BStats Notes

Two types of data: qualitative (categorical) vs quantitative (numerical) data

CATEGORICAL DATA

- Measurement scales
  - Nominal: Arbitrary numbering to represent a label
  - Ordinal: Numerical labels that represent an order/proportional

- Tabulating data: Frequency distribution method
  - Frequency count: The amount of occurrences for each category
  - Relative frequency: The proportion of a category compared to the total data set (decimal between 0 and 1)
  - Percent frequency: The relative frequency expressed as a %

- Methods of visualisation
  - Tabulating the frequency counts (e.g. excel spreadsheet)
  - Bar charts
  - Pie charts

NUMERICAL DATA

- Measurement scales:
  - Interval: Directional difference is meaningful, but the ratio is not (e.g. 15°C is not twice as warm as 30°C). NB: Zero is an arbitrary measurement figure.
  - Ratio: Both direction AND ratio of quantities is meaningful (e.g. Lucy earns twice as much as Fred). NB: Zero is meaningful (e.g. absolute silence is no sound waves)

DEFINITIONS

Random variable (r.v.): A variable factor which independently and randomly occurs i.e. what we are measuring as a success in a trial (e.g. sum of top face for 3 dice)

- Numerical values obtained from experimentation are called the realisation of the r.v.

Population: Total possible things applicable to the study
Sample size: The people selected to be studied

Probability distribution: The general shape of probability (curve of best fit) that a random variable may assume

Outlier: An observed r.v. point which lies outside the range of other realisations. They should be removed so as to not bias the legitimate data

Range: The minimum – maximum values of numerical data

Quartile: Splitting the data set into 4 sections according to medium values

- The quartile is labelled at its upper limit (e.g. Q_1 is at the border with Q_2
- Inter Quartile Range (IQR): Quartile 3 – Quartile 1

NOTATION

- Sets: Capital letters **X** and **Y** represent the random variables (e.g. height) and subscript/ lowercase letters **x** and **y** represent the realisation (e.g. $\mathbf{x_n}$ = height of person n)
- Size: **N** denotes the population size and **n** denotes the sample size
- Mean avg: **μ** or **E(X)** i.e. the 'expectation of X' represents the population mean and **$\bar{X}$** represents the sample mean

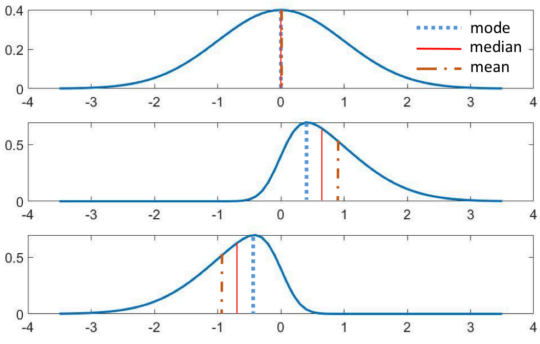
$$\mu = E(X) = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Variance: **σ^2** or **Var (X)** denotes the population variation of X and **s^2** denotes the sample variance. NB: It is squared to measure magnitude i.e. removes the sign so they don't cancel out when added. Therefore end units is actually **u^2** (not just u)

$$\sigma^2 = Var(X) = \frac{(x_1-\mu)^2 + \cdots + (x_N-\mu)^2}{N} \qquad s^2 = \frac{(x_1-\bar{X})^2 + \cdots + (x_n-\bar{X})^2}{n-1}$$

- Standard Deviation: **σ** or **std(X)** is for the population and **s** is for the sample (they are the sq root of variance and the end unit is just **u**)

Probability distribution function (pdf) calculates the probability of the r.v. being a certain number. It's a type of bell curve (characteristic shapes) with variable success rate on the x axis.



Symmetric distribution (skewness = 0):
mode = median = mean

Right-skewed distribution (skewness > 0, positively skewed):
mode < median < mean

Left-skewed distribution (skewness < 0, negatively skewed):
mode > median > mean

Skewedness: the side which has more sparse data (the less steep side of the curve)

Kurtosis: a measure of the tail and hence steepness of distribution (high kurtosis = flatter curve)

TYPES OF DISTRIBUTIONS:

1a) Binomial Distribution: X ~ Bin (n,p) i.e. (# of trials, probability of success)

For a r.v., the number of successful trials is called the 'binomial distribution' and is denoted:

$$P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Meaning: **P** (the probability) **of X** (the r.v.) **= x** (the amount of successes), **given n** (the total amount of trials) **and p** (the success rate for the r.v. for each trial) **is nCr** (out of n trials how many ways can you pick x indistinguishable successes) *** p^x * $(1-p)^{n-x}$** (the probability of a single result chain).

NB: $\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$ i.e. nCr

E(X) = np (number of experiments * avg success rate).
Var(X) = np(1–p) and std (X) = sq rt of var(X)

CDF formula: $P(X \le x; n, p) = \sum_{s=0}^{x} P(X = s; n, p) = \sum_{s=0}^{x} \binom{n}{s} p^s (1-p)^{n-s}$