1. Descriptive statistics

Terminology

Population – the set of all possible measurements of interest
Sample – a subset of measurements from the population

Graphs and tables

Stem and leaf plots

- Don't summarise – present all available info → used for small data sets
- Data split into stem and leaf components → *shape* of distribution
  - Stem
  - Leaf – only final digit of observation
- By default, vertical line = decimal point

Frequency distribution

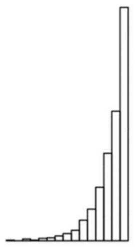
- Summarise → used for large data sets
- Intervals/bins
  - Should have same length

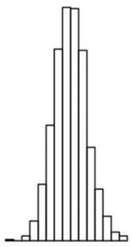
    $lengt \quad of\ bin \quad \frac{range\ of\ data}{number\ of\ bins}$ (round to next integer)

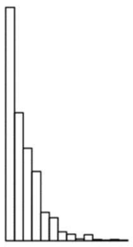
  - Will be right-closed i.e. observation on the boundary between 2 bins will be included in the left bin e.g. (8, 10]
- Cumulative frequency – total frequency up to and including a particular class
- $relative\ frequency = \frac{frequency}{sample\ size}$

Histogram

- Represents frequency distribution graphically
- For continuous data
- Describe main features of data
  - Overall pattern
  - Area of conc.
  - Presence/absence of outliers
  - *Shape* of distribution
    - Skewed to the left – left side stretches further from peak than right side
    - Symmetric – opposing sides approx.. symmetric about the middle
    - Skewed to the right – right side stretches further from peak than left side

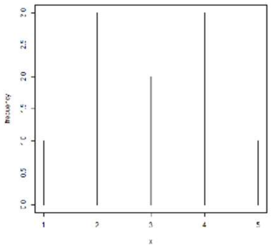

skewed to the left

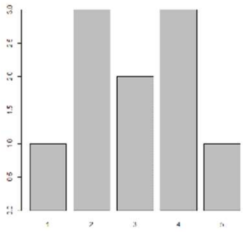

symmetrical


skewed to the right

Ordinate diagram or barpolt

- For discrete data
- Plot of f_x against x





Measures of location

- Attempt to provide a single numerical value which represents whole data set

Mean = $\bar{x}$ – the simple average

- + Simple to calculate
- – Can be greatly affected by extreme observations – pulled towards them
- – Inappropriate when working with skewed distributions

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Median = $\tilde{x}$ – the middle value

- If number of observations is even – $\tilde{x}$ = average of middle 2 observations
- + Not affected by outliers

Mode – the value that appears most frequently

- + Relevant for categorical and numerical data
- – Might not exist
- – Might not be unique – possible to have a bimodal distribution (2 modes)

Shape of distribution

- Mean = median = mode → symmetric
- Mean > median > mode → skewed to the right
- Mean < median < mode → skewed to the left

Measures of spread or dispersion

Variance = s^2 – a measure of the spread around $\bar{x}$

- In terms of squared distances between the observations and $\bar{x}$
- In units^2
- Dividing by (n – 1) → unbiased population variance
- Always non-negative
- Variance = 0 if and only is all observations are equal to each other
- Large values of s^2 → more spread around $\bar{x}$ = highly volatile
- Small values of s^2 → more conc. around $\bar{x}$ = less volatile
- $deviation\ of\ the\ observation = sample\ mean - observation$

$$\boxed{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}$$

Standard deviation = s

- Same units as data

$$s = \sqrt{variance} = \sqrt{s^2}$$