UNIVERSITY OF WARWICK

Lecture 9

Variance-Bias

Gabriele Pergola and Amina Asif

Machine Learning Algorithms and Practice
CS429

October 2026

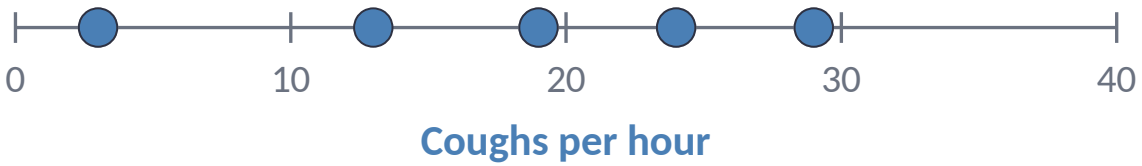
Credits: Some slides adapted from Fayyaz M

Mean, Variance and Standard Deviation

Describing a measurement, and how spread out it is

Our example: coughs in flu patients

We count how many times each of **5** flu patients coughs during one hour.

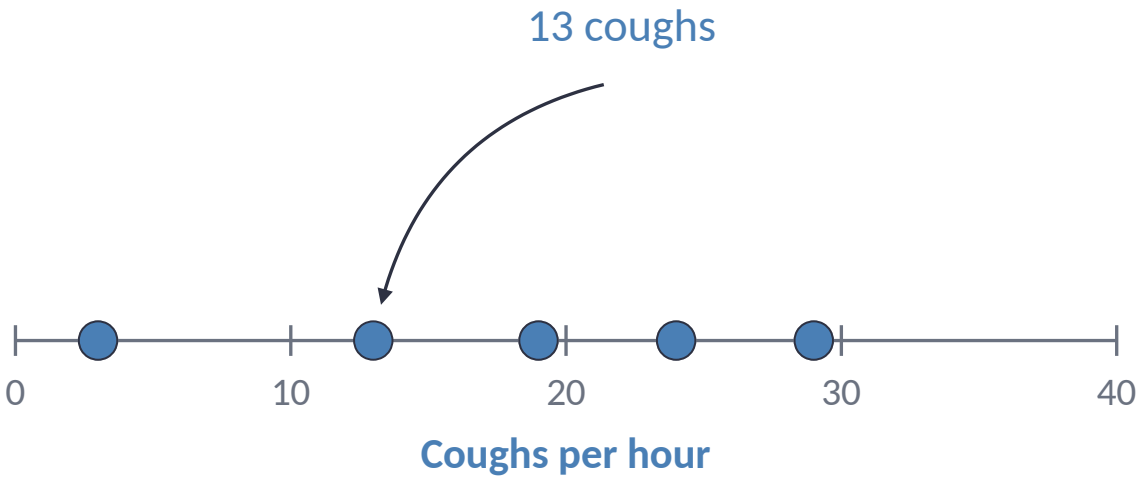


Our example: coughs in flu patients

We count how many times each of **5** flu patients coughs during one hour.

Each dot is one patient. For example, this dot represents a patient who coughed **13** times in that hour.

We will keep using these patients throughout: first for the **mean**, **variance** and **standard deviation**, then for **covariance** and, later, **PCA**.



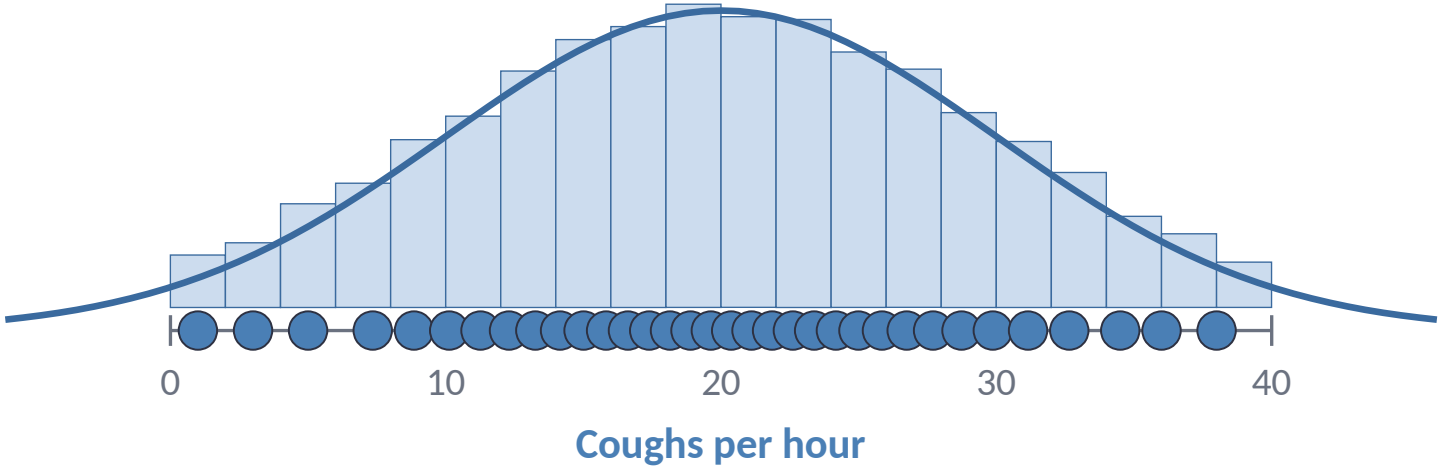
Imagine measuring every patient

Now imagine that we could count the coughs of **every** person who caught flu this winter: say, **2 million** patients.

Now we can draw a **histogram** of the measurements.

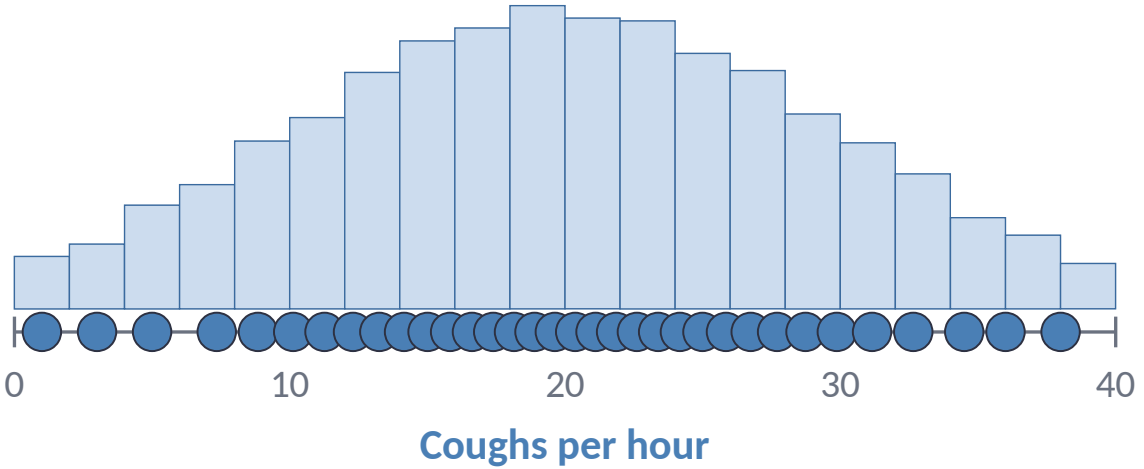
If we wanted to fit a **normal curve** to this histogram…

…then we would need to calculate the **population mean** and the **population variance**, or **population standard deviation**.



Calculating the population mean

Calculating the **population mean** is easy: we take the average of all **2 million** measurements.

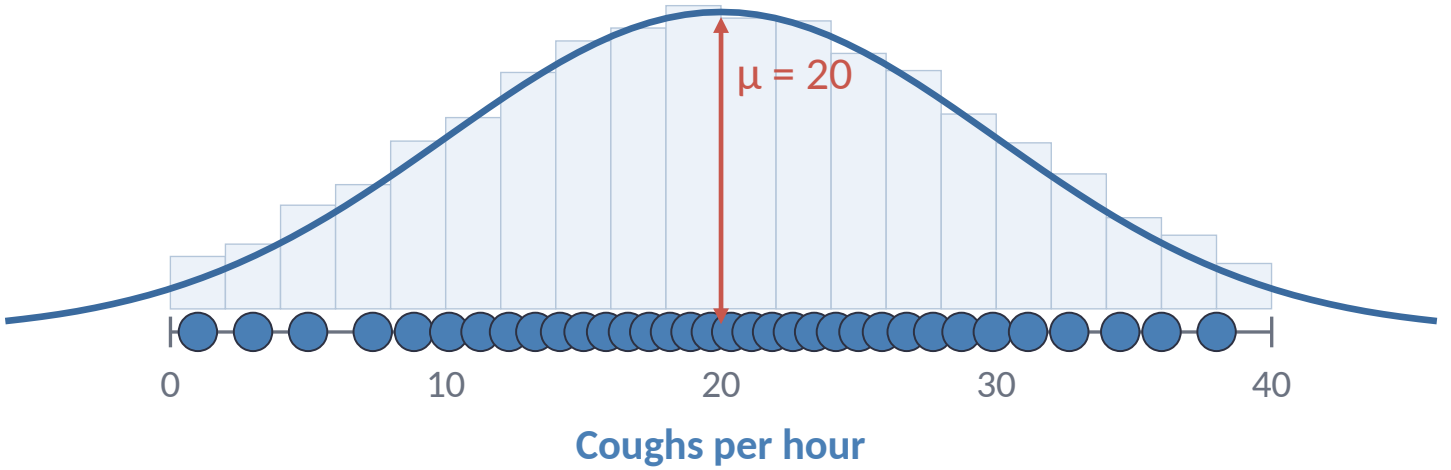


$$\text{Population Mean} = \frac{1 + 3 + 5 + \ldots + 36 + 38}{2{,}000{,}000}$$

Calculating the population mean

Calculating the **population mean** is easy: we take the average of all **2 million** measurements.

The population mean is **20** coughs per hour, and we centre the normal curve on it.

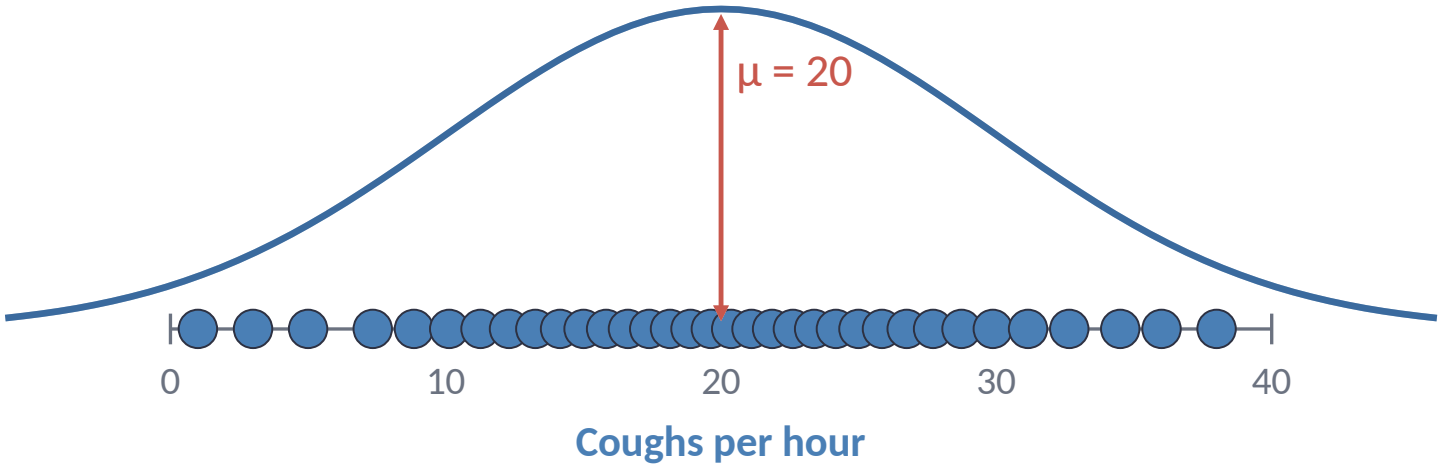


Note: because we used all 2 million measurements in the population, this is not an *estimate* of the population mean. We have *calculated* it.

$$\text{Population Mean} = \frac{1 + 3 + 5 + \ldots + 36 + 38}{2{,}000{,}000} = \boxed{20}$$

Estimating the mean from a sample

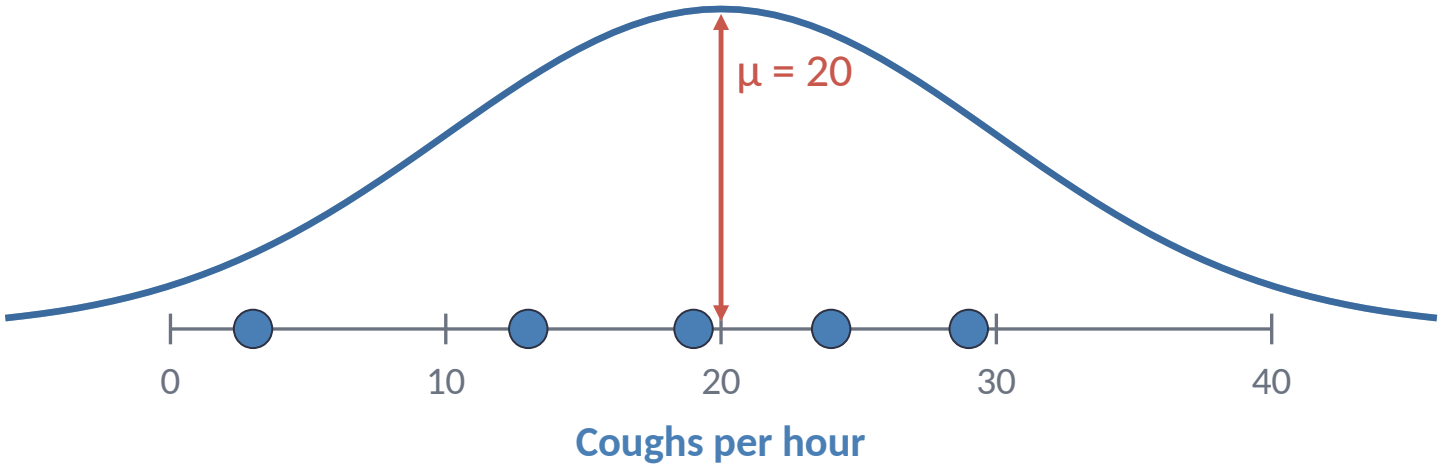
However, we rarely, if ever, have enough time and money to measure every single patient.



Estimating the mean from a sample

However, we rarely, if ever, have enough time and money to measure every single patient.

Instead, we almost always *estimate* the **population mean** from a relatively small sample: here, only **5** of the 2 million patients.



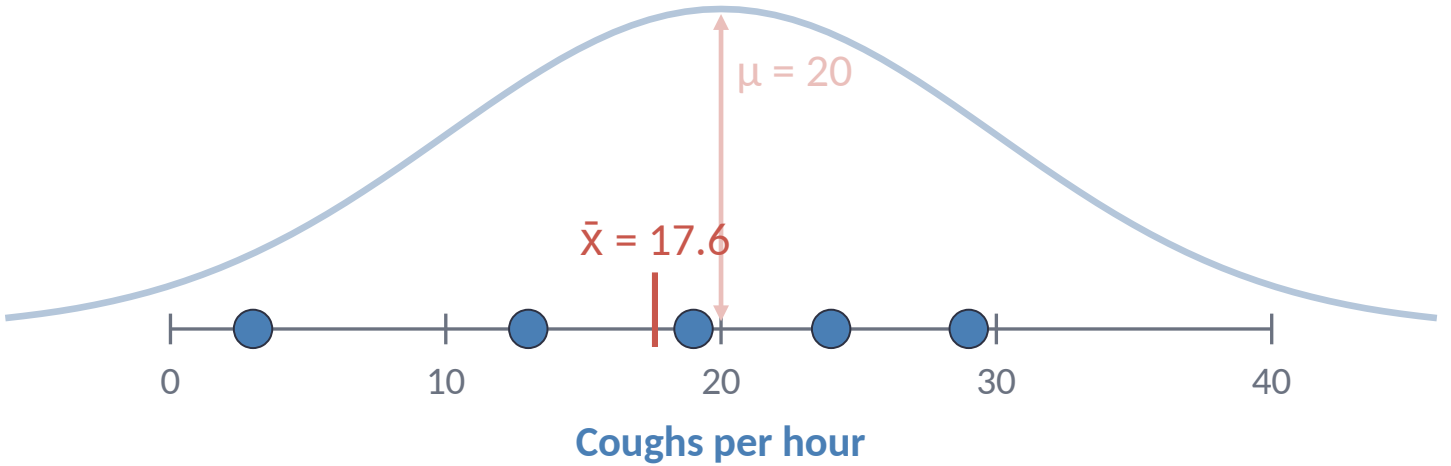
Estimating the mean from a sample

However, we rarely, if ever, have enough time and money to measure every single patient.

Instead, we almost always *estimate* the **population mean** from a relatively small sample: here, only **5** of the 2 million patients.

Estimating the population mean is simple: we calculate the average of the measurements we collected. Here, the **estimated mean** is **17.6**.

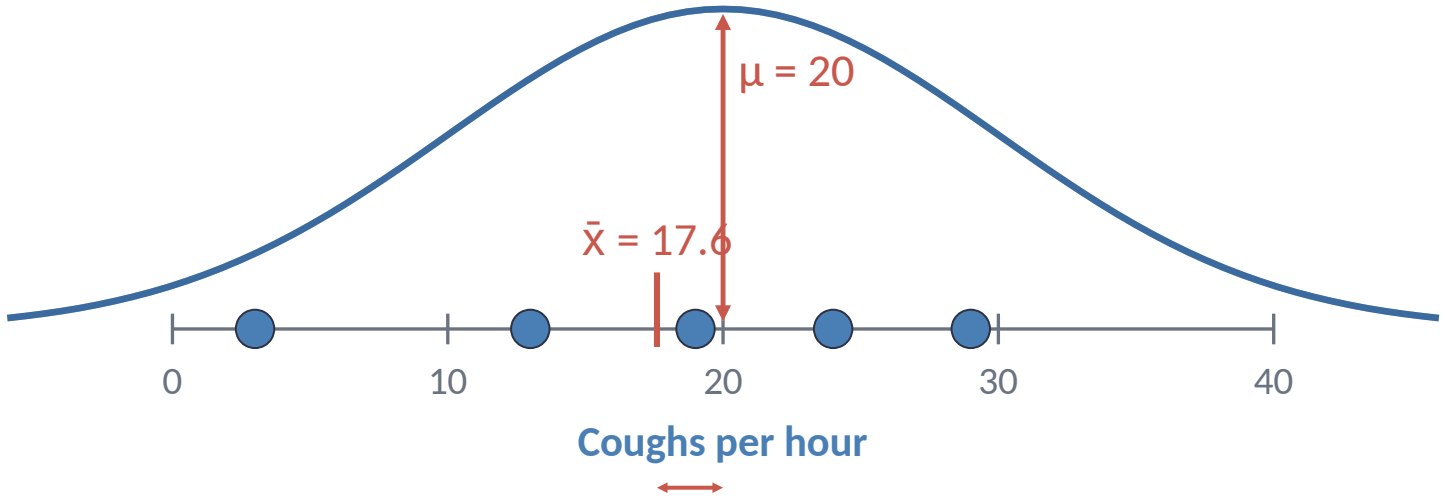
Statisticians often use the symbol $\bar{x}$ ("x-bar") for the **estimated mean**, which is also called the **sample mean**.



$$\text{Estimated Mean} = \frac{3 + 13 + 19 + 24 + 29}{5} = \boxed{17.6}$$

Estimating the mean from a sample

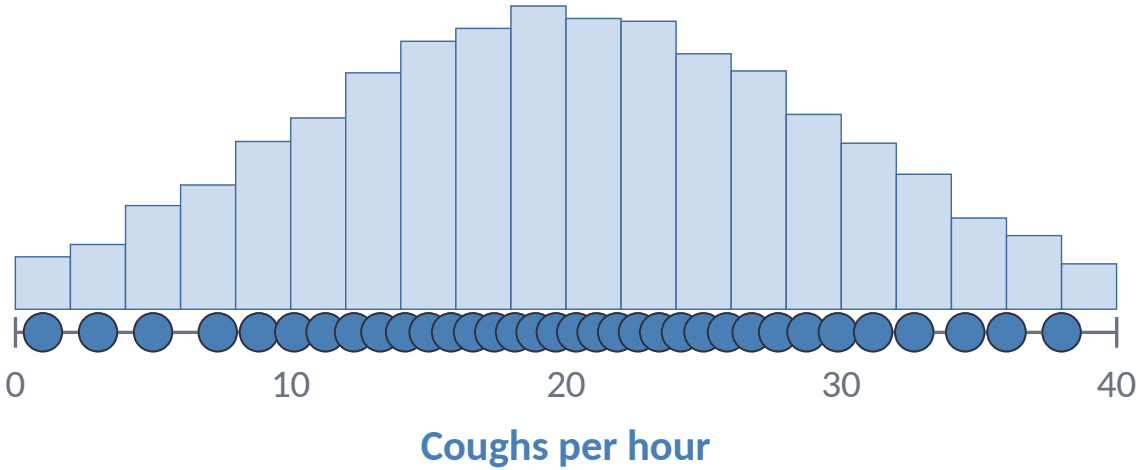
The estimated mean, $\bar{x}$, is different from the population mean, μ. With more and more data, $\bar{x}$ should get closer and closer to μ.



$$\text{Estimated Mean} = \frac{3 + 13 + 19 + 24 + 29}{5} = \boxed{17.6}$$

How spread out are the measurements?

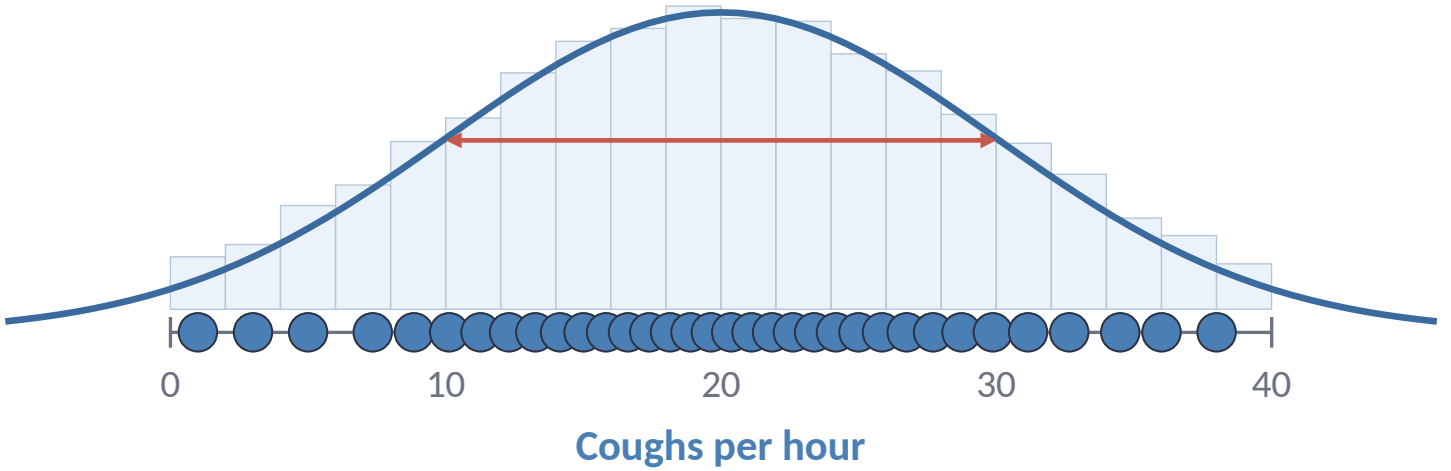
Going back to the full set of population data…



How spread out are the measurements?

Going back to the full set of population data…

…we now want to determine how wide to make the curve, by *calculating*, not estimating, the population **variance** and **standard deviation**.

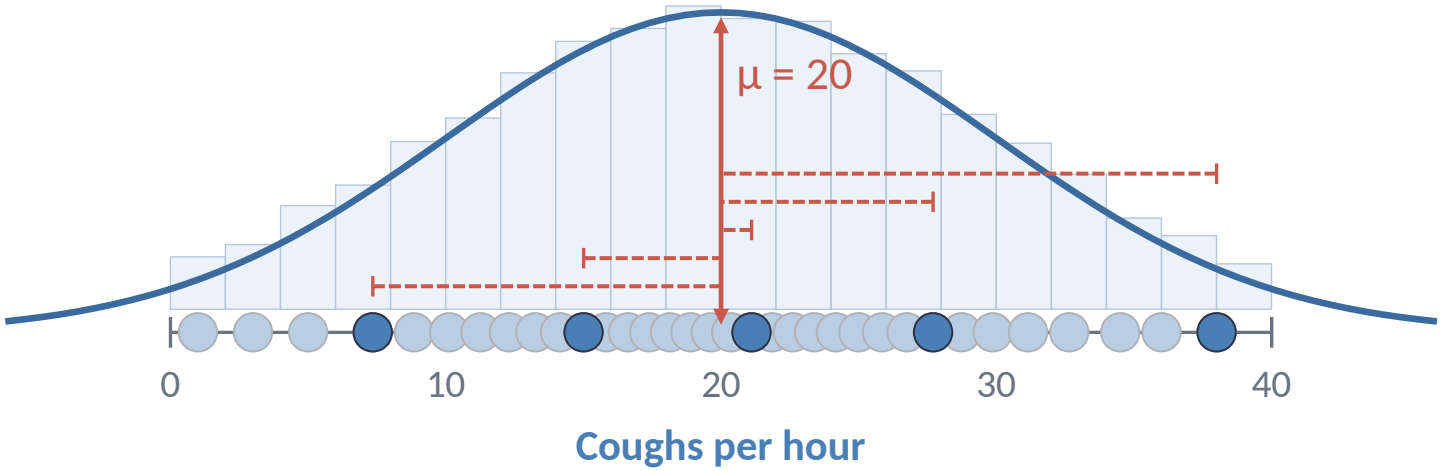


How spread out are the measurements?

Going back to the full set of population data…

…we now want to determine how wide to make the curve, by *calculating*, not estimating, the population **variance** and **standard deviation**.

In other words, we want to calculate how spread out the data are around the **population mean**.

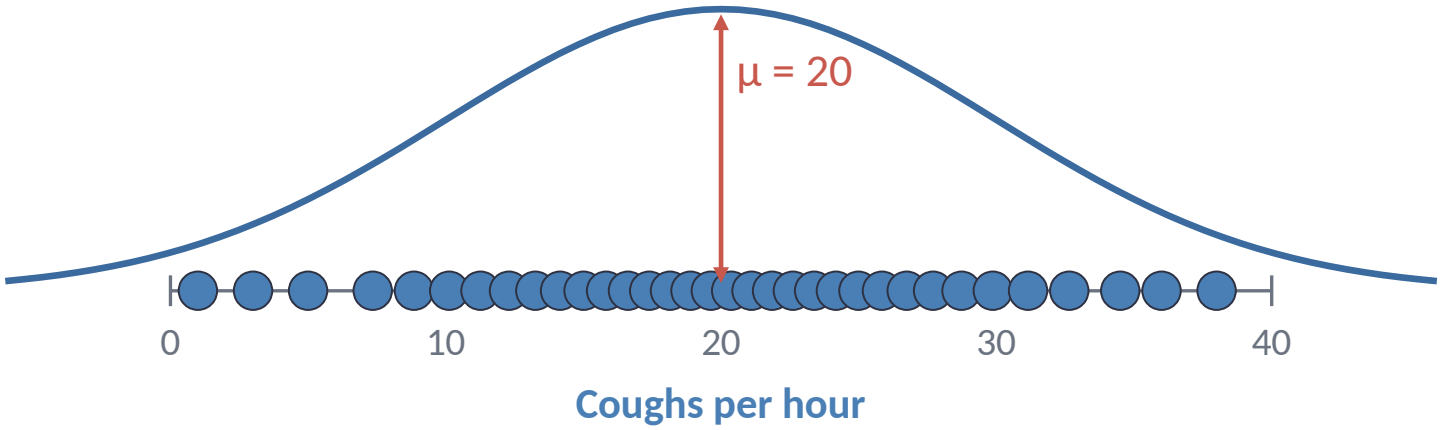


The population variance

This is the formula we use to *calculate*, not estimate, the **population variance**.

Note: the difference between *calculating* and *estimating* the variance matters, as we will see shortly.

The part **x − μ** means that we subtract the population mean, **μ** = 20, from each measurement, **x**.



$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

(1 − 20) (3 − 20) … (36 − 20) (38 − 20)

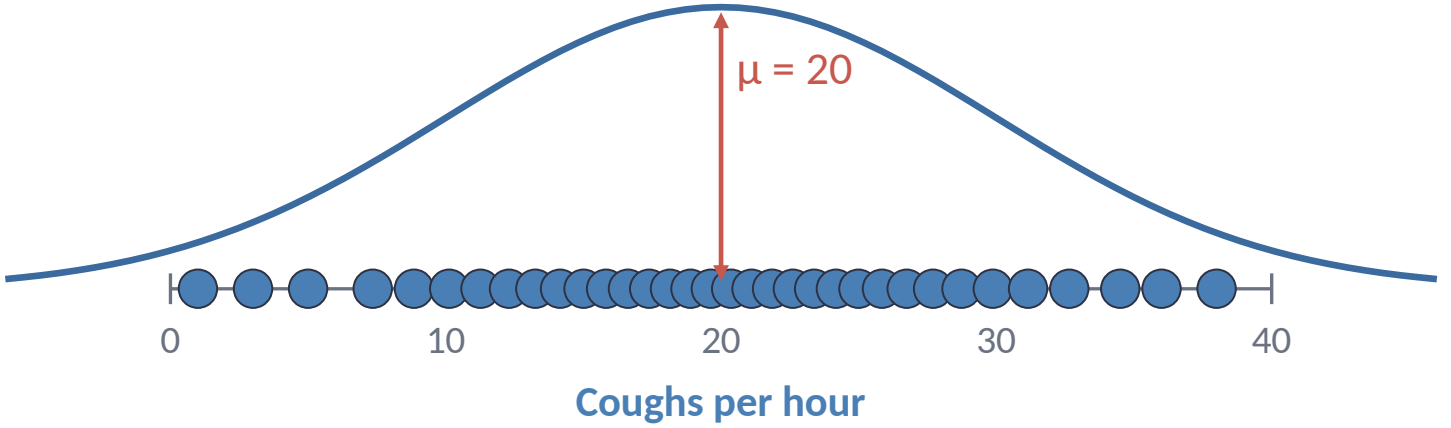
The population variance

This is the formula we use to *calculate*, not estimate, the **population variance**.

Note: the difference between *calculating* and *estimating* the variance matters, as we will see shortly.

The part **x – μ** means that we subtract the population mean, **μ** = 20, from each measurement, **x**.

The square tells us to square each term…



$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$(1 - 20)^2 \quad (3 - 20)^2 \quad \ldots \quad (36 - 20)^2 \quad (38 - 20)^2$

The population variance

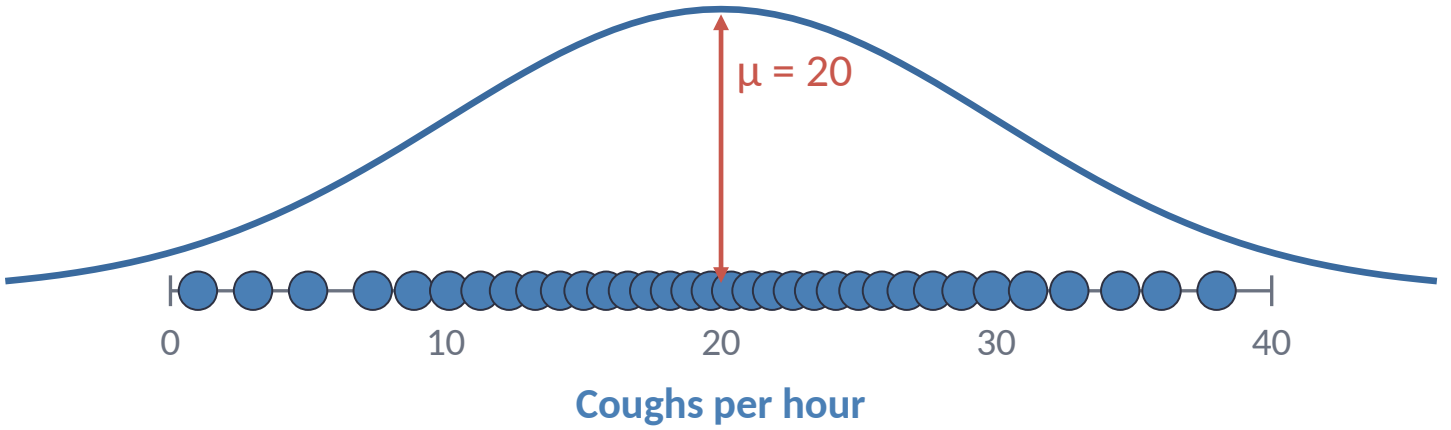
This is the formula we use to *calculate*, not estimate, the **population variance**.

Note: the difference between *calculating* and *estimating* the variance matters, as we will see shortly.

The part **x – μ** means that we subtract the population mean, **μ** = 20, from each measurement, **x**.

The square tells us to square each term…

…and the Greek character **sigma**, **Σ**, tells us to add up all the terms.



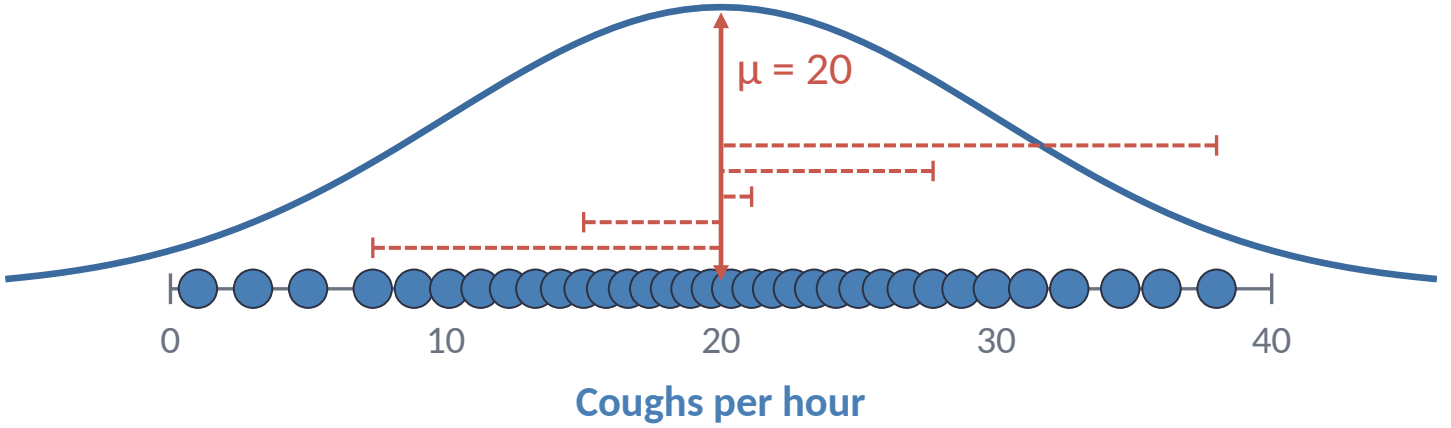
$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$$= (1 - 20)^2 + (3 - 20)^2 + \ldots + (36 - 20)^2 + (38 - 20)^2$$

The population variance

Lastly, we want the **average** of the squared differences, so we divide by the number of measurements, **n**: in this case, 2 million.

Thus, we are just calculating the **average** of the squared differences between the data and the **population mean**.



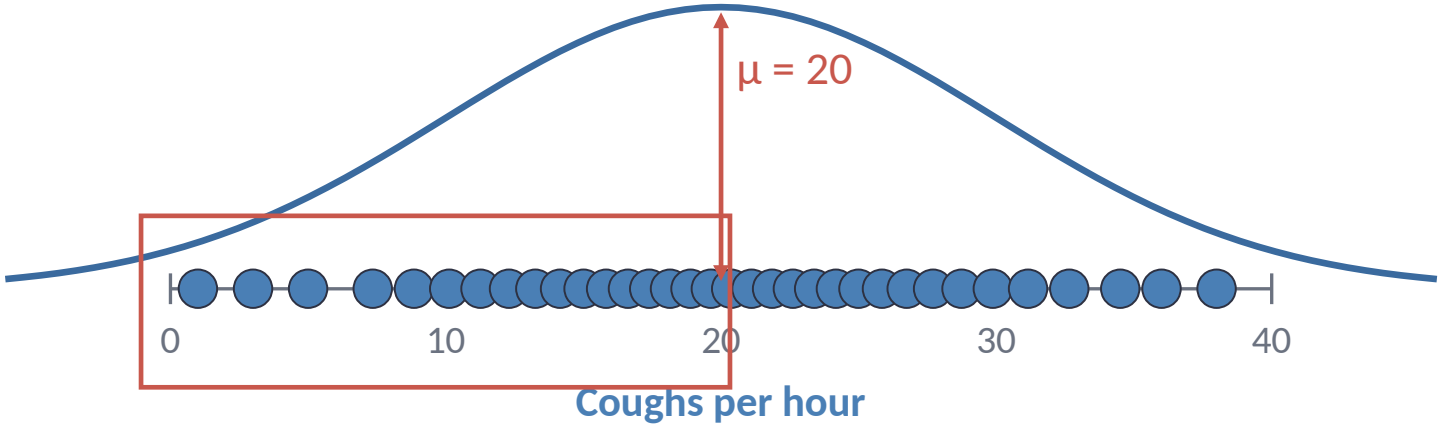
$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$$= \frac{(1 - 20)^2 + (3 - 20)^2 + \ldots + (36 - 20)^2 + (38 - 20)^2}{2{,}000{,}000}$$

Why square the differences?

Note: squaring each term ensures that each difference is positive.

Otherwise, the measurements on the left of the mean would have negative differences…



$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

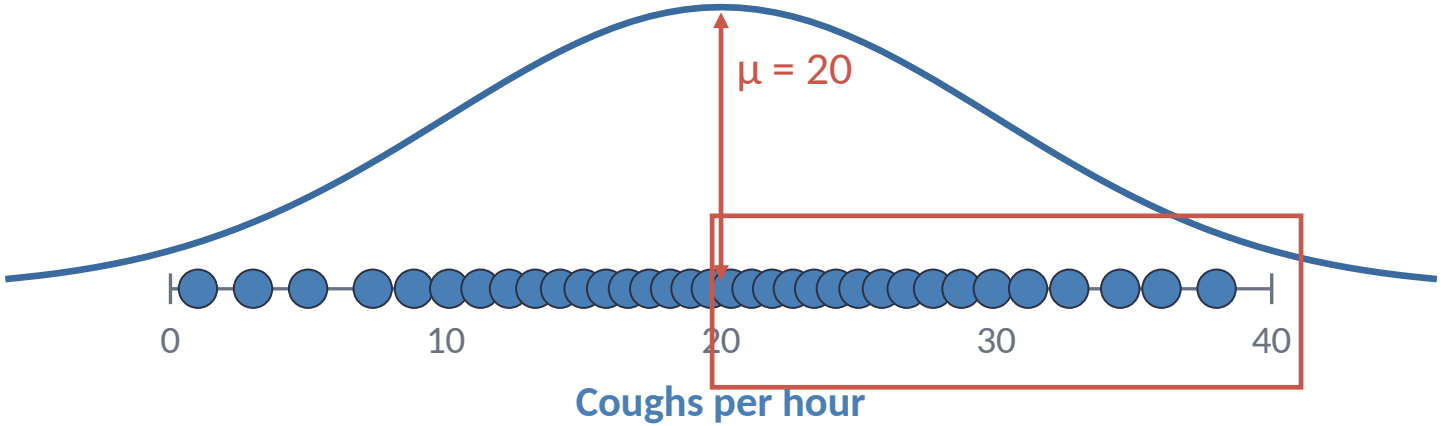
$$= \frac{(1 - 20)^2 + (3 - 20)^2 + \ldots + (36 - 20)^2 + (38 - 20)^2}{2{,}000{,}000}$$

Why square the differences?

Note: squaring each term ensures that each difference is positive.

Otherwise, the measurements on the left of the mean would have negative differences…

…which would cancel out the positive differences from the measurements on the right of the mean.



$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$$= \frac{(1 - 20)^2 + (3 - 20)^2 + \ldots + (36 - 20)^2 + (38 - 20)^2}{2{,}000{,}000}$$

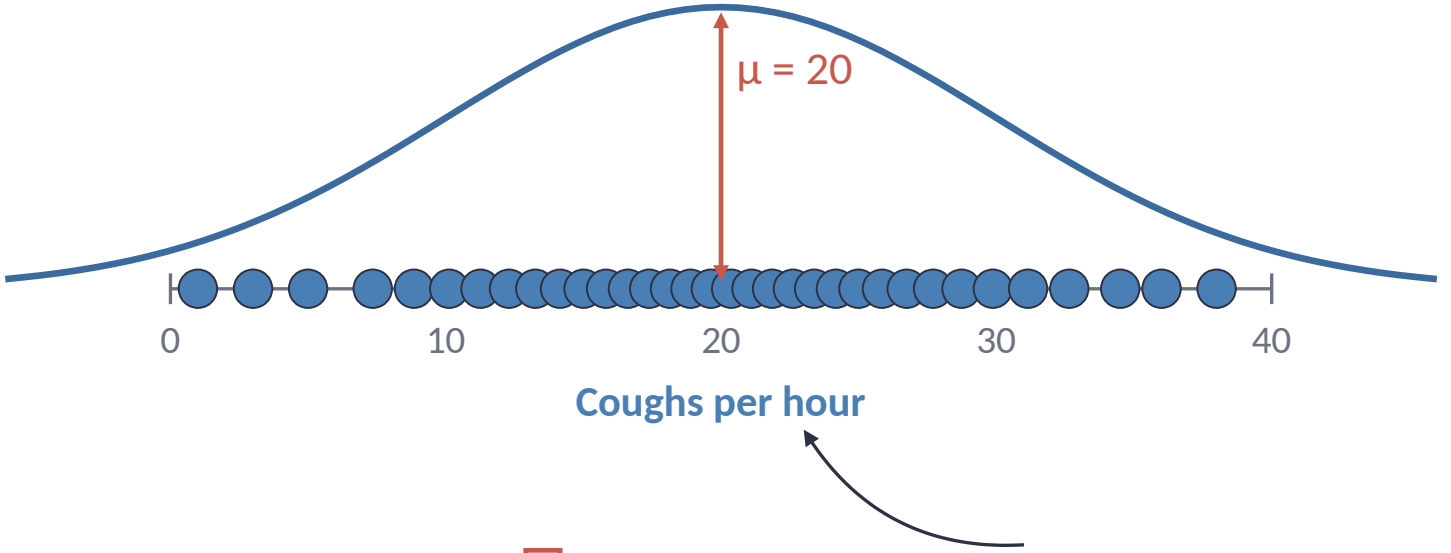
The units of the variance

So we have calculated the **population variance**: **100**. However, there is one thing that is annoying about it.

Because each term is squared…

…the units of the result, 100, are **coughs²**: "coughs per hour, squared".

…so we cannot plot the variance on the graph, since the units on the x-axis are *not* squared.

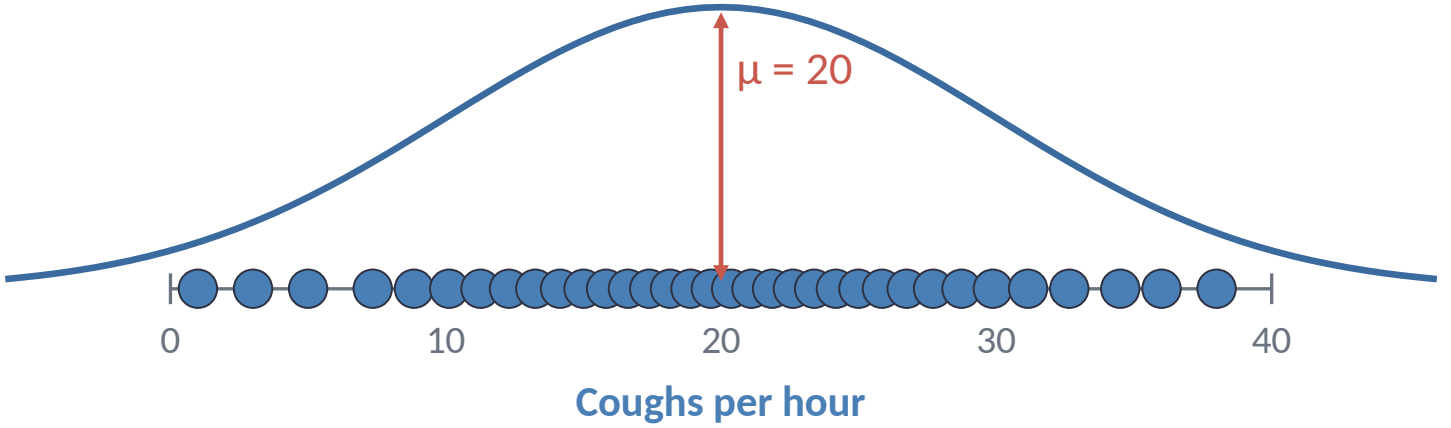


$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$$= \frac{(1 - 20)^2 + (3 - 20)^2 + \ldots + (36 - 20)^2 + (38 - 20)^2}{2{,}000{,}000} = \boxed{100}$$

The population standard deviation

To solve this problem, we take the square root of everything.



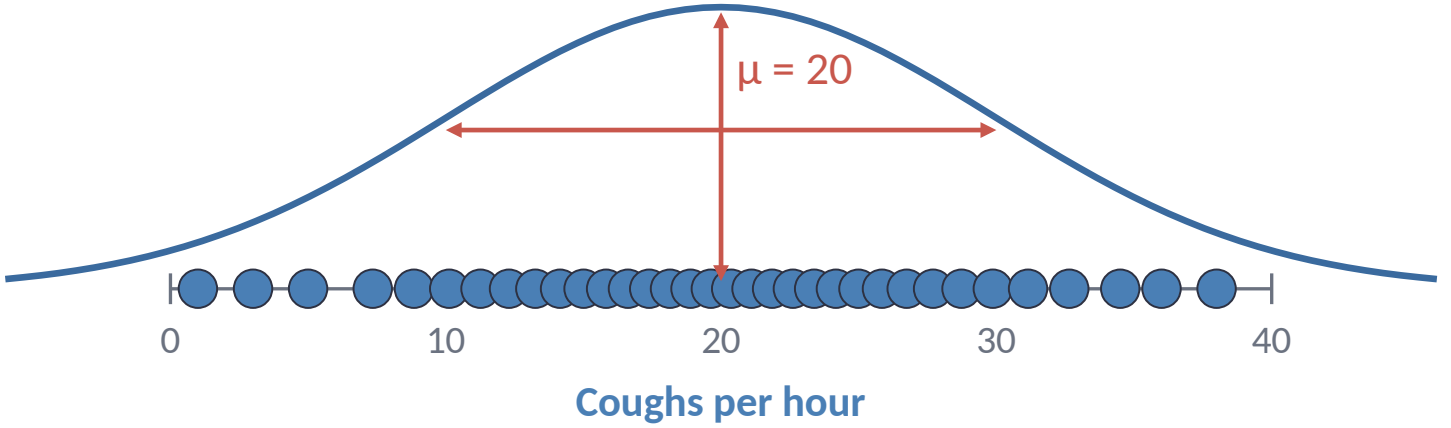
$$\sqrt{\frac{\Sigma(x-\mu)^2}{n}}$$

The population standard deviation

To solve this problem, we take the square root of everything.

So the **population standard deviation** is the square root of **100**, the population variance, which is **10**.

…and, since it is in the same units as the measurements, we can plot it on the graph.

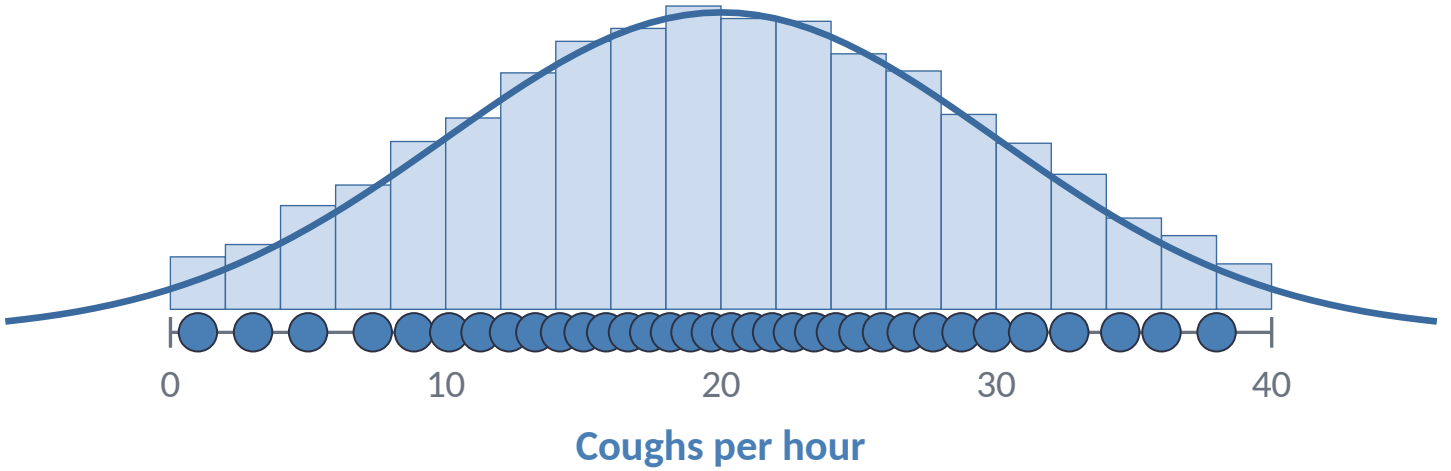


This shows the **mean**, 20, plus and minus the **standard deviation**, 10 coughs per hour.

$$\text{Population Standard Deviation} = \sqrt{\frac{\Sigma(x-\mu)^2}{n}} = \sqrt{\text{Population Variance}} = \sqrt{100} = \boxed{10}$$

Estimating the variance from a sample

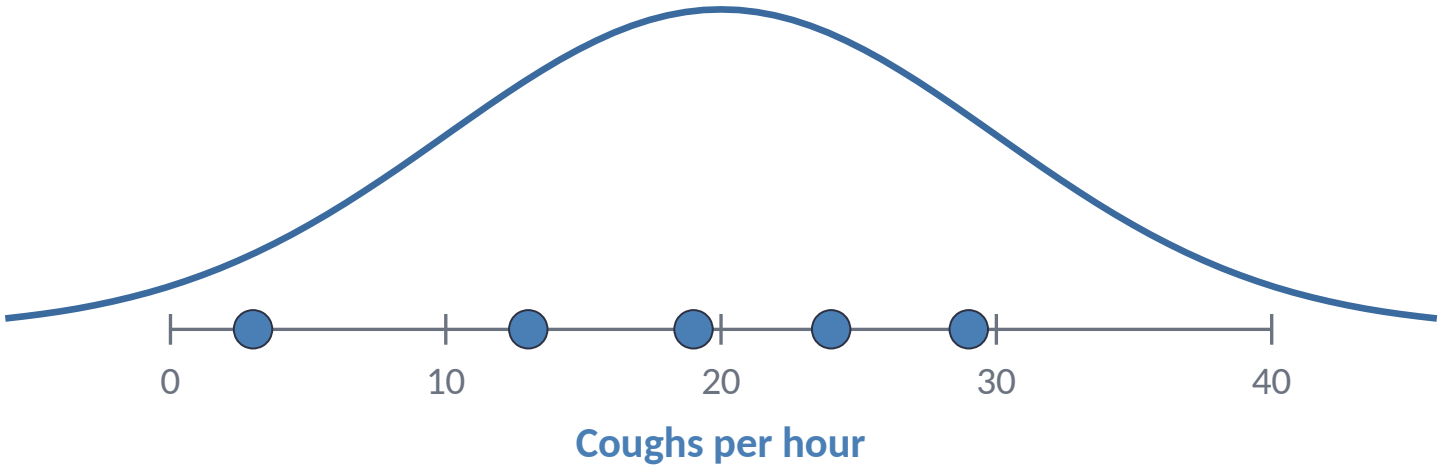
Note: we almost never have the data for the whole population, so we almost never calculate the **population** mean, variance and standard deviation.



Estimating the variance from a sample

Note: we almost never have the data for the whole population, so we almost never calculate the **population** mean, variance and standard deviation.

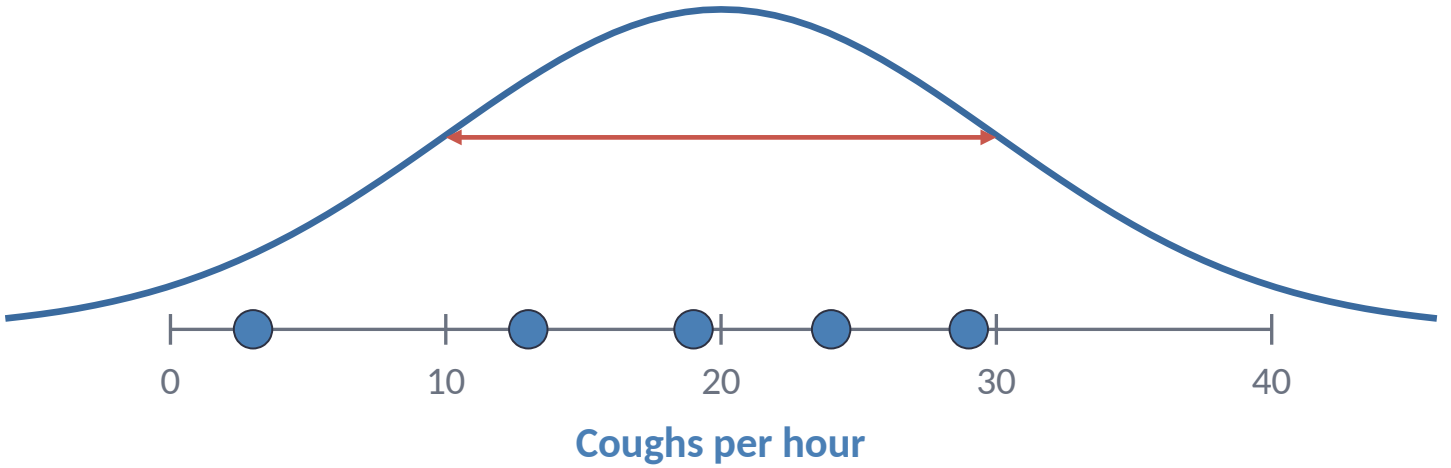
Instead, we ***estimate*** the population variance and standard deviation from the relatively small number of measurements that we have.



Estimating the variance from a sample

Note: we almost never have the data for the whole population, so we almost never calculate the **population** mean, variance and standard deviation.

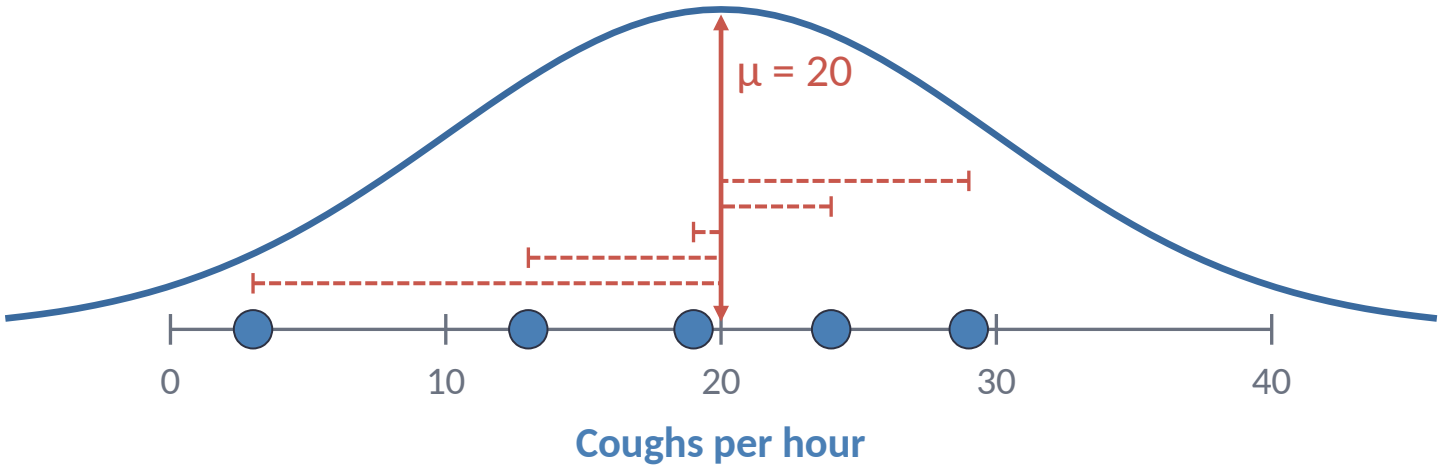
Instead, we *estimate* the population variance and standard deviation from the relatively small number of measurements that we have.



Estimating the variance from a sample

Note: we almost never have the data for the whole population, so we almost never calculate the **population** mean, variance and standard deviation.

Instead, we *estimate* the population variance and standard deviation from the relatively small number of measurements that we have.

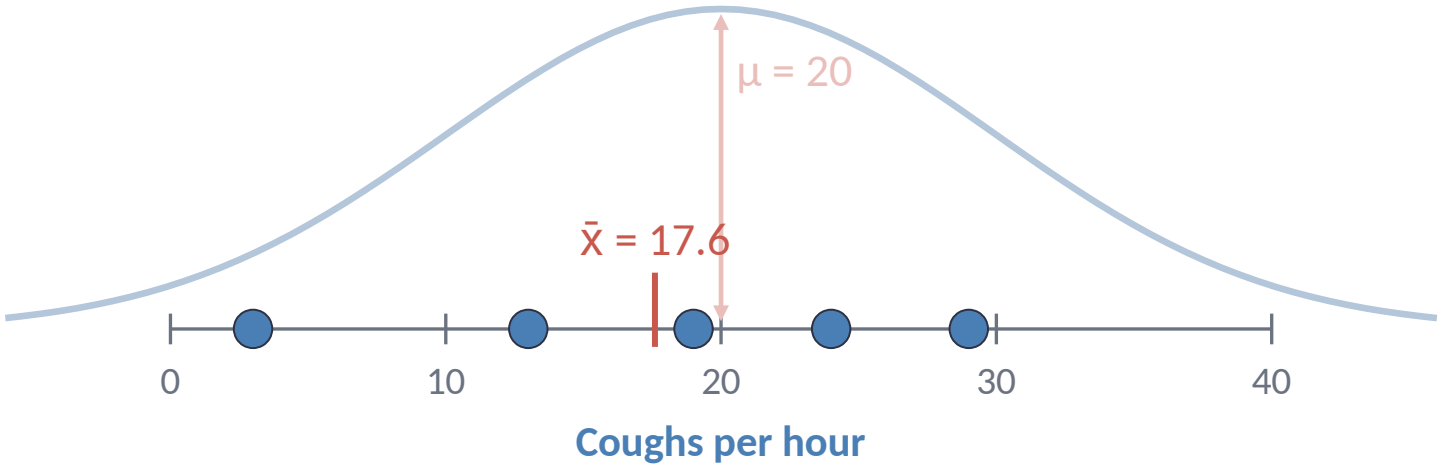


The formula for the estimated variance

This is the formula we use to *estimate* the population variance.

The differences between this formula and the one for the *calculated* population variance are subtle but important.

First, since we do not know the population mean, **μ**, we use the **sample mean**, **$\bar{x}$**.



$$\text{Estimated Population Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

The formula for the estimated variance

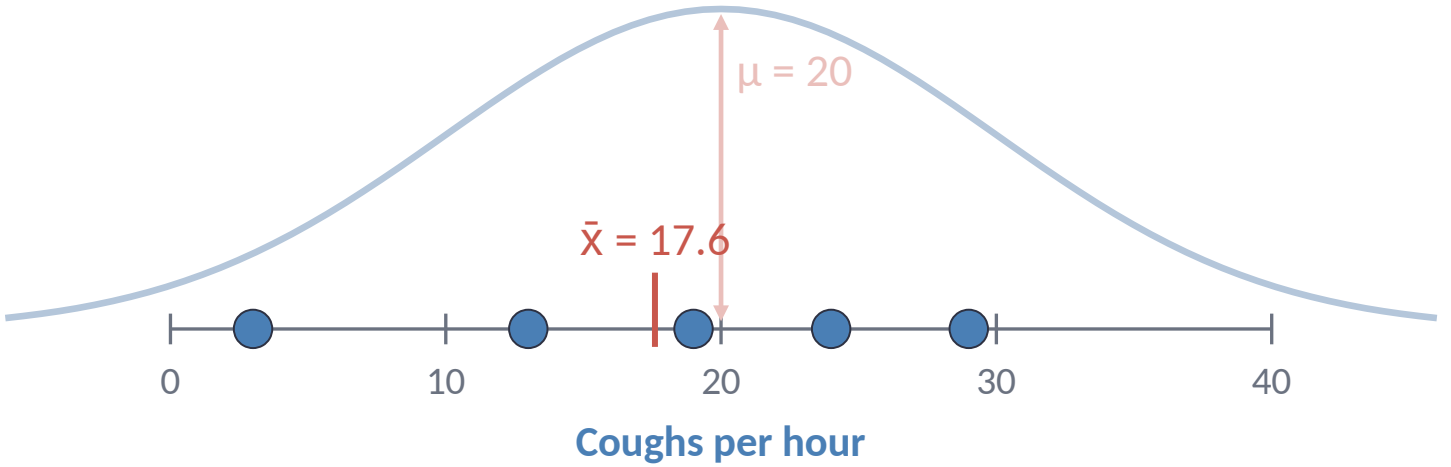
This is the formula we use to *estimate* the population variance.

The differences between this formula and the one for the *calculated* population variance are subtle but important.

First, since we do not know the population mean, **μ**, we use the **sample mean**, **$\bar{x}$**.

Second, we divide by **n − 1** instead of **n**.

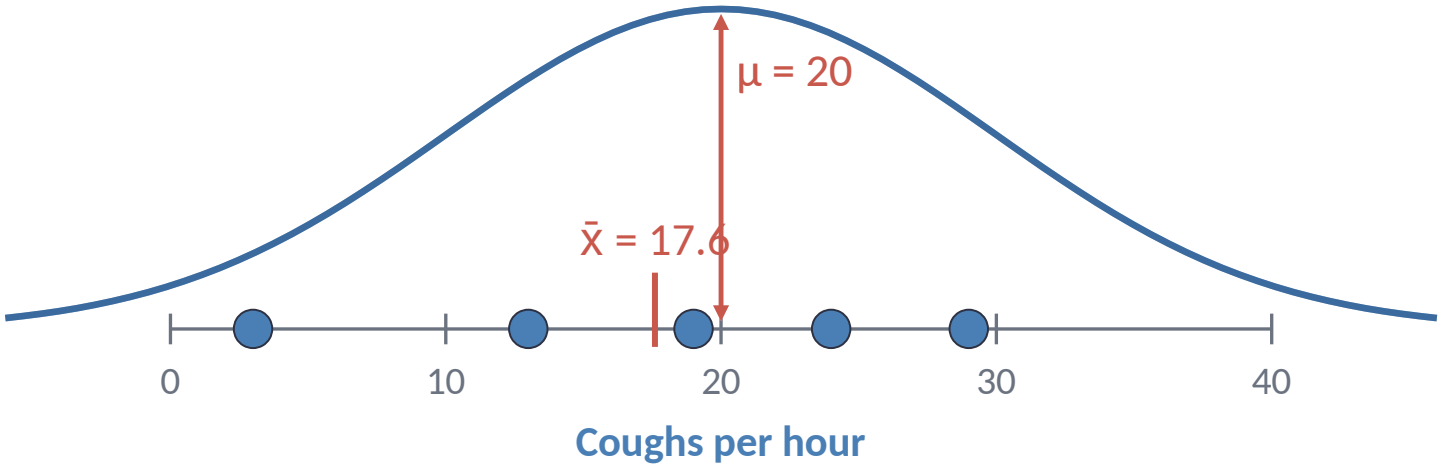
Dividing by n − 1 compensates for the fact that we calculate differences from the **sample mean** instead of the **population mean**…



$$\text{Estimated Population Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

The formula for the estimated variance

…otherwise we would consistently *underestimate* the variance around the population mean.

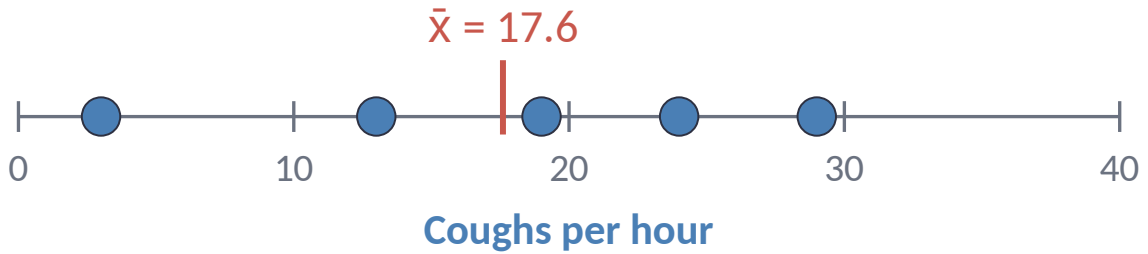


$$\text{Estimated Population Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$$\frac{\Sigma(x - \mu)^2}{n}$$

Calculating the estimates

Now we plug in the numbers. First, we subtract the sample mean, **17.6**, from each measurement…



$$\text{Estimated Population Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$(3 - 17.6)$ $(13 - 17.6)$ $(19 - 17.6)$ $(24 - 17.6)$ $(29 - 17.6)$

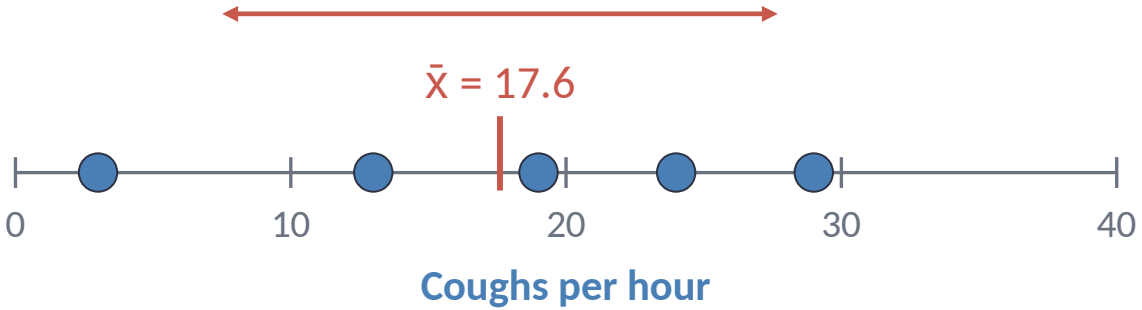
Calculating the estimates

Now we plug in the numbers. First, we subtract the sample mean, **17.6**, from each measurement…

…then we square each difference, add them up and divide by 5 − 1. The **estimated population variance** is **101.8**.

…and the **estimated standard deviation** is its square root: **10.1**.

Since it is in the same units as the measurements, we can draw it on the graph, around the sample mean.



$$\text{Estimated Population Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$$= \frac{213.2 + 21.2 + 2.0 + 41.0 + 130.0}{5 - 1} = \boxed{101.8}$$

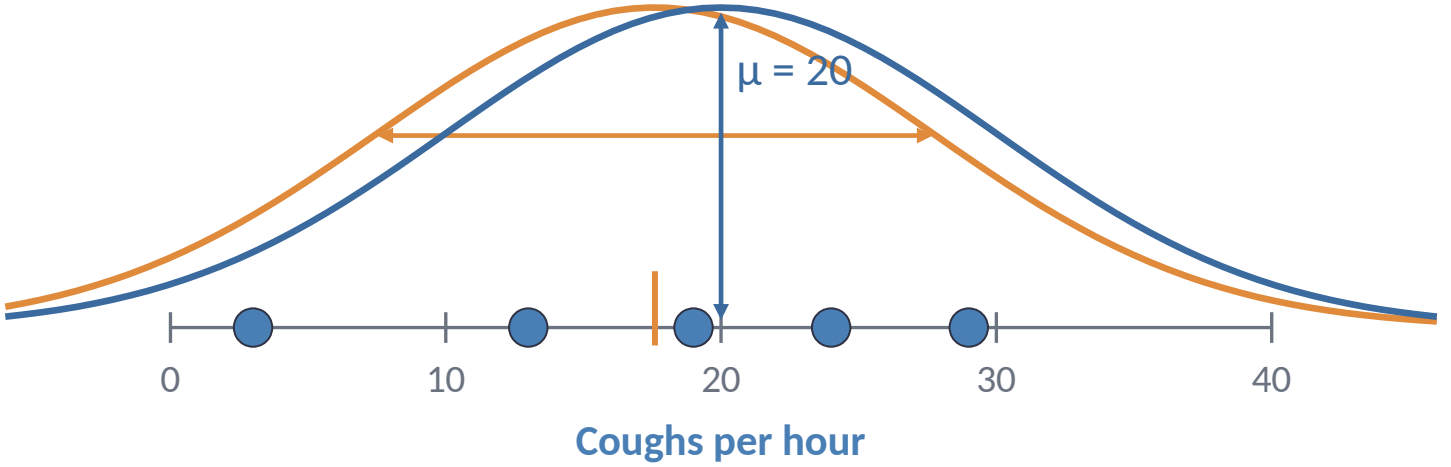
$$\text{Estimated Population Standard Deviation} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{101.8} = \boxed{10.1}$$

Estimates and the true population

The *estimated* population parameters correspond to this **orange curve**, with mean **17.6** and standard deviation **10.1**.

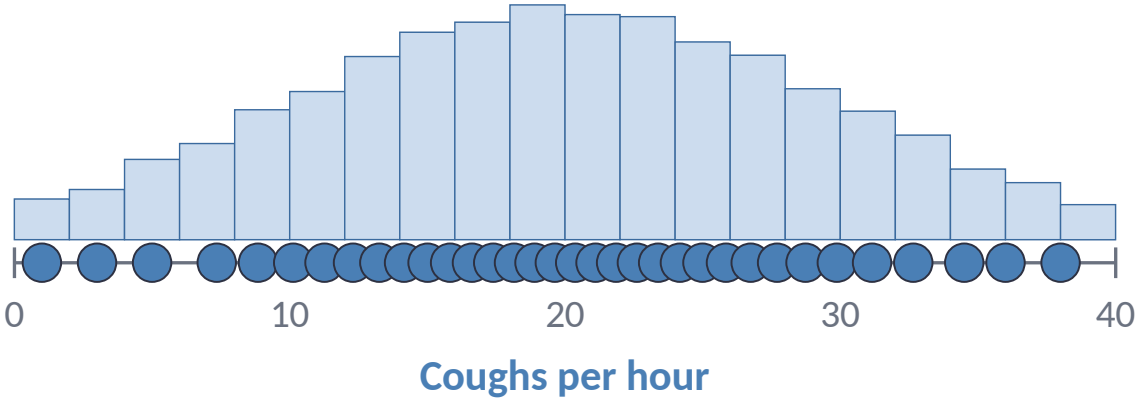
The true population curve, in **blue**, has mean 20 and standard deviation 10.

With more data, the estimated parameters would be more accurate, and we would have more confidence in them.



In summary: the population

When we have data for the whole population, the **population mean** is just the average measurement.



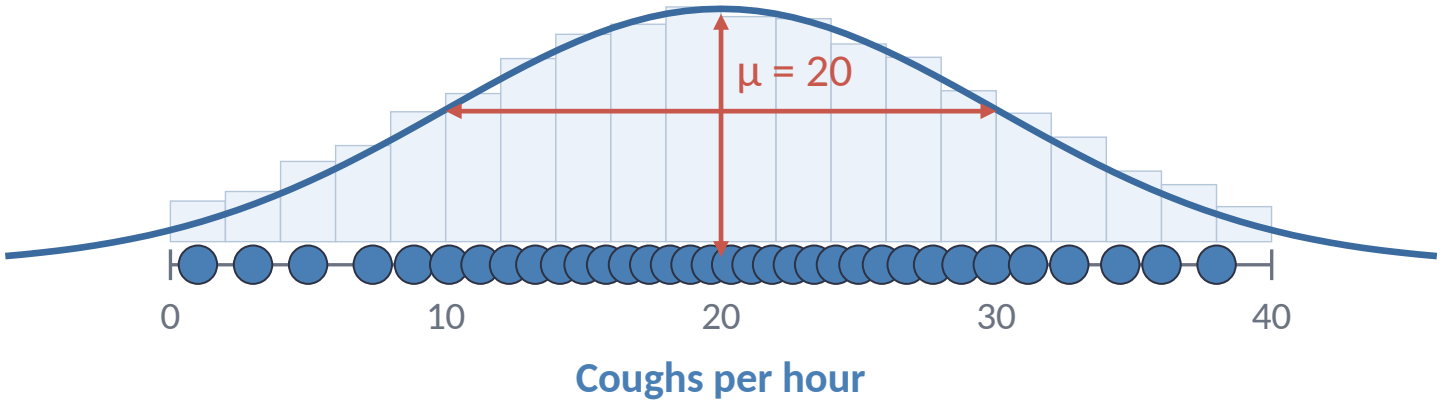
$$\text{Population Mean} = \frac{\text{Sum of the measurements}}{\text{Number of measurements}}$$

In summary: the population

When we have data for the whole population, the **population mean** is just the average measurement.

The **population variance** is the average of the squared differences between the data and the population mean, **μ**. Squaring prevents the differences on the left from cancelling those on the right.

The **population standard deviation** is the square root of the variance. It is in the original units, so we can draw it on the graph.



$$\text{Population Mean} = \frac{\text{Sum of the measurements}}{\text{Number of measurements}}$$

$$\text{Population Variance} = \frac{\Sigma(x - \mu)^2}{n}$$

$$\text{Population Standard Deviation} = \sqrt{\text{Population Variance}} = \sqrt{\frac{\Sigma(x - \mu)^2}{n}}$$

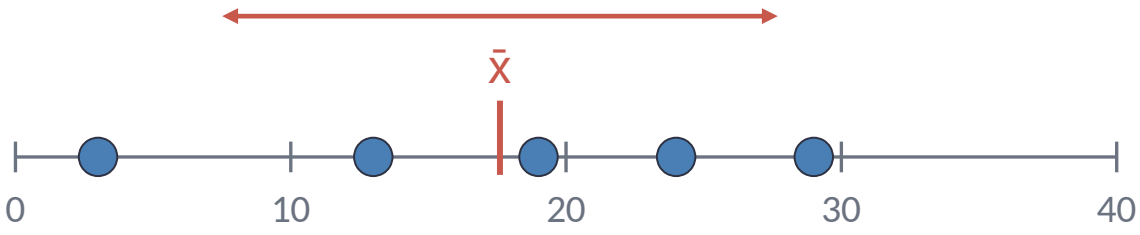
In summary: estimating from a sample

However, we almost never have the population data, so we should rarely use these formulas.

Instead, we almost always *estimate* the mean, the variance and the standard deviation from a sample.

When we estimate the variance, we divide by **n − 1** to compensate for measuring distances from the sample mean instead of the population mean.

The estimated standard deviation is its square root. It is in the same units as the data, so we can draw it on the graph.



~~Population Variance = $\frac{\Sigma(x-\mu)^2}{n}$~~

~~Population Standard Deviation = $\sqrt{\frac{\Sigma(x-\mu)^2}{n}}$~~

Estimated Population Variance = $\frac{\Sigma(x-\bar{x})^2}{n-1}$

Estimated Population Standard Deviation = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

In practice: software

P.S. We have stressed that we rarely have the population data, so we almost always *estimate* the population parameters.

In Python, take care with the defaults: **numpy.var(x)** divides by **n**, while **numpy.var(x, ddof=1)** and pandas' **Series.var()** divide by **n − 1**.

Since we almost always have a relatively small sample, rather than the population data, we should almost always use the **n − 1** version.

Python

```
x = [3, 13, 19, 24, 29]
np.var(x)           # 81.4  (divides by n)
np.var(x, ddof=1)   # 101.8 (divides by n − 1)
pd.Series(x).var() # 101.8 (divides by n − 1)
```

~~Population Variance = $\frac{\Sigma(x-\mu)^2}{n}$~~

Estimated Population Variance = $\frac{\Sigma(x-\bar{x})^2}{n-1}$

Covariance

How two measurements vary together

Conceptual Insight

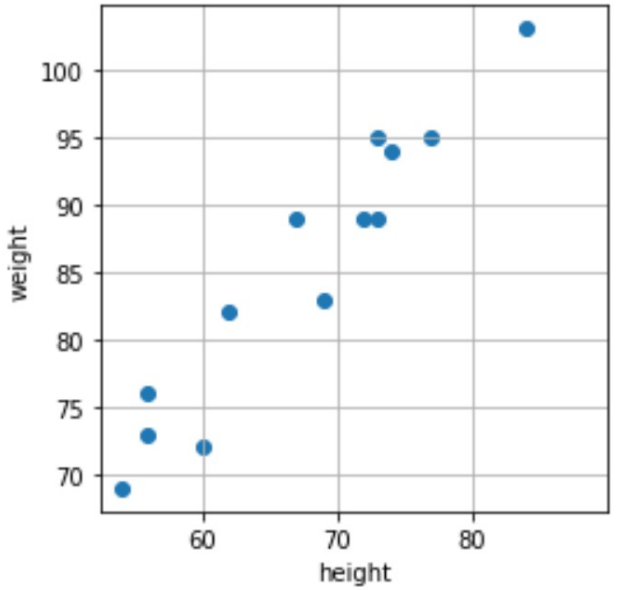
- Suppose we want to guess a variable X.
  If we know nothing else, our uncertainty about X is high.

- Now suppose we are told the value of another variable Y.
  If X and Y are related, then knowing Y gives us **clues** about X.
  Our uncertainty about X goes down.

- The stronger the relationship between them:
  - the more accurately we can guess X from Y
  - the less uncertainty remains
  - the more *information* they share

So:

- **Information is reduction in uncertainty.**
  Knowing one variable helps predict the other.

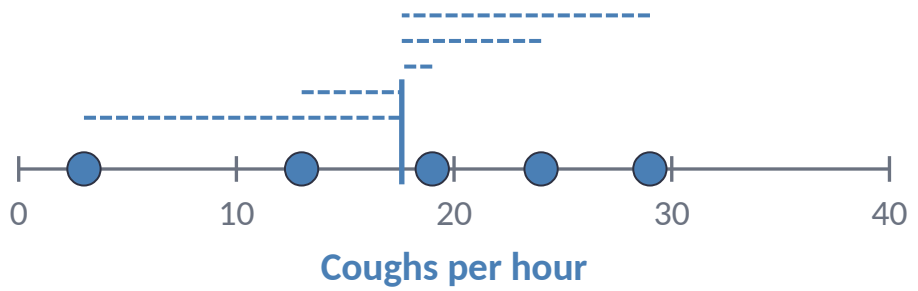


From one measurement to two

Earlier, we counted the coughs of **5** flu patients during one hour…

…and then we estimated the **variance**.

$$\frac{\Sigma(x - \bar{x})^2}{n - 1}$$

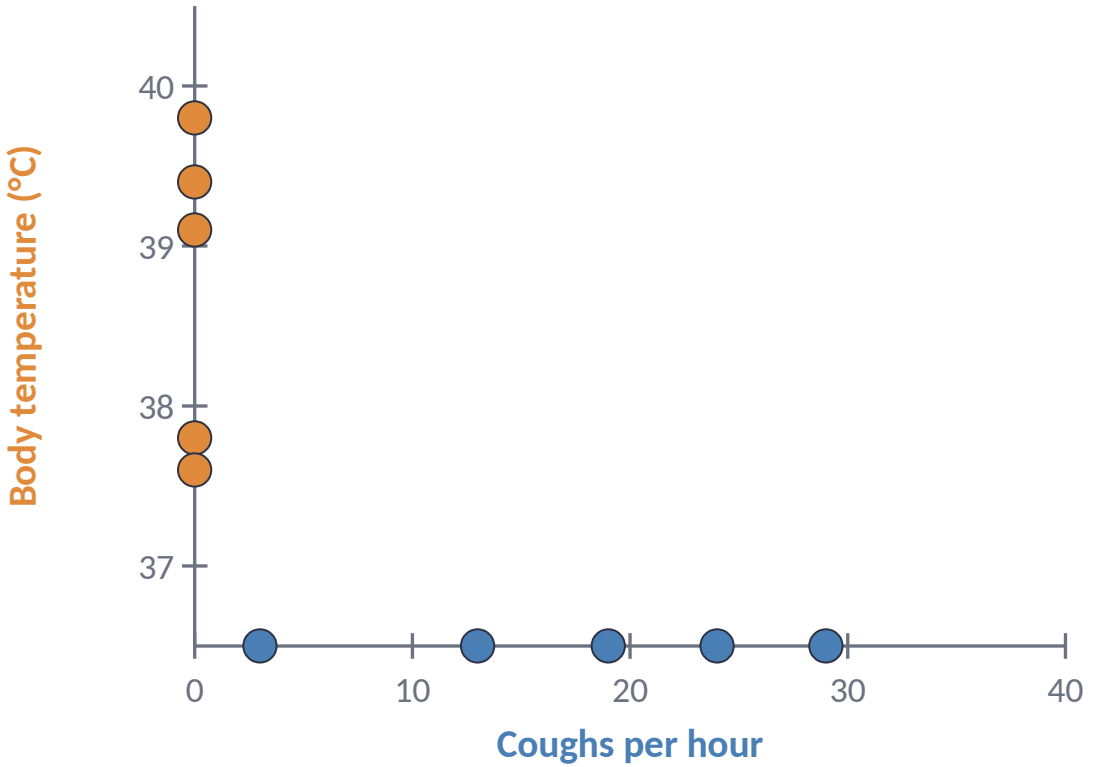


From one measurement to two

Earlier, we counted the coughs of **5** flu patients during one hour…

…and then we estimated the **variance**.

Now imagine that, in addition to counting coughs, we also measured the **body temperature** of the same 5 patients.

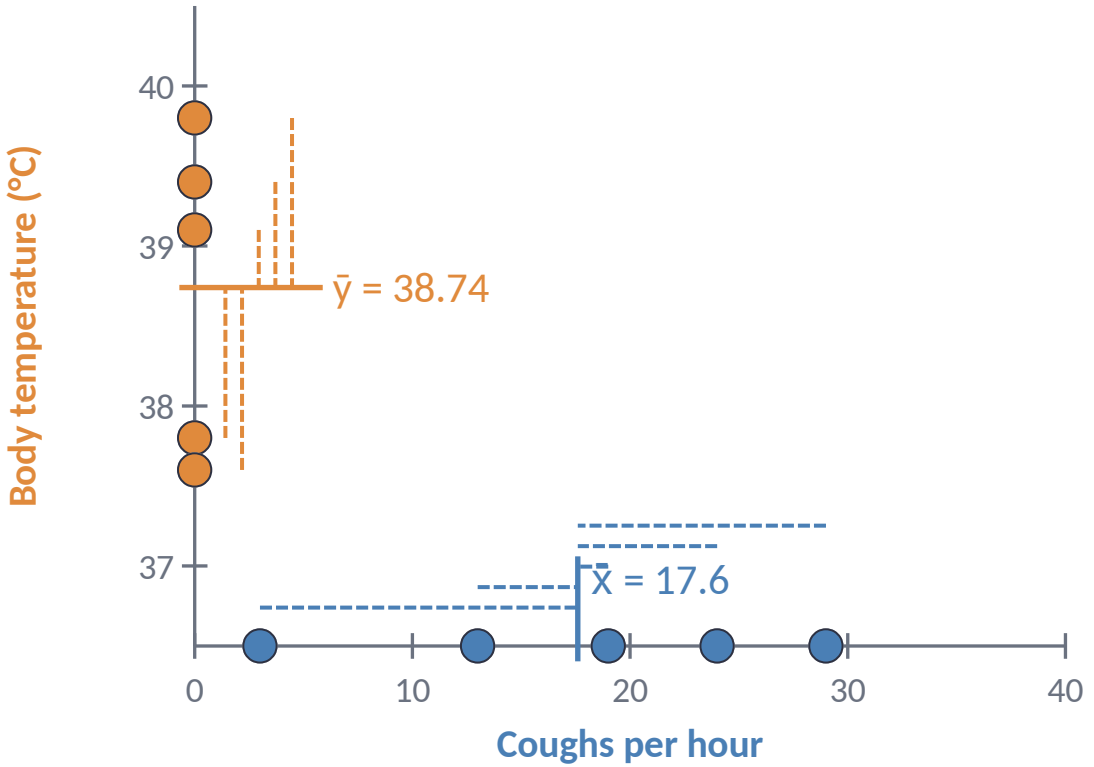


Two means and two variances

Since temperature is on the **y-axis**, we use $\bar{y}$ ("y-bar") for its mean: **38.74 °C**.

So far, we have estimated the **mean** and the **variance** for two different measurements, taken from the **same five patients**.

Since these measurements were taken from the same patients, we can look at them in **pairs**.



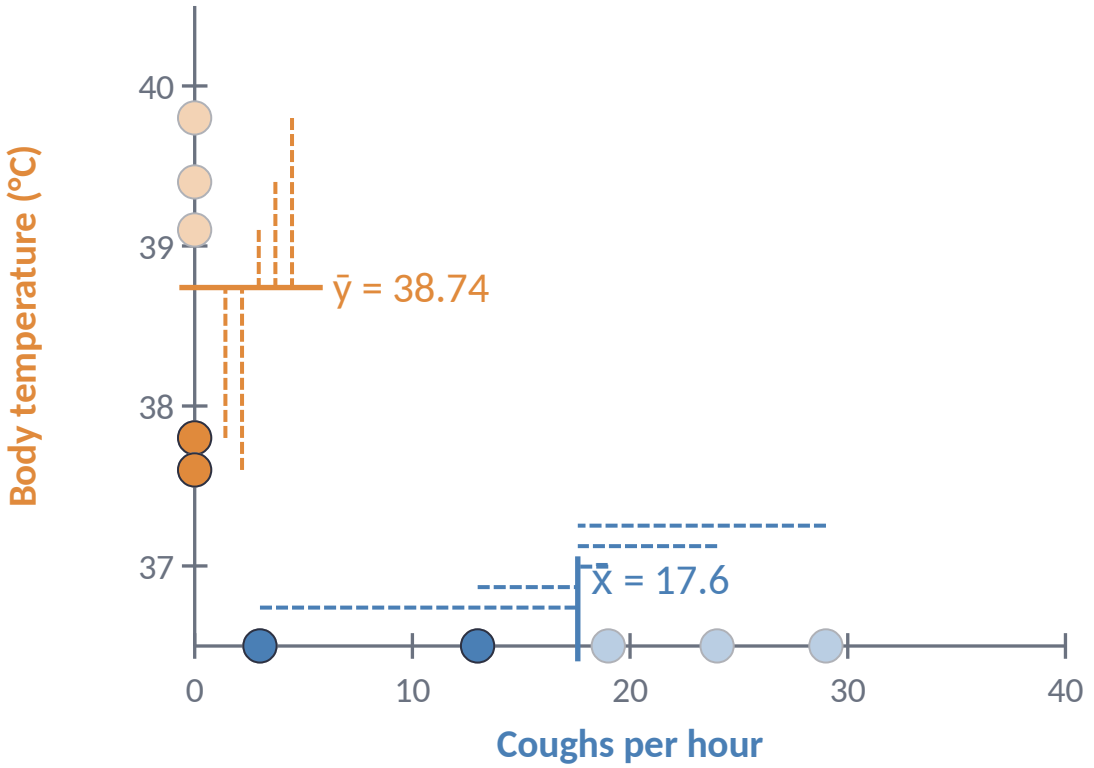
Two means and two variances

Since temperature is on the **y-axis**, we use $\bar{y}$ ("y-bar") for its mean: **38.74 °C**.

So far, we have estimated the **mean** and the **variance** for two different measurements, taken from the **same five patients**.

Since these measurements were taken from the same patients, we can look at them in **pairs**.

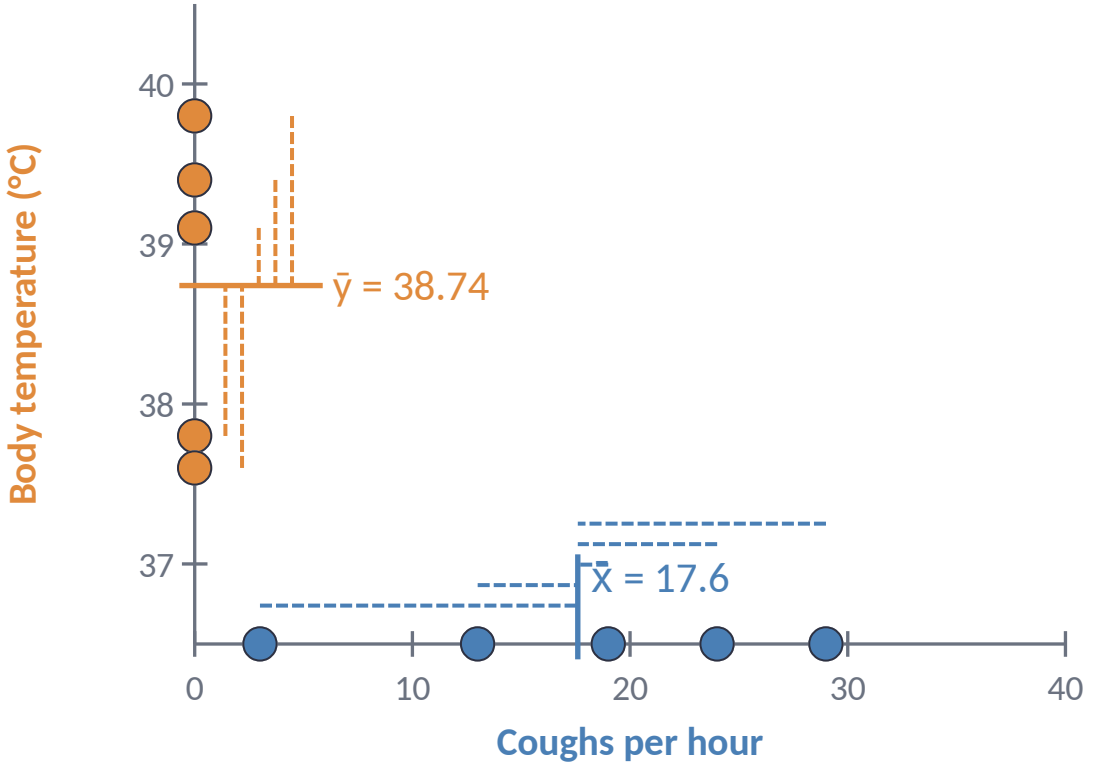
For these two patients, both measurements are less than their respective means…



Two means and two variances

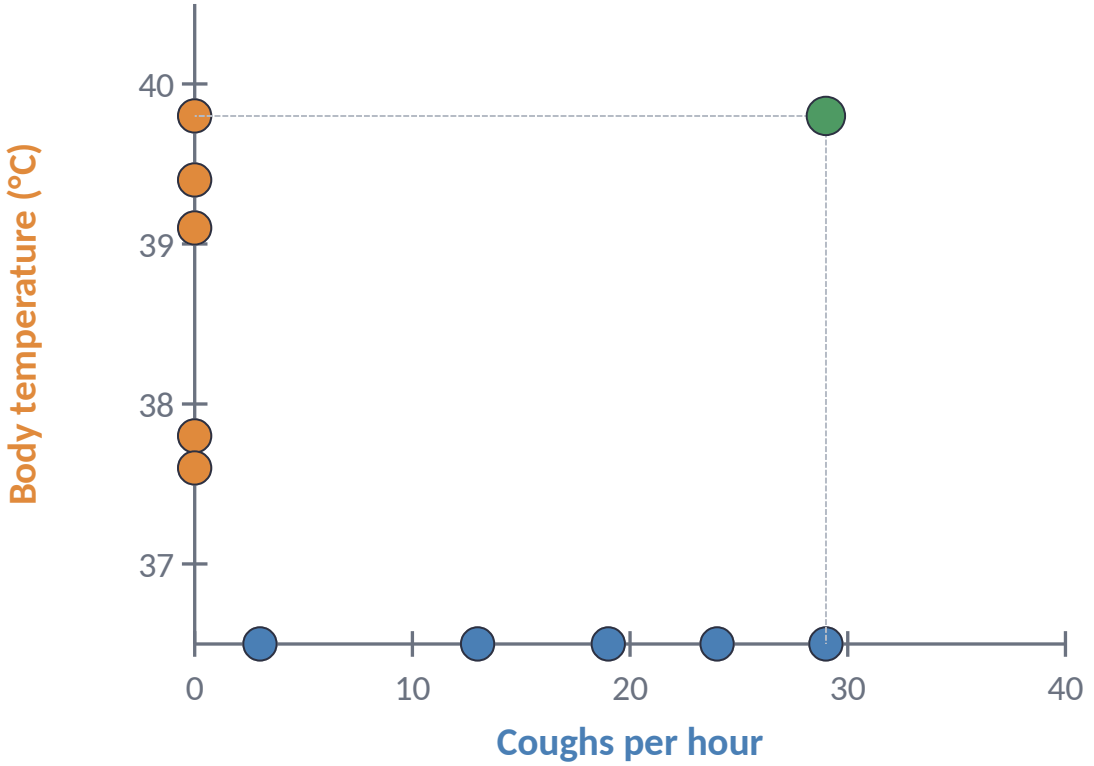
…and for these two patients, both measurements are greater than their respective means.

Since the measurements were taken in pairs, the question is: **“Do the measurements, taken as pairs, tell us something that the individual measurements do not?”**



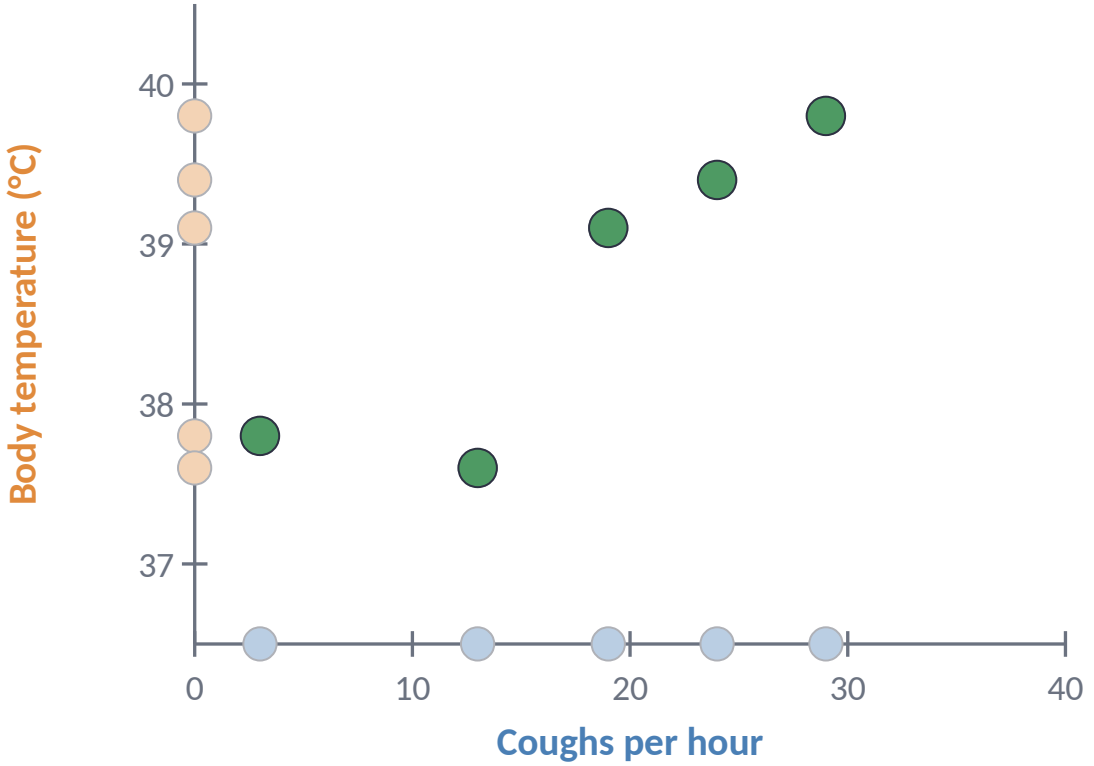
Plotting the pairs

Since the measurements came from the same patients, we can plot each pair as a single dot, by combining the values on the **x-axis** and the **y-axis**.



Plotting the pairs

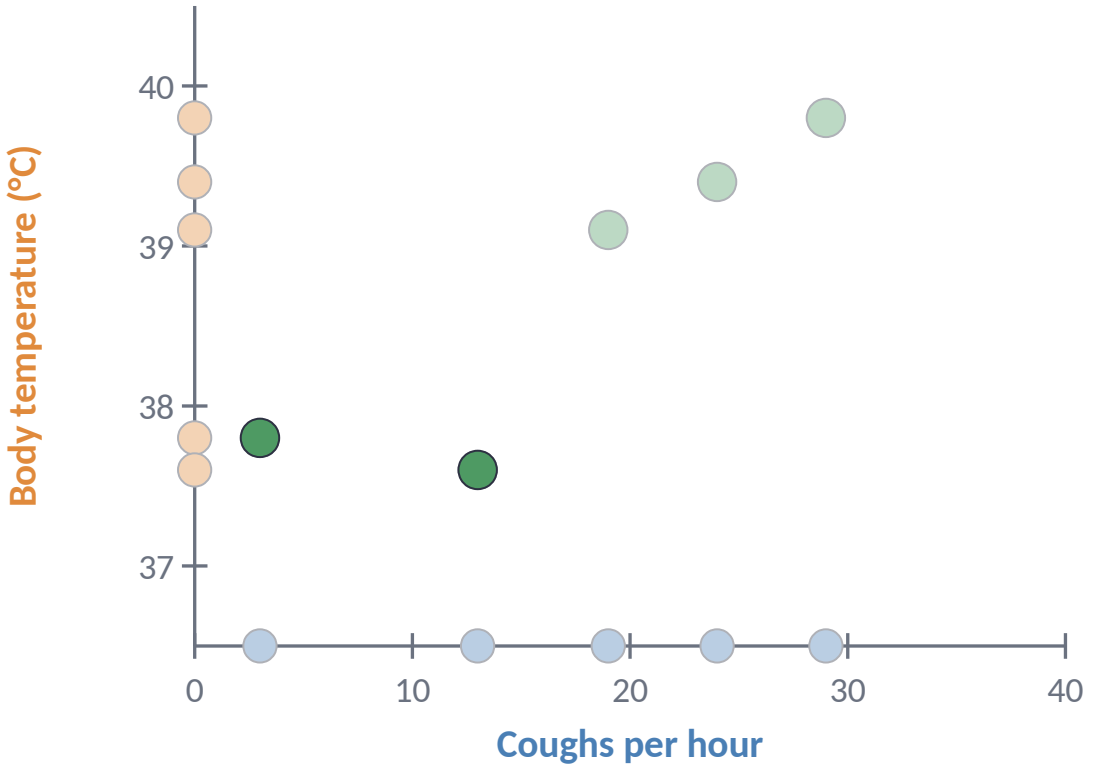
Since the measurements came from the same patients, we can plot each pair as a single dot, by combining the values on the **x-axis** and the **y-axis**.



Plotting the pairs

Since the measurements came from the same patients, we can plot each pair as a single dot, by combining the values on the **x-axis** and the **y-axis**.

Generally speaking, patients with relatively **few** coughs also have relatively **low** temperatures…



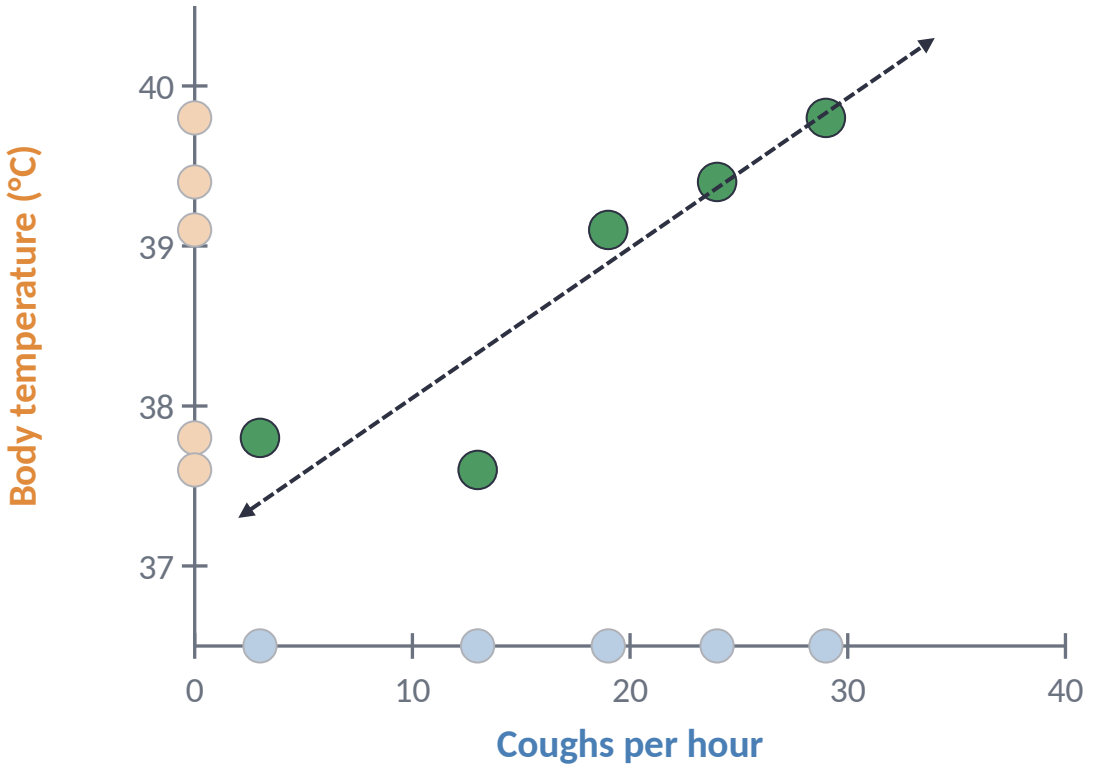
Plotting the pairs

Since the measurements came from the same patients, we can plot each pair as a single dot, by combining the values on the **x-axis** and the **y-axis**.

Generally speaking, patients with relatively **few** coughs also have relatively **low** temperatures…

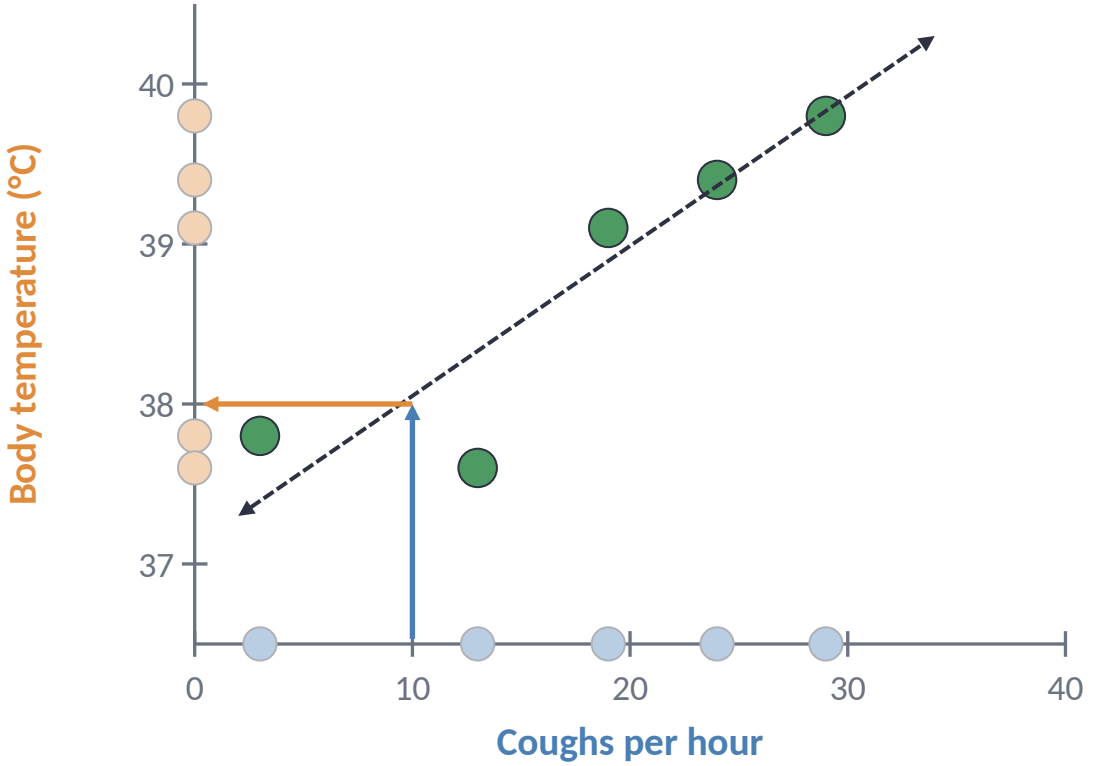
…and patients with relatively **many** coughs also have relatively **high** temperatures.

This reflects the *positive* trend, where the number of coughs and the temperature increase together.



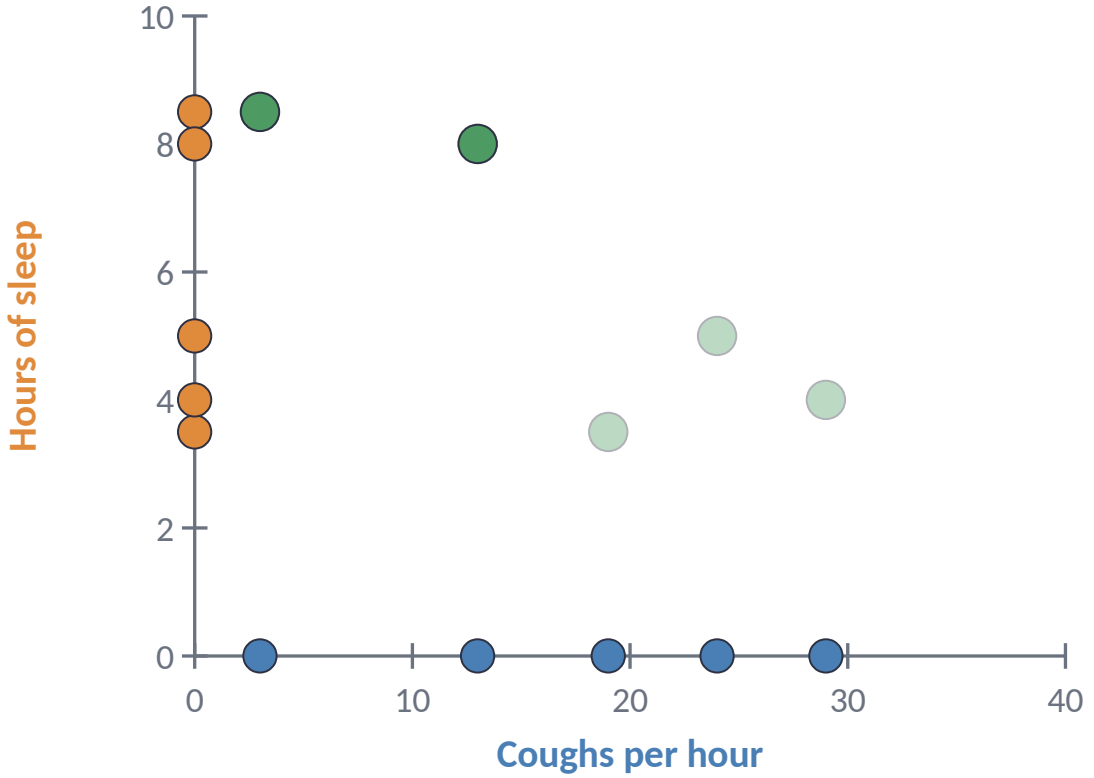
Plotting the pairs

The trend suggests that a patient who coughs only a little is likely to have a relatively low temperature.



A negative trend

Now consider a different measurement: the hours each patient slept the night before. Relatively **few** coughs went with relatively **many** hours of sleep…

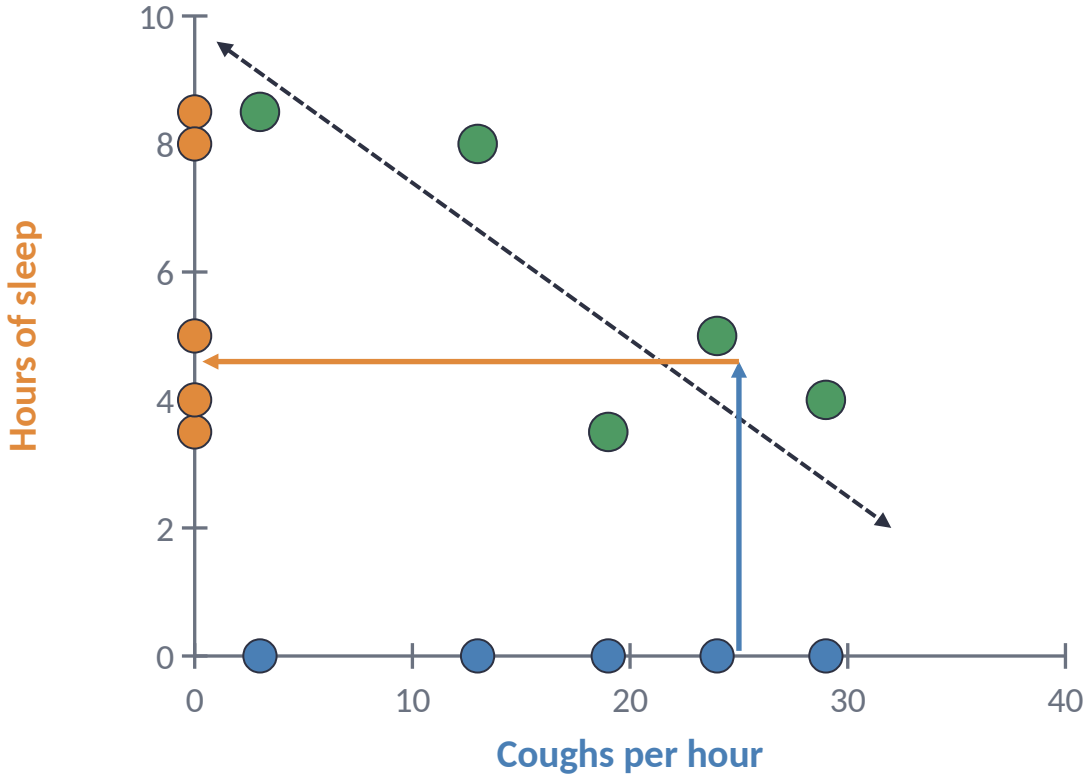


A negative trend

Now consider a different measurement: the hours each patient slept the night before. Relatively **few** coughs went with relatively **many** hours of sleep…

…and relatively **many** coughs went with relatively **few** hours of sleep.

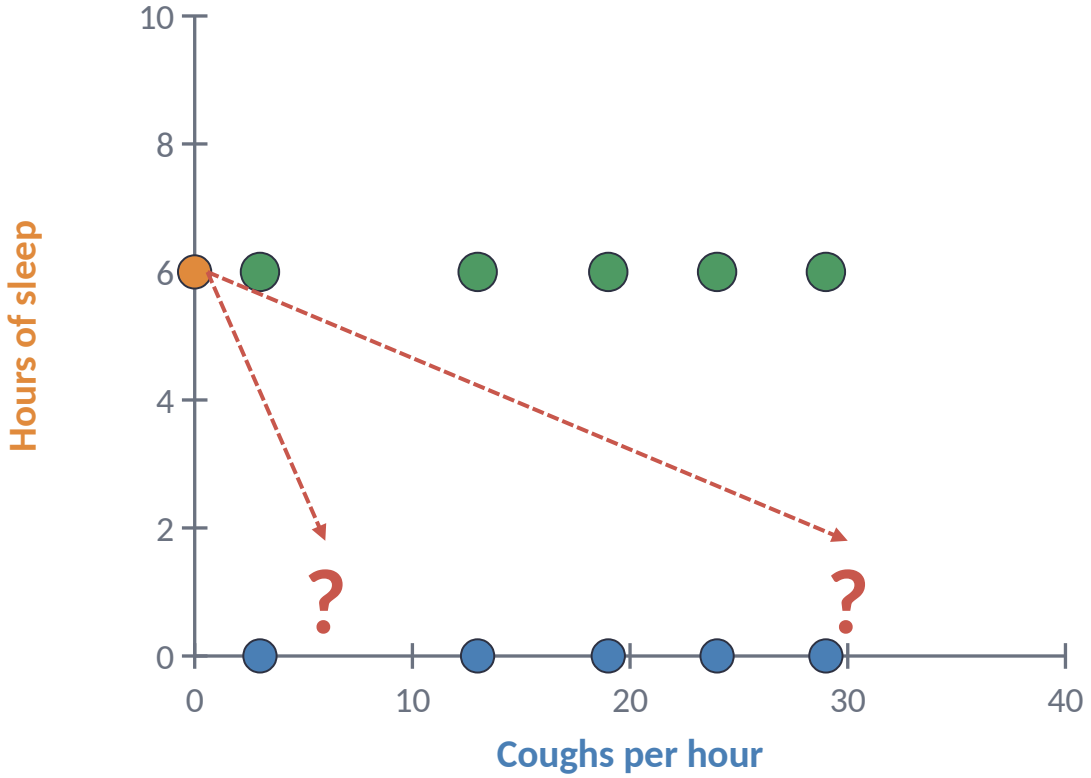
This reflects a *negative* trend: as the number of coughs increases, the hours of sleep decrease.



No relationship

Now imagine that every patient, however much they coughed, slept exactly **6** hours…

…then knowing that a patient slept 6 hours would tell us nothing about their coughs: they could have coughed a little or a lot. There would be **no relationship**.

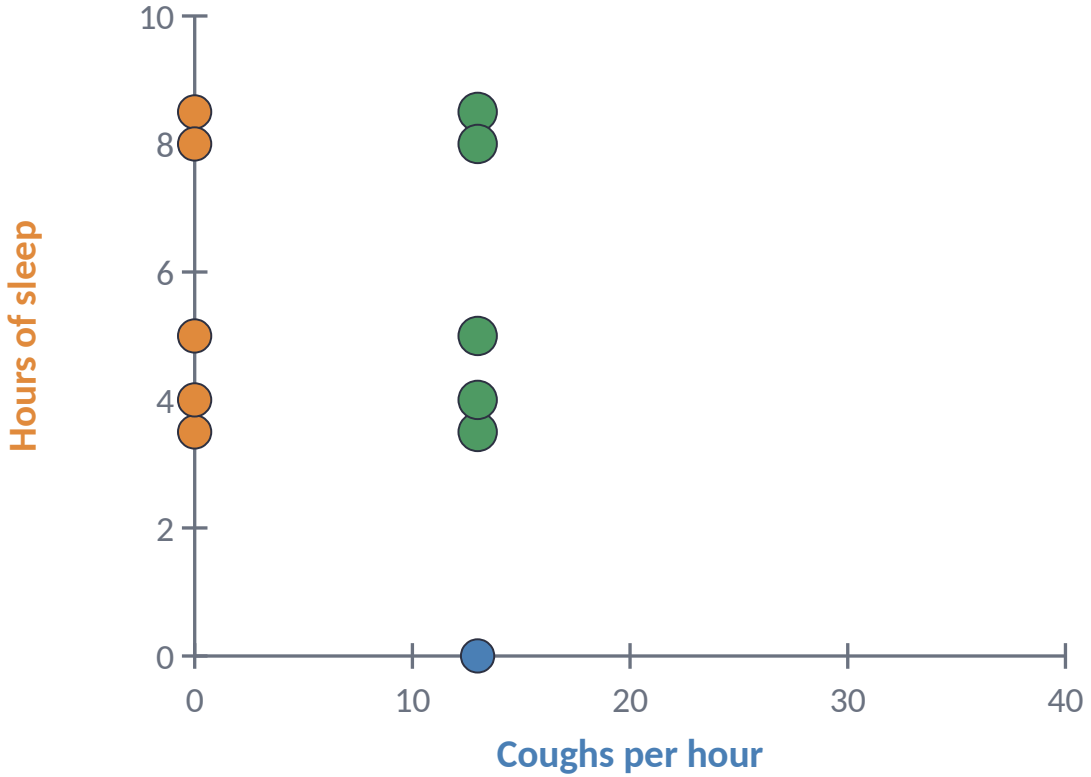


No relationship

Now imagine that every patient, however much they coughed, slept exactly **6** hours…

…then knowing that a patient slept 6 hours would tell us nothing about their coughs: they could have coughed a little or a lot. There would be **no relationship**.

Likewise, if every patient coughed exactly **13** times, whatever their sleep, the number of coughs would tell us nothing about the sleep. Again, there would be **no relationship**.

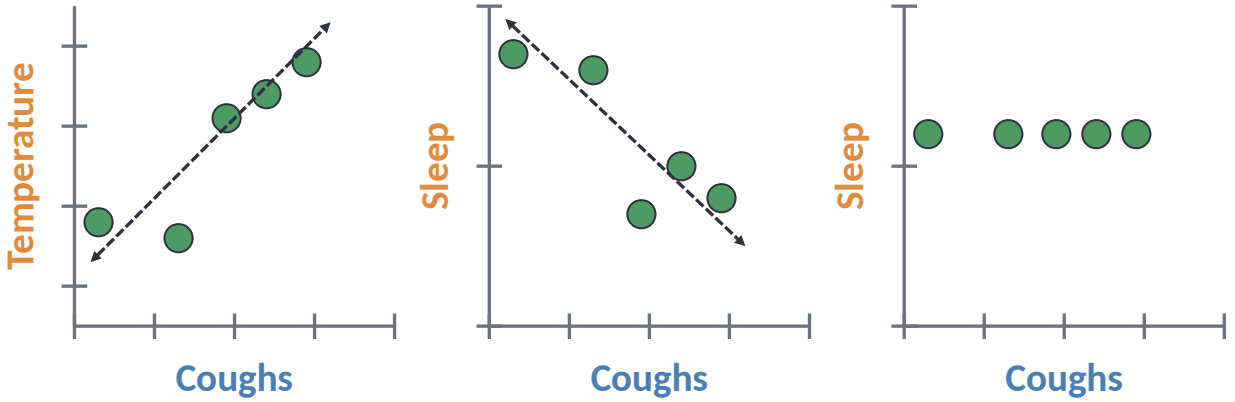


The main idea behind covariance

The main idea behind **covariance** is that it can classify three types of relationship.

This is important!
Cvariance classifies relationships as **positive**, **negative**, or **no relationship**.

Covariance, in and of itself, is not very interesting.



1) Relationships with *positive* trends

2) Relationships with *negative* trends

3) **No relationship**, because there is no trend

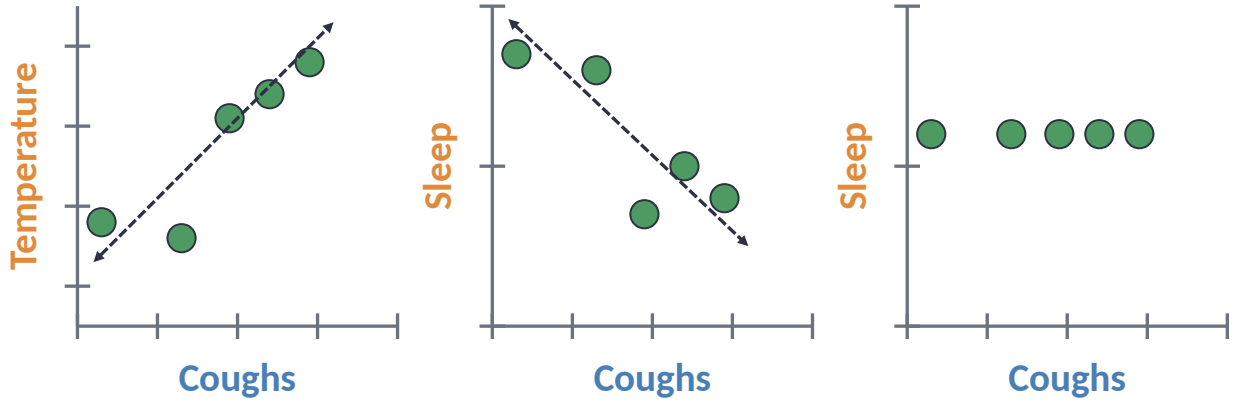
The main idea behind covariance

The main idea behind **covariance** is that it can classify three types of relationship.

This is important!
Cvariance classifies relationships as **positive**, **negative**, or **no relationship**.

Covariance, in and of itself, is not very interesting

Instead, covariance is a **computational step** to something that is interesting, like **correlation** and, later in the module, **PCA**.



1) Relationships with *positive* trends

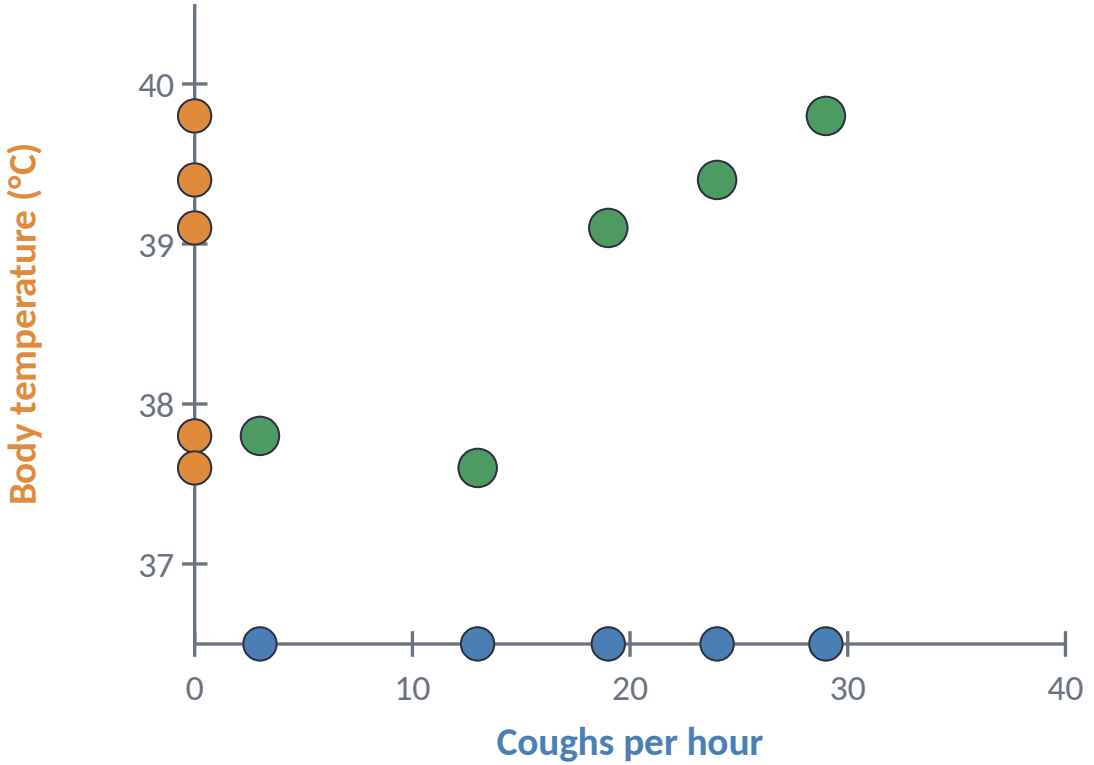
2) Relationships with *negative* trends

3) **No relationship**, because there is no trend

Calculating the covariance

Covariance is calculated with this formula.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

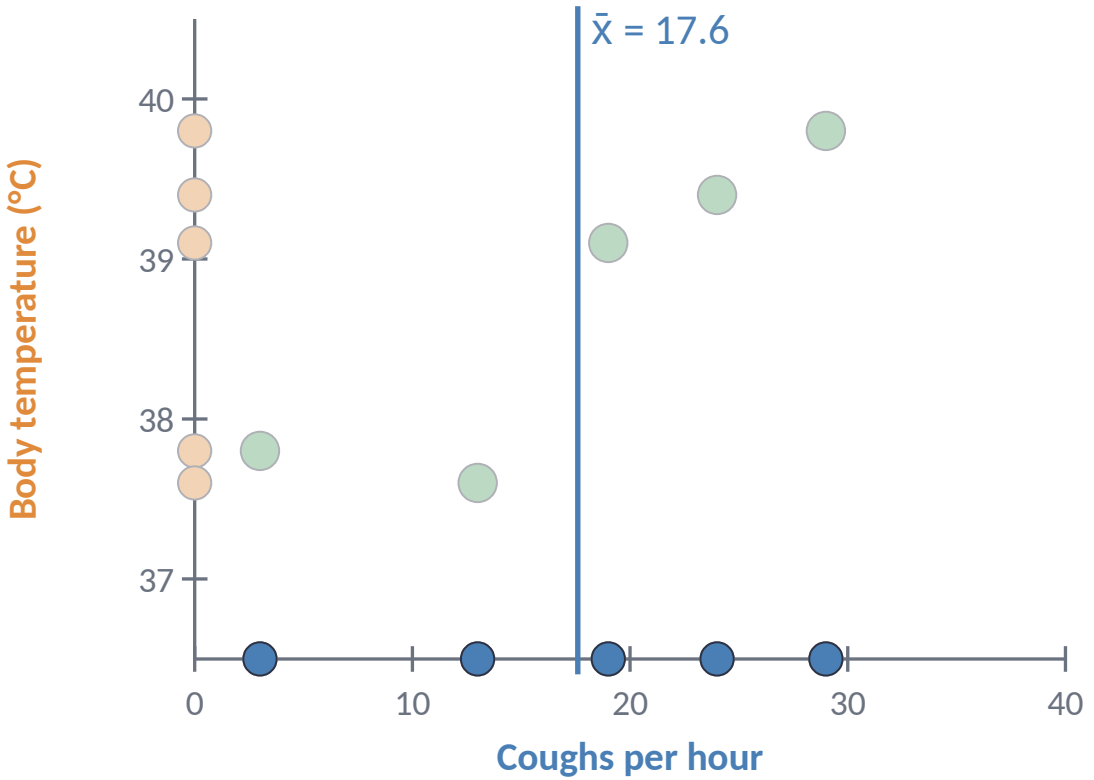


Calculating the covariance

Covariance is calculated with this formula.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

To get an intuitive sense of how it is calculated, we go back to the mean number of coughs, $\bar{x} = 17.6$, and extend the **blue line** to the top of the graph…



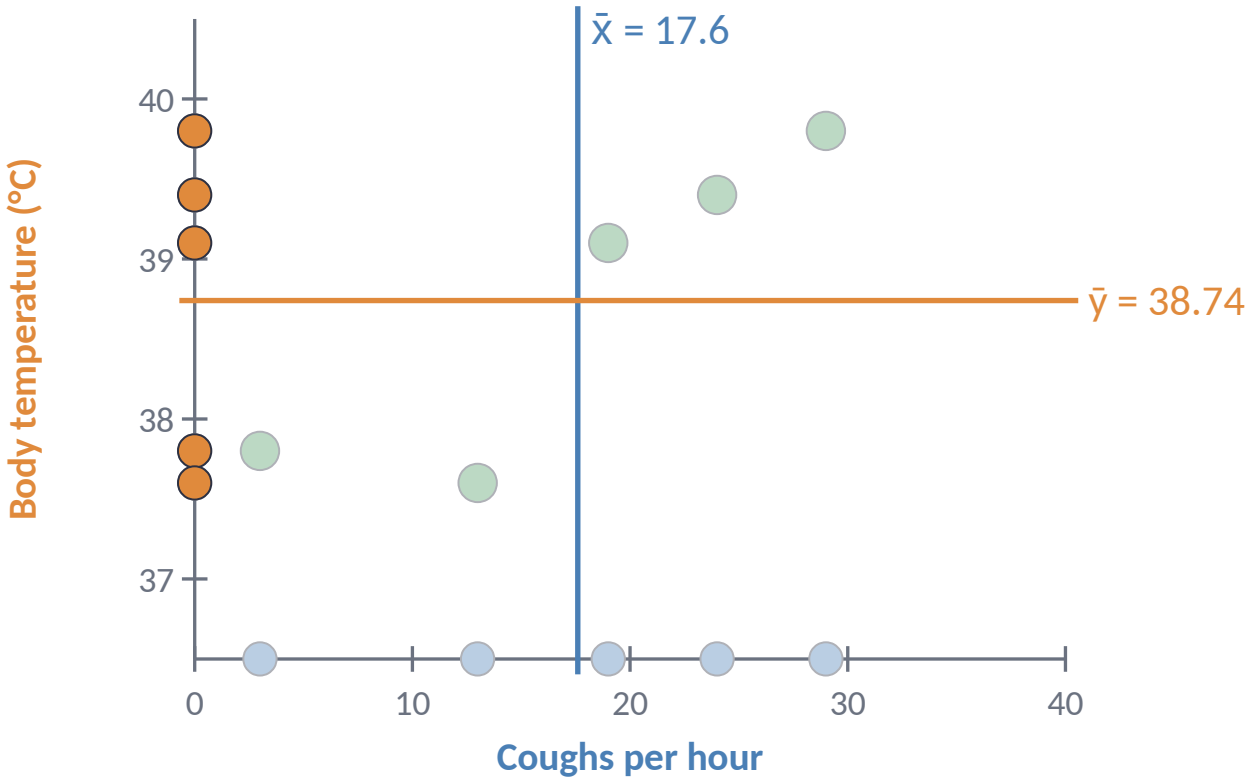
Calculating the covariance

Covariance is calculated with this formula.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

To get an intuitive sense of how it is calculated, we go back to the mean number of coughs, $\bar{x} = 17.6$, and extend the **blue line** to the top of the graph…

…and then extend the **orange line**, which represents the mean temperature, $\bar{y} = 38.74$, to the edge of the graph.



Working through each patient

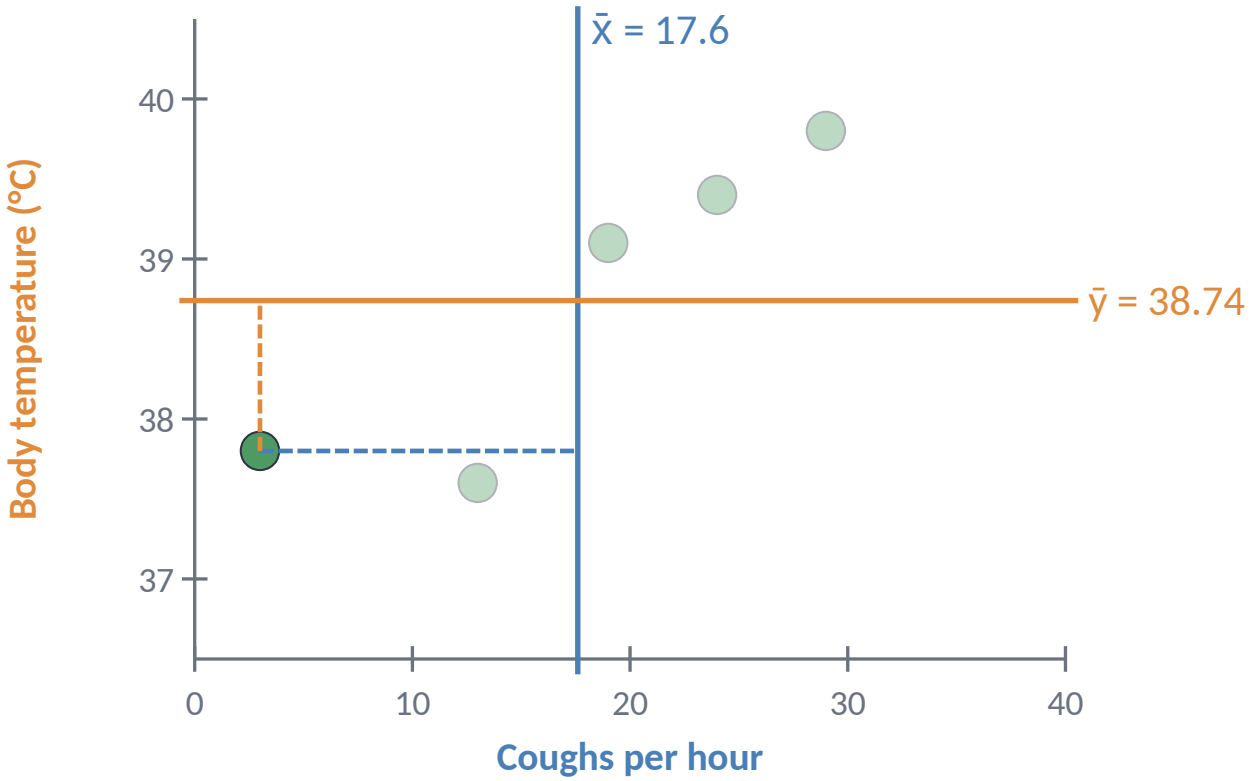
Let us focus on the patient on the far left. They are to the left of the **blue line**, so they coughed *less* than the mean…

…and they are below the **orange line**, so their temperature is *less* than the mean.

Plugging in the numbers, both differences are *negative*…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

$$(3 - 17.6) \times (37.8 - 38.74) = -14.6 \times -0.94$$



Working through each patient

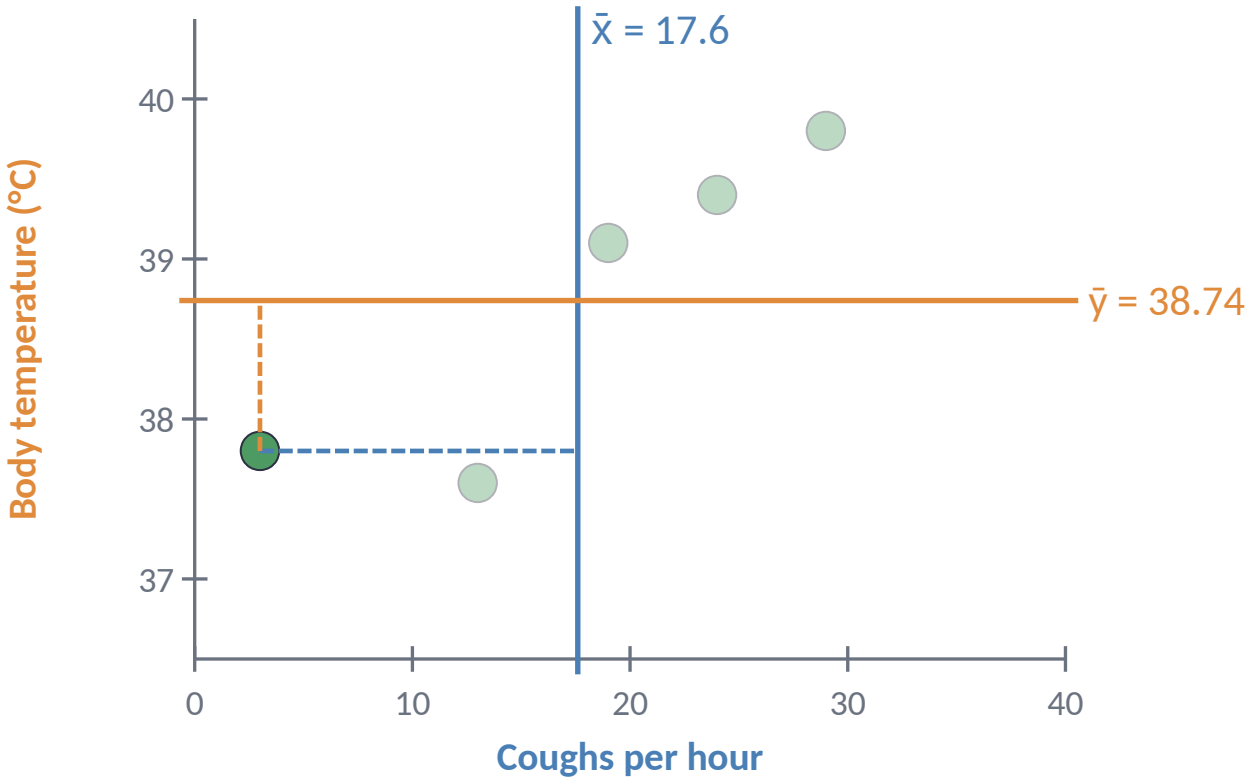
Let us focus on the patient on the far left. They are to the left of the **blue line**, so they coughed *less* than the mean…

…and they are below the **orange line**, so their temperature is *less* than the mean.

Plugging in the numbers, both differences are *negative*…

…so multiplying them together gives a *positive* value: **13.7**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \quad 13.7$$



Working through each patient

Let us focus on the patient on the far left. They are to the left of the **blue line**, so they coughed *less* than the mean…

…and they are below the **orange line**, so their temperature is *less* than the mean.

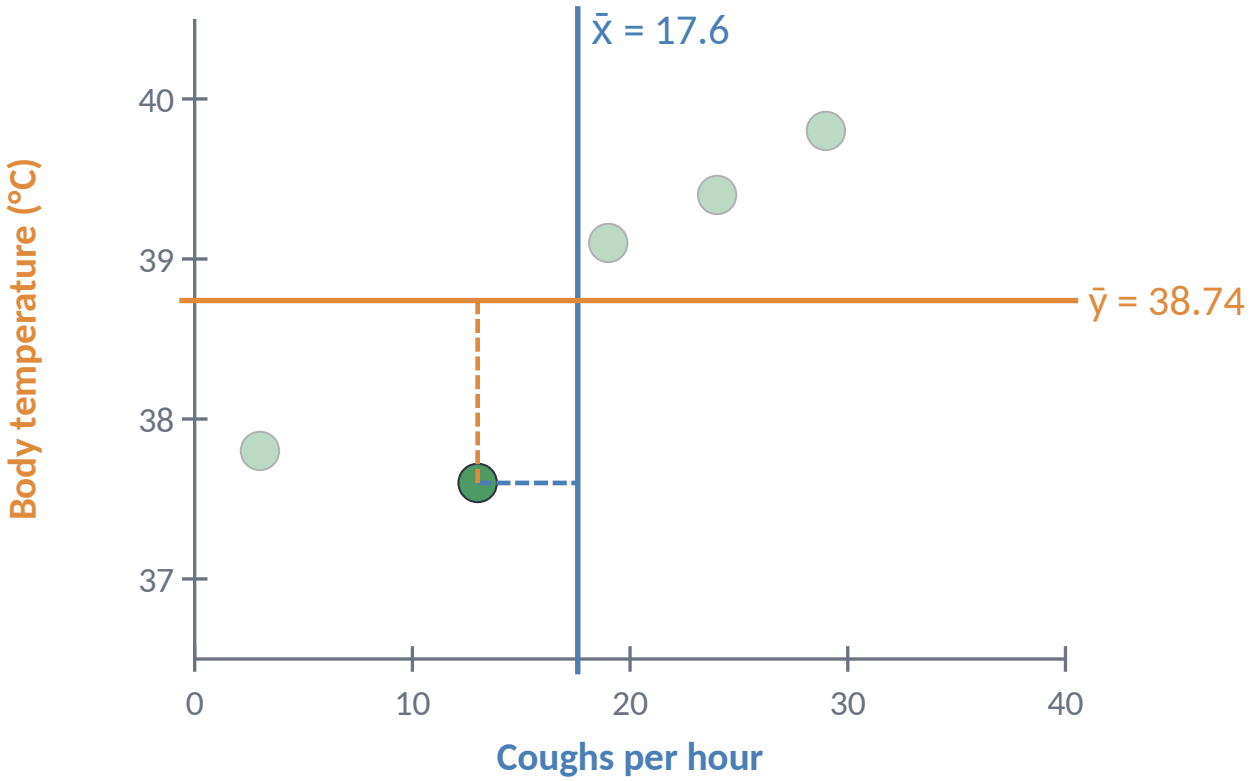
Plugging in the numbers, both differences are *negative*…

…so multiplying them together gives a *positive* value: **13.7**.

We do the same for the next patient. Again, both measurements are less than their means, so both differences are negative…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

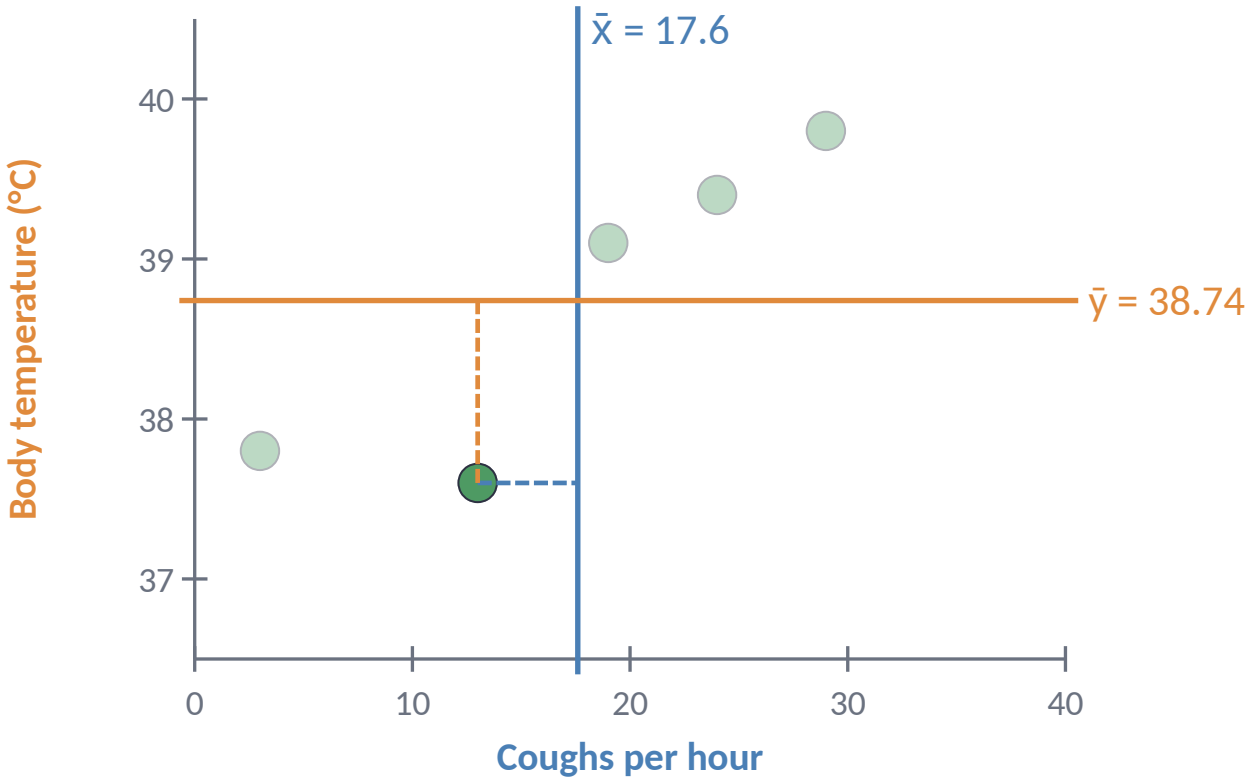
13.7 -4.6×-1.14



Working through each patient

…and their product is *positive*: **5.2**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad 13.7 \qquad 5.2$$

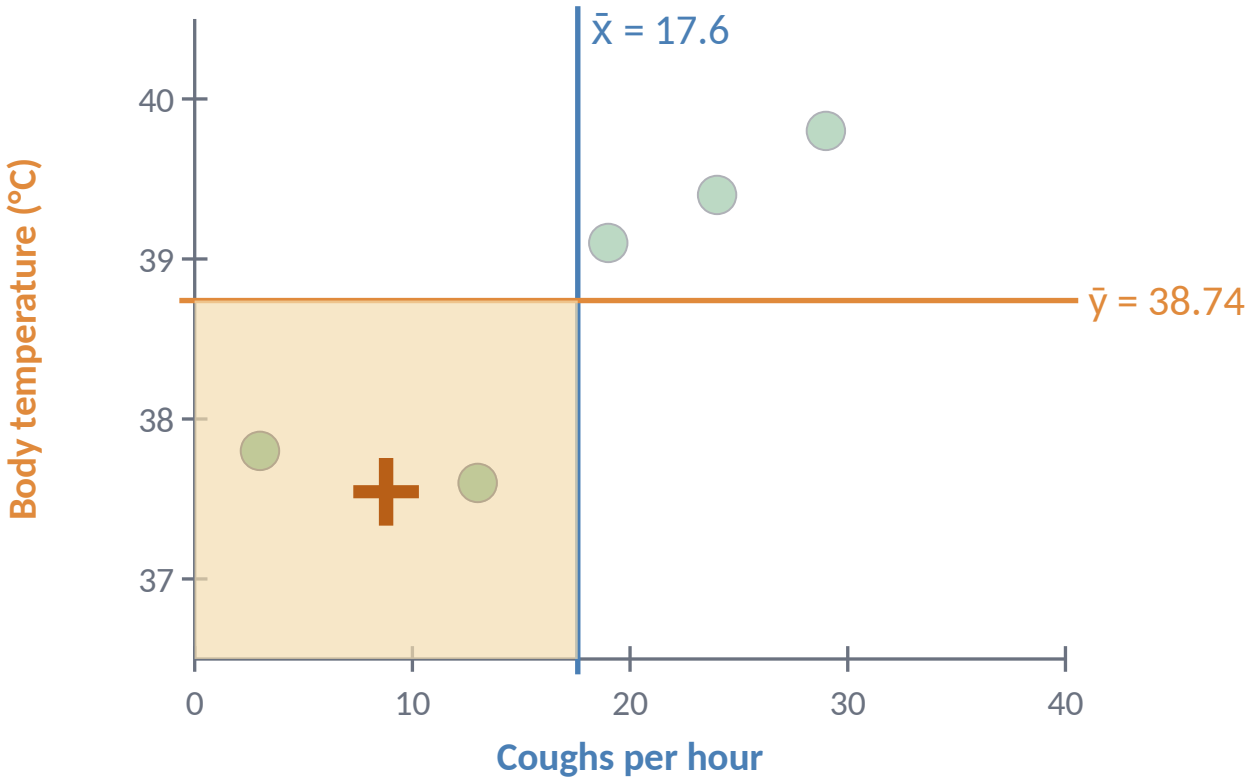


Working through each patient

…and their product is *positive*: **5.2**.

So, when both measurements are *less* than their means, the patient contributes a *positive* value.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad 13.7 \qquad 5.2$$



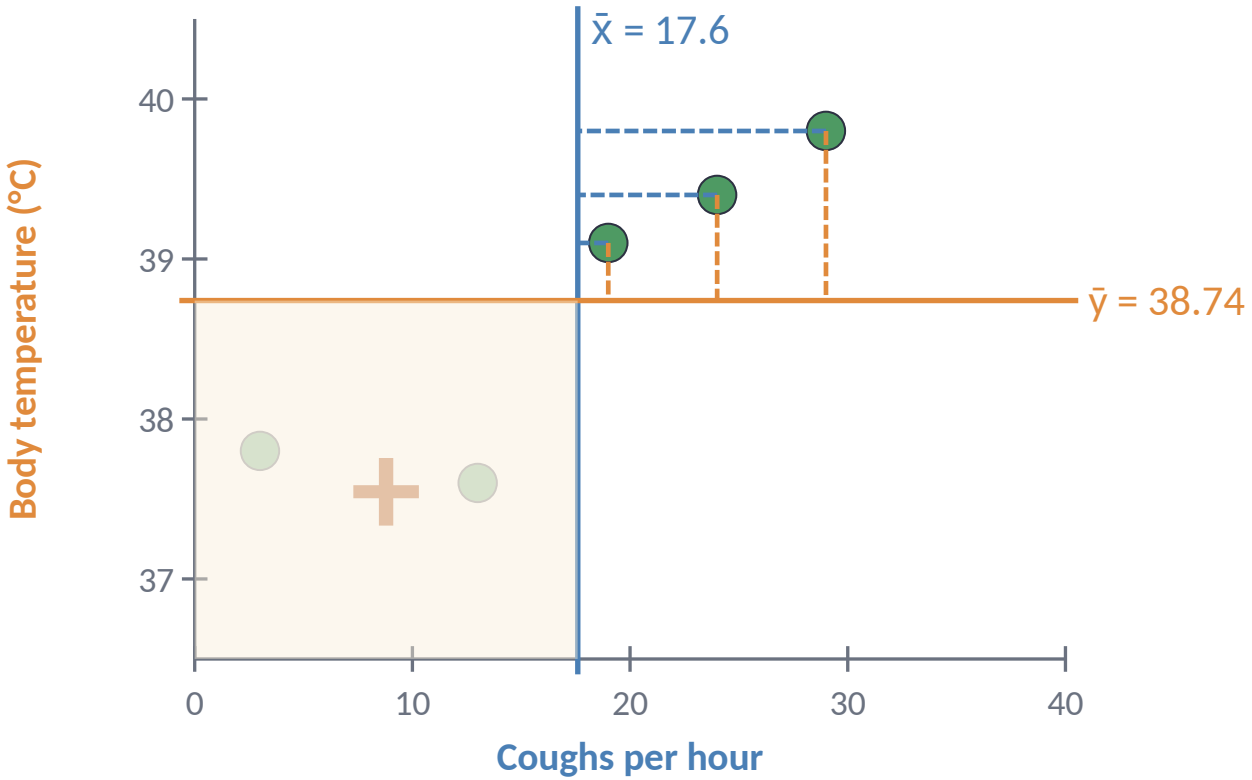
Working through each patient

…and their product is *positive*: **5.2**.

So, when both measurements are *less* than their means, the patient contributes a *positive* value.

The remaining three patients are to the right of the blue line and above the orange line, so both their measurements are *greater* than their means…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad 13.7 \qquad 5.2$$



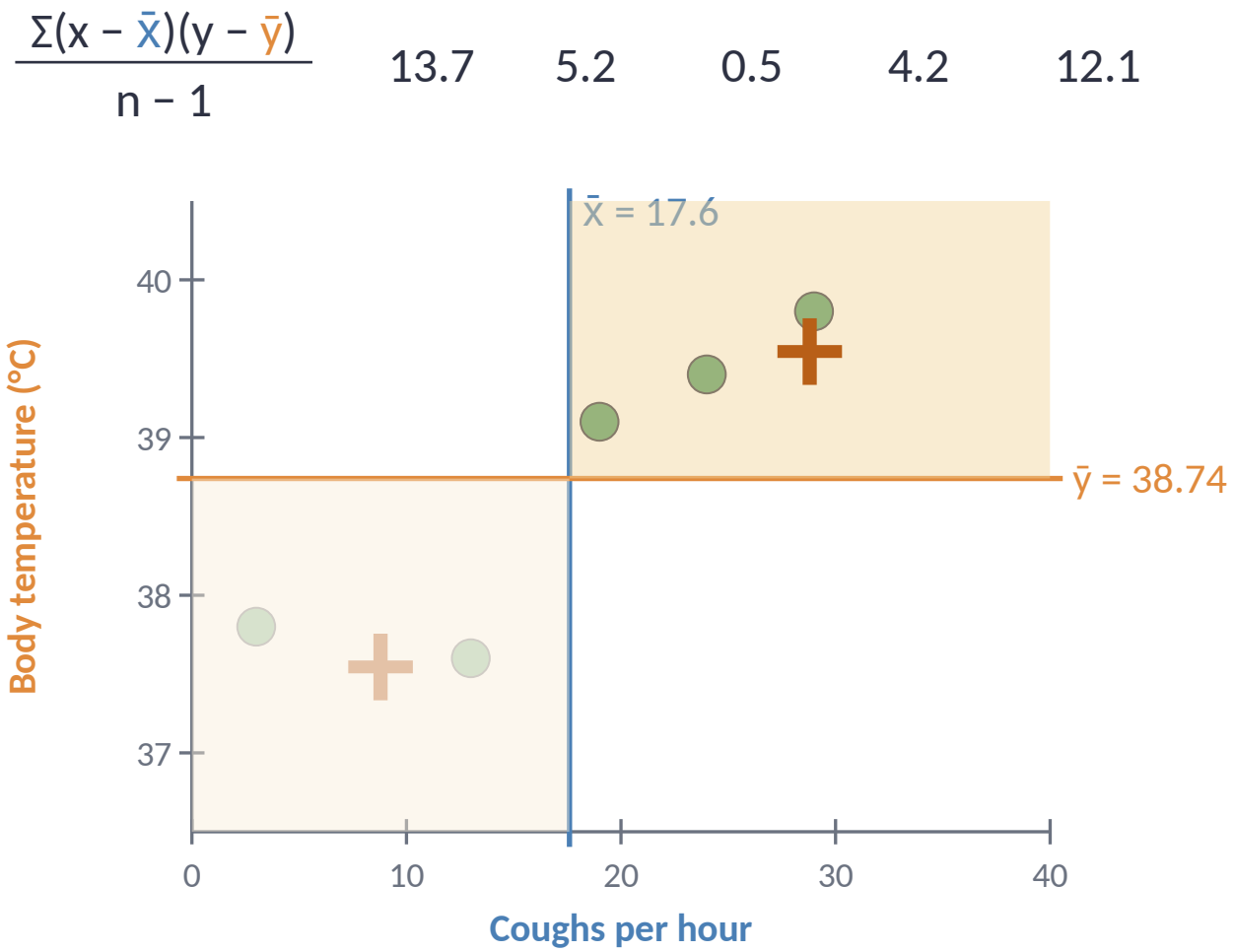
Working through each patient

…and their product is *positive*: **5.2**.

So, when both measurements are *less* than their means, the patient contributes a *positive* value.

The remaining three patients are to the right of the blue line and above the orange line, so both their measurements are *greater* than their means…

…so they also contribute *positive* values. When both measurements are greater than their means, we get positive numbers too.

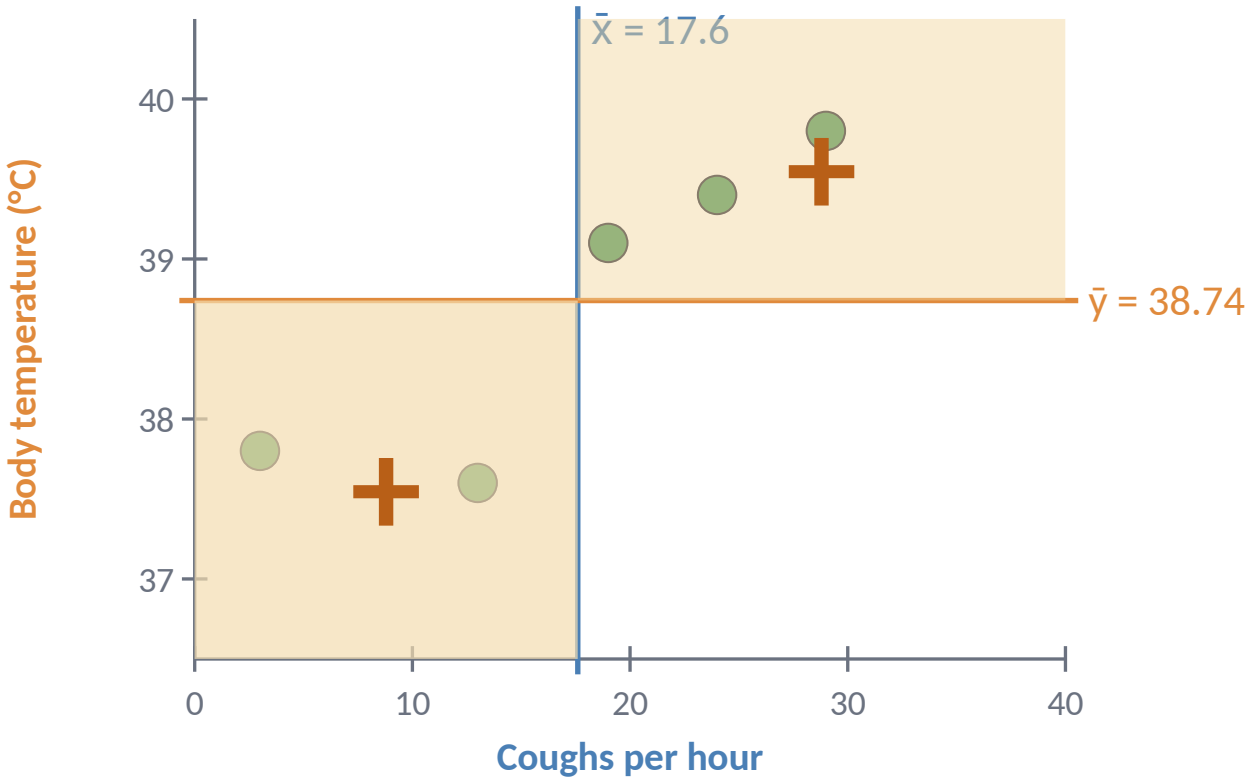


Working through each patient

In summary, the data in these two quadrants contribute **positive** values to the covariance.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

13.7 5.2 0.5 4.2 12.1

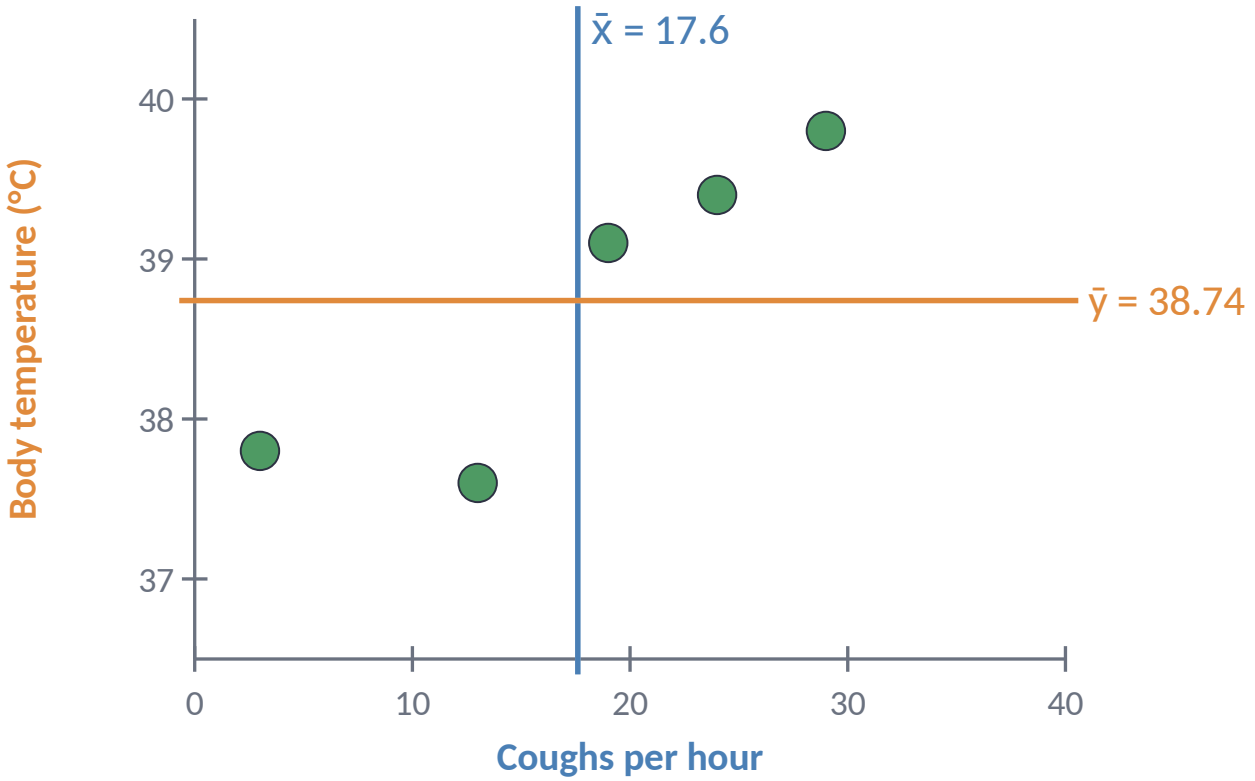


Adding up the terms

The **sigma** tells us to add up each term…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

13.7 5.2 0.5 4.2 12.1



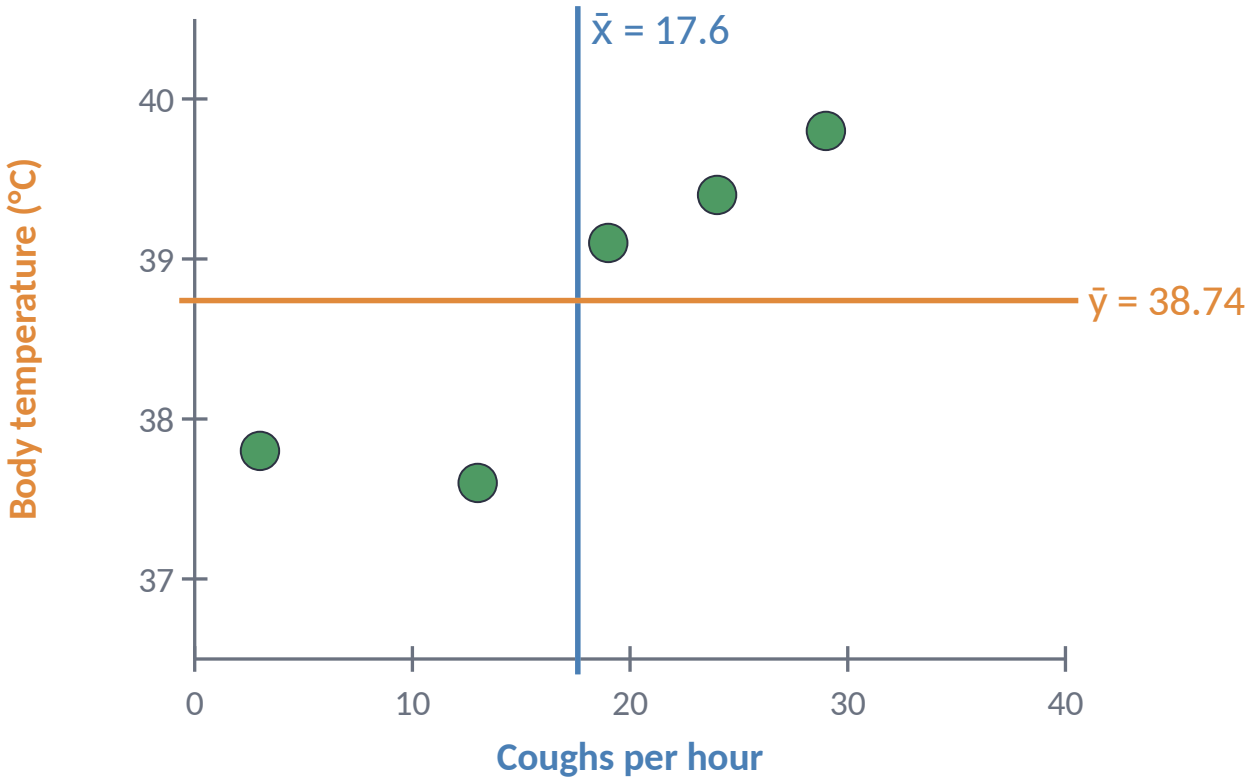
Adding up the terms

The **sigma** tells us to add up each term…

…and then we divide by the number of patients, **n**, minus 1.

So the **covariance** between coughs and temperature is **8.9**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{13.7 + 5.2 + 0.5 + 4.2 + 12.1}{5 - 1} = 8.9$$



Adding up the terms

The **sigma** tells us to add up each term…

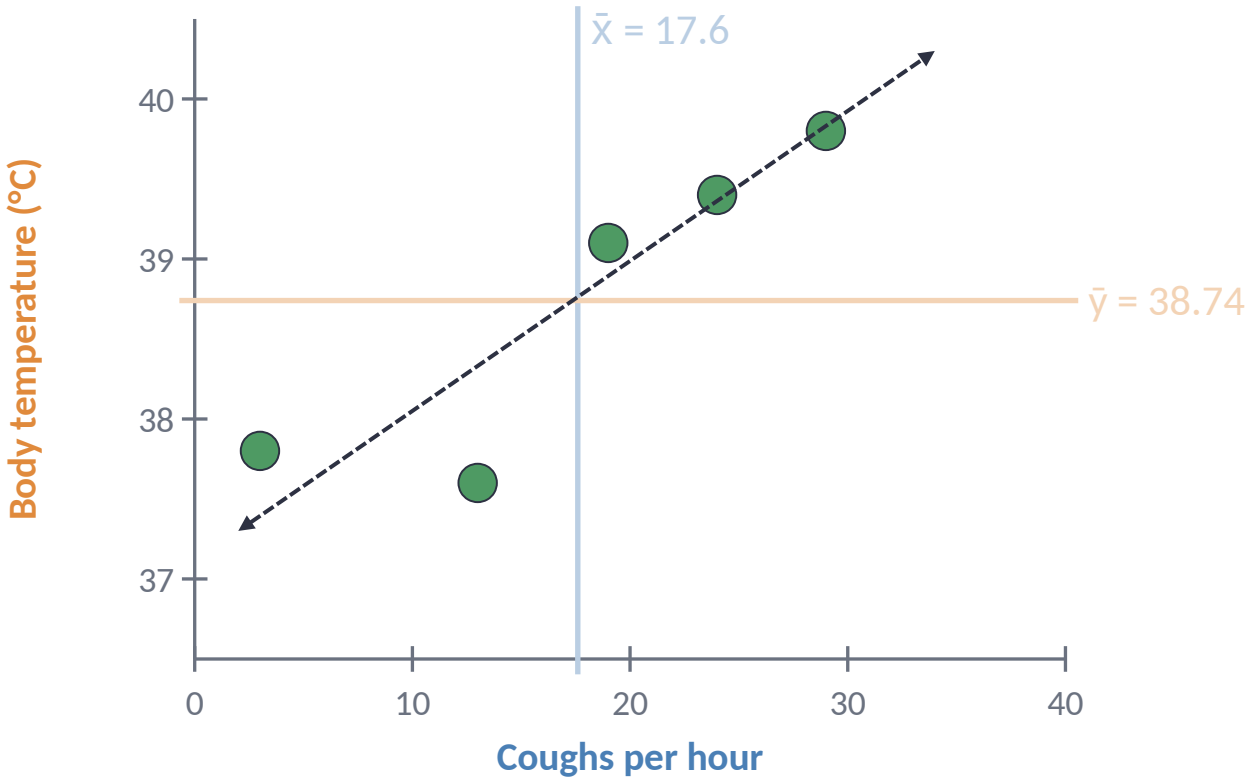
…and then we divide by the number of patients, **n**, minus 1.

So the **covariance** between coughs and temperature is **8.9**.

Since the covariance is *positive*, the relationship between coughs and temperature has a *positive* slope.

Note: the covariance value itself is not very easy to interpret, and it depends on the context. For example, it does not tell us whether the slope is steep…

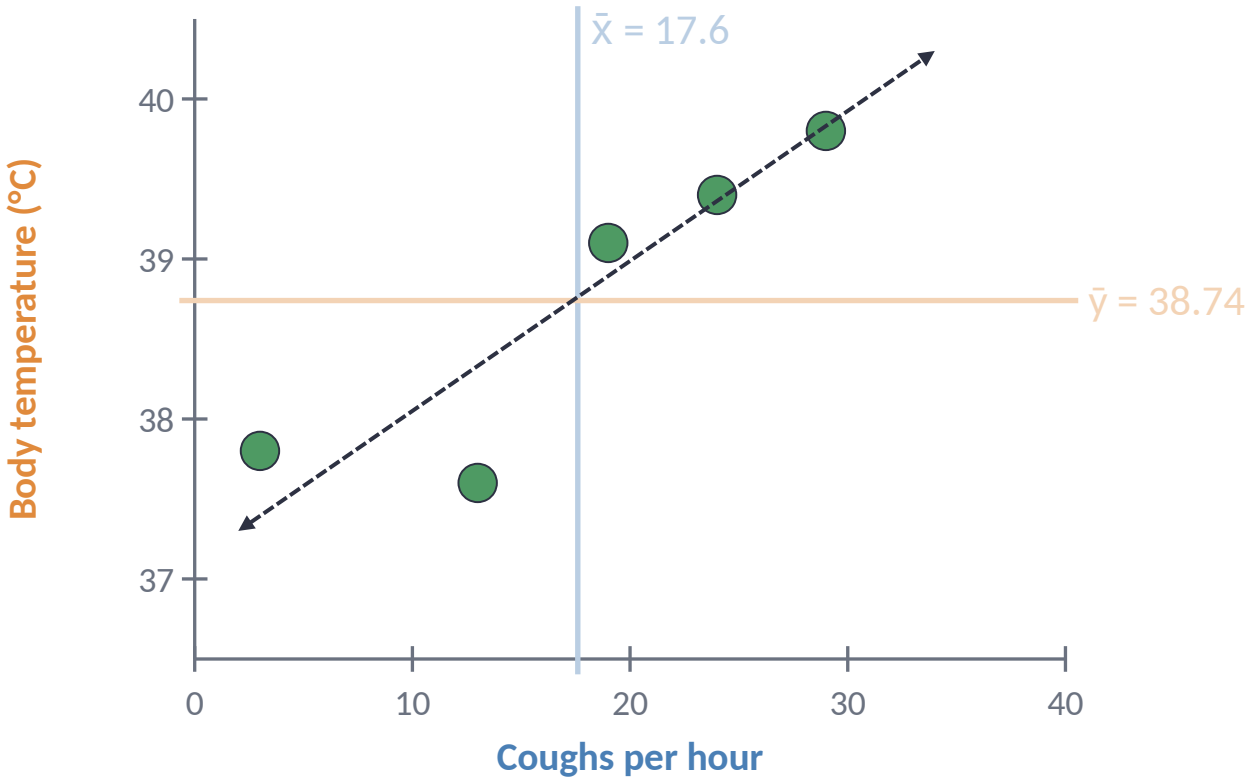
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{13.7 + 5.2 + 0.5 + 4.2 + 12.1}{5 - 1} = 8.9$$



Adding up the terms

…and, more importantly, it does not tell us whether the points are relatively close to the dotted line. We will see why shortly.

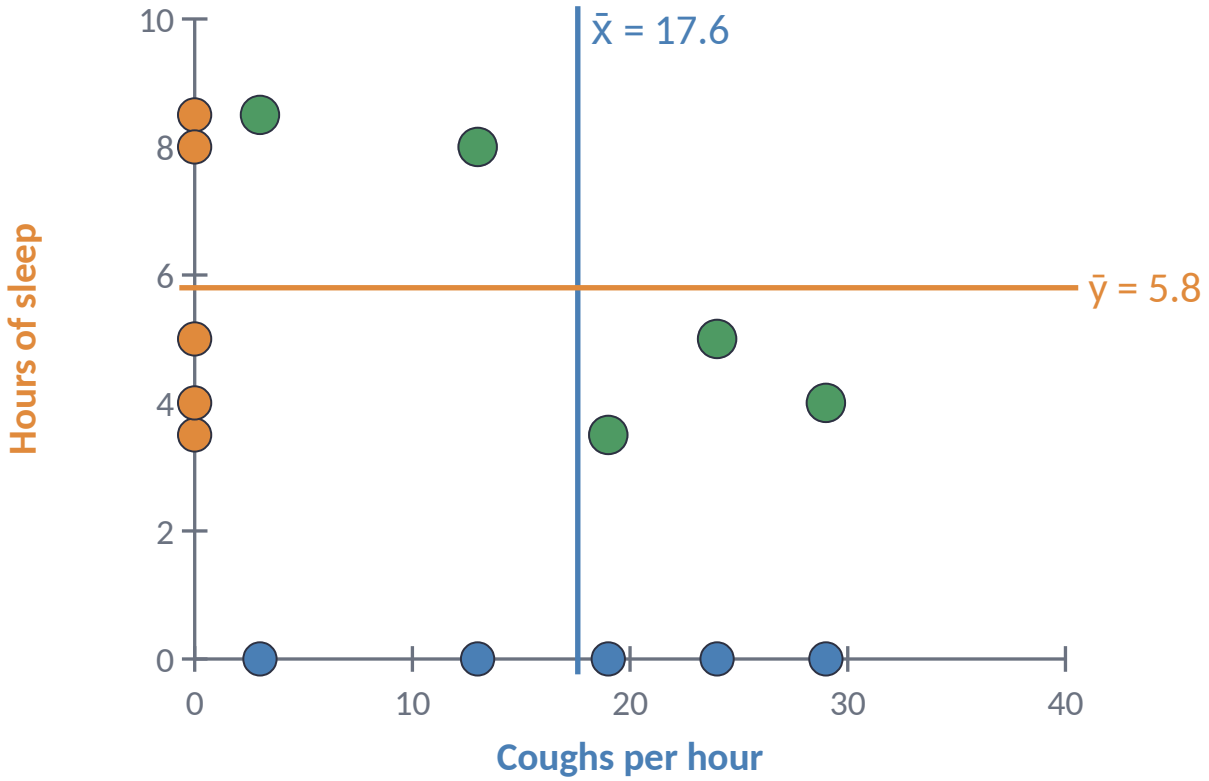
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{13.7 + 5.2 + 0.5 + 4.2 + 12.1}{5 - 1} = 8.9$$



A negative covariance

Now let us calculate the covariance between coughs and sleep. The mean number of coughs is the same as before, **17.6**, and the mean sleep is **5.8** hours.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$



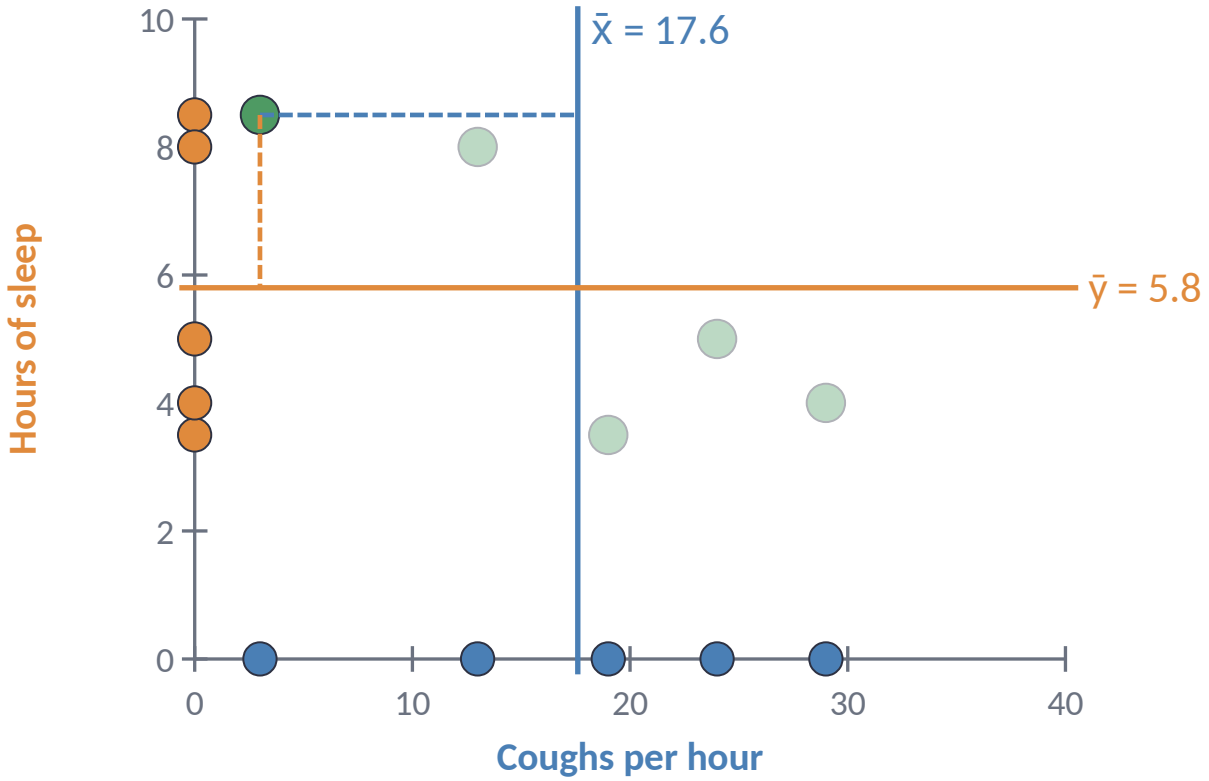
A negative covariance

Now let us calculate the covariance between coughs and sleep. The mean number of coughs is the same as before, **17.6**, and the mean sleep is **5.8** hours.

The patient on the far left coughed *less* than the mean, but slept *more* than the mean…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$

$$(3 - 17.6) \times (8.5 - 5.8) = -14.6 \times 2.7$$



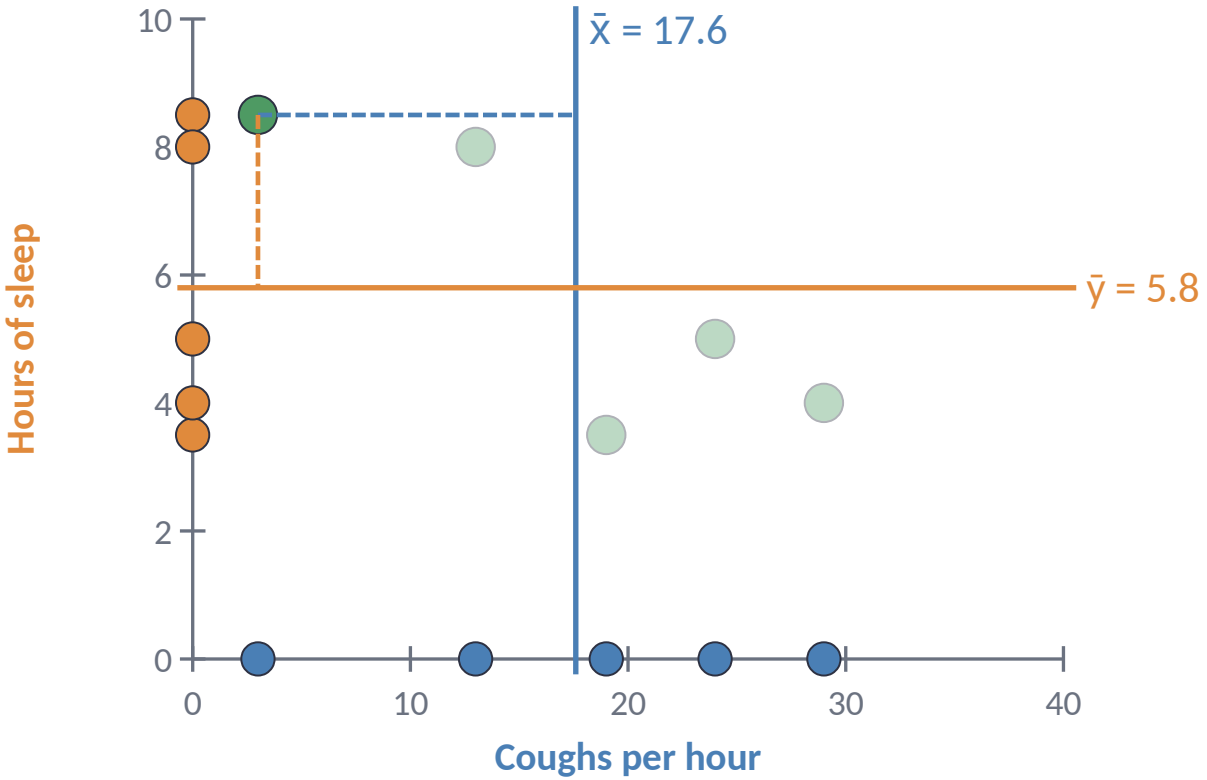
A negative covariance

Now let us calculate the covariance between coughs and sleep. The mean number of coughs is the same as before, **17.6**, and the mean sleep is **5.8** hours.

The patient on the far left coughed *less* than the mean, but slept *more* than the mean…

…so one difference is negative and the other is positive, and their product is *negative*: **−39.4**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \quad -39.4$$



A negative covariance

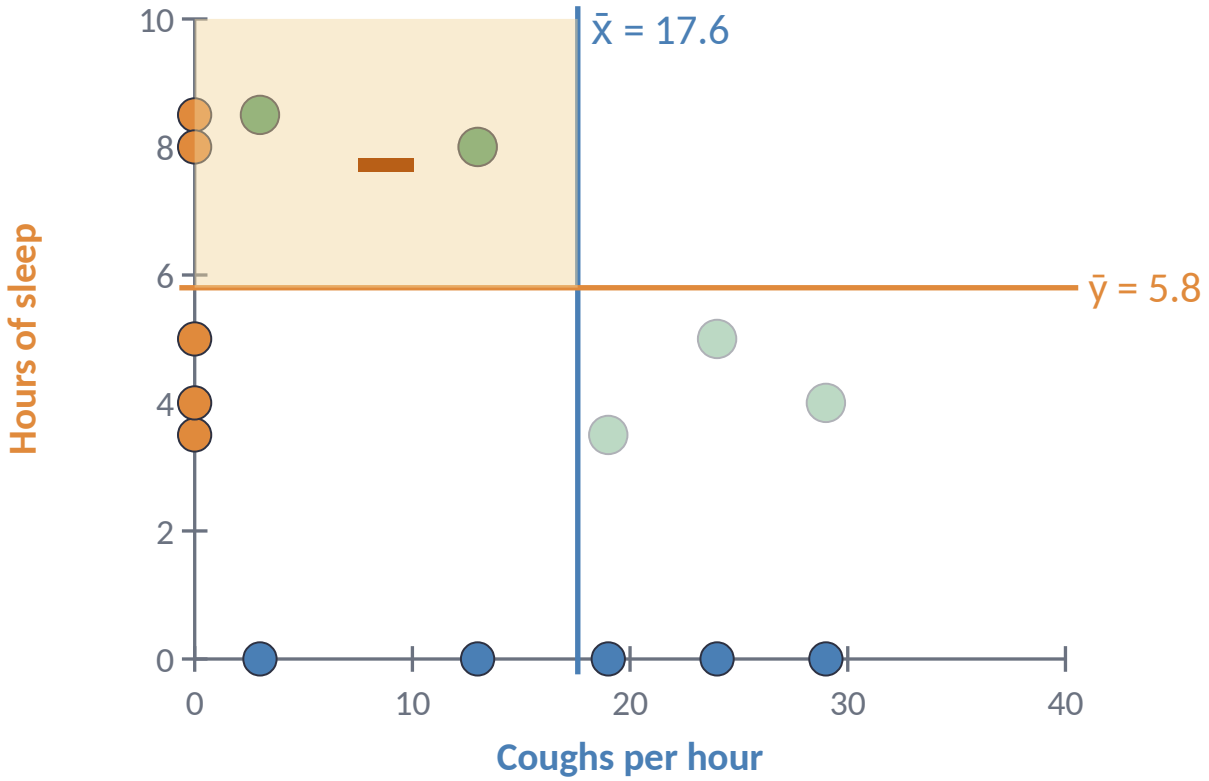
Now let us calculate the covariance between coughs and sleep. The mean number of coughs is the same as before, **17.6**, and the mean sleep is **5.8** hours.

The patient on the far left coughed *less* than the mean, but slept *more* than the mean…

…so one difference is negative and the other is positive, and their product is *negative*: **−39.4**.

The same happens for the next patient: **−10.1**. So if one measurement is less than its mean and the other is more than its mean, we end up with a *negative* number.

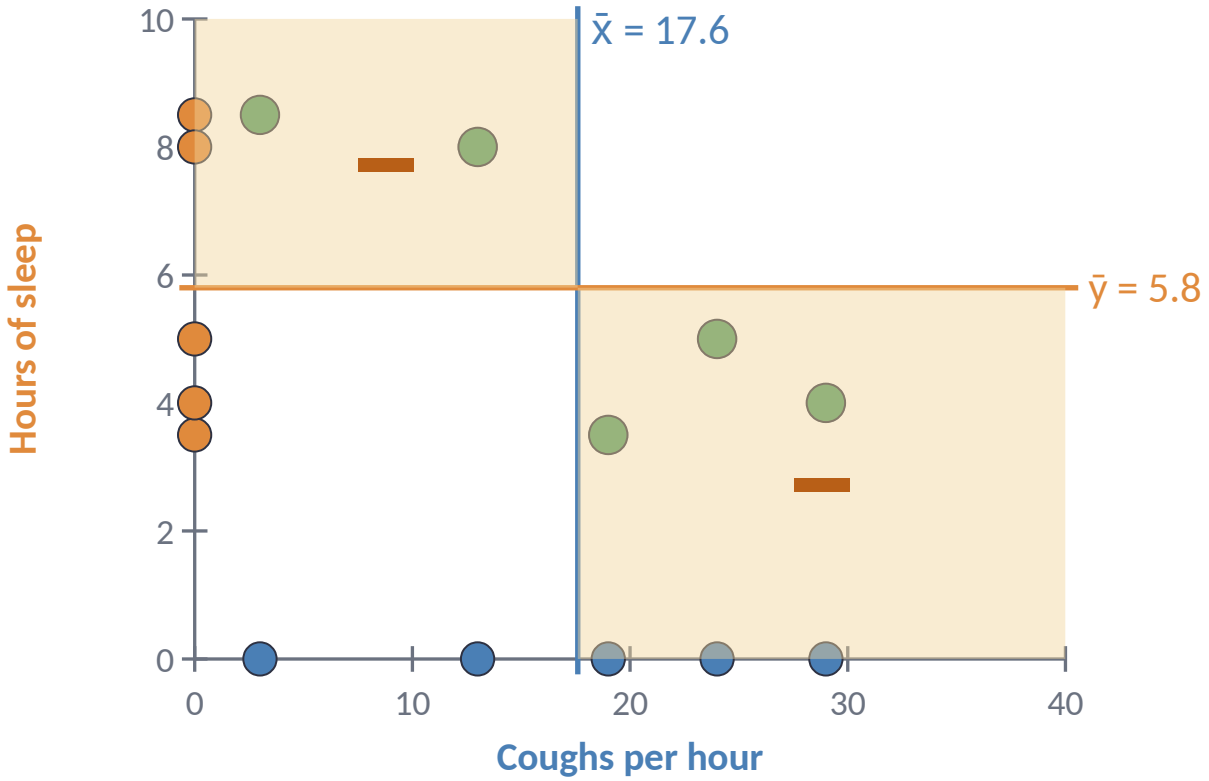
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad -39.4 \qquad -10.1$$



A negative covariance

Likewise, the other three patients coughed *more* than the mean but slept *less*, so they also give negative numbers. Data in these two quadrants contribute **negative** values.

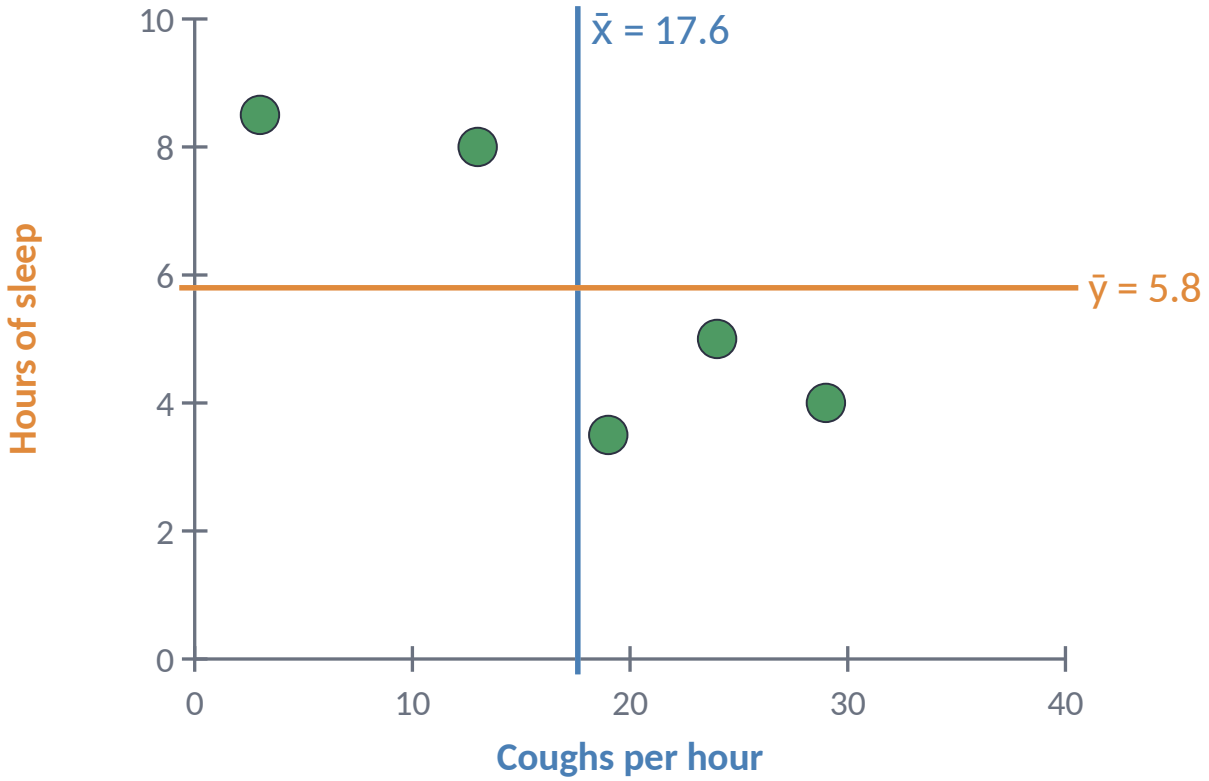
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1}$$ −39.4 −10.1 −3.2 −5.1 −20.5



Adding up the negative terms

Now we add up each term…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad -39.4 \qquad -10.1 \qquad -3.2 \qquad -5.1 \qquad -20.5$$

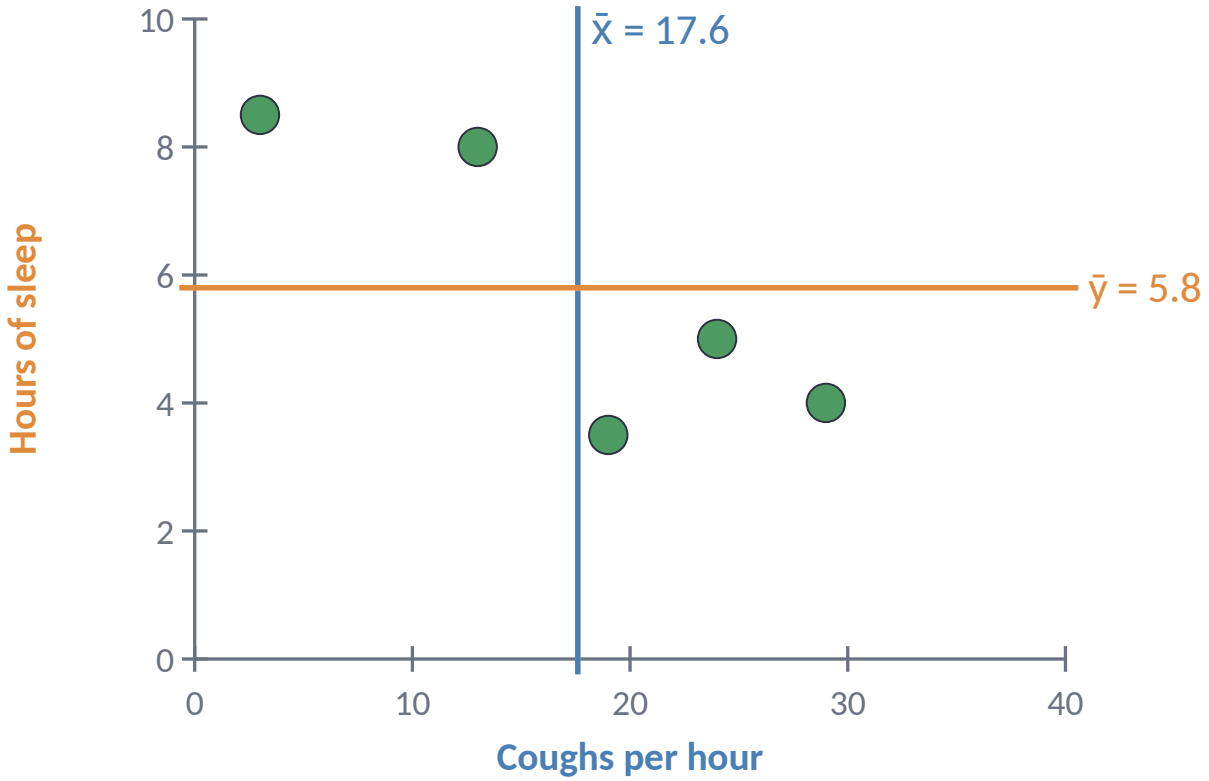


Adding up the negative terms

Now we add up each term…

…and divide by 5 − 1. The **covariance** between coughs and sleep is **−19.6**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{(-39.4) + (-10.1) + (-3.2) + (-5.1) + (-20.5)}{5 - 1} = \boxed{-19.6}$$



Adding up the negative terms

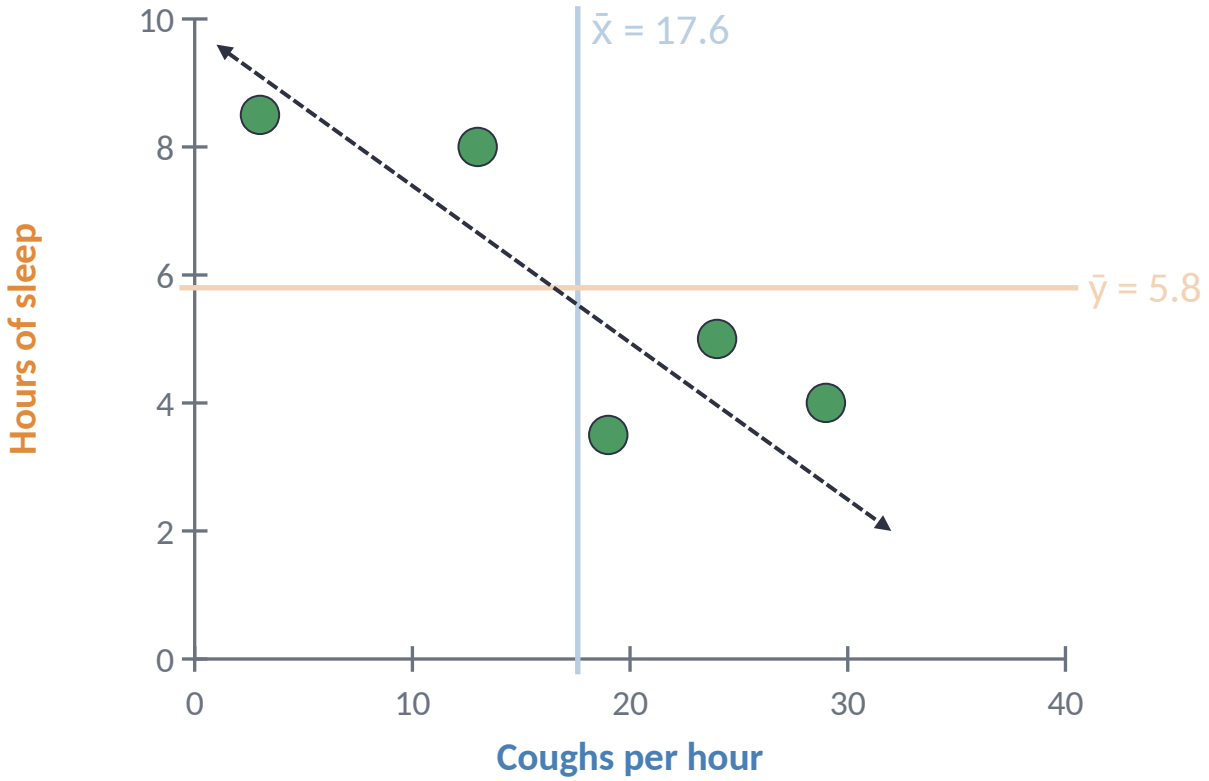
Now we add up each term…

…and divide by 5 − 1. The **covariance** between coughs and sleep is **−19.6**.

Since the covariance is *negative*, the relationship between coughs and sleep has a *negative* slope.

Again, the covariance does not tell us whether the slope is steep or not, or whether the points are close to the line: just that the slope is *negative*.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{(-39.4) + (-10.1) + (-3.2) + (-5.1) + (-20.5)}{5 - 1} = \boxed{-19.6}$$

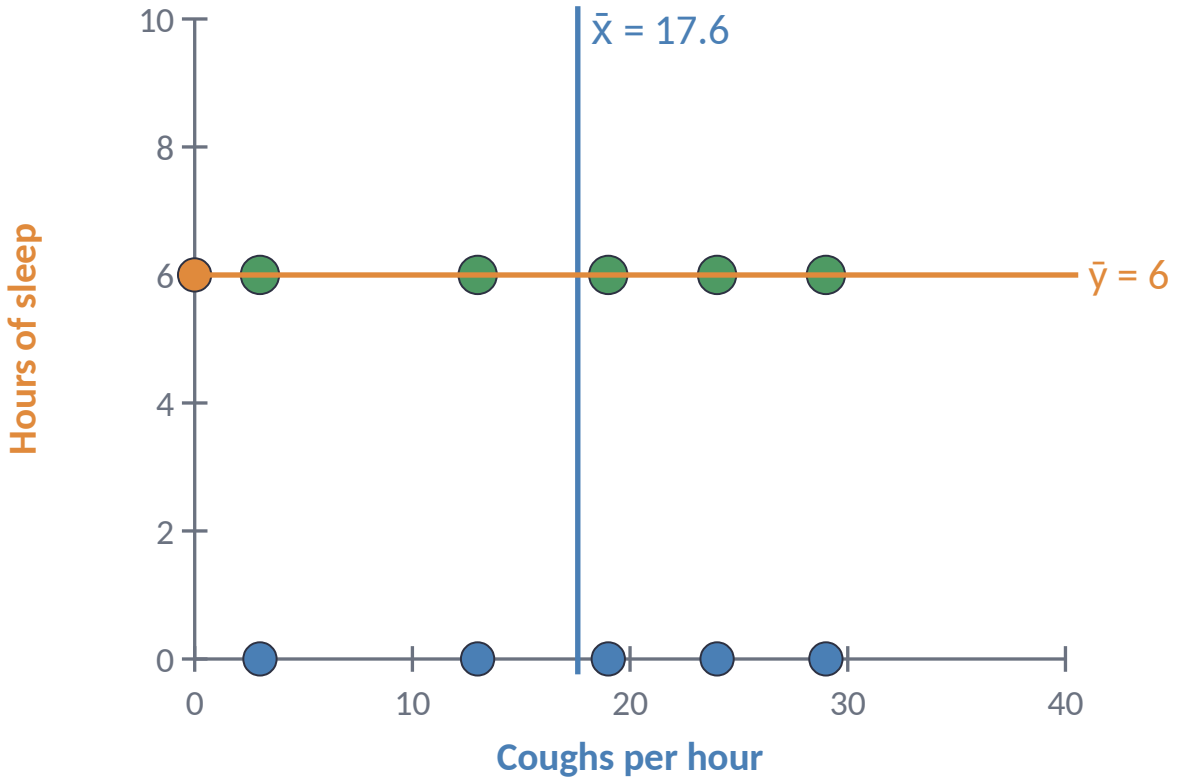


A covariance of zero

Now let us go back to the case where every patient slept exactly **6** hours. First, we calculate the means…

…and, since every patient slept exactly the mean, each difference in sleep is **0**. Anything multiplied by 0 is 0…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \qquad -14.6 \times 0$$



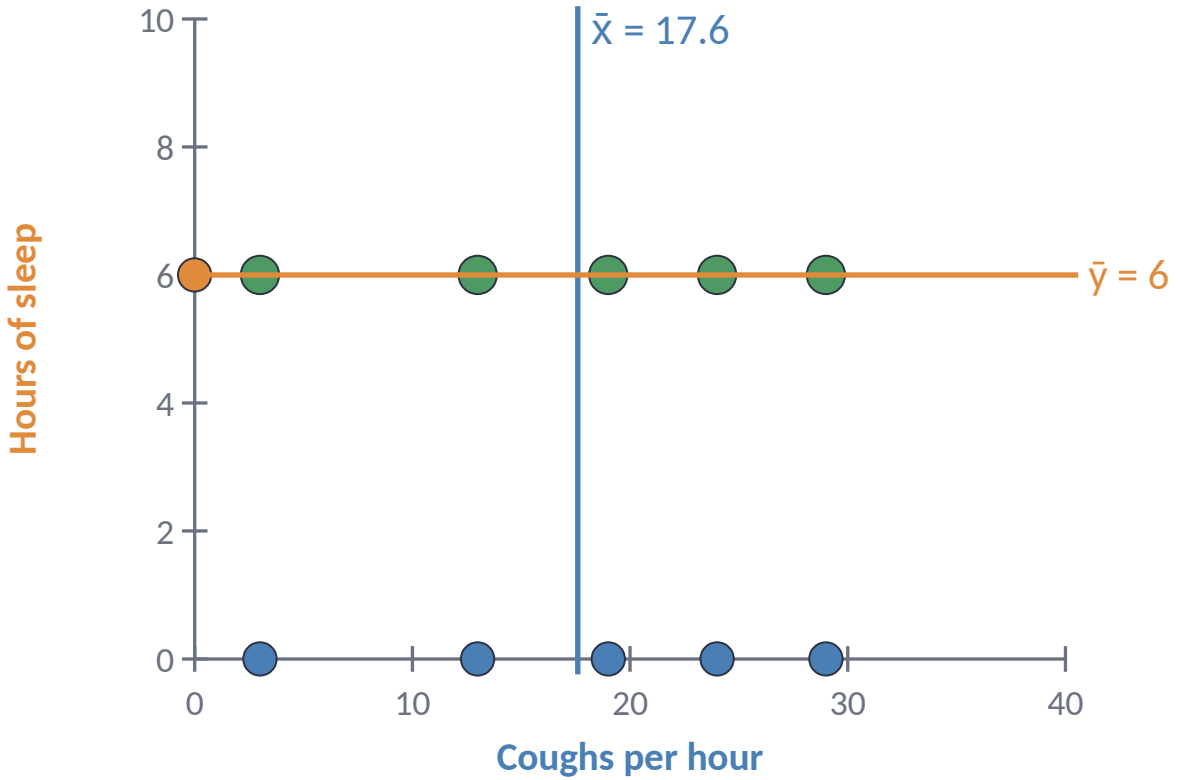
A covariance of zero

Now let us go back to the case where every patient slept exactly **6** hours. First, we calculate the means…

…and, since every patient slept exactly the mean, each difference in sleep is **0**. Anything multiplied by 0 is 0…

…so the numerator is 0, and the whole thing equals **0**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{0 + 0 + 0 + 0 + 0}{5 - 1} = \mathbf{0}$$



A covariance of zero

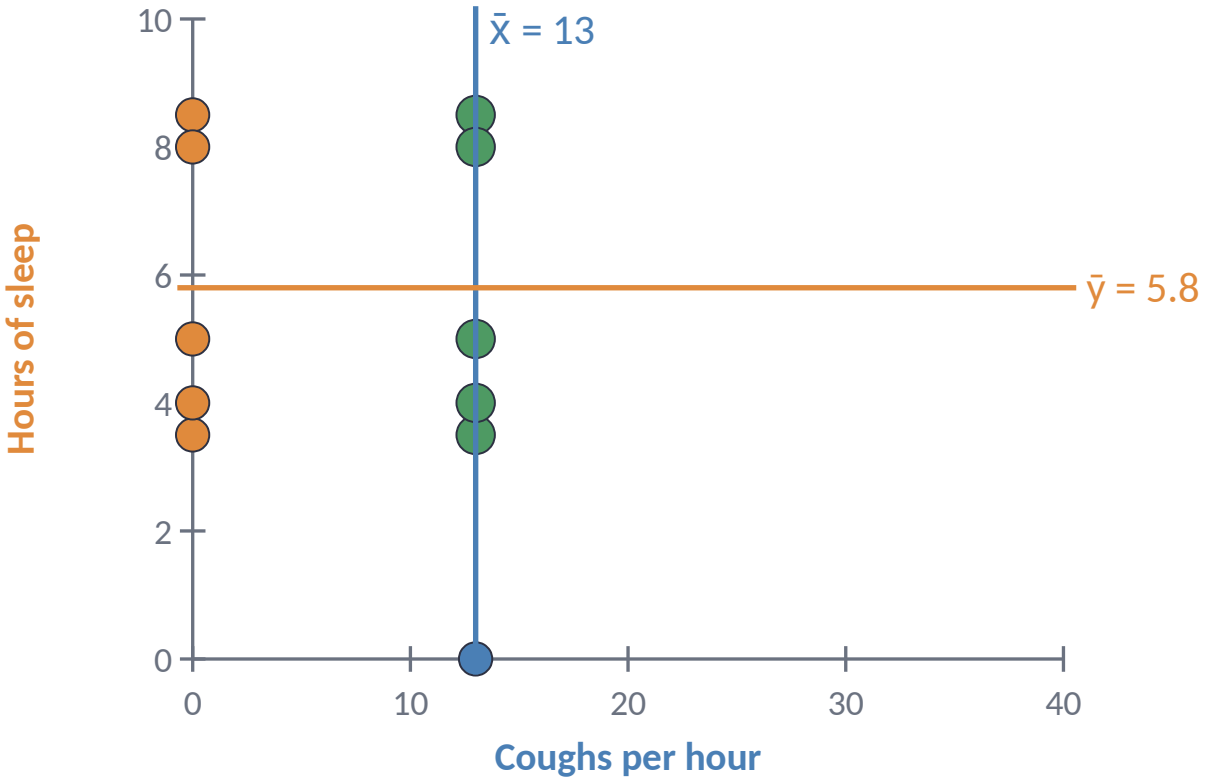
Now let us go back to the case where every patient slept exactly **6** hours. First, we calculate the means…

…and, since every patient slept exactly the mean, each difference in sleep is **0**. Anything multiplied by 0 is 0…

…so the numerator is 0, and the whole thing equals **0**.

Likewise, when every patient coughed the same number of times, every difference in coughs is 0, and the covariance is **0**.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \quad 0 \quad 0 \quad 0 \quad 0 \quad 0$$



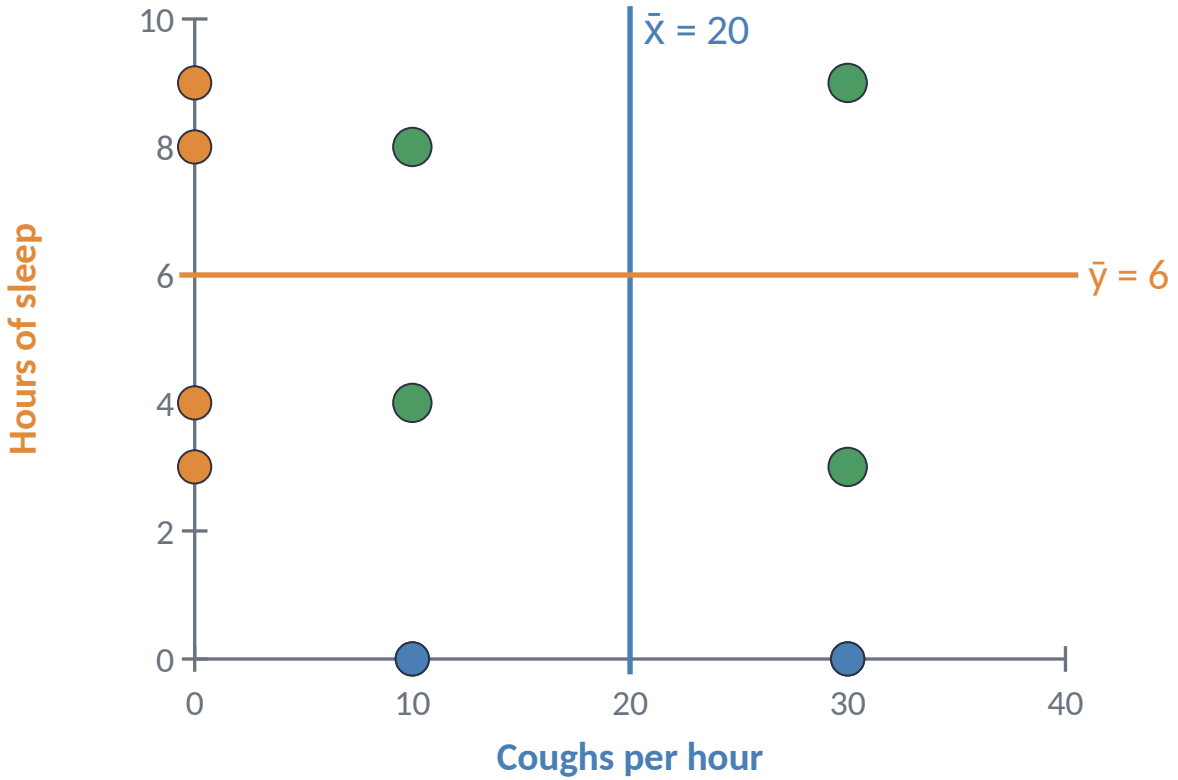
A covariance of zero

In this last case, there are several values for both measurements, but there is still no trend: as coughs increase, sleep goes both up *and* down.

The negative and positive terms cancel out, so the covariance is again **0**.

So **covariance = 0** when there is no relationship between the two measurements.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} \quad -20 \quad 20 \quad 30 \quad -30$$



The covariance of a measurement with itself

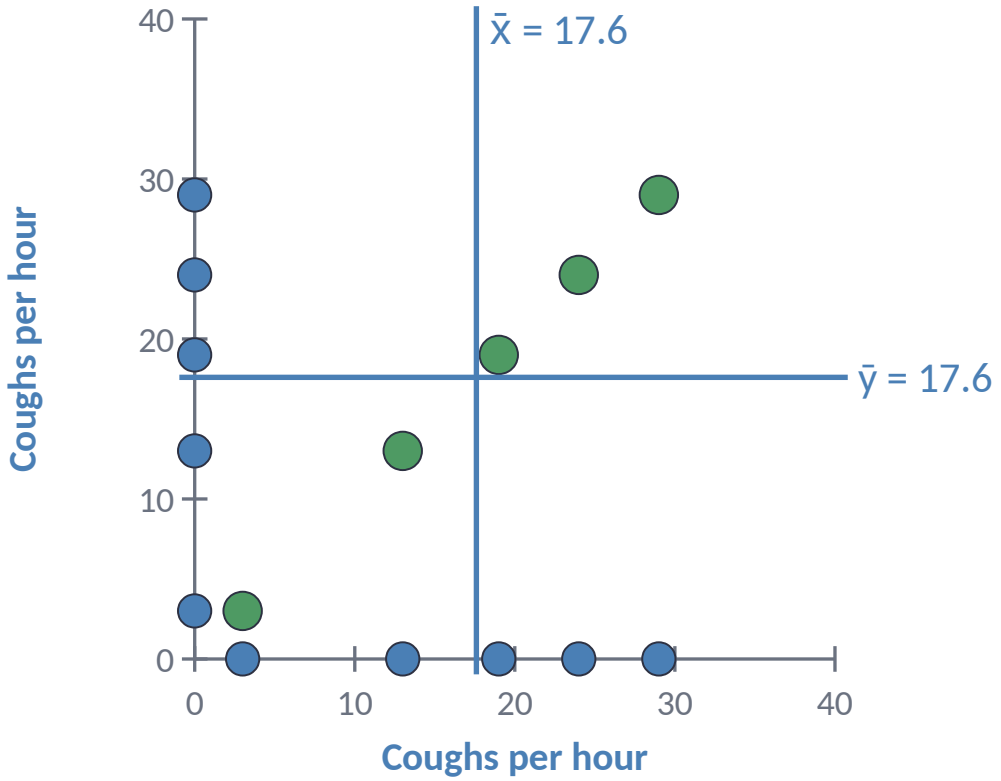
Now let us see why the covariance value is hard to interpret. We go back to looking at the coughs alone…

…and plot them against themselves, using the same data on both the x-axis and the y-axis.

Now we are ready to calculate the covariance. In this case, the same data are on both axes, which means that **y = x** and $\bar{y} = \bar{x}$…

…so we can replace y with x, and $\bar{y}$ with $\bar{x}$. Then we can multiply $(x - \bar{x})(x - \bar{x})$…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\Sigma(x - \bar{x})(x - \bar{x})}{n - 1}$$

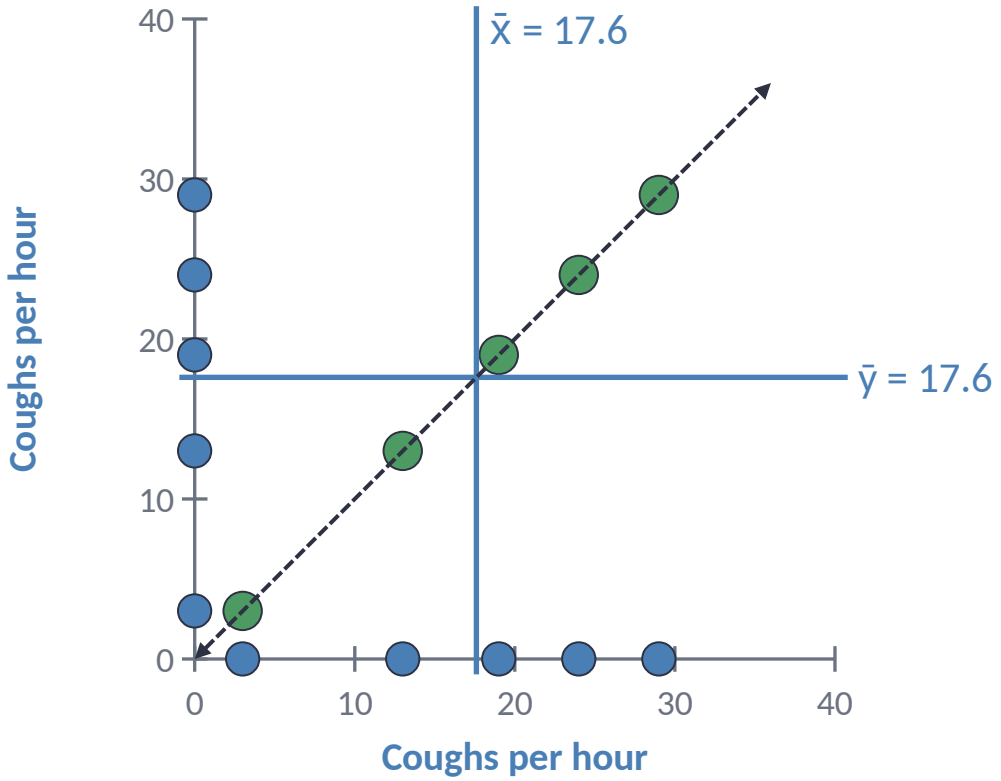


The covariance of a measurement with itself

…and we are left with the formula for the **estimated variance**. In other words, the covariance of a measurement with itself is the same thing as its estimated variance.

Doing the maths, we get **101.8**. Since the covariance is *positive*, the relationship between coughs and themselves has a *positive* slope.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\Sigma(x - \bar{x})(x - \bar{x})}{n - 1} = \boxed{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \boxed{101.8}$$

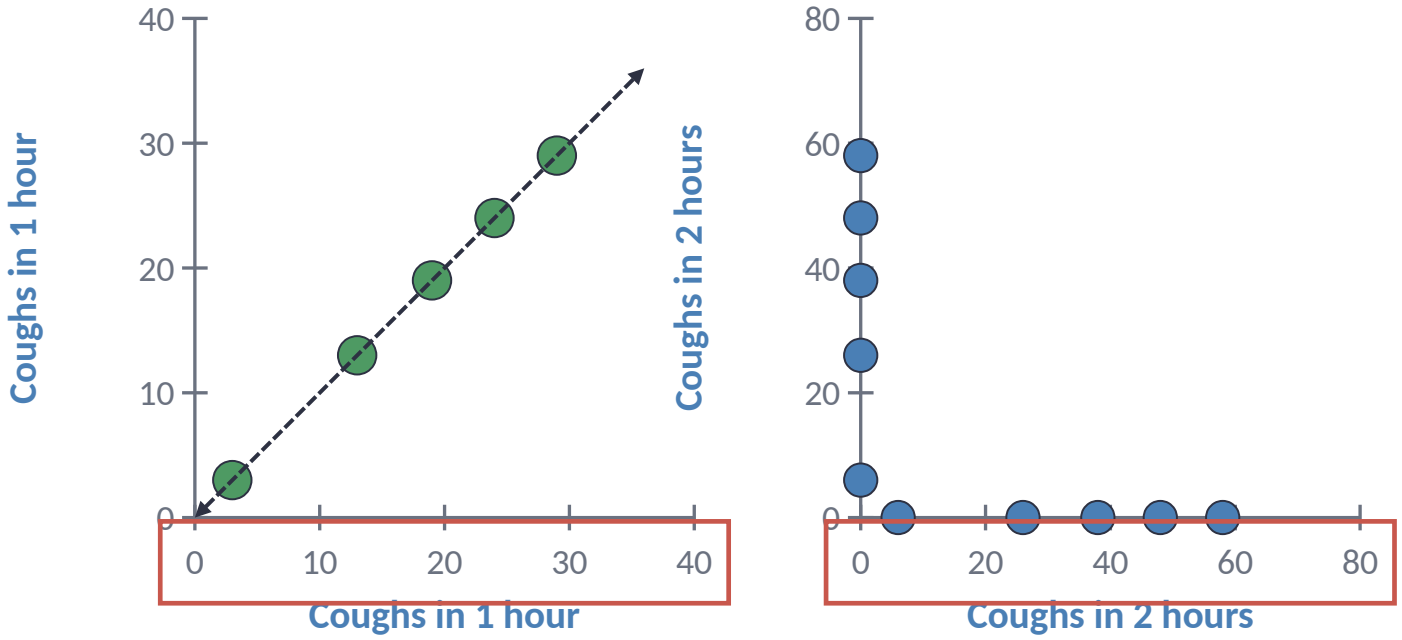


Covariance depends on the scale

Let us move the graph and the covariance value over here, and see what happens when we count coughs over **two hours** instead of one. Every value doubles.

Now the values on the x-axis and the y-axis on the right are twice what they are on the left…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$



Covariance depends on the scale

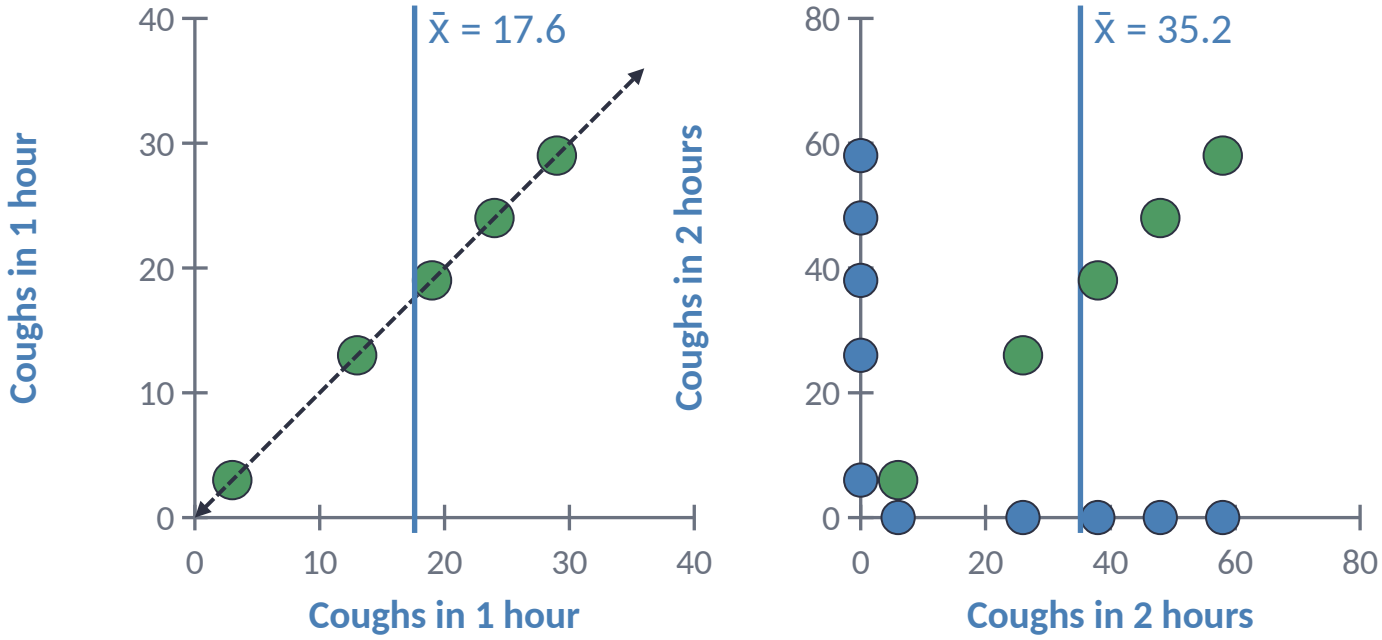
Let us move the graph and the covariance value over here, and see what happens when we count coughs over **two hours** instead of one. Every value doubles.

Now the values on the x-axis and the y-axis on the right are twice what they are on the left…

…and the new means are twice what they were before: **35.2** instead of **17.6**…

…but the relative positions of the data did not change…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$



Covariance depends on the scale

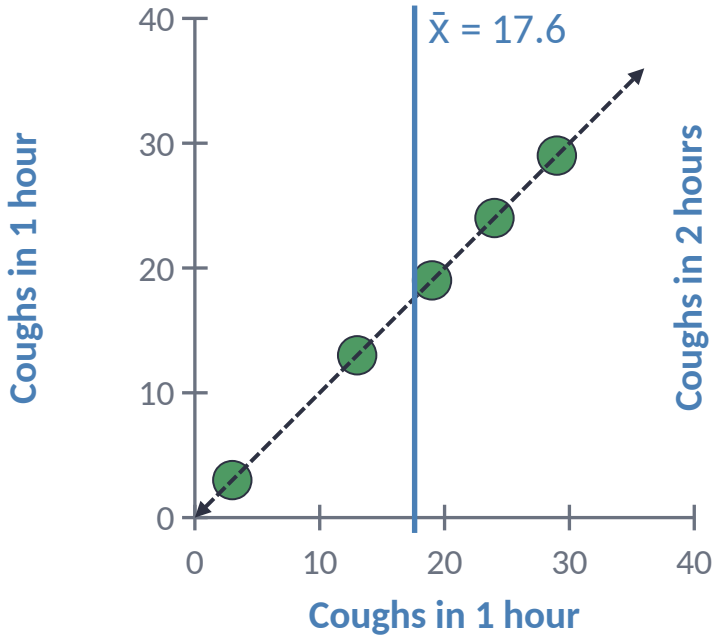
…and each dot still falls on the same straight line with a positive slope. The only thing that changed was the *scale* of the data.

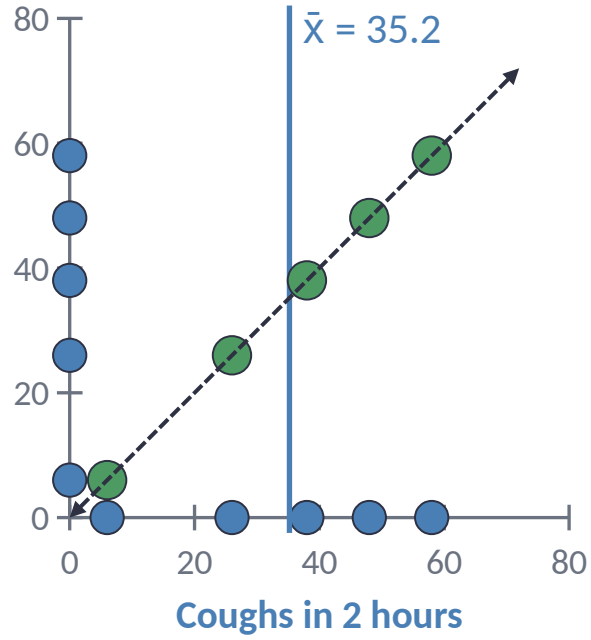
However, when we do the maths, the covariance is **407.2**, which is **4 times** what we got before.

So the covariance value changes even when the *relationship* does not. **Covariance values are sensitive to the scale of the data**, and this makes them difficult to interpret.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{407.2}}$$





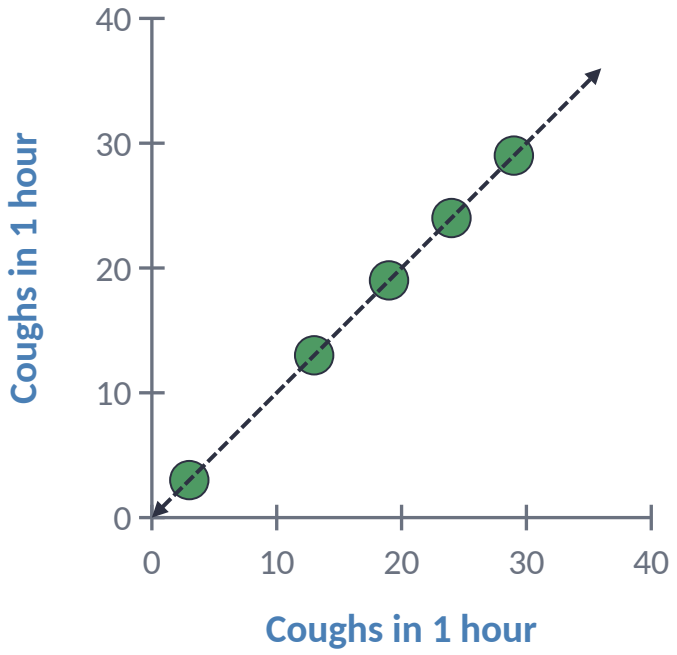
Scale also hides how close points are to the line

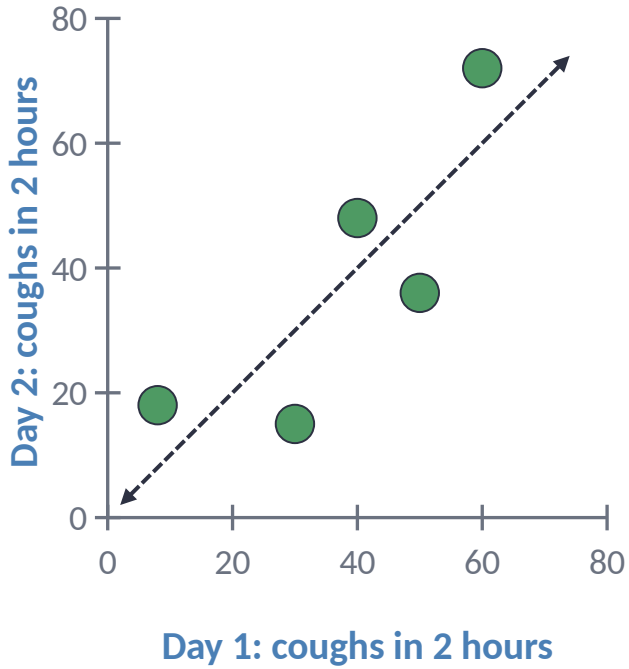
The sensitivity to scale also prevents the covariance value from telling us whether the data are close to the dotted line that represents the relationship…

…or far from it. On the right, five other patients had their coughs counted over two hours on two consecutive days.

In this example, the covariance on the left, where each point is on the dotted line, is **101.8**…

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{101.8}$$





Scale also hides how close points are to the line

The sensitivity to scale also prevents the covariance value from telling us whether the data are close to the dotted line that represents the relationship…

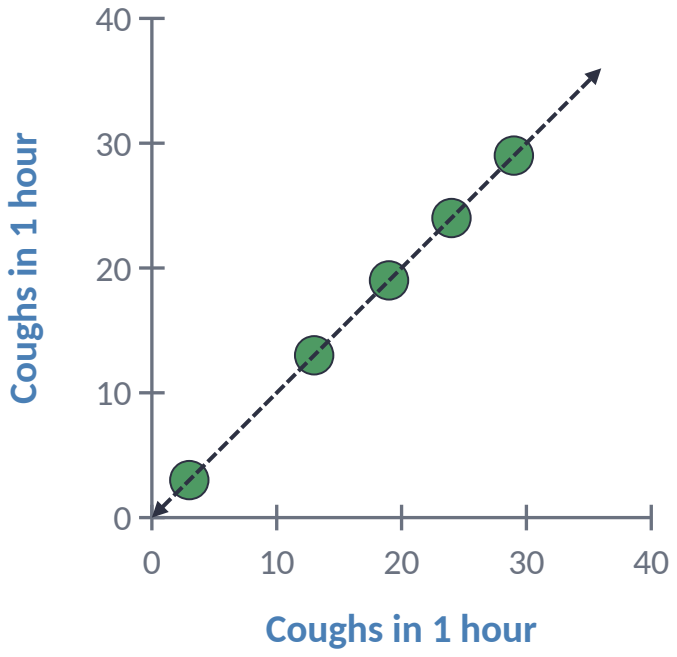
…or far from it. On the right, five other patients had their coughs counted over two hours on two consecutive days.

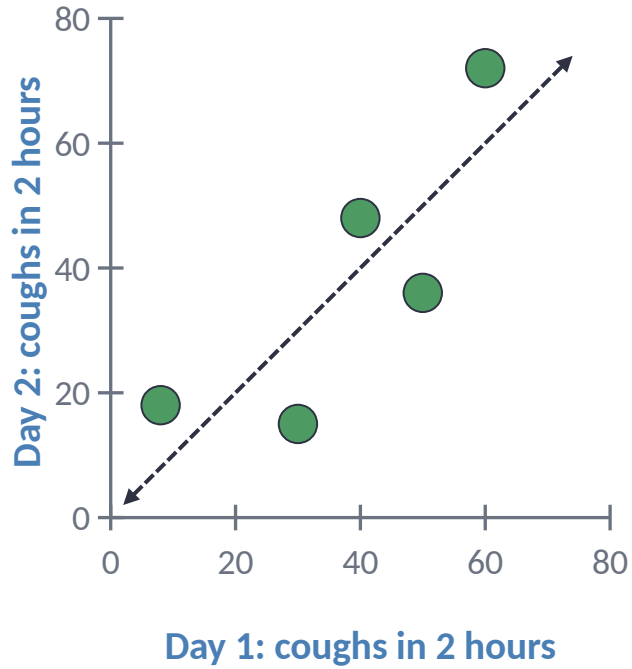
In this example, the covariance on the left, where each point is on the dotted line, is **101.8**…

…and the covariance on the right, where the data are relatively far from the line, is **381.9**. So, in this case, the data far from the line have the *larger* covariance.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{381.9}}$$

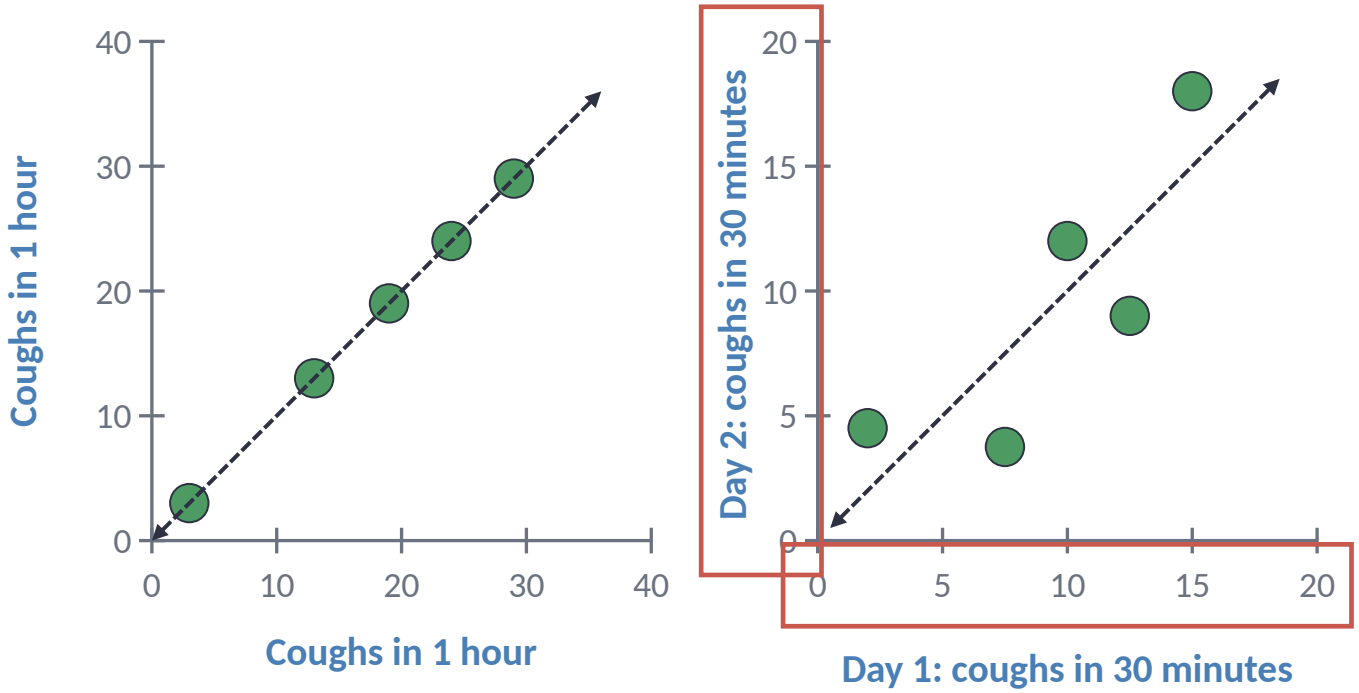




Scale also hides how close points are to the line

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

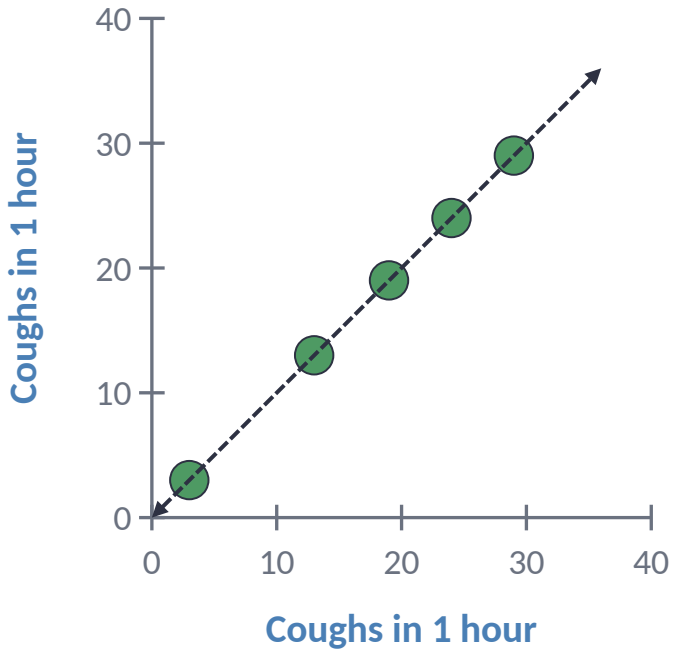
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{23.9}}$$

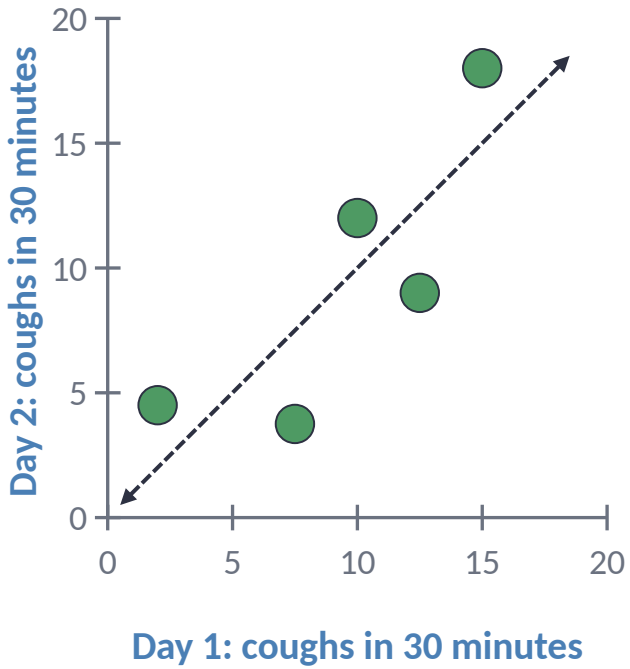


Scale also hides how close points are to the line

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

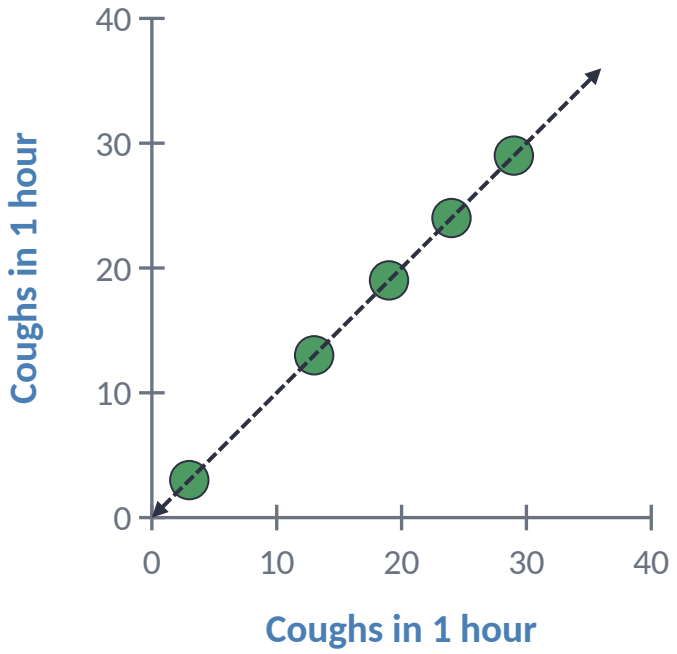
$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{23.9}}$$

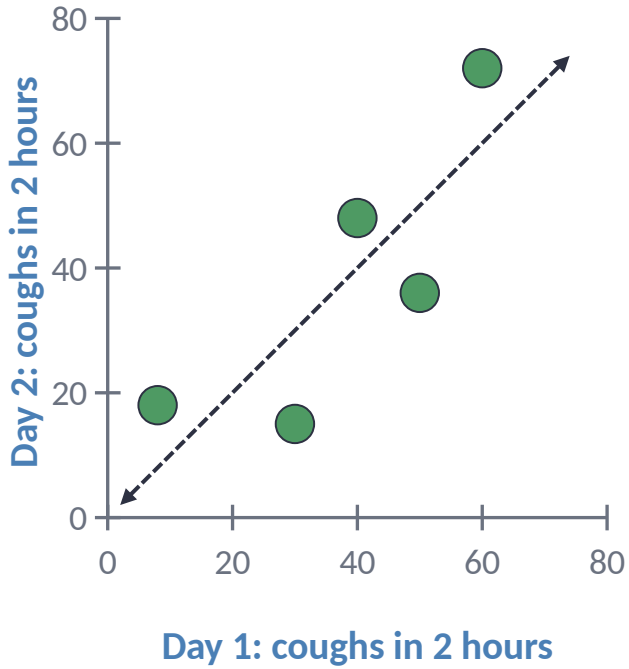




Scale also hides how close points are to the line

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{101.8}$$





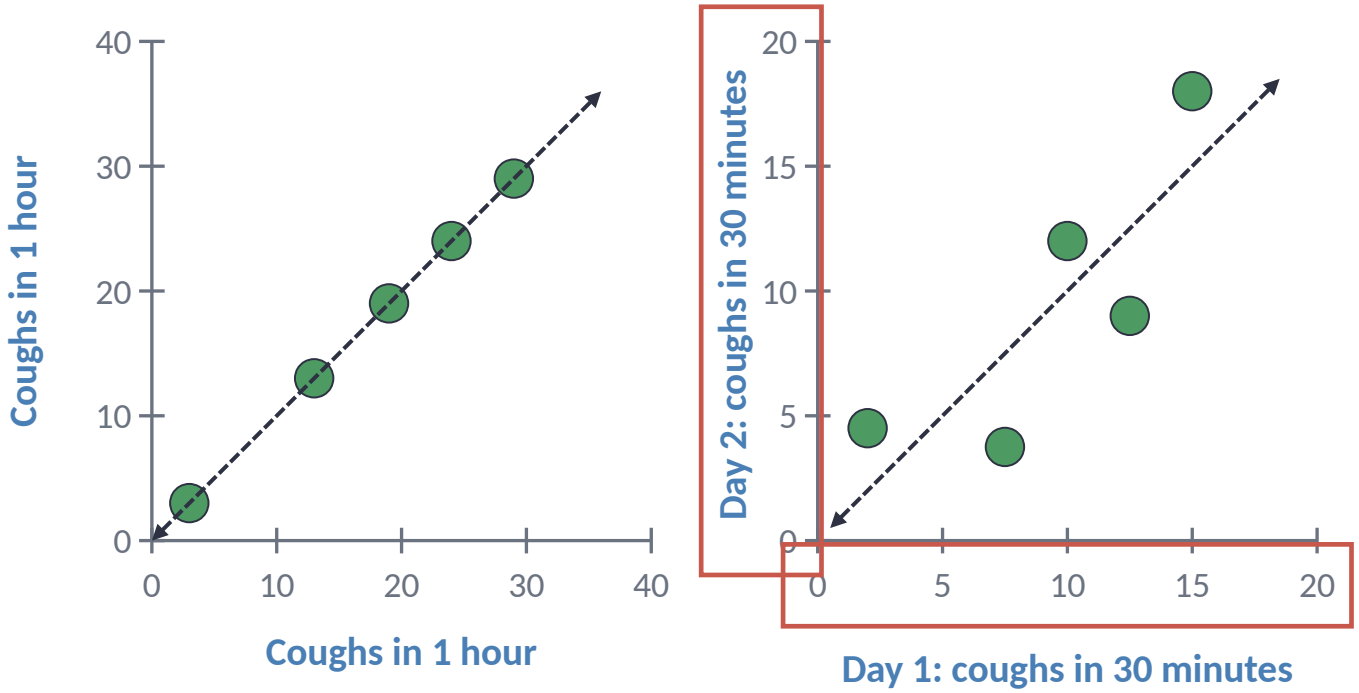
Scale also hides how close points are to the line

Now let us change *only* the scale on the right, by counting the same coughs per 30 minutes instead of per two hours…

…and recalculate the covariance. Now the covariance is *smaller*, **23.9**, for the data that do not fall on the line.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{23.9}}$$



Scale also hides how close points are to the line

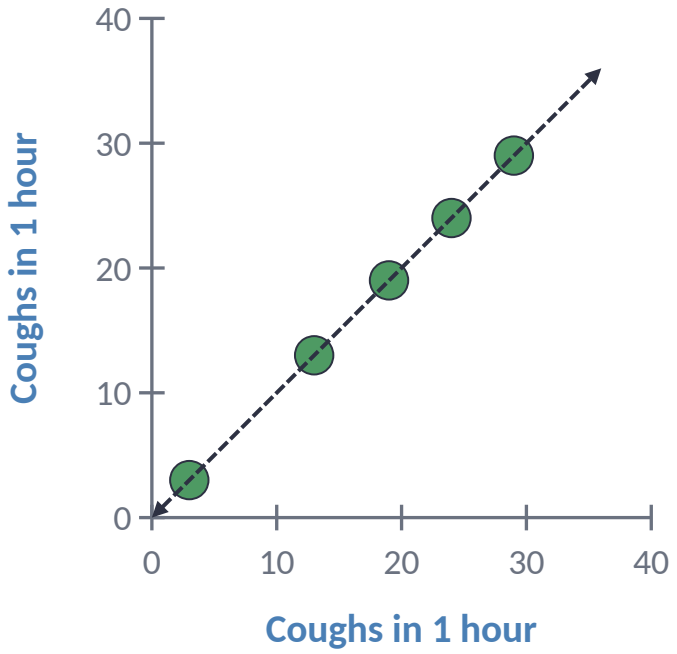
Now let us change *only* the scale on the right, by counting the same coughs per 30 minutes instead of per two hours…

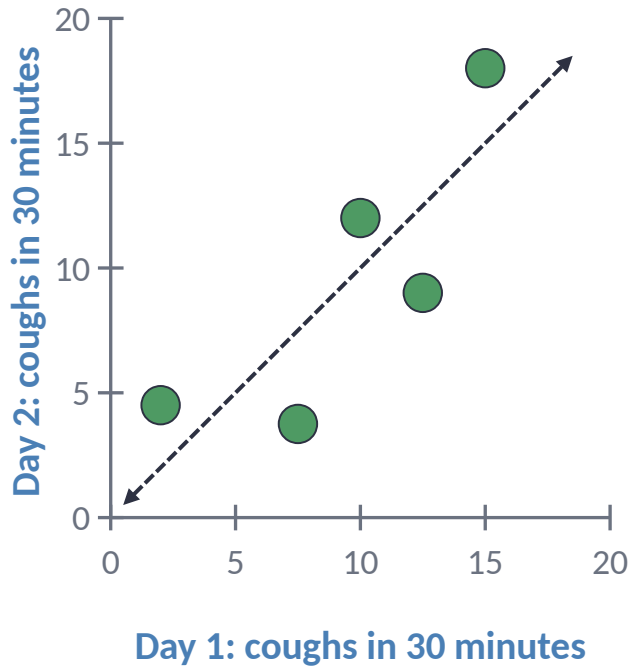
…and recalculate the covariance. Now the covariance is *smaller*, **23.9**, for the data that do not fall on the line.

We need something to describe relationships that was not sensitive to the scale of the data:
that is **correlation**, which we will talk about next.

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \mathbf{101.8}$$

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n - 1} = \boxed{\mathbf{23.9}}$$





Where covariance is used

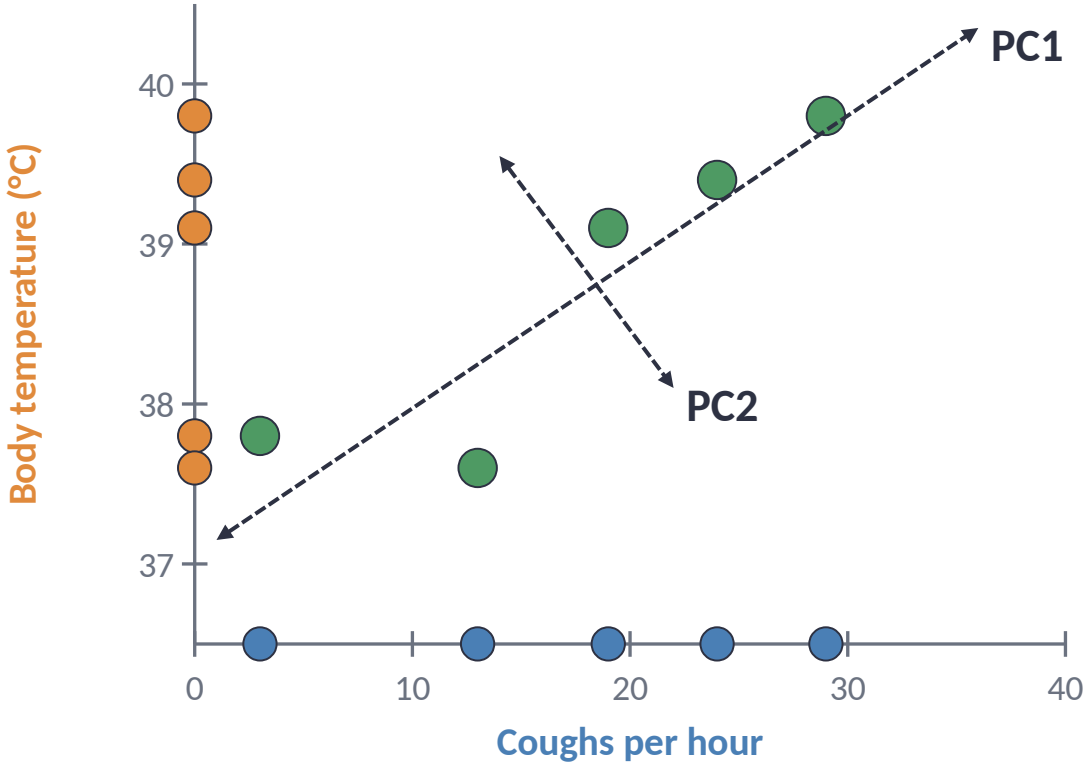
It is also worth mentioning that covariance values are used as a step in a wide variety of analyses.

Where covariance is used

It is also worth mentioning that covariance values are used as step in a wide variety of analyses.

For example, covariance values are used for **Principal Component Analysis (PCA)**.

The covariances between coughs, temperature, sleep and the other measurements we take from patients tell PCA which directions capture most of the variation between them.

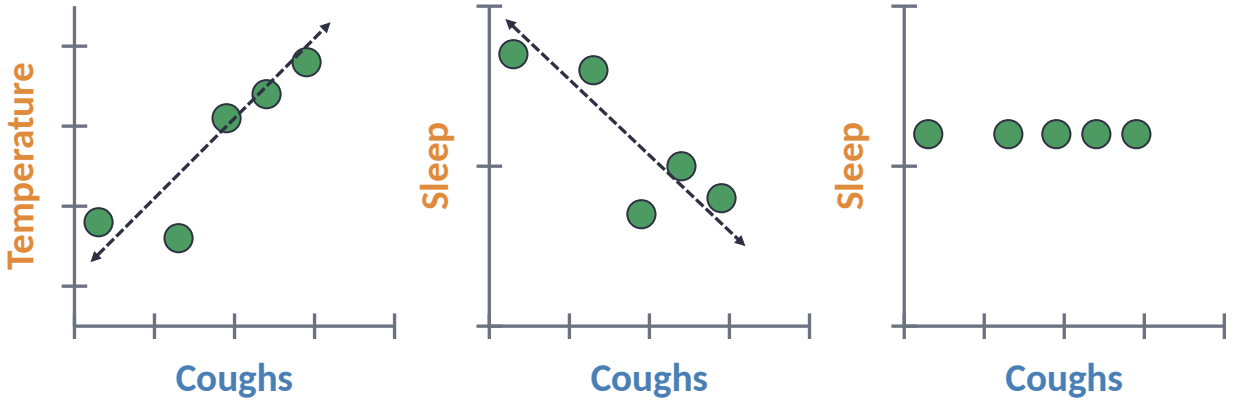


In summary

In summary, **covariance** is a way to classify three types of relationship.

The **covariance value** itself is difficult to interpret, because it depends on the scale of the data.

However, it is useful for calculating **correlations**, and in other computational settings, such as **PCA**.



1) When the covariance is *positive*, the relationship has a *positive* slope.

2) When the covariance is *negative*, the relationship has a *negative* slope.

3) When **covariance = 0**, there is **no relationship**, because there is no trend.

Correlation

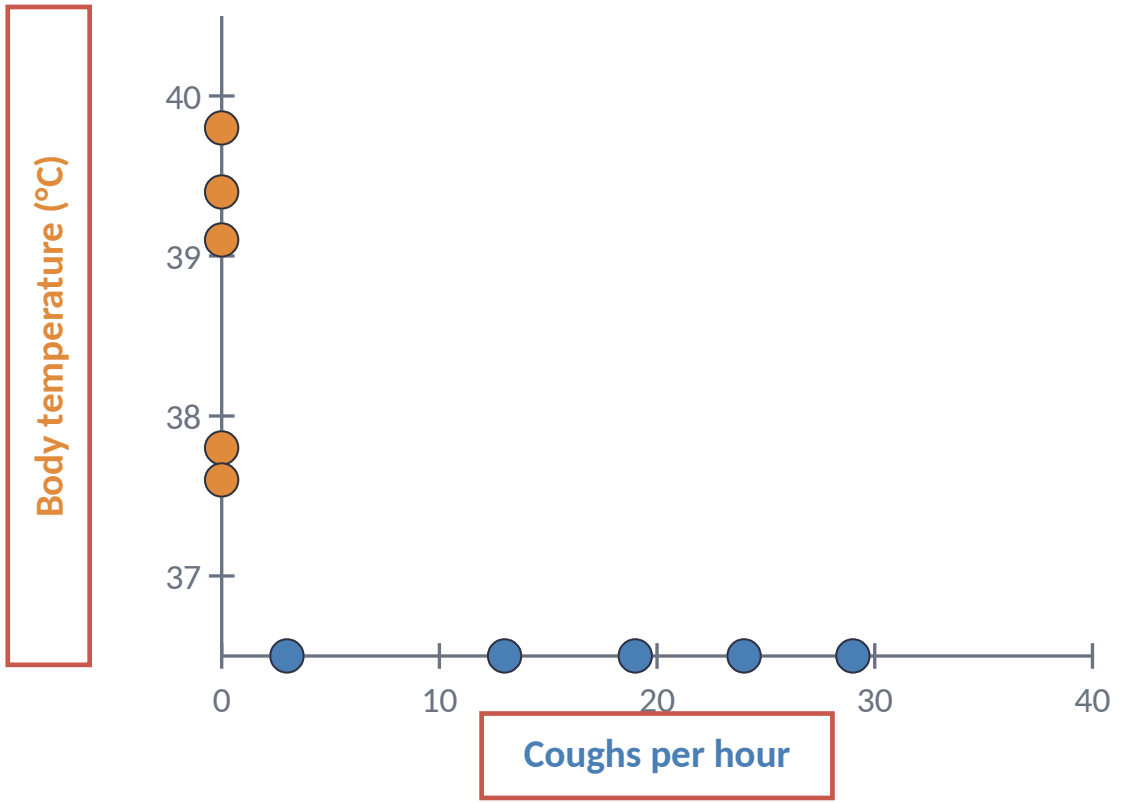
How strong is the relationship between two measurements?

Our patients again

Now we want to talk about the relationship between the data on the **x-axis**…

…which, as in the covariance slides, is the number of coughs during one hour for **5** flu patients…

…and the data on the **y-axis**: the body temperature of the same 5 patients.



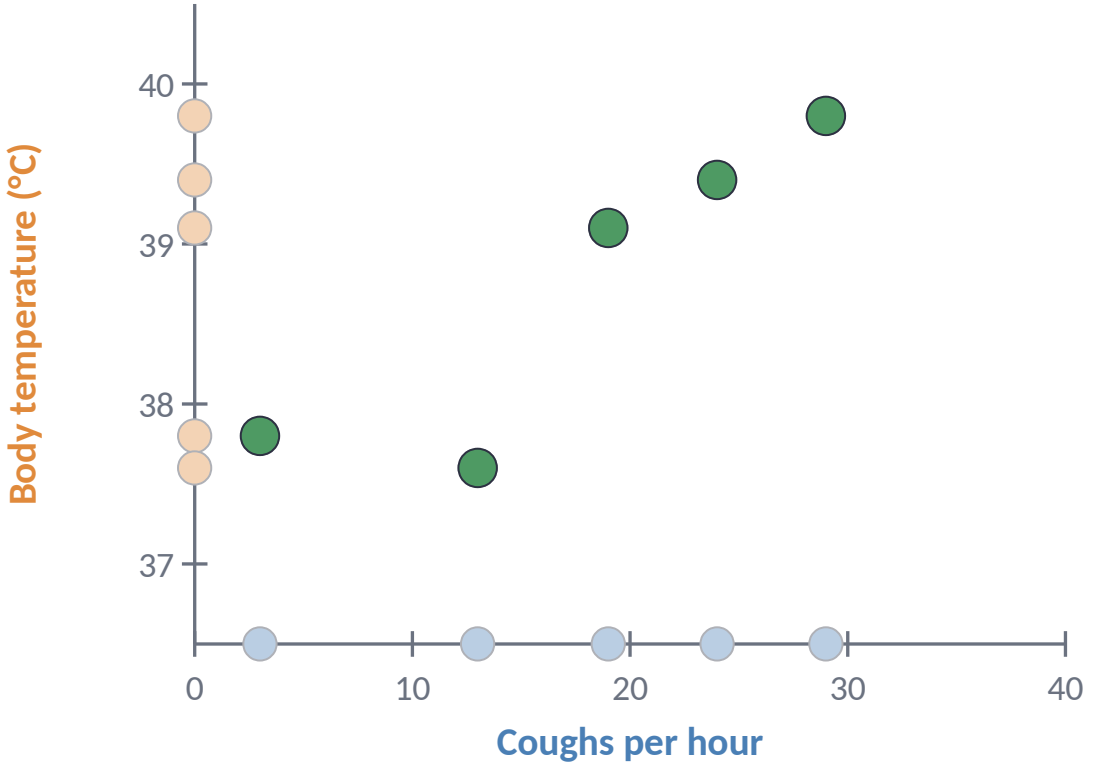
Our patients again

Now we want to talk about the relationship between the data on the **x-axis**…

…which, as in the covariance slides, is the number of coughs during one hour for **5** flu patients…

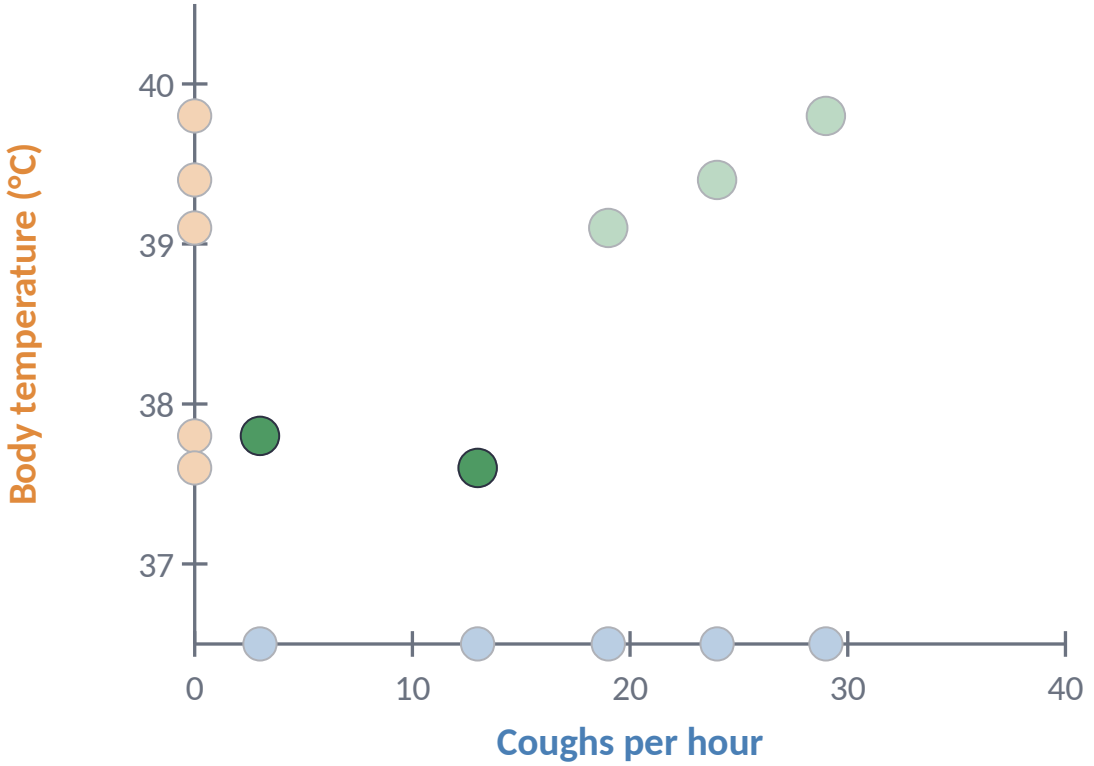
…and the data on the **y-axis**: the body temperature of the same 5 patients.

Each pair of measurements was taken from a single patient, and can be represented by a single **green dot**.



Our patients again

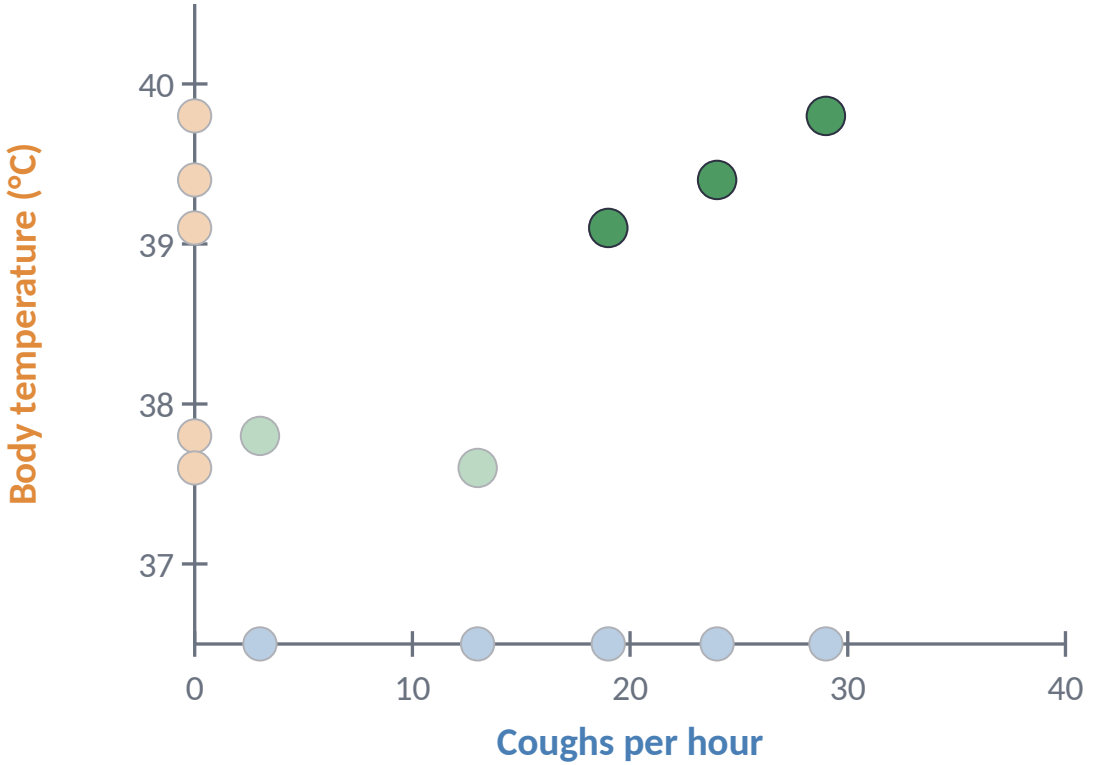
We can see that, in general, relatively few coughs are paired with relatively low temperatures…



Our patients again

We can see that, in general, relatively few coughs are paired with relatively low temperatures…

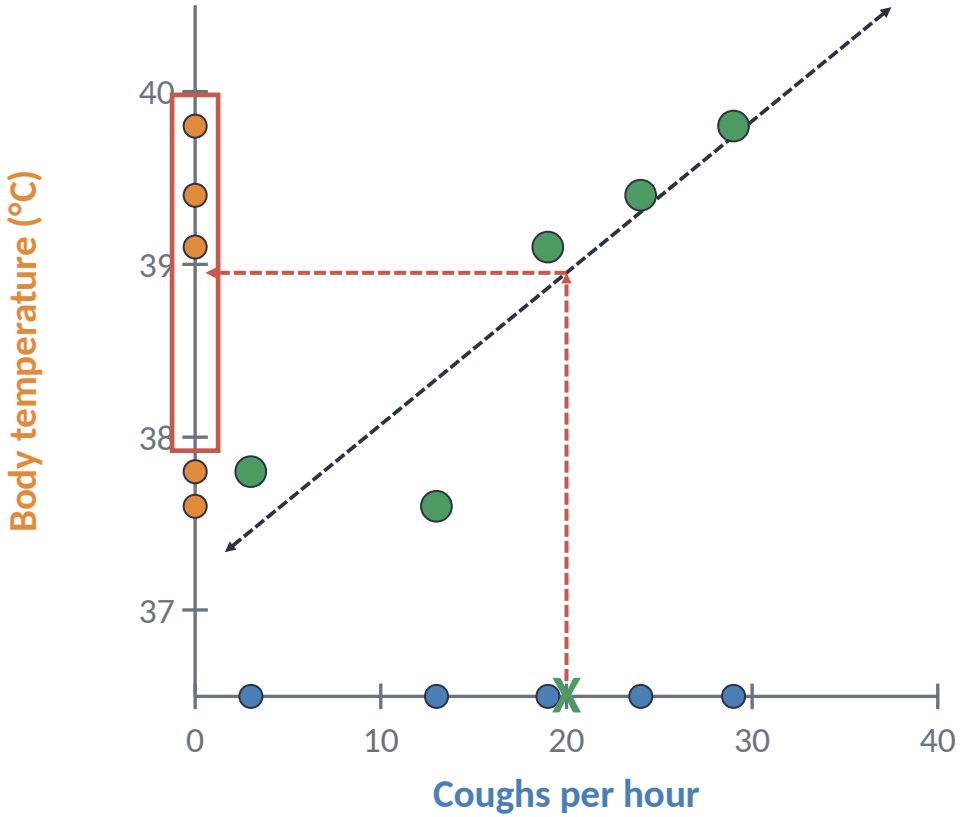
…and relatively many coughs are paired with relatively high temperatures.



Using the trend to make a guess

We can draw a line that represents the trend in the data…

…and if someone told us that a new patient coughed **20** times in an hour, we could use the line to guess that their temperature should be somewhere around **39 ℃**.

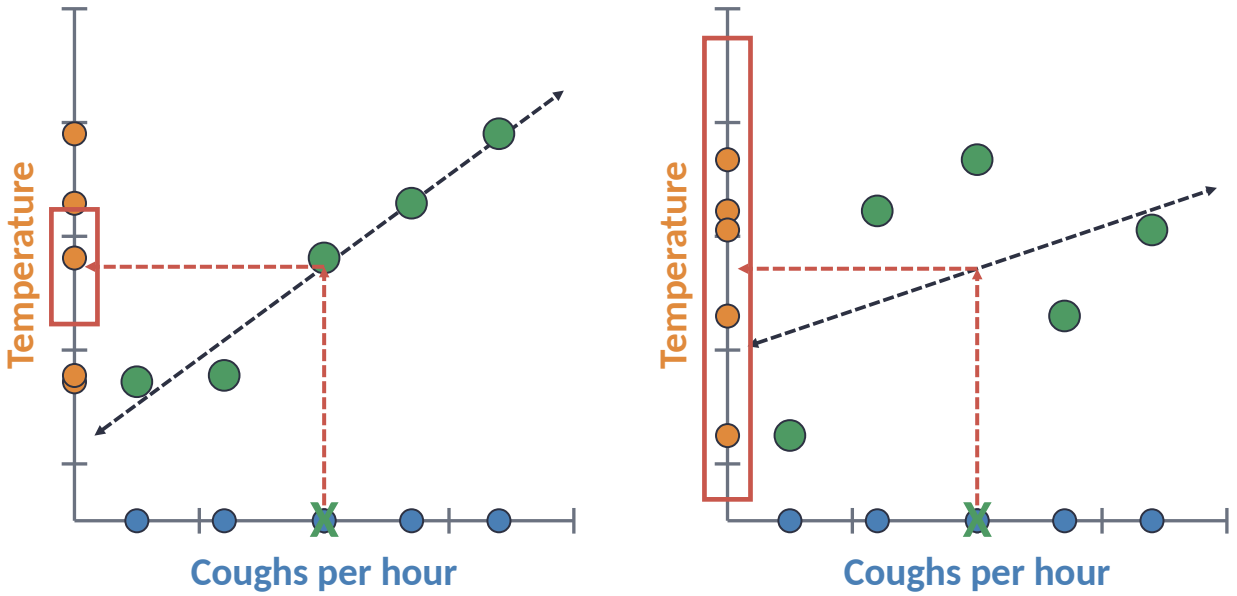


Closer data, better guesses

Here are five other patients. In this case, the data are very close to the line…

…and the closer the data are to the line, the more the coughs can tell us about the temperature: our guess falls within a narrow range.

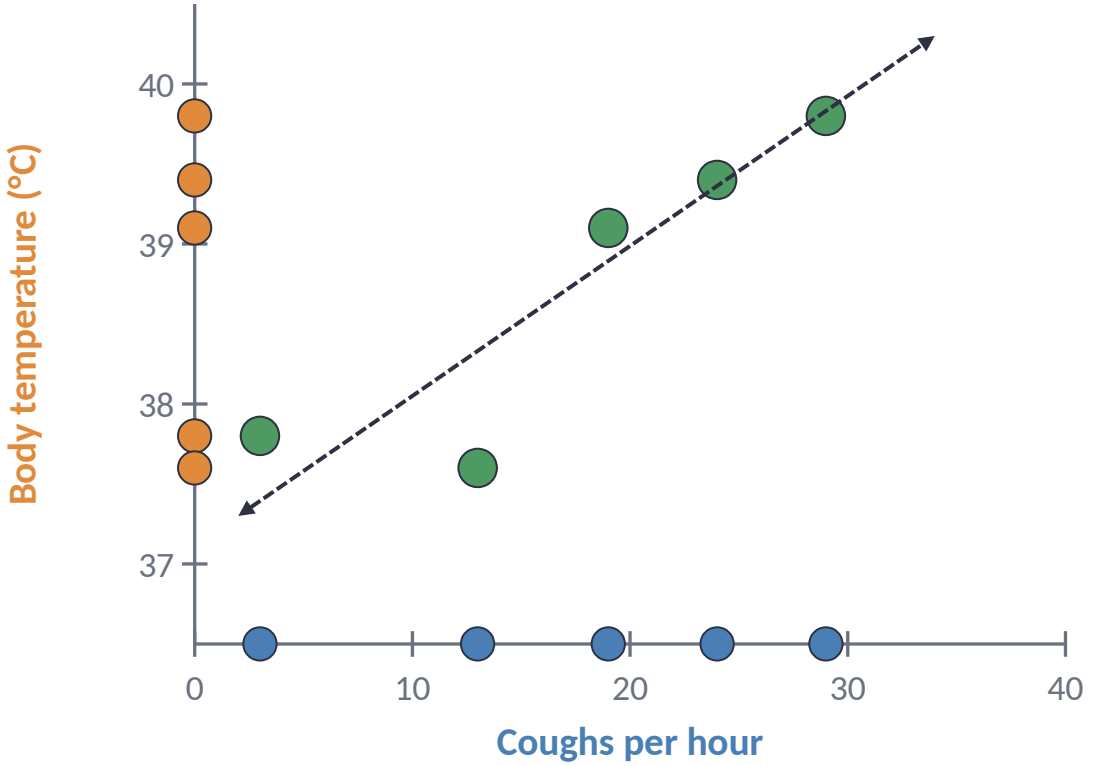
Here, the data are much further from the line, so we could say that the coughs tell us *less* about the temperature: the range of likely values is much wider.



Correlation is not causation

Note: just to be clear, all we are saying is that we *observed* that few coughs tend to be paired with low temperatures…

…and that this observation suggests a trend.



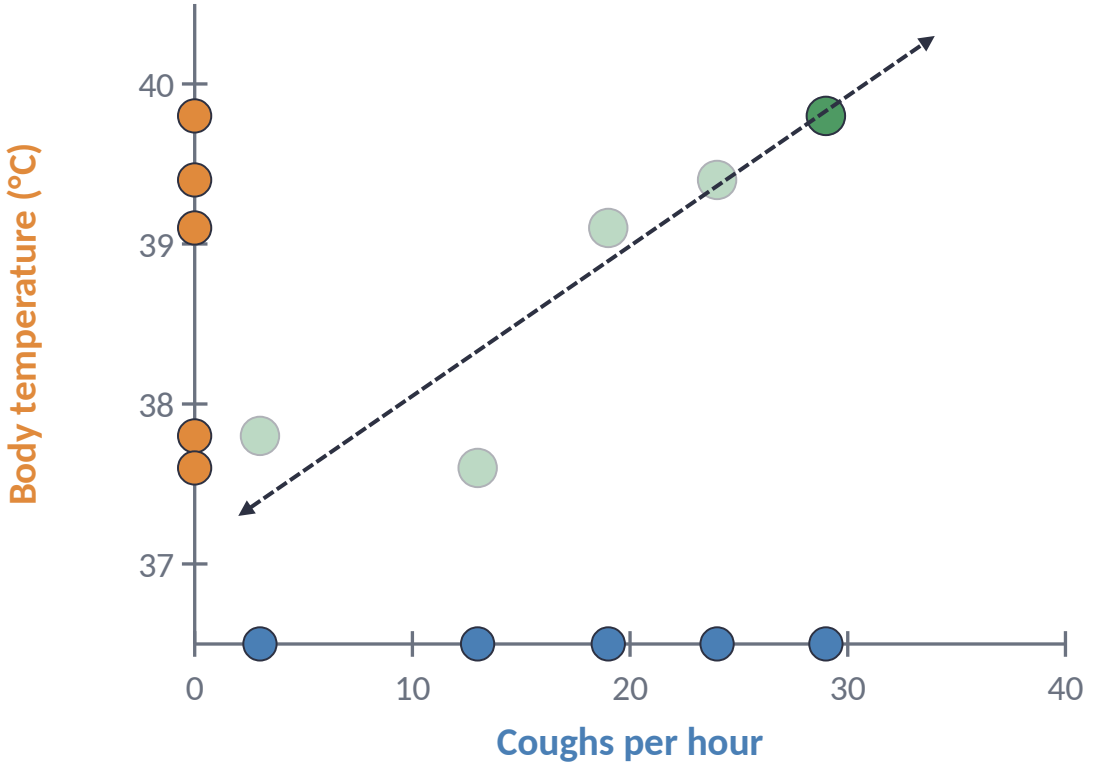
Correlation is not causation

Note: just to be clear, all we are saying is that we *observed* that few coughs tend to be paired with low temperatures…

…and that this observation suggests a trend.

We are *not* saying that coughing *causes* a high temperature, or the other way round.

Here, both are more likely to be consequences of something else: how severe the infection is. Correlation describes the data; it does not explain why.

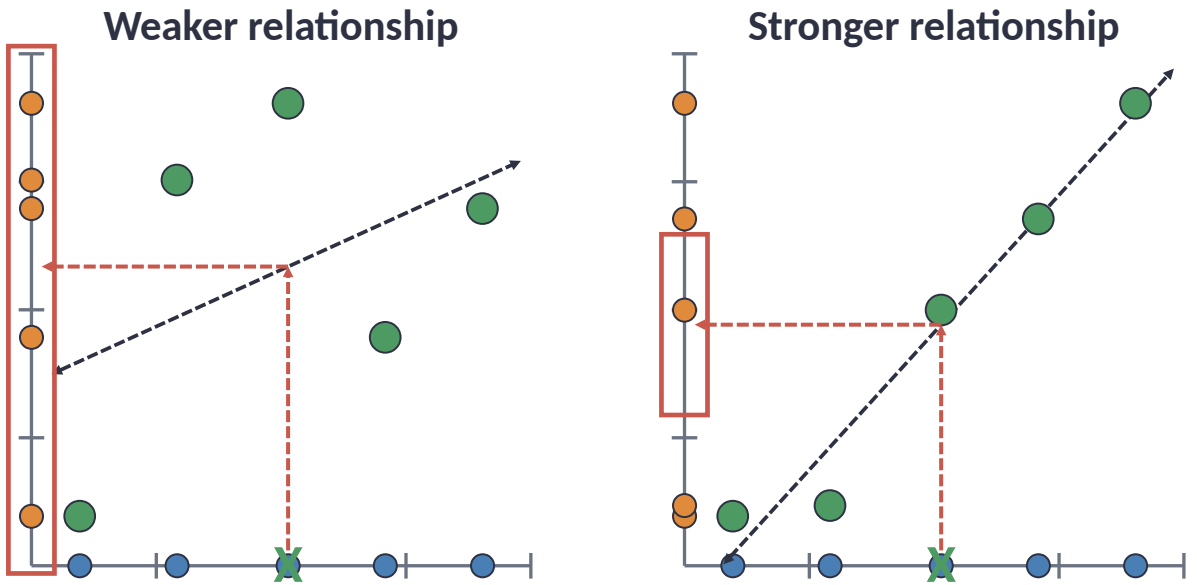


Quantifying the strength of a relationship

So far, we have looked at a relatively weak relationship…

…and a relatively strong relationship.

We can quantify the strength of a relationship with **correlation**.

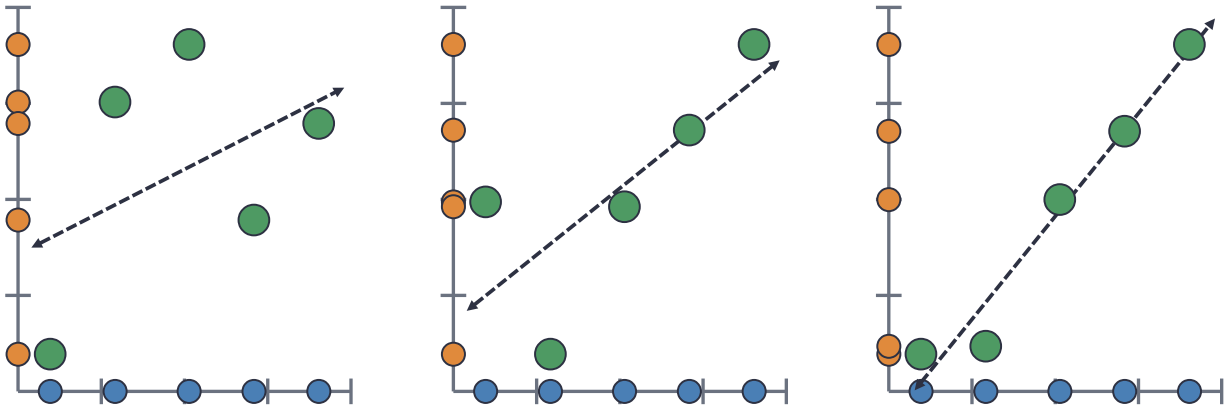


Quantifying the strength of a relationship

In other words, these data, with a relatively weak relationship, have a *small* correlation value…

…these data, with a moderate relationship, have a *moderate* correlation value…

…and these data, with a strong relationship, have a relatively *large* correlation value.



Weak relationship
=
small correlation value

Moderate relationship
=
moderate correlation value

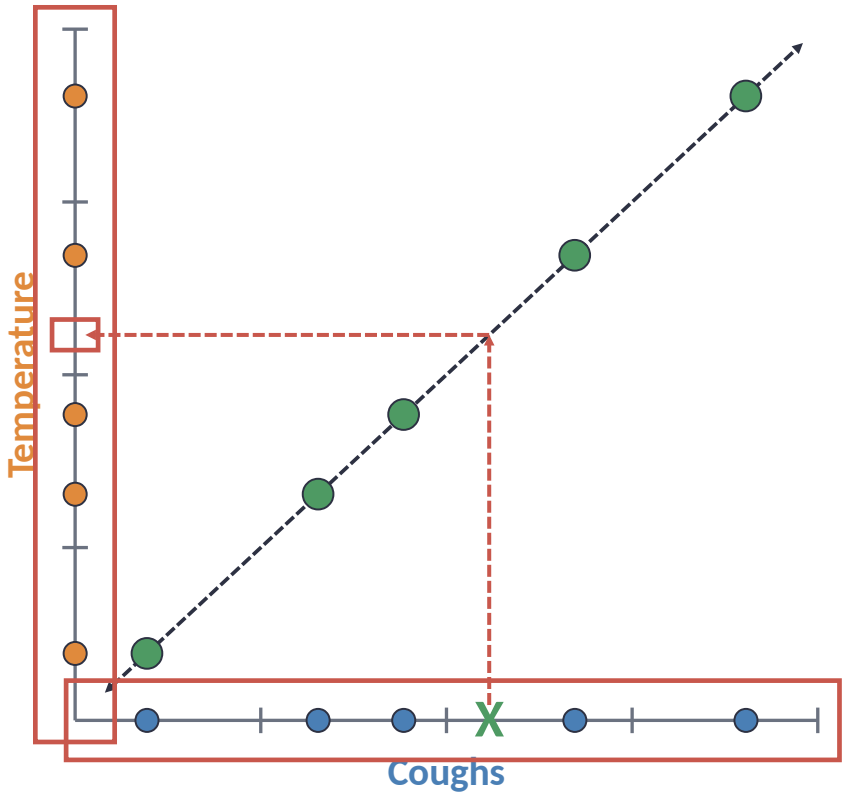
Strong relationship
=
large correlation value

Correlation = 1

Correlation = 1 when a straight line with a positive slope can go through the centre of every data point.

If someone told us how many times a patient coughed, we could guess their temperature within a very, very narrow range.

In fact, the numbers have been left off the axes on purpose, because they do not affect correlation at all.



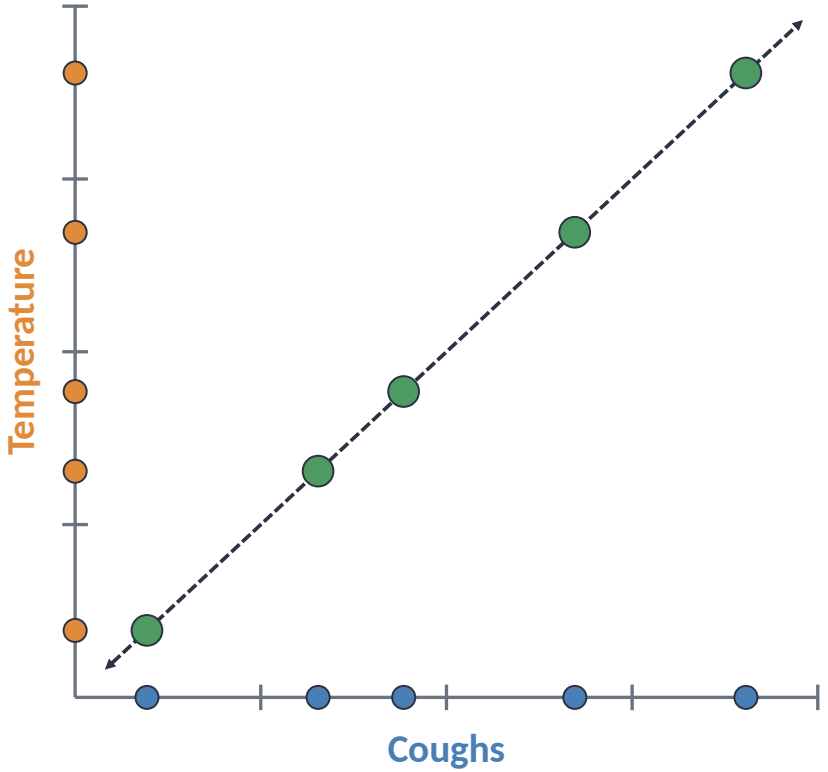
Correlation = 1

Correlation = 1 when a straight line with a positive slope can go through the centre of every data point.

If someone told us how many times a patient coughed, we could guess their temperature within a very, very narrow range.

In fact, the numbers have been left off the axes on purpose, because they do not affect correlation at all.

In other words, regardless of the scale of the data, **correlation = 1** when a straight line with a positive slope can go through all of the data.

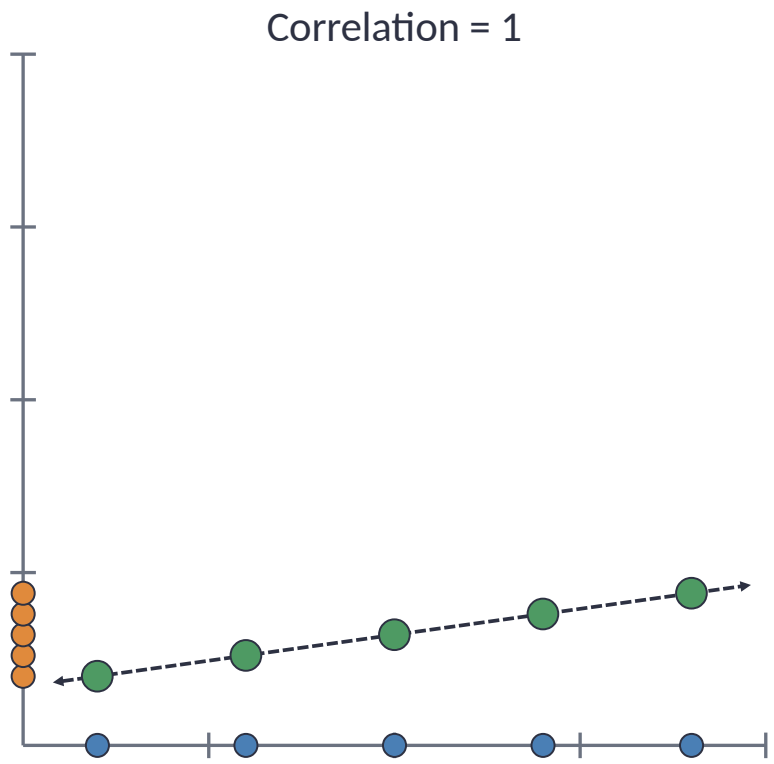


Steep or shallow, it is still 1

That means correlation can equal 1 when the slope is large…

…and when the slope is small.

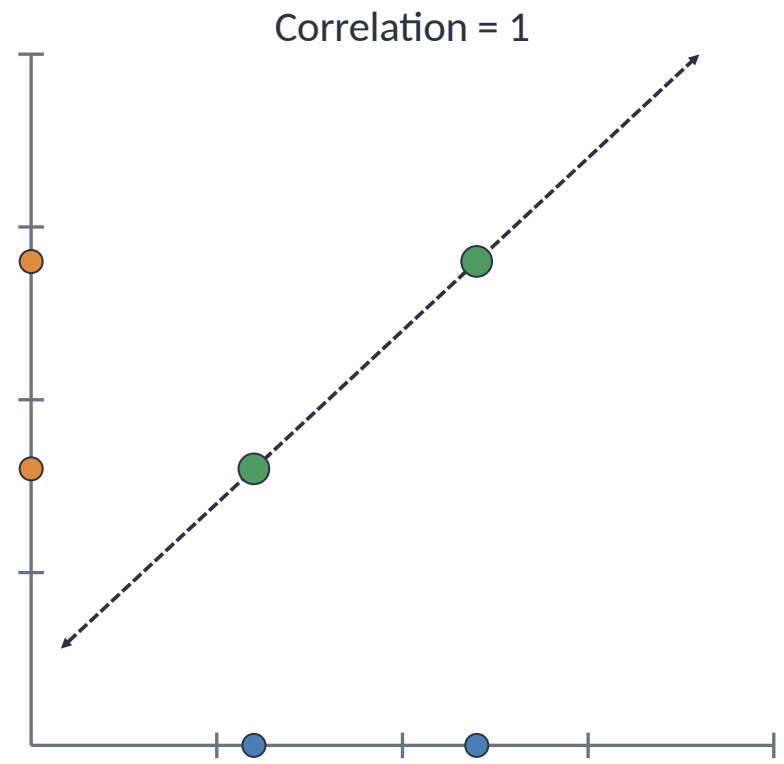
Correlation tells us how well a straight line fits the data. It does *not* tell us how steep the line is.



How much data supports the line?

Note: when a straight line with a positive slope goes through the data, correlation = 1 regardless of how much data we have. With only 2 patients, correlation = 1…

…and that makes the relationship *appear* strong.

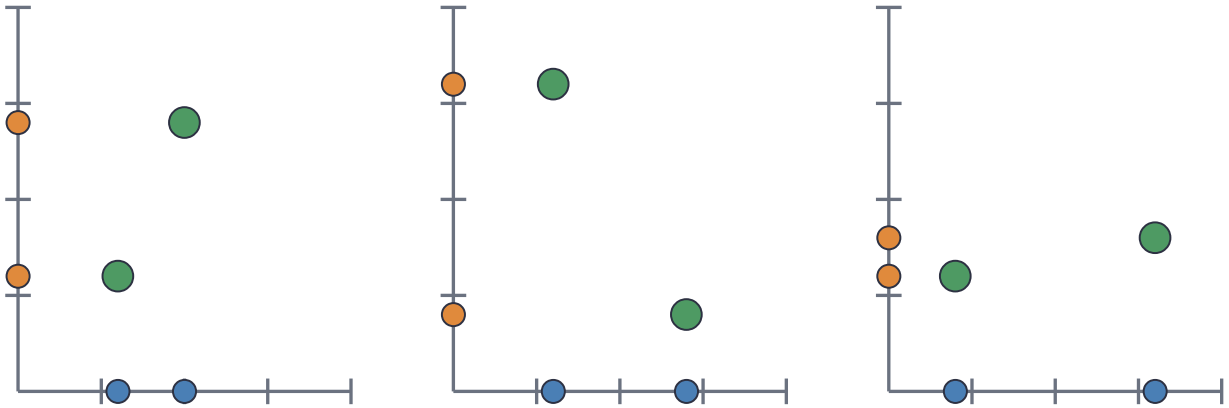


How much data supports the line?

Note: when a straight line with a positive slope goes through the data, correlation = 1 regardless of how much data we have. With only 2 patients, correlation = 1…

…and that makes the relationship *appear* strong.

However, we can always draw a straight line between any two random dots.



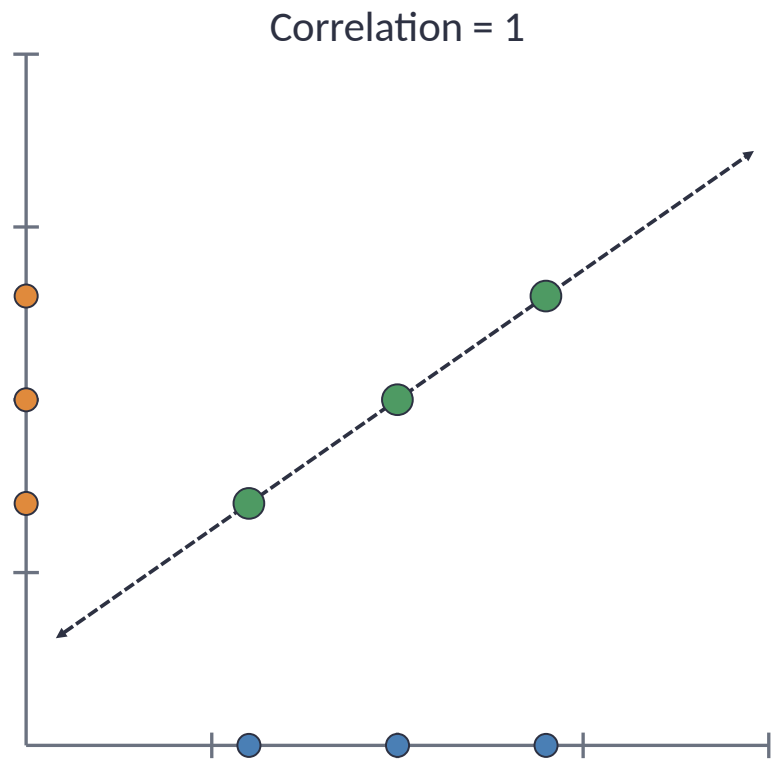
How much data supports the line?

Note: when a straight line with a positive slope goes through the data, correlation = 1 regardless of how much data we have. With only 2 patients, correlation = 1…

…and that makes the relationship *appear* strong.

However, we can always draw a straight line between any two random dots.

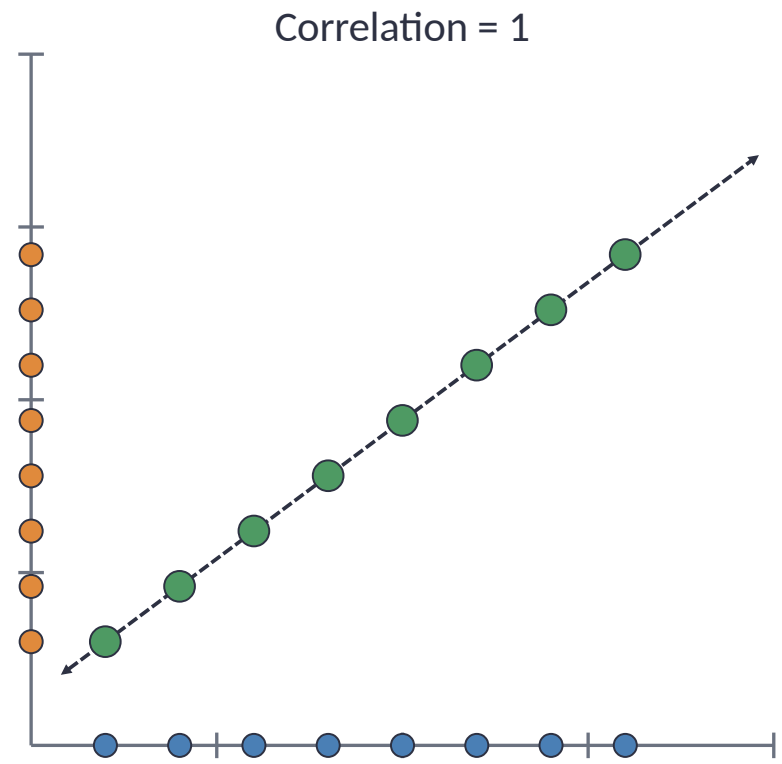
Now, just like before, since we can draw a straight line through all three points, correlation = 1…



How much data supports the line?

…but the probability that we can connect *3* randomly placed points with a straight line is very small.

In general, the more data we have, the more convincing a straight line through them is…



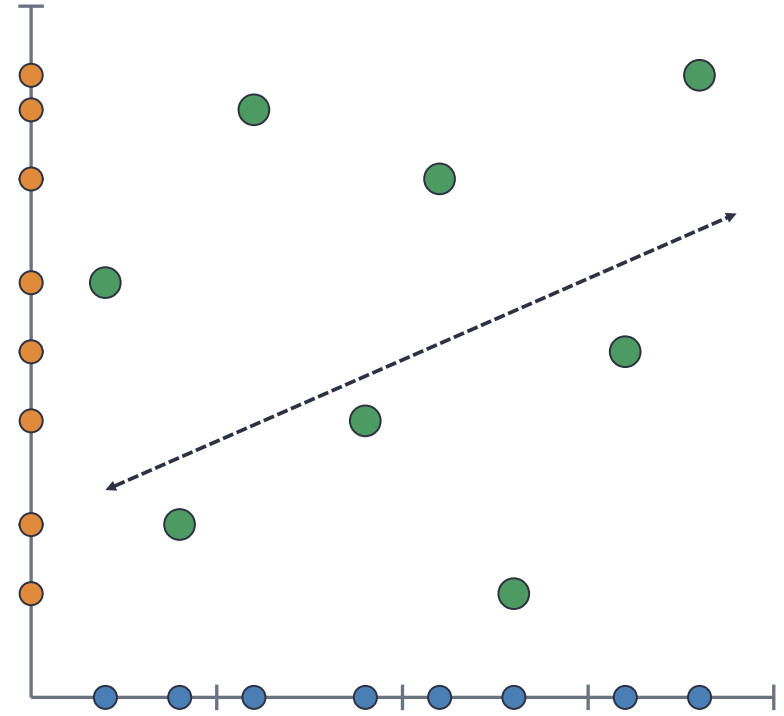
How much data supports the line?

…but the probability that we can connect *3* randomly placed points with a straight line is very small.

In general, the more data we have, the more convincing a straight line through them is…

…because the probability that randomly placed points all fall on the same straight line gets smaller and smaller with each additional point.

Note: when we talk about **correlation**, we are only talking about *straight lines*.

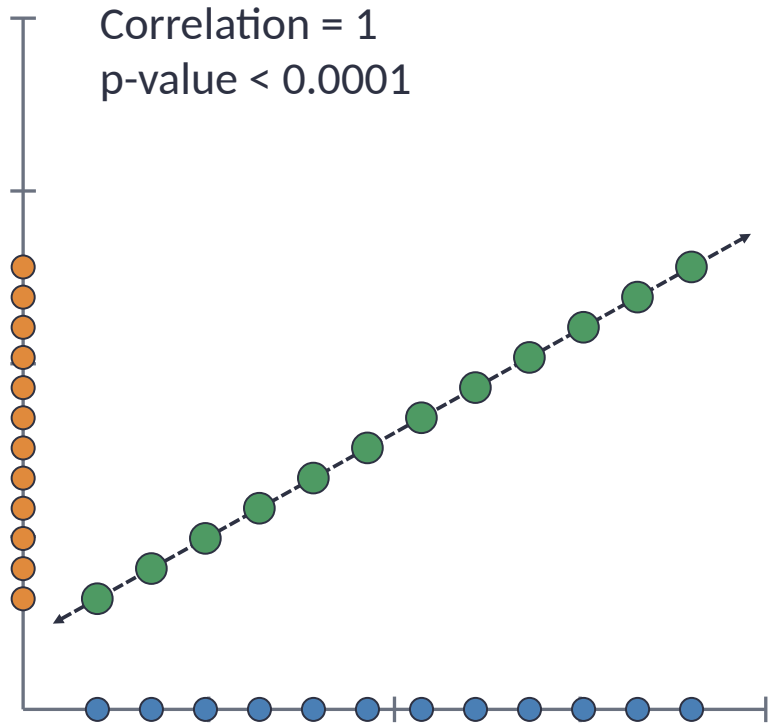


Confidence and the p-value

We can quantify our confidence in a relationship with a **p-value**: the probability of seeing a relationship at least this strong just by chance, if the measurements were really unrelated.

The smaller the p-value, the more confidence we have in the predictions we make with the line.

The maximum value for correlation, 1, occurs whenever a straight line with a positive slope goes through all the data…

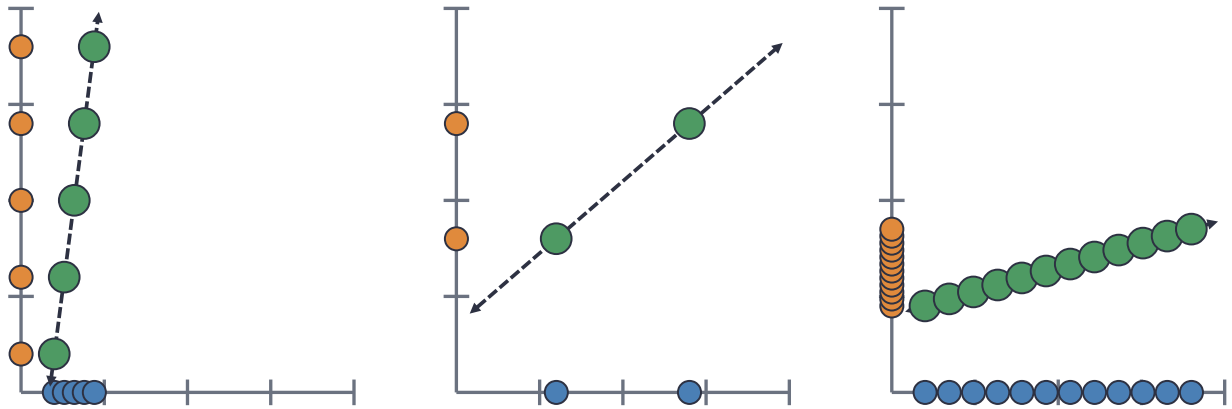


Confidence and the p-value

Here are three examples where correlation = 1.

Of these three, we should have the least confidence in the middle one, since it is supported by the least amount of data: just 2 patients, and a p-value of 1.

…and we should have the most confidence in the one on the right, since it is supported by the most data and has the smallest p-value.



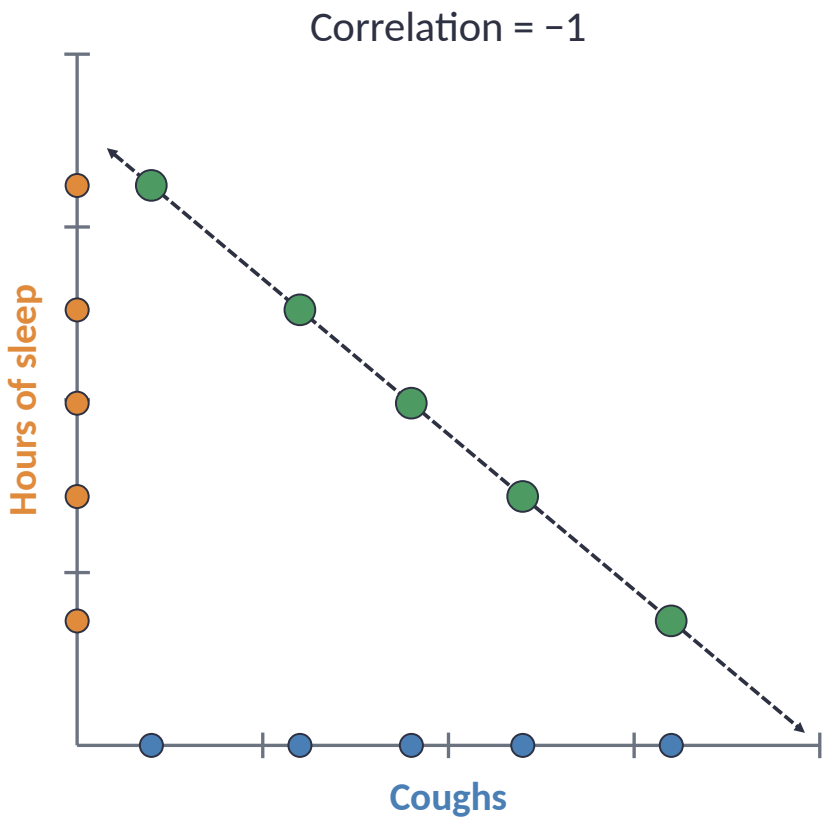
Correlation = 1
p-value < 0.0001

Correlation = 1
p-value = 1

Correlation = 1
p-value < 0.0001

Correlation = −1

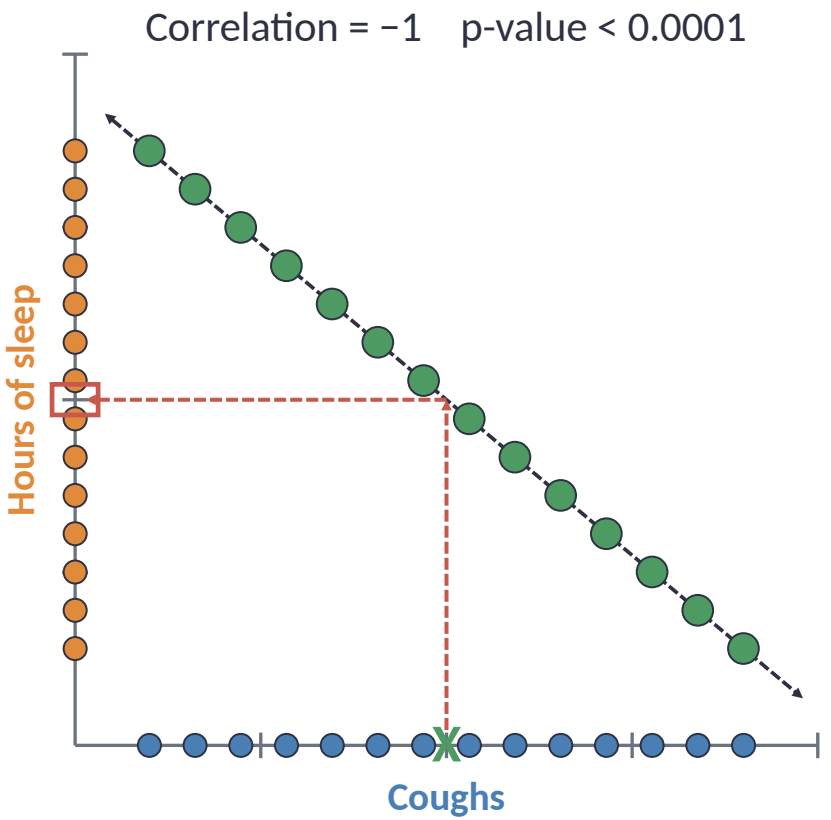
Since a straight line can go through all of the data, **correlation = −1** also implies that there is a strong relationship in the data.



Correlation = −1

Since a straight line can go through all of the data, **correlation = −1** also implies that there is a strong relationship in the data.

If we had a lot of data, we could have a lot of confidence in our guess, because the p-value would be very small…

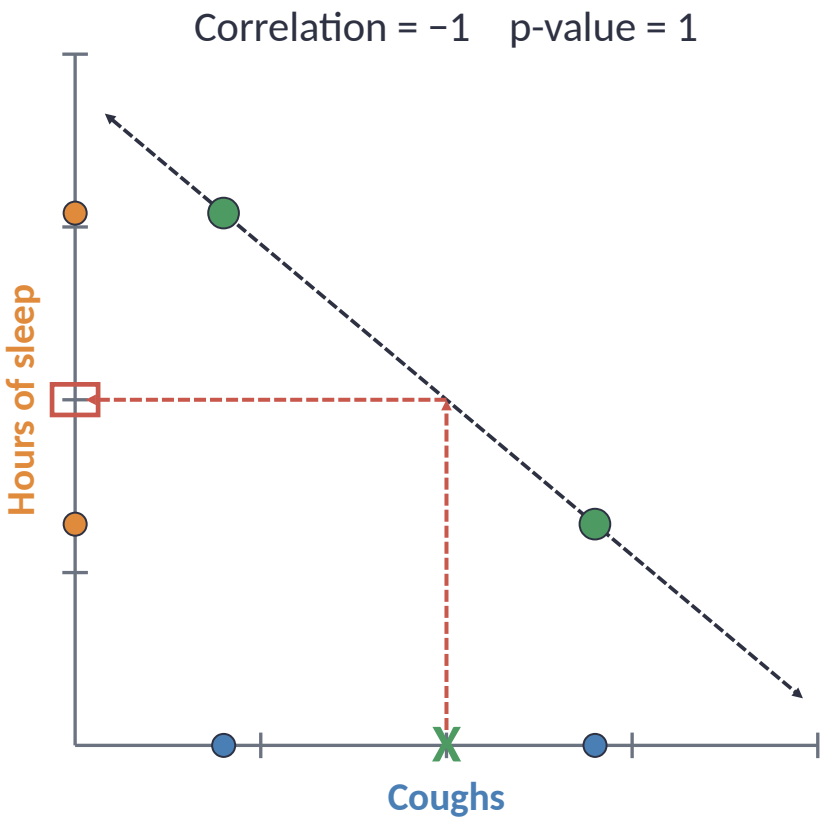


Correlation = −1

Since a straight line can go through all of the data, **correlation = −1** also implies that there is a strong relationship in the data.

If we had a lot of data, we could have a lot of confidence in our guess, because the p-value would be very small…

…and the less data we have, the less confidence we have in the guess, because the p-value gets larger.



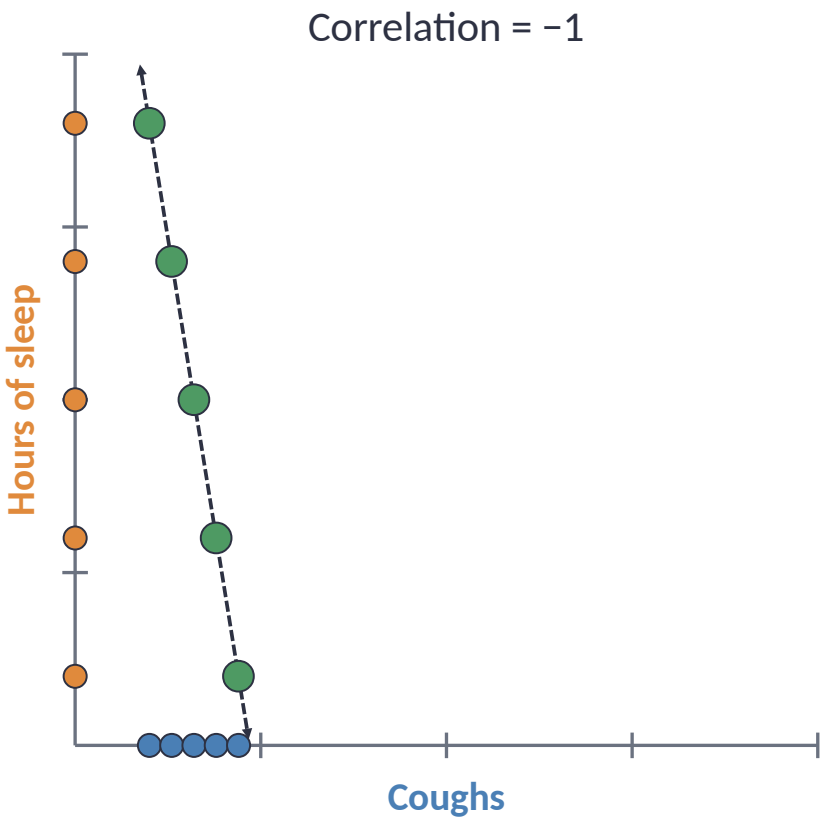
Correlation = −1

Since a straight line can go through all of the data, **correlation = −1** also implies that there is a strong relationship in the data.

If we had a lot of data, we could have a lot of confidence in our guess, because the p-value would be very small…

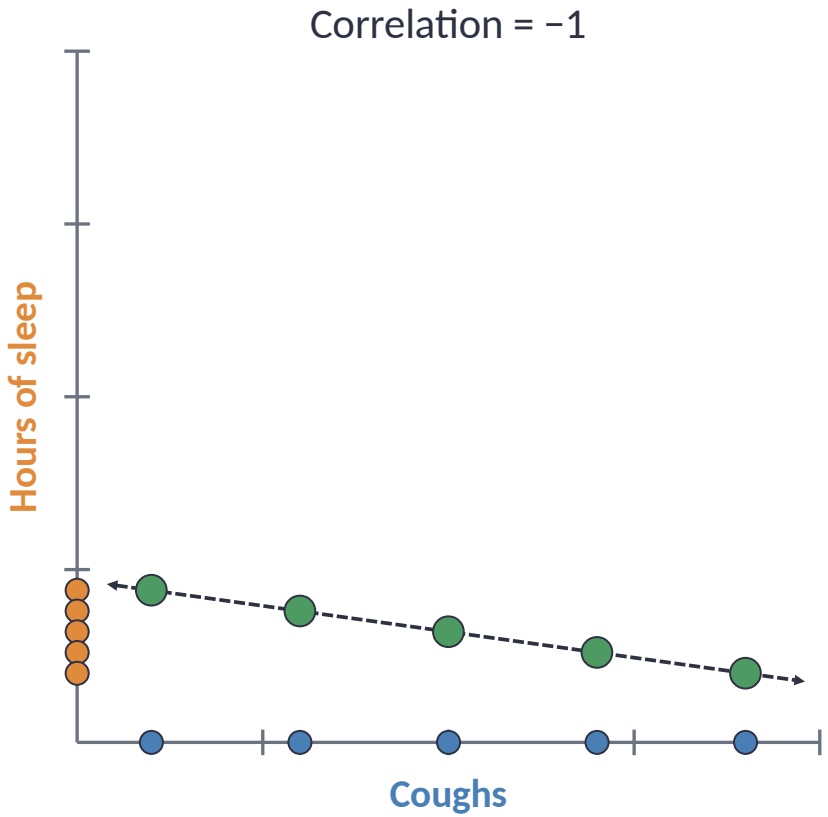
…and the less data we have, the less confidence we have in the guess, because the p-value gets larger.

Just like before, as long as a straight line goes through all the data and the slope is negative, correlation = −1 when the slope is large…



Correlation = −1

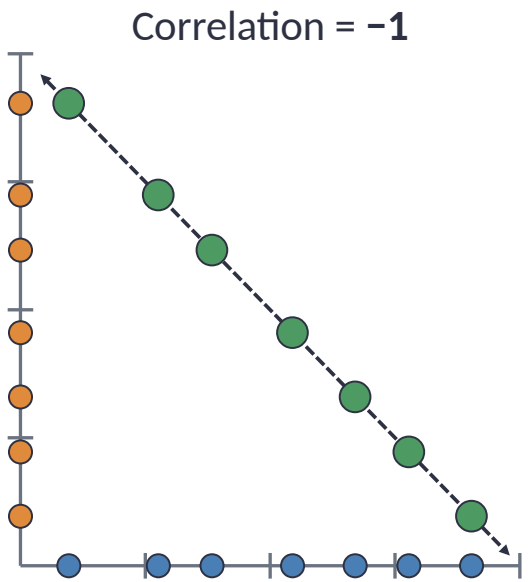
…and when the slope is small.

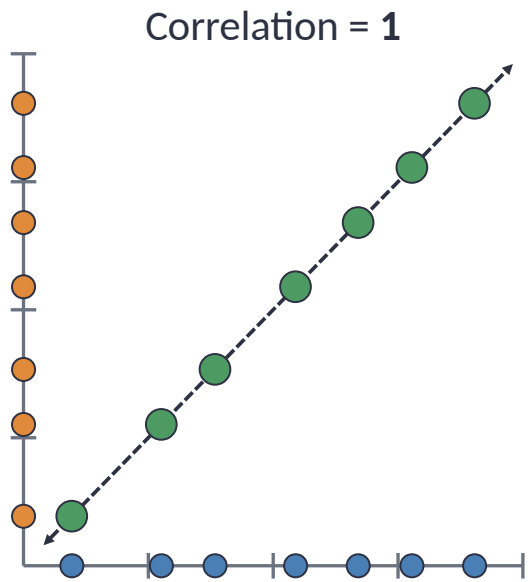


From −1 to 1

So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.



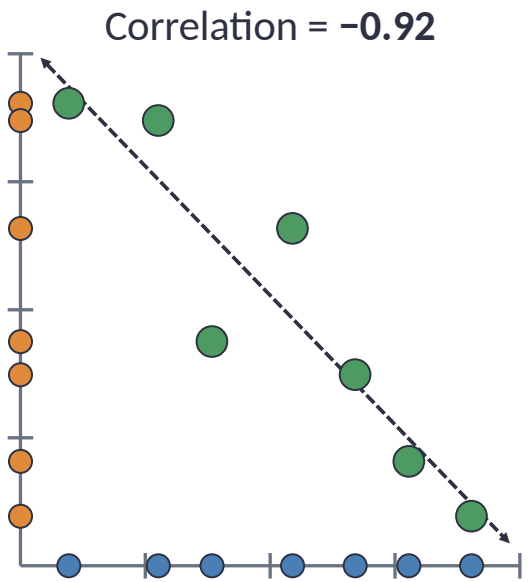


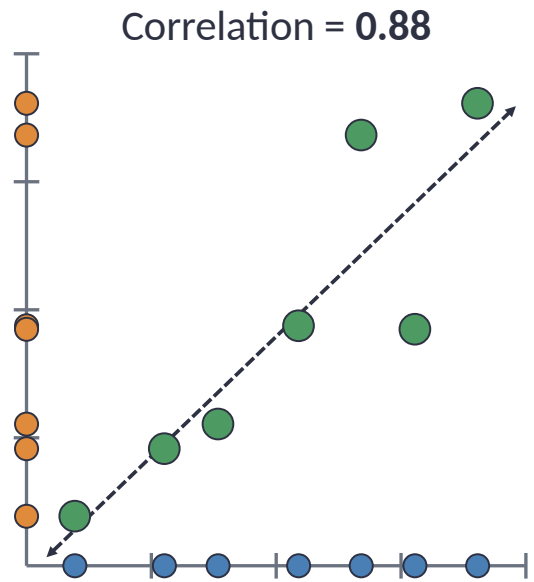
From −1 to 1

So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.

In both cases, if a straight line cannot go through all of the data, then we get correlation values closer to **0**…





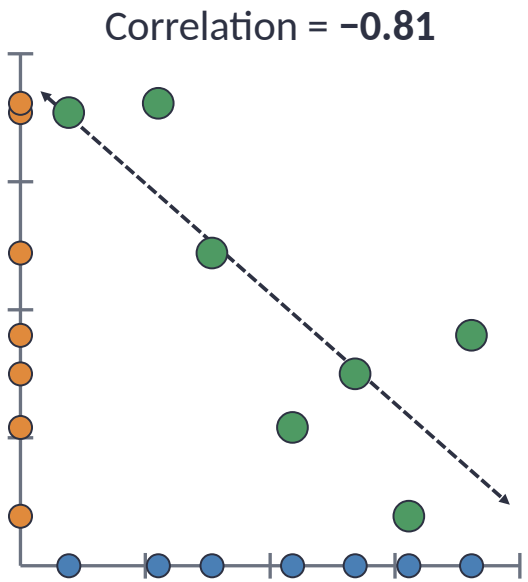
From −1 to 1

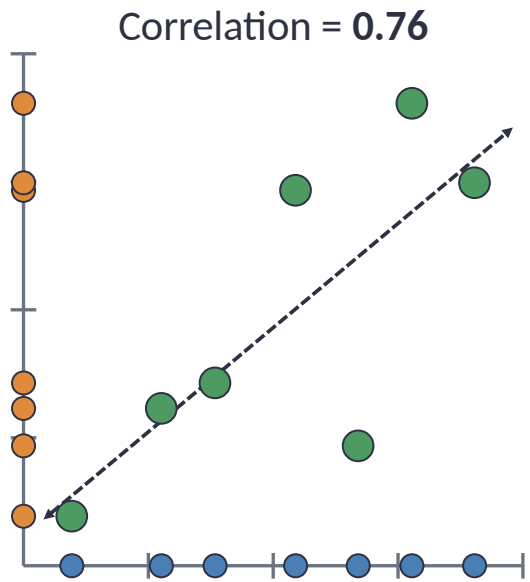
So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.

In both cases, if a straight line cannot go through all of the data, then we get correlation values closer to **0**…

…and the worse the fit, the closer the correlation values get to 0…





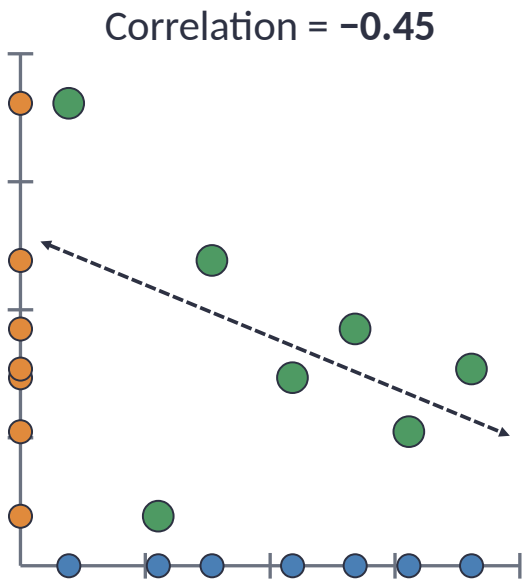
From −1 to 1

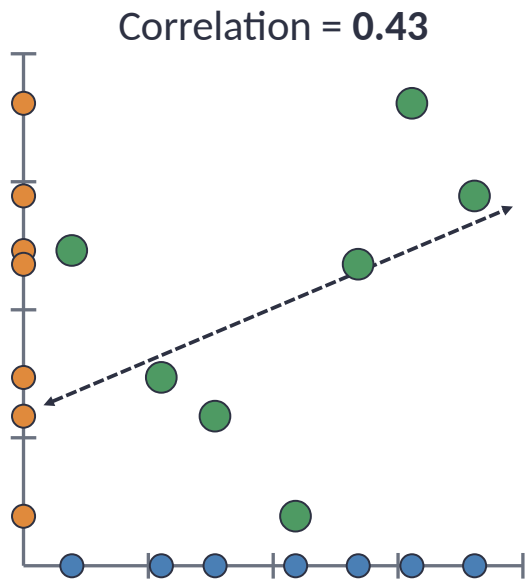
So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.

In both cases, if a straight line cannot go through all of the data, then we get correlation values closer to **0**…

…and the worse the fit, the closer the correlation values get to 0…





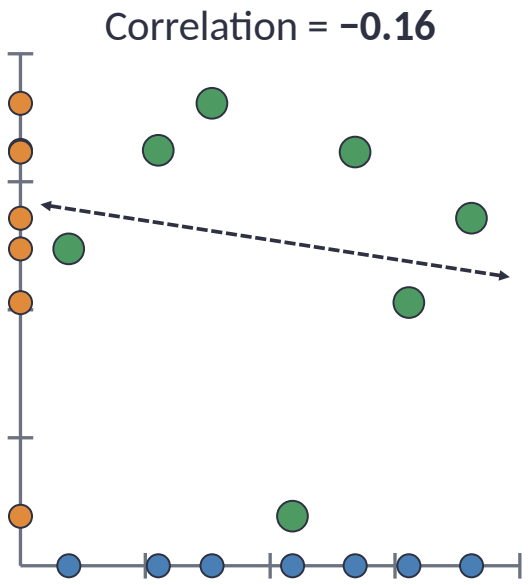
From −1 to 1

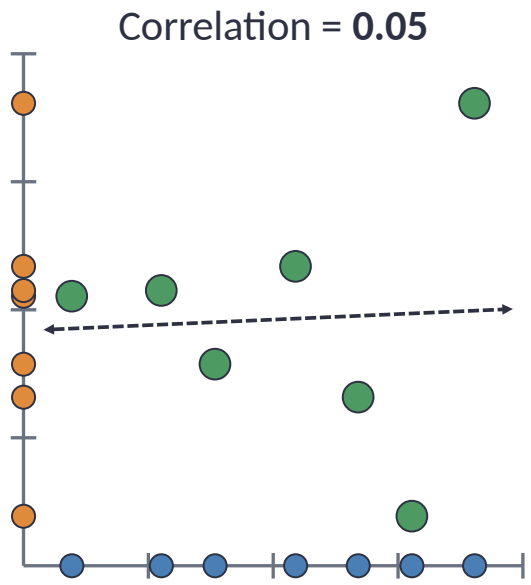
So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.

In both cases, if a straight line cannot go through all of the data, then we get correlation values closer to **0**…

…and the worse the fit, the closer the correlation values get to 0…





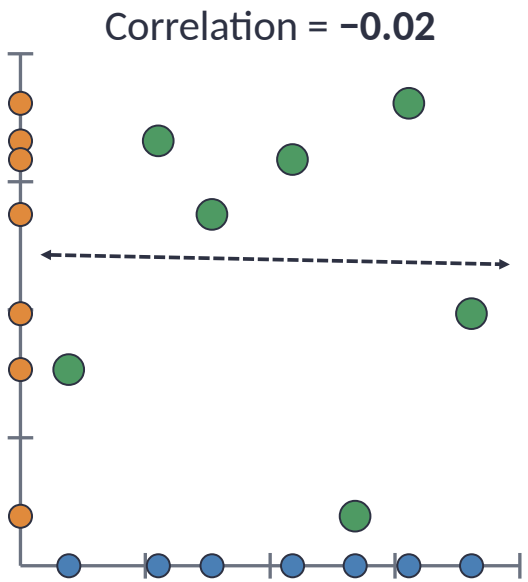
From −1 to 1

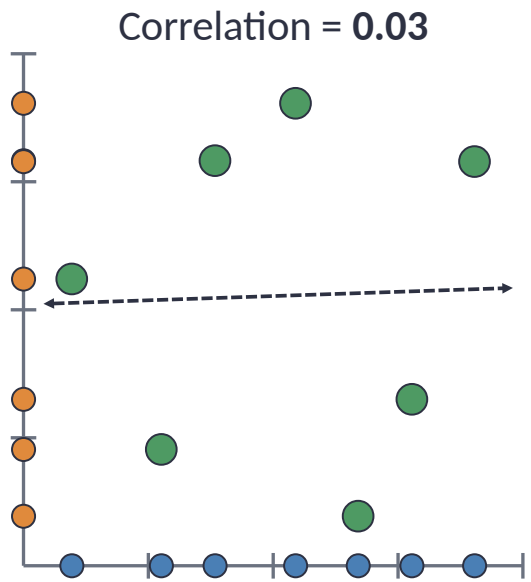
So far, we have seen that when the slope of the line is *negative*, the strongest relationship has **correlation = −1**…

…and when the slope of the line is *positive*, the strongest relationship has **correlation = 1**.

In both cases, if a straight line cannot go through all of the data, then we get correlation values closer to **0**…

…and the worse the fit, the closer the correlation values get to 0…

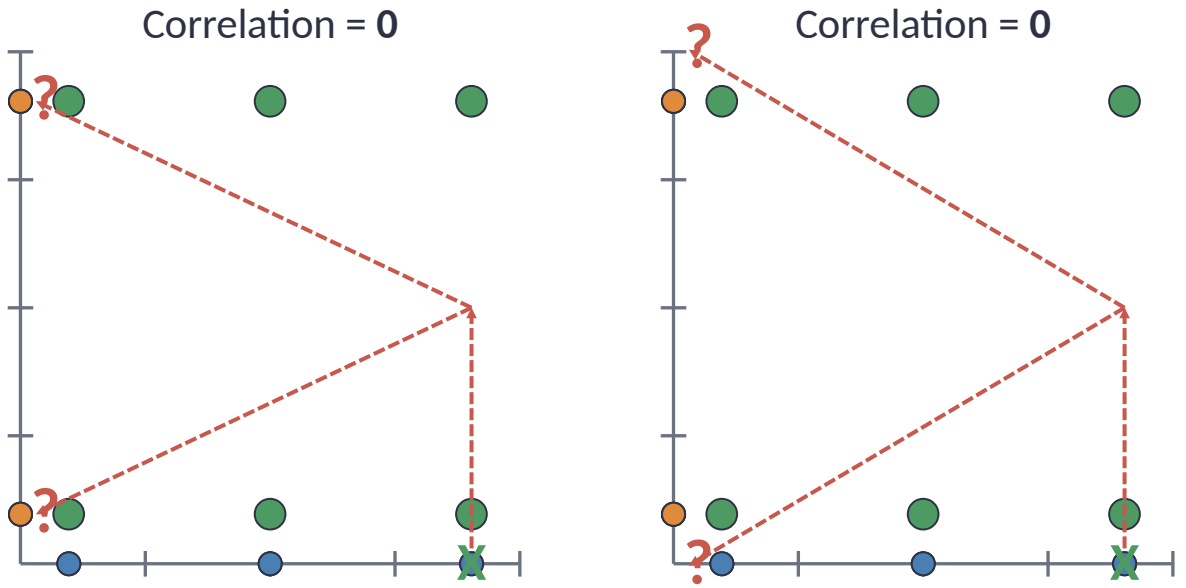




From −1 to 1

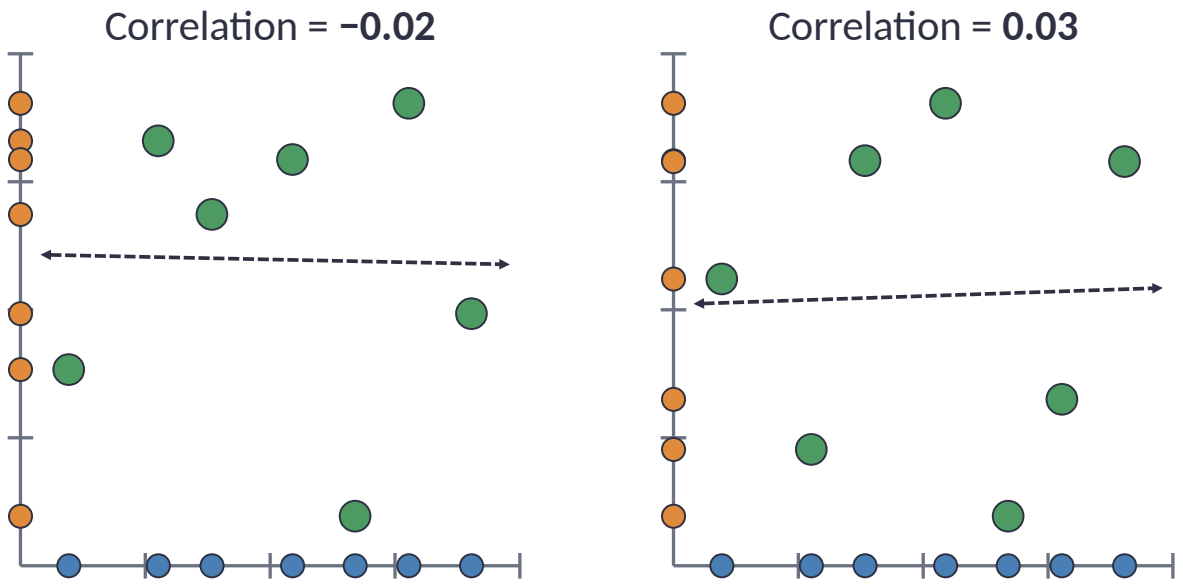
…and when there is no relationship that we can represent with a straight line, **correlation = 0**.

Knowing the value on the x-axis does not tell us anything about what to expect on the y-axis, because there is no reason to choose one value over another.



Guesses improve as correlation approaches ±1

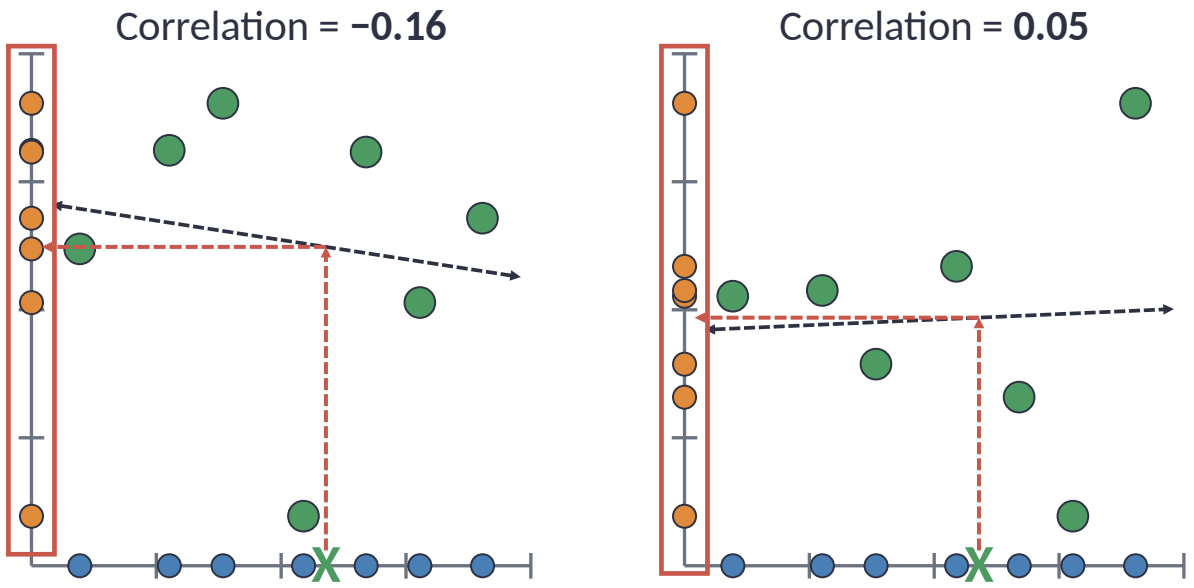
As long as the correlation value is not 0…



Guesses improve as correlation approaches ±1

As long as the correlation value is not 0…

…we can still use the line to make inferences…

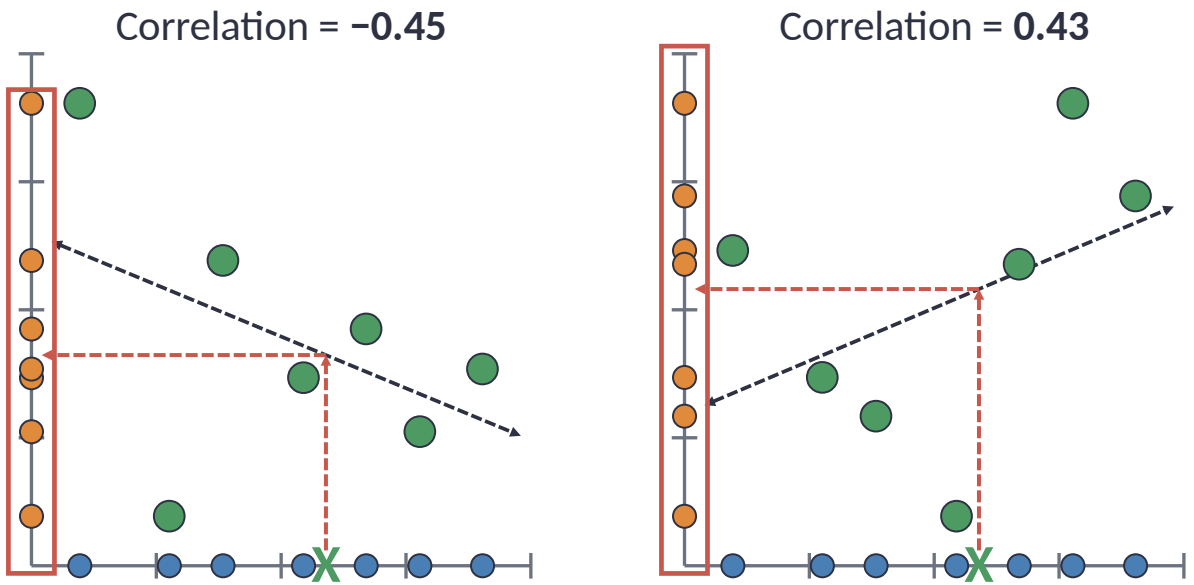


Guesses improve as correlation approaches ±1

As long as the correlation value is not 0…

…we can still use the line to make inferences…

…but our guesses become more refined the closer the correlation values get to **−1** or **1**.

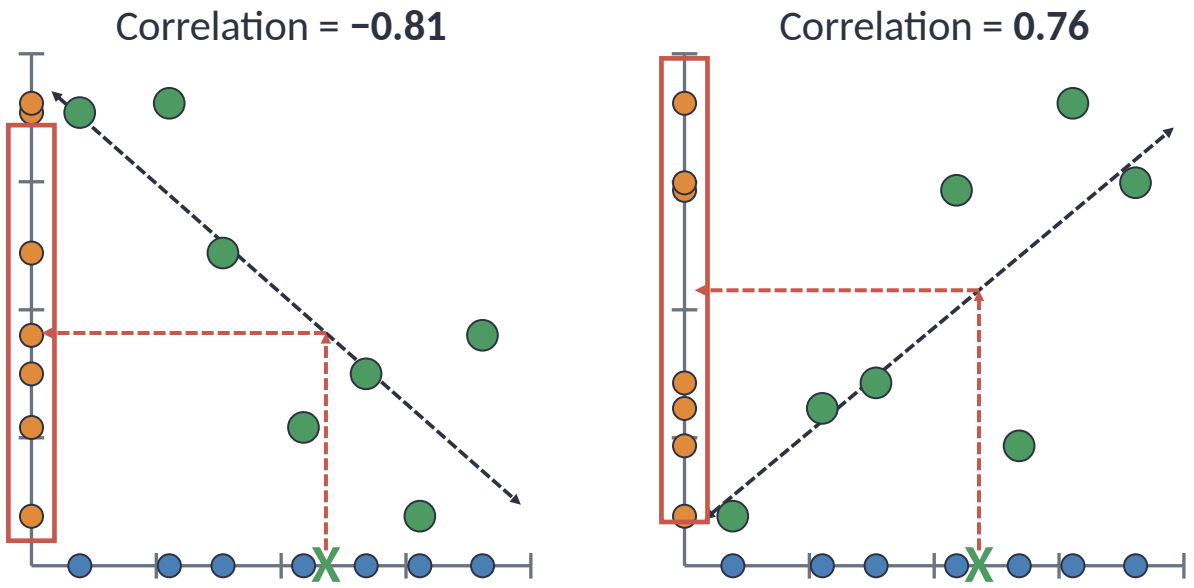


Guesses improve as correlation approaches ±1

As long as the correlation value is not 0…

…we can still use the line to make inferences…

…but our guesses become more refined the closer the correlation values get to **−1** or **1**.

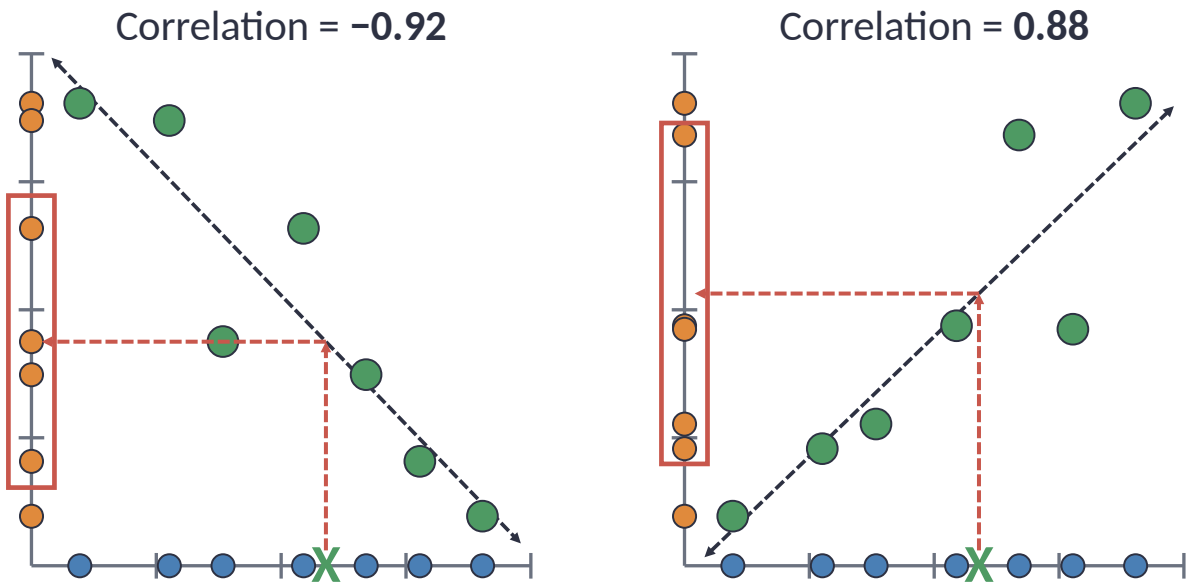


Guesses improve as correlation approaches ±1

As long as the correlation value is not 0…

…we can still use the line to make inferences…

…but our guesses become more refined the closer the correlation values get to **−1** or **1**.

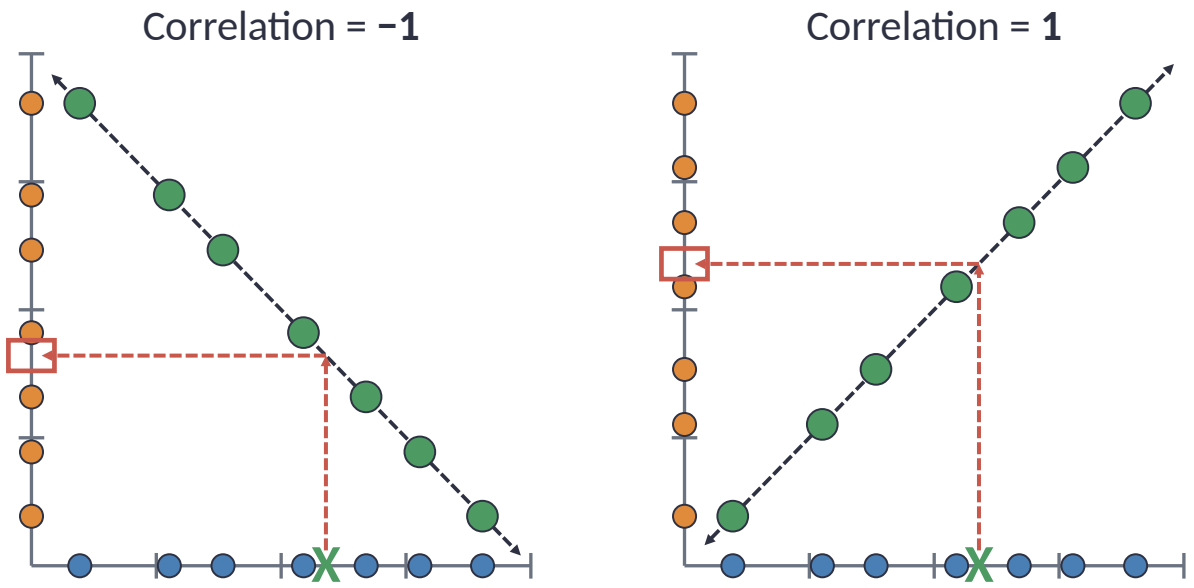


Guesses improve as correlation approaches ±1

As long as the correlation value is not 0…

…we can still use the line to make inferences…

…but our guesses become more refined the closer the correlation values get to **−1** or **1**.



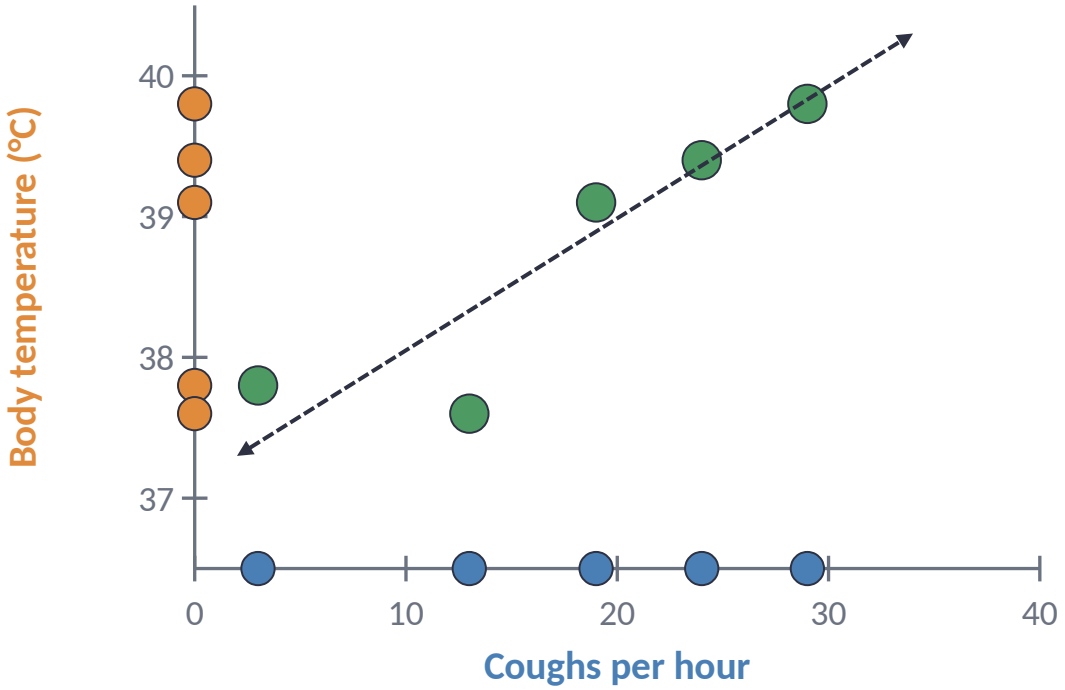
Calculating correlation

If you know how to calculate the **variance** and the **covariance**, calculating the **correlation** is a easy.

As we saw in the covariance slides, the numerator can be any value between minus infinity and plus infinity, depending on…

…whether the slope of the line that represents the relationship is positive or negative…

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$



Calculating correlation

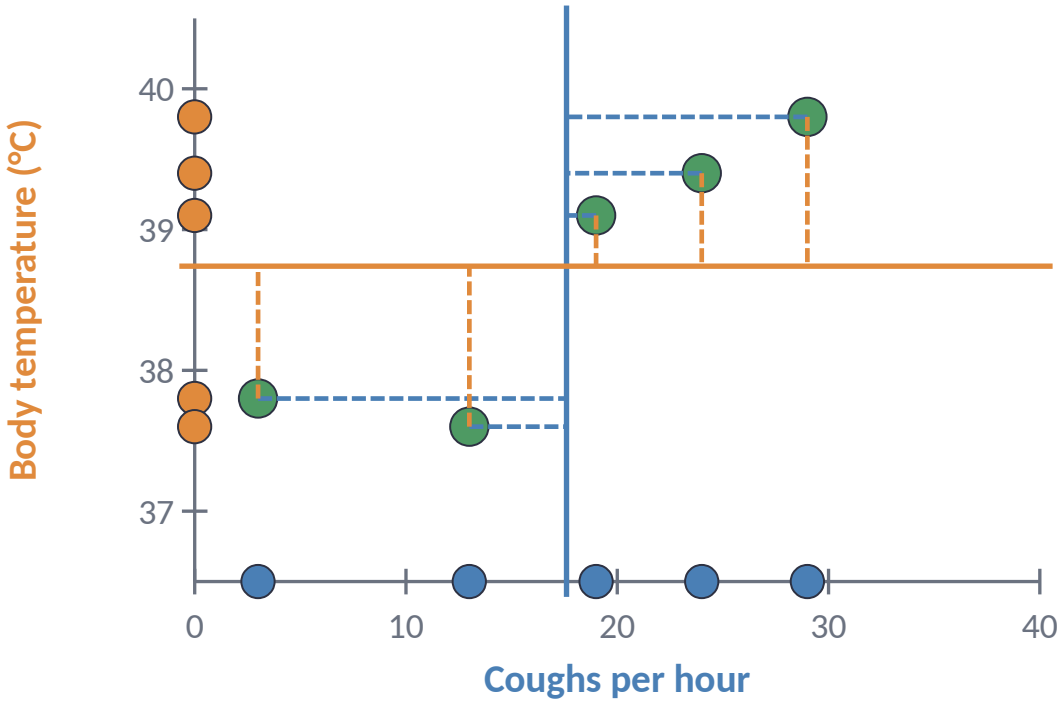
If you know how to calculate the **variance** and the **covariance**, calculating the **correlation** is a easy.

As we saw in the covariance slides, the numerator can be any value between minus infinity and plus infinity, depending on…

…whether the slope of the line that represents the relationship is positive or negative…

…how far the data are spread out around the means…

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$



Calculating correlation

If you know how to calculate the **variance** and the **covariance**, calculating the **correlation** is a easy.

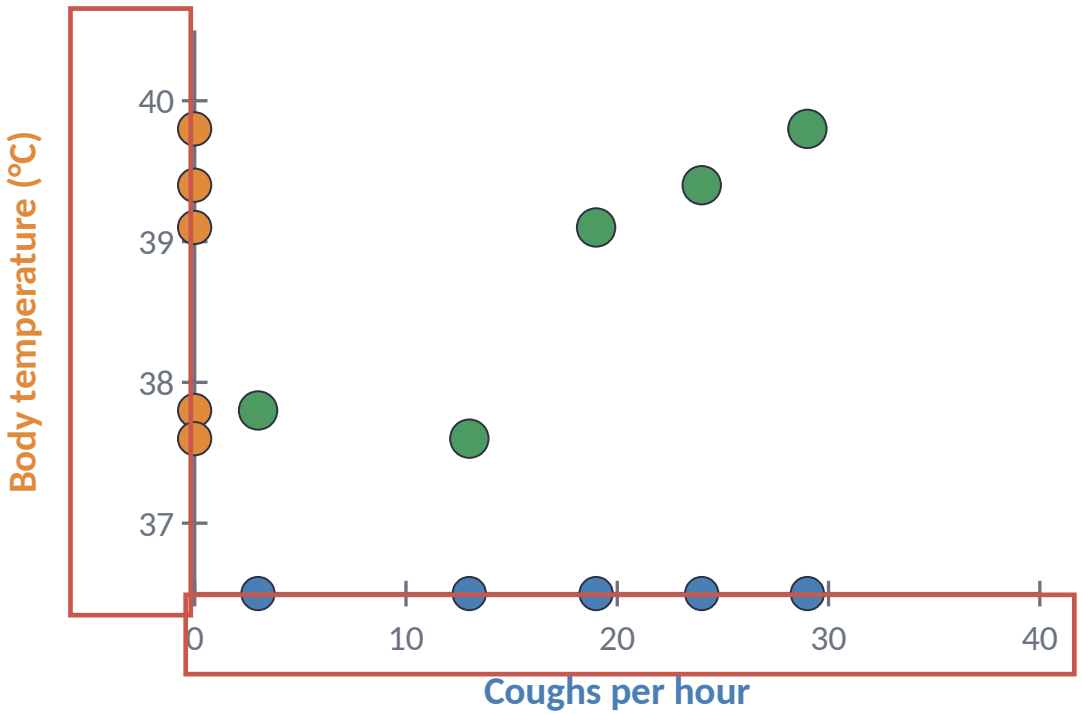
As we saw in the covariance slides, the numerator can be any value between minus infinity and plus infinity, depending on…

…whether the slope of the line that represents the relationship is positive or negative…

…how far the data are spread out around the means…

…and the scale of the data.

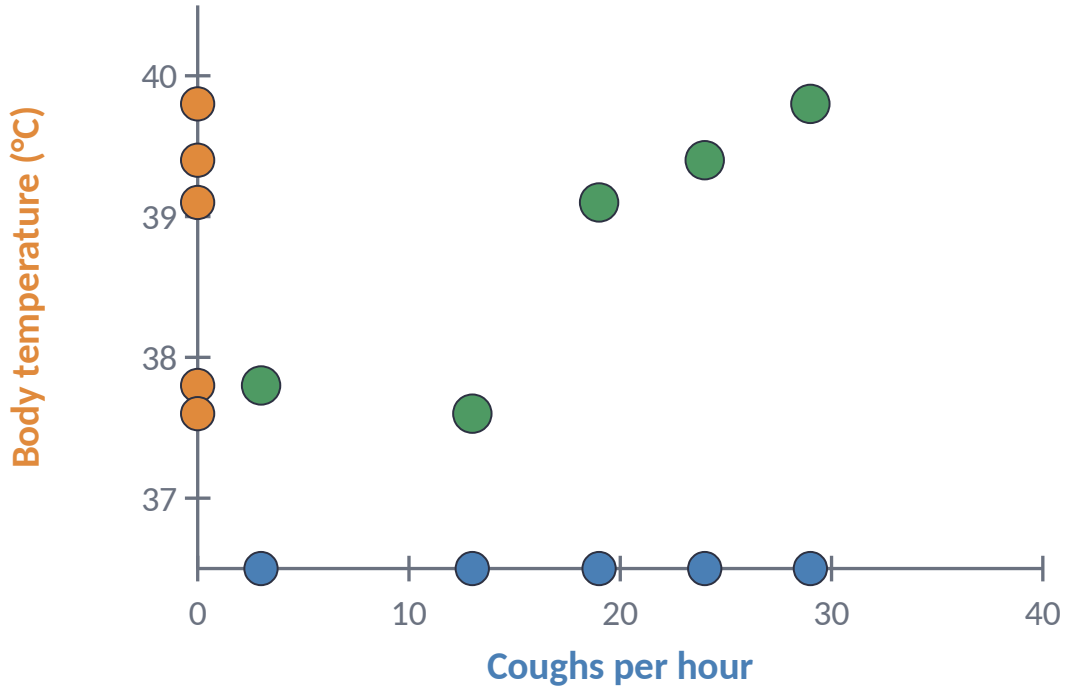
$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$



Calculating correlation

The denominator ensures that the scale of the data does not affect the correlation value, and this makes correlations much easier to interpret.

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$



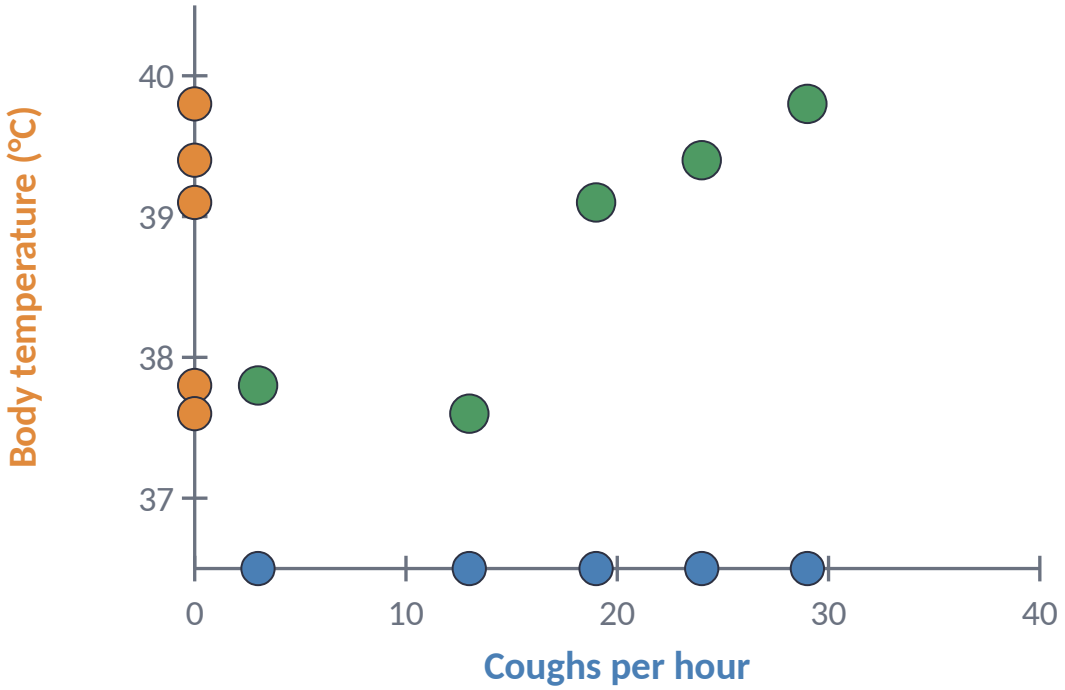
Calculating correlation

The denominator ensures that the scale of the data does not affect the correlation value, and this makes correlations much easier to interpret.

When the data all fall on a straight line, the covariance and the product of the square roots of the variances are the same size, so the division gives **1** or **−1**, depending on the slope.

When the data do not all fall on a straight line, the covariance accounts for less of the variance in the data, and the correlation is closer to **0**.

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$



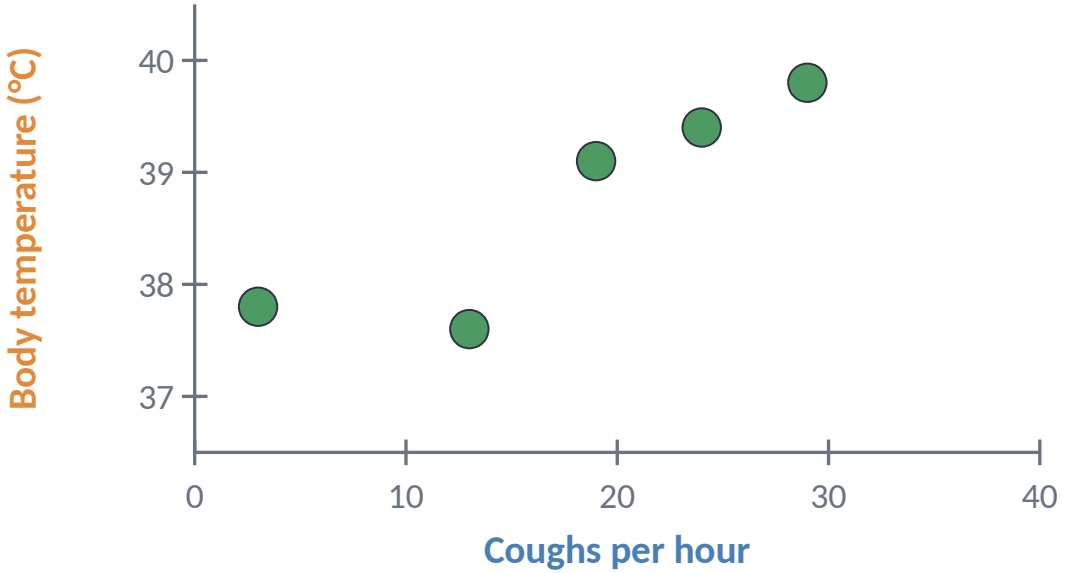
Plugging in the numbers

Let us plug in the numbers for our 5 patients.

As we saw in the covariance slides, the **covariance** between coughs and temperature is **8.9**…

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$

$$= \frac{8.9}{\sqrt{101.8} \times \sqrt{0.97}}$$



Plugging in the numbers

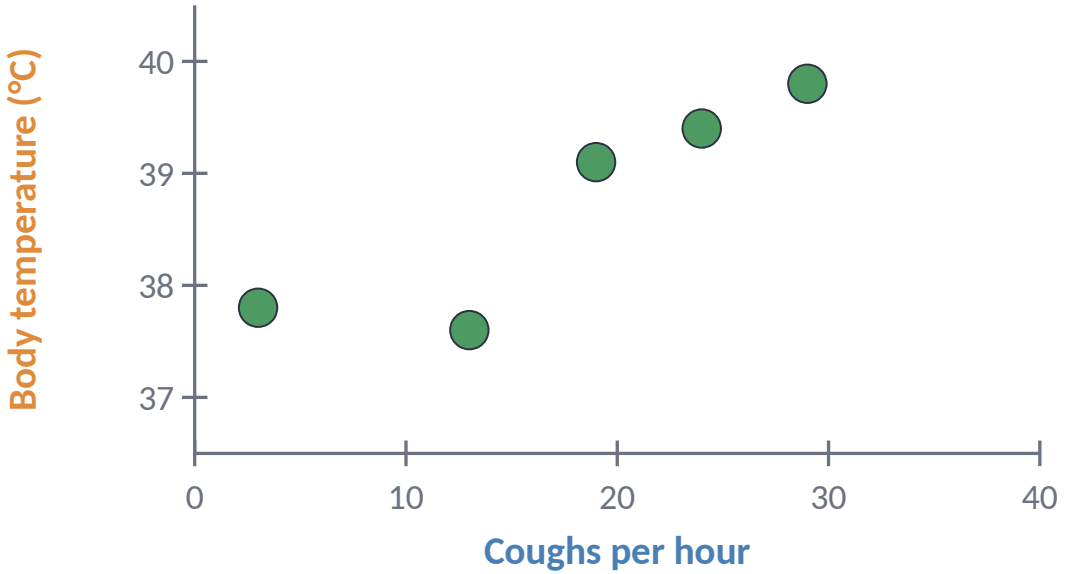
Let us plug in the numbers for our 5 patients.

As we saw in the covariance slides, the **covariance** between coughs and temperature is **8.9**…

…the estimated **variance** of the coughs is **101.8**, and the estimated variance of the temperature is **0.97**…

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$

$$= \frac{8.9}{\sqrt{101.8} \times \sqrt{0.97}}$$



Plugging in the numbers

Let us plug in the numbers for our 5 patients.

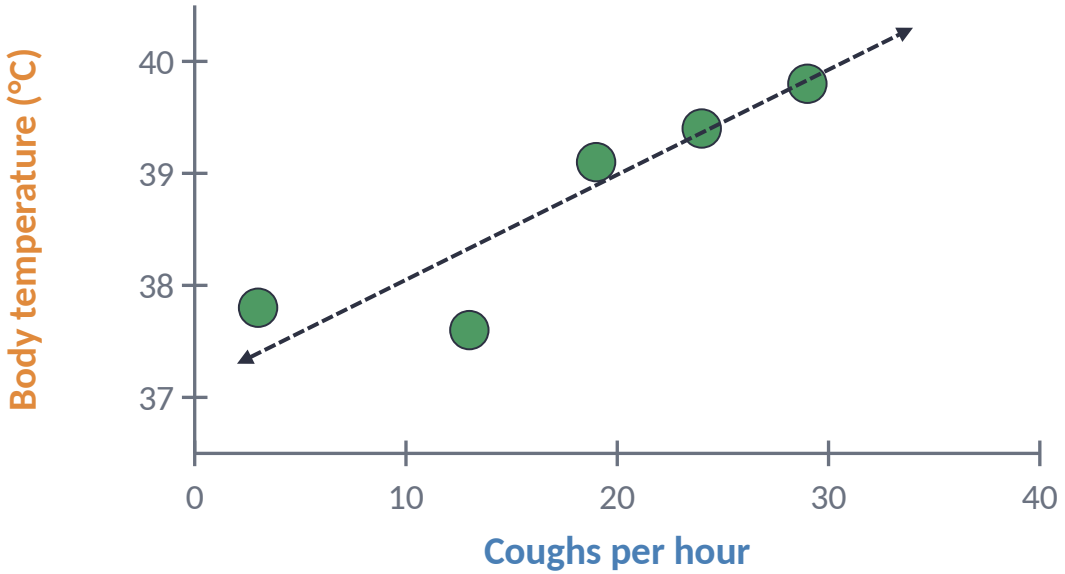
As we saw in the covariance slides, the **covariance** between coughs and temperature is **8.9**…

…the estimated **variance** of the coughs is **101.8**, and the estimated variance of the temperature is **0.97**…

…so the denominator squeezes 8.9 down to a value between −1 and 1. The **correlation** is **0.90**.

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$

$$= \frac{8.9}{\sqrt{101.8} \times \sqrt{0.97}} = \boxed{0.90}$$



Plugging in the numbers

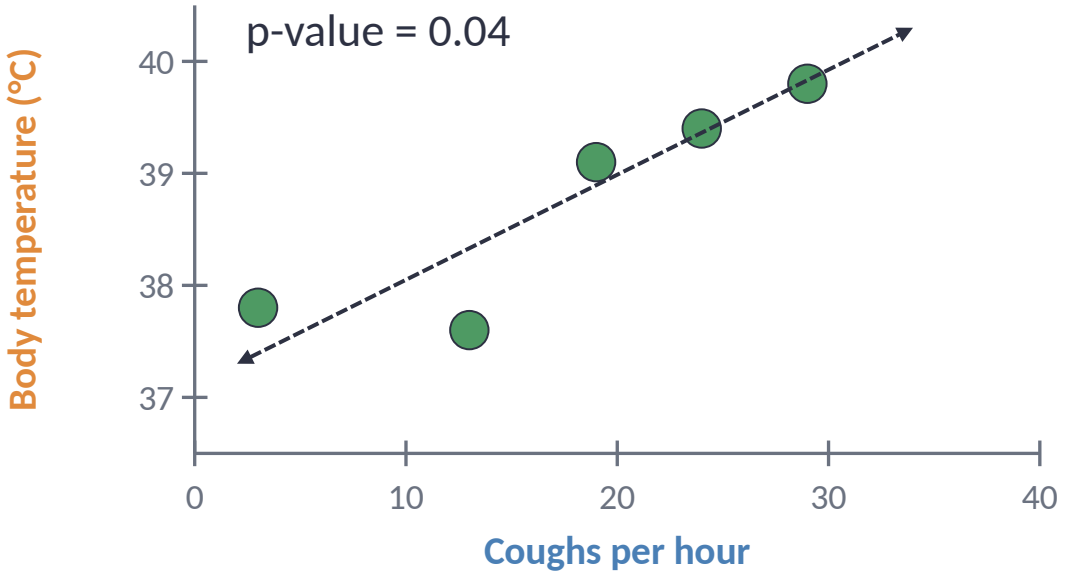
As mentioned earlier, we can quantify our confidence in this relationship with a p-value. Here, the **p-value = 0.04**.

Note: if we counted the coughs over two hours instead of one, the covariance would double to 17.9 and the variance of the coughs would be 4 times larger, 407.2. The correlation would still be **0.90**: unlike covariance, correlation does not depend on the scale.

Likewise, the correlation between coughs and hours of sleep is **−0.84**: a fairly strong negative relationship.

$$\text{Correlation} = \frac{\text{Covariance(Coughs, Temperature)}}{\sqrt{\text{Variance(Coughs)}} \times \sqrt{\text{Variance(Temperature)}}}$$

$$= \frac{8.9}{\sqrt{101.8} \times \sqrt{0.97}} = \boxed{0.90}$$



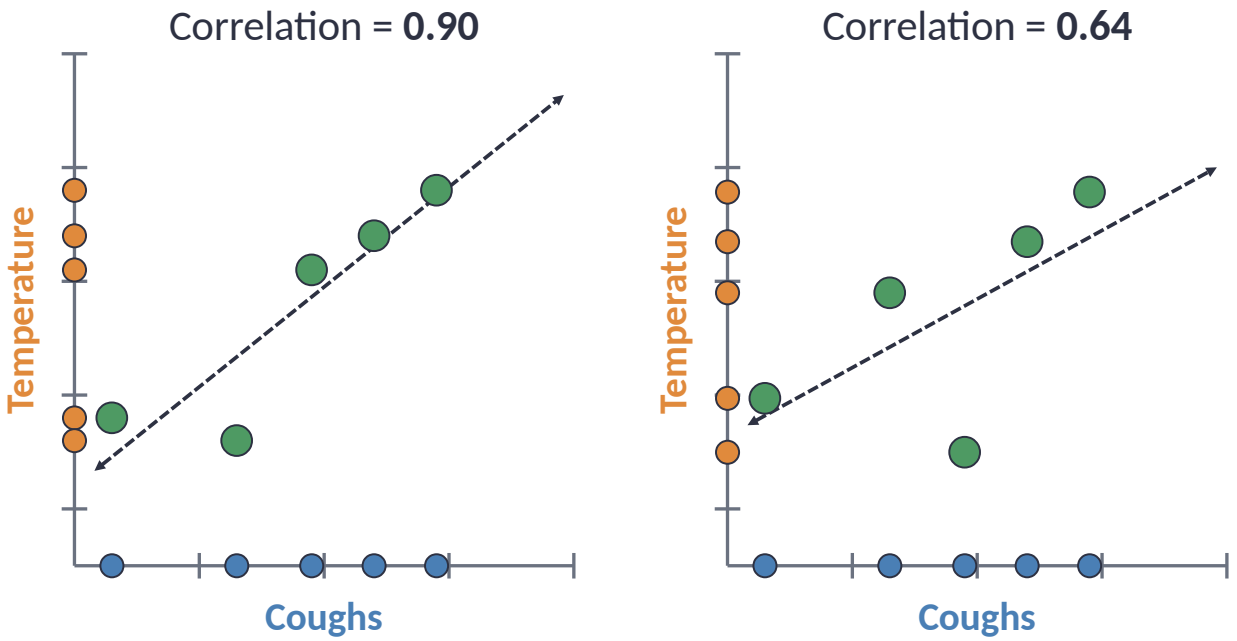
Correlation is still not that easy to interpret

Even though correlation values are much easier to interpret than covariance values, they are still not *that* easy to interpret.

For example, it is not obvious that this relationship, where correlation = 0.90, is about *twice* as good at making predictions as…

…this relationship, where correlation = 0.64.

The good news is that **R^2**, which is related to correlation, solves this problem: $0.90^2 = 0.81$ and $0.64^2 = 0.41$, roughly twice as much.

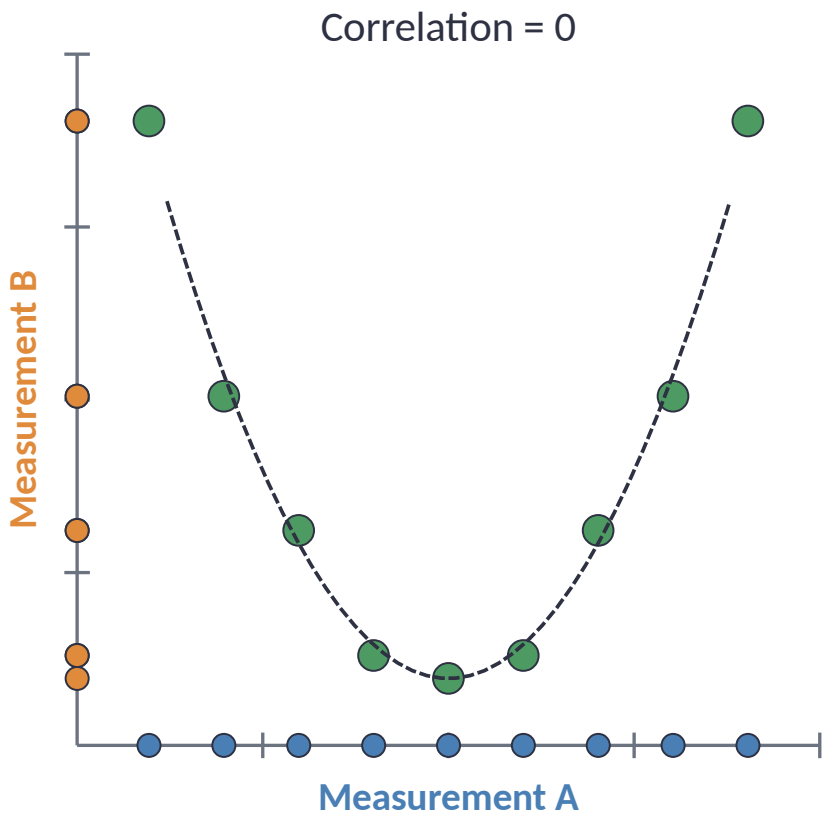


Correlation only measures straight lines

P.S. Here is a clear relationship between two measurements: B is high when A is either low or high.

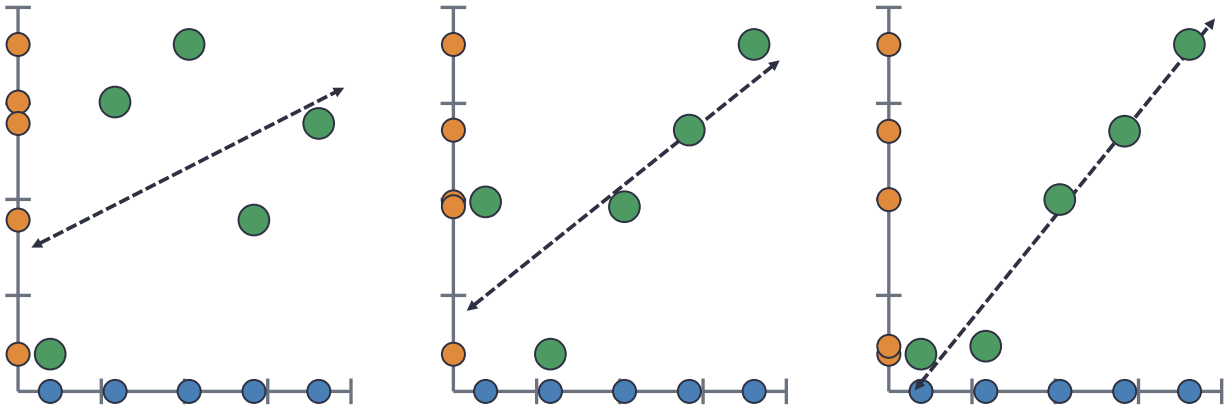
However, the relationship is a curve, not a straight line, so the correlation is **0**.

Correlation = 0 means that there is no *straight-line* relationship; it does not mean that there is no relationship at all. Plotting the data first is always a good idea.



In summary

Correlation quantifies the strength of relationships.



Weak relationship
=
small correlation value

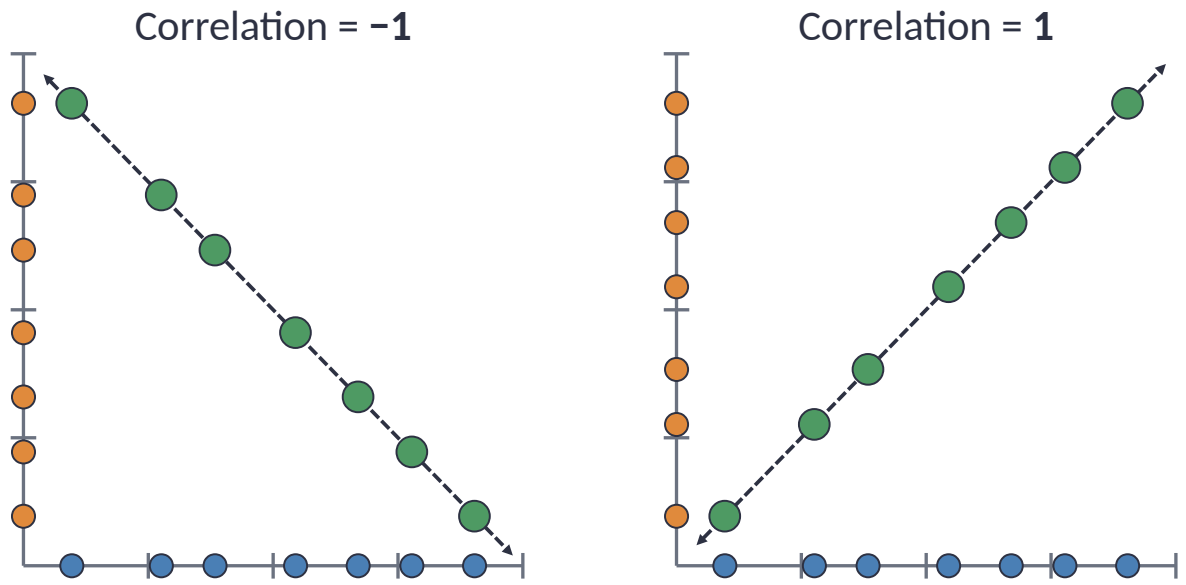
Moderate relationship
=
moderate correlation value

Strong relationship
=
large correlation value

In summary

Correlation quantifies the strength of relationships.

Correlation values go from **−1**, the strongest linear relationship with a negative slope, to **1**, the strongest linear relationship with a positive slope.

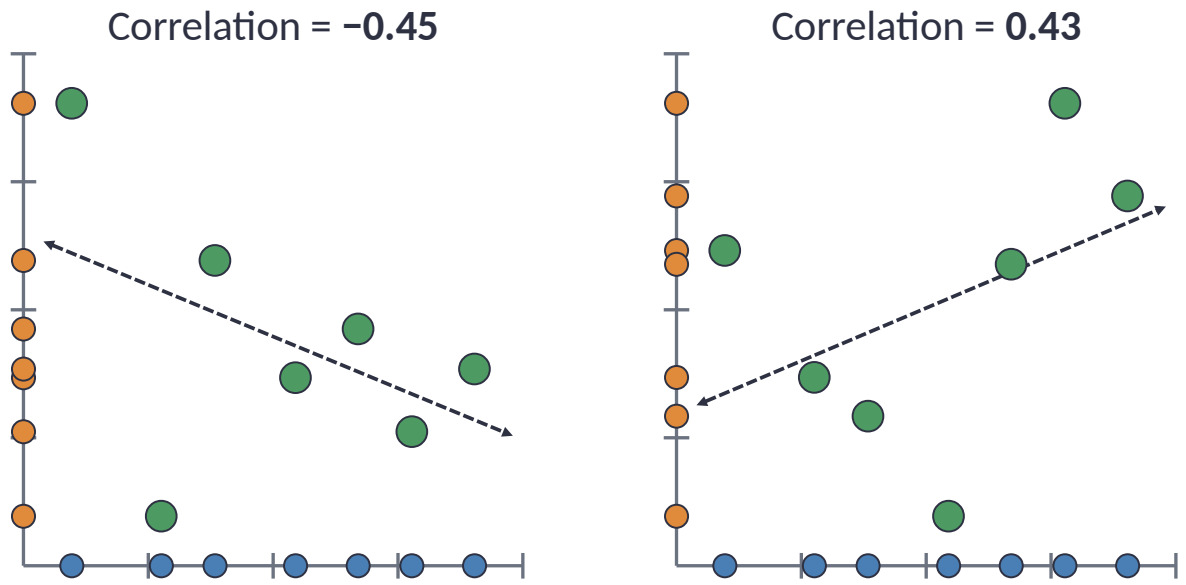


In summary

Correlation quantifies the strength of relationships.

Correlation values go from **−1**, the strongest linear relationship with a negative slope, to **1**, the strongest linear relationship with a positive slope.

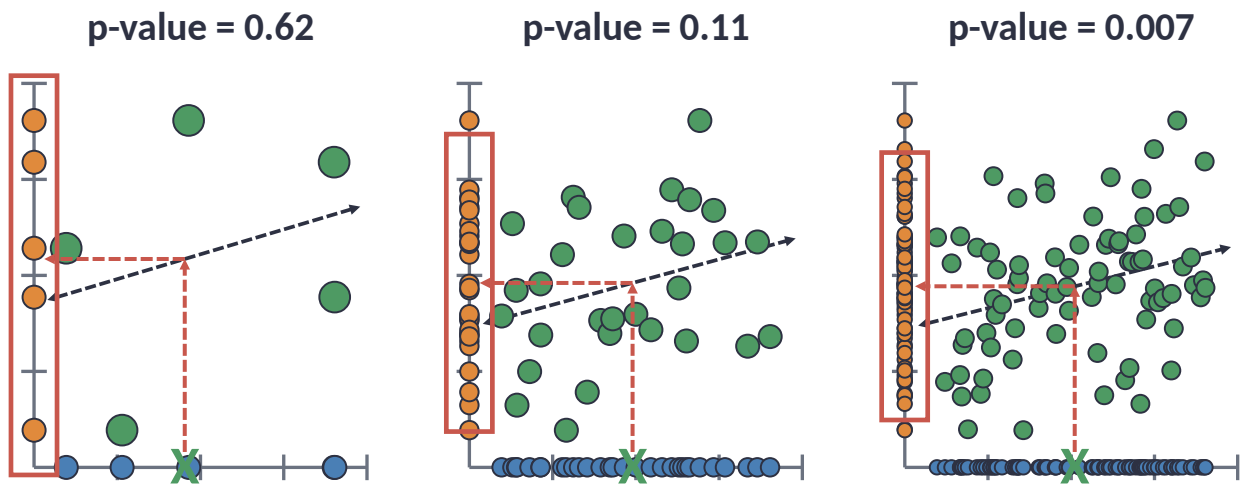
If a straight line cannot go through all the data, the correlation is closer to 0, and the worse the fit, the closer it gets. With no straight-line relationship, correlation = 0.



In summary

Lastly, our confidence in our inferences depends on the amount of data we have collected and on the p-value: the more data, the smaller the p-value and the more confidence we have.

Next, we will see how these ideas, variance, covariance and correlation, come together in **PCA**.



Reading

James, Witten, Hastie, Tibshirani and Taylor (2023). *"**An Introduction to Statistical Learning with Applications in Python (ISLP)**"*.

- 3.1.3

Hastie, Tibshirani and Friedman (2009). "***The Elements of Statistical Learning (ESL), 2nd edition.***"

- 3.2

Murphy, "***Probabilistic Machine Learning: An Introduction"***, MIT Press, 2022. at probml.github.io.

- §3.1, covariance and correlation.