Numerical Measures of Data

The famous trio—the mean, the median, and the mode—represent three different methods for finding the center, or the so-called typical value. When the values are different, they can lead to different interpretations of the data being summarized.

Example: Consider the cost, in $1000s, of five homes in a neighborhood:

180, 180, 200, 240, 1,000.

a) Compute the mean, median, and mode of the home prices.

b) If you were trying to promote that this is an affluent neighborhood, which value would you report?

c) If you were trying to argue against a tax increase, and wanted to argue that income is too low to afford a tax increase, which value(s) might you report?

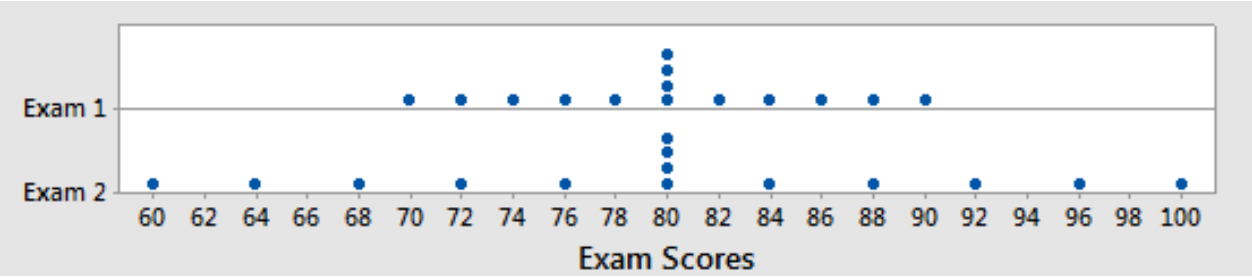
d) If we placed a 1% property tax and wanted to compute how much revenue this would generate, which value could we use to compute this?

Measures of Spread

- Consider the following two sets of data for the scores on two exams in a math class.

  Exam 1: 70, 72, 74, 76, 78, 80, 80, 80, 80, 82, 84, 86, 88, 90
  Exam 2: 60, 64, 68, 72, 76, 80, 80, 80, 80, 84, 88, 92, 96, 100

- The distribution of the exam scores are displayed with dotplots.



- Computing the numerical measures of center for the scores yields:

  Exam 1: Mean = Median = Mode = 80

  Exam 2: Mean = Median = Mode = 80

- Both sets of data have the same mean, median and mode but the values obviously differ in another respect—the variation or *spread* of the values. How can we capture this fact?

Definition: The ***range*** of a data set is the difference between the largest and smallest value.

range = (largest value) – (smallest value)

- The range is the simplest measure of spread.
- The range is often used in weather reports, which give the high and low temperature of the day.
- Compute the range for the two lists of numbers above.

Question: Is the range a resistant measurement?

Variance and Standard Deviation

- Variance and standard deviation are the most commonly used numeric measures of variability. They are both measure how far away, on average, the individual values are from their mean.

- Consider the waiting times in minutes for the Carlisle branch of M&T Bank:

  2 3 5 7 8

- The mean wait time is $\bar{x} = \frac{2+3+5+7+8}{5} = 5$ minutes.

- The ***deviations*** of the waiting times from the mean are portrayed below:

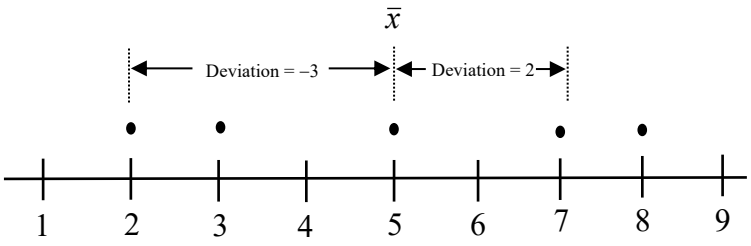


Table of Deviations

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
2	$(2–5) = –3$	3	$(2–5)^2 = 9$
3			
5			
7			
8			
Sum			

- Fill in the second column of the table above, which represent the deviations of the observations from the sample mean. What do you notice about the Sum of the column?

- It turns out that $\sum(x_i - \bar{x})$ will always equal zero for any data set. Therefore, averaging these deviations is meaningless.

Since we are interested in variability, without regard to direction, we could compute the average of the absolute value of $(x - \bar{x})$ to obtain a measure of variability. Fill in the third column of the table.

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$(x-\bar{x})^2$
2	–3	3	$(2–5)^2 = 9$
3	–2		
5	0		
7	2		
8	3		
Sum	0		

Definition: The ***mean absolute deviation*** (MAD), is the average of the absolute deviations of the values from the mean:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

- The mean absolute deviation is an intuitively sensible measure of spread, but it is not widely used because the absolute value presents analytical difficulties.
- Compute the MAD for the wait times of M&T Bank customers.

$$\text{MAD} = \frac{1}{n}\sum|x_i - \bar{x}| =$$

- Instead of using absolute values, we can get another measure of variation by making all the deviations $(x - \bar{x})$ nonnegative by squaring them. Fill in the fourth column of the table.

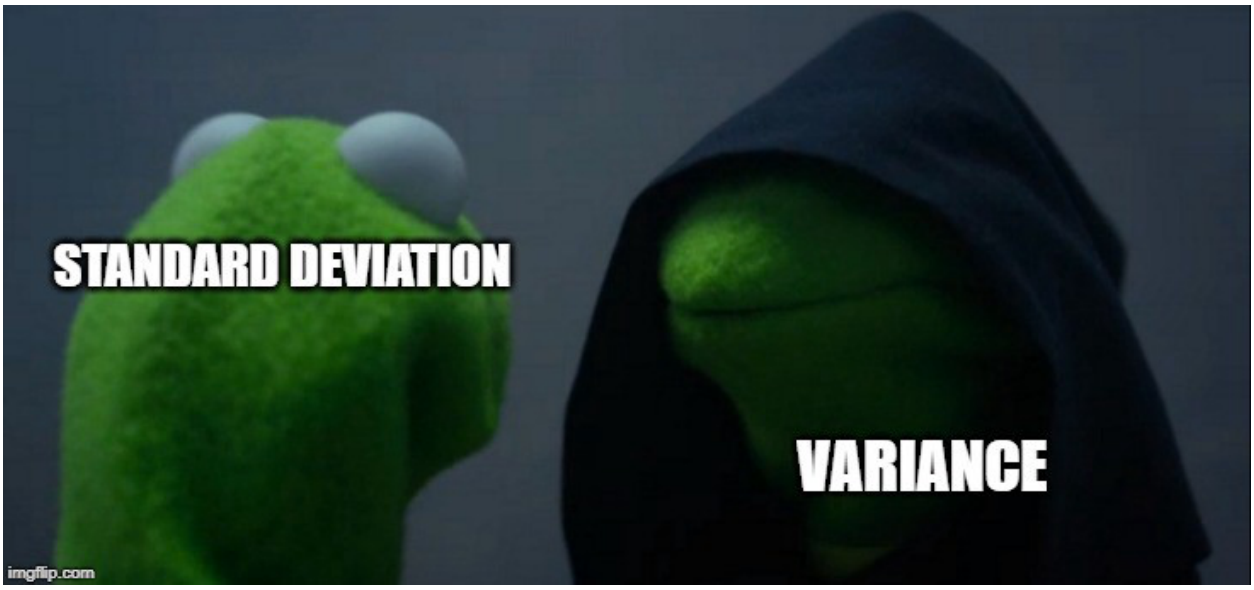
Definition: The ***sample variance***, denoted by s^2, is computed as follows:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

- Note that the sample variance is essentially the average of the squared deviations. However, instead of dividing by the sample size n, we divide by $n - 1$. This will be explained shortly.
- Compute the sample variance for the waiting times of M&T Bank customers.

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} =$$

- What are the units on the sample variance?



Definition: The ***sample standard deviation***, denoted by s, is the square root of the sample variance:

$$s = \sqrt{s^2}$$

- The sample standard deviation has the same units as the original data.

- Compute the sample standard deviation for the waiting times of M&T Bank customers.

$s = \sqrt{s^2} =$

```
> X<-c(2,3,5,7,8)
> sd(X)
[1] 2.54951
```

Interpretation of the Standard Deviation: Think of the standard deviation as roughly an average distance of the observations from their mean. If all of the observations are the same, then the standard deviation is 0 (i.e., no variability). Otherwise, the standard deviation is positive and the more spread out the observations are about their mean, the larger the standard deviation.

Population Variance and Standard Deviation:

If, instead of a sample, you consider the values in a *population*, where N is the total number of units in the population, then the ***population variance***, denoted by σ^2 (sigma-squared), and the ***population standard deviation***, denoted by σ, are computed as follows:

Variance: $\sigma^2 = \dfrac{\sum (x-\mu)^2}{N}$ Standard deviation: $\sigma = \sqrt{\sigma^2}$

- The sample variance, s^2, is an estimate of the population variance, σ^2, while the sample standard deviation s is an estimate of σ.

Remarks:

1. Just as the mean is not a resistant measure of center, the standard deviation is not a resistant measure of spread. It is strongly influenced by extreme values.

2. Why do we divide by $n - 1$ instead of n in the denominator of the sample variance? It can be shown that when the sample variance s^2 is computed by dividing by $n - 1$, s^2 will not systematically over or underestimate the population variance σ^2. That is, if you computed s^2 by dividing by n, you tend to get an inaccurate estimate of σ^2.

3. The variance (and standard deviation) will never be negative. They will equal zero if and only if all the values are the same. Otherwise, they will be positive.

Another Interesting Interpretation of Standard Deviation:

- For most distributions, a majority of the data is within one standard deviation of the mean.