STATISTICS MATH1005 SUMMARY

Contents

3 sections:

Data analysis

- Summarising
- Presenting
- Extraction of information

Probability:

- Provides mathematical models for the data generation process
- Sample is just 1 of a long number of possible samples
- Set of mathematical tools so we can develop methods of analysis

Inference:

- Using probability models and the data to make inferences about the data-generating process.

DATA TYPES:

Overview:

Data types:

- Discrete (can only take on specific values eg integers)/
- Continuous (can take on any values eg all real numbers)

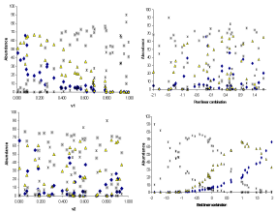
Note: in reality all data we get is discrete, as it has been rounded to an amount of decimal places upon gathering the data. We use continuous data modelling for if the data REPRESENETS something that CAN BE continuous (eg height: even though when measuring height we don't go to the infinite decimal place). Same logic for discrete

Data Summaries

- **Discrete**:
  - Frequency table

Class interval x(weight in kg)	Tally	Frequency f
40 - 44	II	2
45 - 49	IIII	4
50 - 54	𝍸	5
55 - 59	𝍸 III	8
60 - 64	𝍸	5
65 - 69	IIII	4
70 - 74	II	2
		30

- An elementary way to summarise discrete data is to produce a frequency table. (**R: <u>table</u> function)**, used to produce ordinate diagrams

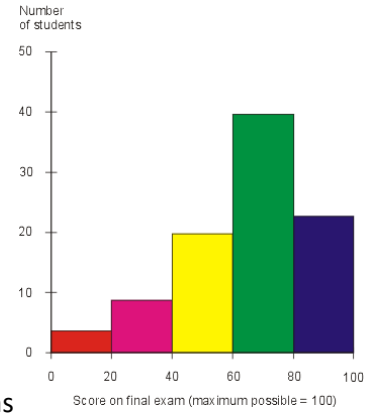
  - Ordinate diagram 

- **Continuous**

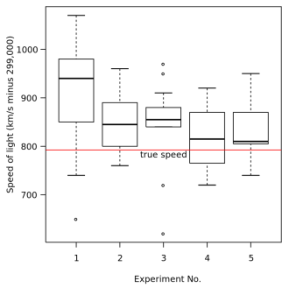
  - Stem and leaf plot

Stem	Leaf
2	2 6 7
3	1 3 5
4	2 4 6
5	7 8 9
6	1 3 4 5 7

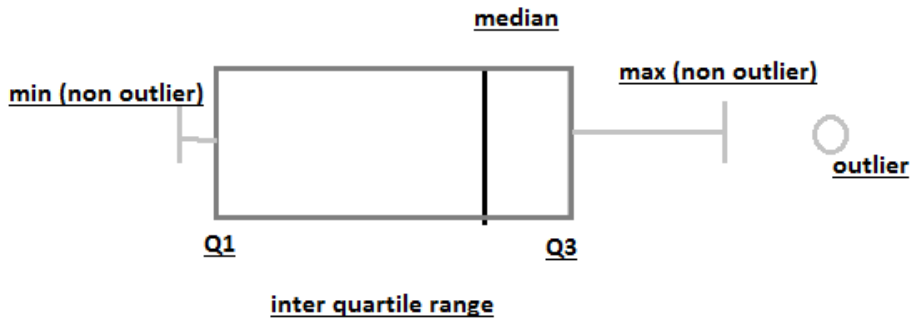
    - For each data value, split along some line (eg. Decimal point, 10's 100's) into stem and leaf, so leaf has only 1 number per data entry (**R: <u>stem( ))</u>**

  - Histograms 

    - A set of non-overlapping intervals is chosen, covers range of data
    - Rectangle is drawn on top of each interval, whose area represents the frequency

  - Box plot (box and whisker plot) 

- Median and quartiles are obtained
- Box drawn between quartiles against a scale
- Median is a line outliers are determined
- Interquartile range determined
- And value more than 1.5 IQR is determined to be an outlier
- Whiskers are drawn to furthest non-outlier
- Outliers marked separately



Statistical Objects

Median:

The middle score:

- Suppose x_1, x_2 … x_n represent the data, and $x_{(1)} <= x_{(2)} <= .... <= x_{(n)}$ represent the corresponging 'ordered statistics' (data in increasing order)
- Then:
  - If n is **ODD:**

$$\tilde{x} = x_{(\frac{n+1}{2})} \ (for\ odd\ n)[middle\ score]$$

  - If n is EVEN

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} \ (for\ even\ n)[average\ of\ middle\ scores]$$

- Properties of $\tilde{x}$**:**

$$if\ z_i = a + bx_i$$
$$\tilde{z} = a + b\tilde{x}$$

Therefore: median is a 'measure of location' (as it gives equivalent medians in linear functions. Eg: if I had a bunch of temperatures in Celsius, and converted it to Fareignheight, the medians would still be the relative same)