Review Session 4

API-201, 10.08.21
Sophie Hill

Review sessions: new & improved!

Thanks to everyone who gave feedback on the Review Sessions in the mid-course evaluation!

Here are some concrete changes I'm making based on your comments (see my post in Slack for more detail):

1. Focus on bridging gap between class & PSets
2. Use small group work to encourage active engagement
3. Collect and address questions more systematically

What questions do you have on this week's material?

Plan:

- I will collect questions at start of session (and throughout)
- In the last 10 minutes of our session, we will return to the list and decide as a group which ones to prioritize

Agenda

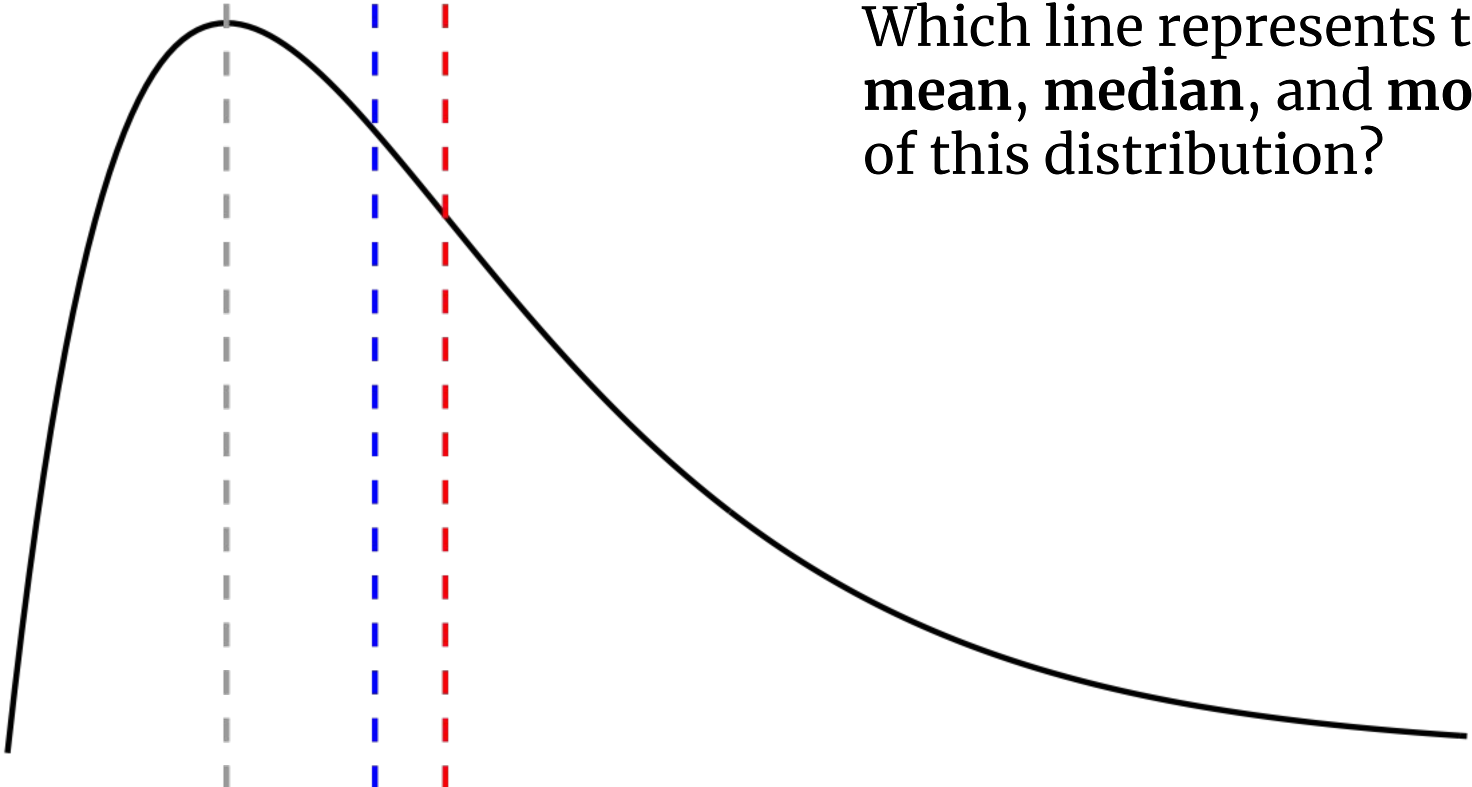
1. Measures of central tendency

2. Missing data

3. Excel skills

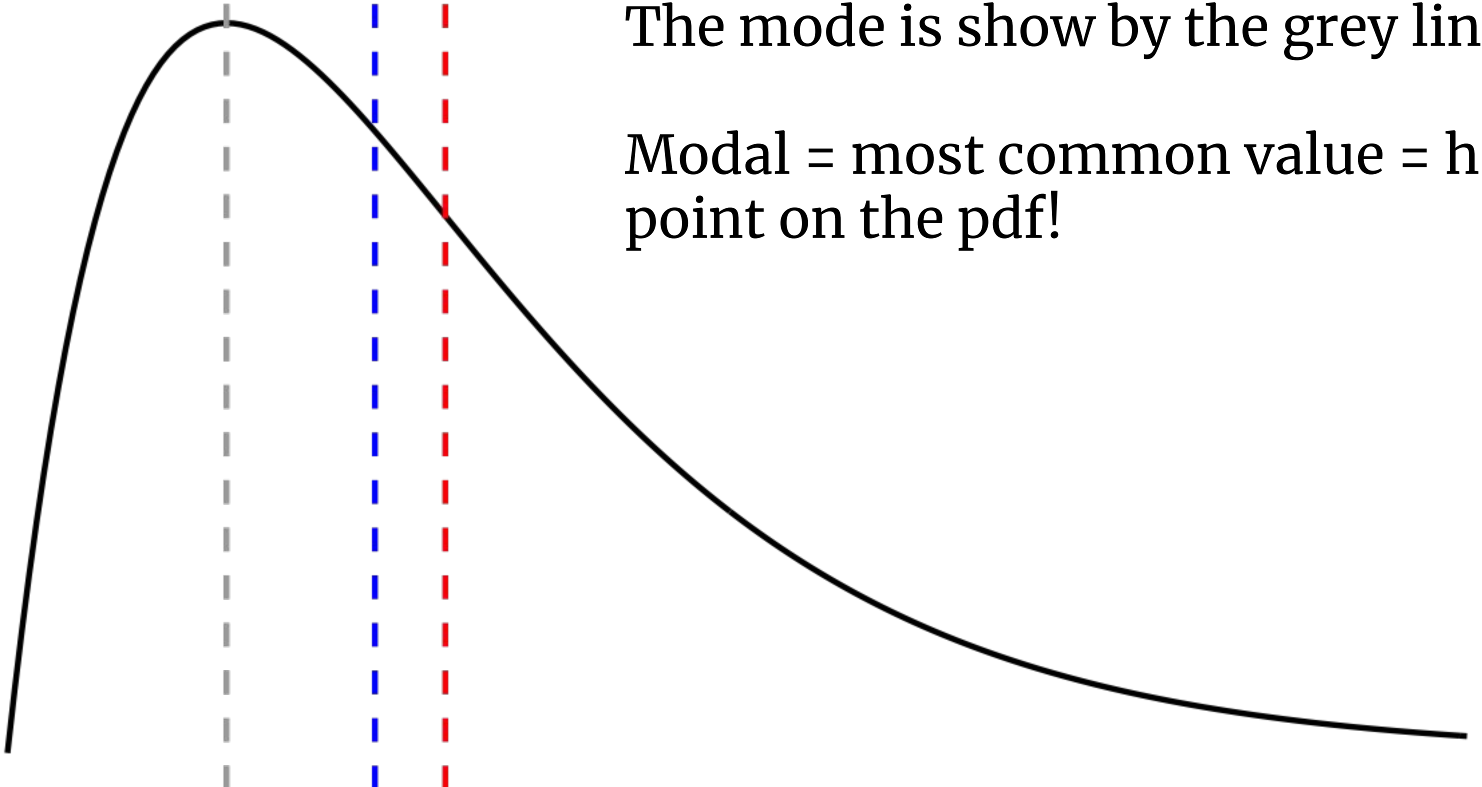
Measures of central tendency

Measures of central tendency



Which line represents the **mean**, **median**, and **mode** of this distribution?

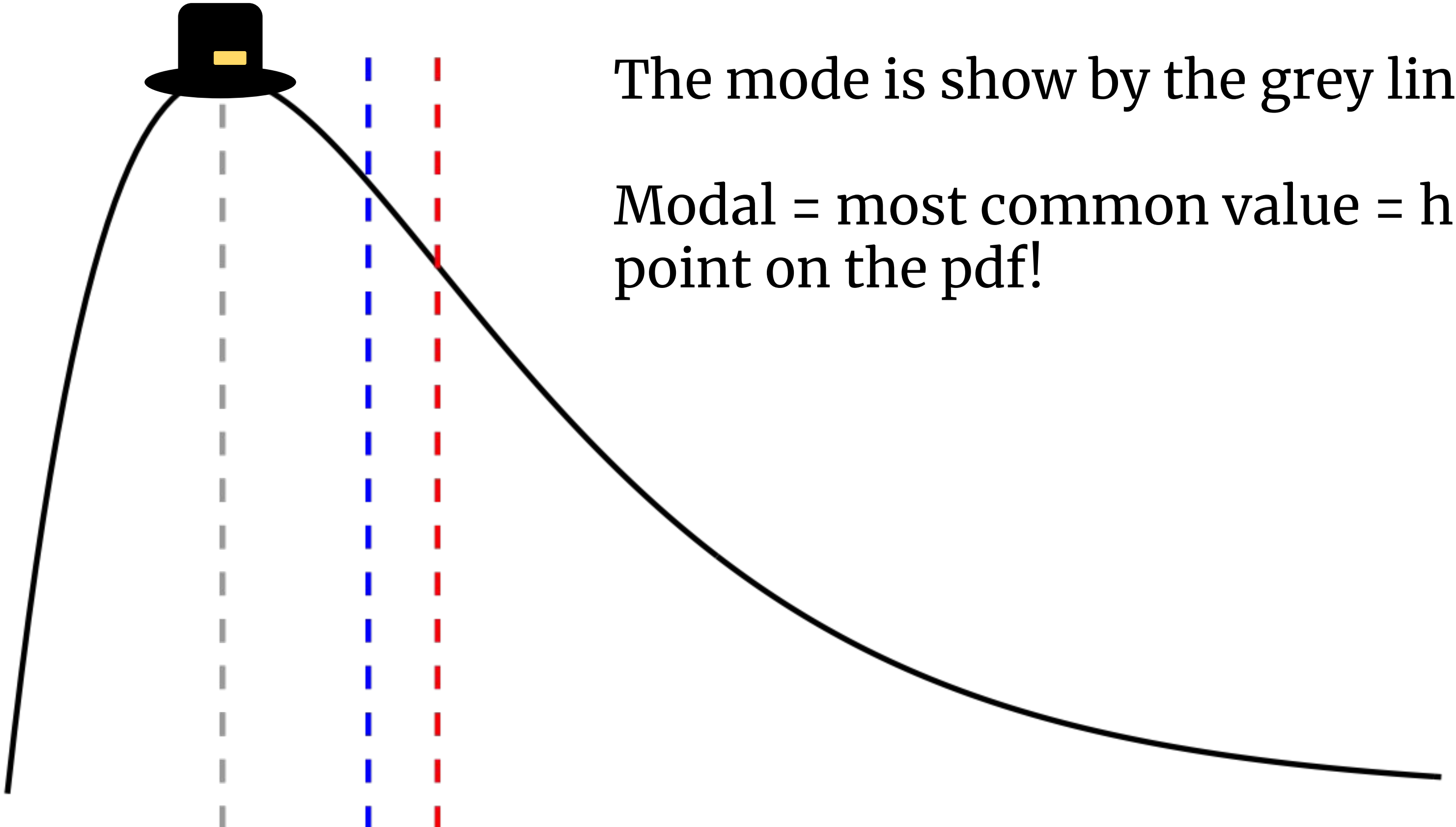
Measures of central tendency



The mode is show by the grey line.

Modal = most common value = highest point on the pdf!

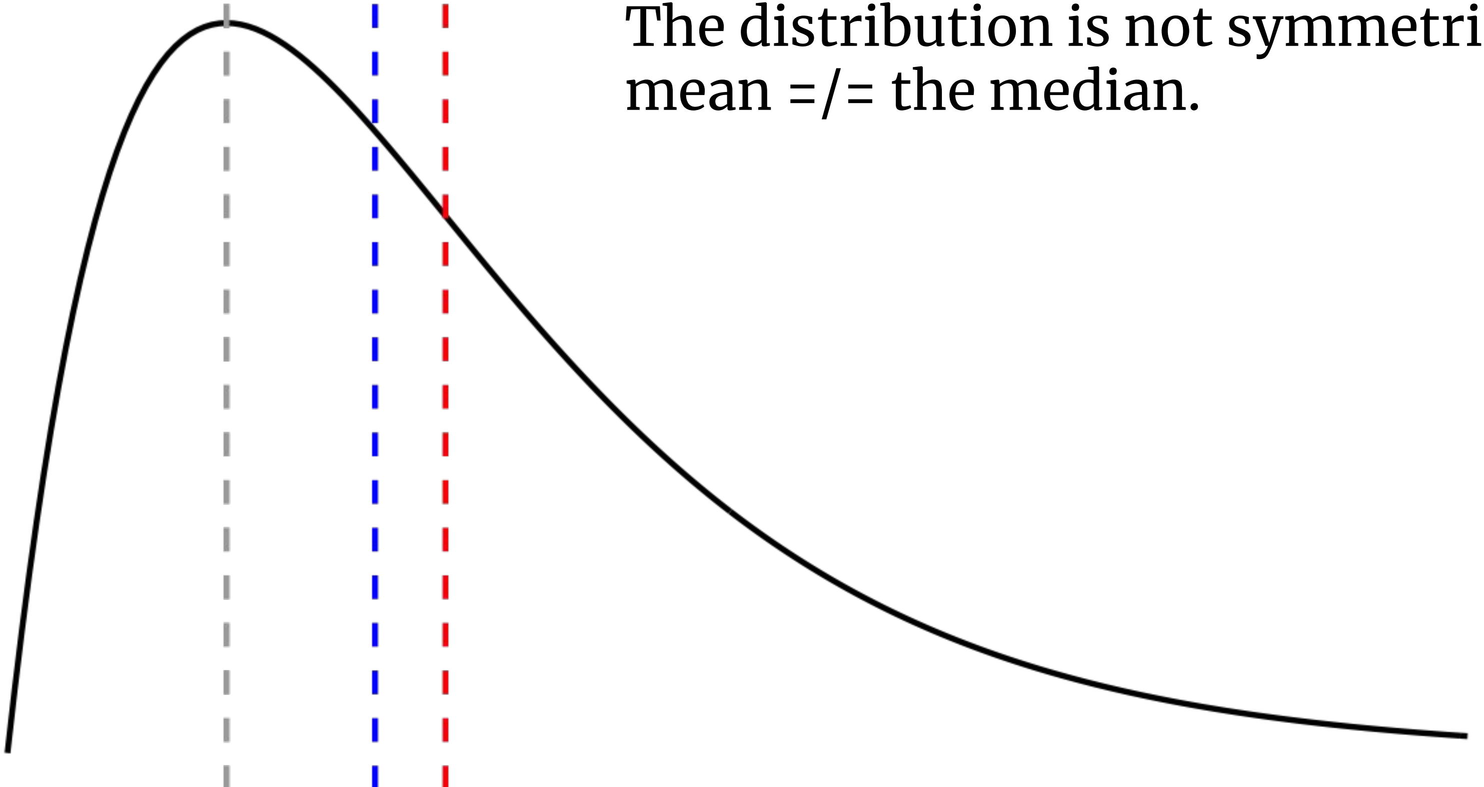
Measures of central tendency



The mode is show by the grey line.

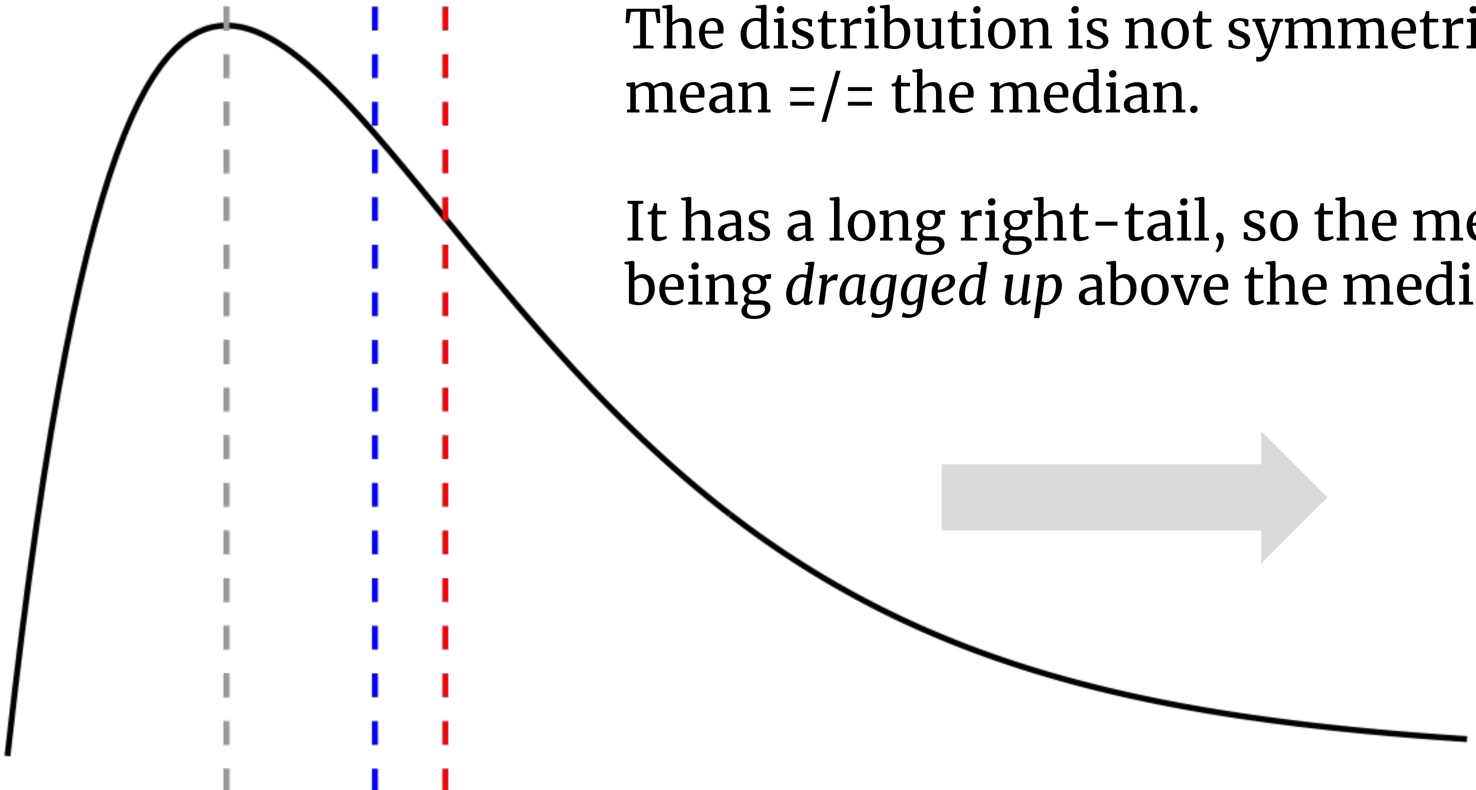
Modal = most common value = highest point on the pdf!

Measures of central tendency



The distribution is not symmetric, so the mean =/= the median.

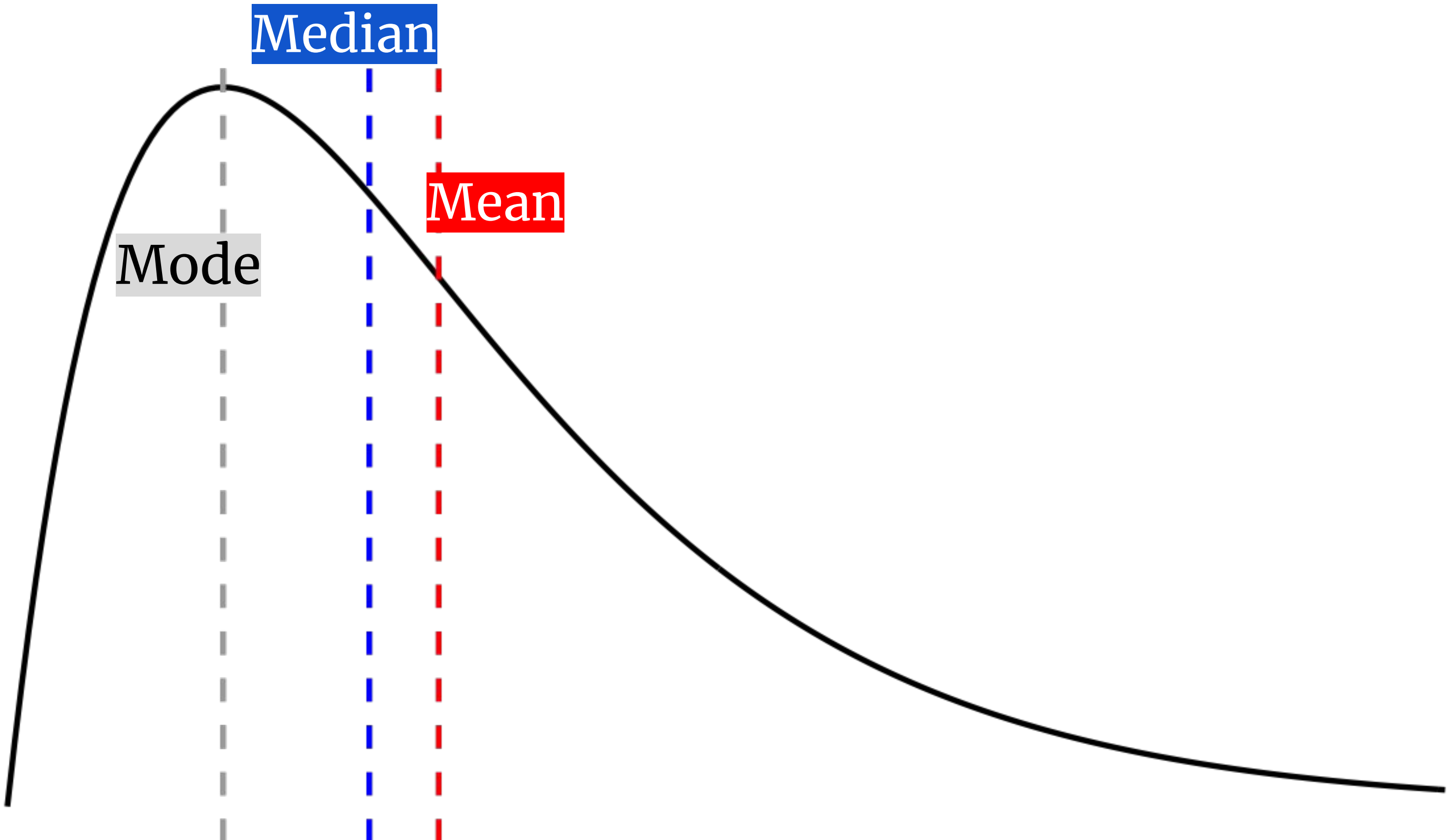
Measures of central tendency



The distribution is not symmetric, so the mean =/= the median.

It has a long right-tail, so the mean is being *dragged up* above the median...

Measures of central tendency



Missing data

Missing data

Suppose we conduct a poll of a representative sample of **1,000 US adults** about their drinking habits. We asked how many alcoholic drinks they have per week - however, the response to this question is **missing for 200** respondents. Among the other 800, we find that they report an average of **4 alcoholic drinks per week**.

Missing data

Suppose we conduct a poll of a representative sample of **1,000 US adults** about their drinking habits. We asked how many alcoholic drinks they have per week - however, the response to this question is **missing for 200** respondents. Among the other 800, we find that they report an average of **4 alcoholic drinks per week**.

Scenario 1: The data are missing because there was glitch in the survey software, which deleted every 5th response.

Scenario 2: The data are missing because some respondents decided to skip this question without answering.

Missing data

Discuss the differences between Scenarios 1 and 2. In particular, assess whether the missing data will cause you to under/over-estimate the average number of drinks per week among the population.

Suppose we conduct a poll of a representative sample of **1,000 US adults** about their drinking habits. We asked how many alcoholic drinks they have per week - however, the response to this question is **missing for 200** respondents. Among the other 800, we find that they report an average of **4 alcoholic drinks per week**.

Scenario 1: The data are missing because there was glitch in the survey software, which deleted every 5th response.

Scenario 2: The data are missing because some respondents decided to skip this question without answering.

Missing data: missing at random

In Scenario 1, the 200 responses are *missing at random.* (It's as if we took a random sample of 1,000 from the population, and then drew another random sample of 800 from that 1,000.)

Put another way: we have no reason to believe that the 200 response that were deleted would have been particularly low or high values.

In this case, 4 drinks per week is still a good (i.e. unbiased) estimate for the whole sample, and thus for the population.

Missing data: non-response bias

In Scenario 2, the 200 responses are *not missing at random.* High levels of alcohol consumption are stigmatized, so heavy drinkers will be more likely to skip the question compared to average drinkers.

In this case, 4 drinks per week would be an *underestimate* for the whole sample, and thus for the population.

Excel skills

You can do *almost* everything in PSet 4 Q1 just using tables...

To convert to table:

- Highlight cells, then go Data → Filter

To find highest/lowest values:

- Sort by variable descending / ascending

To count non-missing values:

- Filter to exclude blanks

… but you might want to impress your future boss by learning how to do it with formulas!

What you want	Formula	Tips
How many non-missing values?	Use COUNT() to count cells with numbers, use COUNTA() to count any non-blank cell (including text).	First, check what the “missing values” look like. Blank cell? “NA”? If necessary, convert them to blank cells. Note: if you use COUNTA() on a whole column, it will count the header row! So you would need to subtract 1.

… but you might want to impress your future boss by learning how to do it with formulas!

What you want	Formula	Tips
The 7th highest / lowest value	=LARGE(B:B, 7) =SMALL(B:B, 7)	Make sure your variable is correctly formatted as a number! Note: in Excel, numbers are right-aligned.

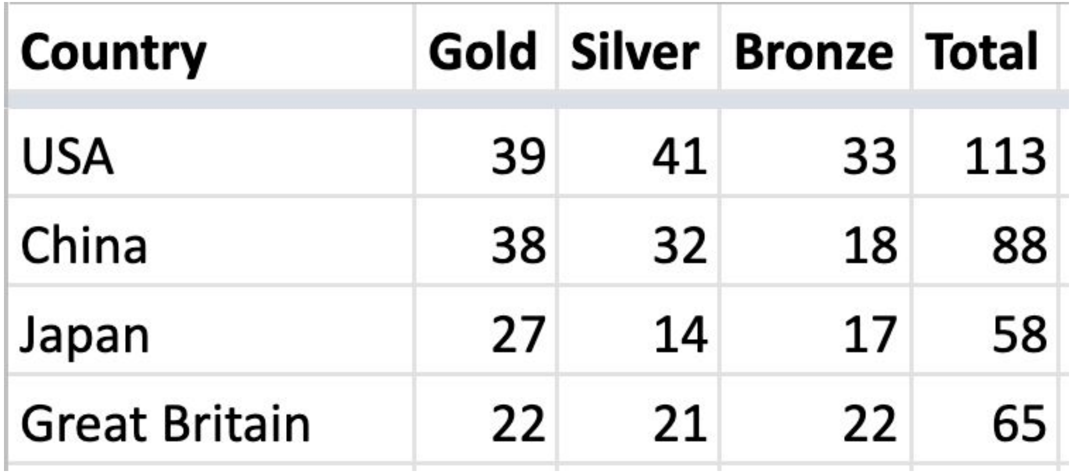
… but you might want to impress your future boss by learning how to do it with formulas!

What you want	Formula	Tips
The name of the observation with a particular value, X	Find row number of the observation that has value X in column B: =MATCH(X, B:B, 0) Look up the value in that row in the name column, say, column A: =INDEX(A:A, rownum)	Take it step by step at first! Manually check that the lookups are working.

Let’s practice!

Example dataset: Medal tables from the Tokyo Olympics!

Link to Excel worksheet



Country	Gold	Silver	Bronze	Total
USA	39	41	33	113
China	38	32	18	88
Japan	27	14	17	58
Great Britain	22	21	22	65

FAQs

FAQs

1. What's the difference between the correlation and the slope?

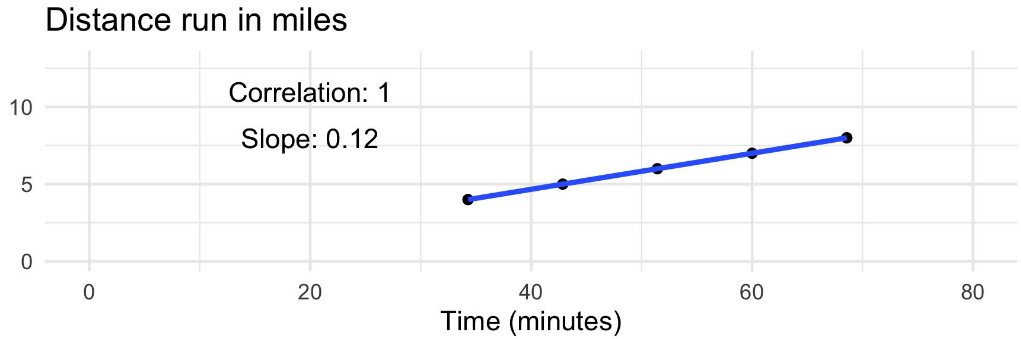
2. When does adding an observation increase/decrease the standard deviation?

3. What is the intuition behind the formulas for variance, correlation, etc.?

What’s the difference between the correlation and the slope of the line of best fit?

Suppose that you go for a 5-10 mile run every Saturday and you decide to plot your latest times vs distance covered.

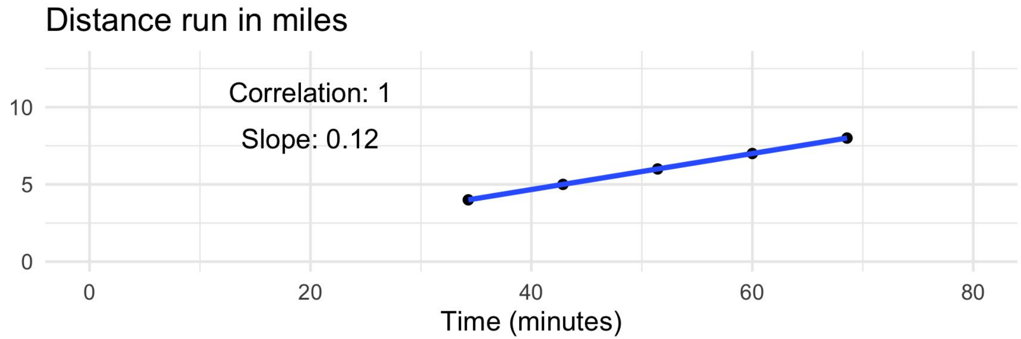
All your times lie exactly on the line of best fit, which slopes upwards, so it’s a perfect positive correlation of +1.



What’s the difference between the correlation and the slope of the line of best fit?

Suppose that you go for a 5-10 mile run every Saturday and you decide to plot your latest times vs distance covered.

All your times lie exactly on the line of best fit, which slopes upwards, so it’s a perfect positive correlation of +1.



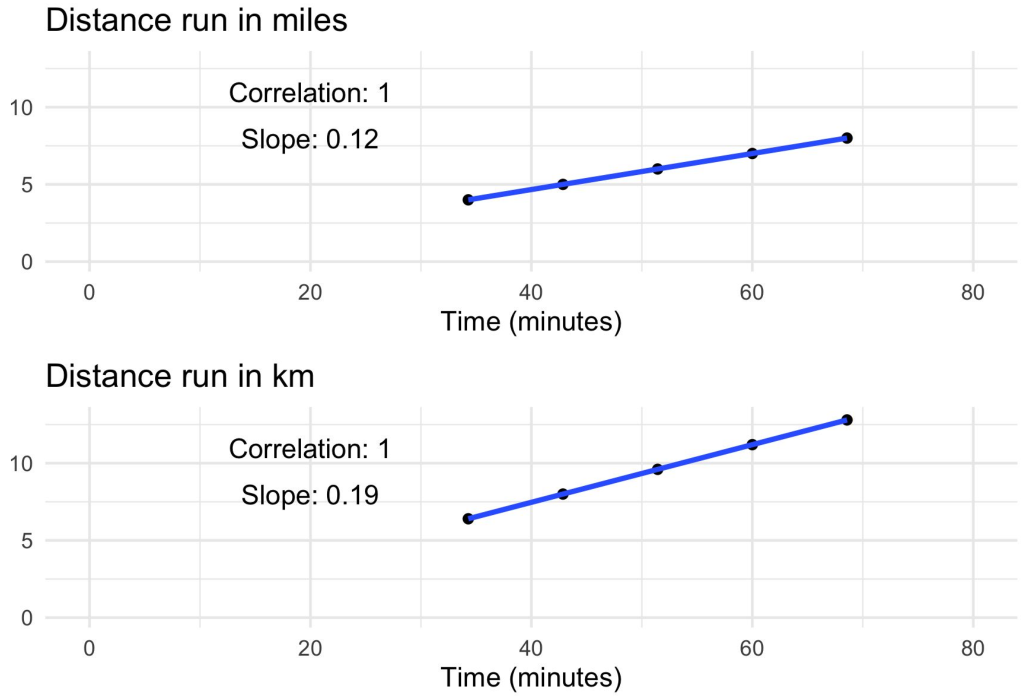
You’d like to compare times with your European friend, but first you’ll need to convert those distances into km.

- What happens to the correlation?
- What happens to the slope?

What’s the difference between the correlation and the slope of the line of best fit?

Changing the unit of distance has changed the slope of the line of best fit (it is steeper now).

But the points still lie exactly on the line, so the correlation is still +1!

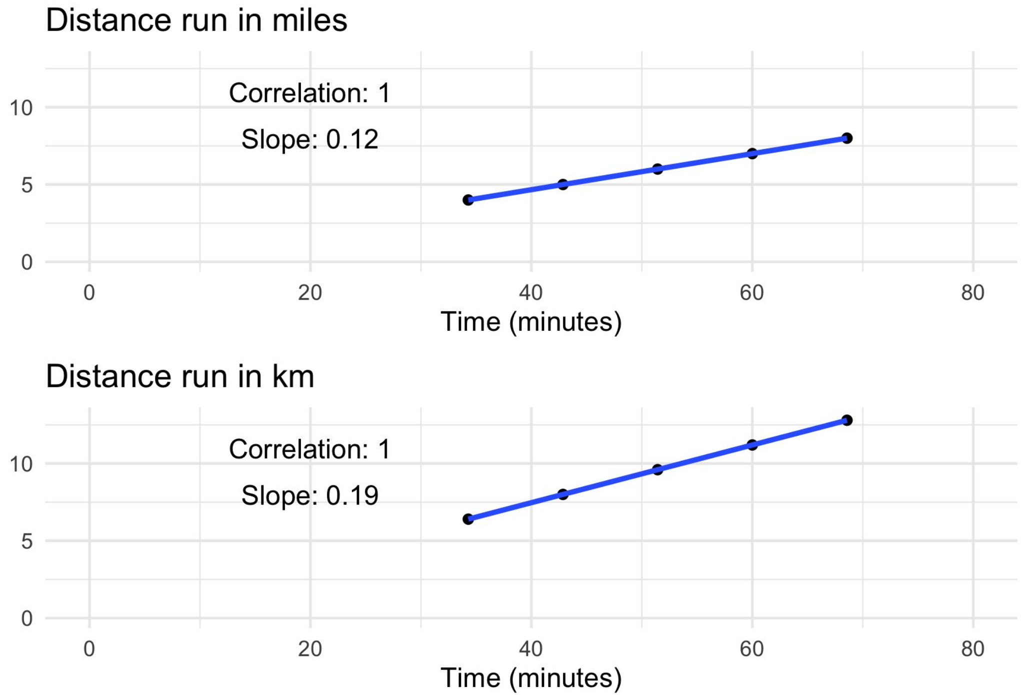


What's the difference between the correlation and the slope of the line of best fit?

Key differences:

- Slope depends on units, correlation does not

- Slope is about the *steepness* of the line, correlation is about the *tightness* to the line



When does adding an observation increase the standard deviation?

Here's a sample of numbers: {5, 2, 5, 8, 3, 7}

Mean = 5

Standard deviation = 2.1

When does adding an observation increase the standard deviation?

Here's a sample of numbers: {5, 2, 5, 8, 3, 7}

Mean = 5

Standard deviation = 2.1

Let's add a new number, 4, so our sample is: {5, 2, 5, 8, 3, 7, 4}

Mean =

Standard deviation =

When does adding an observation increase the standard deviation?

Here’s a sample of numbers: {5, 2, 5, 8, 3, 7}

Mean = 5

Standard deviation = 2.1

Let’s add a new number, 4, so our sample is: {5, 2, 5, 8, 3, 7, 4}

⬇ Mean = 4.9

⬇ Standard deviation = 2.0

When does adding an observation increase the standard deviation?

Here’s a sample of numbers: {5, 2, 5, 8, 3, 7}

Mean = 5

Standard deviation = 2.1

Let’s switch that out for a new number, 1, so our sample is: {5, 2, 5, 8, 3, 7, 1}

Mean =

Standard deviation =

When does adding an observation increase the standard deviation?

Here’s a sample of numbers: {5, 2, 5, 8, 3, 7}

Mean = 5

Standard deviation = 2.1

Let’s switch that out for a new number, 1, so our sample is: {5, 2, 5, 8, 3, 7, 1}

⬇ Mean = 4.4

⬆ Standard deviation = 2.4

When does adding an observation increase the standard deviation?

Intuition:

- Standard deviation = square root of the variance
- Variance = average of the squared deviations
- So the variance increases when the new value has a squared deviation that is above average, i.e. greater than the variance
- So the standard deviation increases when the new value has a deviation that is greater than the old standard deviation

When does adding an observation increase the standard deviation?

Intuition:

- Standard deviation = square root of the variance
- Variance = average of the squared deviations
- So the variance increases when the new value has a squared deviation that is above average, i.e. greater than the variance
- So the standard deviation increases when the new value has a deviation that is greater than the old standard deviation

Note: this is a bit hand-wavy! If we derive this formally we find a slightly stronger condition: SD increases if:

$$|x - \bar{x}| > \mathrm{sd}\sqrt{1 + \frac{1}{n}}$$

What is the intuition behind these formulas?

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

Variance

Visualization

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Correlation

Visualization