# Making Sense of Statistics for Family Practitioners

## "Understanding median, mode and means"

In our previous article in this series, we considered categorical data and noted that this form of data was usually summarised using proportions. The other major data-type is quantitative data. Quantitative data may either be counted as whole numbers (*discrete* data), e.g. the number of cases of meningococcal meningitis in an outbreak on a mine, or be measured using a *continuous* scale, e.g. temperature measurement of a patient with brucellosis.

Quantitative data is traditionally summarised using two measures: one to indicate the centre of the data set while the other reflects the degree of spread about this central point. The four most popular measures used for reflecting the centre of a set of data are the ***average*** (arithmetic mean), ***median***, ***mode*** and ***geometric mean***.

Medical phenomena tend to be congregated about one point with symmetrical spread of data on either side of this point. In this situation the commonly used measure for centrality is the average or *arithmetic mean*. This is calculated by adding all values and then dividing by the total number of values. Take for example the *incubation periods (the time between infection and the onset of malaria symptoms)* in days for twenty South Africans fed on by infectious *Anopheline* mosquitoes: 7, 8, 9, 9, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 13, 14, 18, 25, 80. The ***"arithmetic mean"*** of the data above would be: (7+8+9+9+10+10+10+...)/20 = 299/20 = 14,95 days. The arithmetic mean is amenable to many statistical techniques, as we shall see later in this series. One of the advantages of the arithmetic mean is its consistency i.e. despite the variation in the individual values of a characteristic in samples from the same population, the means for the samples tend to be similar in value. However, one major disadvantage is that the mean – depending on the sample size, can be affected greatly by an extreme value, which differs from the other values.[1]

The second most commonly used measure is the ***"median"***. This is the value that divides a data set in half. One determines the median value by ranking all values in a data set from smallest to largest and then locating the value in the position equal to one plus the total number of values, divided by two. If there are an even number of values, then the median will fall half way between the two central values i.e. their average. The median is of particular value if there are a few extremely high or low values, as the arithmetic mean would be misleading in this situation. This is certainly the case in the data set above where most values are clustered around 10 days but there are a few that are much higher viz. 25 and 80 days. These extreme values are biologically quite plausible for malaria incubation periods due for example to the effect of partial immunity or the use of prophylactic medication.[2] In this case, with 20 observations, the median would be the (20 + 1)/2 observation or the observation halfway between the 10th and 11th value. The 10th value is 10 days and the 11th value is 11 days, therefore the *median* of this data set is 10.5 days. Note the difference between the median and the arithmetic mean. The few extreme values result in a misleading arithmetic mean as an "average" measurement. The main advantage of the median is the lack of impact of extreme values on it, while its disadvantage is that it is not as amenable to statistical testing as the mean.

> "Transformation is an important subject in Medical Statistics"

The ***"mode"*** is simply the value that occurs most often in a data set. If a data set appears to have two modes, then one should suspect that there might actually be two different groups represented in the data. In our example the most common value (mode) is 10 days (six observations). The mode is infrequently used in journal articles, as it does not provide as much information as the mean or median, and frequently there is no particular value in a sample that occurs more often than all other values.

The final measure of centrality that we will consider is the ***"geometric mean"***. The geometric mean is particularly useful in situations where most values are congregated at the lower end of a measurement scale but a few values are scattered towards the higher end of the scale (*positively skewed*), such as in this example. The geometric mean is calculated by taking the logarithm of each value and calculating the arithmetic mean of these *transformed* values (by adding them together and dividing by the total number of observations). One then antilogs (or "back-transforms") this mean value to get the geometric mean.[3] In this case taking the logarithm of the twenty incubation periods we get: 0.85, 0.90, 0.95, 0.95, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.04, 1.04, 1.04, 1.04, 1.08, 1.11, 1.15, 1.26, 1.40, 1.90. Their arithmetic mean is: (0.85+0.90+0.95+...)/20 = 21.72/20 = 1.086. Therefore the geometric mean by antilog is: 12.2. Note that the geometric mean is closer to the median value than to the arithmetic mean. Transformation is an important subject in medical statistics and will be dealt with in more detail later in this series. In this case it can be seen that the logarithmic transformation removed most of the "skewness" even in a very small set of data.

In conclusion, we hope that the reader now understands these useful measures, so that reading and interpreting results of published articles containing such information will be more meaningful.

## References


1 Lilienfeld DE, Stolley PD. *Foundations of Epidemiology*. Oxford University Press, New York. 1994: 293 – 295.

2 Durrheim DN, Ogunbanjo G A, Blumberg L. Malaria - prevention, recognition and cure. *S A Family Practice* 1996; 17: 367-374.

3 Kirkwood BR. *Essentials of Medical Statistics*. Blackwell Scientific Publications, Oxford. 1988: 141-142.



***David N. Durrheim*** *MB, ChB, DTM&H, DCH, FACTM, MPH &TM*
*Consultant: Communicable Disease Control, Department of Health & Welfare, MPUMALANGA*

***Ogunbanjo GA*** *MB, BS, MFGP(SA), M FAM MED (MEDUNSA),*
*Principal Family Physician and Senior Lecturer, Department of Family Medicine, MEDUNSA*


***Table I:*** ***Incubation periods of falciparum malaria in 20 South Africans***