Measures of Spread

Objectives

1. Determine the range of a dataset.
2. Determine the variance and standard deviation of a dataset.

The Range

Range

The **range** of a dataset is found by subtracting the minimum value from the maximum value.

Example 1

During a heat wave one summer, I decided to cool off by drinking milkshakes everyday for a week. The number of milkshakes I had each day is shown:

9, 2, 7, 10, 3, 4, 12

Find the range for the number of milkshakes I drank that week.

Example 1

During a heat wave one summer, I decided to cool off by drinking milkshakes everyday for a week. The number of milkshakes I had each day is shown:

$$9, 2, 7, 10, 3, 4, 12$$

Find the range for the number of milkshakes I drank that week.

Max: 12

Example 1

During a heat wave one summer, I decided to cool off by drinking milkshakes everyday for a week. The number of milkshakes I had each day is shown:

$$9, 2, 7, 10, 3, 4, 12$$

Find the range for the number of milkshakes I drank that week.

Max: 12   Min: 2

Example 1

During a heat wave one summer, I decided to cool off by drinking milkshakes everyday for a week. The number of milkshakes I had each day is shown:

$$9, 2, 7, 10, 3, 4, 12$$

Find the range for the number of milkshakes I drank that week.

Max: 12   Min: 2

Range: $12 - 2 = 10$

Disadvantage to Using Range to Measure Spread of Data

A disadvantage of relying solely on the range as a measure of variation is that it is heavily affected by outliers (extreme values).

Objectives

1. Determine the range of a dataset.
2. Determine the variance and standard deviation of a dataset.

Deviation from Mean

Deviation from the Mean

The ***deviation from the mean** refers to how far a data value, x, is from the mean; found by

$$x - \text{mean}$$

.

Deviation from Mean

Deviation from the Mean

The ***deviation from the mean** refers to how far a data value, x, is from the mean; found by

$$x - \text{mean}$$

.

A data value that is above the mean has a **positive deviation** and one that is below the mean has a **negative deviation**.

Example 2

Find the mean of the following gas prices:

$$\$3.25, \$3.40, \$3.21, \$3.38$$

Example 2

Find the mean of the following gas prices:

$3.25, $3.40, $3.21, $3.38

Total: $13.24

Example 2

Find the mean of the following gas prices:

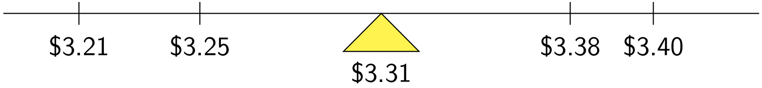
$$\$3.25, \$3.40, \$3.21, \$3.38$$

Total: \$13.24

Mean: $\$13.24/4 = \$3.31$

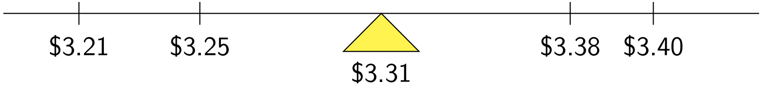
Visual Interpretation of the Mean

We can think of the mean as a “balancing point” for our dataset:



Visual Interpretation of the Mean

We can think of the mean as a “balancing point” for our dataset:



Each data point has a deviation (or *distance*) from the mean.

Example 3

Calculate each data point's deviation from the mean.

\$3.25, \$3.40, \$3.21, \$3.38 Mean: \$3.31

Example 3

Calculate each data point's deviation from the mean.

\$3.25, \$3.40, \$3.21, \$3.38 Mean: \$3.31

Price	**Deviation from Mean**
\$3.25	−\$0.06
\$3.40	\$0.09
\$3.21	−\$0.10
\$3.38	\$0.07

Deviations from Mean

Now, let's get an idea of how much, on average, the data is spread out from the mean.

Deviations from Mean

Now, let's get an idea of how much, on average, the data is spread out from the mean.

We can do that by calculating the mean of the deviations we got in the last example.

Example 4

Find the mean of the deviations in gas prices from the mean:

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Example 4

Find the mean of the deviations in gas prices from the mean:

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Total: 0

Example 4

Find the mean of the deviations in gas prices from the mean:

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Total: 0

Mean: $0/4 = 0$

Now What?

It turns out that this will **always** be the case because the total for the positive deviations will always equal the total for negative deviations.

Now What?

It turns out that this will **always** be the case because the total for the positive deviations will always equal the total for negative deviations.

In order to remedy this issue, of cancelling, we can do one of two things:

Now What?

It turns out that this will **always** be the case because the total for the positive deviations will always equal the total for negative deviations.

In order to remedy this issue, of cancelling, we can do one of two things:

- Take the absolute value of the deviations.

Now What?

It turns out that this will **always** be the case because the total for the positive deviations will always equal the total for negative deviations.

In order to remedy this issue, of cancelling, we can do one of two things:

- Take the absolute value of the deviations.
- Take the squares of the deviations.

Now What?

It turns out that this will **always** be the case because the total for the positive deviations will always equal the total for negative deviations.

In order to remedy this issue, of cancelling, we can do one of two things:

- Take the absolute value of the deviations.
- Take the squares of the deviations.

While the mean of the absolute values of the deviations has its uses (called the *mean absolute deviation*) in terms of calculations, it is better to work with the squares of the deviations instead.

Example 5

Square each of the deviations in gas prices, then find the mean of the squared deviatons.

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Example 5

Square each of the deviations in gas prices, then find the mean of the squared deviatons.

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Squared Deviations:

$$0.0036 \quad 0.0081 \quad 0.01 \quad 0.0049$$

Example 5

Square each of the deviations in gas prices, then find the mean of the squared deviatons.

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Squared Deviations:

$$0.0036 \quad 0.0081 \quad 0.01 \quad 0.0049$$

Total: 0.0266

Example 5

Square each of the deviations in gas prices, then find the mean of the squared deviatons.

$$-\$0.06 \quad \$0.09 \quad -\$0.10 \quad \$0.07$$

Squared Deviations:

$$0.0036 \quad 0.0081 \quad 0.01 \quad 0.0049$$

Total: 0.0266

Mean: $0.0266/4 = 0.00665$

Population Variance

The result of the previous example is known as the **population variance** and is denoted by

$$\sigma^2$$

(the lowercase Greek letter sigma, squared).

Population Variance

The result of the previous example is known as the **population variance** and is denoted by

$$\sigma^2$$

(the lowercase Greek letter sigma, squared).

Just like mean has a sample mean and a population mean, variance also has a **sample variance** and is denoted

$$s^2$$

Population Variance vs. Sample Variance

Sample variance is similar to population variance *except* instead of dividing by the total number of observations, like we did in the gas prices examples, we instead divide by one less than the number of observations (called the **degrees of freedom**).

Population Variance vs. Sample Variance

Sample variance is similar to population variance *except* instead of dividing by the total number of observations, like we did in the gas prices examples, we instead divide by one less than the number of observations (called the **degrees of freedom**).

Remember, the sum of the deviations from the mean must always equal 0.

Population Variance vs. Sample Variance

Sample variance is similar to population variance *except* instead of dividing by the total number of observations, like we did in the gas prices examples, we instead divide by one less than the number of observations (called the **degrees of freedom**).

Remember, the sum of the deviations from the mean must always equal 0.

In a data set with 4 entries, the first 3 entries can be any number we want (note that 4 - 1 = 3). The last entry <u>must</u> make it so that the sum of the deviations from the mean equals 0.

Population Variance vs. Sample Variance

Sample variance is similar to population variance *except* instead of dividing by the total number of observations, like we did in the gas prices examples, we instead divide by one less than the number of observations (called the **degrees of freedom**).

Remember, the sum of the deviations from the mean must always equal 0.

In a data set with 4 entries, the first 3 entries can be any number we want (note that 4 - 1 = 3). The last entry <u>must</u> make it so that the sum of the deviations from the mean equals 0.

Thus, in a data set with n elements, the first $n - 1$ elements can be whatever they want, but that last *n*-th element is forced to cause the deviation from the mean to equal 0.

Population Variance vs. Sample Variance

Because of this, we say that the sample variance is an colorblue**unbiased estimator** of the population variance (that is, the difference between the expected value and the actual value is 0).

Population Variance vs. Sample Variance

Because of this, we say that the sample variance is an colorblue**unbiased estimator** of the population variance (that is, the difference between the expected value and the actual value is 0).

The formulas for population variance and sample variance are below:

Population Variance	**Sample Variance**
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

The Units of Measurement with Variance

The issue with using variance as the primary measure of variation is that variance gives us squared units. The answer to the above example is in square dollars.

The Units of Measurement with Variance

The issue with using variance as the primary measure of variation is that variance gives us squared units. The answer to the above example is in square dollars.

Since we don't normally talk about *squared* dollars, we need to do something to bring those units back down to the same units (dollars) that we started with.

The Units of Measurement with Variance

The issue with using variance as the primary measure of variation is that variance gives us squared units. The answer to the above example is in square dollars.

Since we don't normally talk about *squared* dollars, we need to do something to bring those units back down to the same units (dollars) that we started with.

The **standard deviation** is the square root of the variance:

standard deviation $= \sqrt{\text{variance}}$

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad \text{and} \quad s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$