Week 3

Definitions

- Central Tendency – the extent to which all data group around a central value
- Variation – The level of dispersion of data around the central value
- Shape – The pattern of distribution of data from lowest to highest

Measures Of Central Tendency

Arithmetic Mean ($\bar{x}$)

- Sum of all values, divided by the number of values
- Impacted by extreme values

Median

- Middle number of an ordered array
- Not impacted by extreme values (*'resistant measure'*)
- If odd number of values, median is simply the middle value
- If even number of values, median is the mean of the middle two values

Mode

- Value which occurs must frequently
- Not impacted by extreme values
- Not very useful for continuous data as unlikely to have equal values
- Possible for there to be no mode or multiple modes

Geometric Mean ($\bar{X}_G$)

- Measures the rate of change of a variable over time

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

Geometric Rate Of Return ($\bar{R}_G$)

- Measures average rates of return over time

$$\bar{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

- Where R_i = rate of return for period *i*

Measures Of Variation

Range

- Difference between minimum and maximum values
- Simplest measure of variation
- Highly sensitive to extreme values

Sample Variance (S^2)

- Sum of the differences between each value and the mean, divided by the number of values – 1
- Why n-1? If you have n values, you only have n-1 gaps between those values

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Sample Standard Deviation (S)

- Shows variation about the mean
- Square root of variance
- Has same units as data

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Where $\bar{X}$ = arithmetic mean

n = sample size

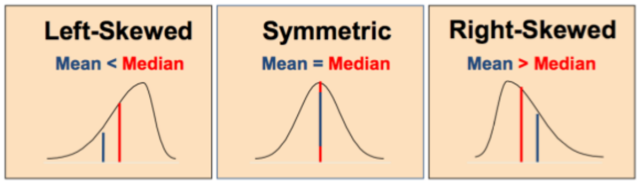
X_i = i^{th} value of the variable X

Coefficient Of Variation

- Measures variation relative to the mean
- Useful in comparing variations between data sets
- Expressed as a percentage

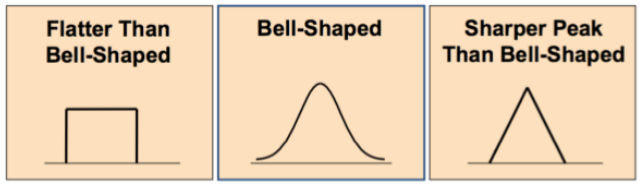
$$CV = \left(\frac{S}{\bar{X}}\right) \cdot 100\%$$

Measures Of Shape



Skewness

- Measures the level of asymmetry in a distribution
- Skewness of 0 means perfectly symmetrical
- Left Skewed: Median > Mean (Skewness <0)
- Right Skewed: Mean > Median (Skewness >0)



Kurtosis

- Measures extent of central tendency
- Kurtosis of 0 means bell shaped
- Lepokurtic: Sharp peak (kurtosis >0)
  - More values in tails
- Platykurtic: Flat peak (kurtosis <0)
  - Less values in tails

Assessing Extreme Values – Z Scores

- A Z-Score describes how many standard deviations a value lies from the mean
- Larger Z-Scores indicate a larger distance from the mean, >3 considered an outlier

$$Z = \frac{X - \overline{X}}{S}$$

where X represents the data value
$\overline{X}$ is the sample mean
S is the sample standard deviation

Quartiles

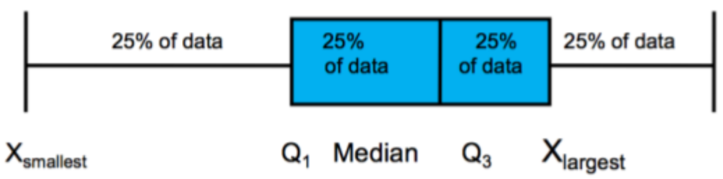
- Splits ranked data into 4 segments, with an equal number of values per segment
- $Q_i = i(n+1)/4$ where i is the quartile (1, 2 or 3)
  - If a whole number, simply use this ranked value
  - If a fractional half, average the two corresponding values
  - If neither of the above, round to nearest whole
- Q_2 is simply the median, with Q_1 and Q_3 being the median between Q_2 and the min/max

Interquartile Range

- Measures the spread of the middle 50% of the data (no indication of tails)
- Simply $Q_3 – Q_1$

Five Number Summary

- Five values describing the center, spread, shape and data:
  - Minimum
  - First Quartile (Q_1)
  - Median (Q_2)
  - Third Quartile (Q_3)
  - Maximum
- Easily communicated using a boxplot



Descriptive Statistics For Populations

- Called parameters

Population Mean (μ)

- Sum of all the values in the population divided by population size

Population Variance (σ^2)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Where μ = population mean
N = population size
X_i = i^{th} value of the variable X

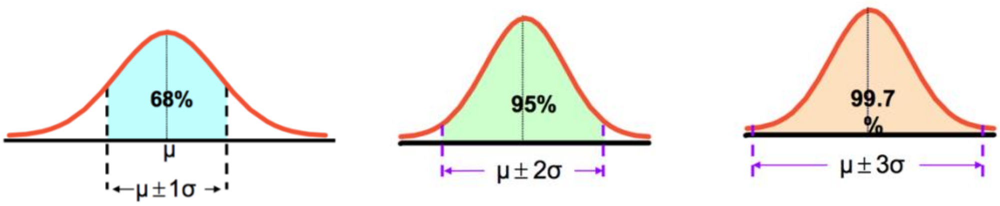
Population Standard Deviation (σ)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Where μ = population mean
N = population size
X_i = i^{th} value of the variable X

The Empirical Rule

- Approximates the distribution of data in a bell curve – ***only works for populations***
- Approximately 68% of the data in a bell curve lies within one standard deviation of the mean
- Approximately 95% of the data in a bell curve lies within two standard deviations of the mean
- Approximately 99.7% of the data in a bell curve lies within three standard deviations of the mean



Chebyshev's Rule

- Regardless of data distribution, Chebyshev's rule estimates the percentage of data within k standard deviations

$$\left(1 - \frac{1}{k^2}\right) \times 100\%, \ k > 1$$

Covariance – WEEK 4