Chapter 3
Describing Data Using Numerical Measures

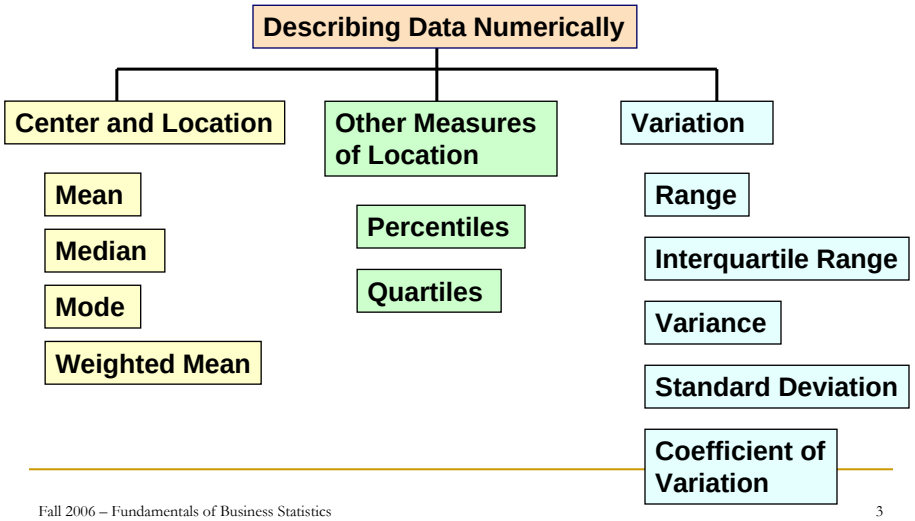
Chapter Goals

To establish the usefulness of summary measures of data.

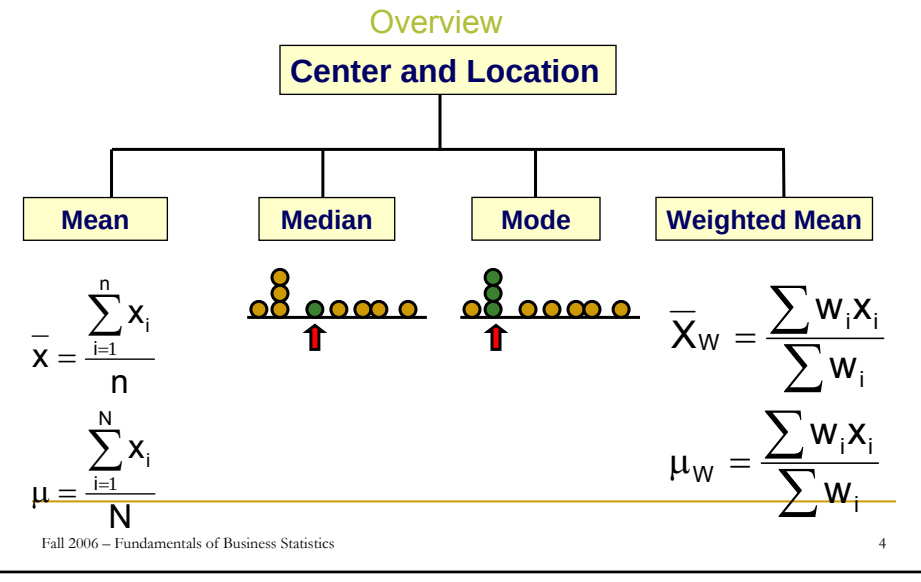
The Scientific Method

1. Formulate a theory
2. Collect data to test the theory
3. Analyze the results
4. Interpret the results, and make decisions

Summary Measures



Measures of Center and Location



Mean (Arithmetic Average)

The mean of a set of quantitative data, $X_1, X_2, \ldots, X_n$, is equal to the sum of the measurements divided by the number of measurements.

n = Sample Size

- Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

N = Population Size

- Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

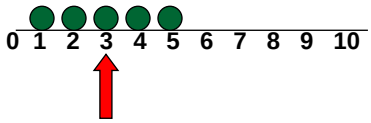
Example

Find the mean of the following 5 numbers 5, 3, 8, 5, 6

Mean (Arithmetic Average)

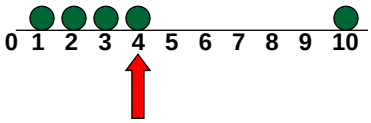
(continued)

- Affected by extreme values (outliers)
- For non-symmetrical distributions, the mean is located away from the concentration of items.



Mean = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



Mean = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

YDI 5.1 and 5.2

- Kim's test scores are 7, 98, 25, 19, and 26. Calculate Kim's mean test score. Does the mean do a good job of capturing Kim's test scores?
- The mean score for 3 students is 54, and the mean score for 4 different students is 76. What is the mean score for all 7 students?

Median

The median Md of a data set is the middle number when the measurements are arranged in ascending (or descending) order.

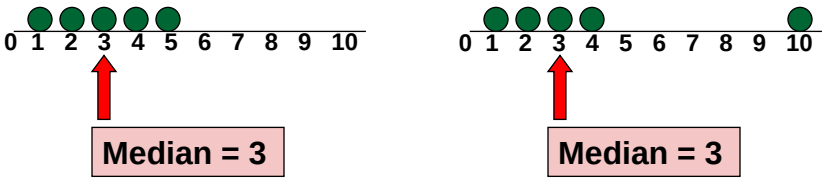
Calculating the Median:

1. Arrange the *n* measurements from the smallest to the largest.
2. If *n* is odd, the median is the middle number.
3. If *n* is even, the median is the mean (average) of the middle two numbers.

Example: Calculate the median of 5, 3, 8, 5, 6

Median

- Not affected by extreme values



- In an ordered array, the median is the "middle" number. What if the values in the data set are repeated?

Mode

Mode is the measurement that occurs with the greatest frequency

Example: 5, 3, 8, 6, 6

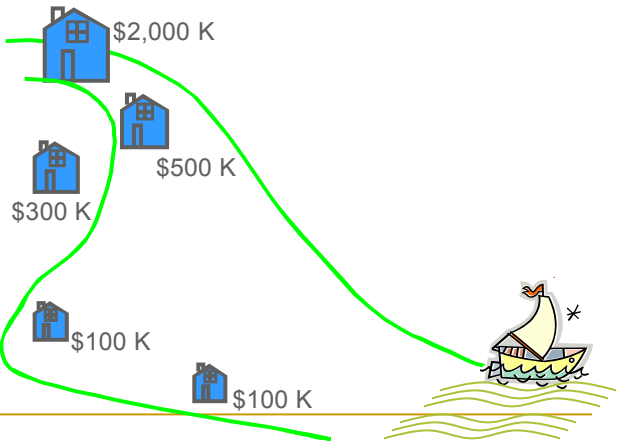
The modal class in a frequency distribution with equal class intervals is the class with the largest frequency. If the frequency polygon has only a single peak, it is said to be unimodal. If the frequency polygon has two peaks, it is said to be bimodal.

Review Example

- Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Summary Statistics

House Prices:
$2,000,000
500,000
300,000
100,000
100,000
Sum 3,000,000

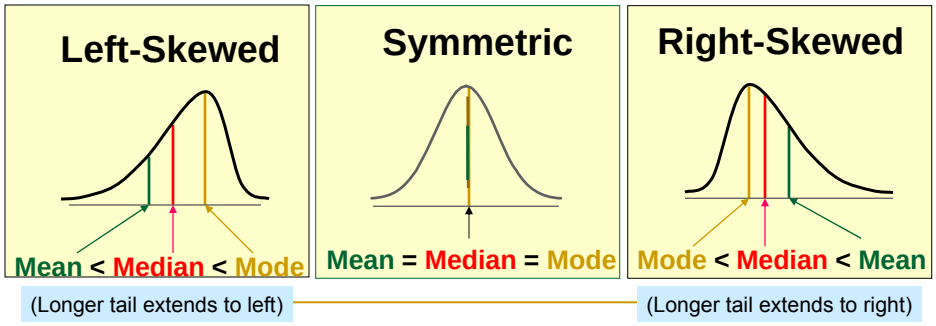
- **Mean:**
- **Median:**
- **Mode:**

Which measure of location is the "best"?

- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed



Other Location Measures

Other Measures of Location

Percentiles

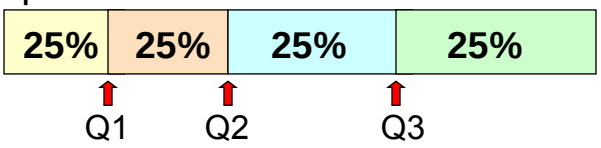
Let $x_1, x_2, \cdots, x_n$ be a set of n measurements arranged in increasing (or decreasing) order. The pth percentile is a number **x** such that p% of the measurements fall below the pth percentile.

Quartiles

- 1st quartile = 25th percentile
- 2nd quartile = 50th percentile = median
- 3rd quartile = 75th percentile

Quartiles

- Quartiles split the ranked data into 4 equal groups



- Example: Find the first quartile

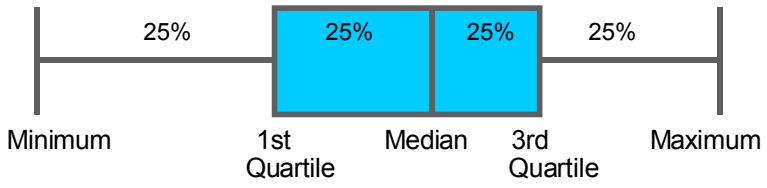
Sample Data in Ordered Array:	11	12	13	16	16	17	18	21	22

Box and Whisker Plot

- A Graphical display of data using 5-number summary:

Minimum -- Q1 -- Median -- Q3 -- Maximum

Example:

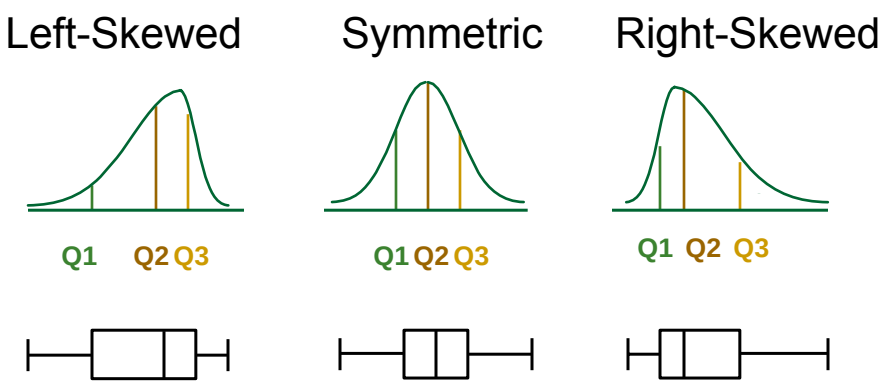


Shape of Box and Whisker Plots

- The Box and central line are centered between the endpoints if data is symmetric around the median

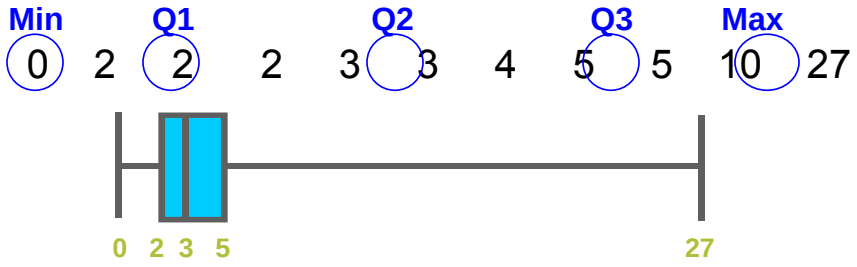
- A Box and Whisker plot can be shown in either vertical or horizontal format

Distribution Shape and Box and Whisker Plot



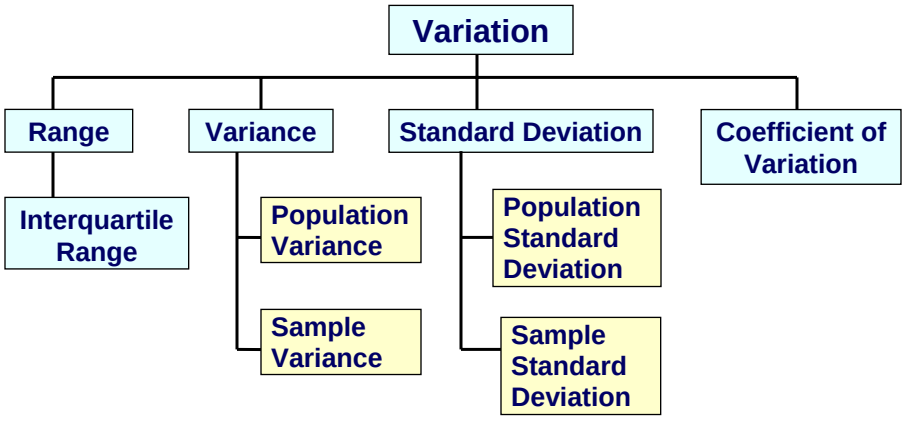
Box-and-Whisker Plot Example

- Below is a Box-and-Whisker plot for the following data:



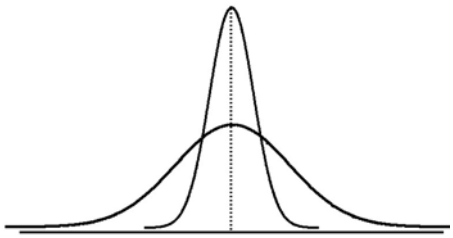
- This data is very right skewed, as the plot depicts

Measures of Variation



Variation

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



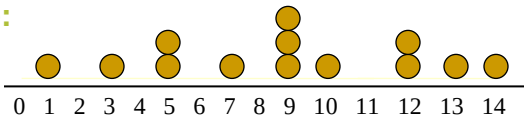
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = x_{\text{maximum}} - x_{\text{minimum}}$$

Example:



Disadvantages of the Range

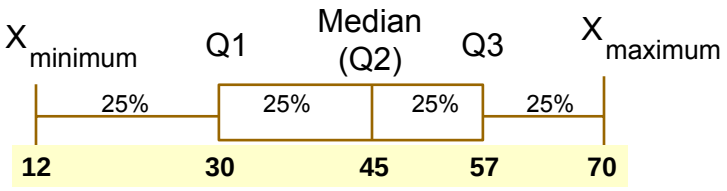
- Considers only extreme values
- With a frequency distribution, the range of original data cannot be determined exactly.

Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high-and low-valued observations and calculate the range from the remaining values.
- Interquartile range = 3rd quartile – 1st quartile

Interquartile Range

Example:

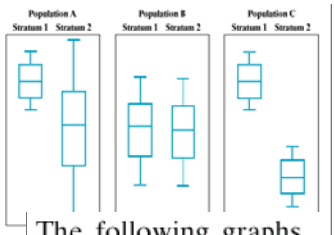


YDI 5.8

Consider three sampling designs to estimate the true population mean (the total sample size is the same for all three designs):

1. simple random sampling
2. stratified random sampling taking equal sample sizes from the two strata
3. stratified random sampling taking most units from one strata, but sampling a few units from the other strata

- For which population will design (1) and (2) be comparably effective?
- For which population will design (2) be the best?
- For which population will design (3) be the best?
- Which stratum in this population should have the higher sample size?



The following graphs are side-by-side box-plots of some variable for two strata in three hypothetical populations, A, B, and C. In each population the units are evenly divided between the two strata.

Variance

- Average of squared deviations of values from the mean
  - **Sample** variance:

Example: 5, 3, 8, 5, 6

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

Variance

- The greater the variability of the values in a data set, the greater the variance is. If there is no variability of the values — that is, if all are equal and hence all are equal to the mean — then $s^2 = 0$.
- The variance s^2 is expressed in units that are the square of the units of measure of the characteristic under study. Often, it is desirable to return to the original units of measure which is provided by the standard deviation.
- The positive square root of the variance is called the sample *standard deviation* and is denoted by *s*

$$s = \sqrt{s^2}$$

Population Variance

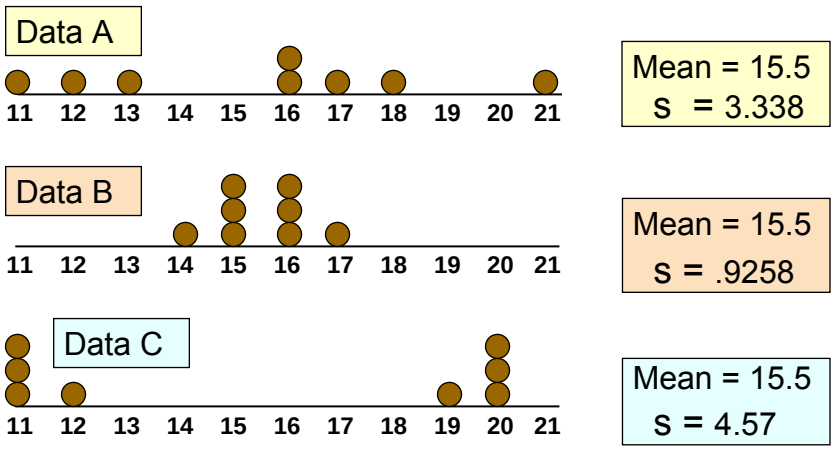
- **Population** variance:

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

- Population Standard Deviation

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

Comparing Standard Deviations



Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Is used to compare two or more sets of data measured in different units

Population

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\%$$

Sample

$$CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$$

YDI

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

Linear Transformations

The data on the number of children in a neighborhood of 10 *households is as follows:* 2, 3, 0, 2, 1, 0, 3, 0, 1, 4.

1. *If there are two adults in each of the above households, what is the mean and standard deviation of the number of people (children + adults) living in each household?*
2. *If each child gets an allowance of* $3*, what is the mean and standard deviation of the amount of allowance in each household in this neighborhood?*

$\overline{X} = 1.6$

$s = 1.43$

Definitions

Let X be the variable representing a set of values, and s_X and $\overline{X}$ be the standard deviation and mean of X, respectively. Let $Y = aX + b$, where a and b are constants. Then, the mean and standard deviation of Y are given by

$$Y = aX + b$$

$$s_Y = |a| s_X$$

Standardized Data Values

- A standardized data value refers to the number of standard deviations a value is from the mean
- Standardized data values are sometimes referred to as z-scores

Standardized Values

A standardized variable Z has a mean of 0 and a standard deviation of 1.

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

where:

- $\bar{x}$ = original data value
- x = sample mean
- s = sample standard deviation
- z = standard score (number of standard deviations x is from the mean)

YDI

- During a recent week in Europe, the temperature X in Celsius was as follows:

Day	M	T	W	H	F	S	S
X	40	41	39	41	41	40	38

- Based on this

$$\overline{X} = 40$$

$$s_X = 1.14$$

- Calculate the mean and standard deviation in Fahrenheit.
- Calculate the standardized score.