Statistical Tests

Quantitative Analysis:
Week 5

Some Housekeeping

- Programming assignments will generally be due on Monday nights.
- Late submissions will not be penalised (within reason).
- This means that a couple of late submissions are no problem. If you submit several assignments late, Moodle will notify you of a grade penalty for future late submissions.

To Recap…

We talked last week ago about **case selection** and **sampling** – and various ways this can create bias, such as **selecting on the dependent variable** or **self-selection bias**.

So, assume you've gathered your data and are confident that you got an unbiased sample.

Now what?

Categories of Analysis

- Statistical **tests** are primarily used to investigate sample data – the most important use being finding out whether two samples came from the same **population**.
- Statistical **models** are mathematical descriptions of the system you are studying. Unlike tests, they are expected to have **predictive power**.

Simple Hypothesis Testing

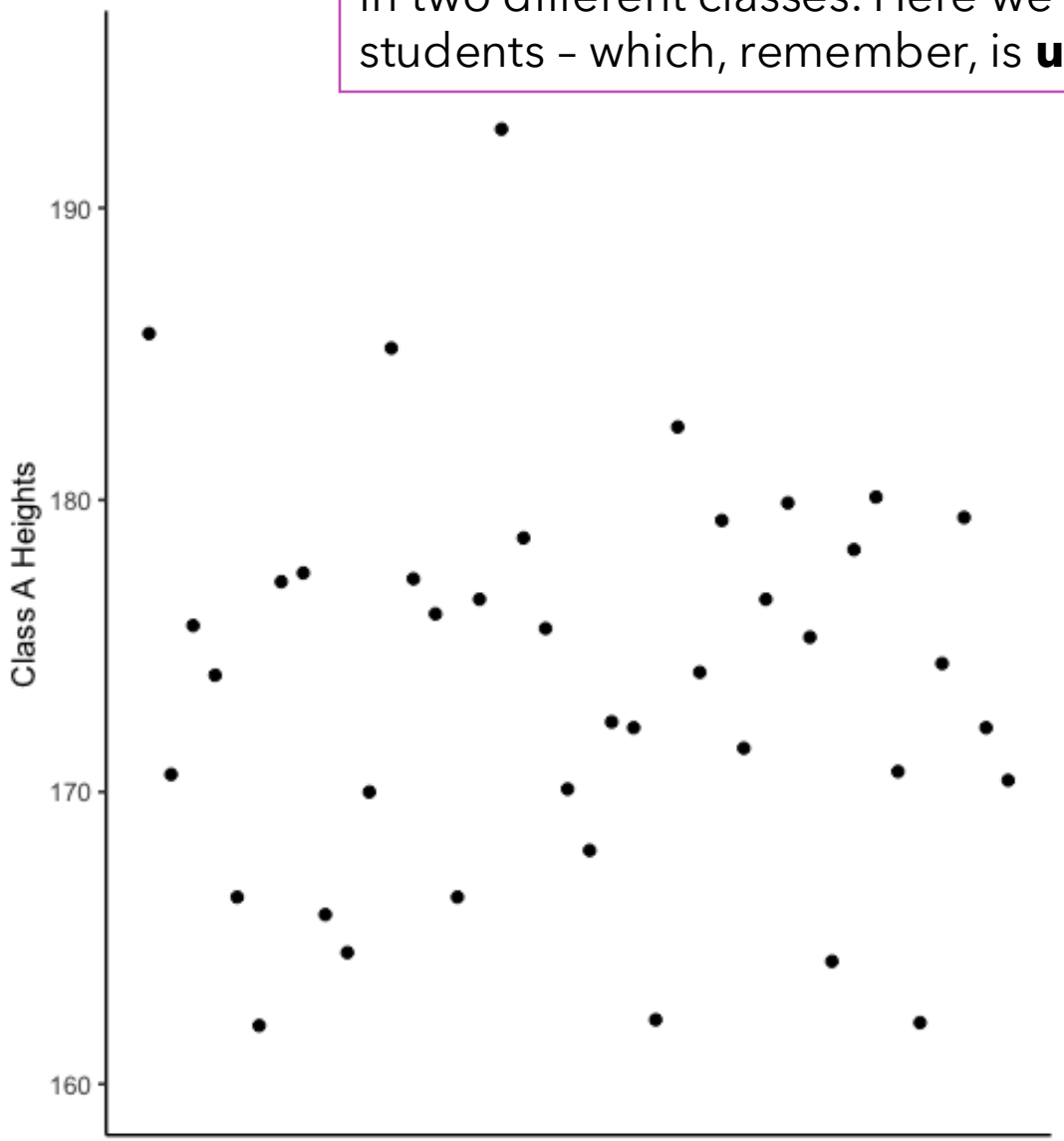
- Statistical tests are used to test simple hypotheses about sample data.
  - "Sample X is from a different population to Sample Y"
  - "Sample X has a different mean to the expected mean μ"
- With non-sample data, these are easy to test; we could just compare the means of the data sets.
- With sample data, we must think **probabilistically**.

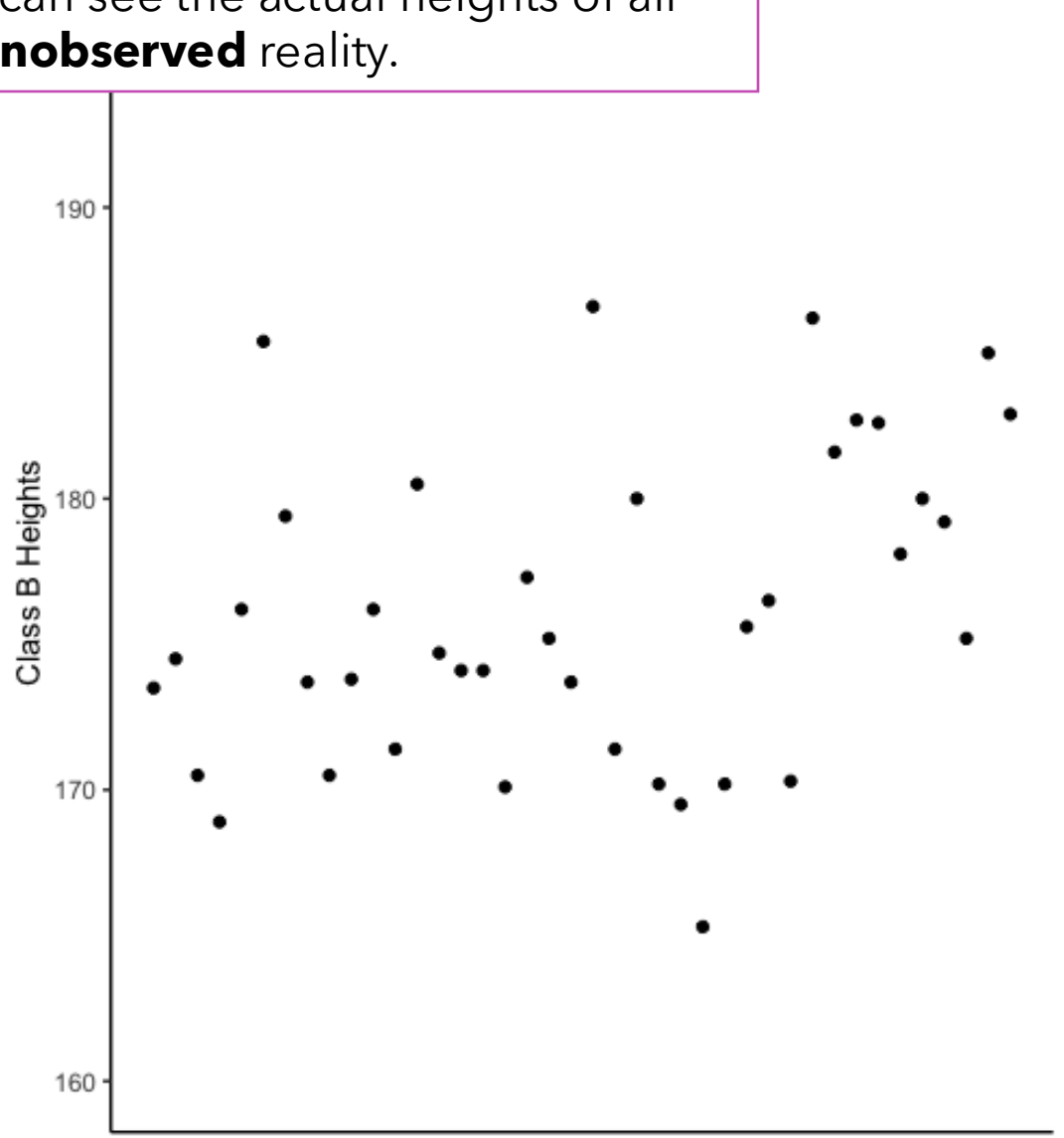
- For example: if I measure the heights of every person in Class A, and every person in Class B, I can calculate the means and say for certain if the classes have a different mean height.
- But what if I only measure 10 randomly chosen people from each class of 40 people?
- The **sample means** I can calculate from this data will be different from the **population means**.
  - Even if Sample Mean A is higher than Sample Mean B, it doesn't necessarily follow that Population Mean A must be higher than Population Mean B.

Sample vs. Population Means

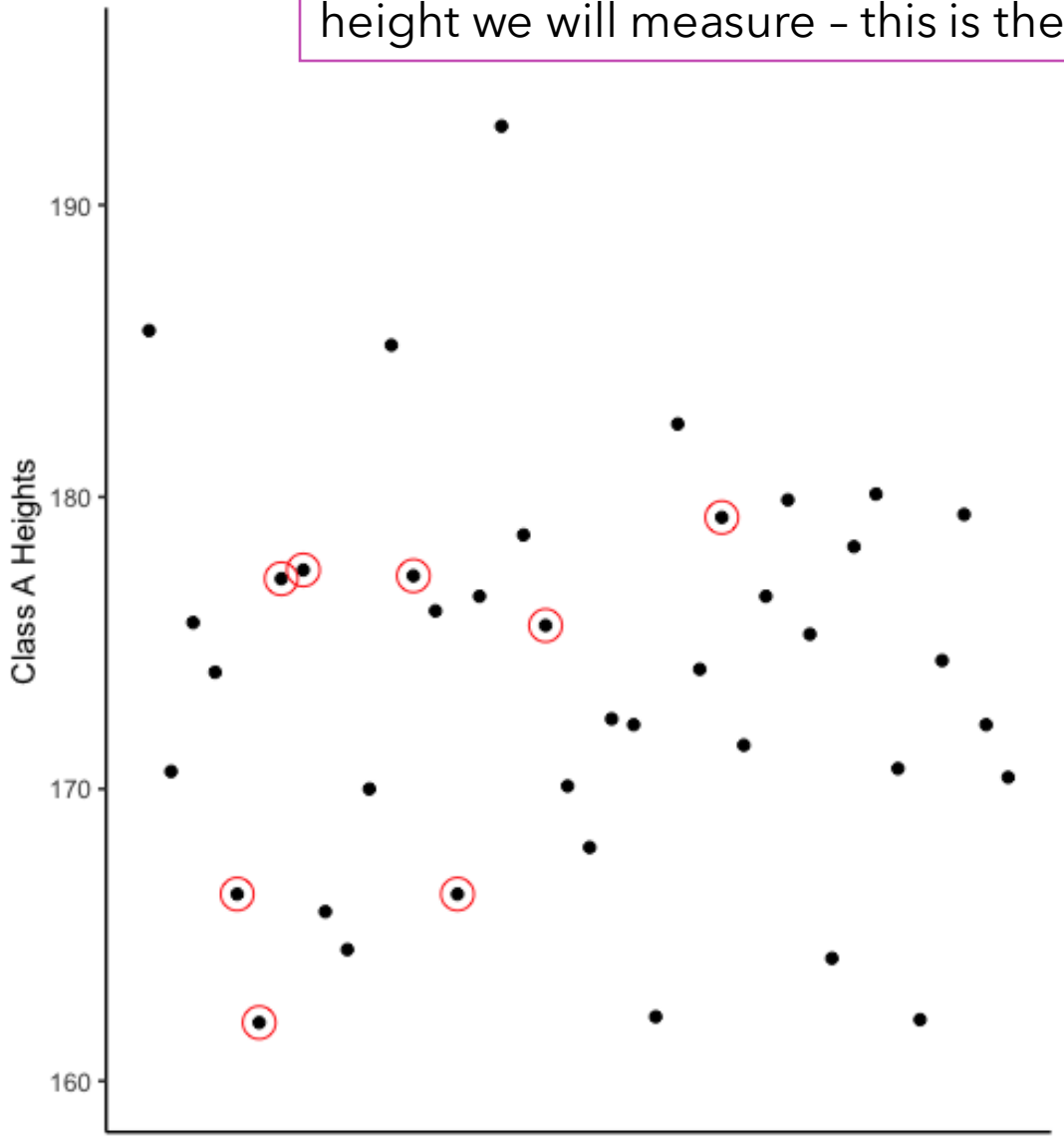
- We can usually only observe **sample means** – but we can use sample data to calculate the probability that the overall **population mean** is within a certain range.
- We can never be totally certain, but we can estimate how **confident** we are that a set of samples comes from a given population.
- By using **confidence levels** like 95% or 99% confidence, we can then reject or retain hypotheses about our data.

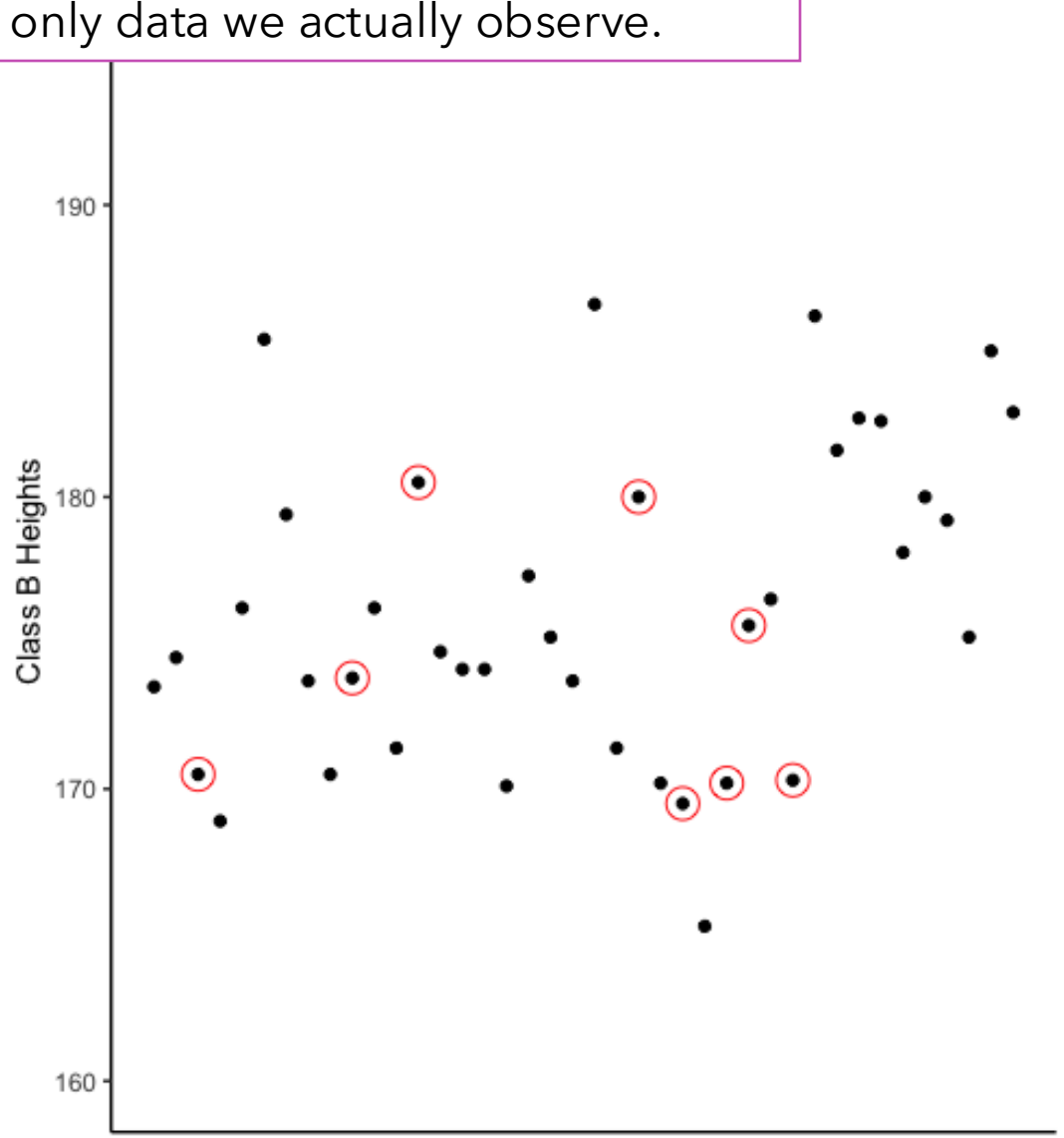
Let's go back to the example of measuring the heights of students in two different classes. Here we can see the actual heights of all students – which, remember, is **unobserved** reality.



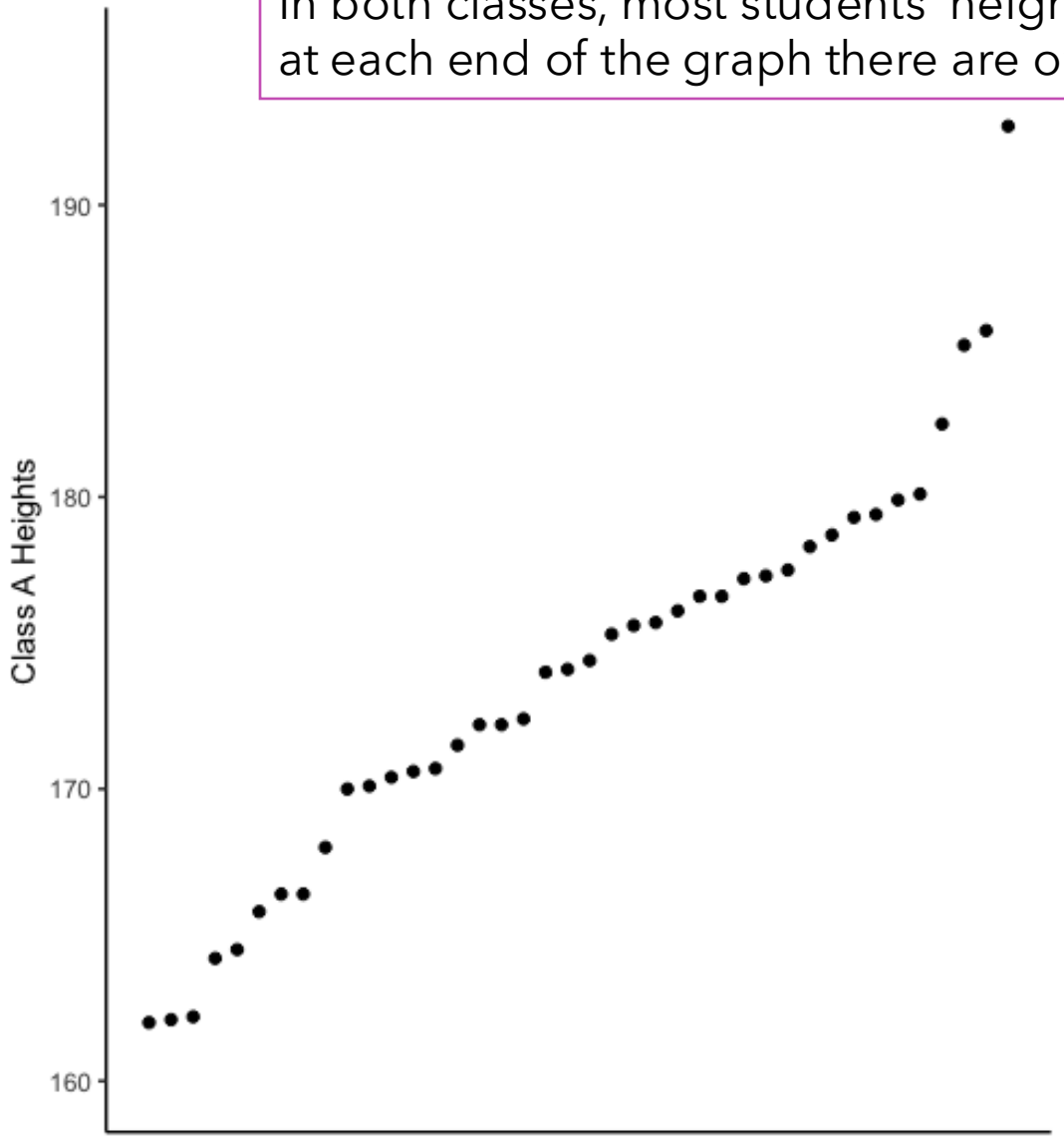


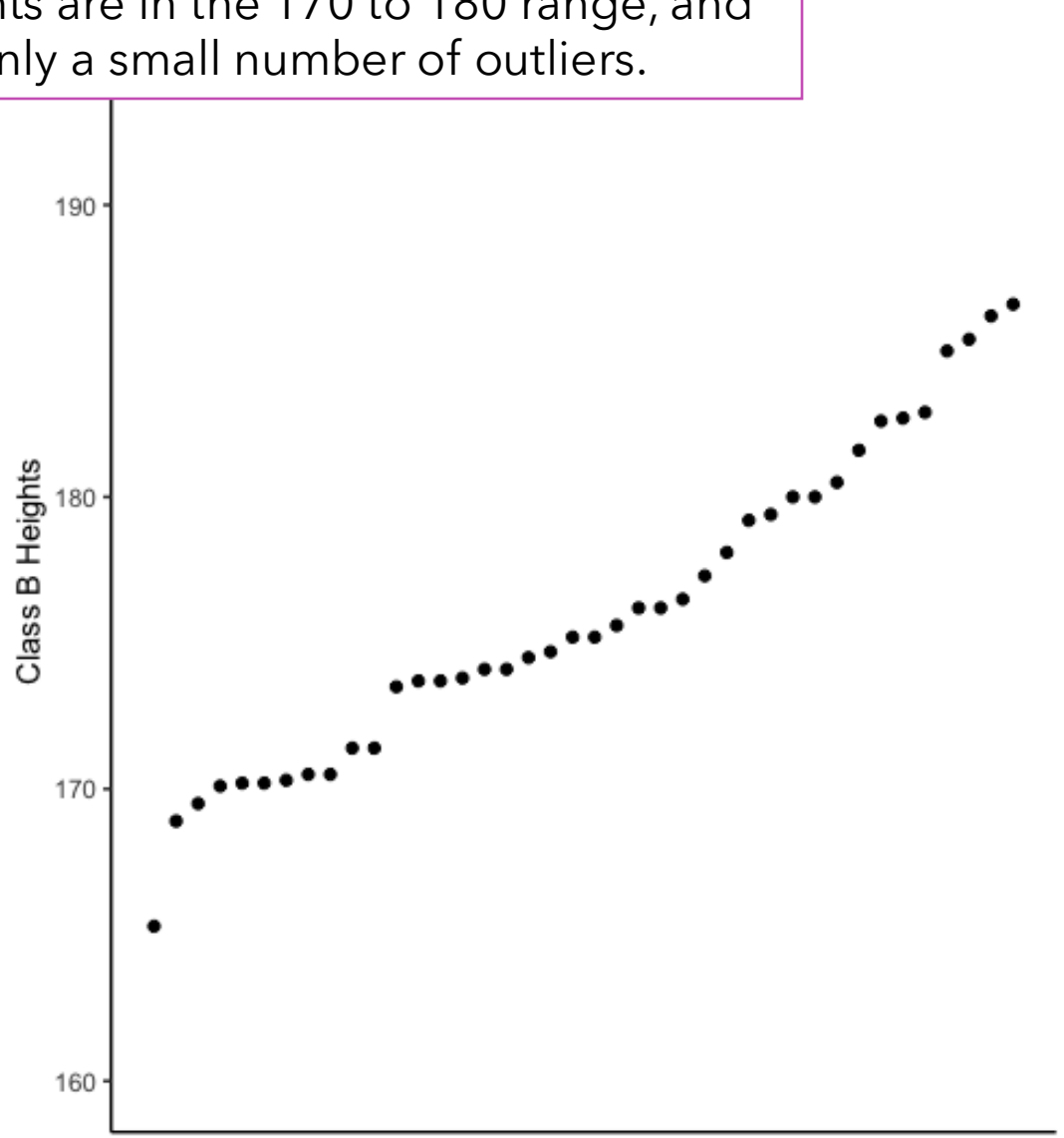
From each class, we select a **random sample** of eight students whose height we will measure – this is the only data we actually observe.



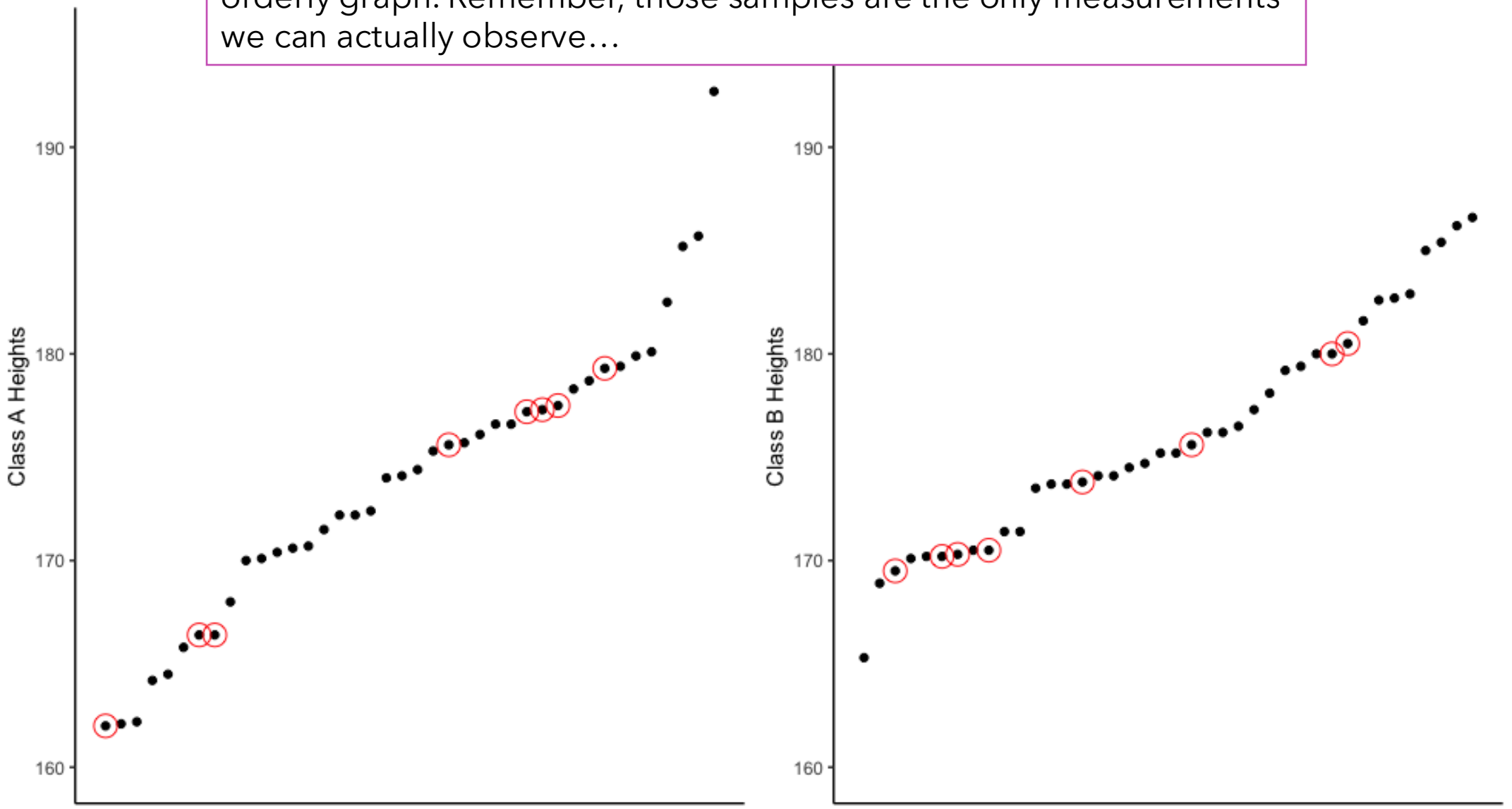


Ordering by height gives a clearer sense of the variable **distribution**. In both classes, most students' heights are in the 170 to 180 range, and at each end of the graph there are only a small number of outliers.

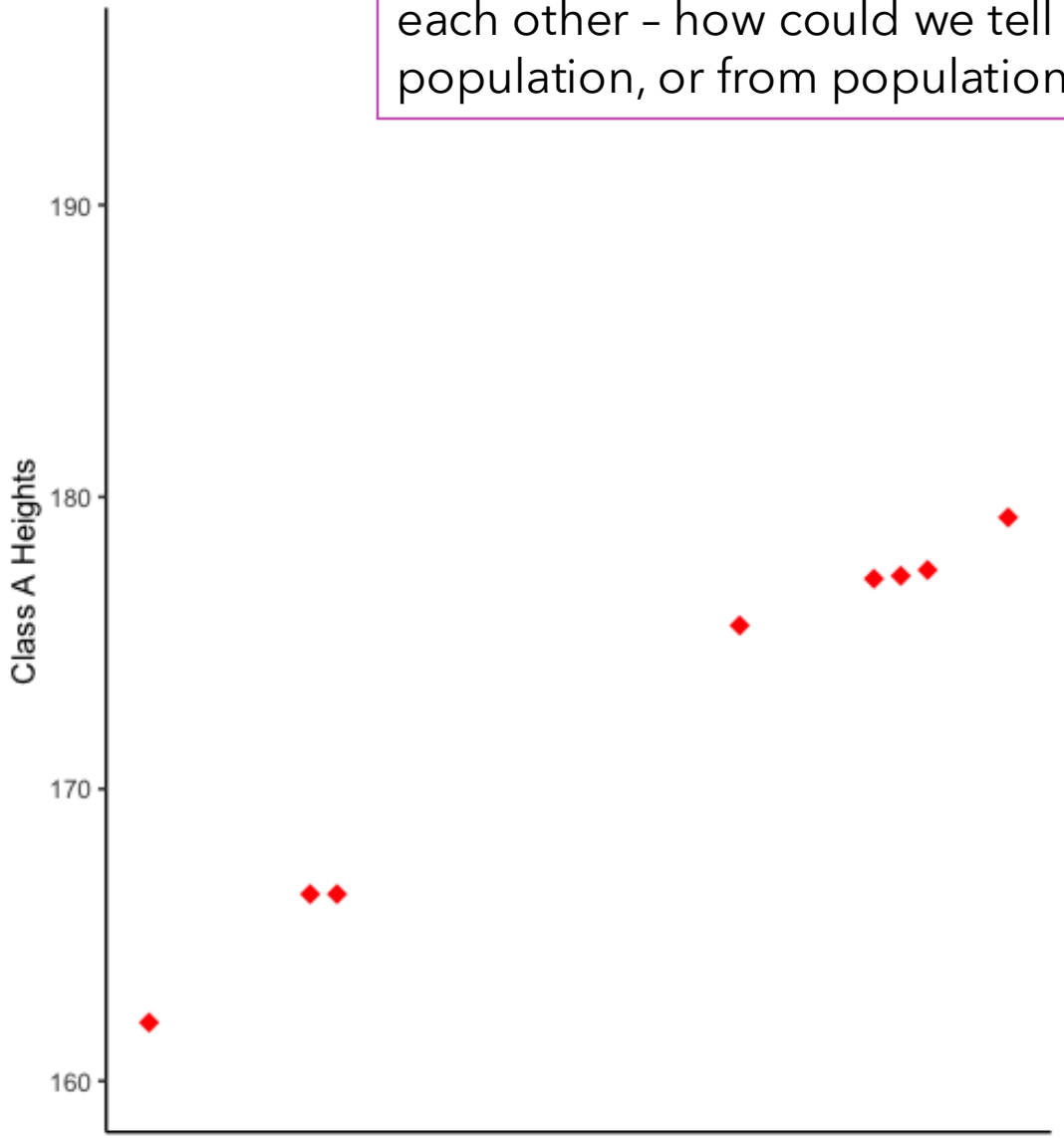


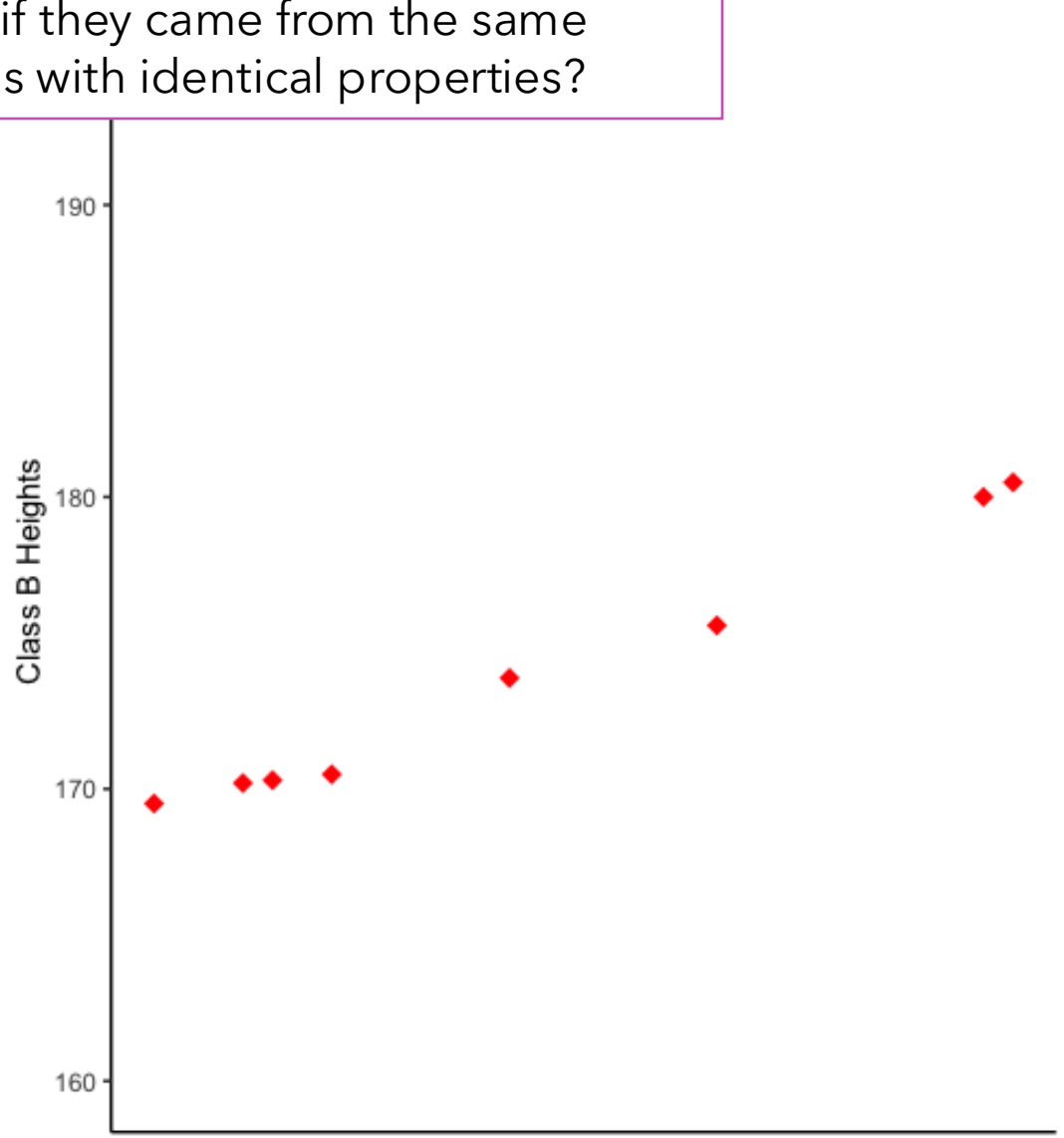


Here are the eight samples we measured, highlighted on the more orderly graph. Remember, those samples are the only measurements we can actually observe…



… So in reality, as a researcher, this would be the only data you have to work with. These two samples look quite different from each other – how could we tell if they came from the same population, or from populations with identical properties?

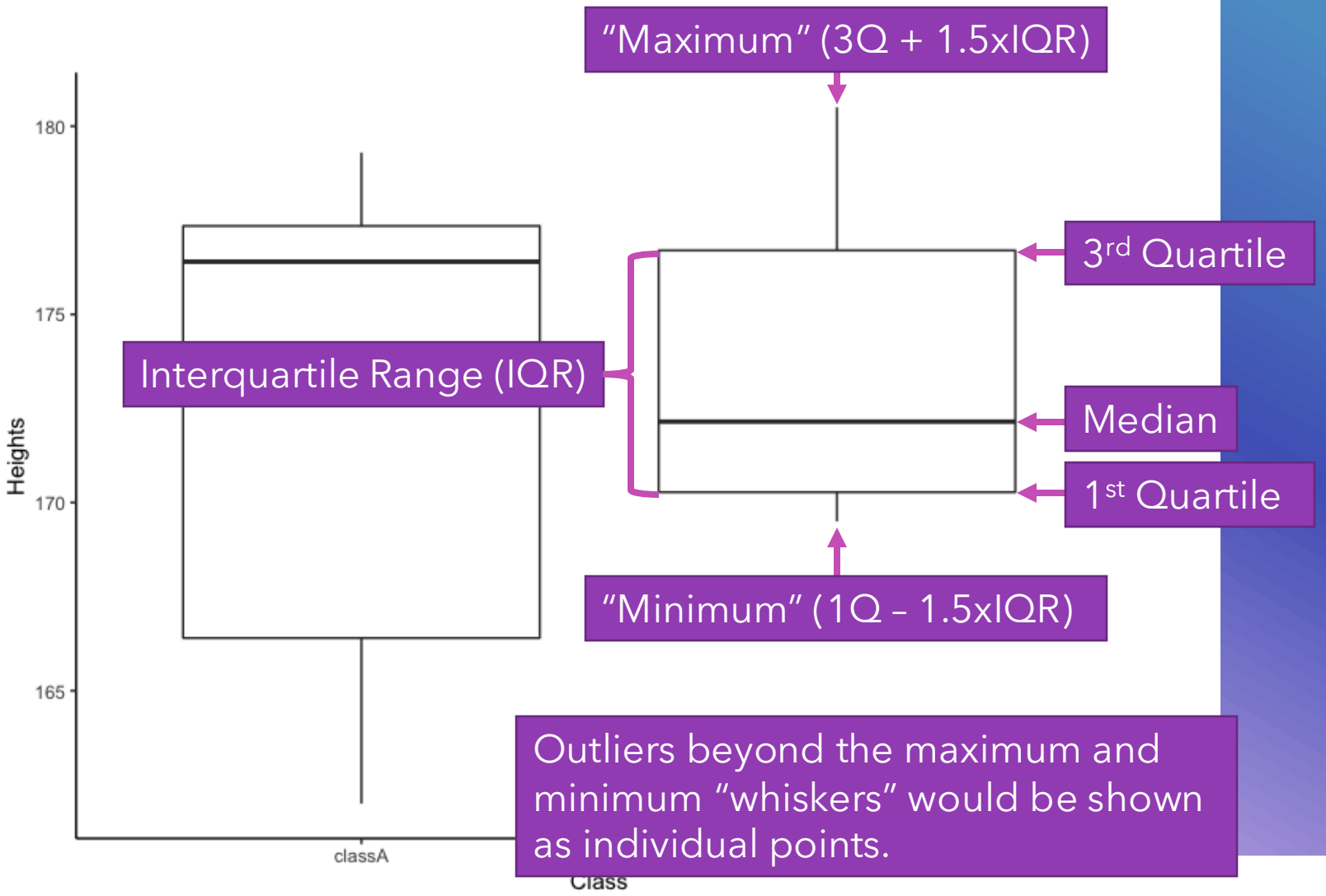




Boxplot

Boxplots are a useful way to understand the distribution of sample data at a glance.

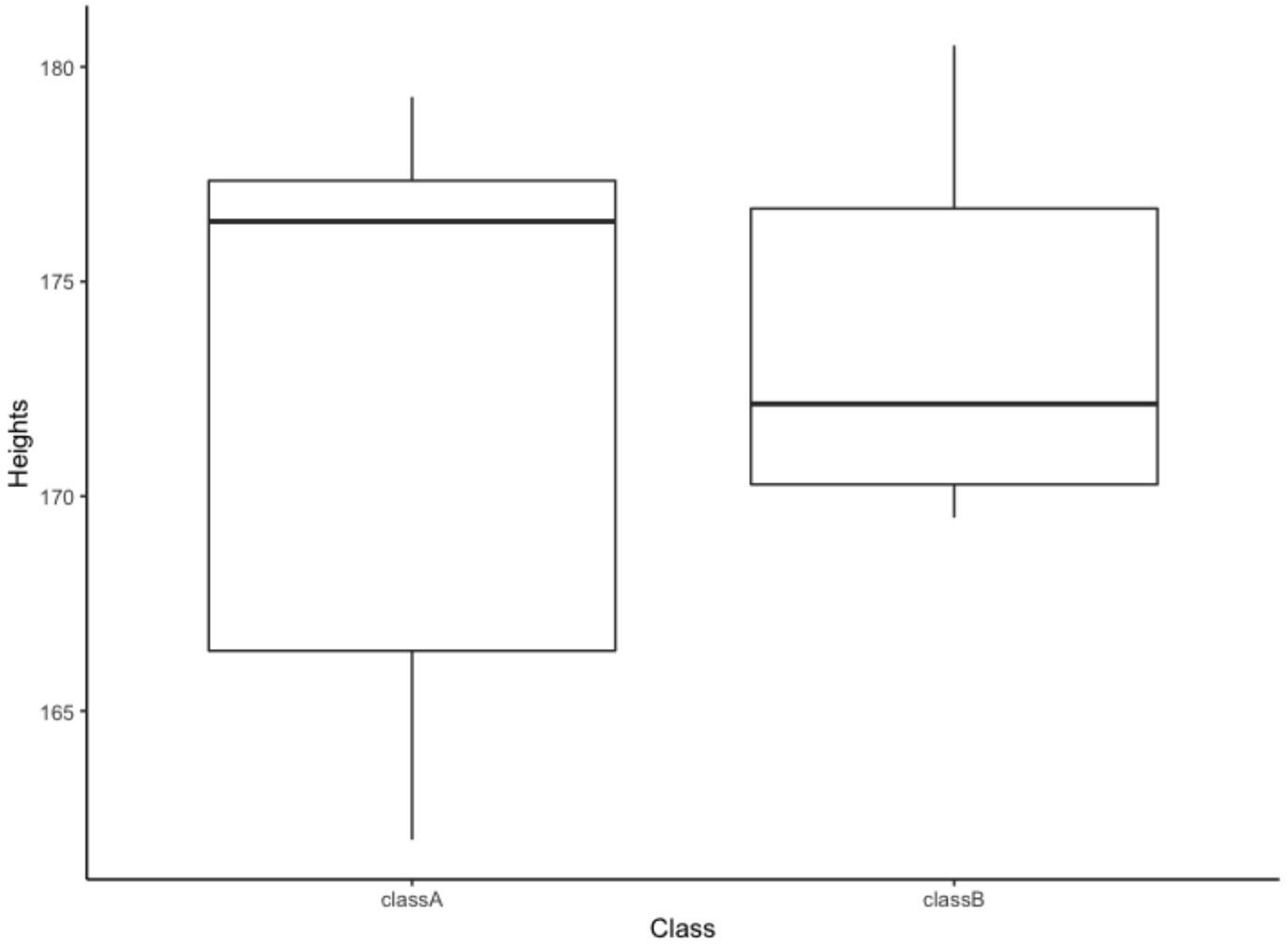
This is the same data as the previous slide, shown in a boxplot graph which makes the sample differences clear.



These two samples are quite different, even though they have significant overlap.

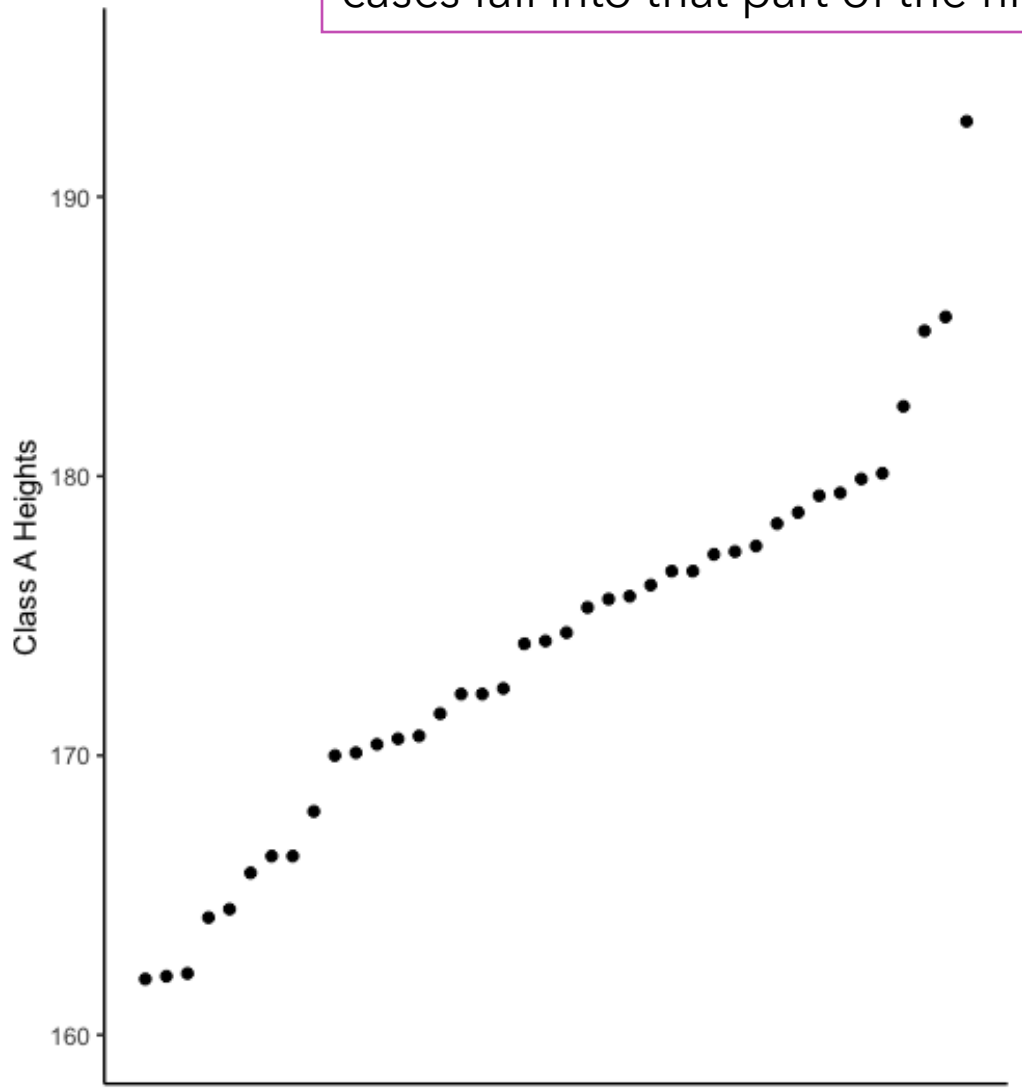
How can we calculate the likelihood that they come from populations with the same characteristics?

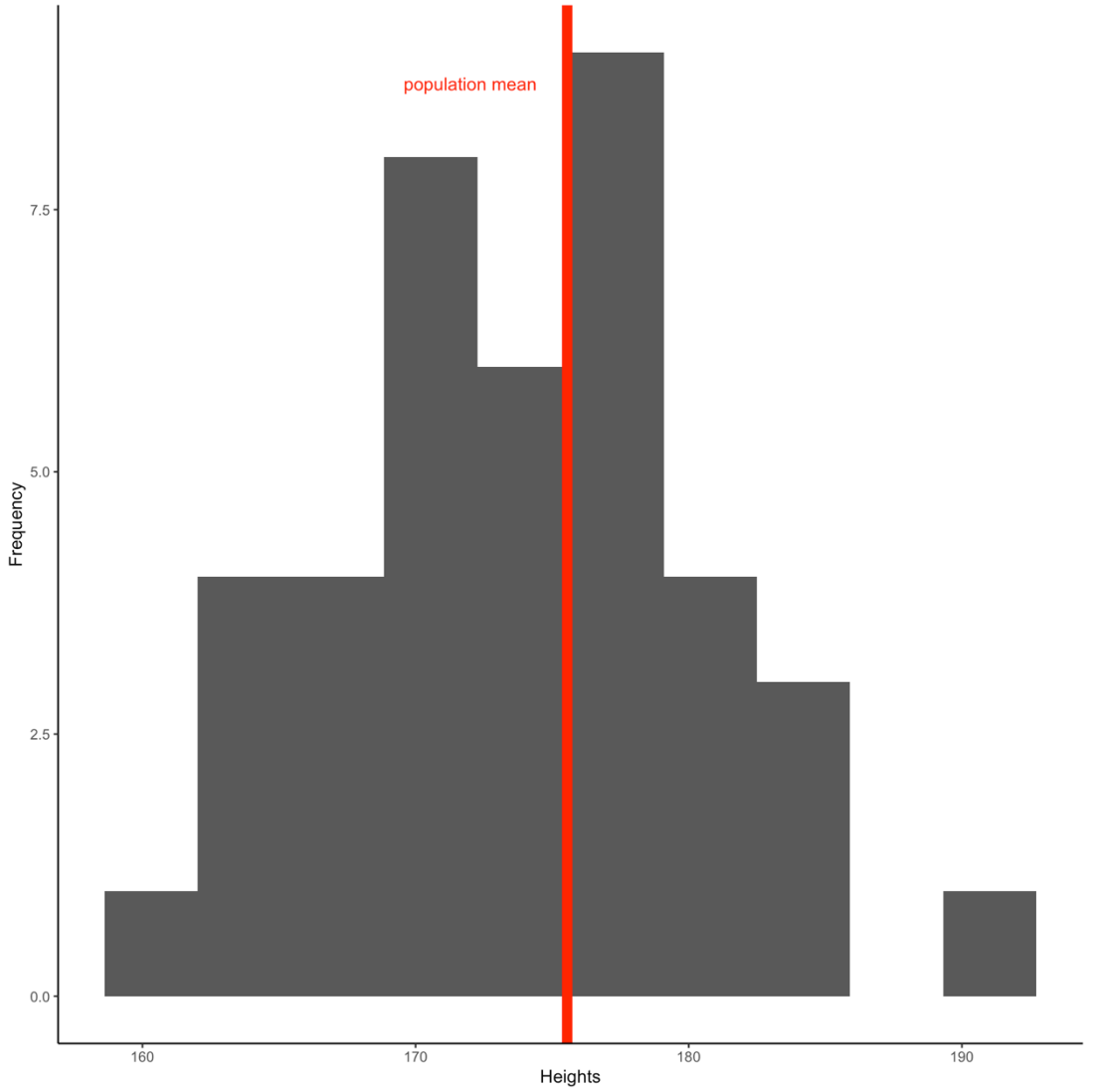
In other words: how can we make inferences about the **unobserved populations** from the **observed samples**?



Population Distributions

Look again at the graph of the heights in the overall population of Class A. We can also represent this as a histogram, which shows the **distribution** of values of this variable – higher bars mean more cases fall into that part of the histogram.

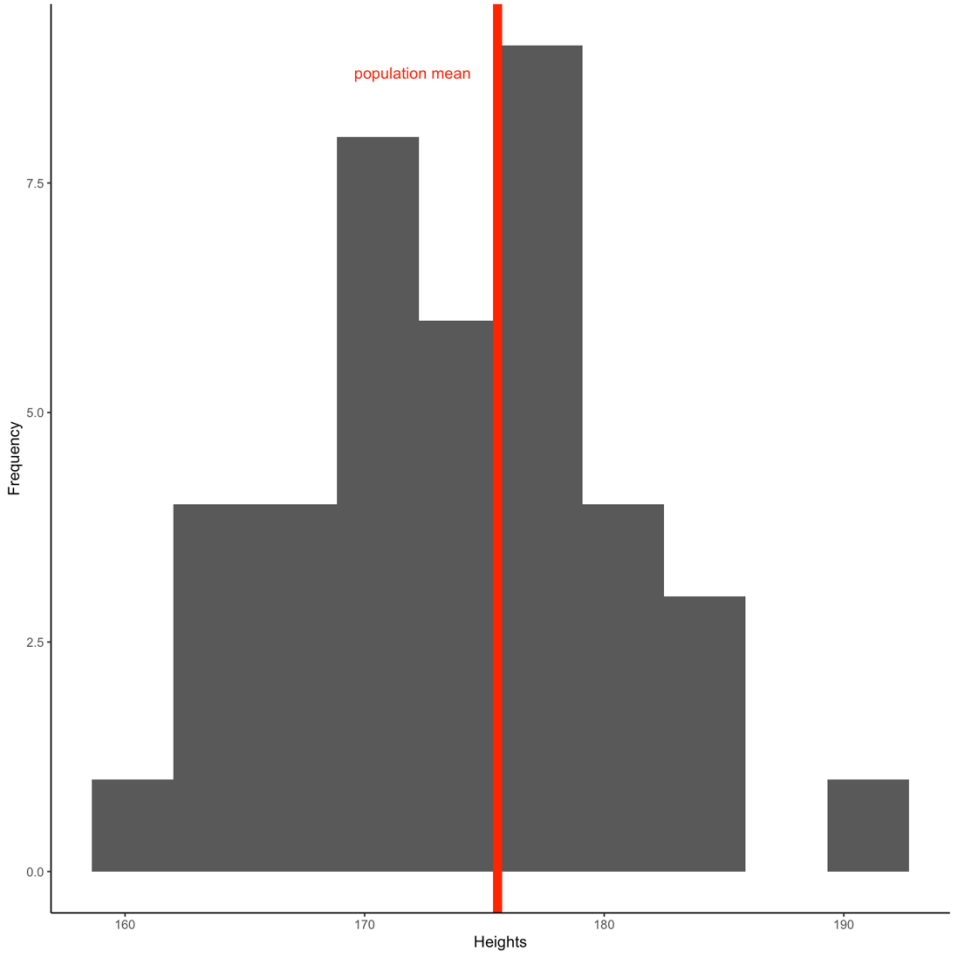




This histogram shows us that the **density** of cases is highest near the mean of the population, and drops off towards the edges, where we find very low density.

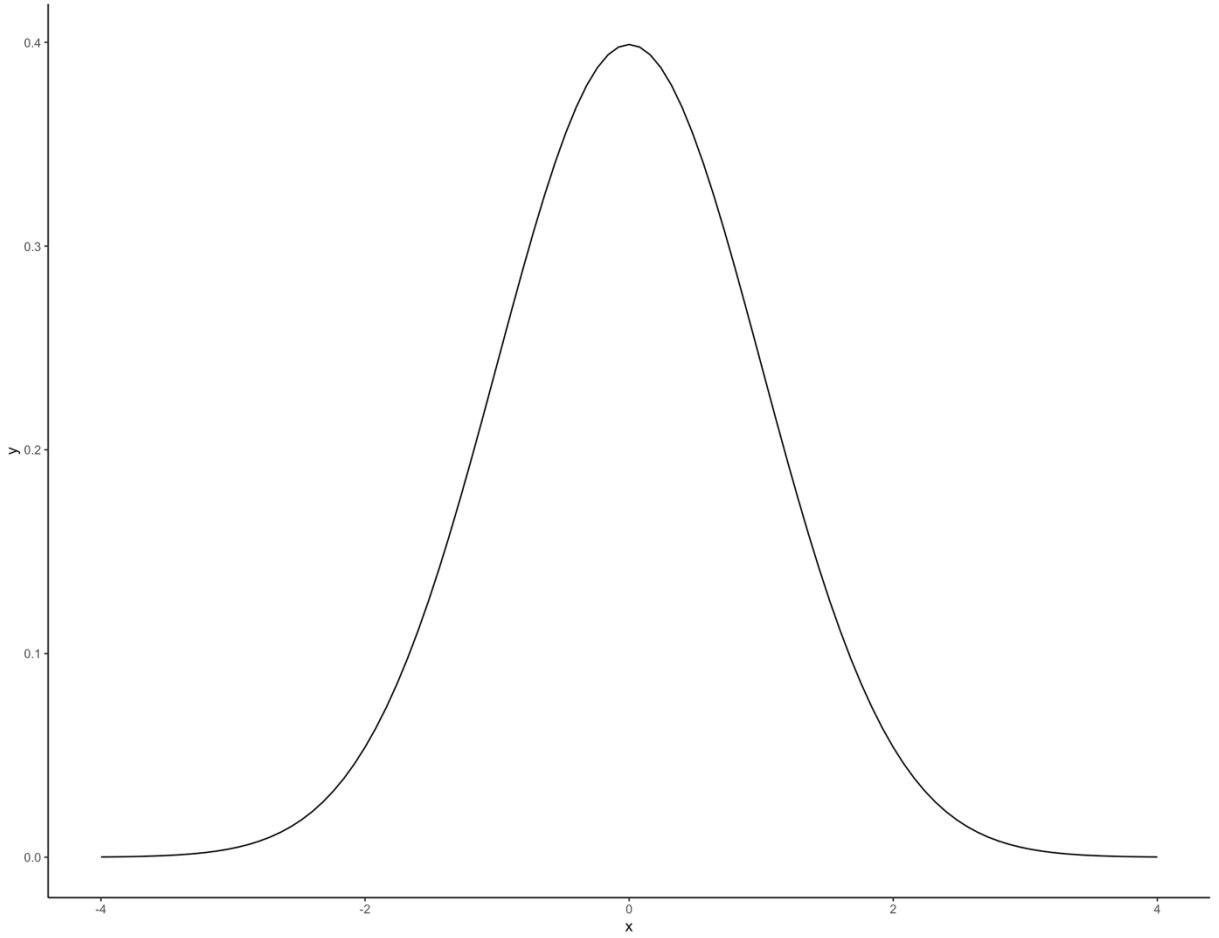
This means that when you sample a random value, the probability of getting one from near the mean is higher than of getting one near the extremes.

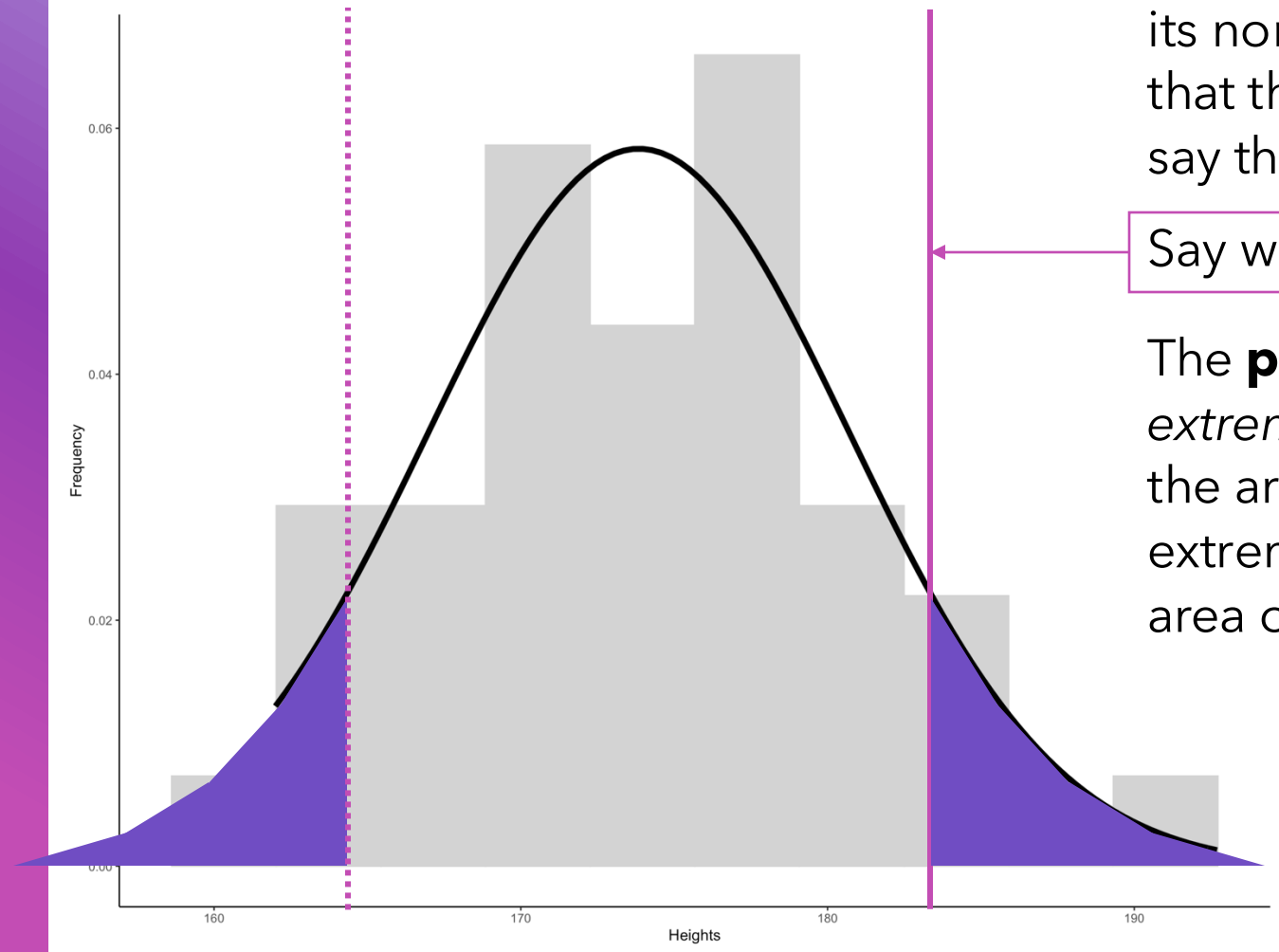
This is called the **Normal Distribution**, and it's found in many, many kinds of data.



The Normal Distribution is often found when we measure a variable that's the result of a complex process with many inputs – like height, income, or vote shares.

Under the **assumption** that the population has a normal distribution, we can estimate the probability that a sample is from that population by calculating the spaces under the curve for the sample measurement.



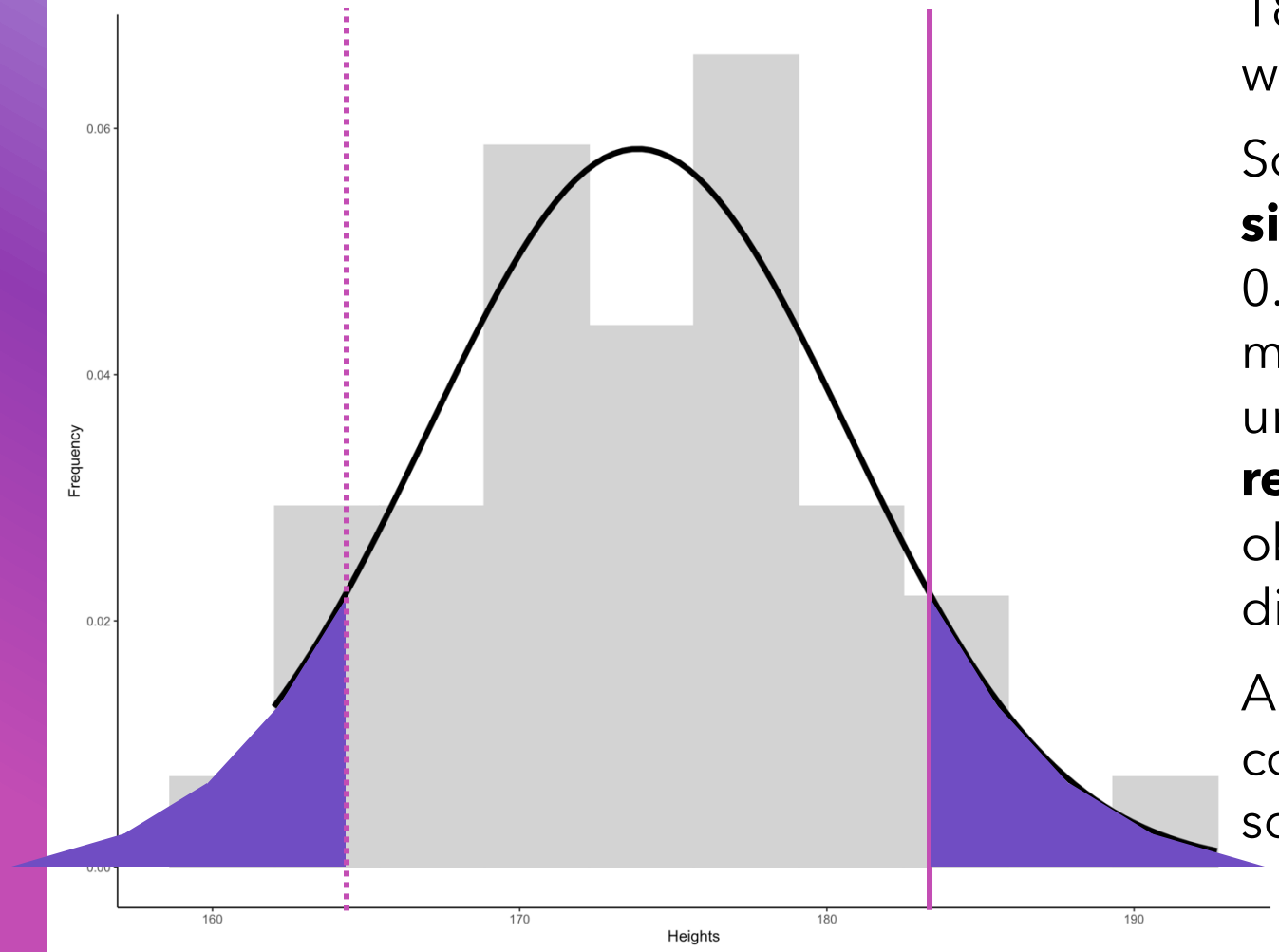


Here's our class population data, with its normal distribution. You can see that the fit doesn't have to be exact to say the data is normally distributed.

Say we measure one sample, at 183cm.

The **probability *p*** of observing a *more extreme* sample in this distribution is the area under the curve towards the extreme end, plus the corresponding area on the other side.

In this case, ***p* = 0.18** – meaning there is an 18% chance of observing a more extreme case.



18% is high: ~1 in 5 observations would be more extreme than 183cm.

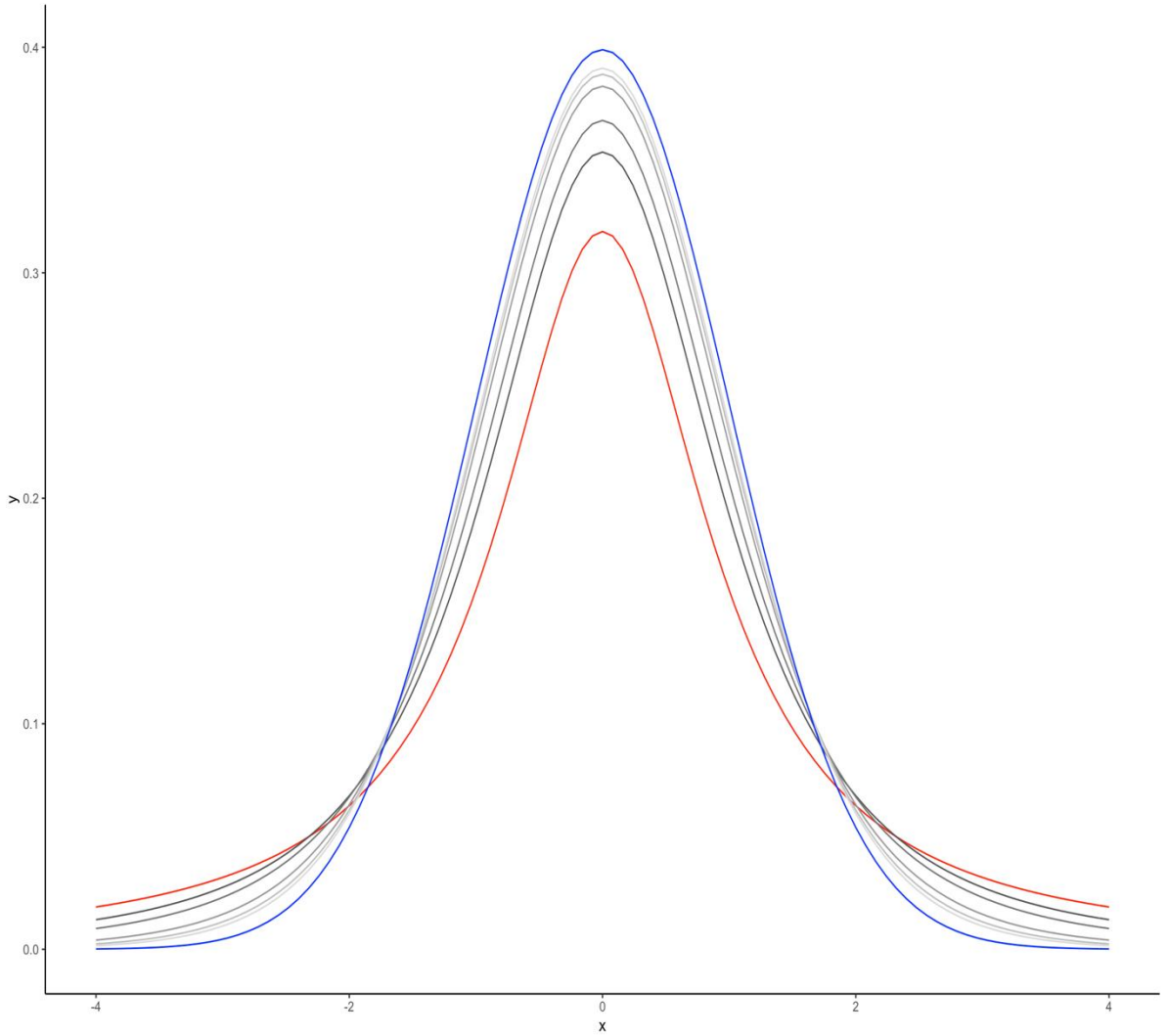
Social scientists usually use a **significance level** of either 0.05 or 0.01 for *p*-values. A value below that means the measurement is highly unusual in this distribution, so we can **reject** the **null hypothesis** that the observation came from the expected distribution.

A higher value – like 0.18 – is likely to come from the expected distribution, so we **retain** the null hypothesis.

So far we've been talking about testing single observations, but we want to see if a set of samples comes from an expected distribution.

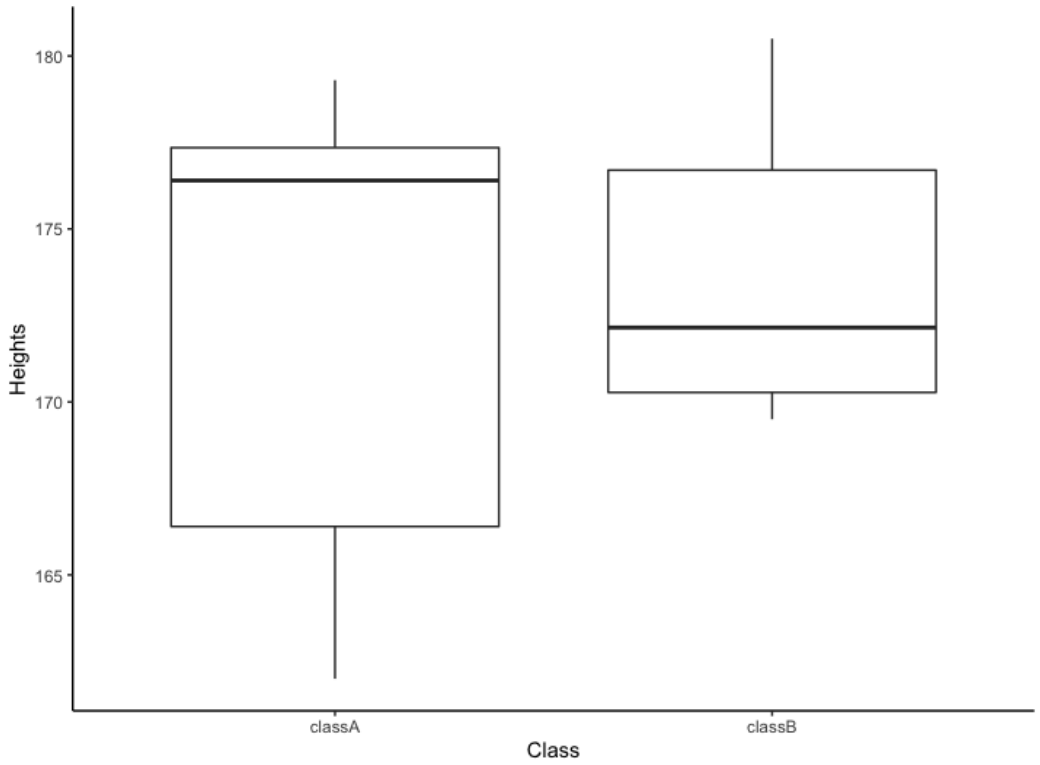
To do this, we use **Student's t test**, which uses the **t distribution.** This is very similar to the normal distribution, but it changes shape depending on how many observations of the variable you have (the **degrees of freedom**).

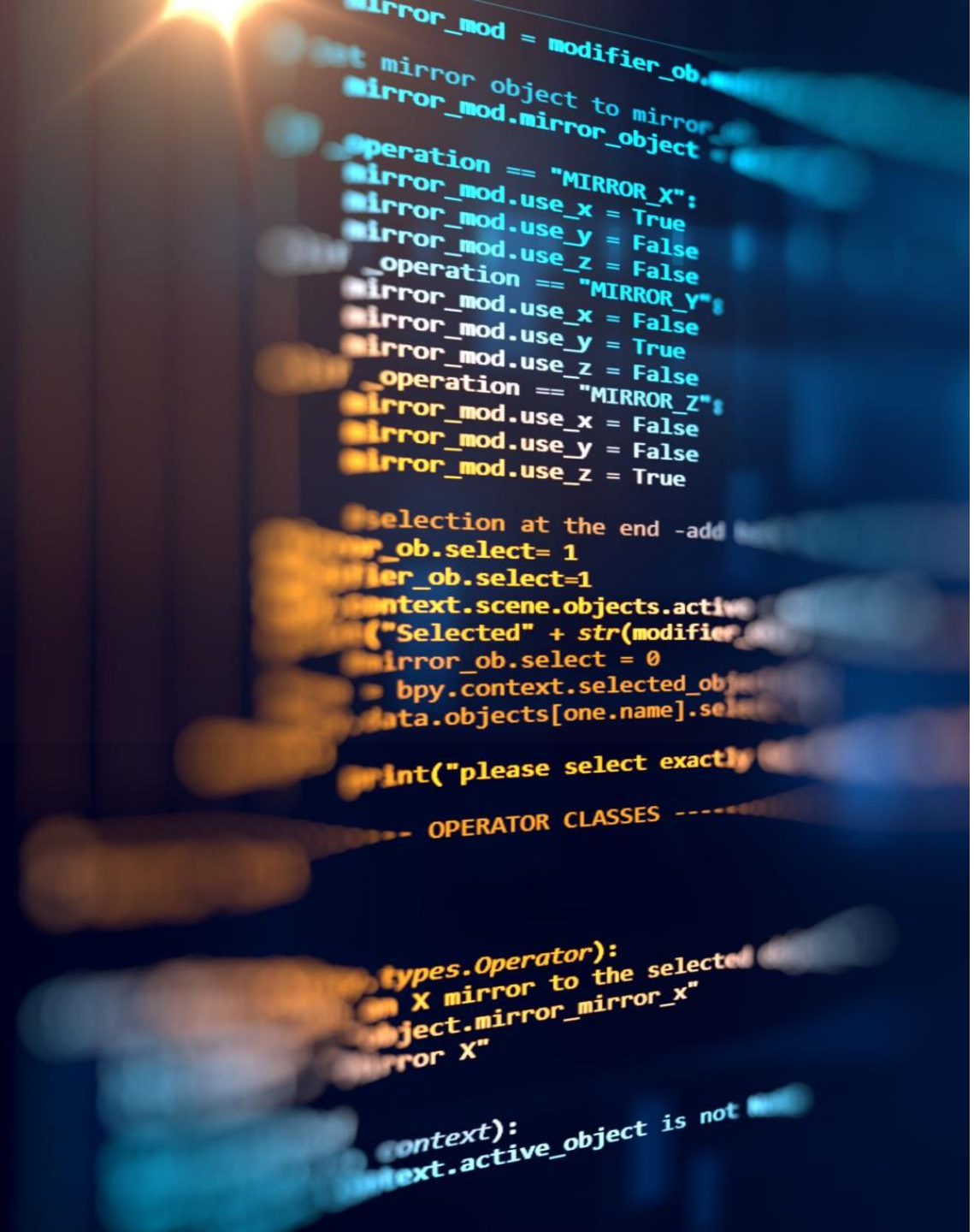
The t distribution is **more uncertain** when you have few observations; as you get more data, it gets more certain and more similar to the normal distribution.



Running t-tests

- Let's go back to our two samples.
- We can use a **two-sample t-test** to find out the probability that there is a real difference between these samples: i.e. whether they come from populations with different distributions.





Move over to RStudio