Summarizing data using mean or median

by Katrien Declerq

Assume we have the following data on the test scores of 30 students who took a programming class:

56 78 84 73 90 44 76 87 92 75 85 67 90 84 74 64 73 78 69 56 87 73 100 54 81 78 69 64 73 65

If we want to summarize these data in one number, we can use different measures of which the most frequently used are the mean and the median. In this example, the mean is 74.6 and the median is 74.5. The mean is obtained by summing up all observations and dividing by the number of students and the median represents the middle observation if there is an odd number of students and the mean of the two middle observations if there is an even number of students, after ordering all observations. In this case, mean and median are very similar and thus give the same information, but this is not always the case. The mean is very sensitive to outliers (observations that are very different from all the others).

If all students had the same test scores except for the one with the lowest score, who had 0 instead of 44, the mean would be 73.2, whereas the median remains unchanged as 74.5. The influence of one or more outliers on the calculation of the mean can even be more dramatic as shown in the following small example. Assume there are ten students of whom nine have a score of 100 and one has a score of 0. The mean score is 90 whereas the median is 100. In the case of outliers, the mean may give misleading information and it is better to use the median instead.

In SAS, PROC UNIVARIATE automatically produces a list of descriptive statistics. In the section on “moments”, the mean can be found (and equals the quotient of the sum of the observations and the number of observations). The median can be found in the section on “Quantiles” as the 50% quantile.

The following program illustrates the first example described.

```
data example;
  input score @@;
  cards;
  56 78 84 73 90 44 76 87 92 75 85 67 90 84 74
  64 73 78 69 56 87 73 100 54 81 78 69 64 73 65
run;

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