Hello Statisticians!

I hope you are having a fantastic summer break! As we prepare for the next school year and AP Statistics, you may be wondering “What exactly is AP Statistics and what did I get myself into?”

Here is what you need to know:

AP Statistics is a college-level statistics class taught in high school. Statistics is much more than making graphs and calculating mean, median, and mode. Rather, it is using data to evaluate claims and make predictions. For example:

- Can you smell Parkinson’s Disease?
- Does Beyonce write her own lyrics?
- Which cookie brand has the most chocolate chips?

This class is a fresh start! AP Statistics isn’t a typical math class. You won’t need to factor a polynomial or prove that triangles are congruent. **But, you will need good communication and critical thinking skills.**

It is a great preparation for college! The skills you learn will help you in a wide variety of college majors and AP courses, including psychology and biology.

We encounter statistics in our daily lives! Examples and exercises are based on real-world studies in a variety of fields. For more information about the application of statistics in the real world check out: thisisstatistics.org

I am looking forward to working with you this year. Your summer packet will preview some of the statistical graphs and numerical statistics used in the course. Watch the videos and complete the practice for each section. See you in August!

What Is Statistics?

Watch Against All Odds: What is Statistics?

Statistics is the art and science of gathering, organizing, analyzing and drawing conclusions from data. And without rudimentary knowledge of how it works, people can't make informed judgments and evaluations of a wide variety of things encountered in daily life.

Stemplots

Watch Against All Odds: Stemplots

The art of looking at stemplots intelligently is as important as the skill of making them. In looking at any distribution, always look first for the overall pattern of the distribution and then for any striking deviations from that pattern. In sizing up the overall pattern, look for and try to describe the following:

- center and spread
- one peak or several
- a regular shape, such as symmetric

For now, identify a center by looking at the stemplot and selecting a number that appears to best measure the middle of the distribution. (In later units, we will cover specific measures of center such as the mean and median).

Key Terms

A **variable** describes some characteristic of interest that can vary in value. Some variables are **categorical** (soldiers' gender – male or female). Others are **quantitative** (soldiers' head circumference or foot length).

The **distribution** of a variable describes the possible values the variable takes and how often it takes these values. Stemplots are one way to graph the distribution of a quantitative variable.

Shape, center, and spread describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **unimodal** (single peak) or **symmetric** (one side is the mirror image of the other).

Outliers are data values that lie outside the overall pattern of the distribution. Always look for gaps in the data and outliers and try to explain them.

A **stemplot** (or **stem-and-leaf plot**) is a useful tool for conveying the shape of relatively small datasets and identifying outliers. It consists of two columns, one for the stems and the other for the leaves (often separated by a vertical line).

Practice #1

Below are the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees, 1920 – 1934.

54 59 35 41 46
25 47 60 54 46
49 46 41 34 22

a. Make a stemplot of the home run data. Then use your stemplot to answer questions (b) and (c).

b. Describe the shape of the distribution. Is it roughly symmetric or not? Is it unimodal (single peak) or multimodal (more than one peak)?

c. What is the center (this is the number of home runs the Babe hit in a typical year)?

d. Ruth's record of 60 home runs in 1927 stood for more than 30 years. Is 60 an observation that falls outside the pattern of the other observations and hence could be considered an outlier?

Histograms

Watch Against All Odds: Histograms

A **frequency distribution** is one method of organizing and summarizing data in a table. The basic idea behind a frequency distribution is to set up categories (class intervals), classify data values into the categories, and then determine the frequency with which data values are placed into each category.

Although a frequency distribution table is a useful tool for extracting information from data, a **histogram** can often convey the same information more effectively.

In describing a histogram, we first look for the overall pattern of the distribution. In sizing up the overall pattern, look for the following:

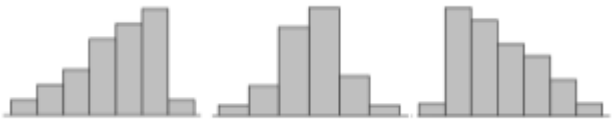
- center and spread;
- one peak or several (unimodal or multimodal);
- a regular shape, such as symmetric or skewed.

Key Terms

A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.



Practice #2

The duration of 40 phone calls (in minutes) for technical support is given below.

12.0	3.3	0.5	48.7	16.7	1.2	14.8	8.2	9.0	5.7
11.5	17.5	3.2	20.8	7.3	8.0	0.2	51.2	3.3	5.2
12.3	24.5	13.3	7.7	13.5	4.3	13.7	10.7	18.8	15.7
3.2	38.7	16.2	23.3	9.7	4.7	6.5	0.5	45.1	5.3

a. Complete the frequency distribution table for the call duration data.

Duration (minutes)	Frequency	Percent
0 – 6		
6 – 12		
12 – 18		
18 – 24		
24 – 30		
30 – 36		
36 – 42		
42 – 48		
48 – 54		

b. What percentage of phone calls lasted less than 12 minutes?

c. What percentage of calls lasted a half hour or more?

d. Represent the frequency distribution with a histogram. Use a percent scale on the vertical axis.

e. Describe the shape of the distribution. Are there any gaps in the data? Outliers?

Measures of Center

Watch Against All Odds: Measures of Center

A graph, such as a stemplot or a histogram, can show us the overall pattern of the data and any striking deviations, such as outliers. The next step is to give a numerical description of some important aspects of the data. The median, mean, and mode are three numerical measures that use different ideas of "center."

We have discussed three measures of center or location, the median, mean, and mode. How do you decide which is best for a given situation? In choosing an appropriate measure of center, start with a graphic display of the data. Consider the overall shape of the data and deviations from that shape before deciding whether to use the mean or median to summarize the location of the data. Keep in mind that the median is a **resistant** measure of center, which is not influenced by a few extreme data values whereas a few extreme outliers can pull the mean in the direction of the extreme values.

For roughly symmetric distributions the mean and median will be close in value. For highly skewed data, or data with extreme outliers, the median is generally the better choice for a measure of the center or location of the data. For data sets with multiple peaks, the modes may give a better indication of location.

Key Terms

The **median** gives the midpoint of a set of data – it separates the upper half of the data from the lower half. To calculate the median, order the data from smallest to largest and count up $(n + 1)/2$ places in the ordered list.

The **mean** is the arithmetic average or balance point of a set of data.
To calculate the mean, sum the data and divide by the number of data: $\bar{x} = \frac{\sum x}{n}$

The **mode** is the data value that occurs most frequently.

A **resistant measure** of some aspect of a distribution (such as its center) is relatively unaffected by a small subset of extreme data values.

Practice #3

Here are the starting salaries, in thousands of dollars, offered to the 20 students who earned degrees in computer science in 2011 at a university.

63 56 66 77 50 53 78 55 90 65 64 69 59 76 48 54 49 68 51 50

a. Make a graph to describe the distribution and write a brief description of its important features.

b. Find the median salary.

c. Find the mean salary.

d. Find the mode of the salaries.

e. Is the mean about the same as the median or not? What feature of the distribution explains the difference between the mean and the median? Is the mode a good measure of the center for these data?

Boxplots

Watch Against All Odds: Boxplots

The topic of this unit is the **five-number summary** and its associated graph, the **box-and whisker plot** or **boxplot**. The five-number summary of a set of data consists of the minimum, **first quartile**, median, **third quartile** and maximum. In its basic form, a **boxplot** (or **box-and-whisker plot**) is a graphical display of the five number summary. It can be drawn either vertically or horizontally depending on your preference.

Key Terms

A **five-number summary** of a set of data consists of the following:

minimum, first quartile (Q_1), median, third quartile (Q_3), maximum.

The **first quartile**, Q_1, is the one-quarter point in an ordered set of data. To compute Q_1, calculate the median of the lower half of the ordered data. The **third quartile**, Q_3, is the three quarter point in an ordered set of data. To compute Q_3, calculate the median of the upper half of the ordered data.

A basic **boxplot** (or **box-and-whisker plot**) is a graphical representation of the five-number summary. A modified boxplot indicates outliers and adjusts the whiskers.

The **interquartile range** or **IQR** measures the spread of the middle half of the data:

$$\text{IQR} = Q_3 - Q_1$$

The **range** measures the spread of the data from its extremes:

range = maximum – minimum

Practice #4

The average SAT Critical Reading scores for each state and the District of Columbia (so 51 total), ordered from smallest to largest, appear below.

469 469 479 479 482 485 485 487 489 493 493 493 494 495 495 499 499 509 512
513 514 515 515 517 520 523 523 539 539 542 546 548 555 563 564 568 570 571
572 575 576 580 583 584 585 586 590 592 593 596 599

The average SAT Math scores, ordered from smallest to largest, appear below.

457 469 487 489 490 490 493 496 499 500 501 501 501 502 502 508 509 511 513 515 516
518 521 523 525 527 529 537 539 541 541 543 545 550 559 565 568 569 570 572 573 591
591 591 593 602 604 606 608 612 617

a. Determine a five-number summary of the average SAT Critical Reading scores.

b. Determine a five-number summary of the average SAT Math scores.

c. Make boxplots to compare the distribution of the Critical Reading and Math scores. (In order to make comparisons, the boxplots must be on the same scale and positioned so that comparisons are easily made.)

Chapter 1 Notes Section:

Please answer the following questions as completely as possible. The answers to these questions will serve as your notes for Chapter 1.

Section 1.1: Analyzing Categorical Data

1. Give an original example of a categorical variable and a quantitative variable for a given individual.

2. What is the difference between a *frequency table* and a *relative frequency table*?

3. What are ways in which to display data of a **single** *categorical variable*?

4. What are ways in which to display data of **two** *categorical variables*?

5. What are the necessary components in constructing a bar graph for a categorical variable?

6. How can pictographs be dangerous?

7. What is the difference in calculations between a *marginal distribution* and a *conditional distribution*?

8. Describe the difference between a *bar graph* and a *segmented bar graph*.

9. What is the 4-step process? List all steps below.

Section 1.2: Displaying Quantitative Data with Graphs

1. What are ways in which to display a **single** *quantitative variable*?

2. What are ways in which to display **two** *quantitative variables*?

3. What are the necessary components in constructing a dotplot for a quantitative variable?

4. What do you need to describe about the distribution of a quantitative variable?

5. What are some key terms used to describe the *shape* of a distribution?

6. What are some key terms used to describe the *center* of a distribution?

7. What are some key terms used to describe the *spread* of a distribution?

8. What are the necessary components in constructing a stem-and-leaf plot for a quantitative variable?

9. What are the necessary components in constructing a histogram for a quantitative variable?

10. How do you create a histogram on the calculator (for TI-83/84)? Describe the steps below.

11. What are four things to be cautious of when using histograms?

Section 1.3: Describing Quantitative Data with Numbers

1. What is the formula for finding the sample mean, $\bar{x}$? Be sure to add this term to #6 from Section 1.2.

2. What is the difference between μ and $\bar{x}$?

3. What do we mean when we say the mean is *not* a **resistant measure** of center?

4. How do you find the median of a set of data (be sure to describe the method for when n is even and odd)? Be sure to add this term to #6 from Section 1.2.

5. What do we mean when we say the median is a **resistant measure** of center?

6. Use the symbols <, >, and = to fill in the blanks below:
    a. When a distribution is skewed to the right, mean ___ median
    b. When a distribution is skewed to the left, mean ____ median
    c. When a distribution is roughly symmetric, mean ___ median

7. Interpret both the median and mean – which one describes the "average" value versus the "typical" value?

8. Which measure of center is a more accurate depiction for a set of data: the mean or the median?

9. What is the first quartile (Q_1) of a set of data? How is it obtained? Be sure to describe the method for when n is even and odd.

10. What is the third quartile (Q_3) of a set of data? How is it obtained? Be sure to describe the method for when n is even and odd.

11. What is the IQR of a set of data? How is it obtained? Be sure to add this term to #7 from Section 1.2.

12. Are the quartiles and interquartile range resistant measures of spread? Explain.

13. What is the formula for identifying outliers?

14. How is a boxplot created (by hand)? For what type of data is a boxplot used to display?

15. How do you create a boxplot on the calculator (for TI-83/84)? Describe the steps below.

16. What is the formula for the sample standard deviation? For the sample variance?

17. Are s_x and s_x^2 resistant measures of spread?

18. Come up with a generic sentence you can use for interpreting the standard deviation.

19. How do you find the mean, standard deviation, variance, Q_1, median, and Q_3 using 1-VAR STATS on the calculator? Describe the steps below.