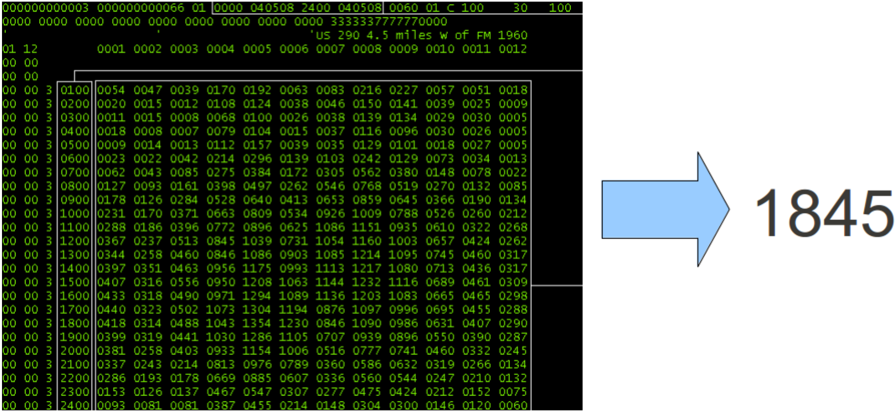
Descriptive Statistics

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.



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

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

Example

The high temperatures for the last week are

76 50 58 67 65 74 74

What is the mean temperature in this sample?

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

Solution

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

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.

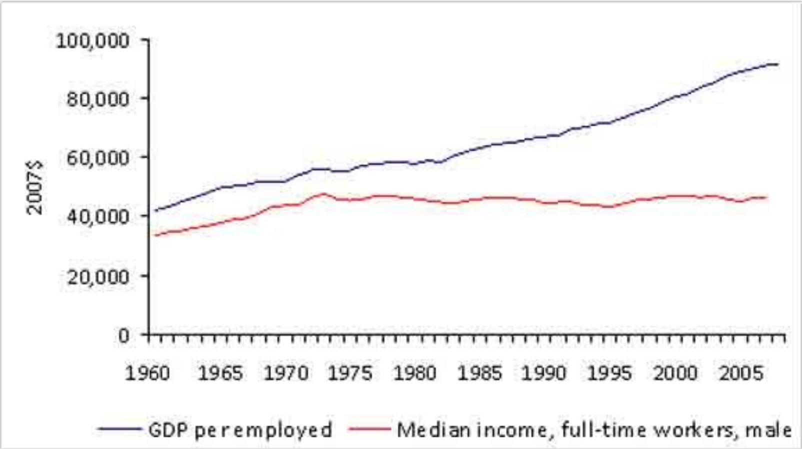
Let's say the proud country of Boylesland has three citizens, with net worths of \$20,000, \$30,000, and \$550,000, respectively.

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

Different measures of location tell different stories.

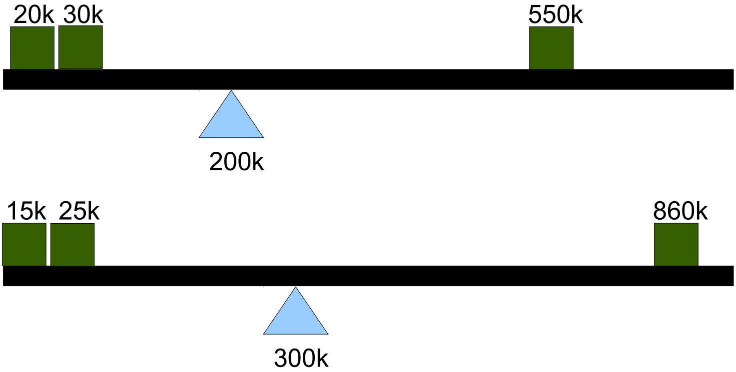
Next year, their net worths are \$15,000, \$25,000, and \$860,000, respectively.

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



100,000
80,000
60,000
40,000
20,000
0
2007$
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
GDP per employed
Median income, full-time workers, male

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
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

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

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

The **sample variance** is defined as

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

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

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

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?

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

Where does the sample variance formula come from?

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

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

Where does the sample variance formula come from?

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

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

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

Where does the sample variance formula come from?

$(x_i - \overline{x})^2$ 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 $x_i - \overline{x}$?

Where does the sample variance formula come from?

$(x_i - \overline{x})^2$ 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 $x_i - \overline{x}$?
- What about taking the average value of $|x_i - \overline{x}|$?

Where does the sample variance formula come from?

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

Where does the sample variance formula come from?

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

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

Where does the sample variance formula come from?

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

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

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

Where does the sample variance formula come from?

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

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

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

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.