

جامعة نيويورك أبوظبي



PSYCH-UH 1004Q: Statistics for Psychology

Class 15: Linear correlation

Prof. Jon Sprouse
Psychology

The fundamental intuition of correlation

What is correlation?

It is when two variables covary with each other. This means that they move together. If one increases, the other either increases or decreases.

A classic example of correlation is **height** and **weight** in humans: as height increases, so does weight. We have intuitions about this because we know that there is a biological mechanism driving this.



person	1	2	3	4	5	6	7	8	9
height (cm)	188	163	201	196	175	152	190	168	180
weight (kg)	76	62	80	95	68	50	88	61	77

What is correlation?

If we sort these values by one of the variables, in this case height, we can begin to see the correlation:

unsorted

person	1	2	3	4	5	6	7	8	9
height (cm)	188	163	201	196	175	152	190	168	180
weight (kg)	76	62	80	95	68	50	88	61	77

sorted by
height

person	6	2	8	5	9	1	7	4	3
height (cm)	152	163	168	175	180	188	190	196	201
weight (kg)	50	62	61	68	77	76	88	95	80

In general, we see that the people with taller heights tend to have heavier weights (though, of course, it is not perfect).

Some important things to keep in mind

Correlation can occur for either **discrete** or **continuous** variables. In this course we will focus correlation of **continuous** variables, but your book also discusses discrete variables.

The way that the two variable covary can either be **linear** or **non-linear**. In this course, we will only study **linear correlation**.

Notice that we are making **two measurements** for each person - their height and their weight. This is how correlation works. It involves two variables that describe different aspects of the same units. In psychology, those units will often be people. But, in principle, it could be anything (like cities on Earth - you could measure their median income and median political alignment).

person	1	2	3	4	5	6	7	8	9
height (cm)	188	163	201	196	175	152	190	168	180
weight (kg)	76	62	80	95	68	50	88	61	77

Some important things to keep in mind

Finally, notice that to see whether the two variables covary, we need to sort them according to one of the variables (here, height) and see if the other variable is also sorted to some degree. That is what it means to covary. When one changes, the other changes in a related way (either increases or decreases).

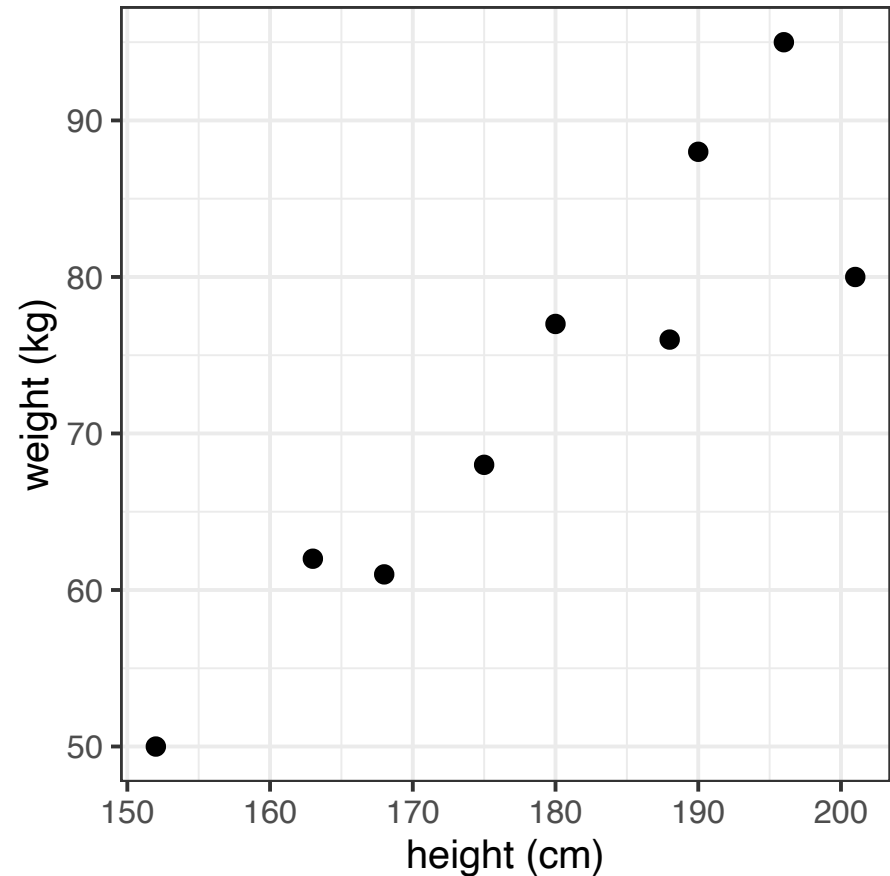
person	6	2	8	5	9	1	7	4	3
height (cm)	152	163	168	175	180	188	190	196	201
weight (kg)	50	62	61	68	77	76	88	95	80

It is easier to see in a scatter plot

The best way to explore correlations is with a scatter plot. So let's spend a few slides becoming familiar with a scatter plot.

The first thing to note is that **each dot is an experimental unit**. In this case, a person. And that dot represents two measurements. The x-axis is one of the measurements (height) and the y is the other (weight).

Here is the data table again so we can see that each point is one of our people:

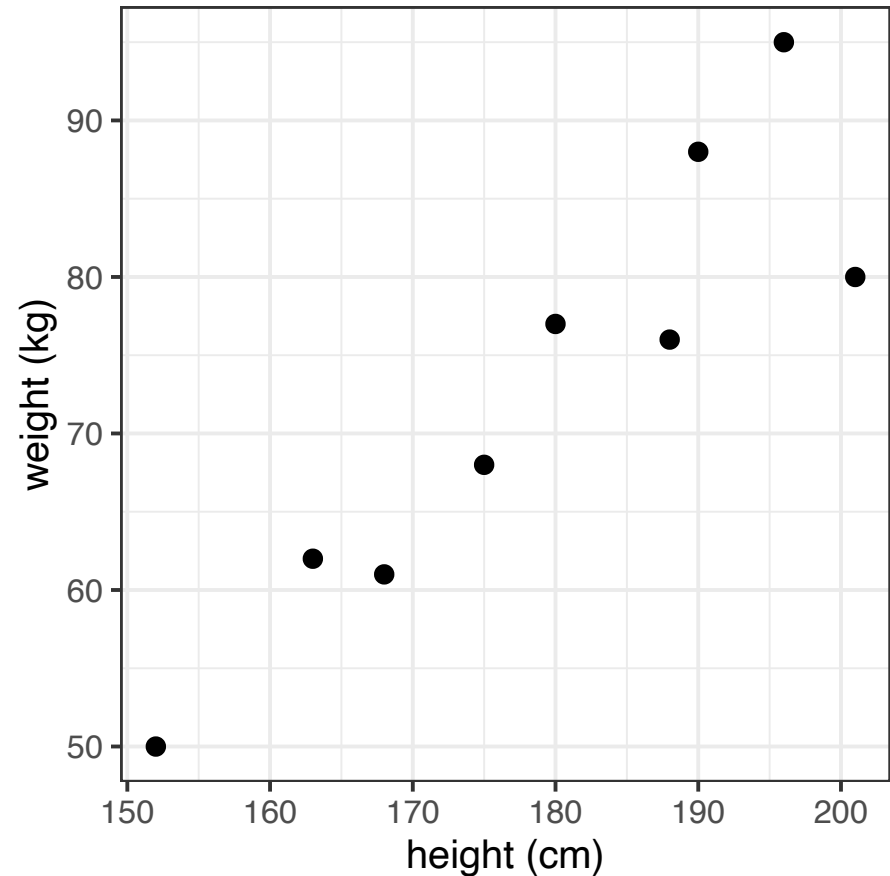


person	6	2	8	5	9	1	7	4	3
height (cm)	152	163	168	175	180	188	190	196	201
weight (kg)	50	62	61	68	77	76	88	95	80

It is easier to see in a scatter plot

The best way to explore correlations is with a scatter plot. So let's spend a few slides becoming familiar with a scatter plot.

The second thing to note is that creating a plot like this **reveals a relationship** between the two measures. In this case, you can see that a dot that is farther to the right on the x-axis (taller) will be higher on the y-axis (heavier). That is one way that two variables can covary - they both increase together.

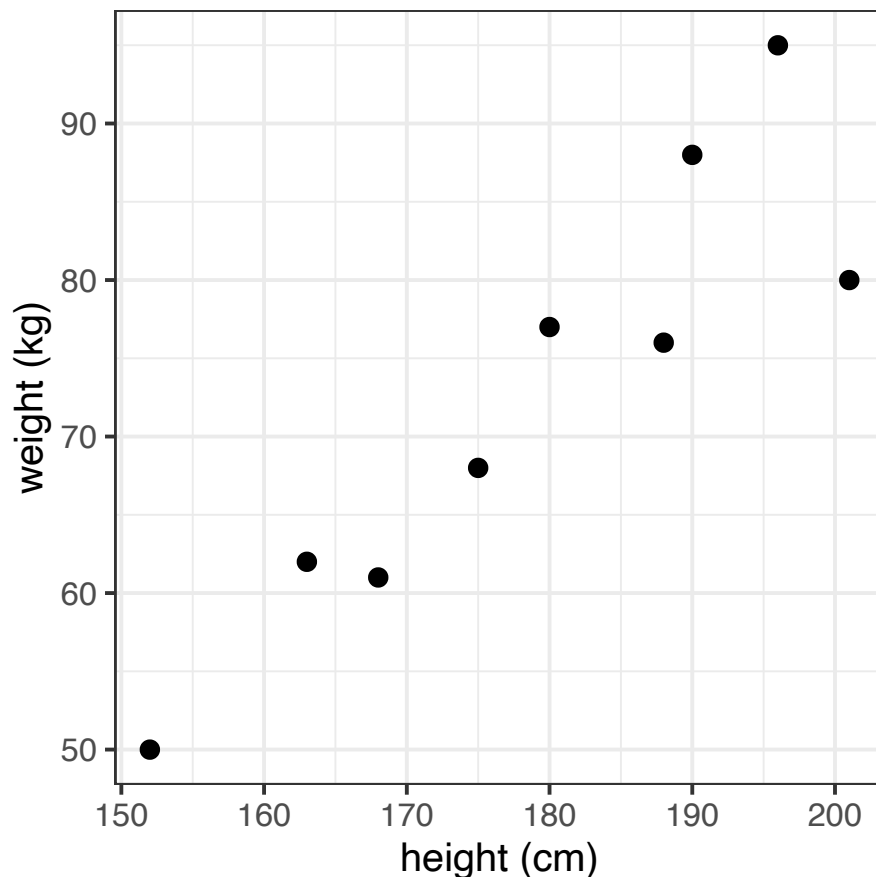


person	6	2	8	5	9	1	7	4	3
height (cm)	152	163	168	175	180	188	190	196	201
weight (kg)	50	62	61	68	77	76	88	95	80

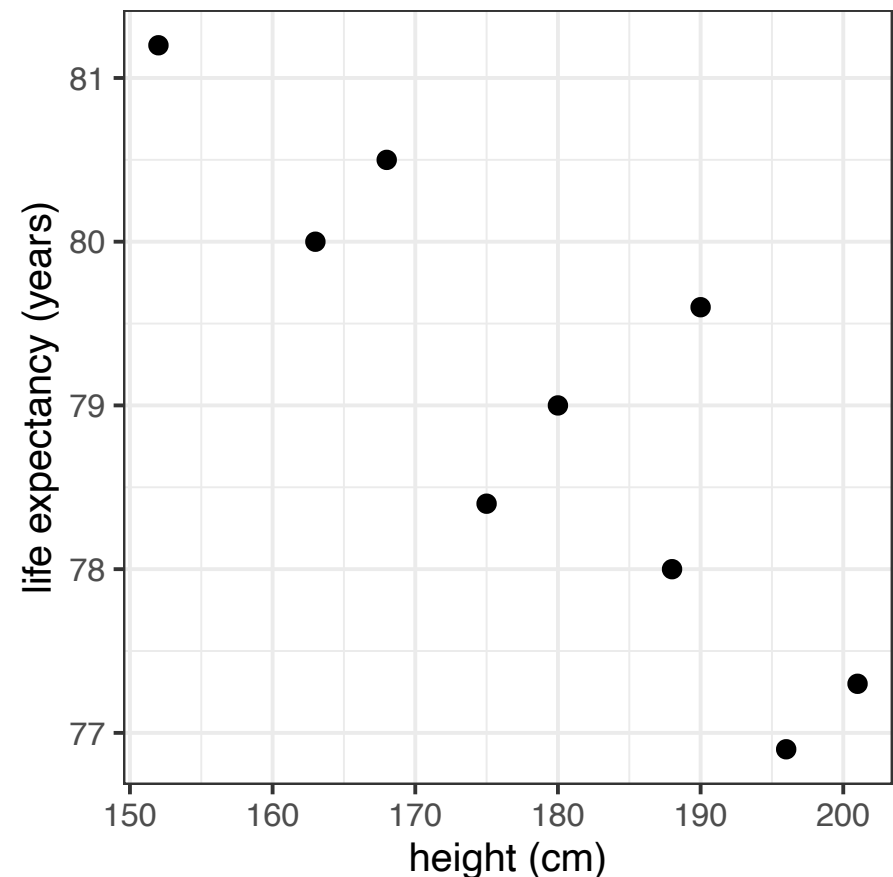
It is easier to see in a scatter plot

The third thing to note is that there are **two directions** that a correlation can go: the two variables can both **increase** together (**positive correlation**), or one can increase while the other decreases (**negative correlation**).

positive correlation



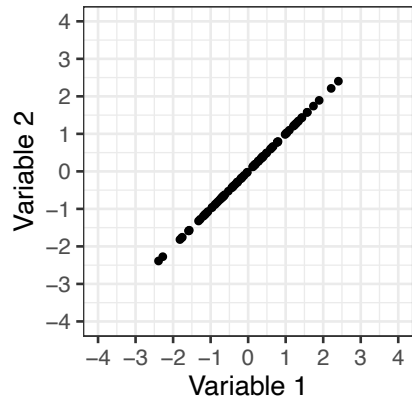
negative correlation



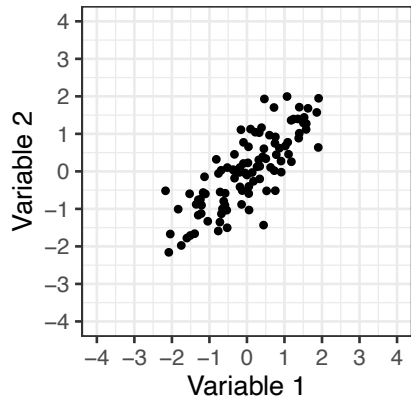
The strength of the relationship

As you can imagine, the strength of the relationship between the two variables can vary. Here are positive (top row) and negative (bottom row) correlations of different strengths:

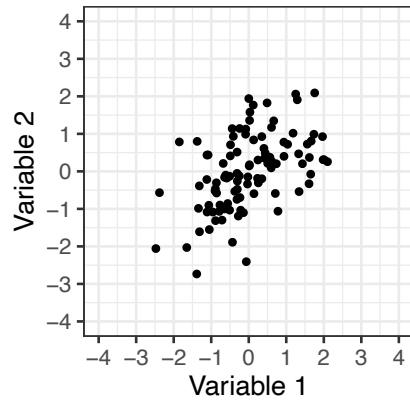
perfect



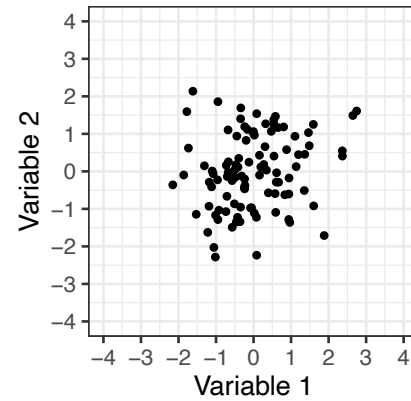
strong



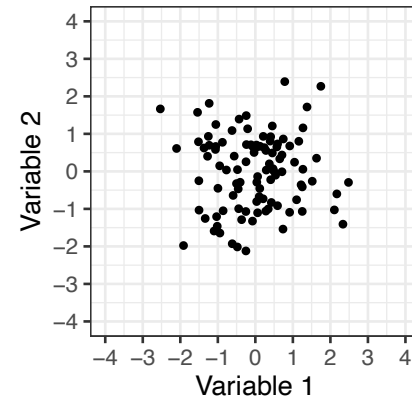
moderate



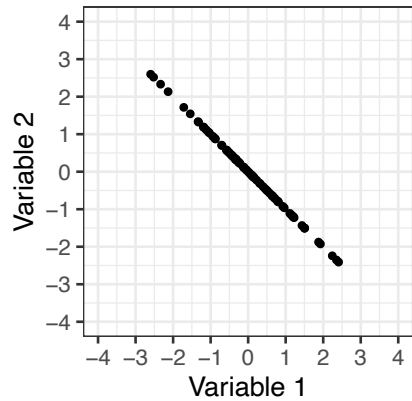
weak



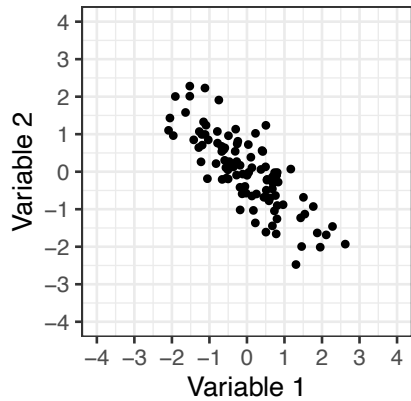
none



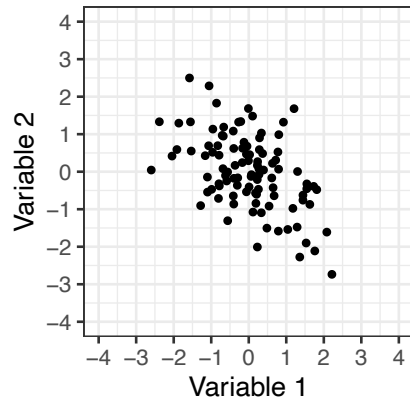
perfect



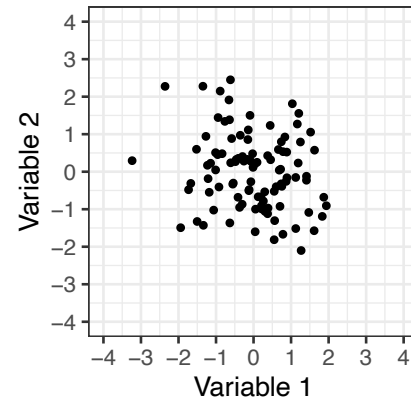
strong



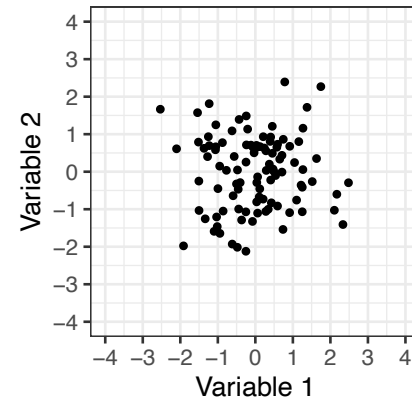
moderate



weak



none

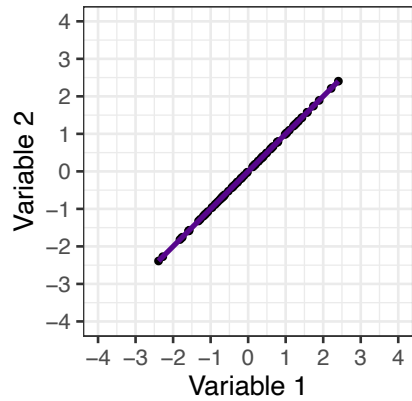


Why do we say it is linear?

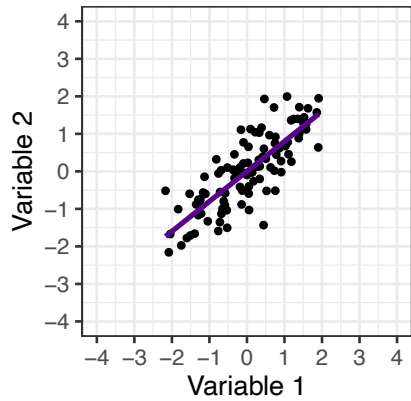
A linear relationship between two variables

The intuitive answer is that when you look at these correlation plots, you can imagine drawing a straight line through them to show the relationship. That is a practical consequence of a linear relationship.

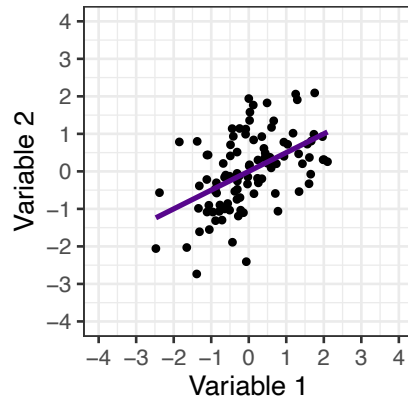
perfect



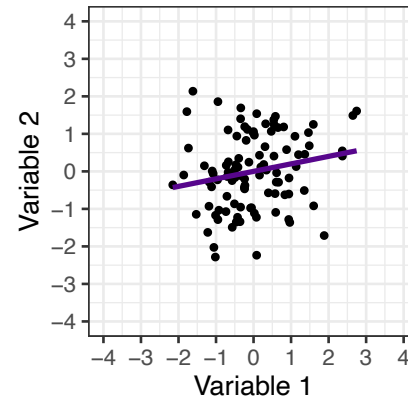
strong



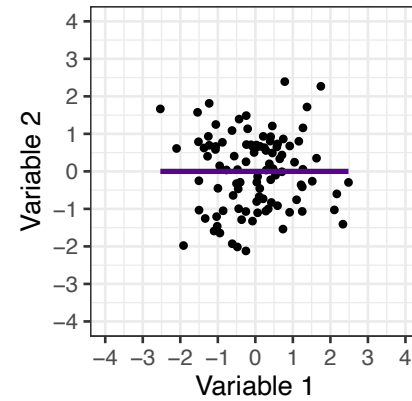
moderate



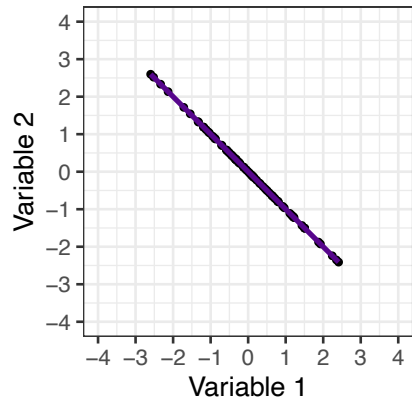
weak



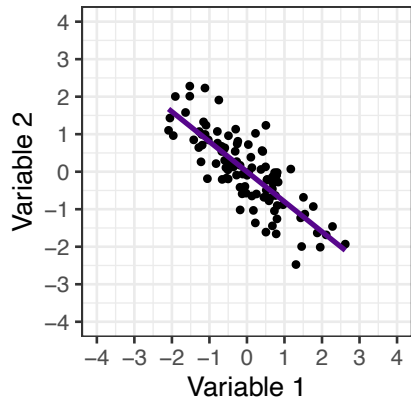
none



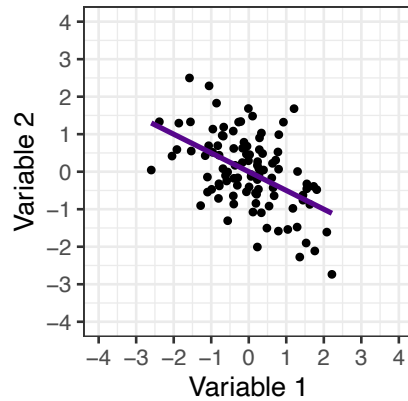
perfect



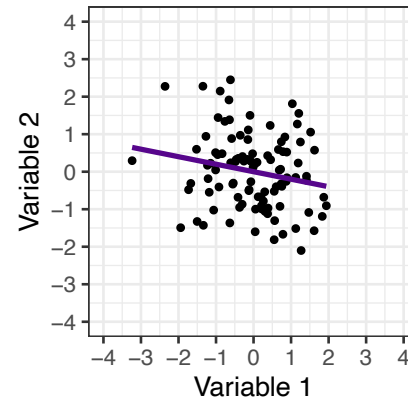
strong



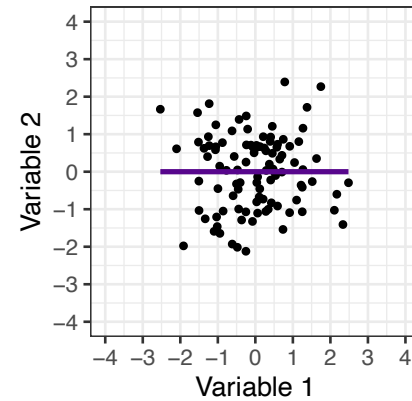
moderate



weak



none



A linear relationship between two variables

The more precise answer is that when two variables are linearly related, you can predict the value of one variable from the value of the other using an equation of the following form:

$$y = mx + b$$

You have seen this before. You know it as the **equation for a line**! It is just algebra all over again. But in this case, x is one of the variables you measured (like height) and y is the other variable you measured (like weight).

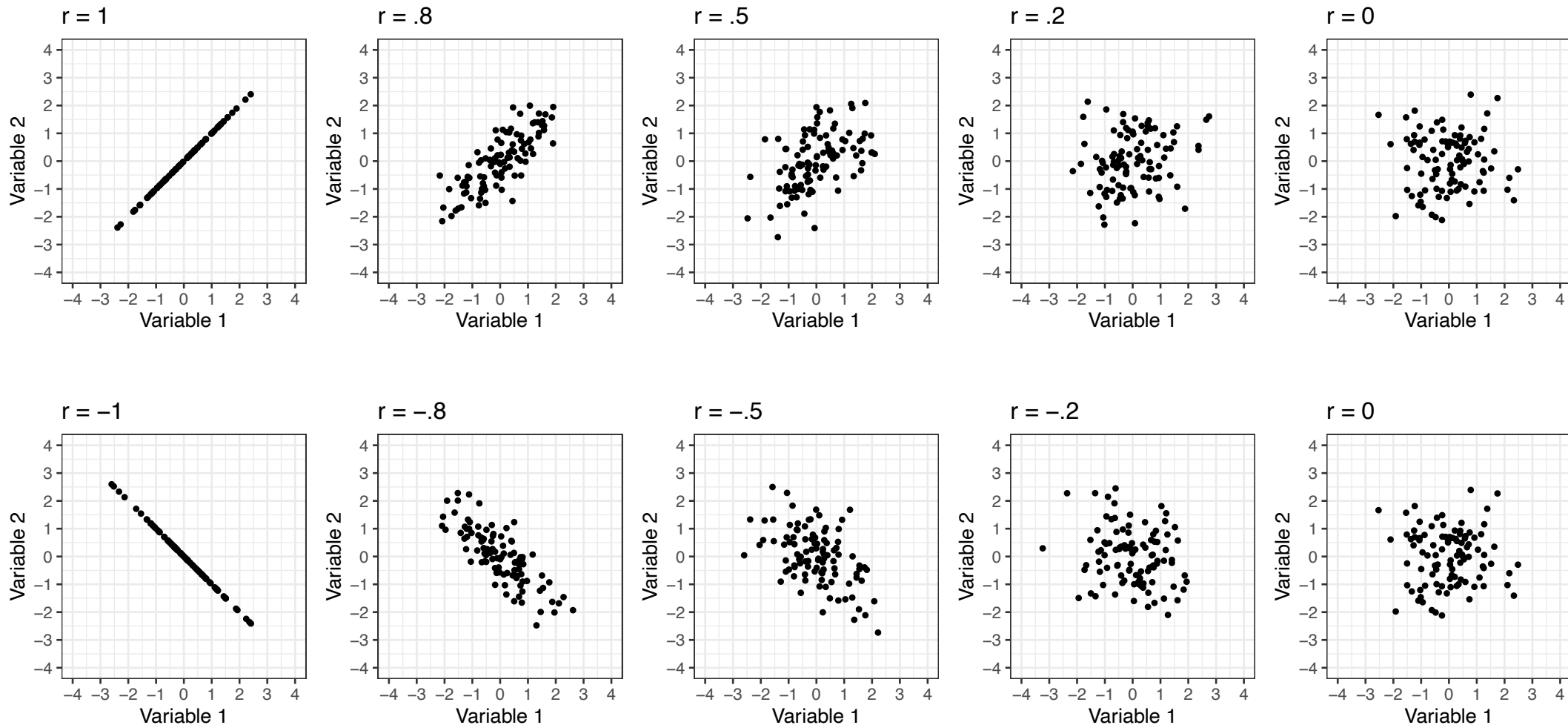
$$\text{weight} = m(\text{height}) + b$$

The idea is that if you found the right values for m (the slope) and b (the y -intercept), you can predict weight based on height.

This is linear regression, which we will cover in detail **next time**.

A linear relationship between two variables

Linear correlation does not go all the way into finding the line that relates the two variables. Instead, linear correlation is just the act of **measuring the strength of the linear relationship**. For that, we use a measure called **Pearson's r**:



Pearson's r

The intuition behind Pearson's r

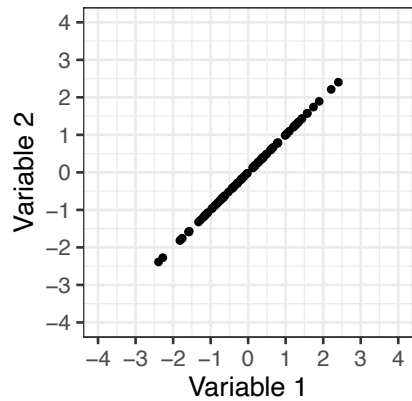
-1 means perfect negative correlation (one goes up, one goes down)

+1 means perfect positive correlation (both go up, or both go down)

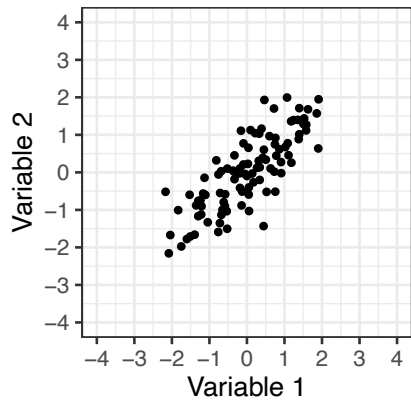
0 means no relationship whatsoever

The magnitude is **strength**, and the sign is **direction**!

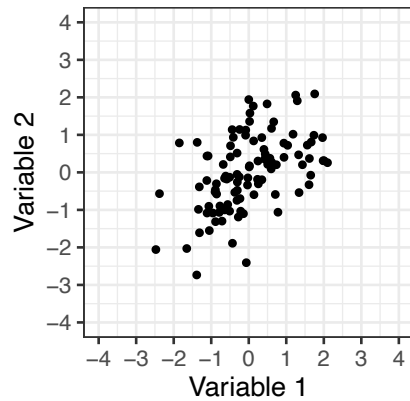
$r = 1$



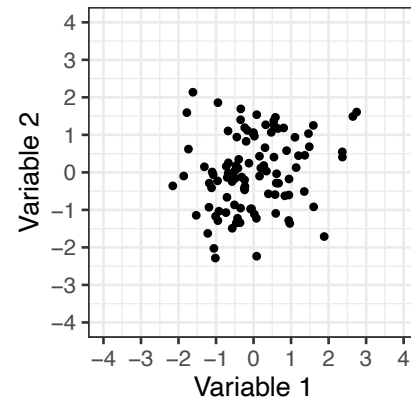
$r = .8$



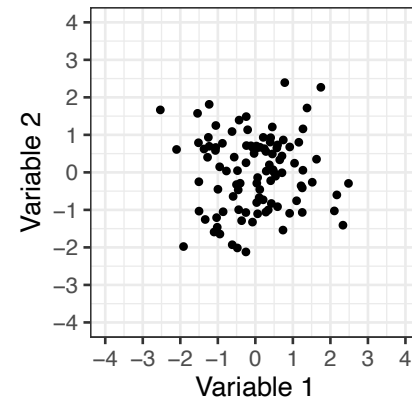
$r = .5$



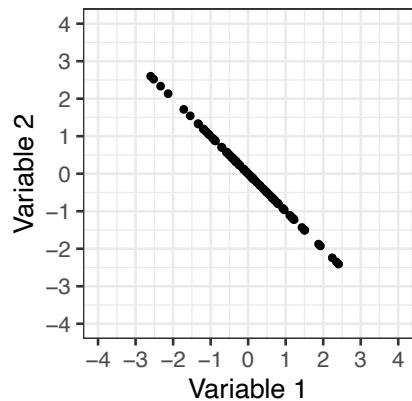
$r = .2$



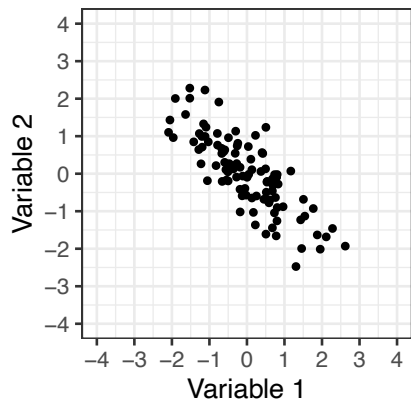
$r = 0$



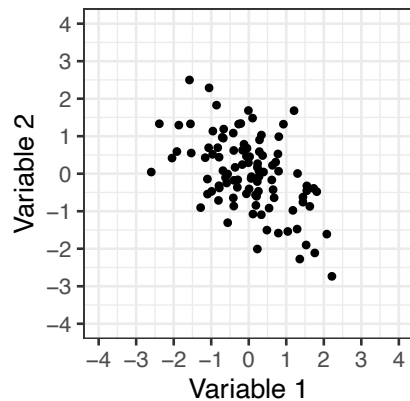
$r = -1$



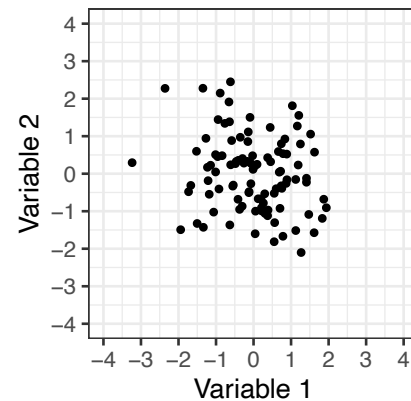
$r = -.8$



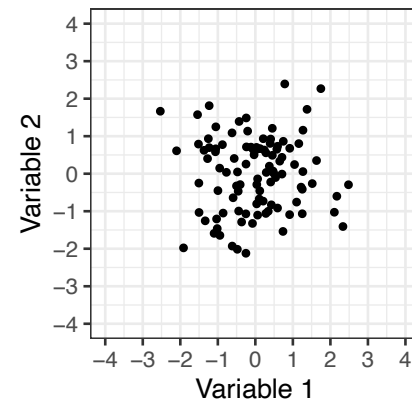
$r = -.5$



$r = -.2$



$r = 0$



The math: covariance

Remember, the thing we are measuring in a linear correlation is whether two variables covary. So we need a mathematical way to measure covariation.

We already have a measure for variation called **variance**. It looks like this.

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

And we can write out the square:

$$\frac{\sum (x_i - \bar{x}) (x_i - \bar{x})}{n-1}$$

So here is the mathematical idea: **covariance** between two variables will be the same, we will just replace one of the x terms with y.



$$\text{COV}_{xy} = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{n-1}$$

This is a small change, but a big conceptual shift. So let's spend some time getting a feel for it.



How does covariance behave?

Just like variance creates squares, covariance creates rectangles, and they have signs. Let's draw some. I'll use color for their signs.

Example 1: x and y are equal. This is a perfect positive correlation, and also identical to variance.



<u>variables</u>	<u>rectangles</u>	<u>sum</u>	<u>covariance</u>
$x = 1, 2, 3, 4$ $y = 1, 2, 3, 4$			1.67
$x - \bar{x} = -1.5, -.5, .5, 1.5$ $y - \bar{y} = -1.5, -.5, .5, 1.5$			

Example 2: y is always twice the size of x . This is a perfect positive correlation. It is not the same as variance anymore.

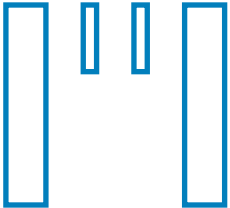

<u>variables</u>	<u>rectangles</u>	<u>sum</u>	<u>covariance</u>
$x = 1, 2, 3, 4$ $y = 2, 4, 6, 8$			3.33
$x - \bar{x} = -1.5, -.5, .5, 1.5$ $y - \bar{y} = -3, -1, 1, 3$			

How does covariance behave?

Example 3: y is always twice the size of x , but opposite in sign. This is a perfect negative correlation.

<u>variables</u>	<u>rectangles</u>	<u>sum</u>	<u>covariance</u>
$x = 1, 2, 3, 4$ $y = -2, -4, -6, -8$			-3.33
$x - \bar{x} = -1.5, -.5, .5, 1.5$ $y - \bar{y} = 3, 1, -1, -3$			


Example 4: y is always 5 times the size of x . This is a perfect positive correlation. Let's see what even larger values do.

<u>variables</u>	<u>rectangles</u>	<u>sum</u>	<u>covariance</u>
$x = 1, 2, 3, 4$ $y = 5, 10, 15, 20$			8.33
$x - \bar{x} = -1.5, -.5, .5, 1.5$ $y - \bar{y} = -7.5, -2.5, 2.5, 7.5$			

The relationship is the same, but the value gets bigger just because of the size of the y values!

How does covariance behave?

Example 5: Here are two perfectly uncorrelated variables. The numbers are decimals because I used a function to generate these for me.

<u>variables</u>	<u>rectangles</u>	<u>sum</u>	<u>covariance</u>
$x = 1.15, 2.19, 3.66, 1.00$ $y = 0.54, 1.77, 2.18, 3.51$		(nothing)	0
$x - \bar{x} = -0.85, 0.19, 1.66, -1.00$ $y - \bar{y} = -1.46, -0.23, 0.18, 1.51$			

What did we learn?

Covariance is positive for positive correlations (both variables increase together).

Covariance is negative for negative correlations (one variable increases and one decreases).

Covariance is zero for uncorrelated variables (no relationship).

But, covariance gets larger when the values of the variables get larger...

Can we just use covariance? No.

Covariance is the right measure for us - it is positive when it should be, negative when it should be, and negative when there is no covariation.

But, it grows based on the scale of the variables. Star temperatures will have higher covariance than earth temperatures, just because they are bigger, not because they have a stronger relationship.

Also, its units are squares, which is odd. In fact, its units are x-unit x y-unit. So, covariance of height and weight would be something like 8.3 cm x kg. That is really odd. I have no idea what that means.

We already know how to fix this! What we want to do is **standardize** this measure.

Pearson's r is standardized covariance

What (Karl) Pearson proposed is dividing covariance by the **maximum value** that covariance could possibly take for two variables.

Think this through. If you divide a value by the maximum it could be, you get a proportion between 0 and 1! That is nicely standardized!

In this case, because covariance can be either positive or negative, we will get a sign too: -1 to 0 to +1.

So what is the maximum that covariance could ever be? It is $s_x \cdot s_y$.

So Pearson's r is:

But you will often see it as this, because people don't like two denominators:

$$r = \frac{\frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{n-1}}{s_x s_y}$$

$$r = \frac{\frac{1}{n-1} \sum (x_i - \bar{x}) (y_i - \bar{y})}{s_x s_y}$$

Why is $s_x s_y$ the maximum?

How do we know this? Well, as always, the answer is calculus. You can find the maximum (or minimum) of a function with calculus. So you can find that the maximum of the covariance equation is $s_x s_y$.

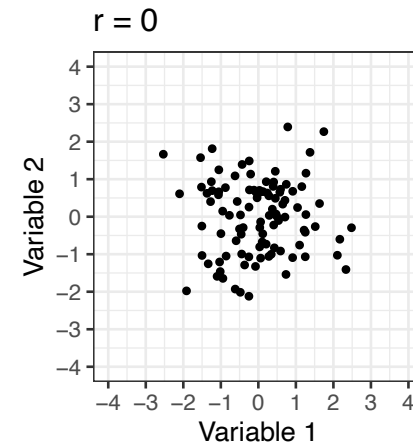
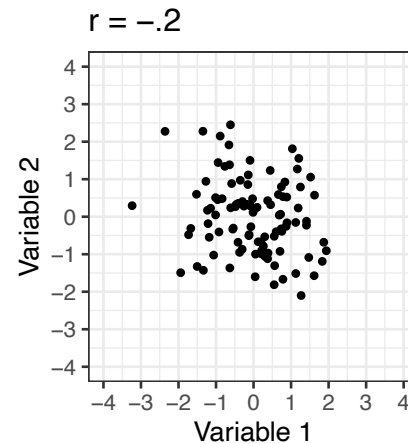
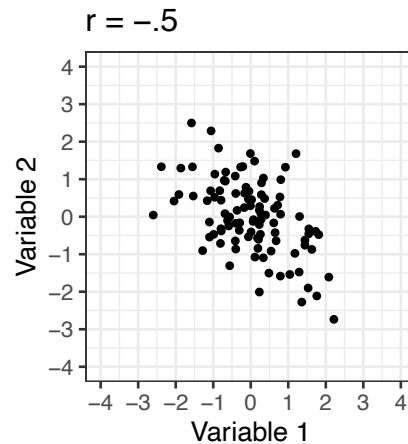
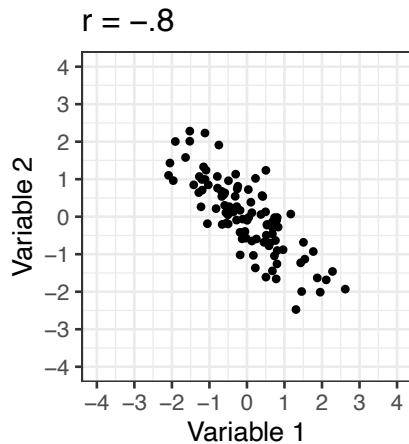
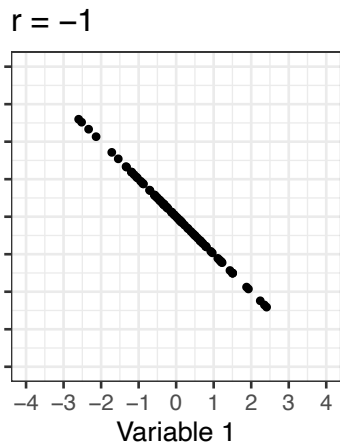
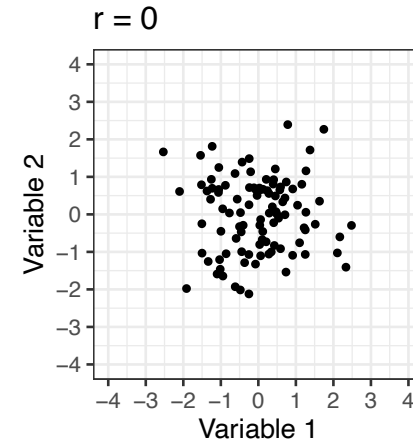
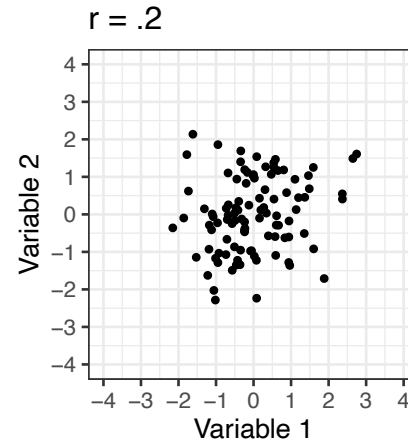
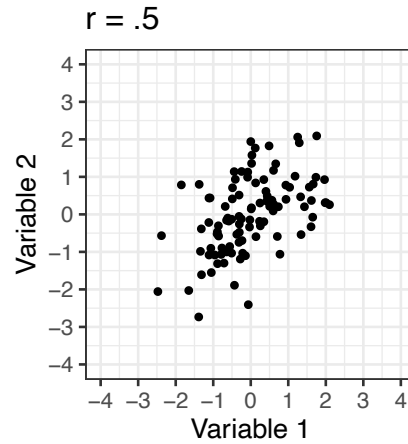
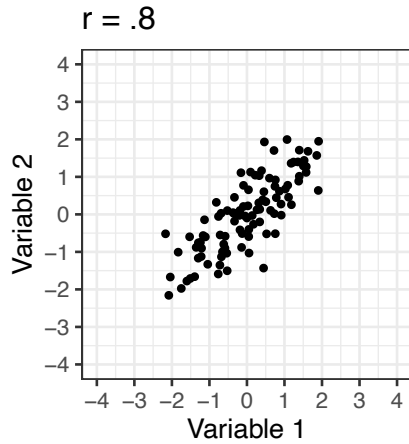
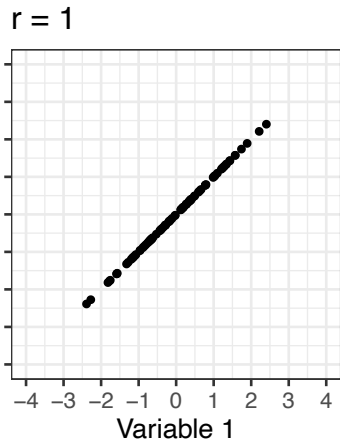
But we can prove it to ourselves with some induction. We know perfect correlation will have the highest covariance. So let's try each of our previous examples of perfect correlation, and compare $s_x s_y$ to cov_{xy} :

1.	$x = 1, 2, 3, 4$ $y = 1, 2, 3, 4$	$\text{sd}(x) = 1.29$ $\text{sd}(y) = 1.29$	$s_x s_y = 1.67$	$\text{cov}_{xy} = 1.67$
2.	$x = 1, 2, 3, 4$ $y = 2, 4, 6, 8$	$\text{sd}(x) = 1.29$ $\text{sd}(y) = 2.58$	$s_x s_y = 3.33$	$\text{cov}_{xy} = 3.33$
3.	$x = 1, 2, 3, 4$ $y = -2, -4, -6, -8$	$\text{sd}(x) = 1.29$ $\text{sd}(y) = 2.58$	$s_x s_y = 3.33$	$\text{cov}_{xy} = -3.33$
4.	$x = 1, 2, 3, 4$ $y = 5, 10, 15, 20$	$\text{sd}(x) = 1.29$ $\text{sd}(y) = 6.45$	$s_x s_y = 8.33$	$\text{cov}_{xy} = 8.33$

Interpreting Pearson's r

Pearson's r is descriptive

Pearson's r is not a test statistic. It is a descriptive statistic. It describes your data. So, interpreting it is up to you. The same way that interpreting the mean is up to you. It is actually probably more like an estimated Cohen's d , because it is a standardized measure.



If you want to test a hypothesis, you need to create a hypothesis test

Like any descriptive statistic (like the mean), if you want to use r to test a hypothesis, you need to create a hypothesis test for it. It turns out that we can treat r very similar to a mean, and create a t -test for it:

one sample t -test

$$t = \frac{\bar{x} - \mu_0}{S_{\bar{x}}}$$

t -test for Pearson's r

$$t = \frac{r - \rho_0}{\sqrt{\frac{1-r^2}{n-2}}}$$

Just like a one-sample t -test tests the null hypothesis that the sample comes from a population with a known mean (typically 0), the t -test for Pearson's r tests the null hypothesis that the sample comes from a population with a known correlation (typically 0).

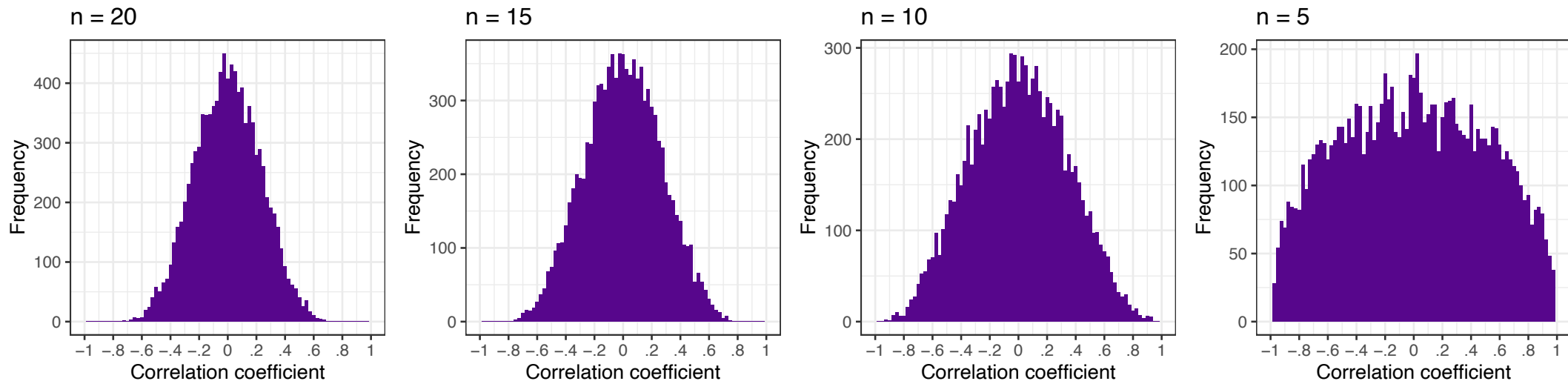
The only difference is in the measure (r and ρ_0) and in the estimated standard error, which for r is the term in orange.

(As always, the reason why is that this equation best approximates the standard deviation of the sampling distribution of r .)

A trap - small sample sizes

It is tempting to interpret all Pearson's r s the same. But you really need to be careful with small sample sizes. Small sample sizes can yield large r s. You really should only use correlation when you have fairly large sample sizes.

Here are simulations of populations where $\rho=0$. For each simulation, I drew a sample (pairs of values) of size n 10,000 times, calculated r each time, and plotted the distribution of r . This is like a null distribution of r for different sample sizes:



As you can see, even with $n=20$, it is possible to get r nearly up to .4 before reaching the .05 part of the tail. For $n=5$, you can almost get an r of 1.

A trap - correlation and causation

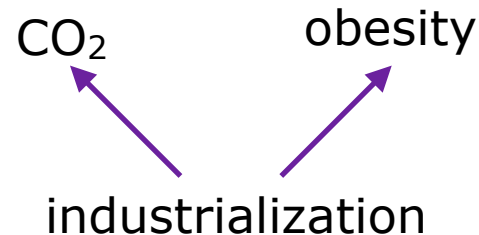
Correlation does not imply causation. Correlation only tells us that there may be a relationship between two variables. It does not tell us what causes it (including “an accident”). Here are the 4 causes of a correlation (from our second class):

1. Variable X **causes** variable Y.
2. Variable Y **causes** variable X.
3. A third variable, Z, **causes** both variable X and Y.
4. Variable X and variable Y correlate **accidentally**.

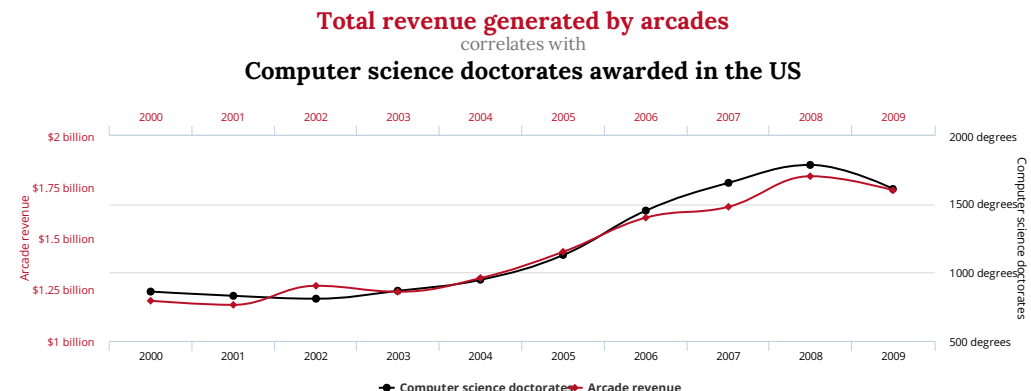
<https://www.tylervigen.com/spurious-correlations>

Examples

High national debt tends to correlate with slower economic growth. Economists argue about which causes which!



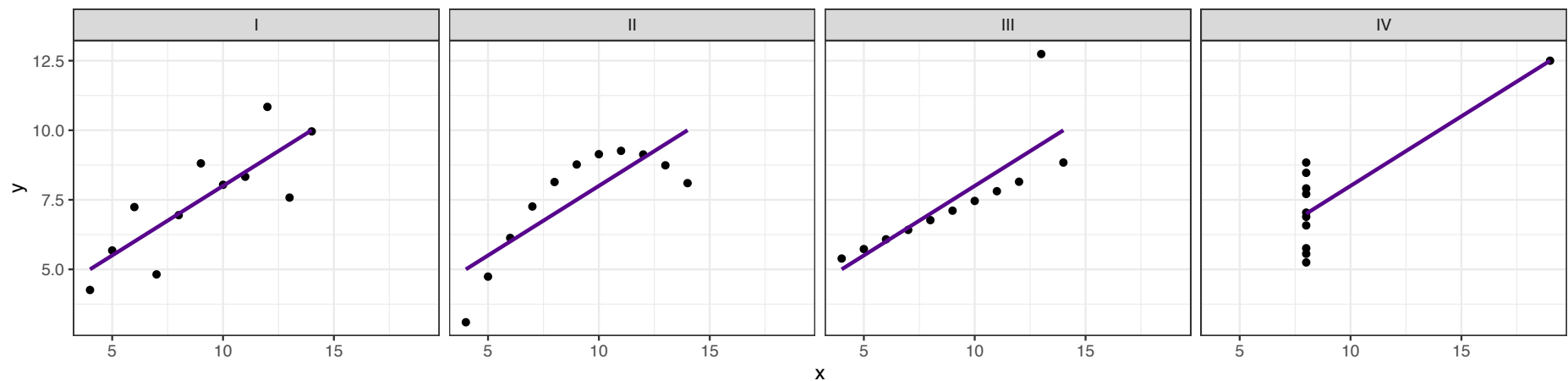
CO₂ and obesity correlate, but neither causes the other!



A trap - always plot your data

There is a famous set of 4 data sets called Anscombe's quartet (made by Francis Anscombe in 1973). These four data sets have nearly identical statistics of various kinds:

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
mean	9	7.5	9	7.5	9	7.5	9	7.5
var	11	4.125	11	4.125	11	4.125	11	4.125
r	0.816		0.816		0.816		0.816	



The lesson here is that descriptive statistics, including r , all miss crucial information. You **must plot your data** to see what its properties truly are.