2.4: Box Plots, Quartiles, and Percentiles

Learning Objectives

- Introduce box plots
- Define quartiles
- Define percentiles
- Calculate percentiles
- Calculate values for a five-number summary

Section 2.4 Excel File (contains all of the data sets for this section)

Using Box Plots to Visualize Data

Frequency distributions and their graphs (bar graphs and histograms) provide insight into data by grouping observations into classes and then determining each class's frequency or relative frequency. The classes depend on the values our data takes on, and there is some freedom regarding the number of classes we might choose to separate our data into.

Another method of graphing ordinal, interval, or ratio level data, called a **box plot** (or a box-and-whisker plot), groups data into four classes based on order, each containing approximately 25% of the observations. Consider the following figure containing three box plots relating students' final grades in statistics across different universities.

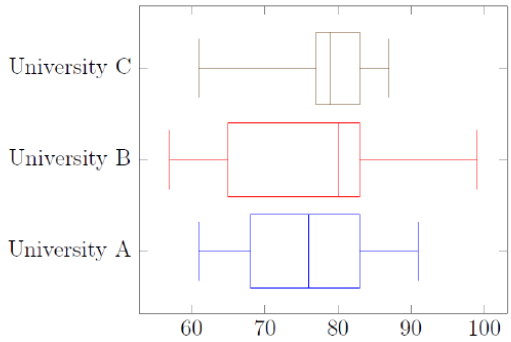

Figure 2.4.1: Box plots of final statistics grades for three universities

We know that there are four classes for each box plot. Note the five vertical lines; these values correspond to the boundaries of the classes. The left-most line corresponds to the data set's smallest value, the minimum. The first class extends from the minimum to the left side of the "box" and includes 25% of the observations; we call the upper bound for this first class, the **first quartile Q_1** (one-quarter of the observations are less than or equal to it). The second class again needs to have 25% of the observations with a lower bound of Q_1 and an upper bound Q_2. We call the upper bound of the second class the **second quartile Q_2**, which is most commonly referred to as the **median,** meaning 50% of the data fall below this value. The third class again needs to have 25% of the observations with a lower bound of Q_2 and an upper bound of the **third quartile Q_3**; 75% of the observations are less than or equal to Q_3. The final class consists of the remaining 25% of observations with lower bound Q_3 and upper bound of the largest value, the **maximum**, of the data set. The first and fourth classes form the whiskers of the box plot, while the second and third form the box. These five numbers (**minimum**, **Q_1, Q_2, Q_3,** and **maximum)** form what we call the **five-number summary** of the data.

Text Exercise 2.4.1

Consider Figure 2.4.1 above and classify each box plot as positively skewed, negatively skewed, or symmetric. Explain.

Answer

We preface our answer by noting that box plots, like histograms and bar graphs from grouped frequency distributions, are formed by grouping observations. Since the graph does not provide information about each data value, we are primarily making a claim about a characteristic of the graphical representation.

Recall that a graph is positively skewed if the right tail extends further than the left tail and negatively skewed if the left tail extends further than the right tail. A graph is symmetric if we can fold it in half so that the left and right sides roughly match. The tails and the whiskers fall in similar parts of the graphs discussed. If one whisker is longer, we can say that the box plot is skewed in the direction of the longer whisker. We classify the box plot of University C as negatively skewed and the box plot of University B as positively skewed. The whiskers seem to be of equal length for University A, but that is not enough to assert symmetry. We also want the two halves of the box to be the same. This is the case for the box plot of University A. We classify the box plot of University A as symmetric.

Given a data set, we can quickly identify the minimum and maximum values. Determining the quartiles, however, presents more of a challenge. With an ordered data set, we understand where a quarter, half, and three-quarters of the data would fall. Consider the following data sets with their five-number summaries.

Observations	2	4	6	8	10	12	14	16
Five Number Summary	min		Q_1		Q_2		Q_3	max

Observations	2	4	4	8	10	12	14	16
Five Number Summary	min		Q_1		Q_2		Q_3	max

Observations	2	4	6	8	10	12	14	16	18
Five Number Summary	min		Q_1		Q_2		Q_3		max

Figure 2.4.2: Three data sets with intuitive placement of the box plot boundaries

We need help determining our five-number summary for these data sets.

The first two data sets each contain 8 observations, while the third contains 9 observations. We can easily group the first two data sets in groups of 2 to get 25% of the observations in each class. With 9 observations, however, we cannot get exactly 25% in each class. However, note that there are equal numbers of observations above and below Q_2 for each data set.

If we try to attach values to quartiles, we face another challenge. Let us begin with Q_1. In the first data set, we see that Q_1 naturally falls between 4 and 6, but what value should we assign? Our box plots would look significantly different if we used 6 as opposed to 4. There is no easy solution to this challenge, and statisticians have developed a variety of approaches. We will provide a simple approach later in this section; please remember that it is not the only approach. However, most of the various approaches produce measures that are reasonably close to each other.

We face another challenge when we try to understand Q_1 in the second data set. We would naturally assign a value of 4 to Q_1 because the only number between 4 and 4 is 4. We wanted 25% of the observations to be less than or equal to Q_1, but we have 3 values that are less than or equal to 4 in our data set, that is 37.5% of our observations.

We highlight these challenges to frame our understanding appropriately. We use box plots, quartiles, and percentiles (which we will get to shortly) to get a general, intuitive feel about our data using methods that may differ from field to field, statistician to statistician, and program to program. When consuming statistics or conducting analysis, know which method is in use.

Quartiles are descriptive statistics that express at what values there will be about 25%, 50%, or 75% of the observations at or below that value. There is nothing extraordinarily unique about 25%, 50%, or 75%. We could choose 10% or 99%. When we expand our ideas to include different percentages of observations, we call them **percentiles**. Q_1 is the 25^{th} percentile. Q_2 is the 50^{th} percentile. Q_3 is the 75^{th} percentile.

Percentiles have utility beyond building summary visualizations; they help us understand how individual observations compare to the entire data set. They measure relative position within an ordered data set. For example, a test score by itself is usually difficult to interpret. For instance, if one of us had a score on a measure of shyness of 35 out of a possible 50, we would have little idea how shy that person was compared to others. It would be helpful to know the percentage of people with equal or lower shyness scores. If 65% of the scores were at or below this person's score, then the score would be at the 65^{th} percentile.

Text Exercise 2.4.2

1. If Helen's score was at the 95^{th} percentile, what percentage of scores are at or below Helen's?

Answer

The percentile means that 95% of the scores are at or below Helen's score.

2. If the scores ranged from 1 to 100 on an exam and Helen earned a score of 95, does this necessarily mean that her score is at the 95^{th} percentile?

Answer

No, the percentile gives a relative position of the scores. The number of scores at or below her score determines the percentile measure. If everyone did well and only 70 of the scores fell at or below Helen's, she would be at the 70^{th} percentile even though she got 95 out of a 100 points.

Calculating Percentiles

We already indicated that there are several different ways to calculate quartiles. This is because quartiles are percentiles, and there are several ways to calculate percentiles, which may lead to different values in different situations. The method that we present is one of the simplest calculations.

The P^{th} percentile is a value such that $P\%$ of the observations fall at or below that value. We need the data to be counted and ordered from smallest to largest. Let n be the number of observations in our data set. Next, we calculate the number of observations that make up $P\%$ of the observations. We call this number the rank R of the percentile.

$$R = \frac{P}{100} \cdot n$$

Now there are two possibilities for the R value; either it will be a natural number $\{1, 2, 3, 4, 5, \dots\}$ or not.

- If R is a natural number, we find the midpoint between the R^{th} and $(R+1)^{st}$ values.
- If R is not a natural number, round R up to the following natural number and take the value in that position.

Note: Classifications of Numbers

The real number system has the following designations.

Natural Numbers: $1, 2, 3, \dots$
Whole Numbers: $0, 1, 2, 3, \dots$
Integers: $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$
Rational Numbers: Any number that can be written as a fraction
Irrational Numbers: Any number that cannot be written as a fraction. Examples include e, π, $\sqrt{2}$

Text Exercise 2.4.3

Consider the 20 quiz scores shown in the table below and compute the five-number summary and 82^{nd} percentile. After completing the calculation by hand, use the Section 2.4 Excel file to calculate each percentile using the functions PERCENTILE.INC and PERCENTILE.EXC. Compare the values.

Table 2.4.1: 20 quiz scores with corresponding rank

Number	**4**	**4**	**4**	**5**	**6**	**6**	**6**	**6**	**6**	**6**	**7**	**7**	**8**	**8**	**8**	**9**	**9**	10	10	10
Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11	12	13	14	15	16	17	18	19	20

Answer

We first note that the data is already ordered from smallest to largest with 20 observations. A secondary row has been created to index the observations. Note that the row heading is Rank; consider how this ties back to rank R in our calculation.

Let us begin with the five-number summary.

The minimum value is 4.

Q_1, the 25^{th} percentile. $R = \frac{25}{100} \cdot 20 = \frac{1}{4} \cdot 20 = \frac{20}{4} = 5$. Note 5 is a natural number. We then look at the 5^{th} and 6^{th} observation values, which are both 6 and find the midpoint between them. Thus $Q_1 = 6$. Using Excel, we get 6; they are all the same.

Q_2, the 50^{th} percentile. $R = \frac{50}{100} \cdot 20 = \frac{1}{2} \cdot 20 = \frac{20}{2} = 10$. Note 10 is a natural number. We then look at the 10^{th} and 11^{th} observation values, which are 6 and 7, respectively and find the midpoint between them. Thus $Q_2 = \frac{1}{2}(6+7) = \frac{13}{2} = 6.5$. Excel computes 6.5 and 6.5;they are all the same.

Q_3, the 75^{th} percentile. $R = \frac{75}{100} \cdot 20 = \frac{3}{4} \cdot 20 = \frac{60}{4} = 15$. Note 15 is a natural number. We then look at the 15^{th} and 16^{th} observation values, which are 8 and 9, respectively and find the midpoint between them. Thus $Q_3 = \frac{1}{2}(8+9) = \frac{17}{2} = 8.5$. Excel gives 8.25 and 8.75;they are all different.

The maximum value is 10.

Let us look at the 82^{nd} percentile. $R = \frac{82}{100} \cdot 20 = \frac{41}{50} \cdot 20 = \frac{820}{50} = \frac{82}{5} = 16.4$. Notice 16.4 is not a natural number. We must look at the 17^{th} observation value showing the 82^{nd} percentile is 9. Excel calculates 9.22 and 9; they are different and the same respectively.

We have seen that different methods of calculation can produce slightly different values. With large data sets, we generally resort to technology to produce our measures and might not have control over the precise methodology employed therein. As such, we remember that percentiles provide rough measures for the distribution of our data sets and nuance our understanding that roughly this percent of observations fall below roughly this value. When large data sets or limited time make hand computation prohibitive, we recommend using functions such as PERCENTILE.INC and QUARTILE.INC.

Consider the preceding example. If we looked at the 83^{rd} percentile $(R = 16.6)$ which would also be the 17^{th} value, which is 9. So, the 83^{rd} percentile is the same as the 82^{nd} percentile. This happens since there are only 20 data points; we cannot subdivide 20 indefinitely. The two percentiles would typically differ with large data sets.

Box Plots: Constructing and Interpreting

? Text Exercise 2.4.4

As part of the "Stroop Interference Case Study," students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. Compare the scores for the 15 men and 30 women who participated in the experiment by making separate box plots for each gender.

Table 2.4.2: Women's times (left) and men's times (right)

14	17	18	19	20	21		16	19	23
15	17	18	19	20	22		17	20	24
16	17	18	19	20	23		18	22	26
16	17	18	20	20	24		19	22	26
17	18	18	20	21	24		19	23	28

Answer

To construct box plots, we need the five-number summaries.

Table 2.4.3 Five number summaries for the data presented in Table 2.4.1

	Females	Males	Box Plot Component
Minimum	14	16	End of Left Whisker
Q_1	17 ($R = \frac{25}{100} \cdot 30 = 7.5$)	19 ($R = \frac{25}{100} \cdot 15 = 3.75$)	Left Side of Box
Q_2=median	18.5 ($R = \frac{50}{100} \cdot 30 = 15$)	22 ($R = \frac{50}{100} \cdot 15 = 7.5$)	Line in Box
Q_3	20 ($R = \frac{75}{100} \cdot 30 = 22.5$)	24 ($R = \frac{75}{100} \cdot 15 = 11.25$)	Right Side of Box
Maximum	24	28	End of Right Whisker

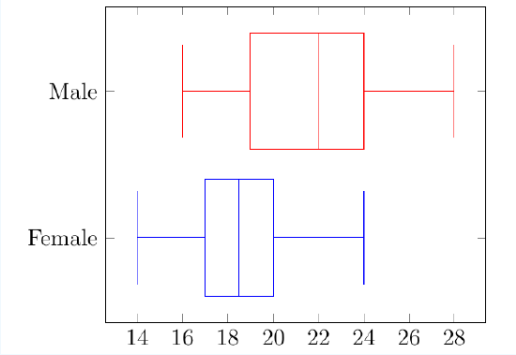

Figure 2.4.3 Box plots for male and female times for naming the colors of various rectangles

The men tended to take longer than the women. About 25% of male times were longer than the maximum female time. At least 75% of the male times were longer than the median female time.

Text Exercise 2.4.5

Suppose data came from a task that aims to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. The time to reach the target was recorded on each trial. The box plots of the two distributions are shown below. What can we conclude by looking at the two box plots?

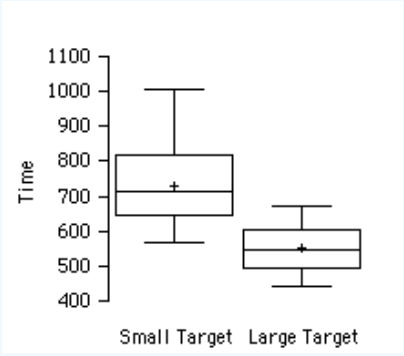

Figure 2.4.4: Box plots for the response times by small and large target

Answer

We can see that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one. The minimum time for the small target is longer than the median time of the large target. At least 50% of times for the small target are longer than all of the large target times.

Text Exercise 2.4.6

Construct two data sets, treated as observations from a discrete quantitative variable, consisting of 12 values each so that the box plots are identical, but the bar graph of one data set is perfectly symmetric while the other is not.

Answer

If the box plots are going to be identical, the five-number summaries must be the same. When arranged from least to greatest, Q_1 is the average of the 3^{rd} and 4^{th} values, Q_2 is the average of the 6^{th} and 7^{th} values, and Q_3 is the average of the 9^{th} and 10^{th} values. We can construct a perfectly symmetric data set by pairing observations by proximity to the center. Since the 6^{th} and 7^{th} values are in the middle, they would be paired together, the 5^{th} with the 8^{th}, and the 4^{th} and 9^{th}, and so forth. We want the values in these positions to be equally distant from the median value.

We can start by picking the first data set that is not perfectly symmetric. We will produce such a data set if we use a pattern of one observation value followed by a different value repeated twice. We picked even numbers starting at 0 to ensure that our quartiles were nice values. Not all procedures would produce a data set suitable for this example because we need our second data set to be symmetric. We must check this because we want the minimum and maximum to be equally far from the median and the first and third quartiles.

$$\{0, 2, 2, 4, 6, 6, 8, 10, 10, 12, 14, 14\}$$

We have fixed our five number summary: $\min = 0$,$Q_1 = 3$, $Q_2 = 7$, $Q_3 = 11$, and $\max = 14$. Our procedure produced values allowing us to produce a symmetric data set with the same five-number summary. The average of two of the same numbers is that number, so, for ease, we can start our symmetric data set with the following numbers.

$$\{0, \ , 3, 3, \ , 7, 7, \ , 11, 11, \ , 14\}$$

All we need to do is fill in the remaining spots, ensuring symmetry and preserving order.

$$\{0, 2, 2, 4, 6, 6, 8, 10, 10, 12, 14, 14\}$$
$$\{0, 1, 3, 3, 4, 7, 7, 10, 11, 11, 13, 14\}$$

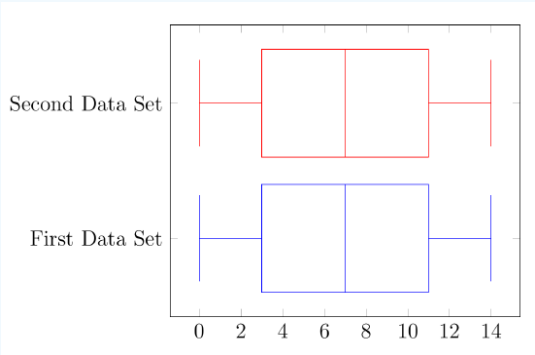

Figure 2.4.5 Two identical box plots

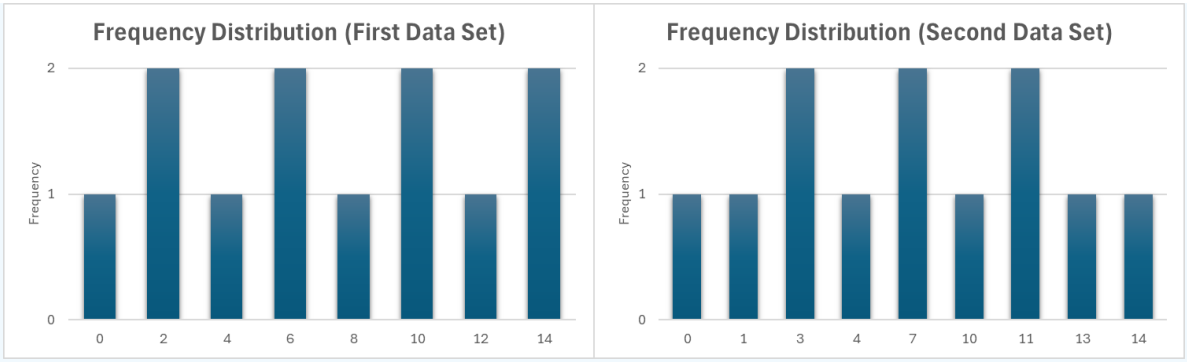

Figure 2.4.6 Non-identical bar graphs of the two data sets that produced identical box plots.
