

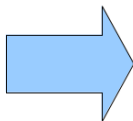
# Descriptive Statistics

CE 311S

# **MEASURES OF LOCATION AND VARIABILITY**

As a starting point, we need a way to briefly summarize an entire sample with simple numerical values.

```
0000000000003 0000000000066 01 |0000 040508 2400 040508|0060 01 c 100 30 100
0000 0000 0000 0000 0000 0000 0000 0000 0000 333333777770000
                                US 290 4.5 miles w of FM 1960
01 12      0001 0002 0003 0004 0005 0006 0007 0008 0009 0010 0011 0012
00 00
00 00
00 00 3 |0100 0054 0047 0039 0170 0192 0063 0083 0216 0227 0057 0051 0018
00 00 3 |0200 0020 0015 0012 0108 0124 0038 0046 0150 0141 0039 0025 0009
00 00 3 |0300 0011 0015 0008 0068 0100 0026 0038 0139 0134 0029 0030 0005
00 00 3 |0400 0018 0008 0007 0079 0104 0015 0037 0116 0096 0030 0026 0005
00 00 3 |0500 0009 0014 0013 0112 0157 0039 0035 0129 0101 0018 0027 0005
00 00 3 |0600 0023 0022 0042 0214 0296 0139 0103 0242 0129 0073 0034 0013
00 00 3 |0700 0062 0043 0085 0275 0384 0172 0305 0562 0380 0148 0078 0022
00 00 3 |0800 0127 0093 0161 0398 0497 0262 0546 0768 0519 0270 0132 0085
00 00 3 |0900 0178 0126 0284 0528 0640 0413 0653 0859 0645 0366 0190 0134
00 00 3 |1000 0231 0170 0371 0663 0809 0534 0926 1009 0788 0526 0260 0212
00 00 3 |1100 0288 0186 0396 0772 0896 0625 1086 1151 0935 0610 0322 0268
00 00 3 |1200 0367 0237 0513 0845 1039 0731 1054 1160 1003 0657 0424 0262
00 00 3 |1300 0344 0258 0460 0846 1086 0903 1085 1214 1095 0745 0460 0317
00 00 3 |1400 0397 0351 0463 0956 1175 0993 1113 1217 1080 0713 0436 0317
00 00 3 |1500 0407 0316 0556 0950 1208 1063 1144 1232 1116 0689 0461 0309
00 00 3 |1600 0433 0318 0490 0971 1294 1089 1136 1203 1083 0665 0465 0298
00 00 3 |1700 0440 0323 0502 1073 1304 1194 0876 1097 0996 0695 0455 0288
00 00 3 |1800 0418 0314 0488 1043 1354 1230 0846 1090 0986 0631 0407 0290
00 00 3 |1900 0399 0319 0441 1030 1286 1105 0707 0939 0896 0550 0390 0287
00 00 3 |2000 0381 0258 0403 0933 1154 1006 0516 0777 0741 0460 0332 0245
00 00 3 |2100 0337 0243 0214 0813 0976 0789 0360 0586 0632 0319 0266 0134
00 00 3 |2200 0286 0193 0178 0669 0885 0607 0336 0560 0544 0247 0210 0132
00 00 3 |2300 0153 0126 0137 0467 0547 0307 0277 0475 0424 0212 0152 0075
00 00 3 |2400 0093 0081 0081 0387 0455 0214 0148 0304 0300 0146 0120 0060
```



1845

This is the realm of **descriptive statistics**.

For now, we consider two main types of descriptive statistic.

**Measures of location** describe what a “typical” value of the variable in a sample.

**Measures of variability** describe how close the variables in the sample are to a “typical” value.

For now, we consider two main types of descriptive statistic.

**Measures of location** describe what a “typical” value of the variable in a sample.

**Measures of variability** describe how close the variables in the sample are to a “typical” value.

We will define three measures of location: the **mean**, **median**, and **mode**.

If we have a sample consisting of  $n$  members of the population, we let  $x_i$  denote the value of the variable for the  $i$ -th member of the sample ( $1 \leq i \leq n$ )

The **(sample) mean** is defined as

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

## Example

The high temperatures for the last week are

76 50 58 67 65 74 74

What is the mean temperature in this sample?

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

## Solution

$$\bar{x} = \frac{76 + 50 + 58 + 67 + 65 + 74 + 74}{7} = 66.3$$

The mean temperature is 66.3 degrees.



## A few notes

- To be technically correct, this is a **sample mean** because it depends on the sample we have. However, it is often called just the **mean** when it is obvious it refers to a sample.

## A few notes

- To be technically correct, this is a **sample mean** because it depends on the sample we have. However, it is often called just the **mean** when it is obvious it refers to a sample.
- The sample mean is also known as the **arithmetic average**.

## A few notes

- To be technically correct, this is a **sample mean** because it depends on the sample we have. However, it is often called just the **mean** when it is obvious it refers to a sample.
- The sample mean is also known as the **arithmetic average**.
- It is often called just the **average**, but there are other types of averages too.

The **(sample) median** is defined as the “middle” value in the data set. Using the temperature data we had before,

76 50 58 67 65 74 74

we can rewrite these in increasing order, and identify the middle value:

50 58 65 **67** 74 74 76

so for this sample, the median is 67 degrees.

If the sample size is **odd**, the “middle” value is well-defined. If the sample size is **even**, there are two middle values.

In this case, the median is defined as the **mean of the two middle values**.

50 53 58 **65 67** 74 74 76

For this sample, the median is  $(65 + 67)/2 = 66$  degrees.

The **(sample) mode** is defined as the most frequently occurring value in the data set.

50 58 65 67 74 74 76

74 occurs most often (twice), so it is the sample mode.

The **(sample) mode** is defined as the most frequently occurring value in the data set.

50 58 65 67 74 74 76

74 occurs most often (twice), so it is the sample mode.

Unlike the mean and median, there can be more than one mode! If there is a tie for the most frequent observation, **all** of these values qualify as modes.

The **(sample) mode** is defined as the most frequently occurring value in the data set.

50 58 65 67 74 74 76

74 occurs most often (twice), so it is the sample mode.

Unlike the mean and median, there can be more than one mode! If there is a tie for the most frequent observation, **all** of these values qualify as modes.

Note: There are other conventions for tiebreaking with modes (e.g., some say that there is no mode if the values are all different). If there are many modes, you should really use something else anyway.



So, which measure of location is best?

So, which measure of location is best?

**IT DEPENDS!**

## Different measures of location tell different stories.

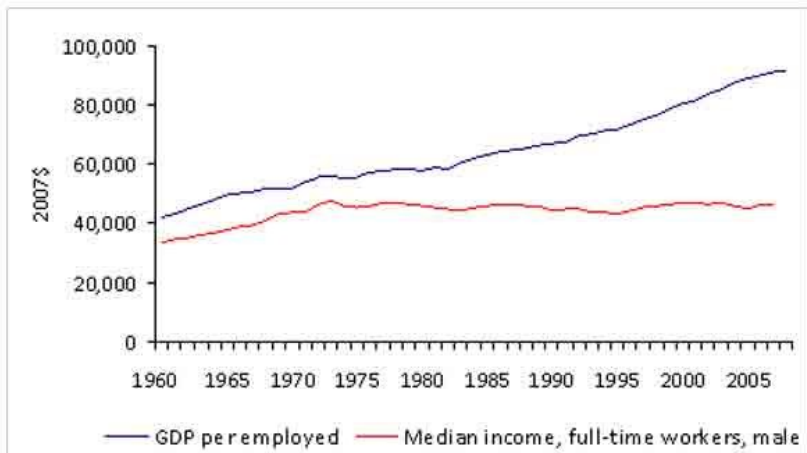
Let's say the proud country of Boylesland has three citizens, with net worths of \$20,000, \$30,000, and \$550,000, respectively.

What is the mean net worth and the median net worth?

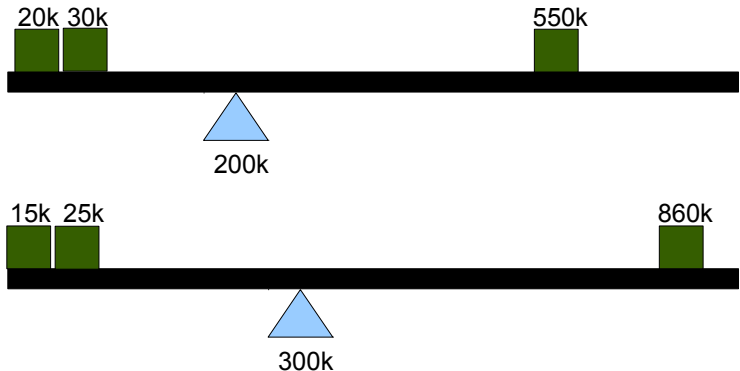
## Different measures of location tell different stories.

Next year, their net worths are \$15,000, \$25,000, and \$860,000, respectively.

What is the mean net worth and the median net worth?



One way to think about the mean is as the centroid or “balance point” of the data set.



One lesson from the story is that the **sample mean is highly affected by outliers**. Even just a few extremely high or extremely low values in a sample can have a dramatic impact on the sample mean. On the other hand, the median is robust to outliers.

So does this mean the median is better?

Let's go back to the first situation, with net worths of \$20,000, \$30,000, and \$550,000, respectively.

Now, next year the net worths are \$30,000, \$30,000, and \$600,000.

What are the new mean net worth and median net worth?



The moral of the story:

You **always** lose information when you reduce a data set to a single number. A single statistic is only one facet of the problem. You can often get a better view of the true situation by looking at multiple statistics.

# **MEASURES OF VARIABILITY**

None of the measures discussed so far address the variation within the data set. All of the following samples have exactly the same mean, median, and mode:

100	100	100	100	100	100
<hr/>					
98	99	100	100	101	102
<hr/>					
90	95	100	100	105	110
<hr/>					
0	50	100	100	150	200
<hr/>					
0	1	100	100	199	200
<hr/>					
0	0	100	100	100	300

None of the measures discussed so far address the variation within the data set. All of the following samples have exactly the same mean, median, and mode:

100	100	100	100	100	100
98	99	100	100	101	102
90	95	100	100	105	110
0	50	100	100	150	200
0	1	100	100	199	200
0	0	100	100	100	300

To distinguish between these, we develop **measures of variability**.

There are a few ways to think about measures of variability.

- They reflect the “consistency” of the sample from one observation to the next.
- They describe the spread in the data.
- If measures of location describe what a “typical” sample element looks like, measures of variability show how “typical” a typical element is.

The **sample variance** is defined as

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

The **sample variance** is defined as

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

The sample variance does not have the same units as the original sample; because it is useful to have a measure of variability with the same units, we define the **sample standard deviation** as

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

## Example

What are the sample variance and sample standard deviation of the following temperature data?

76 50 58 67 65 74 74



## Example

What are the sample variance and sample standard deviation of the following temperature data?

76 50 58 67 65 74 74

Notice that the formula for sample variance ( $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ ) includes the sample mean. We already calculated this to be 66.3. So

$$s^2 = \frac{(76 - 66.3)^2 + (50 - 66.3)^2 + \dots + (74 - 66.3)^2 + (74 - 66.3)^2}{6} = 91$$

and  $s = \sqrt{s^2} = 9.6$  degrees.

## Example

The data sets given at the beginning of these section have dramatically different sample variance and sample standard deviation:

## Example

The data sets given at the beginning of these section have dramatically different sample variance and sample standard deviation:

Data						$s^2$	$s$
100	100	100	100	100	100	0	0
98	99	100	100	101	102	2	1.4
90	95	100	100	105	110	50	7.1
0	50	100	100	150	200	5000	71
0	1	100	100	199	200	?	?
0	0	100	100	100	300	?	?

## Example

The data sets given at the beginning of these section have dramatically different sample variance and sample standard deviation:

## Example

The data sets given at the beginning of these section have dramatically different sample variance and sample standard deviation:

Data						$s^2$	$s$
100	100	100	100	100	100	0	0
98	99	100	100	101	102	2	1.4
90	95	100	100	105	110	50	7.1
0	50	100	100	150	200	5000	71
0	1	100	100	199	200	7920	89
0	0	100	100	100	300	12000	110

Where does the sample variance formula come from?

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

## Where does the sample variance formula come from?

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

This kind of looks like an average... what if it were  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$  instead?

## Where does the sample variance formula come from?

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

This kind of looks like an average... what if it were  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$  instead?

In this case,  $s^2$  would be the average value of  $(x_i - \bar{x})^2$ . What does this mean?



## Where does the sample variance formula come from?

$(x_i - \bar{x})^2$  is a measure of how far away each observation is from the sample mean. But why square it?

- What's wrong with taking the average value of  $x_i - \bar{x}$ ?

## Where does the sample variance formula come from?

$(x_i - \bar{x})^2$  is a measure of how far away each observation is from the sample mean. But why square it?

- What's wrong with taking the average value of  $x_i - \bar{x}$ ?
- What about taking the average value of  $|x_i - \bar{x}|$ ?

Where does the sample variance formula come from?

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

## Where does the sample variance formula come from?

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

So why do we divide by  $n - 1$  instead of  $n$ ?

## Where does the sample variance formula come from?

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

So why do we divide by  $n - 1$  instead of  $n$ ?

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

## Where does the sample variance formula come from?

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

So why do we divide by  $n - 1$  instead of  $n$ ?

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

The sample mean is probably not the **population** mean. (In the temperature example above, maybe the population mean is 66 degrees, while the mean of our sample was 66.3 degrees).

Thus, dividing by  $n$  would bias the sample variance low because our estimate of the mean is inaccurate. We'll talk more about this later in the course.