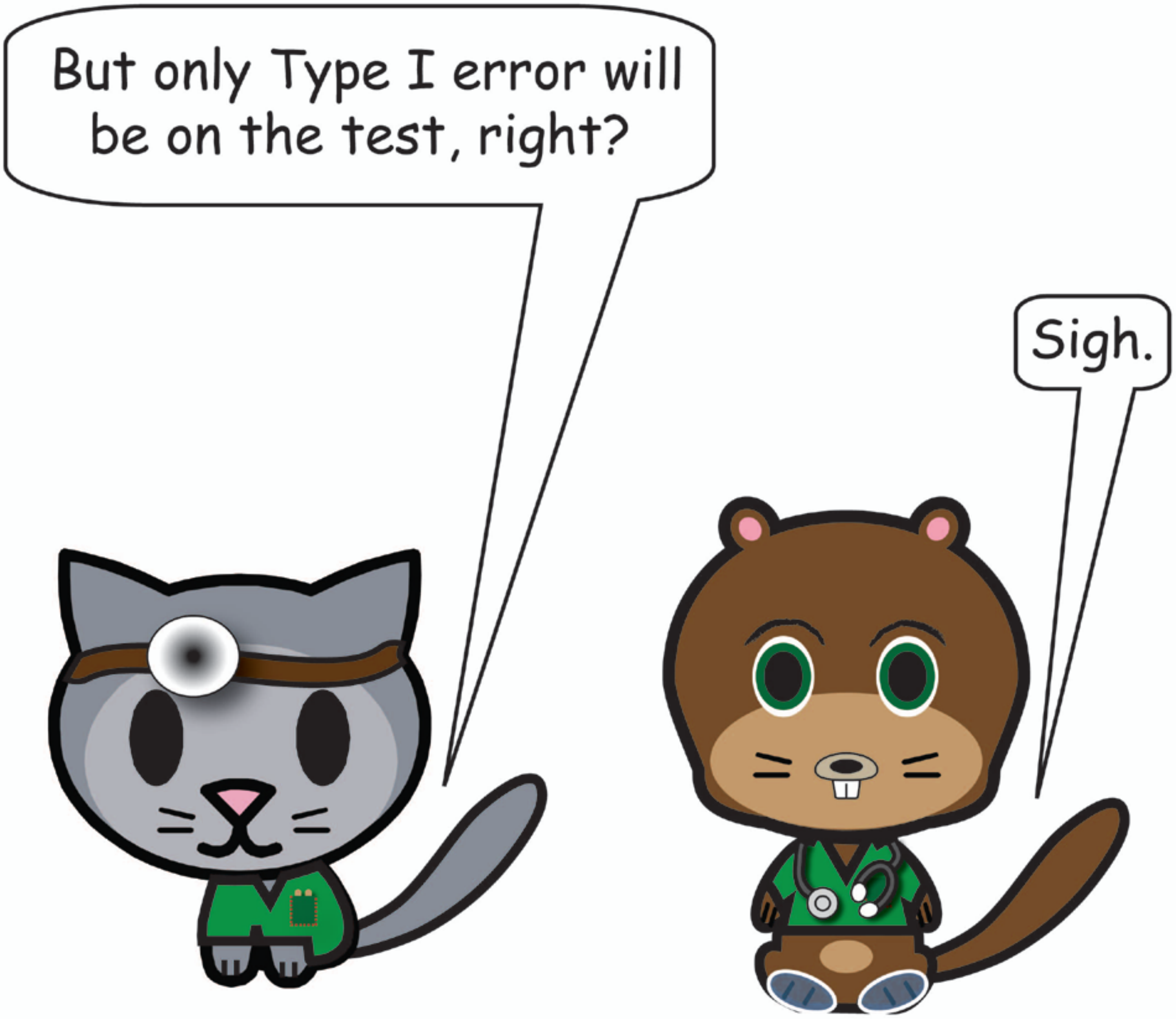
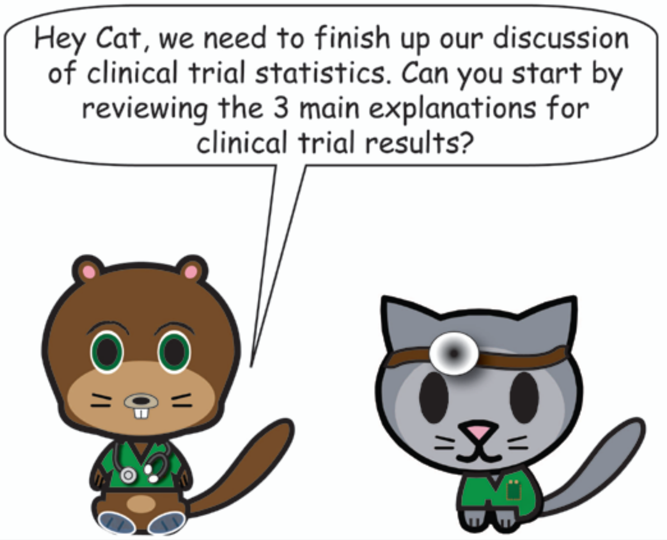
A "Doc Squirrel" and "Kid Cat" Adventure

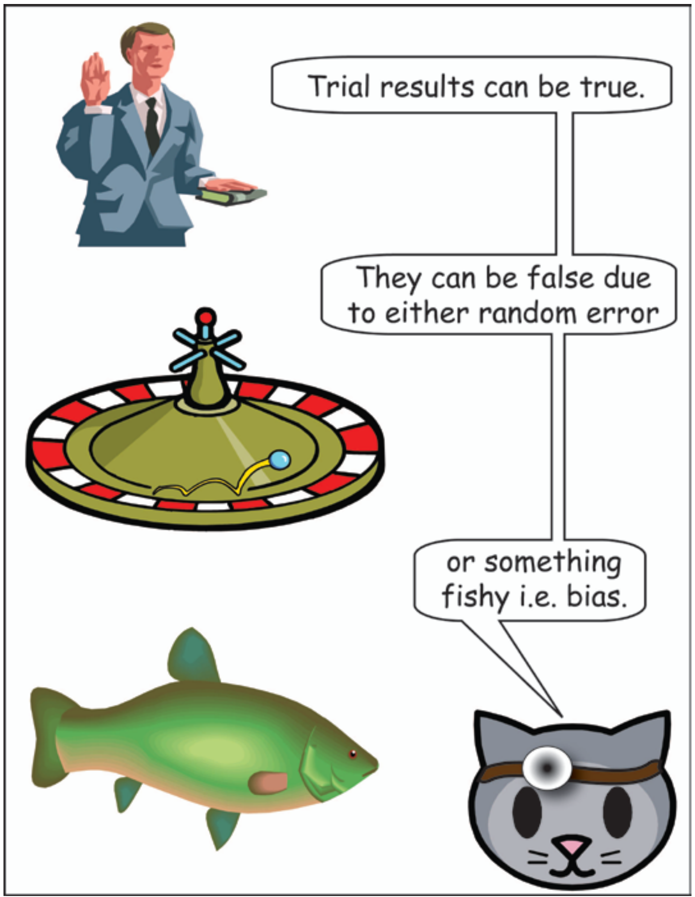
Our Heros Battle Type I *and* II Error

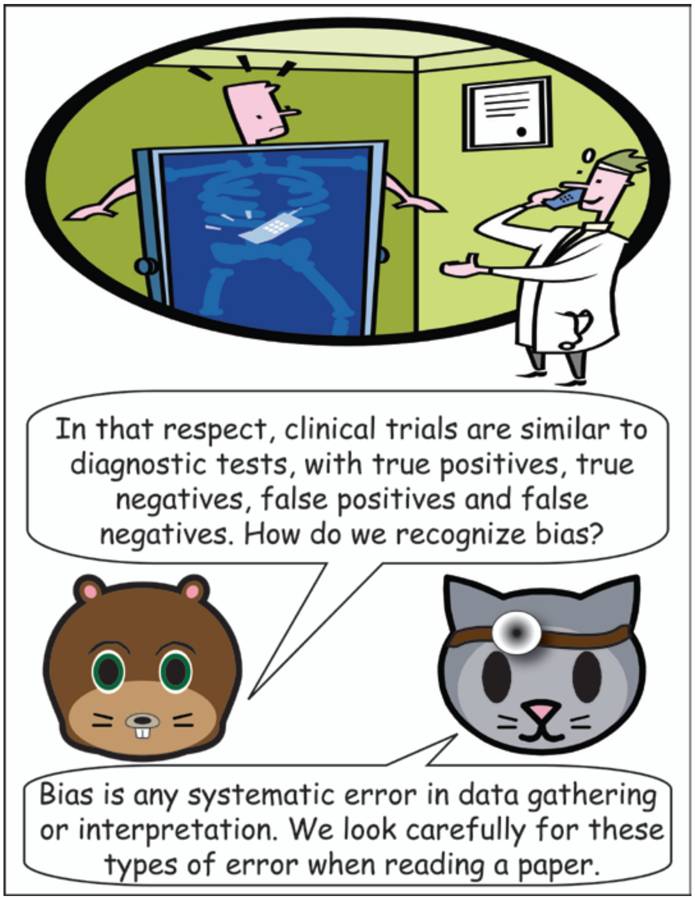


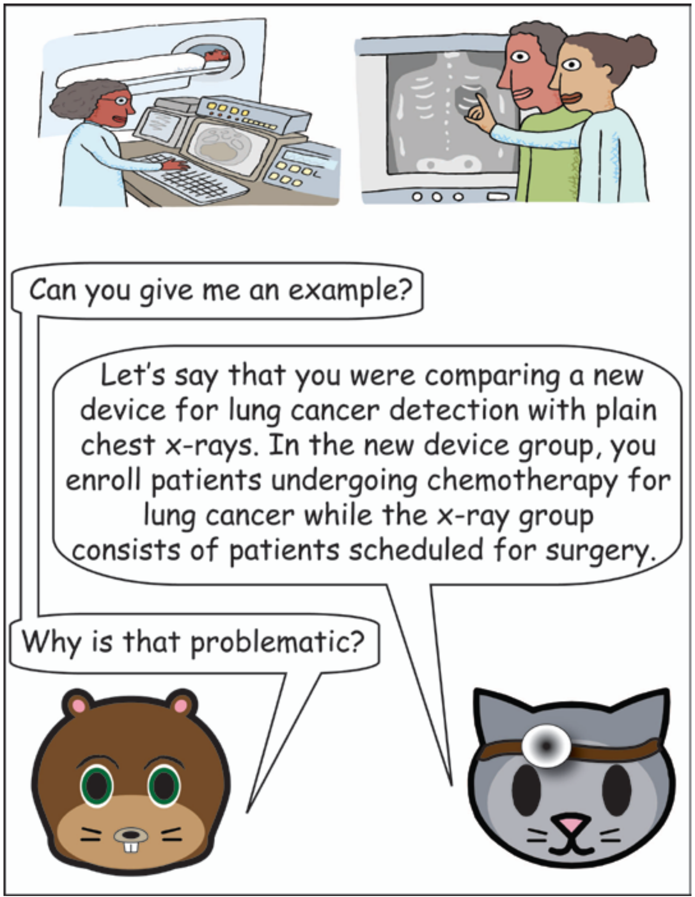
By Stefan Tigges MD MSCR

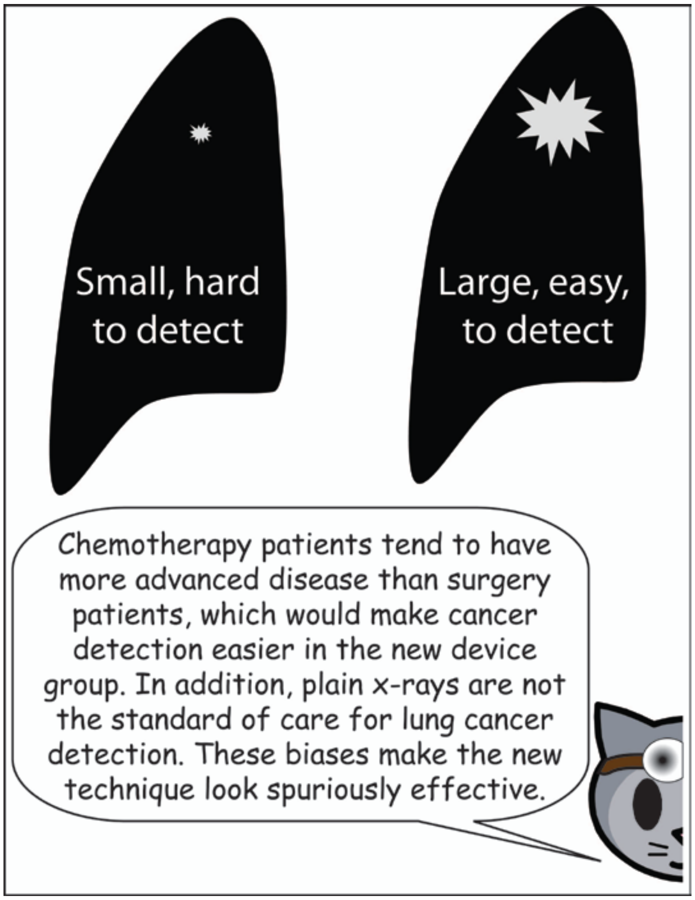
	Actual Rx Difference	No Rx Difference
Trial Positive	True Positive	False Positive
Trial Negative	False Negative	True Negative

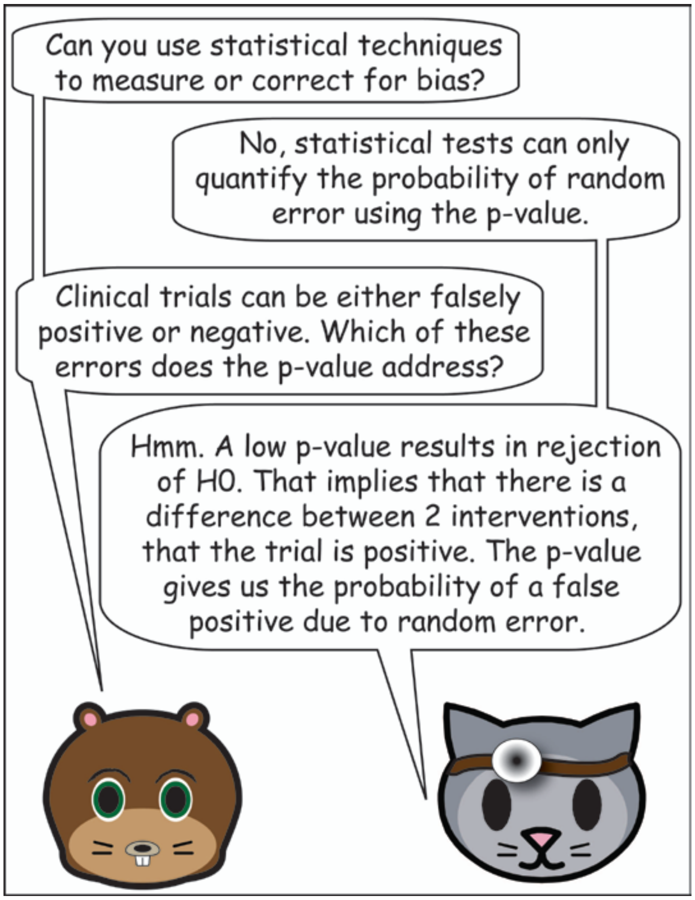

Hey Cat, we need to finish up our discussion of clinical trial statistics. Can you start by reviewing the 3 main explanations for clinical trial results?


Trial results can be true.
They can be false due to either random error
or something fishy i.e. bias.

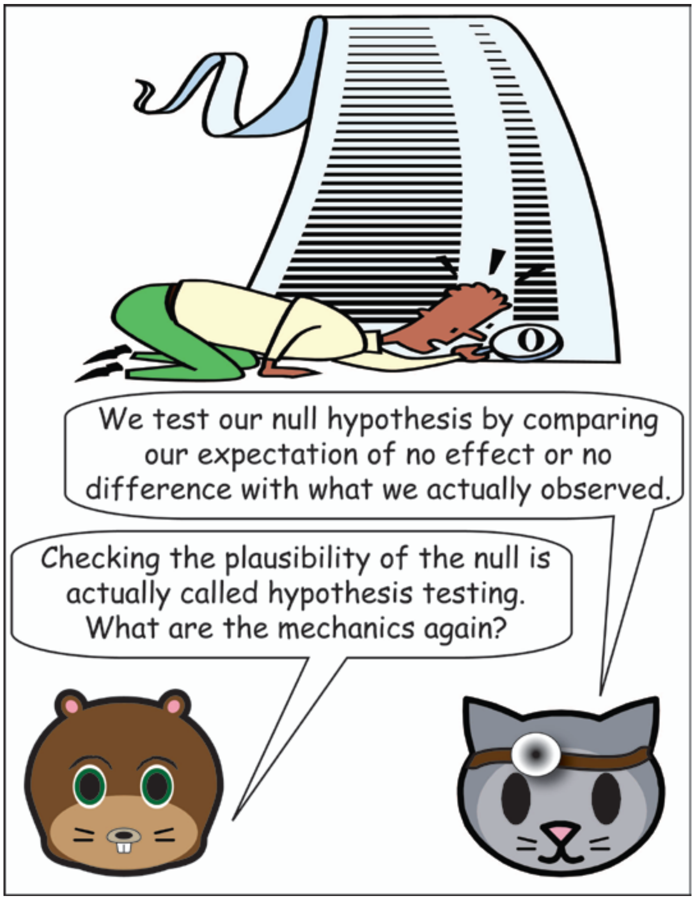

In that respect, clinical trials are similar to diagnostic tests, with true positives, true negatives, false positives and false negatives. How do we recognize bias?
Bias is any systematic error in data gathering or interpretation. We look carefully for these types of error when reading a paper.

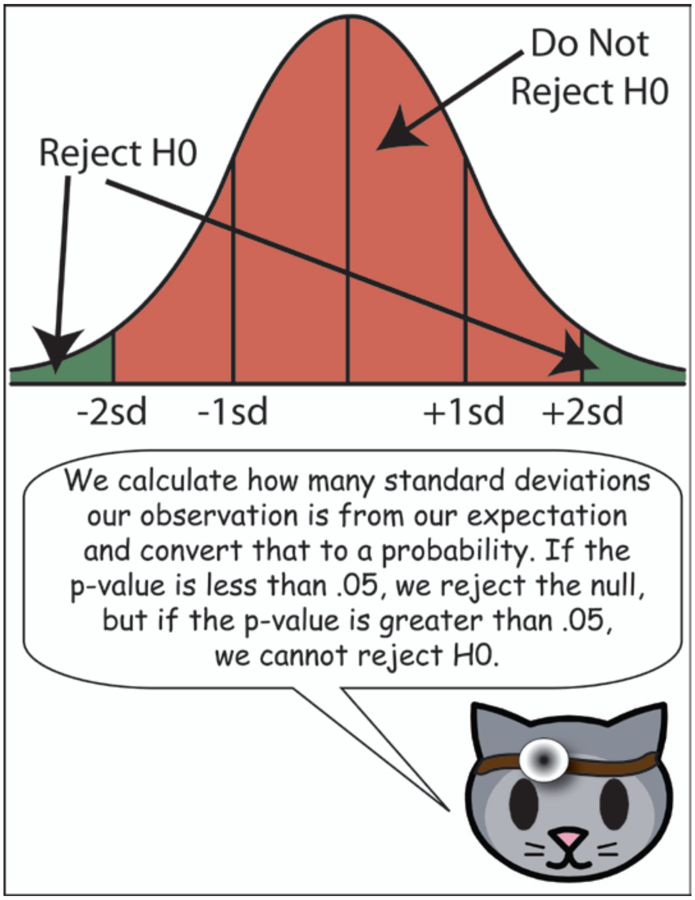

Can you give me an example?
Let's say that you were comparing a new device for lung cancer detection with plain chest x-rays. In the new device group, you enroll patients undergoing chemotherapy for lung cancer while the x-ray group consists of patients scheduled for surgery.
Why is that problematic?

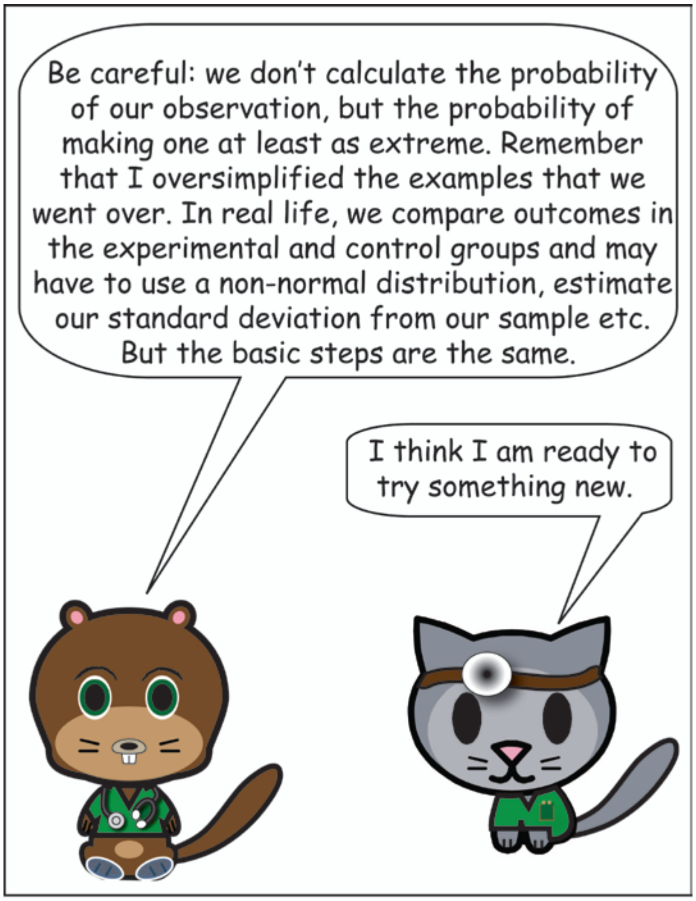

Small, hard to detect
Large, easy, to detect
Chemotherapy patients tend to have more advanced disease than surgery patients, which would make cancer detection easier in the new device group. In addition, plain x-rays are not the standard of care for lung cancer detection. These biases make the new technique look spuriously effective.

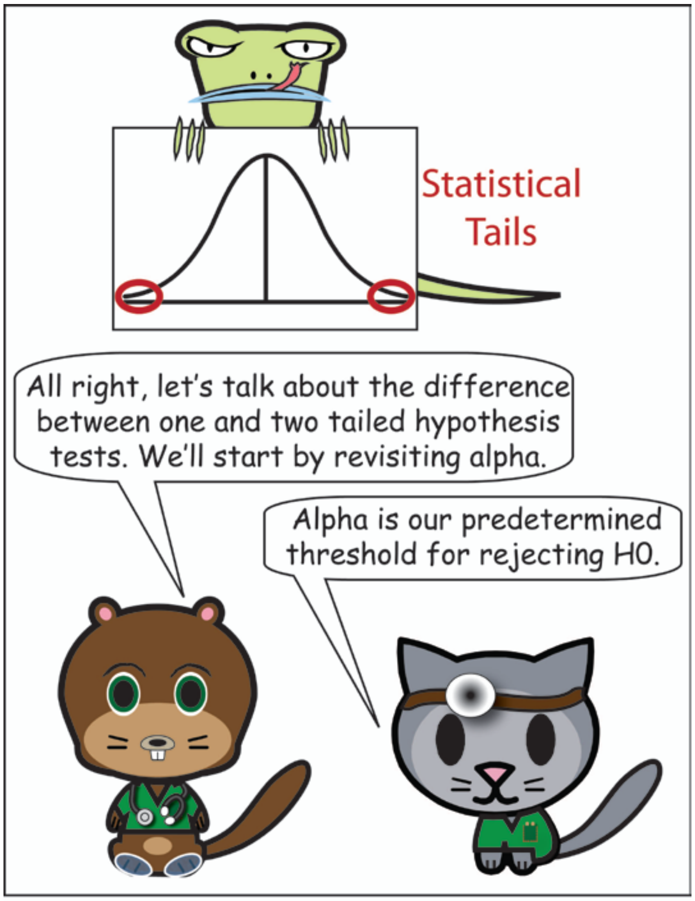

Can you use statistical techniques to measure or correct for bias?
No, statistical tests can only quantify the probability of random error using the p-value.
Clinical trials can be either falsely positive or negative. Which of these errors does the p-value address?
Hmm. A low p-value results in rejection of H0. That implies that there is a difference between 2 interventions, that the trial is positive. The p-value gives us the probability of a false positive due to random error.

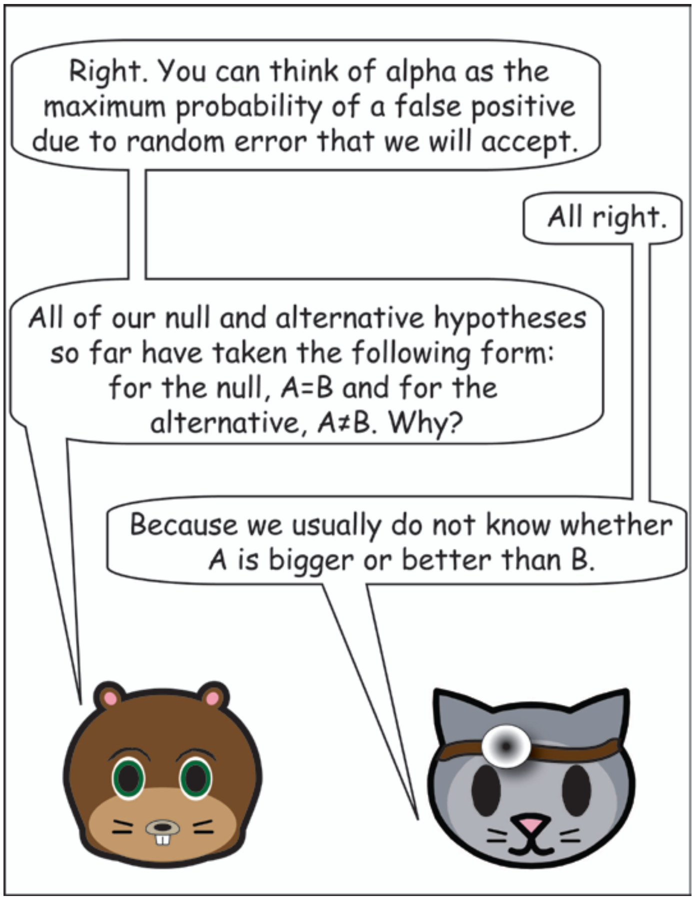
A false positive trial that reports a difference between 2 interventions when there really isn't one can occur due to bias or random error. You told me that we look carefully for bias when we read a clinical trial, but how do we look for random error?

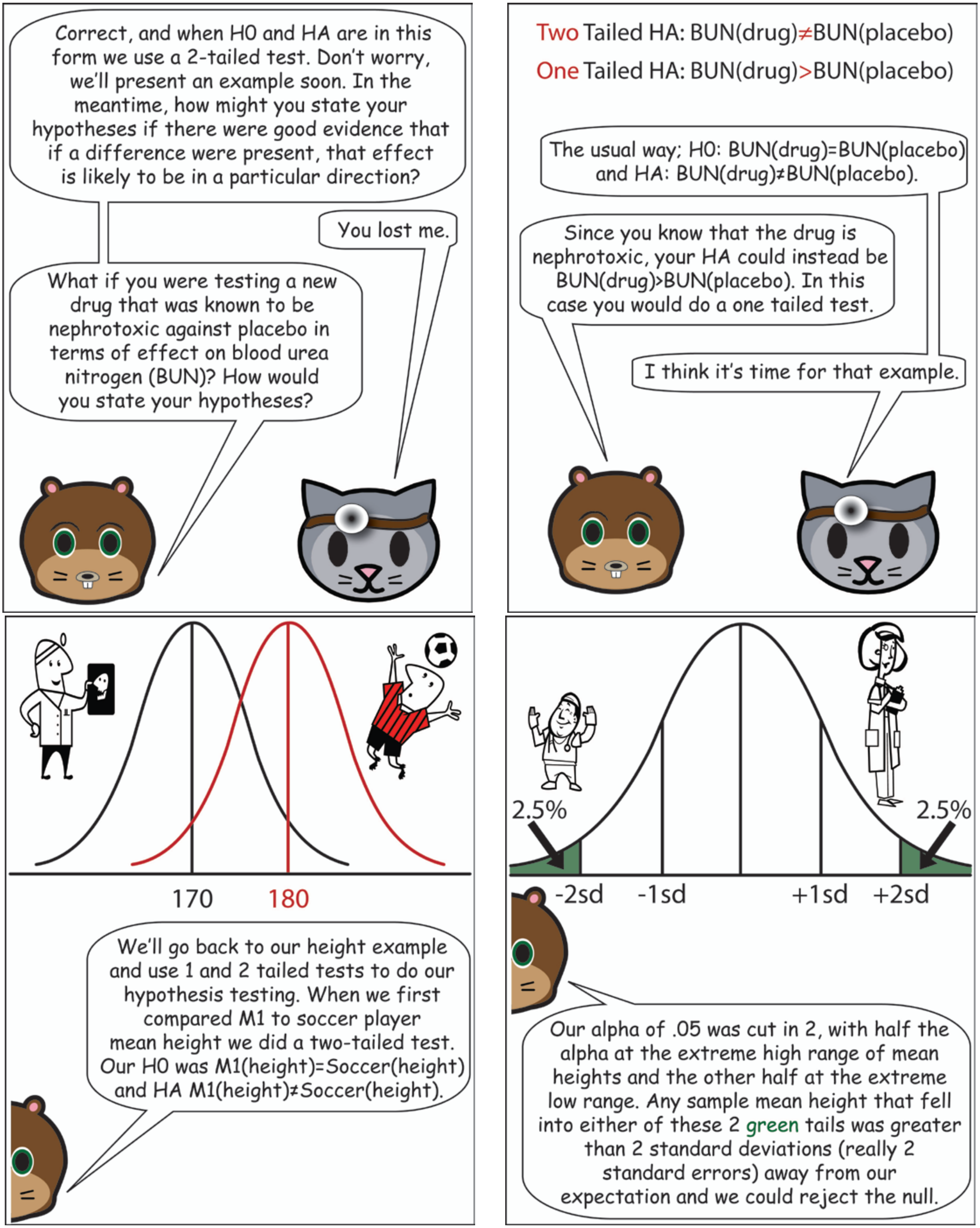

We test our null hypothesis by comparing our expectation of no effect or no difference with what we actually observed.
Checking the plausibility of the null is actually called hypothesis testing. What are the mechanics again?

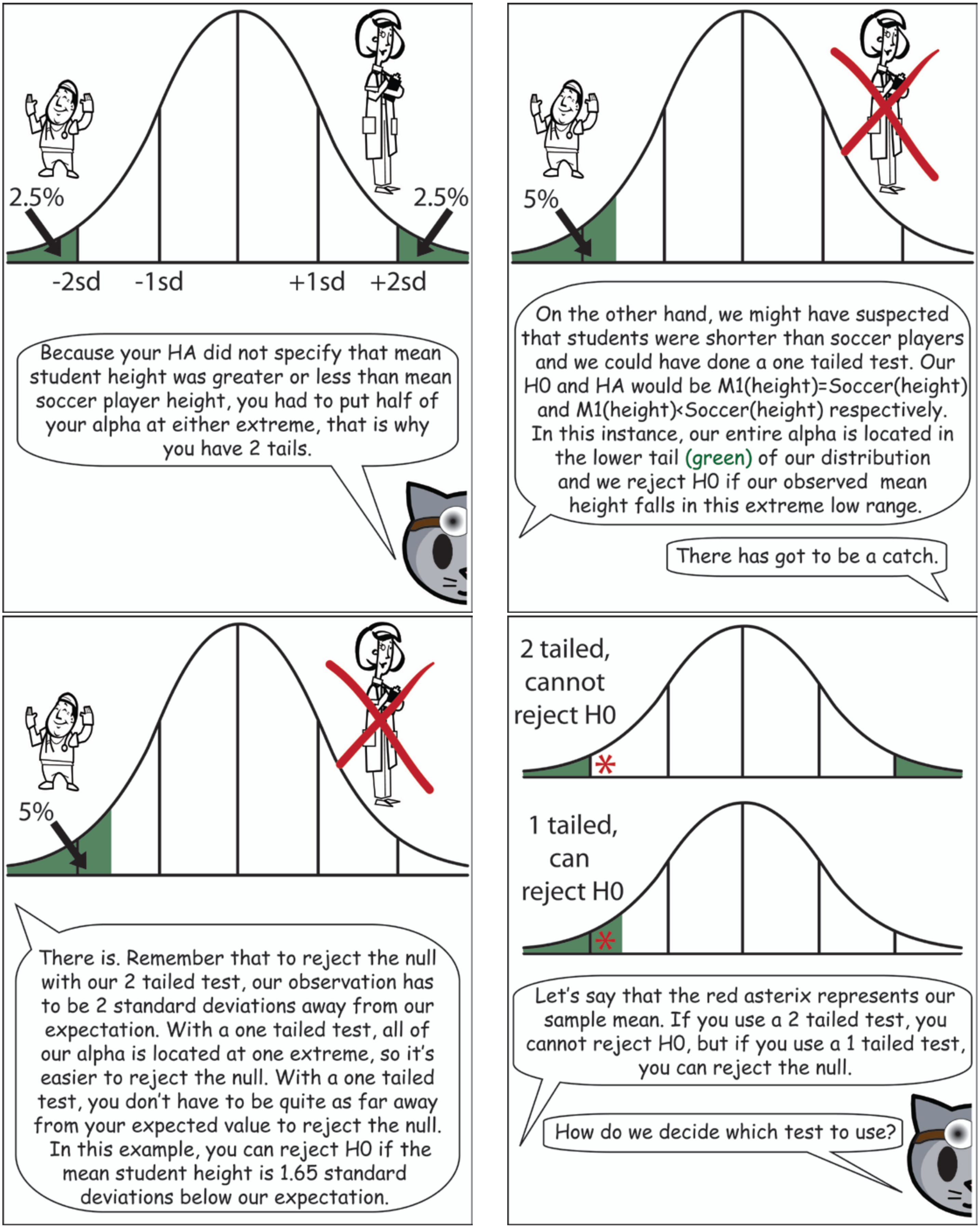

Do Not Reject H0
Reject H0
-2sd
-1sd
+1sd
+2sd
We calculate how many standard deviations our observation is from our expectation and convert that to a probability. If the p-value is less than .05, we reject the null, but if the p-value is greater than .05, we cannot reject H0.

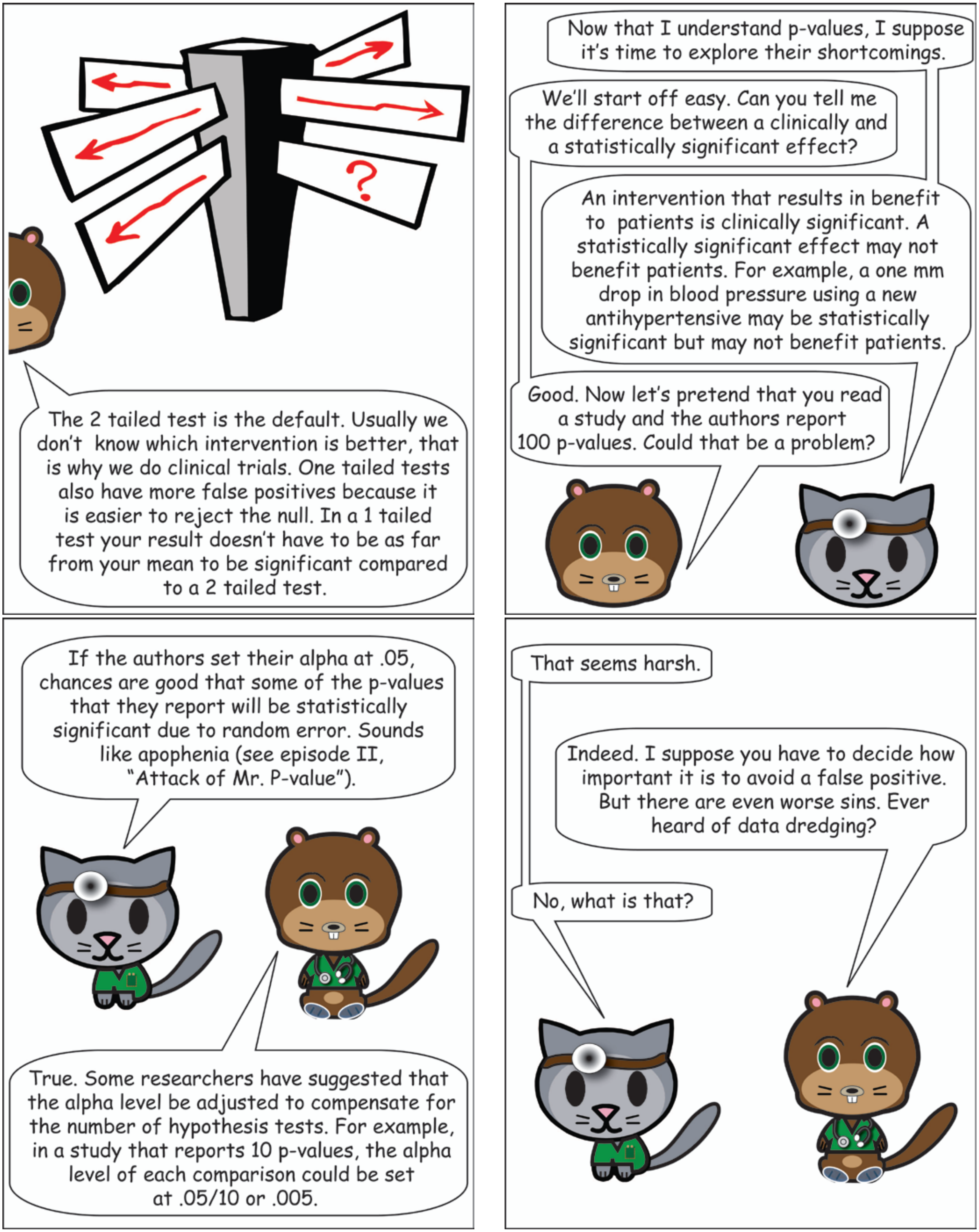

Be careful: we don't calculate the probability of our observation, but the probability of making one at least as extreme. Remember that I oversimplified the examples that we went over. In real life, we compare outcomes in the experimental and control groups and may have to use a non-normal distribution, estimate our standard deviation from our sample etc. But the basic steps are the same.
I think I am ready to try something new.

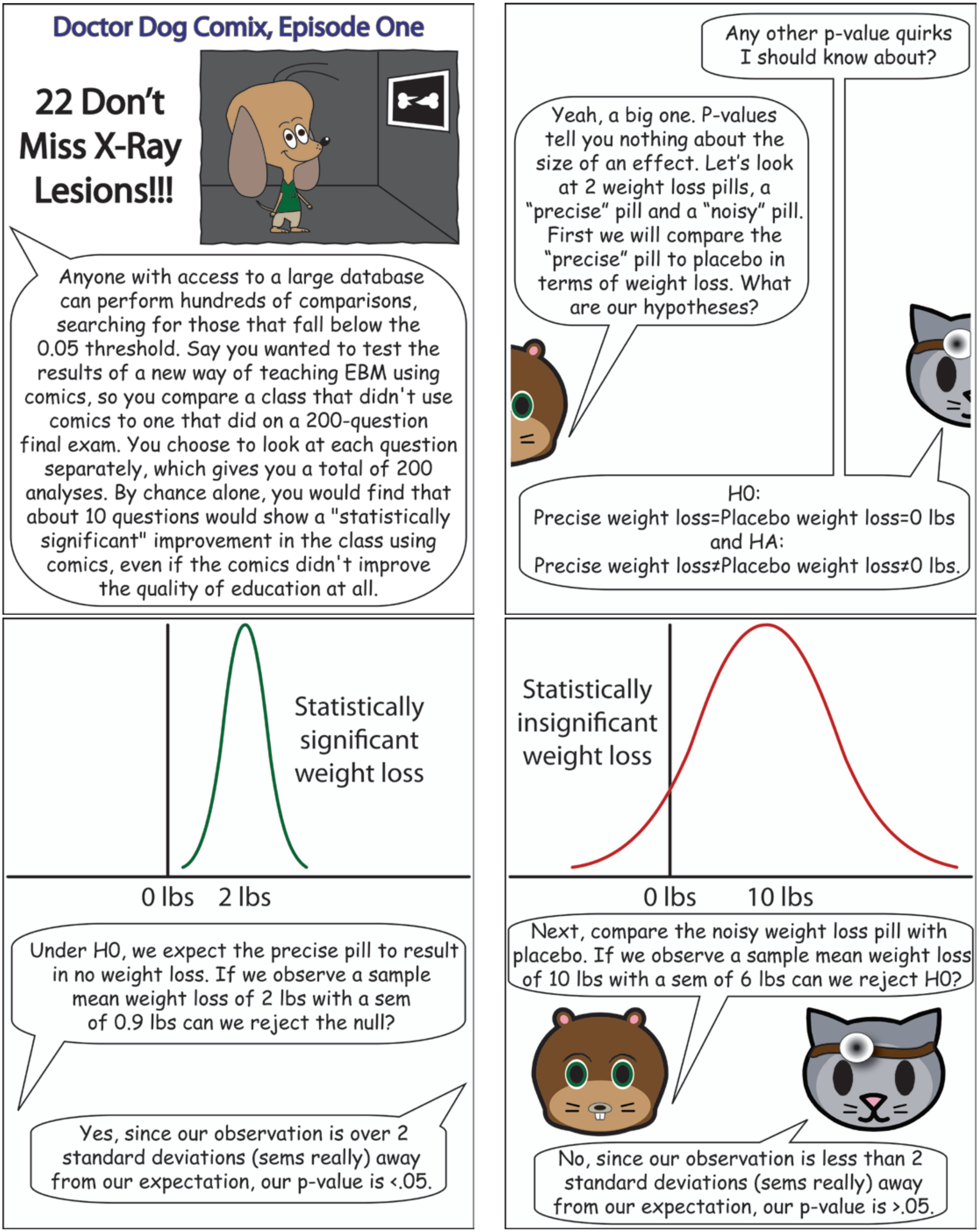

Statistical Tails
All right, let's talk about the difference between one and two tailed hypothesis tests. We'll start by revisiting alpha.
Alpha is our predetermined threshold for rejecting H0.

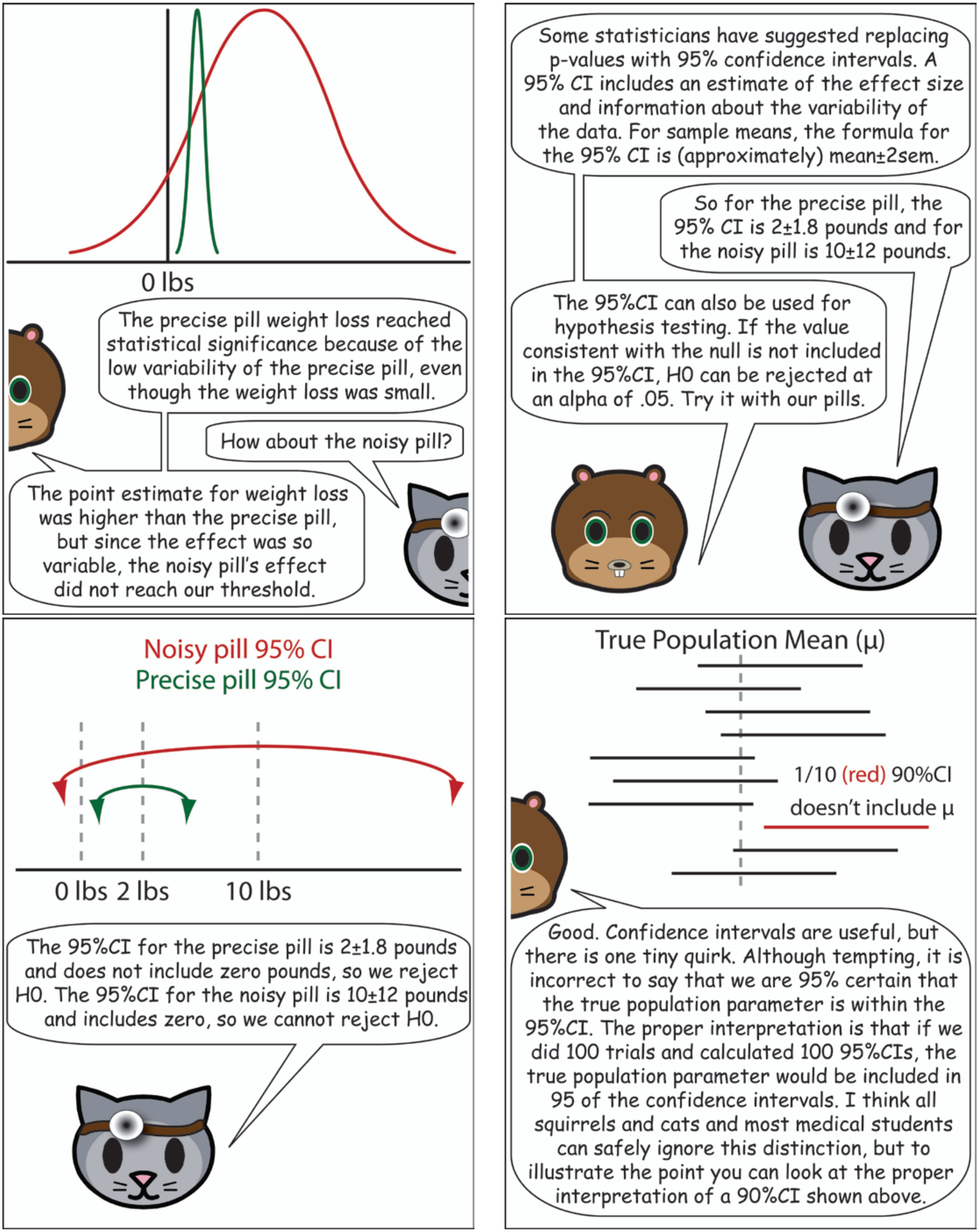

Right. You can think of alpha as the maximum probability of a false positive due to random error that we will accept.
All right.
All of our null and alternative hypotheses so far have taken the following form: for the null, A=B and for the alternative, A≠B. Why?
Because we usually do not know whether A is bigger or better than B.

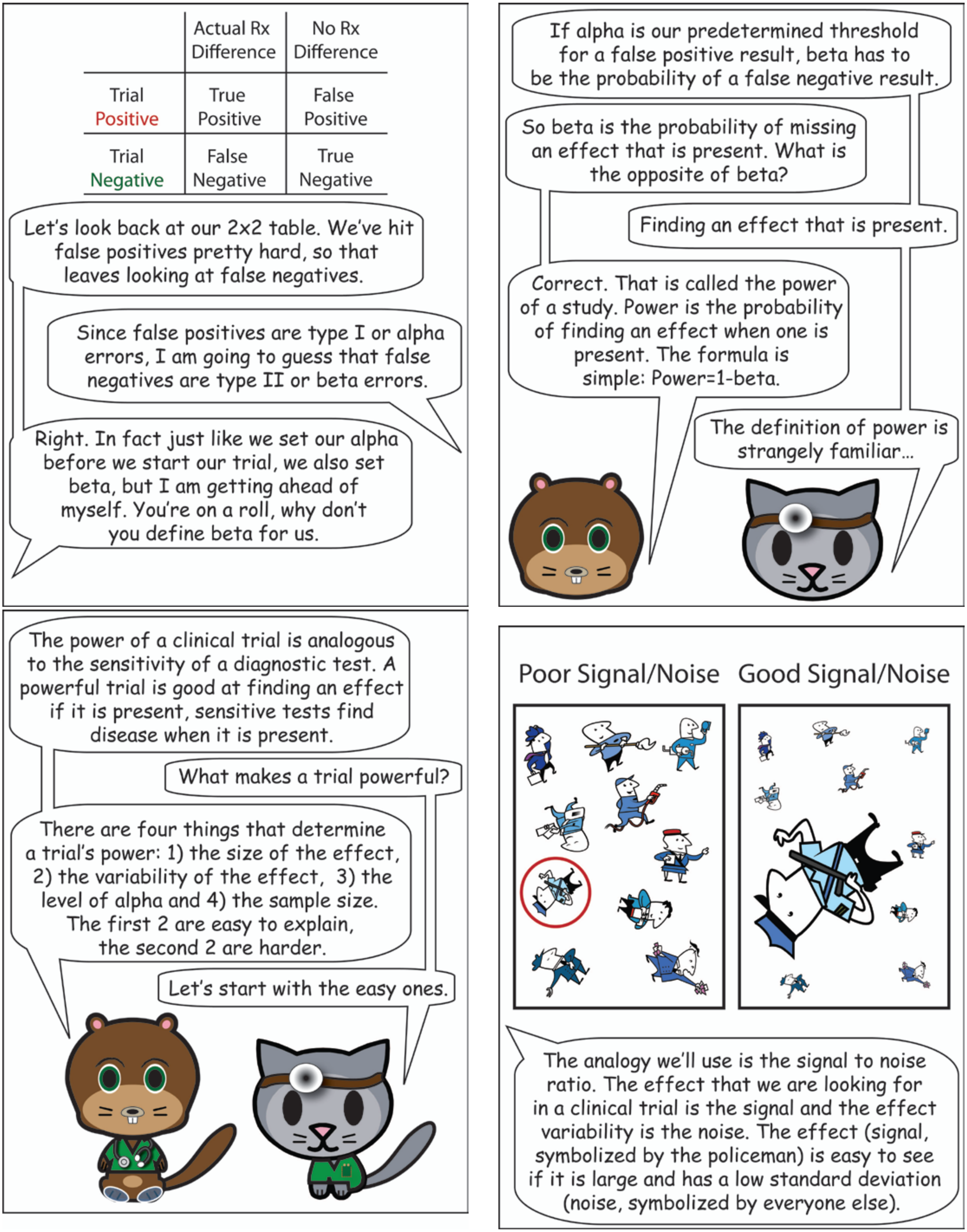

Correct, and when H0 and HA are in this form we use a 2-tailed test. Don't worry, we'll present an example soon. In the meantime, how might you state your hypotheses if there were good evidence that if a difference were present, that effect is likely to be in a particular direction?
You lost me.
What if you were testing a new drug that was known to be nephrotoxic against placebo in terms of effect on blood urea nitrogen (BUN)? How would you state your hypotheses?
Two Tailed HA: BUN(drug)≠BUN(placebo)
One Tailed HA: BUN(drug)>BUN(placebo)
The usual way; H0: BUN(drug)=BUN(placebo) and HA: BUN(drug)≠BUN(placebo).
Since you know that the drug is nephrotoxic, your HA could instead be BUN(drug)>BUN(placebo). In this case you would do a one tailed test.
I think it's time for that example.
170
180
We'll go back to our height example and use 1 and 2 tailed tests to do our hypothesis testing. When we first compared M1 to soccer player mean height we did a two-tailed test. Our H0 was M1(height)=Soccer(height) and HA M1(height)≠Soccer(height).
2.5%
2.5%
-2sd
-1sd
+1sd
+2sd
Our alpha of .05 was cut in 2, with half the alpha at the extreme high range of mean heights and the other half at the extreme low range. Any sample mean height that fell into either of these 2 green tails was greater than 2 standard deviations (really 2 standard errors) away from our expectation and we could reject the null.


2.5%
2.5%
-2sd
-1sd
+1sd
+2sd
Because your HA did not specify that mean student height was greater or less than mean soccer player height, you had to put half of your alpha at either extreme, that is why you have 2 tails.
5%
On the other hand, we might have suspected that students were shorter than soccer players and we could have done a one tailed test. Our H0 and HA would be M1(height)=Soccer(height) and M1(height)<Soccer(height) respectively. In this instance, our entire alpha is located in the lower tail (green) of our distribution and we reject H0 if our observed mean height falls in this extreme low range.
There has got to be a catch.
5%
There is. Remember that to reject the null with our 2 tailed test, our observation has to be 2 standard deviations away from our expectation. With a one tailed test, all of our alpha is located at one extreme, so it's easier to reject the null. With a one tailed test, you don't have to be quite as far away from your expected value to reject the null. In this example, you can reject H0 if the mean student height is 1.65 standard deviations below our expectation.
2 tailed, cannot reject H0
*
1 tailed, can reject H0
*
Let's say that the red asterix represents our sample mean. If you use a 2 tailed test, you cannot reject H0, but if you use a 1 tailed test, you can reject the null.
How do we decide which test to use?


The 2 tailed test is the default. Usually we don't know which intervention is better, that is why we do clinical trials. One tailed tests also have more false positives because it is easier to reject the null. In a 1 tailed test your result doesn't have to be as far from your mean to be significant compared to a 2 tailed test.
Now that I understand p-values, I suppose it's time to explore their shortcomings.
We'll start off easy. Can you tell me the difference between a clinically and a statistically significant effect?
An intervention that results in benefit to patients is clinically significant. A statistically significant effect may not benefit patients. For example, a one mm drop in blood pressure using a new antihypertensive may be statistically significant but may not benefit patients.
Good. Now let's pretend that you read a study and the authors report 100 p-values. Could that be a problem?
If the authors set their alpha at .05, chances are good that some of the p-values that they report will be statistically significant due to random error. Sounds like apophenia (see episode II, "Attack of Mr. P-value").
True. Some researchers have suggested that the alpha level be adjusted to compensate for the number of hypothesis tests. For example, in a study that reports 10 p-values, the alpha level of each comparison could be set at .05/10 or .005.
That seems harsh.
Indeed. I suppose you have to decide how important it is to avoid a false positive. But there are even worse sins. Ever heard of data dredging?
No, what is that?

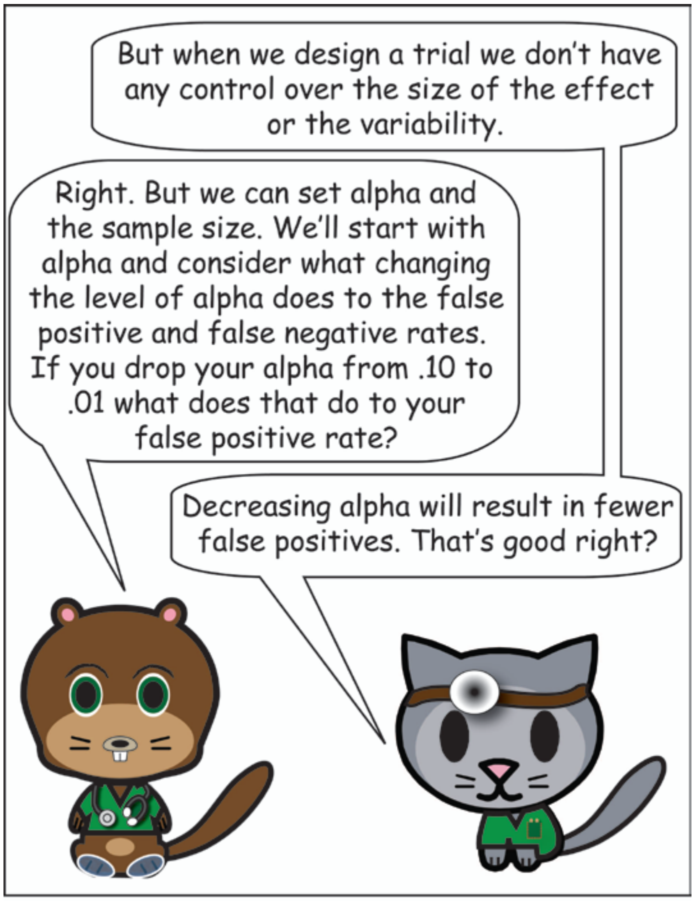

Doctor Dog Comix, Episode One
22 Don't Miss X-Ray Lesions!!!
Anyone with access to a large database can perform hundreds of comparisons, searching for those that fall below the 0.05 threshold. Say you wanted to test the results of a new way of teaching EBM using comics, so you compare a class that didn't use comics to one that did on a 200-question final exam. You choose to look at each question separately, which gives you a total of 200 analyses. By chance alone, you would find that about 10 questions would show a "statistically significant" improvement in the class using comics, even if the comics didn't improve the quality of education at all.
Any other p-value quirks I should know about?
Yeah, a big one. P-values tell you nothing about the size of an effect. Let's look at 2 weight loss pills, a "precise" pill and a "noisy" pill. First we will compare the "precise" pill to placebo in terms of weight loss. What are our hypotheses?
H0: Precise weight loss=Placebo weight loss=0 lbs and HA: Precise weight loss≠Placebo weight loss≠0 lbs.
Statistically significant weight loss
0 lbs
2 lbs
Under H0, we expect the precise pill to result in no weight loss. If we observe a sample mean weight loss of 2 lbs with a sem of 0.9 lbs can we reject the null?
Yes, since our observation is over 2 standard deviations (sems really) away from our expectation, our p-value is <.05.
Statistically insignificant weight loss
0 lbs
10 lbs
Next, compare the noisy weight loss pill with placebo. If we observe a sample mean weight loss of 10 lbs with a sem of 6 lbs can we reject H0?
No, since our observation is less than 2 standard deviations (sems really) away from our expectation, our p-value is >.05.

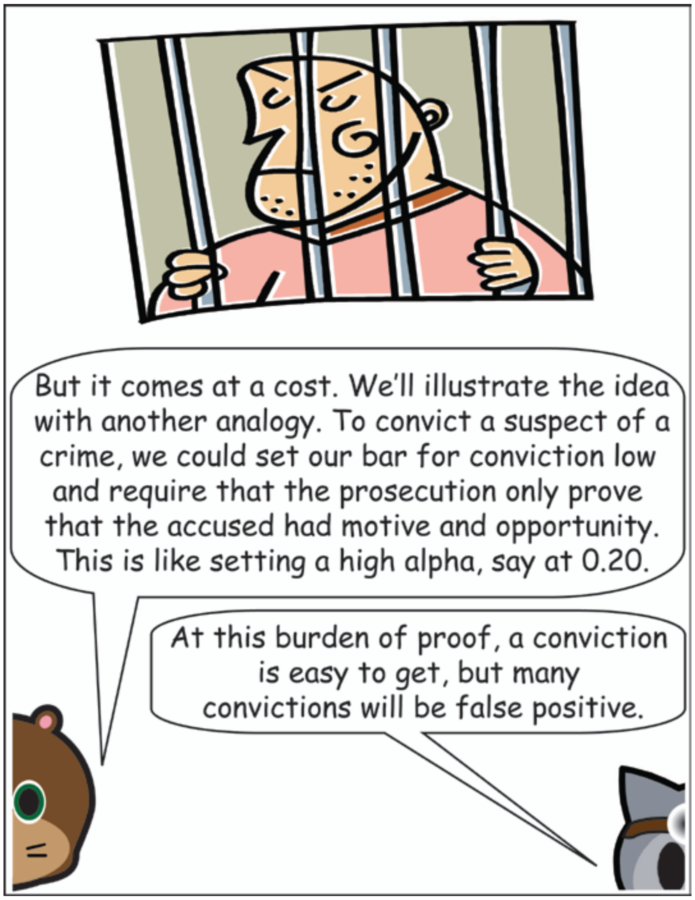

0 lbs
The precise pill weight loss reached statistical significance because of the low variability of the precise pill, even though the weight loss was small.
How about the noisy pill?
The point estimate for weight loss was higher than the precise pill, but since the effect was so variable, the noisy pill's effect did not reach our threshold.
Some statisticians have suggested replacing p-values with 95% confidence intervals. A 95% CI includes an estimate of the effect size and information about the variability of the data. For sample means, the formula for the 95% CI is (approximately) mean±2sem.
So for the precise pill, the 95% CI is 2±1.8 pounds and for the noisy pill is 10±12 pounds.
The 95%CI can also be used for hypothesis testing. If the value consistent with the null is not included in the 95%CI, H0 can be rejected at an alpha of .05. Try it with our pills.
Noisy pill 95% CI
Precise pill 95% CI
0 lbs 2 lbs 10 lbs
The 95%CI for the precise pill is 2±1.8 pounds and does not include zero pounds, so we reject H0. The 95%CI for the noisy pill is 10±12 pounds and includes zero, so we cannot reject H0.
True Population Mean (μ)
1/10 (red) 90%CI doesn't include μ
Good. Confidence intervals are useful, but there is one tiny quirk. Although tempting, it is incorrect to say that we are 95% certain that the true population parameter is within the 95%CI. The proper interpretation is that if we did 100 trials and calculated 100 95%CIs, the true population parameter would be included in 95 of the confidence intervals. I think all squirrels and cats and most medical students can safely ignore this distinction, but to illustrate the point you can look at the proper interpretation of a 90%CI shown above.

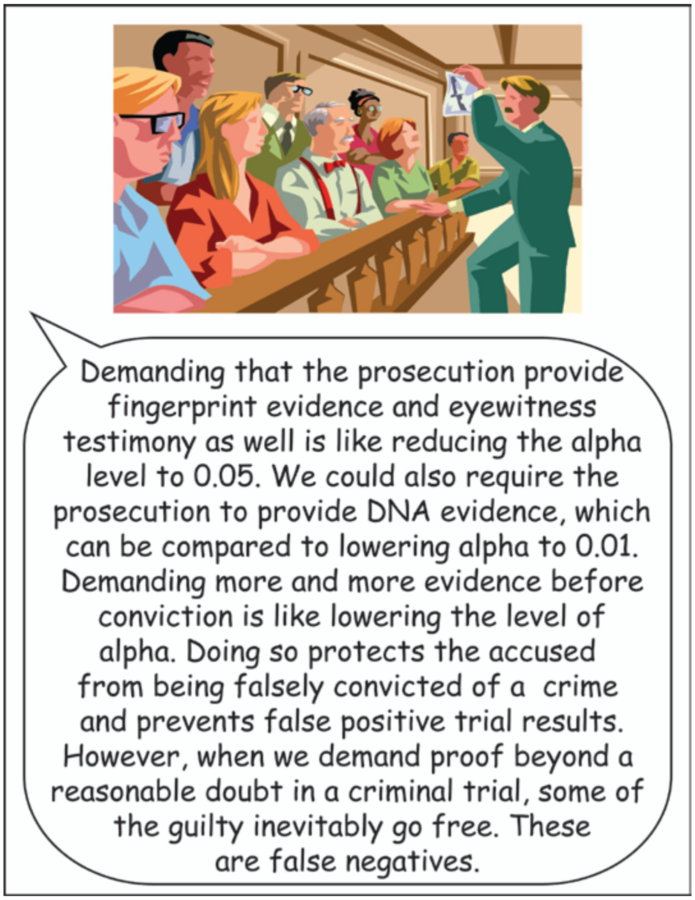


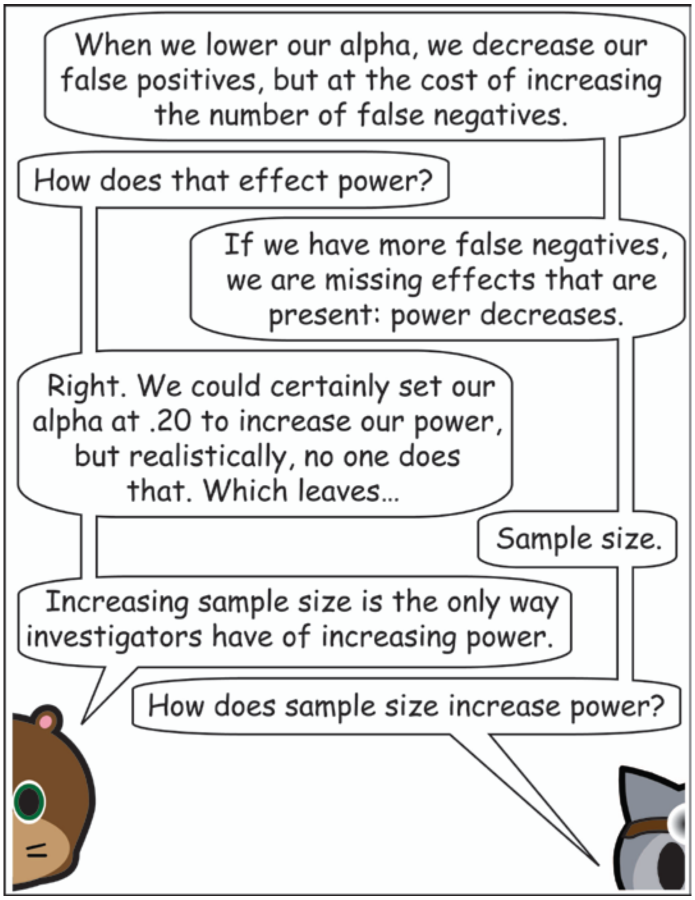
	Actual Rx Difference	No Rx Difference
Trial Positive	True Positive	False Positive
Trial Negative	False Negative	True Negative

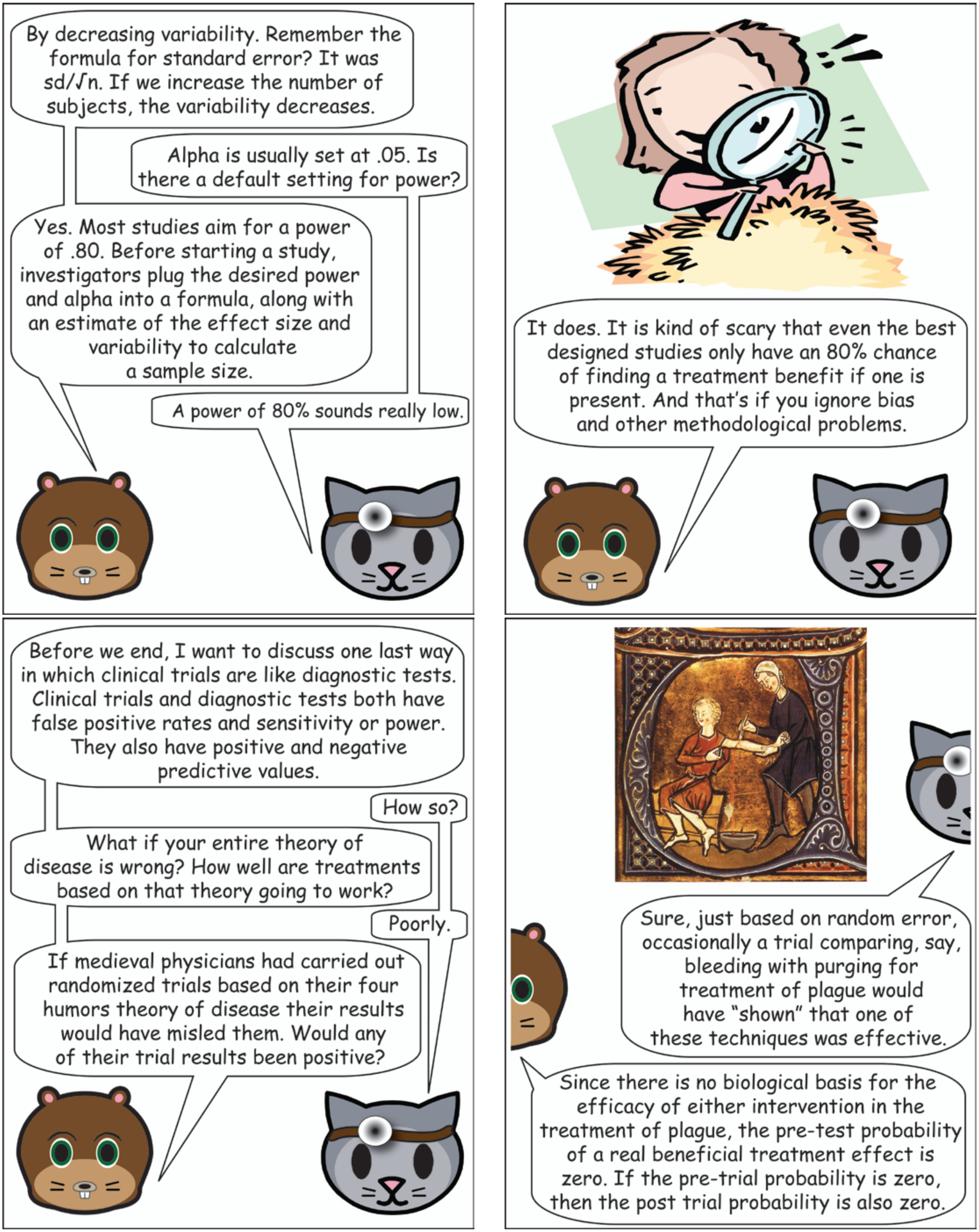
Let's look back at our 2x2 table. We've hit false positives pretty hard, so that leaves looking at false negatives.
Since false positives are type I or alpha errors, I am going to guess that false negatives are type II or beta errors.
Right. In fact just like we set our alpha before we start our trial, we also set beta, but I am getting ahead of myself. You're on a roll, why don't you define beta for us.
If alpha is our predetermined threshold for a false positive result, beta has to be the probability of a false negative result.
So beta is the probability of missing an effect that is present. What is the opposite of beta?
Finding an effect that is present.
Correct. That is called the power of a study. Power is the probability of finding an effect when one is present. The formula is simple: Power=1-beta.
The definition of power is strangely familiar...
The power of a clinical trial is analogous to the sensitivity of a diagnostic test. A powerful trial is good at finding an effect if it is present, sensitive tests find disease when it is present.
What makes a trial powerful?
There are four things that determine a trial's power: 1) the size of the effect, 2) the variability of the effect, 3) the level of alpha and 4) the sample size. The first 2 are easy to explain, the second 2 are harder.
Let's start with the easy ones.
Poor Signal/Noise
Good Signal/Noise
The analogy we'll use is the signal to noise ratio. The effect that we are looking for in a clinical trial is the signal and the effect variability is the noise. The effect (signal, symbolized by the policeman) is easy to see if it is large and has a low standard deviation (noise, symbolized by everyone else).


But when we design a trial we don't have any control over the size of the effect or the variability.
Right. But we can set alpha and the sample size. We'll start with alpha and consider what changing the level of alpha does to the false positive and false negative rates. If you drop your alpha from .10 to .01 what does that do to your false positive rate?
Decreasing alpha will result in fewer false positives. That's good right?


But it comes at a cost. We'll illustrate the idea with another analogy. To convict a suspect of a crime, we could set our bar for conviction low and require that the prosecution only prove that the accused had motive and opportunity. This is like setting a high alpha, say at 0.20.
At this burden of proof, a conviction is easy to get, but many convictions will be false positive.


Demanding that the prosecution provide fingerprint evidence and eyewitness testimony as well is like reducing the alpha level to 0.05. We could also require the prosecution to provide DNA evidence, which can be compared to lowering alpha to 0.01. Demanding more and more evidence before conviction is like lowering the level of alpha. Doing so protects the accused from being falsely convicted of a crime and prevents false positive trial results. However, when we demand proof beyond a reasonable doubt in a criminal trial, some of the guilty inevitably go free. These are false negatives.


When we lower our alpha, we decrease our false positives, but at the cost of increasing the number of false negatives.
How does that effect power?
If we have more false negatives, we are missing effects that are present: power decreases.
Right. We could certainly set our alpha at .20 to increase our power, but realistically, no one does that. Which leaves…
Sample size.
Increasing sample size is the only way investigators have of increasing power.
How does sample size increase power?


By decreasing variability. Remember the formula for standard error? It was sd/√n. If we increase the number of subjects, the variability decreases.
Alpha is usually set at .05. Is there a default setting for power?
Yes. Most studies aim for a power of .80. Before starting a study, investigators plug the desired power and alpha into a formula, along with an estimate of the effect size and variability to calculate a sample size.
A power of 80% sounds really low.
It does. It is kind of scary that even the best designed studies only have an 80% chance of finding a treatment benefit if one is present. And that's if you ignore bias and other methodological problems.
Before we end, I want to discuss one last way in which clinical trials are like diagnostic tests. Clinical trials and diagnostic tests both have false positive rates and sensitivity or power. They also have positive and negative predictive values.
How so?
What if your entire theory of disease is wrong? How well are treatments based on that theory going to work?
Poorly.
If medieval physicians had carried out randomized trials based on their four humors theory of disease their results would have misled them. Would any of their trial results been positive?
Sure, just based on random error, occasionally a trial comparing, say, bleeding with purging for treatment of plague would have "shown" that one of these techniques was effective.
Since there is no biological basis for the efficacy of either intervention in the treatment of plague, the pre-test probability of a real beneficial treatment effect is zero. If the pre-trial probability is zero, then the post trial probability is also zero.

That's all for now. Next time we will cover lung cancer screening.

References, Acknowledgements etc.

Many of the illustrations are modified clipart from Microsoft (Redmond, Washington) Office except "Doc" Squirrel who is a"semi"original creation.
All artwork was created or modified using Adobe Illustrator CS4 (San Jose, California). The precise vs. noisy pill example was adapted from *The Cult of Statistical Significance* by Ziliak and McCloskey.
The article that "showed" associations between astrological signs and disease states was: "Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health." *Journal of Clinical Epidemiology* 2006;59(9):964-969.
The illustration to explain the 90%CI was taken from *The Complete Idiot's Guide to Statistics*. An additional helpful source was *Statistical Reasoning in Medicine* by Moye. The analogy between diagnostic tests and clinical trials comes from: "Are All Significant P Values Created Equal? The Analogy Between Diagnostic Tests and Clinical Research." *JAMA*: 1987;257(18):2459-2463.
Look for another edition, coming soon!