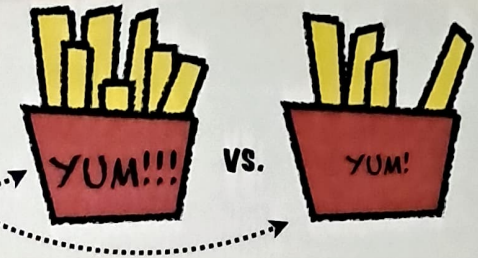
Statistics: Main Ideas

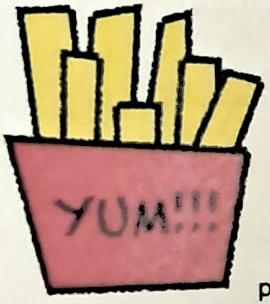
1. **The Problem:** The world is an interesting place, and things are not always the same.

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



2. **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...



Fry Diary

Monday: **21** fries

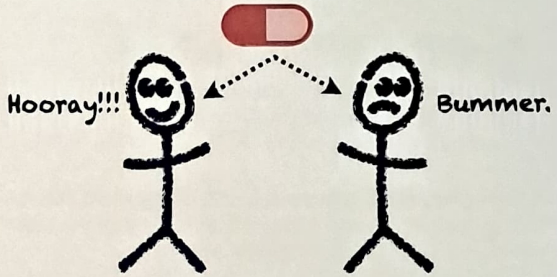
Tuesday: **24** fries

Wednesday: **19** fries

Thursday: ???

...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.

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



...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.

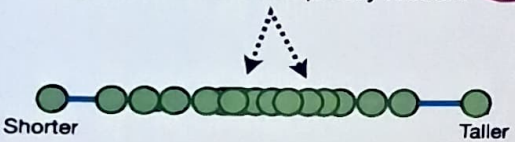
For example, if we predict that the medicine will help, but we're not very confident in that prediction, we might not recommend the medicine and use a different therapy to help the patient.

3. 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**.

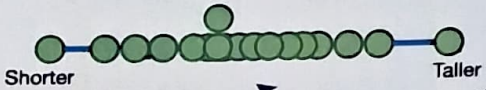
Histograms: Main Ideas

1 **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…

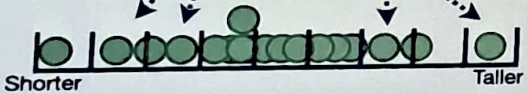


…but measurements that are *exactly* the same are rare, and a lot of the green dots are still hidden.

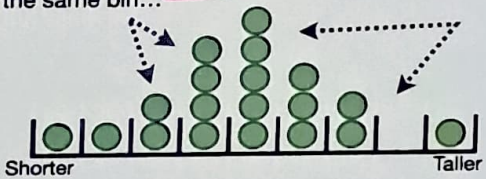
:(

2 **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…



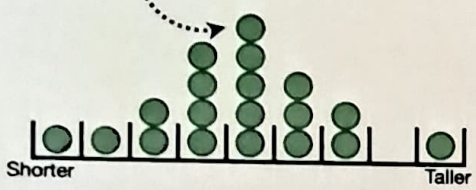
…and we end up with a **histogram!!!**

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

BAM!!!

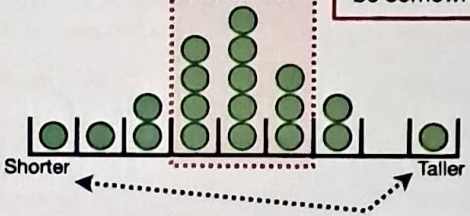
Histograms: Details

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



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

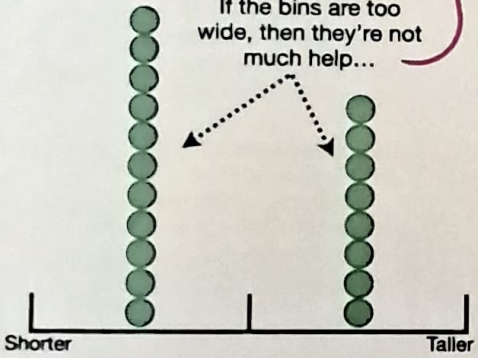
Because most of the measurements are inside this **red box**, we might be willing to bet that the next measurement we make will be somewhere in this range.



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

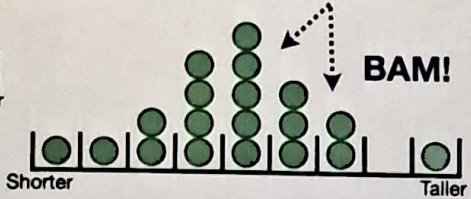
3 **NOTE:** Figuring out how wide to make the bins can be tricky.

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

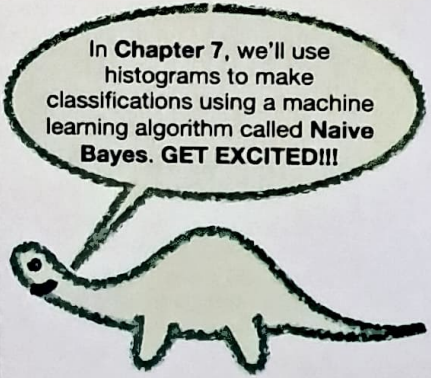


...and if the bins are too narrow, then they're not much help...

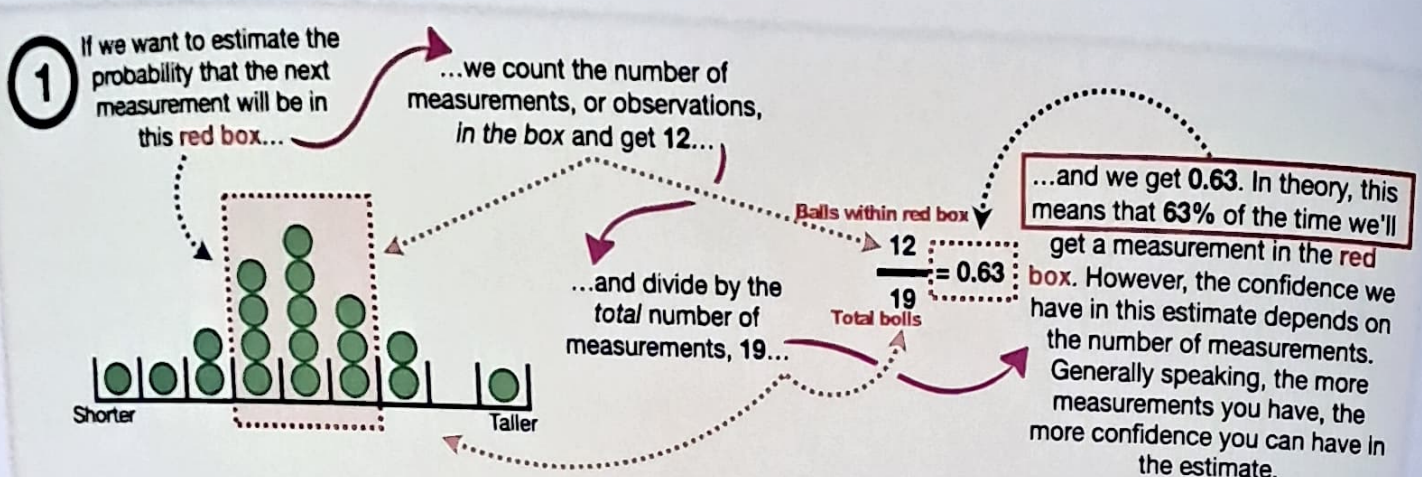
...so, sometimes you have to try a bunch of different bin widths to get a clear picture.

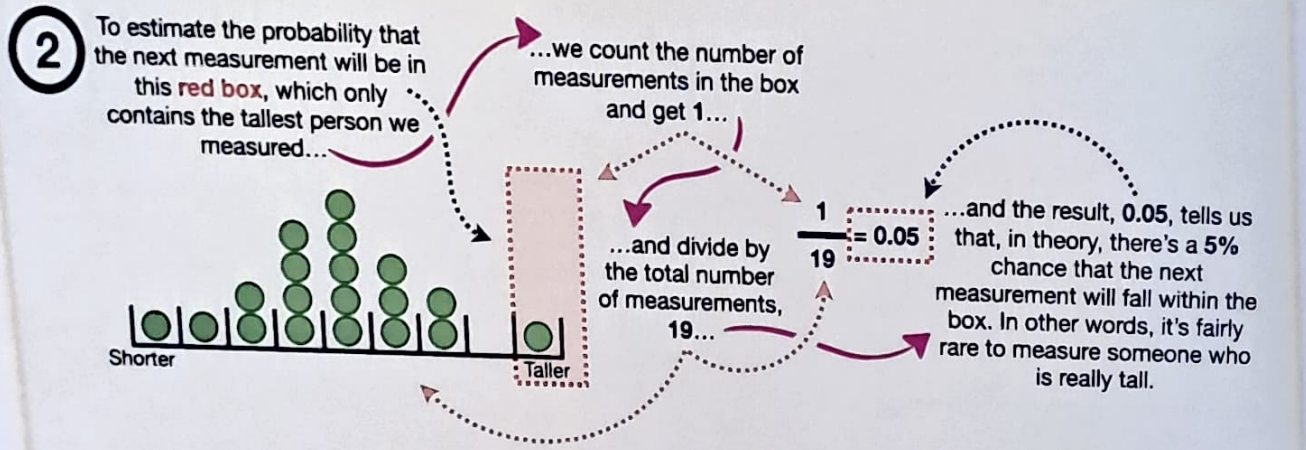


In **Chapter 7**, we'll use histograms to make classifications using a machine learning algorithm called **Naive Bayes. GET EXCITED!!!**



Histograms: Calculating Probabilities Step-by-Step





Histograms: Calculating Probabilities Step-by-Step

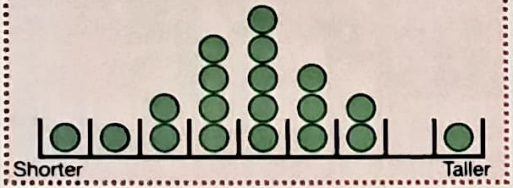
3 To estimate the probability that the next measurement will be in a red box that spans all of the data...

...we count the number of measurements in the box, **19**...

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

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

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



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

...we count the number of measurements in the box, **0**...

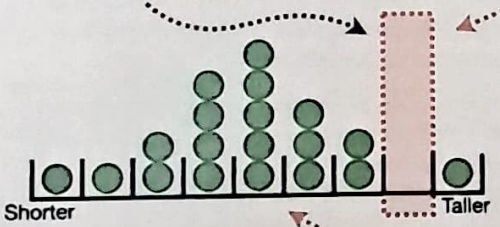
...and divide by the total number of measurements, **19**...

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

...and we get **0**. This is the minimum probability and, in theory, it means that we'll ***never*** get a measurement in this box. However, it could be that the only reason the box was empty is that we simply did not measure enough people.

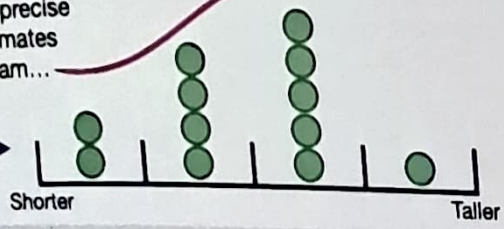
If we measure more people, we may either find someone who fits in this bin or become more confident that it should be empty. However, sometimes getting more measurements can be expensive, or take a lot of time, or both. This is a problem!!!

The good news is that we can solve this problem with a **Probability Distribution**. **Bam!**



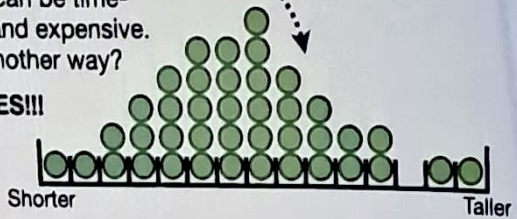
Probability Distributions: Main Ideas

1 **The Problem:** If we don't have much data, then we can't make very precise probability estimates with a histogram...

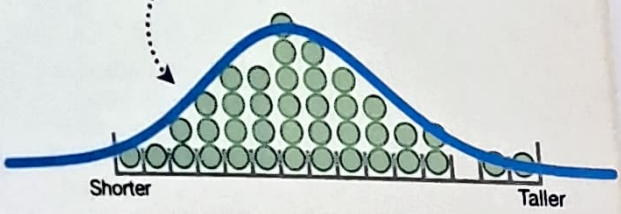


...however, collecting tons of data to make precise estimates can be time-consuming and expensive. Is there another way?

YES!!!

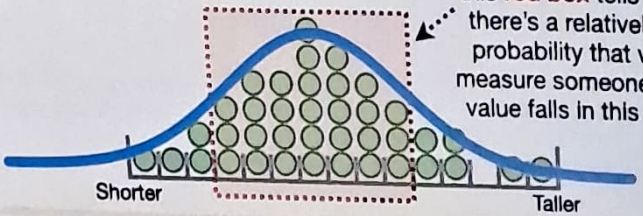


2 **A Solution:** We can use a **Probability Distribution**, which, in this example, is represented by a **blue, bell-shaped curve**, to approximate a histogram.



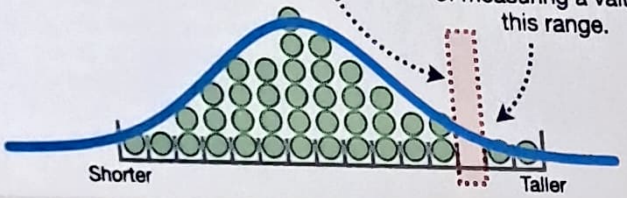
3 This **blue, bell-shaped curve** tells us the same types of things that the histogram tells us.

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.



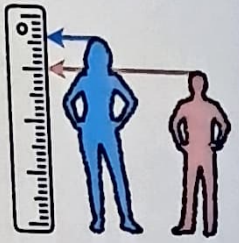
4 Now, even though we never measured someone who's value fell in this range...

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



5 **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.**