Chapter 3
Statistics for Describing, Exploring, and Comparing Data

3-1 Review and Preview
3-2 Measures of Center
3-3 Measures of Variation
3-4 Measures of Relative Standing and Boxplots

Review
• Chapter 1
  - Distinguish between population and sample, parameter and statistic
  - Good sampling methods: simple random sample, collect in appropriate ways
• Chapter 2
  - Frequency distribution: summarizing data
  - Graphs designed to help understand data
  - Center, variation, distribution, outliers, changing characteristics over time

Preview
• Important Statistics
  - Mean, median, standard deviation, variance
• Understanding and Interpreting these important statistics

Preview
• Descriptive Statistics
  - In this chapter we’ll learn to summarize or describe the important characteristics of a known set of data
• Inferential Statistics
  - In later chapters we’ll learn to use sample data to make inferences or generalizations about a population

Key Concept
Characteristics of center. Measures of center, including mean and median, as tools for analyzing data. Not only determine the value of each measure of center, but also interpret those values.
Measures of Center

- **Measure of Center**
  the value at the center or middle of a data set

- **Arithmetic Mean (Mean)**
  the measure of center obtained by adding the values and dividing the total by the number of values

  What most people call an average.

**Notation**

\[ \bar{x} \] pronounced 'x-bar' and denotes the mean of a set of sample values

\[ \mu \] is pronounced ‘mu’ and denotes the mean of all values in a population

Finding the Median

- **Median** - the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

  often denoted by \( \tilde{X} \) (pronounced ‘x-tildetilde’)

  is not affected by an extreme value - is a resistant measure of the center

**Advantages of Mean**

- Is relatively reliable, means of samples drawn from the same population don’t vary as much as other measures of center
- Takes every data value into account

**Disadvantage of Mean**

Is sensitive to every data value, one extreme value can affect it dramatically; is not a resistant measure of center.

**Finding the Median**

First sort the values (arrange them in order), then follow one of these

1. If the number of data values is odd, the median is the number located in the exact middle of the list.
2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.
### Mode
- **Mode**
  - the value that occurs with the greatest frequency
  - Data set can have one, more than one, or no mode

- **Bimodal**
  - two data values occur with the same greatest frequency

- **Multimodal**
  - more than two data values occur with the same greatest frequency

- **No Mode**
  - no data value is repeated

Mode is the only measure of central tendency that can be used with nominal data.

### Midrange
- **Midrange** - the value midway between the maximum and minimum values in the original data set

\[
\text{Midrange} = \frac{\text{maximum value} + \text{minimum value}}{2}
\]

### Round-off Rule for Measures of Center
- **Carry one more decimal place than is present in the original set of values.**

If original data

\[ \text{rounded to } 3 \text{ decimal places} \]

then results are rounded to

\[ \text{number of decimal places} \]

### Mean from a Frequency Distribution

Assume that all sample values in each class are equal to the class midpoint. Use class midpoint of classes for variable \( x \).

\[
\bar{x} = \frac{\sum (f \cdot x)}{\sum f}
\]

### Weighted Mean
- **When data values are assigned different weights, we can compute a weighted mean.**

\[
\bar{x} = \frac{\sum (w \cdot x)}{\sum w}
\]

Think of calculating your GPA. Does \( w \) represent the number of credits per course? Does \( w \) represent the number of quality points per letter grade?
Best Measure of Center

Symmetric and Skewed

• **Symmetric** - distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

• **Skewed** - distribution of data is skewed if it is not symmetric and extends more to one side than the other.

- Skewed to the left - (also called negatively skewed) have a longer left tail, mean and median are to the left of the mode.

- Skewed to the right - (also called positively skewed) have a longer right tail, mean and median are to the right of the mode.

Skewness

Recap

In this section we have discussed:

- Types of measures of center
  - Mean
  - Median
  - Mode

- Mean from a frequency distribution

- Weighted means

- Best measures of center

- Skewness

Tests of Child Booster Seats

The National Highway Traffic Safety Administration conducted crash tests of child booster seats for cars. Listed below are results from those tests, with the measurements given in hic (standard head injury condition units). According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?

774 649 1210 546 431 612

No. Even though all four measures of center fall within accepted guidelines, all the individual values do not. Since one result exceeds the guidelines, it is clear that all the seats in the population do not meet the requirement.

NOTE: Even for a very large sample (e.g., n=10,000), if one sample value exceeds the guidelines then it is clear that all the seats in the population do not meet the requirement.

FICO Scores

The FICO credit rating scores obtained in a simple random sample are listed below. As of this writing, the reported mean FICO score was 678. Do these sample FICO scores appear to be consistent with the reported mean?

714 751 664 789 818 779 698 836 753 834 693 802

No. Even though all four measures of center fall within accepted guidelines, all the individual values do not. Since one result exceeds the guidelines, it is clear that all the seats in the population do not meet the requirement.
102/17. Years to Earn Bachelor’s Degree Listed below are the lengths of time (in years) it took for a random sample of college students to earn bachelor’s degrees (based on data from the U.S. National Center for Education Statistics). Based on these results, does it appear that it is common to earn a bachelor’s degree in four years? 

S32C

4 4 4 4 4 4 4.5 4.5 4.5 4.5 4.5 4.5 6 6 8 9 13 13 15

Although it is common for a student to complete the degree in 4 years, these data indicate that the typical student takes more than 4 years to complete the degree.

102/19. Bankruptcies Listed below are the numbers of bankruptcy filings in Dutchess County, New York State. The numbers are listed in order for each month of a recent year (based on data from the Poughkeepsie Journal). Is there a trend in the data? If so, how might it be explained?

S32D

59 85 98 106 120 117 97 95 143 371 14 15

Arranged in order, the n=12 scores are: 14 15 59 85 95 97 98 106 117 120 143 371

a. mean = \( \frac{\sum x}{n} = \frac{1320}{12} = 110.0 \)

b. med. = \( \frac{97 + 98}{2} = 97.5 \)

c. mode = (none)

d. m.r. = \( \frac{14 + 371}{2} = 192.5 \)

Yes, there does appear to be a trend in the data. The numbers of bankruptcies seems to steadily increase up to a certain point and then drop off dramatically. It could be that new bankruptcy laws (that made it more difficult to declare bankruptcy) went into effect.

103/24. Find the mean and median for each of the two samples, then compare the two sets of results. Customer Waiting Times Waiting times (in minutes) of customers at the Jefferson Valley Bank (where all customers enter a single waiting line) and the Bank of Providence (where customers wait in individual lines at three different teller windows) are listed below. Determine whether there is a difference between the two data sets that is not apparent from a comparison of the measures of center. If so, what is it?

Jefferson Valley (single line): 6.5 6.6 6.7 6.8 7.1 7.3 7.7 9.3

Bank of Providence (individual lines): 4.2 5.4 5.8 6.2 6.7 7.7 7.8 9.3

L1: Jefferson Valley (single line) 
L2: Providence (individual lines)

L1: Mean = 7.15 Median = 7.2 
L2: Mean = 7.15 Median = 7.2

103/30. Find the mean of the data summarized in the given frequency distribution. Also, compare the computed means to the actual means obtained by using the original list of data values, which are as follows: (Exercise 29) 21.1 mg; (Exercise 30) 76.3 beats per minute; (Exercise 31) 46.7 mi/h; (Exercise 32) 1.911 lb.

S32I S32J

See instructions for use of TI with frequency table http://cfcc.edu/faculty/cmoore/TI83-1-VarStatsFD.htm

= 3070 / 40 = 76.75

See instructions for use of TI with frequency table http://cfcc.edu/faculty/cmoore/TI83-1-VarStatsFD.htm

Aug 29-9,29 PM

Aug 29-9,29 PM

Aug 29-9,29 PM

Aug 29-9,29 PM
If you invest $100 at 10% interest rate, you will have $110.5155782 at the end of the first year.

If you invest $110.5155782 at 5% interest rate, you will have $116.1814352 at the end of the second year.

If you invest $116.1814352 at 2% interest rate, you will have $118.5283909 at the end of the third year.

If you invest $100 at 5.6154334% interest rate compounded daily, you will have $118.3458455 at the end of the third year.

The geometric mean is often used in business and economics for finding average rates of change, average rates of growth, or average ratios. Given a set of numbers, the geometric mean is the nth root of their product. The average growth factor for money compounded at annual interest rates of 10%, 5%, and 2% can be found by computing the geometric mean of 1.1, 1.05, and 1.02. Find the average growth factor. What single percentage growth rate would be the same as having these successive growth rates of 10%, 5%, and 2%? This result is the same as the mean of 10%, 5%, and 2%.

\[
\sqrt[3]{1.1 \times 1.05 \times 1.02} = 1.056154334
\]

In each of the following, describe how the mean, median, mode, and midrange of a data set are affected.

a. The same constant \(k\) is added to each value of the data set.

   (a) They all change by the same amount (increase for positive \(k\); decrease for negative \(k\)).

   (b) They are all multiplied by \(k\).

b. Each value of the data set is multiplied by the same constant \(k\).