Lecture 1 – Reading Chapter 1

Terminology

- Individuals/units – objects described by a set of data
  - Can be people, animals or things
- Variable – any characteristic of an individual
  - Can take different values for different units
  - Any random unit will have a random variable
  - Can be categorical (groups) or numerical/quantitative (ordinal (numbered order eg shirt size) or discrete/continuous (counts of a characteristic – can take any number on a number scale))

Distribution of a Random Variable

- A distribution is a summary that indicates:
  - What values a variable takes and
  - How often it takes these values
- Visual summaries can be a table, graph or function
  - Categorical – pie chart, bar chart
  - Numerical – histogram, dot plots, stem and leaf plot (frequency distribution of a quantitative variable)

Examining Distribution of Numerical Continuous Data

- Location – around what value is the data located
- Spread – what is the variability among the data values
  - Range - max and min
  - Limits that most values are in
- Shape – what is the distribution of the data
  - Overall pattern
  - Deviation from the pattern
  - Outliers – any gaps in a histogram indicate that anything above that is an outlier

Histogram

- Frequency distribution of continuous numerical data
- Procedure
  - Divide the values into equal intervals (bins)
  - Count how many observations in each interval
  - Draw chart representing this distribution
  - Aim for between 6 and 12 columns/intervals/bins
  - Right skew = positive skew
- Describing a histograms

    - Shape – symmetric or skewed
    - Centre – around what value is the data grouped
    - Spread – how far spread is the data
    - Outliers – is there an individual value that falls outside the normal pattern (separate)
- Measuring the centre of distribution
    - Mean – average
    - Median – middle
- Measuring the spread of distribution
    - Standard deviation – the variability that individual data values are from the mean
    - Quartiles – quartile 1 is the middle of the lower half, and quartile 3 is the middle of the upper half

Outliers

- An outlier is a data point not consistent with the bulk of the data
- Can have a big influence on conclusions
- Can cause complications in statistical analyses
- Cannot discard without justification
- Possible reasons:
    - Mistake in measurement or data entry
    - Individual in question belongs to a different group than bulk
    - Outlier is legitimate and represents natural variability
- Affect the mean more than the median
-

Lecture 2 – Reading Chapter 2

Measuring the centre of a distribution

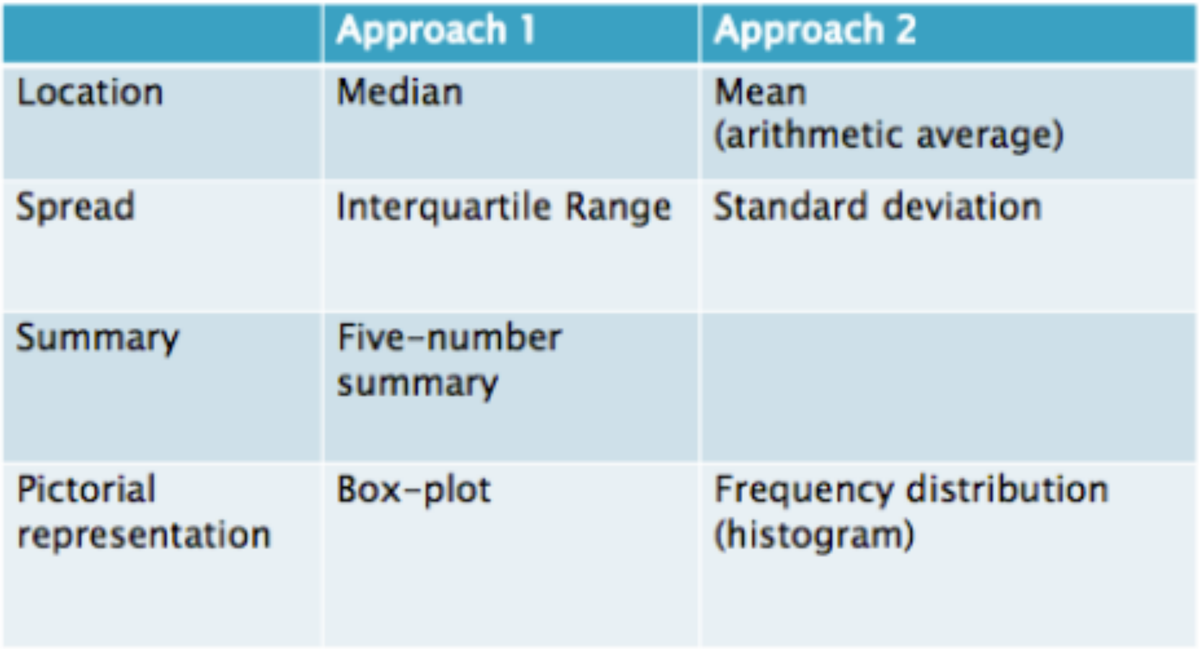
- Mean
    - Arithmetic average of the data value
    - Used when bell shaped distribution is symmetrical
- Median
    - The middle value
    - Location is the (n + 1)/2 position in the ordered (smallest to largest) list
    - Less affected by outliers
    - Used when curve is skewed

Measuring the spread of a distribution

- Standard deviation
    - The variability (on average) that individual data values are from the mean

      $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

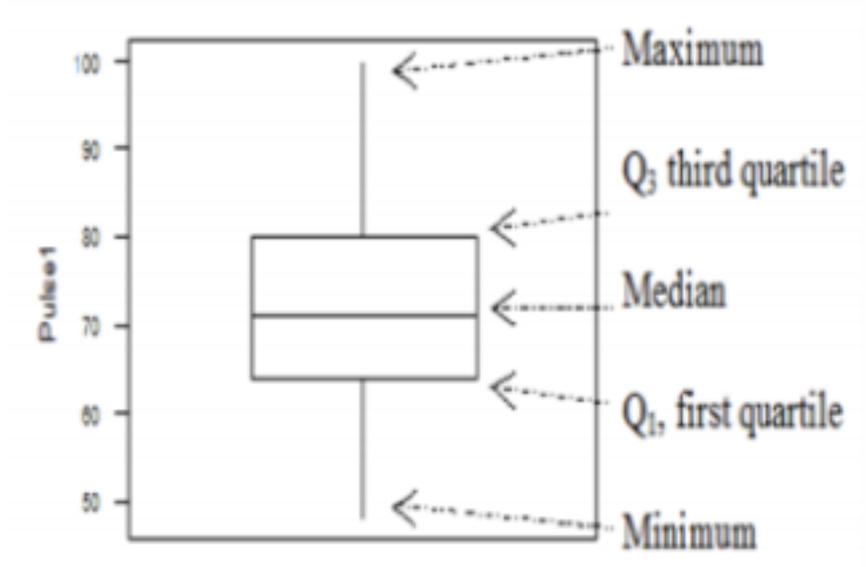
    -
    - Use STDEV.S in excel (measures sample SD rather than population SD)
- Quartiles
    - The 25% and 75% position in the ordered list of data
    - The middle value of each half



	Approach 1	Approach 2
Location	Median	Mean (arithmetic average)
Spread	Interquartile Range	Standard deviation
Summary	Five-number summary	
Pictorial representation	Box-plot	Frequency distribution (histogram)

-

How to draw a boxplot



- Label a vertical (or horizontal) axis with a numbered scale from min to max
- Draw box with lower end at Q1 and upper end at Q3
- Draw a line through the box a the median
- Place a dot at the minimum and the maximum
- Check for outliers
  - Locate the lower boundary (Q1-1.5 x IQR) and upper boundary (Q3 + 1.5 x IQR)
  - All data outside these values are outliers
- Draw line from Q1 end of box to smallest data value inside boundary and from Q3 end to the largest value inside boundary
- IQR -> Q3 – Q1
- When finding 1st and 3rd quartile, exclude the median data point

Comparative Boxplots

- Best way of "picturing" sub-groups in the same measurement
- Location
  - Compare medians and box overlap – is there a difference
- Spread
  - Box covers the middle 50% of the data (the IQR) – are they similar in size
- Possible outliers are marked with an asterisk – are there any in one or both groups?
- Symmetry of distribution
  - Position of median in the box

Outliers