SORBONNE UNIVERSITÉ ABU DHABI

scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Optimization in Statistics

ADIA Executive Program

Day 3 & 4 – Morning Session

Dr. Tanujit Chakraborty

Faculty @ Sorbonne

tanujitisi@gmail.com

Content

- Basic data preprocessing and descriptive statistics
- Fitting a line to data
- Gradient Descent

Content

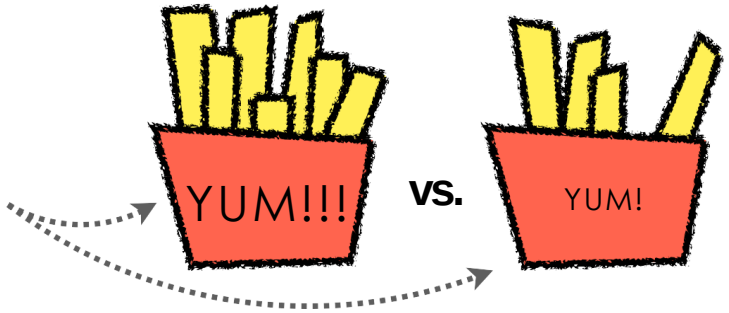
- Basic data preprocessing and descriptive statistics
- Fitting a line to data
- Gradient Descent

Statistics: Main Ideas

The Problem:

The world is an interesting place, and things are not always the same.

For example, every time we order french fries, we don't always get the exact same number of fries.



A Solution:

Statistics provides us with a set of **tools** to **quantify the variation** that we find in everything and, for the purposes of machine learning, **helps us make predictions** and **quantify how confident** we should be in those predictions.

For example, once we notice that we don't always get the exact same number of fries, we can keep track of the number of fries we get each day…

…and statistics can help us predict how many fries we'll get the next time we order them, and it tells us how confident we should be in that prediction.

Fry Diary

Monday: 21 fries

Tuesday: 24 fries

Wednesday: 19 fries

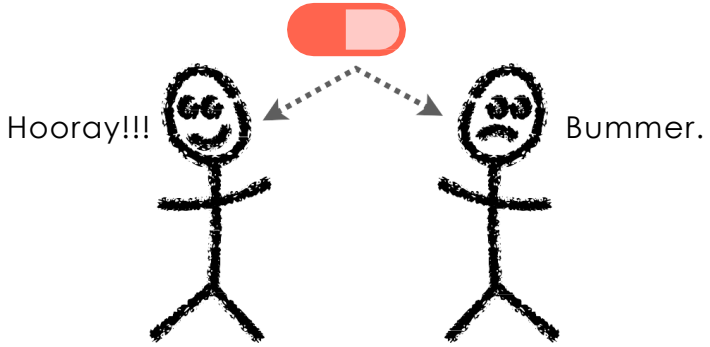
Thursday: ???

Statistics: Main Ideas

The Problem:

The world is an interesting place, and things are not always the same.

Alternatively, if we have a new medicine that helps some people but hurts others…



A Solution:

Statistics provides us with a set of **tools** to **quantify the variation** that we find in everything and, for the purposes of machine learning, **helps us make predictions** and **quantify how confident** we should be in those predictions.

…statistics can help us predict who will be helped by the medicine and who will be hurt, and it tells us how confident we should be in that prediction. This information can help us make decisions about how to treat people.

For example, if we predict that the medicine will help, but we're not very confident in that prediction, we might not recommend the medicine and use a different therapy to help the patient.

Statistics: Main Ideas

The first step in making predictions is to identify trends in the data that we've collected, so let's talk about how to do that with a **Histogram**.

Discrete vs Continuous Data

Discrete Data…

…is ***countable*** and only takes *specific* values.

For example, we can count the number of people that love the color **green** or love the color **blue**.

4 people love **green**

3 people love **blue**

Because we are counting individual people, and the totals can only be whole numbers, the data are **Discrete**.

American shoe sizes are **Discrete** because even though there are half sizes, like **8 1/2**, shoes sizes are never **8 7/36** or **9 5/18**.

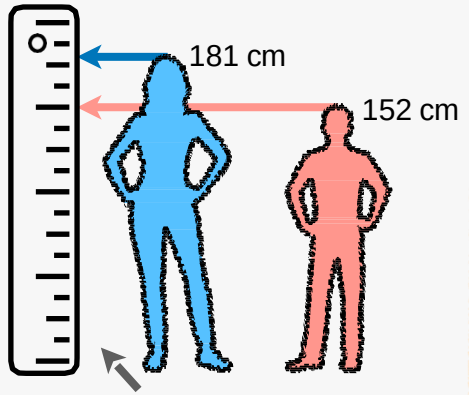
Rankings and other orderings are also **Discrete**. There is no award for coming in **1.68** place.

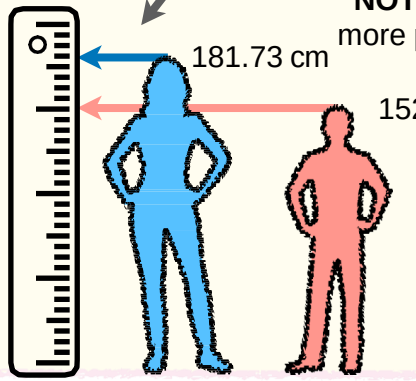
Continuous Data…

…is ***measurable*** and can take *any* numeric value within a range.

For example, Height measurements are **Continuous** data.

Height measurements can be any number between **0** and the height of the tallest person on the planet.





NOTE: If we get a more precise ruler…

…then the measurements get more precise.

So the precision of **Continuous** measurements is only limited by the tools we use.

Statistics: Main Ideas

The first step in making predictions is to identify trends in the data that we've collected, so let's talk about how to do that with a **Histogram**.

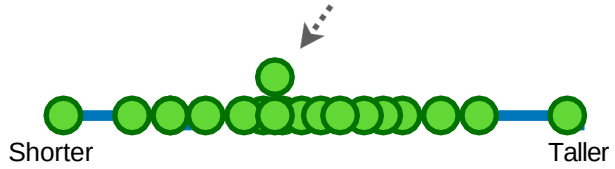
Histogram

The Problem:

We have a lot of measurements and want to gain insights into their hidden trends.

For example, imagine we measured the Heights of so many people that the data, represented by green dots, overlap, and some green dots are completely hidden.

We could try to make it easier to see the hidden measurements by stacking any that are exactly the same…



…but measurements that are exactly the same are rare, and a lot of the green dots are still hidden.

☹

Histogram

The Problem:

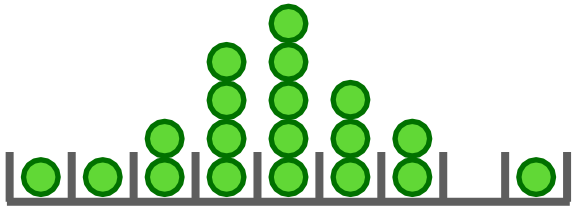
We have a lot of measurements and want to gain insights into their hidden trends.

A Solution:

Histograms are one of the most basic, but surprisingly **useful**, statistical tools that we can use to **gain insights into data**.

Instead of stacking measurements that are exactly the same, we divide the range of values into bins…

and stack the measurements that fall in the same bin…

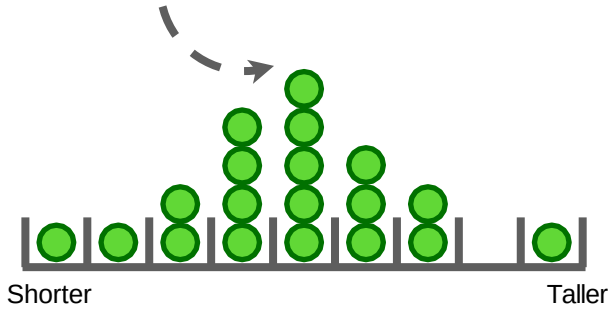


…and we end up with a **histogram!!!**

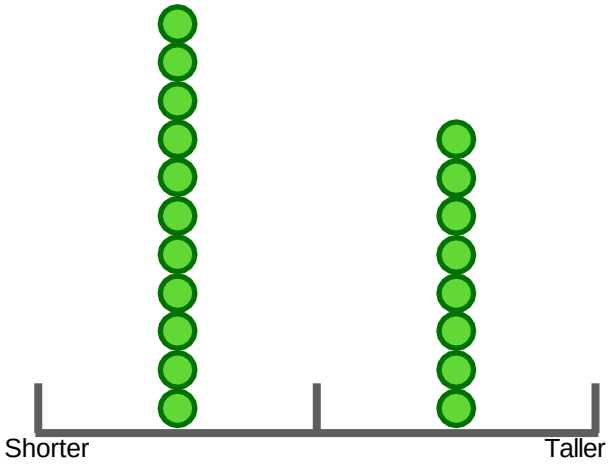
The **histogram** makes it easy to see trends in the data. In this case, we see that most people had close to average heights.

Histogram

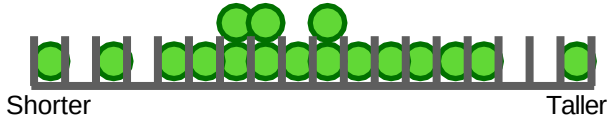
The taller the stack within a bin, the more measurements we made that fall into that bin.



NOTE: Figuring out how wide to make the bins can be tricky.



If the bins are too wide, then they're not much help…

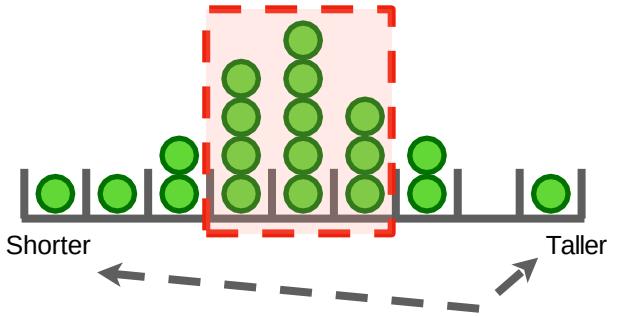


…and if the bins are too narrow, then they're not much help…

…so, sometimes you have to try a bunch of different bin widths to get a clear picture.

Histogram: Calculating Probabilities

We can use the histogram to estimate the **probability** of getting future measurements



Extremely short or tall measurements are rarer and less likely to happen in the future.

Because most of the measurements are inside this red box, we might be willing to bet that the next measurement we make will be somewhere in this range.

If we want to estimate the probability that the next measurement will be in this red box…

…we count the number of measurements, or observations, in the box and get **12**…

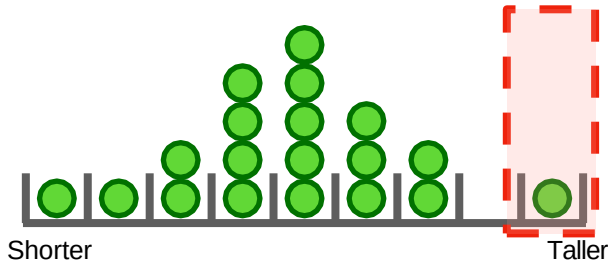
…and divide by the total number of measurements, **19**…

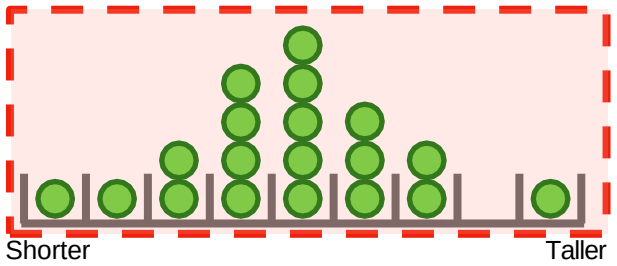
…and we get $\frac{12}{19} \approx 0.63$ .In theory, this means that 63% of the time we'll get a measurement in the red box.

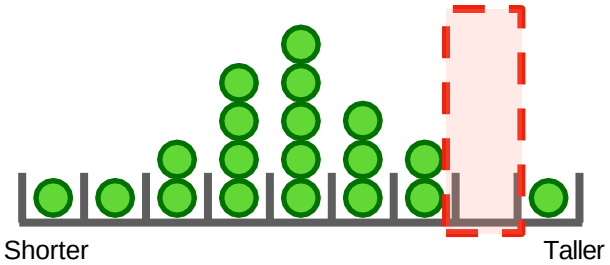
However, the confidence we have in this estimate depends on the number of measurements.

Histogram: Calculating Probabilities

What is the probability that the next measurement will be in this red box…

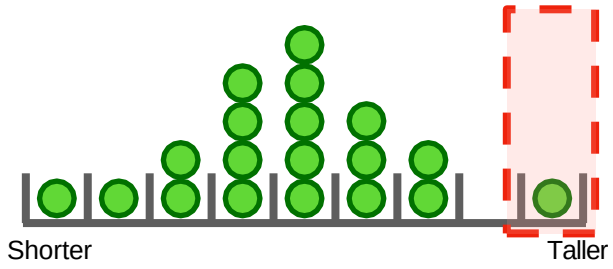






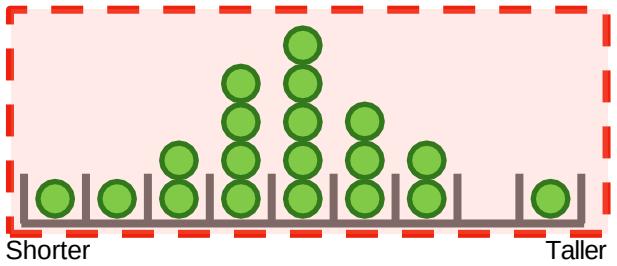
Histogram: Calculating Probabilities

What is the probability that the next measurement will be in this red box…



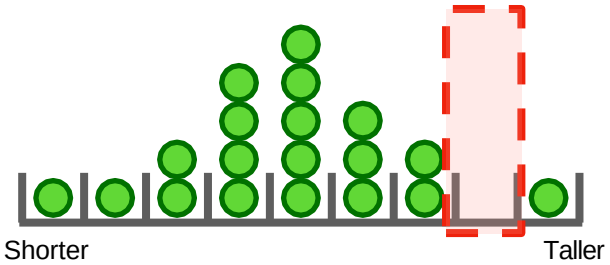
$$\frac{1}{19} \approx 0.05$$

In theory, there's a 5% chance that the next measurement will fall within the box. In other words, it's **fairly rare** to measure someone who is really tall.



$$\frac{19}{19} = 1$$

There's a 100% chance that the next measurement will fall within the box. In other words, the **maximum probability is 1**.

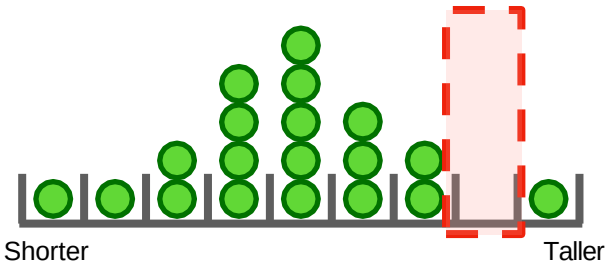


$$\frac{0}{19} = 0$$

…and we get 0. This is the minimum probability and, in theory, it means that we'll never get a measurement in this box.

Histogram: Calculating Probabilities

What is the probability that the next measurement will be in this red box…



- However, it could be that the only reason the box was empty is that we simply did not measure enough people.
- If we measure more people, we may either find someone who fits in this bin or become more confident that it should be empty. However, sometimes getting more measurements can be expensive, or take a lot of time, or both. This is a problem!!!
- The good news is that we can solve this problem with a **Probability Distribution.**

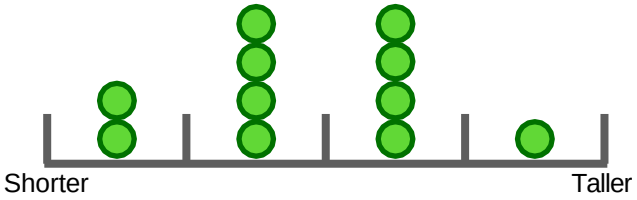
$$\frac{0}{19} = 0$$

…and we get 0. This is the minimum probability and, in theory, it means that we'll never get a measurement in this box.

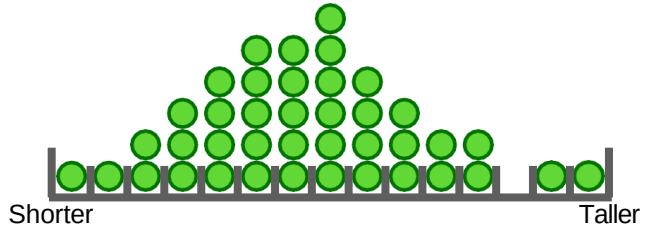
Probability distributions

The Problem:

If we don't have much data, then we can't make very precise probability estimates with a histogram…

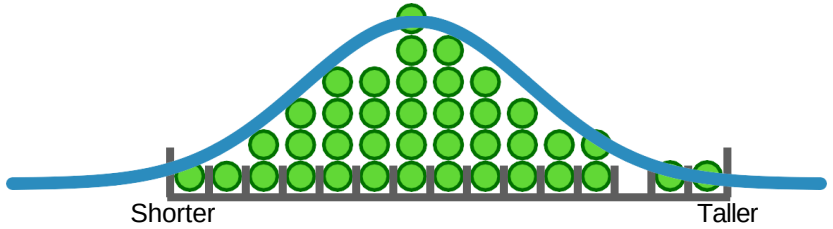


…however, collecting tons of data to make precise estimates can be time-consuming and expensive.
Is there another way?

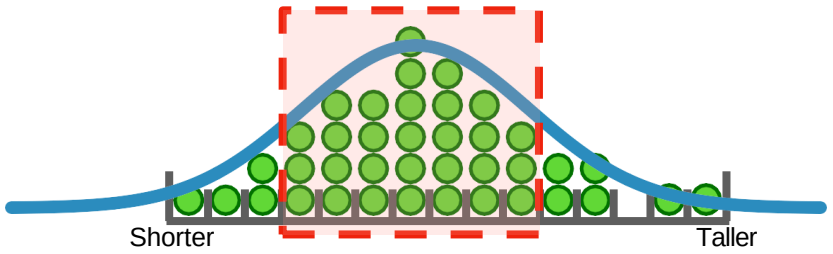


A Solution:

We can use a **Probability Distribution**, which, in this example, is represented by a blue, bell-shaped curve, to approximate a histogram.

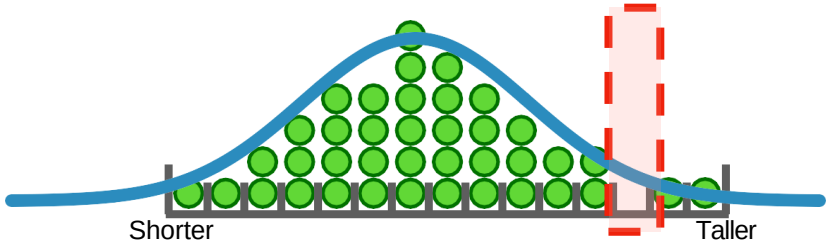


Probability distributions



This blue, bell-shaped curve tells us the same types of things that the histogram tells us.

For example, the relatively large amount of area under the curve in this red box tells us that there's a relatively high probability that we will measure someone whose value falls in this region.

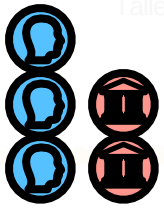


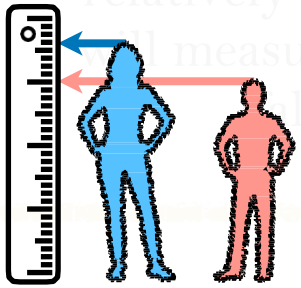
Now, even though we never measured someone who's value fell in this range…

…we can use the area under the curve to estimate the probability of measuring a value in this range.

Probability distributions

NOTE: Because we have Discrete and Continuous data…





…there are Discrete and Continuous Probability Distributions.

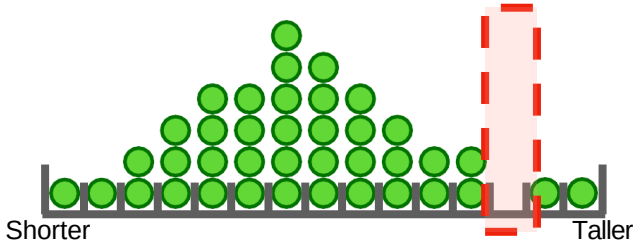
So let's start by learning about **Discrete Probability Distributions**.

Discrete Probability distributions

The Problem:

Although, technically speaking, histograms are **Discrete Distributions**, meaning data can be put into discrete bins and we can use those to estimate probabilities

……they require that we collect a lot of data, and it's not always clear what we should do with blank spaces in the histograms



A Solution:

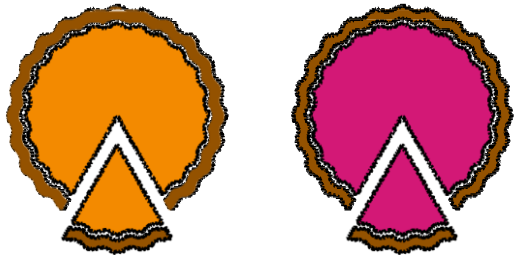
When we have discrete data, instead of collecting a ton of data to make a histogram and then worrying about blank spaces when calculating probabilities, we can let mathematical equations do all of the hard work for us.

Example: Binomial Distribution

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution

First, let's imagine we're walking down the street in **StatLand** and we ask the first 3 people we meet if they prefer pumpkin pie or blueberry pie…



Pumpkin Blueberry

…and the first 2 people say they prefer pumpkin pie and the last person says they prefer blueberry pie

Based on our extensive experience judging pie contests in **StatLand,** we know that **70%** of people prefer pumpkin pie, while **30%** prefer blueberry pie.

So now let's calculate the probability of observing that the first two people prefer pumpkin pie and the third person prefers blueberry.

The probability that the first person will prefer pumpkin pie is **0.7**…

0.7

…and the probability that the first two people will prefer pumpkin pie is **0.49**…

0.7 x 0.7 = **0.49**

…and the probability that the first two people will prefer pumpkin pie and the third person prefers blueberry is **0.147**.

0.7 x 0.7 x 0.3 = **0.147**

NOTE: 0.147 is the probability of observing that the first two people prefer pumpkin pie and the third person prefers blueberry…

…it is ***not*** the probability that **2** out of **3** people prefer pumpkin pie.

Binomial Distribution

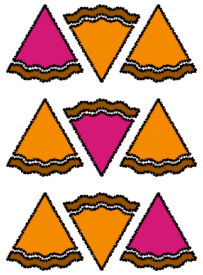
It could have just as easily been the case that the first person said they prefer blueberry and the last two said they prefer pumpkin.

0.3 x 0.7 x 0.7 = **0.147**

Likewise, if only the second person said they prefer blueberry, we would multiply the numbers together in a different order and still get 0.147.

0.7 x 0.3 x 0.7 = **0.147**

So, we see that all three combinations are equally probable…



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**

…and that means that the **probability** of observing that ***2* out of *3*** people **prefer pumpkin** pie is the sum of the 3 possible arrangements of people's pie preferences, 0.441.

Binomial Distribution

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

In our example, k is the number of people who prefer pumpkin pie, so in this case, $k = 2$

…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, $p = 0.7$

Binomial Distribution

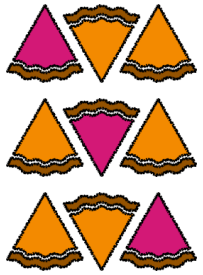
It could have just as easily been the case that the first person said they prefer blueberry and the last two said they prefer pumpkin.

0.3 x 0.7 x 0.7 = **0.147**

Likewise, if only the second person said they prefer blueberry, we would multiply the numbers together in a different order and still get 0.147.

0.7 x 0.3 x 0.7 = **0.147**

So, we see that all three combinations are equally probable…



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**

…and that means that the **probability** of observing that ***2* out of *3*** people **prefer pumpkin** pie is the sum of the 3 possible arrangements of people's pie preferences, 0.441.

Binomial Distribution

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} = \left(\frac{n!}{k!(n-k)!}\right) p^k (1-p)^{n-k}$$

Binomial Distribution

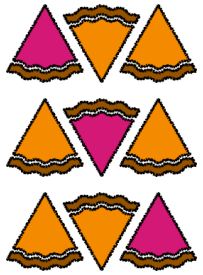
It could have just as easily been the case that the first person said they prefer blueberry and the last two said they prefer pumpkin.

0.3 x 0.7 x 0.7 = **0.147**

Likewise, if only the second person said they prefer blueberry, we would multiply the numbers together in a different order and still get 0.147.

0.7 x 0.3 x 0.7 = **0.147**

So, we see that all three combinations are equally probable…



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**

…and that means that the **probability** of observing that ***2* out of *3*** people **prefer pumpkin** pie is the sum of the 3 possible arrangements of people's pie preferences, 0.441.

Binomial Distribution

In our example, k is the number of people who prefer pumpkin pie, so in this case, $k = 2$

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

probability that **2 out of the 3** people prefer pumpkin pie.

probability that **1 out of 3** people prefers blueberry pie.

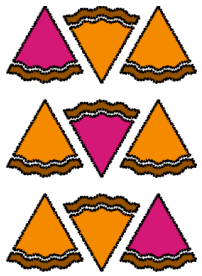
Binomial Distribution

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, $k = 2$, the number of people we asked, $n = 3$, and the probability that someone prefers pumpkin pie, $p = 0.7$

$$p(k = 2|n = 3, p = 0.7) = \left(\frac{n!}{k!(n-k)!}\right) p^k (1-p)^{n-k}$$
$$= \left(\frac{3!}{2!(3-1)!}\right) 0.7^2 (1-0.7)^{3-2} = 3 \times 0.7^2 \times (0.3)^1 = 0.441$$

Recall: The three combinations



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

+

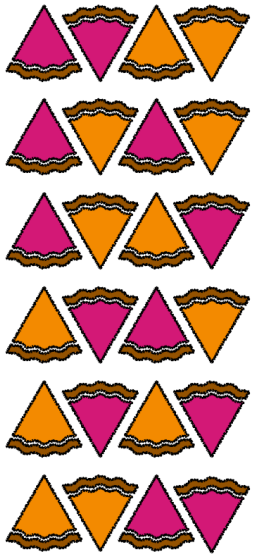
+

= 0.441

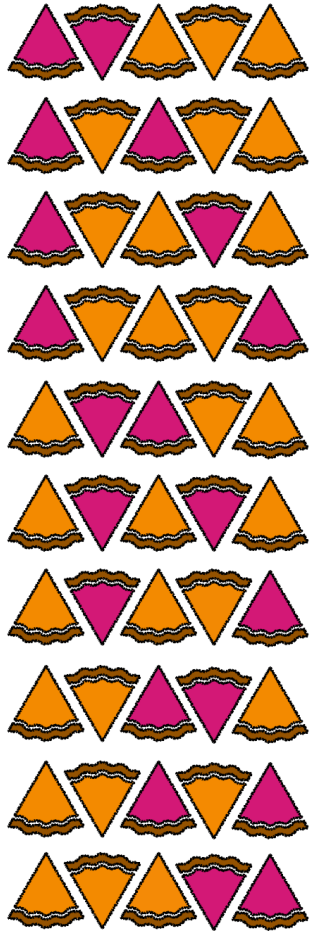
Binomial Distribution

Drawing all the combinations can be very tedious!

For example, if we wanted to calculate the probability of observing that **2** out of **4** people prefer pumpkin pie, we have to calculate and sum the individual probabilities from 6 different arrangements…



…and there are 10 ways to arrange **3** out of **5** people who prefer pumpkin pie.

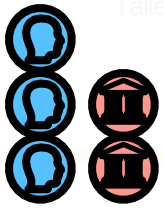


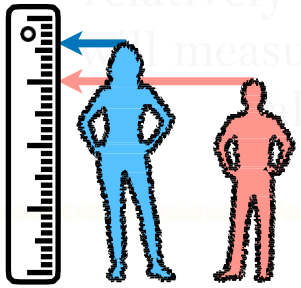
So, instead of drawing out different arrangements of pie slices, we can use the equation for the Binomial Distribution to calculate the probabilities directly.

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability distributions

NOTE: Because we have Discrete and Continuous data…





…there are Discrete and Continuous Probability Distributions.

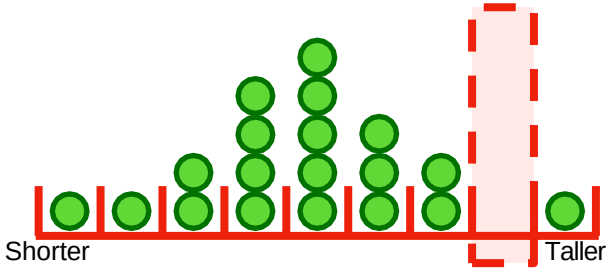
let's now learning about
Continuous Probability Distributions.

Continuous Probability distributions

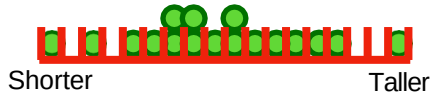
The Problem:

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

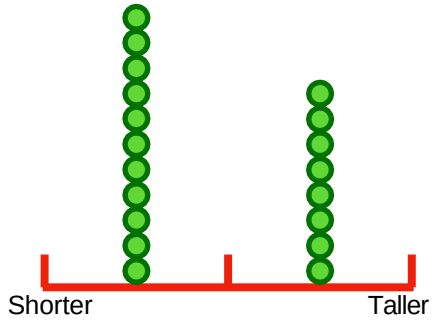
(1) it's not always clear what to do about gaps in the data



(2) Histograms can be very sensitive to the size of the bins



bins too narrow!

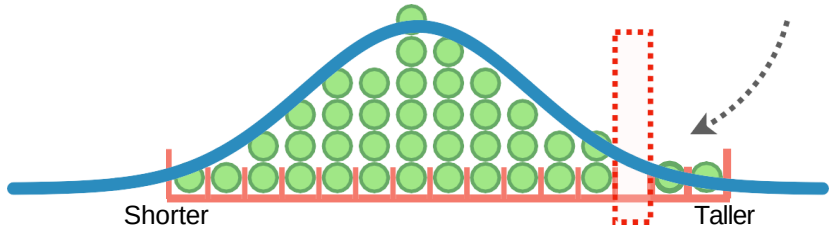


bins too wide!

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.



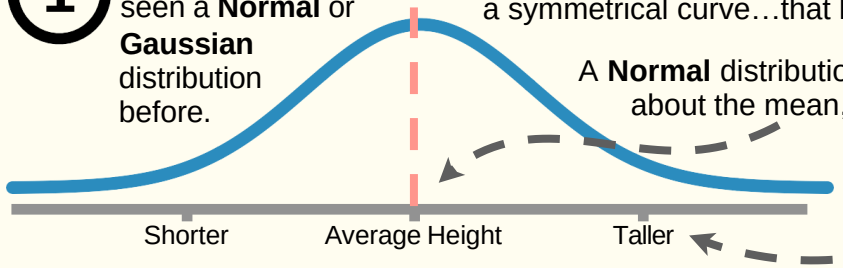
There are lots of commonly used **Continuous Distributions**. Now we'll talk about the most useful of all, the **Normal Distribution**.

The normal (gaussian) distribution

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

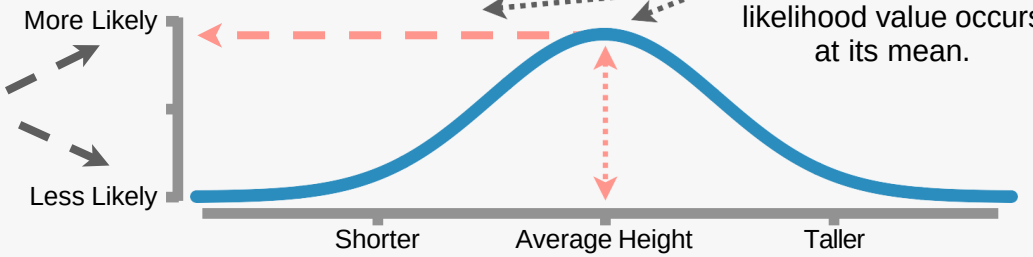
It's also called a **Bell-Shaped Curve** because it's a symmetrical curve…that looks like a bell.

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.



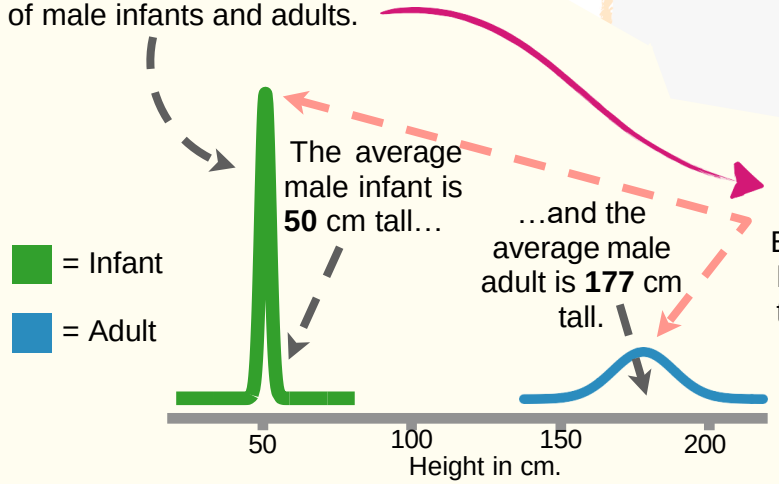
The **Normal Distribution**'s maximum likelihood value occurs at its mean.

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.

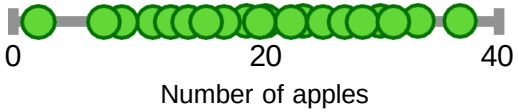


Because the normal distribution for infants has a higher peak than the one for adults, we can see that there's a higher likelihood that an infant will be close to its mean than an adult will be close to its mean. The difference in peak height tells us there's less variation in how tall an infant is compared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

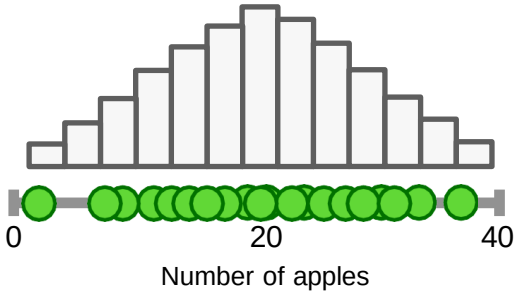
Lots of things can be approximated with **Normal Distributions**: Height, birth weight, blood pressure, job satisfaction, and many more!!!

Mean, Variance and Standard Deviation

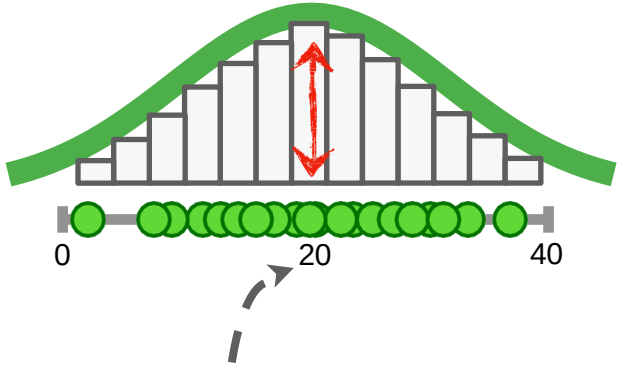
1 Imagine we went to *all* **5,132 Spend-n-Save** food stores and counted the number of **green apples** that were for sale. We could plot the number of **green apples** at each store on this number line…



…but because there's a lot of overlap in the data, we can also draw a **Histogram** of the measurements.



2 If we wanted to fit a **Normal Curve** to the data like this …then, first, we need to calculate the **Population Mean** to figure out where to put the center of the curve



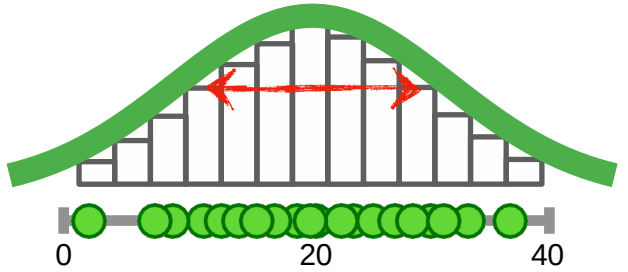
3 Because we counted the number of **green apples** in *all* **5,132 Spend-n-Save** stores, calculating the **Population Mean**, which is frequently denoted with the Greek character ***μ*** (***mu***), is relatively straightforward: we simply calculate the average of all of the measurements, which, in this case, is **20**.

$$\textbf{Population Mean} = \boldsymbol{\mu} = \frac{\text{Sum of Measurements}}{\text{Number of Measurements}}$$

$$= \frac{2 + 8 + \ldots. + 37}{5132} = 20$$

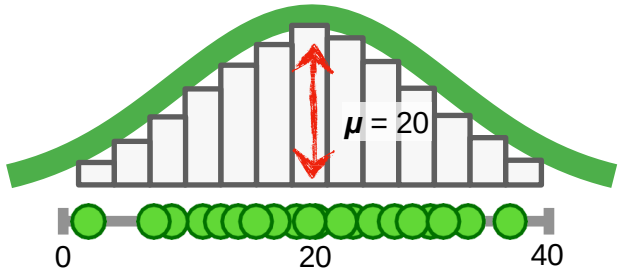
4 Because the **Population Mean**, ***μ***, is **20**, we center the **Normal Curve** over **20**.

5 Now we need to determine with width of the curve by calculating the **Population Variance** (also called the **Population Variation**) and **Standard Deviation**.



Mean, Variance and Standard Deviation

6 In other words, we want to calculate how the data are spread around the **Population Mean** (which, in this example, is **20**).

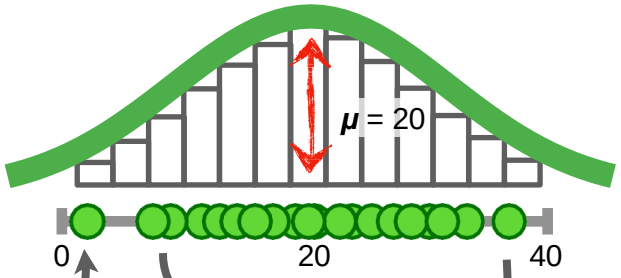


7 The formula for calculating the **Population Variance** is…

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

8 The part in the parentheses, ***x* - *μ***, means we subtract the **Population Mean**, ***μ***, from each measurement, ***x***.

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



For example, the first measurement is **2**, so we subtract ***μ***, which is **20**, from **2**…

(2 - 20) (8 - 20) (28 - 20)

…then the square tells us to square each term…

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

(2 - 20)² (8 - 20)² (28 - 20)²

…and the Greek character **Σ** (**Sigma**) tells us to add up all of the terms…

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

(2 - 20)² + (8 - 20)² + … + (28 - 20)²

Number of Measurements

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

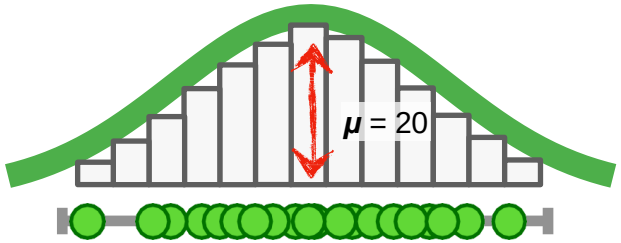
…and lastly, we want the average of the squared differences, so we divide by the total number of measurements, ***n***, which, in this case, is *all* **Spend- n-Save** food stores, **5,132**.

Mean, Variance and Standard Deviation

9 Now that we know how to calculate the **Population Variance**…

$$\text{Population Variance} = \sum \frac{(x-\mu)^2}{n} = \frac{(2-20)^2 + (8-20)^2 + \cdots + (28-20)^2}{5132} = 100$$

…when we do the math, we get **100**.



10 Because each term in the equation for **Population Variance** is squared…

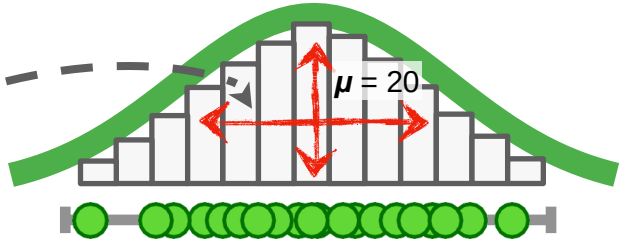
…the units for the result, **100**, are **Number of Apples Squared**…

…and that means we can't plot the **Population Variance** on the graph, since the units on the x- axis are not squared.

11 To solve this problem, we take the *square root* of the **Population Variance** to get the **Population Standard Deviation**…

$$\text{Population standard deviation} = \sigma = \sqrt{\sum \frac{(x-\mu)^2}{n}} = \sqrt{100} = 10$$

…and we can plot that on the graph.

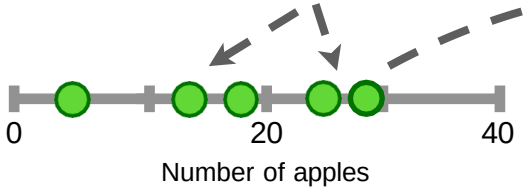


12 **NOTE:** Before we move on, I want to emphasize that we almost never have the population data, so we almost never calculate the **Population Mean**, **Variance**, or **Standard Deviation**.

If we don't usually calculate **Population Parameters**, what do we do???

Mean, Variance and Standard Deviation

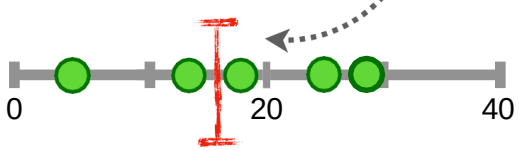
13 Instead of calculating **Population Parameters**, we *estimate* them from a relatively small number of measurements.



14 *Estimating* the **Population Mean** is super easy: we just calculate the average of the measurements we collected…

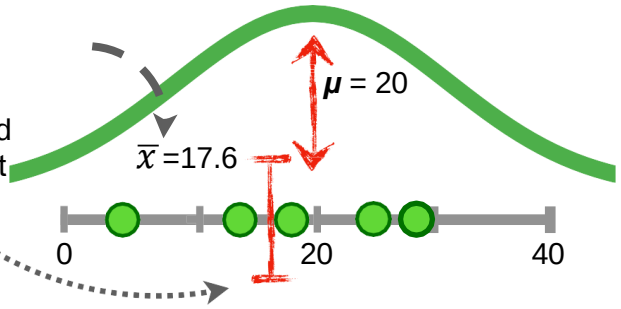
$$\textbf{Estimated Mean} = \frac{\text{Sum of Measurements}}{\text{Number of Measurements}} = \frac{3 + 13 + 19 + 24 + 29}{5} = 17.6$$

…and when we do the math, we get **17.6**.



15 **NOTE:** The **Estimated Mean**, which is often denoted with the symbol $\bar{x}$ (x-bar), is also called the **Sample Mean** and due to the relatively small number of measurements used to calculate the **Estimated Mean**, it's different from the **Population Mean**

A lot of **Statistics** is dedicated to quantifying and compensating for the differences between **Population Parameters**, like the **Mean** and **Variance**, and their *estimated* counterparts.



16 Now that we have an **Estimated Mean**, we can calculate an **Estimated Variance** and **Standard Deviation**. However, we have to compensate for the fact that we only have an **Estimated Mean**, which will almost certainly be different from the **Population Mean**.

17 Thus, when we calculate the **Estimated Variance** and **Standard Deviation** using the **Estimated Mean**…

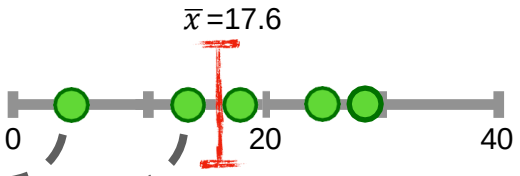
…we compensate for the difference between the **Population Mean** and the **Estimated Mean** by dividing by number of measurements minus **1**, ***n* - 1**, rather than ***n***.

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

$$\textbf{Estimated standard deviation} = \sqrt{\sum \frac{(x - \bar{x})^2}{n - 1}}$$

Mean, Variance and Standard Deviation

18 Now when we plug the data into the equation for the **Estimated Variance**…



NOTE: If we had divided by ***n***, instead of ***n* - 1**, we would have gotten **81.4**, which is a significant *underestimate* of the true **Population Variance**, **100**.

$$\textbf{Estimated Variance} = \sum \frac{(x-\bar{x})^2}{n-1} = \frac{(3-17.6)^2+(13-17.6)^2+(19-17.6)^2+(24-17.6)^2+(29-17.6)^2}{5-1} = 101.8$$

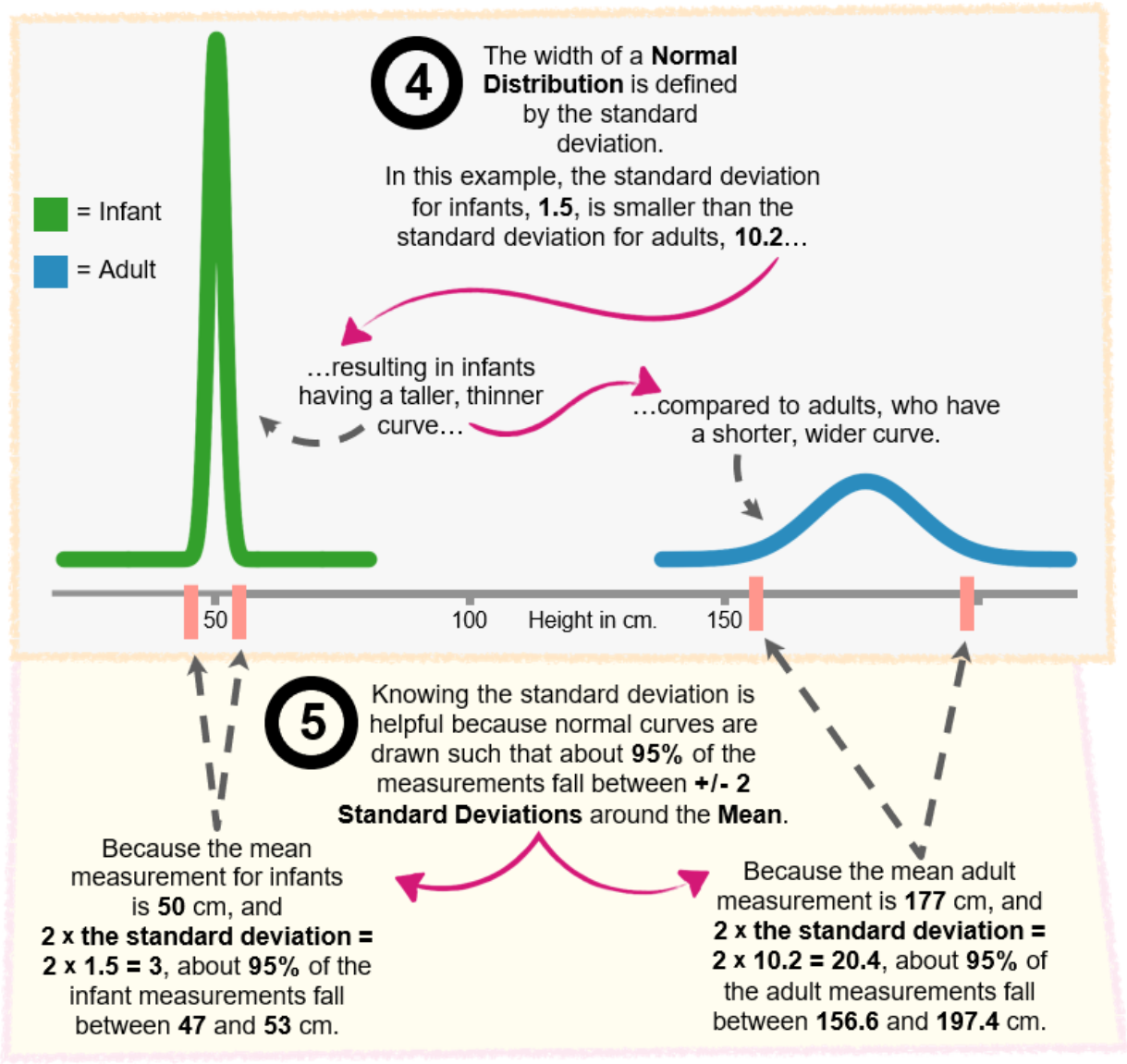
…we get **101.8**, which is a pretty good estimate of the **Population Variance**, which, as we saw earlier, is **100**.

19 Lastly, the **Estimated Standard Deviation** is just the square root of the **Estimated Variance**…

$$\textbf{Estimated standard deviation} = \sqrt{\sum \frac{(x-\bar{x})^2}{n-1}} = \sqrt{\textbf{Estimated Variance}} = \sqrt{101.8} = 10.1$$

…so, in this example, the **Estimated Standard Deviation** is **10.1**. Again, this is relatively close to the **Population** value we calculated earlier.

The normal (gaussian) distribution



The normal (gaussian) distribution

The equation for the **Normal Distribution** is:

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

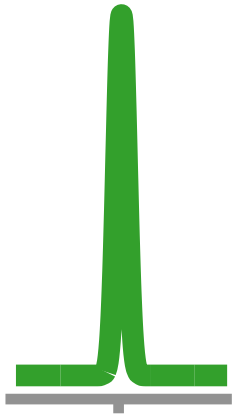
x is the x-axis coordinate. So, in this example, the x-axis represents **Height** and ***x* = 50**.

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**.

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.



50
Height in cm.

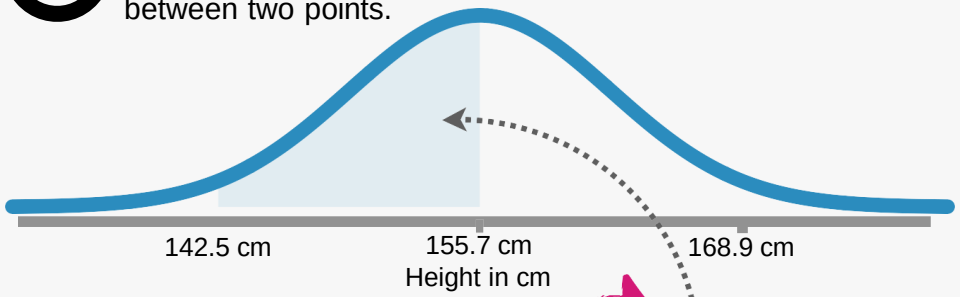
$$f(x=50|\mu=50,\sigma=1.5) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi 1.5^2}} e^{-\frac{(50-50)^2}{2\times 1.5^2}} \approx \frac{1}{\sqrt{14.1}} e^{-\frac{0^2}{4.5}} \approx 0.27$$

Remember, the output from the equation, the y-axis coordinate, is a **likelihood**, **not a probability**.

Basic concepts in Statistics

Calculating probabilities

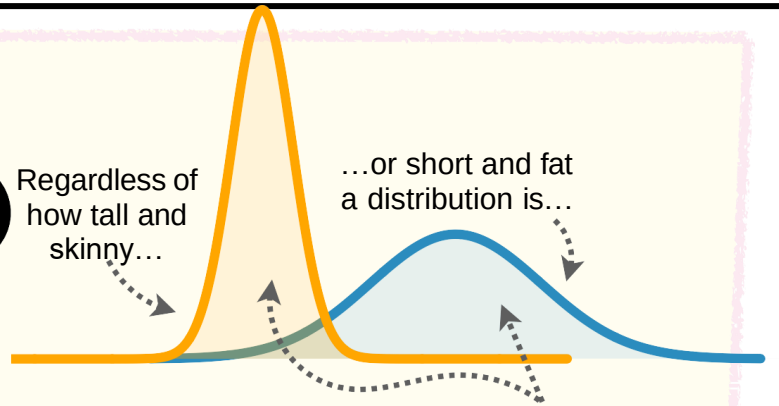
1 For **Continuous Probability Distributions**, probabilities are the **area under the curve** between two points.



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…

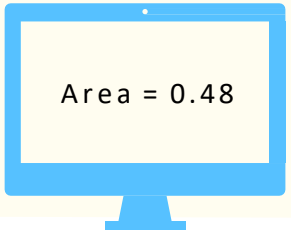
…the total area under its curve is **1**. Meaning, the probability of measuring anything in the range of possible values is **1**.

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

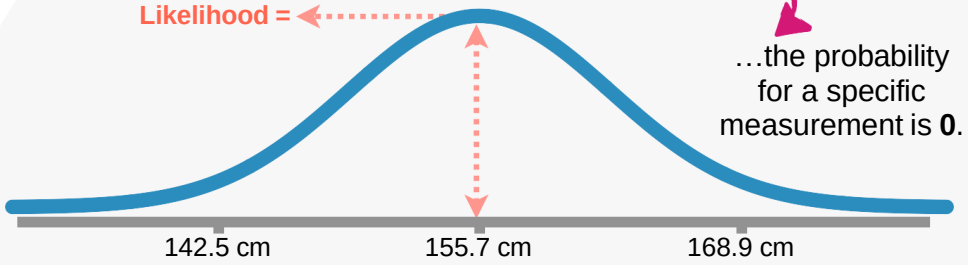
1) By using calculus and integrating the equation between the two points ***a*** and ***b***.

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

2) By using a computer..



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**…

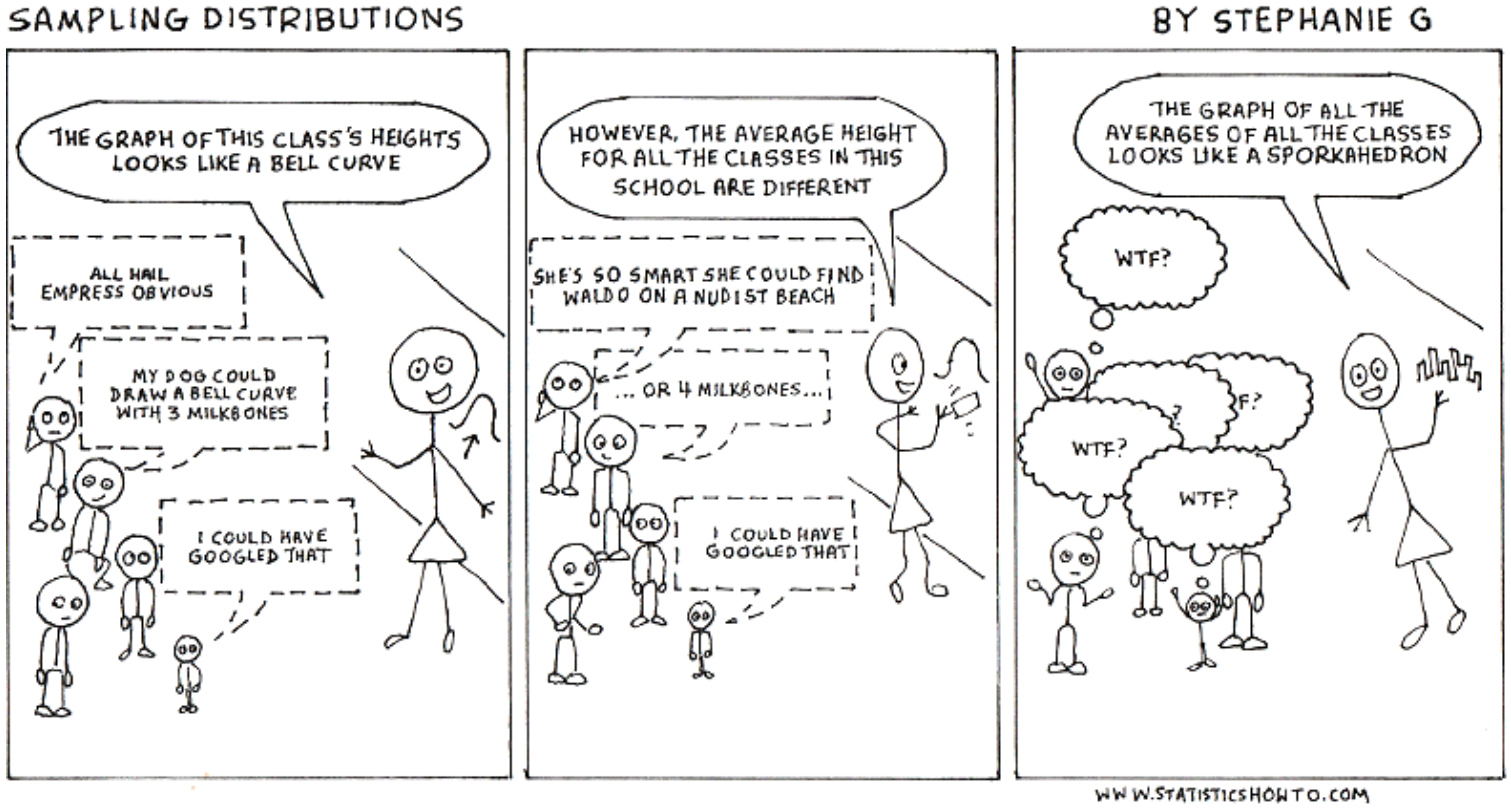


…the probability for a specific measurement is **0**.

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 155.70000000000000000000000000…tall.

Sampling Distribution



Sampling Distribution

Theorem 1: **Sampling distribution of mean and variance**

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $\frac{\sigma^2}{n} = V(\bar{X})$.

Theorem 1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

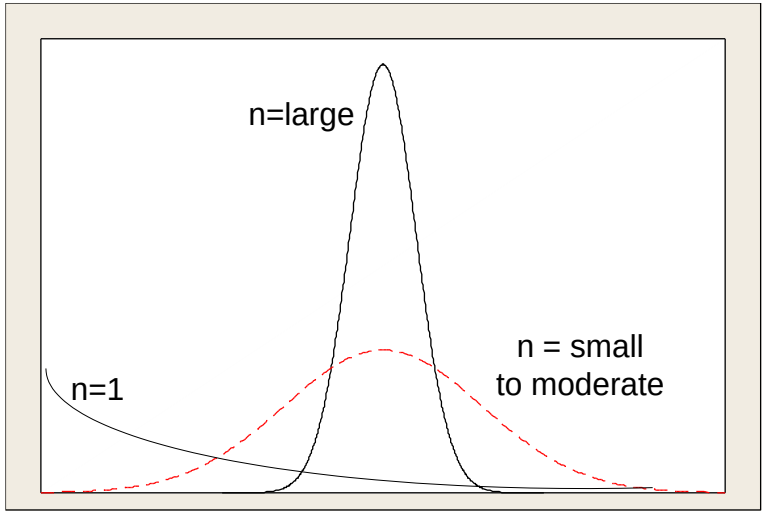
This further, can be established with the famous "central limit theorem", which is stated below.

Central Limit Theorem

Theorem 2: **Central Limit Theorem**

If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$; ***i.e., $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$.*** The approximation becomes better as n increases.

- The normal approximation of $\bar{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of $\bar{X}$ becomes more accurate as n grows larger.

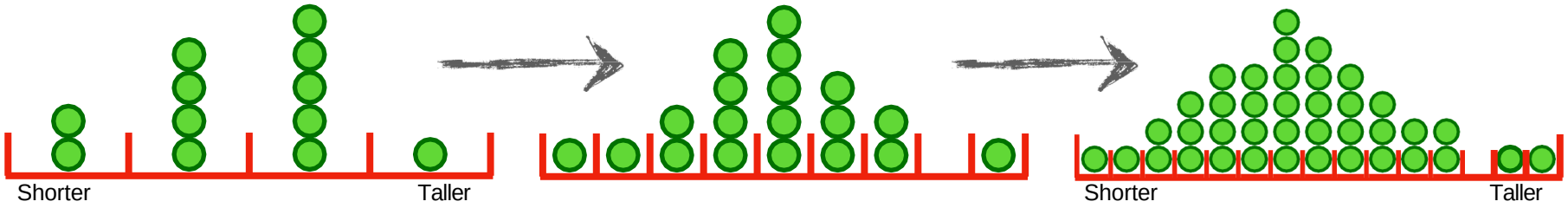


Models

The Problem:

Although we could spend a lot of time and money to build a precise histogram…

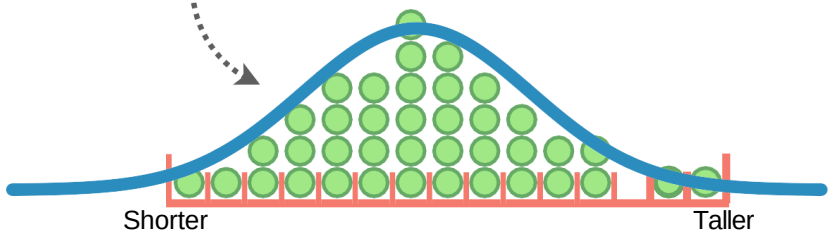
…collecting all of the data in the world is usually impossible.



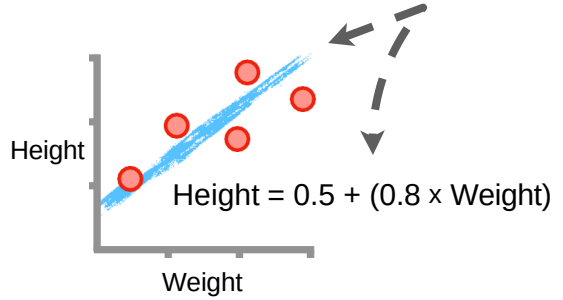
A Solution:

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.

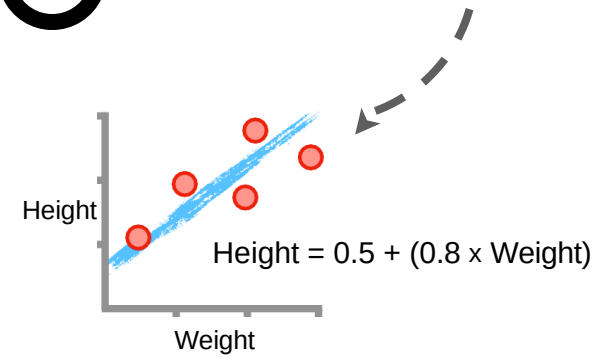


Another commonly used model is the equation for a straight line. Here, we're using a **blue line** to model a relationship between Weight and Height.

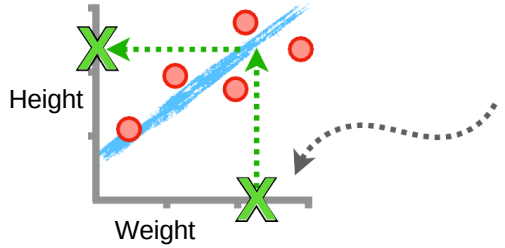


Models

1 Models need data.



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

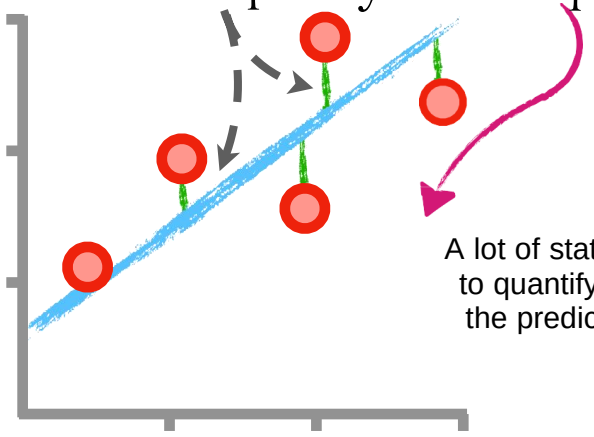


For example, if we want to know how tall someone is who weighs this much…

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

Height = 0.5 + (0.8 x Weight) = 0.5 + (0.8 x 2.1) = **2.18**

3 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.

4 In summary:

(1) Model approximates reality to let us explore relationships and make predictions.

(2) Data is needed to *train* the models

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

Models

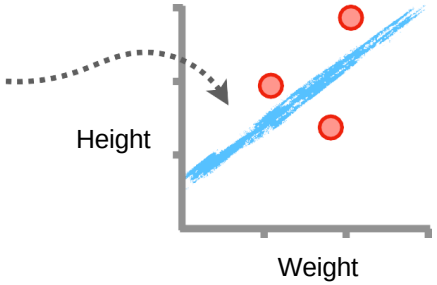
Now let's talk about how statistics can quantify the quality of a model.

The first step is to learn about the **Sum of the Squared Residuals.**

The sum of the squared residuals

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.



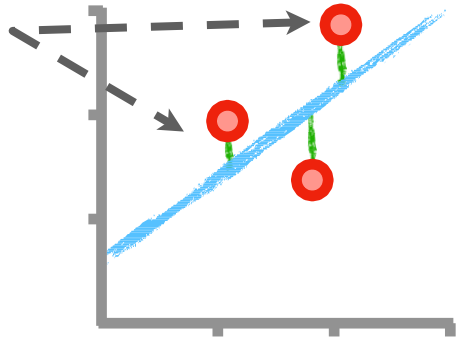
A Solution:

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**).

The Sum of Squared Residuals (SSR)

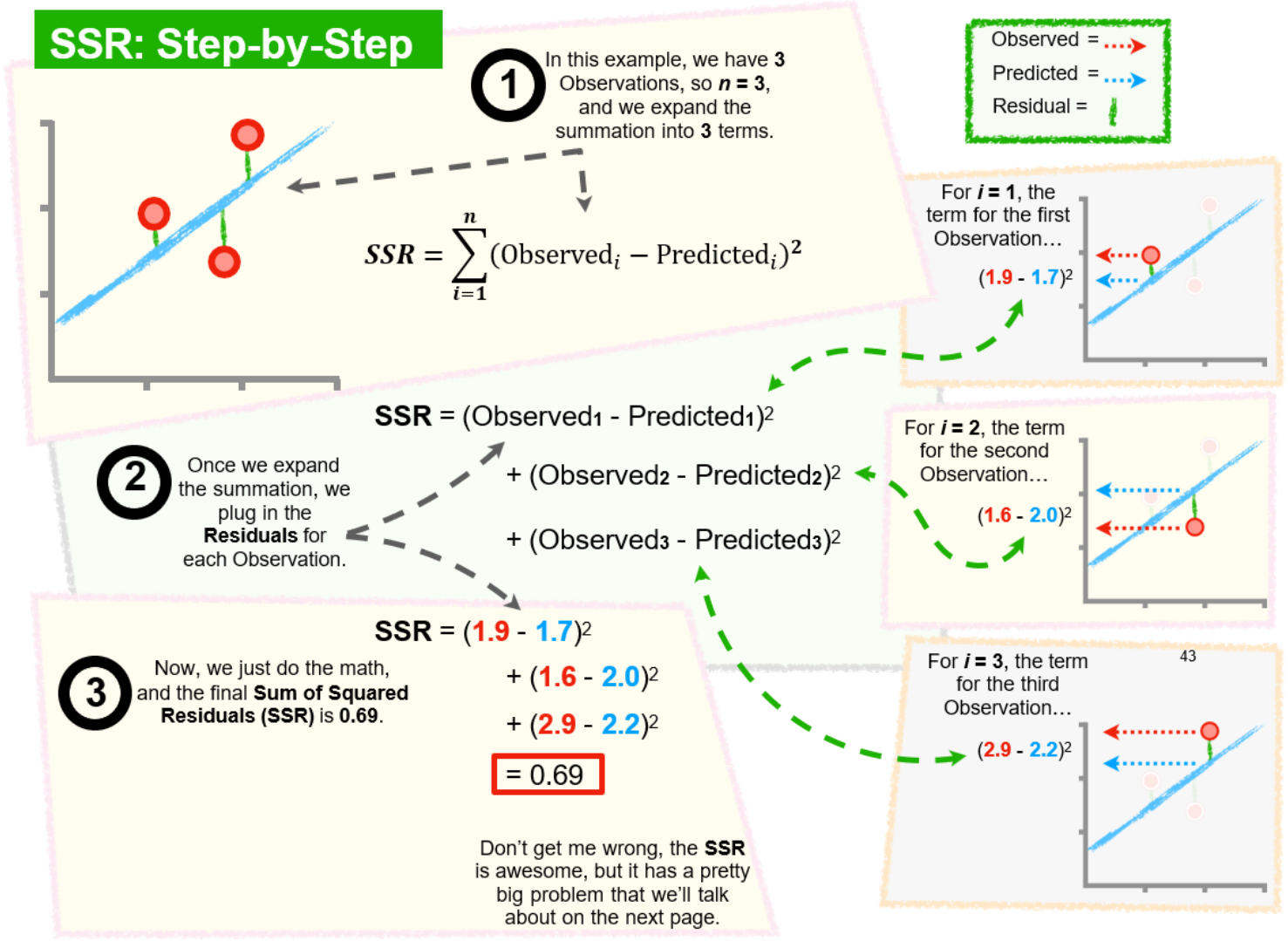
$$SSR = \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2$$

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

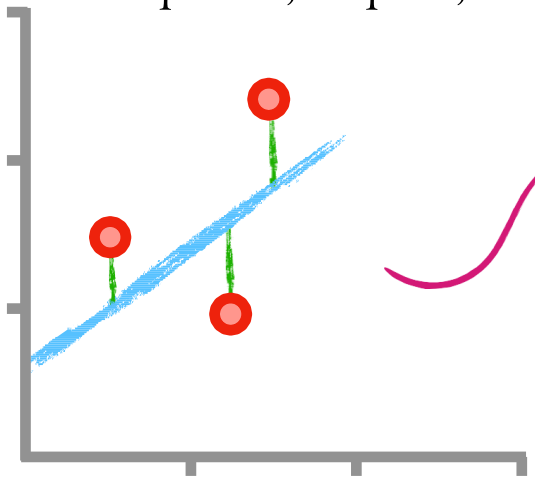


Mean Squared Error (MSE)

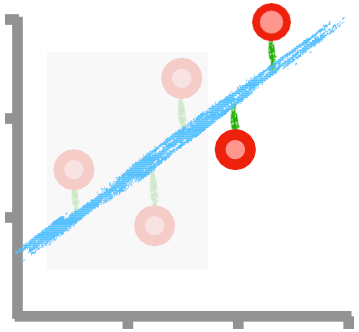
The Problem:

Sum of the Squared Residuals (SSR), although useful, 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**.

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**.

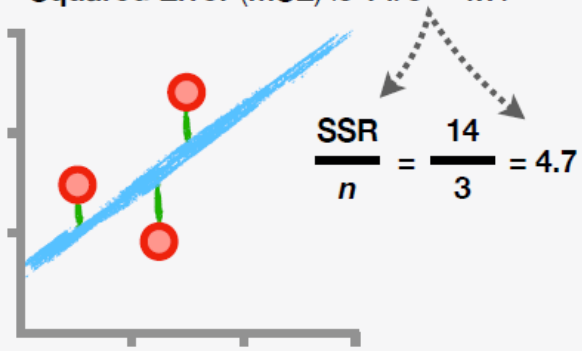
$$\text{Mean Squared Error (MSE)} = \frac{\text{The Sum of Squared Residuals (SSR)}}{\text{Number of Observations, } n} = \sum_{i=1}^{n} \frac{(\text{Observed}_i - \text{Predicted}_i)^2}{n}$$

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

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

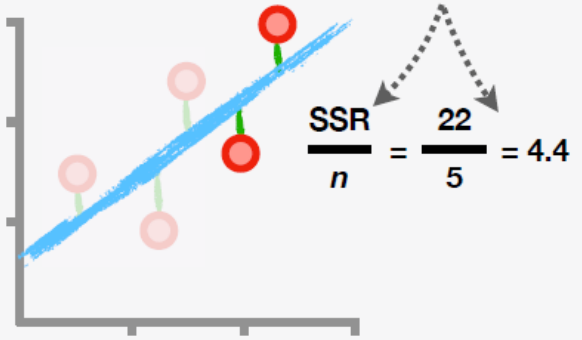
$$\text{Mean Squared Error (MSE)} = \frac{\text{SSR}}{n} = \sum_{i=1}^{n} \frac{(\text{Observed}_i - \text{Predicted}_i)^2}{n}$$

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



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

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

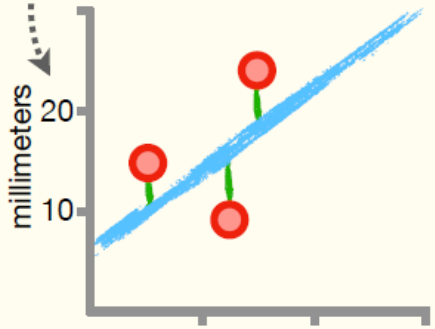


$$\frac{\text{SSR}}{n} = \frac{22}{5} = 4.4$$

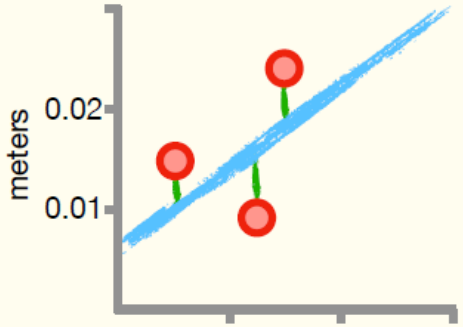
So, unlike the **SSR**, which increases when we add more data to the model, the **MSE** can increase or decrease depending on the average residual, which gives us a better sense of how the model is performing overall.

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!

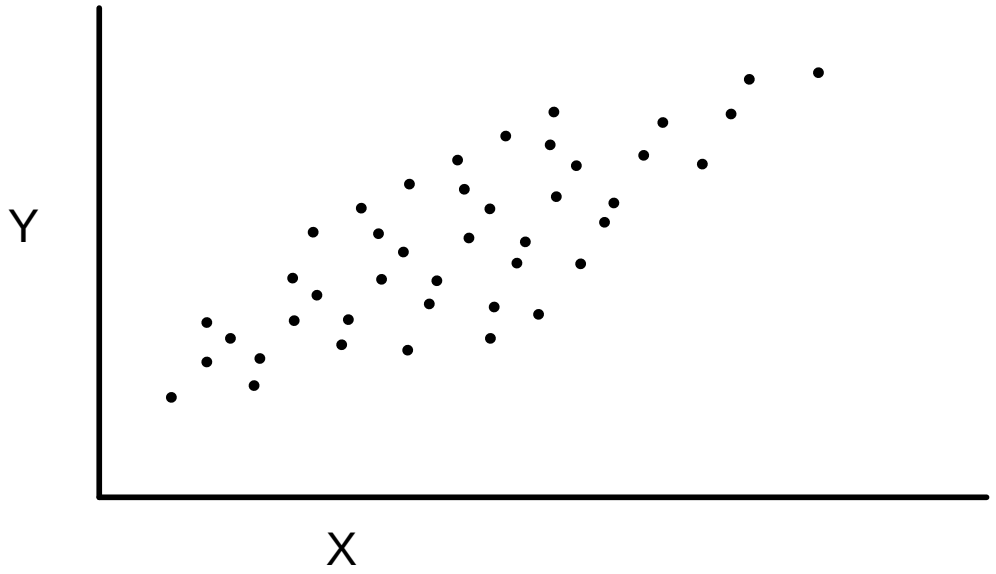


The good news, however, is that both the **SSR** and the **MSE** can be used to calculate something called R^2, which is independent of both the size of the dataset and the scale, so keep reading!

Content

- Basic data preprocessing and descriptive statistics
- Fitting a line to data
- Gradient Descent

Question

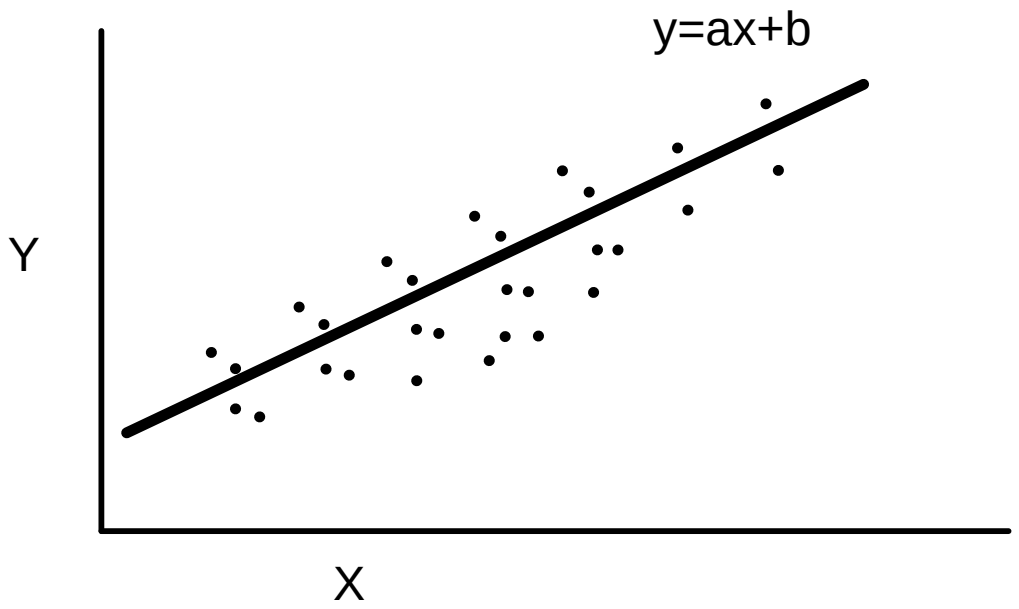


Suppose you have a very large amount (billions!) of data points (e.g.: X:Weights and Y: Heights).

We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Solution



Just decide the values of **a** and **b**

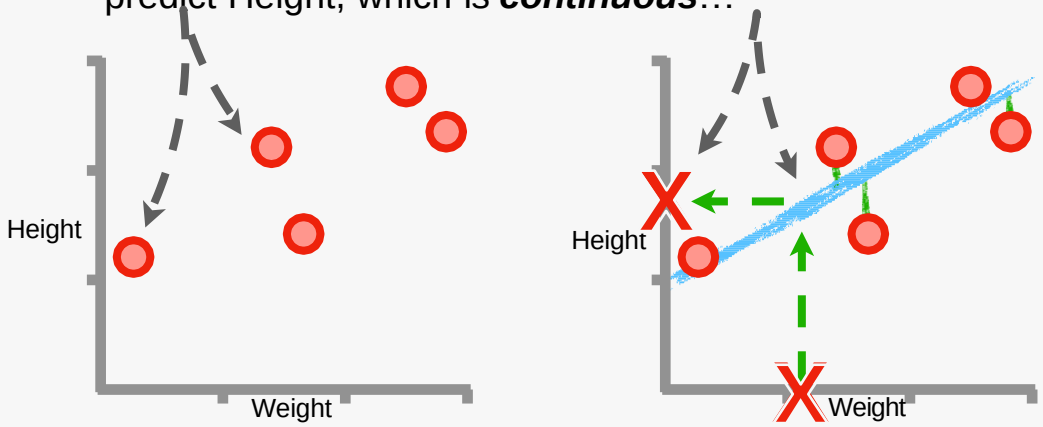
(as if storing one point's data only!)

Note: Here, the trick was to find a relationship among all the points.

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***…

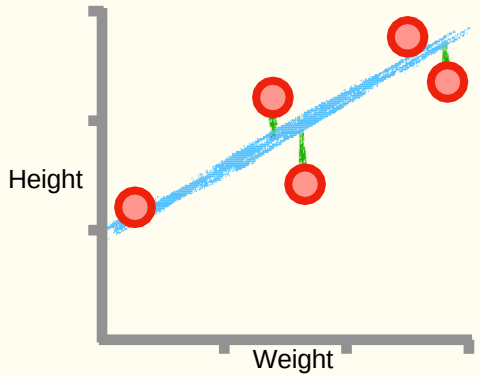


…and, we learned that we could fit a **line** to the data and use it to make predictions.

However, we haven't talked about **how we fit** a **line** to the data…

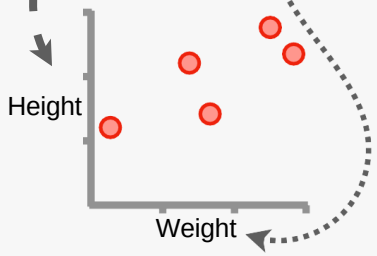
2

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



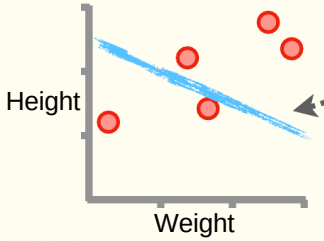
Fitting a Line to Data: Main Ideas

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

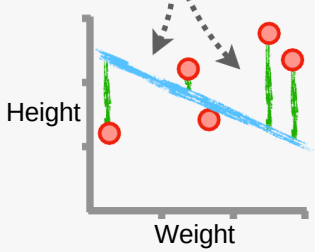


…and we wanted to predict Height from Weight.

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

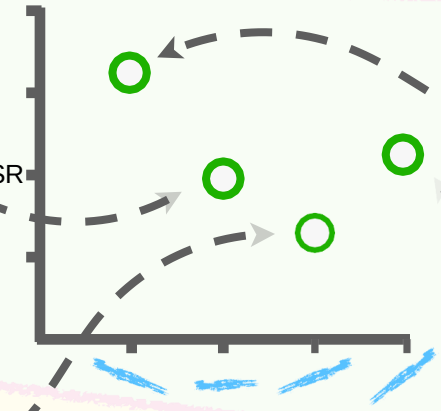


3 We can quantify how bad these predictions are by calculating the **Residuals**, which are the differences between the Observed and Predicted heights…

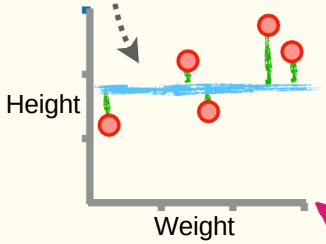


…and using the **Residuals** to calculate the **Sum of the Squared Residuals** (**SSR**).

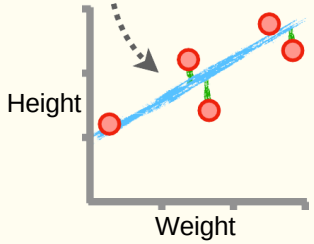
Then we can plot the **SSR** on this graph that has the **SSR** on the y-axis, and different lines fit to the data on the x-axis.



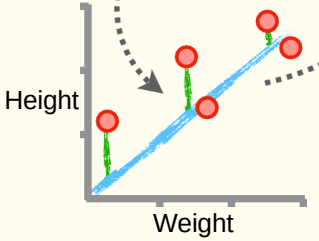
4 This **line**, which has a different y-axis intercept and slope, gives us slightly smaller residuals and a smaller **SSR**…



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



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



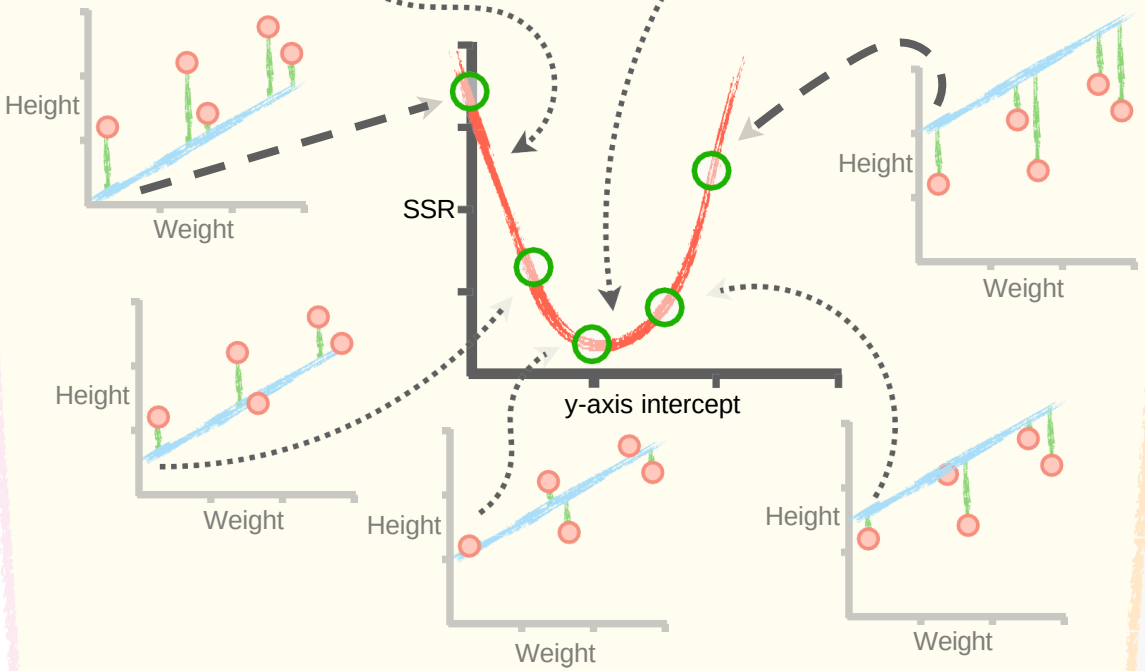
5

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**.

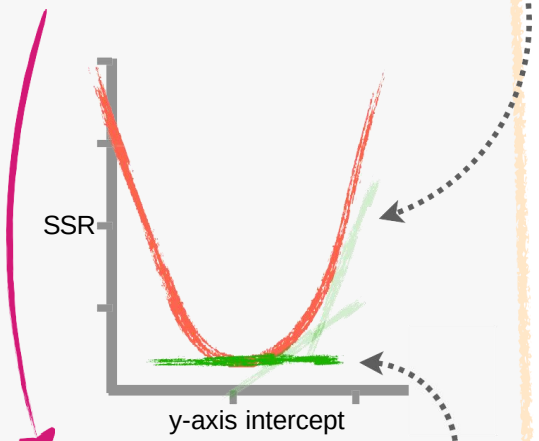
Fitting a Line to Data: Intuition

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**



…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.

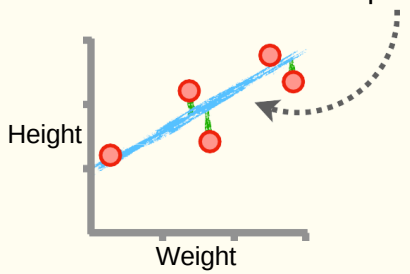
Because **Gradient Descent** is so important, we'll cover it.

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**

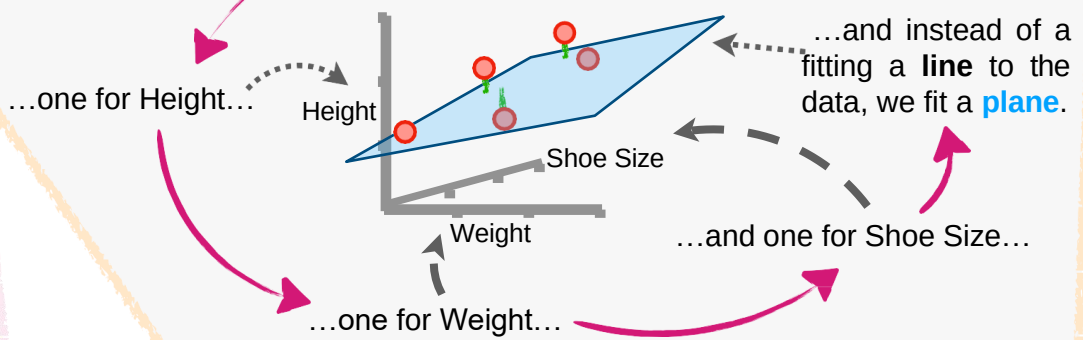
…and, as we've seen, **Simple Linear Regression** fits a line to the data that we can use to make predictions.



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,** 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.

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.

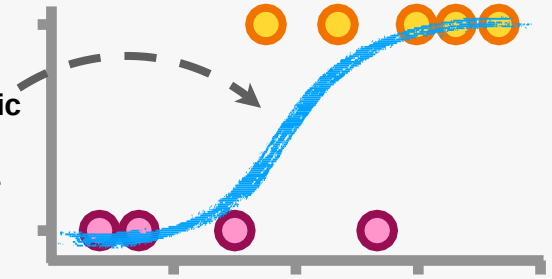
Content

- Basic data preprocessing and descriptive statistics
- Fitting a line to data
- Gradient Descent

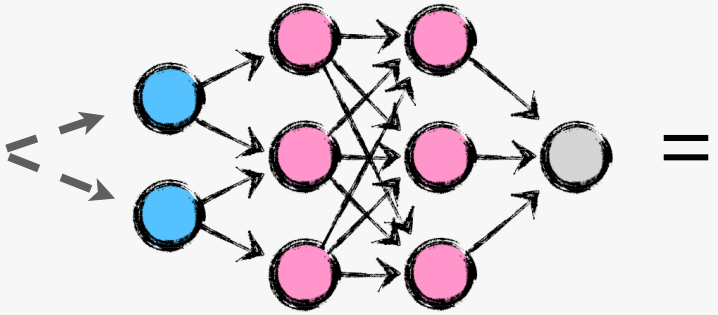
Gradient Descent: Main Ideas

1 **The Problem:** A major part of machine learning is optimizing a model's fit to the data. Sometimes this can be done with an analytical solution, but it's not always possible.

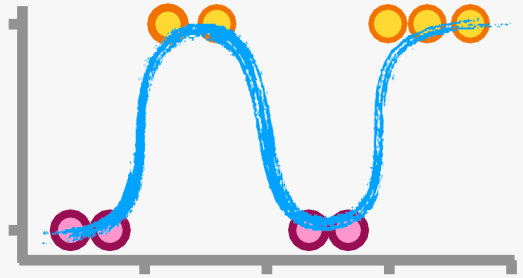
For example, there is no analytical solution for **Logistic Regression,** which fits an **s-shaped squiggle** to data.



Likewise, there is no analytical solution for **Neural Networks**, which fit **fancy squiggles** to data.



=



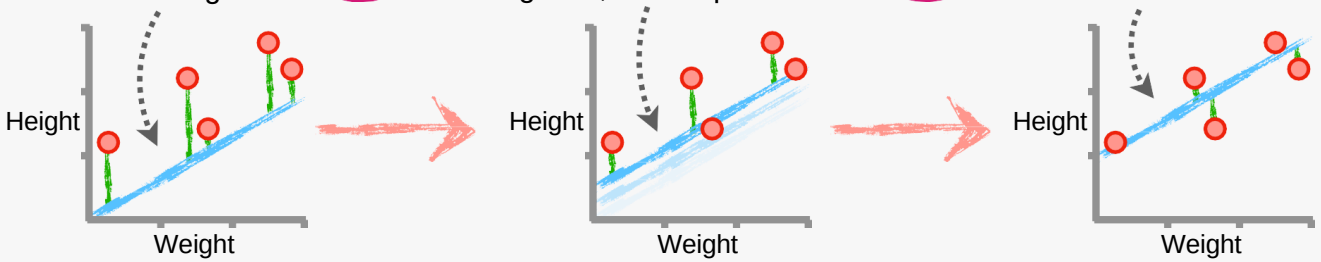
2 **A Solution:** When there's no analytical solution, **Gradient Descent** can save the day!

Gradient Descent is an ***iterative solution*** that incrementally steps toward an optimal solution and is used in a very wide variety of situations.

3 **Gradient Descent** starts with an initial guess…

…and then improves the guess, one step at a time…

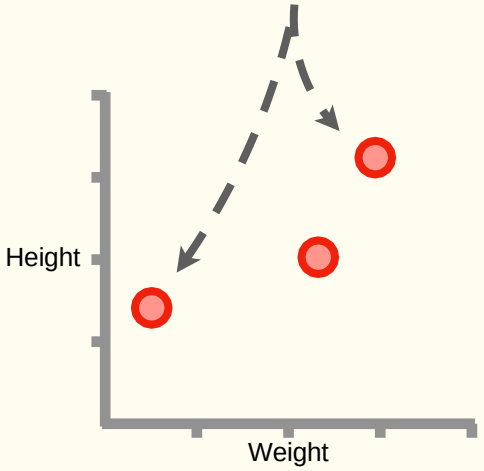
…until it finds an optimal solution or reaches a maximum number of steps.



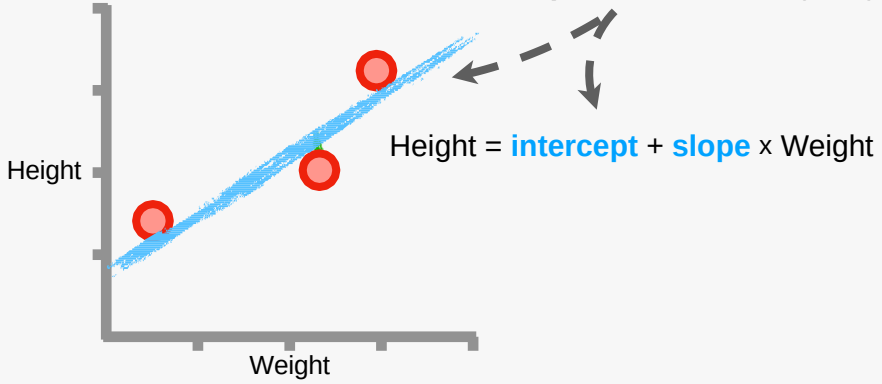
Gradient Descent: Details Part 1

NOTE: Even though there's an analytical solution for **Linear Regression**, we're using it to demonstrate how **Gradient Descent** works because we can compare the output from **Gradient Descent** to the known optimal values.

1 Let's show how **Gradient Descent** fits a line to these Height and Weight measurements.



2 Specifically, we'll show how **Gradient Descent** estimates the **intercept** and the **slope** of this line so that we minimize the **Sum of the Squared Residuals** (**SSR**).

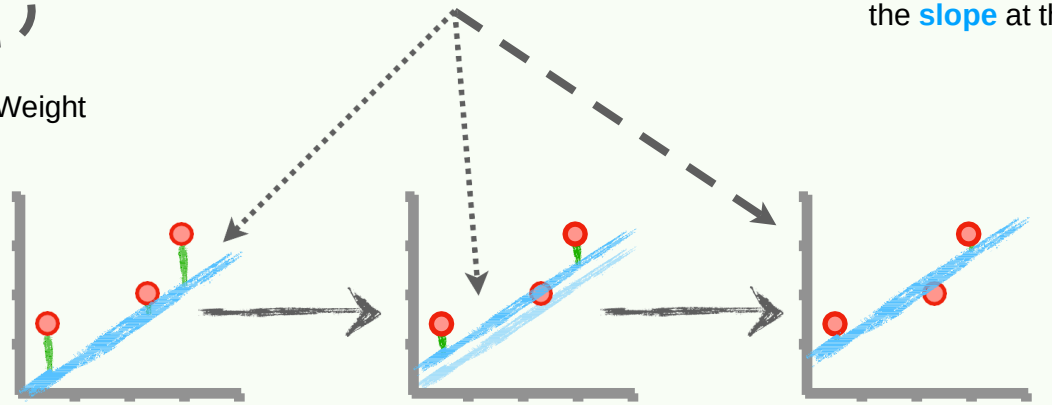


3 To keep things simple at the start, let's plug in the analytical solution for the **slope**, **0.64**…

Height = **intercept** + **0.64** x Weight

…and show how **Gradient Descent** optimizes the **intercept** one step at a time.

Once we understand how **Gradient Descent** optimizes the **intercept**, we'll show how it optimizes the **intercept** *and* the **slope** at the same time.

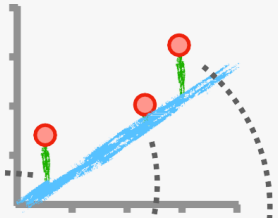


Gradient Descent: Details Part 2

4 In this example, we're fitting a line to data, and we can evaluate how well that line fits with the **Sum of the Squared Residuals** (**SSR**).

Remember, **Residuals** are the difference between the Observed and Predicted values…

5 Now, because we have **3** data points, and thus, **3 Residuals**, the **SSR** has **3** terms.



$$SSR = (Observed\ Height_1 - (\mathbf{intercept} + 0.64 \times Weight_1))^2 + (Observed\ Height_2 - (\mathbf{intercept} + 0.64 \times Weight_2))^2 + (Observed\ Height_3 - (\mathbf{intercept} + 0.64 \times Weight_3))^2$$

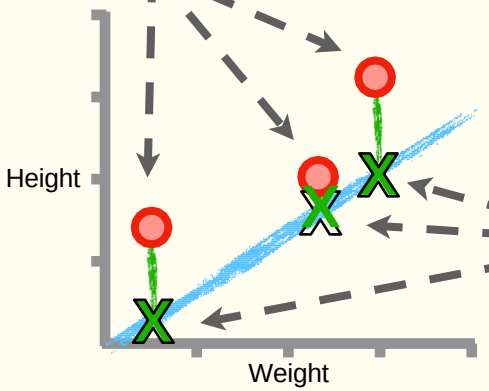
$$Residual = (Observed\ Height - Predicted\ Height) = (Height - (\mathbf{intercept} + 0.64 \times Weight))$$

…and the Observed Heights are the values we originally measured…

…and the Predicted Heights come from the equation for the line…

…so we can plug the equation for the line in for the Predicted value.

Psst! Don't forget to read **Step 5**!!!



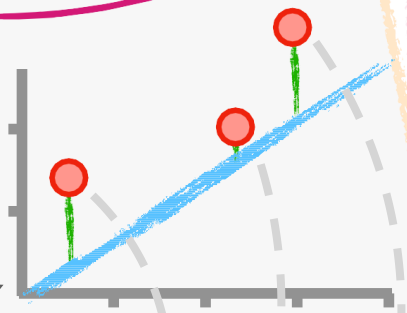
$$Predicted\ Height = \mathbf{intercept} + 0.64 \times Weight$$

Gradient Descent: Details Part 3

Gradient descent

6 In this first example, since we're only optimizing the y-axis **intercept**, we'll start by assigning it a random value. In this case, we'll initialize the **intercept** by setting it to **0**.

Height = **0** + 0.64 x Weight



7 Now, to calculate the **SSR**, we first plug the value for the y-axis **intercept**, **0**, into the equation we derived in **Steps 4** and **5**…

$$SSR = (Observed\ Height_1 - (intercept + 0.64 \times Weight_1))^2 + (Observed\ Height_2 - (intercept + 0.64 \times Weight_2))^2 + (Observed\ Height_3 - (intercept + 0.64 \times Weight_3))^2$$

$$SSR = (Observed\ Height_1 - (0 + 0.64 \times Weight_1))^2 + (Observed\ Height_2 - (0 + 0.64 \times Weight_2))^2 + (Observed\ Height_3 - (0 + 0.64 \times Weight_3))^2$$

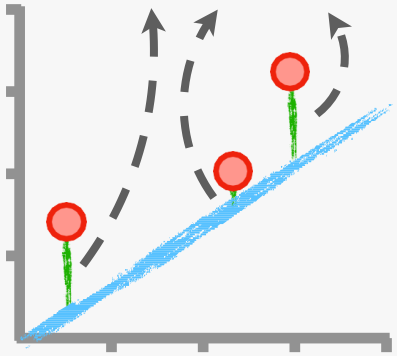
8 …then we plug in the **Observed** values for Height and Weight for each data point.

$$SSR = (Observed\ Height_1 - (0 + 0.64 \times Weight_1))^2 + (Observed\ Height_2 - (0 + 0.64 \times Weight_2))^2 + (Observed\ Height_3 - (0 + 0.64 \times Weight_3))^2$$

$$SSR = (1.4 - (0 + 0.64 \times 0.5))^2 + (1.9 - (0 + 0.64 \times 2.3))^2 + (3.2 - (0 + 0.64 \times 2.9))^2$$

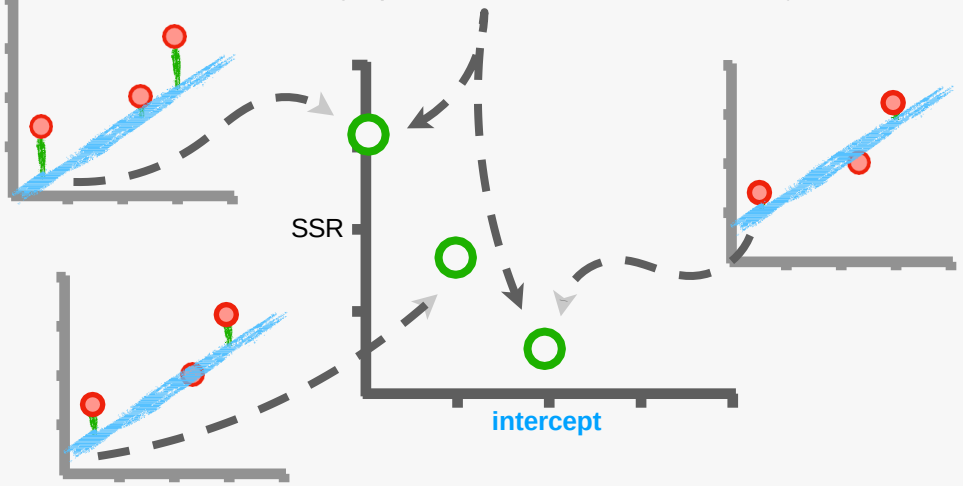
9 Lastly, we just do the math. The **SSR** for when the y-axis **intercept** is set to **0** is **3.1**. Bam!

$$1.1^2 + 0.4^2 + 1.3^2 = 3.1$$



Gradient Descent: Details Part 4

10 Now, because the goal is to minimize the **SSR**, it's a type of **Loss** or **Cost Function** (see **Terminology Alert** on the right). In **Gradient Descent**, we minimize the **Loss** or **Cost Function** by taking steps away from the initial guess toward the optimal value. In this case, we see that as we *increase* the **intercept**, the x-axis of the central graph, we *decrease* the **SSR**, the y-axis.



TERMINOLOGY ALERT!!!

The terms **Loss Function** and **Cost Function** refer to anything we want to optimize when we fit a model to data. For example, we want to optimize the **SSR** or the **Mean Squared Error** (**MSE**) when we fit a straight line with **Regression** or a squiggly line (in **Neural Networks**). That said, some people use the term **Loss Function** to specifically refer to a function (like the **SSR**) applied to *only one data point*, and use the term **Cost Function** to specifically refer to a function (like the **SSR**) applied to *all* of the data. Unfortunately, these specific meanings are not universal, so be aware of the context and be prepared to be flexible. In this book, we'll use them together and interchangeably, as in "The **Loss** or **Cost Function** is the **SSR**."

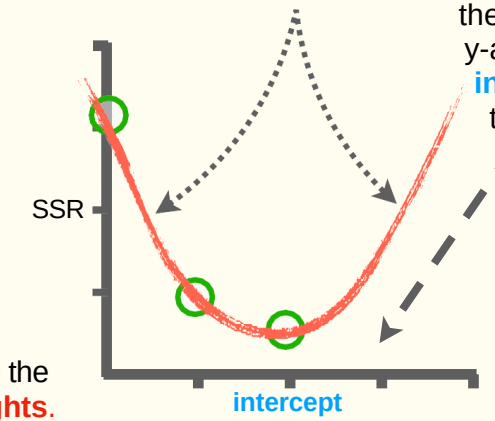
11 However, rather than just randomly trying a bunch of values for the y-axis **intercept** and plotting the resulting **SSR** on a graph, we can plot the **SSR** as a function of the y-axis **intercept**. In other words, this equation for the **SSR**…

…corresponds to this **curve** on a graph that has the **SSR** on the y-axis and the **intercept** on the x-axis.

$$SSR = (1.4 - (intercept + 0.64 \times 0.5))^2 + (1.9 - (intercept + 0.64 \times 2.3))^2 + (3.2 - (intercept + 0.64 \times 2.9))^2$$

Psst! Remember: these are the **Observed Heights**…

…and these are the **Observed Weights**.

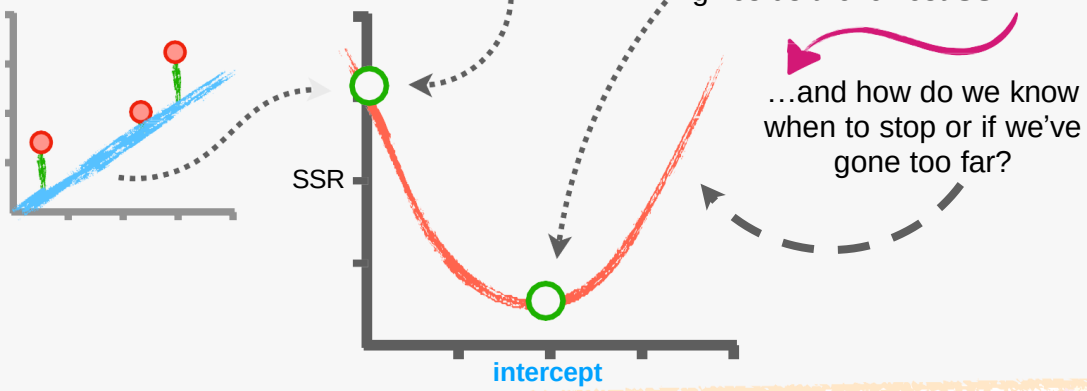


Gradient Descent: Details Part 5

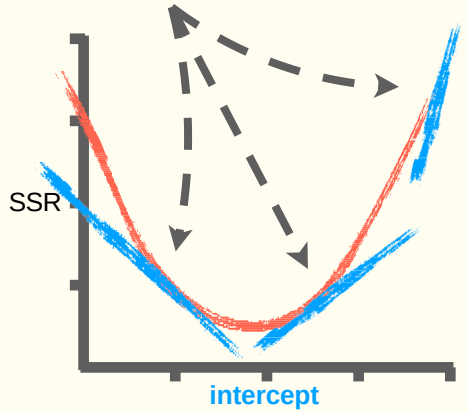
12 Now, when we started with the y-axis **intercept = 0**, we got this **SSR**…

…so how do we take steps toward this y-axis intercept that gives us the lowest **SSR**…

…and how do we know when to stop or if we've gone too far?

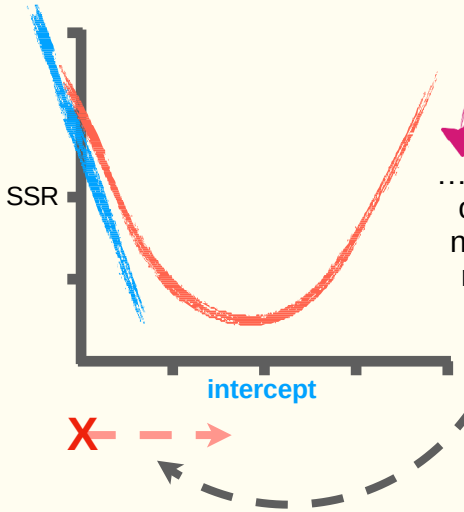


13 The answers to those questions come from the derivative of the curve, which tells us the slope of any **tangent line** that touches it.



14 A relatively large value for the derivative, which corresponds to a relatively steep slope for the **tangent line**, suggests we're relatively far from the bottom of the curve, so we should take a relatively large step…

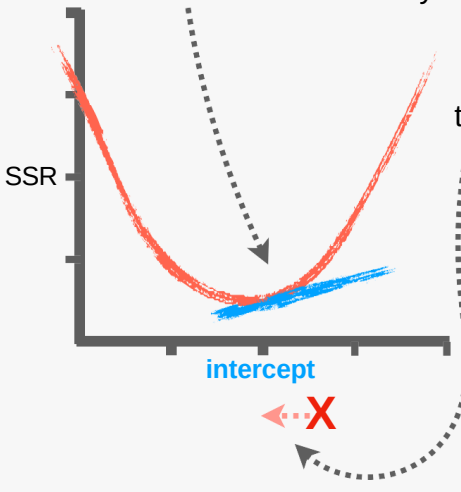
…and a *negative* derivative, or slope, tells us that we need to take a step to the right to get closer to the lowest **SSR**.



15 A relatively small value for the derivative suggests we're relatively close to the bottom of the curve, so we should take a relatively small step…

…and a *positive* derivative tells us that we need to take a step to the left to get closer to the lowest **SSR**.

In summary, the derivative tells us in which direction to take a step and how large that step should be, so let's learn how to take the derivative of the **SSR**!!!



Gradient Descent: Details Part 6

16

Because a single term of the **SSR** consists of a **Residual**…

…wrapped in parentheses and squared…

…one way to take the derivative of the **SSR** is to use **The Chain Rule**.

$$SSR = (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))^2$$

Step 1: Create a *link* between the **intercept** and the **SSR** by rewriting the **SSR** as the function of the **Residual**.

$$SSR = (\text{Residual})^2 \qquad \text{Residual} = \text{Height} - (\text{intercept} + 0.64 \times \text{Weight})$$

Step 2: Because the **Residual** links the **intercept** to the **SSR**, **The Chain Rule** tells us that the derivative of the **SSR** with respect to the **intercept** is…

$$\frac{d\,SSR}{d\,intercept} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,intercept}$$

Because of the subtraction, we can remove the parentheses by multiplying everything inside by **-1**.

Step 3: Use **The Power Rule** to solve for the two derivatives.

$$\frac{d\,Residual}{d\,intercept} = \frac{d}{d\,intercept}(\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))$$

$$= \frac{d}{d\,intercept}(\text{Height} - \text{intercept} - 0.64 \times \text{Weight})$$

$$= 0 - 1 - 0 = -1$$

Because the first and last terms do not include the **intercept**, their derivatives, with respect to the **intercept**, are both **0**. However, the second term is the negative **intercept**, so its derivative is **-1**.

$$\frac{d\,SSR}{d\,Residual} = \frac{d}{d\,Residual}(\text{Residual})^2 = 2 \times \text{Residual}$$

Step 4: Plug the derivatives into **The Chain Rule** to get the final derivative of the **SSR** with respect to the **intercept**.

$$\frac{d\,SSR}{d\,intercept} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,intercept} = 2 \times \text{Residual} \times -1$$

$$= 2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})) \times -1$$

$$= -2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))$$

Multiply this **-1** on the right by the **2** on the left to get **-2**.

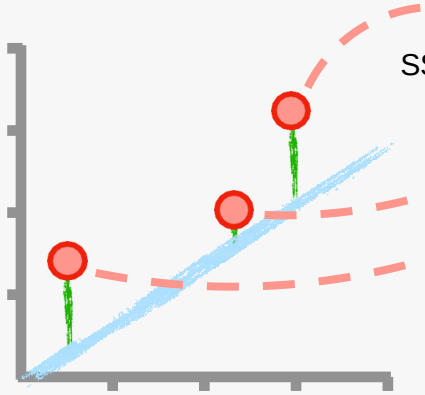
Gradient Descent: Details Part 7

17 So far, we've calculated the derivative of the **SSR** for a single observation.

However, we have three observations in the dataset, so the **SSR** and its derivative both have three terms.

$$SSR = (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))^2$$

$$\frac{d\,SSR}{d\,intercept} = -2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))$$

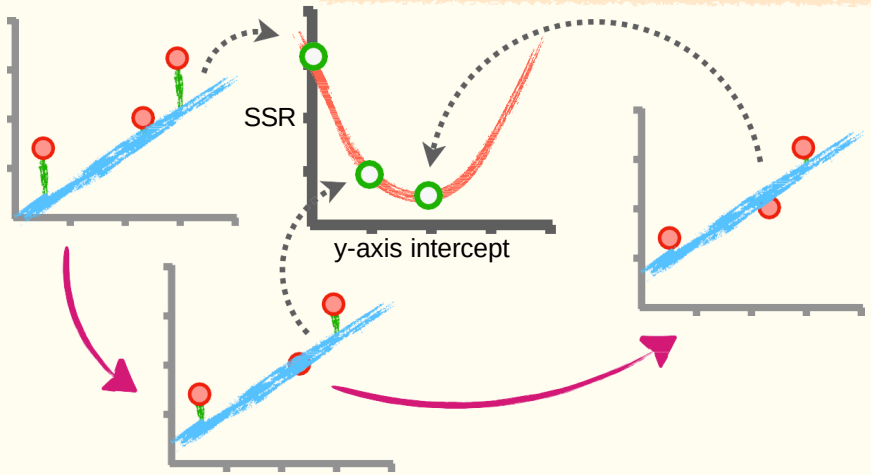


$$SSR = (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))^2 + (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))^2 + (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))^2$$

$$\frac{d\,SSR}{d\,intercept} = -2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})) + -2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})) + -2 \times (\text{Height} - (\text{intercept} + 0.64 \times \text{Weight}))$$

Gentle Reminder: Because we're using **Linear Regression** as our example, we don't actually ***need*** to use **Gradient Descent** to find the optimal value for the intercept. Instead, we could just set the derivative equal to **0** and solve for the **intercept**. This would be an analytical solution. However, by applying **Gradient Descent** to this problem, we can compare the optimal value that it gives us to the analytical solution and evaluate how well **Gradient Descent** performs. This will give us more confidence in **Gradient Descent** when we use it in situations without analytical solutions like **Logistic Regression** and **Neural Networks**.

18 Now that we have the derivative of the **SSR** for all **3** data points, we can go through, step-by-step, how **Gradient Descent** uses this derivative to find the **intercept** value that minimizes the **SSR**. However, before we get started, it's time for the dreaded **Terminology Alert!!!**

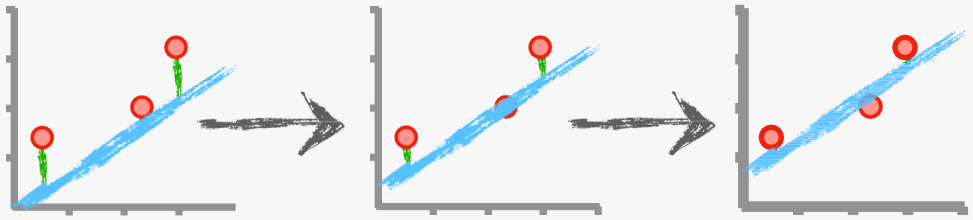


Terminology Alert!!! Parameters

1. In the current example, we're trying to optimize the y-axis **intercept**.

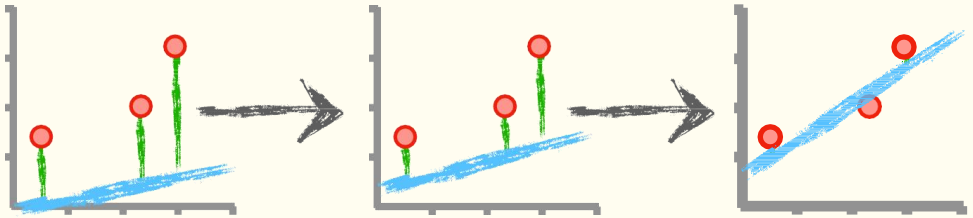
In machine learning lingo, we call the things we want to optimize ***parameters***. So, in this case, we would call the y-axis **intercept** a **parameter**.

Predicted Height = **intercept** + 0.64 x Weight



2. If we wanted to optimize both the y-axis **intercept** and the **slope**, then we would need to optimize two **parameters**.

Predicted Height = **intercept** + **slope** x Weight

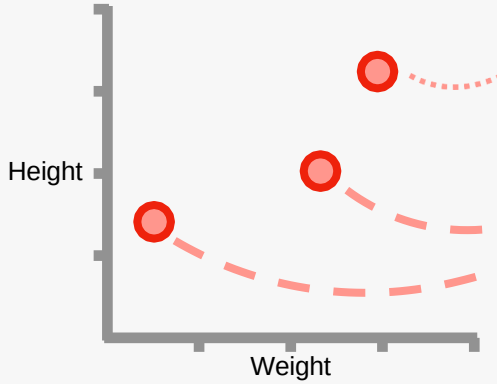


3. Now that we know what we mean when we say **parameter**, let's see how **Gradient Descent** optimizes a single **parameter**, the **intercept**, one step at a time!!!

Gradient Descent for One Parameter: Step-by-Step

1 First, plug the **Observed** values into the derivative of the **Loss** or **Cost Function**. In this example, the **SSR** is the **Loss** or **Cost Function**…

…so that means plugging the **Observed Weight** and **Height** measurements into the derivative of the **SSR**.



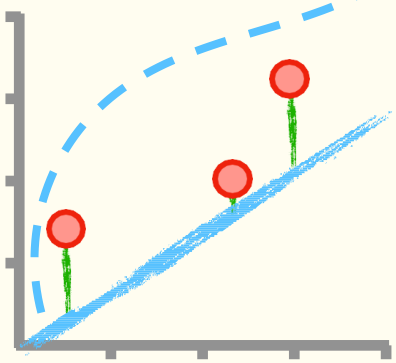
$$\frac{d\,SSR}{d\,intercept} = -2 \times (\,\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})\,)$$
$$+ -2 \times (\,\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})\,)$$
$$+ -2 \times (\,\text{Height} - (\text{intercept} + 0.64 \times \text{Weight})\,)$$

$$\frac{d\,SSR}{d\,intercept} = -2 \times (\,3.2 - (\text{intercept} + 0.64 \times 2.9)\,)$$
$$+ -2 \times (\,1.9 - (\text{intercept} + 0.64 \times 2.3)\,)$$
$$+ -2 \times (\,1.4 - (\text{intercept} + 0.64 \times 0.5)\,)$$

2 Now we initialize the parameter we want to optimize with a random value. In this example, where we just want to optimize the y-axis **intercept**, we start by setting it to **0**.

Height = intercept + 0.64 x Weight

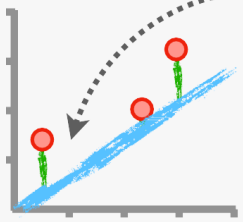
= 0 + 0.64 x Weight



$$\frac{d\,SSR}{d\,intercept} = -2 \times (\,3.2 - (0 + 0.64 \times 2.9)\,)$$
$$+ -2 \times (\,1.9 - (0 + 0.64 \times 2.3)\,)$$
$$+ -2 \times (\,1.4 - (0 + 0.64 \times 0.5)\,)$$

Gradient Descent for One Parameter: Step-by-Step

3 Now evaluate the derivative at the current value for the **intercept**. In this case, the current value is **0**.

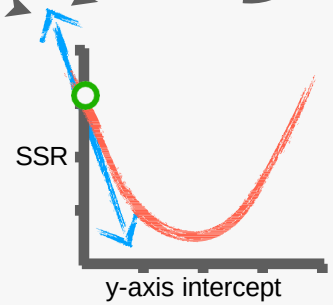


$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (0 + 0.64 \times 2.9)) + -2 \times (1.9 - (0 + 0.64 \times 2.3)) + -2 \times (1.4 - (0 + 0.64 \times 0.5))$$

When we do the math, we get **-5.7**…

$$= -5.7$$

…thus, when the **intercept = 0**, the slope of this **tangent line** is **-5.7**.



4 Now calculate the **Step Size** with the following equation:

Gentle Reminder: The magnitude of the derivative is proportional to how big of a step we should take toward the minimum. The sign (+/-) tells us which direction.

Step Size = **Derivative** x **Learning Rate**

= **-5.7** x **0.1**

= **-0.57**

NOTE: The **Learning Rate** prevents us from taking steps that are too big and skipping past the lowest point in the curve. Typically, for **Gradient Descent**, the **Learning Rate** is determined automatically: it starts relatively large and gets smaller with every step taken. However, you can also use **Cross Validation** to determine a good value for the **Learning Rate**. In this case, we're setting the **Learning Rate** to **0.1**.

5 Take a step from the **current intercept** to get closer to the optimal value with the following equation:

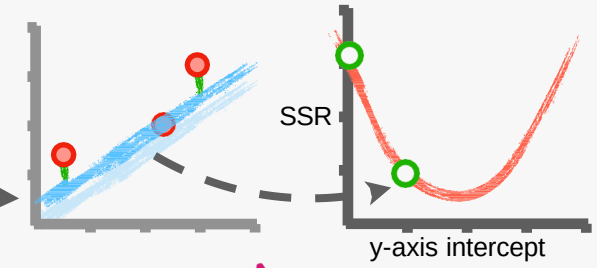
New intercept = **Current intercept** - **Step Size**

Remember, in this case, the **current intercept** is **0**.

= **0** - (**-0.57**)

= **0.57**

The **new intercept**, **0.57**, moves the line up a little closer to the data…



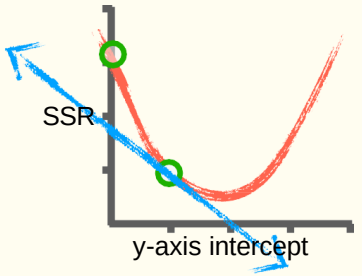
…and it results in a lower **SSR**. Bam!

Gradient Descent for One Parameter: Step-by-Step

6 Now repeat the previous three steps, updating the **intercept** after each iteration until the **Step Size** is close to **0** or we take the maximum number of steps, which is often set to **1,000** iterations.

a Evaluate the derivative at the current value for the intercept…

$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (0.57 + 0.64 \times 2.9)) + -2 \times (1.9 - (0.57 + 0.64 \times 2.3)) + -2 \times (1.4 - (0.57 + 0.64 \times 0.5)) = -2.3$$



b Calculate the **Step Size**…

Step Size = **Derivative** x **Learning Rate**

= **-2.3** x **0.1**

= **-0.23**

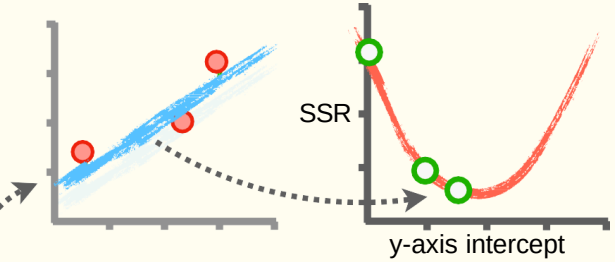
NOTE: The **Step Size** is smaller than before because the slope of the **tangent line** is not as steep as before. The smaller slope means we're getting closer to the optimal value.

c Calculate the **new intercept** value…

New intercept = **Current intercept** - **Step Size**

= **0.57** - (**-0.23**)

= **0.8**



The **new intercept**, **0.8**, moves the line up a little closer to the data…

…and it results in a lower **SSR**

Gradient Descent for One Parameter: Step-by-Step

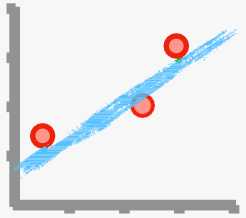
7 After **7** iterations…

a Evaluate the derivative at the current value…

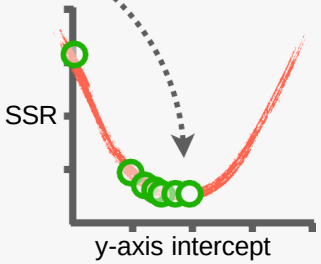
b Calculate the **Step Size**…

c Calculate the new value…

…the **Step Size** was very close to **0,** so we stopped with the **current intercept** = **0.95**…



…and we made it to the lowest **SSR**.

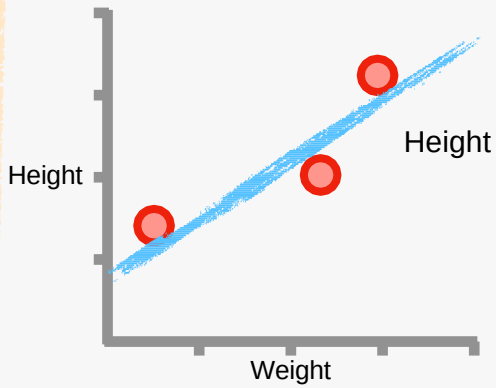


8 If, earlier on, instead of using **Gradient Descent**, we simply set the derivative to **0** and solved for the **intercept**, we would have gotten **0.95**, which is the same value that **Gradient Descent** gave us. Thus, **Gradient Descent** did a decent job.

Now let's see how well **Gradient Descent** optimizes the **intercept** *and* the **slope**!

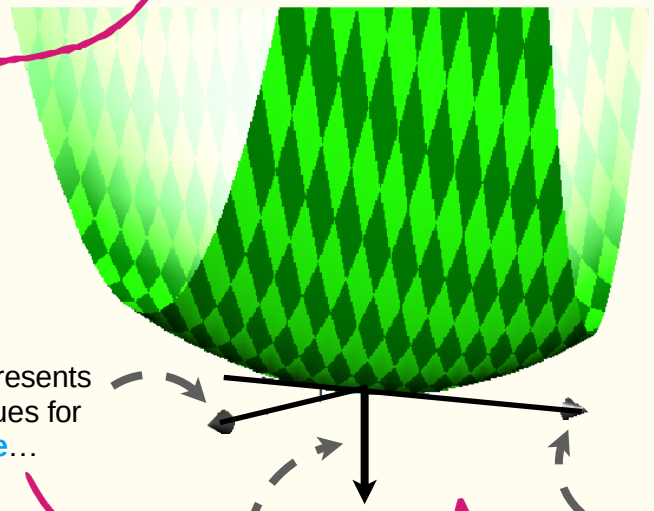
Optimizing Two or More Parameters: Details

1 Now that we know how to optimize the **intercept** of the line that minimizes the **SSR**, let's optimize both the **intercept** and the **slope**.



Height = **intercept** + **slope** x Weight

2 When we optimize two parameters, we get a 3-dimensional graph of the **SSR**.

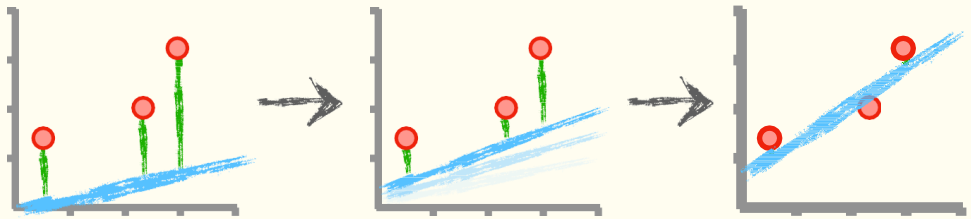


This axis represents different values for the **slope**…

…the vertical axis is for the **SSR**…

…and this axis represents different values for the **intercept**.

3 Just like before, the goal is to find the parameter values that give us the lowest **SSR**. And just like before, **Gradient Descent** initializes the parameters with random values and then uses derivatives to update those parameters, one step at a time, until they're optimal.



4 So, now let's learn how to take derivatives of the **SSR** with respect to both the **intercept** and the **slope**.

$$SSR = (\text{Height} - (\textbf{intercept} + \textbf{slope} \times \text{Weight}))^2$$

Taking Multiple (Partial) Derivatives of the SSR: Part 1

1. The good news is that taking the derivative of the **SSR** with respect to the **intercept** is exactly the same as before.

We can use **The Chain Rule** to tell us how the **SSR** changes with respect to the **intercept**.

$$SSR = (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))^2$$

Step 1: Create a *link* between the **intercept** and the **SSR** by rewriting the **SSR** as the function of the **Residual**.

$$SSR = (\text{Residual})^2 \qquad \text{Residual} = \text{Observed Height} - (\text{intercept} + \text{slope} \times \text{Weight})$$

Step 2: Because the **Residual** links the **intercept** to the **SSR**, **The Chain Rule** tells us that the derivative of the **SSR** with respect to the **intercept** is…

$$\frac{d\,SSR}{d\,intercept} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,intercept}$$

Because of the subtraction, we can remove the parentheses by multiplying everything inside by **-1**.

Step 3: Use **The Power Rule** to solve for the two derivatives.

$$\frac{d\,Residual}{d\,intercept} = \frac{d}{d\,intercept}(\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))$$

$$= \frac{d}{d\,intercept}(\text{Height} - \text{intercept} - \text{slope} \times \text{Weight})$$

$$= 0 - 1 - 0 = -1$$

$$\frac{d\,SSR}{d\,Residual} = \frac{d}{d\,Residual}(\text{Residual})^2 = 2 \times \text{Residual}$$

Because the first and last terms do not include the **intercept**, their derivatives, with respect to the **intercept**, are both **0**. However, the second term is the negative **intercept**, so its derivative is **-1**.

Step 4: Plug the derivatives into **The Chain Rule** to get the final derivative of the **SSR** with respect to the **intercept**.

$$\frac{d\,SSR}{d\,intercept} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,intercept} = 2 \times \text{Residual} \times -1$$

$$= 2 \times (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight})) \times -1$$

$$= -2 \times (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))$$

Multiply this **-1** on the right by the **2** on the left to get **-2**.

Taking Multiple (Partial) Derivatives of the SSR: Part 2

2 The other good news is that taking the derivative of the **SSR** with respect to the **slope** is very similar to what we just did for the **intercept**.

We can use **The Chain Rule** to tell us how the **SSR** changes with respect to the **slope**.

$$SSR = (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))^2$$

NOTE: A collection of derivatives of the same function but with respect to different parameters is called a **Gradient**, so this is where **Gradient Descent** gets its name from. We'll use the *gradient* to *descend* to the lowest **SSR**.

Step 1: Create a *link* between the **slope** and the **SSR** by rewriting the **SSR** as the function of the **Residual**.

$$SSR = (\text{Residual})^2 \qquad \text{Residual} = \text{Observed Height} - (\text{intercept} + \text{slope} \times \text{Weight})$$

Step 2: Because the **Residual** links the **slope** to the **SSR**, **The Chain Rule** tells us that the derivative of the **SSR** with respect to the **slope** is…

$$\frac{d\,SSR}{d\,slope} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,slope}$$

Because of the subtraction, we can remove the parentheses by multiplying everything inside by **-1**.

Step 3: Use **The Power Rule** to solve for the two derivatives.

$$\frac{d\,Residual}{d\,slope} = \frac{d}{d\,slope}(\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))$$

$$= \frac{d}{d\,slope}(\text{Height} - \text{intercept} - \text{slope} \times \text{Weight})$$

$$= 0 - 0 - \text{Weight} = -\text{Weight}$$

Because the first and second terms do not include the **slope**, their derivatives, with respect to the **slope**, are both **0**. However, the last term is the negative **slope** times **Weight**, so its derivative is **-Weight**.

$$\frac{d\,SSR}{d\,Residual} = \frac{d}{d\,Residual}(\text{Residual})^2 = 2 \times \text{Residual}$$

Step 4: Plug the derivatives into **The Chain Rule** to get the final derivative of the **SSR** with respect to the **slope**.

$$\frac{d\,SSR}{d\,slope} = \frac{d\,SSR}{d\,Residual} \times \frac{d\,Residual}{d\,slope} = 2 \times \text{Residual} \times -\text{Weight}$$

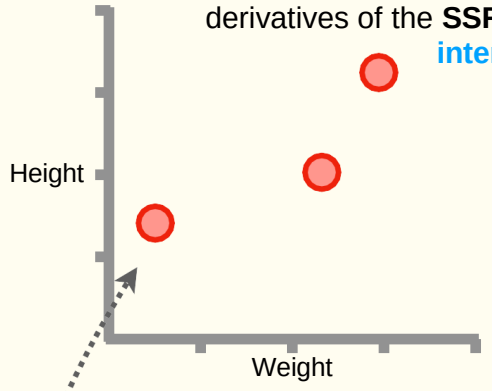
$$= 2 \times (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight})) \times -\text{Weight}$$

Multiply this **-Weight** on the right by the **2** on the left to get **-2 x Weight**.

$$= -2 \times \text{Weight} \times (\text{Height} - (\text{intercept} + \text{slope} \times \text{Weight}))$$

Gradient Descent for Two Parameters: Step-by-Step

1 Plug the **Observed** values into the derivatives of the **Loss** or **Cost Function**. In this example, the **SSR** is the **Loss** or **Cost Function**, so we'll plug the **Observed Weight** and **Height** measurements into the two derivatives of the **SSR**, one with respect to the **intercept**…



….and one with respect to the **slope**.

$$\frac{d\,SSR}{d\,intercept} = -2 \times (Height_1 - (intercept + slope \times Weight_1)) + -2 \times (Height_2 - (intercept + slope \times Weight_2)) + -2 \times (Height_3 - (intercept + slope \times Weight_3))$$

$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (intercept + slope \times 2.9)) + -2 \times (1.9 - (intercept + slope \times 2.3)) + -2 \times (1.4 - (intercept + slope \times 0.5))$$

$$\frac{d\,SSR}{d\,slope} = -2 \times Weight_1 \times (Height_1 - (intercept + slope \times Weight_1)) + -2 \times Weight_2 \times (Height_2 - (intercept + slope \times Weight_2)) + -2 \times Weight_3 \times (Height_3 - (intercept + slope \times Weight_3))$$

$$\frac{d\,SSR}{d\,slope} = -2 \times 2.9 \times (3.2 - (intercept + slope \times 2.9)) + -2 \times 2.3 \times (1.9 - (intercept + slope \times 2.3)) + -2 \times 0.5 \times (1.4 - (intercept + slope \times 0.5))$$

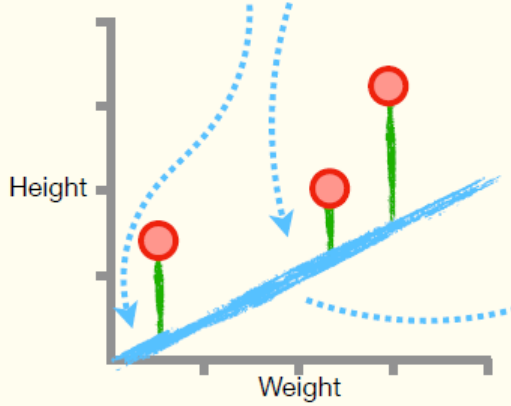
Gentle Reminder: The Weight and Height values that we're plugging into the derivatives come from the raw data in the graph.

Gradient Descent for Two Parameters: Step-by-Step

2 Now initialize the parameter, or parameters, that we want to optimize with random values. In this example, we'll set the intercept to **0** and the slope to **0.5**.

Height = intercept + slope x Weight

Height = 0 + 0.5 x Weight



$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (intercept + slope \times 2.9)) + -2 \times (1.9 - (intercept + slope \times 2.3)) + -2 \times (1.4 - (intercept + slope \times 0.5))$$

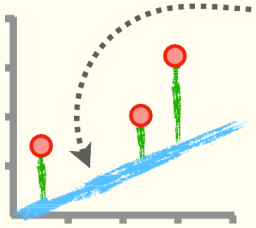
$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (0 + 0.5 \times 2.9)) + -2 \times (1.9 - (0 + 0.5 \times 2.3)) + -2 \times (1.4 - (0 + 0.5 \times 0.5))$$

$$\frac{d\,SSR}{d\,slope} = -2 \times 2.9 \times (3.2 - (intercept + slope \times 2.9)) + -2 \times 2.3 \times (1.9 - (intercept + slope \times 2.3)) + -2 \times 0.5 \times (1.4 - (intercept + slope \times 0.5))$$

$$\frac{d\,SSR}{d\,slope} = -2 \times 2.9 \times (3.2 - (0 + 0.5 \times 2.9)) + -2 \times 2.3 \times (1.9 - (0 + 0.5 \times 2.3)) + -2 \times 0.5 \times (1.4 - (0 + 0.5 \times 0.5))$$

Gradient Descent for Two Parameters: Step-by-Step

3 Evaluate the derivatives at the current values for the **intercept**, **0**, and **slope**, **0.5**.



$$\frac{d\,SSR}{d\,intercept} = -2 \times (3.2 - (0 + 0.5 \times 2.9)) + -2 \times (1.9 - (0 + 0.5 \times 2.3)) + -2 \times (1.4 - (0 + 0.5 \times 0.5)) = -7.3$$

$$\frac{d\,SSR}{d\,slope} = -2 \times 2.9 \times (3.2 - (0 + 0.5 \times 2.9)) + -2 \times 1.9 \times (2.3 - (0 + 0.5 \times 1.9)) + -2 \times 0.5 \times (1.4 - (0 + 0.5 \times 0.5)) = -14.8$$

4 Calculate the **Step Sizes**: one for the **intercept**…

…and one for the **slope**.

$$\text{Step Size}_{Intercept} = \text{Derivative} \times \text{Learning Rate}$$

$$\text{Step Size}_{Intercept} = -7.3 \times 0.01$$

$$\text{Step Size}_{Intercept} = -0.073$$

$$\text{Step Size}_{Slope} = \text{Derivative} \times \text{Learning Rate}$$

$$\text{Step Size}_{Slope} = -14.8 \times 0.01$$

$$\text{Step Size}_{Slope} = -0.148$$

NOTE: We're using a smaller **Learning Rate** now (**0.01**) than before (**0.1**) because **Gradient Descent** can be very sensitive to it. However, as we said earlier, usually the **Learning Rate** is determined automatically.

5 Take a step from the current **intercept**, **0**, and **slope**, **0.5**, to get closer to the optimal values…

$$\text{New intercept} = \text{Current intercept} - \text{Step Size}_{Intercept}$$

$$= 0 - (-0.073)$$

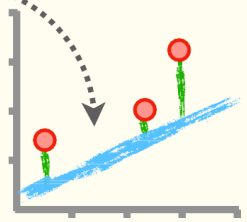
$$= 0.073$$

…and the **intercept** increases from **0** to **0.073**, the **slope** increases from **0.5** to **0.648**, and the **SSR** decreases. **BAM**!

$$\text{New slope} = \text{Current slope} - \text{Step Size}_{Slope}$$

$$= 0.5 - (-0.148)$$

$$= 0.648$$



Gradient Descent for Two Parameters: Step-by-Step

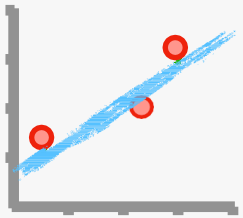
6 And after **475** iterations…

a Evaluate the derivatives at their current values…

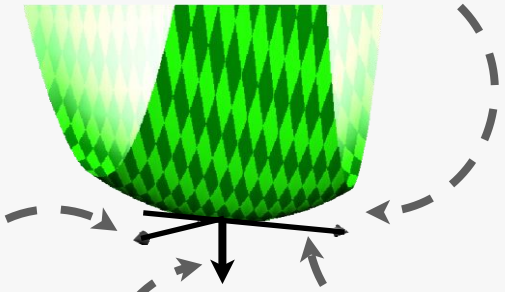
b Calculate the **Step Sizes**…

c Calculate the new values…

…the **Step Size** was very close to **0**, so we stopped with the **current intercept** = **0.95** and the **current slope** = **0.64**…



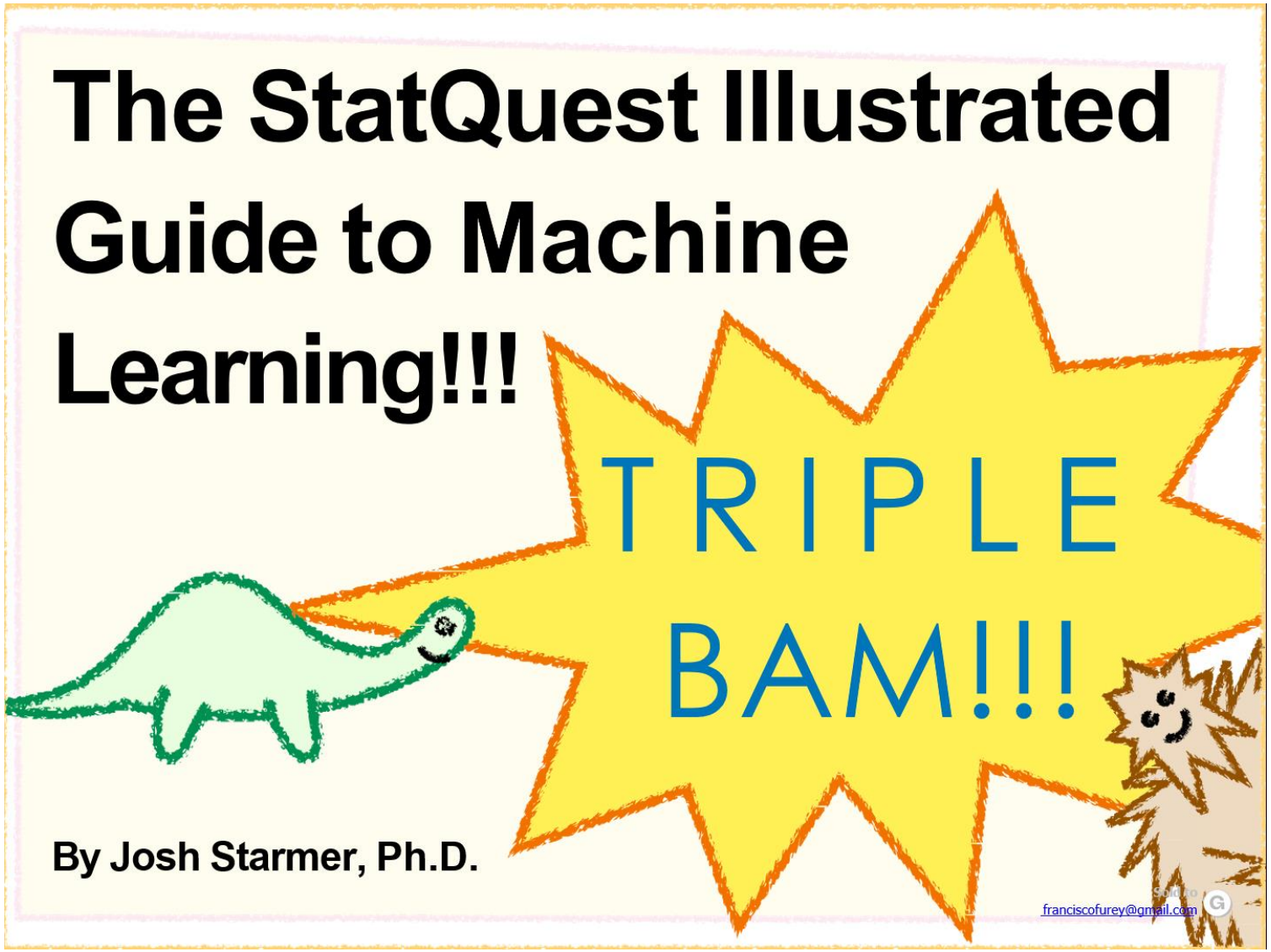
…and we made it to the lowest **SSR**.



7 If, earlier on, instead of using **Gradient Descent**, we simply set the derivatives to **0** and solved for the **intercept** and **slope**, we would have gotten **0.95** and **0.64**, which are the same values **Gradient Descent** gave us. Thus, **Gradient Descent** did a great job, and we can confidently use it in situations where there are no analytical solutions, like **Logistic Regression** and **Neural Networks**.

Gradient Descent is awesome, we will use it in many ML Algorithms.

References


The StatQuest Illustrated Guide to Machine Learning!!!
TRIPLE BAM!!!
By Josh Starmer, Ph.D.
franciscofurey@gmail.com