**AP STATISTICS**

Summer FUN

2016-2017 School Year

**Brief Description of Summer Assignment:** This packet that contains information and examples of basic statistics problems, and also exercises for the student to complete. In addition students will read a novel and answer statistical, summary questions.

**Resources Necessary to Complete Assignment:** Graphing calculator and “How to Lie With Statistics” by Darrell Huff

**Objective of Summer Assignment:** For students to gain understanding in basic statistical topics that should be known before starting AP Statistics. Also, students should learn important vocabulary that will be used throughout the year.

**Approximate time commitment during the Summer:** 6 - 8 hours

**Due Date:** Per FCPS regulations, the summer assignment cannot be due until after the second week of school. That said, much of content of this packet is background material designed to prepare you for the first unit of AP Statistics. The first until will be very challenging without having completed this work first. The first quiz of AP Statistics will be on the 3rd day of class. This assignment will be due on the 6th day of class.

**Value of Assignment:** Quiz Grade - out of 65 points.

**Collaboration:** This is an individual project, NOT a group project! You may discuss with others, but everyone is to complete their own original work. In statistics, there are no two answers that are ever identical.

**For questions over the summer, please contact:**

Ms. Jaffa- ebjaffa@fcps.edu
Welcome to AP Statistics! This course is built around four main topics: exploring data, planning a study, probability as it related to distributions of data, and inferential reasoning. Among leaders of industry, business, government, and education, almost everyone agrees that some knowledge of statistic is necessary to be an informed citizen or a productive worker.

**This assignment is due the FIRST day of class and will count as your first quiz grade (65 pts).**

**Summer Packet Guidelines:**

1. Start summer assignment early to allow for time to receive clarification (if necessary) and to complete it by the 6th day of class. If you have any questions, you may contact me. Please do not wait until the last minute to contact me and I will be busy preparing for the upcoming school year and may not be able to respond as quickly to your last minute questions!!

   **E-mail for Questions:** Ms. Jaffa- ebjaffa@fcps.edu

2. I have provided a small resource of information on statistical basics at the end of this packet. However, if you are still stuck and cannot complete the problems on your own it is okay to use math reference books and websites to help. Google is a wonderful thing! You can Google any term or concepts if you want to find more information. I also recommend the following websites:

   - [http://stattrek.com/](http://stattrek.com/)
   - [http://calculator.maconstate.edu/calc_topics.html](http://calculator.maconstate.edu/calc_topics.html) (*Calculator help*)

3. **DO NOT DISCARD THIS SUMMER ASSIGNMENT!** Write, highlight, take notes throughout this packet! Email with any questions you have! There are 3 parts to this packet with 26 practice problems. **Record your answers to the provided answer sheet at the very end. Be sure to review the checklist on page 20 to make sure you’ve done EVERYTHING and purchased any necessary materials for AP Statistics.**

4. **YOU MUST HAVE YOUR OWN GRAPHING CALCULATOR AND BRING IT TO CLASS EVERYDAY!!** A TI-83 is the minimum calculator needed for this course. TI-84 or TI-84 + is better. The TI-84 will be the calculator demonstrated in class. Do not discard the owner’s manual that is included when you purchase a calculator. If you choose not to use the TI-84+ (or TI-83) it will be your responsibility to learn where to located the functions we use in class. Our current textbook does give instructions on using the TI-89.

5. You are to complete your OWN work. Collaboration includes discussion, but not copying anyone’s work. You may discuss, but everyone still does his/her own work.

   Remember, this is an AP Course! **Do not expect this to be an “easy course”**. Although it may not seem as difficult computationally as calculus, it required a great deal of outside reading and homework, and it required a thorough understanding of many abstract concepts. **This is as much a writing course as it is a math course!** **Explaining in complete sentences is required on this assignment and throughout the course.** You cannot just write down numbers and be done, you must use numbers in context – what they mean to that particular problem using appropriate units like feet or $, for example.
PART 1: READING ASSIGNMENT

Read “How to Lie with Statistics” by Darrell Huff. Complete your responses on a separate sheet of paper. You may hand write (must be neat and legible) or type. Please READ the book. Do NOT try to “cherry pick” for the answers. It’s a short book and meant to be a fun, summer read. You will get a lot more out of it if you read it through first then answer the questions. The questions are to “guide” you in your reading, but I really just want thoughtful responses to show me you read the book.

Chapter 1:
1. Describe 3 different types of bias discussed. Give examples.
2. Discuss the characteristics of a good sample. Give specific examples described in the chapter.

Chapter 2:
3. Discuss at least 3 problems that arise with the word “average”. Give examples.
4. What are the differences between mean and median? When should you use each?

Chapter 3:
5. What is the first “little figure” that is discussed? Summarize the major problems that arise?
6. P.39-p.40 discusses a concept called the “law of large numbers.” (it is not labeled this in the book) Summarize this concept.
7. Summarize what went wrong with the polio vaccine experiment?
8. Summarize the meaning, purpose and problems that arise with tests of significance.
9. What is the third “little figure” that is discussed starting on bottom of p.42 – p. 46? Summarize the major problems that arise.
10. Identify three other “little figures” that can mislead readers of statistics.

Chapter 4:
11. Give a brief summary of this chapter, including at least two statistical concepts.

Chapter 5:
12. Name three ways in which graphs can be misleading.

Chapter 6:
13. Summarize (with mathematical reasoning) on how pictures can be used to mislead.

Chapter 7:
14. Summarize the overall concept of a semi-attached figure.
15. There are several “semi-attached figures” mentioned in this chapter. List at least 4 and describe in detail (ie. Use examples) two of them.
16. On p.84, it talks about polio. Can you think of a similar issue today that has a similar effect as described on this page.

Chapter 8:
17. List and describe three different misuses of correlation.
18. Explain why causation cannot be proved with and observational study, but only through a controlled experiment.

Chapter 9:
19. Explain how percentages can be used to mislead. Give at least two examples

Chapter 10:
This is a summary chapter. Briefly discuss as many statistical principles as you can. What are your overall impressions of the book? What are your take-aways?
PART 2 & 3: VOCABULARY & PRACTICE!

Complete Topics 1 through 5. Record your answers on the given answer sheet located at the very end of the packet. Bring the answer sheet and writing assignment to the first day of class.

Topic 1: Basic Vocabulary & Symbols

As you work through the summer assignment, record any definitions with a “★” next to it on a separate sheet of paper. The list of “★” words, vocabulary below, and symbols will be checked on the first day of class.

★ Individual- the objects described by a set of data. Individuals may be people but they may also be animals or things.
★ Variable - any characteristic of an individual. A variable can take different values for different individuals.
★ Data - the actual observations or measurements of a variable
★ Discrete data- quantitative data consists of data that are a listable set of values
★ Continuous data- quantitative data consist of data that can take on any values in the domain of the variable
★ Percentile- data value below which the specified percent of data values occurs
★ 2 number summary of data - mean and standard deviation (for data that symmetric and without outliers)
★ Bivariate Data (two variable) – data that describes a two characteristics of a population

★ Important Symbols you will see throughout this school year

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>µ</td>
<td>Population mean</td>
</tr>
<tr>
<td>\bar{x}</td>
<td>Sample mean</td>
</tr>
<tr>
<td>\rho</td>
<td>Population proportion</td>
</tr>
<tr>
<td>\hat{\rho}</td>
<td>Sample proportion</td>
</tr>
<tr>
<td>\sigma</td>
<td>Standard deviation of the population</td>
</tr>
<tr>
<td>s</td>
<td>Standard deviation of the sample</td>
</tr>
<tr>
<td>n</td>
<td>Number of observations</td>
</tr>
<tr>
<td>R</td>
<td>Correlation coefficient</td>
</tr>
<tr>
<td>\hat{y}</td>
<td>Estimated value of y</td>
</tr>
<tr>
<td>B</td>
<td>Slope of regression</td>
</tr>
<tr>
<td>A</td>
<td>y-intercept of regression</td>
</tr>
<tr>
<td>P(A)</td>
<td>Probability of event A</td>
</tr>
<tr>
<td>P(A \cap B)</td>
<td>Probability of events A and B</td>
</tr>
<tr>
<td>P(A</td>
<td>B)</td>
</tr>
</tbody>
</table>
**Topic 2: Important Comparisons**

*Categorical versus Quantitative Data*

★ Quantitative Variables: takes on numerical values, measurable quantities (Ex. Weight)
★ Categorical Variables: takes on values that are names and descriptions (Ex. Color)

*Population versus Sample*

★ Population: The entire group of individuals intended to be studied (Ex. Every individual living in Fairfax County)
★ Sample: Part of a population that is examined in order to gather information (Ex. 200 individuals living in Fairfax County)

*Statistics versus Parameter*

★ Parameter: number that describes a population (Ex. The true average height of 16-year-old girls, the true proportion of people that wear their seat belts)
★ Statistic: a number that describes a sample (Ex. The average test score for a class, The proportion of students at MHS that take AP classes)

**PRACTICE**

Identify as Quantitative (Q), Categorical (C), Population (Pop), Sample (S), Statistic (ST), or Parameter (P):
(Need help? See quick reference pg. 10)

1. _____ Types of Dog Breeds
2. _____ Names of Students in a Class
3. _____ True mean height of everyone living in California
4. _____ Students in a School
5. _____ Heights of Students in a Class
6. _____ Daily Temperature in a Given Month
7. _____ 50 Dogs in a City
8. _____ Proportion of Test Grades for 30 Students in a Class
9. _____ Favorite Breakfast Cereal
10. _____ Mean amount of Liquid in 100 selected Bottles of a certain Juice
11. _____ 80 Families in a County
12. _____ True Mean Number of Family Members in Wisconsin
13. _____ Colors of Shirts
14. _____ True Proportion of Students Wearing Glasses in a School
**Topic 3: Univariate Distributions**

★ **Distribution**- of a variable tells us what values the variable takes and how often it takes these values

★ **Univariate data (one-variable)** – data that describes a single characteristic of a population

★ **Resistant measure**- a measure that is not sensitive to extreme values

How do we describe univariate distributions? There are 4 characteristics to look for: (SOCS)

1. **Shape** – what form does the distribution take?
2. **Outliers (and other unusual features)** – identify any extreme values and gaps or cluster (clumps) in the distribution
3. **Center** – Where is it centered? Which measure of center should you use?
4. **Spread** – How dispersed is the data? Which measure of spread should you use?

<table>
<thead>
<tr>
<th>Shape</th>
<th>Measures of Center</th>
<th>Measures of Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal</td>
<td>★ Median (M)</td>
<td>★ Interquartile Range (IQR) = Q3 – Q1</td>
</tr>
<tr>
<td></td>
<td>– Resistant to outliers</td>
<td>Q1 is the first 25% of the data</td>
</tr>
<tr>
<td></td>
<td>– Use if distribution is skewed</td>
<td>Q3 is the last 25% of the data</td>
</tr>
<tr>
<td></td>
<td>– calculate by placing observations in numerical order and find middle observation</td>
<td>IQR always goes with the median</td>
</tr>
<tr>
<td>Bimodal</td>
<td>★ Mean (µ)</td>
<td>★ Standard deviation (s) – measures the spread or dispersion about the mean and should only be used when the mean is the chosen measure of center</td>
</tr>
<tr>
<td></td>
<td>– NOT resistant to outliers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Use if distribution is symmetric</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– calculate by taking the average (Σ x_i)/ n</td>
<td></td>
</tr>
<tr>
<td>Multimodal</td>
<td>µ is the PARAMETER (population) mean (use with population standard deviation, σ)</td>
<td>★ Range – measures max – min. The range is a single value. However, when describing a distribution it’s best to give the range and then state the actual values of min and max.</td>
</tr>
<tr>
<td>Uniform</td>
<td>x-bar (x̄) is the SAMPLE mean (use with sample standard deviation, s)</td>
<td></td>
</tr>
</tbody>
</table>

**Skewness (part of SHAPE)**

★ **Symmetric**

★ **Right Skew**

★ **Left Skew**

<table>
<thead>
<tr>
<th>★Symmetric</th>
<th>★Right Skew</th>
<th>★Left Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median = mean</td>
<td>Median &lt; mean “Skewed right, the mean is MIGHT”</td>
<td>Median &gt; mean “Skewed left, the mean is LESS”</td>
</tr>
</tbody>
</table>
*Note symmetric does not have to be “bell shaped”. It just means the data is balanced on the right & left side

**PRACTICE** - Describe each distribution by its center, shape, spread and outliers (if they exist). (Need help? See quick reference pg. 14)

Examples of  
Shape: symmetric, skewed right, skewed left  
Center: approximate where the center of the data is (median) Use frequencies to count and find the middle number or the interval that contains the middle number.  
Spread: Calculate the range of the data  
Outliers: approximate the location of extreme values, if they exist

### Practice Problems

**15.**

- **Shape:** ____________  
- **Center:** ____________  
- **Outliers:** _________  
- **Spread:** ____________

![Graph 15](image)

**16.**

- **Shape:** ____________  
- **Center:** ____________  
- **Outliers:** _________  
- **Spread:** ____________

![Graph 16](image)

**17.**

- **Shape:** ____________  
- **Center:** ____________  
- **Outliers:** _________  
- **Spread:** ____________

![Graph 17](image)

**18.**

- **Shape:** ____________  
- **Center:** ____________  
- **Outliers:** _________  
- **Spread:** ____________

![Graph 18](image)
Topic 4: Normal Distributions (review from Algebra 2)

★Properties of a normal distribution curve
a. Mean and median are the same
b. Bell Shape (perfectly symmetrical) and follows the empirical rule
   68% of everything in the population is within 1 Standard deviation
   95% of everything in the population is within 2 Standard deviations
   99.7% of everything in the population is within 3 Standard deviations
c. The total area under the curve is 1 or 100%

Notation:
For normal distributions, a short notation is helpful. We abbreviate
the normal distribution with mean and standard deviation as ~N(μ, σ). For example, the distribution of
young women’s heights is ~N(64.5, 2.5). This means that the average heights of young women are 64 inches with a
standard deviation of 2.5 inches.

To show the proportion or probability of a certain data falling below, above, or at a specific value can be depicted
through your choice of notation. For example, P(X ≤ μ), means, “the probability that the sample value, X, is less than or
equal to the population mean, μ. It is extremely important to use correct probability notation.

Example: The duration of a flight between 2 cities is normally distributed with a mean of 3.6 hours and a standard
deviation of .15 hour. What proportions of flights will be less than 3.8 hours long?

1) NOTATION: ~N(3.6, .15)
2) P(X ≤ 3.8) = ?
3) First calculate the z-score: \( z = \frac{x - \mu}{\sigma} = \frac{3.8 - 3.6}{.15} = 1.33 \)
   This means that the value 3.8 hours is 1.33 standard deviations above
   the mean.
4) Shaded Area = P(z < 1.33). You can look this up on a z-table or use
   your calculator. This probability is the area under the standard normal
distribution:
   \( \text{normalcdf}(-1E99, 1.33) = .9087 \)
5) Write a contextual statement:
   The proportion of flights that will take less than 3.8 hours,
   when the mean flight length is 3.6 hours, is 90.9%.

(PRACTICE! Use the example above to properly construct your answer, record your answers on the Summer Assignment answer
sheet located at the very end of this packet. For each problem, write out each of the steps as shown in the example.)

19. X is a normally distributed variable with mean μ = 30 and standard deviation σ = 4. Draw a labeled sketch and
determine the probabilities (\( P(x \)) listed     a) P(x < 40)     b) P(x > 21)     c) P(30 < x < 35)
20. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What proportion of cars are travelling more than 100 km/hr?

21. For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

22. The annual salaries of employees in a large company are approximately normally distributed with a mean of $50,000 and a standard deviation of $20,000.
   a) What proportion of people earn less than $40,000?
   b) What proportion of people earn more than $70,000?

**Topic 5: Graphing Distributions**

★ A **dotplot** is a type of graphic display used to compare frequency counts within categories or groups. As you might guess, a dotplot is made up of dots plotted on a graph. Here is how to interpret a dotplot. Each dot can represent a single observation from a set of data, or a specified number of observations from a set of data.

★ A **stemplot** is used to display quantitative data, generally from small data sets (50 or fewer observations).

23. **WEATHER!** The data below gives the number of hurricanes that happened each year from 1944 through 2000 as reported by *Science* magazine.

<table>
<thead>
<tr>
<th>3 2 1</th>
<th>4 3 7 2 3</th>
<th>3 2 5 2 2 4 2 2 6 0 2 5 1 3 1 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1 0</td>
<td>1 2 3 2 1</td>
<td>2 2 3 1 1 1 3 0 1 3 2 1 2 1</td>
</tr>
<tr>
<td>1 0 5 6 1 3 5 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On your answer key, make a dotplot to display these data. Include appropriate labels, title, and scale.

(Need help? See quick reference pg. 12)  (Hint: x-axis is NOT the year. We want to know the frequency of each # of hurricanes per year)

24. **SHOPPING SPREE!** A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (round to the nearest dollar), arranged in increasing order:

3 9 9 11 13 14 15 16 17 17
18 18 19 20 20 20 21 22 23 24
25 25 26 26 28 28 28 28 32 35
36 39 39 41 43 44 45 45 47 49
50 53 55 59 61 70 83 86 86 93

On your answer key, make a stemplot using tens of dollars as the stem and dollars as the leaves. Include appropriate labels, title and key.  (Need help? See quick reference page 12)
25. WHERE DO OLDER FOLKS LIVE? This table gives the percentage of residents aged 65 or older in each of the 50 states.

<table>
<thead>
<tr>
<th>State</th>
<th>Percent</th>
<th>State</th>
<th>Percent</th>
<th>State</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>13.1</td>
<td>Louisiana</td>
<td>11.5</td>
<td>Ohio</td>
<td>13.4</td>
</tr>
<tr>
<td>Alaska</td>
<td>5.5</td>
<td>Maine</td>
<td>14.1</td>
<td>Oklahoma</td>
<td>13.4</td>
</tr>
<tr>
<td>Arizona</td>
<td>13.2</td>
<td>Maryland</td>
<td>11.5</td>
<td>Oregon</td>
<td>13.2</td>
</tr>
<tr>
<td>Arkansas</td>
<td>14.3</td>
<td>Massachusetts</td>
<td>14.0</td>
<td>Pennsylvania</td>
<td>15.9</td>
</tr>
<tr>
<td>California</td>
<td>11.1</td>
<td>Michigan</td>
<td>12.5</td>
<td>Rhode Island</td>
<td>15.6</td>
</tr>
<tr>
<td>Colorado</td>
<td>10.1</td>
<td>Minnesota</td>
<td>12.3</td>
<td>South Carolina</td>
<td>12.2</td>
</tr>
<tr>
<td>Connecticut</td>
<td>14.3</td>
<td>Mississippi</td>
<td>12.2</td>
<td>South Dakota</td>
<td>14.3</td>
</tr>
<tr>
<td>Delaware</td>
<td>13.0</td>
<td>Missouri</td>
<td>13.7</td>
<td>Tennessee</td>
<td>12.5</td>
</tr>
<tr>
<td>Florida</td>
<td>18.3</td>
<td>Montana</td>
<td>13.3</td>
<td>Texas</td>
<td>10.1</td>
</tr>
<tr>
<td>Georgia</td>
<td>9.9</td>
<td>Nebraska</td>
<td>13.8</td>
<td>Utah</td>
<td>8.8</td>
</tr>
<tr>
<td>Hawaii</td>
<td>13.3</td>
<td>Nevada</td>
<td>11.5</td>
<td>Vermont</td>
<td>12.3</td>
</tr>
<tr>
<td>Idaho</td>
<td>11.3</td>
<td>New Hampshire</td>
<td>12.0</td>
<td>Virginia</td>
<td>11.3</td>
</tr>
<tr>
<td>Illinois</td>
<td>12.4</td>
<td>New Jersey</td>
<td>13.6</td>
<td>Washington</td>
<td>11.5</td>
</tr>
<tr>
<td>Indiana</td>
<td>12.5</td>
<td>New Mexico</td>
<td>11.4</td>
<td>West Virginia</td>
<td>15.2</td>
</tr>
<tr>
<td>Iowa</td>
<td>15.1</td>
<td>New York</td>
<td>13.3</td>
<td>Wisconsin</td>
<td>13.2</td>
</tr>
<tr>
<td>Kansas</td>
<td>13.5</td>
<td>North Carolina</td>
<td>12.5</td>
<td>Wyoming</td>
<td>11.5</td>
</tr>
<tr>
<td>Kentucky</td>
<td>12.5</td>
<td>North Dakota</td>
<td>14.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

★Histograms are a way to display groups of quantitative data into bins (the bars). These bins have the same width and scale and are touching because the number line is continuous. To make a histogram you must first decide on an appropriate bin width and count how many observations are in each bin.

The bins for percentage of residents aged 65 or older have been started for you on your answer key.

On your answer key, create a histogram using those bins on the grid. Include appropriate labels, title and scale. (Need help? See quick reference page 13)

★Boxplot, sometimes called a box and whisker plot, is a type of graph used to display patterns of quantitative data. A boxplot splits the data set into quartiles. The body of the boxplot consists of a "box" (hence, the name), which goes from the first quartile (Q1) to the third quartile (Q3).

★5 number summary of data - minimum, maximum, Q1, Q3, and median (use when skewed and outlier resistant)

26. SSHA SCORES
Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

| 154 | 109 | 137 | 115 | 152 | 140 | 154 | 178 | 101 | 103 | 126 | 126 | 137 | 165 | 165 | 129 | 200 | 148 |

and for 20 first-year college men:

| 108 | 140 | 114 | 91  | 180 | 115 | 126 | 92  | 169 | 146 | 109 | 132 | 75  | 88  | 113 | 151 | 70  | 115 | 187 | 104 |

a. Put the data values in order for each gender. Using the chart on your answer sheet, compute numeral summaries for each gender. (Need help? See quick reference pages 11 & 12)

b. Using the minimum, Q1, Median, Q3, and Maximum from each gender, make parallel boxplots to compare the distributions. Be sure to check for outliers. (Read note in reference p. 14) Graph is located on your answer sheet.

c. On your answer sheet, give a summary comparison of each gender.
I. Types of Data

Quantitative (or measurement) Data

These are data that take on numerical values that actually represent a measurement such as size, weight, how many, how long, score on a test, etc. For these data, it makes sense to find things like “average” or “range” (largest value – smallest value). For instance, it doesn't make sense to find the mean shirt color because shirt color is not an example of a quantitative variable. Some quantitative variables take on discrete values, such as shoe size (6, 6 ½, 7, ...) or the number of soup cans collected by a school. Other quantitative variables take on continuous values, such as your height (60 inches, 72.99999923 inches, 64.039 inches, etc.) or how much water it takes to fill up your bathtub (73.296 gallons or 185.4 gallons or 99 gallons, etc.). The “easiest” way to differentiate between discrete and continuous is to think of continuous as measurable items (height, weight, volume, time, etc.). Discrete values fit into a “bins” (either your shoe size is a size 6, size 7, etc...).

Categorical (or qualitative) Data

These are data that take on values that describe some characteristic of something, such as the color of shirts. These values are “categories” of a population, such as M or F for gender of people, Don't Drive or Drive for the method of transportation used by students to get to school. These are examples of binary variables. These variables only have two possible values. Some categorical variables are not binary and have more than two values, such as hair color, brand of jeans, and so on.

Two types of variables:

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>Binary</td>
</tr>
<tr>
<td>Continuous</td>
<td>More than 2 categories</td>
</tr>
</tbody>
</table>
II. Numerical Descriptions of Quantitative Data

Measures of Center

Mean: The sum of all the data values divided by the number (n) of data values.

\[
\text{Mean} = \overline{x} = \frac{\sum x_i}{n} = \frac{4 + 36 + 10 + 22 + 9}{5} = \frac{81}{5} = 16.2
\]

Median: The middle element of an ordered set of data.

\[
\text{Median} = 10
\]

Measures of Spread:

Range: Maximum value - Minimum value

Interquartile Range (IQR): The difference between the 75th percentile (Q_3) and the 25th percentile (Q_1). This is Q_3 – Q_1. Q_1 is the median of the lower half of the data and Q_3 is the median of the upper half. In neither case is the median of the data included in these calculations.

The IQR contains 50% of the data. Each quartile contains 25% of the data.

One Example

Data: 4, 36, 10, 22, 9

\[
\text{Mean} = \overline{x} = \frac{\sum x_i}{n} = \frac{4 + 36 + 10 + 22 + 9}{5} = \frac{81}{5} = 16.2
\]

\[
\text{Median} = 10
\]

Data: 4, 36, 10, 22, 9, 43

\[
\text{Median} = \frac{10 + 22}{2} = 16
\]

\[
\text{Range} = \text{Max.} - \text{Min.} = 36 - 4 = 32
\]

Examples

1. Data: 4, 36, 10, 22, 9

So, the IQR = 29 - 6.5 = 22.5

2. Data: 4 9 10 | 22 36 43

So, the IQR = 36 - 9 = 27
TO MAKE A STEMPLot:

1. Put the data in ascending order. Make a key!
2. Use only the last digit of the number as a leaf (see the numbers to the right of the line – each digit is the last digit of a larger number).
3. Use one, two, or more digits as the stem. (Sometimes, you can truncate data when there are too many digits in each data value – i.e. the number 20,578 would become 20 | 5, where the “20” is in thousands. Note that this is different from rounding.)
4. Place the “stem” digit(s) to the left of the line and the leaf digit to the right of the line. Do this for each data value. You should then arrange the “leaves” in ascending order.
5. Sometimes, there are many numbers with the same “stem.” In this situation it might be useful to break the numbers with the same stem into either two distinct groups (each on a separate line; say, “leaves” from 0 – 4 on the first line and 5 – 9 on the second.) or into five distinct groups as is shown in the graph to the right. Here, the first line for each stem contains all the 0 – 1 leaves, the next line contains the 2 – 3 leaves and so on. This technique is called “splitting the stems.” It is useful in some cases in order to show the shape of the data more clearly.

Five-number summary: consists of Minimum , Q1, Median, Q3, and Maximum. To find these statistics, enter the data you have into your calculator using the list function:

STAT → ENTER → type the data into L1. If you make a mistake, you can go to the error and DELETE. If you forget an item, you can go to the line below where it is supposed to be and press 2nd DEL to insert it. To find the each value of the five-number summary, go to 2nd STAT → MATH → 5 and then type in L1 by typing 2nd → 1

NOTE: If the lists you are using already have numbers in them before you start, you can clear them this way: Arrow up (↑ ) to the line where L1 is shown. Press CLEAR, then the down arrow (↓ ).

III. Graphical Displays of Univariate (one variable) Data

• Dotplot
• Boxplot (Box and Whiskers)
• Stemplot (Stem and Leaf)
• Histogram

To make a Dotplot:

1. Draw and label a number line so that all the values in your dataset will fit.
2. Graph each of the data values with a dot.
   Be sure to line the dots up vertically as well as horizontally so that you can really see the shape of the graph.

Student GPA’s

GPA
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Student GPA’s

Dot Plot

TO MAKE A STEMPLot:
Stemplot of Student GPAs

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>444</td>
</tr>
<tr>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>1</td>
<td>8888999</td>
</tr>
<tr>
<td>2</td>
<td>0000000000000011111111</td>
</tr>
<tr>
<td>2</td>
<td>3333333333333333333333</td>
</tr>
<tr>
<td>2</td>
<td>44444444444444444555555555</td>
</tr>
<tr>
<td>2</td>
<td>6666666666666666677777</td>
</tr>
<tr>
<td>2</td>
<td>88888888899999999999999</td>
</tr>
<tr>
<td>3</td>
<td>0000000000000000000001111111</td>
</tr>
<tr>
<td>3</td>
<td>22333333333333333</td>
</tr>
<tr>
<td>3</td>
<td>44444444444444445</td>
</tr>
<tr>
<td>3</td>
<td>666666666666666667</td>
</tr>
<tr>
<td>3</td>
<td>888888888889999999999999</td>
</tr>
</tbody>
</table>

Key:
3|4 = 3.4

To make a Boxplot:

1. **Draw and label a number line** that includes the minimum and the maximum values for the set of data.
2. Calculate the five-number summary and make a dot for each of these summary numbers above the number line.
3. Draw a line between the 1st and 2nd dot, showing the “lower quartile”; and then draw a line from the 4th to the 5th dot to show the “upper quartile.” These are commonly called the “whiskers.”
4. Draw a rectangular box from the 2nd to the 4th dot and draw a line through the box on the middle dot – the median.

**NOTE:** In AP Statistics, a “modified boxplot” is used. This shows any “outliers.” An outlier is a data point that does not fit the pattern of the rest of the data. When your calculator or computer software graphs a modified boxplot, an algorithm is used to determine what it takes to “not fit the pattern of the rest of the data.” This algorithm is:

\[1.5 \times (\text{IQR}) \text{ away from the “box” part of the graph. (above and below the box). These outliers are shown with dots or stars, or any other small symbol. Any value above } Q_3 + 1.5\text{IQR} \text{ or any value below } Q_1 - 1.5\text{IQR}\]
Categorical Data:

- Bar Graph
- Circle Graph (Pie Chart)

I'm assuming that you already know how to make these two types of graphs.

To make a histogram:

1. Put the data into ascending order.
2. Decide upon evenly spaced intervals into which to divide the set of data (such as 0, 10, 20, 30, etc.) and then count the number of values that fall within each interval. This number is called the “frequency.” If you divide each of these frequencies by the size of the data set, \( n \), making percents, then you have what are called “relative frequencies.”
3. Draw and label a 1st quadrant graph using scales appropriate for the data. Be sure to include a title for the x- and for the y-axes.
4. Graph the frequencies that you calculated in step 2.
IV. Assessing the Shape of a Graph

There are two basic shapes that we will examine: **Symmetric** and **Skewed**.

**Symmetric:** One can tell if a graph is symmetric if a vertical line in the “center” divides the graph into two fairly congruent shapes. (A graph does not have to be “bell-shaped” to be considered symmetric.)

Mean ~ Median in a symmetric distribution

**Skewed:** One can tell that a graph is skewed if the graph has a big clump of data on either the left (skewed right) or on the right (skewed left) with a tendency to get flatter and flatter as the values of the data increase (skewed right) or decrease (skewed left). A common misconception is that the “skewness” occurs at the big clump. The direction of the skew = direction of the “tail”.

Relationship between Mean and Median in a skewed distribution:

Skewed Left, the mean is Less.
Skewed Right, the mean is Might.

**Gathering Information from a Graphical Display**

The first thing that should be done after gathering data is to examine it graphically and numerically to find out as much information about the various features of the data as possible. These will be important when choosing what kind of procedures will be appropriate to use to find out an answer to a question that is being investigated.

The features that are the most important are Shape, Outliers, Center, Spread: **SOCS**. Most of these can only be seen in a graph. Outliers are stated as “potential outliers” unless you have data and can prove them mathematically using $1.5 \times \text{IQR}$ rule. Use complete sentences and state what you are using to describe each. For example: Center described by the median is 45mg. Spread described by the IQR is 15, where 50% of the data falls between 40 and 55mg.
V. Calculating normal probabilities using a TI-83 and TI-84 graphing calculator

**DISTR 2: normalcdf(**
- The input for the command is the minimum value for area, the maximum value, the mean, \( \mu \), the standard deviation, \( \sigma \).
- The keystrokes below calculate the area between -1 and 1 for a normal distribution with \( \mu=0 \) and \( \sigma=1 \).
  - (the STANDARD normal distribution)
- Press ENTER.
- The shaded image will not appear, on the screen, but is the sketch that should accompany the solution.

**DISTR 3: invNorm(**
- The input for the command is the area as a decimal, the mean, \( \mu \), the standard deviation, \( \sigma \).
- The syntax shown calculates the observation with an area 0.025 or 2.5% below its value.
  - First a z-score for \( N(\mu=0, \sigma=1) \) the STANDARD normal distribution.
SUMMER ASSIGNMENT CHECK-LIST

Have you completed the following....

Part 1: Reading Assignment

- Read the book and answer each question with thoughtful responses. Typed is preferred, but as long as it’s legible I don’t mind hand written.

Part 2: Vocabulary & Symbols

- Record the vocabulary and symbols onto a separate piece of paper. This will be checked for completion and then you will keep it in your binder to study from.

Part 3: Complete the practice problems

- Complete the practice problems in the packet.
  Record your answers on the Summer Break Packet Answer Sheet.
TOPIC 2: IMPORTANT CHARACTERISTICS (Page 5)
Identify as Quantitative (Q), Categorical (C), Population (Pop), Sample (S), Statistic (ST), or Parameter (P):

1. _____  
2. _____  
3. _____  
4. _____  
5. _____  
6. _____  
7. _____  
8. _____  
9. _____  
10. _____  
11. _____  
12. _____  
13. _____  
14. _____

TOPIC 3: UNIVARIATE DISTRIBUTIONS (Page 7)
Describe each distribution by its center, shape, spread and outliers (if they exist, identify where they are).

15. Shape _________  
Outliers _________  
Center _________  
Spread _________

16. Shape _________  
Outliers _________  
Center _________  
Spread _________

17. Shape _________  
Outliers _________  
Center _________  
Spread _________

18. Shape _________  
Outliers _________  
Center _________  
Spread _________

TOPIC 4: NORMAL DISTRIBUTIONS (Page 8)
Answer question #19 through 22. Use the example on page 8 to properly construct your answers. Include notation, work, curve and concluding sentence in context. #19 you do not need to include a sentence.

19. a)  
b)
19. c)

20.

21.

22. a) b)

**TOPIC 5: GRAPHING DISTRIBUTIONS (Pages 9 & 10)**

23. Construct a dotplot to display these data. Include appropriate labels, title, and scale.
24. Construct a stemplot using tens of dollars as the stem and dollars as the leaves. Include appropriate labels, title and key.

25. Create a histogram using those bins on the grid below. Include appropriate labels, title and scale.
26. 

a. Put the data values in order for each gender. Compute numeral summaries for each gender.

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th></th>
<th>Men</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>Minimum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td></td>
<td>Q1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>Median</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td></td>
<td>Q3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td></td>
<td>Maximum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>Range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQR</td>
<td></td>
<td>IQR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Using the minimum, Q1, Median, Q3, and Maximum from each gender, make parallel boxplots to compare the distributions. Don’t forget to check for outliers!

```
[Boxplot 1]
[Boxplot 2]
[Boxplot 3]
```

c. Give a summary comparison of each gender. Write complete sentences and be sure to address SOCS!!! (See quick reference p.15 for an example)

27. Estimate how long did it take you to complete this summer assignment? ________________