CHAPTER 2

Graphical method may not always be sufficient for describing data. You can use the data to calculate a set of numbers that will convey a good mental picture of the frequency distribution. Numerical descriptive measures associated with a population of measurement are called **parameters**; those computed from sample measurements are called **statistics**.

- **Measures of Center**
  - Mean
    This is the usual arithmetic mean or average and is equal to the sum of the measurements divided by number of measurements.

$$\begin{aligned} Sample\ Mean &= \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \\ Population\ Mean &= \mu = \text{the average of the population} \end{aligned}$$

  - Median
    This is the middle of the measurements when one orders them.
    How to obtain the median of the sample $x_1, \ldots, x_n$.
    First order the measurements from the lowest to the highest to get $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Note that $x_{(1)}$ is the smallest measurement, $x_{(2)}$ is the second smallest, ... , $x_{(n)}$ is the largest measurement.

$$\begin{aligned} \text{The median when } n \text{ is odd} &= x_{(\frac{n+1}{2})} \\ \text{The median when } n \text{ is even} &= \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}. \end{aligned}$$

  - Mode
    The mode is the measurement or the class which occurs most frequently.
    Note: Mean and median are equal when distribution of data is symmetric, mean is greater when distribution is skewed to right and is less than median when distribution is skewed to left.

Example: The prices for 14 different brands of water-packed light tuna are 0.99, 1.92,1.23, 0.85, 0.65, 0.53, 1.41, 1.12, 0.63, 0.67, 0.69, 0.60, 0.60, 0.66.

a. Find the average price for the 14 different brands of tuna.
b. Find the median price for the 14 different brands of tuna.
c. c. Based on your findings in parts a and b, do you think that the distribution of prices is skewed? Hint: Compare the mean to the median.

- **Measures of Variability**
  Data sets may have the same center but look different because of the way the numbers spread out from the center. Measures of variability can help you create a mental picture of the spread of data.

  – Range

  $$\textit{Range} = \textit{largest measurement - smallest measurement}$$

  – Variance
  It measures the average deviation of the measurements about their mean.

  $$\begin{aligned} \textit{Sample variance} &= S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}, \end{aligned}$$

  where $\sum_{i=1}^{n} x_i^2 = x_1^2 + \ldots + x_n^2$.
  Note: Further important details on the sample variance and also on the POPULATION VARIANCE σ^2 are to be discussed in class.

  – Standard deviation

  $$\begin{aligned} \textit{Sample standard deviation} = S &= \sqrt{S^2} \\ \textit{Population standard deviation} = \sigma &= \sqrt{\sigma^2} \end{aligned}$$

Example: You are given n = 8 measurements: 3, 1, 5, 6, 4, 4, 3, 5.
a. Calculate the range.
b. Calculate the sample mean.
c. Calculate the sample variance and standard deviation.

d. Compare the range and the standard deviation. The range is approximately how many standard deviations?

Relation between the Range and the Standard deviation

$$Standard\ deviation = S \approx \frac{Range}{4}.$$

Tchebysheff's Theorem

Given a number k greater than or equal to 1 and a set of n measurements, *at least* $1 - \frac{1}{k^2}$ of the measurement will lie within k standard deviations of the mean. In other words, the proportion of the measurements that lie inside the interval $[\bar{x} - k.S, \bar{x} + k.S]$ (in case of sample) or the interval $[\mu - k.\sigma, \mu + k.\sigma]$ (in case of population) is *at least* $1 - \frac{1}{k^2}$.

The Empirical Rule

When a distribution of measurements is approximately *mound-shaped* and *symmetrical* then

- The interval $[\mu - \sigma, \mu + \sigma]$ or $[\bar{x} - S, \bar{x} + S]$ contains approximately 68% of the measurements.
- The interval$[\mu - 2\sigma, \mu + 2\sigma]$ or $[\bar{x} - 2S, \bar{x} + 2S]$ contains approximately 95% of the measurements.
- The interval $[\mu - 3\sigma, \mu + 3\sigma]$ or $[\bar{x} - 3S, \bar{x} + 3S]$ contains approximately 99.7% of the measurements.

Examples:

1. The ages of 50 tenured faculty at a state university are 34, 48, 70, 63, 52, 52, 35, 50, 37, 43, 53, 43, 52, 44, 42, 31, 36, 48, 43, 26, 58, 62, 49, 34, 48, 53, 39, 45, 34, 59, 34, 66, 40, 59, 36, 41, 35, 36, 62, 34, 38, 28, 43, 50, 30, 43, 32, 44, 58, 53.

a. Do the data agree with those given by Tchebysheff's Theorem?
b. Do they agree with the Empirical Rule? Why?

2. The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?

Measures of Relative Standing
Where does one particular measurement stand in relation to the other measurements in the set of data?

- z-score
  How many standard deviations away from the mean does the measurement lie? This is measured by the z-score. Consider a particular measurement x. Then the z-score for this measurement is defined as follows.

  $$z - score = \frac{x - \bar{x}}{S}$$

  z-score between -2 and 2 are *not* unusual. z-score should not be more than 3 in absolute value. z-scores larger than 3 in absolute value would indicate a possible **outlier**.

- Percentiles

  How many measurements lie below the measurement of interest? This is measured by p^{th} percentile. p^{th} percentile is the value of measurement that is more than $p\%$ of the measurements in ordered data. Particular and important form of percentiles are the *Quartiles*.

  - Lower Quartile (Q_1) is the 25^{th} percentile. It is the value of x which is larger than 25% and less than 75% of the ordered measurements.
  - Second Quartile (Q_2) 50^{th} percentile. It is the value of x which is larger than 50% and less than 50% of the ordered measurements. Q_2 is the same as the *Median*.
  - Upper Quartile (Q_2) 75^{th} percentile. It is the value of x which is larger than 75% and less than 25% of the ordered measurements.

Note: Details and examples concerning how to obtain quartiles is to be discussed in class.

Interquartile range (IQR)

$$IQR = Q_3 - Q_1 = Upper\ Quartile - Lower\ Quartile$$

Example: The prices of 18 brands of walking shoes: 50, 60, 65, 65, 65, 68, 68, 70, 70, 70, 70, 70, 70, 74, 75, 75, 90, 95 Find IQR and the median.

Box Plot

Box plot describes center of data, how spread the data, the extend and nature of any departure from symmetry, and identification of outliers.
In general, box plot is based on the so=called **five number summary** which are:

Smallest value, Q_1, median, Q_3 and largest value.

Constructing Box Plot

- Compute five number summary and also the IQR.
- Show five numbers on horizontal line and draw a box above the horizontal line from Q_1 to Q_3 and determine median by a vertical line through the box.
- Draw 2 vertical lines from *lower fence* and *upper fence*

$$lower\ fence = Q_1 - 1.5(IQR)$$
$$Upper\ fence = Q_3 + 1.5(IQR)$$

- Determine the outliers (any observation beyond the fences) by $*$.
- Draw two horizontal lines from the end of the box to largest and smallest observations which are not outliers (whiskers).

Interpreting a Box plot

- Median line in center of box and whiskers of equal length then we have a set of measurement with a symmetrical distribution

- Median line left of center and long right whisker then we have a set of measurements with a skewed to the right distribution
- Median line right of center and long left whisker then we have a set of measurements with a skewed to the left distribution

Example: Construct a box plot for these data and identify any outliers: 25, 22, 26, 23, 27, 26, 28, 18, 25, 24, 12