

Data Description

Chapter 3

- 3-1 Introduction
- 3-2 Measures of Central Tendency
- 3-3 Measures of Variation
- 3-4 Measures of Position
- 3-5 Exploratory Data Analysis
- 3-6 Summary

Outline

- Center: a representative or average value that indicates where the middle of the data set is located
- Variation: a measure of the amount that the values vary among themselves
- Distribution: the nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)
- Outliers: Sample values that lie very far away from the majority of other sample values
- Time: Changing characteristics of data over time

Important Characteristics of Data

- The most common characteristic to measure is the center the dataset. Often people talk about the AVERAGE.
- “Average’ when you stop to think about it is a funny concept. Although it describes all of us it describes none of us... While none of us wants to be the average American, we all want to know about him or her.” Mike Feinsilber & William Meed, American Averages
- Examples
 - The average American man is five feet, nine inches tall; the average woman is five feet, 3.6 inches
 - On the average day, 24 million people receive animal bites

Section 3-1 Introduction

- “Average” Is ambiguous, since several different methods can be used to obtain an average
- Loosely stated, the average means the center of the distribution or the most typical case
- Measures of Average are also called the **Measures of Central Tendency**
 - Mean
 - Median
 - Mode
 - Midrange

Is an average enough to describe a data set?

- NO!
 - Consider: A shoe store owner knows that the average size of a man’s shoe is size 10, but she would not be in business very long if she ordered only size 10 shoes
- So, what else do we need to know?
 - We need to know how the data are dispersed—do they cluster around the center or are they spread more evenly throughout the distribution . How spread out are the data points?
- **Measures of Variation** or Measures of Dispersion
 - Range
 - Variance
 - Standard Deviation

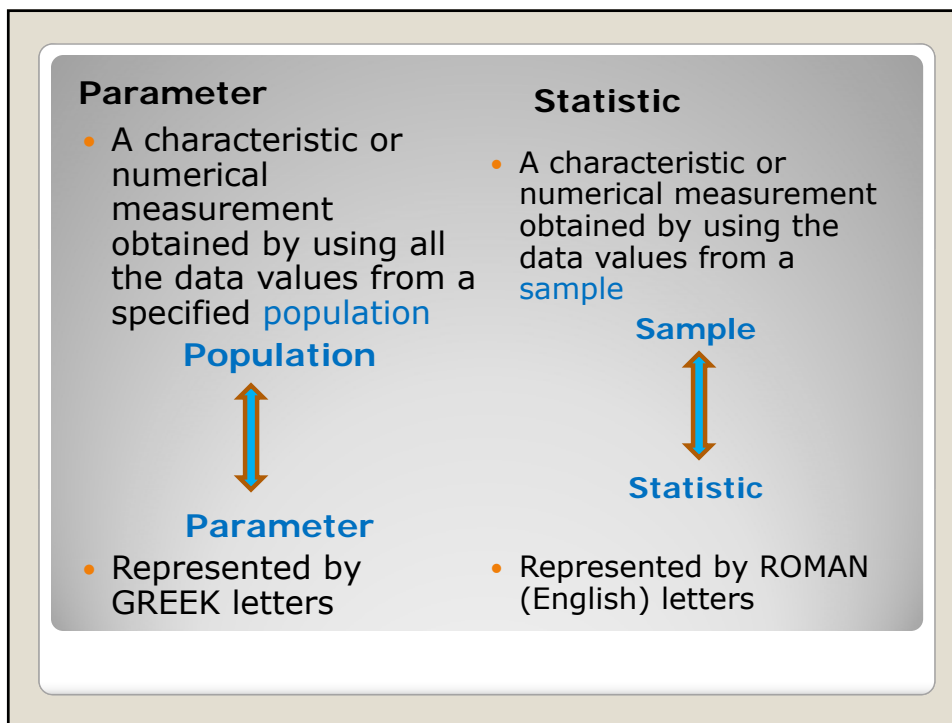
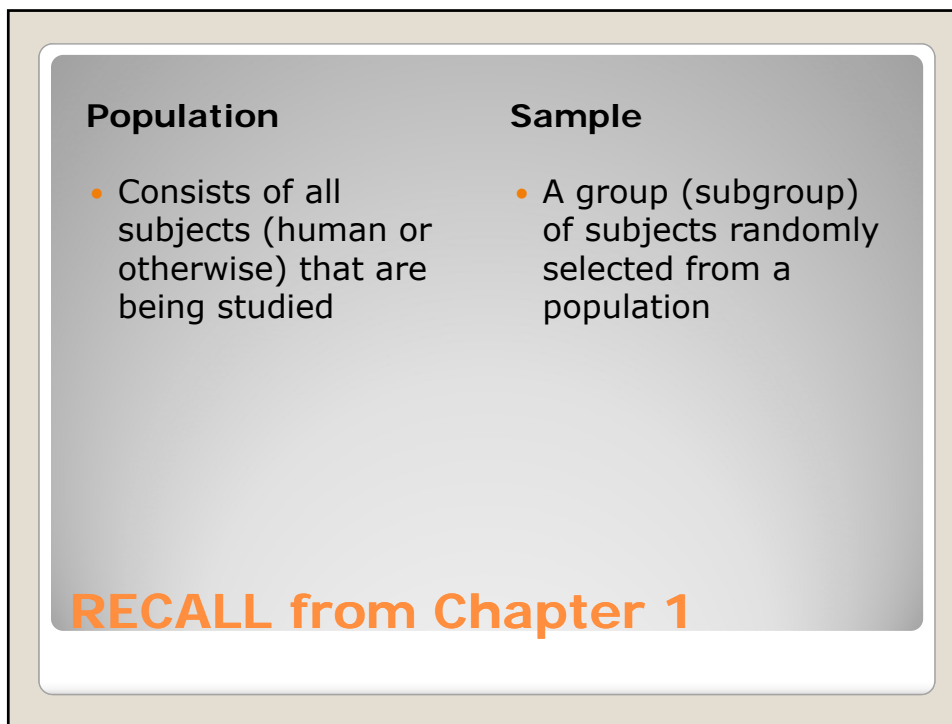
Measures of Variation

- We also need to know the **Measures of Position**
 - Percentiles, Deciles, and Quartiles
 - Used extensively in Psychology and Education, referred to as "Norms"
- These tell use where a specific data value falls within the data set or its relative position in comparison with other data values

Measures of Position

- Objective(s)
 - Summarize data using measures of central tendency, such as the mean, median, mode, and midrange

Section 3-2 Measures of Central Tendency



- When calculating the measures of central tendency, variation, or position, do NOT round intermediately. Round only the final answer
 - Rounding intermediately tends to increase the difference between the calculated value and the actual "exact" value
- Round measures of central tendency and variation to one more decimal place than occurs in the raw data
 - For example, if the raw data are given in whole numbers, then measures should be rounded to nearest tenth. If raw data are given in tenths, then measures should be rounded to nearest hundredth.

General Rounding Guidelines

- Measures of Center is the data value(s) at the center or middle of a data set
- Mean
- Median
- Mode
- Midrange
 - We will consider the definition, calculation (formula), advantages, disadvantages, properties, and uses for each measure of central tendency

Measures of Central Tendency)

- AKA Arithmetic Average
- Is found by adding the data values and dividing by the total number of values
- In general, mean is the most important of all numerical measurements used to describe data
 - Is what most people call an “average”
 - Is unique and in most cases, but is not necessarily an actual data value
 - Varies less than the median or mode when samples are taken from the same population and all three measures are computed for those samples
 - Is used in computing other statistics, such as variance
 - Is affected by extremely high or low values (outliers) and may not be the appropriate average to use in those situations

Mean

Mean ----Formula

- Notation
 - Σ (sigma) denotes the sum of a set of values
 - x is the variable usually used to represent the individual data values
 - n represents the number of values in a *sample*
 - N represents the number of values in a *population*
- Mean of a set of sample values (read as x-bar)

$$\bar{x} = \frac{\sum x}{n}$$
- Mean of all values in a population (read as “mu”)

$$\mu = \frac{\sum x}{N}$$

- The number of highway miles per gallon of the 10 worst vehicles is given:

12 15 13 14
 15 16 17 16
 17 18

- Find the mean.

Mean ---Example

- Is the middle value when the raw data values are arranged in order from smallest to largest or vice versa
- Is used when one must find the center or midpoint of a data set
- Is used when one must determine whether the data values fall into the upper half or lower half of the distribution
- Is affected less than the mean by extremely high or low values
- Does not have to be an original data value
- Various notations:
 - MD, Med,

\tilde{x}

Median

Odd Number of Data Values (n is odd)

- Arrange data in order from smallest to largest
- Find the data value in the "exact" middle

Even Number of Data Values (n is even)

- Arrange data in order from smallest to largest
- Find the mean of the TWO middle numbers (there is no "exact" middle)

Finding the Median

- The number of highway miles per gallon of the 10 worst vehicles is given:

| | | | |
|----|----|----|----|
| 12 | 15 | 13 | 14 |
| 15 | 16 | 17 | 16 |
| 17 | 18 | | |

- Find the median.

Median ---Example

- Measured amounts of lead (in $\mu\text{g}/\text{m}^3$) in the air are given:

5.40 1.10 0.42 0.73

0.48 1.10 0.66

- Find the median

Median – Example #2

- Is the data value(s) that occurs most often in a data set
- Sometimes said to be the most typical case
- Is the easiest average to compute
- Can be used when the data are nominal, such as religious preference, gender, or political affiliation
- Is not always unique. A data set can have more than one mode, or the mode may not exist for a data set
- Has no "special" symbol
- Look for the number(s) that occur the most often in the data set

Mode

- The number of highway miles per gallon of the 10 worst vehicles is given:

12 15 13 14
15 16 17 16
17 18

- Find the mode.

Mode ---Example

- Measured amounts of lead (in $\mu\text{g}/\text{m}^3$) in the air are given:

5.40 1.10 0.42 0.73
0.48 1.10 0.66

- Find the mode.

Mode – Example #2

- Is a rough estimate of the midpoint for the data set
- Is found by adding the lowest and highest data values and dividing by 2
- Is easy to compute
- Gives the midpoint
- Is affected by extremely high or low data values
- Is rarely used
- Is denoted by MR

$$MR = \frac{\text{highest value} + \text{lowest value}}{2}$$

Midrange

- The number of highway miles per gallon of the 10 worst vehicles is given:

| | | | |
|----|----|----|----|
| 12 | 15 | 13 | 14 |
| 15 | 16 | 17 | 16 |
| 17 | 18 | | |

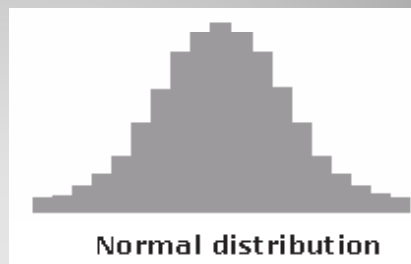
- Find the midrange.

Midrange --- Example

- There is no single best answer to that question because there are no objective criteria for determining the most representative measure for all data sets
- Avoid the term “average”, instead use the actual measure of central tendency that is calculated (mean, median, mode, or midrange)
- Use the advantages and disadvantages stated above to decide which measure of central tendency is best.

Which Measure of Central Tendency is best?

- A comparison of the mean, median, and mode can reveal information about the distribution shape
- RECALL: (p. 56) A bell-shaped (normal) distribution is **symmetric**
 - Data values are evenly distributed on both sides of the mean
 - Unimodal (one peak)
 - Mean \approx Median \approx Mode



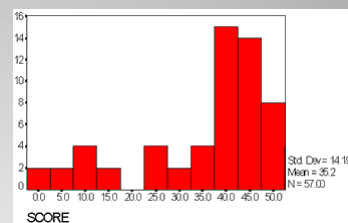
Making Connections



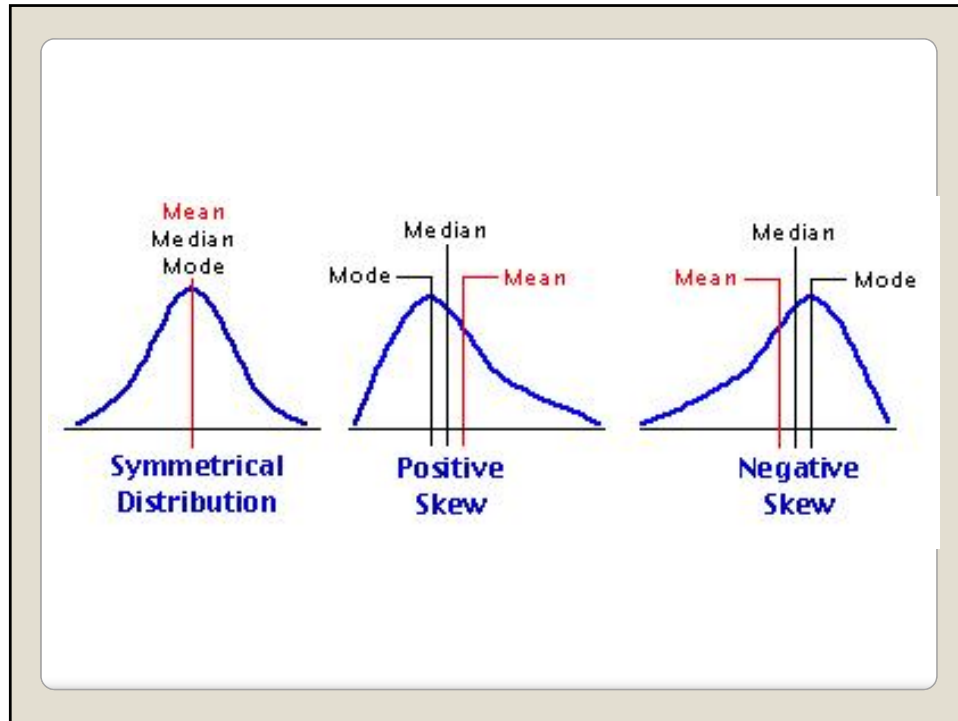
- Right-skewed (or positively) distribution has the majority of data values fall to the left of the mean and cluster at the lower end of the distribution; the "tail" is to the right
- Mode < Median < Mean
- Median is the "center" point

Making Connections

- Left-skewed (or negatively) distribution has the majority of data values to the right of the mean and cluster at the upper end of the distribution, with the tail to the left
- Mean < Median < Mode



Making Connections



- Example- Find the mean, median, mode, and midrange for the ages of NASCAR Nextel Cup Drivers
- Is the distribution symmetric, left-skewed, or right-skewed?

Using Technology

| Ages of NASCAR Nextel Cup Drivers in Years (NASCAR.com) (Data is ranked---Collected Spring 2008) | | | | | | | |
|--|----|----|----|----|----|----|----|
| 21 | 21 | 21 | 23 | 23 | 23 | 24 | 25 |
| 25 | 26 | 26 | 26 | 26 | 27 | 27 | 28 |
| 28 | 28 | 28 | 29 | 29 | 29 | 29 | 30 |
| 30 | 30 | 30 | 31 | 31 | 31 | 31 | 31 |
| 32 | 34 | 35 | 35 | 35 | 36 | 36 | 37 |
| 37 | 38 | 38 | 39 | 41 | 42 | 42 | 42 |
| 43 | 43 | 43 | 44 | 44 | 44 | 44 | 45 |
| 45 | 46 | 47 | 48 | 48 | 48 | 49 | 49 |
| 49 | 50 | 50 | 51 | 51 | 65 | 72 | |