# Standard Deviation Introduction

**Learning Objective: Relate measures of center and spread to the shape of the distribution, and choose the appropriate measures in different contexts.**

## Introduction

So far, we have introduced two measures of spread; the range (covered by all the data) and the inter-quartile range (IQR), which looks at the range covered by the middle 50% of the distribution. We also noted that the IQR should be paired as a measure of spread with the median as a measure of center. We now move on to another measure of spread, the **standard deviation**, which quantifies the spread of a distribution in a completely different way.

## Idea

The idea behind the standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean, $\bar{x}$. The standard deviation gives the average (or typical distance) between a data point and the mean, $\bar{x}$.

## Notation

There are many notations for the standard deviation: SD, s, Sd, StDev. Here, we'll use **SD** as an abbreviation for standard deviation, and use **s** as the symbol.

## Calculation

In order to get a better understanding of the standard deviation, it would be useful to see an example of how it is calculated. In practice, we will use a computer to do the calculation.

### Example: Video Store Customers

The following are the number of customers who entered a video store in 8 consecutive hours:

7, 9, 5, 13, 3, 11, 15, 9

To find the standard deviation of the number of hourly customers:

1. Find the mean, $\overline{x}$ of your data: $\frac{(7 + 9 + 5 + \ldots + 9)}{8} = 9$
2. Find the deviations from the mean: the difference between each observation and the mean

   (7 - 9), (9 - 9), (5 - 9), (13 - 9), (3 - 9), (11 - 9), (15 - 9), (9 - 9)

   -2, 0, -4, 4, -6, 2, 6, 0

   Since the standard deviation is the average (typical) distance between the data points and their mean, it would make sense to average the deviations we got. Note, however, that the sum of the deviations from the mean, $\overline{x}$ is 0 (add them up and see for yourself). This is always the case, and is the reason why we have to do a more complicated calculation to determine the standard deviation:
3. Square each of the deviations:
   The first few are

   $(-2)^2$ = **4**, $(0)^2$ = **0**, $(-4)^2$ = **16**, and the rest are **16, 36, 4, 36, 0.**
4. Average the square deviations by adding them up, and dividing by n - 1, (one less than the sample size):

   $\frac{(4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)}{(8 - 1)} = \frac{112}{7} = 16$

   - the reason why we "sort of" average the square deviations (divide by n - 1) rather than take the actual average (divide by n) is beyond the scope of the course at this point, but will be addressed later.
   - This average of the squared deviations is called the **variance** of the data.
5. The SD of the data is the square root of the variance: $\text{SD} = \sqrt{16} = 4$

   - Why do we take the square root? Note that 16 is an average of the squared deviations, and therefore has different units of measurement. In this case 16 is measured in "squared customers," which obviously cannot be interpreted. We therefore take the square root in order to compensate for the fact that we squared our deviations, and in order to go back to the original unit of measurement.

   Recall that the average number of customers who enter the store in an hour is 9. The interpretation of SD = 4 is that on average, the actual number of customers that enter the store each hour is 4 away from 9.

---

**Comment:**

The importance of the numerical figure that we found in #4 above called the variance (=16 in our example) will be discussed much later in the course when we get to the inference part.

---

## Scenario: Instructor Ratings

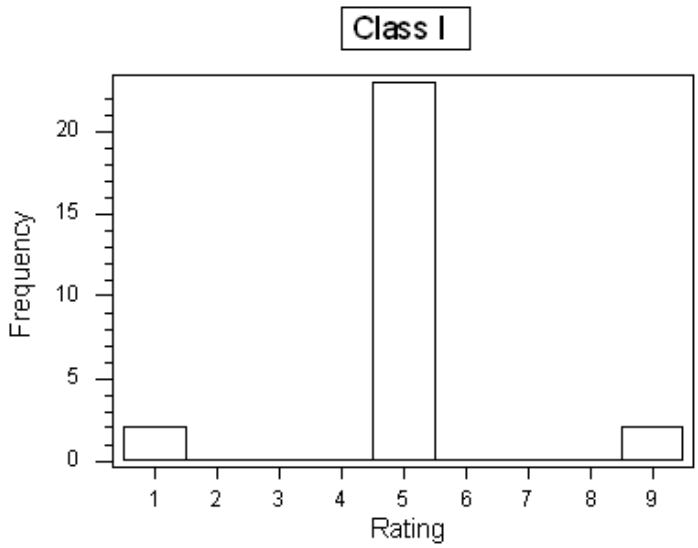
The concept of standard deviation is less intuitive as a measure of spread than the range or the IQR. The following activity is designed to help you develop a better intuition for the standard deviation.

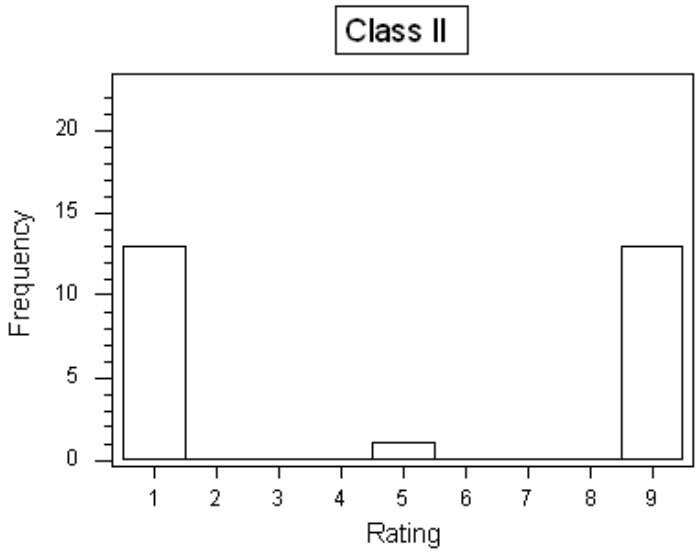
### Background

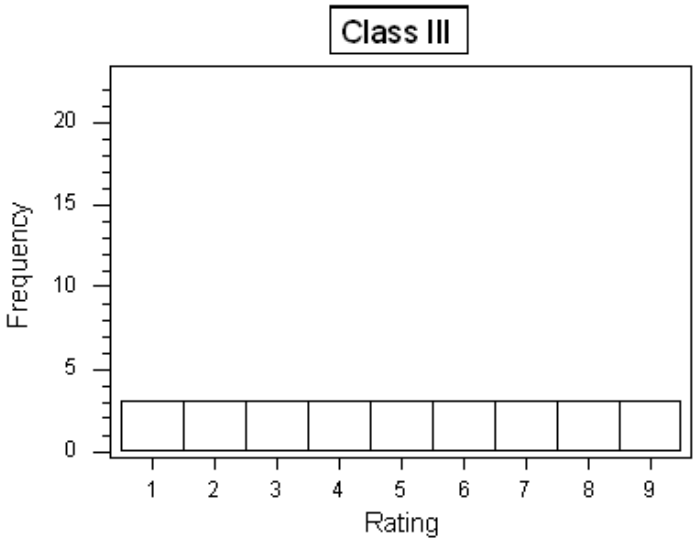
At the end of a statistics course, the 27 students in the class were asked to rate the instructor on a number scale of 1 to 9 (1 being "very poor," and 9 being "best instructor I've ever had"). The following table provides three hypothetical rating data:

| Rating | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| **Class I** | 1 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 1 |
| **Class II** | 12 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 12 |
| **Class III** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

And here are the histograms of the data:






## Learn By Doing

1/1 point (graded)

Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a "typical" or "average" distance between the data points and their mean.

Judging from the table and the histograms, which class would have the largest standard deviation?

- ○ Class I
- ◉ Class II ✔
- ○ Class III

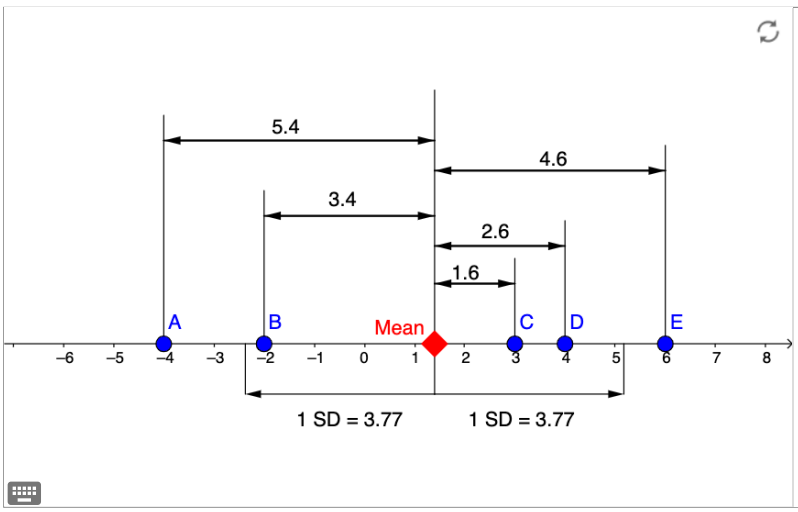
**Answer**

Correct:
In class II most of the observations are far from the mean (at 1 or 9). The average distance between the observations and the mean in this case would be the largest.

Submit

## Interactive Simulation

Drag points A, B, C, D, and E around and observe how the standard deviation changes. The standard deviation is the average squared distance from the mean. Ignore the z-score column.




| 9999 | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Data | x-Mean | (x-Mean)² | z-score | |
| 2 | -4 | -5.4 | 29.16 | -1.43 | |
| 3 | -2 | -3.4 | 11.56 | -0.9 | |
| 4 | 3 | 1.6 | 2.56 | 0.42 | |
| 5 | 4 | 2.6 | 6.76 | 0.69 | |
| 6 | 6 | 4.6 | 21.16 | 1.22 | |
| 7 | | | | | |
| 8 | | SSE = | 71.2 | | |
| 9 | | | | | |
| 10 | | Mean = | 1.4 | | |
| 11 | | SD = | 3.77 | | |
| 12 | | | | | |
| 13 | | | | | |
| 14 | | | | | |
| 15 | | | | | |
| 16 | | | | | |



## Learn By Doing

1/1 point (graded)

Using the standard deviation simulation above, if you set points A, B, C, D, and E all equal to 1, what is the resulting standard deviation? If necessary, round your answer to the closest whole number.

0 ✔

0

**Answer**
Correct:
When all the data points are exactly the same and therefore all equal to the mean, the standard deviation will be 0.

Submit

## Learn By Doing

1/1 point (graded)
Using the standard deviation simulation above, by increasing point E the value of the standard deviation _____.

increases ✔

**Answer**
Correct: Moving a data point further from the mean causes the standard deviation to increase.

Submit

## Properties of the Standard Deviation

1. It should be clear from the discussion thus far that the SD should be paired as a measure of spread with the mean as a measure of center.

2. Note that the only way, mathematically, in which the SD = 0, is when all the observations have the same value (Ex: 5, 5, 5, ... , 5), in which case, the deviations from the mean (which is also 5) are all 0. This is intuitive, since if all the data points have the same value, we have no variability (spread) in the data, and expect the measure of spread (like the SD) to be 0. Indeed, in this case, not only is the SD equal to 0, but the range and the IQR are also equal to 0. Do you understand why?

3. Like the mean, the SD is strongly influenced by outliers in the data. Consider the example concerning video store customers: 3, 5, 7, 9, 9, 11, 13, 15 (data ordered). If the largest observation was wrongly recorded as 150, then the average would jump up to $\bar{x} = 25.9$, and the standard deviation would jump up to SD = 50.3. Note that in this simple example, it is easy to see that

while the standard deviation is strongly influenced by outliers, the IQR is not. The IQR would be the same in both cases, since, like the median, the calculation of the quartiles depends only on the order of the data rather than the actual values.

The last comment leads to the following very important conclusion:

## Choosing Numerical Summaries

Use $\overline{x}$ (the mean) and the standard deviation as measures of center and spread **only** for reasonably symmetric distributions with no outliers.

Use the five-number summary (which gives the median, IQR and range) for all other cases.

---