

# STATISTICS MATH1005 SUMMARY

---

## Contents

- DATA TYPES:..... 4
  - Data types: ..... 4
  - Data Summaries ..... 4
  - Median* ..... 6
  - Quartiles and Quantiles: ..... 7
    - Quartiles:..... 7
  - Interquartile Range (IQR) ..... 7
  - MEAN: (arithmetic mean) ..... 7
  - Variance and Standard Deviation (SD)..... 8
- Analysing Bivariate Data: ..... 9
  - Inner minimisation:..... 10
  - Outer Minimisation:..... 10
- Computing formulae ..... 10
- Correction ..... 11
  - Properties of  $r$  correlation coefficient: ..... 11
- Axioms of Probability: ..... 13
- Special sample spaces:..... 14
  - Sequences: ..... 14
- Product Sets: ..... 14
  - Permutations: ..... 14
  - Combinations: ..... 15
- Conditional Probability: ..... 15
- BAYES' RULE: ..... 15
- Independence: ..... 16
- Non-Equally Likely Outcomes: ..... 17
- Combing Experiments independently: ..... 17
- Binomial Distribution: ..... 18
- Infinite Sample spaces: ..... 18

Defining a probability of subsets of $\mathbb{N}$ .....	18
Uncountably infinite sample space.....	18
Random Variables .....	19
Hypergeometric distribution: .....	19
Geometric distribution:.....	19
Expectation:.....	19
Interpretation of expectation: .....	20
Expectation of a sum:.....	20
Expectation of $g(x)$ .....	21
Important examples:.....	21
Linear functions of $X$ .....	21
expectation .....	21
Variance .....	21
Standardised version of $X$ .....	21
Probability generating functions.....	21
Binomial: .....	22
Geometric: .....	22
Variance of a sum: .....	22
Joint Distribution of discrete random variables .....	22
Expectation of $g(X,Y)$ .....	22
Independent random variables.....	23
Expectation: .....	23
Covariance: .....	23
Variance: .....	23
PGF of sum of independent random variables .....	23
Continuous random variables: .....	23
Multinomials: .....	24
Poisson Distribution .....	24
Poisson distribution: .....	24
PGF of Poisson distribution:.....	24
Sum of independent Poissons :.....	25
Continuous Random variables: .....	25

Standard Normal Distribution:.....	26
Expectation: .....	26
Bivariate Continuous Distribution.....	26
Random ordered pairs: .....	26
Joint distribution:.....	27
Marginal distribution .....	27
Independent (Continuous) Random Variables .....	27
Sums of independent continuous random variables .....	27
Random sample sums and means .....	27
Sum of independent normal random variables.....	27
Markov and Chebyshev Inequalities .....	28
Random sample sums and mean .....	28
Normal as limiting binomial:.....	29
Sample variance .....	29
Chi squared distribution $\chi^2$ .....	29
Central Limit Theorem: .....	30
Detection problems: .....	31
interpretation of p values .....	31
Setting up the experiment: .....	31
1 sample t-test .....	31
Paired t test.....	31
Independent:.....	31
Inference for linear regression models with normal error.....	32
Pearson's chi squared statistic (multinomials) .....	33

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 REPRESENTS 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

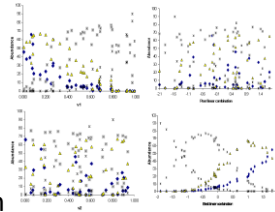
### Data Summaries

- **Discrete:**

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

- Frequency table

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



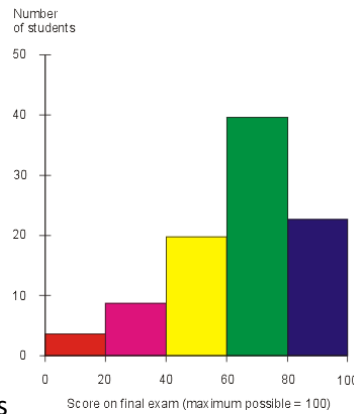
- Ordinate diagram

- **Continuous**

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

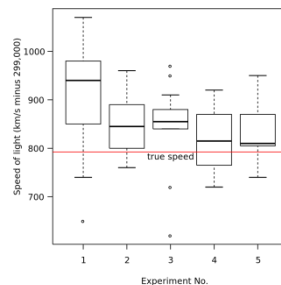
- Stem and leaf plot

- 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: stem()**)



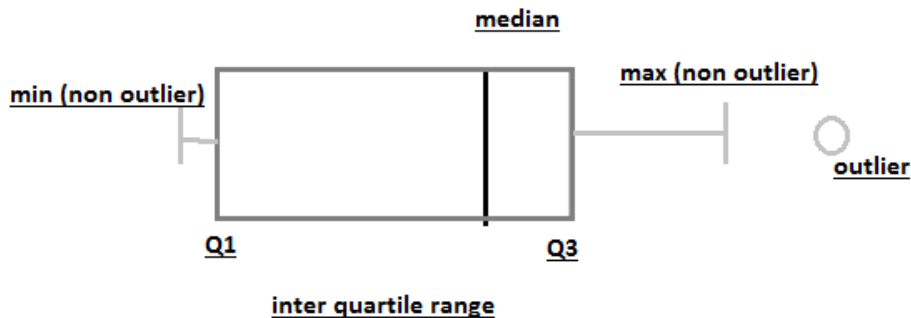
- Histograms

- 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  $x_1, x_2 \dots x_n$  represent the data, and  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  represent the corresponding 'ordered statistics' (data in increasing order)
- Then:

- If  $n$  is **ODD**:

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

- If  $n$  is **EVEN**

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

- Properties of  $\tilde{x}$ :

$$\text{if } z_i = a + bx_i$$

$$\tilde{z} = a + b\tilde{x}$$

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 Fahrenheit, the medians would still be the relative same)