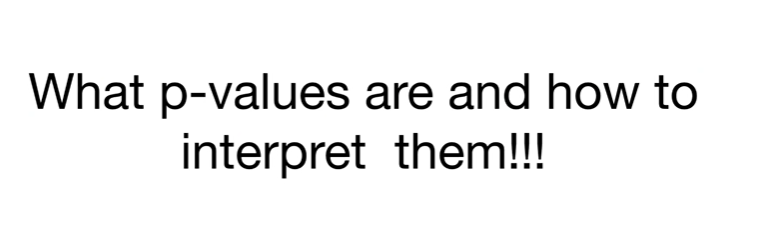
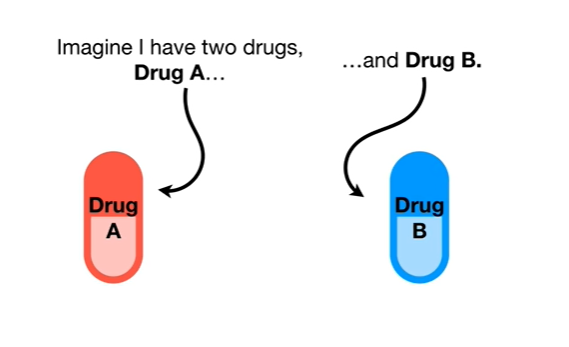
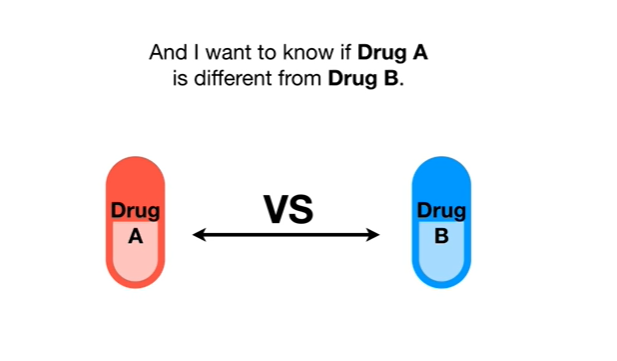
<https://www.youtube.com/watch?v=vemZtEM63GY&list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9&index=10>



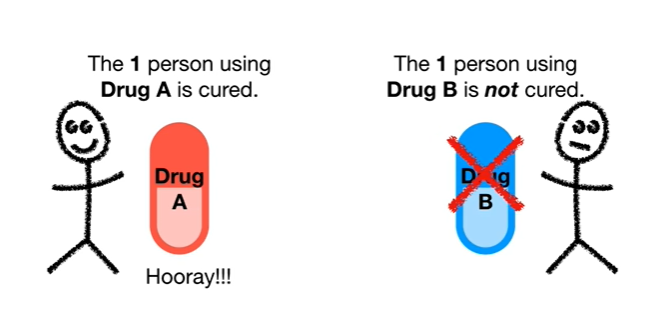
Today we're gonna talk about what p-values are and how to interpret them.



Imagine I have two drugs drug a and drug B.



And I want to know if drug a is different from drug B.

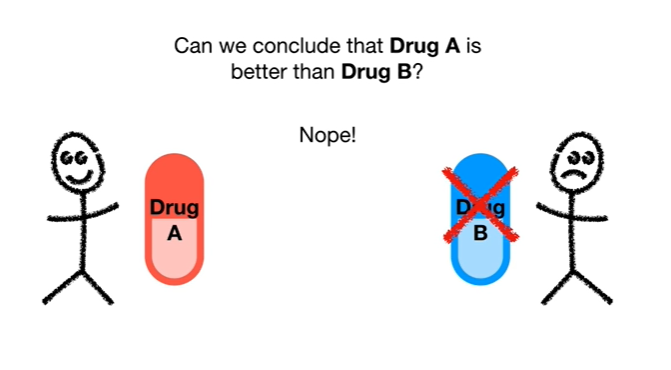


So I give one person drug a and I give one other person drug B the one person using drug a is cured.

Hooray.

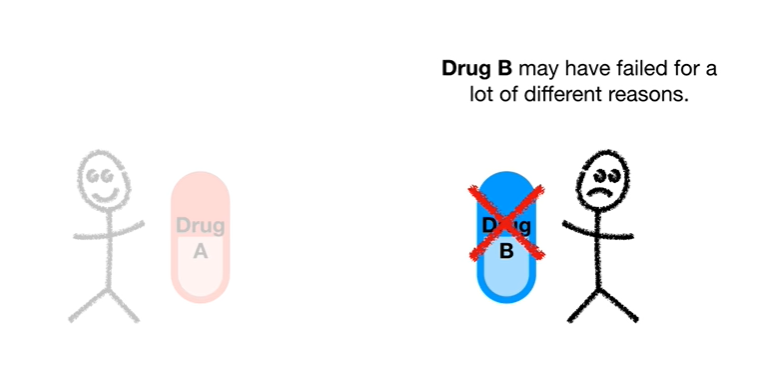
The one person using drug B is not cured.

Bummer.

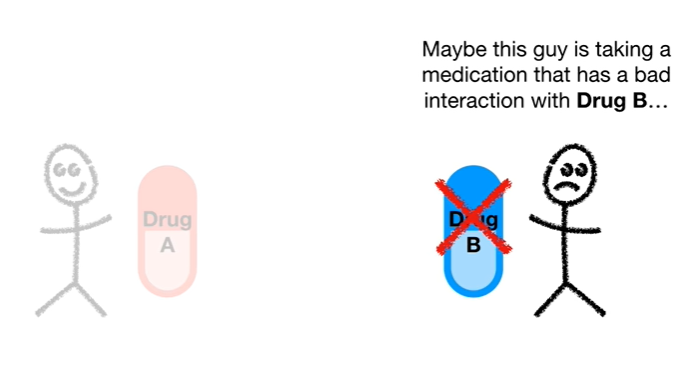


Can we conclude that drug a is better than drug B ?

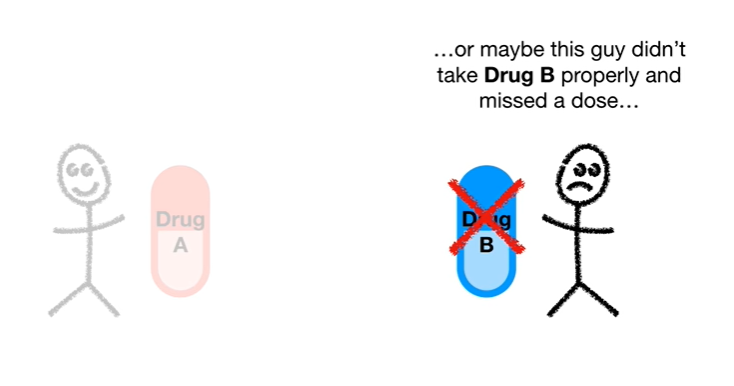
Nope !



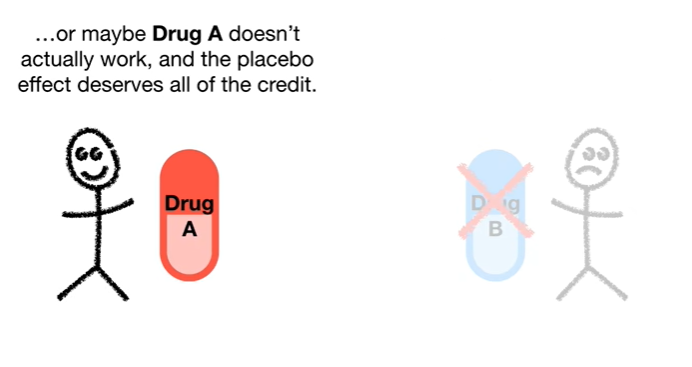
Drug B may have failed for a lot of different reasons.



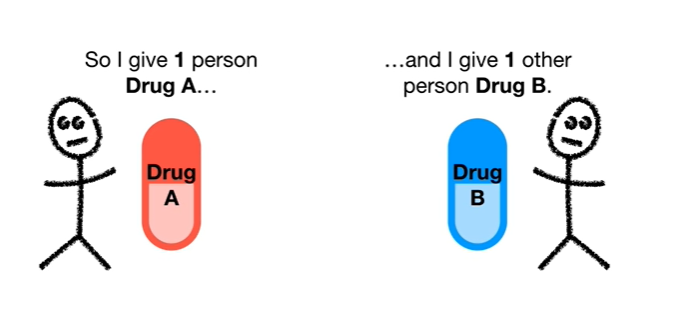
Maybe this guy is taking a medication that has a bad interaction with drug B

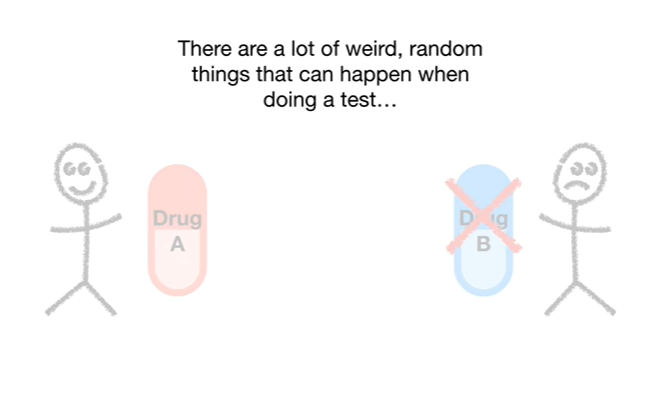


or maybe this guy has a rare allergy to drug B or maybe this guy didn't take drug B properly and missed a dose

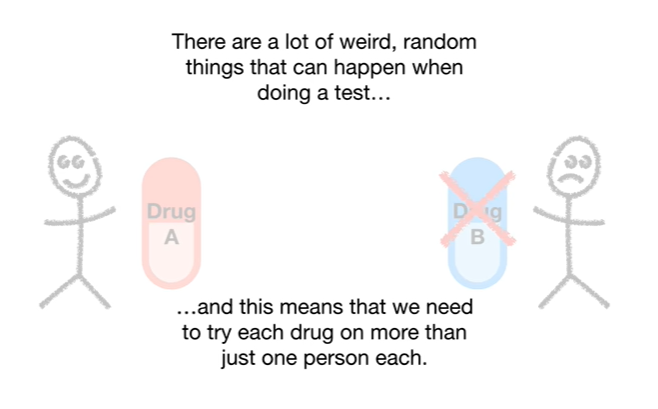


or maybe drug A doesn't actually work and the placebo effect deserves all of the credit.

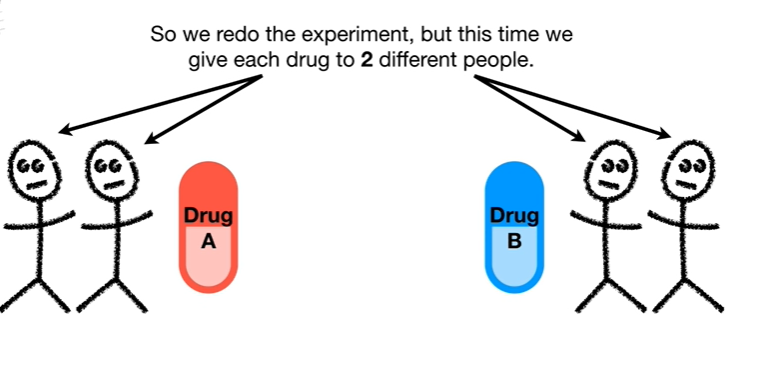




There are a lot of weird random things that can happen when doing a test



and this means that we need to try each drug on more than just one person each.

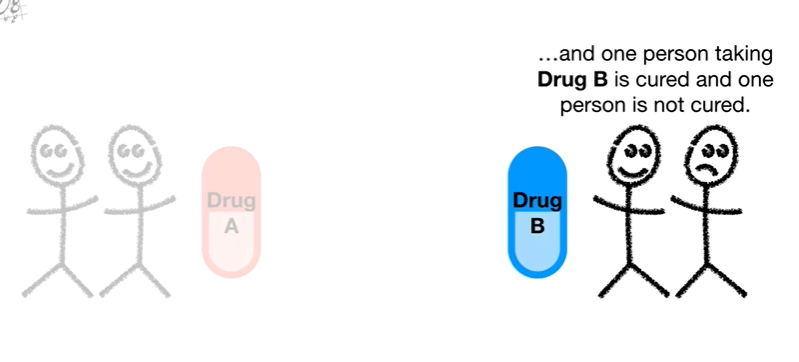


So we redo the experiment but this time we give each drug to two different people.



This time both people taking drug a are cured.

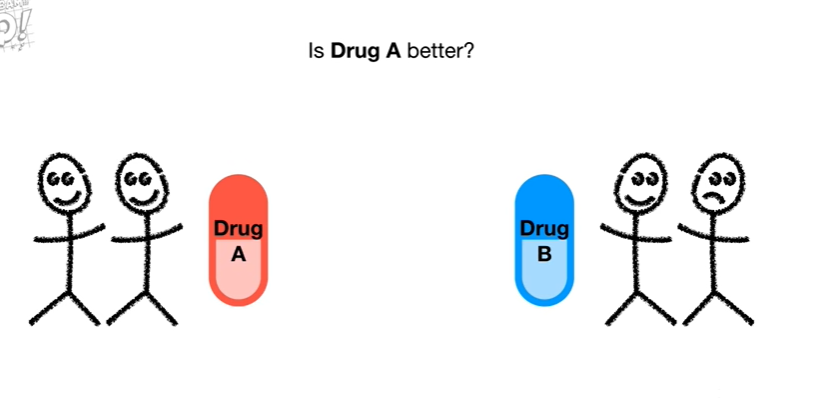
Hooray !!!



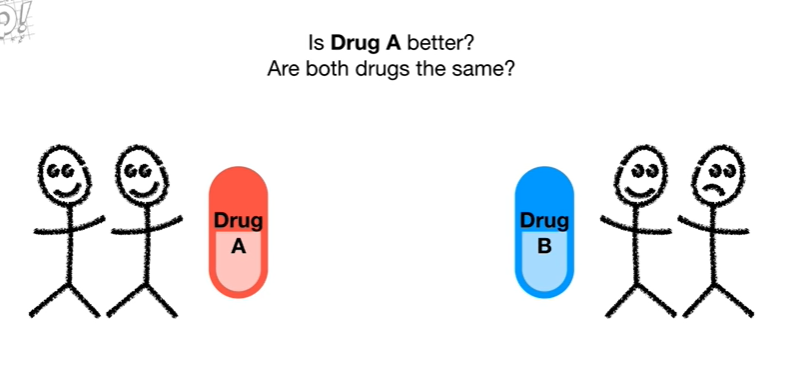
And one person taking drug B is cured and one person is not cured.

Hooray !!!

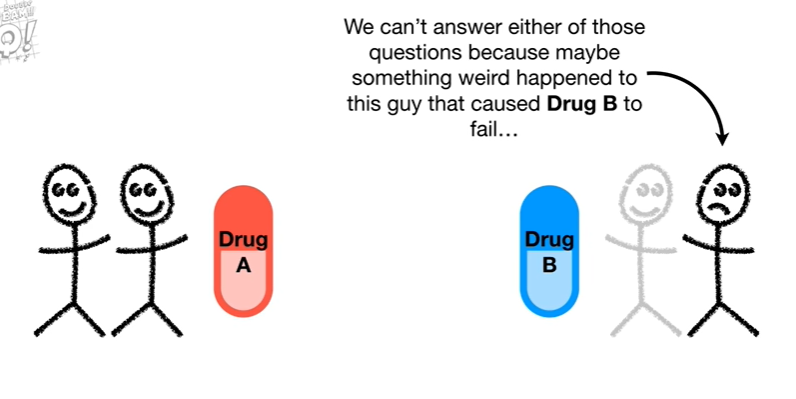
And bummer !!



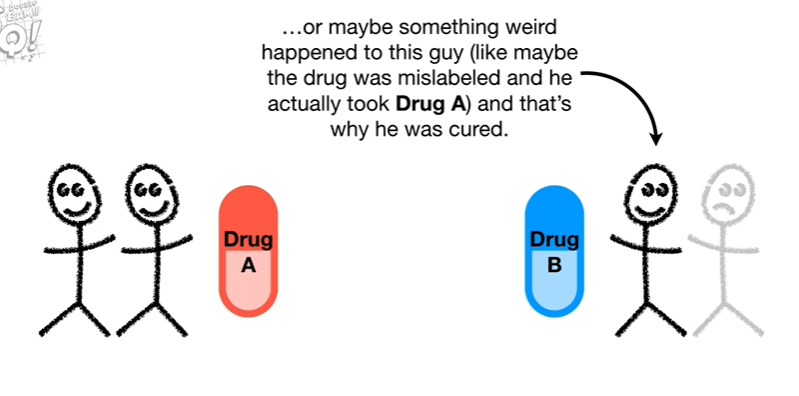
Is drug A better ?



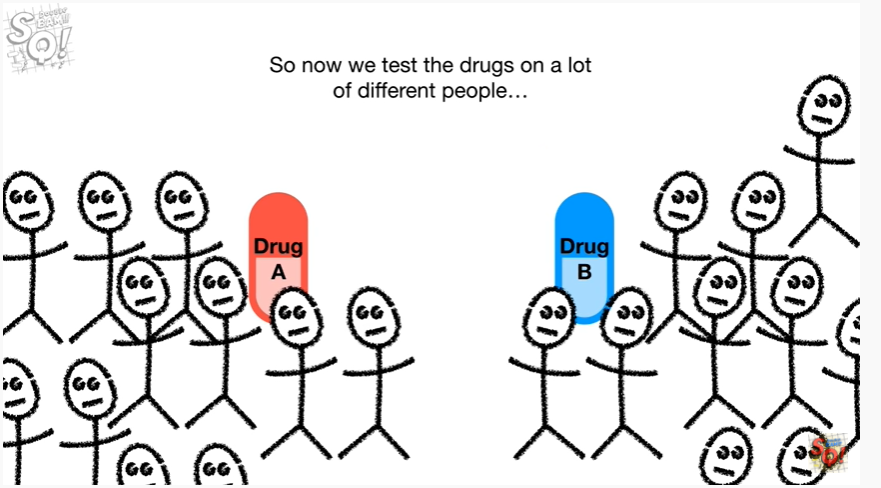
Are both drugs the same ?



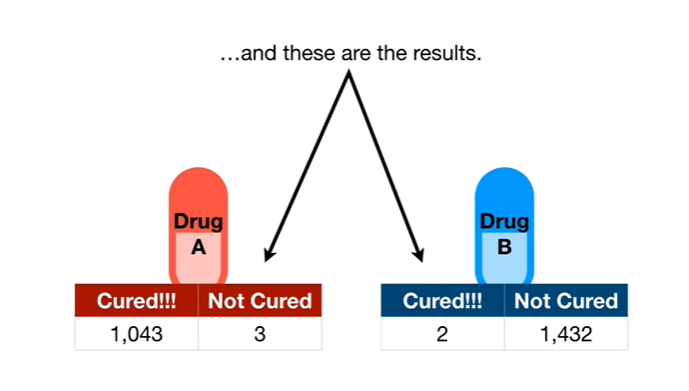
We can't answer either of those questions because maybe something weird happened to this guy that caused drug B to fail



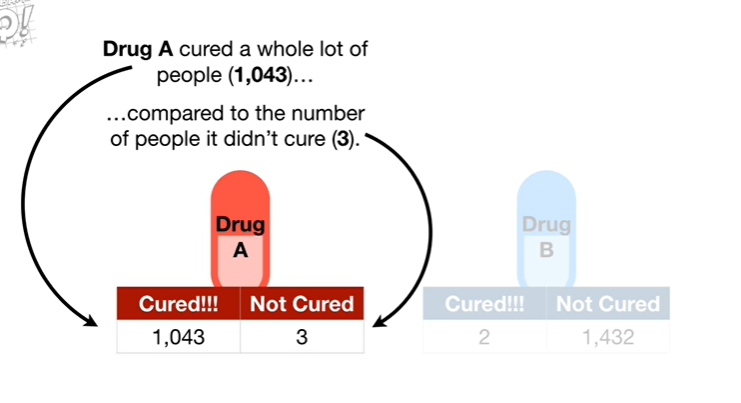
or maybe something weird happened to this guy like maybe the drug was mislabeled and he actually took drug a and that's why he was cured ;



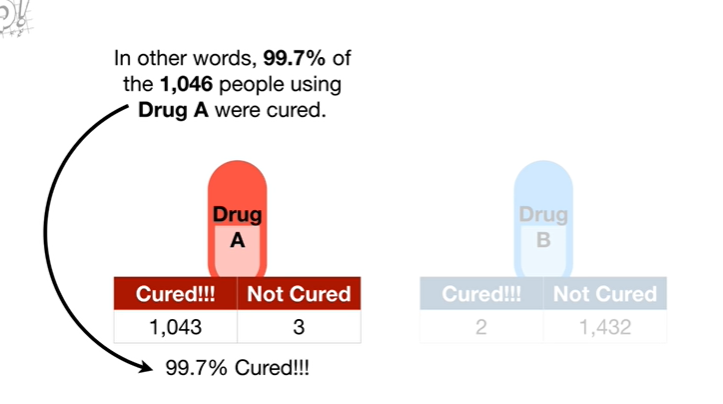
So now we test the drugs on a lot of different people



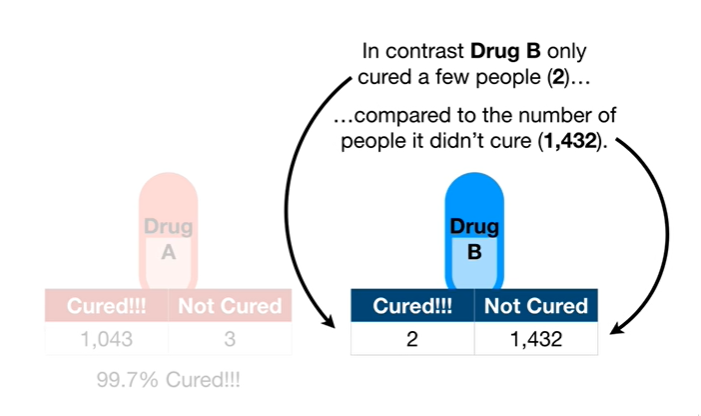
and these are the results.



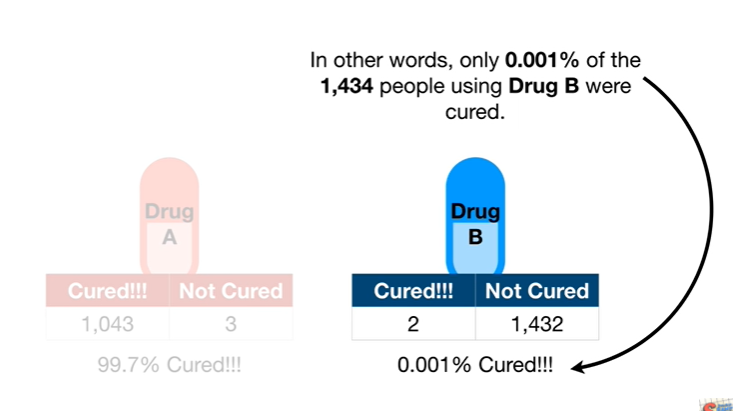
Drug A cured a whole lot of people 1043 compared to the number of people it didn't cure 3.



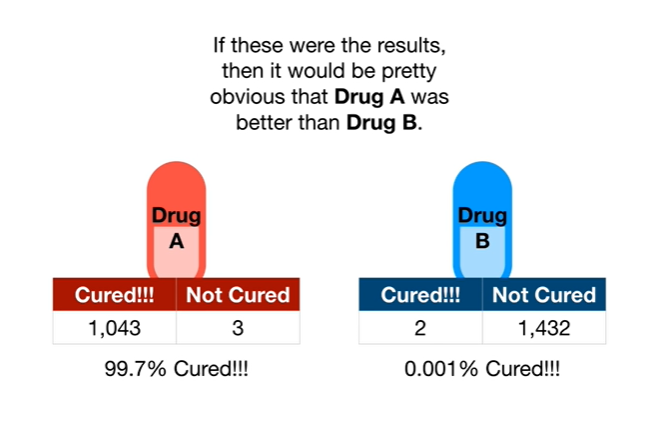
In other words 99.7% of the 1046 people using drug a were cured.



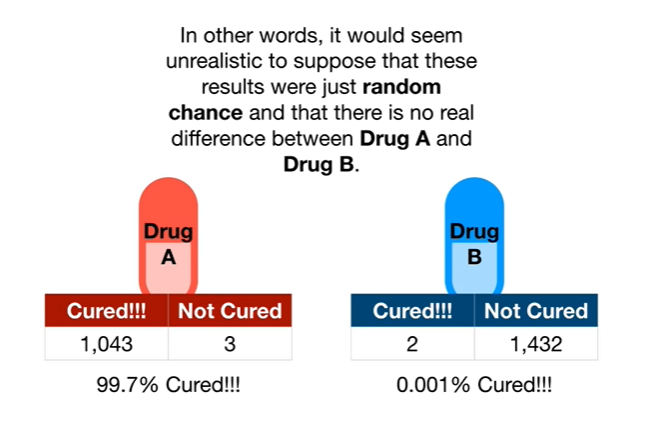
In contrast drug B only cured a few people - compared to the number of people it didn't cure 1432.



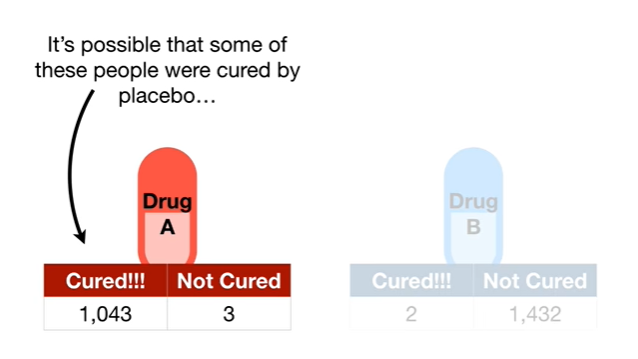
In other words only 0.001 percent of the 1434 people using drug B were cured.



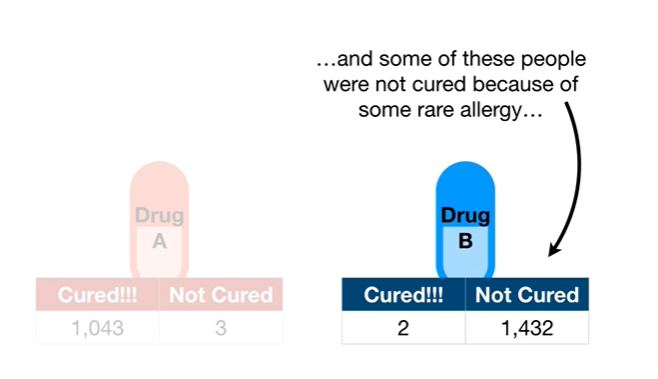
If these were the results then it would be pretty obvious that drug a was better than drug B.



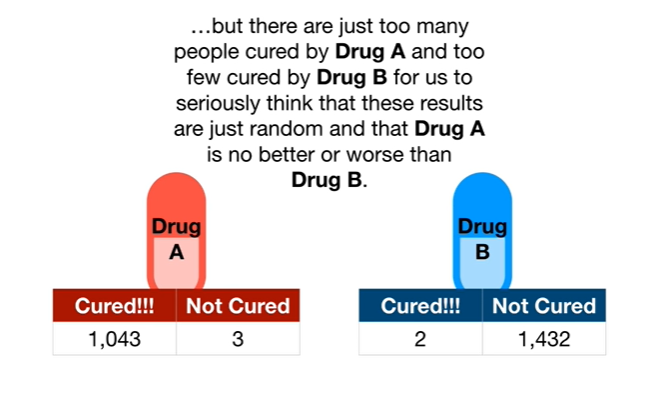
In other words it would seem unrealistic to suppose that these results were just random chance and that there is no real difference between drug A and drug B.



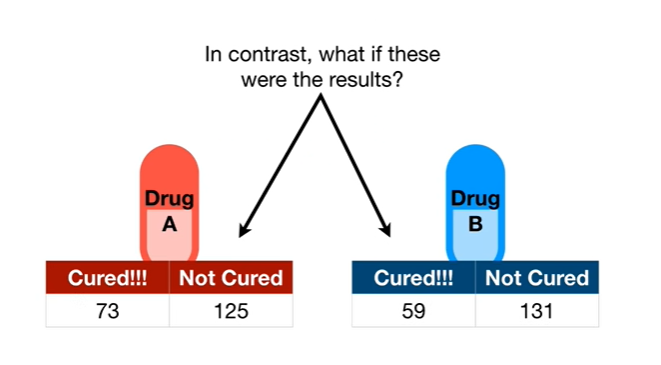
It's possible that some of these people were cured by placebo



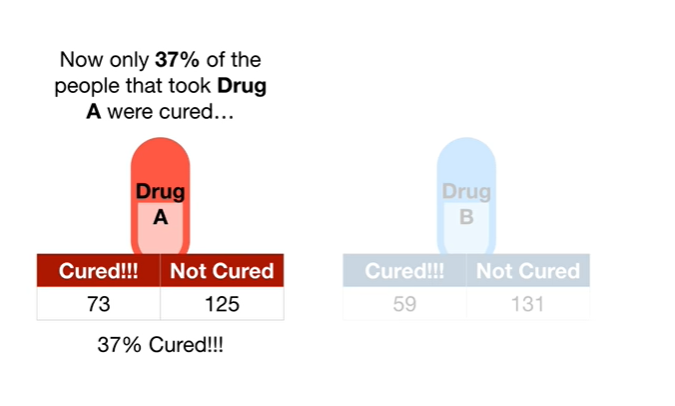
and some of these people were not cured because of some rare allergy



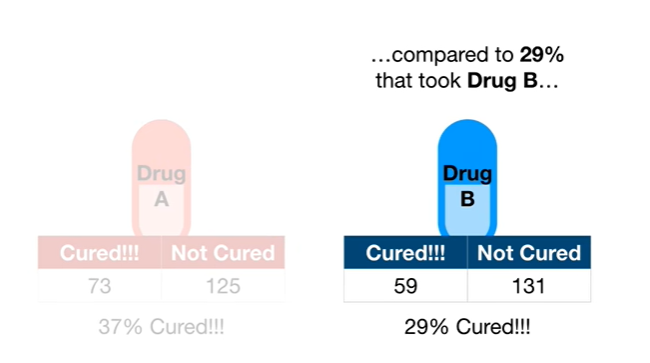
but they are just too many people cured by drug a and too few cured by drug B for us to seriously think that these results are just random and that drug A is no better or worse than drug B.



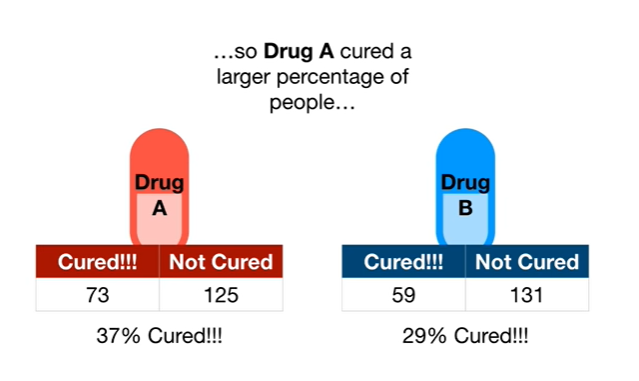
In contrast what if these were the results ?



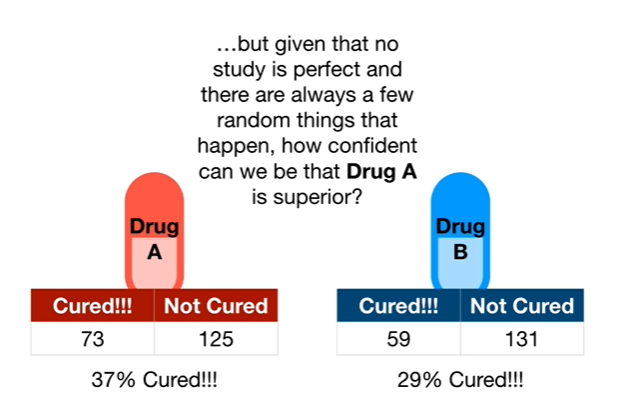
Now only 37% of the people that took drug a were cured



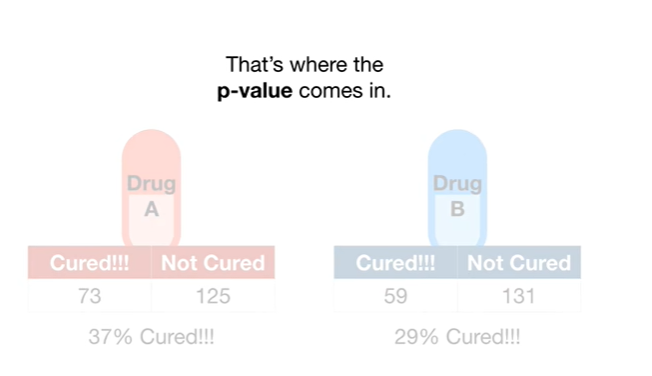
compared to 29% that took drug B



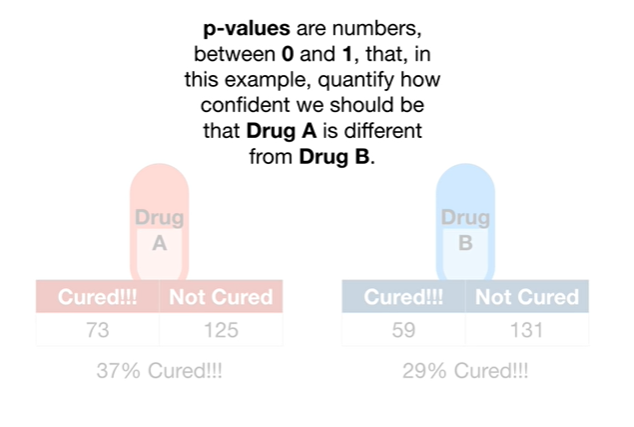
so drug A cured a larger percentage of people



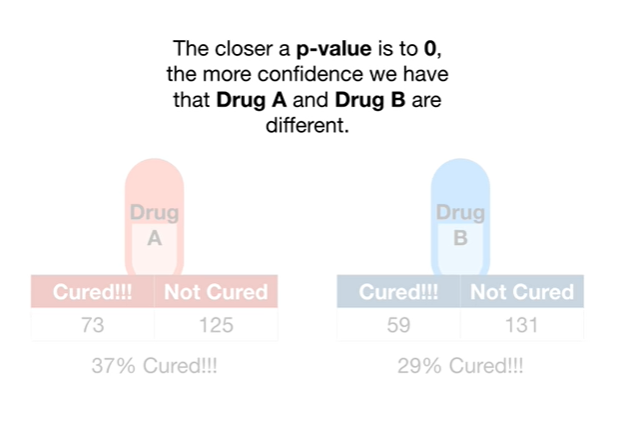
but given that no study is perfect and there are always a few random things that happen how confident can we be that drug a is superior ?



That's where the p-value comes in.



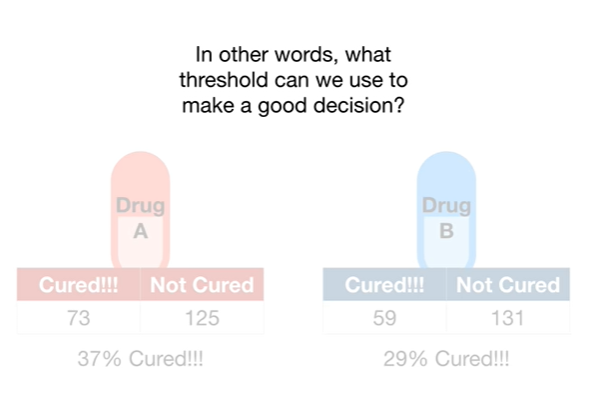
p-values are numbers between zero and one that in this example quantify how confident we should be that drug A is different from drug B.



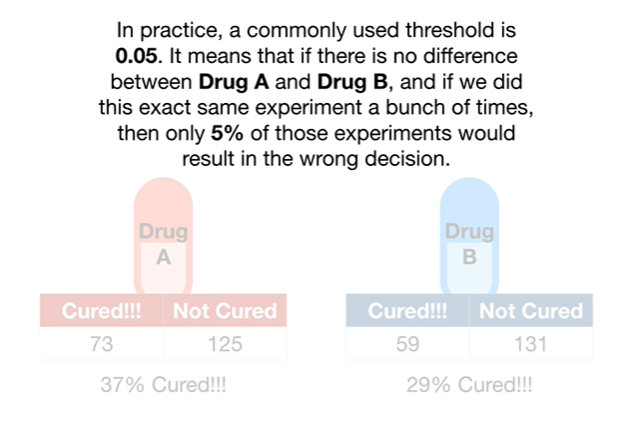
The closer a p-value is to zero, the more confidence we have the drug A and drug B are different.



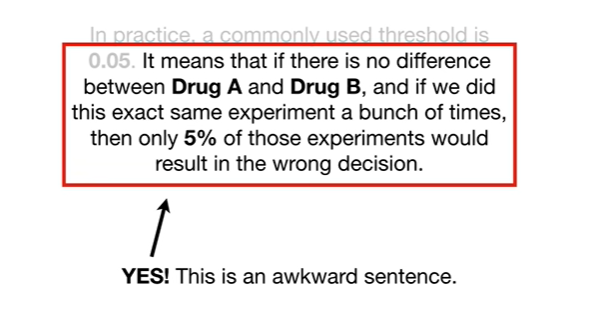
So the question is how small does a p-value have to be before we are sufficiently confident that drug A is different from drug B ?



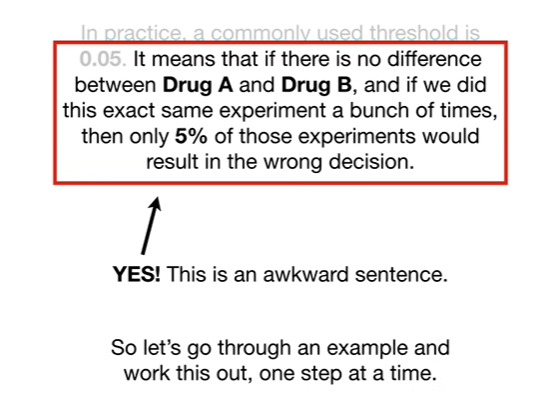
In other words what threshold can we use to make a good decision ?



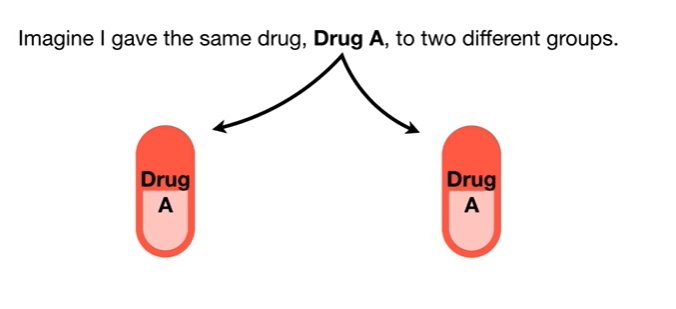
In practice a commonly used threshold is 0.05 it means that if there is no difference between drug a and drug B and if we did this exact same experiment a bunch of times that only 5% of those experiments would result in the wrong decision.



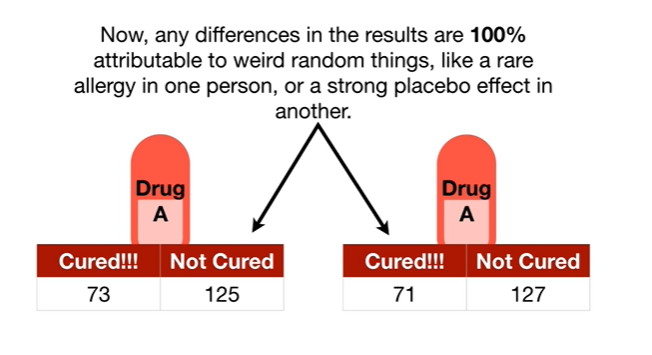
Yes ! This is an awkward sentence.



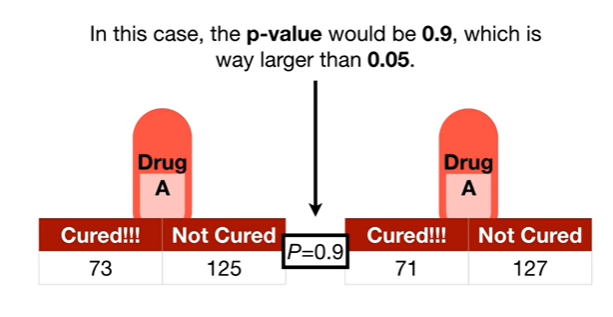
So let's go through an example and work this out one step at a time.



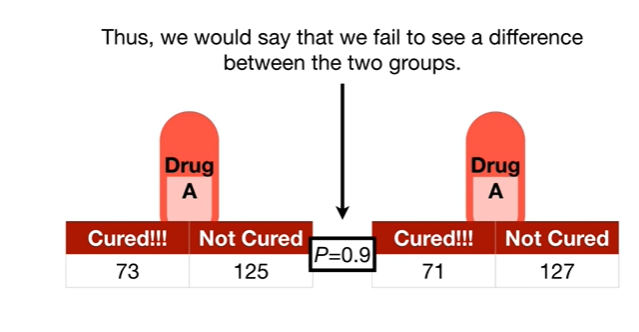
Imagine I gave the same drug drug A to two different groups.



Now any differences in the results are 100% attributable to weird random things like a rare allergy in one person or a strong placebo effect in another.



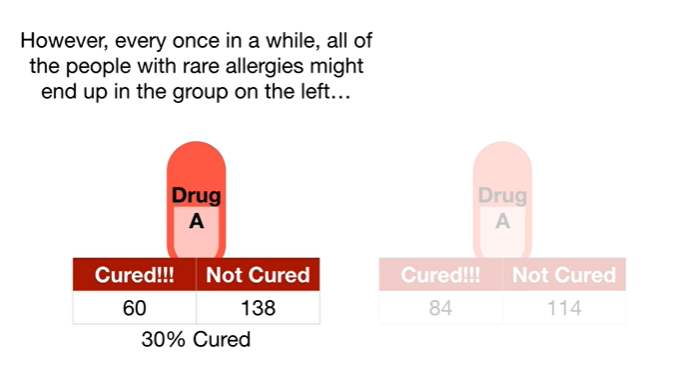
In this case the p-value would be 0.9 which is way larger than 0.05.



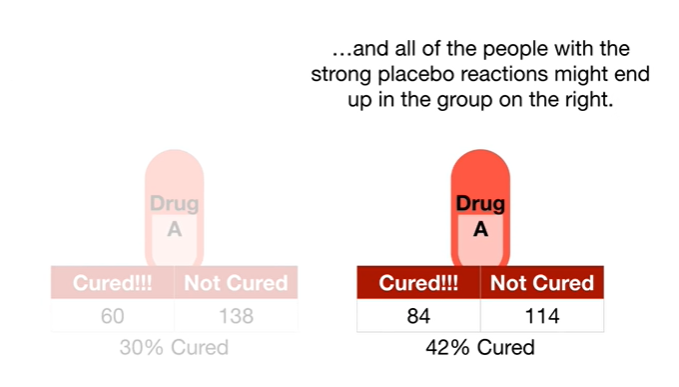
Thus we would say that we fail to see a difference between the two groups.



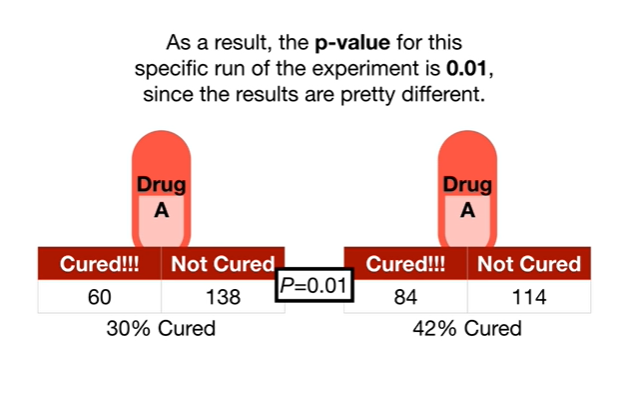
If we repeated this same experiment a lot of times most of the time we would get similarly large p-values.



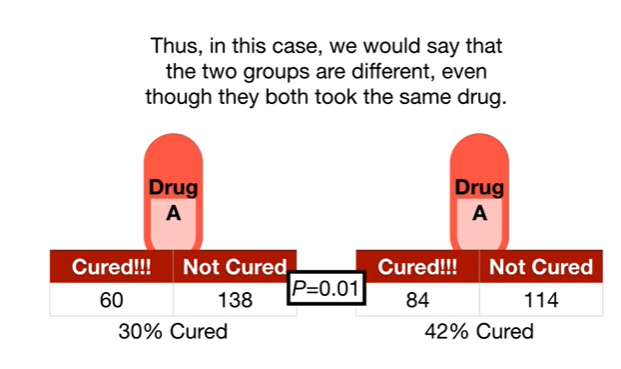
However every once in a while all of the people with rare allergies might end up in the group on the left



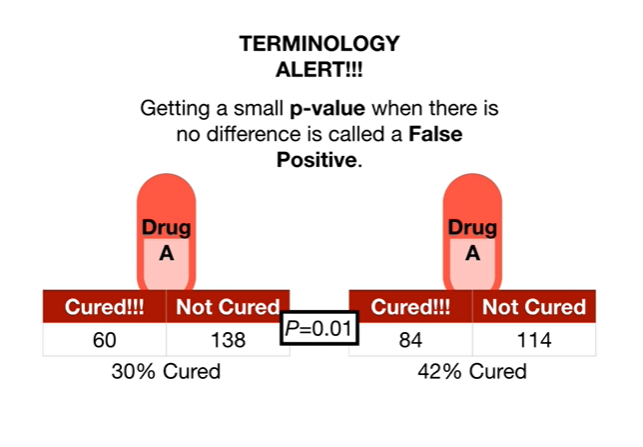
and all of the people with the strong placebo reactions might end up in the group on the right.



As a result the p-value for this specific run of the experiment is 0.01 since the results are pretty different.

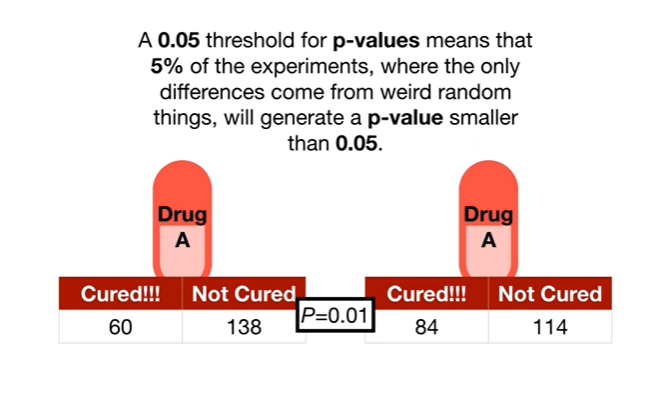


Thus, in this case, we would say that the two groups are different even though they both took the same drug.

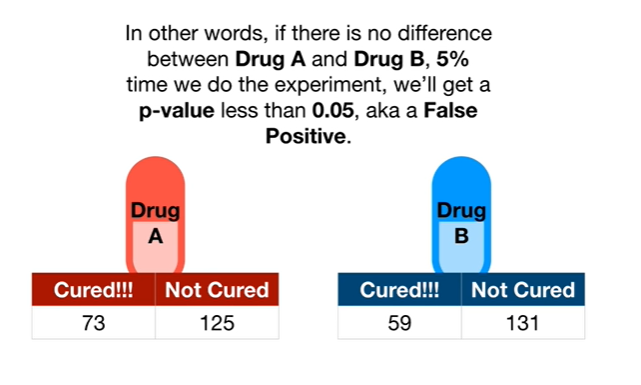


Oh no it's the dreaded terminology alert.

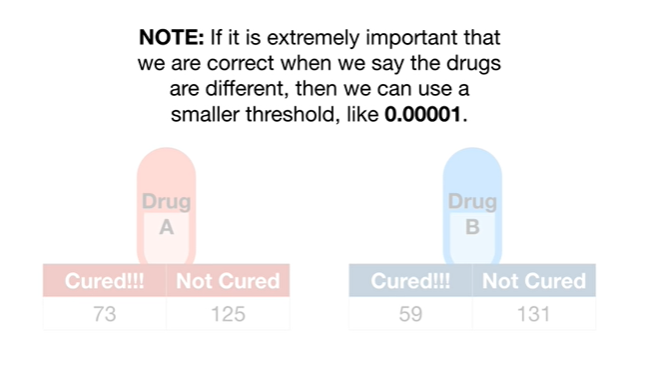
Getting a small p-value when there is no difference is called a false positive.



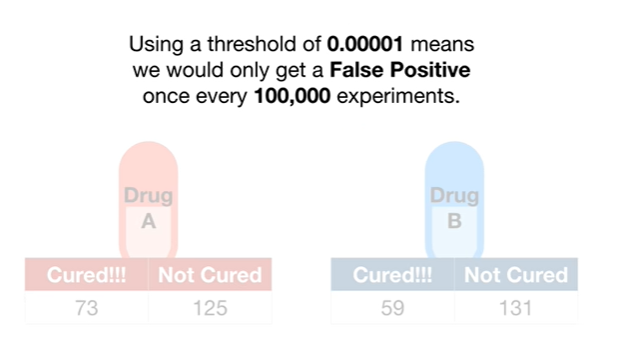
A 0.05 threshold for p-values means that 5% of the experiments where the only differences come from weird random things will generate a p-value smaller than 0.05.



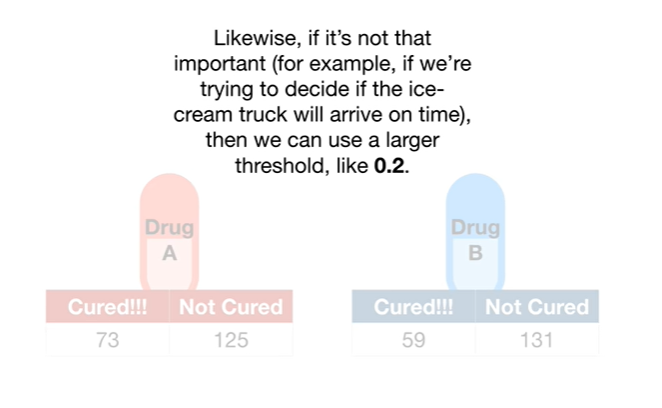
In other words if there is no difference between drug a and drug B 5% of the time we do the experiment we will get a p-value less than 0.05 aka a false positive.



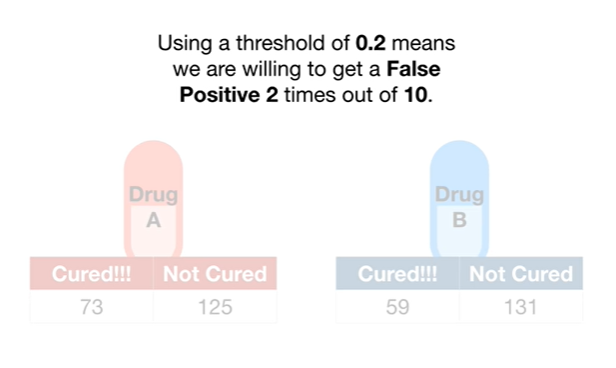
Note : if it is extremely important that we are correct when we say the drugs are different then we can use a smaller threshold like zero point zero zero zero zero one.



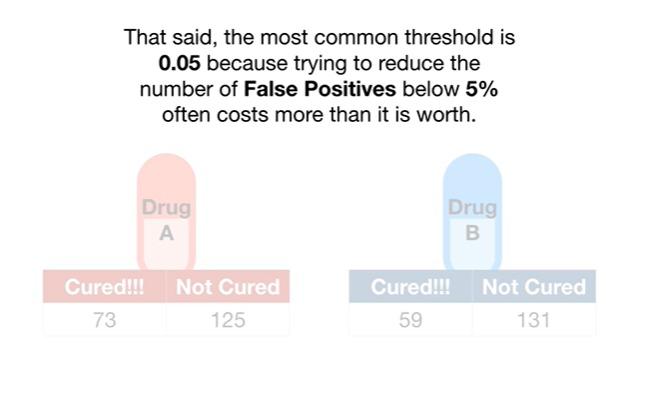
Using a threshold of 0.00001 means we would only get a false positive once every 100,000 experiments.



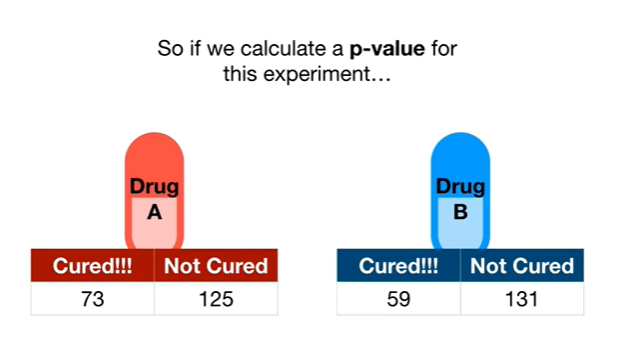
Likewise if it's not that important for example if we're trying to decide if the ice cream truck will arrive on time then we can use a larger threshold like 0.2.



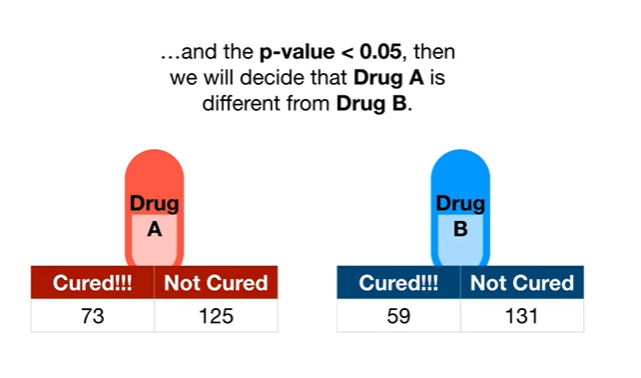
Using a threshold of 0.2 means we are willing to get a false positive two times out of ten.



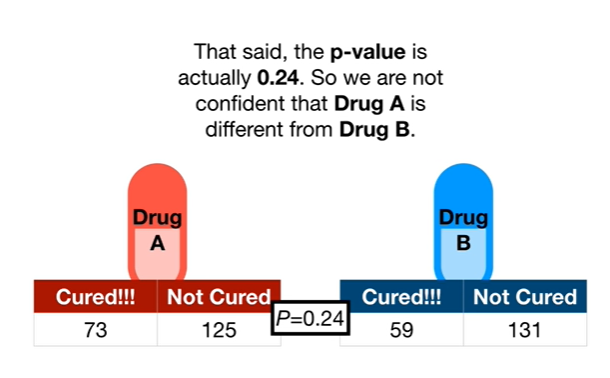
That said the most common threshold is 0.05 because trying to reduce the number of false positives below 5% often costs more than it's worth.



So if we calculate a p-value for this experiment



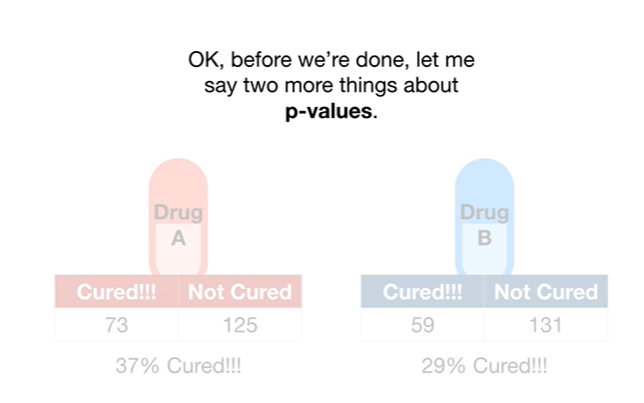
and the p-value is less than 0.05 then we will decide that drug A is different from drug B.



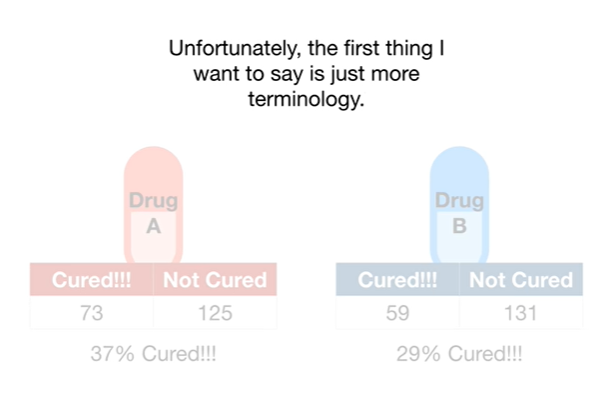
That said, the p-value is actually zero point two four.

So we are not confident that drug A is different from drug B.

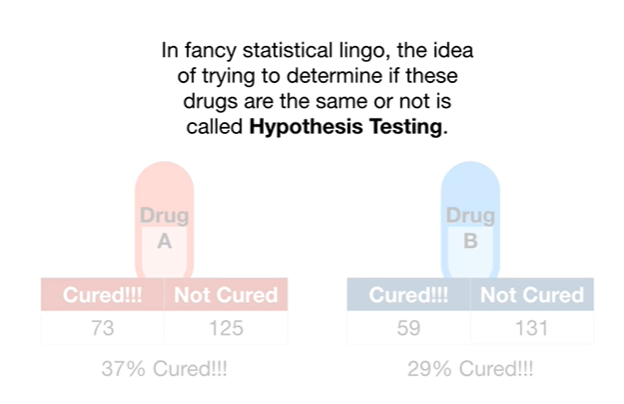
BAM !!!



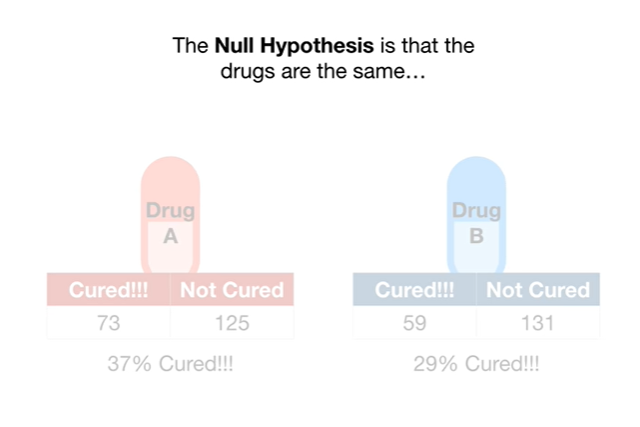
Okay before we're done let me say two more things about p-values.



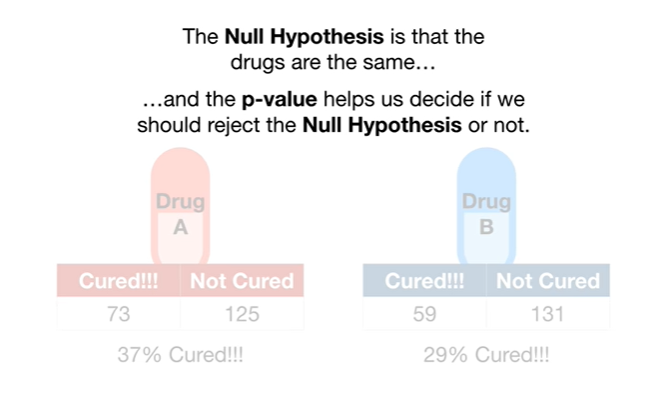
Unfortunately the first thing I want to say is just more terminology.



In fancy statistical lingo the idea of trying to determine if these drugs are the same or not is called hypothesis testing.

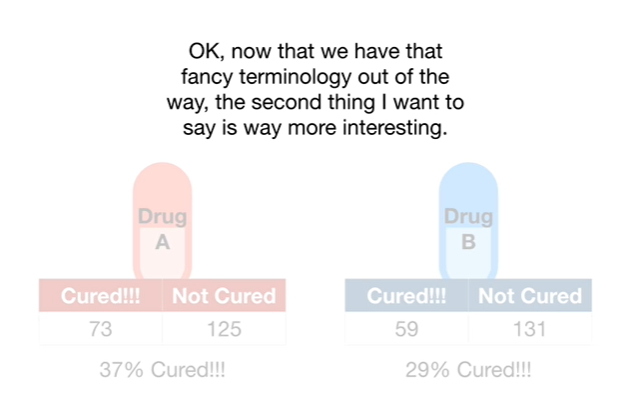


The null hypothesis is that the drugs are the same

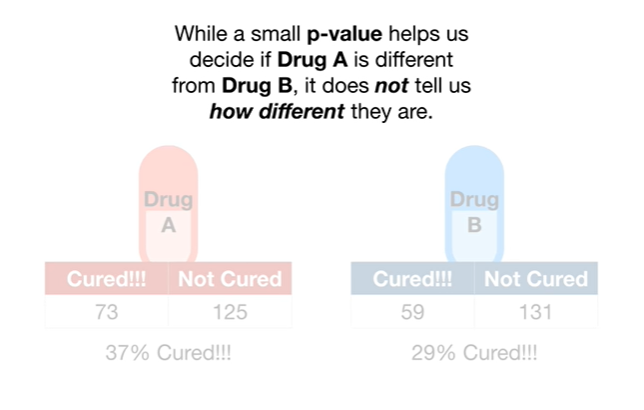


and the p-value helps us decide if we should reject the null hypothesis or not.

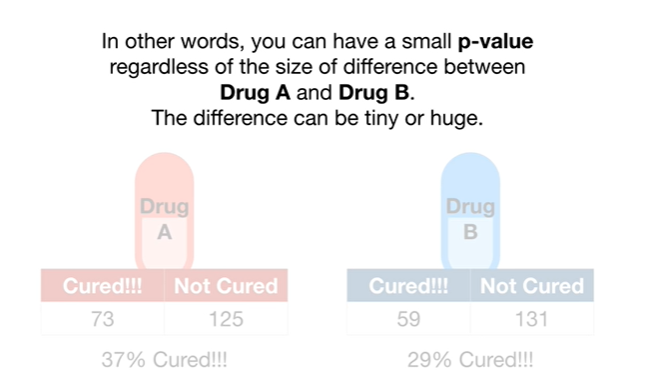
Small BAM !



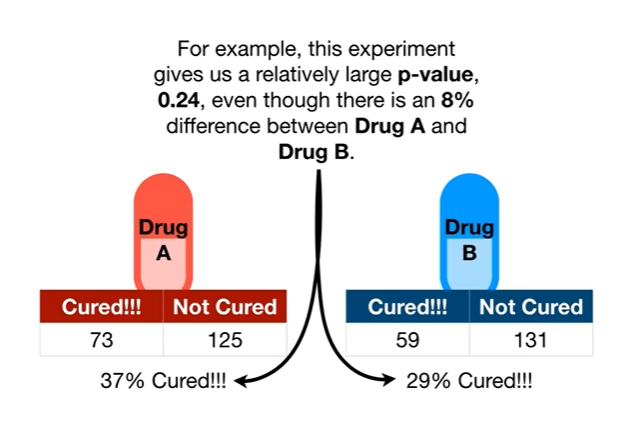
Okay now that we have that fancy terminology out of the way the second thing I want to say is way more interesting.



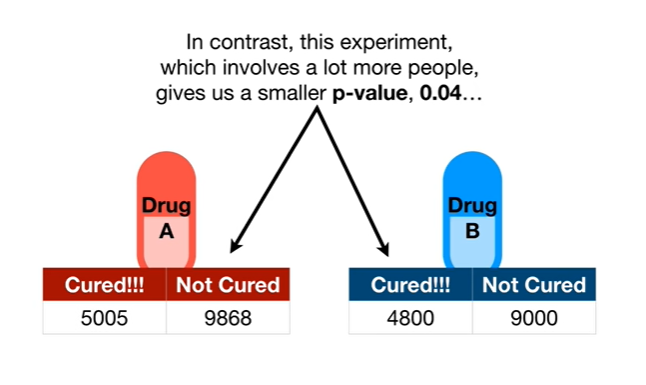
While a small p-value helps us decide if drug A is different from drug B, it does not tell us how different they are.



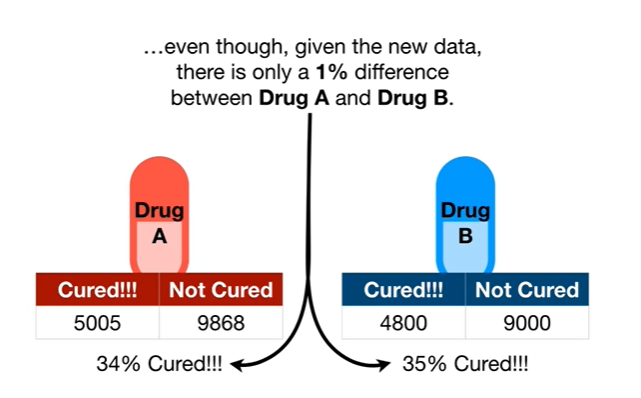
In other words you can have a small p-value regardless of the size of difference between drug a and drug B the difference can be tiny or huge.



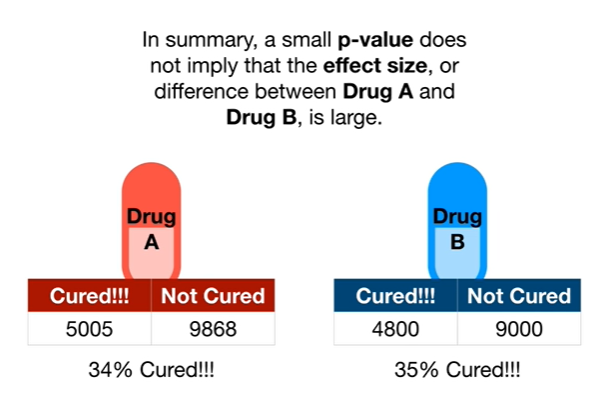
For example this experiment gives us a relatively large p-value 0.2 for even though there is an 8 percent difference between drug A and drug B.



In contrast this experiment which involves a lot more people gives us a smaller p-value 0.04



even though given the new data there is only a 1% difference between drug A and drug B.



In summary a small p-value does not imply that the effect size or difference between drug a and drug B is large.

Double BAM !!!