Introduction to Numerical Statistics

Colm Mulcahy

Math 107-03, Spring 2020, Spelman College

1 Apr 2020

We discuss different measures of average and spread

We discuss different measures of average and spread

The word "average" (in reference to a collection of numbers) is ambiguous in English and has three totally different meanings.

We discuss different measures of average and spread

The word “average” (in reference to a collection of numbers) is ambiguous in English and has three totally different meanings.

It can refer to **mean**, **median**, or **mode**, each of which will be discussed.

We discuss different measures of average and spread

The word “average” (in reference to a collection of numbers) is ambiguous in English and has three totally different meanings.

It can refer to **mean**, **median**, or **mode**, each of which will be discussed.

There are also several ways to measure the spread of a collection of numbers.

We discuss different measures of average and spread

The word "average" (in reference to a collection of numbers) is ambiguous in English and has three totally different meanings.

It can refer to **mean**, **median**, or **mode**, each of which will be discussed.

There are also several ways to measure the spread of a collection of numbers.

The most important of these is the **standard deviation**: it's messy to work out especially if the **mean** (or the original collection of numbers) involves many decimals.

We discuss different measures of average and spread

The word "average" (in reference to a collection of numbers) is ambiguous in English and has three totally different meanings.

It can refer to **mean**, **median**, or **mode**, each of which will be discussed.

There are also several ways to measure the spread of a collection of numbers.

The most important of these is the **standard deviation**: it's messy to work out especially if the **mean** (or the original collection of numbers) involves many decimals.

Today we will learn how to compute it; in future classes we will learn what it signifies and how to use it to answer interesting questions.

Statistics: Averages (6A in text)

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

The **mean** (add them up and divide by 6) is roughly 81.66666667

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

The **mean** (add them up and divide by 6) is roughly 81.66666667

The **median** (take the middle one, or the mean of the two middle ones) is 85.

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

The **mean** (add them up and divide by 6) is roughly 81.66666667

The **median** (take the middle one, or the mean of the two middle ones) is 85.

The **mode** ("the most popular one") is 90.

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

The **mean** (add them up and divide by 6) is roughly 81.66666667

The **median** (take the middle one, or the mean of the two middle ones) is 85.

The **mode** ("the most popular one") is 90.

Repeat for:

60, 70, 80, 90, 90

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 70, 80, 90, 90, 100

The **mean** (add them up and divide by 6) is roughly 81.66666667

The **median** (take the middle one, or the mean of the two middle ones) is 85.

The **mode** ("the most popular one") is 90.

Repeat for:

60, 70, 80, 90, 90

The mean is 78, the median is 80 and the mode is 90.

Statistics: Averages (6A in text)

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 65, 70, 75, 80, 90, 100

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 65, 70, 75, 80, 90, 100

The mean is roughly 77.142857143, the median is 75 and there is no mode!

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 65, 70, 75, 80, 90, 100

The mean is roughly 77.142857143, the median is 75 and there is no mode!

Repeat for:

65, 65, 85, 70, 75

Statistics: Averages (6A in text)

What is the average of these test scores?

60, 65, 70, 75, 80, 90, 100

The mean is roughly 77.142857143, the median is 75 and there is no mode!

Repeat for:

65, 65, 85, 70, 75

The mean is 72, the median is 70 and the mode is 65.

The Mean Formula

Given a list of n numbers $x_1, x_2, \ldots, x_n$ we can compute their mean using the formula:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

The Mean Formula

Given a list of n numbers $x_1, x_2, \ldots, x_n$ we can compute their mean using the formula:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ugly decimals may well arise, depending on the denominator.

The Mean Formula

Given a list of n numbers $x_1, x_2, \ldots, x_n$ we can compute their mean using the formula:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ugly decimals may well arise, depending on the denominator.

While we can report the answer correct to 2 decimal places in many siutations, we will still need to use 8 or 9 decimal places when using the mean to work out another important "summmary" number, the standard deviation.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

1. First we list the numbers vertically, in the left column of a table.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

1. First we list the numbers vertically, in the left column of a table.

2. Next, we add them up and compute the mean $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

1. First we list the numbers vertically, in the left column of a table.
2. Next, we add them up and compute the mean $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$.
3. Then we work out the deviations: subtracting $\bar{x}$ from each x_i.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

1. First we list the numbers vertically, in the left column of a table.

2. Next, we add them up and compute the mean $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$.

3. Then we work out the deviations: subtracting $\bar{x}$ from each x_i. This always gives some negative and some positive numbers.

Standard Deviation

The **Standard Deviation** of a collection of numbers is a number obtained from them by a multi-stage process.

It is usually calculator intensive, and it helps to organize our work in a three-column table.

1. First we list the numbers vertically, in the left column of a table.

2. Next, we add them up and compute the mean $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$.

3. Then we work out the deviations: subtracting $\bar{x}$ from each x_i. This always gives some negative and some positive numbers. We put those in the second colum of the table.

Standard Deviation

5. We now work out the squares of the deviations

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table.

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

6. Now we add up the squares of the deviations

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

6. Now we add up the squares of the deviations, write that at the bottom of the third column of the table.

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

6. Now we add up the squares of the deviations, write that at the bottom of the third column of the table.

7. Divide that by $n - 1$.

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

6. Now we add up the squares of the deviations, write that at the bottom of the third column of the table.

7. Divide that by $n - 1$. (dividing by n would have yielded the mean of the squares ofthe deviations, but that's not quite what we do!)

Standard Deviation

5. We now work out the squares of the deviations: this always gives positive numbers (or zero), which we place in the third column of the table. *Make sure there are no negative numbers here!*

6. Now we add up the squares of the deviations, write that at the bottom of the third column of the table.

7. Divide that by $n - 1$. (dividing by n would have yielded the mean of the squares ofthe deviations, but that's not quite what we do!)

8. Take the square root of the number obtained in Step 7.

WE'RE DONE!

A simple example of standard deviation computation

The data 40, 75, 95 have sum 210, and hence mean 70.

A simple example of standard deviation computation

The data 40, 75, 95 have sum 210, and hence mean 70.

data	data - mean	$(\text{data - mean})^2$
40	-30	900
75	5	25
95	25	625
210	0	1550

A simple example of standard deviation computation

The data 40, 75, 95 have sum 210, and hence mean 70.

data	data - mean	$(\text{data - mean})^2$
40	-30	900
75	5	25
95	25	625
210	0	1550

Divide 1550 by $3 - 1 = 2$ to get 775.

A simple example of standard deviation computation

The data 40, 75, 95 have sum 210, and hence mean 70.

data	data - mean	(data - mean)^2
40	-30	900
75	5	25
95	25	625
210	0	1550

Divide 1550 by $3 - 1 = 2$ to get 775. Finally, sqrt(775) is about 27.83882181415, which is 27.84 to 2 decimals of accuracy.

A simple example of standard deviation computation

The data 40, 75, 95 have sum 210, and hence mean 70.

data	data - mean	(data - mean)^2
40	-30	900
75	5	25
95	25	625
210	0	1550

Divide 1550 by $3 - 1 = 2$ to get 775. Finally, sqrt(775) is about 27.83882181415, which is 27.84 to 2 decimals of accuracy.

This was easy as only whole numbers were involved until the end.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Show that the standard devation for the data 79, 80, 81 is 1.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Show that the standard devation for the data 79, 80, 81 is 1.

Don't be fooled by those! Generally one can't guess the answer.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Show that the standard devation for the data 79, 80, 81 is 1.

Don't be fooled by those! Generally one can't guess the answer.

Show that the standard devation for 78, 80, 81 is about 1.53.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Show that the standard devation for the data 79, 80, 81 is 1.

Don't be fooled by those! Generally one can't guess the answer.

Show that the standard devation for 78, 80, 81 is about 1.53.

Show that the standard devation for 60, 70, 80, 90, 90, 100 is about 14.72.

Simple examples to check for yourself

Show that the standard devation for the data 70, 80, 90 is 10.

Show that the standard devation for the data 60, 80, 100 is 20.

Show that the standard devation for the data 79, 80, 81 is 1.

Don't be fooled by those! Generally one can't guess the answer.

Show that the standard devation for 78, 80, 81 is about 1.53.

Show that the standard devation for 60, 70, 80, 90, 90, 100 is about 14.72.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880. The mean is about 125.71428571.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880. The mean is about 125.71428571.

data	data - mean	$(\text{data - mean})^2$
98	-27.71428571	768.08163265
140	14.28571429	204.08163265
130	4.28571429	18.36734694
120	-5.71428571	32.65306122
130	4.28571429	18.36734694
102	-23.71428571	562.36734694
160	34.28571429	1175.51020408
880	0	2779.43000000

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880. The mean is about 125.71428571.

data	data - mean	(data - mean)^2
98	-27.71428571	768.08163265
140	14.28571429	204.08163265
130	4.28571429	18.36734694
120	-5.71428571	32.65306122
130	4.28571429	18.36734694
102	-23.71428571	562.36734694
160	34.28571429	1175.51020408
880	0	2779.43000000

Divide 2779.43 by $7 - 1 = 6$ to get 463.23833333.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880. The mean is about 125.71428571.

data	data - mean	(data - mean)^2
98	-27.71428571	768.08163265
140	14.28571429	204.08163265
130	4.28571429	18.36734694
120	-5.71428571	32.65306122
130	4.28571429	18.36734694
102	-23.71428571	562.36734694
160	34.28571429	1175.51020408
880	0	2779.43000000

Divide 2779.43 by $7 - 1 = 6$ to get 463.23833333. Finally, sqrt(463.23833333) is about 21.52297222, which is 21.52 to 2 decimals of accuracy.

Another example

Systolic blood pressure data for 7 people: 98, 140, 130, 120, 130, 102, 160. These sum to 880. The mean is about 125.71428571.

data	data - mean	(data - mean)^2
98	-27.71428571	768.08163265
140	14.28571429	204.08163265
130	4.28571429	18.36734694
120	-5.71428571	32.65306122
130	4.28571429	18.36734694
102	-23.71428571	562.36734694
160	34.28571429	1175.51020408
880	0	2779.43000000

Divide 2779.43 by $7 - 1 = 6$ to get 463.23833333. Finally, sqrt(463.23833333) is about 21.52297222, which is 21.52 to 2 decimals of accuracy.

One key fact about standard deviation

For large data sets, abour two thirds (actually 68%) of the data is within one standard devatrion of the mean.

One key fact about standard deviation

For large data sets, abour two thirds (actually 68%) of the data is within one standard devatrion of the mean.

If 3000 women's heights had mean 5 feet 5 inches, with standard devation 2 inches, then about 2000 of the women's heights would be between 5 feet 3 inches and 5 feet 7 inches.

One key fact about standard deviation

For large data sets, abour two thirds (actually 68%) of the data is within one standard devatrion of the mean.

If 3000 women's heights had mean 5 feet 5 inches, with standard devation 2 inches, then about 2000 of the women's heights would be between 5 feet 3 inches and 5 feet 7 inches.

If 6000 men's heights had mean 5 feet 9 inches, with standard devation 3 inches, then about XXXXXX of the men's heights would be between YYYYYY and ZZZZZZ.