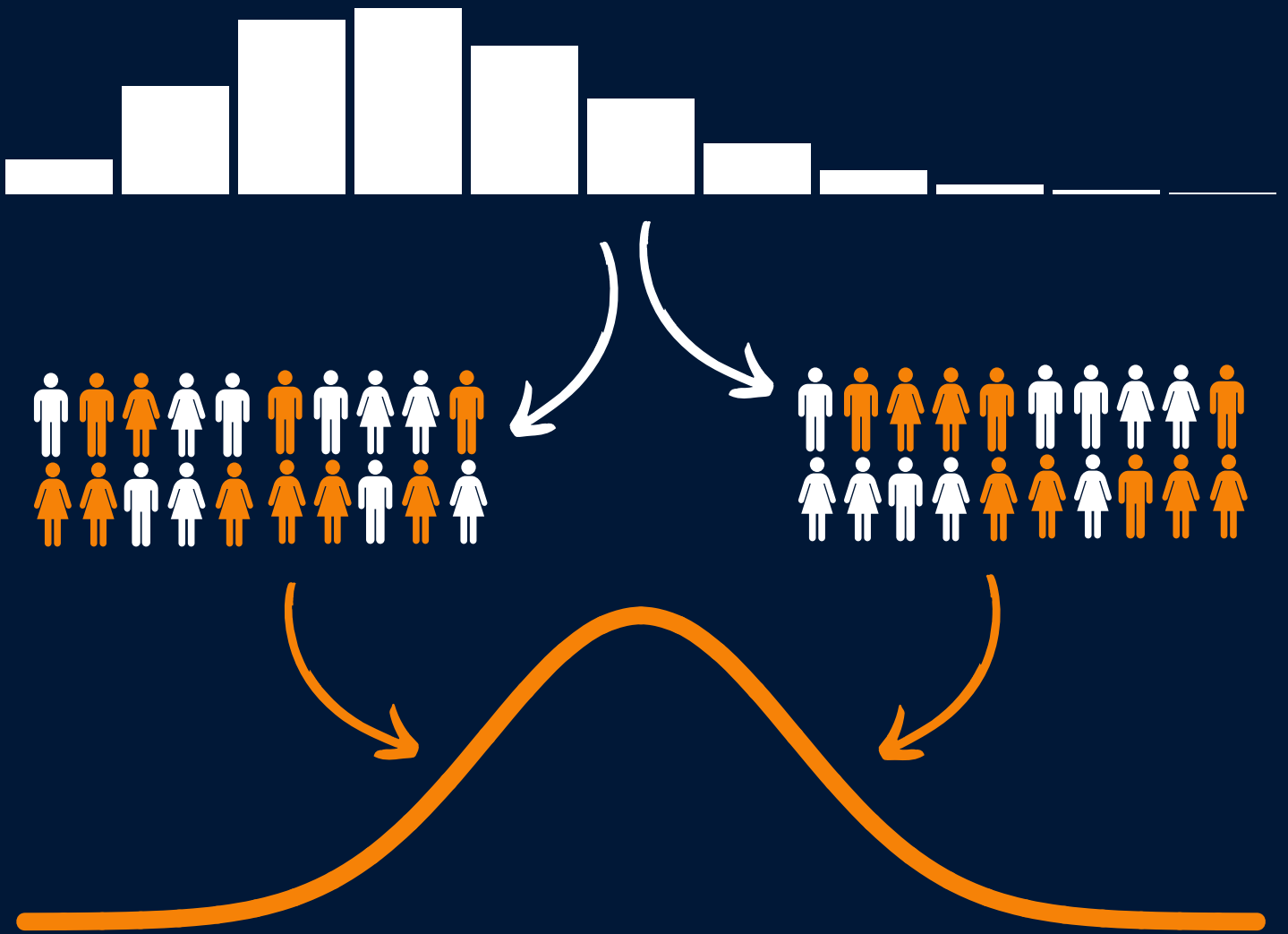


WHAT THE HECK IS THE CENTRAL LIMIT THEOREM?

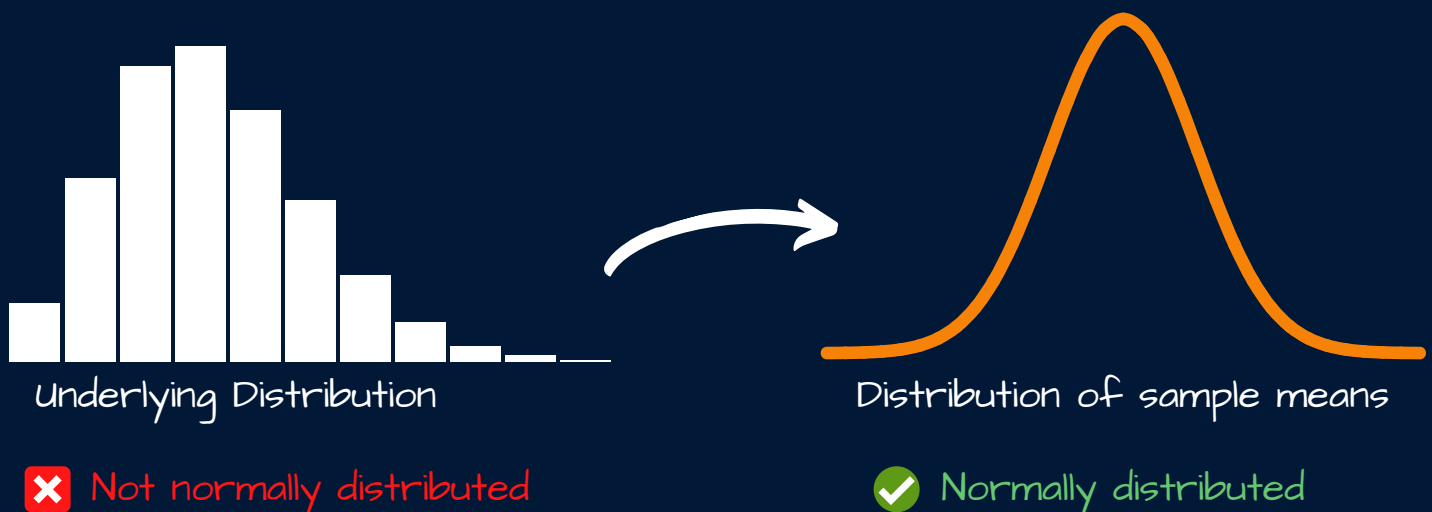


AN OVERVIEW

What?

In Statistics, the **Central Limit Theorem** is the idea that regardless of the **underlying distribution** of our data, the **distribution of the means** of many **samples** from within that distribution will be **normally distributed** (or at least, will tend towards being normally distributed).

Sample sizes ≥ 30 are often considered sufficient for the CLT to hold (although sometimes they will need to be larger)



Why does this matter?

Using the CLT to create a normal (or near normal) distribution allows us to run many types of statistical tests (where normal distribution assumptions need to be met)!

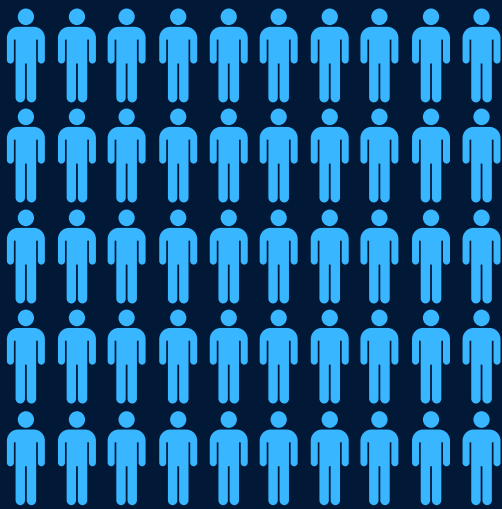
A worked example...

Imagine that we've been tasked with understanding the distribution of heights for **all men** in the USA...

The idea that we could know the **true mean** and **true standard deviation** of around 160 million men is completely implausible...but through the clever use of **sampling** and a little bit of the **Central Limit Theorem** we can get a pretty good estimation for them!

So, instead of attempting to measure the whole population, we start by collecting a **random sample of 50 men**...

Random Sample 01



Within this sample there will be men of differing heights, some shorter, some taller, but as it's a random sample, chances are that most will be somewhere around average height!

For our sample, we measure the height of each of the 50 men, and calculate the **mean height for the sample** which for our example turns out to be **174cm!**

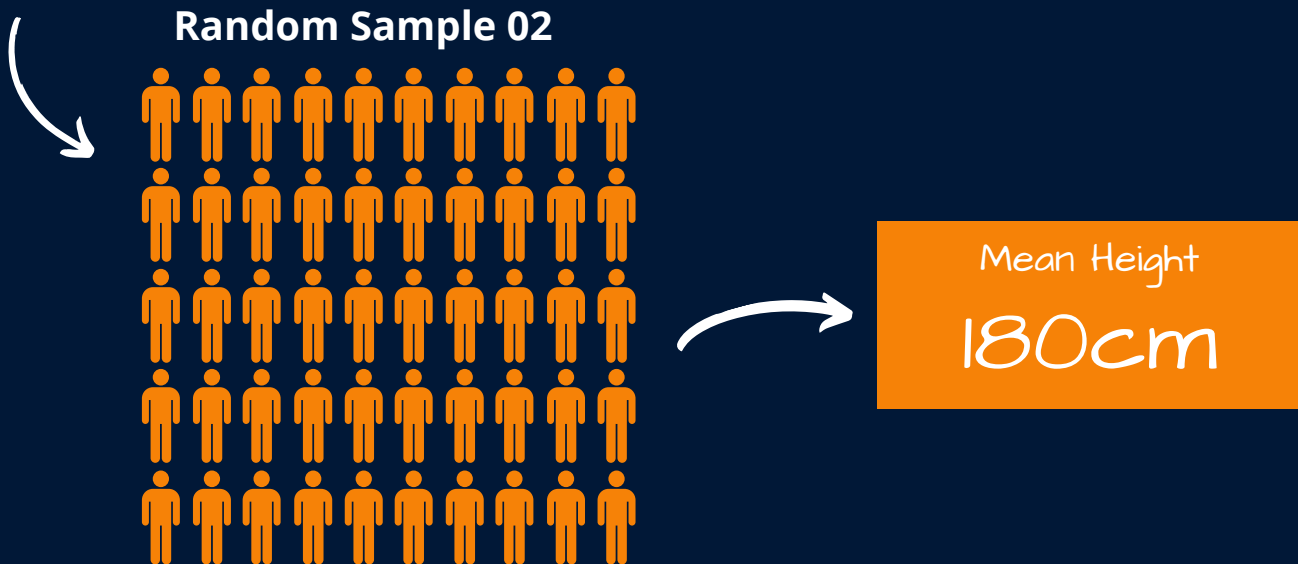
Mean Height

174cm

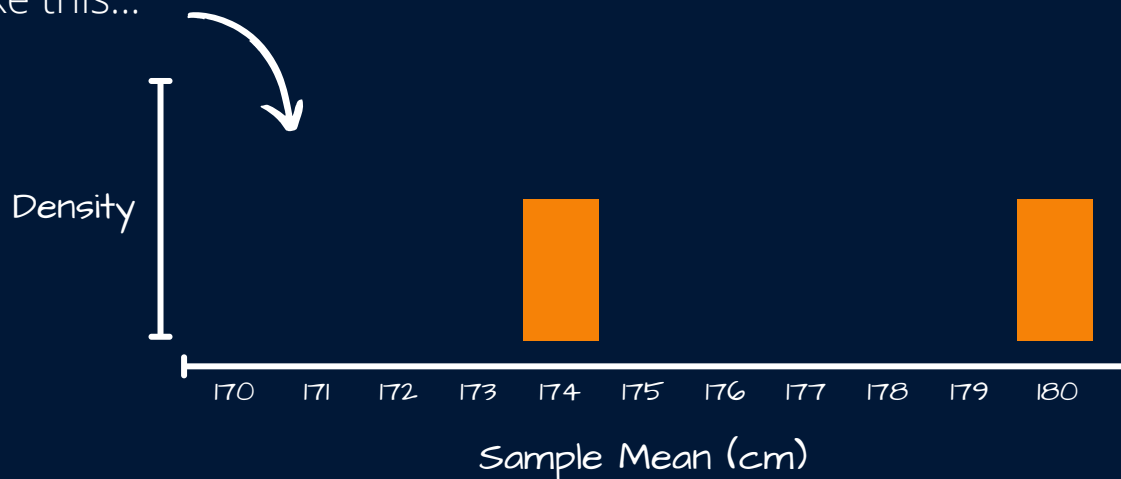
Let's now take a second sample...

A worked example...

We take a second sample, again made up of a **random selection of 50 men**...



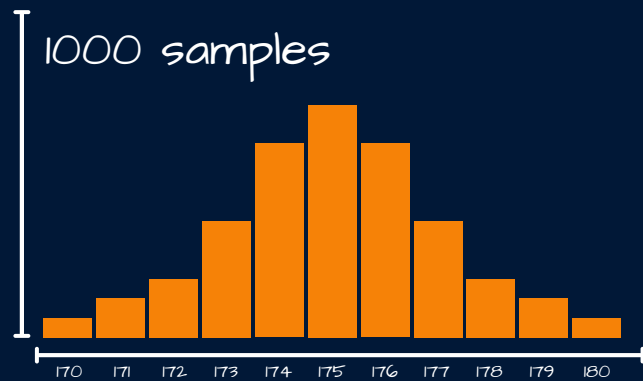
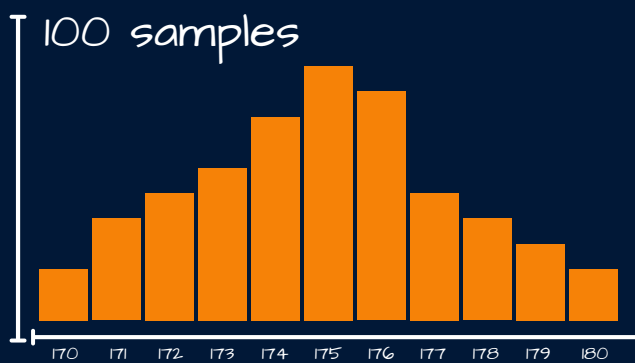
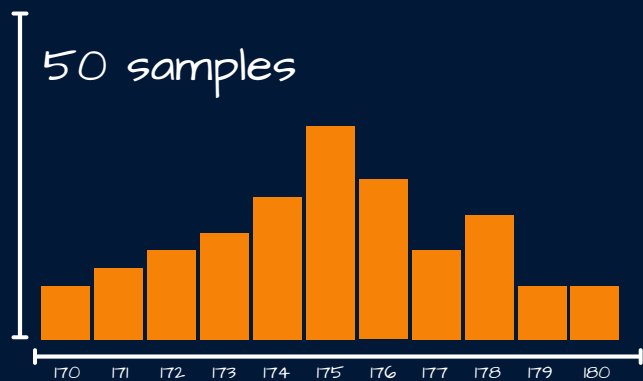
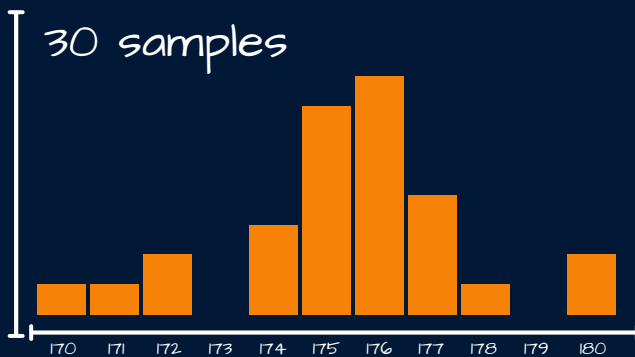
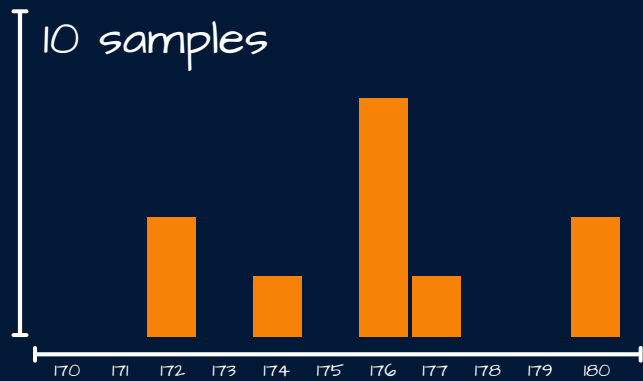
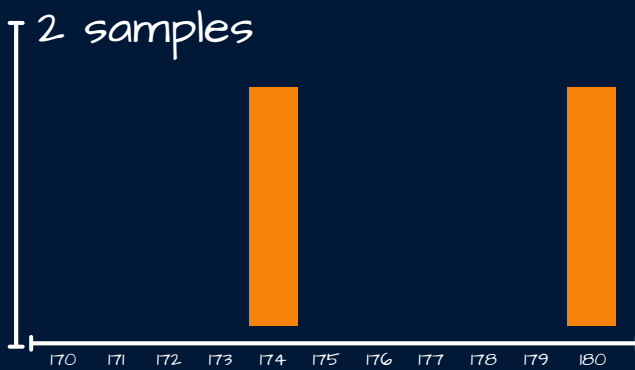
Just like with Sample 01, we measure these 50 men and calculate the mean height - which this time turns out to be **180cm!** At this point, we've measured **two samples** - and therefore we have **two sample means** of 174cm and 180cm. If we plotted these sample means, it would look a little bit like this...



Not much to write home about... But, as we obtain **more and more random sample means** - watch what starts to happen...

DATA SCIENCE INFINITY

A worked example...



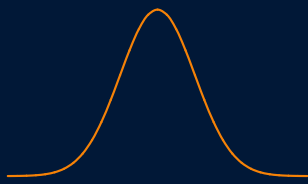
As you can see - the more samples we take, just like the Central Limit Theorem said, the closer and closer this **distribution of sample means** gets to the shape of a true normal distribution.

Now, that's not to say that the full population mean was definitely a normal distribution. **We don't know** the actual distribution of the full population and that's the reason we're sampling in the first place...

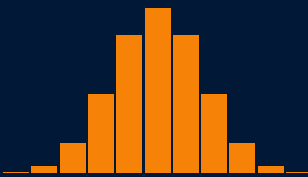
Always to normal...

One very interesting thing about the Central Limit Theorem is that it doesn't actually matter what the underlying distribution is - the distribution of the **sample means** will *always* tend towards a normal distribution...

**Underlying
Distribution**



Normal



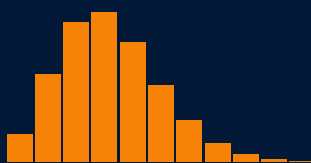
Binomial



uniform

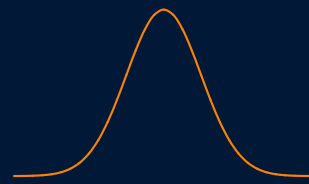


Bernoulli

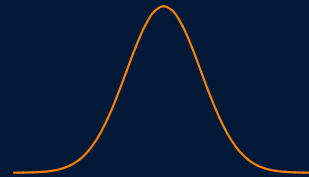


Poisson

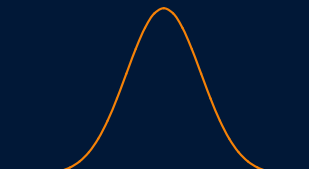
**Distribution of
Sample Means**



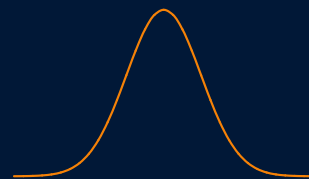
Normal



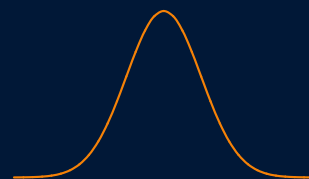
Normal



Normal



Normal



Normal

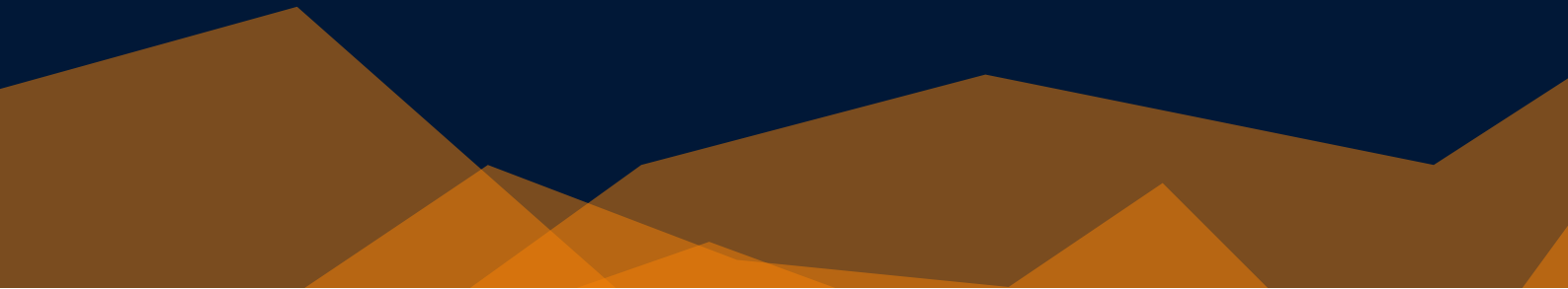
Always to normal...

This idea that the **distribution of sample means will tend to normal, no matter the underlying distribution** may just seem like an interesting fact but it's actually really powerful - and here's why...

In practice, we often have no idea what type of distribution our data comes from, and the Central Limit Theorem has our back in this case, as it doesn't really care. It just states that if we take enough samples from the data, the distribution of the **sample means** will tend to normal.

Still, maybe you're saying "so what?"

Well, running statistical tests on **unknown distribution types** can be really tricky and problematic, but if we always have a way to transform how we think of our data (from non-normal to normal) it allows us the freedom to robustly run all sorts of statistical tests that allow us understand our data and compare our data to other data in a fair and accurate way!

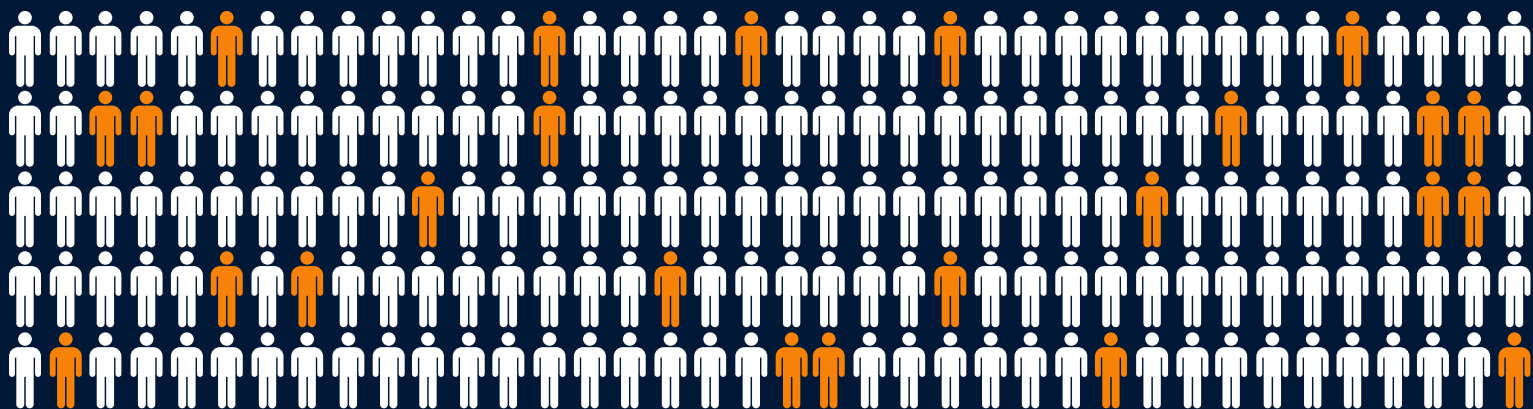
The bottom of the slide features a decorative graphic consisting of several overlapping, semi-transparent geometric shapes in shades of brown and orange, creating a layered, mountain-like effect.

Bootstrap Sampling

In our example of men's heights - we ended up needing around 100 random samples of 50 men before we ended up with a distribution that started to look somewhat normal.

This would still require us to go out and measure **5,000 men** which might be quite an expensive and time consuming task...

Another common way to do this is using a process called **bootstrapping** or **bootstrap sampling**. With this approach we could instead start with a random sample of say, **500 men**



From that initial random sample, we randomly select a **smaller sub-sample** - for examples sake let's say this was of size 30 (shown in orange).

We calculate the mean height of this sub-sample - which we can see here is equal to **178cm**...

Mean Height (Sub-Sample)

178cm

Bootstrap Sampling

Next, from this **same sample of 500 men** we select **another** random sub-sample of 30. This is done *with replacement* meaning each sub-sample could contain some of the same men



Just like the first sub-sample, we take the mean height of this next sub-sample. This time this is equal to **174cm**

Mean Height (Sub-Sample)

174cm

Each time we do this, our sub-sample will contain a different combination of men - and thus will have a different mean height.

We can repeat this sampling process as many times as we like - and can make the computer do all of the hard work for us. We could even ask for something like 10,000 sub-samples - we can then plot the distribution of the 10,000 sample means - form a normal distribution, and go from there!

This is a much more efficient way of getting a distribution of sample means, as we need less observations to begin with, we just use the clever bootstrap sampling technique to get a variation of sample means!