CHAPTER PROBLEM

Do women really talk more than men?

A common belief is that women talk more than men. Is that belief founded in fact, or is it a myth? Do men actually talk more than women? Or do men and women talk about the same amount?

In the book *The Female Brain,* neuropsychiatrist Louann Brizendine stated that women speak 20,000 words per day, compared to only 7,000 for men. She deleted that statement after complaints from linguistics experts who said that those word counts were not substantiated.

Researchers conducted a study in an attempt to address the issue of words spoken by men and women. Their findings were published in the article "Are Women Really More Talkative Than Men?" (by Mehl, Vazire, Ramirez-Esparza, Slatcher, and Pennebaker, *Science,* Vol. 317, No. 5834). The study involved 396 subjects who each wore a voice recorder that collected samples of conversations over several days. Researchers then analyzed those conversations and counted the number of spoken words for each of the subjects. Data Set 8 in Appendix B includes male/female word counts from each of the six different sample groups, but if we combine all of the male word counts and all of the female word counts in Data Set 8, we get two sets of sample data that can be compared. A good way to begin to explore the data is to construct a graph that allows us to visualize the samples. See the relative frequency polygon shown below. Based on that figure, the samples of word counts from men and women appear to be very close, with no substantial differences.

When comparing the word counts of the samples of women, one step is to compare the **means** from the two samples. Shown below are the values of the means and the sample sizes. The graph and the sample means give us considerable insight into a comparison of the numbers of words spoken by men and women. In this section, we introduce other common statistical methods that are helpful in making comparisons. Using the methods of this chapter and of other chapters, we will determine whether women actually do talk more than men, or whether that is just a myth.

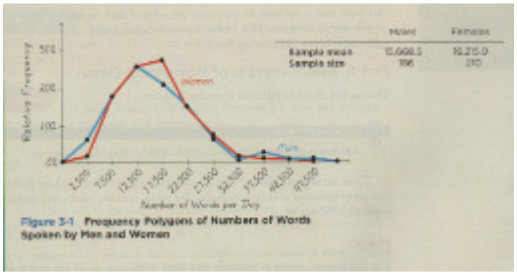

Figure 3-1 Frequency Polygons of Numbers of Words Spoken by Men and Women

MATH 103 CHAPTER 3 HOMEWORK			
3.1	NA	**3.3**	1-5, 7, 9, 11, 13, 17, 20, 21, 25, 29, 31, 33, 35
3.2	1-5, 7, 9, 11, 17, 20, 21, 25, 29, 31, 33	**3.4**	1, 3, 4, 6, 7, 8, 10

3.1 REVIEW AND PREVIEW

Chapter 1 discussed methods of collecting _____ data, and Chapter 2 presented the _____ distribution as a tool for _____ data. Chapter 2 also presented graphs designed to help us understand some _____ of the data, including the _____. We noted in Chapter 2 that when _____, _____, and _____ data sets, these characteristics are usually extremely important: (1) _____, (2) _____, (3)_____, (4) _____, and (5) _____ characteristics of data over time. Upon completing this chapter, you should be able to find the _____, _____, standard _____, and _____ from a data set, and you should be able to clearly understand and _____ such values.

3.2 MEASURES OF CENTER

Key Concept…

In this section, we discuss the characteristic of _____.

In particular, we present measures of center, including _____

and _____, as tools for _____ data.

DEFINITION

A **measure of center** is a value at the _____ or

_____ of a data set.

DEFINITION

The **arithmetic mean (aka mean)** of a set of data is the _____

of _____ found by _____ the

_____ values and _____ the total by the

_____ of data values.

$$\text{mean} = \frac{\sum x}{n} = \text{_____}$$

**One advantage of the mean is that it is relatively _____, so that when samples are selected from the same population, sample means tend to be more consistent than other measures of center. Another advantage of the mean is that it takes every data value into account. However, because the mean is _____ to every value, just one _____ value can affect it dramatically. Because of this fact, we say the mean is not a _____ measure of center.

NOTATION

$\sum$

x

n

N

$\bar{x} = \frac{\sum x}{n}$

$\tilde{x}$

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

Example 1: Find the mean of the following numbers:

17 23 17 22 21 34 27

DEFINITION

The **median** of a data set is the measure of center that is the _____ value when the original data values are arranged in _____ of increasing (or decreasing) magnitude. The median is often denoted _____ (pronounced "x-tilde"). To find the median, first _____ the values, then follow one of these two procedures:

1. If the number of data values is _____, the median is the number located in the exact _____ of the list.
2. If the number of data values is _____, the median is the _____ of the _____ two numbers.

**The median is a _____ measure of center, because it does not change by _____ amounts due to the presence of just a few _____ values.

Example 2:

a. Find the median of the following numbers:

17 23 17 22 21 34 27

b. Find the median of the following numbers

17 23 17 22 34 27

DEFINITION

The **mode** of a data set is the value that occurs with the greatest

_____. A data set can have more than one mode, or no mode.

- When two data values occur with the same greatest frequency, each one is a _____ and the data set is _____.
- When more than two data values occur with the same greatest frequency, each is a _____ and the data set is said to be _____.
- When no data value is repeated, we say there is no _____.

**The mode is the only measure of center that can be used with data at the _____ level of measurement.

Example 3:

a. Find the mode of the following numbers:

17 23 17 22 21 34 27

b. Find the mode of the following numbers

17 23 17 22 21 34 27 22

DEFINITION

The **midrange** of a data set is the measure of center that is the value

_____ between the _____

and _____ values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by two.

$$\text{midrange} = \frac{\qquad\qquad}{\qquad\qquad}$$

**The midrange is rarely used because it is too sensitive to extremes since it uses only the minimum and maximum data values.

Example 4: Find the midrange of the following numbers:

17 23 17 22 21 34 27

ROUND-OFF RULE FOR THE MEAN, MEDIAN, AND MIDRANGE

Carry _____ more decimal place than is present in the original data set. Because values of the mode are the same as some of the original data values, they can be left without any rounding.

MEAN FROM A FREQUENCY DISTRIBUTION

When working with data summarized in a frequency distribution, we don't know the _____ values falling in a particular _____. To make calculations possible, we assume that all sample values in each class are equal to the class _____. We can then add the _____ from each _____ to find the total of all sample values, which we can the _____ by the sum of the frequencies, $\sum f$.

mean from frequency distribution: $\overline{x} = \dfrac{\sum (f \cdot x)}{\sum f}$

Example 5: Find the mean of the data summarized in the given frequency distribution.

Tar (mg) in nonfiltered cigarettes	Frequency
10-13	1
14-17	0
18-21	15
22-25	7
26-29	2

WEIGHTED MEAN

When data values are assigned different weights, we can compute a weighted mean.

$$\text{weighted mean: } \bar{x} = \frac{\sum(w \cdot x)}{\sum w}$$

Example 6: A student earned grades of 92, 83, 77, 84, and 82 on her regular tests. She earned grades of 88 on the final and 95 on her class project. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn?

SKEWNESS

A comparison of the _____,

_____, and _____ can reveal information about the characteristic of **skewness**. A distribution of data is said to

be _____ if it is not _____ and extends more to one side than the other.

3.3 MEASURES OF VARIATION

Key Concept…

In this section, we discuss the characteristic of _____.

In particular, we present measures of variation, such as

_____ _____, as tools for

_____ data.

DEFINITION

The **range** of a set of data values is the _____

between the _____ and the

_____ data value.

DEFINITION

The **standard deviation** of a set of sample values, denoted by s, is a measure of

_____ of values about the _____.

It is a type of _____ deviation of values from the mean that is calculated by using either of the following formulas:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

or

$$s = \sqrt{\frac{n\sum (x)^2 - (\sum x)^2}{n(n-1)}}$$

- The standard deviation is a measure of _____ of all values from the _____.
- The value of the standard deviation is usually _____.
  - It is zero only when all of the data values are the same
  - _____.

- It is never _____.

π Larger values of the standard deviation indicate _____

amounts of _____.

π The value of the standard deviation can increase dramatically with the

inclusion of one or more _____.

π The units of the standard deviation are the same units as the original

_____ values.

General Procedure for Finding Standard Deviation (1st formula)	Specific Example Using the Following Numbers: 2, 4, 5, 16
Step 1: Compute the mean $\overline{x}$	
Step 2: Subtract the mean from each individual sample value	
Step 3: Square each of the deviations obtained from Step 2.	

Step 4: Add all of the squares obtained from Step 3.

Step 5: Divide the total from Step 4 by the number $n-1$, which is one less than the total number of sample values present.

Step 6: Find the square root of the result from Step 5. The result is the standard deviation.

STANDARD DEVIATION OF A POPULATION

The definition of standard deviation and the previous formulas apply to the standard deviation of _____ data. A slightly different formula is used to calculate the standard deviation σ of a _____: instead of dividing by $n-1$, we divide by the population size N.

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

DEFINITION

The **variance (aka dispersion aka spread)** of a set of values is a measure of _____ equal to the _____ of the _____ _____.

Sample variance: s^2

Population variance: σ^2

**The sample variance is an unbiased estimator of the _____ variance, which means that values of s^2 tend to target the value σ^2 of instead of systematically tending to _____ or underestimate σ^2.

USING AND UNDERSTANDING STANDARD DEVIATION

One simple tool for understanding standard deviation is the _____ _____ of _____, which is based on the principle that for many data sets, the vast majority (such as 95%) lie within ____ standard deviations of the _____.

RANGE RULE OF THUMB

Interpreting a known value of the standard deviation: We informally defined _____ values in a data set to be those that are typical and not too _____. If the standard deviation of a collection of data is _____, use it to find rough estimates of the

_____ and _____ values as follows:

minimum "usual " value = (mean) – 2 x (standard deviation)
maximum "usual " value = (mean) + 2 x (standard deviation)

Estimating a value of the standard deviation *s*: To roughly estimate the standard deviation from a collection of _____ sample data, use

$$s \approx \frac{\text{range}}{4}$$

Example 1: Use the range rule of thumb to estimate the ages of all instructors at MiraCosta if the ages of instructors are between 24 and 60.

EMPIRICAL (OR 68-95-99.7) RULE FOR DATA WITH A BELL-SHAPED DISTRIBUTION

Another concept that is helpful in interpreting the value of a standard deviation is the _____ rule. This rule states that for data sets having a _____ that is approximately _____, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean
- About 95% of all values fall within 2 standard deviations of the mean
- About 99.7% of all values fall within 3 standard deviations of the mean

Example 2: The author's Generac generator produces voltage amounts with a mean of 125.0 volts and a standard deviation of 0.3 volt, and the voltages have a bell-shaped distribution. Use the empirical to find the approximate percentage of voltage amounts between

a. 124.4 volts and 125.6 volts

b. 124.1 volts and 125.9 volts

CHEBYSHEV'S THEOREM

The _____(or fraction) of any data set lying within *K* standard deviations of the mean is always _____ _____

$1-\frac{1}{K^2}$, $K \geq 1$. For *K* = 2 or *K* = 3, we get the following statements:

- At least ¾ or 75% of all values lie within 2 standard deviations of the mean.
- At least 8/9 or 89% of all values lie within 3 standard deviations of the mean.

COMPARING VARIATION IN DIFFERENT POPULATIONS

When comparing _____ in _____ different sets of _____, the _____ deviations should be compared only if the two sets of data use the same _____ and _____ and they have approximately the same _____.

DEFINITION

The **coefficient of variation (aka CV)** for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation _____ to the _____, and is given by the following:

Sample : $CV = \frac{s}{\bar{x}} \cdot 100\%$

Population: $CV = \frac{\sigma}{\mu} \cdot 100\%$

Example 3: Find the coefficient of variation for each of the two sets of data, then compare the variation.

The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winnersfrom two different time periods.

BMI (from the 1920s and 1930s): 20.4 21.9 22.1 22.3 20.3 18.8 18.9 19.4 18.4 19.1

BMI (from recent winners): 19.5 20.3 19.6 20.2 17.8 17.9 19.1 18.8 17.6 16.8

3.4 MEASURES OF RELATIVE STANDING AND BOXPLOTS

Key Concept…

In this section, we introduce measures of _____

_____, which are numbers showing the

_____ of data values _____

to the other values within a data set. The most important concept is the

_____, which will be used often in following

chapters. We will also discuss _____ and

_____ , which are common statistics, as well as

a statistical graph called a _____.

BASICS OF *Z*-SCORES, PERCENTILES, QUARTILES, AND BOXPLOTS

A _____ (aka standard value) is found by converting a

value to a _____ scale.

DEFINITION

The ***z* score (aka standard value)** is the number of _____ deviations a given value x is above or below the _____. The z score is calculated by using one of the following:

$$\text{Sample: } z = \frac{x - \bar{x}}{s} \qquad \text{Population: } z = \frac{x - \mu}{\sigma}$$

ROUND-OFF RULE FOR *Z* SCORES

Round z scores to _____ decimal places. This rule is due to the fact that the standard table of z scores (Table A-2 in Appendix A) has z scores with two decimal places.

Z SCORES, UNUSUAL VALUES, AND OUTLIERS

In Section 3.3 we used the _____ _____ of _____ to conclude that a value is _____ if it is more than 2 standard deviations away from the _____. It follows that unusual values have z scores less than _____ or greater than _____.

Example 1: The U.S. Army requires women's heights to be between 58 inches and 80 inches. Women have heights with a mean of 63.6 inches and a standard deviation of 2.5 inches. Find the z score corresponding to the minimum height requirement and find the z score corresponding to the maximum height requirement. Determine whether the minimum and maximum heights are unusual.

PERCENTILES

Percentiles are one type of _____ or _____

which _____ data into groups with roughly the

_____ number of values in each group.

DEFINITION

Percentiles are measures of _____, denoted

_____, which divide a set of data into

_____ groups with about _____ of the values in each group.

The process of finding the percentile that corresponds to a particular data value x is given by the following:

Percentile of x = ----------

Example 2: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years.

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a. Find the percentile corresponding to the given number of points.
   i. 65

ii. 41

b. Find the indicated percentile or quartile.

i. Q_1

ii. P_{80}

iii. P_{95}

NOTATION

n

k

L

P_k

DEFINITION

Quartiles are measures of _____, denoted

_____, which divide a set of data into

_____ groups with about _____ of the values in each group.

FIRST QUARTILE:

SECOND QUARTILE:

THIRD QUARTILE:

5 NUMBER SUMMARY AND BOXPLOT

The values of the three _____ are used for the

_____ _____ _____ and

the construction of _____ graphs.

DEFINITION

For a set of data, the **5-number summary** consists of the _____ value, the _____ _____, the _____ (aka _____ _____), the _____ _____, and the _____ value.

A **boxplot (aka box-and-whisker diagram)** is a graph of a data set that consists of a _____ extending from the _____ value to the _____ value, and a _____ with lines drawn at the _____ _____, the _____, and the _____.

OUTLIERS

When _____ data, it is important to _____ and _____ outliers because they can strongly affect values of some important statistics, such as the _____ and _____ _____. In _____ _____, a data value is an _____ if it is…

above quartile 3 by an amount greater than 1.5 x inner quartile range or **below quartile 1 by an amount greater than 1.5 x inner quartile range**

_____ _____ are called

_____ or _____ boxplots,

which represent _____ as special points. A **modified boxplot** is a boxplot constructed with these modifications: (1) A special

symbol, such as an _____ or point is used to identify

_____ and (2) the solid horizontal line extends only as far as the minimum and maximum values which are not outliers.

Example 3: Use the given sorted values, which are the number of points scored in the Super Bowl for a recent period of 24 years to construct a boxplot. Are there any outliers?

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

PUTTING IT ALL TOGETHER

We have discussed several basic tools commonly used in statistics. When designing an _____, _____ data, reading an article in a professional journal, or doing anything else with data, it is important to consider certain key factors, such as:

- _____ of the data
- _____ of the data
- _____ method
- Measures of _____
- Measures of _____
- _____
- _____
- Changing _____ over _____
- _____ implications