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!

Ms. Kroenke

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).

2	2 5
3	4 5
4	1 1 6 6 6 7 9
5	4 4 9
6	0

KEY: 2|2 = 22 home runs

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

The distribution of Babe Ruth's home runs with the NY Yankees is roughly symmetric and unimodal.

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

The number of home runs the Babe hit in a typical year is about 46.

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?

The maximum number of Home Runs is 60, but the Babe also had 59 home runs for one year and 54 home runs for two years with the NY Yankees. Thus 60 does not fall outside the pattern and is not 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:

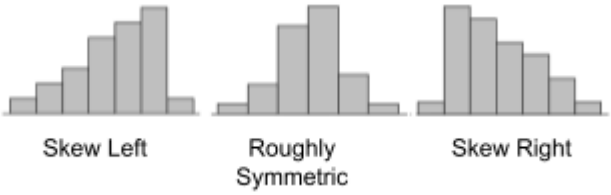
- center and spread;
- one peak or several (unimodal or multimodal);
- a regular shape, such as symmetric or skewed.

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	13	32.5%
6 – 12	9	22.5%
12 – 18	10	25%
18 – 24	3	7.5%
24 – 30	1	2.5%
30 – 36	0	0%
36 – 42	1	2.5%
42 – 48	1	2.5%
48 – 54	2	5%

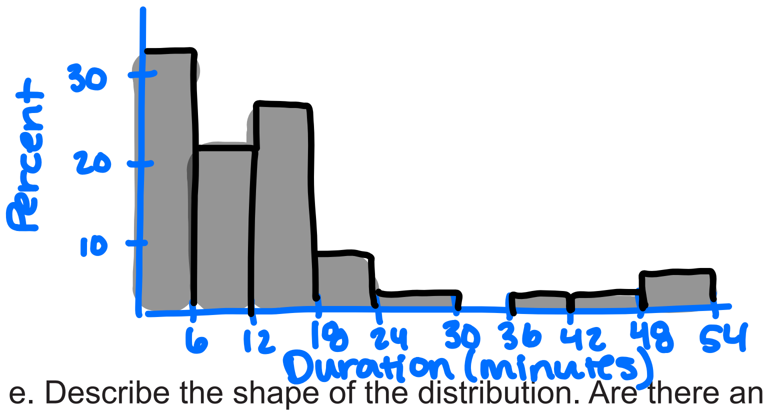
b. What percentage of phone calls lasted less than 12 minutes?

55%

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

10%

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?

The distribution of call duration (minutes) for technical support is right skew. There is a gap in the data between 30 and 36. The majority of calls last less than 30 minutes, but there are a few calls that last longer than 36 minutes. There does not appear to be any 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.

4 | 89
5 | 00134569
6 | 345689
7 | 678
8 |
9 | 0

KEY: 4|8 = 48 thousand dollars

b. Find the median salary. $\frac{59+63}{2} = 61$ thousand dollars ($61,000)

c. Find the mean salary. $\frac{1241}{20} = 62.05$ thousand dollars ($62,050)

d. Find the mode of the salaries.

50 thousand ($50,000)

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?

The mean starting salary is slightly higher than the median. The starting salary of $90,000 is a potential outlier / the distribution of starting salaries is right skew which causes the mean to increase. The mode is not a good measure for center because it closer to the minimum.

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:

$$\text{range} = \text{maximum} - \text{minimum}$$

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.

$\overset{\text{Min}}{469}$ 469 479 479 482 485 485 487 489 493 493 493 $\overset{\text{Q1}}{\underline{494}}$ 495 495 499 499 509 512 513 514 515 515 517 520 $\overset{\text{Med}}{\underline{523}}$ 523 539 539 542 546 548 555 563 564 568 570 571 $\underset{\text{Q3}}{\underline{572}}$ 575 576 580 583 584 585 586 590 592 593 596 $\underset{\text{Max}}{\underline{599}}$

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

$\overset{\text{Min}}{\underline{457}}$ 469 487 489 490 490 493 496 499 500 501 501 $\overset{\text{Q1}}{\underline{501}}$ 502 502 508 509 511 513 515 516 518 521 523 525 $\underset{\text{Med}}{\underline{527}}$ 529 537 539 541 541 543 545 550 559 565 568 569 $\overset{\text{Q3}}{\underline{570}}$ 572 573 591 591 591 593 602 604 606 608 612 $\underset{\text{Max}}{\underline{617}}$

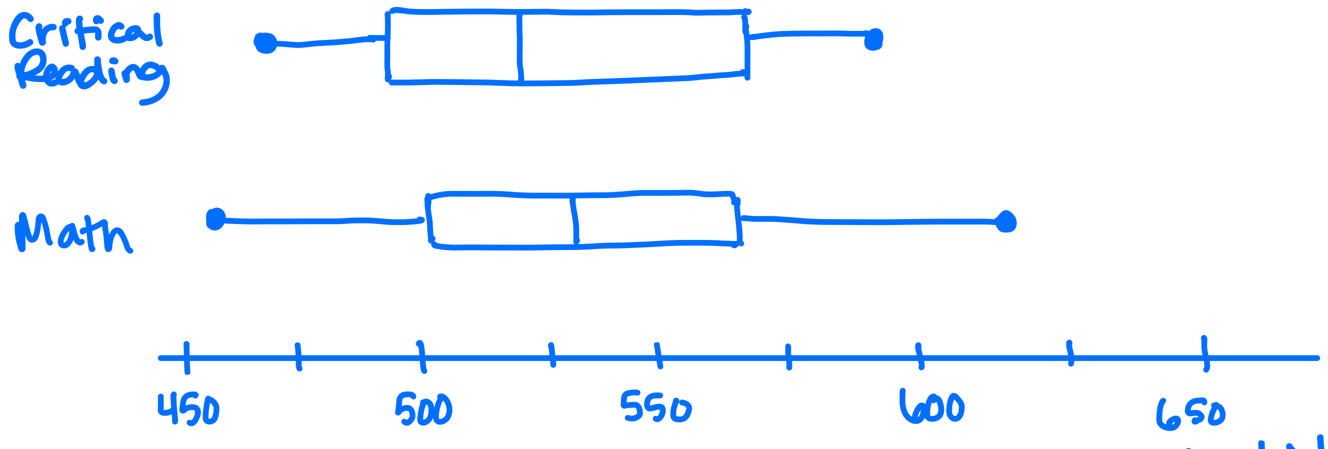
a. Determine a five-number summary of the average SAT Critical Reading scores.

Min: 469 Q1: 494 Med: 523 Q3: 572 Max: 599

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

Min: 457 Q1: 501 Med: 527 Q3: 570 Max: 617

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.)



The median of the average SAT Math scores is slightly higher than the median of the average critical reading score. The range for math is also higher than the range for critical reading, although the IQR for critical reading is larger than the IQR for math.