The Binomial Distribution: Details Part 1

1 First, let's focus on just the left-hand side of the equation.

In our pie example, ***x*** is the number of people who prefer pumpkin pie, so in this case, ***x* = 2**…

…***n*** is the number of people we ask. In this case, ***n* = 3**…

…and ***p*** is the probability that someone prefers pumpkin pie. In this case, ***p* = 0.7**…

$$p(x \mid n, p) = \left(\frac{n!}{x!(n-x)!}\right) p^x (1-p)^{n-x}$$

2 …***p*** means **probability**…

…the vertical bar or pipe symbol means **given** or **given that**…

…and the comma between ***n*** and ***p*** means **and**…

So, the left-hand side reads: "The probability we meet ***x* = 2** people who prefer pumpkin pie, given that we ask ***n* = 3** people *and* the probability of someone preferring pumpkin pie is ***p* = 0.7**."

BAM!

Gentle Reminder: We're using the equation for the **Binomial Distribution** to calculate the probability that **2** out of **3** people prefer pumpkin pie…

0.3 x **0.7** x **0.7** = **0.147**

+

0.7 x **0.3** x **0.7** = **0.147**

+

0.7 x **0.7** x **0.3** = **0.147**

= 0.441

The Binomial Distribution: Details Part 2

3 Now, let's look at the first part on the right-hand side of the equation. **StatSquatch** says it looks scary because it has factorials (the exclamation points; see below for details), but it's not that bad.

Despite the factorials, the first term simply represents the number of different ways we can meet **3** people, **2** of whom prefer pumpkin pie…

…and, as we saw earlier, there are **3** different ways that **2** out of **3** people we meet can prefer pumpkin pie.

$$p(x \mid n, p) = \left(\frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}$$

Gentle Reminder: We're using the equation for the **Binomial Distribution** to calculate the probability that **2** out of **3** people prefer pumpkin pie…

0.3 x **0.7** x **0.7** = **0.147**

+

0.7 x **0.3** x **0.7** = **0.147**

+

0.7 x **0.7** x **0.3** = **0.147**

= 0.441

4 When we plug in ***x* = 2**, the number of people who prefer pumpkin pie…

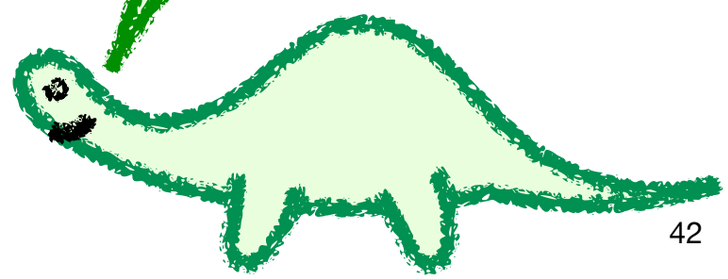
…and ***n* = 3**, the number of people we asked, and then do the math…

…we get **3**, the same number we got when we did everything by hand.

$$\frac{n!}{x!\,(n-x)!} = \frac{3!}{2!\,(3-2)!} = \frac{3!}{2!\,(1)!} = \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = 3$$

NOTE: If ***x*** is the number of people who prefer pumpkin pie, and ***n*** is the total number of people, then **(*n* - *x*)** = the number of people who prefer blueberry pie.

A factorial—indicated by an exclamation point—is just the product of the integer number and all positive integers below it. For example, **3! = 3 x 2 x 1 = 6**.



The Binomial Distribution: Details Part 3

5 Now let's look at the second term on the right hand side.

The second term is just the probability that **2** out of the **3** people prefer pumpkin pie.

In other words, since ***p***, the probability that someone prefers pumpkin pie, is **0.7**…

…and there are ***x* = 2** people who prefer pumpkin pie, the second term **=** 0.7^2 **=** **0.7** **x** **0.7**.

$$p(x \mid n, p) = \left(\frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}$$

6 The third and final term is the probability that **1** out of **3** people prefers blueberry pie…

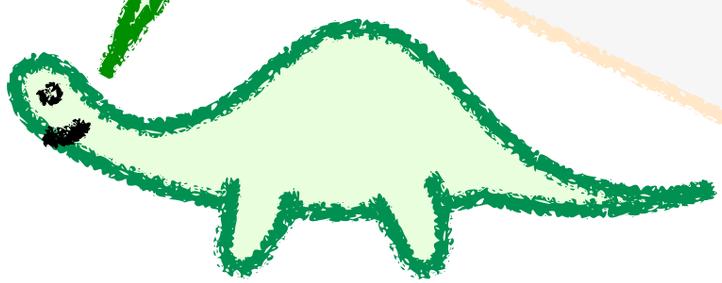
…because if ***p*** is the probability that someone prefers pumpkin pie, **(1 - *p*)** is the probability that someone prefers blueberry pie…

…and if ***x*** is the number of people who prefer pumpkin pie and ***n*** is the total number of people we asked, then ***n* - *x*** is the number of people who prefer blueberry pie.

So, in this example, if we plug in ***p* = 0.7**, ***n* = 3**, and ***x* = 2**, we get **0.3**.

$$(1 - p)^{n - x} = (1 - 0.7)^{3-2} = 0.3^1 = 0.3$$

Just so you know, sometimes people let ***q* = (1 - *p*)**, and use ***q*** in the formula instead of **(1 - *p*)**.



Gentle Reminder: We're using the equation for the **Binomial Distribution** to calculate the probability that **2** out of **3** people prefer pumpkin pie…

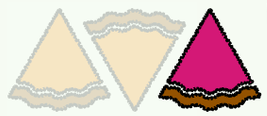
0.3 x **0.7** **x** **0.7** = **0.147**

+

0.7 x **0.3** x **0.7** = **0.147**

+

0.7 x **0.7** x **0.3** = **0.147**



= **0.441**

The Binomial Distribution: Details Part 4

7 Now that we've looked at each part of the equation for the **Binomial Distribution**, let's put everything together and solve for the probability that **2** out of **3** people we meet prefer pumpkin pie.

We start by plugging in the number of people who prefer pumpkin pie, ***x* = 2**, the number of people we asked, ***n* = 3**, and the probability that someone prefers pumpkin pie, ***p* = 0.7**…

$$p(x = 2 \mid n = 3, p = 0.7) = \left(\frac{n!}{x!(n-x)!}\right) p^x (1-p)^{n-x}$$

…then we just do the math…

$$= \left(\frac{3!}{2!(3-2)!}\right) 0.7^2 (1-0.7)^{3-2}$$

(Psst! Remember: the first term is the number of ways we can arrange the pie preferences, the second term is the probability that **2** people prefer pumpkin pie, and the last term is the probability that **1** person prefers blueberry pie.)

$$= 3 \times 0.7^2 \times (0.3)^1$$

$$= 3 \times 0.7 \times 0.7 \times 0.3$$

…and the result is **0.441**, which is the same value we got when we drew pictures of the slices of pie.

$$= 0.441$$

Gentle Reminder: We're using the equation for the **Binomial Distribution** to calculate the probability that **2** out of **3** people prefer pumpkin pie…

0.3 x **0.7** x **0.7** = **0.147**

+

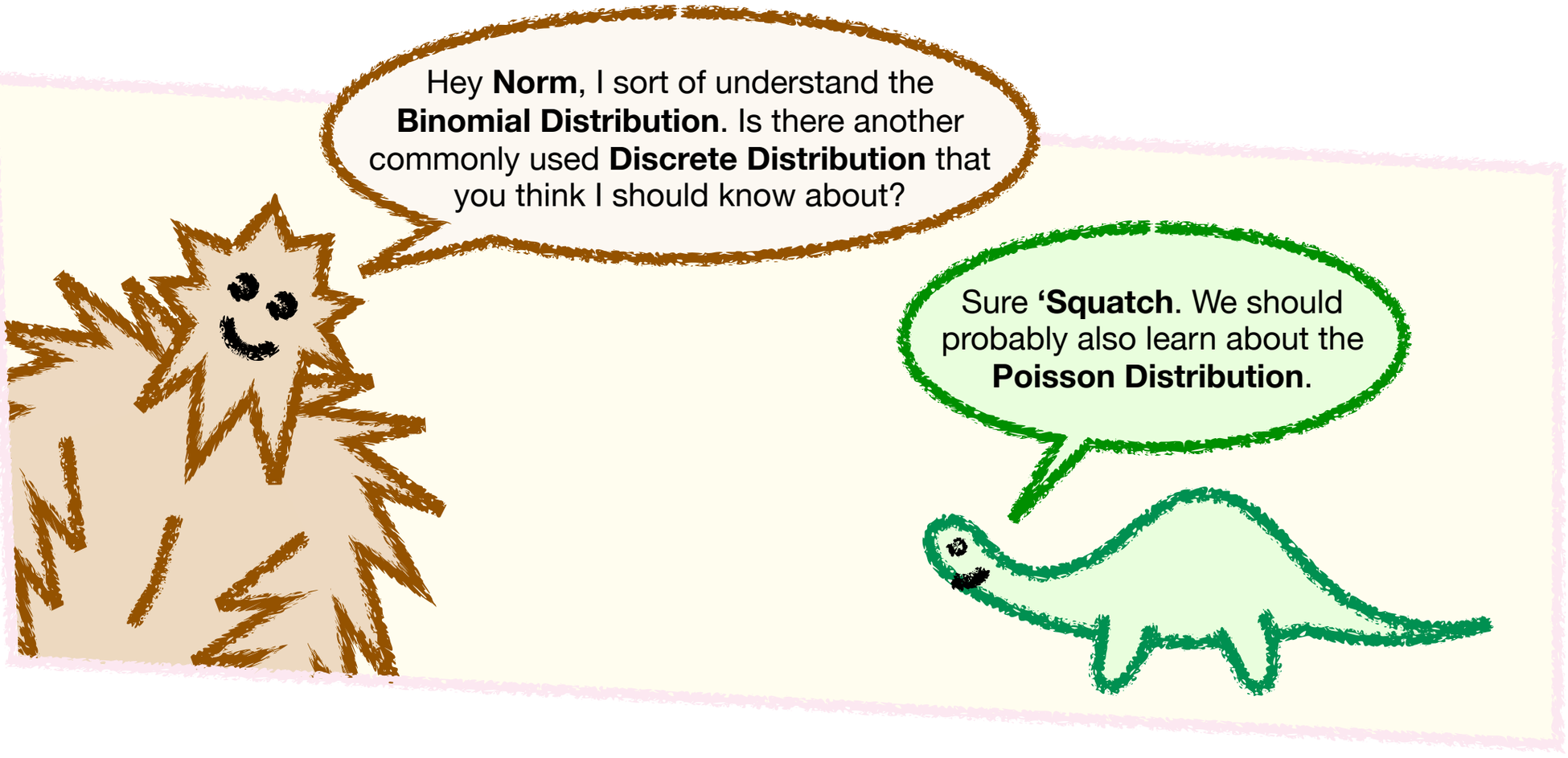
0.7 x **0.3** x **0.7** = **0.147**

+

0.7 x **0.7** x **0.3** = **0.147**

= 0.441

TRIPLE BAM!!!


Hey **Norm**, I sort of understand the **Binomial Distribution**. Is there another commonly used **Discrete Distribution** that you think I should know about?
Sure **'Squatch**. We should probably also learn about the **Poisson Distribution**.

The Poisson Distribution: Details

1 So far, we've seen how the **Binomial Distribution** gives us probabilities for sequences of binary outcomes, like **2** out of **3** people preferring pumpkin pie, but there are lots of other **Discrete Probability Distributions** for lots of different situations.

2 For example, if you can read, on average, **10** pages of this book in an hour, then you can use the **Poisson Distribution** to calculate the probability that in the next hour, you'll read exactly **8** pages.

The equation for the **Poisson Distribution** looks super fancy because it uses the Greek character ***λ***, ***lambda***, but lambda is just the average. So, in this example, ***λ*** = **10** pages an hour.

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

x is the number of pages we think we might read in the next hour. In this example, ***x* = 8**.

NOTE: This '**e**' is **Euler's number**, which is roughly **2.72**.

3 Now we just plug in the numbers and do the math…

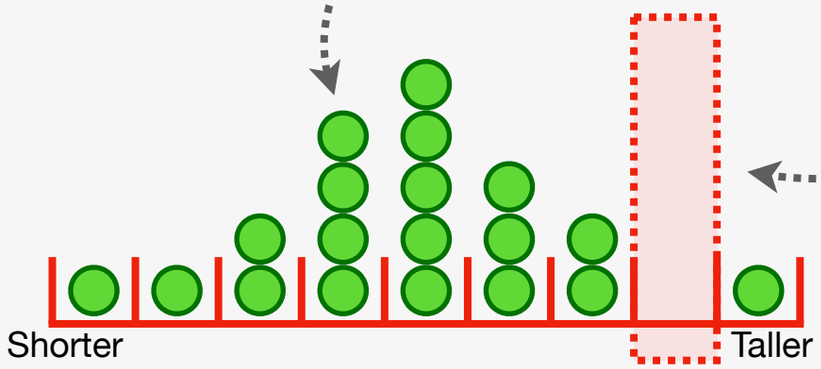
$$p(x = 8 \mid \lambda = 10) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-10}10^{8}}{8!} = \frac{e^{-10}10^{8}}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = 0.113$$

…and we get **0.113**. So the probability that you'll read exactly **8** pages in the next hour, given that, on average, you read **10** pages per hour, is **0.113**.

BAM!!!

Discrete Probability Distributions: Summary

1 To summarize, we've seen that **Discrete Probability Distributions** can be derived from histograms…



…and while these can be useful, they require a lot of data that can be expensive and time-consuming to get, and it's not always clear what to do about the blank spaces.

2 So, we usually use ***mathematical equations***, like the equation for the **Binomial Distribution**, instead.

$$p(x \mid n, p) = \left(\frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}$$

The **Binomial Distribution** is useful for anything that has binary outcomes (wins and losses, yeses and noes, etc.), but there are lots of other **Discrete Probability Distributions**.

3 For example, when we have **events** that happen in discrete units of time or space, like reading **10** pages an hour, we can use the **Poisson Distribution**.

$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

4 There are lots of other **Discrete Probability Distributions** for lots of other types of data. In general, their equations look intimidating, but looks are deceiving. Once you know what each symbol means, you just plug in the numbers and do the math.

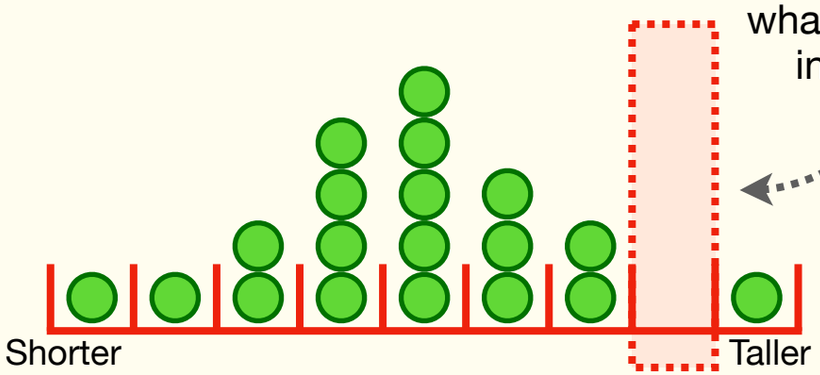
BAM!!!

Now let's talk about **Continuous Probability Distributions**.

Continuous Probability Distributions: Main Ideas

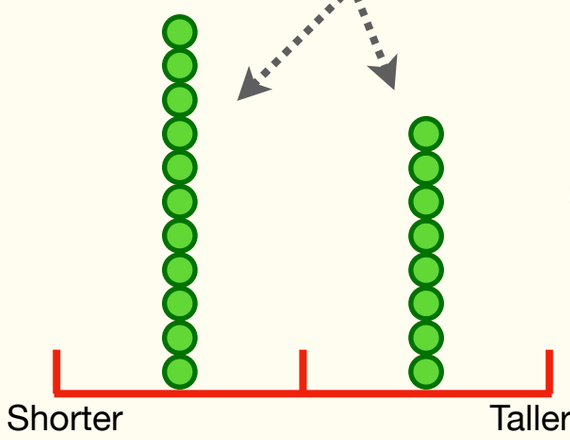
1 **The Problem:** Although they can be super useful, beyond needing a lot of data, histograms have two problems when it comes to continuous data:

1) it's not always clear what to do about gaps in the data and…

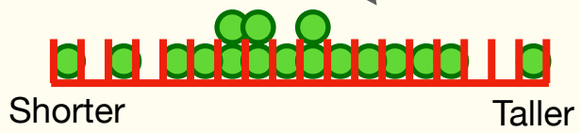


…**2)** histograms can be very sensitive to the size of the bins.

If the bins are too wide, then we lose all of the precision…

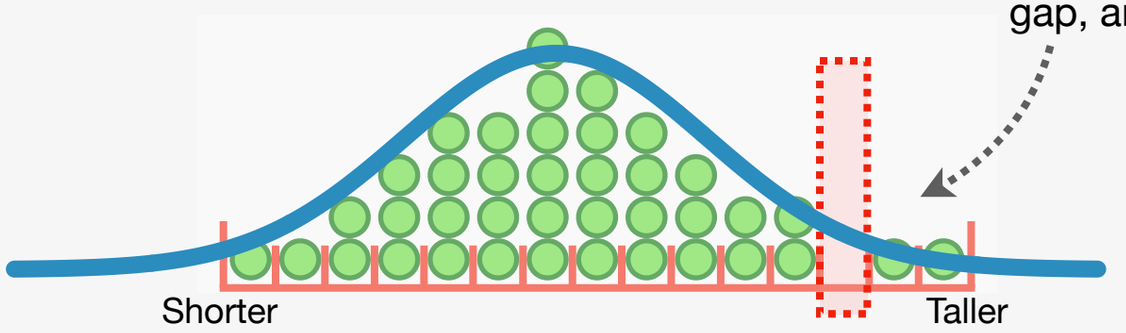


…and if the bins are too narrow, it's impossible to see trends.



2 **A Solution:** When we have continuous data, a **Continuous Distribution** allows us to avoid all of these problems by using mathematical formulas just like we did with **Discrete Distributions**.

In this example, we can use a **Normal Distribution**, which creates a **bell-shaped curve**, instead of a histogram. It doesn't have a gap, and there's no need to fiddle with bin size.



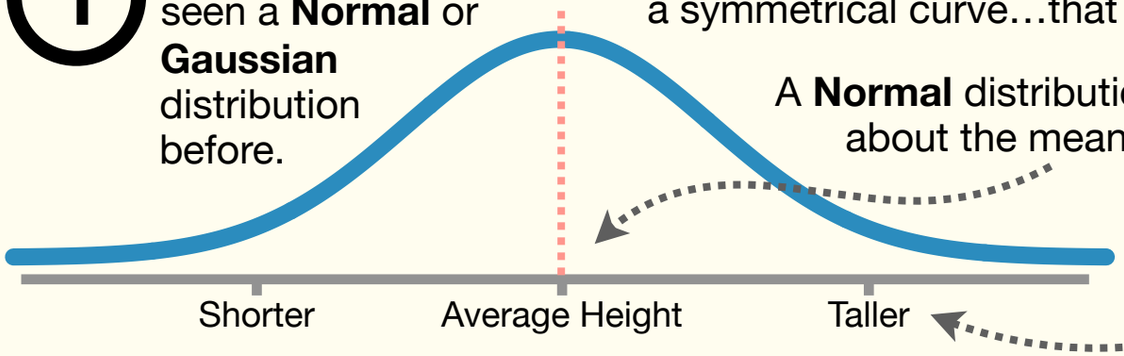
There are lots of commonly used **Continuous Distributions**. Now we'll talk about the most useful of all, the **Normal Distribution**.

The Normal (Gaussian) Distribution: Main Ideas Part 1

1 Chances are you've seen a **Normal** or **Gaussian** distribution before.

It's also called a **Bell-Shaped Curve** because it's a symmetrical curve…that looks like a bell.

A **Normal** distribution is symmetrical about the mean, or average, value.



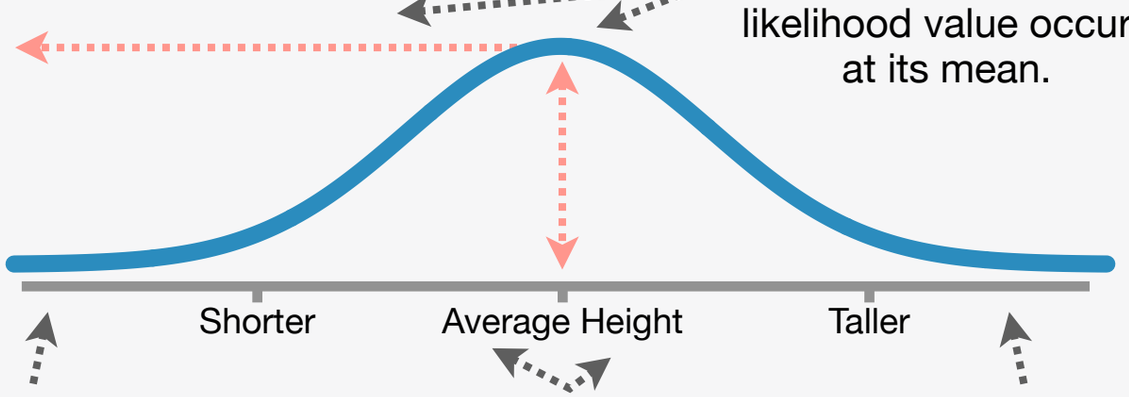
In this example, the curve represents human Height measurements.

2 The y-axis represents the **Likelihood** of observing any specific Height.

More Likely

Less Likely

The **Normal Distribution**'s maximum likelihood value occurs at its mean.

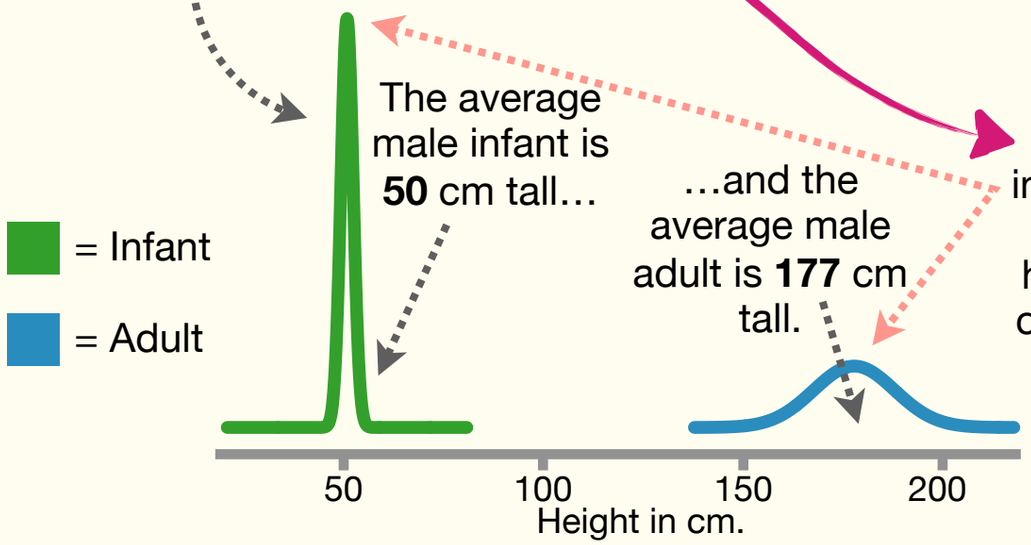


For example, it's relatively rare to see someone who is super short…

…relatively common to see someone who is close to the average height…

…and relatively rare to see someone who is super tall.

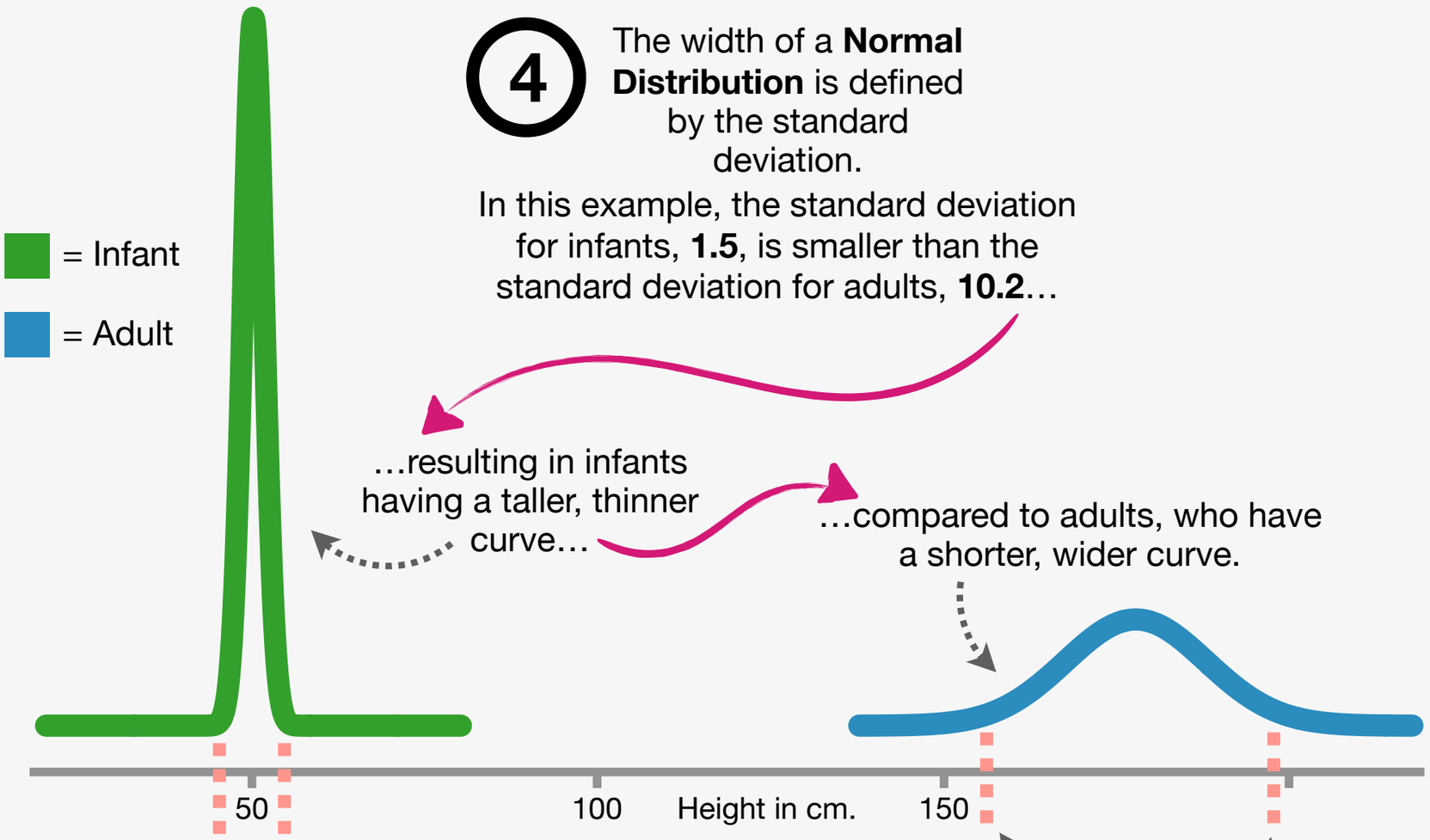
3 Here are two **Normal Distributions** of the heights of male infants and adults.



Because the normal distribution for infants has a higher peak than the one for adults, we can see that there's a higher likelihood that an infant will be close to its mean than an adult will be close to its mean. The difference in peak height tells us there's less variation in how tall an infant is compared to how tall an adult is.

Lots of things can be approximated with **Normal Distributions**: Height, birth weight, blood pressure, job satisfaction, and many more!!!

The Normal (Gaussian) Distribution: Main Ideas Part 2



4 The width of a **Normal Distribution** is defined by the standard deviation.

In this example, the standard deviation for infants, **1.5**, is smaller than the standard deviation for adults, **10.2**…

…resulting in infants having a taller, thinner curve…

…compared to adults, who have a shorter, wider curve.

5 Knowing the standard deviation is helpful because normal curves are drawn such that about **95%** of the measurements fall between **+/- 2 Standard Deviations** around the **Mean**.

Because the mean measurement for infants is **50** cm, and **2 x the standard deviation = 2 x 1.5 = 3**, about **95%** of the infant measurements fall between **47** and **53** cm.

Because the mean adult measurement is **177** cm, and **2 x the standard deviation = 2 x 10.2 = 20.4**, about **95%** of the adult measurements fall between **156.6** and **197.4** cm.

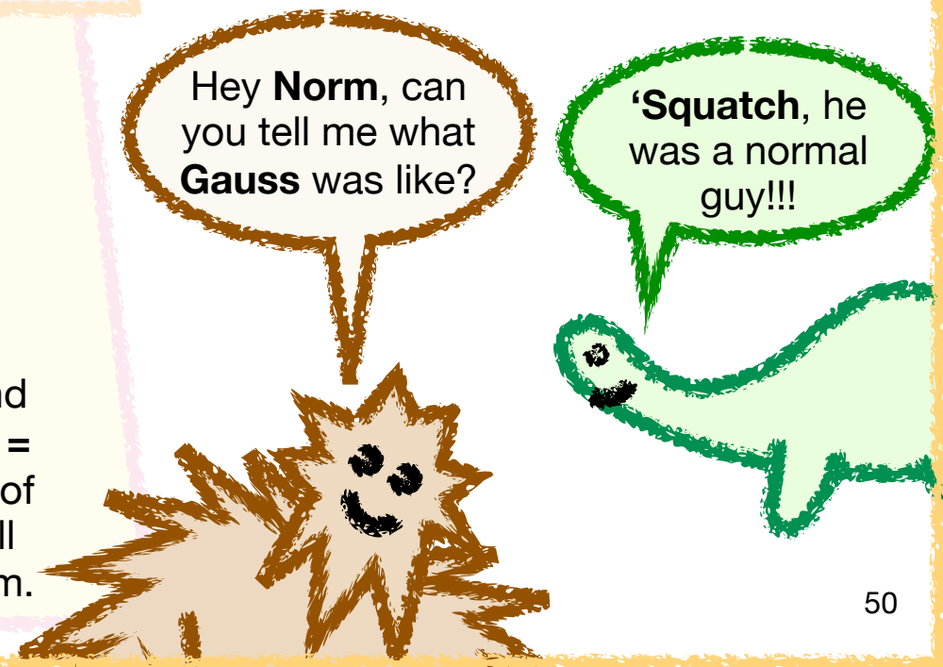
To draw a **Normal Distribution**, you need to know:

1) The **Mean** or average measurement. This tells you where the center of the curve goes.

2) The **Standard Deviation** of the measurements. This tells you how tall and skinny, or short and fat, the curve should be.

If you don't already know about the **Mean** and **Standard Deviation**, check out **Appendix B**.

BAM!!!



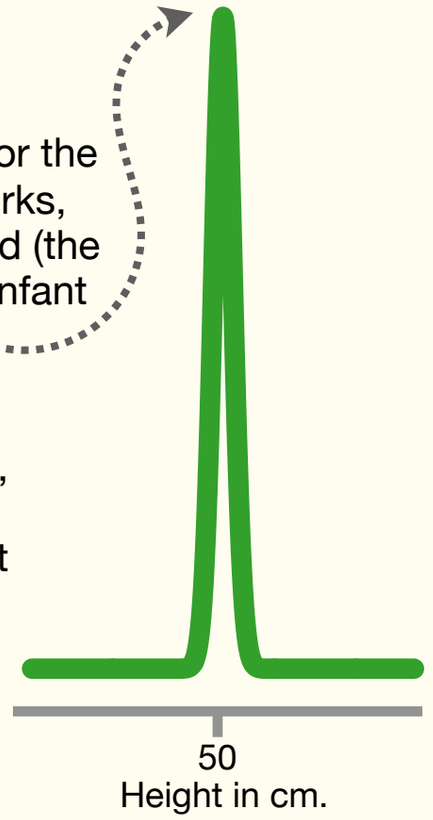
The Normal (Gaussian) Distribution: Details

1 The equation for the **Normal Distribution** looks scary, but, just like every other equation, it's just a matter of plugging in numbers and doing the math.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

2 To see how the equation for the **Normal Distribution** works, let's calculate the likelihood (the y-axis coordinate) for an infant that is **50** cm tall.

Since the mean of the distribution is also **50** cm, we'll calculate the y-axis coordinate for the highest part of the curve.



3

x is the x-axis coordinate. So, in this example, the x-axis represents **Height** and ***x* = 50**.

The Greek character ***μ***, ***mu***, represents the mean of the distribution. In this case, ***μ* = 50**.

Lastly, the Greek character ***σ***, ***sigma***, represents the standard deviation of the distribution. In this case, ***σ* = 1.5**.

$$f(x = 50 \mid \mu = 50, \sigma = 1.5) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Now, we just do the math….

$$= \frac{1}{\sqrt{2\pi 1.5^2}} e^{-(50-50)^2/(2\times 1.5^2)}$$

$$= \frac{1}{\sqrt{14.1}} e^{-0^2/4.5}$$

$$= \frac{1}{\sqrt{14.1}} e^{0}$$

$$= \frac{1}{\sqrt{14.1}}$$

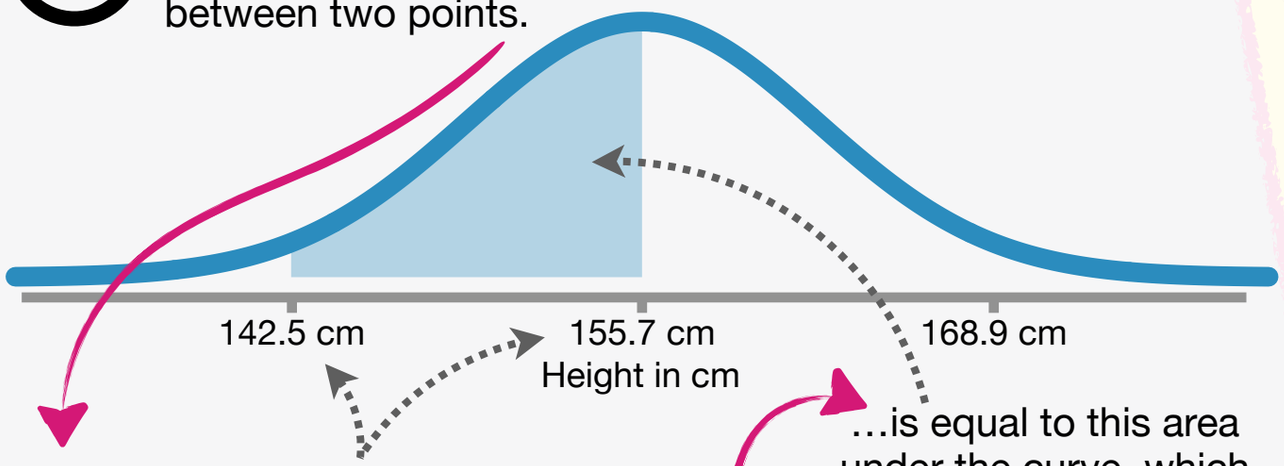
$$= 0.27$$

…and we see that the likelihood, the y-axis coordinate, for the tallest point on the curve, is **0.27**.

Remember, the output from the equation, the y-axis coordinate, is a **likelihood**, ***not*** a **probability**. In **Chapter 7**, we'll see how likelihoods are used in **Naive Bayes**. To learn how to calculate probabilities with **Continuous Distributions**, read on…

Calculating Probabilities with Continuous Probability Distributions: Details

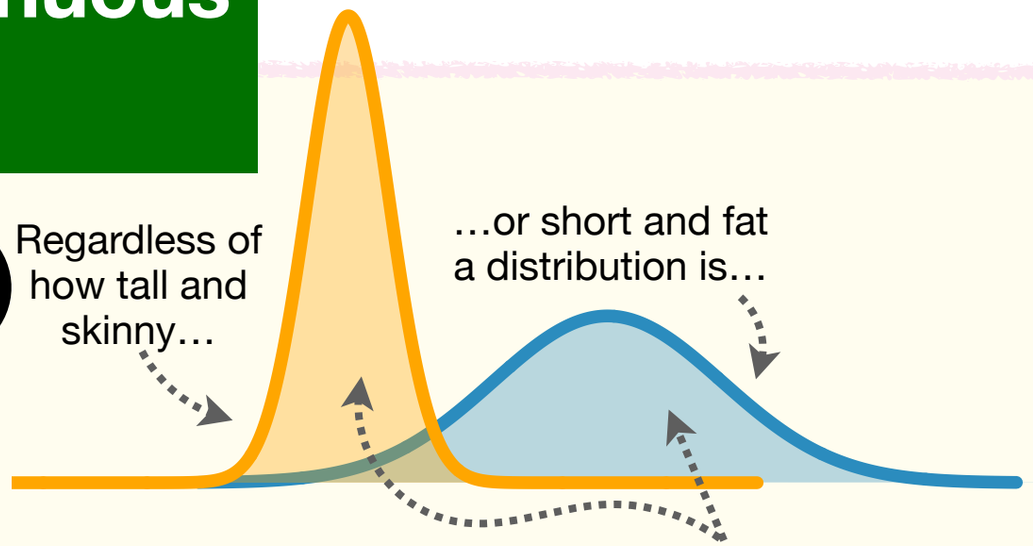
1 For **Continuous Probability Distributions**, probabilities are the **area under the curve** between two points.



For example, given this **Normal Distribution** with **mean = 155.7** and **standard deviation = 6.6**, the probability of getting a measurement between **142.5** and **155.7** cm…

…is equal to this area under the curve, which in this example is **0.48**. So, the probability is **0.48** that we will measure someone in this range.

2 Regardless of how tall and skinny…



…or short and fat a distribution is…

…the total area under its curve is **1**. Meaning, the probability of measuring anything in the range of possible values is **1**.

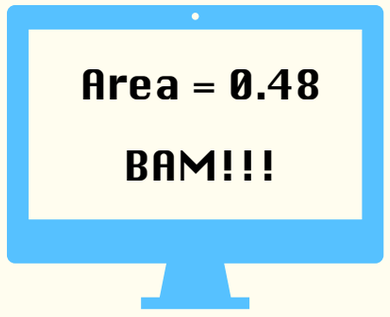
3 There are two ways to calculate the area under the curve between two points:

1) The hard way, by using calculus and integrating the equation between the two points ***a*** and ***b***.

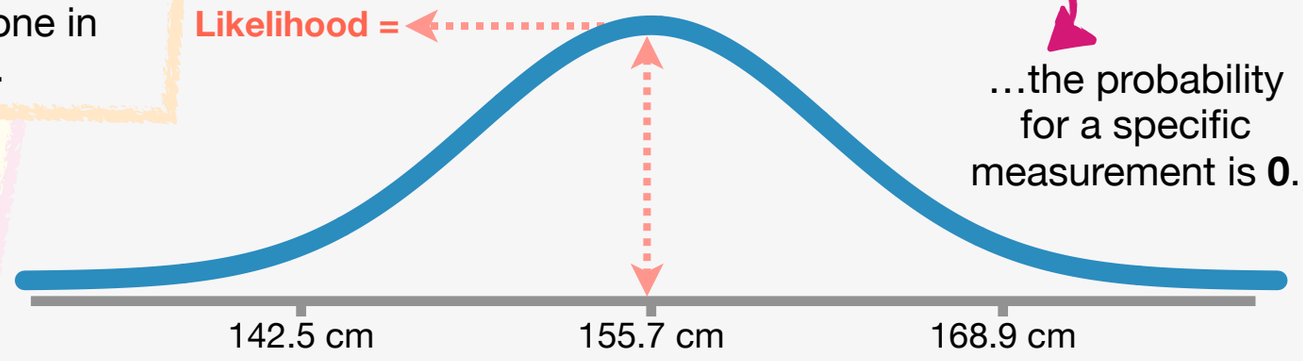
$$\int_a^b f(x)\,dx$$

UGH!!! NO ONE ACTUALLY DOES THIS!!!

2) The easy way, by using a computer. See **Appendix C** for a list of commands.



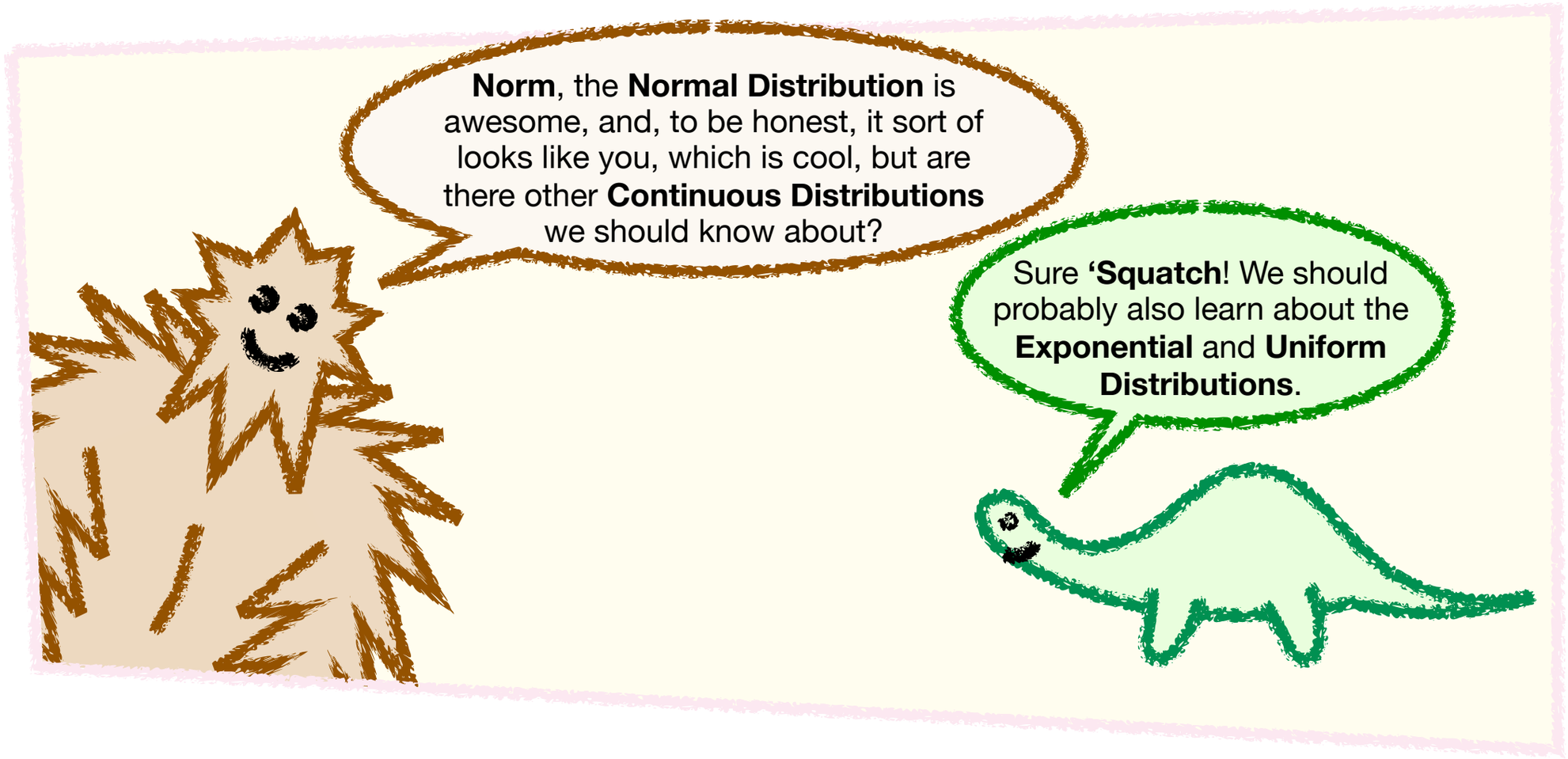
4 One confusing thing about **Continuous Distributions** is that the while the likelihood for a specific measurement, like **155.7**, is the y-axis coordinate and **> 0**…



…the probability for a specific measurement is **0**.

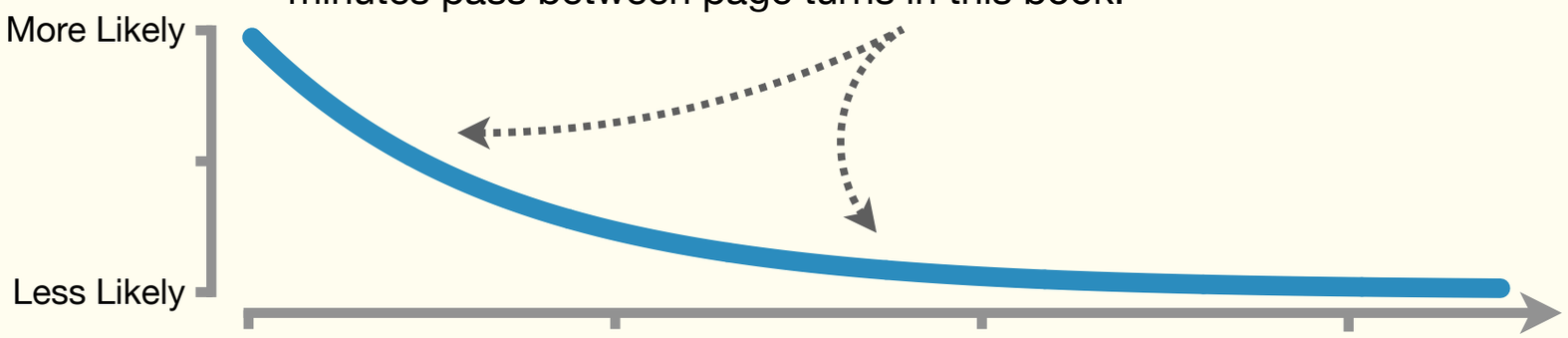
One way to understand why the probability is **0** is to remember that probabilities are areas, and the area of something with no width is **0**.

Another way is to realize that a continuous distribution has infinite precision, thus, we're really asking the probability of measuring someone who is *exactly* **155.700000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000**… tall.


Norm, the **Normal Distribution** is awesome, and, to be honest, it sort of looks like you, which is cool, but are there other **Continuous Distributions** we should know about?
Sure **'Squatch**! We should probably also learn about the **Exponential** and **Uniform Distributions**.

Other Continuous Distributions: Main Ideas

1 **Exponential Distributions** are commonly used when we're interested in how much time passes between events. For example, we could measure how many minutes pass between page turns in this book.

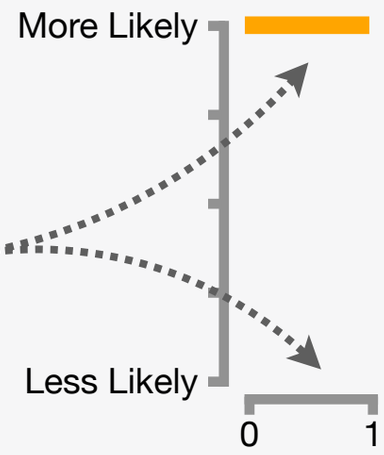


Using Distributions To Generate Random Numbers

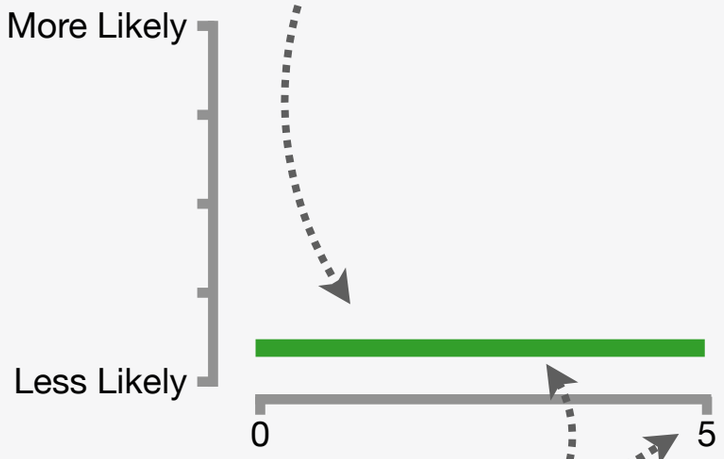
We can get a computer to generate numbers that reflect the likelihoods of any distribution. In machine learning, we usually need to generate random numbers to initialize algorithms before training them with **Training Data**. Random numbers are also useful for randomizing the order of our data, which is useful for the same reasons we shuffle a deck of cards before playing a game. We want to make sure everything is randomized.

2 **Uniform Distributions** are commonly used to generate random numbers that are equally likely to occur.

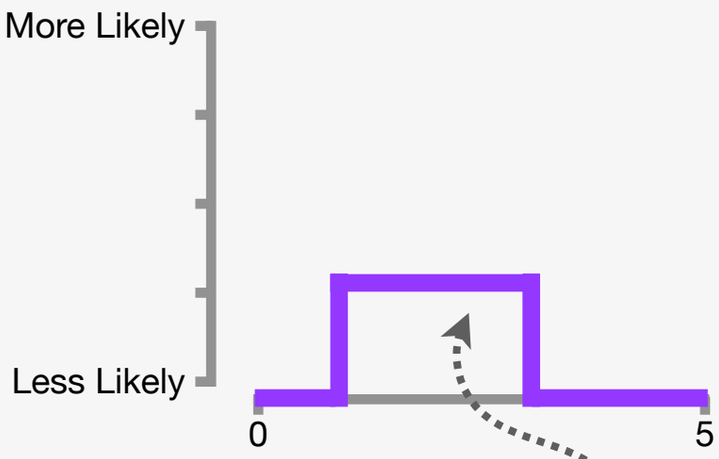
For example, if I want to select random numbers between **0** and **1**, then I would use a **Uniform Distribution** that goes from **0** to **1**, which is called a **Uniform 0,1 Distribution**, because it ensures that every value between **0** and **1** is equally likely to occur.



NOTE: Because there are fewer values between **0** and **1** than between **0** and **5**, we see that the corresponding likelihood for any specific number is higher for the **Uniform 0,1 Distribution** than the **Uniform 0,5 Distribution**.



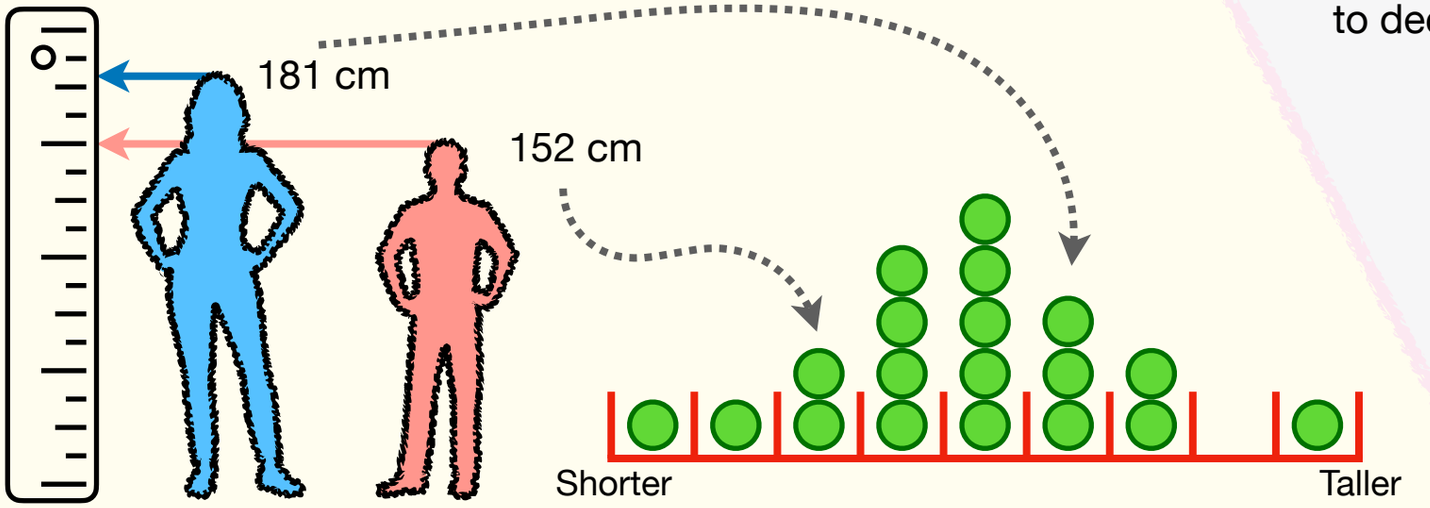
In contrast, if I wanted to generate random numbers between **0** and **5**, then I would use a **Uniform Distribution** that goes from **0** to **5**, which is called a **Uniform 0,5 Distribution**.



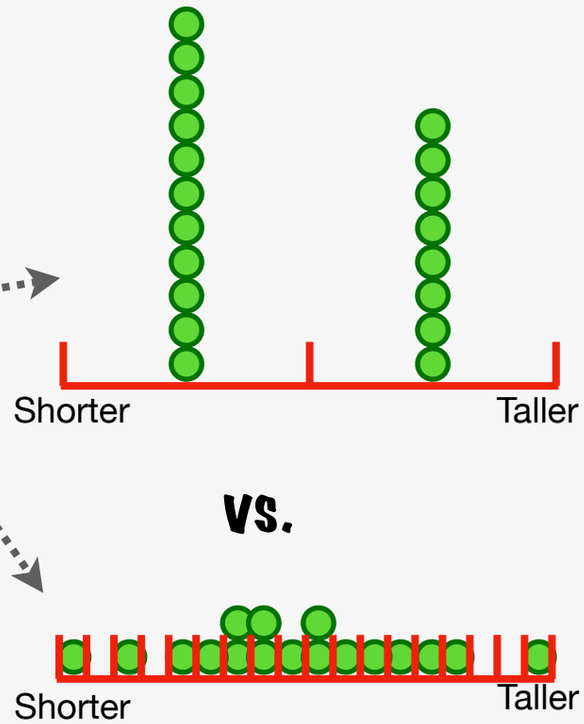
Uniform Distributions can span any **2** numbers, so we could have a **Uniform 1,3.5 Distribution** if we wanted one.

Continuous Probability Distributions: Summary

1 Just like **Discrete Distributions**, **Continuous Distributions** spare us from having to gather tons of data for a histogram…



2 …and, additionally, **Continuous Distributions** also spare us from having to decide how to bin the data.



3 Instead, **Continuous Distributions** use equations that represent smooth curves and can provide likelihoods and probabilities for all possible measurements.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

4 Like **Discrete Distributions**, there are **Continuous Distributions** for all kinds of data, like the values we get from measuring people's height or timing how long it takes you to read this page.

In the context of machine learning, both types of distributions allow us to create **Models** that can predict what will happen next.

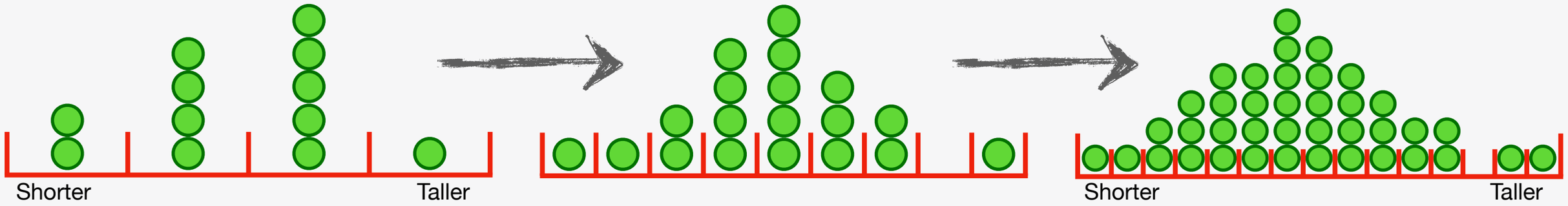
So, let's talk about what **Models** are and how to use them.

(small but mighty) **BAM!!!**

Models: Main Ideas Part 1

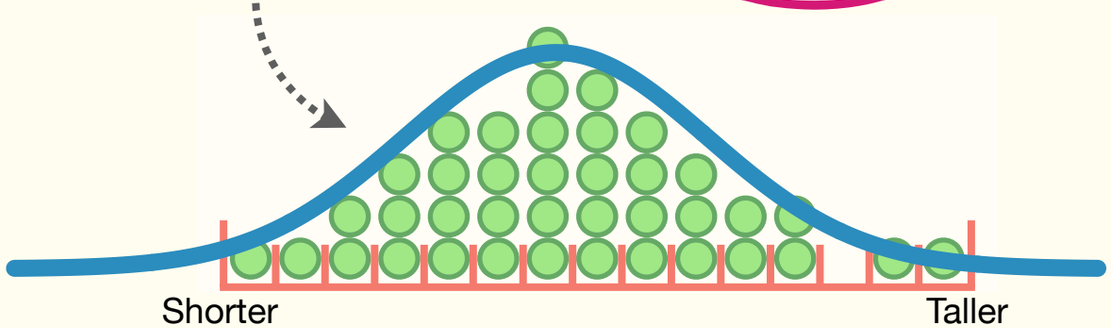
1 **The Problem:** Although we could spend a lot of time and money to build a precise histogram…

…collecting *all* of the data in the world is usually impossible.

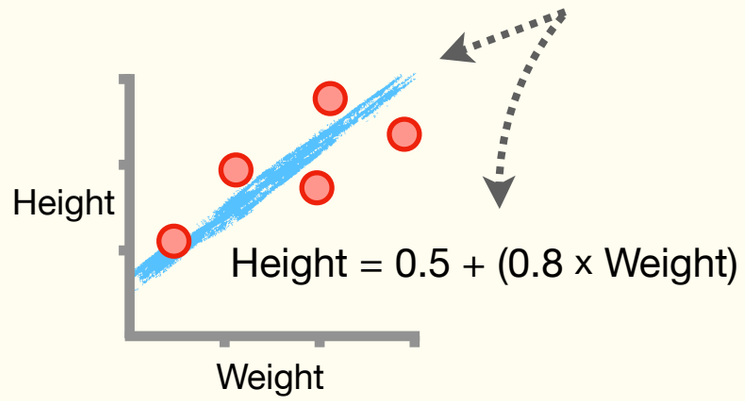


2 **A Solution:** A statistical, mathematical, or machine learning **Model** provides an *approximation* of reality that we can use in a wide variety of ways.

A **Probability Distribution** is a type of model that approximates a histogram with an infinite amount of data.

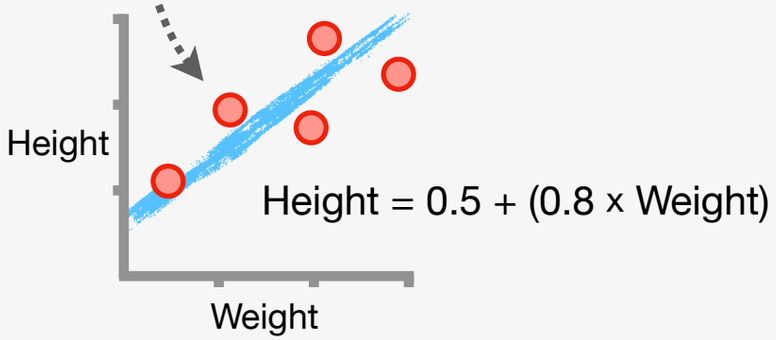


Another commonly used model is the equation for a straight line. Here, we're using a **blue line** to model a relationship between Weight and Height.



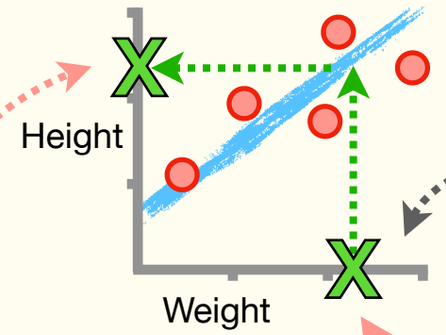
Models: Main Ideas Part 2

3 As we saw in **Chapter 1**, models need **Training Data**. Using machine learning lingo, we say that we *build models by training machine learning algorithms*.



4 Models, or equations, can tell us about people we haven't measured yet.

For example, if we want to know how tall someone is who weighs this much…



…we plug the Weight into the equation and solve for Height…

Height = 0.5 + (0.8 x Weight)

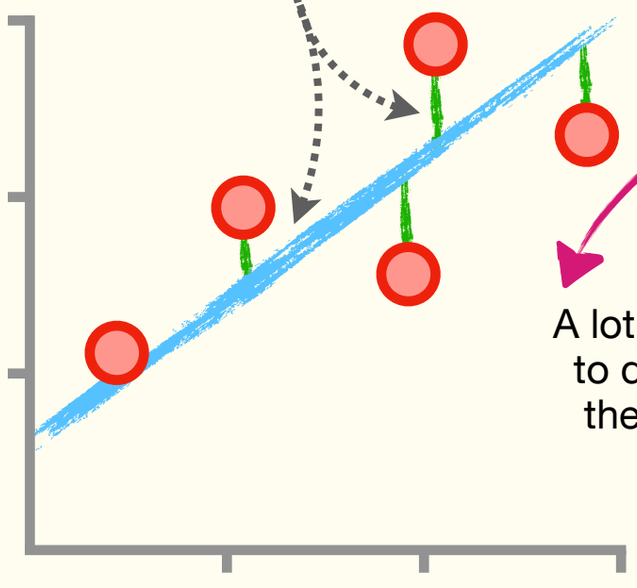
Height = 0.5 + (0.8 x 2.1)

…and get **2.18**.

Height = 2.18

5 Because models are only approximations, it's important that we're able to measure the quality of their predictions.

These **green lines** show the distances from the model's predictions to the actual data points.



A lot of statistics is dedicated to quantifying the quality of the predictions made by a model.

6 In summary:

1) Models approximate reality to let us explore relationships and make predictions.

2) In machine learning, we build models by training machine learning algorithms with **Training Data**.

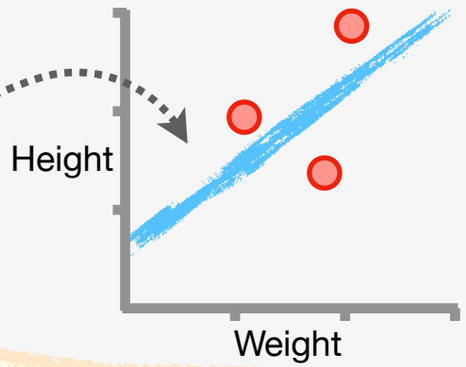
3) Statistics can be used to determine if a model is useful or believable.

Bam!

Now let's talk about how statistics can quantify the quality of a model. The first step is to learn about the **Sum of the Squared Residuals**, which is something we'll use throughout this book.

The Sum of the Squared Residuals: Main Ideas Part 1

1 **The Problem:** We have a model that makes predictions. In this case, we're using Weight to predict Height. However, we need to quantify the quality of the model and its predictions.

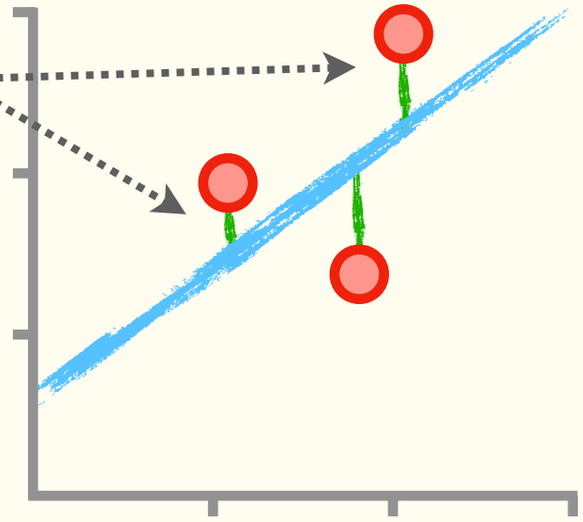


2 **A Solution:** One way quantify the quality of a model and its predictions is to calculate the **Sum of the Squared Residuals**.

As the name implies, we start by calculating **Residuals**, the differences between the **Observed** values and the values **Predicted** by the model.

Residual = Observed - Predicted

Visually, we can draw **Residuals** with these **green lines**.



Since, in general, the smaller the **Residuals**, the better the model fits the data, it's tempting to compare models by comparing the sum of their **Residuals**, but the **Residuals** below the **blue line** would cancel out the ones above it!!!

So, instead of calculating the sum of the **Residuals**, we square the **Residuals** first and calculate the **Sum of the Squared Residuals** (**SSR**).

NOTE: ***Squaring***, as opposed to taking the ***absolute value***, makes it easy to take the derivative, which will come in handy when we do **Gradient Descent** in **Chapter 5**.

The Sum of Squared Residuals (SSR)

is usually defined with fancy **Sigma** notation and the right-hand side reads: "The sum of all observations of the squared difference between the observed and predicted values."

$$SSR = \sum_{i=1}^{n} (Observed_i - Predicted_i)^2$$

n = the number of **Observations**.

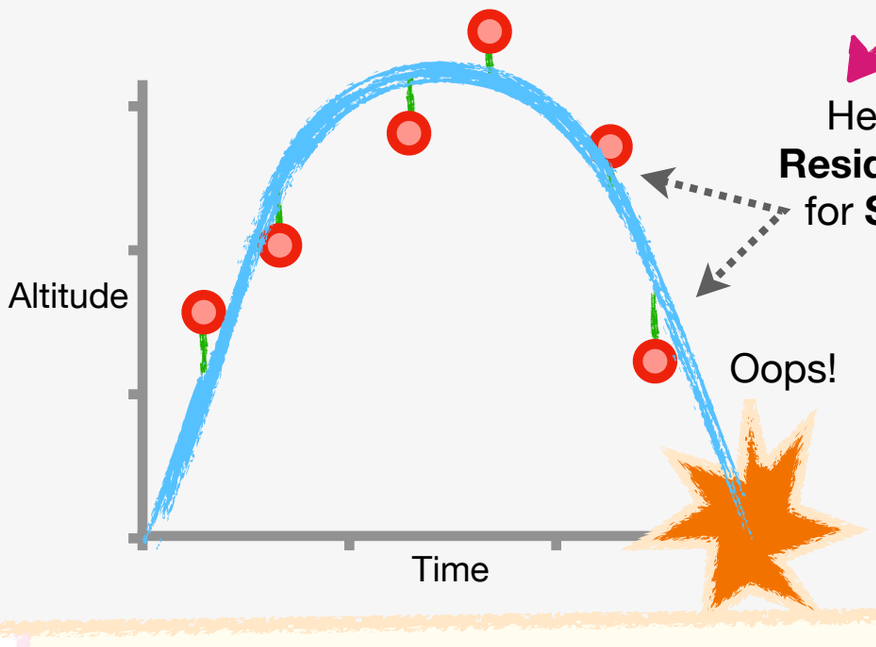
i = the index for each **Observation**. For example, ***i* = 1** refers to the first **Observation**.

The **Sigma** symbol, **Σ**, tells us to do a **summation**.

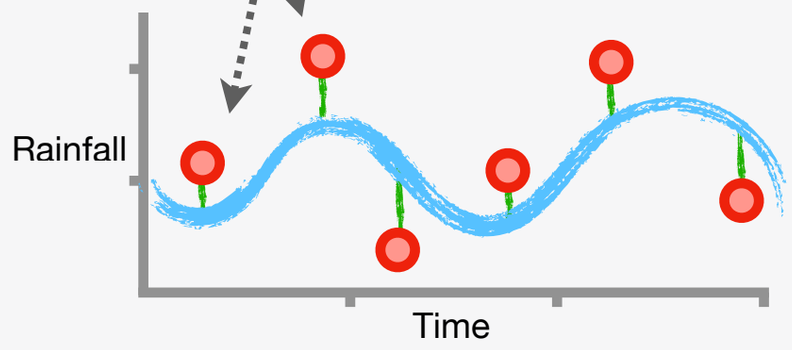
The Sum of the Squared Residuals: Main Ideas Part 2

3 So far, we've looked at the **SSR** only in terms of a simple straight line model, but we can calculate it for all kinds of models.

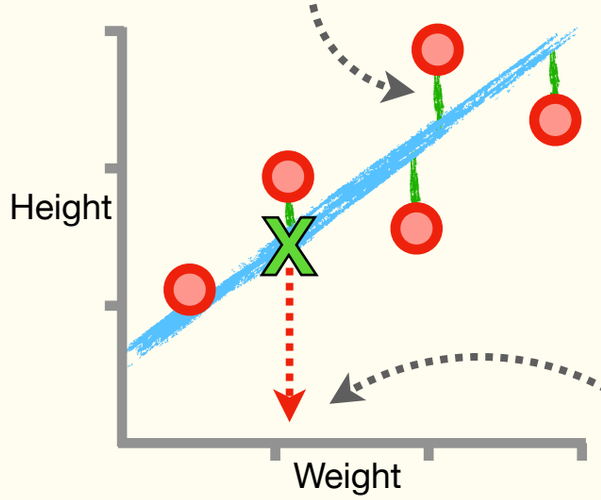
Here's an example of the **Residuals** for altitude vs. time for **SpaceX's SN9** rocket…



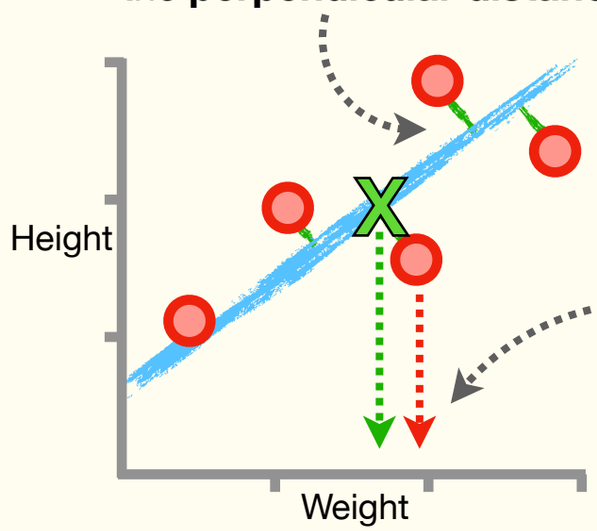
…and here's an example of the **Residuals** for a sinusoidal model of rainfall. Some months are more rainy than others, and the pattern is cyclical over time. If you can calculate the **Residuals**, you can square them and add them up!



4 **NOTE:** When we calculate the **Residuals**, we use the **vertical distance** to the **model**…



…instead of the shortest distance, the **perpendicular distance**…

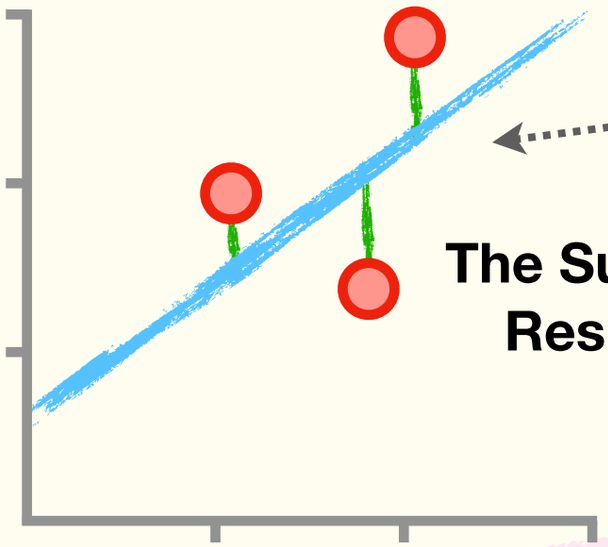


…because, in this example, perpendicular lines result in different Weights for the Observed and Predicted Heights.

In contrast, the vertical distance allows both the Observed and Predicted Heights to correspond to the same Weight.

5 Now that we understand the main ideas of the **SSR**, let's walk through an example of how it's calculated, step-by-step.

SSR: Step-by-Step



1 In this example, we have **3** Observations, so ***n* = 3**, and we expand the summation into **3** terms.

The Sum of Squared Residuals (SSR) $= \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2$

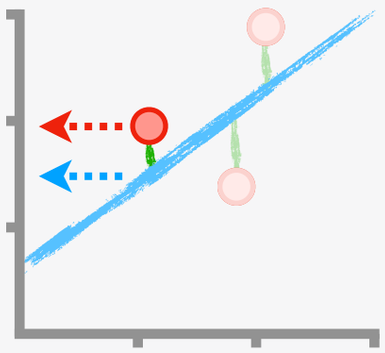
Observed = ····▸

Predicted = ····▸

Residual = |

For ***i* = 1**, the term for the first Observation…

$(1.9 - 1.7)^2$



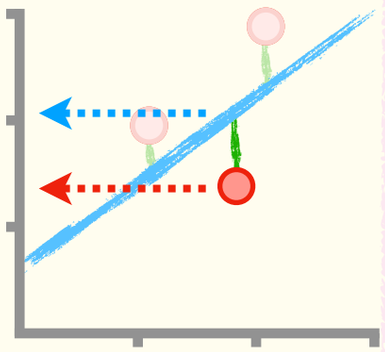
SSR $= (\text{Observed}_1 - \text{Predicted}_1)^2$

$+ (\text{Observed}_2 - \text{Predicted}_2)^2$

$+ (\text{Observed}_3 - \text{Predicted}_3)^2$

For ***i* = 2**, the term for the second Observation…

$(1.6 - 2.0)^2$



2 Once we expand the summation, we plug in the **Residuals** for each Observation.

SSR $= (1.9 - 1.7)^2$

$+ (1.6 - 2.0)^2$

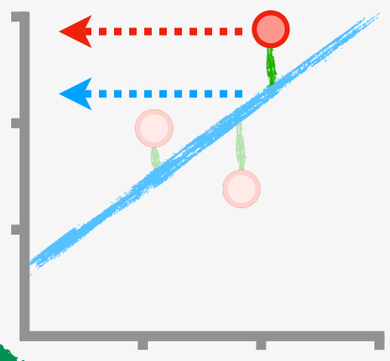
$+ (2.9 - 2.2)^2$

$= 0.69$

3 Now, we just do the math, and the final **Sum of Squared Residuals (SSR)** is **0.69**.

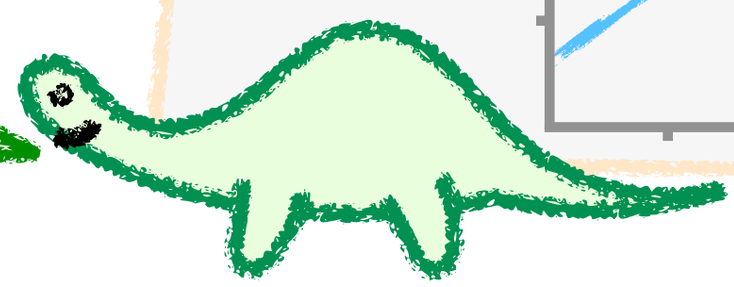
For ***i* = 3**, the term for the third Observation…

$(2.9 - 2.2)^2$



BAM!!!

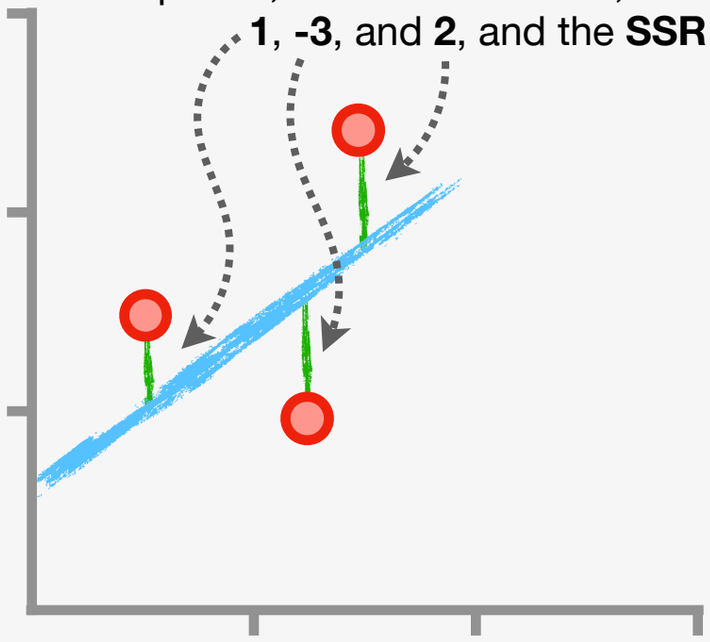
Don't get me wrong, the **SSR** is awesome, but it has a pretty big problem that we'll talk about on the next page.



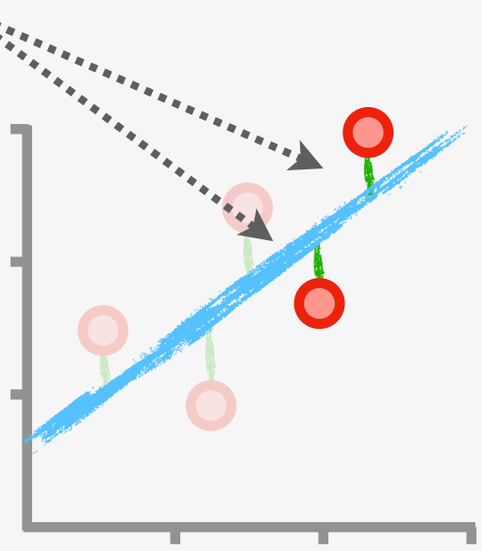
Mean Squared Error (MSE): Main Ideas

1 **The Problem: Sum of the Squared Residuals** (**SSR**), although awesome, is not super easy to interpret because it depends, in part, on how much data you have.

For example, if we start with a simple dataset with **3** points, the **Residuals** are, from left to right, **1**, **-3**, and **2**, and the **SSR = 14**.



Now, if we have a second dataset that includes **2** more data points added to the first one, and the **Residuals** are **-2** and **2**, then the **SSR** increases to **22**.



However, the increase in the **SSR** from **14** to **22** *does not* suggest that the second model, fit to the second, larger dataset, is worse than the first. It only tells us that the model with more data has more **Residuals**.

2 **A Solution:** One way to compare the two models that may be fit to different-sized datasets is to calculate the **Mean Squared Error** (**MSE**), which is simply the average of the **SSR**.

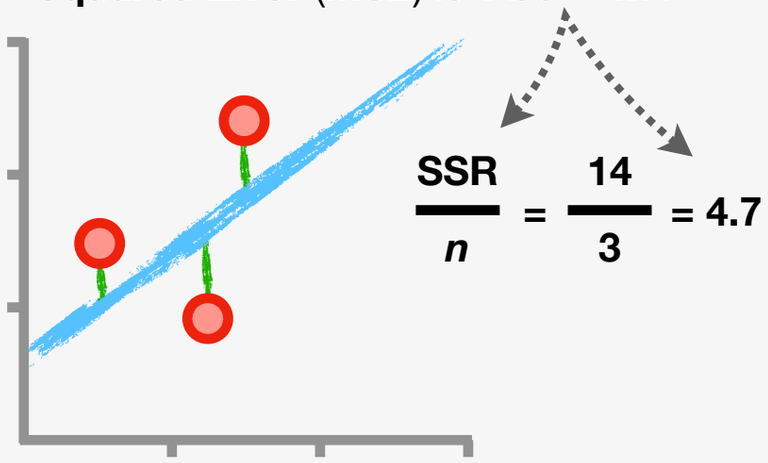
$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{The Sum of Squared Residuals (SSR)}}{\textbf{Number of Observations, } n} = \sum_{i=1}^{n} \frac{(\text{Observed}_i - \text{Predicted}_i)^2}{n}$$

Mean Squared Error (MSE): Step-by-Step

1 Now let's see the **MSE** in action by calculating it for the two datasets!!!

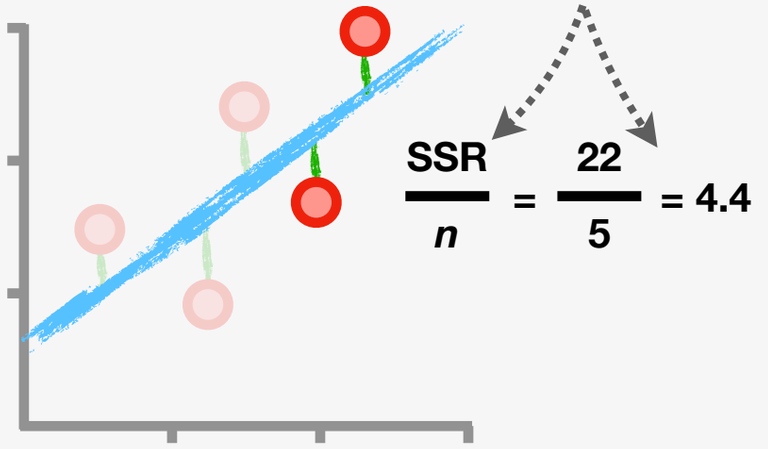
$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{SSR}}{n} = \sum_{i=1}^{n} \frac{(\text{Observed}_i - \text{Predicted}_i)^2}{n}$$

2 The first dataset has only **3** points and the **SSR = 14**, so the **Mean Squared Error** (**MSE**) is **14/3 = 4.7**.



$$\frac{\textbf{SSR}}{n} = \frac{14}{3} = 4.7$$

The second dataset has **5** points and the **SSR** increases to **22**. In contrast, the **MSE**, **22/5 = 4.4**, is now slightly lower.

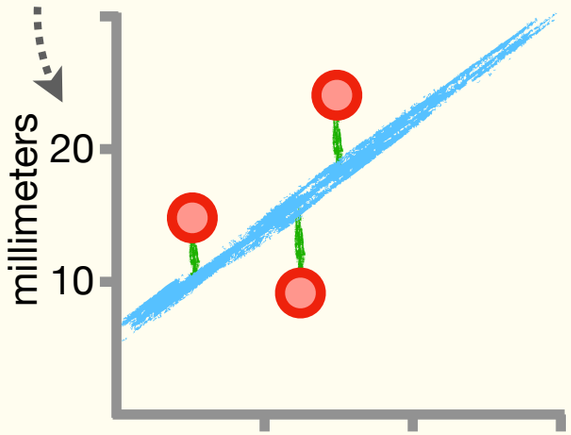


$$\frac{\textbf{SSR}}{n} = \frac{22}{5} = 4.4$$

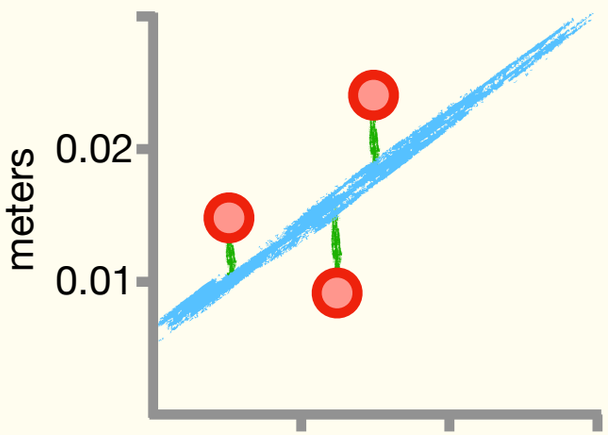
So, unlike the **SSR**, which increases when we add more data to the model, the **MSE** can increase or decrease depending on the average residual, which gives us a better sense of how the model is performing overall.

3 Unfortunately, **MSEs** are still difficult to interpret on their own because the maximum values depend on the scale of the data.

For example, if the y-axis is in *millimeters* and the **Residuals** are **1**, **-3**, and **2**, then the **MSE = 4.7**.



However, if we change the y-axis to *meters*, then the **Residuals** for the exact same data shrink to **0.001**, **-0.003**, and **0.002**, and the **MSE** is now **0.0000047**. It's tiny!

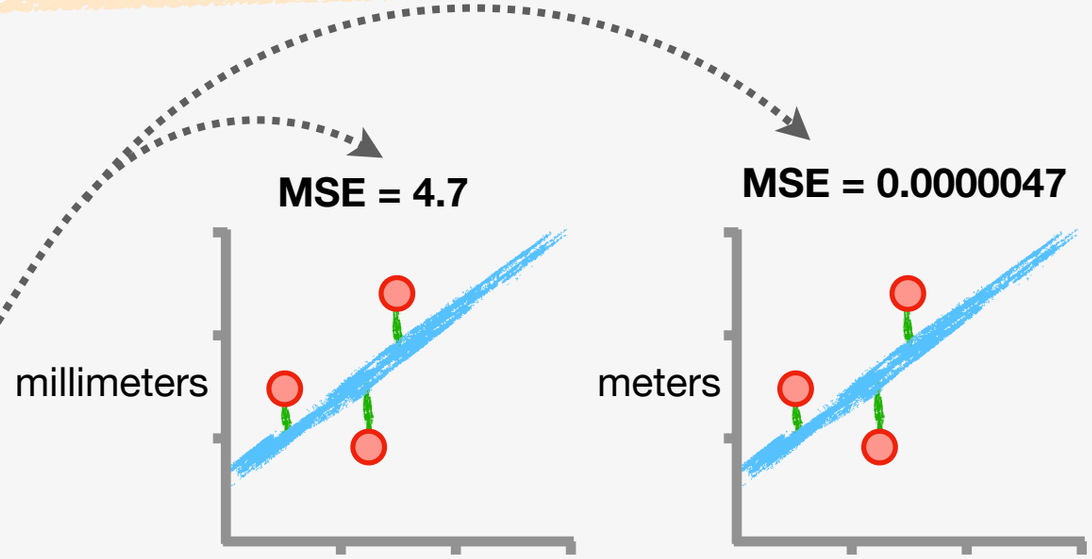


The good news, however, is that both the **SSR** and the **MSE** can be used to calculate something called R^2, which is independent of both the size of the dataset and the scale, so keep reading!

R^2: Main Ideas

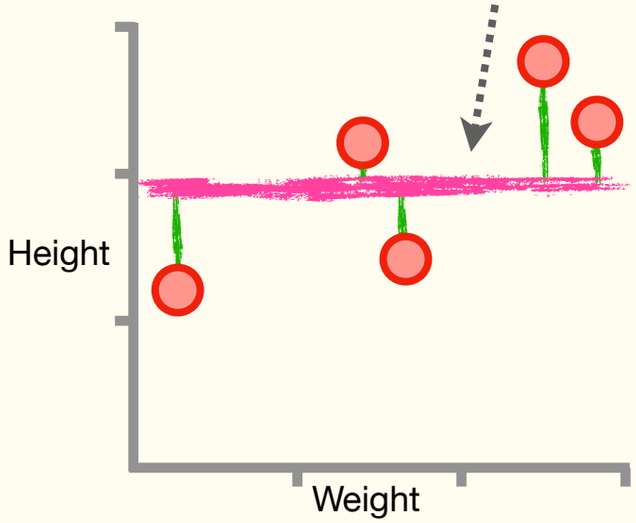
1 **The Problem:** As we just saw, the **MSE**, although totally cool, can be difficult to interpret because it depends, in part, on the scale of the data.

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

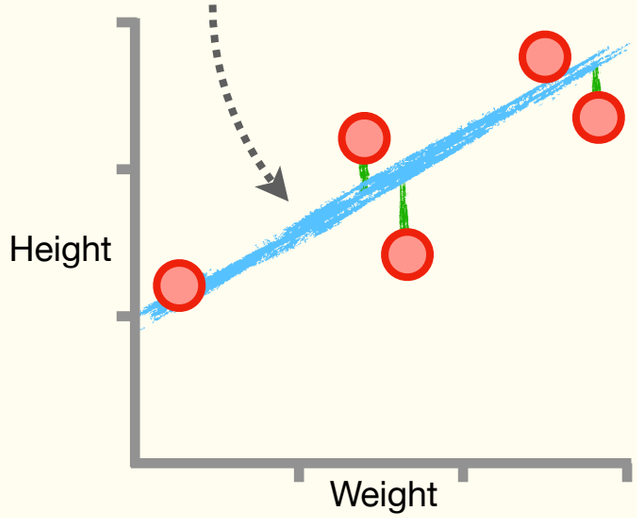


2 **A Solution:** R^2, pronounced **R squared**, is a simple, easy-to-interpret metric that does not depend on the size of the dataset or its scale.

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**.

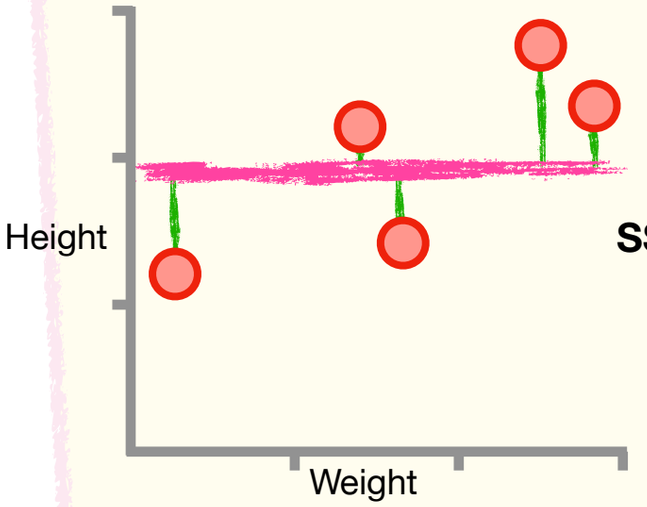
In this example, R^2 would tell us how much better our predictions are when we use the **blue line**, which uses Weight to predict Height, instead of predicting that everyone has the **average** Height.

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

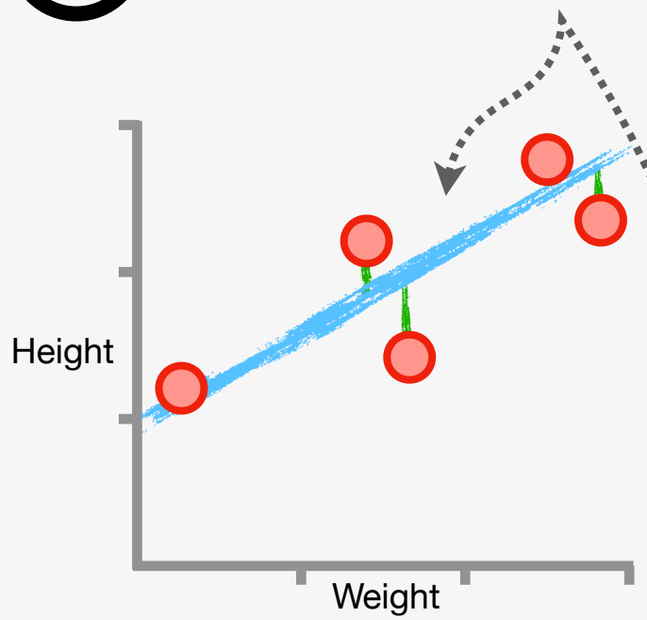
1 First, we calculate the **Sum of the Squared Residuals** for the **mean**. We'll call this **SSR** the **SSR(mean)**. In this example, the mean Height is **1.9** and the **SSR(mean) = 1.6**.



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

= 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) = $(1.2 - \mathbf{1.1})^2 + (2.2 - \mathbf{1.8})^2 + (1.4 - \mathbf{1.9})^2 + (2.7 - \mathbf{2.4})^2 + (2.3 - \mathbf{2.5})^2$

= 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 $R^2 = 0$

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

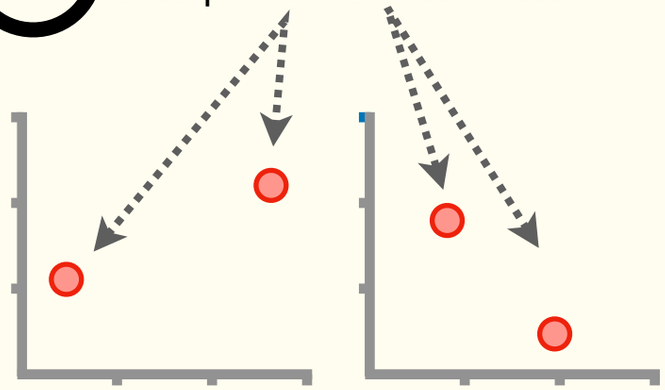
When **SSR(fitted line) = 0**, meaning that the **fitted line** fits the data perfectly, then $R^2 = 1$.

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

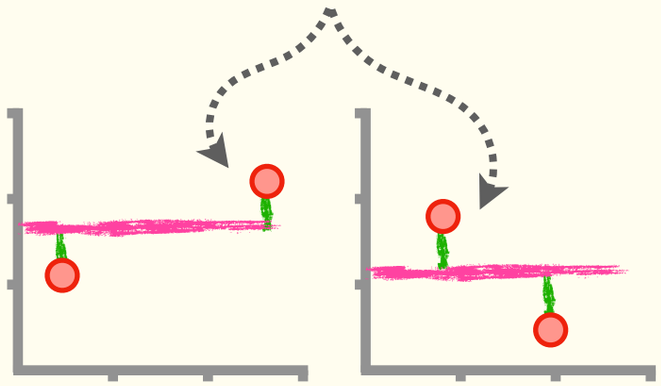
R^2: Details Part 2

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

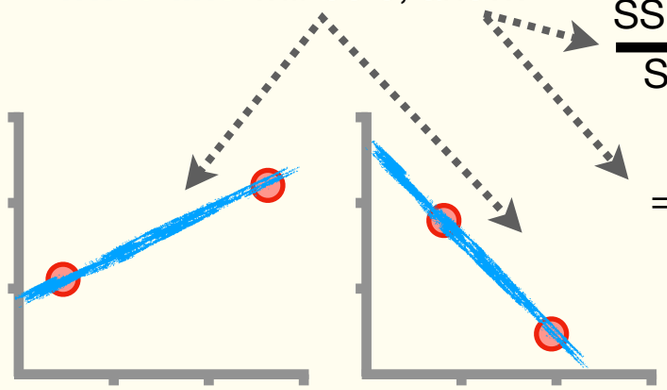
5 **NOTE:** Any **2** random data points have $R^2 = 1$…



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



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

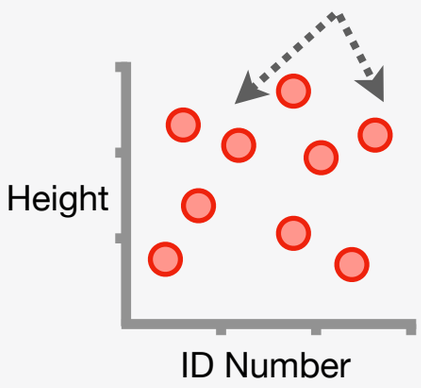


$$\frac{\text{SSR(mean)} - 0}{\text{SSR(mean)}}$$

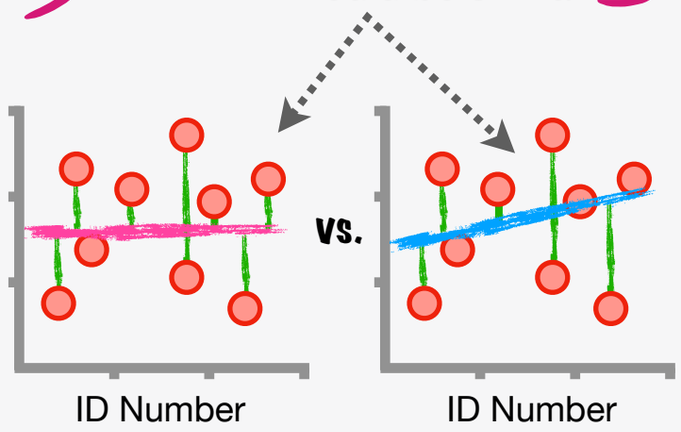
$$= \frac{\text{SSR(mean)}}{\text{SSR(mean)}} = 1$$

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.

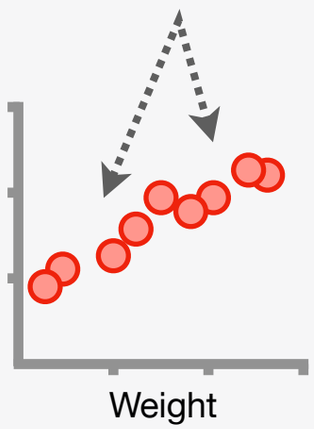
6 If we had a lot of data organized randomly using a random **ID Number**, we would expect the graph to look like this…



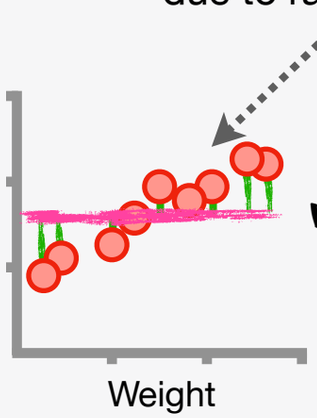
…and have a relatively small (close to **0**) R^2 because the **Residuals** would be similar.

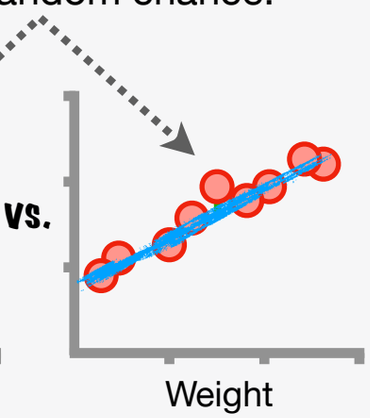


In contrast, when we see a trend in a large amount of data like this…



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





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

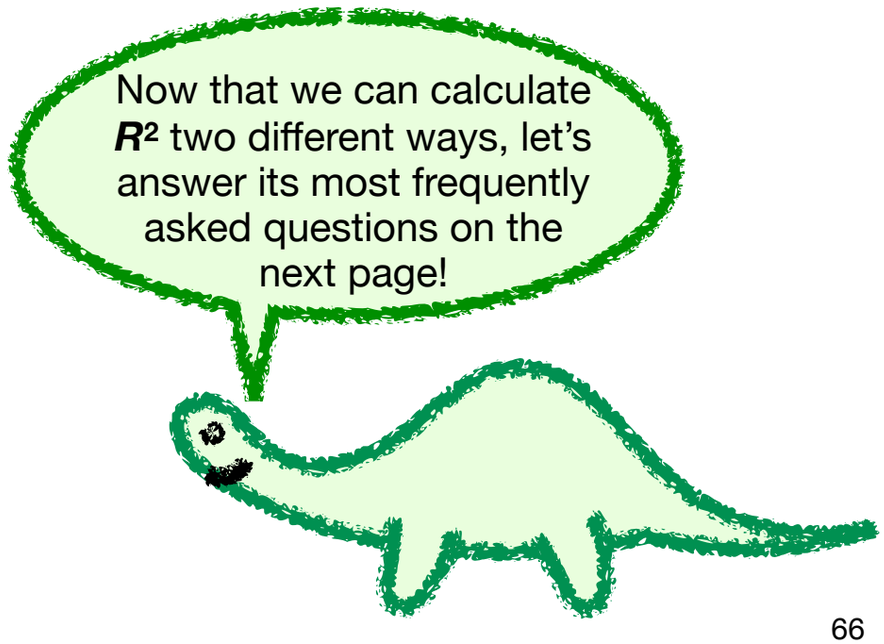
SSR = Sum of Squared Residuals

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

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

…where ***n*** is the sample size

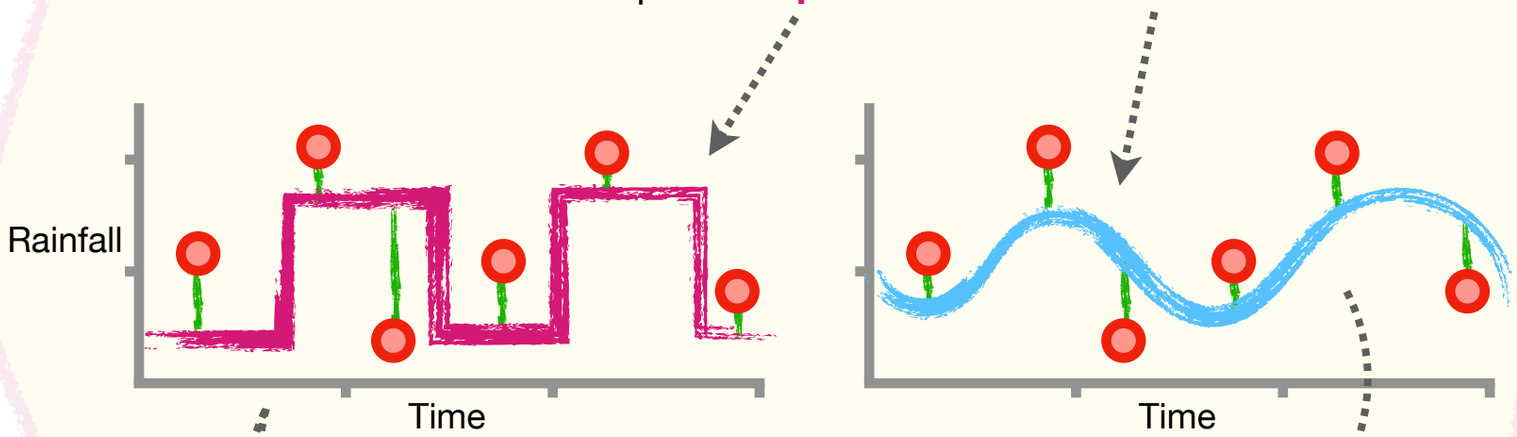
$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$



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**.



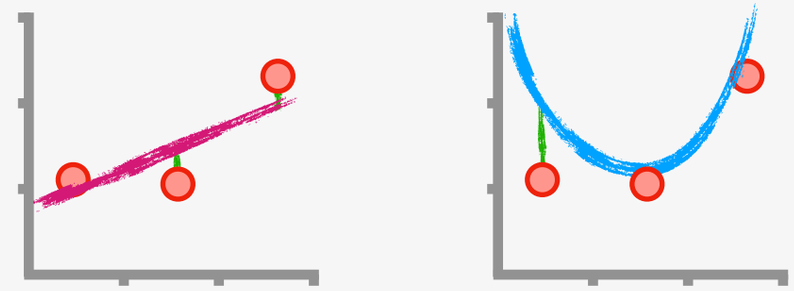
In this case, we calculate R^2 based on the **Sum of the Squared Residuals** around the **square** and **sine** waves.

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

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!!!

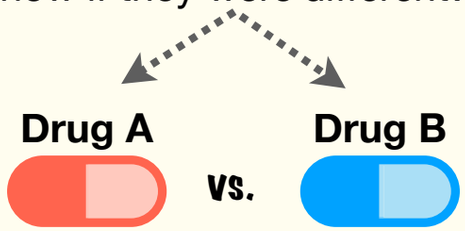
p-values: Main Ideas Part 1

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

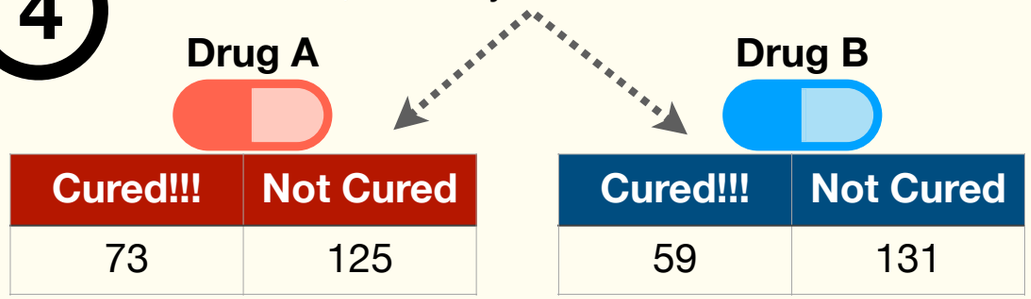
Cured!!!	Not Cured
2	1,432

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.

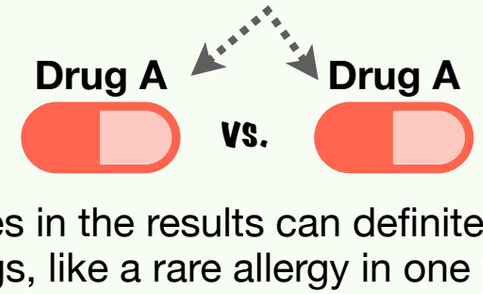
So, the question is, "how small does a ***p*-value** have to be before we're sufficiently confident that Drug A is different from Drug B?"

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

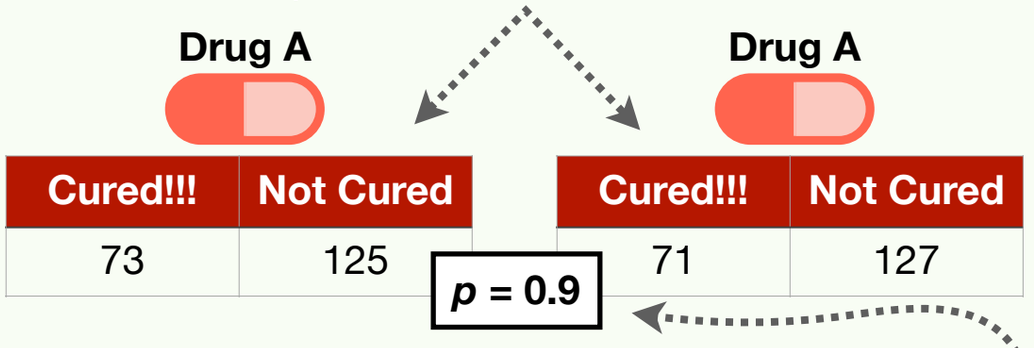
5 In practice, a commonly used threshold is **0.05**. It means that if there's no difference between Drug A and Drug B, and if we did this exact same experiment a bunch of times, then only **5%** of those experiments would result in the wrong decision.

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

Cured!!!	Not Cured
71	127

***p* = 0.9**

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.

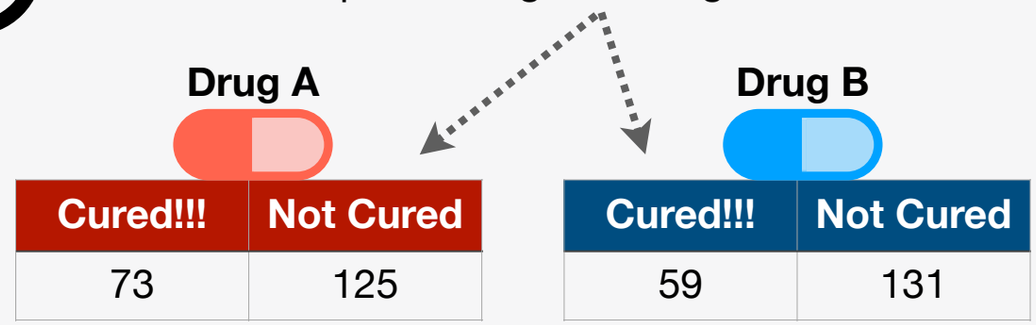
Using a threshold of **0.001** would get a **False Positive** only once in every **1,000** experiments.

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		Drug B	
Cured!!!	**Not Cured**	**Cured!!!**	**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**.

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

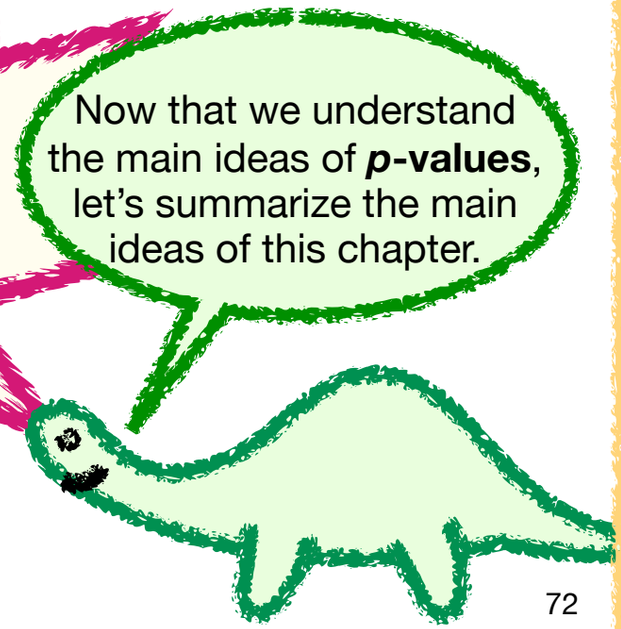
Drug B

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

35% Cured

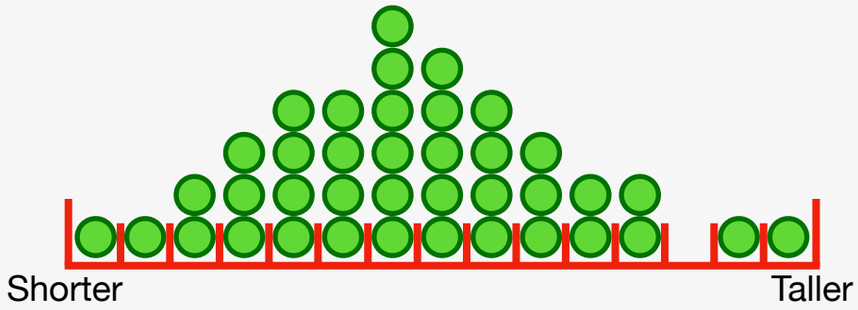
In summary, a small ***p*-value** *does not imply* that the effect size, or difference between Drug A and Drug B, is large.

DOUBLE BAM!!!

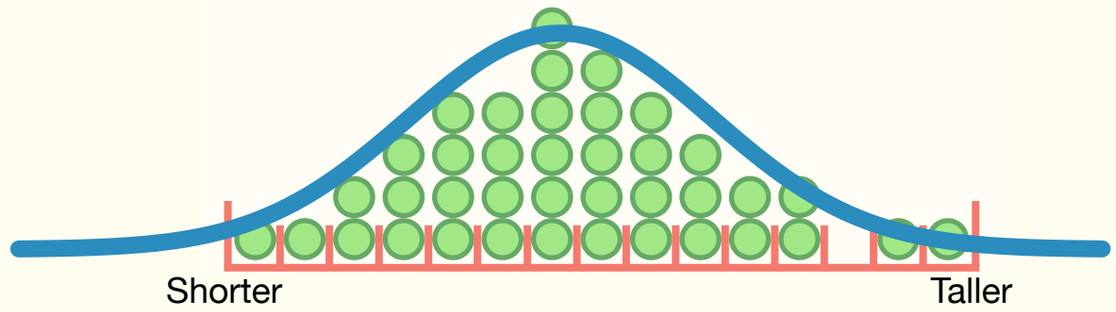


The Fundamental Concepts of Statistics: Summary

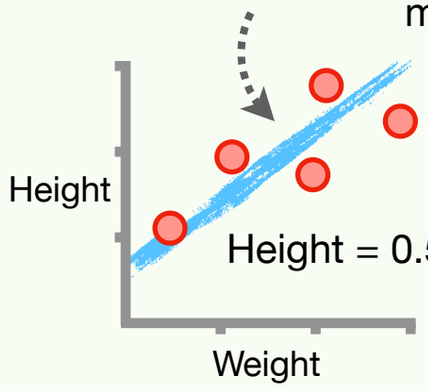
1 We can see trends in data with **histograms**. We'll learn how to use **histograms** to make classifications with **Naive Bayes** in **Chapter 7**.



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

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

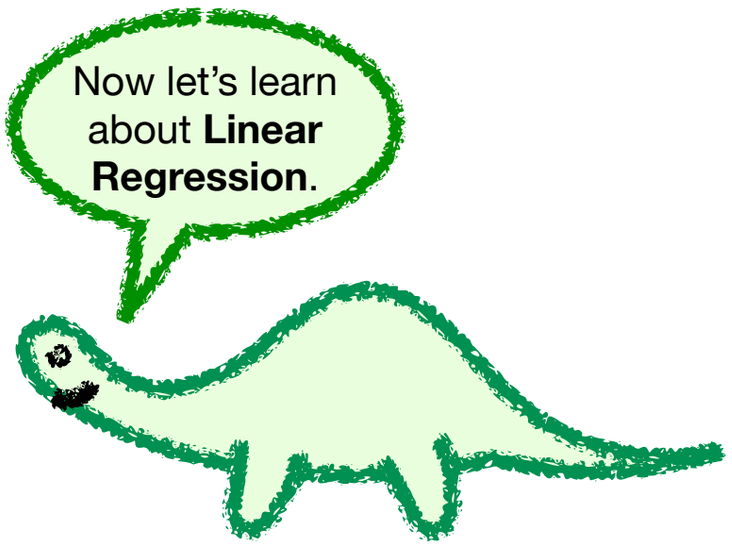
$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{SSR}}{n}$$

…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!!!

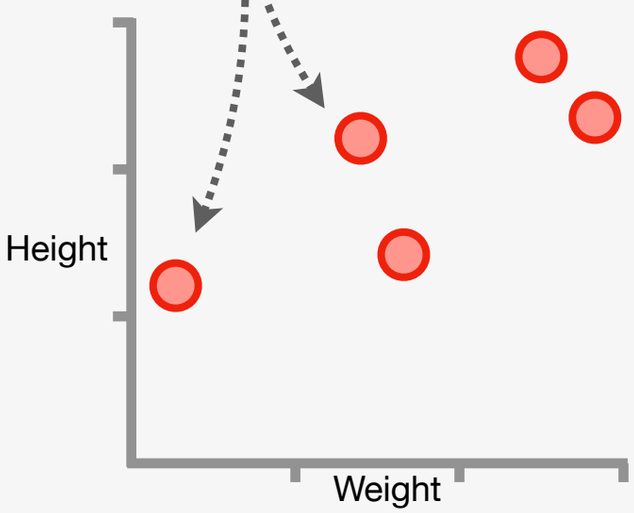


Chapter 04

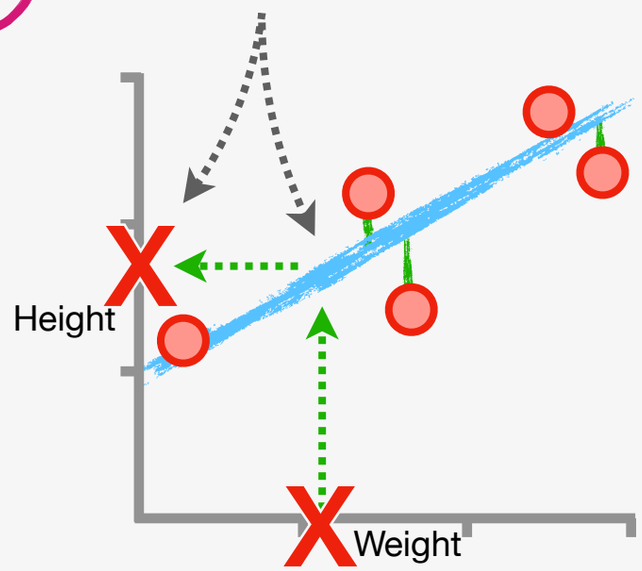
Linear Regression!!!

Linear Regression: Main Ideas

1 **The Problem:** We've collected Weight and Height measurements from **5** people, and we want to use Weight to predict Height, which is ***continuous***…



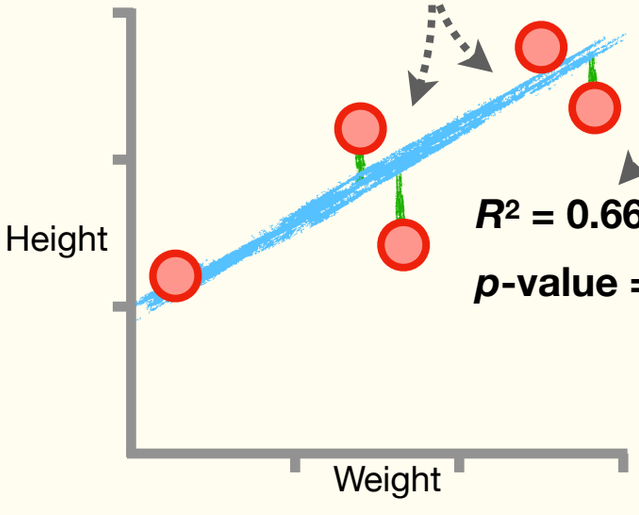
…and in **Chapter 3**, we learned that we could fit a **line** to the data and use it to make predictions.



However, **1)** we didn't talk about how we fit a **line** to the data…

…and **2)** we didn't calculate a ***p*-value** for the **fitted line**, which would quantify how much confidence we should have in its predictions compared to just using the **mean** y-axis value.

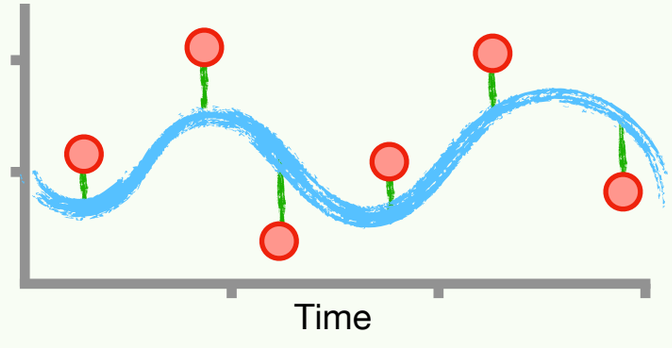
2 **A Solution: Linear Regression** fits a **line** to the data that *minimizes* the **Sum of the Squared Residuals** (**SSR**)…



…and once we fit the line to the data, we can easily calculate R^2, which gives us a sense of how accurate our predictions will be…

…and **Linear Regression** provides us with a ***p*-value** for the R^2 value, so we can get a sense of how confident we should be that the predictions made with the **fitted line** are better than predictions made with the **mean** of the y-axis coordinates for the data.

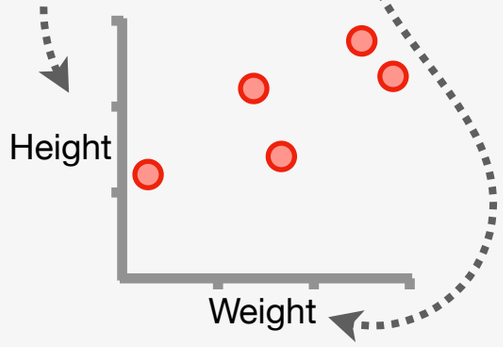
NOTE: Linear Regression is the gateway to a general technique called **Linear Models**, which can be used to create and evaluate models that go way beyond fitting simple straight lines to data!!!



BAM!!!

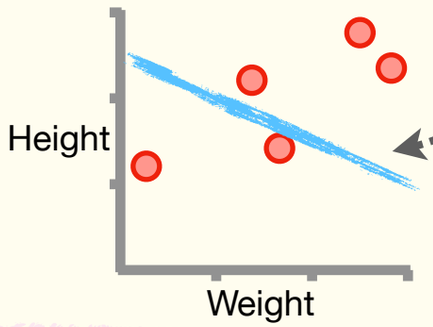
Fitting a Line to Data: Main Ideas

1 Imagine we had Height and Weight data on a graph…

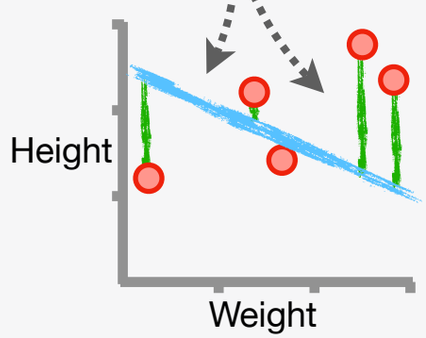


…and we wanted to predict Height from Weight.

2 Because the heavier Weights are paired with taller Heights, this **line** makes terrible predictions.



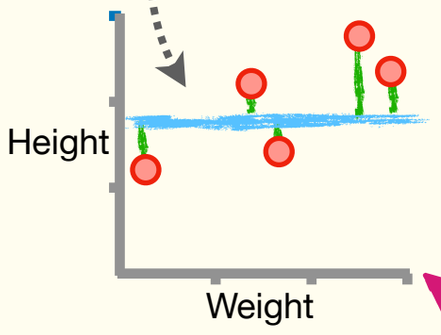
3 We can quantify how bad these predictions are by calculating the **Residuals**, which are the differences between the Observed and Predicted heights…



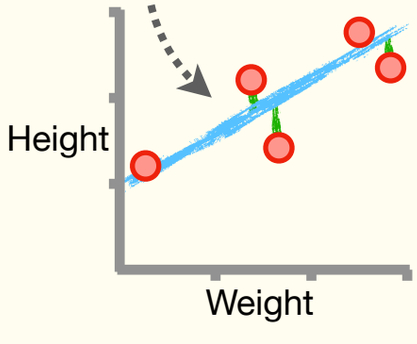
…and using the **Residuals** to calculate the **Sum of the Squared Residuals** (**SSR**).

Then we can plot the **SSR** on this graph that has the **SSR** on the y-axis, and different lines fit to the data on the x-axis.

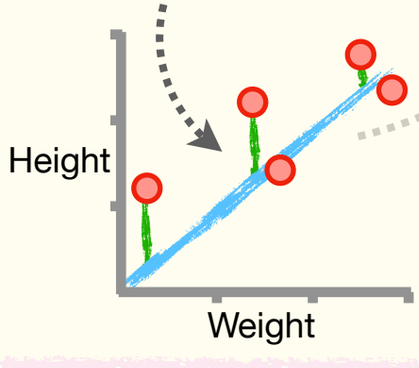
4 This **line**, which has a different y-axis intercept and slope, gives us slightly smaller residuals and a smaller **SSR**…

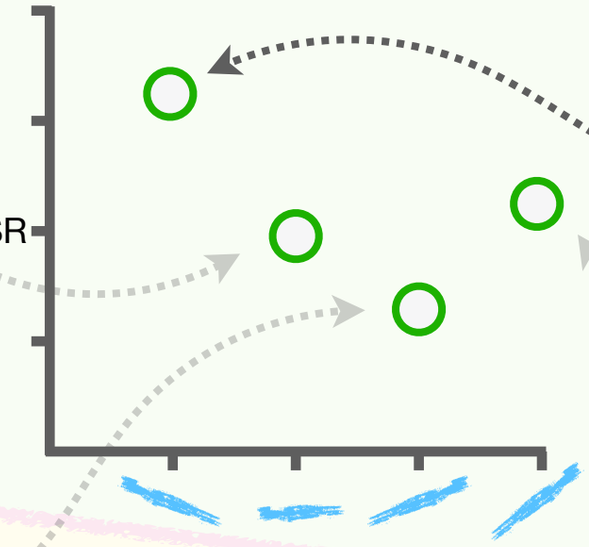


…and this **line** has even smaller residuals and a smaller **SSR**…



…and this **line** has larger residuals and a larger **SSR**.





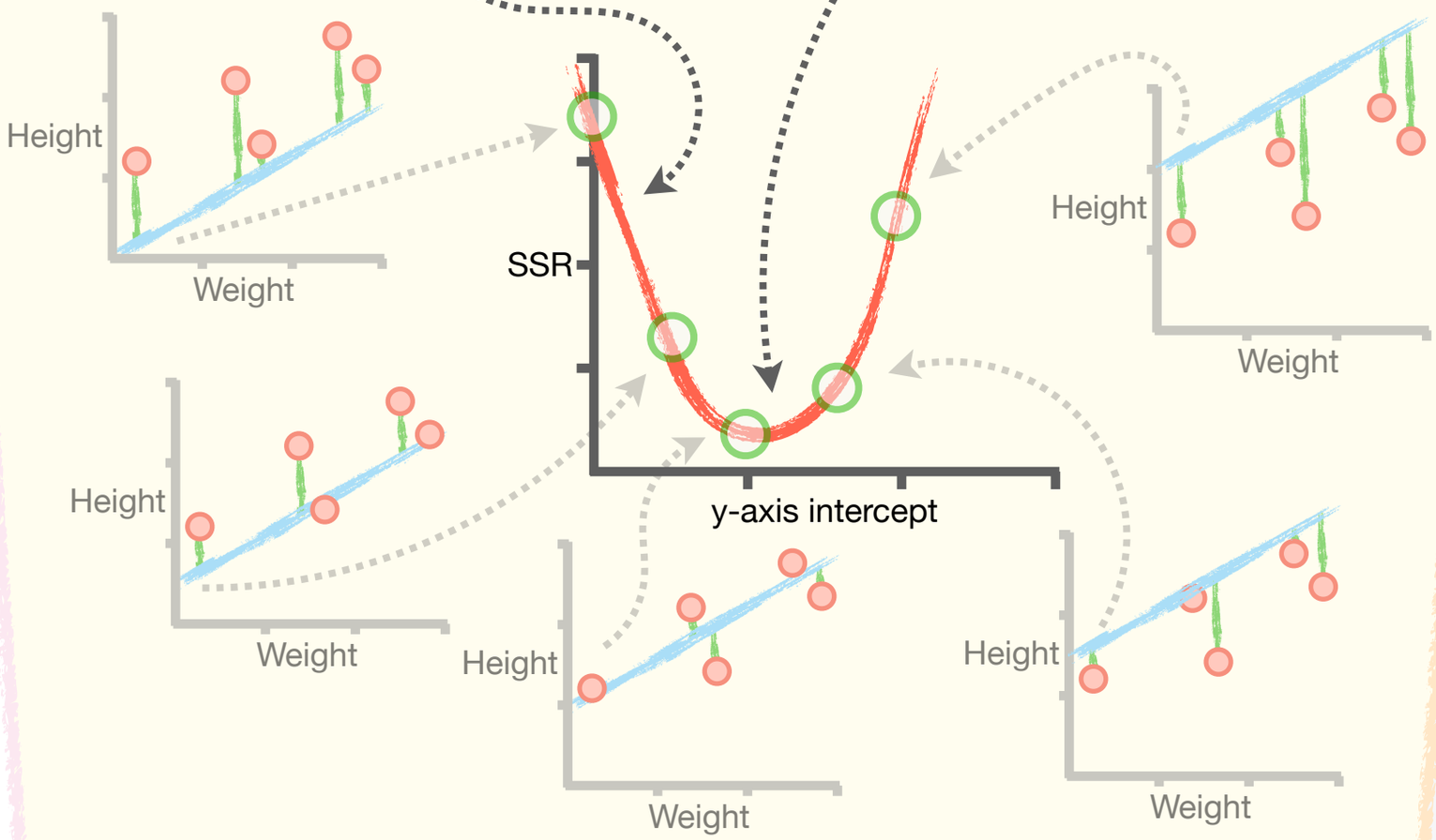
5 As we can see on the graph, different values for a **line**'s y-axis intercept and slope, shown on the x-axis, change the **SSR**, shown on the y-axis. **Linear Regression** selects the **line**, the y-axis intercept and slope, that results in the minimum **SSR**.

BAM!!!

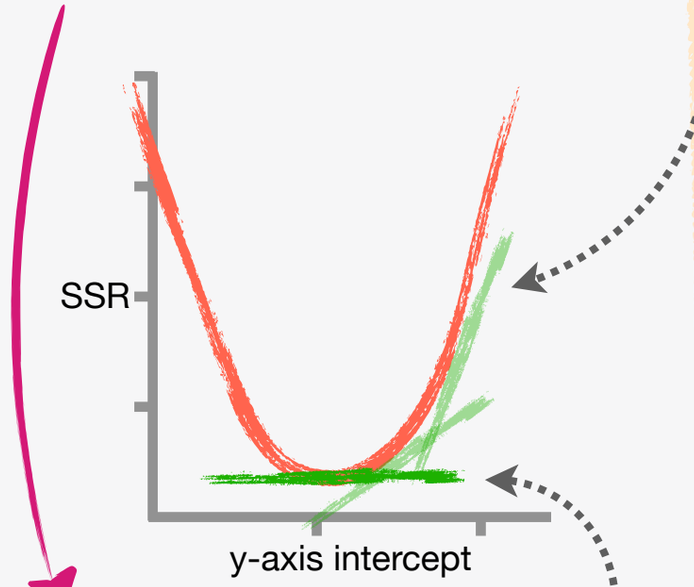
Fitting a Line to Data: Intuition

1 If we don't change the slope, we can see how the **SSR** changes for different y-axis intercept values…

…and, in this case, the goal of **Linear Regression** would be to find the y-axis intercept that results in the lowest **SSR** at the bottom of this curve.



2 One way to find the lowest point in the **curve** is to calculate the **derivative** of the **curve** (**NOTE:** If you're not familiar with derivatives, see **Appendix D**).



…and solve for where the **derivative** is equal to **0**, at the bottom of the **curve**.

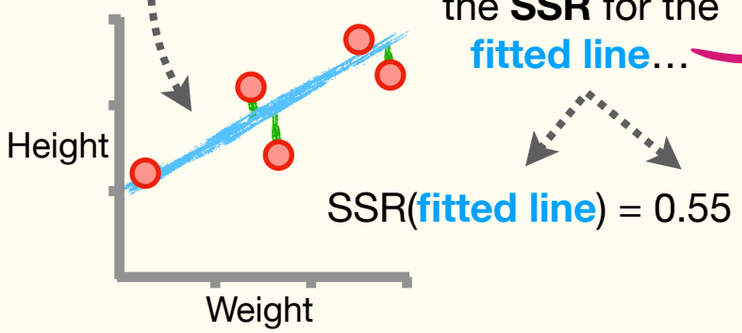
Solving this equation results in an **Analytical Solution**, meaning, we end up with a formula that we can plug our data into, and the output is the optimal value. Analytical solutions are awesome when you can find them (like for **Linear Regression**), but they're rare and only work in very specific situations.

3 Another way to find an optimal slope and y-axis intercept is to use an **Iterative Method** called **Gradient Descent**. In contrast to an **Analytical Solution**, an **Iterative Method** starts with a guess for the value and then goes into a loop that improves the guess one small step at a time. Although **Gradient Descent** takes longer than an analytical solution, it's one of the most important tools in machine learning because it can be used in a wide variety of situations where there are no analytical solutions, including **Logistic Regression**, **Neural Networks**, and many more.

Because **Gradient Descent** is so important, we'll spend all of **Chapter 5** on it. **GET EXCITED!!!**

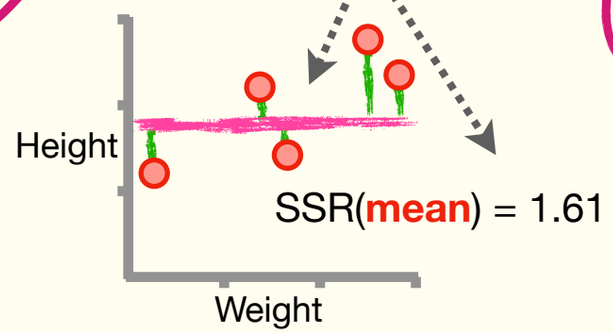
p-values for Linear Regression and R^2: Main Ideas

1 Now, assuming that we've fit a line to the data that minimizes the **SSR** using an analytical solution or **Gradient Descent**, we calculate R^2 with the **SSR** for the **fitted line**…



SSR(**fitted line**) = 0.55

…and the **SSR** for the **mean** height…



SSR(**mean**) = 1.61

…and plug them into the equation for R^2 and get **0.66**.

Gentle Reminder:

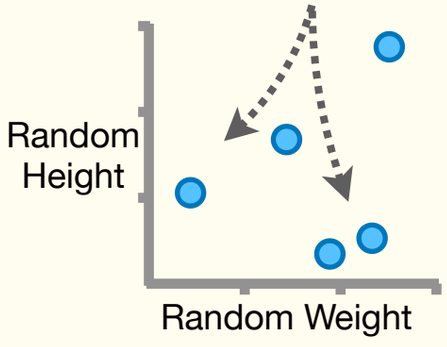
$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$

$$R^2 = \frac{1.61 - 0.55}{1.61} = 0.66$$

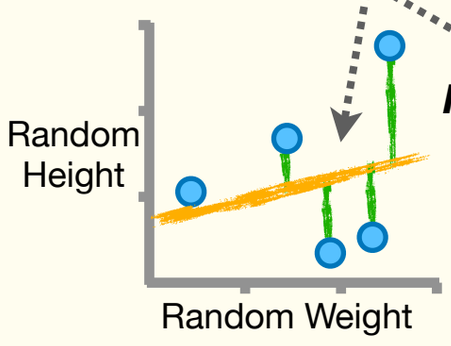
2 The R^2 value, **0.66**, suggests that using Weight to predict Height will be useful, but now we need to calculate the ***p*-value** to make sure that this result isn't due to random chance.

In this context, a ***p*-value** tells us the probability that random data could result in a similar R^2 value or a better one. In other words, the ***p*-value** will tell us the probability that random data could result in an $R^2 \geq 0.66$.

3 Because the original dataset has **5** pairs of measurements, one way* to calculate a ***p*-value** is to pair **5** *random* values for Height with **5** *random* values for Weight and plot them on a graph…



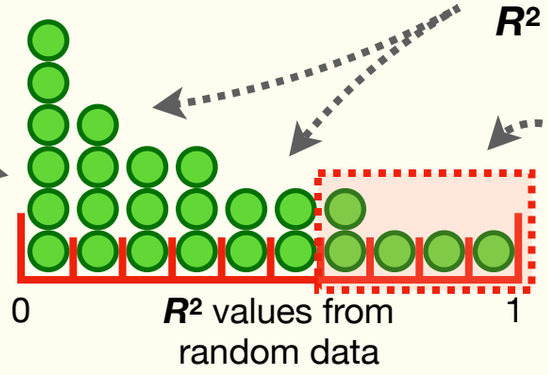
…then use **Linear Regression** to fit a line to the random data and calculate R^2…



$R^2 = 0.03$

…and then add that R^2 to a histogram…

…and then create **>10,000** more sets of random data and add their R^2 values to the histogram…



…and use the histogram to calculate the probability that random data will give us an $R^2 \geq 0.66$.

4 In the end, we get ***p*-value = 0.1**, meaning there's a **10%** chance that random data could give us an $R^2 \geq 0.66$. That's a relatively *high* ***p*-value**, so we might not have a lot of confidence in the predictions, which makes sense because we didn't have much data to begin with.

small bam.

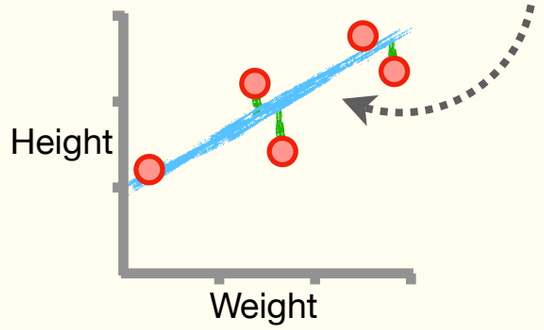
*** NOTE:** Because **Linear Regression** was invented before computers could quickly generate random data, **this is not the traditional way to calculate *p*-values**, but it works!!!

Multiple Linear Regression: Main Ideas

1 So far, the example we've used demonstrates something called **Simple Linear Regression** because we use one variable, Weight, to predict Height…

Height = 1.1 + 0.5 x **Weight**

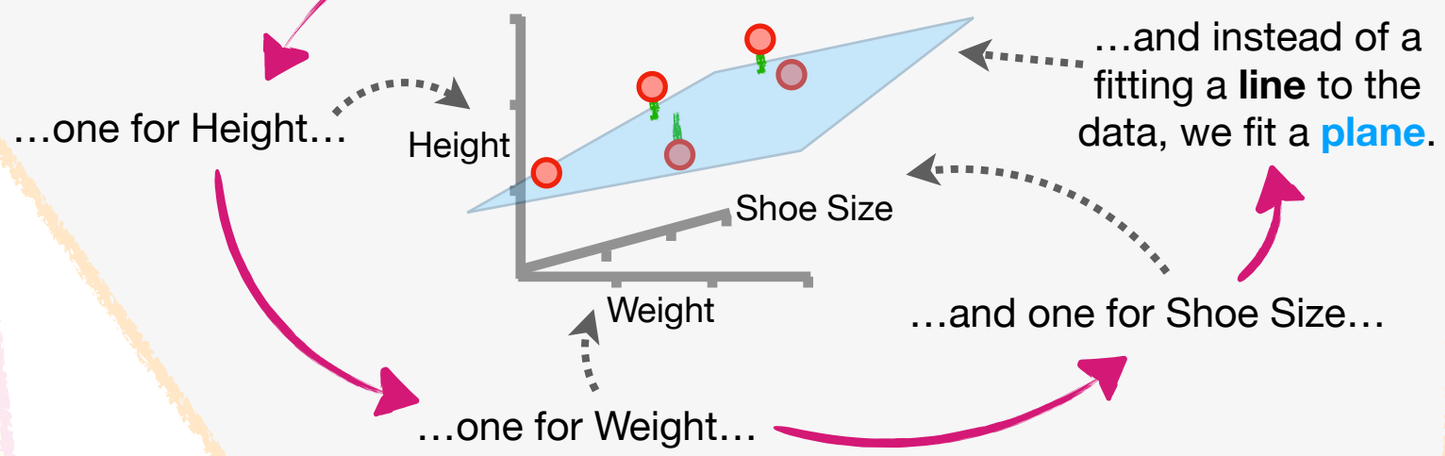
…and, as we've seen, **Simple Linear Regression** fits a line to the data that we can use to make predictions.



2 However, it's just as easy to use **2** or more variables, like Weight and Shoe Size, to predict Height.

Height = 1.1 + 0.5 x **Weight** + 0.3 x **Shoe Size**

This is called **Multiple Linear Regression**, and in this example, we end up with a **3**-dimensional graph of the data, which has **3** axes…



3 Just like for **Simple Linear Regression**, **Multiple Linear Regression** calculates R^2 and ***p*-values** from the **Sum of the Squared Residuals** (**SSR**). And the **Residuals** are still the difference between the **Observed Height** and the **Predicted Height**.

The only difference is that now we calculate **Residuals** around the **fitted plane** instead of a line.

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

4 And when we use **3** or more variables to make a prediction, we can't draw the graph, but we can still do the math to calculate the **Residuals** for R^2 and its ***p*-value**.

Bam.