p-Values and significance levels (false positive rates)

Let's say 123 people in the class toss a coin. Call it "Coin A." There are 65 heads.

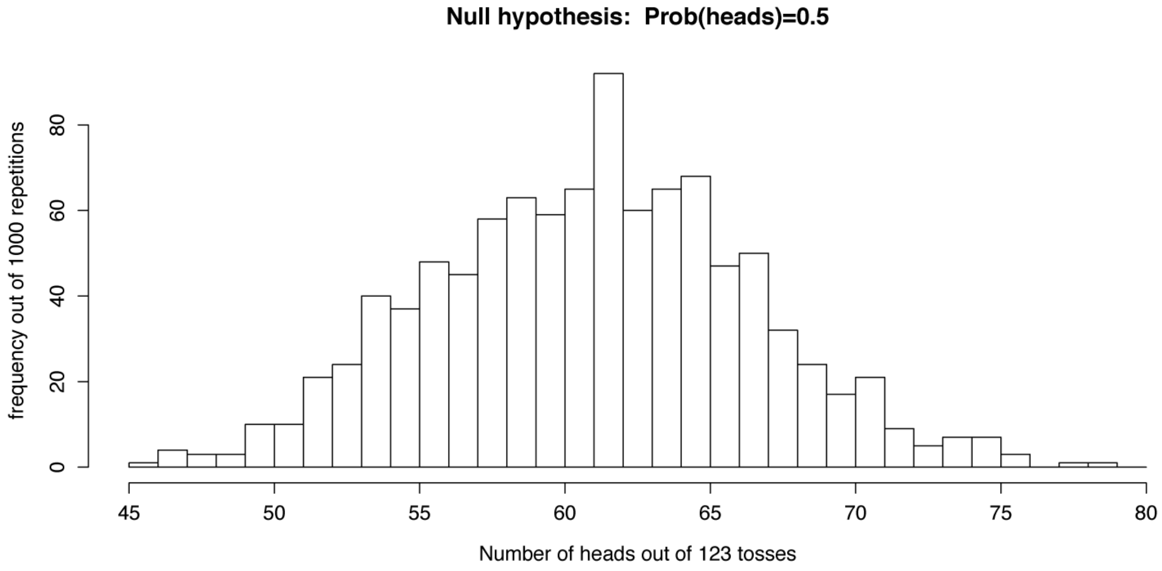
Then they toss another coin. Call it "Coin B." There are 72 heads.

Unbeknownst to the class, one of these coins is biased towards heads and one is fair. We should expect around half heads with the fair coin and more for the biased coin. But of course it is *possible* that we could get 72 heads with a fair coin and it is *possible* to get 65 heads with a biased coin. So knowing this doesn't really tell us which one is which.

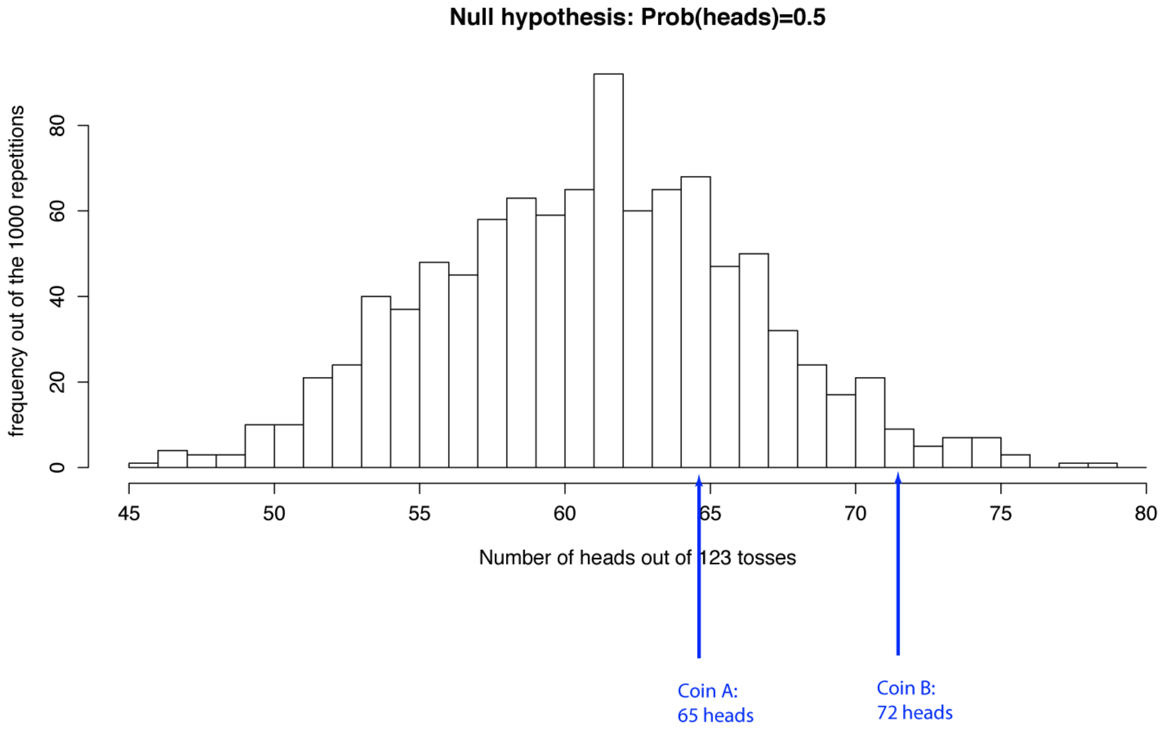
Our null hypothesis is that the coins are fair – this means that the probability of getting heads is 50%.

Our alternative hypothesis is that the coins are biased towards heads. This means that the probability of getting heads is >50%.

We want to see if we can reject the null hypothesis based on the data, so we generate a null distribution. We use the computer to simulate 123 people tossing fair coins and count how many heads there were and repeat this 1000 times. This gives us the following histogram:



Now we plot our data on here:



Should we reject our null hypothesis in each of these cases?

The decision to reject the null hypothesis depends on a cutoff. We need to decide on an acceptable false positive rate, also called a significance level. If the probability of getting our statistic or more extreme is less than that significance level cutoff, then we will reject the null hypothesis.

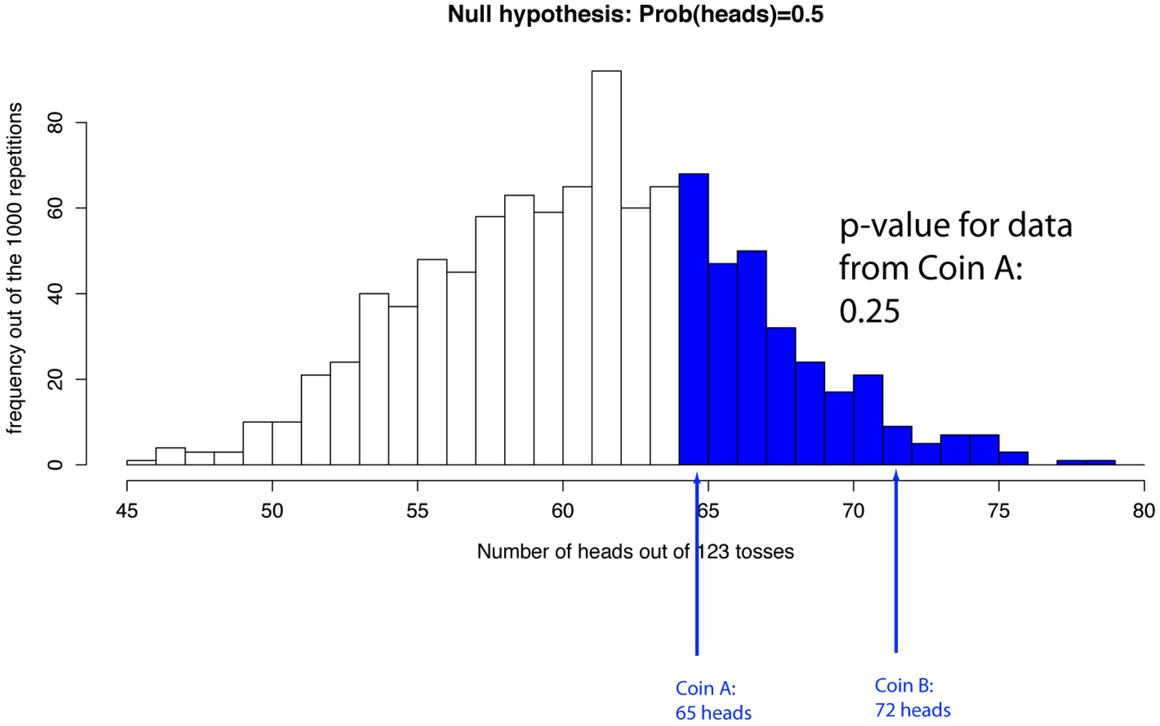
So we need to decide on a significance level cutoff (acceptable false positive rate) and then see if the p-values for our actual data are more or less than this cutoff.

The significance level is the threshold p-value. If the p-value for the data is less than this *significance level*, then we will reject the null hypothesis.

Normally, you would decide on a significance level at which to reject the null hypothesis. There is a convention to have a significance level of 5%, but this is ultimately an arbitrary choice. We want to pick a threshold low enough that we think we would be unlikely to get our data result if the null hypothesis were true.

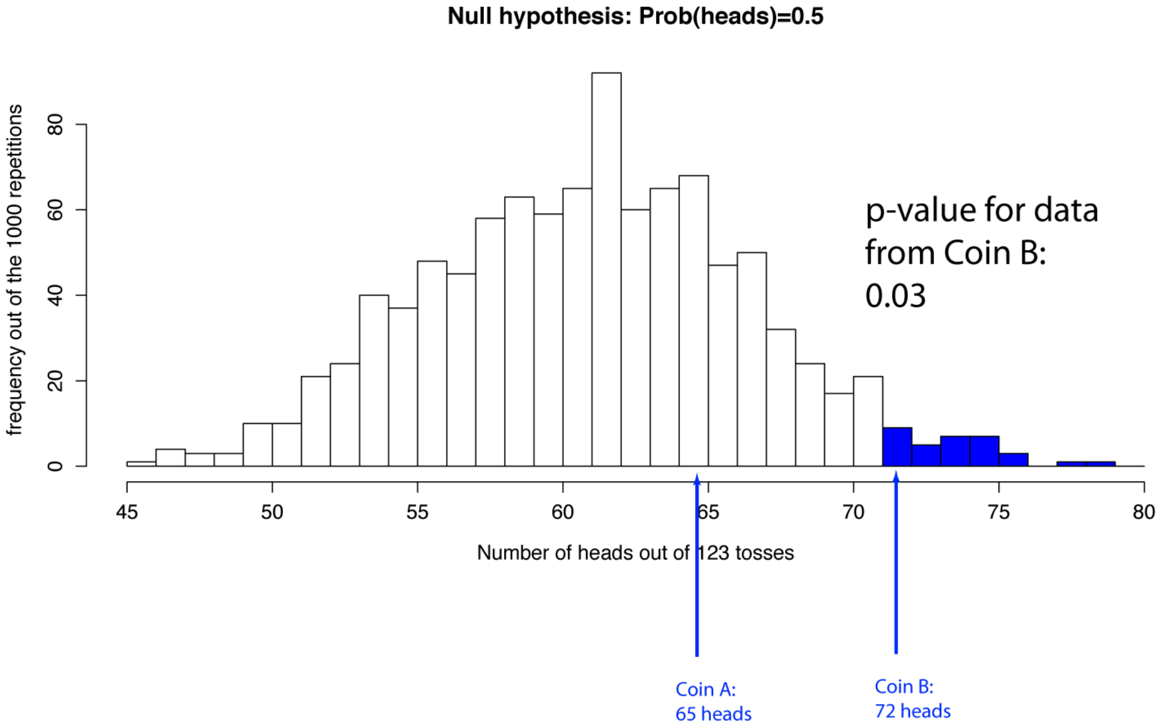
Let's first calculate the p-values for each of our samples.

First, for Coin A:



So if the null hypothesis were true, we would get 65 heads or more 25% of the time.

Now for Coin B:



p-value for data from Coin B: 0.03

If the null hypothesis were true, we would get 72 heads or more 3% of the time.

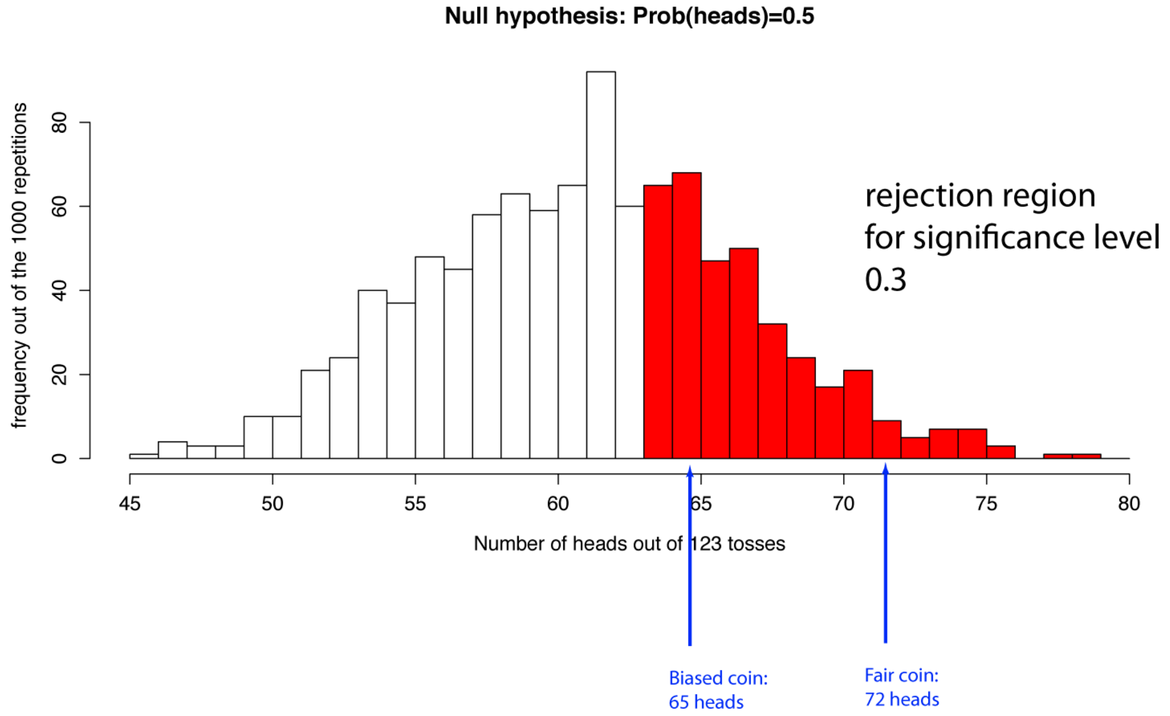
So what to do? We need to pick a significance level. Remember that we don't know which of our coins is fair and which is biased. In fact, the students who flipped them didn't know if either were fair or either were biased! So we need to pick a single significance level that we'll use to test all our samples. In a real situation, the decision of what significance level to use should be made *before* you see your data.

Let's explore the consequences of this choice.

Before we do this, I'll tell you which coin is which...

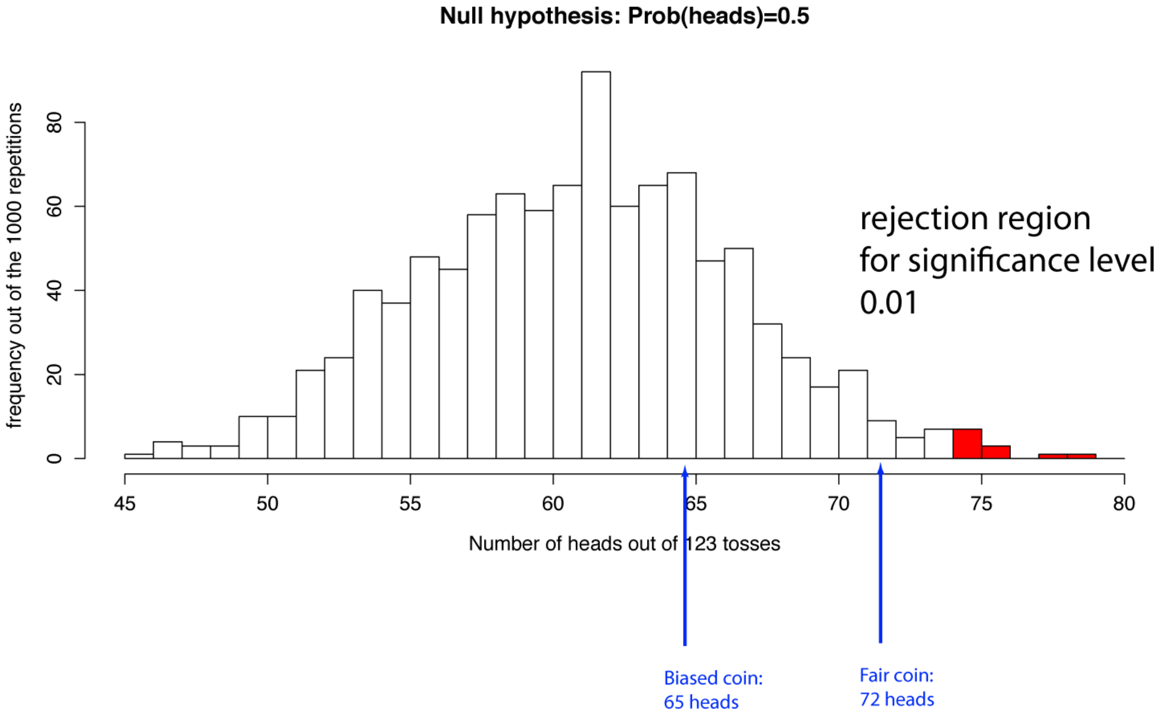
Coin A is biased and Coin B is fair!

Let's say we choose a significance level of 0.3



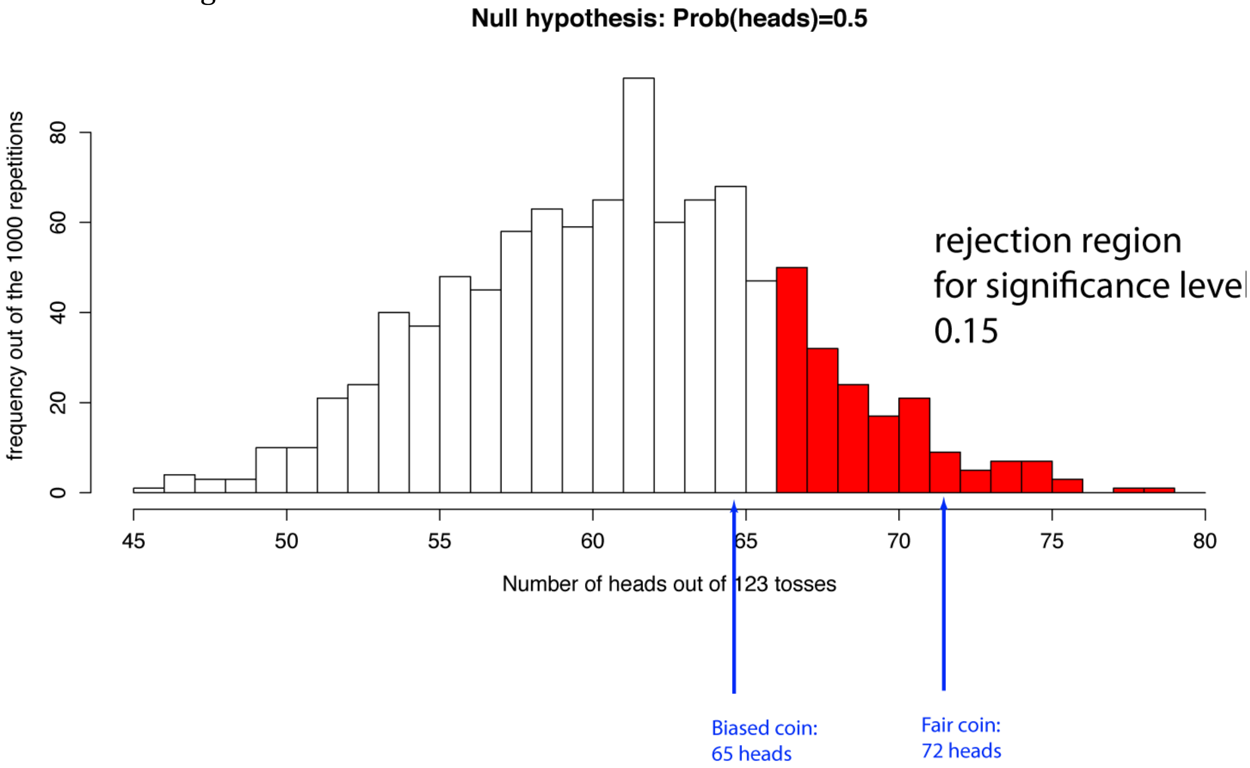
The red region is now the region where we reject our null hypothesis. With this significance level, we would correctly reject the null hypothesis for the biased coin but incorrectly reject it for the fair coin (Type I error).

Let's say we choose significance level of 0.01



Now we correctly fail to reject the null hypothesis for the fair coin, but we also fail to reject it for the biased coin (Type II error).

How about a significance level of 0.15?



Here we are wrong about both coins! We fail to reject for the biased coin (Type II) and reject for the fair coin (Type I).

Ultimately, when we are doing science, we need to make a statement about our data, so we have to pick a single significance level.

Let's choose the standard 0.05 significance level cutoff. Then we would make a statement like this:

We did two experiments with two different coins In the first, we got 65 heads and in the second we got 72 heads. The p-value for 65 heads is 0.25. The p-value for 72 heads is 0.03. We chose a significance level (a p-value *cutoff*) of 0.05. Based on this significance level, we fail to reject the null hypothesis for the coin that gave us 65 heads. We reject the null hypothesis for the coin that gave us 72 heads. Therefore, our data is consistent with the coin that gave us 65 heads being fair and consistent with the coin that gave us 72 heads being biased.

When we do a real experiment, we wouldn't know whether the coins were fair or biased. So we could get unlucky (as I set us up to be here) and just be completely wrong! Setting a significance level lets us control the tradeof between false alarms (false positives, Type I errors) and missed opportunities (false negatives, Type II errors).