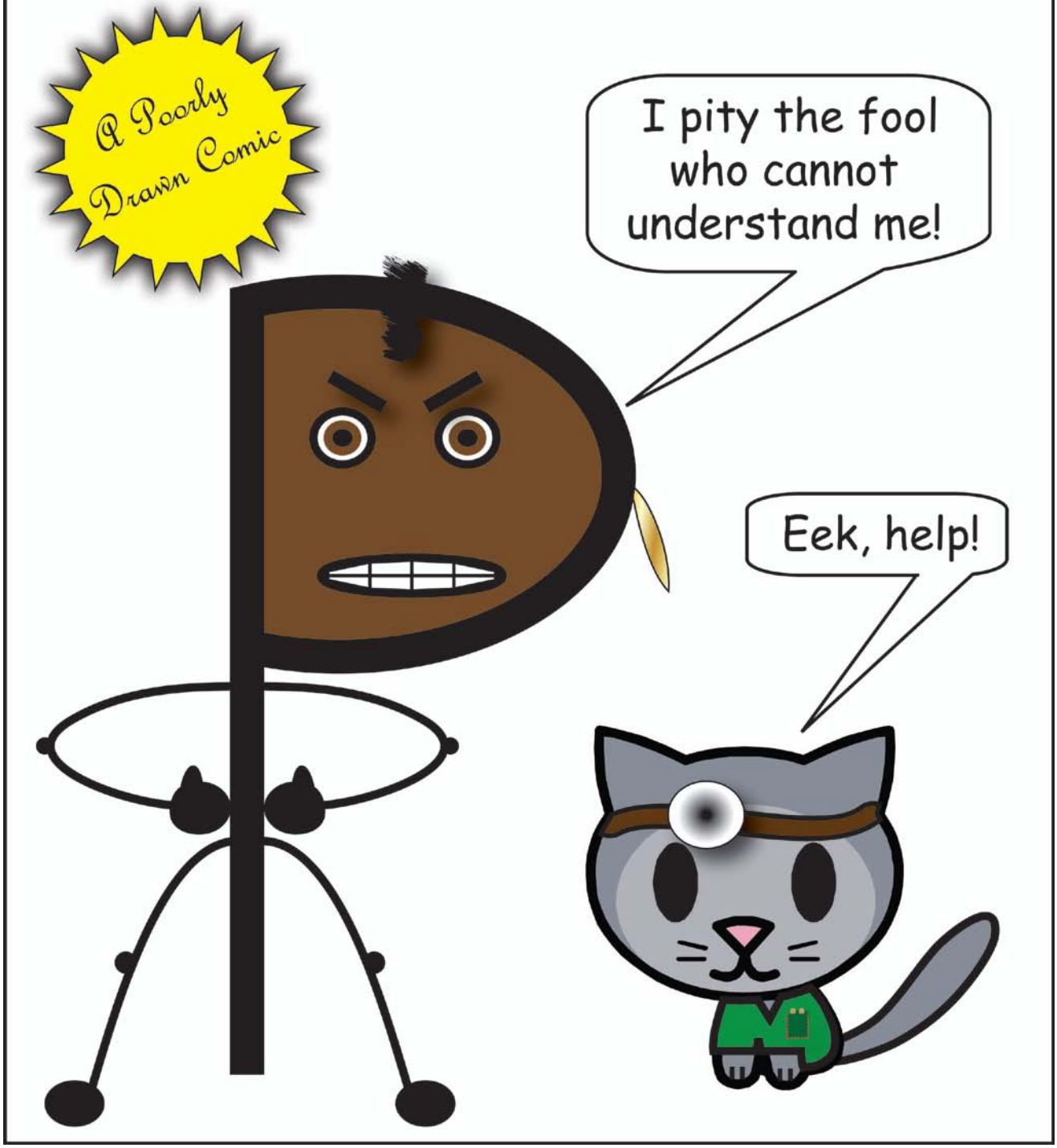
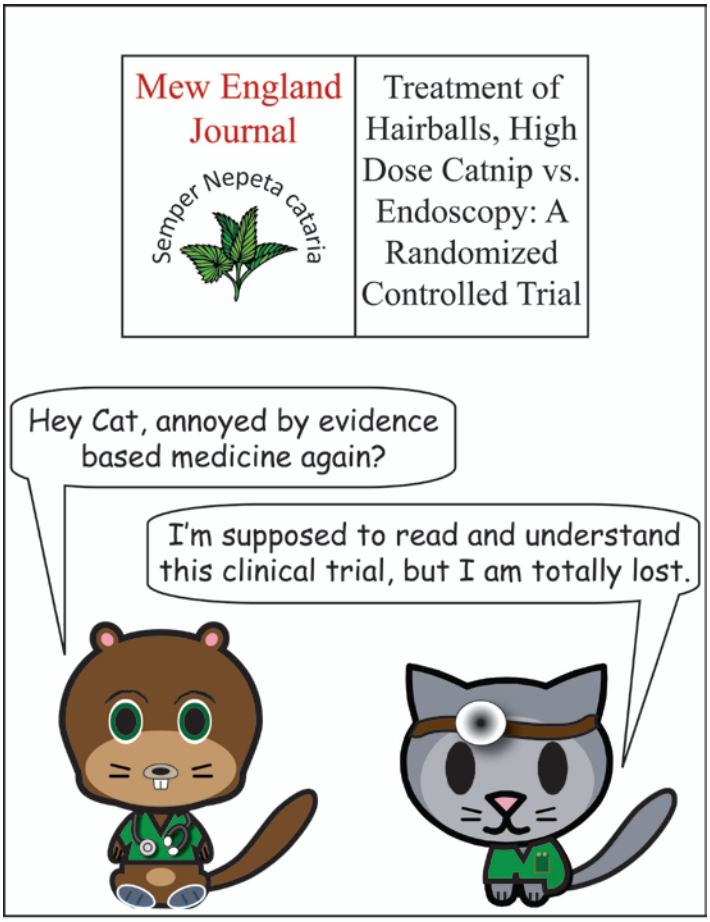
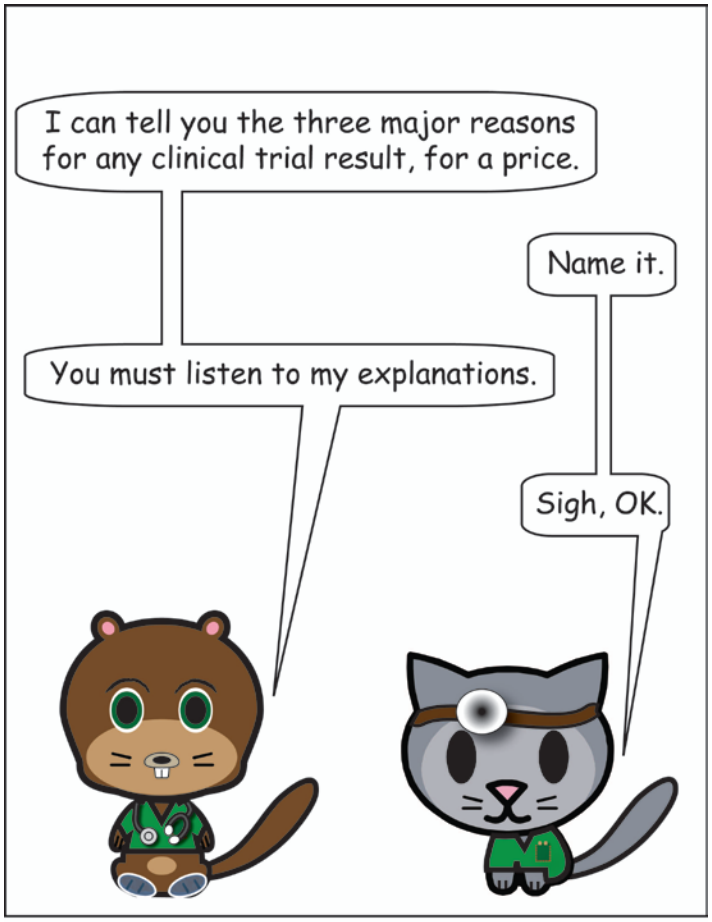
A "Doc Squirrel" and "Kid Cat" Adventure

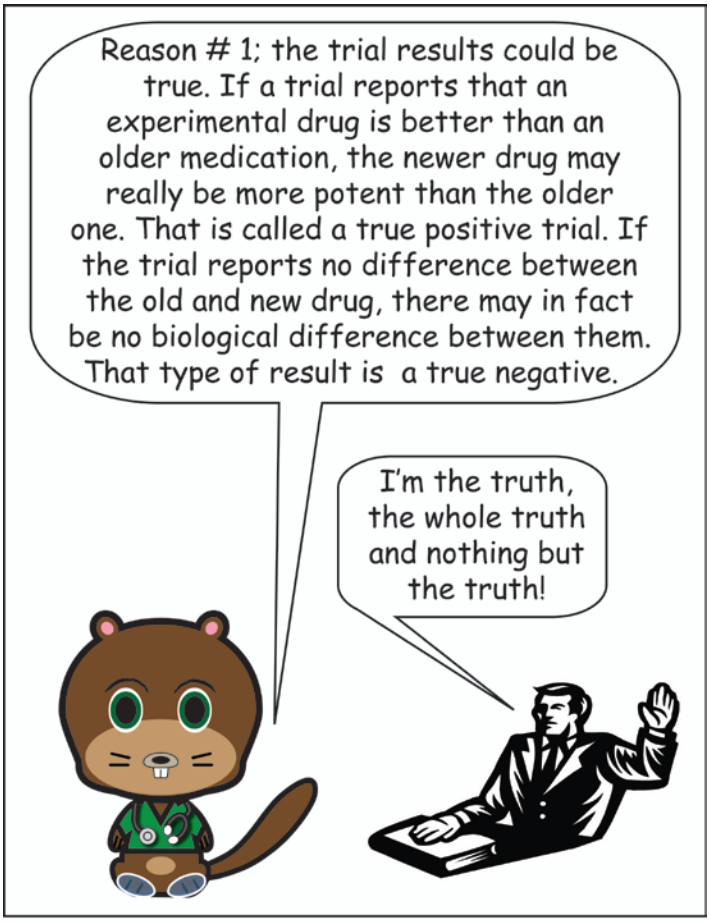
Attack of Mr. P-value!

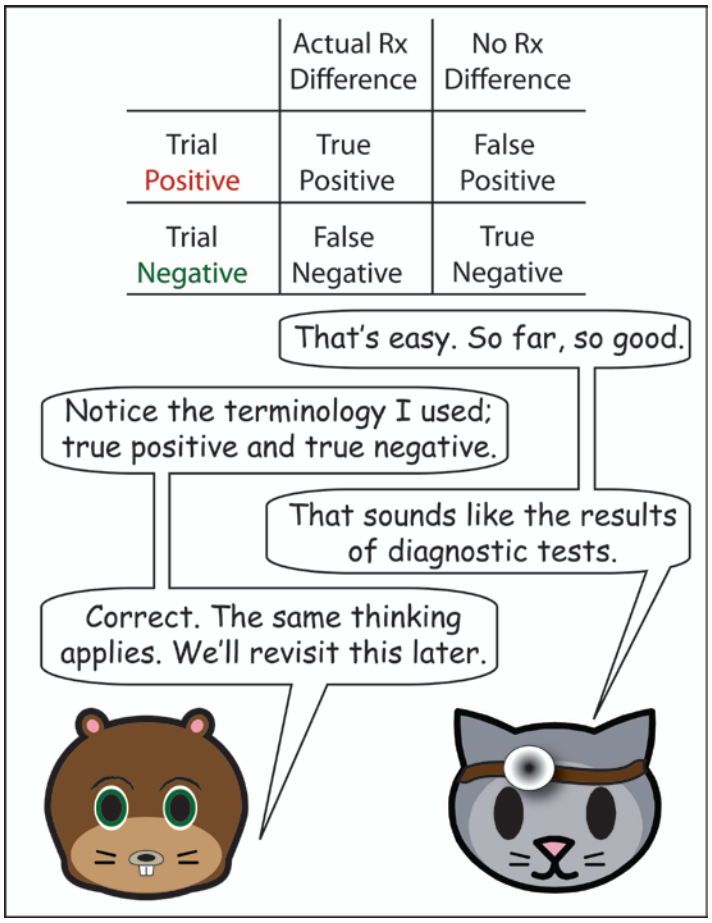


By Stefan Tigges MD MSCR

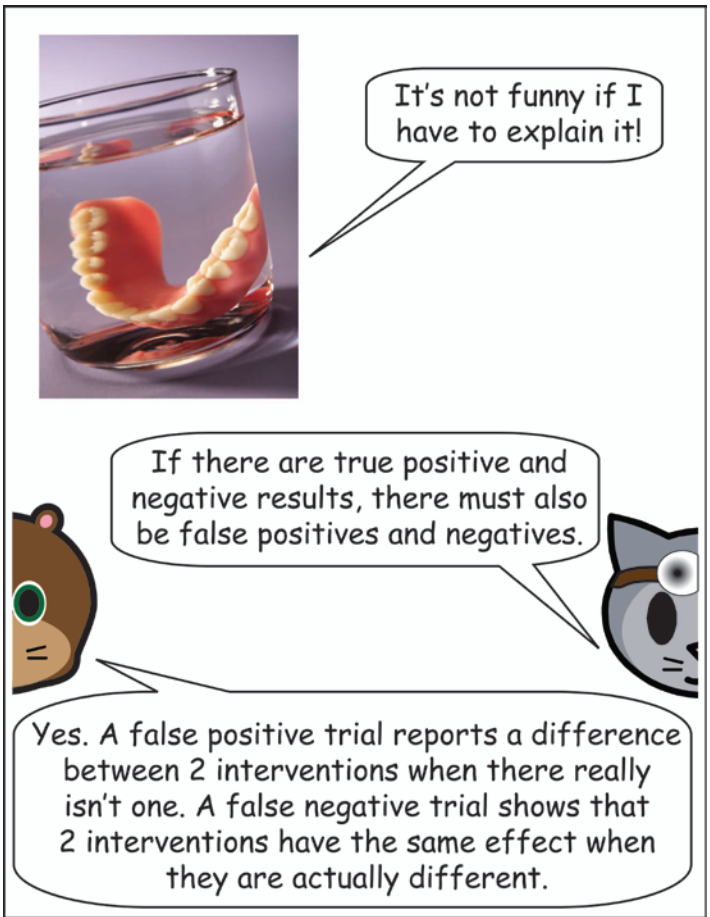




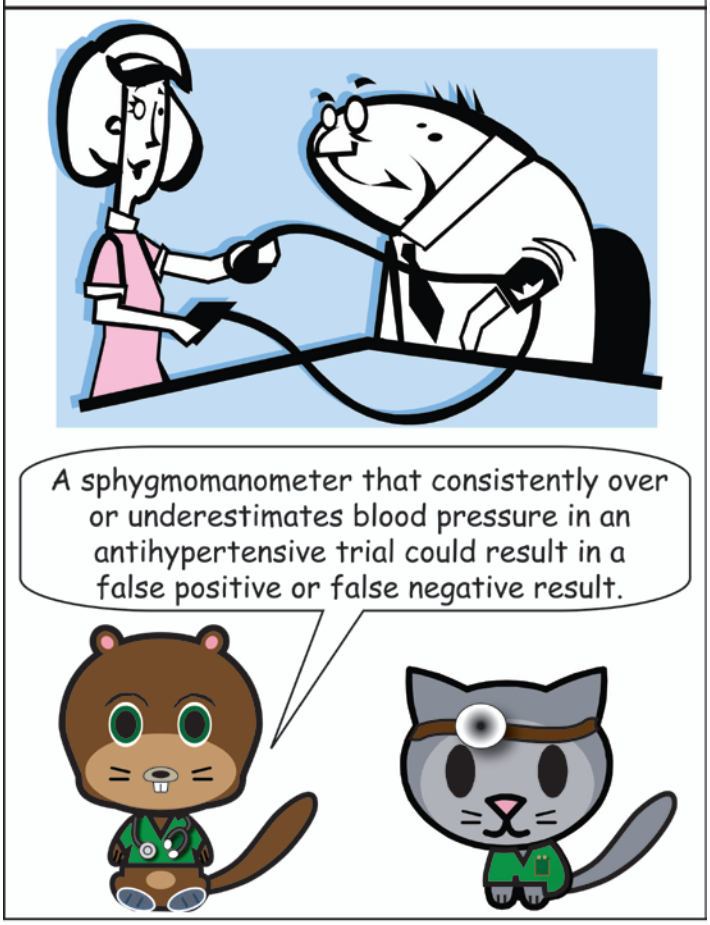


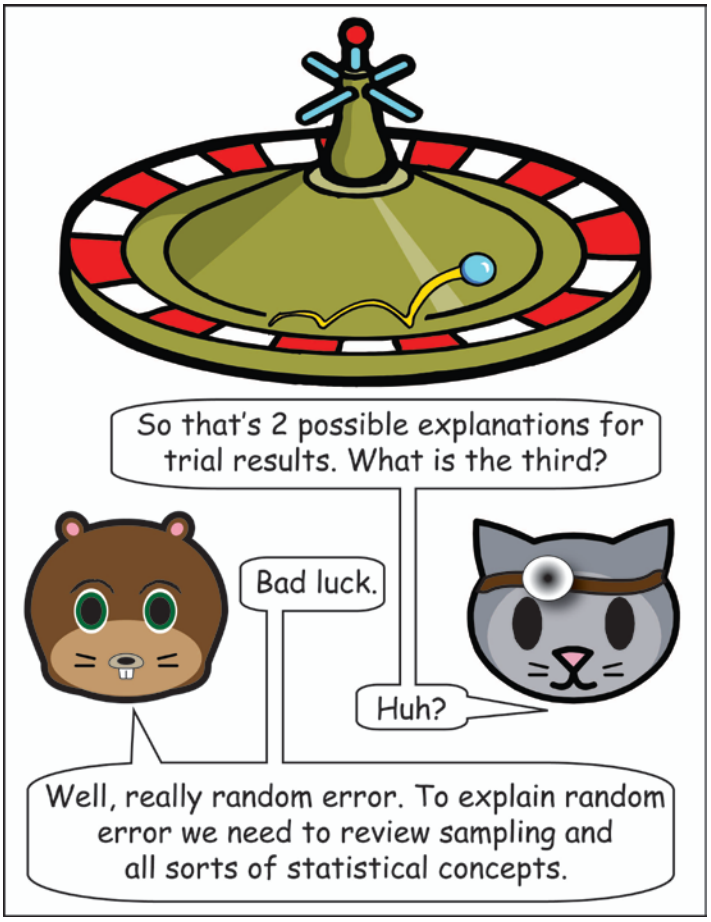


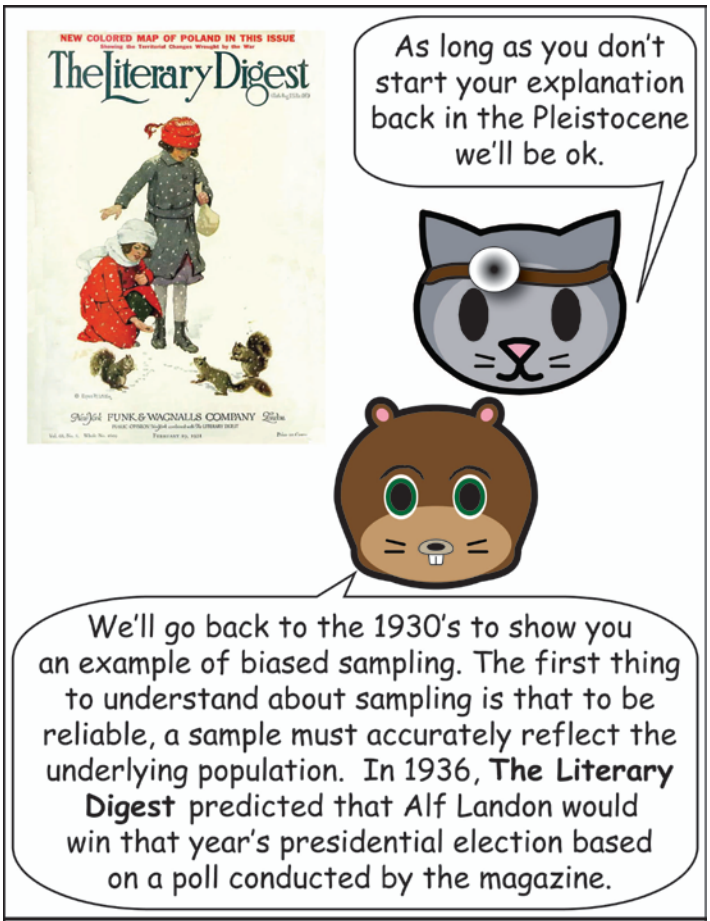
	Actual Rx Difference	No Rx Difference
Trial Positive	True Positive	False Positive
Trial Negative	False Negative	True Negative

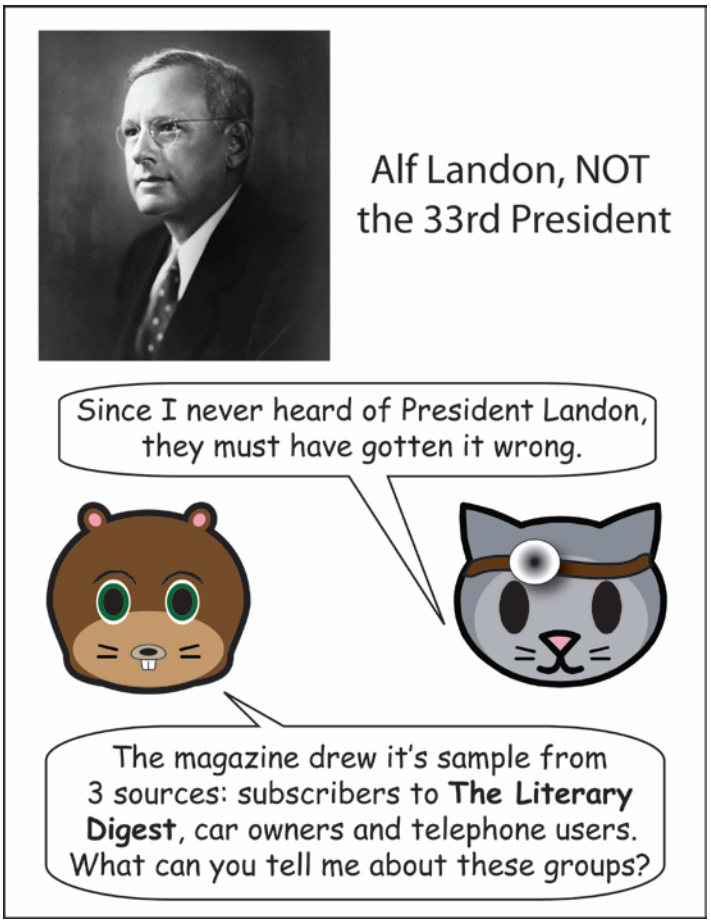

It's not funny if I have to explain it!
If there are true positive and negative results, there must also be false positives and negatives.
Yes. A false positive trial reports a difference between 2 interventions when there really isn't one. A false negative trial shows that 2 interventions have the same effect when they are actually different.

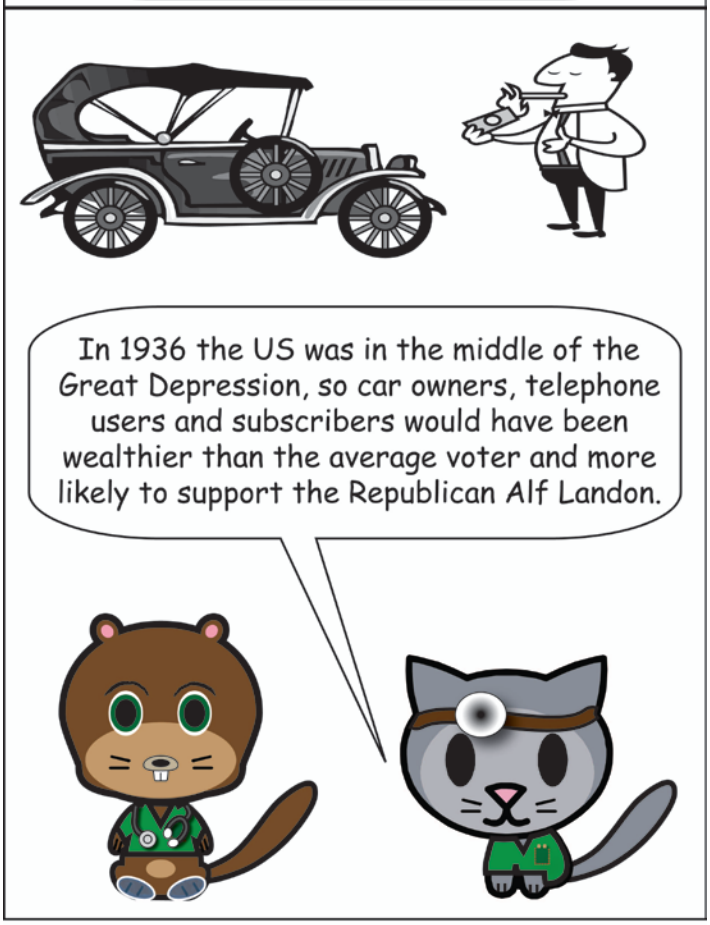

So how do you get false trial results?
One major reason is bias, which is any systematic error in data collection or interpretation.
Example please.

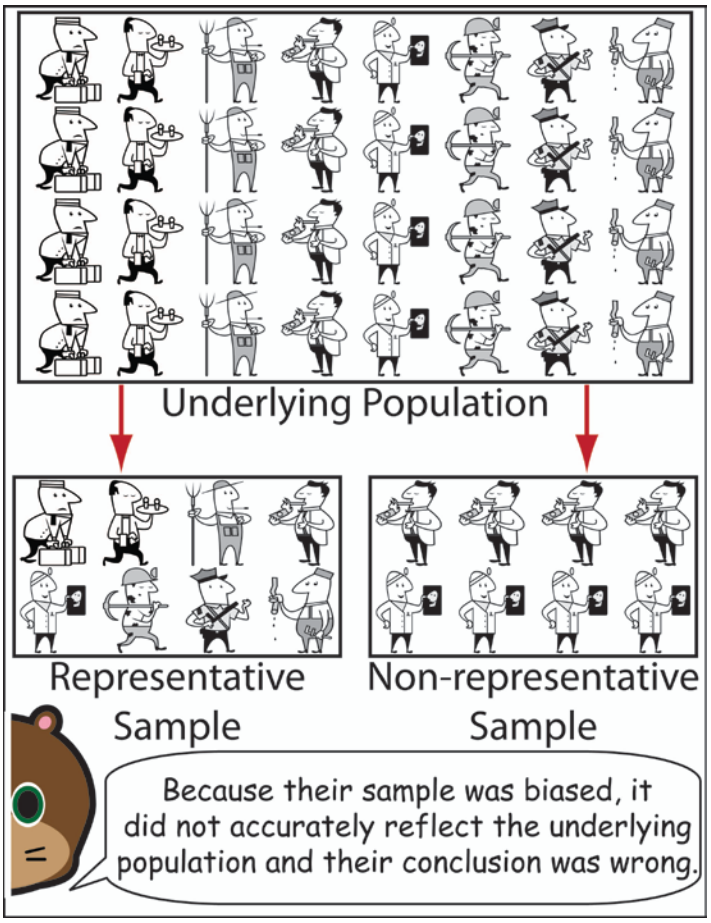

A sphygmomanometer that consistently over or underestimates blood pressure in an antihypertensive trial could result in a false positive or false negative result.


So that's 2 possible explanations for trial results. What is the third?
Bad luck.
Huh?
Well, really random error. To explain random error we need to review sampling and all sorts of statistical concepts.

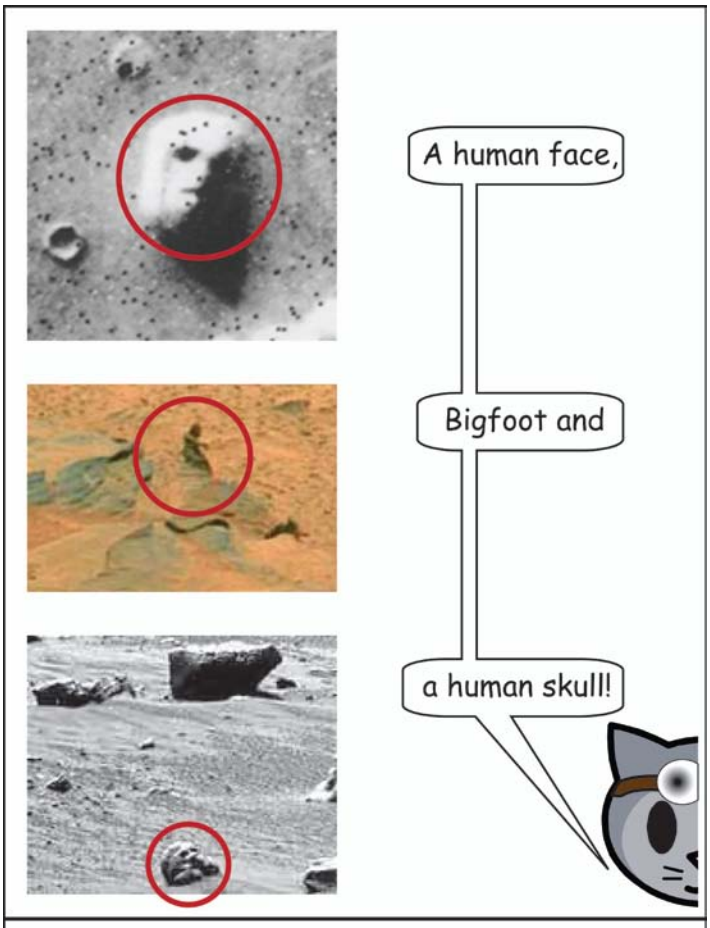

NEW COLORED MAP OF POLAND IN THIS ISSUE
The Literary Digest
FUNK & WAGNALLS COMPANY
As long as you don't start your explanation back in the Pleistocene we'll be ok.
We'll go back to the 1930's to show you an example of biased sampling. The first thing to understand about sampling is that to be reliable, a sample must accurately reflect the underlying population. In 1936, **The Literary Digest** predicted that Alf Landon would win that year's presidential election based on a poll conducted by the magazine.

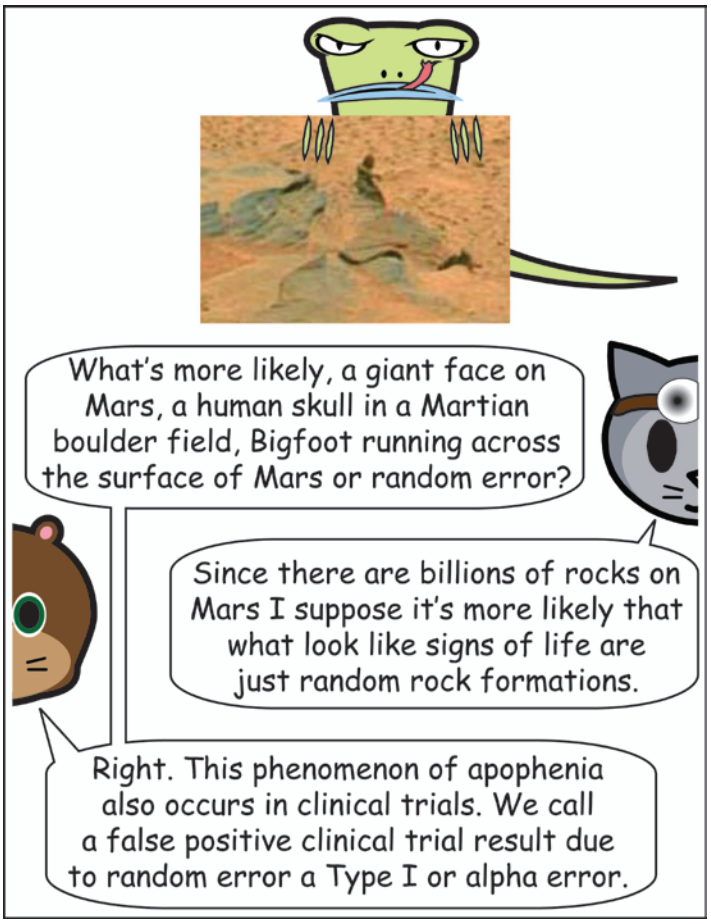

Alf Landon, NOT the 33rd President
Since I never heard of President Landon, they must have gotten it wrong.
The magazine drew it's sample from 3 sources: subscribers to **The Literary Digest**, car owners and telephone users. What can you tell me about these groups?

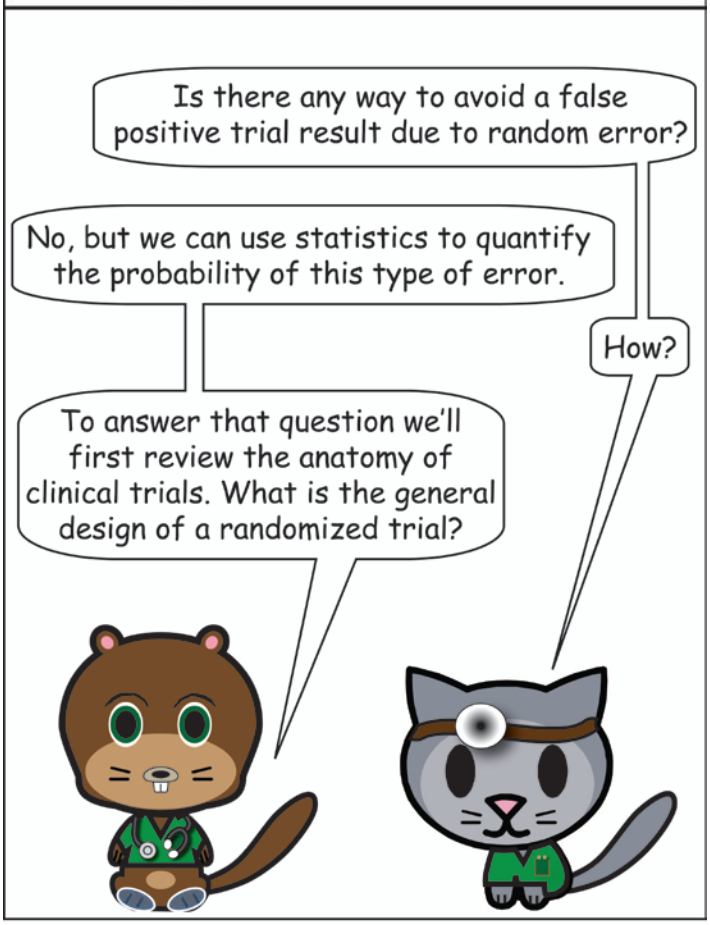

In 1936 the US was in the middle of the Great Depression, so car owners, telephone users and subscribers would have been wealthier than the average voter and more likely to support the Republican Alf Landon.

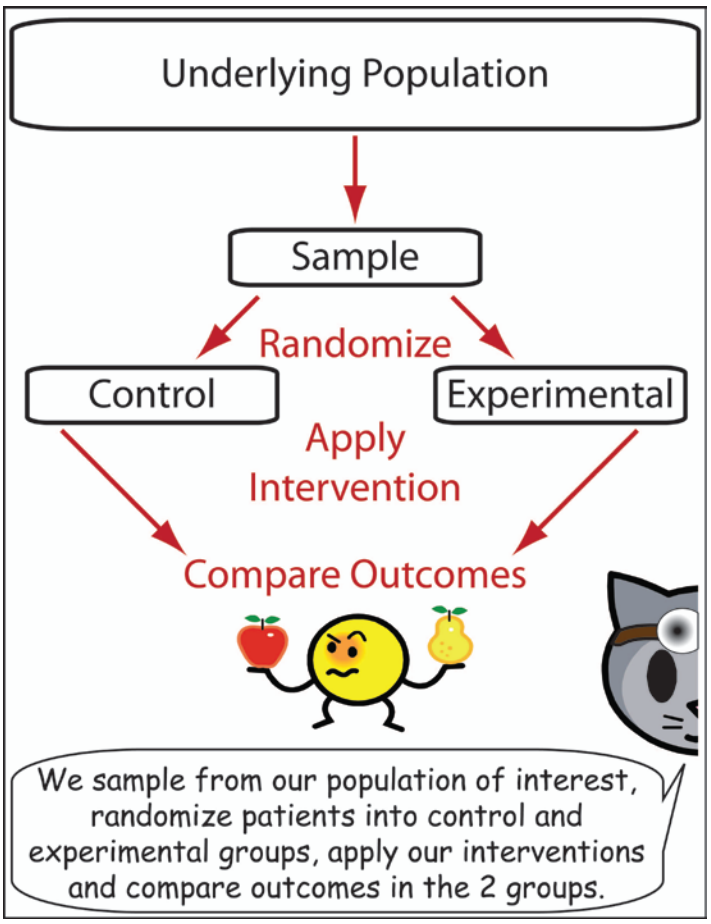

Underlying Population
Representative Sample
Non-representative Sample
Because their sample was biased, it did not accurately reflect the underlying population and their conclusion was wrong.

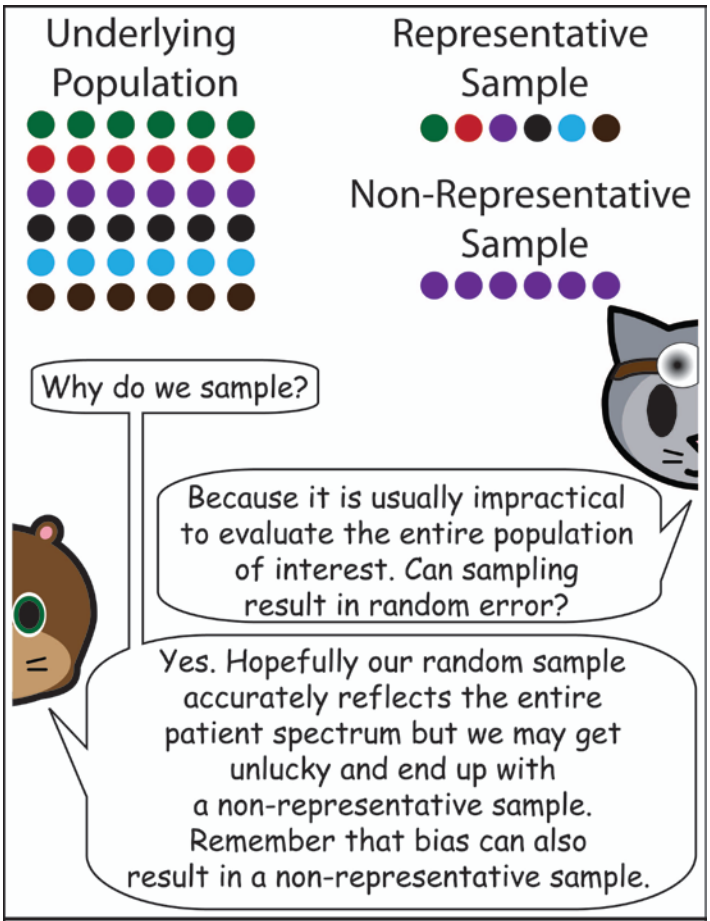

But they had a sample size over 2 million people!
It doesn't matter. No mathematical manipulation or sample size can compensate for bias. Now we will have to tackle random error and I am afraid we will have to return to the educational equivalent of the Pleistocene.
But I will start in the year 2004 to give you a feel for random error. Do you remember how much that grilled cheese sandwich with the Virgin Mary sold for on eBay?
Why?
$28,000!
I suppose some people thought the image was an actual miracle.
I agree, but how many grilled cheese sandwiches are made in the US every year?
Billions, probably.
Isn't it likely that the image of the Virgin Mary is the result of a random burn pattern on the bread that just happens to look vaguely human? After all, if there are billions of sandwiches made every year, chances are good that at least one will look like a person, or a goat,chicken, walrus or whatever.
I guess that means we won't be bidding on this pancake with Mother Theresa on it. It's a bargain!
The tendency to find spuriously meaningful patterns in random data is called apophenia. Check out these pictures from Mars. What do you see?

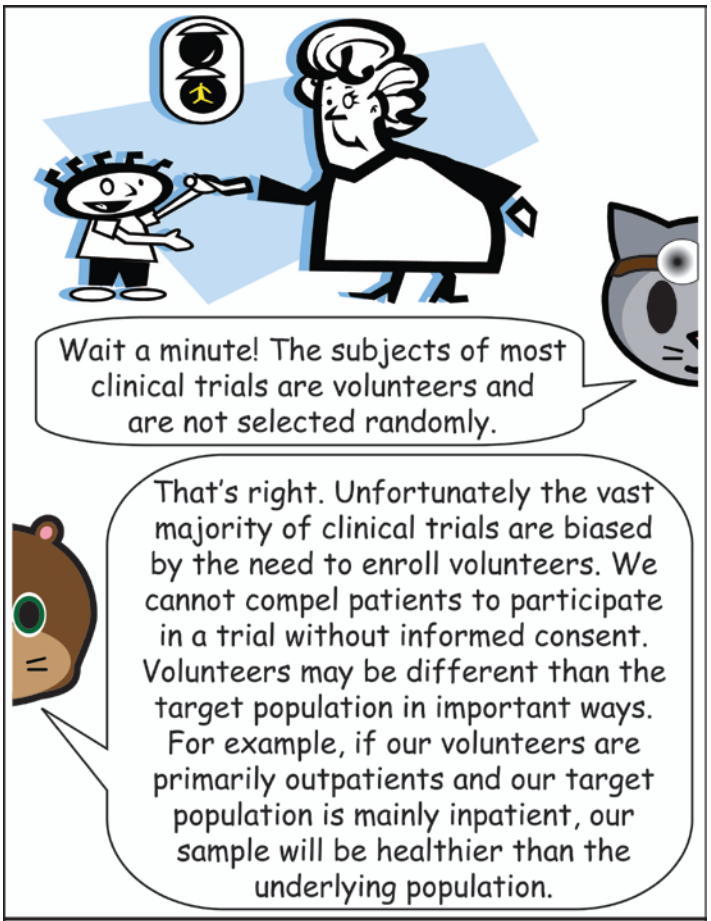

A human face,
Bigfoot and
a human skull!

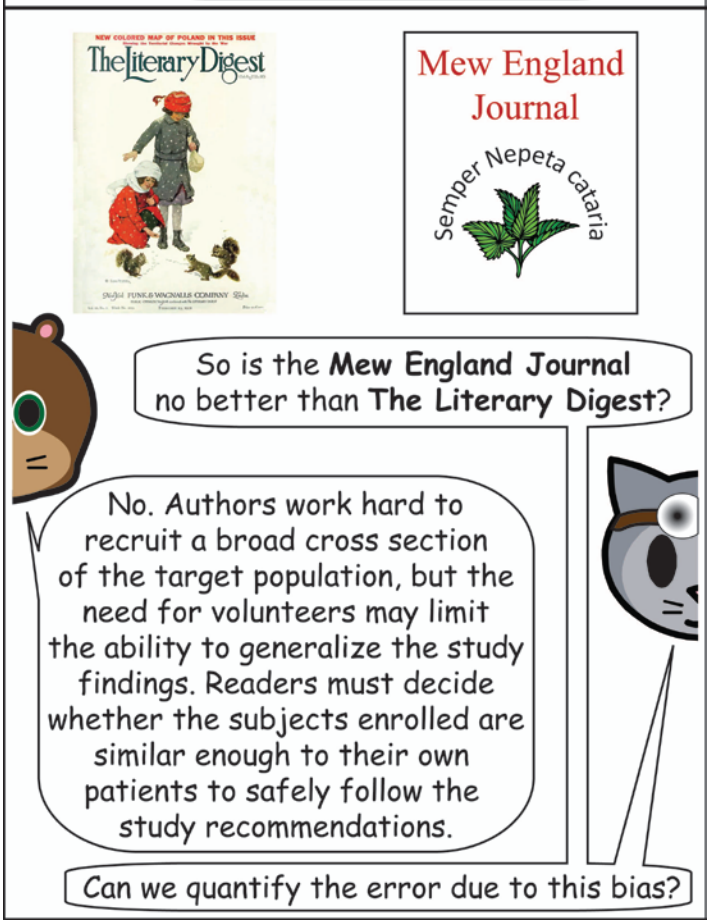

What's more likely, a giant face on Mars, a human skull in a Martian boulder field, Bigfoot running across the surface of Mars or random error?
Since there are billions of rocks on Mars I suppose it's more likely that what look like signs of life are just random rock formations.
Right. This phenomenon of apophenia also occurs in clinical trials. We call a false positive clinical trial result due to random error a Type I or alpha error.

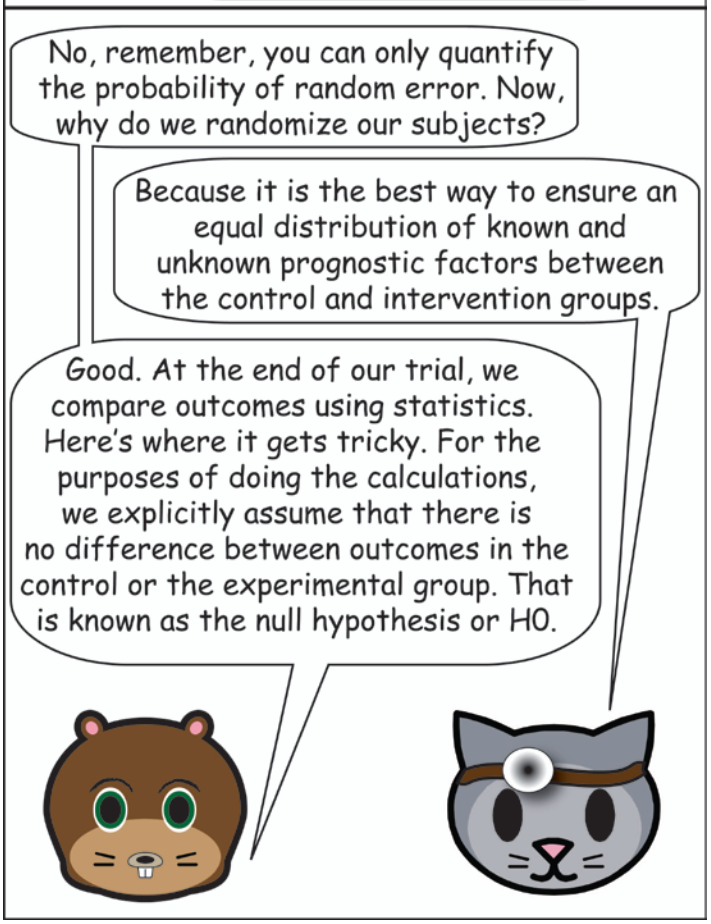

Is there any way to avoid a false positive trial result due to random error?
No, but we can use statistics to quantify the probability of this type of error.
How?
To answer that question we'll first review the anatomy of clinical trials. What is the general design of a randomized trial?

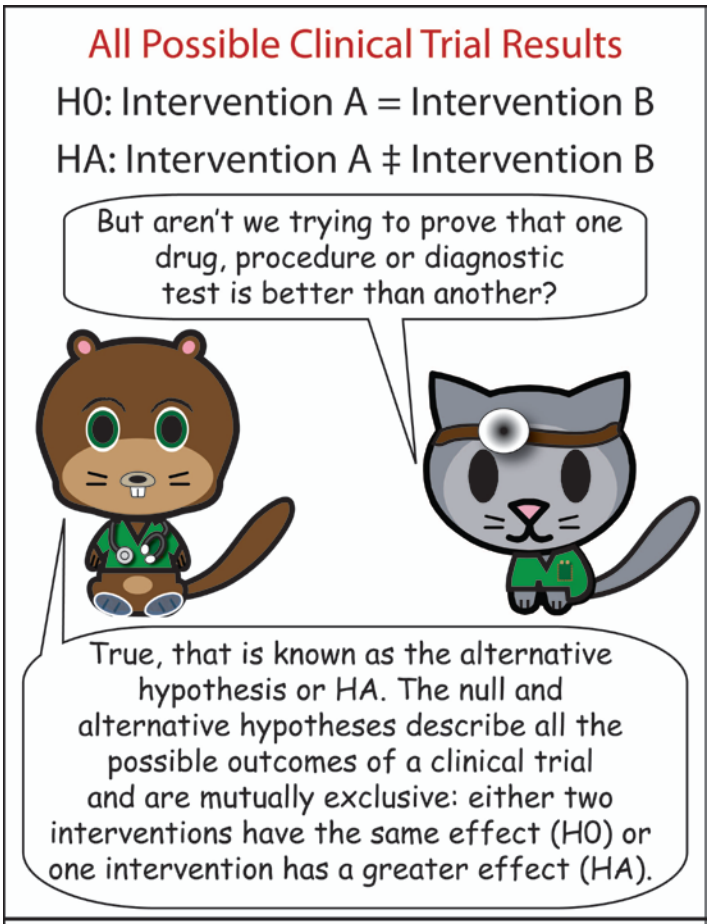

Underlying Population
Sample
Randomize
Control
Experimental
Apply Intervention
Compare Outcomes
We sample from our population of interest, randomize patients into control and experimental groups, apply our interventions and compare outcomes in the 2 groups.

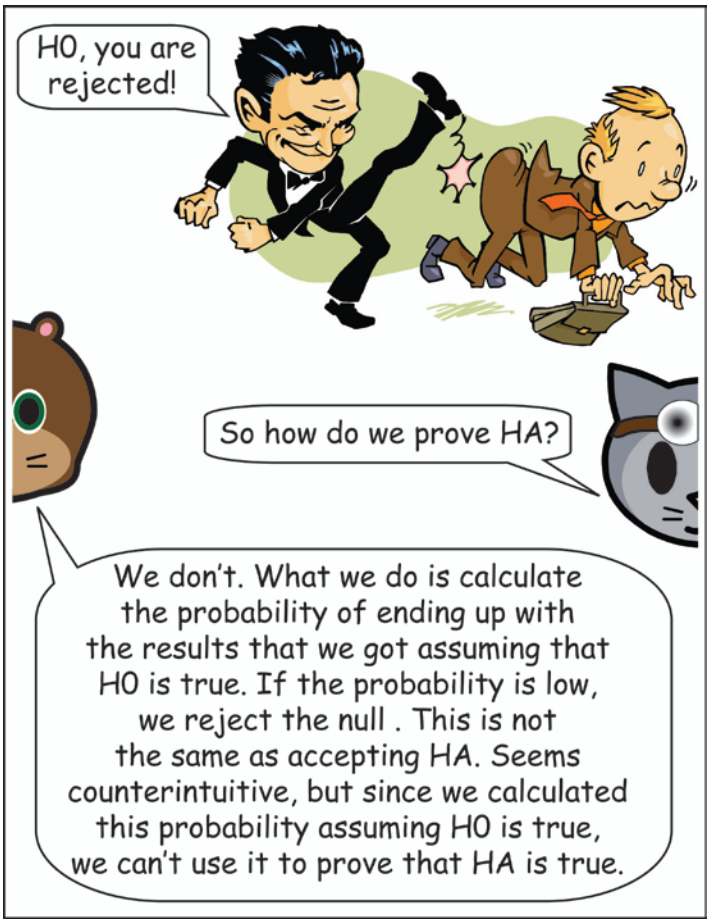

Underlying Population
Representative Sample
Non-Representative Sample
Why do we sample?
Because it is usually impractical to evaluate the entire population of interest. Can sampling result in random error?
Yes. Hopefully our random sample accurately reflects the entire patient spectrum but we may get unlucky and end up with a non-representative sample. Remember that bias can also result in a non-representative sample.

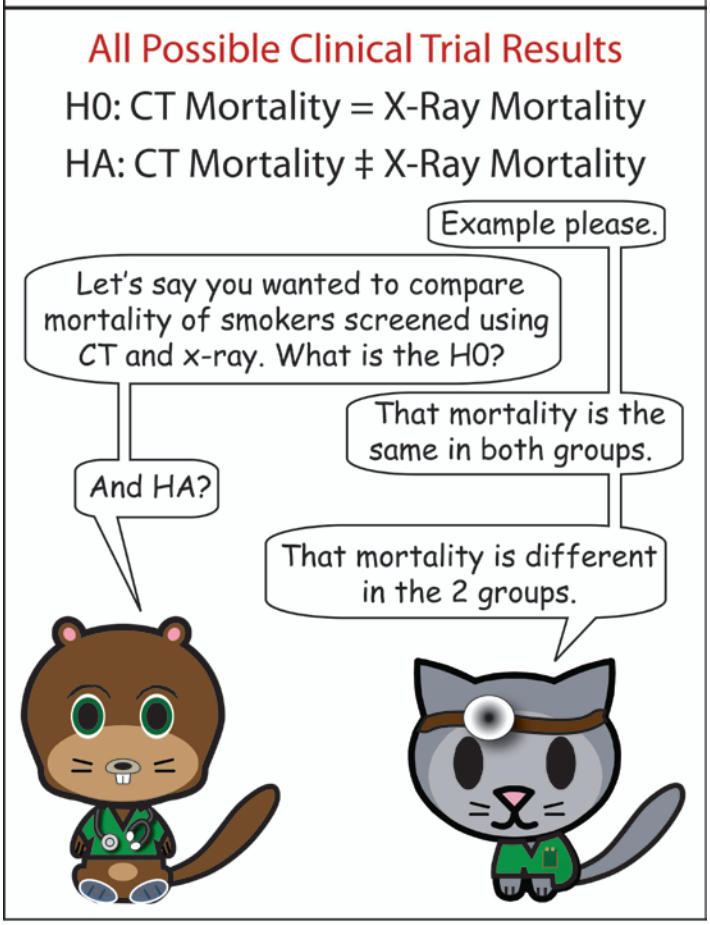

Wait a minute! The subjects of most clinical trials are volunteers and are not selected randomly.
That's right. Unfortunately the vast majority of clinical trials are biased by the need to enroll volunteers. We cannot compel patients to participate in a trial without informed consent. Volunteers may be different than the target population in important ways. For example, if our volunteers are primarily outpatients and our target population is mainly inpatient, our sample will be healthier than the underlying population.

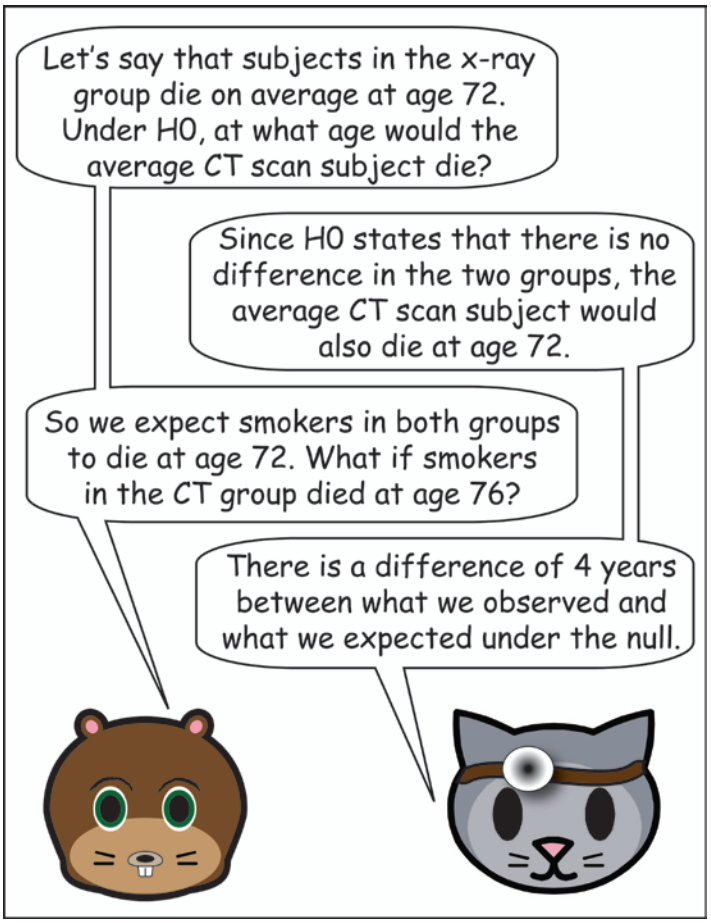

The Literary Digest
Mew England Journal
Semper Nepeta cataria
So is the Mew England Journal no better than The Literary Digest?
No. Authors work hard to recruit a broad cross section of the target population, but the need for volunteers may limit the ability to generalize the study findings. Readers must decide whether the subjects enrolled are similar enough to their own patients to safely follow the study recommendations.
Can we quantify the error due to this bias?

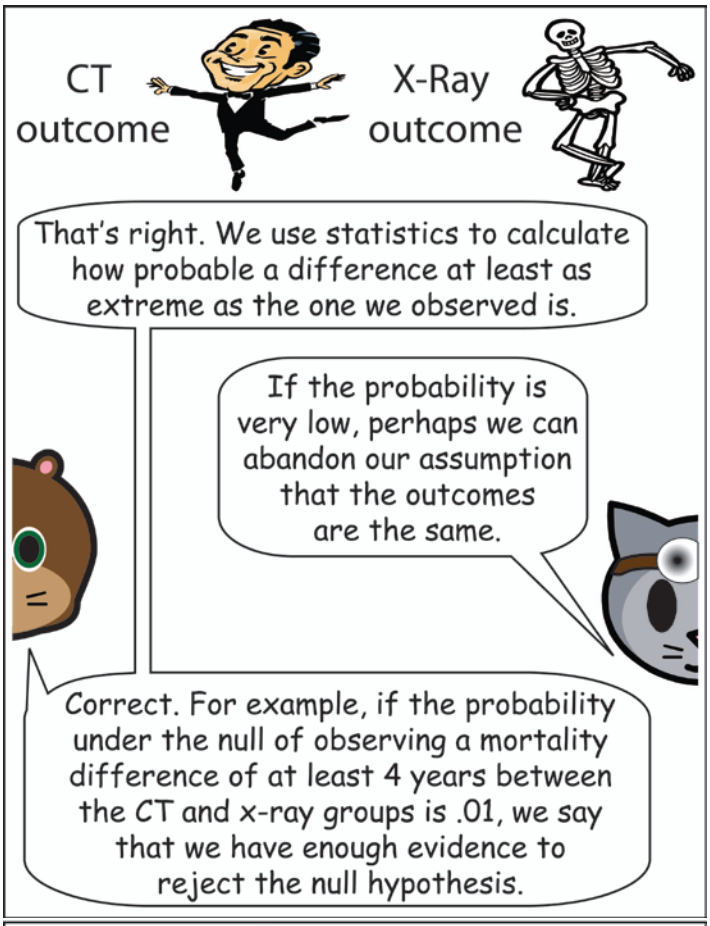

No, remember, you can only quantify the probability of random error. Now, why do we randomize our subjects?
Because it is the best way to ensure an equal distribution of known and unknown prognostic factors between the control and intervention groups.
Good. At the end of our trial, we compare outcomes using statistics. Here's where it gets tricky. For the purposes of doing the calculations, we explicitly assume that there is no difference between outcomes in the control or the experimental group. That is known as the null hypothesis or H0.

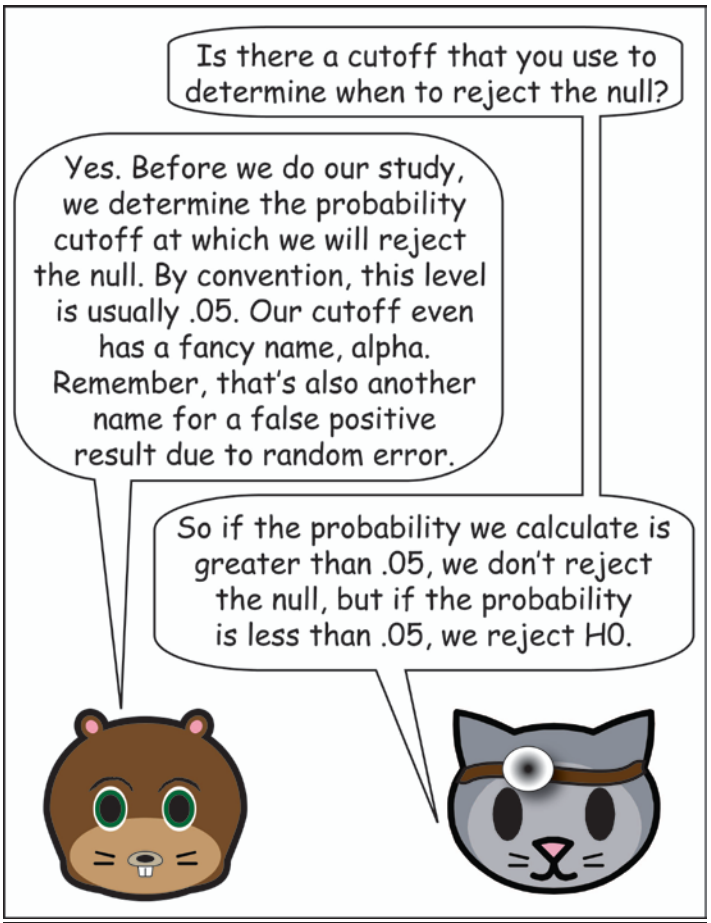

All Possible Clinical Trial Results
H0: Intervention A = Intervention B
HA: Intervention A ‡ Intervention B
But aren't we trying to prove that one drug, procedure or diagnostic test is better than another?
True, that is known as the alternative hypothesis or HA. The null and alternative hypotheses describe all the possible outcomes of a clinical trial and are mutually exclusive: either two interventions have the same effect (H0) or one intervention has a greater effect (HA).

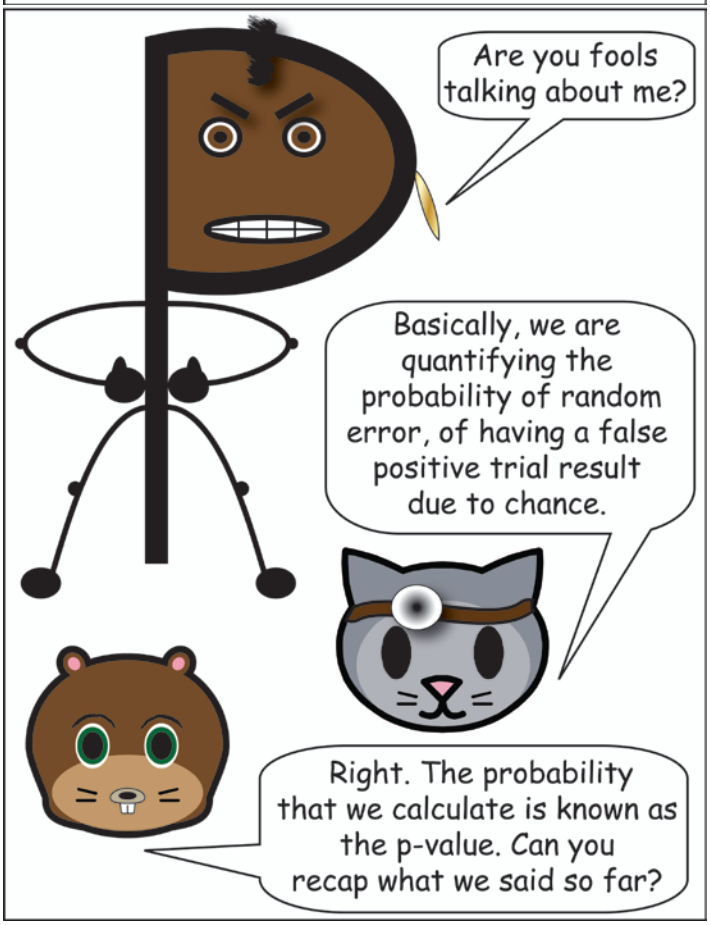

H0, you are rejected!
So how do we prove HA?
We don't. What we do is calculate the probability of ending up with the results that we got assuming that H0 is true. If the probability is low, we reject the null . This is not the same as accepting HA. Seems counterintuitive, but since we calculated this probability assuming H0 is true, we can't use it to prove that HA is true.

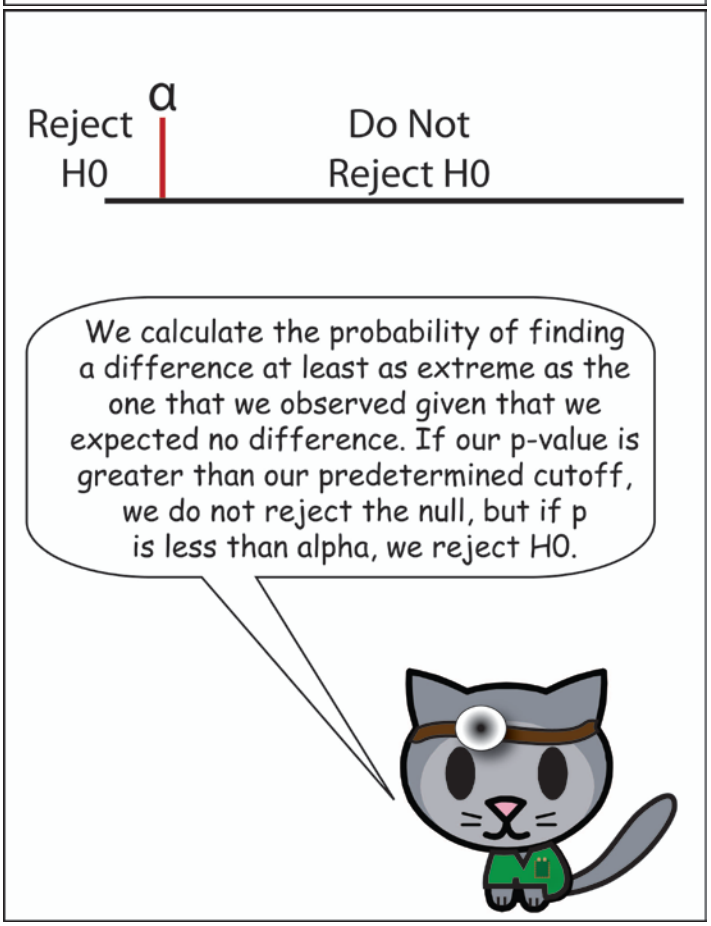

All Possible Clinical Trial Results
H0: CT Mortality = X-Ray Mortality
HA: CT Mortality ‡ X-Ray Mortality
Example please.
Let's say you wanted to compare mortality of smokers screened using CT and x-ray. What is the H0?
That mortality is the same in both groups.
And HA?
That mortality is different in the 2 groups.


Let's say that subjects in the x-ray group die on average at age 72. Under H0, at what age would the average CT scan subject die?
Since H0 states that there is no difference in the two groups, the average CT scan subject would also die at age 72.
So we expect smokers in both groups to die at age 72. What if smokers in the CT group died at age 76?
There is a difference of 4 years between what we observed and what we expected under the null.

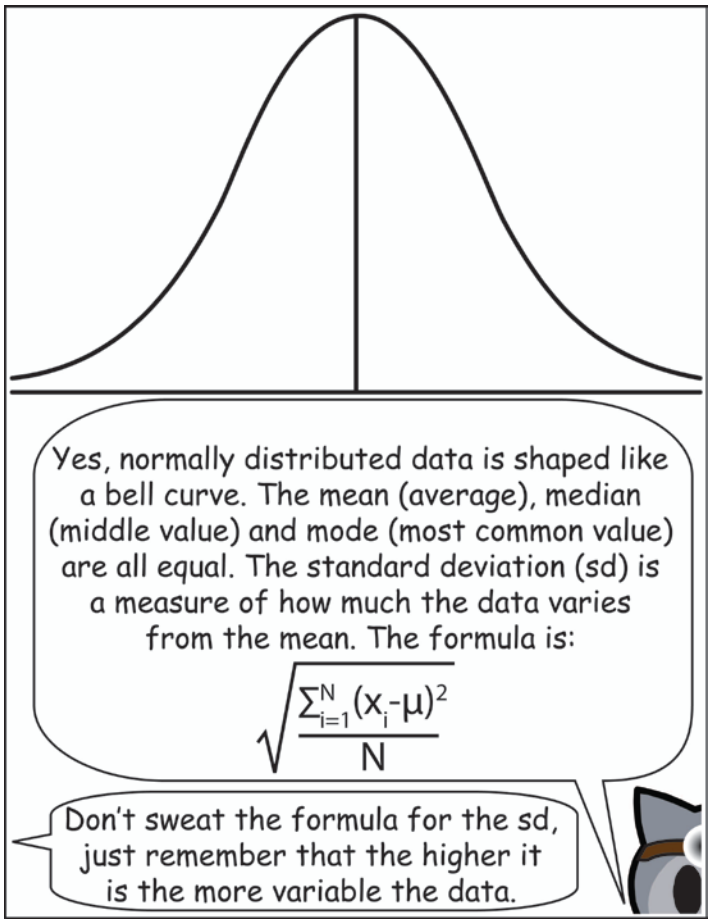

CT outcome
X-Ray outcome
That's right. We use statistics to calculate how probable a difference at least as extreme as the one we observed is.
If the probability is very low, perhaps we can abandon our assumption that the outcomes are the same.
Correct. For example, if the probability under the null of observing a mortality difference of at least 4 years between the CT and x-ray groups is .01, we say that we have enough evidence to reject the null hypothesis.

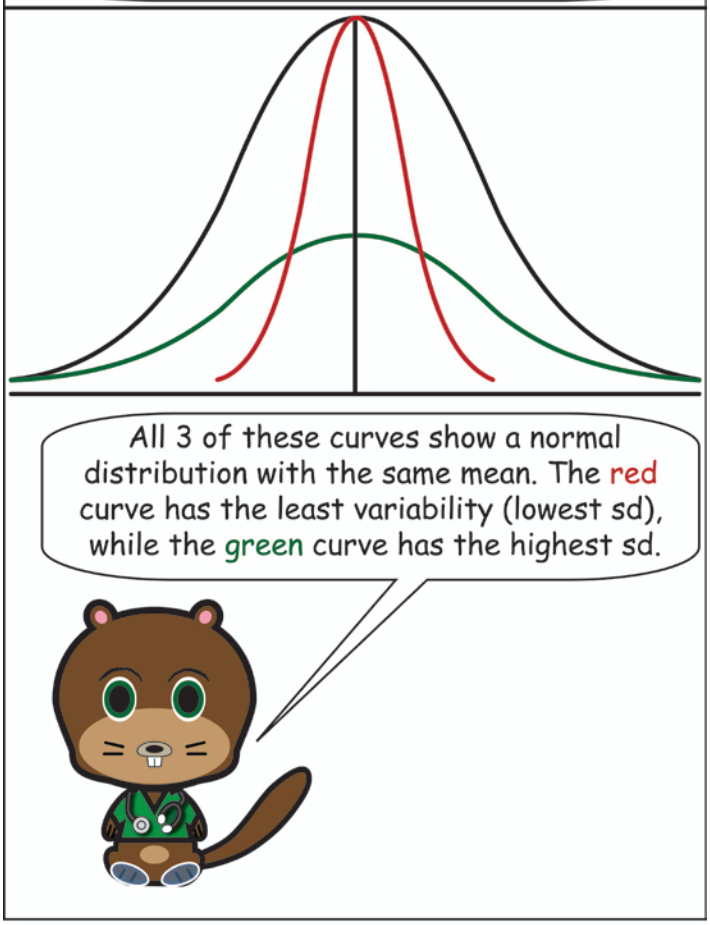

Is there a cutoff that you use to determine when to reject the null?
Yes. Before we do our study, we determine the probability cutoff at which we will reject the null. By convention, this level is usually .05. Our cutoff even has a fancy name, alpha. Remember, that's also another name for a false positive result due to random error.
So if the probability we calculate is greater than .05, we don't reject the null, but if the probability is less than .05, we reject H0.

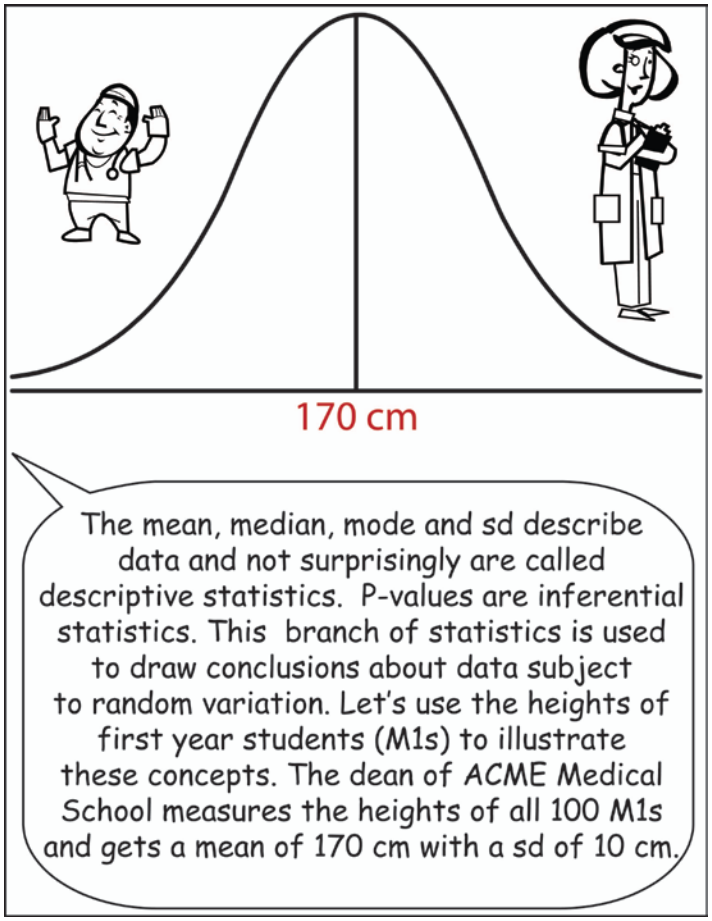

Are you fools talking about me?
Basically, we are quantifying the probability of random error, of having a false positive trial result due to chance.
Right. The probability that we calculate is known as the p-value. Can you recap what we said so far?

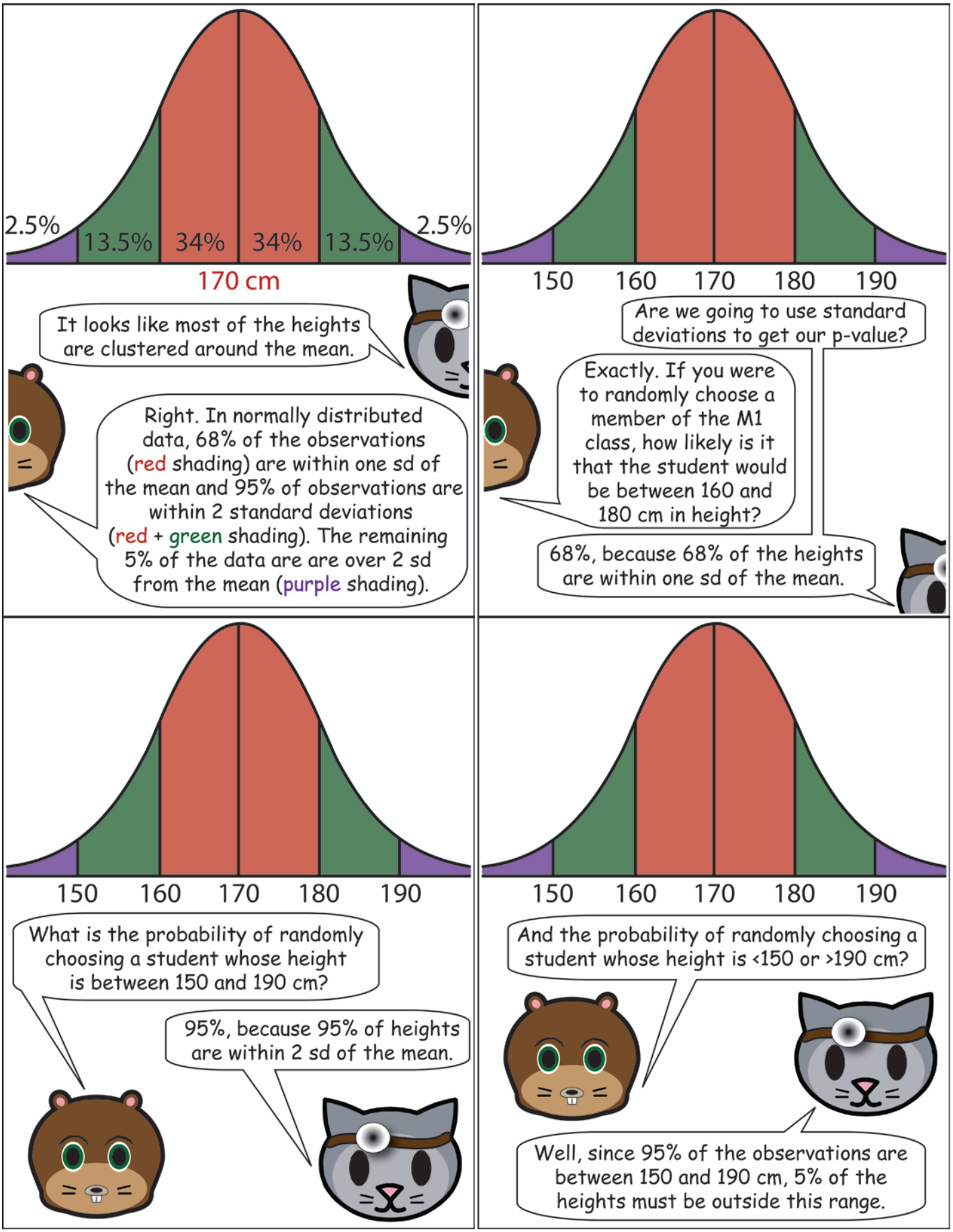

Reject H0
α
Do Not Reject H0
We calculate the probability of finding a difference at least as extreme as the one that we observed given that we expected no difference. If our p-value is greater than our predetermined cutoff, we do not reject the null, but if p is less than alpha, we reject H0.

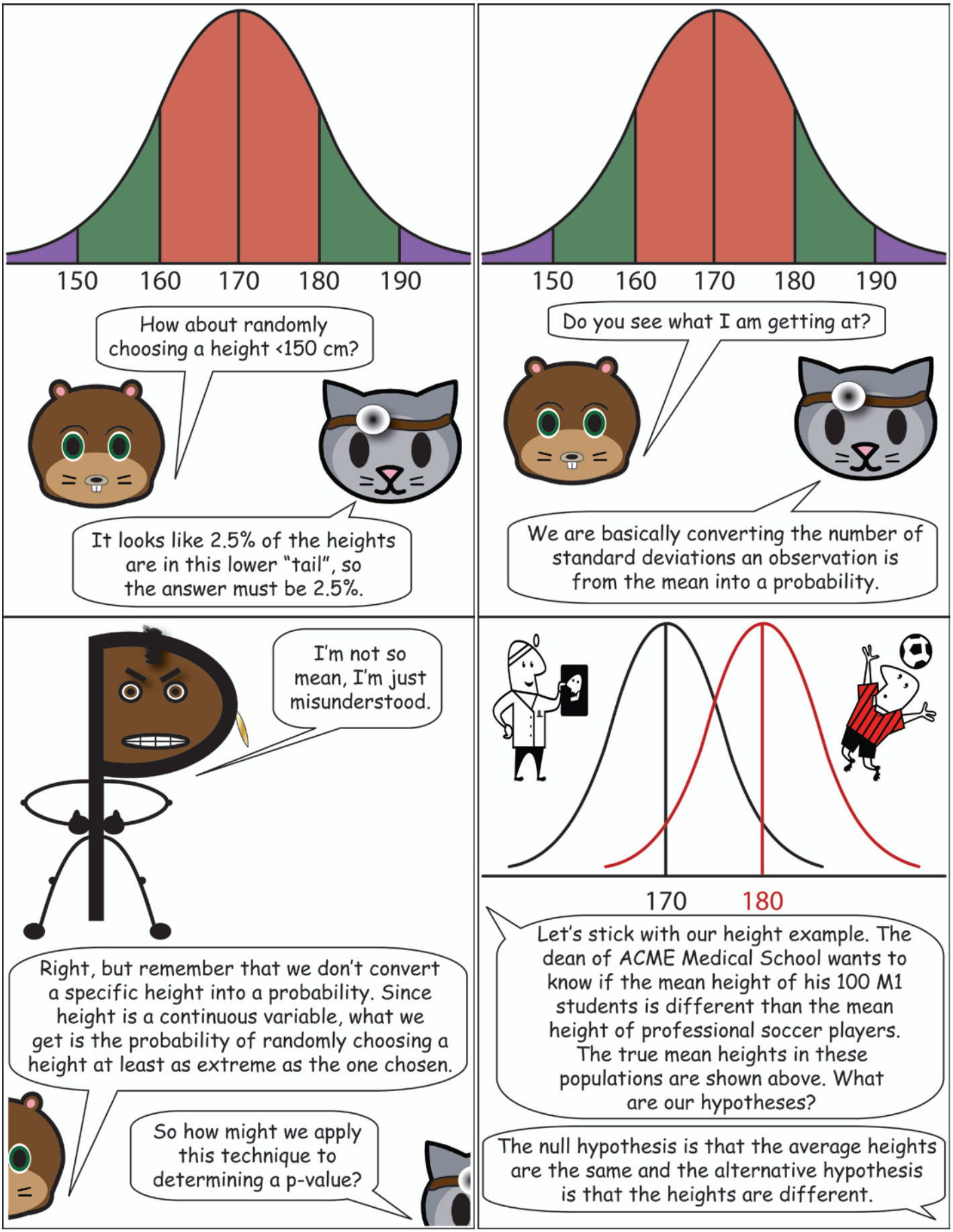

H0 False!
In other words, we decide whether the null hypothesis is plausible when we compare our expectation of no difference with what we actually observed.
So how do we determine a p-value?
It's not difficult, but we'll have to start by reviewing the normal distribution. Do you remember what that is?

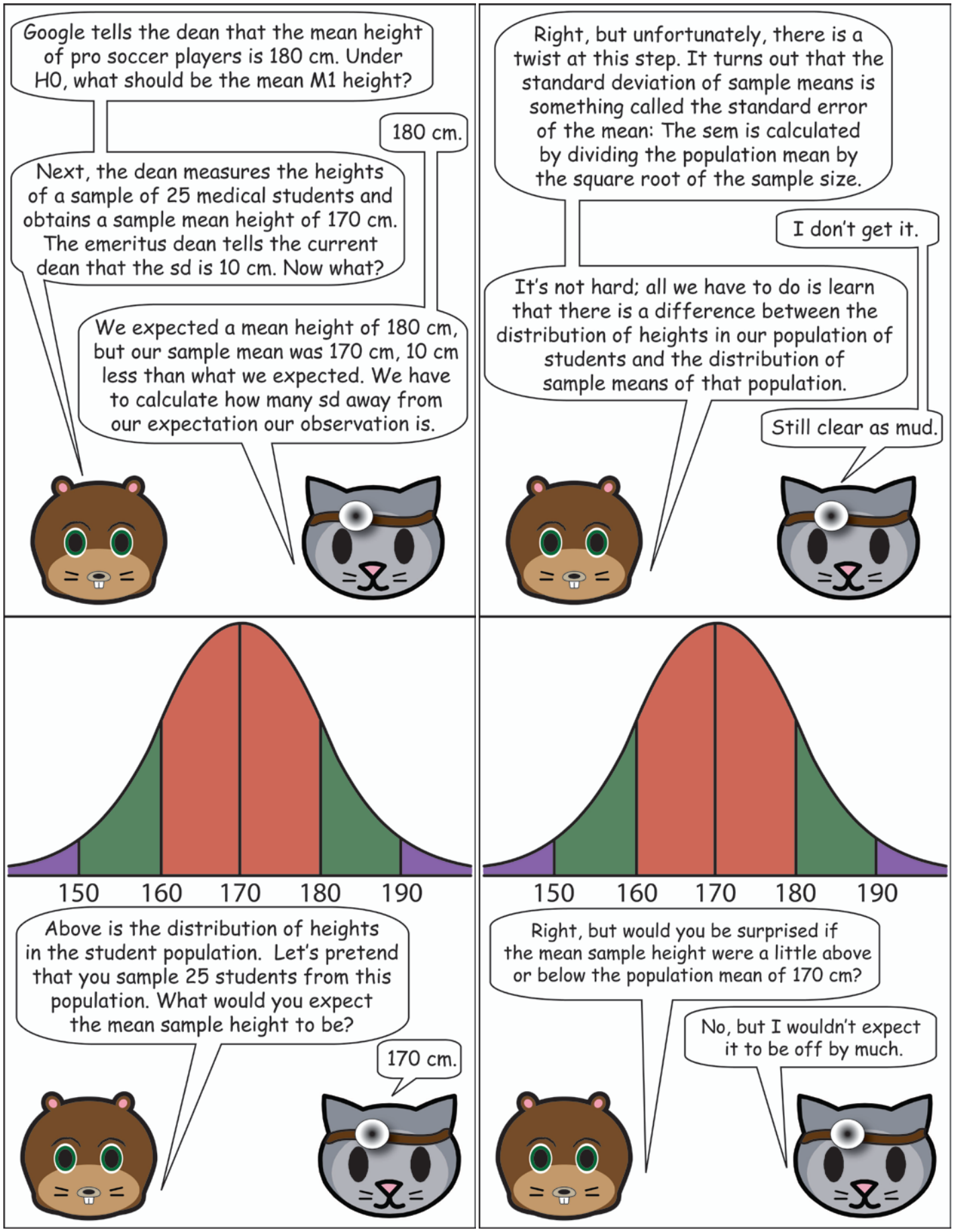

Yes, normally distributed data is shaped like a bell curve. The mean (average), median (middle value) and mode (most common value) are all equal. The standard deviation (sd) is a measure of how much the data varies from the mean. The formula is:
$$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$
Don't sweat the formula for the sd, just remember that the higher it is the more variable the data.

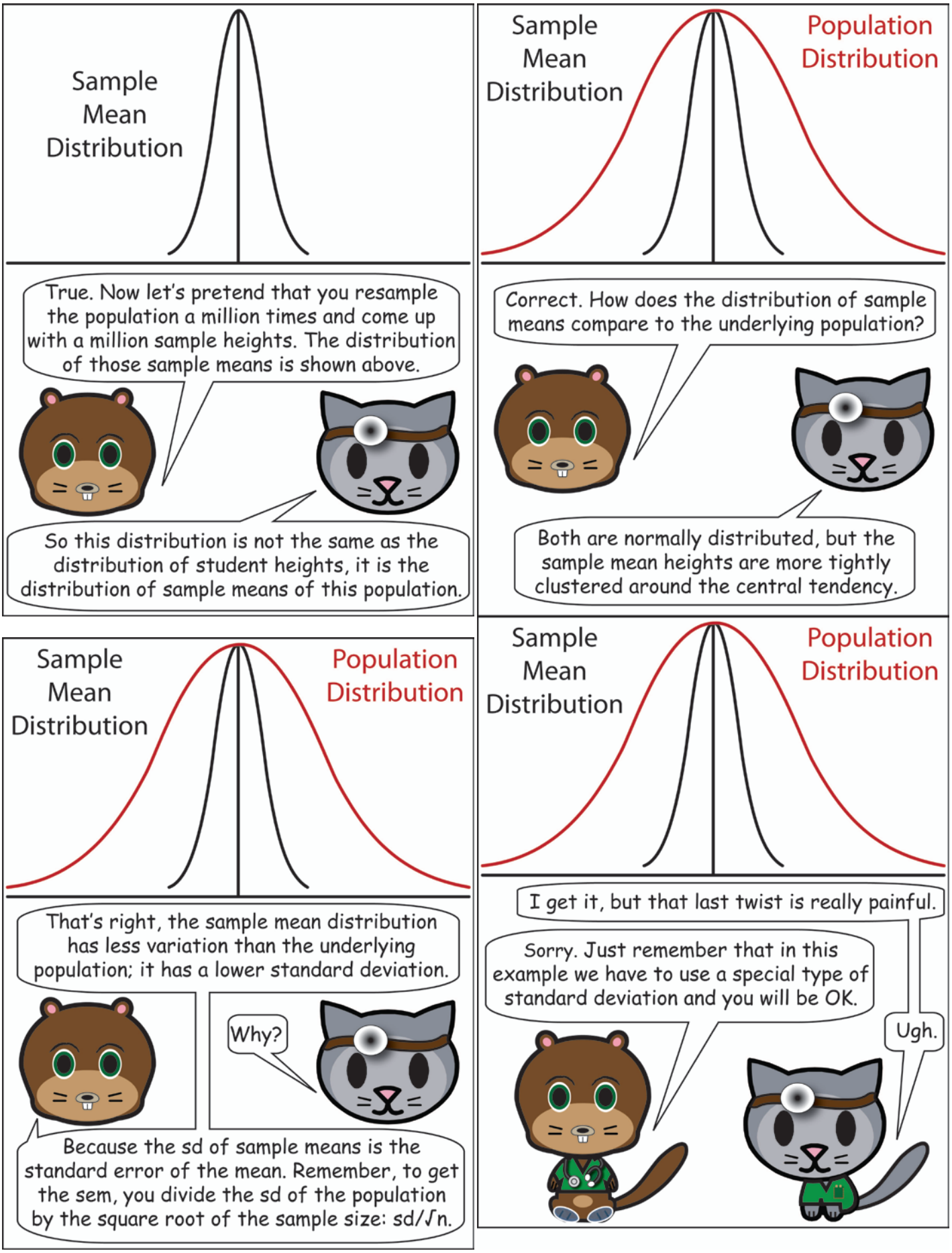

All 3 of these curves show a normal distribution with the same mean. The red curve has the least variability (lowest sd), while the green curve has the highest sd.

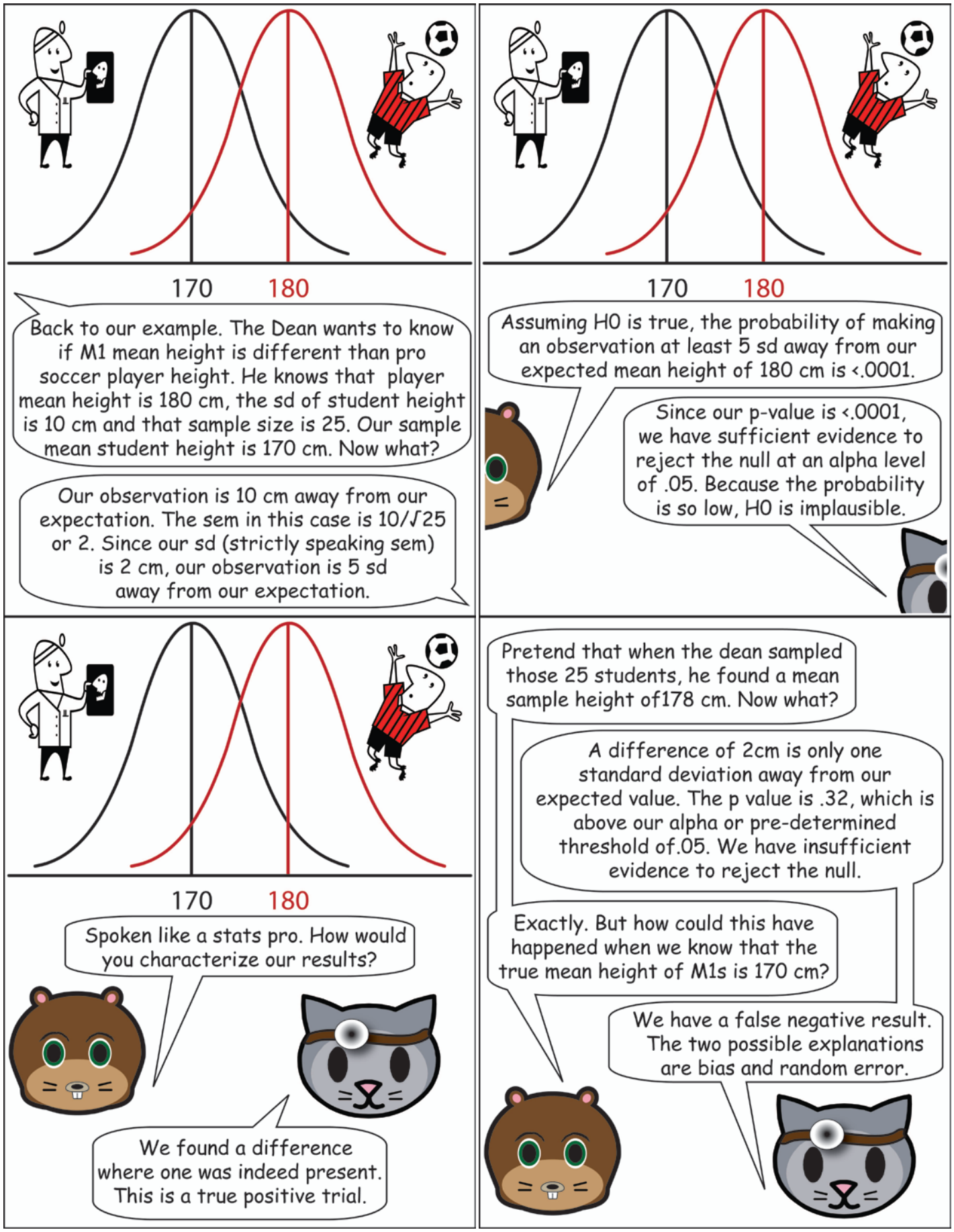

170 cm
The mean, median, mode and sd describe data and not surprisingly are called descriptive statistics. P-values are inferential statistics. This branch of statistics is used to draw conclusions about data subject to random variation. Let's use the heights of first year students (M1s) to illustrate these concepts. The dean of ACME Medical School measures the heights of all 100 M1s and gets a mean of 170 cm with a sd of 10 cm.


2.5%
13.5%
34%
34%
13.5%
2.5%
170 cm
It looks like most of the heights are clustered around the mean.
Right. In normally distributed data, 68% of the observations (red shading) are within one sd of the mean and 95% of observations are within 2 standard deviations (red + green shading). The remaining 5% of the data are are over 2 sd from the mean (purple shading).
150 160 170 180 190
Are we going to use standard deviations to get our p-value?
Exactly. If you were to randomly choose a member of the M1 class, how likely is it that the student would be between 160 and 180 cm in height?
68%, because 68% of the heights are within one sd of the mean.
150 160 170 180 190
What is the probability of randomly choosing a student whose height is between 150 and 190 cm?
95%, because 95% of heights are within 2 sd of the mean.
150 160 170 180 190
And the probability of randomly choosing a student whose height is <150 or >190 cm?
Well, since 95% of the observations are between 150 and 190 cm, 5% of the heights must be outside this range.


150 160 170 180 190
How about randomly choosing a height <150 cm?
It looks like 2.5% of the heights are in this lower "tail", so the answer must be 2.5%.
150 160 170 180 190
Do you see what I am getting at?
We are basically converting the number of standard deviations an observation is from the mean into a probability.
I'm not so mean, I'm just misunderstood.
Right, but remember that we don't convert a specific height into a probability. Since height is a continuous variable, what we get is the probability of randomly choosing a height at least as extreme as the one chosen.
So how might we apply this technique to determining a p-value?
170 180
Let's stick with our height example. The dean of ACME Medical School wants to know if the mean height of his 100 M1 students is different than the mean height of professional soccer players. The true mean heights in these populations are shown above. What are our hypotheses?
The null hypothesis is that the average heights are the same and the alternative hypothesis is that the heights are different.

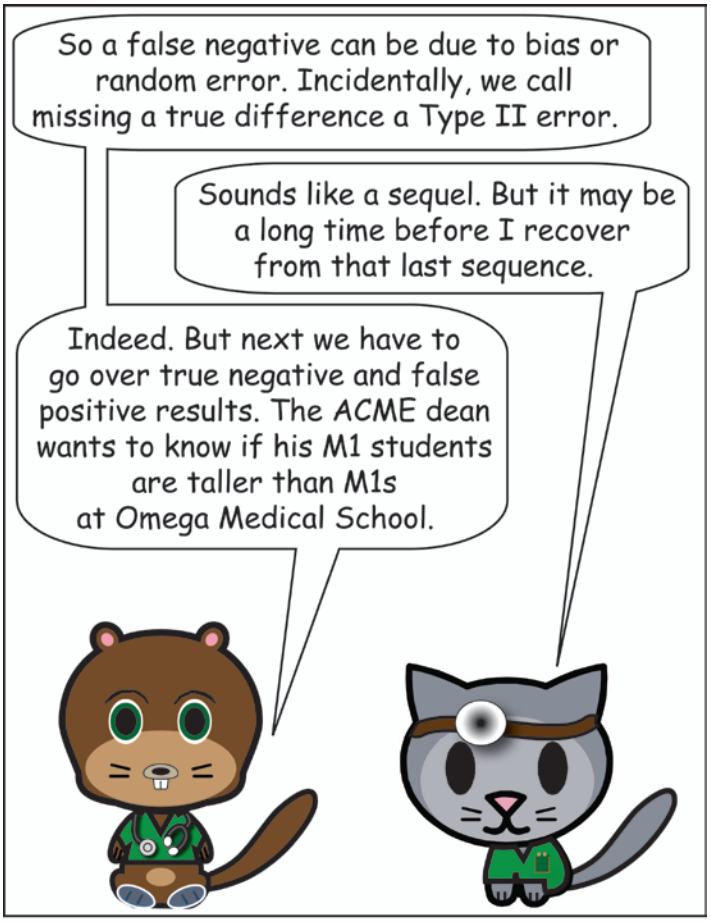

Google tells the dean that the mean height of pro soccer players is 180 cm. Under H0, what should be the mean M1 height?
180 cm.
Next, the dean measures the heights of a sample of 25 medical students and obtains a sample mean height of 170 cm. The emeritus dean tells the current dean that the sd is 10 cm. Now what?
We expected a mean height of 180 cm, but our sample mean was 170 cm, 10 cm less than what we expected. We have to calculate how many sd away from our expectation our observation is.
Right, but unfortunately, there is a twist at this step. It turns out that the standard deviation of sample means is something called the standard error of the mean: The sem is calculated by dividing the population mean by the square root of the sample size.
I don't get it.
It's not hard; all we have to do is learn that there is a difference between the distribution of heights in our population of students and the distribution of sample means of that population.
Still clear as mud.
150 160 170 180 190
Above is the distribution of heights in the student population. Let's pretend that you sample 25 students from this population. What would you expect the mean sample height to be?
170 cm.
150 160 170 180 190
Right, but would you be surprised if the mean sample height were a little above or below the population mean of 170 cm?
No, but I wouldn't expect it to be off by much.

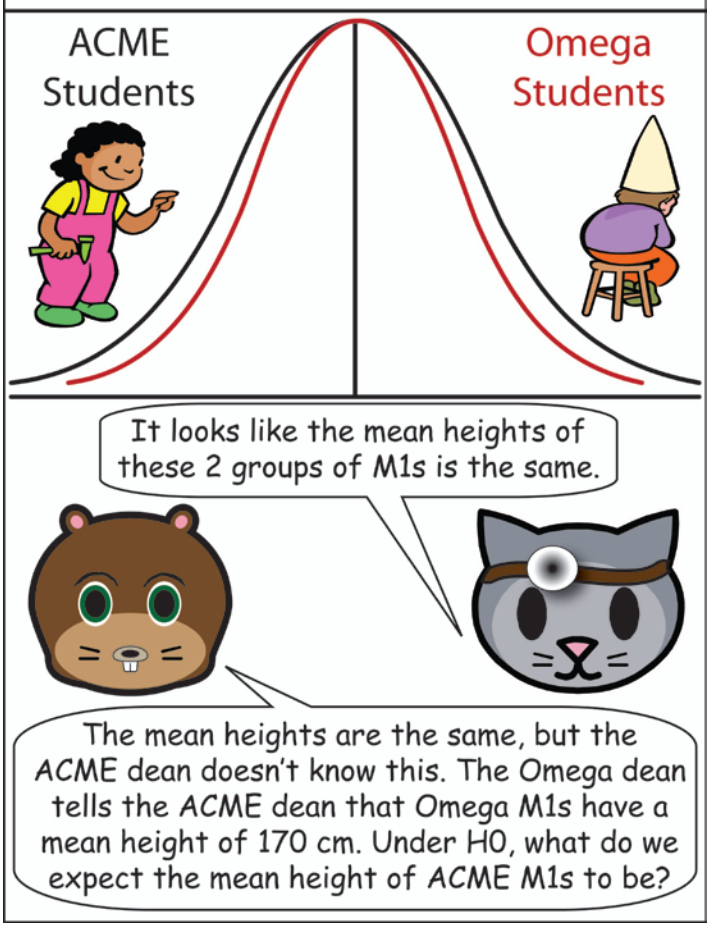

Sample Mean Distribution
True. Now let's pretend that you resample the population a million times and come up with a million sample heights. The distribution of those sample means is shown above.
So this distribution is not the same as the distribution of student heights, it is the distribution of sample means of this population.
Sample Mean Distribution
Population Distribution
Correct. How does the distribution of sample means compare to the underlying population?
Both are normally distributed, but the sample mean heights are more tightly clustered around the central tendency.
Sample Mean Distribution
Population Distribution
That's right, the sample mean distribution has less variation than the underlying population; it has a lower standard deviation.
Why?
Because the sd of sample means is the standard error of the mean. Remember, to get the sem, you divide the sd of the population by the square root of the sample size: sd/√n.
Sample Mean Distribution
Population Distribution
I get it, but that last twist is really painful.
Sorry. Just remember that in this example we have to use a special type of standard deviation and you will be OK.
Ugh.

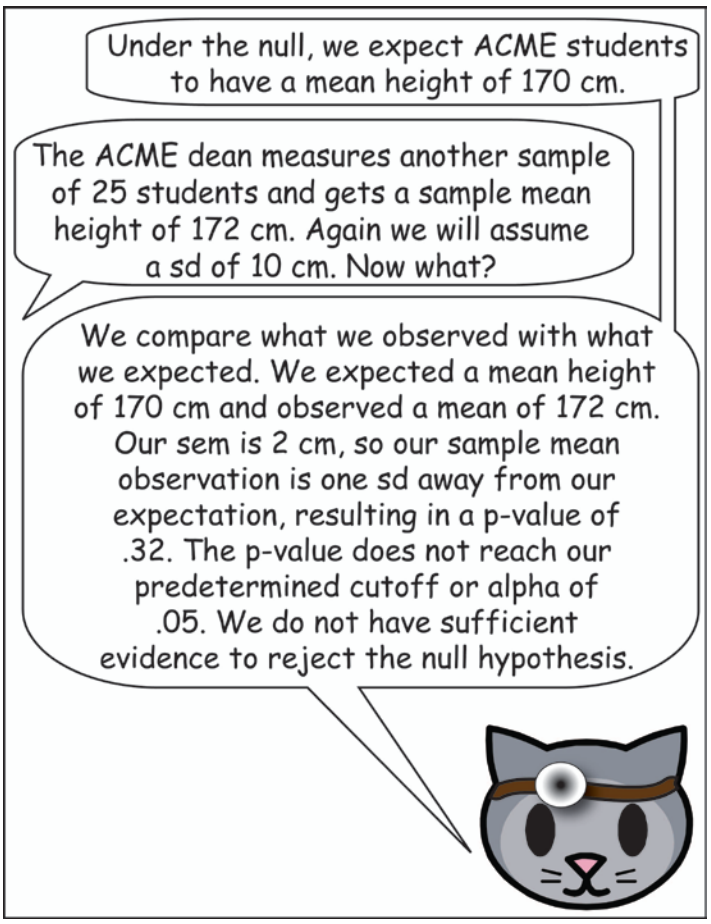

170
180
Back to our example. The Dean wants to know if M1 mean height is different than pro soccer player height. He knows that player mean height is 180 cm, the sd of student height is 10 cm and that sample size is 25. Our sample mean student height is 170 cm. Now what?
Our observation is 10 cm away from our expectation. The sem in this case is 10/√25 or 2. Since our sd (strictly speaking sem) is 2 cm, our observation is 5 sd away from our expectation.
170
180
Assuming H0 is true, the probability of making an observation at least 5 sd away from our expected mean height of 180 cm is <.0001.
Since our p-value is <.0001, we have sufficient evidence to reject the null at an alpha level of .05. Because the probability is so low, H0 is implausible.
170
180
Spoken like a stats pro. How would you characterize our results?
We found a difference where one was indeed present. This is a true positive trial.
Pretend that when the dean sampled those 25 students, he found a mean sample height of178 cm. Now what?
A difference of 2cm is only one standard deviation away from our expected value. The p value is .32, which is above our alpha or pre-determined threshold of.05. We have insufficient evidence to reject the null.
Exactly. But how could this have happened when we know that the true mean height of M1s is 170 cm?
We have a false negative result. The two possible explanations are bias and random error.

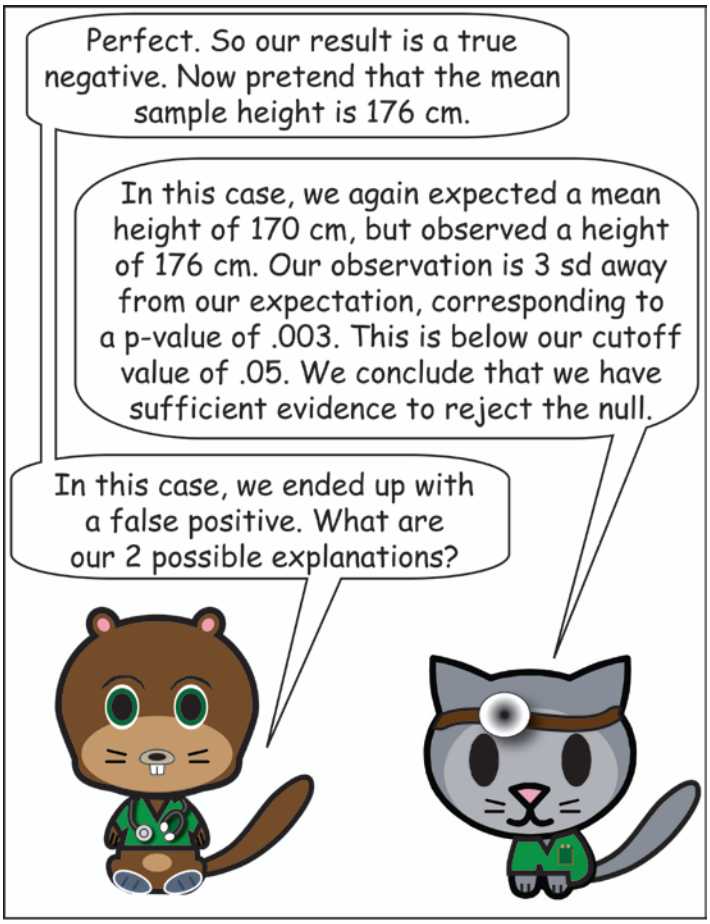

In terms of bias, maybe the dean got lazy and decided to measure the members of the M1 basketball club instead of obtaining a truly random sample.
How about random error?
Sometimes you just get unlucky. Even a truly random sample can give you misleading results. It's sort of like getting 20 heads in a row tossing a fair coin; possible, but unlikely. Occasionally, a sample will just happen to include only data at one extreme of a distribution. We calculate a p-value that tells us how likely we are to obtain a result at least as extreme as the one we got. Let's revisit random error vs. bias.
The athletes above represent different height categories as shown by their position on the graph. In this population there will be many baseball players and runners within one sd of the mean, fewer swimmers and football players more than one sd from the mean and very few boxers and basketball players more than 2 sd from the mean.
So what would our population look like?
With that distribution, our population would look like this, with most heights close to the mean and fewer extreme heights, symbolized by the boxer (height more than 2 sd below the mean) and the basketball player (height more than 2 sd above the mean).
A representative sample looks like this, with most observations about the mean.

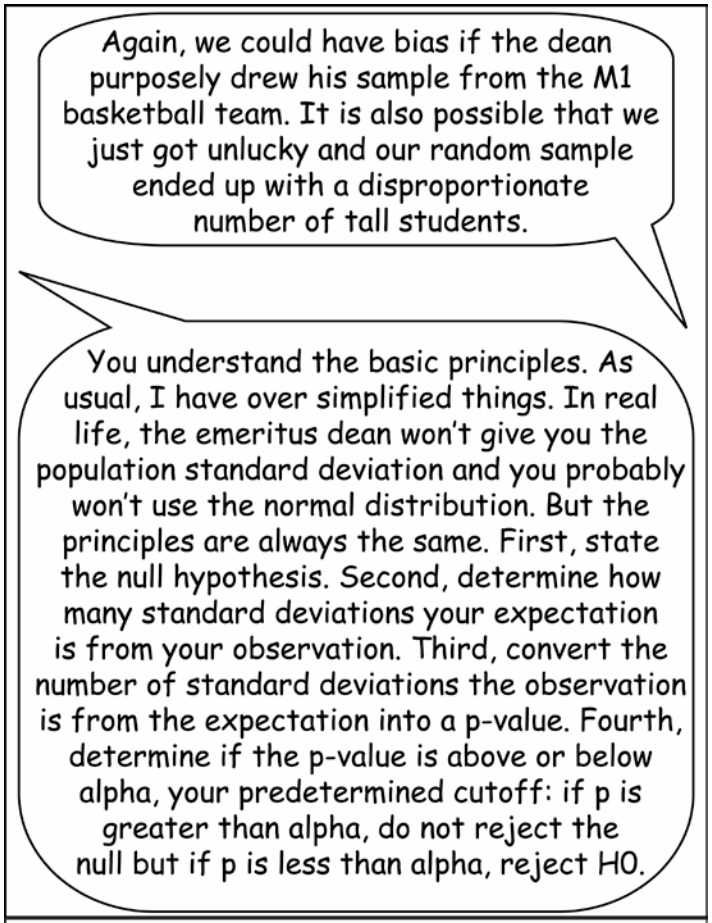

A non-representative sample looks like this, with a disproportionate number of extreme observations and can be due to bias or random error.

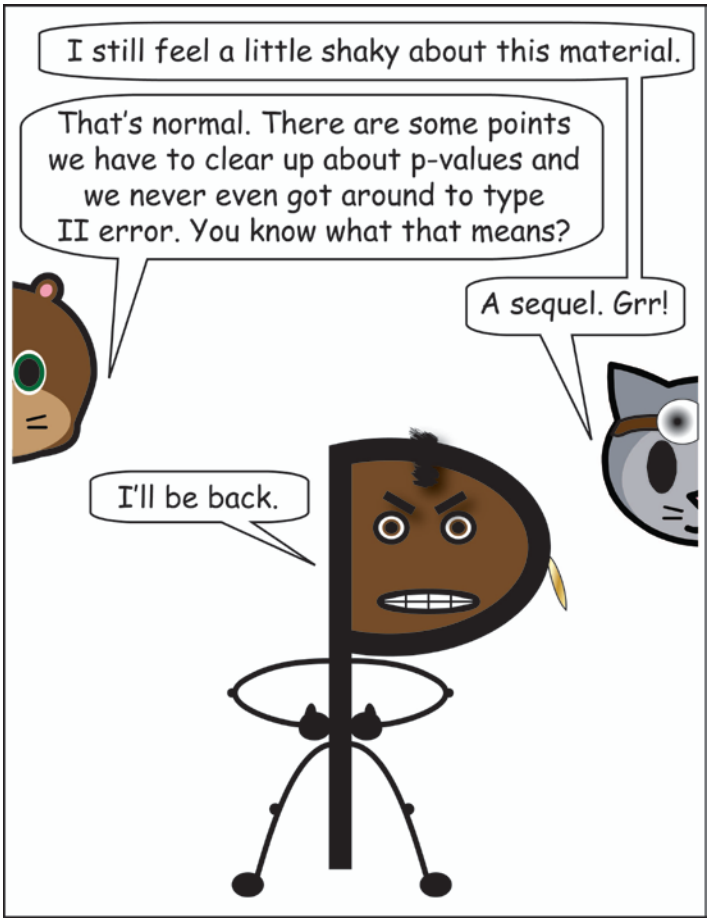

So a false negative can be due to bias or random error. Incidentally, we call missing a true difference a Type II error.
Sounds like a sequel. But it may be a long time before I recover from that last sequence.
Indeed. But next we have to go over true negative and false positive results. The ACME dean wants to know if his M1 students are taller than M1s at Omega Medical School.


ACME Students
Omega Students
It looks like the mean heights of these 2 groups of M1s is the same.
The mean heights are the same, but the ACME dean doesn't know this. The Omega dean tells the ACME dean that Omega M1s have a mean height of 170 cm. Under H0, what do we expect the mean height of ACME M1s to be?


Under the null, we expect ACME students to have a mean height of 170 cm.
The ACME dean measures another sample of 25 students and gets a sample mean height of 172 cm. Again we will assume a sd of 10 cm. Now what?
We compare what we observed with what we expected. We expected a mean height of 170 cm and observed a mean of 172 cm. Our sem is 2 cm, so our sample mean observation is one sd away from our expectation, resulting in a p-value of .32. The p-value does not reach our predetermined cutoff or alpha of .05. We do not have sufficient evidence to reject the null hypothesis.







References, Acknowledgements etc.

Many of the illustrations are modified clipart from Microsoft (Redmond, Washington) Office except "Doc" Squirrel and Mr. P who are "semi" original creations. Pictures of the Martian surface are from NASA and are in the public domain. The Literary Digest Cover and Alf Landon's portrait are from Wikipedia and are in the public domain. All artwork was created or modified using Adobe Illustrator CS4 (San Jose, California). Look for the sequel, coming soon!