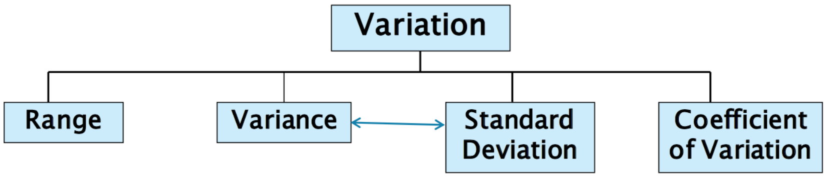
2. Variation And Shape:



- Measures of variation give information on the spread or variability or dispersion of the data values.

- Range:
  - The simplest. Can be used for all ordered data
  - The difference between the largest and smallest value
  - The range can be misleading because it ignores the way the data is distributed; it can be positively or negatively skewed.
  - It is also highly sensitive to outliers.

- Sample Variance:
  - Average (approx.) of squared deviation of values from the mean.

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

Where $\overline{X} =$ arithmetic mean

$n =$ sample size

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

- Variance Finance And Risk:
  - Variance forms a major component of modern financial investment and risk management practice.

- Sample Standard Deviation:
  - Most commonly used measure of variation
  - Shows variation about the mean
  - Is the square root of the variance
  - Has the same units as the original data

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

  - Data may have large or small standard deviations

- Summary Characteristics:
  - The more the data are spread out, the greater the range, variance, and standard deviation.

  - There more the data are concentrated, the smaller the range, variance, and standard deviation.
  - If the values are all the same (no variation), all these measures will all be zero.
  - None of these measures are ever negative.

- Coefficient Of Variation:
  - Measures the scatter of data relative to the mean.
  - As a percentage %
  - Can be used to compare the variability of two or more sets of data measured in different units.
  - Example: two stocks may have the same standard deviation, but their coefficient of variation may be larger or smaller – relative to the average value of the stock for that year.

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

Assessing Extreme Observations:

- Sample Z-Score:
  - To compute the Z-score of a data value, subtract the mean and divide by the standard deviation.
  - The Z-score is the number of standard deviations a data value is from the mean.
  - A data value could be considered extreme (an outlier) if its Z-score is less than -3.0 or greater than 3.0.
  - The larger the absolute value of the Z-score, the further away the data value is from the mean.
  - You can find the proportion of scores that are considered outliers by subtracting or adding three standard deviations from the mean.

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

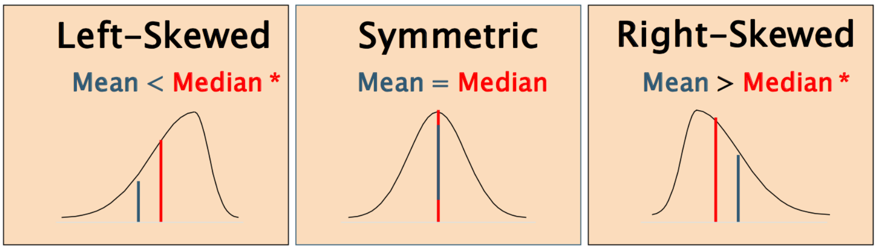
Shape Of A Distribution:

- The pattern to the distribution of values throughout the entire range of all the values is called the shape.
- Measures how data are distributed
- Two useful shape related statistics
  - Skewness:

        - Measures the amount of asymmetry around the mean in a distribution
    - Kurtosis:
        - Measures the relative concentration, or “peakedness” of values in the center of a distribution, as compared with the “tails”.

$$Skewness = \tfrac{1}{n}\sum_{i=1}^{n}\frac{\left(X_i-\overline{X}\right)^3}{S^3} \quad ; \quad \underset{\textit{statistic}}{Kurtosis} = \tfrac{1}{n}\sum_{i=1}^{n}\frac{\left(X_i-\overline{X}\right)^4}{S^4}-3$$

- Skewness:
    - Describes the amount of asymmetry in distribution
        - Symmetric or skewed
        - Left-Skewed: skewedness statistic < 0
            - Negatively skewed: mean < median
        - Right-Skewed: skewedness statistic > 0
            - Positively skewed: mean > median
        - Symmetric: skewedness statistic = 0
            - Symmetrical: mean = median



- Kurtosis:
    - Describes relative concentration of values in the center as compared to the tails. It measures the extent to which values that are very different from the mean affect the shape of the distribution of a set of data.
        - Flatter: kurtosis statistic < 0 (platykurtic)
        - Bell-Shaped: kurtosis statistic = 0
        - Sharper: kurtosis statistic > 0 (lepokurtic)

3. Exploring Numerical Data

- Quartile Measures:
    - Quartiles split the ranked data into 4 segments with an equal number of values per segment.
    - The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger.
    - Q_2 is the median.
    - Only 25% of the observations are greater than the third quartile Q_3.
    - They sit at the 25th, 50th and 75th percentile.

- Locating Quartiles:
  - Find a quartile by determining the value in the appropriate position in the ranked data.
  - Where n is the number of observed values:

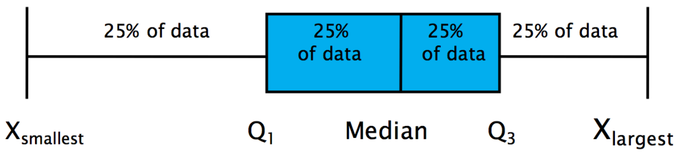
First quartile position: $Q_1 = (n+1)/4$ ranked value
Second quartile position: $Q_2 = (n+1)/2$ ranked value
Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

- Calculation Rules:
  - When calculating the ranked position use the following rules:
    - If the result is a whole number then it is the ranked position to use.
    - If the result is a fractional half, then average the two corresponding data values.
    - If neither, round the result to the nearest integer to find the ranked position.

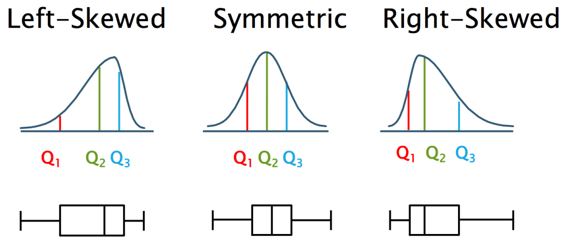
- The Interquartile Range:
  - The IQR is $Q_3 - Q_1$
  - IQR measures the spread in the middle 50% of the data.
  - IQR is also called the mid-spread
  - It is a measure of variability, which is not influenced by outliers or extreme values.
  - Measure like Q_1, Q_3, and IQR that are not influence by outliers are called "resistant" or "robust" meausres.

- The Five Number Summary:
  - $X_{smallest}$, Q_1, Q_2, Q_3, $X_{largest}$

- The Boxplot:
  - A graphical display of the data based on the five-number summary.



- Box plot distribution:



4. Numerical Descriptive Measures For A Population:

Numerical Descriptive Measures For A Population:

- Descriptive statistics discussed previously described a sample, not the population.
- Summary measures describing a population, called parameters, are denoted with Greek letters.
- Important population parameters are the population mean, variance and standard deviation.
- Note that n (or n-1) is replaced by N = population size.

Measure	Population Parameter	Sample Statistic
Mean	μ	$\bar{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

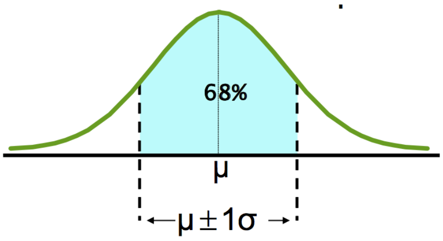
- Population Mean:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

- Population Variance:

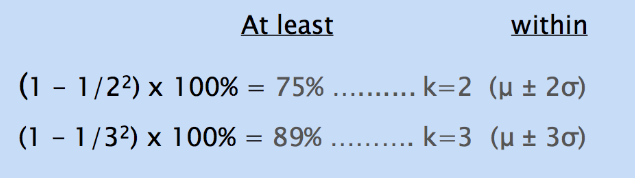
$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

- The Empirical Rule:
    - Approximates the variation of data in a bell-shaped distribution
    - Approximately 68% of the data in a bell shaped distribution is within one standard deviation of the mean.



    - Approximately 95% of the data will lie within two standard deviations, and approximately 99.7% of the data will lie within three standard deviations.

- Chebyshev's Rule:
  - For any data set, regardless of its shape, at least (1-1 / k^2) x 100% of the values will fall within k standard deviations of the mean (for $k > 1$).



	At least	within
$(1 - 1/2^2) \times 100\% =$	75% .... $k=2$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% =$	89% .... $k=3$	$(\mu \pm 3\sigma)$

5. The Covariance And Coefficient Of Correlation

- The Covariance:
  - Covariance measures the strength of the linear relationship between two numerical variables.

- The Coefficient Of Correlation:
  - The coefficient of correlation measures the relative strength of a linear relationship between two numerical variables.
  - The values can range from -1.0 (perfect negative correlation) to 1.0 (perfect positive correlation)

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

where

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

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

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

6. Descriptive Statistics: Pitfalls And Ethical Issues:

- Numerical descriptive measures:
  - Should document both good and bad results
  - Should be presented in a fair, objective and neutral manner
  - Should not use inappropriate summary measures to distort facts
  - Need to look at what 'average' measure is used and compare it to others as the results may be skewed.
  - Need to look at all the other measures also.