Unit 5: Boxplots

SUMMARY OF VIDEO

Hot dogs are an American icon – and we eat billions of them every year. It seems like there is a hot dog to fit just about every taste out there… all beef, some pork, turkey, skinless, even tofu for the vegetarian hot dog lover. Not all hot dogs are created equal though, at least in terms of calories. The calorie count varies quite a bit depending on the type of hot dog and also from brand to brand of a given type. The video gives us an inside view at Vallid Labs in Agawam, Massachusetts, and the calorie counting techniques they use to find the number of calories in particular hot dogs. After turning a hot dog into mush and then applying a series of treatments with acids and bases, distillations and titrations, the number of grams of fat, protein, and carbohydrates per hot dog is determined. Then using the information from the Nutrition Facts in Figure 5.1, we can determine a hot dog's calories.

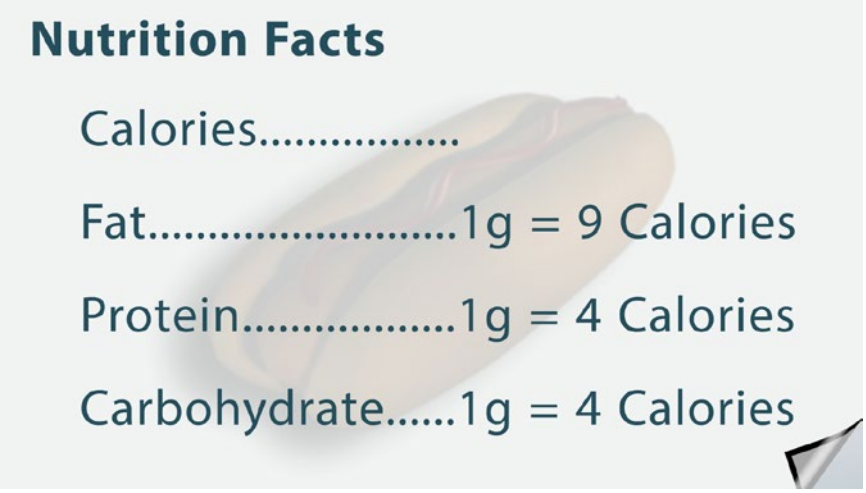

Figure 5.1. Determining the calories in a hot dog.

Despite similar appearances, hot dogs vary widely in nutritional content. For those calorie-conscious among us, statistics can help suggest the healthiest choice. We start with the calorie counts (listed in order from smallest to largest) of 20 different brands of all-beef hot dogs.

110	110	130	130	140	150	160	160	170	170
175	180	180	180	190	190	190	200	210	230

You can see that all-beef hot dogs range from 110 calories in the lowest brand to 230 calories in the highest. One way we can describe this distribution numerically is with the median – the number of calories in a typical beef hot dog. Since we have 20 brands, the location of the median is (10 + 1)/2, which is 10.5. So, we find the median by averaging the 10th and 11th calorie count in the ordered list:

$$\text{median} = \frac{170 + 175}{2} = 172.5 \text{ Calories}$$

So, your typical beef hot dog has 172.5 calories.

Next, we add some more numbers to our analysis of the spread of the calorie counts of beef hot dogs. We know the minimum is 110 calories and the maximum is 230 calories. However, we also need some information about the numbers in between. For that we can determine the quartiles. These are values one-quarter and three-quarters up the ordered list of calories. The first quartile – also known as Q_1 – has 25% of the ordered observations at or below it. It is the median of the lower half of the data:

110 110 130 130 140 150 160 160 170 170

The median of this 10 number set is between the fifth and sixth data value: $Q_1 = 145$. Now, we turn our attention to the third quartile – also known as Q_3 – which has 75% of the ordered observations at or below it. To find Q_3, take the median of the upper half of the data:

175 180 180 180 190 190 190 200 210 230

The median of these 10 upper numbers gives us $Q_3 = 190$.

So, with our minimum, first quartile, median, third quartile, and maximum, we have what is called the five-number summary, which we've summarized in Figure 5.2.

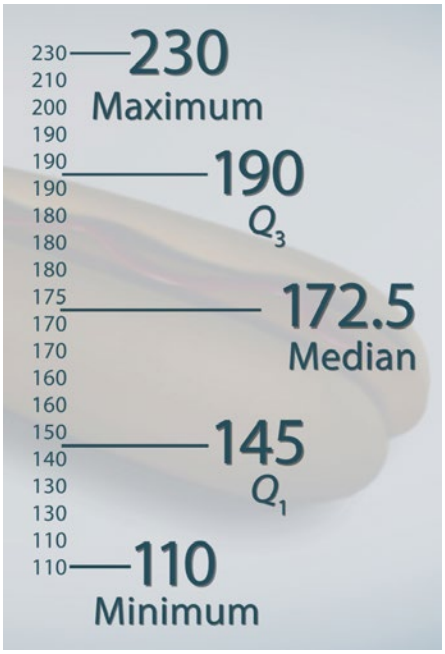

Figure 5.2. The five-number summary of all-beef hot dog calories.

The five-number summary gives us a nice snapshot of both the center and spread of the data. The median marks the center. The first and third quartiles contain between them the middle half of the data. We can measure the spread of the inner 50% of the data by the interquartile range (IQR):

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

The two extremes show how far out the data extends. We can measure that spread using the range:

range = maximum – minimum

In statistics, the best description of data often combines the precision of numbers with the clarity of pictures. A boxplot (or box-and-whisker plot) is a graphic display of the five-number summary. Figure 5.3 shows a boxplot of the all-beef hot dog calories. The box spans the first and third quartiles, the median is marked inside the box, and whiskers extend out to the extremes.

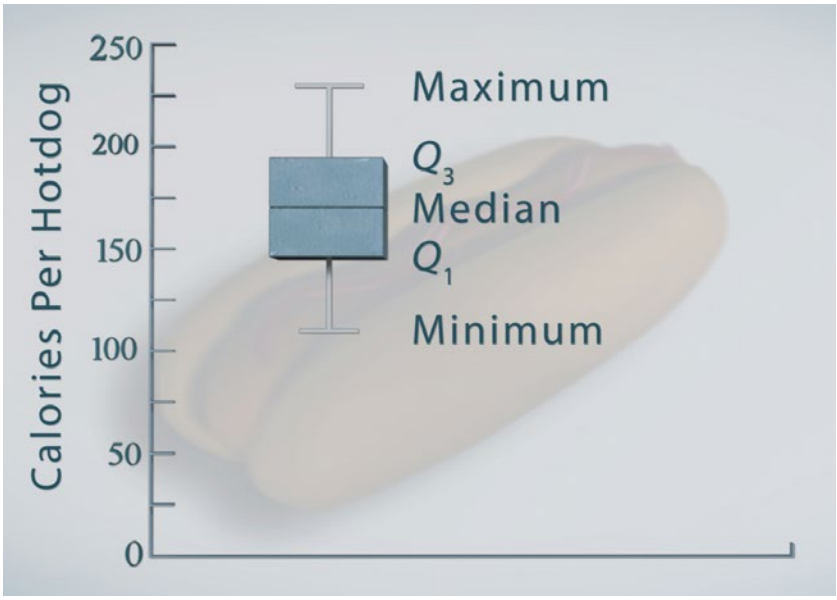

Figure 5.3. Boxplot of all-beef hot dog calories.

Boxplots don't show a distribution in detail the same way that a stemplot or histogram would, but boxplots can be a great way to make a quick side-to-side comparison of a few distributions. Take a look at Figure 5.4 comparing the calories of beef, poultry, and veggie hot dogs.

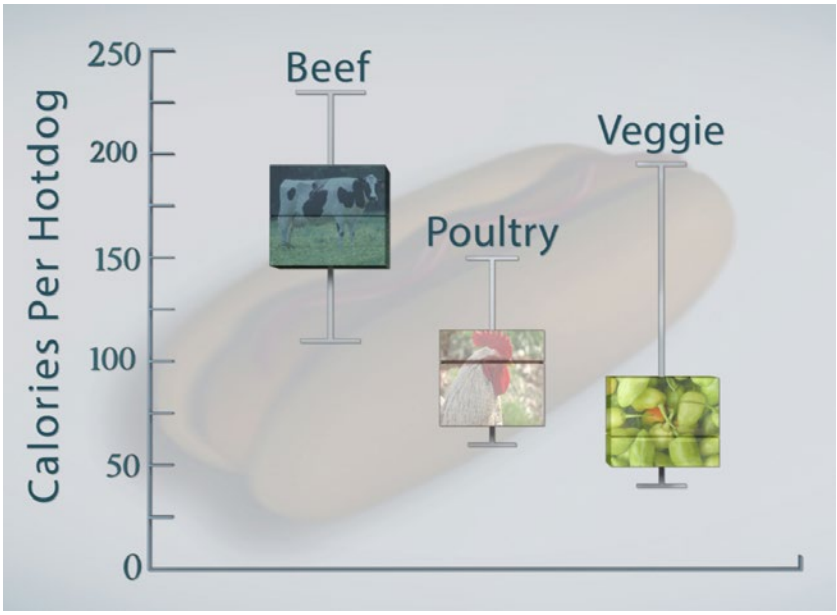

Figure 5.4. Comparing calories of beef, poultry, and veggie hot dogs.

Notice that the median of the poultry hot dogs is below the minimum for the beef hot dogs. So, half of the brands of poultry hot dogs have fewer calories than the lowest calorie brand of all-beef hot dogs. Now, check out the boxplot for the vegetarian hot dog. Notice that at least one

veggie brand has more calories than three quarters of the beef hot dogs!

So, now we can add boxplots to the list of ways we can graphically represent data – and, as shown by the hot dogs, this method allows for easy comparisons between groups.

STUDENT LEARNING OBJECTIVES

A. Recognize that a basic numerical description of a distribution requires both a measure of center and a measure of spread.

B. Use the quartiles and the extremes to provide information about the unequal spread in the two sides of a skewed distribution.

C. Use the 1.5 × IQR rule to identify outliers.

D. Be able to calculate the quartiles and give the five-number summary of a data set of moderate size (say $n \leq 100$).

E. Understand that boxplots provide less detail than stemplots or histograms but are especially useful for comparing several distributions.

CONTENT OVERVIEW

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. You already know how to calculate the minimum, median, and maximum. In this overview, we will provide algorithms for calculating the quartiles. It should be noted, however, that there are several different algorithms for calculating the quartiles. So, check with your textbook or software to see how it calculates quartiles.

First, we discuss a rationale for the five-number summary for describing a data set. The five-number summary provides information on both the center of a distribution and its spread. The median is a useful measure of the *center* of a set of observations. The median is the midpoint, the point with half of the data at or below it and half above. However, the median alone is not an adequate description of a set of data. For example, it is not enough to know that the median number of candies in bags of candy is 60 pieces. It is quite a different story if (1) some bags have as few as 40 and others have as many as 75 compared to (2) some bags have as few as 55 and others have as many as 65. To quantify these two situations, we'll need information about the *spread* or *variability* of the data.

Because the median is the "halfway" point in a data set, one way to show spread is by giving the two quartiles along with the median. The first quartile is the one-quarter point in the data: one-fourth of the data values are at or below the first quartile and three-quarters above. The third quartile is the three-quarters point, with three-quarters of the data at or below it. The two quartiles capture the middle half of the data between them. So, the distances from the median out to the quartiles and between the quartiles show how spread out the data are, or at least how spread out the middle 50% of the data are. The distances between the first and third quartiles, $Q_3 - Q_1$, is called the **interquartile range** or **IQR**.

To calculate the quartiles, first locate the median in an ordered data list. The median divides the ordered data into a lower half and an upper half.

- If there is an odd number of data values, the median is the middle data value in the ordered list. Omit this value when forming the lower half and upper half of the ordered data.
- If there is an even number of observations, the median is between the middle two data values. So, the ordered data can be divided into a lower half and upper half about the median.

The first quartile, Q_1, is the median of the lower half of the ordered data and the third quartile, Q_3, is the median of the upper half of the ordered data.

So far, we have discussed using the median to describe the center of a distribution and the interquartile range to describe the spread of the middle half of the data. We can add information about how far the data are spread by giving the distances from the median out to the minimum and maximum data values and between the minimum and maximum. The distance between the minimum and maximum, maximum – minimum, is called the **range**.

Now, we work through an example. Grades from an exam are displayed in the stemplot in Figure 5.5. We use the stemplot to order the test scores from smallest to largest.

```
4 | 01
5 |
6 | 89
7 | 2699
8 | 0256788
9 | 01236
```

Figure 5.5. Stemplot of test scores.

Since n = 20, the median is at the (20 + 1)/2 or 10.5 position, midway between 82 and 85; hence, the median = 83.5. The median divides the data into a lower half and upper half:

Lower half:	40	41	68	69	72	76	79	79	80	82
Upper half:	85	86	87	88	88	90	91	92	93	96

The median of the lower half is at the (10 + 1)/2 or 5.5 position, midway between 72 and 76; so, $Q_1 = 74$. The median of the upper half is midway between 88 and 90; so, $Q_3 = 89$.

Here's our five-number summary of the exam grades:

minimum = 40, Q_1 = 74, median = 83.5, Q_3 = 89, maximum = 96

We can use the median, 83.5, as a measure of center for the test scores. The spread of the middle 50% of the test scores is given by the interquartile range, IQR = 89 – 74 = 15. The spread as measured from the smallest test score to the largest is given by the range = 96 – 40 = 56. Notice that the overall spread of the test scores is more than three times the spread of the middle 50% of the test scores.

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. Once you have the five-number summary, it takes only three steps to draw a basic boxplot as outlined below.

Constructing a Basic Boxplot

The instructions below are for horizontal boxplots but easily can be adapted for vertical boxplots.

Step 1: Draw a number line. Add a scale that begins at or below the minimum and ends at or above the maximum.

Step 2: Directly above the number line, draw a rectangular box that extends from Q_1 to Q_3. Divide the box with a vertical line at the median.

Step 3: Draw two whiskers: one from the middle left side of the box to the minimum and the other from the middle right side of the box to the maximum.

Figure 5.6 shows the result of applying these steps to create a basic boxplot from the five-number summary for the test scores.

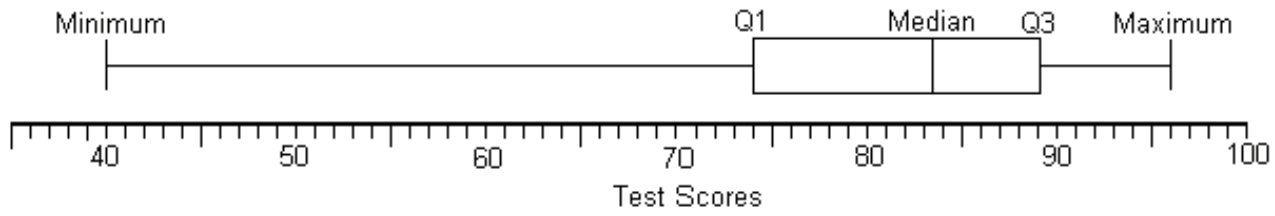

Figure 5.6. Basic boxplot of exam grades.

$Q_1 = 74$, median = 83.5, $Q_3 = 89$, maximum = 96

Each part of the boxplot – the left whisker, the box from Q_1 to the median, the box from the median to Q_3, and the right whisker – represents the spread of one quarter of the data. So, for example, because the box from Q_1 to the median is longer than the box from the median to Q_3, we know that the second quarter of the test scores are more spread out than the third quarter of the test scores.

Notice also the long left whisker that extends from $Q_1 = 74$ all the way down to the minimum test score of 40. We don't know if that long whisker is the result of a single low grade, an outlier, or if the pattern of the lower quarter of the test scores spreads out over the interval from 40 to 74. A modified boxplot, which separates out the outliers and adjusts the lengths of the whiskers so that they are unaffected by outliers, will help us sort out this issue. Here are the steps needed to convert a basic boxplot into a modified boxplot (the generally preferred plot).

Constructing a Modified Boxplot

Step 1: After making a basic boxplot, remove the whiskers.

Step 2: Compute the IQR = $Q_3 - Q_1$; compute a step = 1.5 × IQR.

Step 3: Calculate the inner fences (one step on either side of the box ends):

$$Q_1 - \text{ 1 step} \text{ and } Q_3 - \text{ 1 step}.$$

Calculate the outer fences (two steps on either side of the box ends):

$$Q_1 - 2 \text{ steps} \text{ and } Q_3 + 2 \text{ steps}.$$

Step 4: Identify the mild outliers. Use an asterisk (*) to plot any data values that lie between the two fences. Identify the extreme outliers. Use another symbol, such as an open circle, to plot any data values that are more extreme than the outer fences.

Step 5: Attach a whisker from the left end of the box to the smallest data value that is not an outlier. Then attach a whisker from the right end of the box to the largest data value that is not an outlier.

Next, we convert the basic boxplot from Figure 5.6 into the modified boxplot shown in Figure 5.7. We begin by removing the whiskers from the basic boxplot. Then we calculate the inner and outer fences as follows:

IQR = 89 – 74 = 15
Step = 1.5 × IQR = 1.5(15) = 22.5
Inner fences: Q_1 - 1 step; Q_3 + 1 step: 74 - 22.5 = 51.5; 89 + 22.5 = 111.5
Outer fences: Q_1 - 2 steps; Q_3 + 2 steps: 74 - 2(22.5) = 29; 89 + 2(22.5) = 134

Two test scores, 40 and 41, fall between the lower inner fence and lower outer fence and hence are classified as mild outliers. Mark each of their locations with an asterisk. (There are no extreme outliers.) Attach the left end of the box at Q_1 to 68, the smallest test score that is not an outlier. Redraw the original right whisker (since all test scores were smaller than the upper fences).

The completed modified boxplot appears in Figure 5.7.

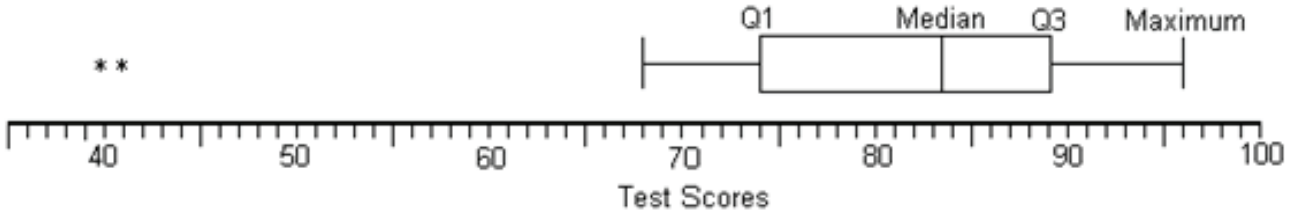

Figure 5.7. Modified boxplot of test scores.

Notice that in the modified boxplot, the length of the lower whisker is about the same as the upper whisker, which indicates that with outliers removed the lower quarter of the test scores had about the same spread as the upper quarter of the test scores. The long left whisker in the basic boxplot was due to two students whose grades were outliers.

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

THE VIDEO: BOXPLOTS

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What *variable* is used to compare different brands of hot dogs?

2. What name do we give to the value for which one-quarter of the data values falls at or below it?

3. What numbers make up a five-number summary?

4. How do you calculate the interquartile range?

5. Boxplots show that poultry hot dogs as a group differ from all-beef hot dogs. Compare the distribution of calories between the two types of hot dogs.

UNIT ACTIVITY:
USING BOXPLOTS TO ANALYZE DATA

Return to the class survey data collected for Unit 2's Activity. The questions are restated at the end of this activity so that you don't have to refer back to Unit 2.

ANALYSIS OF THE SURVEY

1. Make modified boxplots for the data from questions 1 – 5. Describe the key features of each of your boxplots.

2. How do estimates of waiting time compare for males and females? Make comparative modified boxplots of the data from survey question 1 for males and females.

3. Do males or females spend more time studying for exams? Make comparative modified boxplots of the average time spent studying for an exam for males and females. Interpret your graphs in the context of study times.

4. Make comparative modified boxplots for the amount of time that males and females exercise on a typical day.

SURVEY QUESTIONS (UNIT 2, ACTIVITY)

1. How long (in seconds) did you wait while your instructor was getting ready for this activity?

2. How much money in coins are you carrying with you right now?

3. To the nearest inch, how tall are you?

4. How long (in minutes) do you study, on average, for an exam?

5. On a typical day, how many minutes do you exercise?

6. Circle your gender: Male Female

EXERCISES

1. A consumer testing laboratory measured the calories per hot dog in 20 brands of beef hot dogs. Here are the results:

186	181	176	149	184	190	158	139	175	148
152	111	141	153	190	157	131	149	135	132

a. Find the five-number summary of this distribution. Explain how you arrived at your answer.

b. Compute the range and interquartile range. Explain what these numbers tell you about the variability in calories in different brands of all-beef hot dogs.

c. Would a beef hot dog with 175 calories be in the top quarter of the data?
Support your answer.

2. Return to the data on all-beef hot dog calories from exercise 1.

a. Draw a basic boxplot for the calories per hot dog.

b. In which quarter – the first, second, third, or fourth – are the data most concentrated? Explain how you can answer this question based on the boxplot from (a).

c. In which quarter – the first, second, third, or fourth – is the data most spread out? Explain how you can answer this question based on the boxplot from (a).

d. If a data value is more than 1.5 × IQR below the first quartile or more than 1.5 × IQR above the third quartile, it is considered an outlier. Should any of the calorie counts for the beef hot dogs be classified as outliers? Explain.

3. Make a stemplot of the calories in the sample of beef hot dogs from exercise 1. What do you learn from the stemplot that you could not learn from the boxplot?

4. The calories for 20 brands of veggie dogs are given below. (Notice these data have been ordered from smallest to largest.)

40	45	45	45	50	50	55	57	60	60
70	80	80	81	90	95	100	100	110	190

a. Make a five-number summary of the veggie dog calories.

b. Make a modified boxplot for the veggie dog data. Use asterisks (*) to indicate any mild outliers and open circles to indicate any extreme outliers. (Leave room to add another graph to this graphic display.)

c. Add a modified boxplot for the beef hot dog calories next to your display in (b). This will allow you to compare the calorie distributions of the two types of hot dogs.

d. Based on your displays in (c), compare the distributions of calories for beef dogs and veggie dogs.

5. Refer to the baseball data from Table 3.4, Unit 3. Focus on the variable career home runs. (Keep in mind there are 104 players listed because of ties in career batting averages.)

a. Order the number of career home runs from smallest to largest. (You may want to use software such as Excel or your graphing calculator to do the ordering. Otherwise, try using a stemplot to help you order these data.)

b. Create a five-number summary of the career number of home runs.

c. Make a modified boxplot of the career number of home runs. Mark mild outliers with asterisks (*) and extreme outliers with open circles. Show your calculations for the fences. Write the names of the players above each of the outliers.

d. Would you describe the shape of the distribution as symmetric, skewed to the right, or skewed to the left? Justify your choice.

REVIEW QUESTIONS

Recall that Table 2.1 in Unit 2 gives data on state average SAT scores and the percent of high school graduates in each state taking the SATs. Refer to Table 2.1 as needed for questions 1 and 2.

1. The state average SAT Critical Reading scores, 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	

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

b. Does California (average score 499) fall in the top half of the states in the SAT Critical Reading score? Does it fall above the bottom quarter? Support your answer.

c. Roughly what percentage of the states have scores higher than Wyoming's 572? How many states would that be?

d. Make a basic boxplot of the states' average SAT Critical Reading scores. Which quarter of the data, the first, second, third or fourth, shows the most amount of spread?

2. The states' 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. Give the five-number summary of the 51 state average SAT Math scores.

b. 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.)

c. Write a brief description comparing the distributions. Include in your descriptions comparisons of both center and spread.

3. The stemplot video in Unit 2 included data on the fuel economy information on Toyota's 2012 vehicle line. A stemplot of the city miles per gallon (mpg) data appears below.

1	3 4
1	67799
2	0112
2	5678
3	0
3	
4	34
4	
5	1
5	

a. Make a five-number summary of the mpg data.

b. Make a basic boxplot of the mpg data.

c. Based on the stemplot, how many of the data values are potential outliers?

d. Make a modified boxplot of the mpg data. Show the calculations for the fences. Based on your plot, how many data values were identified as outliers?