CS109A Week 8 Notes

Ian Tullis

May 17, 2022

I. Beta: setting the scene

I was introduced to the beta distribution in a (frankly pretty boring) stats class long ago. All I remembered about it was that – unlike the normal distribution – it is supported only within a limited range $[0, 1]$, and so it can be a good model for things like exam scores that can't be negative.[1] But there are many other distributions with that property, and the math behind the beta is scary – it has a whole separate gamma function in it! What even is that? So I mostly forgot about it.

Watch out, Cloud! This giant snake knows probability!

Big mistake! The beta distribution is AWESOME[2], and it, along with its multinomial cousin the Dirichlet distribution, has applications all over the place when we need to quantify uncertainty about uncertainty. If you take CS238 (Decision Making Under Uncertainty), which I recommend, you will get much more practice with those.

[1] To use the beta to model scores on an exam out of 100, for instance, we can just multiply it by 100.

[2] You may recall that the last thing I said was AWESOME was linearity of expectation, and I meant it!

Suppose we are the main character in a spy movie. It's still early on in the film, so for some reason we're gambling with the supervillain in a Monte Carlo casino. We'll be fighting them later in the movie, probably on top of a train on a different continent, but for now we have to be superficially polite while trying to crush them at this game.

As is often the case with CS109A casino games, the rules are rather silly. In each round of the game, the dealer – who is in a tuxedo and wearing white gloves – flips a fancy coin made of platinum or something. If it comes up heads, we have to give \$10000 to the supervillain. If it comes up tails, the supervillain has to give \$10000 to us.[3]

We have been playing for a little while, and so far we have lost more rounds than we have won. We don't trust this game *or* the supervillain. They keep smirking at us, and they seem awfully smug about their chances. As they sip from a six-figure glass of 100-year-old port, they assure us that our losses are *simply due to chance*. And, irritatingly, they have a point: we can't completely rule out that possibility. But we think the coin might be unfair – that is, it has some probability p of coming up heads, and we think p is not 0.5. How do we express our uncertainty about this in some way other than punching? (since it is too early in the movie for that)

II. A frequentist approach

We could keep track of the number N of rounds played and the number H of heads seen, and then declare that $p = \frac{H}{N}$. This is the "maximum likelihood" estimate, as we will see in class, and surely that's the only sensible estimate, right? And if that's not precise enough, we can just play more rounds of the game to get more information, our pocketbook be damned.

A shortcoming of this method is that it only gives us a single value as an estimate. What if we want some notion of confidence in that estimate? That is, if we think $p = 0.54$, for example, are we 95% sure that the value is within 0.02 of that, or is there still a decent chance that it could be farther away, maybe even *below* 0.5? (Then we would just look like sore losers!)

Before we dive into the beta distribution, let's investigate the situation through the lens of frequentist statistics – which, broadly speaking, asks "what is the probability of the observed data, given the hypothesis?" In this case, our hypothesis is that the coin is unfair, i.e., $p \neq 0.5$. Specifically, we think it is unfair in the supervillain's favor, i.e., $p > 0.5$. But just showing that the coin is unfair (without saying in whose favor) should embarrass the supervillain (and the casino) publicly. So we declare a *null hypothesis* – basically, that the coin is fair and nothing fishy is going on – and then proceed to argue that that null

[3] The real game of baccarat is awfully close to this.

hypothesis is unlikely to be true, given what we actually saw. And if the null hypothesis is unlikely to be true, then all that remains are the unsavory alternatives...

Problem 1. Suppose that so far, we have played 10 rounds of the game, and we have seen 8 heads and 2 tails.

(a) For a fair coin ($p = 0.5$), what is the probability of seeing 8 heads in 10 flips? (Use Python / Wolfram Alpha / etc. to find the value to 3 decimal places or so.)

(b) This value is pretty small – less than 0.05. We remember from reading scientific journal articles (in our downtime between spy missions) that, by convention, results are thought to be "significant" when $P < 0.05$, i.e. when they have a less than 5% probability of occurring purely by chance. Why is it not a convincing argument to just point to your answer from (a) and say that it is less than 0.05, and so it is very unlikely to have happened by chance?

(Hint: imagine, instead, that we had flipped 1000 coins and seen 500 heads.)

(c) How could we modify the approach in (b) to argue more convincingly? (That is, we want to calculate something other than $P(X = 8)$. What other outcomes should we include?)

(d) Using this new method, what is the probability of seeing the observed data, given that the coin is fair? Does it fall under the 0.05 threshold?

(e) If we had to make a guess at the value of p, based on the results that you have seen so far, what would we say? Why?

Solutions to Problem 1.

(a) The number of heads in 10 flips has the distribution $Bin(10, 0.5)$, and $P(X = 10) = \binom{10}{8}(0.5)^8(1 - 0.5)^{10-8}$. This comes out to $\boxed{\approx 0.044}$.

(b) One problem is that just because an event is low-probability doesn't mean that it can't happen. The supervillain can always counter with this, no matter how we argue, and indeed, they're not wrong. But we have to draw a line somewhere, where a reasonable person would find it implausible that chance was the only explanation. (If this makes you uncomfortable because it feels subjective, well, that's statistics for you! We can try to make the subjective feel more objective, but all the math in the world can never make it completely objective.)

But there is a more subtle issue here, which is that our approach is kind of unfair to the poor supervillain. Suppose we had flipped 1000 coins and gotten 500 heads. The probability of this (for a fair coin) is around 0.025, which is less than 0.05, so we could claim that this was not due to chance. But this is an absurd claim – if we flip 1000 times, what result could be *more* like a fair coin than 500 heads? In fact, in this case, there is *no* outcome that occurs with probability ≥ 0.05. So no matter what happens, we will accuse the supervillain of cheating, even if the setup is fair!

(c) Therefore, in our 10-flip case, we should instead be finding the probability of seeing *at least* 8 heads. That is, our argument will be: even if the coin were fair, the probability of seeing at least this extreme a result is very small.

(d) The values for $P(X = 9)$ and $P(X = 10)$ turn out to be about 0.010 and 0.001. So the overall $P(X \geq 8)$ is $\boxed{\approx 0.055}$, which is over the 0.05 threshold (and therefore not "statistically significant"). So maybe we shouldn't go accusing the supervillain just yet!

(e) It seems most sensible to conclude that $p = \frac{8}{10}$, and we will make this more rigorous in a future CS109 lecture on maximum likelihood. But we might have a bad feeling about this. Do we really feel confident concluding, on the basis of a small amount of data, that p is so large? As an extreme case, what if we had seen three heads in three flips? Does it really make sense to conclude that $p = 1$?

III. A Bayesian approach

Our frequentist methods above gave us an estimate of the coin's probability p of coming up heads (0.8), and a way to argue that $p \neq 0.5$. But what if we want to explicitly quantify our beliefs about, say, how much more likely p is to be 0.8 than 0.79? What if we want to see a distribution of these beliefs, to get a sense of whether they are tightly centered around 0.8 (in which case we can be more

confident) or more diffuse (in which case we probably need more information)? We can do this with frequentist methods as well, but this is where Bayesian methods really shine.

As we've already seen in CS109, Bayesian methods involve bringing in a prior set of beliefs. We might believe that the person who just walked into our ice cream shop has an 0.7 probability of buying some ice cream, just based on overall trends about customers (maybe yesterday we saw 70 out of 100 customers make a purchase). But then we see this particular customer look long and hard at the mint chocolate chip, and maybe our posterior probability goes up to 0.95. Notice that Bayesian methods ask "what is the probability of the hypothesis, given the observed data?", whereas frequentist methods ask the opposite.

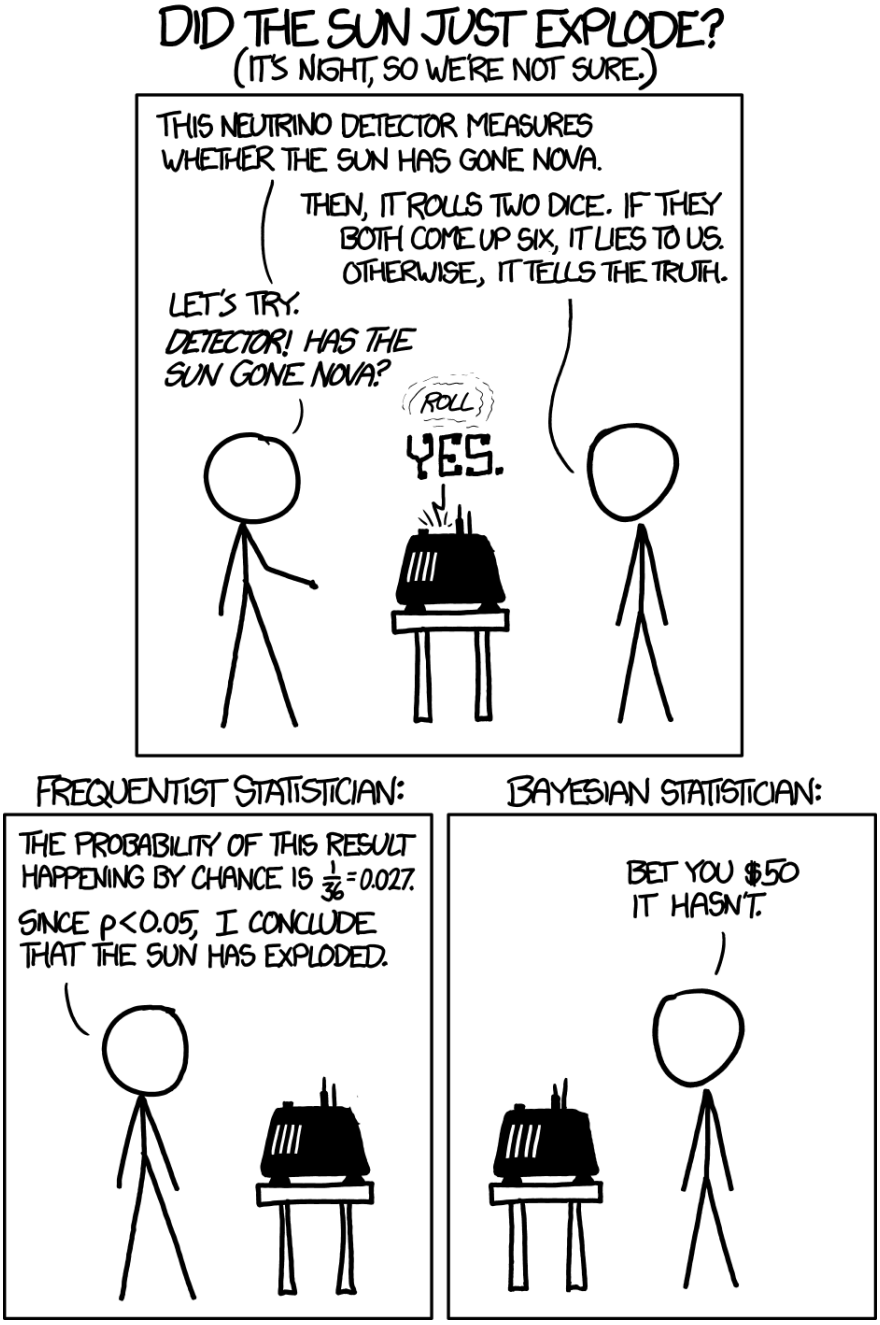
There have been acrimonious disputes between frequentists and Bayesians in the literature and elsewhere. To oversimplify a couple of the points of contention:

- Frequentists think that explicitly bringing your own beliefs into an analysis makes everything hopelessly subjective.
- Bayesians argue that what we really want to know is the probability of the hypothesis given the observed data, and doing it the other way around is artificial and misguided.

My own position is that – as much as I usually dislike lazy both-sides-ism – both approaches really do have their merits. There is no canonically correct way to do statistics, because uncertainty can only be described, not eliminated. Probably the contemporary ethos is to use whatever drives progress forward and brings in the most sweet, sweet cash from venture capitalist investors. Of course, it is dangerous to uncritically just try everything and see what works – this makes it more likely that some approach will only *appear*, by chance, to work well! But I think it's best to understand both frequentist and Bayesian methods well enough to be able to use them in situations that seem appropriate.

So, back to the beta distribution and our spy movie example. The beta distribution itself is not *inherently* Bayesian, but the way we use it in CS109 is. Specifically, we have some set of beliefs about the value of p, and then each time we make an observation (the outcome of one round), we update those beliefs.

What set of beliefs should we have even before the game begins? In this case we might reasonably come in *expecting* the supervillain to be a cheater. Or, we might be unusually magnanimous (for a secret agent), and come in very confident that the coin is fair. A third approach (which is maybe less objectionable to frequentists) is to come in giving every possible value of p equal likelihood, i.e. "flat priors", and this is what we will often do when working with betas.



XKCD comic number 1132.

In the following problem, we will derive the beta distribution and see that it is not such a big scary snake after all...

Problem 2. Say we came into the game believing that every possible value of p was equally likely. Then we saw 8 heads and 2 tails.

(a) First of all, we need to turn "every possible value of p is equally likely" into a PDF – call it $f(x)$ or $f(p = x)$. That is, if we want the probability density of our belief that $p = 0.9$, we evaluate $f(0.9)$. Here we use x to avoid confusion with p, which is a single fixed (but unknown) value representing how unfair the coin really is.

This PDF should just look like a horizontal line floating somewhere above the x-axis; it should have the same value c everywhere. It should also only be supported on the range [0, 1], since p is a probability and can only take on values in that range. What is $f(x)$?

(b) Using Bayes' Rule, we can express $f(p = x|8 \text{ heads out of } 10)$ as

$$\frac{P(8 \text{ heads out of } 10|p = x)f(p = x)}{P(8 \text{ heads out of } 10)}$$

One of those terms is your distribution from part (a) of this problem. Another is very similar to what you found in part (a) of problem 1. Replace both of those terms with expressions in terms of x and/or constants.

(c) We need to use a version of the Law of Total Probability in the denominator, but here we can't write out a finite sequence of terms that look like the numerator. What integral should we use instead? (Use your numerator from this problem, but introduce a new letter like y as the integration variable, to avoid confusion with the x in the numerator).

(d) Evaluate this integral to get a constant, then plug it into our expression from (b) to get our final Bayesian posterior distribution.

(e) Look at the Wolfram Alpha result for `beta distribution with alpha = 9, beta = 3`. You should see that it gives the same PDF. Visually find the value of x for which $f(x)$ is maximized; is it what you expect?

(f) What is the CDF of the beta distribution that you just found? (Feel free to use Wolfram Alpha to do the integral.) Using this CDF, what is the probability that the coin is quite unfair – i.e. has $p \geq 0.55$, for example?

(g) Why could we not have answered part (f) directly using frequentist methods? (Put differently, what did we include in our Bayesian method that let us answer that question?)

(h) A uniform distribution is $Beta(1, 1)$. How different would our posterior beta distribution look if we had instead started with Laplace smoothing, i.e., $Beta(2, 2)$? (Don't go through all the steps of the problem again – just consider what the final beta distribution's parameters α and β would be in this case. Then use Wolfram Alpha to plot that.)

Solutions to Problem 2.

(a) This uniform distribution is a PDF, so it has to have $\int_0^1 c dx = 1$. Integrating $\int_0^1 c dx$, we get $[cx]_0^1 = c(1) - c(0) = c$. So $c = 1$, and therefore $\boxed{f(x) = 1}$.

(b) For $P(8 \text{ heads out of } 10)|p = x)$, we again use the binomial distribution, but now with a success probability of x. So this evaluates to $\binom{10}{8}x^8(1-x)^2$, which is $45x^8(1 - 2x + x^2) = 45x^8 - 90x^9 + 45x^{10}$. (This form will be easier to integrate later on.)

Plugging in that expression and our prior from part (a) (which is just 1, so it goes away), we now have

$$f(p = x|8 \text{ heads out of } 10) = \frac{45x^8 - 90x^9 + 45x^{10}}{P(8 \text{ heads out of } 10)}$$

(c) The integral is

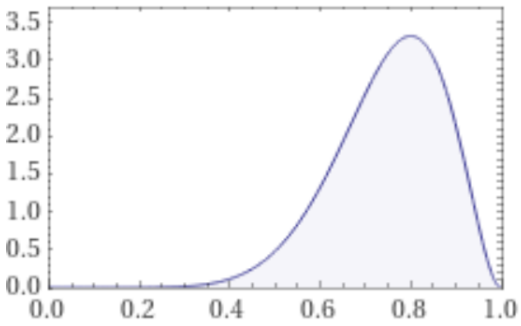
$$\int_0^1 (45y^8 - 90y^9 + 45y^{10})dy$$

(d) Evaluating this, we get

$$[5y^9 - 9y^{10} + \frac{45}{11}y^{11}]_0^1 = 5 - 9 + \frac{45}{11} = \boxed{\frac{1}{11}}$$

$$f(p = x|8 \text{ heads out of } 10) = \frac{45x^8 - 90x^9 + 45x^{10}}{\frac{1}{11}} = \boxed{495x^8 - 990x^9 + 495x^{10}}$$

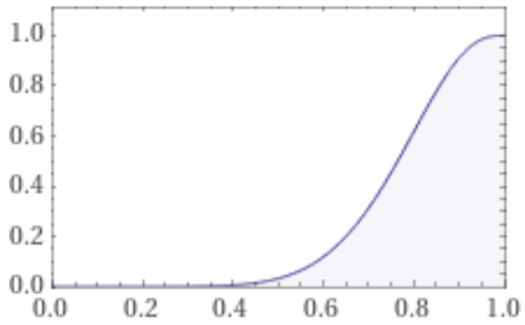
(e) The PDF looks like this:



and the maximum is at 0.8, which is what we would expect. (It is exactly 0.8, as you could check directly via calculus.)

(f) As usual, the CDF is the integral of the PDF from the lower end of the support range to some stopping point y.

$$\int_0^y (495x^8 - 990x^9 + 495x^{10})dx = [55x^9 - 99x^{10} + 45x^{11}]_0^y = \boxed{55y^9 - 99y^{10} + 45y^{11}}$$

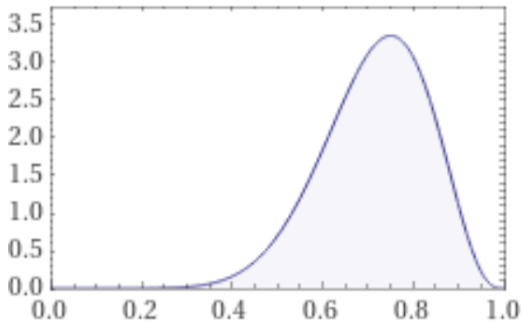


To find the area of the part of the PDF that is *below* 0.55, we evaluate the CDF at 0.55 to get $55(0.55)^9 - 99(0.55)^{10} + 45(0.55)^{11} \approx 0.065$. Then the area that is above 0.55 is $\approx 1 - 0.065\boxed{\approx 0.935}$. That is, we have a pretty strong belief that the true probability p of the coin coming up heads is 0.55 or higher.

(g) The critical piece of information that the Bayesian method brought in was the assumption that, in the absence of other information, all values of p were equally likely. Without such an assumption, it doesn't even make sense to talk about the *probability* of p taking on any particular value. It depends on how evil the supervillain was feeling this morning, what fake coins they own, and so on. This seems hopelessly complicated to quantify. So maybe it's not so bad to make a very basic assumption (flat priors) and then iterate from there?

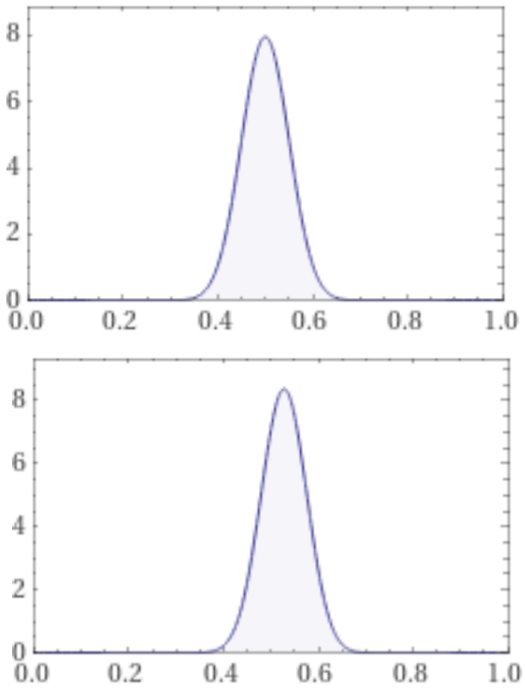
However, we shouldn't lose sight of the fact that our probability estimate includes an assumption. That is, we have not objectively determined that the true probability of $p \geq 0.55$ is 0.935. If we had used a non-flat prior distribution, we would have gotten very different results...

(h) Starting with $Beta(2, 2)$ and adding 8 "successes" to α and 2 "failures" to β, in this case we would end up with $Beta(10, 4)$:



This is fairly close to our original $Beta(9, 3)$, but now our best estimate of p is slightly below 0.8 – it has been dragged down a bit because the Laplace smoothing essentially adds one "bonus" head and one "bonus" tail into the mix.

On the other hand, suppose that we come in naively believing that the coin is very fair, i.e., something like $Beta(50, 50)$. Then after our 10 flips, we believe $Beta(58, 52)$. Our prior beliefs were so strong that the new data barely changes them. So – as a frequentist would point out – the choice of prior distribution can *dramatically* change the results!



IV. That Cal Chip On The Shoulder

Suppose that our friend from UC Berkeley[4] claims that Cal students sleep less than Stanford students.[5] She interviewed 9 Cal students and 7 Stanford students, and asked each person how many hours of sleep they got in the last week.[6] Suppose that the responses were:

- Cal students: 36, 59, 40, 53, 48, 48, 28, 36, 48 (mean = 44)
- Stanford students: 55, 40, 60, 48, 53, 50, 37 (mean = 49)

Our friend points out the large difference in means. As in the spy problem, we are skeptical. For one thing, these are very small sample sizes. How do we know the results aren't just due to chance? That is, what if she unwittingly interviewed Cal students who don't happen to sleep as much, and Stanford students who happen to get more sleep?

We want tell our friend to go out and collect more data, but she is currently asleep. (Suspicious!) Luckily, we just learned about bootstrapping in CS109. Bootstrapping seems like a way to wring more truth out of a limited set of data. Is it?

When we bootstrap, we are making a massive assumption, which is so massive that I'm going to break out the LaTeX
`Large` environment:

So if the overall set of data we collected was badly biased, this assumption is illegitimate and we are already out of luck! In a small enough dataset, this might happen due to chance, which is the very thing we are trying to use bootstrapping to argue about!

We can proceed with our analysis, but we (and our friend) should be aware of the inherent limitations. There is no way to magically get more data from less data without paying a price, and even though bootstrapping can give us objective-sounding values, we should never forget that bolded assumption above. If we don't believe it, then we shouldn't believe the results of bootstrapping either.

That said, let's look at how we would use bootstrapping in this situation. We

[4] I have a Master's degree from Cal, so I am allowed to make fun of it.

[5] The r/berkeley Reddit certainly does seem to be more stressed out, on average, than r/stanford, but Stanford students also seem to go to greater lengths to hide their stress. So I'm honestly not sure what the real answer is here.

[6] The official position of CS109A is that getting more sleep is a *good* thing. Sometimes it is worth setting the pset aside and letting your subconscious make sense of things overnight!

are going to assume that the *combined* set of Cal and Stanford data accurately represents the *combined* Cal and Stanford population. We think of this combined population as a distribution rather than a set of 16 people. The draws are independent and identically distributed, i.e., drawing a person with 59 hours of sleep doesn't "use up" that value or make it less or more likely on future draws.

We will repeatedly do the following: draw a new fake "Cal" sample of 9 students from that distribution, draw a new fake "Stanford" sample of 7 students from that distribution, find the difference between the "Stanford" mean and the "Cal" mean, and compare it to the actual difference between the means of the Stanford and Cal samples.

Problem 3.

(a) Why don't we draw our "Cal" sample from only the original Cal students, and the "Stanford" sample from only the original Stanford students?

(b) Why is it important that our "Cal" and "Stanford" samples have the same sizes as the real ones?

(c) Suppose we do the following: run 100000 trials, and count the number of trials in which the difference in means between the "Stanford" and "Cal" samples equals 5 (which is the real difference). We then divide this number of trials by 100000, find that the result is very small, and conclude that the actual observed difference is unlikely to have arisen by chance. What's wrong with this argument? How do we fix it? (This should remind you of something from last week's 109A...)

(d) Also thinking back to last week's 109A, is this a frequentist or Bayesian method?

(e) When we run the corrected version of the method in part (c), suppose we get a value of 0.13046. What should we conclude from this? (What can we say to our friend?)

(f) Is it ever possible for a bootstrap setup like this to be *unable* to see a difference as large as the one observed in the real data?

Solutions to Problem 3.

(a) The "null hypothesis" underlying this method is that there is no difference between the Cal and Stanford samples, i.e. they are part of the same overall group. Then we ask: if this null hypothesis is true, how often would we see the same kind of difference that we actually saw in the real data? If that turns out to happen commonly by chance, then we should be skeptical that the difference in means is based on any real difference between Cal and Stanford.

However, if we choose a fake Cal from the Cal samples and a fake Stanford from the Stanford samples, we are just reproducing the original (and possibly unrepresentative) difference from the data! That defeats the purpose of what we are trying to do: see how often such a difference would arise by chance.

(b) A smaller sample is inherently less likely to look representative, in a way that contributes to an artificial (chance-based) difference, so sample sizes do matter when we do our bootstrap.

As a thought experiment, suppose that when bootstrapping, we generated "Cal" and "Stanford" distributions of 50000 students each. (Why not? We can keep drawing as many people as we want!) But then these two samples would pretty much never be very different, so we would pretty much always conclude that the observed real differences couldn't have arisen by chance.

Or, on the other hand, suppose that we generated "Cal" and "Stanford" distributions of 1 student each. Then we would see differences ≥ 5 very often, and we would be very prone to concluding that the original difference was due to chance.

(c) The probability of the difference in the fake sample means turning out to be *exactly* 5 is very small, so we will always conclude that the observed difference is unlikely to have arisen by chance, and is therefore significant/real. But this is the wrong metric – we want to know the probability of seeing a difference *at least* as extreme as the real difference. That is, the real argument here is about whether the difference of 5 counts as large enough to not be just noise... *not* about whether that *exact* value, 5, is likely.

(d) This is a frequentist method; we are finding a p-value, i.e., the probability of seeing (at least this extreme of) a result due to chance alone (i.e. under the null hypothesis). It is not Bayesian; we are never bringing in a prior belief.

(e) This result is a p-value: $p = 0.13046$. It is not less than the standard threshold of 0.05, so we conclude that the observed difference could be just noise – we have a reasonable doubt.

Beware of those many digits of precision, though! For one thing, there is inherent randomness in the bootstrap procedure itself, so we would almost certainly get a different (but pretty close) p-value from another set of 100000 trials. For another thing, the assumption underlying the bootstrap method is probably questionable here – the combined sample size of 16 is still unlikely to be very representative of the entire population. (Where do we draw the line? What counts as "sufficiently representative"? This might be a fun topic to explore for a challenge project!)

(f) No, because a bootstrap trial can always (in theory) reproduce exactly the original data, just by chance. So there is at least some positive probability of seeing a difference at least as large as the real one.

By the way, here is the code I wrote to do the bootstrapping:

```
import numpy as np

cal = [36, 59, 40, 53, 48, 48, 28, 36, 48]
stanford = [55, 40, 60, 48, 53, 50, 37]
true_diff = np.mean(stanford) - np.mean(cal)
combined = cal + stanford  # you can concatenate two lists with +

num_trials = 100000
at_least_as_extreme = 0
for _ in range(num_trials):
    fake_cal = np.random.choice(combined, size=len(cal), replace=True)
    fake_stanford = np.random.choice(combined, size=len(stanford), replace=True)
    fake_diff = np.mean(fake_stanford) - np.mean(fake_cal)
    if fake_diff >= true_diff:
        at_least_as_extreme += 1

print("p-value is: ", at_least_as_extreme/num_trials)
```

V. An only mildly scary convolution

This is a small stretch past CS109 material, but it can be satisfying to see how two normal distributions add.

Problem 4.

(a) Let X_1 and X_2 be independent standard normal random variables. Let $Y = X_1 + X_2$. Without using any normal PDFs, integrals, etc. yet – based just on what you have learned about adding independent normal random variables – what distribution and parameters would you expect Y to have?

(b) Now, write an expression for $f(Y = y)$ in terms of $f(X_1)$ and $f(X_2)$ – don't use the normal PDF yet. It should be an integral from $-\infty$ to ∞.

As a hint, recall a similar discrete case: let Z_1 and Z_2 be the results of rolling (independent) single 6-sided dice, and let $W = Z_1 + Z_2$. Then $P(W = w) = \sum_{z=1}^{6} P(Z_1 = z)P(Z_2 = w - z)$. This is the sum over all the ways that the two dice can add up to w, where we call the result of the first die z, and therefore the second die must be $w - z$.

(c) Recall that a normal distribution with mean μ and variance σ^2 has the PDF:

$$f(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

What is the PDF of a *standard* normal distribution?

(d) Replace $f(X_1)$ and $f(X_2)$ in your expression with standard normal PDFs. Then rearrange the terms so that the parts that do not depend on y are outside the integral. Finally, use the fact, which you can verify on Wolfram Alpha, that – with k being any constant and u being the integration variable –

$$\int_{-\infty}^{\infty} e^{-u^2 - ku} du = \sqrt{\pi} e^{\frac{k^2}{4}}$$

What does $f(Y)$ end up being, in terms of y? What distribution is this, and what are its parameters? Does this match what you expect? Isn't this neat?

There are very few distributions that have this property; it's not important for CS109, but they are called *stable*. Another example is the Cauchy distribution, which is also not important for CS109, but has some weird and fun properties like an undefined mean and variance (and it comes up in actual applications, e.g., in CS261).

Solutions to Problem 4.

(a) We have seen in class that if we add two independent normal random variables distributed as $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, the result is another normal random variable distributed as $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. In this case, since we are adding two standard normal RVs, this becomes $\mathcal{N}(0, 2)$.

(b) The analogous expression to the die example is

$$f(Y = y) = \int_{-\infty}^{\infty} f(X_1 = x) f(X_2 = y - x) dx$$

That is, we are taking a kind of "sum" (an integral) over all of the (infinite) ways that the two variables X_1 and X_2 can add up to $Y = y$.

(c) Plugging in $\mu = 0$ and $\sigma^2 = \sigma = 1$, we get

$$f(X = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(d) Using $f(X_1 = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $f(X_2 = y - x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}}$ in our integral from part (b), we get

$$f(Y = y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} dx$$

Simplifying this somewhat, we have

$$f(Y = y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2 - 2xy + x^2}{2}} dx$$

Simplifying further to move the y-only term out:

$$f(Y = y) = \frac{1}{2\pi} e^{-\frac{y^2}{2}} \int_{-\infty}^{\infty} e^{-x^2 - xy} dx$$

Using the integral given in part (d), this becomes

$$f(Y = y) = \frac{1}{2\pi} e^{-\frac{y^2}{2}} \sqrt{\pi} e^{\frac{y^2}{4}}$$

and all this boils down to

$$f(Y = y) = \frac{1}{2\sqrt{\pi}} e^{-\frac{y^2}{4}}$$

which is the PDF for a normal distribution with mean 0 and variance 2, just as we expected in (a). HELL YEAH