

## Biostatistics: DESCRIPTIVE STATISTICS: 2, VARIABILITY

### 1. Introduction

Besides arriving at an appropriate expression of an average or consensus value for observations of a population, it is important to quantify the amount of variation within the sample – commonly called either the '**variability**' or '**dispersion**'. Consider three datasets:

Set 1	3.9	4.2	4.4	4.5	4.7	4.9	5.1
Set 2	2.3	2.5	3.4	4.5	6.7	7.4	9.3
Set 3	2.3	3.4	6.7	6.9	7.4	8.1	9.3

You will see that Sets 1 and 2 have identical median values (shaded column). However, the lowest value in Set 1 is 3.9 and the highest value is 5.1 (a difference of 1.2), whilst for Set 2, the corresponding values are 2.3 and 9.3 (difference of 7.0). Set 2 is **more variable** than Set 1, even though the median value is the same for each.

Set 3 has the same minimum and maximum values as Set 2 – by simple measures the variability of the two datasets is the same. However their median values differ. If we were comparing the datasets, we would probably consider that the variability in the two is too large to accept that the differences in median values is meaningful.

This helpsheet contains references to the companion helpsheet on descriptive statistics, which covers different measures of averages, such as means and medians. Please refer to that helpsheet if you do not understand these concepts. All measures of variability covered here are appropriate for scale data (continuous or discrete), some are appropriate for ordinal data, but none can be used with nominal data.

### 2. Range

For a set of observations, the simplest measure of variability is the range - the difference between the smallest and largest values. Consider the datasets introduced earlier. For Set 1, the maximum value,  $x_{\max} = 5.1$ , and the minimum value,  $x_{\min} = 3.9$ . The range is given by:

$$\text{range} = x_{\max} - x_{\min} = 5.1 - 3.9 = 1.2$$

Note that the range is a quantity, a measure of variability, so does not have a sign – there is no such thing as 'negative' variability. The equation given here still applies to instances where one or both values are negative. For instance, if all of the values in Set 1 were negative,  $x_{\max} = -3.9$  and  $x_{\min} = -5.1$ .

$$\text{range} = x_{\max} - x_{\min} = -3.9 - (-5.1) = -3.9 + 5.1 = 1.2$$

Looking at our three example datasets and some others:

	$x_{\min}$	$x_{\max}$	Range	Median	Mean
Set 1	3.9	5.1	1.2	4.5	4.53
Set '1A'	-5.1	-3.9	1.2	-4.5	-4.53
Set '4'	-5.1	3.9	9		
Set 2	2.3	9.3	7	4.5	5.16
Set 3	2.3	9.3	7	6.9	6.3

The range is suitable as a measure of variability for scale (continuous as in the examples here, or discrete) or ordinal data.

### 3. Interquartile range

The range of observations within a dataset is the simplest measure of variability. Look at another dataset:

Set 5	2	3	5	6	6	7	9	9	13	28	50
-------	---	---	---	---	---	---	---	---	----	----	----

These are discrete scale data. There are eleven observations, and the range is given by:

$$\text{range} = x_{\max} - x_{\min} = 50 - 2 = 48$$

Notice that the nine lowest values are all under ten, and that the range is 'inflated' by the three very high values. This can be seen by comparing the median value of 7 with the mean value of 12.55. Extreme values, often called '**outliers**', will make the range larger, and if there is only a few of them they cause an exaggerated view of the variability. Note that outliers could be atypically large, atypically small or both.

You have already seen the use of the median value – simply the middle value of an ordered data series. Calculation of the interquartile range takes this principle further by finding the 'medians' of the value groups either side of the median. So for Set 5:

Set 5	2	3	5	6	6	7	9	9	13	28	50
-------	---	---	---	---	---	---	---	---	----	----	----

The median of this set of observations is the middle value, observation 6. The **lower quartile** is the midpoint of the group below the median, in this case observation 3. The **upper quartile** is the midpoint of the group above the median, here observation 9.

The interquartile range is calculated simply by subtracting the value of the lower quartile from the value of the upper quartile:

$$\textit{interquartile range} = \textit{upper quartile value} - \textit{lower quartile value}$$

In this case, the interquartile range is  $13 - 5 = 8$ . Even if the highest value had been 500, or the lowest value -500, the interquartile range would have been the same if the other values had remained as shown. As with the range, the value of the interquartile range is always a positive number.

Interquartile range can be used with scale or ordinal data, but not with nominal data. It is an appropriate measure of variability to use when the median is used to describe the population average, especially where the distribution of observation values is skewed as in this example.

#### 4. Variance

Whilst range and interquartile range are defined simply by particular points in the spread of data, neither measure captures the contribution of all data points to the overall variability. The quantity called '**variance**' is based on the difference between the value for each observation and the mean of all of the observations. It can only be used with scale data, and would normally be associated with a mean as the description of the population average.

The calculation of variance,  $s^2$ , can be represented by the word equation:

$$s^2 = \textit{sum of squared deviations from the mean} / \textit{degrees of freedom}$$

The deviation of a data point from the mean is simply written as:

$$(x - x_{\text{mean}})$$

where  $x_{\text{mean}}$  represents the mean value of all observations. If a data point is similar in value to the mean, the deviation will be small, whilst a large deviation arises from a data point that is very different from the mean. The squared deviation is written as:

$$(x - x_{\text{mean}})^2$$

Note that whilst a deviation can be negative, its square is always positive. So when we add the deviations together, they reflect the differences between all points and the mean – positives and negatives don't cancel out. The sum of squared deviations is written as:

$$SS_d = \sum (x - x_{\text{mean}})^2$$

where the symbol  $\Sigma$  represents the sum, in this case of all of the deviations calculated for all data points.

The denominator in the variance word equation was called the '**degrees of freedom**'. This expression is common in statistics, and is related clearly to the number of

observations or categories. In the case of variance, the number of degrees of freedom is simply the number of observations minus one:

$$df = (n - 1)$$

So we can re-write the word equation in terms of the two expressions above:

$$s^2 = \frac{SS_d}{df}$$

or:

$$s^2 = \Sigma(x - x_{\text{mean}})^2 / (n - 1)$$

Looking at the formula, you will see that a high average deviation, or a few very high values, will increase variance. For large samples, the number of degrees of freedom is very similar to the number of observations ( $df \approx n$ ), and the variance is almost the same as the average squared deviation from the mean. However, for very small samples the number of degrees of freedom is significantly smaller than the number of observations ( $df \ll n$ ), so that the variance is larger than the average squared deviation from the mean.

You can use a spreadsheet to calculate variance, or it can be generated using a statistical software package such as SPSS. The purpose of describing the calculation here is simply to help you to understand how it works.

## 5. Standard deviation

Variance is a measure of variability or spread in a set of observations that uses the square of the deviation (difference from the mean value). Whilst variance is a measure of overall variability, it is difficult to relate it directly to the observations, as is possible for the range, because the calculations are based on squares of deviations.

The '**standard deviation**' is given simply by the square root of the variance, and has the same units the observations:

$$s = \sqrt{s^2} = \sqrt{\{\Sigma(x - x_{\text{mean}})^2 / (n - 1)\}}$$

or using power notation:

$$s = \{\Sigma(x - x_{\text{mean}})^2 / (n - 1)\}^{0.5}$$

The square root of a number has two values - a positive and a negative ('minus times minus equals plus'). So the typical way to write the standard deviation is in association with the mean as in:

$$\text{Mean } (\pm 1 \text{ SD}) = 11.34 \pm 1.06 \text{ mm}$$

This indicates that the mean value for a series of observations of lengths (of something) is 11.34 mm, and that the standard deviation is 1.06 mm either side of the mean (ie 10.28 – 12.40 mm).

As with variance, it is only appropriate to use standard deviation with scale data, and there is an underlying assumption that data are distributed normally about the mean (the interval bounded by the standard deviation is symmetrical about the mean).

## 6. Standard error and confidence intervals

If we wished to find the overall mean value for a population, we could take several subsamples and calculate an independent value for the mean of each. It can be shown that the standard deviation of the means of  $N$  measurements from a population with an overall standard deviation of  $\sigma$  (Greek lowercase 'sigma') is given by:

$$\sigma/\sqrt{N}$$

This quantity is termed the '**standard error**' of the population mean, and defines a range either side of the estimated population mean that is likely to contain the true value.

Where a population has been sampled several times, and the samples are normally distributed, the standard error of the mean of the sample ( $s_{x_{\text{mean}}}$ ) can be estimated from the standard deviation calculated in section 5:

$$s_{x_{\text{mean}}} = s/\sqrt{N}$$

The standard error of a sample mean can be multiplied by a value, usually denoted by the letter  $t$ , to provide what is termed a '**confidence interval**' (CI). This assigns a probability that the interval between the sample mean minus the CI and the sample mean plus the CI will include the value of the population mean.

The confidence interval is calculated as:

$$\text{CI} = \text{mean} \pm t \cdot s_{x_{\text{mean}}}$$

For example, if a sample size of 15 has a mean value of 100 and a standard error of 10, then the value of  $t$  for 15-1 = 14 degrees of freedom and a probability of 0.95 is 2.145. The upper CI is given by 100 plus (10 × 2.145) equals 121.45, and the lower CI is given by 100 minus (10 × 2.145) equals 78.55, more conventionally written as:

$$\text{95\% confidence interval} = 100 \pm 21.45$$

This indicates that the true population mean has a 95% probability of being within the specified confidence interval around the sample mean (or a 5% chance of falling outside this range).

The value of  $t$  decreases as sample size increases, so that the ability to predict the population mean from the sample mean improves with larger sample sizes for a given

standard deviation. As an illustration, for the same mean of 100 and standard error of 10:

Number of samples ( $df = n-1$ ) <sup>*</sup>	90% CI <sup>†</sup> $\alpha = 0.10$ <sup>†</sup>	95% CI $\alpha = 0.05$	99% CI $\alpha = 0.01$
5	79.68 - 121.32	72.24 - 127.76	53.96 - 146.04
10	81.67 - 118.33	77.38 - 122.62	67.50 - 132.50
15	82.39 - 117.61	78.55 - 121.45	70.23 - 129.77
20	82.71 - 117.29	79.07 - 120.93	71.39 - 128.61
30	83.01 - 116.99	79.55 - 120.45	72.44 - 127.46

\*  $df$  denotes degrees of freedom for critical values of  $t$ ,

† a confidence interval (CI) of 90% equates to a significance level,  $\alpha$ , of 0.10, i.e. a probability of 0.1 (10%) of the mean falling outside the CI.