Data Analysis Grab-Bag

percent change

$$\frac{\text{NEW - OLD}}{\text{OLD}}$$

If teacher salaries were $31,500 in 2017 and $32,000 in 2018, we can say:

- Teacher salaries increased by $500.
  - 32000-31500 = 500
- Teacher salaries increased by 1.6 PERCENT.
  - (32000-31500)/31500 = 0.01587

It works the same way with a decrease. If teacher salaries were $31,500 in 2017 and $30,000 in 2018, we can say:

- Teacher salaries decreased by $1,500.
  - 30000-31500 = -1500
- Teacher salaries decreased by 4.8 PERCENT.
  - (30000-31500)/31500 = -0.04762

percent change of a percent

$$\boxed{\frac{\text{NEW - OLD}}{\text{OLD}}}$$

What if we're dealing with changes to something that's ALREADY measured as a percentage?

If 25% of teachers had a masters degree in 2017 and 30% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree increased by 5 PERCENTAGE POINTS.
  - 30-25 = 5
- The share of teachers with a masters degree increased by 20 PERCENT.
  - (30-25)/25 = 0.20

Or, if we're in a decrease situation: If 25% of teachers had a masters degree in 2017 and 18% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree decreased by 7 PERCENTAGE POINTS.
  - 18-25 = -7
- The share of teachers with a masters degree decreased by 28 PERCENT.
  - (18-25)/25 = -0.28

per-capita

How many murders were there in New York City versus Austin, Texas?

To get a reasonable comparison, be sure to account for how many people live in each place!

city	homicide_rate_2017	population_2017	homicides_per_capita	
New York City	290	8,622,698	3.4	per 100,000
Austin, Texas	29	931,830	3.1	per 100,000
Detroit	267	672,795	39.7	per 100,000

** numbers not fact-checked!*

choosing your denominator wisely

How should we measure participation in an election in a particular county?

Some options:

- votes cast / registered voters in the county
- votes cast / eligible voters in the county
- votes cast / U.S. citizens who are at least 18 yrs old
- votes cast / people who live in the county

There's not necessarily a RIGHT answer. You're answering a different question with each option.

risk ratio

group	population	event	risk	
group 1	x	a	a/x	<- risk for group 1
group 2	y	b	b/y	<- risk for group 2

risk ratio = (risk for group 1) / (risk for group 2)

Risk ratios can take any value from 0 to Infinity.

- RR > 1: group 1 has higher risk than group 2
- RR = 1: the two groups have the same risk
- RR < 1: group 1 has lower risk than group 2

risk ratio, cont.

group	population	number died	percent died		
men	142582	4526	3.17%	<- risk of death for men	
women	251440	12573	5.00%	<- risk of death for women	

In this case, the risk ratio for women relative to men is 5.00/3.17 = 1.58. Women are 58% more likely to die than men.

If you like the framing better, you can calculate the risk ratio for men relative to women, which is 3.17/5.00 = 0.63. You could write this up two ways:

- Men are 63% as likely to die as women.
- Men are 37% less likely to die than women. (1 - 0.63 = 0.37)

risk ratio, cont.

group	population	number died	percent died	
men	142582	4526	3.17%	<- risk of death for men
women	251440	24573	9.77%	<- risk of death for women

What if the numbers are further apart? Now the risk ratio for women relative to men is 9.77/3.17 = 3.08.

Here's options for how to write about it:

- Women are about three times AS likely to die as men.
- Women are two times MORE likely to die than men.
- The death rate for women is three times that of men.

correlation

Calculate the correlation between two columns in Google Sheets with

```
=CORREL(A2:A,B2:B)
```

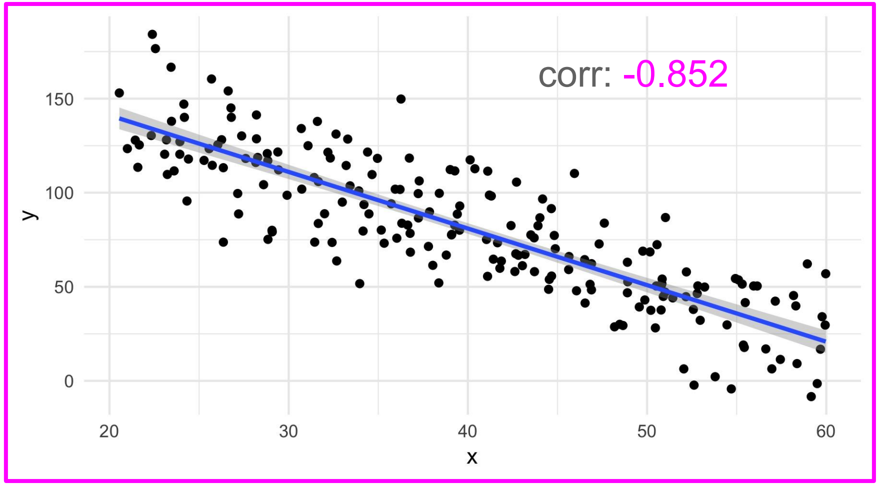
This will return a single value, a correlation coefficient. The value measures how close the two variables are to having a perfectly linear relationship with each other.

It will always be between -1 and 1.

- -1 : perfectly negatively correlated
- 0 : no correlation
- 1 : perfectly positively correlated

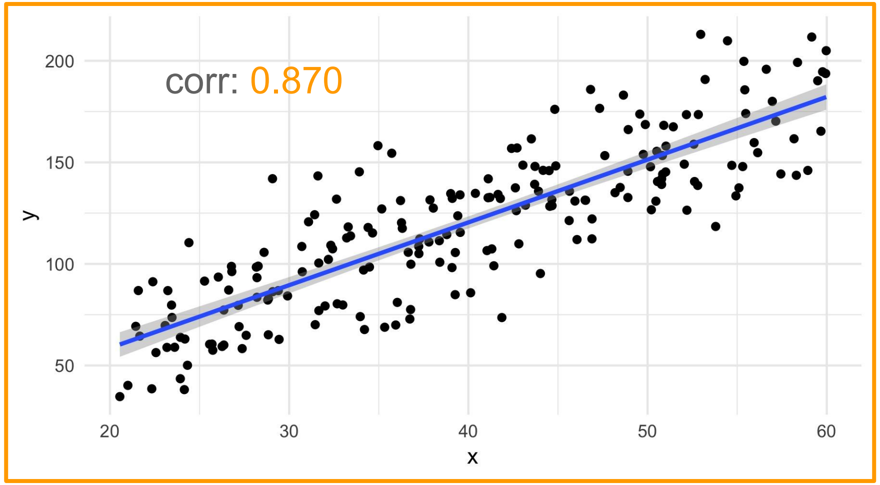
Negative correlation:

as one variable goes up, the other goes down



Positive correlation:

as one variable goes up, the other also goes up



Slope and correlation are different concepts