

Exploratory Data Analysis Using PROC UNIVARIATE

Robert E. Maitland Jr., Bank of America, Jacksonville, FL

ABSTRACT

Exploratory Data Analysis (EDA) is an important first step in any data analysis. PROC UNIVARIATE generates a number of statistics useful for EDA. This paper explores two samples. The focus is on measures of location, measures of dispersion, measures of shape, and plots. An emphasis is placed on the new features available in version 8.

INTRODUCTION

Exploratory Data Analysis (EDA) is an important first step in any type of data analysis. It is the process by which we learn about our data before we use it. Effective EDA improves the quality of our analysis. It aids us in drawing the correct conclusions based on the data.

PROC UNIVARIATE is part of base SAS. It provides a number of statistics and plots for numeric variables that are useful for EDA. This paper will focus on measures of location, dispersion, and shape. We will then discuss leaf and stem plots, histograms, and box and whisker plots.

It is beyond the scope of this paper to provide detailed statistical and mathematical explanations. These can be found in the *Version 8 SAS OnLine Documentation*, "The UNIVARIATE Procedure: Statistical Computations".

METHODOLOGY

Two random samples of 1,000 observations using the Poisson distribution with different seeds and a mean of 25 were created. For the second sample, we added more distortion by multiplying a hundred of the observations by five. This is the code that produced the data:

```
data sample;
  keep v d;
  seed1 = 100;
  seed2 = 1;
  do n = 1 to 1000;
    if mod(n,100) = 0 then m=5; else m = 1;
    call ranpoi(seed1,25,v); v = v;
    d = "Poisson1"; output;
    call ranpoi(seed2,25,v); v = v*m;
    d = "Poisson2"; output;
  end;
  label d = "Distribution Used to Generate
          Variable";
        v = "Variable";
run;
proc sort data = sample;
  by d;
run;
```

The two data sets are compared using EDA. An emphasis is placed on the new features SAS added with version 7. Since these new features have been carried over to version 8 and version 8 is the version now in production, the comparison is made between version 6 and version 8. This is the code used to produce the statistics:

```
proc univariate data=sample alpha=.025 cibasic
  cipctldf(type = asymmetric alpha = .05)
  loccount mu0 = 25 modes nexttrval = 10
  plots plotsize = 24 robustscale
  trimmed = .1 (type = upper alpha = .05)
  winsorized = .025;
  var v;
  by d;
run;
```

The parameters will be discussed in the sections related to the statistics generated by those parameters.

MEASURES OF LOCATION

Measures of location are used to find the central tendency of the data. The ones most commonly used are the mean, median, and mode (where the mean is the average, the median is the middle value with half of the observations having a value less than it and the other half are greater than it, and the mode is the most frequent value).

Percentiles are also useful for a study of tendency; especially the upper and lower quartiles. The upper quartile is the 75th percentile (75% of the values fall below this value) and the lower quartile is the 25th percentile). Therefore, 50% of the observations fall between the upper and lower quartile.

The median is considered more robust than the mean, because it is not distorted by extreme values. If the data is symmetric then the mean and the median will be the same. The mode's usefulness is based on its percent of the population.

In version 8, SAS has techniques to determine the robustness of these measures of location. PROC UNIVARIATE now produces confidence limits on these measures. It also produces counts of the number of observations less than, equal to, and greater than the mean, and for all possible modes. Two robust methods for computing the mean have been added – trimmed and Winsorized. For a trimmed mean values in the upper and lower tails are dropped from the calculation. The Winsorized method calculates the mean replacing the replacing the n+1, n+2, ... with the nth largest value and i-1, i-2,... with the ith smallest value.

These are the parameters used to generate the statistics:

- alpha=.025 sets default confidence limits to 97.5%
- cibasic requests two sided confidence on the means
- cipctldf(type = asymmetric alpha = .05) requests two sided 95% asymmetric confidence limits on the quantiles
- loccount requests a table with the counts of the values greater than, equal to, and less than the mean; where the mean is the value specified by the mu0 parameter
- mu0 = 25 sets 25 as the value of the mean for the null hypothesis and table produced by loccount
- modes requests a table of all possible modes
- nexttrval = 10 requests a list of ten extreme values
- trimmed = .1 (type = upper alpha = .05) requests the a 10% trimmed mean (the top and bottom 10% of the data dropped) with a 95% upper confidence limit
- winsorized = .025 requests a Winsorized mean with values greater than the value at 97.5% and values less than the value at 2.5% be replace with those respective measures. A two sided 97.5% confidence limit is also produced.

Table 1 contains the measures of location for samples 1 and 2. The range for sample 2 is much larger than for sample 1 (175 to 12 vs. 42 to 9). The upper quartiles, medians, and lower quartiles are the

same for both samples. The mean for sample 2 is slightly larger than sample 1; which, indicates a few extreme values are affecting the mean. This is born out by the robust calculations of the means. The trimmed and Winsorized means are the same for both samples. For the trimmed mean, 101 observations are dropped from the tails. Twenty six observations are Winsorized in the tails for the Winsorized mean. The confidence limits for the mean also indicate that the larger spread is impacting the calculation of the mean (24.8 – 25.5 for sample 1 vs. 25.2 –26.9 sample 2). The confidence limits for the Winsorized means are similar (24.8 – 25.5 sample 1 vs. 24.8 – 25.6 sample 2)

	Sample 1	Sample2
Mean	25.1	26.1
Median	25	25
Mode	23	27
Trimmed Mean	25.0	25.0
Winsorized Mean	25.2	25.2
Maximum	42	175
Upper Quartile	28	28
Lower Quartile	22	22
Minimum	9	12

Table 1 Measures of Location

The location counts around the sample means are very similar for both samples. Roughly the same number of observations are less than the mean and greater than the mean.

Location Counts: Mu0=25.00
Count Sample 1 Sample 2
Num Obs > Mu0 458 461
Num Obs ^= Mu0 923 923
Num Obs < Mu0 465 462

This indicates that sample 2 has more tail weight. The mode counts (83 for sample 1 and 100 for sample 2) indicate the mode does not provide much information.

Figures 1 and 2 contain the quantiles and their 95% distribution free confidence limits along with the respective ranks in the data sets.

Quantile	Estimate	LL	UL	UR	LR	Coverage
100% Max	42					
99%	38	36	40	984	997	96.35
95%	34	33	35	937	964	95.04
90%	32	31	33	882	920	95.46
75% Q3	28	28	29	724	778	95.13
50% Median	25	25	25	470	532	95.00
25% Q1	22	21	22	223	277	95.13
10%	19	18	19	81	119	95.46
5%	17	16	18	37	64	95.04
1%	14	12	15	4	17	96.35
0% Min	9					

Fig. 1 – Sample 1 quantiles

Quantile	Estimate	LL	UL	UR	LR	Coverage
100% Max	175.0					
99%	60.5	38	130	984	997	96.35
95%	34.0	34	35	937	964	95.04
90%	32.0	32	33	882	920	95.46
75% Q3	28.0	28	29	724	778	95.13
50% Median	25.0	25	25	470	532	95.00
25% Q1	22.0	21	22	223	277	95.13
10%	19.0	18	19	81	119	95.46
5%	17.0	17	18	37	64	95.04
1%	14.0	13	15	4	17	96.35
0% Min	12.0					

Fig. 2 – Sample 2 quantiles

There is very little difference in either the quantiles or the range of the confidence limits until the 99th percentile (38 and 36 to 40 for sample 1 vs. 60.5 and 38 to 130 for sample2). This is another indicator that sample 2 has a heavier tail weight.

The ten extreme values for samples 1 and 2 are displayed in figures 3 and 4.

Extreme Values					
-----Lowest-----			-----Highest-----		
Order	Value	Freq	Order	Value	Freq
1	9	1	24	33	29
2	10	1	25	34	17
3	12	2	26	35	9
4	13	1	27	36	13
5	14	8	28	37	3
6	15	12	29	38	4
7	16	13	30	39	4
8	17	17	31	40	1
9	18	27	32	41	1
10	19	44	33	42	2

Fig. 3 – Sample 1 extreme values

Extreme Values					
-----Lowest-----			-----Highest-----		
Order	Value	Freq	Order	Value	Freq
1	12	1	30	41	1
2	13	4	31	80	1
3	14	8	32	85	1
4	15	11	33	90	1
5	16	12	34	110	1
6	17	25	35	115	1
7	18	29	36	120	1
8	19	38	37	130	1
9	20	51	38	170	2
10	21	58	39	175	1

Fig. 4 – Sample 2 extreme values

Here, we see that the nine largest values in sample 2 are two or more times higher than the highest value in sample 1. Otherwise, the extreme values are similar for the two samples. It appears that sample 2 is slightly skewed to the right, with a heavy right hand tail.

MEASURES OF DISPERSION

Measures of dispersion measure the spread or variability of the data. The most commonly used measures are the range (the maximum – the minimum), the interquartile range (upper quartile – lower quartile), the variance (squared differences from the mean), and the standard deviation (the square root of the variance). PROC UNIVARIATE now has several additional measures of scale that are not as sensitive to extreme values as the variance/standard deviation. The Gini's mean difference, the MAD, Qn, and Sn. The MAD and Sn are based on the median. The ROBUSTSCALE parameter produces a table with robust estimates of scale.

The measures of dispersion are displayed in Table 2. As can be expected based on the measures of location, sample 2 has a much larger spread than sample 1 (163 vs. 33). This is also reflected in the variances and standard deviations (134.2 and 11.6 vs. 25.6 and 5.1). However, the interquartile ranges are the same (6). This indicates that sample 2 has a heavier tail weight.

	Sample 1	Sample 2
Range	33.0	163.0
Variance	25.6	134.2
Standard Deviation	5.1	11.6
Interquartile Range	6	6

Table 2 Measures of Dispersion

The robust measures for sample 1 and sample 2 are displayed in figures 5 and 6. These estimators can be used to estimate the standard deviation if multiplied by a correction factor (see SAS *version8 Online Documentation*, "The UNIVARIATE PROCEDURE: Statistical Computations, Robust Measures of Scale.")

Estimate	Measure	Value of Sigma
Interquartile Range	6.000000	4.447805
Gini's Mean Difference	5.686372	5.039416
MAD	3.000000	4.447800
Sn	4.770400	4.770400
Qn	4.443800	4.426977

Fig. 5 Sample 1 robust measures of dispersion

Estimate	Measure	Value of Sigma
Interquartile Range	6.000000	4.447805
Gini's Mean Difference	7.455483	6.607250
MAD	3.000000	4.447800
Sn	4.770400	4.770400
Qn	4.443800	4.426977

Fig. 6 Sample 2 robust measures of dispersion

The robust estimators indicate that when values in the tails are removed, sample 1 and sample 2 have similar dispersions. All the robust estimators except Gini's Mean Difference generate the same values. However, the magnitude of that difference is much less than the comparisons of the standard deviations (11.6/5.1 vs. 7.5/5.7).

MEASURES OF SHAPE

PROC UNIVARIATE generates two measures of shape : skewness and kurtosis. The skewness statistic is the third moment around the mean. It is a measure of the tendency of the data to be larger in one direction or the other. A value of zero indicates that the data is symmetrical. Positive values indicate that the data is skewed to the right; while negative values indicate that the data is skewed to the left.

Kurtosis is estimated using the fourth moment around the mean. It measures the tail heaviness of the data set. Small values indicate that tail heaviness is not a factor.

The estimates for skewness and kurtosis for samples one and two are displayed in table 3.

	Sample 1	Sample 2
Skewness	.2	8.4
Kurtosis	.1	91.8

Table 3 – Measures of Shape

Based on the skewness and kurtosis statistics we conclude that sample one is skewed slightly to the right. The tails for sample one have little weight. Sample 2 is skewed to the right and that tail has considerable weight.

PLOTS

The PLOT option in PROC UNIVARIATE generates either a Stem and Leaf Plot (if no single interval contains more than 49 observations) or histogram (horizontal bar chart), a box and whisker plot, and a normal probability plot. The PLOTSIZE=24 tells PROC UNIVARIATE to use 24 rows to display the plot. The normal probability plot is useful when fitting distributions. Since it doesn't apply to EDA, we will not include it in our discussion.

Stem and Leaf plots and histograms provide methods to visualize the distribution of the data. Both these methods group the data around an interval. Stem and Leaf plots use the first part of the number as the stem and the second part of the number as the leaf. Therefore, the numbers 20 and 25 would be displayed as 2: 0 5. The histogram is a horizontal bar chart. SAS computes the intervals internally. Neither of these methods handles extreme observations well.

The box plot is displayed next to the stem and leaf plot / histogram. Box plots handle extreme values better than stem and leaf plots / histograms. The upper quartile is the top of the box and the lower quartile is the bottom of the box. The box's length is the interquartile range. A line represents the median, and the mean is denoted by a plus sign. Vertical lines (whiskers) projected from the box represent data within 1.5 interquartile ranges. Values between 1.5 and 3 interquartile ranges are represented with a zero. Observations outside of that range are represented with an asterisk. When a by variable is used, PROC UNIVARIATE generates side-by-side box plot comparisons for the by variable.

Figures 7 and 8 display the plots for samples 1 and 2. Figure 9 has the side-by-side box plot comparison.

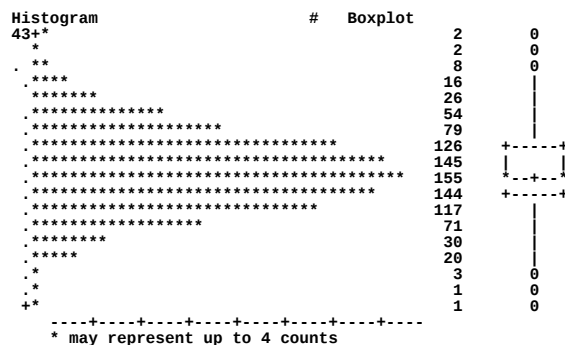


Fig. 7 Plots for sample 1

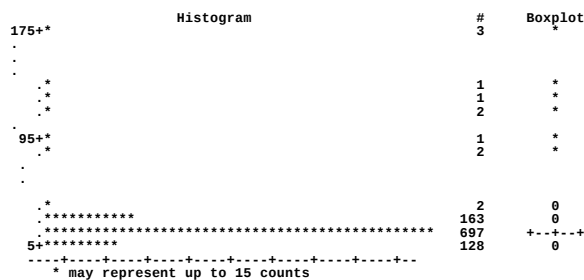


Figure 8 Plots for sample 2

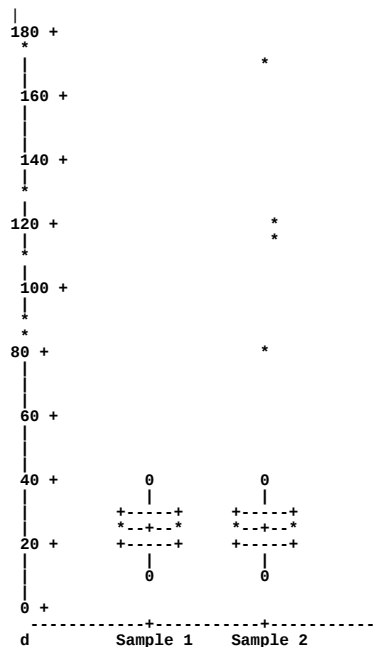


Figure 9 Box plots for sample 1 and sample 2

Figure 7 indicates that sample one is fairly symmetrical with no observations greater than 3 interquartiles. Figure 8 demonstrates the problem one can have viewing plots in PROC UNIVARIATE. The few extreme values in sample 2 have distorted even the box plot. The data sets look completely different. Sample 2 appears extremely skewed.

The plotsize option produces a larger plot for the side-by-side box plot comparison (figure 9). Viewing the side-by-side comparison indicates that sample 1 and 2 are similar. Sample 2 has a few extreme values.

CONCLUSION

PROC UNIVARIATE generates a number of statistics. Many of these are useful for fitting distributions. This paper has focused on those of interest for EDA. The statistics used here should not be considered inclusive for EDA. The examples have been provided to encourage and help users to explore their data as a first step in data analysis.

REFERENCES

SAS Institute Inc. (1999) "SAS Elementary Statistics Procedures: Statistical Background", *SAS Version 8 Online Documentation*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999) "The UNIVARIATE Procedure", *SAS Version 8 Online Documentation*, Cary, NC: SAS Institute Inc.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading Massachusetts: Addison-Wesley Publishing Company.

CONTACT INFORMATION

YOUR COMMENTS AND QUESTIONS ARE VALUED AND ENCOURAGED.
CONTACT THE AUTHOR AT:

Robert E. Maitland Jr.
Bank of America
9000 Southside Blvd, FL9-400-05-04
Jacksonville, FL 32256
Work Phone: (904) 987-3742
Fax: (904) 987-3747
Email: bob.maitland@bankofamerica.com