Tidsskriftet Tidsskrift for Den norske legeforening

Mean and standard deviation or median and quartiles?

MEDICINE AND NUMBERS

STIAN LYDERSEN

E-mail: stian.lydersen@ntnu.no
Stian Lydersen dr.ing. and professor of medical statistics at the Regional Centre for Child and Youth Mental Health and Child Welfare (RKBU Central Norway) at the Department of Mental Health, Norwegian University of Science and Technology.
The author has completed the ICMJE form and declares no conflicts of interest.

Mean and standard deviation are frequently used measures of central tendency and variability in data from scale variables. If data are not normally distributed, some researchers prefer reporting median and quartiles instead. But the mean and standard deviation have useful properties and can be relevant also when data are not normally distributed.

Let us first consider the normal distribution, which is shown in Figure 1. If data are normally distributed, approximately 16 % of the observations will be lower than one standard deviation under the mean, and approximately 84 % of the observations will be lower than the mean plus the standard deviation. So, for normally distributed data, the standard deviation will be related to the 16^{th} percentile and the 84^{th} percentile. How about the median and the quartiles? Since the distribution is symmetric, the median will equal the mean. The quartiles are defined as the 25^{th} percentile and the 75^{th} percentile. Hence, for the normal distribution, these define a narrower interval than does one standard deviation on each side of the mean.

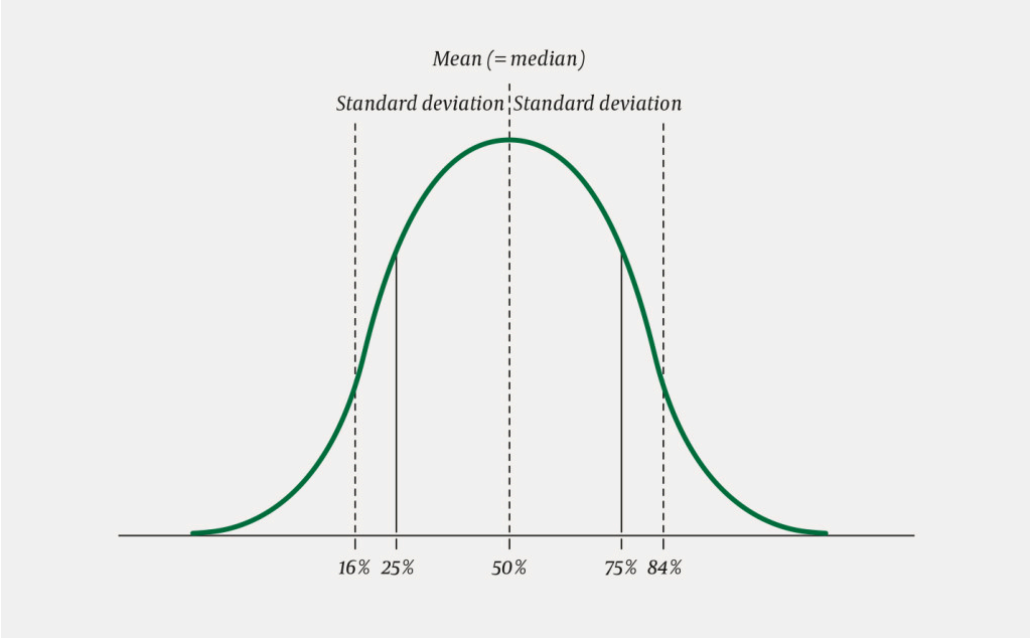

Figure 1 The normal distribution, with mean (=median), standard deviation (SD), and quartiles (25 % and 75 %).

Skewed distributed data

Figure 2 shows a right-skewed distribution. Such distributions may originate from measurements which cannot be negative, for example plasma concentration. In a right-skewed distribution, the mean will be larger than the median, and the standard deviation is not related to specific percentiles, as was the case for the normal distribution.

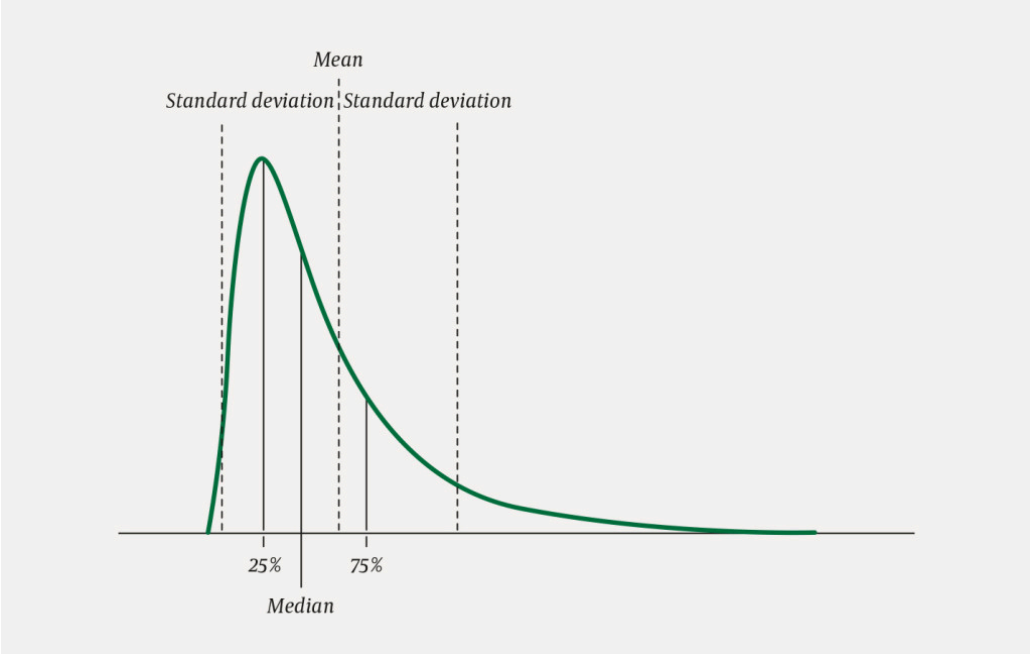

Figure 2 A right skewed distribution, with mean, standard deviation (SD), median, and quartiles (25 % and 75 %).

What are relevant measures of central tendency and variability if data are not normally distributed? The mathematical expressions for mean and standard deviation are defined regardless of whether data are normally distributed. Now, consider a fictitious data set, from (1): Assume we have recorded the length of stay in hospital for 13 patients with a given diagnosis: (respectively 3, 9, 10, 10, 10, 12, 13, 14, 18, 21, 27, 38, and 62 days). The average is 19 days, and the median is 13 days. The standard deviation is 15.8 days, and the quartiles are 10 days and 24 days. If we intend to estimate cost or need for personnel, the mean is more relevant than the median. If we want to state a 'typical' length of stay for a single patient, the median may be more relevant.

Some authors report only the interquartile range, which is 24–10=14 days in this example, instead of the quartiles. This is less informative than reporting the quartiles, which in combination with the median also indicate the skewness of the distribution. In our example, the median of 13 days is closer to the lower quartile of 10 days than the upper quartile of 24 days, indicating a right-skewed distribution, similar to the one illustrated in Figure 2. In some contexts, it may be useful to report the minimum and maximum values instead of, or in addition to, the quartiles. But one should bear in mind that unlike the interquartile range, the distance between minimum and maximum is expected to increase with increasing sample size.

What should be reported?

Which measures should be reported when data are not normally distributed? One criterion may be the relevance in the application at hand, such as length of stay in hospital. But how about descriptive statistics for background data in a study? Some researchers claim that it is generally wrong to report mean and standard deviation when data are not normally distributed. Such a point of view is difficult to defend. These measures are not only well defined for all types of distributions; they are also the measures needed for summarising data, for example in future meta analyses. This is a good reason for reporting mean and standard deviation for scale variables, also when they are not normally distributed, and one can report median and quartiles in addition, when relevant.

When data are categorical with few categories, for example the values 1, 2, 3, and 4, the median and quartiles will not be suitable for describing the distribution. We will discuss this in the next article in this column.

LITERATURE

1. Skovlund E. Bootstrapping – å løfte seg selv etter håret? Tidsskr Nor Legeforen 2019; 139. doi: 10.4045/tidsskr.19.0413. [PubMed][CrossRef]

Publisert: 11 June 2020. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.20.0032