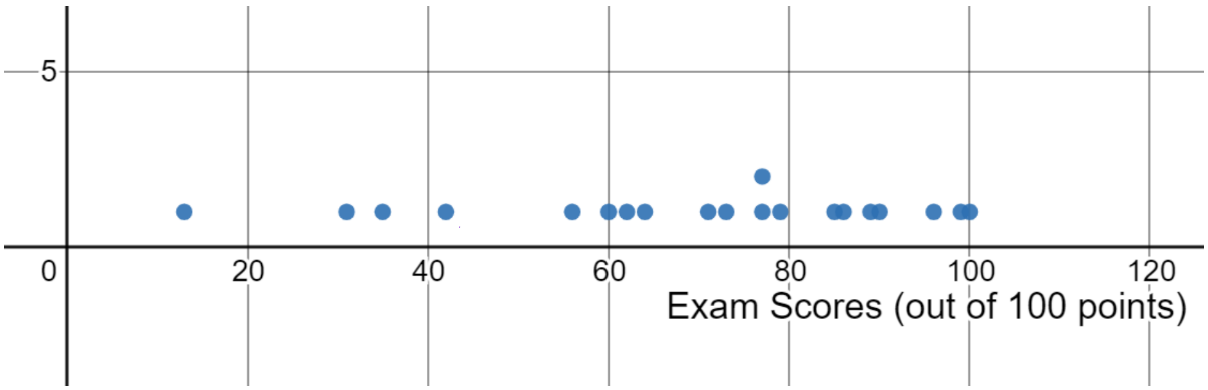
2.3: Quantifying Variability Relative to the Median

In a previous lesson, we estimated the center, shape, and spread of distributions. We quantified the center by computing the mean and median for a data set. In this section, we will examine the spread or variability of a distribution.

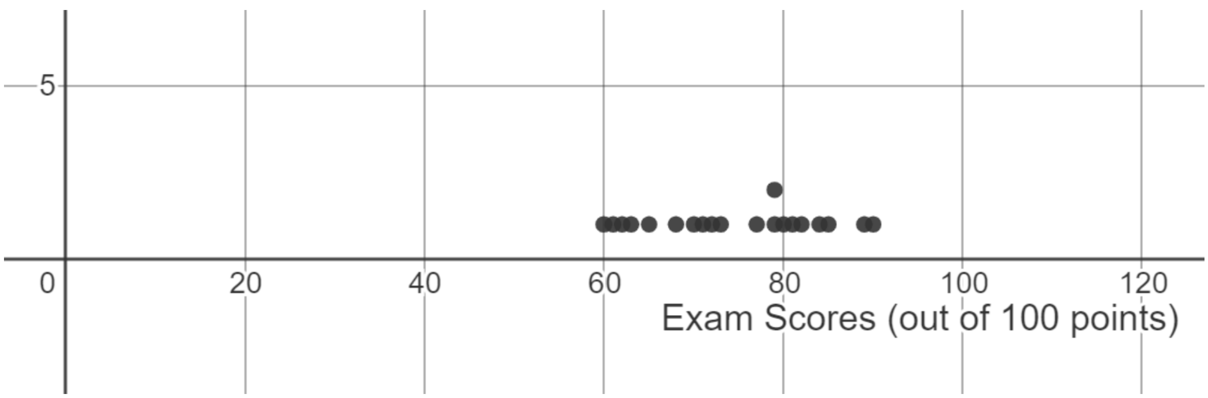
1. Given below are the dotplots for two samples of quiz scores.

   Scores for class A: [13,31,35,42,56,60,62,64,71,73,77,77,79,85,86,89,90,96,99,100]

   Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   Scores for class B: [60,61,62,63,65,68,70,71,72,73,77,79,79,80,81,82,84,85,89,90]

   Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   Estimate the center, shape, and spread of the distributions. Then compare the two distributions. What do you notice is similar about the distributions? What do you notice is different about the distributions?

2. In a previous section, we summarized the center of a distribution by using the sample mean and median. This is an important aspect of reading a graph, but it is not enough. The center of the distributions above are similar, but one distribution has more variability than the other distribution. Choose between the following two numbers to represent the variability for the class A distribution. Then explain your choice: 85 or 30

Range

3. One way to represent the variability in data is with the **range**. The range is the difference between the maximum and minimum data values.

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

Compute the range for the two classes. Which class has the larger range?

One issue with the range is that it is influenced by outliers. In other words, the range is **not a resistant** measure of spread. This is because the calculation of the range always uses the two most extreme data values in the set.

Five Number Summary

Another way to describe variability is with **quartiles**. When the data is sorted from lowest to greatest, the quartiles are numbers that divide the data into four equal parts. Recall that the median is the number that divides the data into two equal parts. 50% of the data is below the median, and 50% of the data is above the median.

Let's compute the quartiles for the class A scores. To find the quartiles, we should first compute the median.

First, we make sure that the data are in order from smallest to largest.

$$[13, 31, 35, 42, 56, 60, 62, 64, 71, 73, 77, 77, 79, 85, 86, 89, 90, 96, 99, 100]$$

There are 20 values in this set, so we divide 20 by 2 to find the location of the median. 20 divided by 2 is 10 so the median is between the 10th and 11th data values in the ordered list.

$$[13, 31, 35, 42, 56, 60, 62, 64, 71, 73] \text{ Median } [77, 77, 79, 85, 86, 89, 90, 96, 99, 100]$$

$$\text{Median} = \frac{73+77}{2} = 75$$

Notice that the median is *not* an existing value in the set. We see that the median divides the set into two *new* sets. To find the quartiles, we find the median of the lower half and upper half of the data set. There are 10 values to the left of the median. To find the location of the **first quartile Q1**, we divide 10 by 2 to get 5 and therefore, there will be 5 data values on each side of Q1. Similarly, to find the location of the **third quartile Q3**, there will be 5 data values on each side of Q3.

$$[13, 31, 35, 42, 56] \text{ Q1 } [60, 62, 64, 71, 73] \text{ Median } [77, 77, 79, 85, 86] \text{ Q3 } [89, 90, 96, 99, 100]$$

$$\text{Q1} = \frac{56+60}{2} = 58 \qquad \text{Q3} = \frac{86+89}{2} = 87.5$$

$$[13, 31, 35, 42, 56]58[60, 62, 64, 71, 73]75[77, 77, 79, 85, 86]87.5[89, 90, 96, 99, 100]$$

The median, first quartile Q1, and third quartile Q3, together with the minimum and maximum values in the set, make up **the five number summary**. The five number summary is given in the table below:

Class A Five Number Summary

Minimum	13
Q1	58
Median = Q2	75
Q3	87.5
Maximum	100

4. Compute the five number summary from Class B's scores.

Class B Five Number Summary

Minimum	
Q1	
Median	
Q3	
Maximum	

We have learned that quartiles are medians! When a set contains an odd number of values, the middle value is not included in the lower or upper half of the given data set to compute the quartiles.

5. Another way to describe variability in a data set is to find the distance between the first and third quartiles (Q1 and Q3). This distance is called the **interquartile range**, abbreviated as **IQR**.

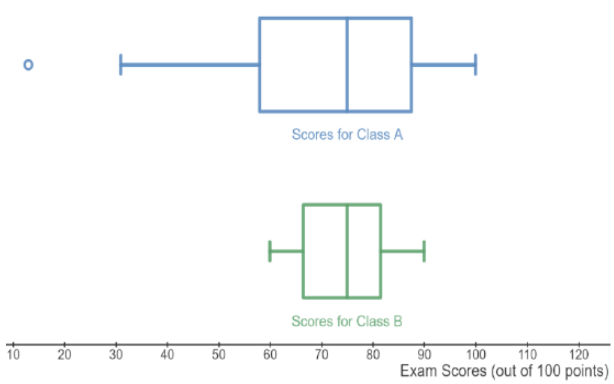
$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$$

   The IQR gives the range of the middle 50% of the data. For class A, the IQR is $\mathrm{Q3} - \mathrm{Q1} = 87.5 - 58 = 29.5$ points. Calculate the IQR for the scores from class B.

6. Compare the IQRs for scores for class A and class B. Does this comparison agree with your intuition about the spread of the distributions? Do you think the IQR is highly affected by extreme values?

Boxplots (Box and Whisker Plots)

The values in the five-number summary can be represented in a graph called a **boxplot** (sometimes referred to as a box and whiskers plot). Given below are the boxplots for the scores for each of the classes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

7. Mark the minimum, Q1, median, Q3, and maximum on each of the boxplots above.

8. Are the boxplots consistent with your interpretation of the spread of the distributions?

9. What do you think the open circle on the boxplot for scores for class A indicates?

Recall, an *outlier* is a value that is extremely far away from the rest of the values in the set of data. When we are working with boxplots, values that deviate further than 1.5 IQRs from the quartiles are considered outliers. Outliers are outside of the range that extends from $\mathrm{Q1}-(1.5\cdot\mathrm{IQR})$ to $\mathrm{Q3}+(1.5\cdot\mathrm{IQR})$.

These boundaries are called the **fences**. Any value that is below the lower fence, $\mathrm{LF}=\mathrm{Q1}-(1.5\cdot\mathrm{IQR})$, or above the upper fence, $\mathrm{UF}=\mathrm{Q3}+(1.5\cdot\mathrm{IQR})$, is considered an outlier. On the graph we represent the outlier with an open circle, or an asterisk. The attached whisker is connected to the next non-outlier in the data set.

10. Explain why the score 13 from the class A data is an outlier. Compute the fences in your answer.

11. You try! The ages of 15 Academy Award winners for best actress are given below:

$$[48, 47, 54, 35, 45, 65, 77, 36, 33, 37, 36, 44, 21, 37, 38]$$

a. Sort the ages from smallest to largest and compute the five number summary.

Ages of Academy Award winners - Five Number Summary

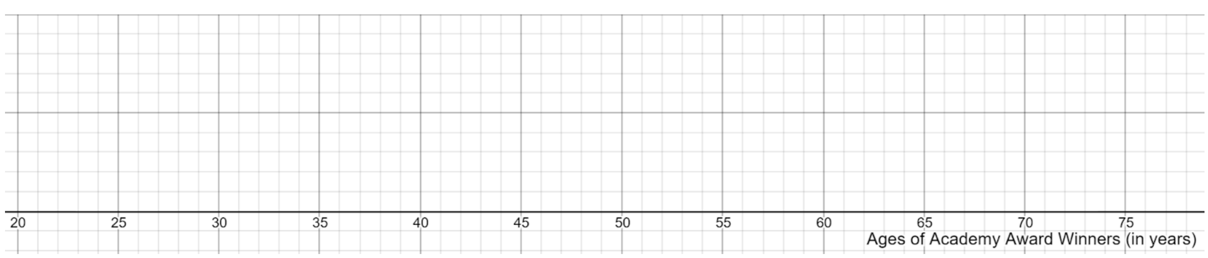
Minimum	
Q1	
Median	
Q3	
Maximum	

b. Compute the range of the ages of academy award winners.

c. Compute the IQR of the ages of academy award winners.

d. Compute the fences and use them to identify any outliers.

e. Graph the boxplot for the age of academy award winners.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

f. Go to https://www.desmos.com/calculator and complete the following steps:

i. $A = [48, 47, 54, 35, 45, 65, 77, 36, 33, 37, 36, 44, 21, 37, 38]$ to the first line.
ii. Tap Enter on your keyboard.
iii. Type boxplot(A).
iv. Click the Zoom Fit icon and confirm the box next to “Exclude Outliers” is checked.

Use this tool to check your work.

This page titled 2.3: Quantifying Variability Relative to the Median is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.