MMaD: Lecture 2 handout

Types of data

Data is either:

- **Qualitative** - non-numeric data.
- **Quantitative** - data that can be represented by a number.

Quantitative data can be further classed in two subgroups:

- **Discrete** - a variable that can be counted or that has a fixed set of values.
- **Continuous** - a variable that can be measured on a continuous scale.

Measures of centrality

- The **arithmetic mean** (i.e. the average) is defined as the sum of the data $\{x_i\}$ divided by the number N of data points:

  $$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  μ denotes the mean of the entire population from which we have this sample of N values.

- The **median** is the value that separates an ordered list into two. For an odd number of values this is the middle number. If there are an even number of values, the median is the halfway point between the middle two numbers.
- The **mode** (or modal class) is the value (or class) of the sample that occurs the most.

Measures of spread of data

- The **range** of the data is simply the difference between the largest and smallest values.

  $$\text{Range} = \text{maximum value} - \text{minimum value}$$

- The **interquartile range** (IQR) is the difference between the upper and lower quartiles:

  $$IQR = U_{25} - L_{25}$$

  where:

  The lower quartile (L_{25}) is the median of the lower half of the data.

  The upper quartile (U_{25}) is the median of the upper half.

- The **variance** is the average of the square of the deviation of each data point from the arithmetic mean.

  Variance if the data set is the entire population:

  $$\sigma^2 \quad = \quad \overline{\Delta x^2} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

  Variance of the population based on a sample:

  $$\sigma^2 \quad = \quad \overline{\Delta x^2} \quad = \quad \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

- The **standard deviation** is the square root of the variance, so it also depends on whether we have the population or just a sample.

  S.D. if the data set is the entire population:

  $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2}$$

  S.D. of the population based on a sample:

  $$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2}$$

Standard Error

For a sample of N measurements, from which we have estimated a standard deviation of σ, the standard error is defined as:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

This is essentially an estimate of the confidence of how accurately we have measured the mean.