



Academic

## Measures of spread

by Katrien Declercq

Assume we have the following three sets of observations:

(A) 0 20 40 50 60 80 100

(B) 0 48 49 50 51 52 100

(C) 0 1 2 50 98 99 100

As discussed in our previous newsletter, these three sets of observations may be summarized in one measure by calculating their mean and median, which are both 50 for all three datasets. However, the three datasets look quite different: the way the data are spread around their central value of 50 is quite different between the three sets. There are different statistical measures that describe this spread.

The simplest measure to describe spread in datasets is the range, which is the difference between the largest and smallest observation. For the example above, we obtain 100 for all three datasets. So, using the range, we would conclude that all three datasets show the same amount of variability. This doesn't really reflect the impression we get by looking at the three datasets. Note that the range is also very sensitive to outliers.

A better measure of spread is the (sample) variance which is obtained by subtracting the sample mean from each individual observation, squaring this difference, adding these differences for all observations and dividing this total by the number of observations minus 1. So, the further an observation is away from the mean, the larger its contribution to the variance. This means that datasets that show a large spread around the mean will have large values for the sample variance, compared to datasets with a smaller spread. In the example datasets, the variances are respectively 1166.67, 835 and 2401.67. Immediately linked to the variance is the standard deviation, which is the square root of the variance and expresses the spread of the data in the same units as the data. Again large values correspond to large variability. For the example, we observe as standard deviations 34.16, 28.90 and 49.01.

In order to be able to compare the variability of observations that are on different scales (e.g. like body weight and height in a group of 10-year olds), one can use the coefficient of variation (C.V.), which is obtained by expressing the standard deviation as a percentage of the (sample) mean. If the standard deviation of the weight in a group of 10-year olds were 5 kg on a mean of 20 kg, and the standard deviation of the height were 9 cm on a mean of 150 cm, it is quite hard to tell which of both variables shows the largest variability by comparing the variances (or standard deviations). This can be answered by comparing the C.V., which is 25% for the weight and 6% for the height in the group of 10-year olds. So, the measurement of the weight shows much more variability compared to the height in this group (fictive example).

Another approach to describe the spread in datasets comes from the description of the distribution of data using percentiles (or quantiles). For example the 40th percentile is that value such that 40% of the observations are smaller than or equal to that value (so 60% of the observations are larger). The most regularly used percentiles are the 25%, 50% and 75% percentiles, which are also called the first quartile (Q1), the median (Med) and the third quartile (Q3). Note that half of the observations are situated between the first and the third quartile. The interquartile range, which is the difference between the third and the first quartile, is then another regularly used measure of spread in data. In our example with the three datasets, we obtain 60, 4 and 98. This gives us the same message as the comparison of the variances and standard deviations.

In SAS, PROC UNIVARIATE automatically produces a list of descriptive statistics. In the section on “moments”, the variance and standard deviation are listed. In the section on “Quantiles”, the maximum and minimum observation are listed (Max (100% percentile) and Min (0% percentile)), as well as Q1 (25%) Med (50%) and Q3 (75%). Besides these percentiles, SAS also gives by default the 1%, 5%, 10%, 90%, 95% and 99% percentiles. At the bottom of the “Quantiles” table, SAS shows the range (Max – Min) and the interquartile range (Q3 – Q1).

The following program illustrates the example with the three datasets.

```
data example;
input set $ obs @@;
cards;
A 0 A 20 A 40 A 50 A 60 A 80 A 100
B 0 B 48 B 49 B 50 B 51 B 52 B 100
C 0 C 1 C 2 C 50 C 98 C 99 C 100
;
run;

title1 'Measures of Spread in PROC UNIVARIATE';
proc univariate data=example;
  var obs;
  by set;
run;
```