Making Sense of Statistics for Family Practitioners

"The Spread"

Quantitative data is traditionally summarised by two measures: a measure of the centre (average) of a data set and one of spread around this central point. In the previous article we noted that there are four useful measures for depicting the centre of a set of data namely the "arithmetic mean", "median", "mode" and "geometric mean".

For the spread of a data set around the central point, there are three measures that are commonly used viz. the range, inter-quartile range and standard deviation.

The ***range*** is the difference between the highest and lowest values in a data set and is a measure of variability that is not very useful since it can be greatly influenced by one extremely large or small value. It is obvious that this measure is completely dependent on only two values and if these are extreme, or *outliers,* then the range will provide an inaccurate summary of the dispersion of data. As an example, the time delay (in days) between onset of malaria symptoms and seeking assistance from the formal health sector for thirty-eight patients who subsequently died of malaria in Mpumalanga Province during 1996 were:

0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,4,4,4,6,7,11,12,14,16,20,21 days.[1] The range in this example is 21 - 0 = 21 days and although it provides an idea of the extreme spread of delays, the more common scatter around the centre point is not captured as the range is dominated by one value viz. 21 days.

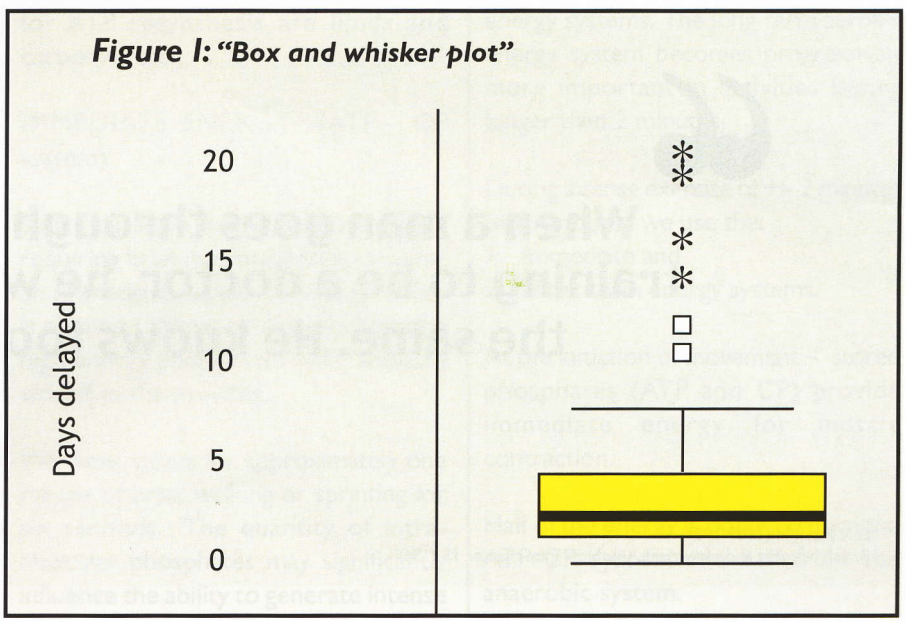
The ***inter-quartile range*** is a similar measure, but it takes into account the spread of the central data values. It measures the numerical difference between the values that occupy the quarter (or 25th centile) and three-quarter (or 75th centile) marks of a data set, i.e. 75th centile minus 25th centile that equals the middle 50 percent.[2] Although appropriate statistical software packages easily generate this value, it is useful to understand that all centiles are derived in a similar way to the median, which is also the 50th centile. There are as many values less than the 25th centile, as there are between the 25th centile and the median, between the median and the 75th centile and greater than the 75th centile. Where a data set such as the one above, is so small that a histogram is not meaningful, then a great deal can be learnt about that data set by constructing a "box-and-whisker" plot. The "box and whisker" plot is the summary plot based on the median, quartiles, and extreme values.

The box represents the interquartile range that contains the middle 50% of values. The whiskers are lines that extend from the box to the highest and lowest values, excluding outliers and a line across the box indicates the median". In this case, it is a rectangle that extends from the 25th centile to the 75th centile, with lines ("whiskers") drawn from the smallest value (0) to the 25th centile and from the 75th centile to the highest value excluding the outliers (7) (see Figure 1).

The "box-and-whisker" plot provides an immediate visual indication of the skewness of a set of data. If the median is near the centre of the rectangle and the "whiskers" are of similar length, then the data is "symmetrical". If the median is appreciably off-centre or the whiskers are markedly different in length, as in this example, then the data is "skewed". Visualisation is very important before deciding what statistical tests should be used, as we will see later in this series.

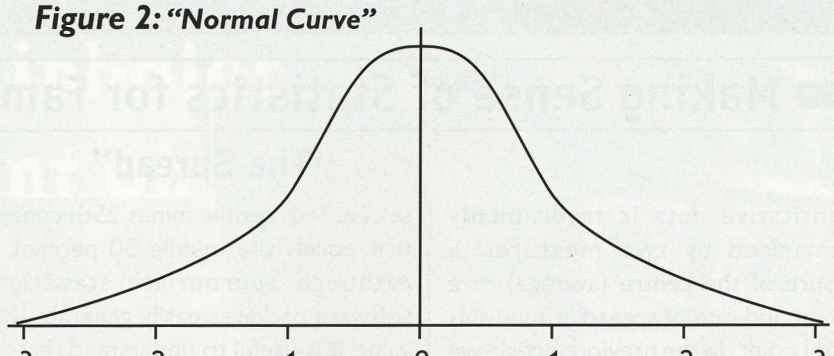
The ***standard deviation*** is the most useful measure for quantifying the spread of biological measurements around the arithmetic mean. It gives an indication of the average distance from the mean. If values are closely gathered about their mean, then there is little "dispersion" of a data set, and vice versa. It is therefore logical that we should measure the variation in a set of data by the amount with which the data varies from the mean. To derive the standard deviation, the distance of every data item from the mean is calculated and then these

Figure 1: ***"Box and whisker plot"***



differences are averaged. You will appreciate that just to consider an average difference is meaningless. Some values are greater than the mean (positive) and others are less than the mean (negative) and so the differences will cancel each other out with the average difference from the mean being zero (because the mean by definition is central). Since we are actually interested in the amount of spread, and not whether it is negative or positive, we work with the squares of the distances from the mean, which effectively transforms all negative values into positive measures. This is done by averaging the squared differences from the mean and then taking the square root of the result to get the standard deviation. The standard deviation is one of the most important statistics routinely calculated for a data set and is extensively used, particularly when determining required sample sizes or confidence intervals of a mean calculated from a sample. It is of interest to note that 68% of values in a large symmetrical data set(normal curve) lie within one standard deviation of the mean, while 95% lie within two standard deviations of the mean and 99% lie within three standard deviations of the mean (Figure 2). Fortunately most statistical software packages will also perform this convoluted procedure at the touch of a key. But, for all those family physicians that may feel offended by us coyly concealing the mathematics, a simplified formula for calculating the standard deviation for a sample is:

Figure 2: "Normal Curve"



$$s = \sqrt{\left[\frac{\sum x^2 - \left(\sum x\right)^2 / n}{(n-1)}\right]}$$

Where s is the "sample standard deviation", sigma is "the sum of", x is each data value, and n is the number of values in the data set.

What are the important messages we want you to remember from this article?

- The ***range*** is not very useful as it maybe greatly influenced by one extremely large or small value.
- The "box-and-whisker" plot is an immediate visual indication of the skewness of a small set of data and adds more meaning to the ***inter-quartile range***.
- The ***standard deviation*** is the most useful measure for quantifying the dispersion of a data set symmetrically spread around the arithmetic mean.

References

1. Dürrheim DN, Fieremans S. Profile of patients dying with *Plasmodium falciparum* malaria in Mpumalanga. *S Afr J Epidemiol Inf* 1999; 14: 24-25.
2. Altman DG, Bland JM. Quartiles, quintiles, centiles and other quantiles. *BMJ* 1994; 309: 996.

David N. Durrheim *MBChB, DTM&H, DCH, FACTM, MPH &TM*
Consultant: Communicable Disease Control, Department of Health, Mpumalanga

Gboyega A. Ogunbanjo *MBBS, MFGP(SA), M FAM MED (MEDUNSA)*
Principal Family Physician & Senior Lecturer, Department of Family Medicine, MEDUNSA

"When a man goes through six years' training to be a doctor, he will never be the same. He knows too much."

Enid Bagnold (1889-1981)
British novellist and playwright. Autobiography, chapter 15 (1969)