MAST SUMMARY NOTES

CHAPTER 1: Software, data handling, descriptive statistics and graphical methods

Objectives:

1. To use R & its basic data handling facilities.
2. To calculate appropriate descriptive statistics to summarise data.
3. To use a range of graphical methods to explore univariate & bivariate data.

The use of a command-line interface (CLI) has many advantages:

- It allows us to keep a record of the steps undertaken;
- It allows us to easily re-run commands without intervention;
- It allows us to make templates of often-used graphics or reports;
- It aids in the communication of the analysis to others;
- It protects us against change: when software changes, menus change but syntax rarely does.

Descriptive Statistics: The first step in understanding the data, before inference

There are two main types of variable, **categorical** (qualitative) & **numerical** (quantitative), each of which can be further divided:

- CATEGORICAL: **nominal** (e.g. gender) and **ordered** (e.g. severity of disease: mild/moderate/severe)
- NUMERICAL: **discrete** (e.g. number of children) and **continuous** (e.g. temperature).

Main roles of descriptive statistics are to:

- Detect anomalies (find **outliers**); It is *likely* (but not certain) entered wrong number. Tend to rid outliers in a '**summary**'.
- Examine & summarise data (type of variables or info); Allow for inferences. Location & spread give an idea of shape.
- Communicate results (summarise & get a result from that).

Tables of counts by category usually provide the best summaries of **categorical variables**. Most obvious features of **numerical variables**' data are (i) typical or central value, considered as a measure of location of the data & (ii) the spread or variability.

Measures of location:

- Sample mean or average: The sum of the observations divided by the number of measurements. Suppose we have *n* observations, and let x_1 denote the first observation, x_2 the second, & so on till x_n. Then the sample mean, $\bar{x}$, is given by:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample median: The middle measurement when the measurements are arranged in order. If there is an even number of measurements (*n*), it is average of the middle two ($x_{(\frac{n}{2})}$ & $x_{(\frac{n}{2}+1)}$). Let $x_{(1)}$ denote the smallest observation, $x_{(2)}$ the second smallest & so on up to the largest, $x_{(n)}$. The sample median, *m*, can be defined in terms of these order statistics as:

$$m = x_{(\frac{n+1}{2})}$$

  - Since it is not affected by a few wild values (like the mean is) the median is said to be **robust to outliers**.
  - But it is a less efficient estimator: Thus, all other things equal, the uncertainty about the true median, given an estimate of the median, is **larger** than the uncertainty about the true mean, given an estimate of the mean.
- Winsorised mean: Mean of sample with extreme values removed. It is robust to outliers, but not as robust as median.

Measures of spread: Sample standard deviation. Calculate differences between each observation & the average & square these differences to make them positive. Then we add them up & divide by n-1. Then take the square root of this.

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

- ~ 68% data lies within 1s of the mean; ~ 95% data lies within 2s of the mean. It is based on the normal distribution.
- Variance: Square of standard deviation. More difficult to interpret than s since it is not in the same units as the data.

Distribution of the data: Five number summary: An overview of the distribution of the data: median, first & third quartiles, minimum & maximum. 1Q is the value below ~25% data lie; 3Q is value above ~25% data lie. Basis of boxplot.

Correlation coefficient: A measure of the strength of the *linear* relationship between two **numerical** variables e.g. x & y:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The value of *r* always lies between -1 and 1. Positive *r* indicates positive association between the variables (as x increases, so does y) and negative *r* indicates negative association. The extreme values *r* = -1 and *r* = 1 only occur when the data lie exactly on a straight line, a perfect linear relationship. A correlation of 0 indicates **no** linear relation between x & y.