SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES

R for Arts: Basic Statistics

Welcome to R for Arts! This document provides an overview of the definitions for some basic statistical terminology. We will be using this terminology in the R for Arts course, so please read this before attending.

Mean

To calculate the mean of a dataset, add up the values of all observations, then divide by the number of observations. For example, given this simple dataset:

$$4,\ 3,\ 4,\ 3,\ 2,\ 4,\ 13,\ 2,\ 1 \tag{1}$$

To calculate the mean first, add them together:

$$4 + 3 + 4 + 3 + 2 + 4 + 13 + 2 + 1 = 36 \tag{2}$$

Then divide by the number of observations (there are nine):

$$36/9 = \mathbf{4} \tag{3}$$

Median

The median is simply the value of the central observation, when the data are ordered by magnitude. So when the data are ordered, the middle value is the median (***highlighted***):

$$1,\ 2,\ 2,\ 3,\ \boldsymbol{3},\ 4,\ 4,\ 4,\ 13 \tag{4}$$

Mode

The mode is the most frequently-occurring number in the data: in this dataset the mode is **4**.

Range

The range is the overall extent of the data. This is found by subtracting the value of the smallest observation from the largest:

$$13 - 1 = \mathbf{12} \tag{5}$$

Interquartile range

The interquartile range offers a view of the range in the data that excludes extreme values (in this dataset: 1 and 13). It excludes the top and bottom 25% of the data, and calculates the range of the middle 50%.

First, we need to calculate the 'quartiles' (the three values that split the data into four equal parts). The 'second quartile' (also called Q2) is the same as the median – in this dataset the median is 3. The lower quartile (Q1) is the median of the lower half of the data (= 2), and the upper quartile (Q3) is the median of the upper half of the data (= 4).

Then the lower quartile is subtracted from the upper quartile:

$$IQR = Q3 - Q1 = 4 - 2 = \mathbf{2} \tag{6}$$

Standard Deviation

The standard deviation is a measure of how "spread out" the data is. The larger the standard deviation, the more spread out and 'flatter' your data would look on a graph. It is calculated by taking the square root of the 'variance' (the average of the squared differences from the mean).

First, calculate the variance. For each observation in the dataset, subtract the mean (in this dataset the mean is 4) then square the result (the 'squared difference'). Then calculate the average of these squared differences (divide by the number of observations: 9):

$$Variance = \frac{(-3)^2 + (-2)^2 + (-2)^2 + (-1)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 9^2}{9} = 11.11 \tag{7}$$

Finally, to get the standard deviation we take the square root of the variance:

$$SD = \sqrt{11.11} = \mathbf{3.33} \tag{8}$$

Standard error of the mean

When performing an analysis on a large dataset, it's usually more practical to take a small sample of the much larger population. The standard error of a sample provides an indication of how closely the sample represents the population. Generally, the larger your sample, the smaller the standard error, and the more representative of the whole population it will be.

Imagine that this dataset of 9 observations is a sample of a much larger population. To find the standard error of the mean, divide the sample standard deviation (defined as the sample estimate of the *population* standard deviation; s = 3.54) by the square root of the number of observations in the sample (n = 9).

$$SEM = \frac{s}{\sqrt{n}} = \frac{3.54}{\sqrt{9}} = \mathbf{1.18} \tag{9}$$

That's it for now – we're looking forward to seeing you on the course, and remember to bring your laptop with R fully installed!