

8.6: Box Plots

Learning Objectives

- Find the 5-number summary of a data set
- Construct and interpret a box plot for a data set

5-Number Summary

While quartiles are not a 1-number summary of spread like standard deviation, the quartiles are used with the median, minimum, and maximum values to form a **5-number summary** of the data.

5-Number Summary

The **5-number summary** takes the form:

Minimum, Q_1 , Median, Q_3 , Maximum

The 5-number summary is written as a list of numbers. It is understood by context what each value in the list represents.

✓ Example 8.6.1

Find the 5-number summaries for the 9 female sample and the 8 female sample from the previous section.

59 60 62 64 66 67 69 70 72

59 60 62 64 66 67 69 70

Solution

For the 9 female sample, the median is 66, the minimum is 59, and the maximum is 72. The 5-number summary is: 59, 61, 66, 69.5, 72.

For the 8 female sample, the median is 65, the minimum is 59, and the maximum is 70, so the 5-number summary would be: 59, 61, 65, 68, 70.

✓ Example 8.6.2

Find the 5-number summary of the quiz score data from the previous section.

- section A: 5 5 5 5 5 5 5 5 5
- section B: 0 0 0 0 10 10 10 10 10
- section C: 4 4 4 5 5 5 5 6 6
- section D: 0 5 5 5 5 5 5 5 10

Solution

In each section, the median is the average of 5th and 6th data values. The first quartile is the 3rd data value, and the third quartile is the 8th data value. Creating the 5-number summaries:

section and data	5-number summary
section A: 5 5 5 5 5 5 5 5 5	5, 5, 5, 5, 5
section B: 0 0 0 0 10 10 10 10 10	0, 0, 5, 10, 10
section C: 4 4 4 5 5 5 5 6 6	4, 4, 5, 6, 6
section D: 0 5 5 5 5 5 5 5 10	0, 5, 5, 5, 10

Of course, with a relatively small data set, finding a 5-number summary is a bit silly, since the summary contains almost as many values as the original data.

Try It 8.6.1

The total cost of textbooks for the term was collected from 36 students. Find the 5-number summary of this data.

\$140 \$160 \$160 \$165 \$180 \$220 \$235 \$240 \$250 \$260 \$280 \$285
 \$285 \$285 \$290 \$300 \$300 \$305 \$310 \$310 \$315 \$315 \$320 \$320
 \$330 \$340 \$345 \$350 \$355 \$360 \$360 \$380 \$395 \$420 \$460 \$460

Answer

The data is already in order, so we don't need to sort it first.

The minimum value is \$140 and the maximum is \$460.

There are 36 data values so $n = 36$. 36 is an even number, so the median is the average of the 18th and 19th data values, \$305 and \$310. The median is \$307.50.

To find the first quartile, the number of data in the first half is $n = 18$. Since 18 is an even number, we know Q_1 is the average of the 9th and 10th data values, \$250 and \$260. $Q_1 = \$255$

To find the third quartile, the number of data in the second half is also $n = 18$. Since 18 is an even number, we know Q_3 is the average of the 9th and 10th data values of the second half of the data set (or the 18 + 9 = 27th and 28th data values in the original data set), \$345 and \$350. $Q_3 = \$347.50$

The 5-number summary of this data is: \$140, \$255, \$307.50, \$347.50, \$460

✓ Example 8.6.3

Returning to the household income data from earlier, find the 5-number summary.

Income (thousands of dollars)	Frequency
15	6
20	8
25	11
30	17
35	19
40	20
45	12
50	7

Solution

By adding the frequencies, we can see there are $n = 100$ data values represented in the table. In a previous example, we found the median was the mean of the 50th and 51st data values, \$35 thousand. We can see in the table that the minimum income is \$15 thousand, and the maximum is \$50 thousand.

To find Q_1 , we find the median of the first 50 data values. It will be the mean of the 25th and 26th data values.

Counting up in the data as we did before,

There are 6 data values of \$15, so values 1 to 6 are \$15 thousand

The next 8 data values are \$20, so values 7 to $(6 + 8) = 14$ are \$20 thousand

The next 11 data values are \$25, so values 15 to $(14 + 11) = 25$ are \$25 thousand

The next 17 data values are \$30, so values 26 to $(25 + 17) = 42$ are \$30 thousand

The 25th data value is \$25 thousand, and the 26th data value is \$30 thousand, so Q_1 will be the mean of these:

$$\frac{(25 + 30)}{2} = \$27.5 \text{ thousand.}$$

To find Q_3 , we find the median of the second 50 data values. It will be the mean of the 75th and 76th data values. Continuing our counting from earlier,

The next 19 data values are \$35, so values 43 to $(42 + 19) = 61$ are \$35 thousand

The next 20 data values are \$40, so values 61 to $(61 + 20) = 81$ are \$40 thousand

Both the 75th and 76th data values lie in this group, so Q_3 will be \$40 thousand.

Putting these values together into a 5-number summary, we get: 15, 27.5, 35, 40, 50 (in thousands of dollars).

Note that the 5 number summary divides the data into four intervals, each of which will contain about 25% of the data. In the previous example, that means about 25% of households have income between \$40 thousand and \$50 thousand.

Box Plots

For visualizing data, there is a graphical representation of a 5-number summary called a **box plot**, or box-and-whisker graph.

Box plot

A **box plot** is a graphical representation of a 5-number summary.

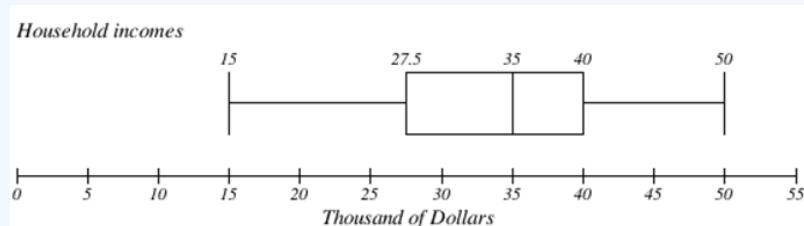
To create a box plot, a scaled number line is first drawn. A box is drawn from the first quartile to the third quartile, and a line is drawn through the box at the median. “Whiskers” are extended out to the minimum and maximum values. Be sure to give the graph a title and label the number line.

✓ Example 8.6.4

Construct a box plot for the household income data.

Solution

The box plot below is based on the household income data with 5-number summary: 15, 27.5, 35, 40, 50

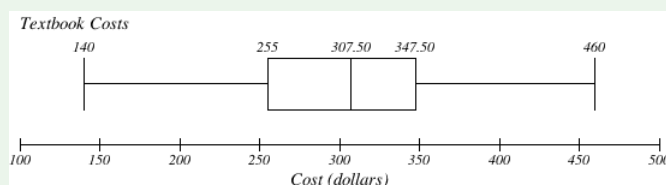


The box plot illustrates how spread or concentrated the data is. Since each part of the box plot represents a quarter of the data set, we can see that the data in the first quarter is more spread out, but the data between the second and third quartiles is more concentrated. There are the same number of data values in each part: 25 values between \$15 to \$27.5, 25 values between \$27.5 to \$35 thousand, 25 values between \$35 to \$40 thousand and 25 values between \$40 to \$50 thousand.

✎ Try It 8.6.2

Create a box plot based on the textbook price data from the last Try It.

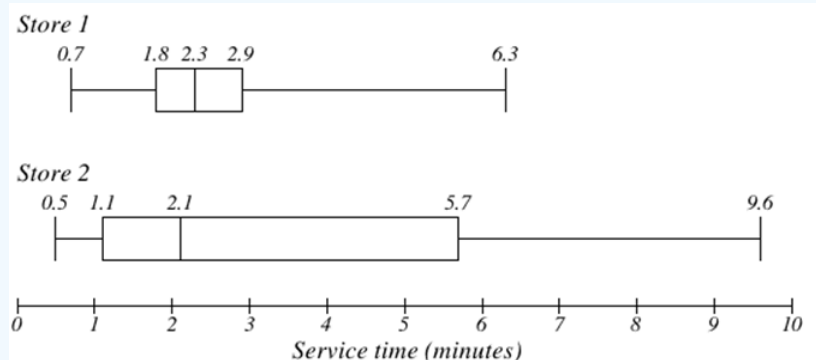
Answer



Box plots are particularly useful for comparing data from two populations.

✓ Example 8.6.5

The box plots of service times for two fast-food restaurants are shown below. Discuss the differences between the stores based on the box plots.



Solution

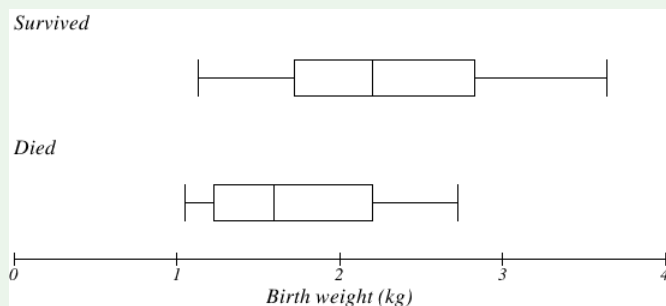
While store 2 had a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), store 2 is less consistent, with a wider spread of the data.

At store 1, 75% of customers were served within 2.9 minutes, while at store 2, 75% of customers were served within 5.7 minutes.

Which store should you go to in a hurry? That depends upon your opinions about luck: 25% of customers at store 2 had to wait between 5.7 and 9.6 minutes!

✎ Try It 8.6.3

The box plot below is based on the birth weights in kilograms of infants with severe idiopathic respiratory distress syndrome (SIRDS) [1]. There are separate box plots to show the birth weights of infants who survived and those that did not. Compare and contrast the groups based on the box plots.



Answer

Comparing the two groups, the box plots reveal that the birth weights of the infants that died appear to be, overall, less than the weights of infants that survived. In fact, we can see that the median birth weight of infants that survived is the same as the third quartile of the infants that died.

Similarly, we can see that the first quartile of the survivors is greater than the median weight of those that died, meaning that over 75% of the survivors had a birth weight greater than the median birth weight of those that died.

Looking at the maximum value for those that died and the third quartile of the survivors, we can see that over 25% of the survivors had birth weights greater than the heaviest infant that died.

The box plots give us a quick, albeit informal, way to determine that birth weight is quite likely linked to survival of infants with SIRS.

[1] van Vliet, P.K. and Gupta, J.M. (1973) Sodium bicarbonate in idiopathic respiratory distress syndrome. *Arch. Disease in Childhood*, **48**, 249–255. As quoted on <http://openlearn.open.ac.uk/mod/ouco...§ion=1.1>

This page titled [8.6: Box Plots](#) is shared under a [CC BY-SA 3.0](#) license and was authored, remixed, and/or curated by [David Lippman & Jeff Eldridge](#) ([The OpenTextBookStore](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [11.6: Measures of Variation](#) by David Lippman & Jeff Eldridge is licensed [CC BY-SA 3.0](#). Original source: <http://www.opentextbookstore.com/mathinsociety>.