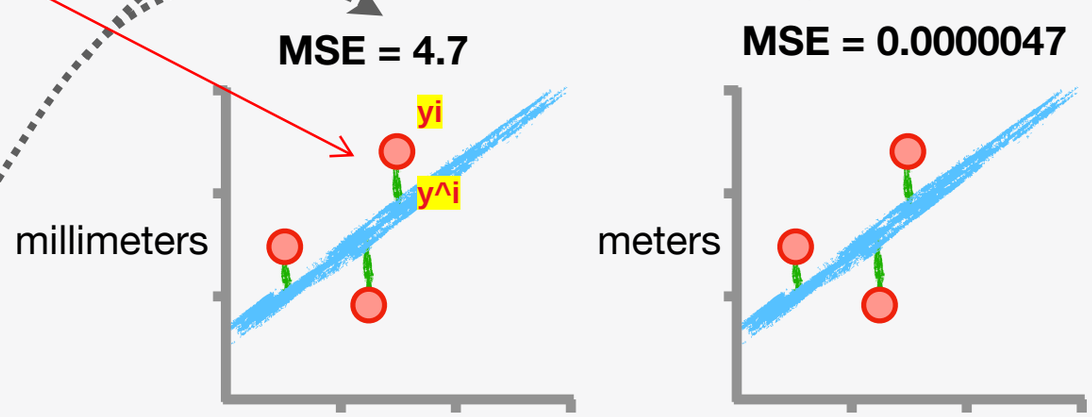
R^2: Main Ideas

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

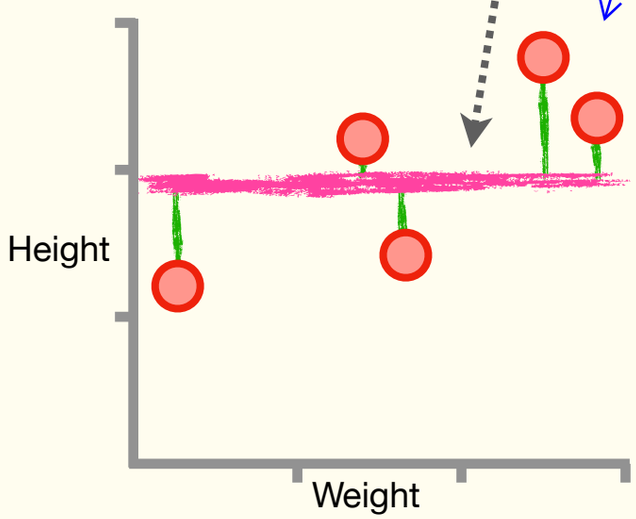
1 **The Problem:** As we just saw, the **MSE**, although totally cool, can be difficult to interpret because it depends, in part, on the scale of the data.

In this example, changing the units from millimeters to meters reduced the **MSE** by a lot!!!

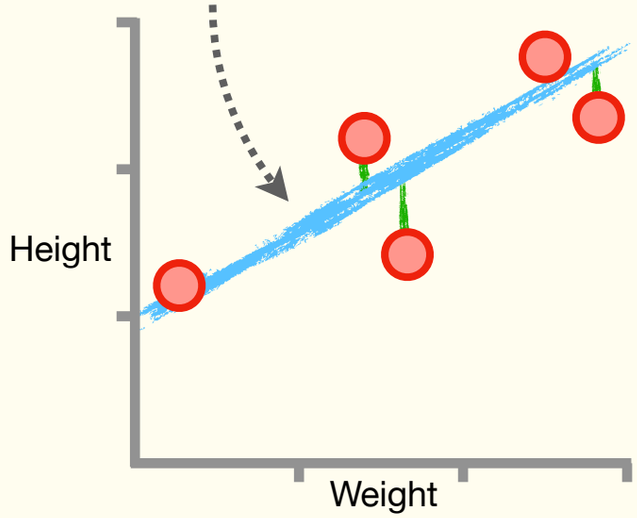


2 **A Solution:** ***R*²**, pronounced **R squared**, is a simple, easy-to-interpret metric that does not depend on the size of the dataset or its scale.

Typically, ***R*²** is calculated by comparing the **SSR** or **MSE** around the **mean** y-axis value. In this example, we calculate the **SSR** or **MSE** around the **average Height**…



…and compare it to the **SSE** or **MSE** around the model we're interested in. In this case, that means we calculate the **SSR** or **MSE** around the **blue line** that uses **Weight** to predict **Height**.



***R*²** then gives us a percentage of how much the predictions improved by using the model we're interested in instead of just the **mean**.

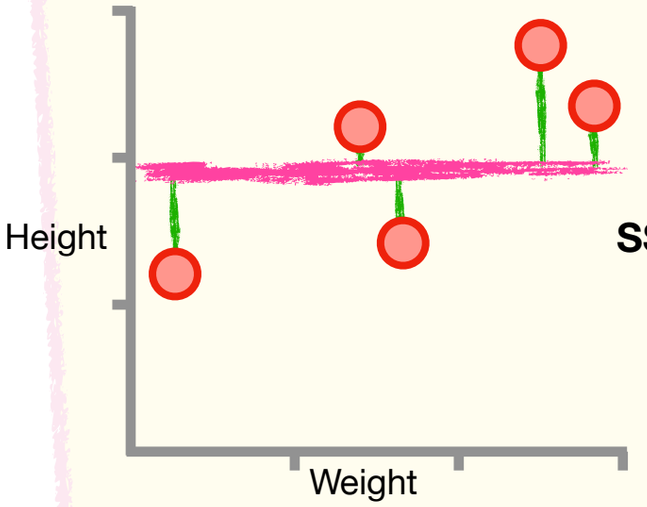
In this example, ***R*²** would tell us how much better our predictions are when we use the **blue line**, which uses Weight to predict Height, instead of predicting that everyone has the **average** Height.

***R*²** values go from **0** to **1** and are interpreted as percentages, and the closer the value is to **1**, the better the model fits the data relative to the mean y-axis value.

Now that we understand the main ideas, let's dive into the details!!!

R^2: Details Part 1

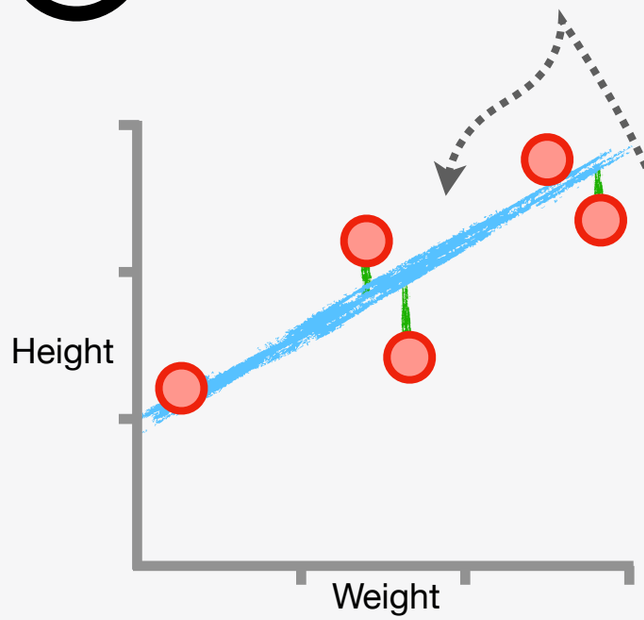
1 First, we calculate the **Sum of the Squared Residuals** for the **mean**. We'll call this **SSR** the **SSR(mean)**. In this example, the mean Height is **1.9** and the **SSR(mean) = 1.6**.



SSR(mean) = $(2.3 - 1.9)^2 + (1.2 - 1.9)^2 + (2.7 - 1.9)^2 + (1.4 - 1.9)^2 + (2.2 - 1.9)^2$

= 1.6

2 Then, we calculate the **SSR** for the **fitted line**, **SSR(fitted line)**, and get **0.5**.



NOTE: The smaller **Residuals** around the **fitted line**, and thus the smaller **SSR** given the same dataset, suggest the **fitted line** does a better job making predictions than the **mean**.

SSR(fitted line) = $(1.2 - 1.1)^2 + (2.2 - 1.8)^2 + (1.4 - 1.9)^2 + (2.7 - 2.4)^2 + (2.3 - 2.5)^2$

= 0.5

3 Now we can calculate the R^2 value using a surprisingly simple formula…

$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$

$$= \frac{1.6 - 0.5}{1.6}$$

= 0.7

…and the result, **0.7**, tells us that there was a **70%** reduction in the size of the **Residuals** between the **mean** and the **fitted line**.

4 In general, because the numerator for R^2…

SSR(mean) - SSR(fitted line)

…is the amount by which the **SSRs** shrank when we fitted the line, R^2 values tell us the percentage the **Residuals** around the **mean** shrank when we used the **fitted line**.

When **SSR(mean) = SSR(fitted line)**, then both models' predictions are equally good (or equally bad), and $R^2 = 0$

$$\frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}} = \frac{0}{\text{SSR(mean)}} = 0$$

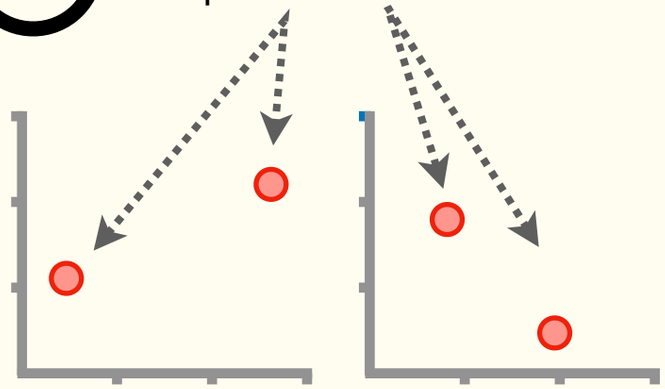
When **SSR(fitted line) = 0**, meaning that the **fitted line** fits the data perfectly, then $R^2 = 1$.

$$\frac{\text{SSR(mean)} - 0}{\text{SSR(mean)}} = \frac{\text{SSR(mean)}}{\text{SSR(mean)}} = 1$$

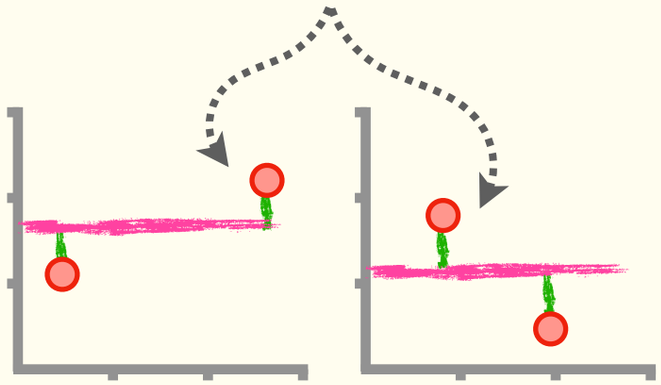
R^2: Details Part 2

$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$

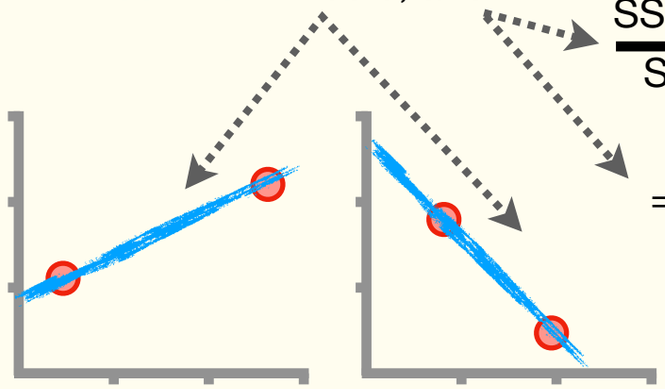
5 **NOTE:** Any **2** random data points have $R^2 = 1$…



…because regardless of the **Residuals** around the **mean**…



…the **Residuals** around a **fitted line** will be **0**, and…

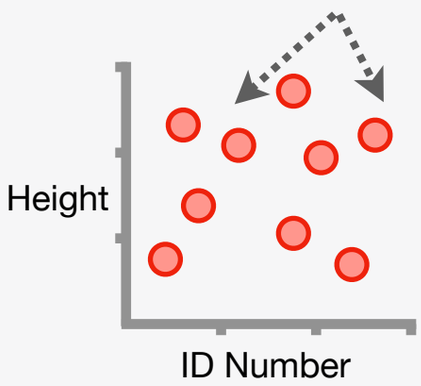


$$\frac{\text{SSR(mean)} - 0}{\text{SSR(mean)}}$$

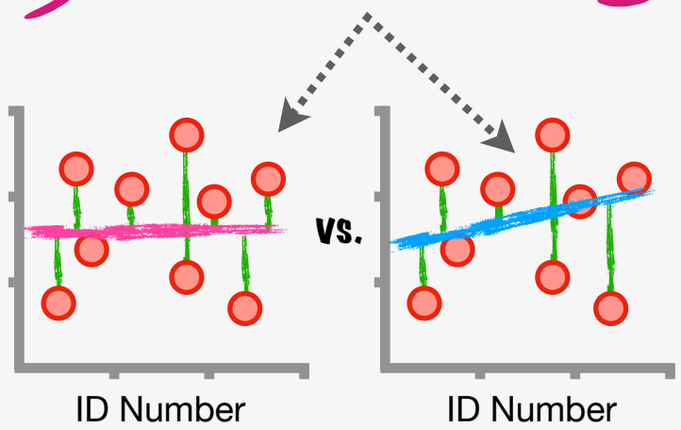
$$= \frac{\text{SSR(mean)}}{\text{SSR(mean)}} = 1$$

Because a small amount of random data can have a high (close to **1**) R^2, any time we see a trend in a small dataset, it's difficult to have confidence that a high R^2 value is not due to random chance.

6 If we had a lot of data organized randomly using a random **ID Number**, we would expect the graph to look like this…

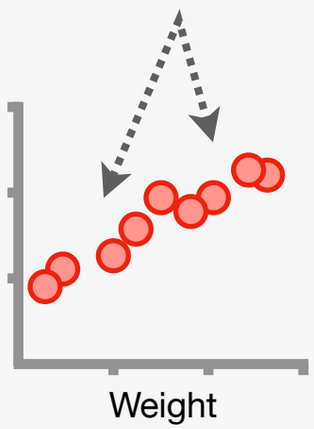


…and have a relatively small (close to **0**) R^2 because the **Residuals** would be similar.

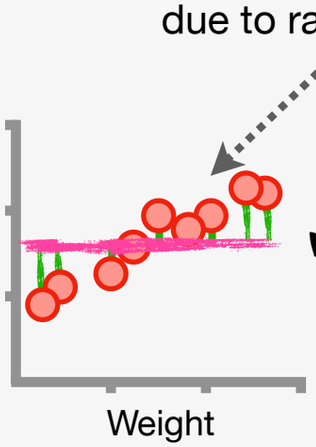


vs.

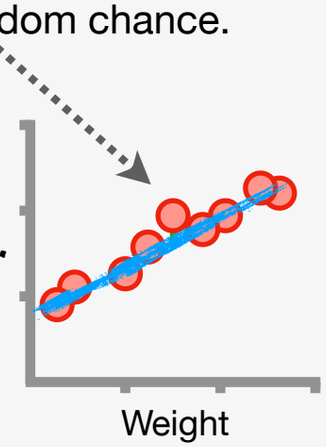
In contrast, when we see a trend in a large amount of data like this…



…we can, intuitively, have more confidence that a large R^2 is not due to random chance.



vs.



7 Never satisfied with intuition, statisticians developed something called ***p*-values** to quantify how much confidence we should have in R^2 values and pretty much any other method that summarizes data. We'll talk about ***p*-values** in a bit, but first let's calculate R^2 using the **Mean Squared Error** (**MSE**).

Calculating R^2 with the Mean Squared Error (MSE): Details

So far, we've calculated R^2 using the **Sum of the Squared Residuals** (**SSR**), but we can just as easily calculate it using the **Mean Squared Error** (**MSE**).

$$\frac{\text{MSE(mean)} - \text{MSE(fitted line)}}{\text{MSE(mean)}}$$

First, we rewrite the **MSE** in terms of the **SSR** divided by the size of the dataset, ***n***…

$$= \frac{\frac{\text{SSR(mean)}}{n} - \frac{\text{SSR(fitted line)}}{n}}{\frac{\text{SSR(mean)}}{n}}$$

…then we consolidate all of the division by ***n*** into a single term…

$$= \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}} \times \frac{n}{n}$$

…and since ***n*** divided by ***n*** is **1**…

$$= \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}} \times 1$$

$$= R^2$$

…we end up with R^2 times **1**, which is just R^2. So, we can calculate R^2 with the **SSR** or **MSE**, whichever is readily available. Either way, we'll get the same value.

BAM!!!

Gentle Reminders:

Residual = Observed - Predicted

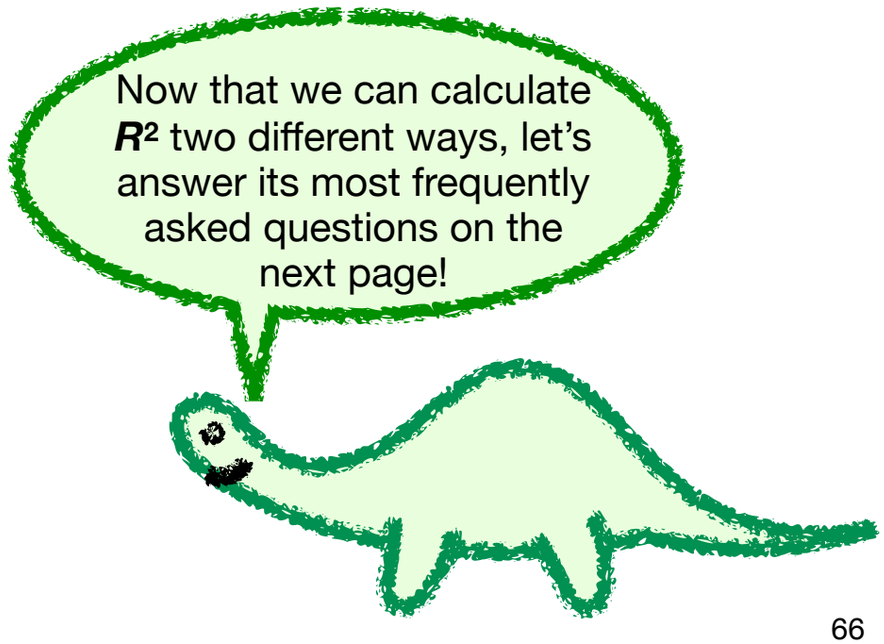
SSR = Sum of Squared Residuals

$$\textbf{SSR} = \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2$$

$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{SSR}}{n}$$

…where ***n*** is the sample size

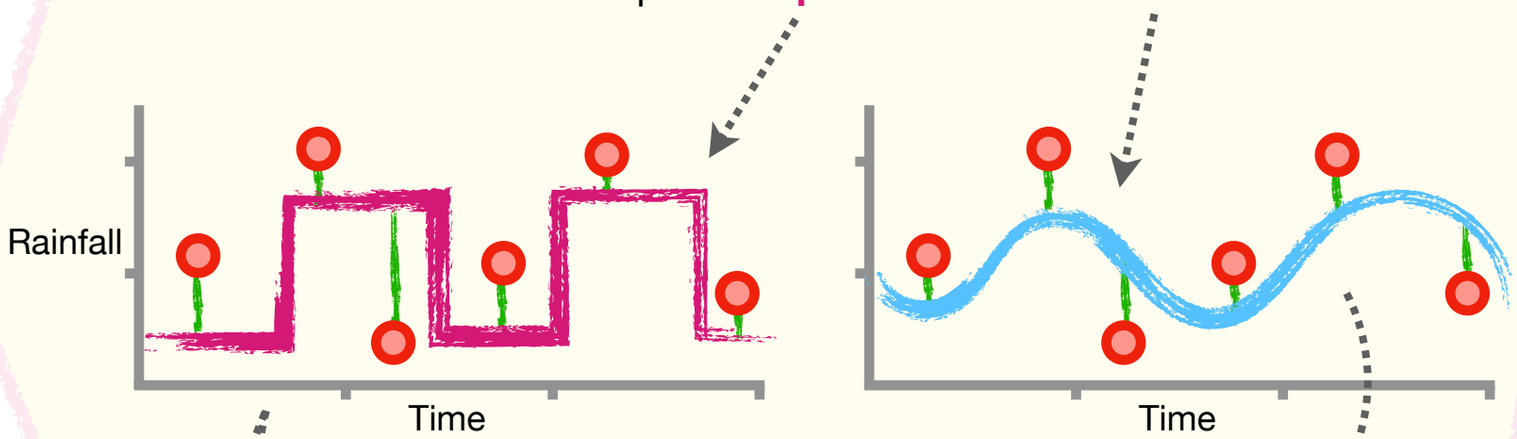
$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$



R^2: FAQ

Does R^2 always compare the mean to a straight fitted line?

The most common way to calculate R^2 is to compare the **mean** to a **fitted line**. However, we can calculate it for anything we can calculate the **Sum of the Squared Residuals** for. For example, for rainfall data, we use R^2 to compare a **square wave** to a **sine wave**.



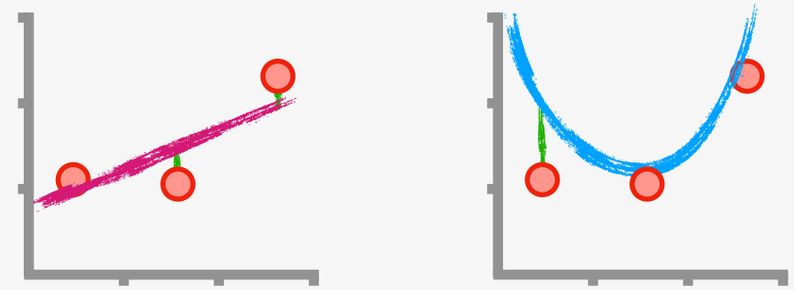
In this case, we calculate R^2 based on the **Sum of the Squared Residuals** around the **square** and **sine** waves.

$$R^2 = \frac{\text{SSR(square)} - \text{SSR(sine)}}{\text{SSR(square)}}$$

Can R^2 be negative?

When we're only comparing the **mean** to a **fitted line**, R^2 is positive, but when we compare other types of models, anything can happen.

For example, if we use R^2 to compare a **straight line** to a **parabola**…



RSS(straight line) = 5 **RSS(parabola) = 12**

$$R^2 = \frac{\text{SSR(line)} - \text{SSR(parabola)}}{\text{SSR(line)}}$$

$$R^2 = \frac{5 - 12}{5} = -1.4$$

…we get a negative R^2 value, **-1.4**, and it tells us the **Residuals** *increased* by **140%**.

Is R^2 related to Pearson's correlation coefficient?

Yes! If you can calculate **Pearson's correlation coefficient**, ρ (the Greek character **rho**) or r, for a relationship between two things, then the square of that coefficient is equal to R^2. In other words…

$$\rho^2 = r^2 = R^2$$

…and now we can see where R^2 got its name.

BAM!!!

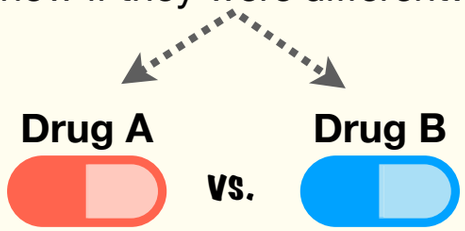
p-values: Main Ideas Part 1

1 **The Problem:** We need to quantify how confident we should be in the results of our analysis.

2 **A Solution: *p*-values** give us a measure of confidence in the results from a statistical analysis.

NOTE: Throughout the description of ***p*-values**, we'll only focus on determining whether or not Drug A is *different* from Drug B. If a ***p*-value** allows us to establish a difference, then we can worry about whether Drug A is better or worse than Drug B.

Imagine we had two antiviral drugs, **A** and **B**, and we wanted to know if they were *different*.



So we gave Drug A to **1** person and they were cured…

…and we gave Drug B to another person and they were not cured.

Can we conclude that Drug A is different from Drug B?

No!!! Drug B may have failed for a lot of reasons. Maybe this person is taking a medication that has a bad interaction with Drug B, or maybe they have a rare allergy to Drug B, or maybe they didn't take Drug B properly and missed a dose.

Or maybe Drug A doesn't actually work, and the placebo effect deserves all of the credit.

There are a lot of weird, random things that can happen when doing a test, and this means that we need to test each drug on more than just one person.

3 So, we redid the experiment with lots and lots of people, and these were the results: Drug A cured a lot of people compared to Drug B, which hardly cured anyone.

Drug A

Cured!!!	Not Cured
1,043	3

Drug B

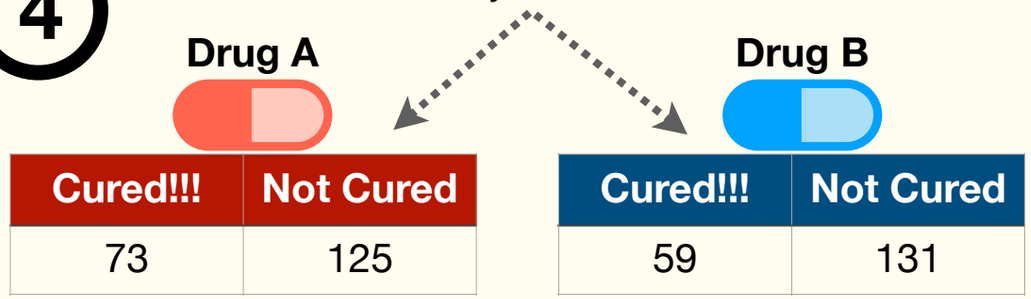
Cured!!!	Not Cured
2	1,432

Now, it's pretty obvious that Drug A is different from Drug B because it would be unrealistic to suppose that these results were due to just random chance and that there's no real difference between Drug A and Drug B.

It's possible that some of the people taking Drug A were actually cured by placebo, and some of the people taking Drug B were not cured because they had a rare allergy, but there are just too many people cured by Drug A, and too few cured by Drug B, for us to seriously think that these results are just random and that Drug A is no different from Drug B.

p-values: Main Ideas Part 2

4 In contrast, let's say that these were the results…



Cured!!!	Not Cured
73	125

Cured!!!	Not Cured
59	131

…and **37%** of the people who took Drug A were cured compared to **31%** who took Drug B.

Drug A cured a larger percentage of people, but given that no study is perfect and there are always a few random things that happen, how confident can we be that Drug A is different from Drug B?

This is where ***p*-values** come in. ***p*-values** are numbers between **0** and **1** that, in this example, quantify how confident we should be that Drug A is different from Drug B. The closer a **p-value** is to **0**, the more confidence we have that Drug A and Drug B are different.

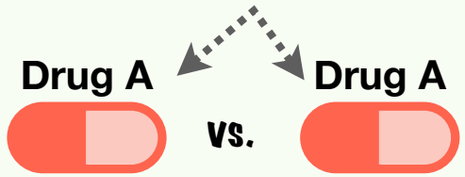
So, the question is, "how small does a ***p*-value** have to be before we're sufficiently confident that Drug A is different from Drug B?"

In other words, what threshold can we use to make a good decision about whether these drugs are different?

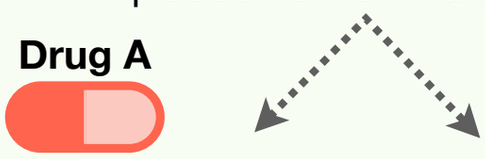
5 In practice, a commonly used threshold is **0.05**. It means that if there's no difference between Drug A and Drug B, and if we did this exact same experiment a bunch of times, then only **5%** of those experiments would result in the wrong decision.

Yes! This wording is awkward. So, let's go through an example and work this out, one step at a time.

6 Imagine we gave the *same* drug, Drug A, to two different groups.



Now, the differences in the results can definitely be attributed to weird, random things, like a rare allergy in one person or a strong placebo effect in another.



Cured!!!	Not Cured
73	125

***p* = 0.9**

Cured!!!	Not Cured
71	127

When we calculate the ***p*-value** for these data using a **Statistical Test** (for example, **Fisher's Exact Test**, but we'll save those details for another day) we get **0.9**…

…which is larger than **0.05**. Thus, we would say that we fail to see a difference between these two groups. And that makes sense because both groups are taking Drug A and the only differences are weird, random things like rare allergies.

p-values: Main Ideas Part 3

7 If we repeated this same experiment over and over again, most of the time we would get similarly large ***p*-values**…

Drug A

Cured!!!	Not Cured
73	125

***p* = 0.9**

Drug A

Cured!!!	Not Cured
71	127

Cured!!!	Not Cured
71	127

***p* = 1**

Cured!!!	Not Cured
72	126

Cured!!!	Not Cured
75	123

***p* = 0.7**

Cured!!!	Not Cured
70	128

etc. etc. etc.

etc. etc. etc.

etc. etc. etc.

Cured!!!	Not Cured
69	129

***p* = 0.9**

Cured!!!	Not Cured
71	127

8 However, every once in a while, by random chance, all of the people with rare allergies might end up in the group on the left…

Drug A

Cured!!!	Not Cured
60	138

30% Cured

Drug A

Cured!!!	Not Cured
84	114

42% Cured

…and by random chance, all of the people with strong (positive) placebo reactions might end up in the group on the right…

…and, as a result, the ***p*-value** for this specific run of the experiment is **0.01** (calculated using **Fisher's Exact Test**, but we'll save those details for another day), since the results are pretty different.

Thus, because the ***p*-value** is **< 0.05** (the threshold we're using for making a decision), we would say that the two groups are different, even though they both took the same drug!

TERMINOLOGY ALERT!!!

Getting a small ***p*-value** when there is no difference is called a **False Positive**.

A **0.05** threshold for ***p*-values** means that **5%** of the experiments, where the only differences come from weird, random things, will generate a ***p*-value** smaller than **0.05**.

In other words, if there's no difference between Drug A and Drug B, in **5%** of the times we do the experiment, we'll get a ***p*-value** less than **0.05**, and that would be a **False Positive**.

p-values: Main Ideas Part 4

9 If it's extremely important that we're correct when we say the drugs are different, then we can use a smaller threshold, like **0.01** or **0.001** or even smaller.

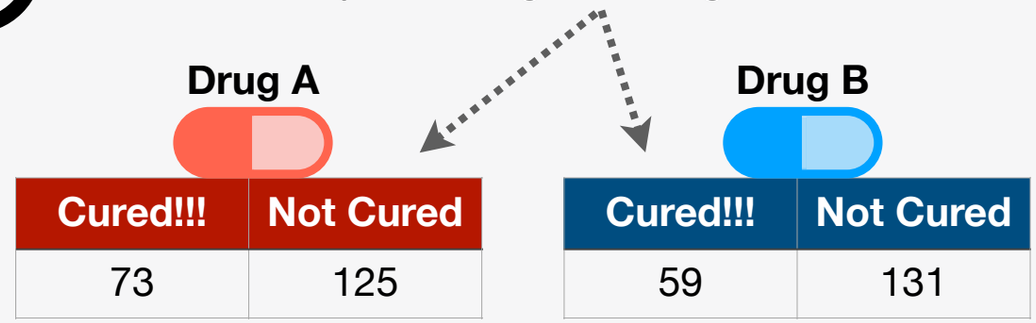
Using a threshold of **0.001** would get a **False Positive** only once in every **1,000** experiments.

Likewise, if it's not that important (for example, if we're trying to decide if the ice-cream truck will arrive on time), then we can use a larger threshold, like **0.2**.

Using a threshold of **0.2** means we're willing to get a **False Positive 2** times out of **10**.

That said, the most common threshold is **0.05** because trying to reduce the number of **False Positives** below **5%** often costs more than it's worth.

10 Now, going back to the original experiment, where we compared Drug A to Drug B…



Drug A: Cured!!!	Drug A: Not Cured	Drug B: Cured!!!	Drug B: Not Cured
73	125	59	131

…if we calculate a ***p*-value** for this experiment and the ***p*-value < 0.05**, then we'll decide that **Drug A** is different from **Drug B**.

That said, the ***p*-value = 0.24**, (again calculated using **Fisher's Exact Test**), so we're not confident that **Drug A** is different from **Drug B**.

TERMINOLOGY ALERT!!!

In fancy statistical lingo, the idea of trying to determine if these drugs are the same or not is called **Hypothesis Testing**.

The **Null Hypothesis** is that the drugs are the same, and the ***p*-value** helps us decide if we should *reject* the **Null Hypothesis**.

BAM!!!

p-values: Main Ideas Part 5

11 While a small ***p*-value** helps us decide if Drug A is different from Drug B, it does not tell us *how different* they are.

In other words, you can have a small ***p*-value** regardless of the size of the difference between Drug A and Drug B.

The difference can be tiny or huge.
For example, this experiment gives us a relatively large ***p*-value**, **0.24**, even though there's a **6**-point difference between Drug A and Drug B.

Drug A

Cured!!!	Not Cured
73	125

37% Cured

Drug B

Cured!!!	Not Cured
59	131

31% Cured

In contrast, this experiment involving a lot more people gives us a smaller ***p*-value**, **0.04**, even though there's only a **1**-point difference between Drug A and Drug B.

Drug A

Cured!!!	Not Cured
5,005	9,868

34% Cured

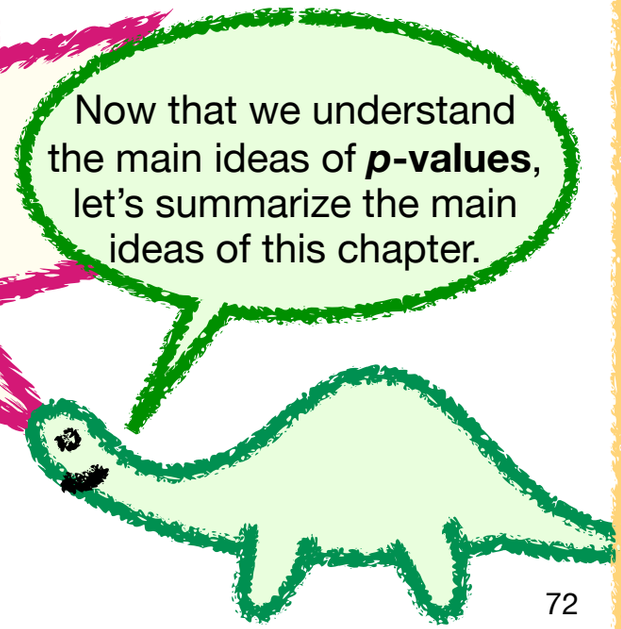
Drug B

Cured!!!	Not Cured
4,800	9,000

35% Cured

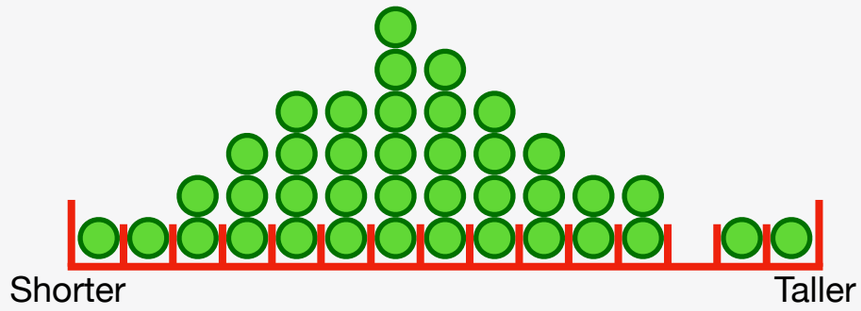
In summary, a small ***p*-value** *does not imply* that the effect size, or difference between Drug A and Drug B, is large.

DOUBLE BAM!!!

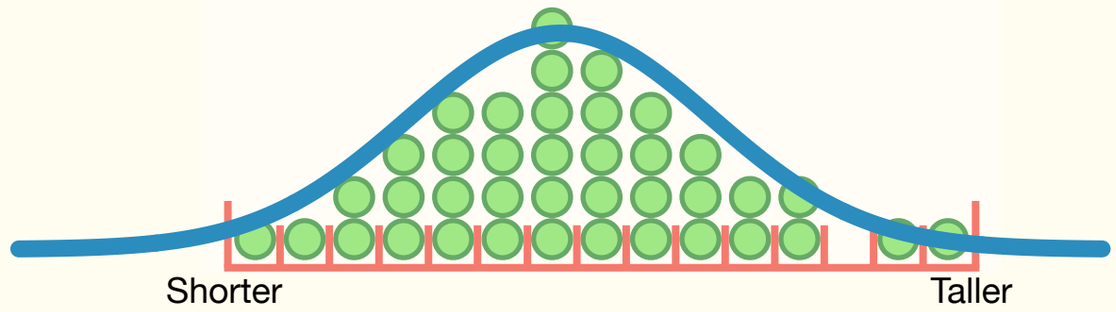


The Fundamental Concepts of Statistics: Summary

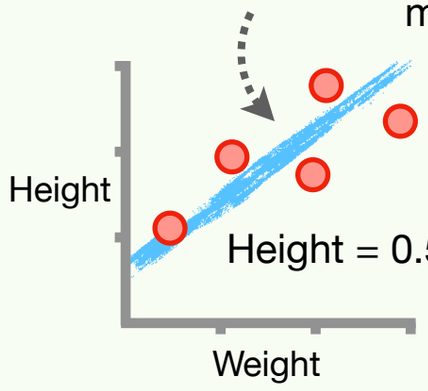
1 We can see trends in data with **histograms**. We'll learn how to use **histograms** to make classifications with **Naive Bayes** in **Chapter 7**.



2 However, **histograms** have limitations (they need a lot of data and can have gaps), so we also use **probability distributions** to represent trends. We'll learn how to use **probability distributions** to make classifications with **Naive Bayes** in **Chapter 7**.



3 Rather than collect all of the data in the whole wide world, which would take forever and be way too expensive, we use **models** to approximate reality. **Histograms** and **probability distributions** are examples of **models** that we can use to make predictions. We can also use a **mathematical formula**, like the equation for the **blue line**, as a **model** that makes predictions.



Throughout this book, we'll create machine learning **models** to make predictions.

Height = 0.5 + (0.8 x Weight)

4 We can evaluate how well a model reflects the data using the **Sum of the Squared Residuals** (**SSR**), the **Mean Squared Error** (**MSE**), and R^2. We'll use these metrics throughout the book.

Residual = Observed - Predicted

SSR = Sum of Squared Residuals

$$\textbf{SSR} = \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2$$

$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{SSR}}{n}$$

…where ***n*** is the sample size

$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$

5 Lastly, we use ***p*-values** to give us a sense of how much confidence we should put in the predictions that our **models** make. We'll use ***p*-values** in **Chapter 4** when we do **Linear Regression**.

TRIPLE BAM!!!