



Basic Graphical Representations using PROC UNIVARIATE

by Katrien Declercq

Academic

Assume we want to explore the following data on test scores of a group of students:

14 20 39 58 60 63 65 66 68 70 70 71 73 74 75 75 76 76 77 77 79 79 80 80 81 82 83 84 87 87 88
89 90 91 93 93 98

In the two previous newsletters, we discussed simple measures of location (mean and median) and spread (range, variance, C.V., interquartile range). Besides these descriptive measures, a visual representation may also be very helpful in the exploration of your data.

By simple means of the PLOT option in PROC UNIVARIATE, SAS produces a stem-and-leaf diagram, a boxplot and a normal probability plot for the variable(s) you are looking at. These are no high quality graphs, but provide you with a rough visual impression of your data.

A stem-and-leaf diagram is a kind of histogram that shows the shape of the distribution as well as the raw data. Every single observation is split in two parts: a stem and a leaf. The place where the split-up occurs depends on different factors and is determined by an internal algorithm in SAS. For the plot, all stems are put vertically (sorted) and all occurring leaves are put to the right of the correct stem (in an ordered fashion). This gives a visual impression of the distribution of the data. SAS always indicates how to convert the stem-and-leaf plot to actual data values by a legend. SAS also adds a column to the plot giving the number of observations in each line of the plot. Note that SAS produces a horizontal bar chart if more than 48 observations fall in a single interval. In the example shown here, it is clear that the lowest three observations are quite different than the other observations. The other observations show a quite symmetric distribution.

A boxplot (or box-and-whisker plot) provides information on the location (median and mean), the variability, the outliers and symmetry of the distribution of your data. The plot consists of a box that extends from the first (Q1) to the third quartile (Q3), with a line indicating the median and a plus-sign showing the mean. Outside this box, lines are drawn on either side up to the last observation that is no outlier. Possible outliers are indicated by a '0' and are those observations that are further away from the box than 1.5 interquartile ranges (Q3-Q1). Most probable outliers are those observations that are further away from the box than 3 interquartile ranges and these are indicated by an asterisk in the box plot. In the example, the median of the distribution is 77, the mean is 73.8. The box extends from 70 (Q1) to 84 (Q3) with 77 being exactly in the middle (so the middle half of the data is symmetric). Mean and median are slightly different, mainly due to the three outliers and indicating the data on the whole is not perfectly symmetric). The observations 14 and 20 are indicated by an asterisk as most probably being outliers and 39 is indicated by '0' as possible outlier. Besides these outliers, the distribution seems to be quite symmetric.

A normal probability plot is a visual method for determining whether or not your data come from a distribution that is approximately normal. This may be useful to know, since a whole series of statistical tests is based on the assumption of normality. The vertical axis represents the actual data values whereas the horizontal axis shows the expected percentiles if your data came from a normal distribution. SAS provides a normal probability plot with plus signs and asterisks, where the plus signs represent where the data values would fall if they came from a normal distribution, whereas the asterisks represent the observed data values. If the asterisks follow a fairly straight line and cover up many plus signs, you can conclude that there does not appear to be any severe departure from normality. In the example, however, this is not really true, so the data don't seem to follow a normal distribution.

If you are using a BY statement in the UNIVARIATE procedure, SAS will do the analysis by group and will also produce a schematic plot of the box plots by group, which is very useful to compare the groups visually in terms of location, spread and symmetry.

The following code may be used for analyzing the example:

```
data example;
  input score @@;
  cards;
  14 20 39 58 60 63 65 66 68 70 70 71 73 74 75 75
  76 76 77 77 79 79 80 80 81 82 83 84 87 87 88 89
  90 91 93 93 98
  ;
run;

title 'Data on Test Scores';
proc univariate data=example plot;
  var score;
run;
```

New statements for histograms, probability plots and quantile-quantile plots are available for PROC UNIVARIATE in SAS V8 and produce high resolution plots. The HISTOGRAM statement allows to plot a histogram and superimpose several theoretical distributions and the PROBLOT and QQPLOT statements allow to obtain visual tests for several distributions (normal probability plots for normal distribution, but many other distributions may be chosen)!