

Comparing Data Distributions



Objective

In this lesson, you will

Comparing Data Sets in Tables

Organizing data makes it easier to compare and find similarities and differences between two data sets. To analyze data sets, use measures of _____ and measures of _____.

Measures of Center	Measures of Variation
<ul style="list-style-type: none">• Mean (average): the _____ of all values in a data set _____ by the number of values• Median: the _____ value of an ordered set of data when the number of items in the data set is _____ or the average of the two _____ values if the number of items is even	<ul style="list-style-type: none">• Mean absolute deviation (mean absolute variation): the _____ of the absolute difference of each item in the data set from the data set’s mean.

To compute mean absolute deviation:

1. Compute the _____ of the data set.
2. Find the absolute difference, or _____, of each item in the data set from the mean.
3. Find the _____ of the distances you obtained in step 2.
4. Compute the average of the distances by _____ the sum you found in step 3 by the total number of observations in the data set.

Example: The table shows the number of people who visited a week-long art exhibit. The organizers wanted to find the average daily attendance for the event and how close the data values were to the mean.

Day	Visitors
Monday	25
Tuesday	54
Wednesday	30
Thursday	40
Friday	50
Saturday	73
Sunday	92

The average daily attendance is the same as the median mean.

$$\begin{aligned} \text{average daily attendance} &= \frac{\text{total number of visitors}}{\text{number of days}} \\ &= \frac{25 + 54 + 30 + 40 + 50 + 73 + 92}{7} \end{aligned}$$

The average daily attendance for the exhibit was _____ visitors.

Find the absolute difference, or deviation, between each value and the mean.

The mean absolute deviation is the sum of those values divided by the number of days.

$$\begin{aligned} \text{mean absolute deviation} &= \frac{\text{total absolute deviation}}{\text{number of days}} \\ &= \frac{27 + 2 + 22 + 12 + 2 + 21 + 40}{7} \\ &= \boxed{} \end{aligned}$$

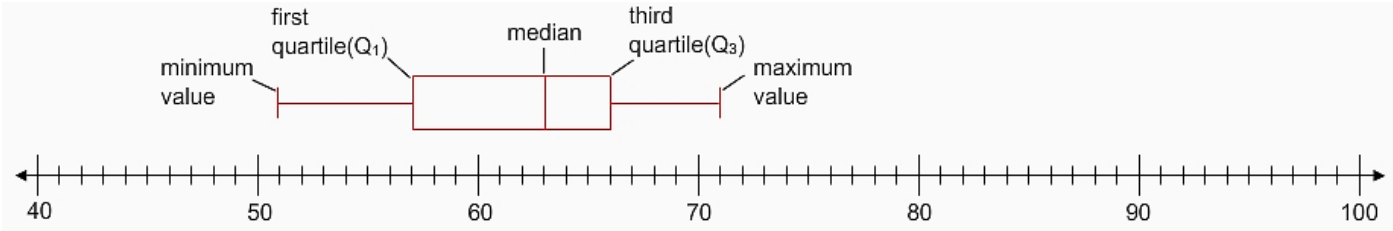
Day	Visitors	Absolute Deviation
Monday	25	$ 52 - 25 = 27 $ $= 27$
Tuesday	54	$ 52 - 54 = \underline{\hspace{2cm}}$
Wednesday	30	$ 52 - 30 = \underline{\hspace{2cm}}$
Thursday	40	$ 52 - 40 = \underline{\hspace{2cm}}$
Friday	50	$ 52 - 50 = \underline{\hspace{2cm}}$
Saturday	73	$ 52 - 73 = \underline{\hspace{2cm}}$
Sunday	92	$ 52 - 92 = \underline{\hspace{2cm}}$




The average difference of each data value and the mean value is _____ people, which is a relatively low high value.

So, we can say that the data values _____ are _____ are not very close to the mean.

Comparing Data Sets in Box Plots






Key Term

The **interquartile range** of a data set is the difference of its first quartile (Q_1) and its third quartile (Q_3):

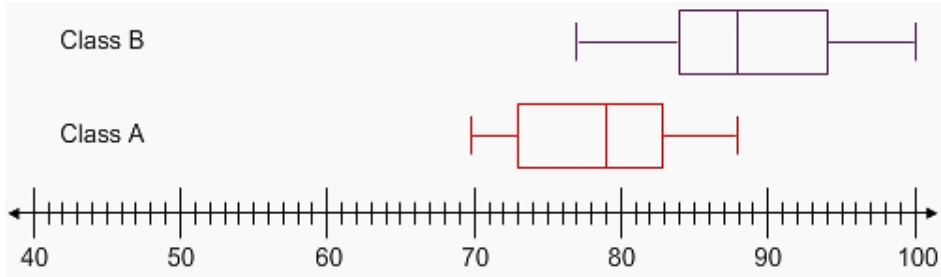
$$\text{interquartile range} = Q_3 - Q_1$$

If the data does not have outliers and is not overly spread out, use:	mean	median
	mean absolute deviation	interquartile range
If the data does have outliers or is overly spread out, use:	mean	median
	mean absolute deviation	interquartile range



Question

The box plots show the test scores of students in class A and class B.



- The vertical line inside each box represents the _____ test score.
 - For class A, the _____ test score is _____.
 - For class B, the _____ test score is _____.
- The difference of the _____ test scores is _____.
- For either data set, the left edge of the box represents the _____ first third quartile and the right edge of the box represents the _____ first third quartile.
 - For class A, the interquartile range is _____.
 - For class B, the interquartile range is _____.

- The difference of the means (____) expressed as a multiple (m) of the interquartile range of either data set is shown below.
 - class A: ____ = m (____) $\rightarrow m = 0.9$
 - class B: ____ = m (____) $\rightarrow m = 0.9$

Choose the statements that are true about the box plots.

The median score of class A is 79.

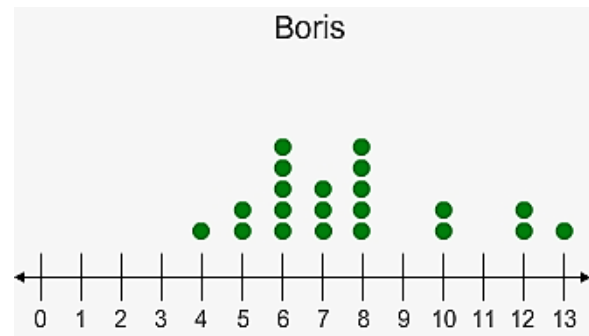
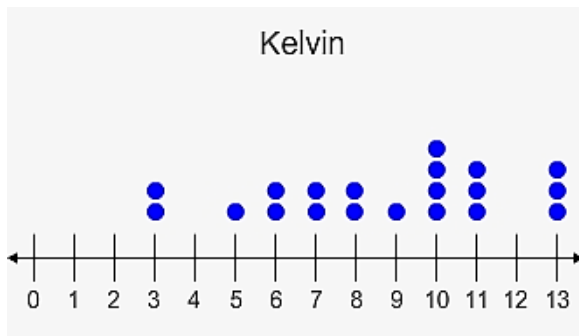
The difference of the median scores for class A and class B is about 1 times the interquartile range of either data set.

The median score of class B is 84.

The difference of the median scores for class A and class B is about half the interquartile range of either data set.

Comparing Data Sets in Dot Plots

Boris and Kelvin have decided to eat out less often to save up for a beach vacation next summer. They're keeping a record of the number of meals they eat at home each week. The dot plots show the data they have gathered.



First, compute the medians and compare them with the interquartile ranges of the data sets to see if they are similar.

Median: _____

Q_1 : _____

Q_3 : _____

IQR: _____

Median: _____

Q_1 : _____

Q_3 : _____

IQR: _____

The difference of Boris's median and Kelvin's median is _____ - _____ = _____.

→ The difference of the medians is less greater than half of either interquartile range but less greater than one times either interquartile range.

Find the mean of each data set and use the mean absolute deviation as a measure of variability.

$$\begin{aligned}\text{mean} &= \frac{\text{sum of Kelvin's data}}{\text{number of data items}} \\ &= \boxed{}\end{aligned}$$

$$\begin{aligned}\text{mean} &= \frac{\text{sum of Boris's data}}{\text{number of data items}} \\ &= 7.7\end{aligned}$$

$$\text{mean absolute deviation} = \frac{\text{total absolute deviation}}{\text{number of data items}}$$

$$\frac{50.6}{20} = \boxed{}$$

mean absolute deviation is 1.97

The difference of Boris's mean and Kelvin's mean is _____ - _____ = _____.

→ This value is approximately equal to _____ of the mean absolute deviation for either data set.

The results from either method show that the difference of the two data sets is not very noticeable.

Summary

To compare the measures of center of two data sets as a multiple of their measures of variation, their measures of variation must be similar. In your own words, why do you think this is true?