The language of research (part 13): research methodologies: descriptive statistics — measures of central tendency

KEY WORDS
- Substantive theory
- Grounded theory
- Interview
- Qualitative
- Sociology

PETER ELLIS
Nursing Director, Hospice in the Weald, Pembury, Tunbridge Wells

Quantitative studies that use numbers rather than words usually report their findings using one of two forms of statistics (Ellis, 2014). These are known as descriptive statistics, which give some idea of overall numbers, spread of numbers and variability; and inferential statistics, from which conclusions, inferences, are drawn and p values and the like calculated that are more complex to calculate and understand.

The most frequently used and easily interpreted of the descriptive statistics are 'measures of central tendency'. Measures of central tendency present figures that are different types of averages from the data collected (Ellis, 2016). Measures of central tendency are the arithmetic mean (what many of us refer to as the average or plain mean), the median, the mode and variance (usually recalculated as the standard deviation, SD).

Each of these measures is calculated in a different way and each provides information that informs our thinking and understanding of what we have read and how it might apply to where we work and our patients This paper will explore each of these measures, including an example of how they are each calculated.

THE ARITHMETIC MEAN

When using the term arithmetic mean, most people are talking about the average. This has various other names and denotations in statistics; it is simply the mean and is denoted by the symbol $\bar{x}$ also called x-bar. The arithmetic mean is calculated by adding all of the numbers in the data set together (the sum of all numbers, sometimes denoted as Σ, which is also called sigma) and dividing by the number of observations (usually denoted as n). It is worth noting that n is almost always used to denote the word 'number' (so if ten people respond to a questionnaire and six say they are happy with a service, n is 10 for the total number of people replying, but n is 6 for those satisfied with the service).

The numbers could represent anything; for this example, it is the satisfaction score given to a clinic out of **10**: So the mean of: **8, 10, 7, 3, 4, 9, 5, 7, 2** is:

8+10+7+3+4+9+5+7+2

Number of observations (n) is **9**

55 / 9 = 6.11

The arithmetic mean score given to the clinic is:

6.11

The mean uses all the observations in the data set and each observation affects the mean. The mean is sensitive to extreme values — that is to say, extremely large or small data points can cause the mean to be pulled toward the extreme data.

So for our example, if we replace the first **8** with a **58** we get:

58+10+7+3+4+9+5+7+2= 105

Number of observations (n) is 9

105 / 9 = 11.66

Here, one extreme value (**58**) has almost doubled the value of the arithmetic mean.

The mean has valuable mathematical properties that make it convenient for use with inferential statistics analysis. For example, the sum of the deviations of the numbers in a set of data from the mean is zero. We will return to this later in the series when we look at inferential statistics.

THE MEDIAN

The median is the middle value in an ordered (and this is important) array of observations (if there is an even number of observations in the array, the median is the arithmetic mean of the two middle numbers). In many cases, the median is a superior measure of central tendency over the arithmetic mean; this is especially true when there are some extremely large or small observations in the data set.

Our data set ordered by size is: **2, 3, 4, 5, 7, 7, 8, 9, 10**; the middle number (the median) is **7**.

With the same extreme value added to the data set, removing one of the sevens (ordered by size): **2, 3, 4, 5, 7, 8, 9, 10, 58**; the middle number is still **7** and so the measure of central tendency in this case (the median) is not affected by the extreme value.

The median is widely used in predicting survival. For example, when we say median survival with x disease is 5 years, we are actually saying 50% of people diagnosed with x will die

before 5 years and 50% will live beyond 5 years. In essence, the median can be thought as the mid-way point in a data set.

It is worth noting that the median has no further statistical properties and is therefore not used in inferential statistics.

THE MODE

The term mode applies to the number that occurs most frequently in a data set; in this case, the word mode is like the use in à la mode (the most fashionable). Where two figures occur with the same level of high frequency, the data set is aid to be bi-modal. Where two or more figures occur with the same amount of frequency, the data is aid to be multimodal. For example, if we were to measure the low-density lipoproteins (LDL) cholesterol levels in patients with diabetes (in mmol/l) and we got the results:

4.1, 4.5, 4.6, 5.0, 3.4, 3.5, 4.1, 3.9, 4.2 and 3.8, the mode would be 4.1 mmol/l, as this figure occurs twice and no other figure occurs more than once.

If we took the same data: 4.1, 4.5 4.6, 5.0, 3.4, 3.5, 4.1, 3.9, 4.2 and 3.8, and added another reading (3.9), then the date would be bi-modal at 4.1 and 3.9 mmol/l.

VARIANCE AND THE STANDARD DEVIATION

Variance is the average (mean) of the squared deviations from the arithmetic mean of the numbers in a set of data. The following five steps are used to calculate the variance:

- Find the arithmetic mean
- Find the difference between each observation and the mean
- Square these differences
- Sum (add together) the squared differences
- Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e. n-1 (where n is equal to the number of observations in the data set)

So, for example:

- Find the arithmetic mean: **7, 5, 4, 6, 7, 11, 5, 9, 10, 8 = 6.8**
- Find the difference between each observation and the mean: **0.2, -1.8, -2.8, -0.8, +0.2, +4.2, -1.8, +2.2, +3.2, +1.2**
- Square these differences: **0.4, 3.24, 7.84, 0.64, 0.4, 17.64, 3.24, 4.84, 10.24, 1.44**
- Sum the squared differences = **49.92**
- Divide this number by the number of observations minus one = **5.55**

Variance is hard to understand because the deviations from the mean are squared, making it too large for logical explanation. These problems can be solved by working with the square root of the variance, which is called standard deviation.

In our example, therefore, the standard deviation is the square root of **5.55** = **2.345**

Standard deviation gives a good feel of how widely spread around the arithmetic mean the data are in a data set; this is perhaps most important in large data sets that contain many numbers, so the reader does not have to look at each number individually. The Empirical Rule, which generally applies to data that are normally distributed, i.e. bell-shaped states that: approximately 68% of the measurements (data) will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 97.7% (or almost 100%) will fall within three standard deviations (University of Utah, 2017). Knowing this will help us to understand where the majority of data lie in a very large data set.

CONCLUSION

Understanding some of the basic rules of the use of statistics to present data is helpful to the nurse when reading research and other papers. Readers are encouraged to look at some data and undertake their own calculations, as in doing so, the nature of the measure of central tendency and the usefulness of it in understanding data becomes more apparent. WUK

REFERENCES

University of Utah (2017) *Descriptive Statistics*. Available at: https://www.che.utah.edu/~tony/course/material/Statistics/12_descriptive.php (accessed 1.06.2017)

Ellis P (2014) Decoding Science: the Language of research (part 1): Research Paradigms. *Wounds UK* 10(2): 118–9

Ellis P (2016) *Understanding Research for Nursing Students. 3rd edn*. London; Sage