minimum.

Quantiles (Definition 5)

Quantile	Estimate
100% Max	104.11
99%	104.11
95%	101.60
90%	101.60
75% Q3	97.79
50% Median	93.98
25% Q1	83.82
10%	82.55
5%	82.55
1%	77.47
0% Min	77.47

ERRORS OR OUTLIERS

Unusual values are frequently of interest or concern in data analysis. This table identifies the 5 smallest and the 5 largest values in the data set. Unusually large gaps between some values may indicate an outlier.

Extreme Observations

----Lowest-----		----Highest-----	
Value	Obs	Value	Obs
77.47	14	101.60	36
78.24	6	101.60	38
82.55	13	101.60	41
82.55	12	104.10	25
82.55	5	104.11	35

STEM-AND-LEAF/BOX-AND-WHISKER

The stem-and-leaf plot is a horizontal histogram and was generated by the PLOT option of the PROC UNIVARIATE statement. The first two or three digits of each value are used as the stem and the leaf is the next digit. This forms a histogram that not only provides the frequency for each class, but the actual value or, in this case, slightly rounded values for each observation. For example, the stem 78 and leaf 2 is the value of 78.24. The 77.47 is in the stem marked 76, which includes values in the 76's and 77's.

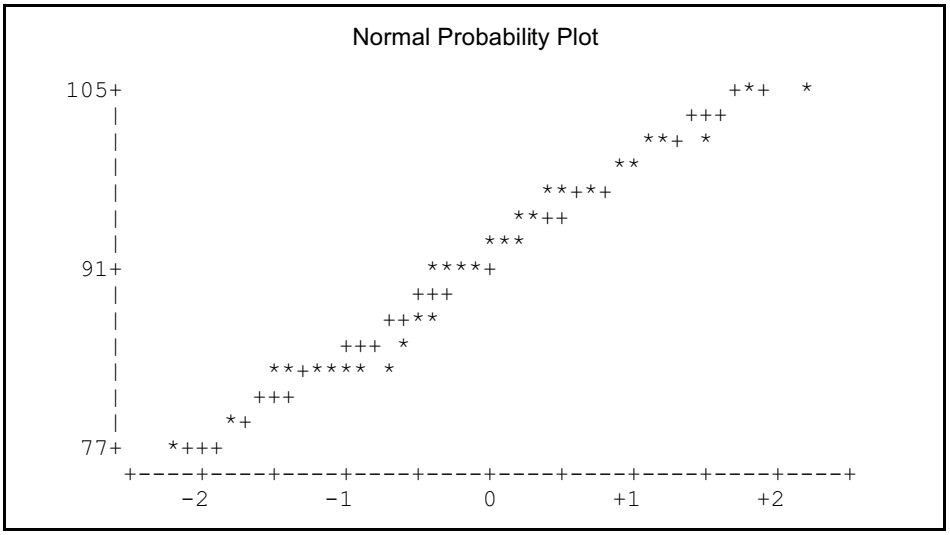
The vertical axis of the box-and-whisker plot (also requested by the PLOT option) is also labeled by the stem of the stem and leaf plot. The bar in the center anchored with the * represents the median, which falls in stem 94 (median=93.98) and the + in the center of the box represents the stem where the mean (91.88) is located. The ends of the box are the stems for the 75th and 25th percentiles (interquartile range) and the whiskers on each end extend to the stem containing the minimum and maximum values in the data set. The whiskers will not extend more than 1½ interquartiles. Observations between 1½ and 3 interquartiles are marked with a "0" and those beyond 3 interquartiles are marked with an "*", which may be extreme enough to be considered an outlier. In this case, none of the observations were considered unusual by these criteria, however, they will be evident in later plots.

Stem	Leaf	#	Boxplot
104	11	2	
102			
100	3666	4	
98	11	2	
96	2588888	7	+-----+
94	000022	6	*-----*
92			| |
90	222444	6	| + |
88			| |
86	46	2	| |
84	1	1	| |
82	666668888	9	+-----+
80			|
78	2	1	|
76	5	1	|

----+----+----+----+

NORMAL PROBABILITY PLOTS

The Normal Probability Plot (also generated by the PLOT option) provides a visual check of normality. On the vertical axis are the values of hetight. The horizontal axis represents the number of standard deviations above or below the mean. When plotted on these scales, a normal distribution forms a straight line running from the lower left corner to the upper right. UNIVARIATE forms this plot by first laying down a band of + signs to form the expected straight line where normal data would lie. The sample data are then represented with * signs laid down over top of the + signs. If the sample data are from a normal distribution, we would expect most of the observations to land on the + signs with only a few observations off the diagonal band and those most commonly occur at the extremes of the data.



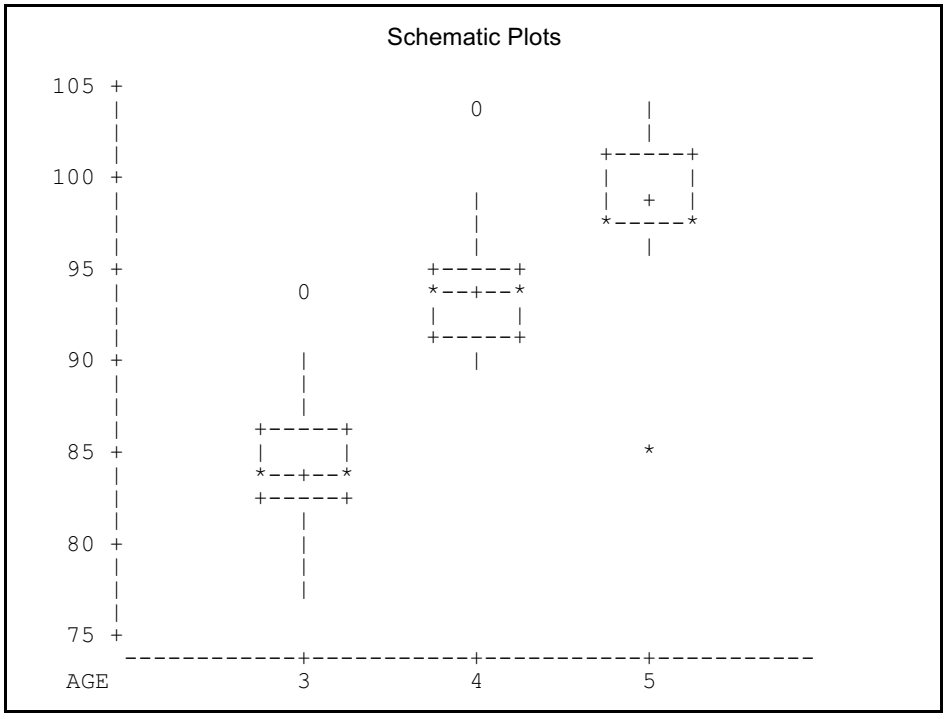
The same output was generated for HAZ (height adjusted for age Z score), because the VAR statement included both HT and HAZ.

CLASS AND BY STATEMENTS

The following SAS statements can be used to generate the above output for each level of a categorical variable in the data set. In the first PROC we use the CLASS statement to request the desired output for each age (3, 4 and 5 year olds). The second PROC uses the BY statement to produce the same result. In this later case, a SORT procedure was required before the Univariate procedure to arrange the data in order by age. The only difference in the resulting output is one additional plot, commonly referred to as a side-by-side box-and-whisker plot (labeled as a "Schematic Plot" by SAS). This plot is the only portion of the resulting output included below.

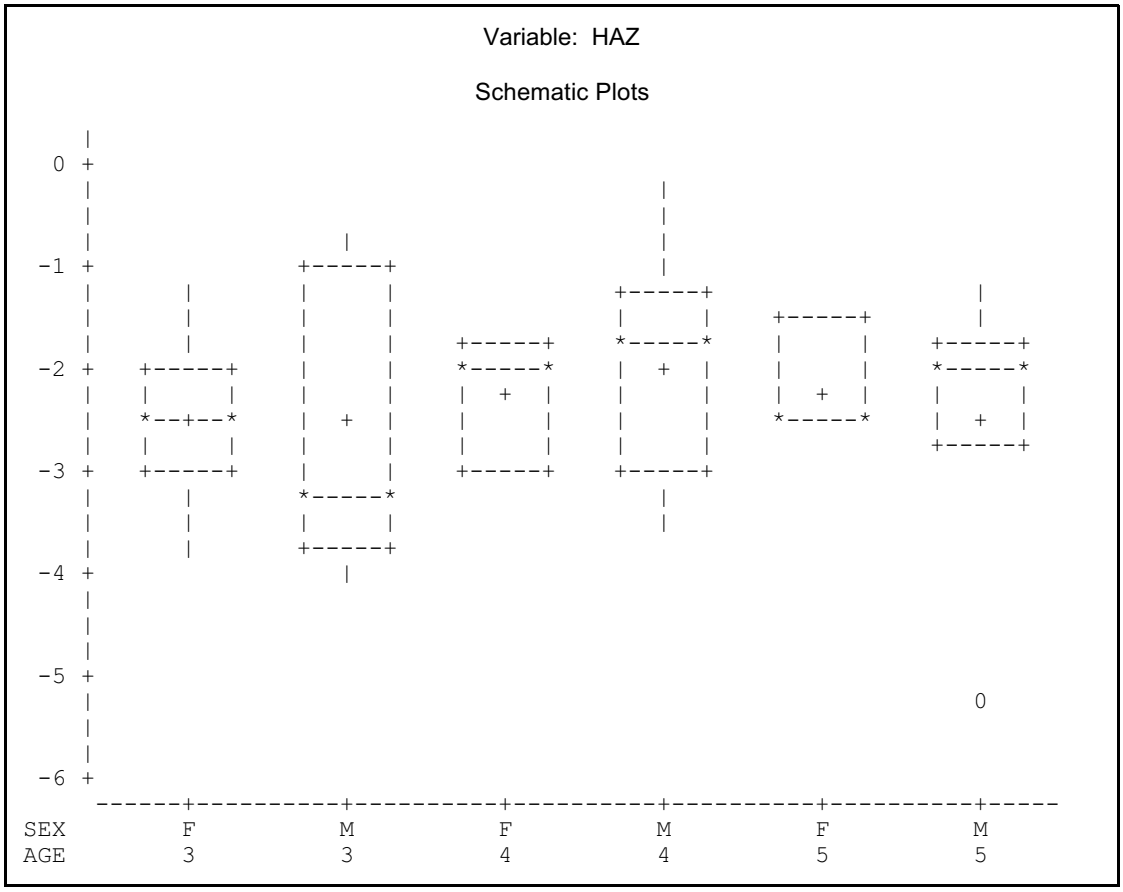
```
TITLE2 Using the Class Statement;
PROC UNIVARIATE DATA=height NORMAL PLOT;
   CLASS age;
   VAR ht;

TITLE2 Using the By Statement;
PROC UNIVARIATE DATA=height NORMAL PLOT;
   BY age;
   VAR ht;
```

You can now see the effect of increasing age, which as expected is associated with increasing height. You will also notice that we now observe some observations that are a little unusual - "0", and one observation that is very unusual - "*", (a small 5 year old). This child had a height of 85 cm, while all other 5 year olds had heights between 95 and 105.



You may also use the BY statement to generate a two-way plot, in this case for age and gender. The variable plotted here is the standardized height, which appears to have corrected for the age effects. Also notice that all the children had heights below average as indicated by the negative Z scores.

```
TITLE2 By age and sex;
PROC UNIVARIATE DATA=height PLOT;
   BY age sex;
   VAR haz;
```



ENHANCED UNIVARIATE PLOTS

Now to illustrate how the Univariate can be used to easily generate high quality plots. We are introducing three new statements (HISTOGRAM, PROBPLOT and INSET) in this Univariate Procedure. The HISTOGRAM statement, with the NORMAL option, requests that a histogram be generated and overlaid with a normal distribution with the sample mean and standard deviation. The PROBPLOT statement, with the NORMAL option, requests a normal probability plot. The INSET statement identifies statistics to be printed on the figure. Presented below are the new output and four new figures (2 for HT and 2 for HAZ) generated by this PROC UNIVARIATE.

```
TITLE2 Histogram and ProbPlot Graphics Statements;
PROC UNIVARIATE DATA=height NORMAL;
   VAR ht haz;
   HISTOGRAM / NORMAL;
       INSET N MEAN STD;
   PROBPLOT / NORMAL;
       INSET N MEAN STD NORMALTEST PNORMAL;
```

STATISTICS USED TO GENERATE GRAPHICS

The output window contains the estimates of the parameters (mean and standard deviation) used to determine the fitted distribution, goodness of fit statistics for the normal distribution and a table of percentiles for the sample and for the normal distribution.

Histogram and ProbPlot Graphics Statements

The UNIVARIATE Procedure
Fitted Distribution for HT

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	91.88341
Std Dev	Sigma	7.424191

Goodness-of-Fit Tests for Normal Distribution

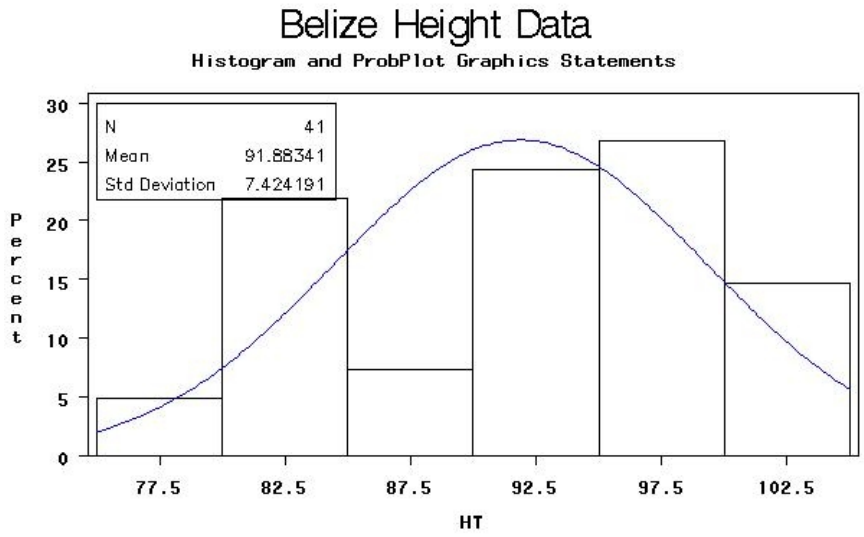
Test		---Statistic---		-----p Value------
Kolmogorov-Smirnov	D	0.12947432	Pr > D	0.083
Cramer-von Mises	W-Sq	0.12483254	Pr > W-Sq	0.050
Anderson-Darling	A-Sq	0.77862781	Pr > A-Sq	0.041

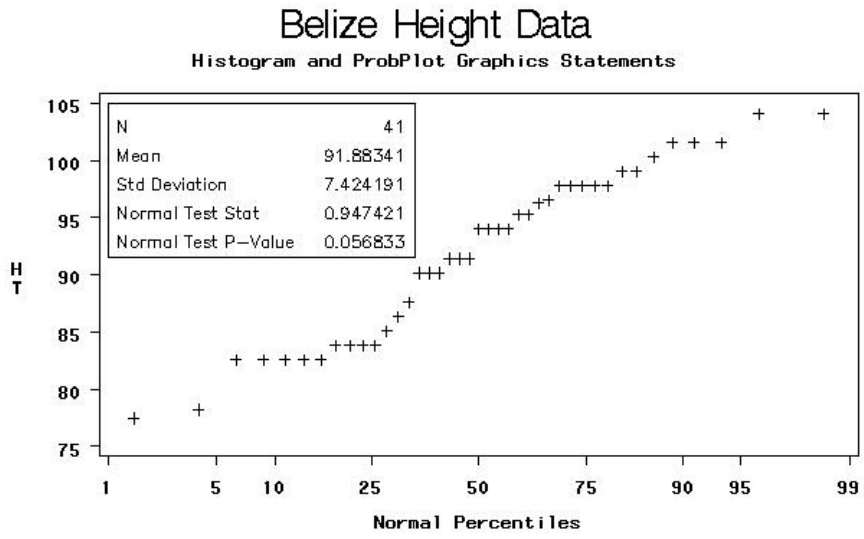
Quantiles for Normal Distribution

	----------Quantile----------	
Percent	Observed	Estimated
1.0	77.4700	74.6122
5.0	82.5500	79.6717
10.0	82.5500	82.3689
25.0	83.8200	86.8759
50.0	93.9800	91.8834
75.0	97.7900	96.8910
90.0	101.6000	101.3979
95.0	101.6000	104.0951
99.0	103.1100	109.1547

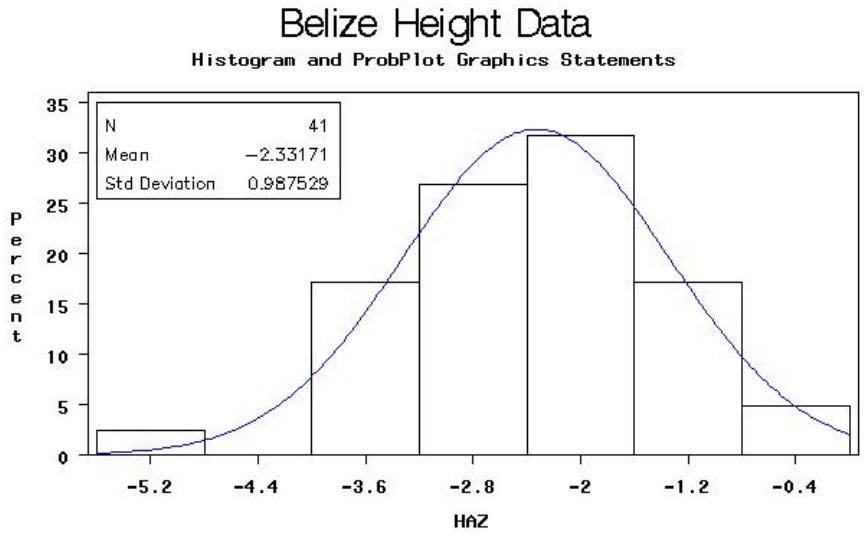
HISTOGRAMS AND NORMAL DISTRIBUTION PLOTS

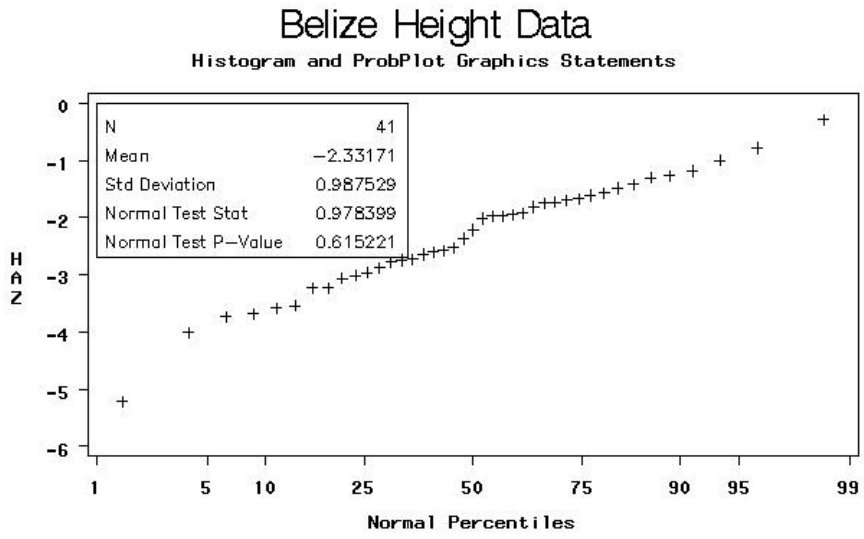
The height data do not appear to fit the normal distribution. However, the normal test statistic indicates that, given a sample of 41 observations the observed sample distribution could have come from a normal distribution.





It is not surprising that the corresponding plots for the HAZ appears to be much more normal, because the HAZ values are adjusted for differences in age.





THE NEOSPORA DATA

The Neospara data contains sera titer values on 1048 animals of different genders and ages, housed at four locations. The infection causes abortion in dairy cattle and can be a serious problem in some herds. The following is a small portion of the data, where Elisa are the sera titer values and Log Elisa are the $\log_{10}$ transformation of the titer values.

Dairy Cattle Neospara Data

Obs	ID	Elisa	Location	Sex	Log Elisa
1	1015HNB	474	1	F	2.75891
2	1018HNB	474	1	F	2.75891
3	1041HNB	948	1	F	3.02036
4	1067HNB	441	1	F	2.73320
5	1086HNB	640	1	F	2.86923
6	1101HNB	731	1	F	2.91960
7	1115HNB	1707	1	F	3.25696
8	1117HNB	492	1	F	2.77232
9	1126HNB	200	1	F	2.47712
.					
.					
.					

CROSS TABULATION

In large data sets it is frequently helpful to look at the distribution of animals by various categories of the classification variables. The Frequency Procedure can be used to generate a two-way cross tabulation of the number animals for combinations of location and sex. We included the NOROW and NOCOL options to suppress the printing of column and row percentages to simplify the output.

```
TITLE1 Dairy Cattle Neospara Data;
TITLE2 Two-Way Frequencies;
PROC FREQ DATA=neospara;
   TABLE location*sex / NOROW NOCOL;
```

Location identifies the barn housing the animals and is confounded with age. Location 1 housed calves younger than six months of age. Calves form six to 15 months were housed in location 2. Location 3 housed non-lactating heifers, while location 4 housed the mature lactating cows. Absence of males in locations 3 and 4 must be considered when analyzing and presenting the results of the study.

Dairy Cattle Neospara Data
Two-Way Frequencies
The FREQ Procedure

Table of Location by Sex

Location Frequency Percent	Sex F	M	Total
1	68 6.61	45 4.37	113 10.98
2	62 6.03	171 16.62	233 22.64
3	218 21.19	0 0.00	218 21.19
4	465 45.19	0 0.00	465 45.19
Total	813 79.01	216 20.99	1029 100.00

HISTOGRAMS AND PROBABILITY DENSITY FUNCTIONS

Those having experience with sera titer data would probably have a good idea of how the sera titer values are likely to be distributed. However, suppose that you have not analyzed such data or are unsure if these data are distributed as expected. The Univariate Procedure, with a few additional options, will often help you identify a distribution that will adequately describe the data. The HISTOGRAM statement allows you to fit a number of probability density functions, including normal, exponential, lognormal, Weibull, beta, gamma and a non-parametric kernel estimation procedure. The normal, exponential and lognormal are illustrated in the following Univariate example.

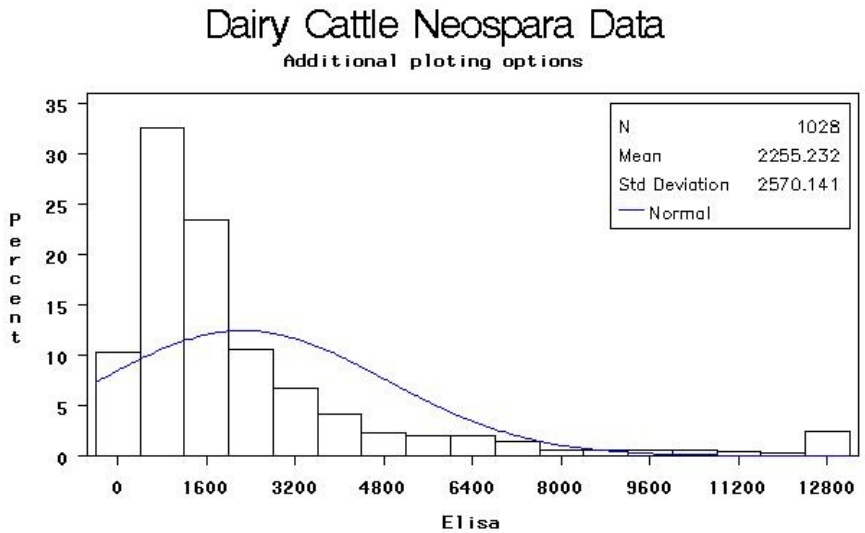
Because we discussed and presented the univariate descriptive statistics and the low density graphics on the previous Univariate Procedures, we will present only the high density graphics in this section. The following Univariate Procedure produces four histograms and one probability plot, three are for the titer levels (elisa) and the last two are for the log of the titer levels (logelisa). The POSITION option is used for some plots to place the descriptive statistics in the upper right-hand corner (NorthEast), rather than in the default upper left-hand corner (NorthWest).

```
TITLE2 Additional ploting options;
PROC UNIVARIATE DATA=neospara normal;
   VAR elisa logelisa;
   HISTOGRAM elisa / NORMAL;
       INSET N MEAN STD NORMAL / POSITION=NE;
   HISTOGRAM elisa / EXPONENTIAL;
       INSET N MEAN STD EXPONENTIAL / POSITION=NE;
   HISTOGRAM elisa / LOGNORMAL(THETA=est);
       INSET N MEAN STD LOGNORMAL / POSITION=NE;
   HISTOGRAM logelisa / NORMAL;
       INSET N MEAN STD NORMAL;
   PROBPLOT logelisa / NORMAL;
       INSET N MEAN STD NORMALTEST PNORMAL;
```

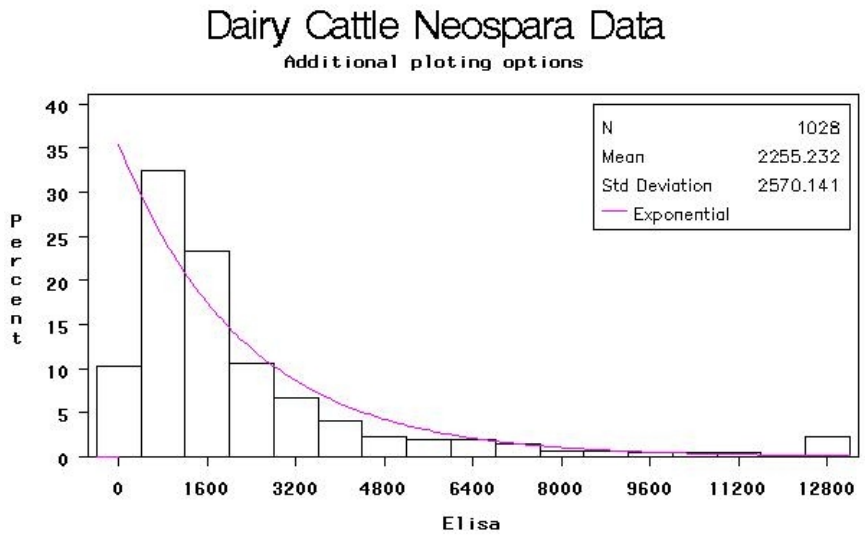
From a comparison of the first three figures, we can make the following points. 1) Of the three probability density functions, the log normal distribution best fits the histogram of the elisa values. 2) Comparing the log normal distribution to the histogram there is an excess of titer values in the right-hand tail of the distribution at the highest titer level.

This box and what follows is the part of the program used for that specific output;

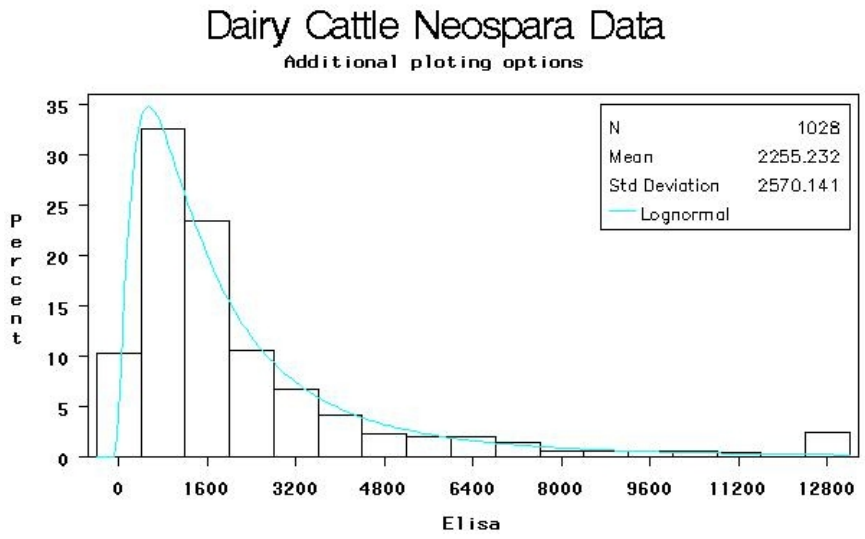
```
HISTOGRAM elisa / NORMAL;
   INSET N MEAN STD NORMAL / POSITION=NE;
```



```
HISTOGRAM elisa / EXPONENTIAL;
   INSET N MEAN STD EXPONENTIAL / POSITION=NE;
```



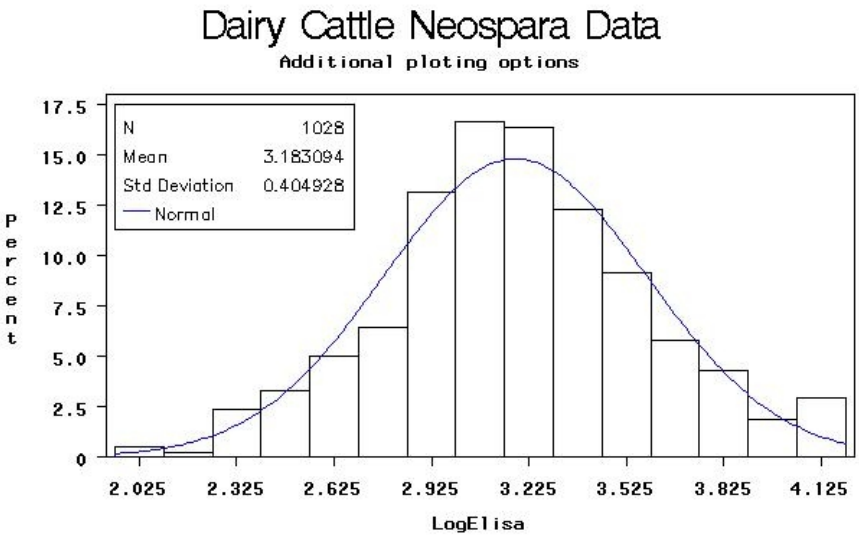

```
HISTOGRAM elisa / LOGNORMAL(THETA=est);
   INSET N MEAN STD LOGNORMAL / POSITION=NE;
```



```
HISTOGRAM logelisa / NORMAL;
    INSET N MEAN STD NORMAL;
```

Because the titer values appear to be log normally distributed, we would expect the data to be normally distributed on a log scale.

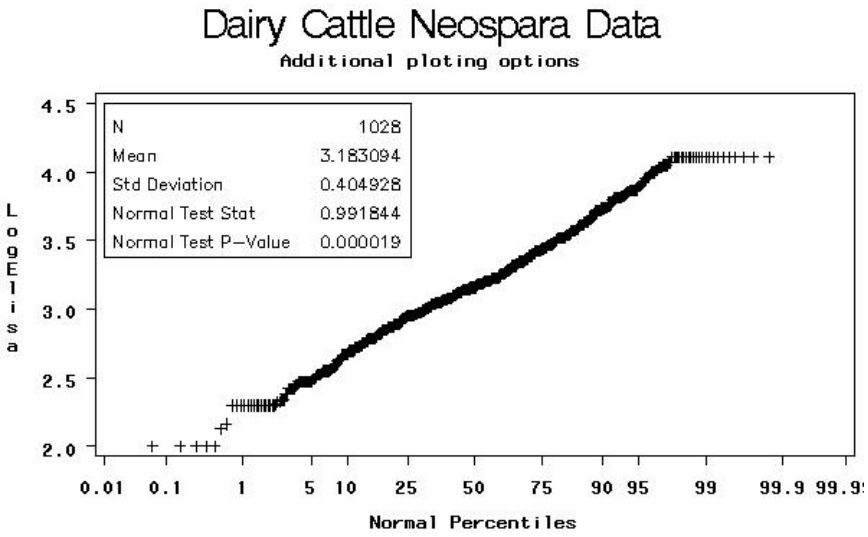
The last histogram and the normal probability plot are on the log of titer values. From a comparison of the histogram and the normal density function it appears that the sample distribution is slightly peaked and heavy tailed.



NORMAL PROBABILITY PLOT

Examination of the normal probability plot may explain the heavy tails' problem. We suspect that the problem is a laboratory sensitivity problem. There are several identical values at the upper end of the distribution and at the lower end, several small titer values. The rest of the distribution appears to be fairly normal as indicated by the nearer linear relationship. The normality statistic(W) is very close to 1, but the probability indicates that it is not likely that the sample data come from a truly normal distribution. We would consider the normal distribution to adequately describe the log titer values. This is probably a case where the sample size is so large that tests of hypotheses indicate non-normality, while the distribution differs very little from a normal distribution.

```
PROBPLOT logelisa / NORMAL;
   INSET N MEAN STD NORMALTEST PNORMAL;
```

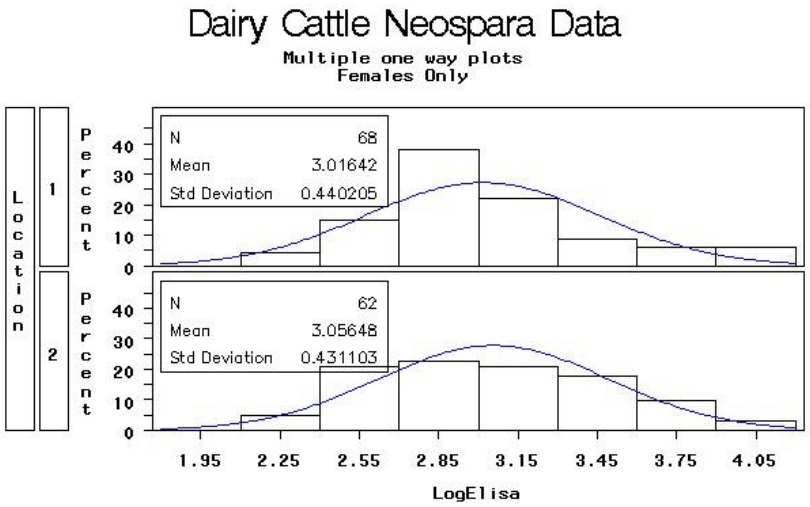


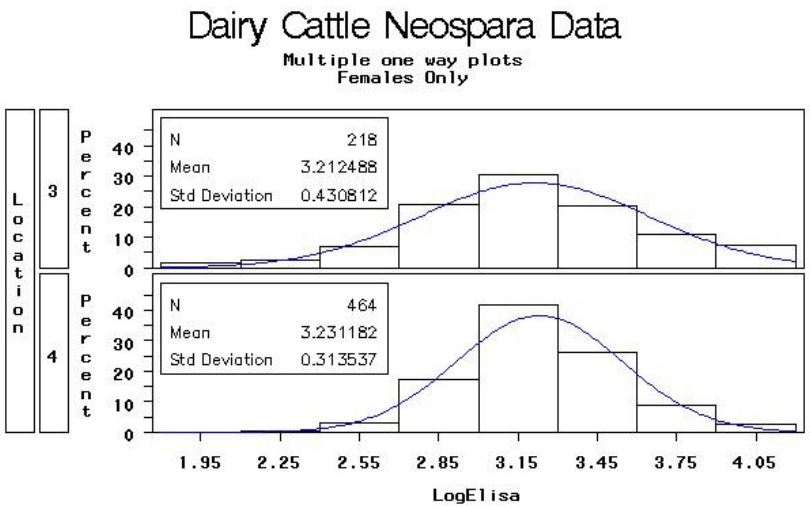
USING ONE-WAY CLASS STATEMENTS

Because of the absence of males at locations 3 and 4, two other presentations are suggested. The first one looks at the distribution of "Females Only" using the WHERE statement to select the females from the data set and the CLASS statement to generate plots for each location. The following Univariate Procedure generates four histograms overlaid with the normal distribution using the sample mean and standard deviation at each location and on the same scale for easy comparison.

```
TITLE2 Multiple one way plots;
TITLE3 Females Only;
PROC UNIVARIATE DATA=neospara;
WHERE sex='F';
CLASS location;
HISTOGRAM logelisa / NORMAL;
       INSET N MEAN STD;
```

Examination of these plots seems to indicate that the distributions are nearly identical for locations 1 and 2, while at location 3 there is a slight increase in titer levels and at location 4 the mean is similar to location 3, but the standard deviations is smaller.



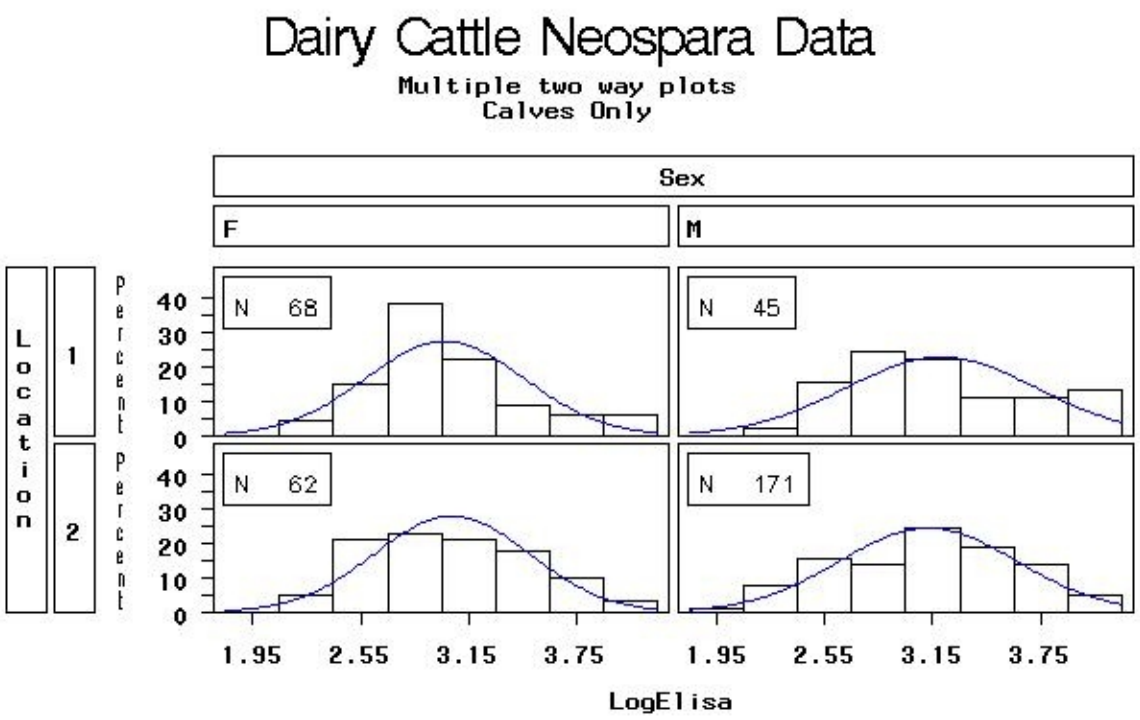


USING A TWO-WAY CLASS STATEMENT

The last Univariate Procedure produces a two-way histogram plot of gender and location for locations 1 and 2 only. The following Univariate Procedure illustrates the appropriate syntax.

```
TITLE2 Multiple two way plots;
TITLE3 Calves Only;
PROC UNIVARIATE DATA=neospara NORMAL;
   WHERE location LE 2;
   CLASS location sex;
   VAR logelisa;
   HISTOGRAM logelisa / NORMAL;
       INSET N;
```

No apparent trends are obvious among the following plots.



CONCLUSIONS

The PROC UNIVARIATE is an easy-to-use procedure that alows you to quickly examine and visualize your data. Descriptive statistics, low density graphics and high density graphics can be obtained by using very few statements and/or options. The Univariate Procedure should be the tool of choice for getting to know your data.

CONTACT INFORMATION

For further information, you may contact:

Larry Douglass
Phone: 301-405-1405
Email: ldouglas@umd.edu

SAS and all other SAS Institute Inc. Products or service names are registered trademarks or trademarks of SAS Institute Inc. In the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.