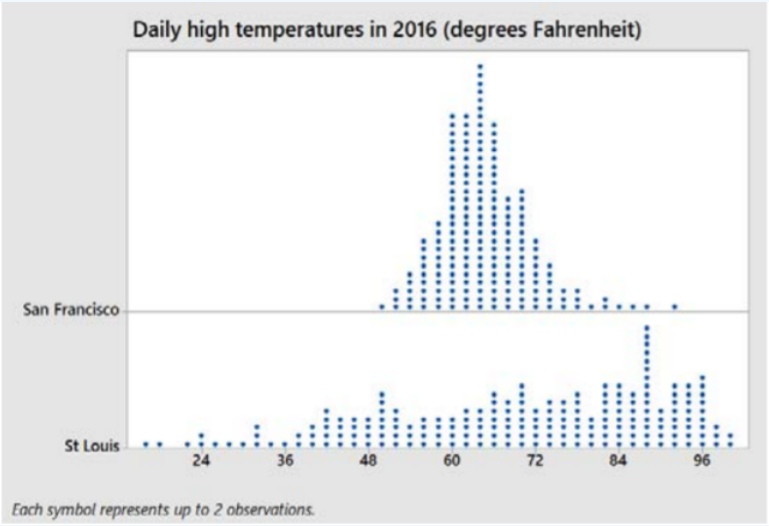
3.2: Measures of Variability

When analyzing data, it is also important to describe the spread or variability of the data.

Example: Comparing high temperatures between San Francisco and St. Louis

Here are the daily high temperatures for every day in 2016 for the cities of San Francisco and St. Louis.[272829]



Even though both cities seem to have approximately the same center, it's obvious that the spread of daily high temperatures in San Francisco is much lower than it is in St. Louis. San Francisco temperatures are mostly mild all year long, while St. Louis has some very hot and very cold days. This section will explore statistics that are used to measure variability in data.

Range

Definition: Range

The easiest measure of variability to calculate is the range of the data.

Range = maximum value - minimum value

Here are the extreme high temperatures in 2016 for San Francisco and St. Louis.

City	Minimum High Temperature	Maximum High Temperature	Range
San Francisco	50˚F	92˚F	42˚F
St. Louis	16˚F	101˚F	85˚F

The range for San Francisco high temperatures is about half of the range for St. Louis.

Example: Students browsing the web

Let's return to the example of daily minutes spent on the internet by 30 students and find the difference of the two most extreme values.

67	71	78	82	85	86	87	87	92	95	97	99	99	100	101
102	103	103	104	105	105	107	108	109	112	116	118	122	124	125

Range = 125 - 67 = 58 minutes

The advantage of the range is that it is easy to calculate. The main disadvantage is that the range only uses two points and is extremely affected by outliers. For example, on September 1, 2017 San Francisco set an all time high temperature record of 106˚F! If this had occurred in 2016, an outlier of 106˚F would have changed the range for San Francisco from 42˚F to 56˚F. Therefore, statisticians prefer to use measures of variability that use all the data, not simply the outliers.

Variance and Standard Deviation

Statisticians wanted to develop a measure of spread that showed variability with respect to the center of the data, call it an "average deviation from center". This section will explore deviations from the sample mean and a later section will explore variability with respect to the sample median.

Example: Pizza delivery

Let's Return to the Anthony's pizza example in which a sample of 5 delivery runs by a driver showed that the total number of pizzas delivered on each run were {2, 2, 5, 9, 12}. Recall that the sample mean X for this data was 6, so we can calculate the deviation from the sample mean for each point:

Record number i	Pizzas delivered X_i	Deviation from mean $X_i - \bar{X}$
1	2	2 - 6 = - 4
2	2	2 - 6 = - 4
3	5	5 - 6 = -1
4	9	9 - 6 = +3
5	12	12 - 6 = +6
Total $\sum$		0

The sum of deviations from the mean will always equal zero, so we need a way to calculate an "average" deviation from the mean. Statisticians realized the sign of the deviation doesn't really matter so they explored statistics such as the absolute value of the deviation from the mean:

Record number i	Pizzas delivered X_i	Deviation from mean $X_i - \bar{X}$	Absolute value of Deviation from mean $\lvert X_i - \bar{X} \rvert$
1	2	2 - 6 = - 4	4
2	2	2 - 6 = - 4	4
3	5	5 - 6 = -1	1
4	9	9 - 6 = +3	3
5	12	12 - 6 = +6	6
Total $\sum$		0	18

Dividing by the sample size, we can find the "average absolute deviation from the mean" to be 18/5 = 3.6 pizzas. For reasons that will be explained in a later section, this measure was not found to be ideal.

Another method of eliminating negative signs from data is to square the numbers, since any negative numbers raised to an even power will become positive.

Record number i	Pizzas delivered X_i	Deviation from mean $X_i - \bar{X}$	Squared Deviation from the mean $(X_i - \bar{X})^2$

Record number i	Pizzas delivered X_i	Deviation from mean $X_i - \bar{X}$	Squared Deviation from the mean $(X_i - \bar{X})^2$
1	2	2 - 6 = - 4	16
2	2	2 - 6 = - 4	16
3	5	5 - 6 = -1	1
4	9	9 - 6 = +3	9
5	12	12 - 6 = +6	36
Total $\sum$		0	78

The quantity $\sum (X_i - \bar{X})^2 = 78$ is called the **sum of squared deviations** from the mean. To calculate an "average" square deviation, it is best for the sum of squared deviations to be divided by n-1 instead of by n (n is the sample size). This statistic is called the **sample variance** and referred to by the symbol s^2 .

$$\text{Sample Variance:} \quad s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

You might be asking “Since this is an average of squared deviations, why are we dividing by n-1 instead of by n?” The reason is that $\bar{X}$, the sample mean, uses the same data $X_1, X_2, \cdots, X_n$ so you can show mathematically that you only need to know n-1 points plus the sample mean to determine the sample variance. In statistics this is called n-1 **degrees of freedom**, and they will be explored in a later section.

For the pizza data, the sample variance is: $s^2 = \frac{78}{5-1} = 19.5$

Although the sample variance uses all the data and measures variability from the mean, the units of this statistic are squared when the deviations are squared. In our example, the sample variance is 19.5 pizzas-squared. To solve this problem, we can simply take the square root of the variance to return to the original units. This statistic is called **sample standard deviation** and is represented by the symbol s.

$$\text{Sample Standard Deviation:} \quad s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Example: Comparing high temperatures between San Francisco and St. Louis

Calculating the variance and standard deviation manually is tedious, so we will use technology to calculate summary statistics for 2016 daily high temperatures in San Francisco and St. Louis.

City	Sample Size	Median	Mean	Variance	Standard Deviation
San Francisco	366	64	64.3	44.0	6.64
St. Louis	366	73	69.6	391.3	19.78

The means and medians show that on average St. Louis is somewhat warmer than San Francisco. The variances and standard deviations show that there is much more variability in high temperatures for St. Louis, consistent with the dot plot shown at the beginning of this section.

Interpreting the Standard Deviation

A student once asked me about the distribution of score from a statistics midterm after she saw her score of 82 out of 100. I told her the distribution of test scores had a mean score of 70 and a standard deviation of 10. Most people would have an intuitive grasp of

the mean score as being the “average student’s score” and would say this student did better than average. However, having an intuitive grasp of standard deviation is more challenging. Fortunately, there is a tool to help us.

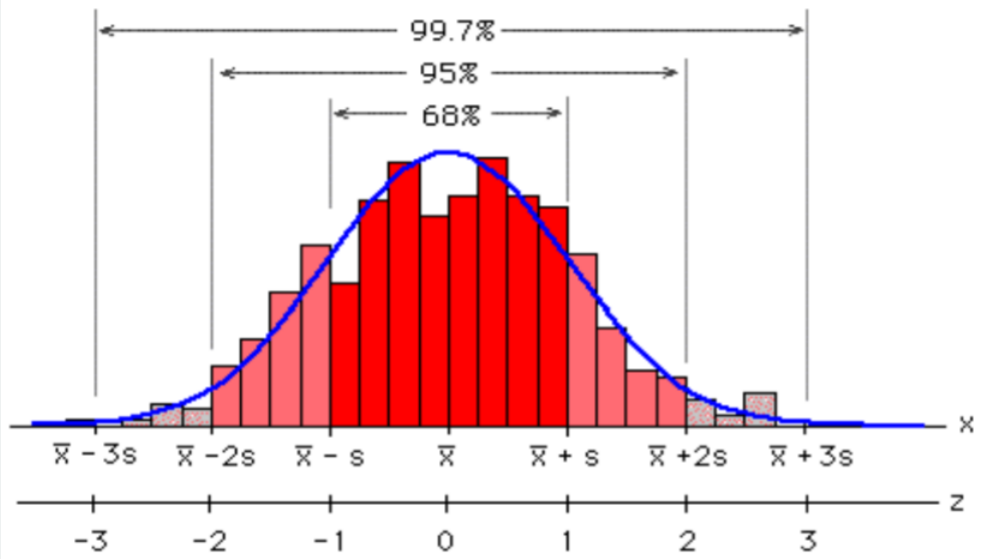
The Empirical Rule (68 – 95 – 99.7 Rule)

The Empirical Rule is a helpful tool in explaining standard deviation if you have data that is clustered towards the mean and not heavily skewed.

The standard deviation is a measure of variability or spread from the center of the data as defined by the mean.

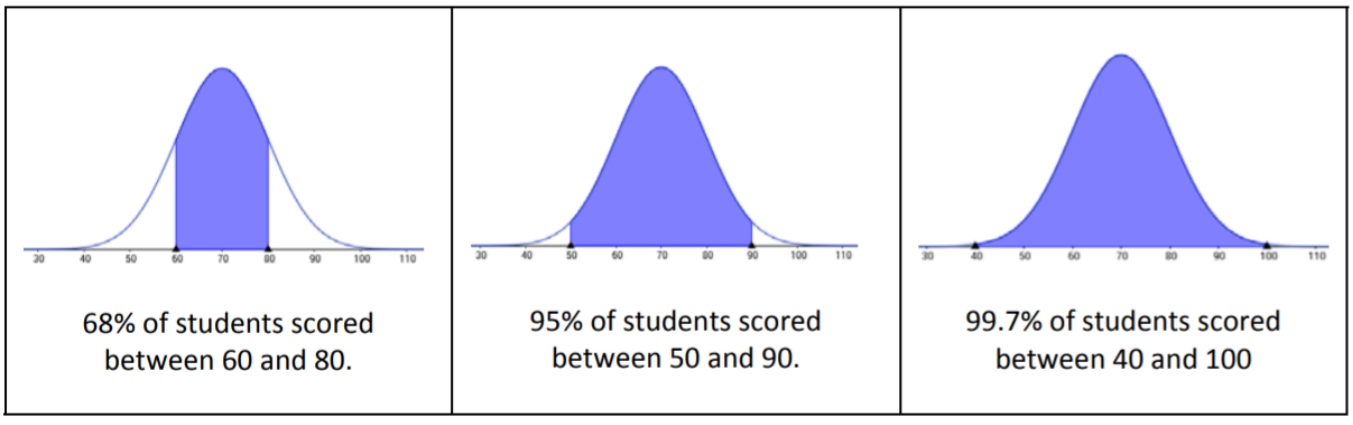
Note

The Empirical Rule states that for bell-shaped data:



- 68% of the data is within 1 standard deviation of the mean.
- 95% of the data is within 2 standard deviations of the mean.
- 99.7% of the data is within 3 standard deviations of the mean.

Here is an interpretation of the exam grades for the class in which the sample mean was 70 and the standard deviation was 10 using the Empirical Rule.



The student who scored an 82 would be in the upper 16% of the class, more than one standard deviation above the mean score.

Example: Students browsing the web

Let's return to the example of daily minutes spent on the internet by 30 students and use the empirical rule to find values between which 68%, 95% and 99.7% of the data lie. Compare these results to the actual results from the data.

67	71	78	82	85	86	87	87	92	95	97	99	99	100	101
102	103	103	104	105	105	107	108	109	112	116	118	122	124	125

Recall that the shape of this data is slightly skewed, but the data values cluster to the center. Let's see how close the Empirical Rule is to actual results.

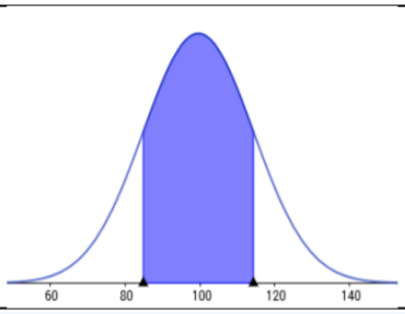
To use the Empirical Rule, we need to first calculate the sample mean and standard deviation.

$$\bar{X} = 99.6 \quad s = 14.7$$

The Empirical Rule says that about 68% of the data is within 1 standard deviation of the mean, between 84.9 and 114.3 minutes. The actual result for the data is 21/30 or 70% of the data.

Lower Bound:
$\bar{X} - 1s = 99.6 - 14.7 = 84.9$

Upper Bound
$\bar{X} + 1s = 99.6 + 14.7 = 114.3$

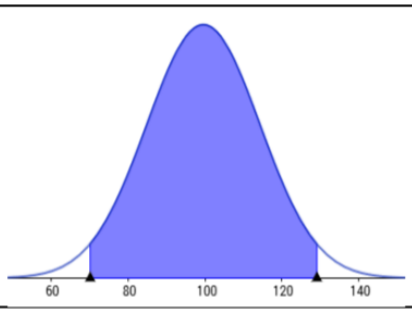


67	71	78	82	85
86	87	87	92	95
97	99	99	100	101
102	103	103	104	105
105	107	108	109	112
116	118	122	124	125

The Empirical Rule says that about 95% of the data is within 2 standard deviations of the mean, between 70.2 and 129.1 minutes. The actual result for the data is 29/30 or 96.7% of the data.

Lower Bound:
$\bar{X} - 2s = 99.6 - 2(14.7) = 70.2$

Upper Bound
$\bar{X} + 2s = 99.6 + 2(14.7) = 129.1$

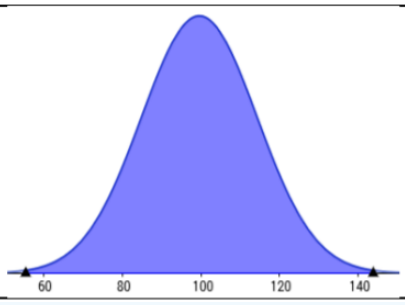


67	71	78	82	85
86	87	87	92	95
97	99	99	100	101
102	103	103	104	105
105	107	108	109	112
116	118	122	124	125

The Empirical Rule says that about 99.7% of the data is within 3 standard deviations of the mean, between 55.5 and 143.8 minutes. The actual result for the data is 30/30 or 100% of the data.

Lower Bound:
$\bar{X} - 3s = 99.6 - 3(14.7) = 55.5$

Upper Bound
$\bar{X} + 3s = 99.6 + 3(14.7) = 143.8$



67	71	78	82	85
86	87	87	92	95
97	99	99	100	101
102	103	103	104	105
105	107	108	109	112
116	118	122	124	125

So even though the time on internet data has some negative skewness, the actual percentages of data within 1, 2 and 3 standard deviations of the mean are close to the percentages from the Empirical Rule.

Using the range to estimate sample standard deviation.

The Empirical Rule also gives a very quick rule for making a rough estimate of the standard deviation.

Rough estimate of Sample Standard Deviation using Range

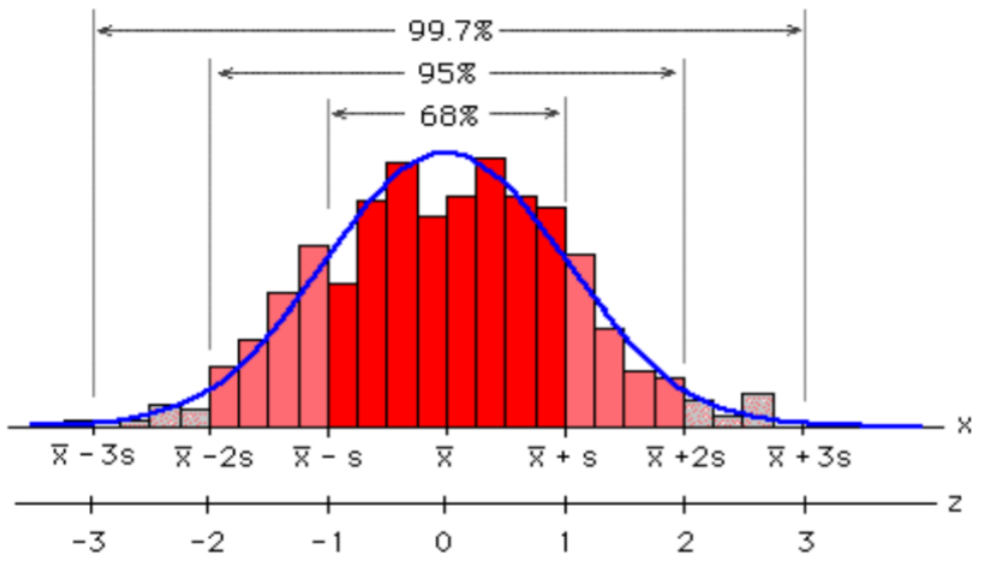
For small sample sizes (between 15 and 70): $s \approx \text{Range}/4$

For intermediate sample sizes (between 70 and 500): $s \approx \text{Range}/5$

For large sample sizes (over 500): $s \approx \text{Range}/6$

Example: Students browsing the web

In the prior example of time spent on the Internet by 30 students, we determined the Range to be 58. Using this rule, we would estimate the sample standard deviation to be 58/4 = 14.5 minutes. This rough estimate is actually quite close to the calculated sample standard deviation of 14.7 minutes.



This rule should not be used to determine the actual standard deviation, but can be used to check the reasonableness of a calculated or presented sample standard deviation.
