Foundations of Statistics

A Tale of Two Statistics

The Dilemma at the Heart of Statistics

- We want to understand the world
- We may form a hypothesis, H. This is a model or a description for how the world might work.
  - It might be about what the big bang was like, or it might be about how customers like our new web widget

- Of course, we want to know if our model is correct – the true state of the world!
- So we collect evidence, and from this evidence, we want to figure out if our model is true.
  - In practice, evidence is rarely conclusive, so we really want Pr(H|D)
  - Think of this as the probability of the event that our hypothesis is true.

- But this is something we can never know!
  - Because the real world is not a perfectly controlled laboratory.
  - The evidence that we collect contains information, but it's just a few numbers (a handful of bytes, if you will) – that doesn't begin to identify a unique model out of all the possible models that could govern the real world.
  - We don't know how to weight all the possibilities that are out there.
  - A few examples will make this clear…

Example 1

- Say you flip a coin once, and it lands heads. What is the probability that it is a double-headed coin?
  - Is there really enough information to answer this?
  - How did the coin get there?
    - Is it from the US Mint?
    - Or is it from a magician?
    - And did this magician select it at random from a stash of coins, or deviously to trick us?
  - The context is missing - how do we choose between the different models for how the coin got there? How do we weight all the alternatives?
  - Of course, you might know more of the context, but you can never know it completely.

Example 2

- Isaac Newton observes measurements of the trajectories of the planets
  - And, if you believe the legend, the falling of an apple on his head
  - He notices that both of these are consistent with a gravitational attraction which is proportional to the square of the distance between two objects.

$$F = G\frac{m_1 m_2}{r^2}$$

- What is the probability that Newton's theory of gravity is correct?
  - Problem: Newton's theory seemed to work up to the precision of 17th century instruments.
    - It was only later that scientists developed instruments precise enough to show Newton's laws were incorrect.
  - So how could Newton decide how probable his model was compared to general relativity, when that hadn't been imagined yet?
  - What if he could imagine another theory? How could Newton decide whether gravity was causing the motions he observed, or a guy in a toga riding a chariot around the sky?
    - Do the numbers we get equip us to compare these two very different ideas?
  - The number of models we could come up with that are consistent with observations of the planet is literally infinite, so we could never ever write them down to assign each a probability.
  - You might ask, does it even make sense to assign a probability to Newton's gravity? It's not like a coin flip that could come up heads or tails – we know that it is the true state of the world or it isn't

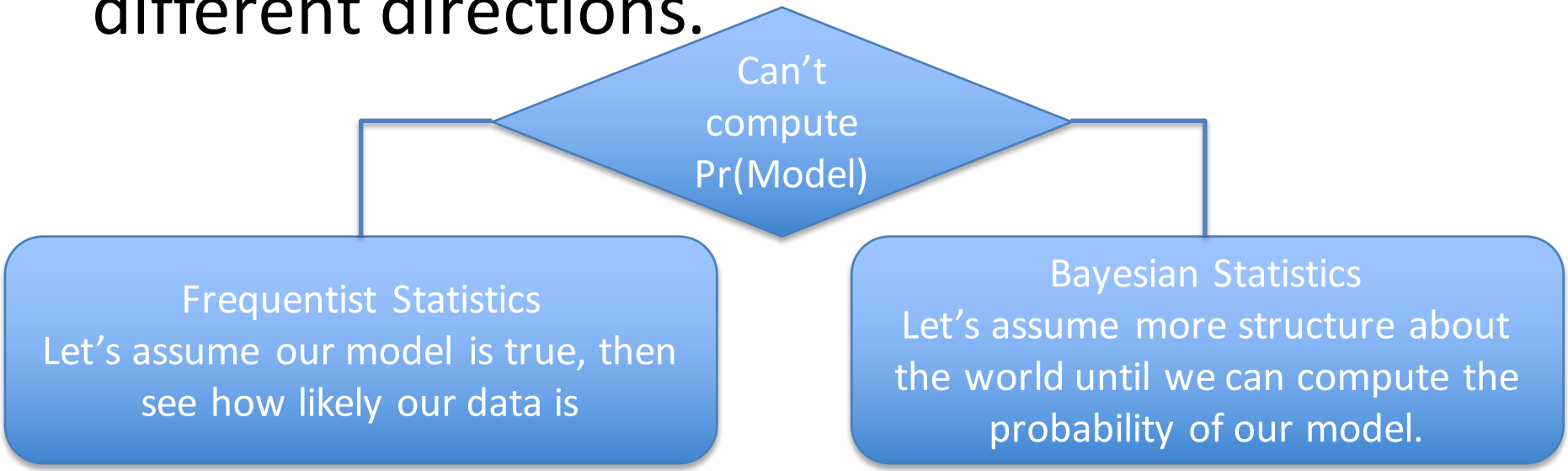
Example 3

Artist: could we get a gnarly picture of a squid here? Perhaps fighting a whale?

- You discover a new species of squid in an undersea trench (paulus statisticus). The three specimens you find are 3.2, 3.3, and 4.0 feet long.
- What is the probability that the average length among the entire species is 3.5 feet?
  - Ok, that's a single number, so probably zero. A point estimate doesn't seem like it can be true in this case
- Better question: What is the probability that the average length is between 3 and 4 feet?
- It might seem like we have a better shot at answering this. The average length has to be a positive number, and we know all the positive numbers.
  - There's no new numbers out there we haven't imagined yet.
  - We have a better grasp of the possibilities
- But we don't know how representative the three specimens we found are.
  - What if one out of 10 squid are female, and the females average 200 feet long?
- And we still don't know all the relevant information.
  - What if there's something about the deep water pressure that makes anything longer than 4 feet collapse?
  - Or so little light that anything under 3 feet long can't metabolize enough food to stay alive?

A Fork in the Road

- So we can't deduce the probability of our model, because we don't know enough about the structure of the world.
- At this point, statistics branches in two very different directions.



The Frequentist Approach

The Birth of Modern Statistics

- Most of what we consider modern statistics can be traced to two scientists: Neyman and Pearson, who published influential papers in the 1930's.
- Before this, there were a lot of statistical procedures, but no coherent account of how to choose the right one.
- Neyman and Pearson added a rigorous mathematical treatment, forming the basis of what we call Frequentist statistics

Jerzy Neyman

April 16, 1894- August 5, 1981

Egon Pearson

11 August 1895 - 12 June 1980

Objective Probability

- Remember the central dilemma of statistics: We observe data, D, and given that the data occurs, we want the probability our hypothesis is true, P(H|D)

- To a strict frequentist, this isn't just impossible to compute, it doesn't even make sense to assign a probability to a hypothesis.
- A frequentist defines probability as a matter of long-run frequencies.
  - Need to specify a **collective** of elements – like throws of a dice. This is a frame of observations that can happen over and over.
  - In the long run – as number of observations goes to infinity – the proportion of throws of a dice showing a 3 is 1/6
  - The probability of a '3' is 1/6 because that is the long run frequency of '3's relative to all throws
- This is what we call objective probability.

Objective Probability and Hypotheses

- If you view probability as objective, you cannot talk about the probability of a hypothesis
  - Newton's law of universal gravitation – this either governs the motion of planets or it doesn't. There's no collective here, can't run the universe over and over and see how often gravity holds.
  - "genes are coded by DNA" is not true 2/3 of the time in the long run – it is just true. There is no relevant long run collective.
- A hypothesis is just true or false.
- When we say what the probability of a hypothesis is, we are referring to a **subjective** probability (and you and I can disagree about what that probability is).
- This usually reflects our lack of information.
- For example, we may say that there's a 50% chance of rain tomorrow, but that doesn't mean that if you could make lots of identical copies of the Earth, it will rain in half of them. It means that we don't have enough information – the movement of every atmospheric particle to make the determination.

Pr(D|H)

- So if we can't talk about the probability of our hypothesis, what probabilities can we study?
- We need a long-run collective, and we can get one using a thought experiment.

- Again, H is our hypothesis and D is our data.
- Frequentists assume the hypothesis is true.
  - From here on, we'll call it the null hypothesis
- Now this null has to be quite specific, because that's the only extra assumption we're going to make. And based on just that assumption we need to make predictions about what data should come out of our experiment.
  - So "vitamin W kicks the bad toxins right smack out of your system" probably won't work. But "Vitamin W decreases blood pressure by 12 mmHg" might (depending on what we're measuring).
- We imagine running our experiment over and over, with the null hypothesis governing how it behaves,
- Now we have a meaningful collective.
  - So we can look at the relative frequencies of different outcomes
  - Specifically, In how many of these hypothetical experiments would we get data at least as extreme as D?
- This is captured by the p-value.
  - p-value – the probability of getting data as extreme as our observations assuming the null hypothesis is true.

Example

- Let's say you want to know whether Vitamin W reduces acne
- You randomly assign participants to a treatment group which gets Vitamin W, and a control group which gets a placebo, count the number of pimples that each participant has a week later
- As you know from your beginner stats class, the t-statistic is a measure of how different two group means are.
  - Let's say we get a t-statistic of 2. This is D.
- Our null hypothesis has to be specific enough to predict how probable each value of the t-statistic is.
- In this case, the natural choice is zero difference between Vitamin W and the placebo.
- So we can imagine running an infinite number of experiments in which Vitamin W has the same effect as the placebo.
  - In how many of these would we get a t-statistic above 2?
  - This is Pr(t>2 | drug is a placebo)

Evaluating Hypotheses

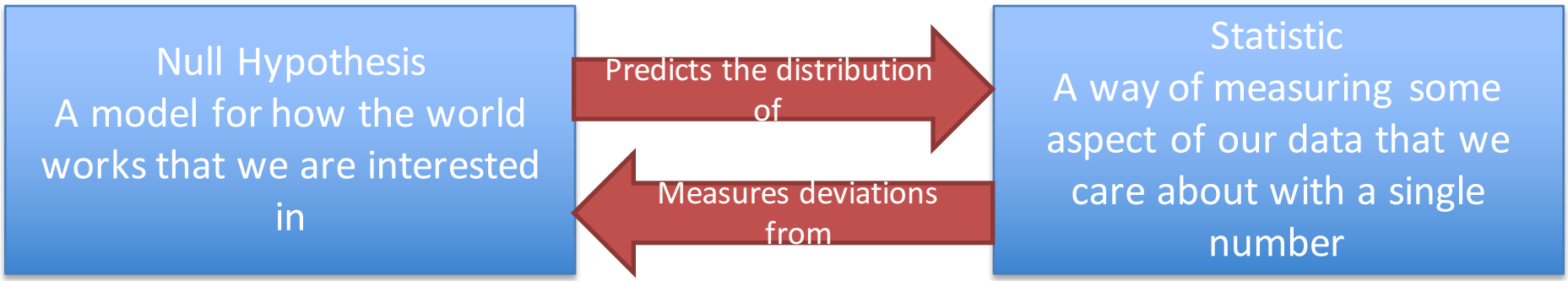
- It doesn't make sense to ask Pr(drug is a placebo | t>2)
  - What's the reference class? this statement is either true or false.

- Pr(H|D) is the inverse of the conditional probability Pr(D|H)
  - As we know from Bayes' rule, inverting conditional probabilities makes a big difference

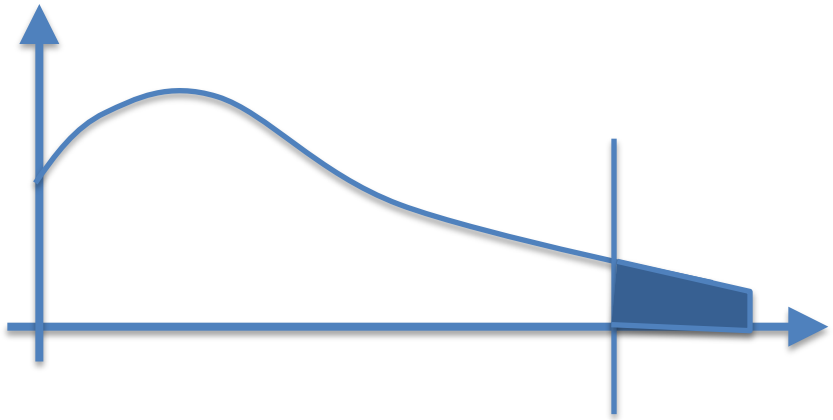
- Example from Dienes:
  - P('dying within two years'|'head bitten off by shark') = 1
  - P('head was bitten off by shark'|'died in the last two years') ~ 0
  - <artist: I hope you have a good graphic for this one☺)

- P(H|D) can have a very different value from P(D|H)

Statistics and Hypotheses



With the Null, we can predict how likely each possible value of the statistic is



Here, we'll assume "more extreme" is towards the right, so the p-value is the area under the curve, from the statistic we get to infinity.

p-values and Decision Rules

What p-values are not

- The p-value is not the probability our null is true!
  - This is a very tempting interpretation, but it's misapplying Bayesian intuition.
- An example from Dienes: Say you're totally convinced that a coin is weighted unfairly.
  - You contracted with a specialist in unfair coins, that designed it to land heads exactly 60.0% of the time.
  - He tested it rigorously, using the very best computer simulation, and flipping it thousands of times.
- You run an experiment, flipping the coin 6 times.
  - Your null hypothesis is that it's a fair coin.
  - You get 3 heads out of 6.
  - In this case, every possible outcome is at least as extreme as 3 heads. 4 heads seems farther from a fair coin, so is 2 heads. This tells us that p=1.
- But you made the coin to be unfair – this experiment is very unlikely to convince you that the coin is actually fair.
  - In this case, you have a strong prior belief that the coin is unfair, and the frequentist framework has no way of recognizing this.

p-values and Continuous Parameters

- A further problem comes when we have a parameter that can vary continuously. The probability that a coin comes up heads could be anything between 0 and 1.
- So you get a very high p-value
  - maybe you flip the coin 100 times and get 50 heads
  - This might convince you that the probability of heads is close to 50%. You're probably convinced that the probability of heads is not 90%
  - But what about 51%? Or 50.001%?
  - No matter how high the p-value is, or how many times you toss the coin, your experiment can't establish that the parameter is 50% with zero error.
- This is often an issue in social research. Socio-demographic variables can be connected by all kinds of complex mechanisms, so their relationship is basically never zero.
  - Liking the color red may seem unrelated to your opinion on GMO foods, but surely there are subtle factors in our culture that influence what your favorite color is.

Decision Rules

- Frequentist stats can't tell you the probability your hypothesis is true, because it doesn't recognize what other hypotheses are out there, or how probable they are before you gather data.
- All we can do is set up decision rules for certain behaviours
- –rejecting or not rejecting hypotheses –
- such that in following those rules in the long run we will not often be wrong.

- E.g. Decision procedure:
- Run 40 subjects and reject null hypothesis if t-value larger than a critical value

Type 1 Error rates

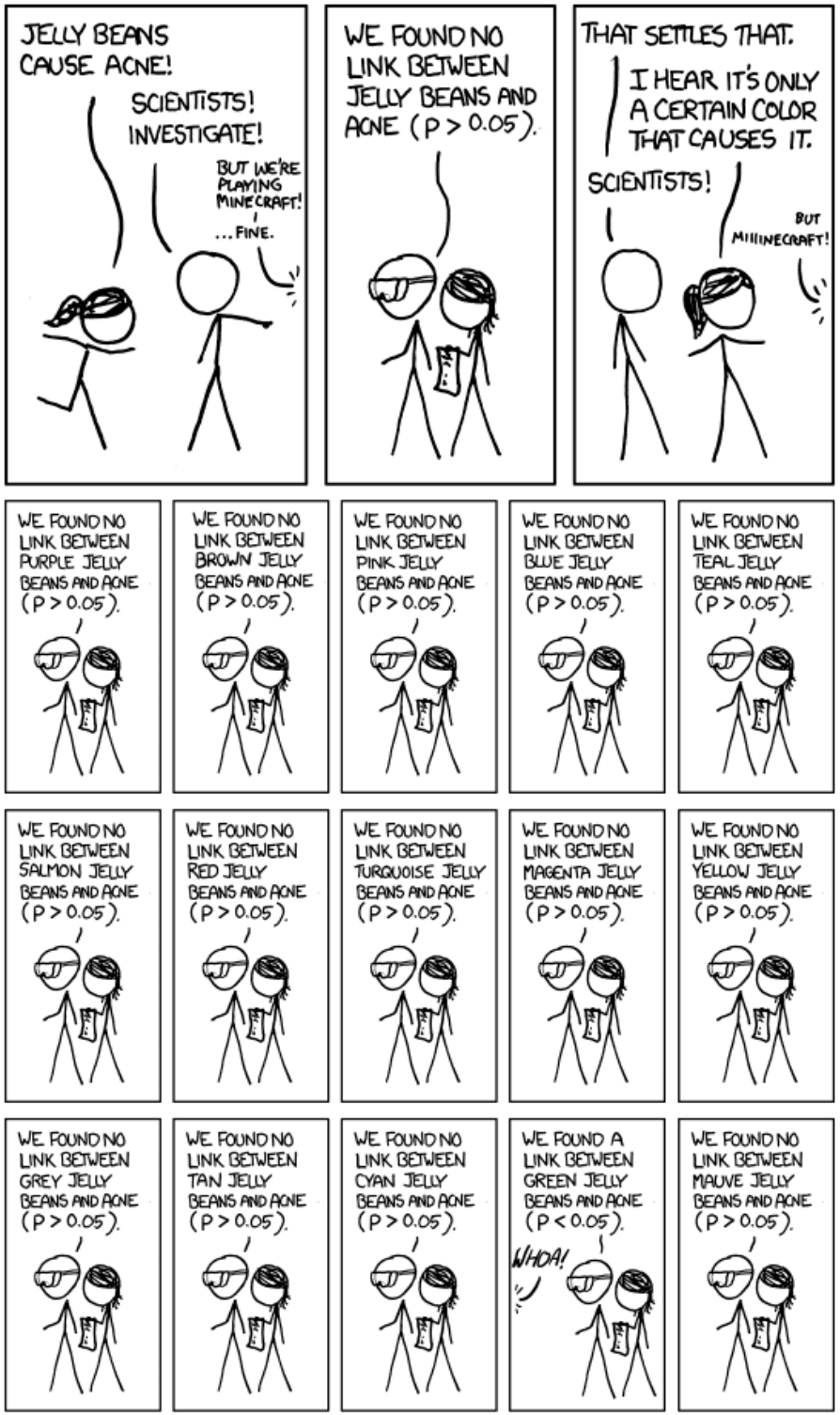
- In particular, there is one type of error that statisticians focus on.
- A Type 1 error. This is when we reject the null hypothesis, even though it's true.
  - In practice, this is usually (but not always) the most damaging type of error.
  - You discover that there is an effect: your drug improves health. It gets in the papers, there's no way to take the news back.
- We can define the type 1 error rate as $\alpha = Pr(\text{rejecting } H_0 | H_0)$
- Ronald Fisher, a founding father of statistics who held public and ill-mannered debates with Neyman and Pearson, suggested setting a critical value so that $\alpha = .05$.
  - This is an arbitrary number, but it has become an ironclad and universal standard for statistical significance.

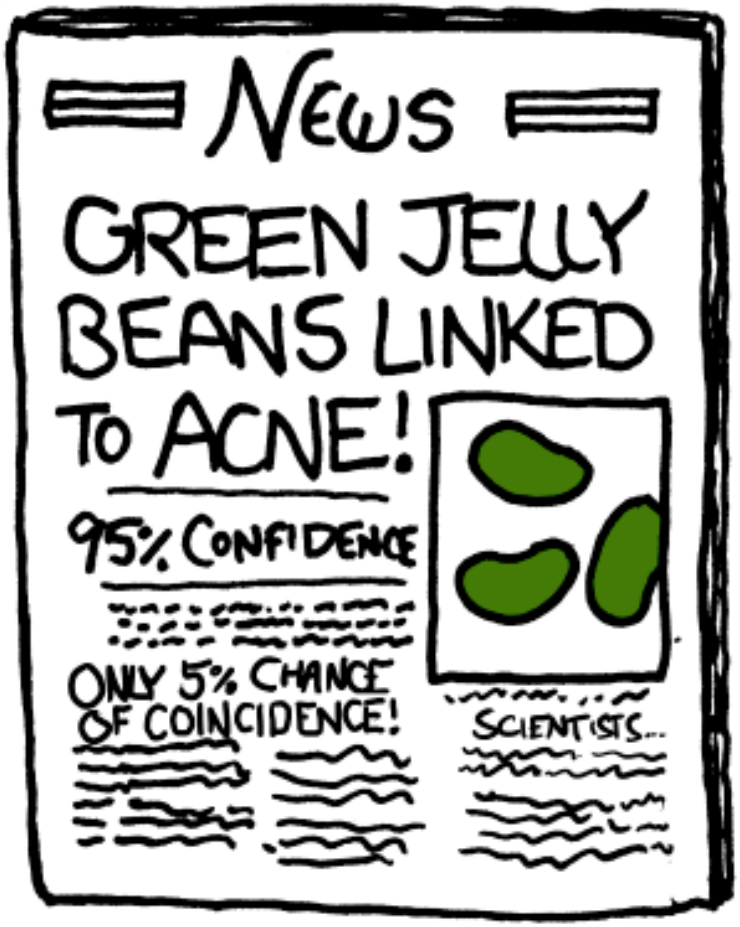
- Note that we do NOT ever accept the null hypothesis. That's because accepting a null that's actually false is a type 2 error, and most studies do not control type 2 errors as strictly as type 1 errors (or don't control them at all!)
  - Also remember that we often have a continuous parameter,
  - Let's say the null is that the coin is fair – it lands heads 50% of the time.
  - You get a high p-value, so you can't reject the null. 50% is plausible.
  - But this doesn't mean that it's not 51% or 50.001%, etc.

Multiple Comparisons

The Multiple Comparison Problem

- The key to the Frequentist framework is controlling error rates
  - Anything that affects your error rates affects the decisions you can draw.
- The more opportunities you have to make an error, the higher your error rate will be.
- Example: If you perform several t-tests, the overall probability of an error is increased





http://xkcd.com/882/

Family-Wise Error Rate

- An important concept is the family-wise error rate.
- This is the probability, assuming all null hypotheses are true, of getting at least one type 1 error.
- One possible response is to run a correction on your p-values, so the family-wise error rate returns to .05.
- The simplest is the Bonferroni correction.
  - If you perform n comparisons, multiply each p-value by n, then compare to the normal value of .05.
  - It would be equivalent to leave the p-values the same, but check them against a threshold of .05/n.
- The Bonferroni correction is considered conservative (type 1 error rates may actually be quite a bit lower than .05)
  - This increases the odds of a type 2 error, and it may become quite difficult to have a significant result with a large number of comparisons.
- This may seem strange – your interpretation of your test depends on what other tests you run.
  - Critics of the Neyman-Pearson approach often complain that this doesn't make sense. Shouldn't the evidence we get from a test only depend on the results of that test? Shouldn't the green jellybean data be all that matters when studying green jellybeans?
  - But of course, we really should have much less trust in Researcher A's conclusions, seeing that she was testing everything in sight. We instinctively know that the other tests matter, because it could have easily been another color that was significant.

What is a Family?

- To control a family-wise error rate, you have to decide what constitutes a family – and this question is driven by theory.
- The jelly bean comic makes it easy – the conversation makes it clear the researchers have no idea what color might work.
- But of course there could be a theory out there that predicts that green jellybeans, and only green ones cause acne.
  - Maybe there's some spectral-dermatologist out there who's been studying evolutionary links between our color and taste receptors, and he's not at all surprised that the green jelly beans cause acne.
  - If this is your belief, then you probably don't feel like the other experiments have any bearing on the evidence gathered for green jelly beans, and you would argue that your p-value should not be adjusted.
- You have to decide whether to define your family of tests in terms of the green jelly bean theory or not.
  - One factor to consider is when you came up with the theory.

Planned vs Post Hoc Comparisons

- The timing of a theory can make a difference in interpreting frequentist results
  - After I see the data on green jellybeans, maybe I can come up with a compelling theory to support the green jellybean-acne link.
  - But you would probably find that much less convincing than if I announced my theory before running the study.
- A lot of statisticians therefore recommend a correction whenever comparisons are made post hoc – in an exploratory fashion after gathering data.
- If I came up with a well-motivated green jellybean theory before gathering data, I could instead run a planned comparison
  - This is easy to do in a regression framework, as we'll explore in coming weeks.
  - I could create a dummy variable to represent green jellybeans, and add other planned comparisons that I was interested in.
- Many statisticians recommend no correction when tests are planned in advance.
  - For example, we usually don't apply a correction when we estimate multiple coefficients in a linear regression.

Other Considerations

- What if your theory is about green jellybeans, but you have the data on other colors and don't want to throw it away.
  - That's fine: run the planned comparison for green at the nominal level, but separately run the post hoc tests with a correction.
- What if you put several tests into your research paper, but they seem to be testing totally different theories?
  - You test green jellybeans, and classical ballet, and you believe the mechanisms that link each to acne are unrelated.
  - You still run an inflated risk of reporting at least one type 1 error.
  - But if you feel they are testing different theories, what bearing does one result have on another? Shouldn't each one be considered in isolation?
  - Can't you present them to the reader as separate experiments, and avoid a correction that may be overly conservative?

Deciding When to Correct

- In practice, the question of whether to correct your p-values or not can be quite complicated
  - And the problem is partly cultural: Different fields of research employ very different standards of when to run a correction.
  - Clinical studies tend to use corrections quite a bit.
    - Moye (2008) recommends a disciplined approach in which all primary endpoints of a study are corrected to bring the type 1 error rate to .05
      - These results are considered confirmatory.
      - Secondary endpoints should be uncorrected (but interpreted only as supportive evidence)
  - Corrections are rare in economics
    - Here, transparency is emphasized: researchers should report all tests conducted and it is up to the reader to evaluate results in context.

Takeaways

- The question of when to correct is contextual and often difficult to answer.
- This means that it's important to look beyond p-values:
  - Did the researchers go looking for a specific effect, or is the study more of a fishing expedition?
  - What other results, including those in other studies, could be understood to fall in the same family?
  - Has a hypothesis been supported by multiple studies?
- Decide how convincing the results are, and whether you would run a correction yourself.

Stopping Rules

A Bit More Data

Here's a scenario that's probably all too common

- Researcher A gathers 30 subjects to see if his drug reverses hair loss
- He compares the treatment and control group and his t-test is not quite significant at the .05 level
  - Say p=.06
  - But the difference is in the right direction
- Researcher A realizes how close he is, and gathers 10 more subjects for his study
  - Now he has a total of 40 subjects to test, and maybe a nominal t-test comes out significant, p=.04
- Does this result convince you that the drug works?
- What's the type 1 error rate?
- Assuming the drug has no effect, there's already a 5% chance of a type 1 error after 30 subjects. Testing again after 40 subjects can only increase the risk of a type 1 error.
- So Researcher A cannot test his 40 subjects without a correction.
  - In fact, since he already accrued a 5% chance of error, he can not reject the null, no matter how many more subjects he gathers!
- A better strategy: correct both p-values, at 30 subjects as well as 40 subjects, so that the total type 1 error rate is 5%.

An Alternate Study

- Now imagine that Researcher B gathers all 40 subjects at once, and gets the exact same data as Researcher A (p=.04). Researcher B does not need a correction and can reject the null hypothesis.

- Think about how surprising this is: Researcher A has the same data as Researcher B, and Researcher A did not actually stop gathering data after 30 subjects. But the fact that Researcher A would have stopped if her first t-test were significant changes our interpretation of the results.
- A critic of the Neyman-Pearson approach might say that this doesn't make sense – our interpretation of the data depends on events that didn't actually happen! Shouldn't the same data always lead to the same conclusion?
  - On the other hand, shouldn't the fact that Researcher A has more chances to get a significant result make you doubt his results?
  - An extreme example may help convince you

A Determined Researcher

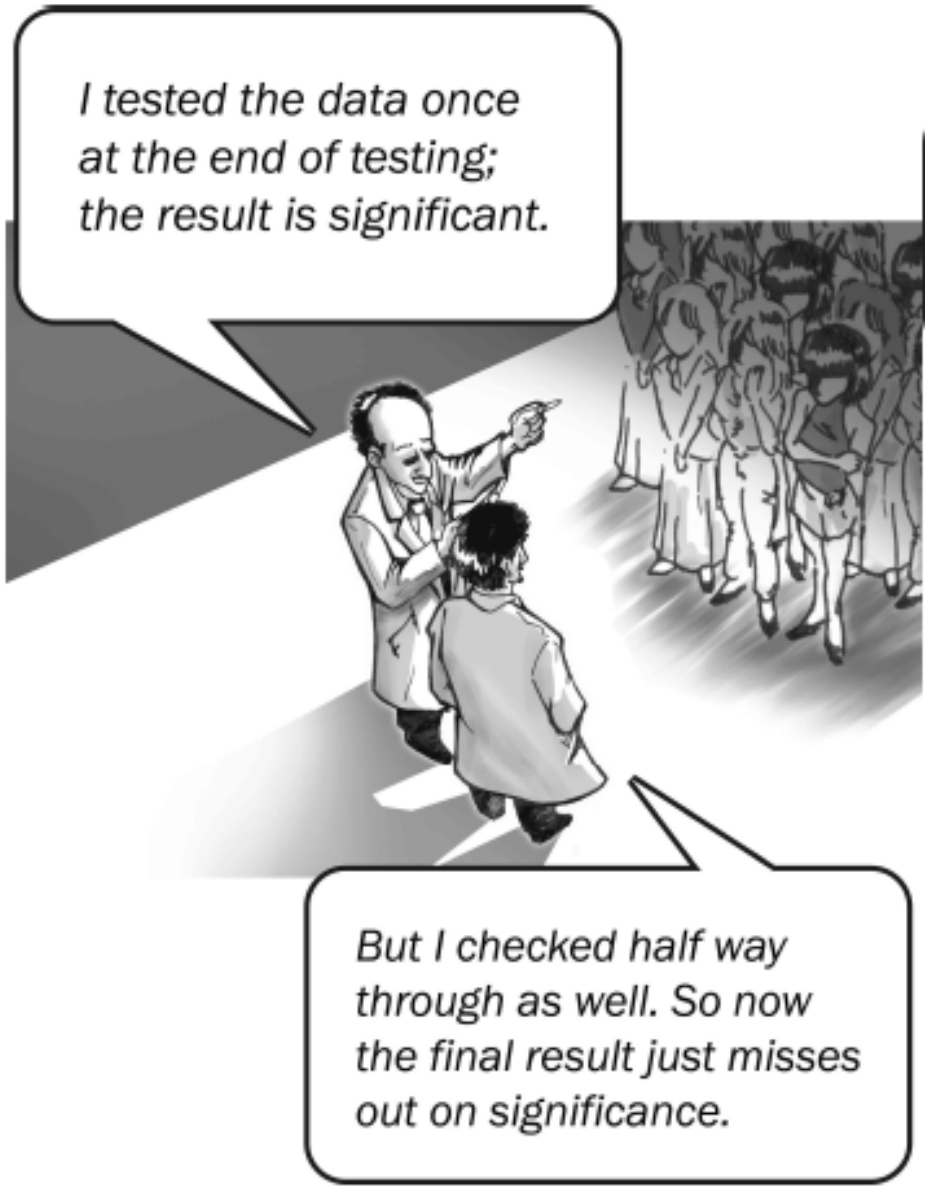
- Suppose that Researcher A would have kept going.
- She recruits her first 30 subjects, computes a t-test at the nominal level.
- If the t-test is non-significant, she collects 10 more subjects, and runs a t-test on the whole group.
- As long as her last t-test is non-significant, she keeps repeating the process.
- Researcher A comes to you and announces that she ended her study with a significant result. How convinced should you be that the drug works?
- What's the type 1 error rate?
- In fact, Researcher A is guaranteed to eventually get a significant t-test!
  - This surprising fact is a basic mathematical property of infinite sequences
- This procedure has an alpha of 1.
  - You always knew Researcher A was going to come and claim significance, so there's no information in this fact.

Another Explanation

- Remember that we're computing the probability of our data (or more extreme data) given our null hypothesis.
  - This is an objective probability, it doesn't just depend on the single experiment, it depends on what collective the experiment is part of.
    - Remember that we imagine an infinite sequence of hypothetical experiments conducted assuming the null hypothesis
    - Each experiment is conducted according to the stopping rules the experimenter uses.
    - So for researcher A, some hypothetical experiments will end after 30 subjects, some after 40, and so forth.
    - For researcher B, all experiments in the collective end after 40 subjects.
  - So events that don't happen are important, because we have to see how likely our results were compared to all the other results that are possible.
    - This tells us how significant the data is.
  - Hopefully this helps convince you that events that don't occur really can affect what your data is telling you.
    - Even so, this result is paradoxical and some clever critics have a great deal of fun with it, at the expense of Frequentists.

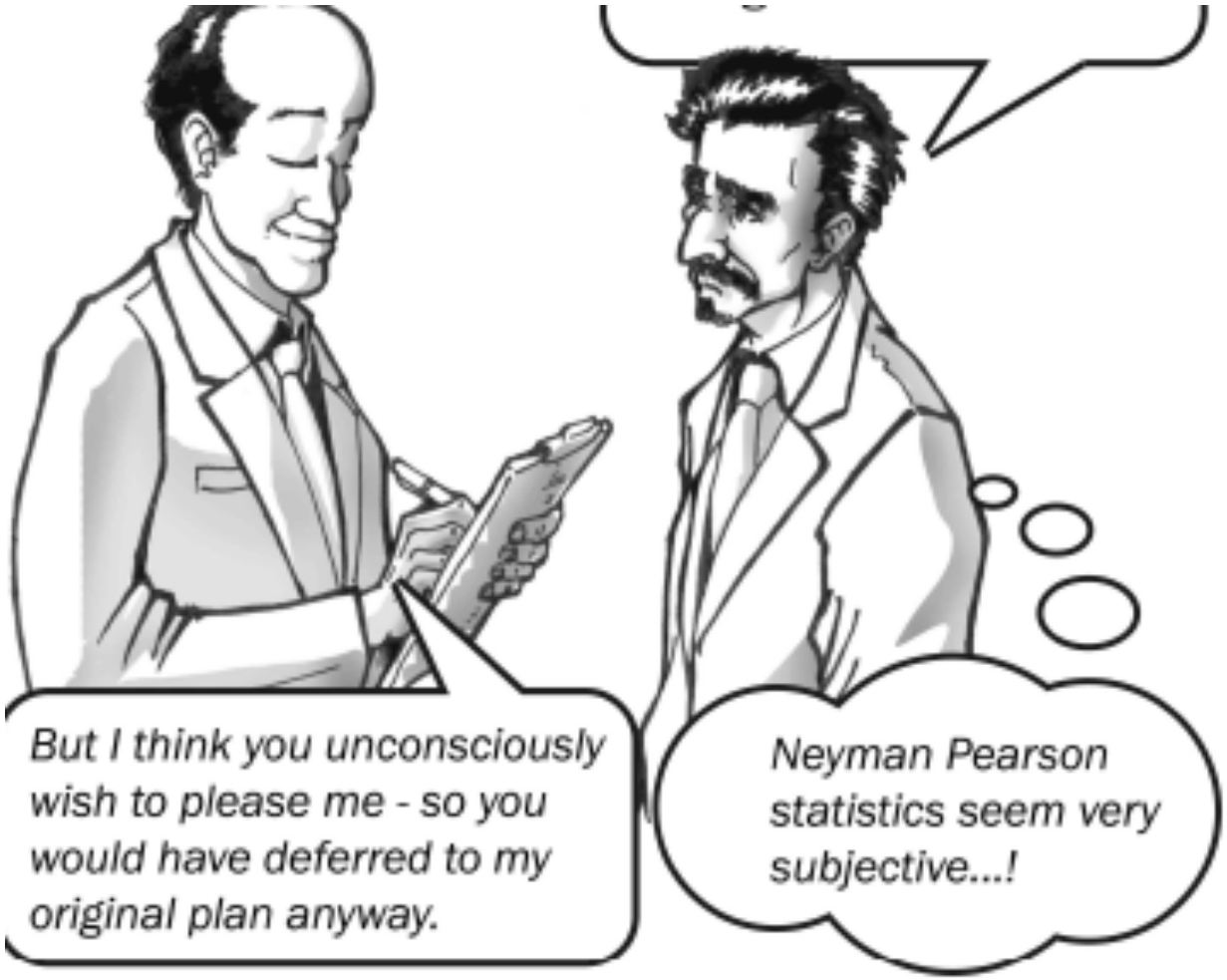
p-values and Kung Fu

- In this comic taken from Dienes, two scientists realize *after the fact* that they have different stopping rules.
- As we know, the final p-value will have to be corrected if the experiment would have stopped halfway through if the second scientist's test was significant.
  - But would it have stopped?
  - It seems the scientists didn't agree to this in advance



- They try to figure out whether the experiment would have stopped.
- The result, it seems, depends on who would convince the other – and therefore, on who's kung fu is better!

- But it gets even better: maybe the kung fu outcome will depend on the researchers' desires to please each other
  - Which are unconscious, so how can we compare them?



Takeaways

- The sensitivity of Frequentist stats to events that don't occur is controversial.
  - Many critics – Bayesian statisticians in particular – think that the same data should always lead to the same conclusion.
  - You can decide if this is a strength or a weakness of the approach, but I hope that you understand the intuition behind why events that don't occur may change your understanding of data.

In Defense of Frequentists

In Defense of Frequentists

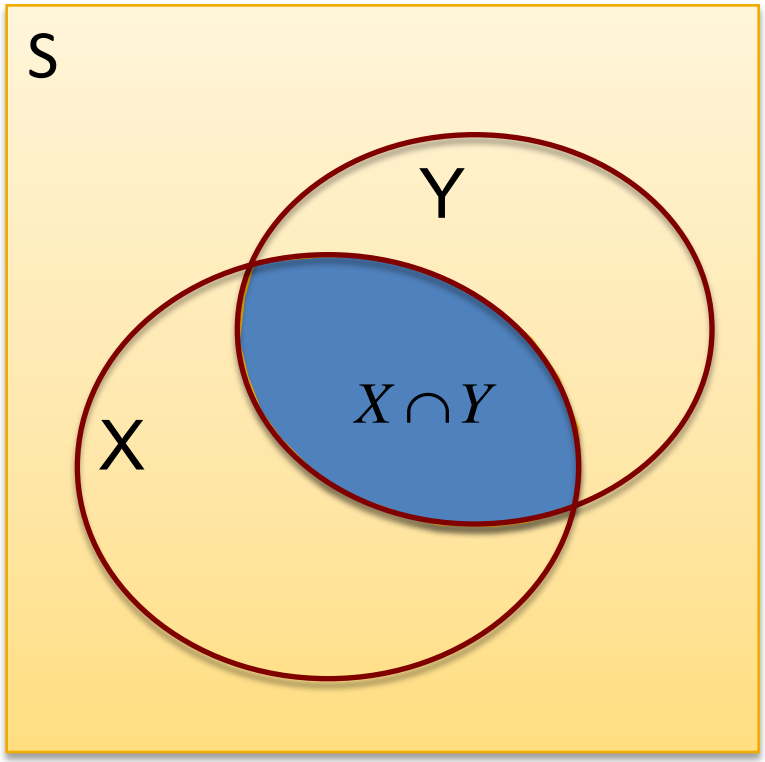
- There are certainly many quirks to the Neyman Pearson approach
  - And it has many critics
  - What we conclude from a test may depend on things like other tests we conduct, when we came up with our theory, and what we would have done in case of events that never happened.
  - This is unsettling to many people.
- But take a moment to think of how clever all this is.
- We want a system for learning truths about the world, knowing that we may not agree on how likely hypotheses are
  - And we may not even be able to imagine what alternate hypotheses are out there!
- So we define probability in terms of long-run frequencies.
- And even though we can't discuss the probability that our null is true, we know the probability of accidentally rejecting it if it is.
  - That's because we've also managed to define our error rate in terms of long-run frequencies.
- We have a system that makes very minimal assumptions about the world, just looking at one hypothesis in isolation
  - The p-value is a measure of how consistent it is with the data
- p-values depend on many things, researcher intentions, stopping rules, etc.
  - But we've seen that these are all responses to things that inflate our error rate.
  - There will always be type-1 errors.
  - Science progresses when we come up with hypotheses that survive repeated testing.

Bayes Rule

Bayes' Rule

- Before we take a closer look at Bayesian statistics, we need to review Baye's rule
- This is the famous property of probability that underlies Bayesian statistics.
- To derive Baye's rule, we can take the definition of conditional probability and rearrange it to get:

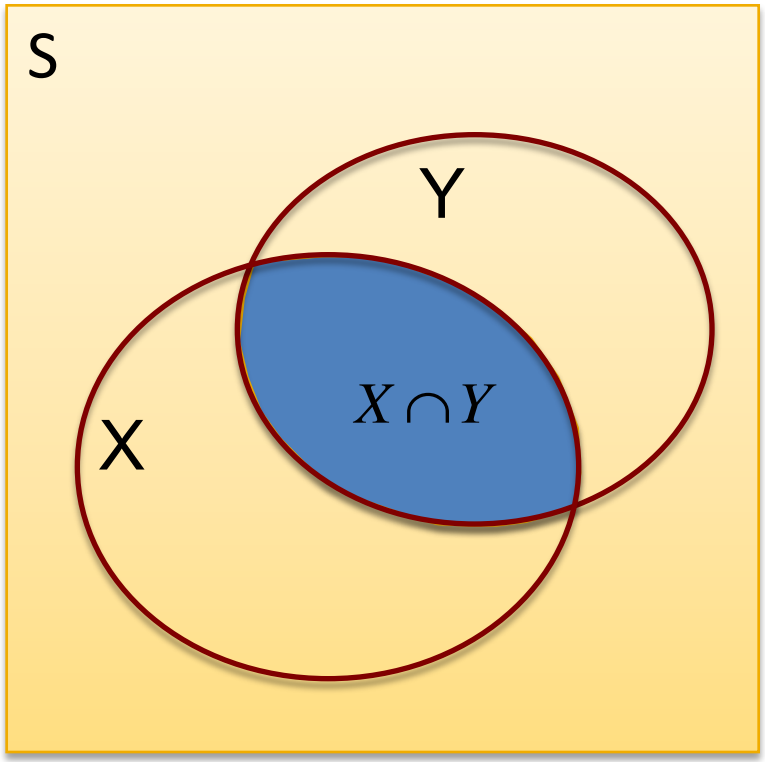
  Pr(X and Y) = P(Y) Pr(X | Y)
  - In other words, the probability that our dart lands in the intersection is the probability that it lands in Y, times the probability that it lands in X, given that it already lands in Y.
  - This is called the multiplication rule.
- Notice that we could have done this the other way around, first assuming that X occurs, then adding the fact that Y occurs.
- Pr(X ∩ Y) = P(X) Pr(Y | X)
- You can combine these two equations to get
- P(X) Pr(Y | X) = P(Y) Pr(X | Y)
- Or P(Y|X) = Pr(X | Y) Pr(Y)/ Pr(X)
  - This is a famous relationship known as Bayes' Rule
  - It relates P(X|Y) to its inverse, P(Y|X)
  - If you ever forget it, it's easy to derive.



Bayes' Rule Quiz

- $P(Y|X) = Pr(X \mid Y)\, Pr(Y) / Pr(X)$

- Suppose that 0.2% of the US population suffers from coca colic.
- There's a test for the disease, but it gives a false positive 10% of the time for healthy individuals, and a false negative 10% of the time for sick individuals.
- Suppose you take the test and the result comes back positive.
- What is the probability you have coca colic?

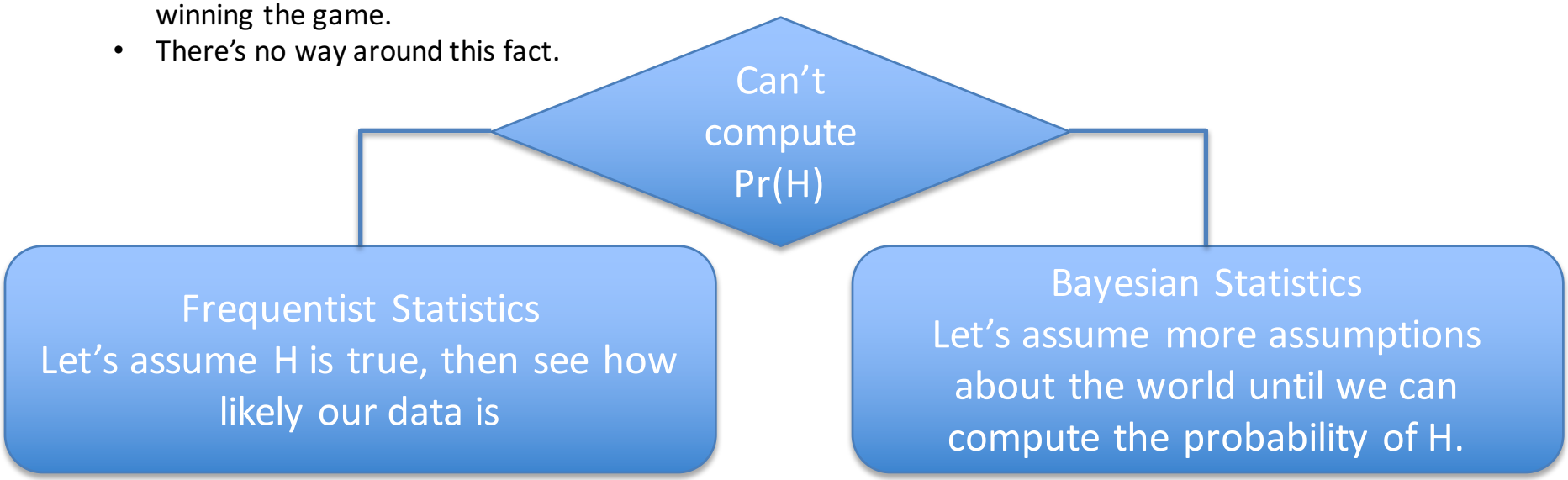


- Hint (should be hidden until students click something) Let X be the event that the test comes back positive. Let Y be the event that you have coca colic.
- Ans: 0.01771653543 (or .018, .0177, 1.77%, etc)

Embracing Subjectivity

A Fork in the Road

- The Neyman-Pearson approach is the dominant framework for statistical analysis
- But there are other forms of inference, and the main competitor is called Bayesian Analysis.
- Remember the problem we started with: how can we know the probability our hypothesis is true?
  - A frequentist would say it doesn't make sense to assign a probability to a hypothesis. Probability is considered objective and a hypothesis is either true or false. There's no collective.
- A Bayesian would say that we ***can*** give our hypothesis a probability.
  - We often do this when we talk:
    - "There's a 30% chance I'll make it to your dinner."
    - "Berkeley has a 50-50 chance of winning the football game."
  - This is what we call a subjective probability
    - It expresses how confident we are in the given statement.
    - The catch is that you and I may have different degrees of confidence. You may think Berkeley has a 60% chance of winning the game.
    - There's no way around this fact.



Bayesian statistics

Because probabilities are subjective, Bayesian statistics can't tell you how much to believe in your hypothesis.

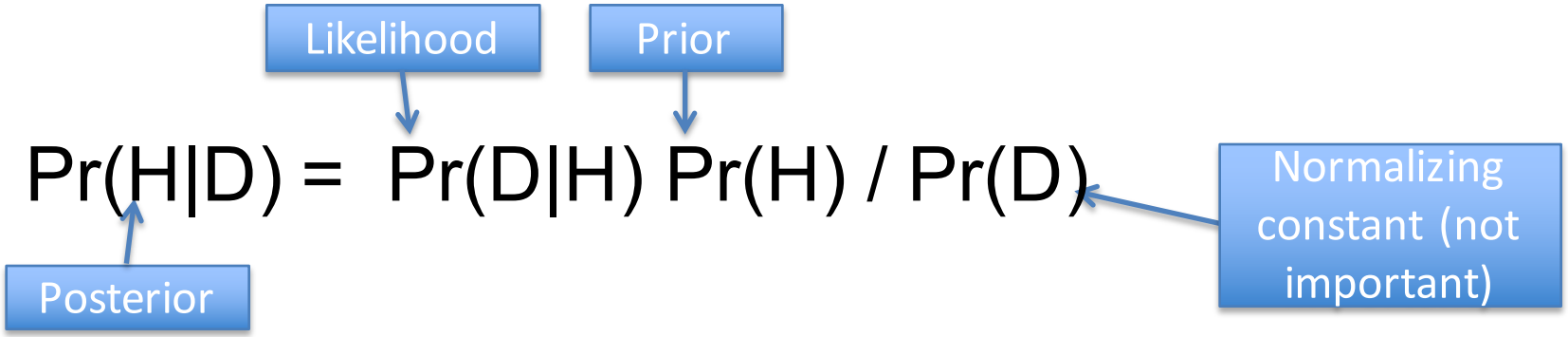
However, if you know how confident you are in the hypothesis to start with, Bayesian statistics can tell you how to change your beliefs in light of Data.

As before, let H be your hypothesis. You begin with some amount of confidence in the hypothesis, so you can assign it a probability Pr(H)

Let D be the event we get the data we observe (not some other data our study could produce)

We want the probability the hypothesis is true, given the data, Pr(H|D). We can compute this using Bayes' rule…

Anatomy of Bayes' Rule



In this context, the parts of Bayes' Rule have special names:

Prior: This is the belief that you start with, before you do your study. Your prior might be different than mine, and that means that our posterior beliefs will also be different, even given the same evidence.

Posterior: This is your updated belief, stronger than your prior if the evidence supports your hypothesis, but weaker if it doesn't.

Likelihood: Assuming your hypothesis is correct, this is how likely the evidence is. Just like before, your hypothesis has be quite specific - detailed enough that you can compute how likely your data is.

Normalizing constant: we usually don't worry about this. The basic idea is that you multiply your prior by the likelihood. Then if the total probability doesn't equal 1, you divide so that it does.

So another way of phrasing Bayes' rule is your posterior is proportional to your prior times likelihood.

A Bayesian Analysis

A Bayesian analysis has three steps:

1. We have to define a set of hypotheses, and assign a prior probability to each one.
   - This is what we call a total probability model
   - Our probably model must include all possible states of the world
     - So there's a problem if there's a state out there that we didn't imagine.
       - If you're studying planetary motion, and you forget to include general relativity as a model, or the guy in the chariot, Bayesian stats won't discover them for you.
     - But if you're interested in the average length of your newly discovered squid species, you know all the possible values, so at least they all get considered
       - But your prior beliefs about them could be very different than mine! We have no way of judging who's prior beliefs are better.
2. We gather evidence
3. We update our priors according to Bayes' rule, yielding posterior probabilities.

- The output of a Bayesian analysis is not a single hypothesis or a single decision, it's a detailed set of beliefs over all possible hypotheses.

Possible example of a Bayesian Analysis

Maybe the coin flip example, where probability of heads can be anything between 0 and 1

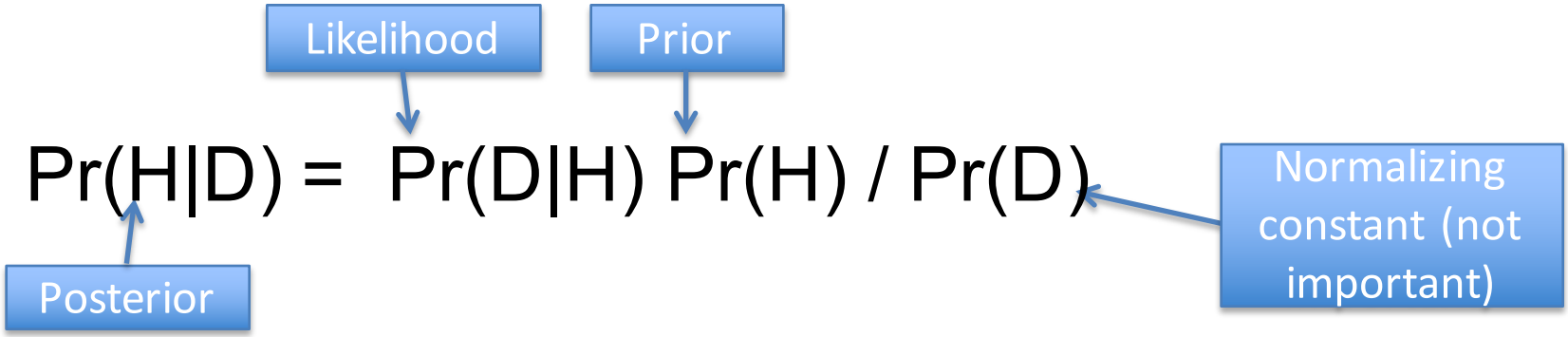
The Likelihood Principle

Bayes versus the Frequentists

- You may wonder why you'd want to use a Bayesian analysis.
  - After all, you have to select prior beliefs, and these are difficult to justify.
- But the Bayesian framework has one very attractive feature…

- As we saw in the Frequentist framework, our decision of whether to reject a null hypothesis depends on many things besides the test statistic:
  - when you planned to stop running subjects,
  - whether you conduct other tests,
  - or whether the test is planned or post hoc.
- You may view these as advantages, but there is no doubt that some people are uncomfortable with these features.
- A Bayesian analysis does not depend on any of these things.

The Likelihood Function



- Our beliefs are updated by multiplying them by the likelihood (and normalizing)
  - Nothing else enters the equation.
  - no other tests, stopping rules, events that never occur etc…
- Researcher A and Researcher B could be using very different stopping rules, come up with their theories at different times, etc. But if they observe the same data, they'll compute the same likelihoods
  - and you'll end up with the same beliefs.
  - The only thing that matters is the data we collect, none of that other stuff
- This idea, that the likelihood should be the only thing to guide our inference, is known as the likelihood principle

Likelihood principle: The idea that all the information in a study relevant to statistical analysis is contained in the likelihood for each hypothesis.

The Appeal of Likelihood

- Many people find the likelihood principle appealing
  - If you see the same data, you come to the same conclusions. Everything besides the data seems extraneous

- Bayesian statistics obeys the likelihood principle.
  - So does a lesser-known school of statistics known as likelihood inference.
- Frequentist statistics does not obey the likelihood principle
  - Two researchers with the exact same data may have to reach different decisions.
  - There's no way to fix Frequentist stats to obey the likelihood principle
    - Decision rules are different than beliefs, and they have to be sensitive to all the things that affect our error rates.
- The likelihood principle is often cited as a major advantage of Bayesian inference
  - And something to consider when you choose a framework.

Comparing Frameworks

Bayes and the Frequentists

Let's list some key differences between our frameworks.

- The output:
  - The output of a Frequentist analysis is a binary decision– reject or don't reject. We can't know how much to believe an individual hypothesis, all we know is that if we follow the rules, our error rates will be controlled.
  - The output of a Bayesian analysis is a set of posterior probabilities.
    - It actually addresses how much we should believe things
    - And it provides a continuous spectrum of beliefs between hypotheses, not just a binary decision.
    - But we can't say how often errors occur
- Parameters (how we model the world):
  - In a Frequentist framework, only one hypothesis is considered at a time. This means that parameters describing the world have a single value. The mean height of flying deer is 40 inches. The probability of a coin landing heads is 50%.
  - To a Bayesian, model parameters are typically random variables, taking on different values with some probability. A Bayesian analysis refines these probabilities in light of data.

Bayes and the Frequentists

- Things that affect our results:
  - The frequentist framework depends on other tests we run, whether our test is planned or post hoc, our stopping rule
  - A Bayesian analysis depends on your prior beliefs.

- Some statisticians are strong proponents of one framework over the other
- Others move between both, or mix elements of both
  - Some problems are more geared towards a Bayesian analysis
    - Well-defined parameter space.
    - Multiple sources of information.
    - Need for a detailed output
  - Others are more geared towards a Frequentist analysis
    - Incomplete knowledge of possible hypotheses
    - Disagreement about priors
    - Clearly motivated null hypothesis

- Now you're aware of a few of the key issues when choosing a statistical framework.
- This was just a very quick look at Bayesian statistics.
- To keep learning, a good first step is Chapter 4 of the text by Dienes. This is a very high-level overview of Bayesian stats and really helps with building up your intuition.