Summarising Data

Presenting data

- Listing- not summarised in any way
- Summarised

Listing can be useful for small amounts of data but not for many observations

Summarising data

- To communicate findings and prepare for data analysis
- Univariate vs bivariate (two variables) summaries
- Two types of summary displays
    1 Numeric summary – numbers
    2 Graphical summary – visual (graphs)

Appropriation summary depends on the measurement type (quantitate and quantitative) (categorical and numeric)

Categorical Data – discrete groups or groups (gender, race)

- Frequency tables and bar chart/pie chart

Numeric data – score on a scale (Age, distance driven to uni)

Numeric summary stats (mean/mode/median, standard deviation) and histogram

Think about:

- Typicality – most common score
- Variability – how much variety in scores
- Shape – pattern of the distribution

Categorical Data

Frequency Table: lists of categories and the count frequency of how many patterns fall into them.

Bar chart: X (horizontal) axis is distinct categories, Y (Vertical) axis is count or %

Pie chart: Size of the slices show the relative proportion of the category. – relative numbers, not literal numbers.

Numeric Data

Greater number of possible scores for numeric data, so need to summarise it differently

Freq Table: not as useful

Typicality: If we had to pick one score to best summarise the entire dataset, what score would we pick? – average

Types of typicality:

- Mean – Average
- Median – Middle – most score (ranked from lowest to highest)
- Mode- most common score

Mean- advantage: easy to calculate, most common, represents all the data

Disadvantage: easily affected by extreme , not always actual score in the dataset

Median – Rank from smallest to biggest, if not odd number, take the mean of the 2 middle scores. Advantages: easy to find, not affected by extreme values

Disadvantages: may not represent the data if is unbalances, not always actual score in the dataset.

Mode- most common score, can have multiple modes. Advantages: always an actual value, easy to find.

Disadvantages: can be multiple, doesn't take into account all of the data.

Variability

Measure of how much spread on scores or range of scores.