Week 2 - Review

Thursday, January 18, 2018 3:55 AM

Measures of Central Tendency

- **Mean**
    - This is the average of a group of numbers and is calculated by summing all numbers and dividing by the count of numbers in that group.
    - The mean can be highly sensitive to outliers.
    - ***Population (ungrouped)***
        - $$\mu = \frac{\sum x_i}{N}$$
    - ***Sample (ungrouped)***
        - $$\bar{x} = \frac{\sum x_i}{n}$$
    - ***Grouped Mean***
        - $$\mu_{grouped} = \frac{\sum (f_i M_i)}{N}$$
- **Median**
    - This is the middle value in an ordered array of numbers.
        - Odd number of terms - the median is the middle number
        - Even number of terms - the median is the average of the two middle numbers
    - The median can be highly insensitive to outliers.
    - ***Grouped Median***
        - $$Median_{grouped} = L + \frac{((\frac{N}{2}) - cf_p)}{f_{med}} * W$$
- **Mode**
    - This is the most frequently occurring value in a set of data.
- **Percentiles**
    - This divides a group of data into 100 parts.
        - $$i = \frac{P}{100}(N)$$
- **Quartiles**
    - This divides a group of data into four subgroups or parts.
        - These are typically denoted as Q_1, Q_2, and Q_3.
        - ***The second quartile (Q2) is the median of the data.***

- $Q_1 = \frac{1}{4}(n+1)$
- $Q_2 = \frac{2}{4}(n+1)$
- $Q_3 = \frac{3}{4}(n+1)$

Measure of Variability - describes the spread of the dispersion of a data set

- **Range**
    - The range is the difference between the largest value and the smallest value of the data set.
- **Interquartile Range**
    - The interquartile range is the range of values between the first (Q1) and third (Q3) quartile.
    - The interquartile ranges measures variability without being sensitive to the more extreme values.
        - ***This property makes it well suited to detecting outliers.***
        - $IQR = Q_3 - Q_1$
- **Mean Absolute Deviation**
    - Also known as MAD - is the average of the absolute values of the deviations around the mean for a set of numbers
    - $MAD = \frac{\sum |x_i - \mu|}{N}$
- **Variance**
    - Variance is the average of the squared deviations around the arithmetic mean for a set of numbers.
        - ***Population Variance (ungrouped)***
            - $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$
        - ***Population Variance (grouped)***

$$\sigma^2_{grouped} = \frac{\left(\sum f_i M_i^2 - \frac{\left(\sum f_i M_i\right)^2}{N}\right)}{N}$$

- ***Sample Variance (ungrouped)***

    - $$s^2 = \frac{\sum (x_i - \overline{x})^2}{(n-1)}$$
  - ***Sample Variance (grouped)***
    - $$s_{grouped}^2 = \frac{\sum f_i M_i^2 - \frac{\left(\sum f_i M_i\right)^2}{n}}{n-1}$$

- **Standard Deviation**
  - The standard deviation is the square root of the variance.
    - ***Population Standard Deviation (ungrouped)***
      - $$\sigma = \sqrt{\left(\frac{\sum (x_i - \mu)^2}{N}\right)}$$
    - ***Population Standard Deviation (grouped)***
      - $$\sigma_{grouped} = \sqrt{\frac{\sum f_i M_i^2 - \frac{\left(\sum f_i M_i\right)^2}{N}}{N}}$$
    - ***Sample Standard Deviation (ungrouped)***
      - $$s = \sqrt{\left(\frac{\sum (x_i - \overline{x})^2}{(n-1)}\right)}$$
    - ***Sample Standard Deviation (grouped)***

      $$s_{grouped} = \sqrt{\frac{\sum f_i M_i^2 - \frac{\left(\sum f_i M_i\right)^2}{n}}{n-1}}$$
  - ***Empirical Rule***
    - It is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.
      - Mean +/- 1 standard deviation = 68%
      - Mean +/- 2 standard deviation = 95%
      - Mean +/- 3 standard deviation = 99.7%
  - ***Chebyshev's Theorem***
    - At least 1 - 1/k^2 values will fall within +/-k standard deviations of the mean regardless of the shape of the distribution.
    - This theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution is unknown or is non-normal.

        - $1-\frac{1}{k^2}$

- **Z Scores**
    - A z score represents the number of standard deviations a values (x) is above or below the mean of a set of numbers when the data are normally distributed.
    - $z=\frac{\left(x_i-\overline{x}\right)}{s}$
    - $z=\frac{\left(x_i-\mu\right)}{\sigma}$
- **Coefficient of Variation**
    - This is a statistic that is the ratio of the standard deviations to the mean which is expressed in percentages .
    - This can be useful when comparing standard deviations were computed from data with different means.
    - $CV=\frac{\sigma}{\mu}*100$

Measure of Shape - these are tools that can be used to describe the shape of a data distribution

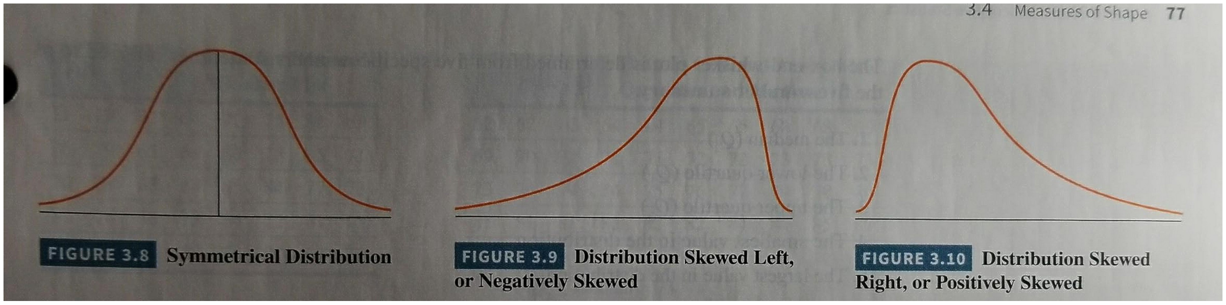
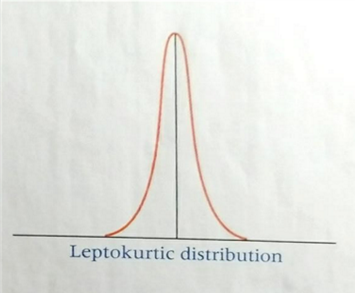
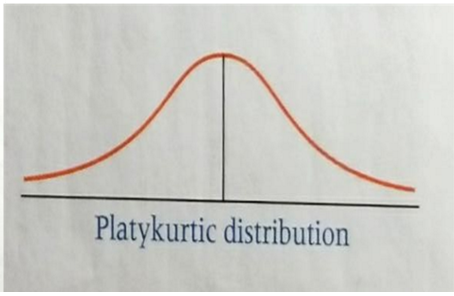
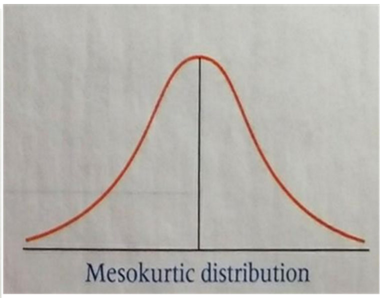
- **Skewness**
    - This describes when a distribution is asymmetrical or lakes symmetry.
    - 

FIGURE 3.8 Symmetrical Distribution

FIGURE 3.9 Distribution Skewed Left, or Negatively Skewed

FIGURE 3.10 Distribution Skewed Right, or Positively Skewed

- **Kurtosis**
    - Kurtosis describes the amount of peakedness of a distribution.
    - 
Leptokurtic distribution

Platykurtic distribution

Mesokurtic distribution

- **Box-and-whisker plots**
  - Also known as box plot - a diagram that utilizes the upper and lower quartiles along with the mean and the two most extreme values to depict a distribution graphically.
  - ***Five-Number Summary***
    - *The median (Q2)*
    - *The lower quartile (Q1)*
    - *The upper quartile (Q3)*
    - *The smallest value in the distribution*
    - *The largest value in the distribution*

Calculations to trim data

- ***Trimmed Mean***
  - This removes the top and bottom percent of the data set.
  - According to Wilcox in Basic Statistics, 20% trimming is often a good choice.
- ***Winsorized data***
  - This is similar to the trimmed mean, however instead of removing the data from the set. They are set equal to the smallest value not trimmed and the largest values are set equal to the largest value not trimmed.