CS109A Week 8 Notes

Ian Tullis

February 22, 2022

I. More Stressed Than Thou

Suppose that our friend from UC Berkeley claims that Cal students sleep less than Stanford students.[1] She interviewed 9 Cal students and 7 Stanford students, and asked each person how many hours of sleep they got in the last week.[2] 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. 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:

the observed distribution of our data is the same as the real distribution of the entire underlying population

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

[1] 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.

[2] 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!

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 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 1.

(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 1.

(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)
```

II. Let's ask about something else

Why do a boring survey about sleep when there are more important matters to ask about? This is CS109 – we should be engaging with real-world issues!

Problem 2.

(a) Suppose that any random person is 30% likely to prefer Squirtle as a Gen 1 starter, 20% likely to prefer Bulbasaur, and 50% likely to prefer Charmander[3]. We go out on the streets of Palo Alto and ask 12 people what their favorite starter is. What is the probability that there will be a three-way tie? (Hint: there is a distribution that is perfect for this!)

(b) Now suppose instead that we go around asking people what their favorite Pokémon is. As of the time that these notes were written, there are 905 Pokémon. For now, assume, somewhat unrealistically, that each person we talk to is equally likely to prefer any of them. What is the expected number of people we will need to talk to in order to get *every* Pokémon as an answer at least once? (Come up with an expression, and then use Wolfram Alpha or Python to evaluate it.)

Hint: Break this up into a series of checkpoints. The first checkpoint is that we are trying to get our first answer that we haven't heard yet. This is trivial; no matter what the first person says, we satisfy this requirement. Then the second checkpoint is that we are trying to get our second answer that we haven't heard yet. So we just need to keep talking to people until we find one who doesn't give the answer we already got – what is the expected number of people we will need to ask? And so on.

(c) The assumption in part (b) is obviously unrealistic, since more people are going to prefer, e.g., Garchomp to Stunfisk. Roughly how would you expect this to influence the answer to (b)? As a specific example, suppose that the preferences follow a *power law* distribution in which the second-place choice is half as likely as the first-place choice, the third-place choice is one-third as likely as the first-place choice, and so on. Or, what about an extreme case in which there is some Pokémon (*cough* Mr. Mime *cough*) who is *by far* the least popular?

[3] and therefore be correct

Solutions to Problem 2.

(a) This is a job for the multinomial distribution! Specifically, let S and B be random variables for the numbers of people (out of the 12) who prefer Squirtle and Bulbasaur, respectively. Then

$$P(S = s, B = b) = \binom{12}{s, b, 12 - s - b} 0.3^s 0.2^b 0.5^{12-s-b}$$

Plugging in $s = 4$ and $b = 4$, we have $\binom{12}{4,4,4} 0.3^4 0.2^4 0.5^4 = \frac{12!}{4!4!4!}(0.3 \cdot 0.2 \cdot 0.5)^4 \boxed{\approx 0.028}$.

What if we don't remember the form of the multinomial coefficient? Well, we have 12 people. We first want to pick 4 of them to be Squirtle fans. Then we pick 4 of the remaining 8 to be Bulbasaur fans, and all of the leftover 4 are Charmander fans. So the total number of ways is $\binom{12}{8}\binom{8}{4}\binom{4}{4}$. But this is $\frac{12!}{8!4!} \cdot \frac{8!}{4!4!} \cdot \frac{4!}{4!0!}$, which simplifies to $\frac{12!}{4!4!4!}$.

(b) This problem is very similar to the card shuffling problem (1d) on the Spring 2016 practice midterm. It is also an instance of an important and ubiquitous phenomenon in combinatorics and algorithms: the *coupon collector problem*. The name is supposed to suggest a contest in which there are many types of coupon, and you get a coupon of a uniformly randomly selected type e.g. each time you make a purchase, and you need to collect at least one coupon of each type to win. Intuitively, it is easy to get "new" (i.e., previously unseen) types early on, but then those last few that you don't have become harder and harder to get, as you keep (frustratingly) getting types you already have!

Let's see what the math says. Proceeding in accordance with the hint: the first person surely names a Pokémon we haven't heard yet, and then we just have to hear a Pokémon other than that first kind. The chances of this are $\frac{904}{905}$, but we could get unlucky and hear the first Pokémon again (chances $\frac{1}{905}$) before asking another person, and have a $\frac{904}{905}$ chance once again... We see that this is a geometric distribution with $p = \frac{904}{905}$, so the expected number of people we will need to get our next different Pokémon is $\frac{1}{p} = \frac{905}{904}$.

Proceeding in this way, reaching the next milestone (a third different Pokémon) is another geometric distribution with $p = \frac{903}{905}$, so we need to ask an expected $\frac{905}{903}$ people, and so on. When we are looking for our 905th and final distinct Pokémon, we only have a $\frac{1}{905}$ chance of getting that one, so in expectation, it takes 905 people to get through this phase!

Therefore the answer is $\boxed{\sum_{i=1}^{905} \frac{905}{i}}$, which means we need to talk to ≈ 6684

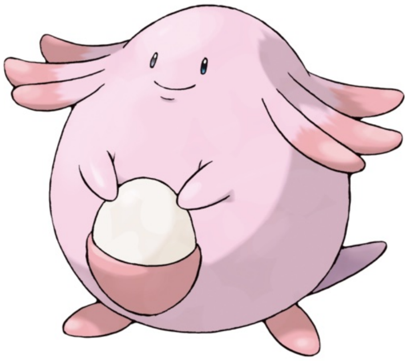
people (in expectation) to catch 'em all. Notice that the summation looks like it goes in the opposite order of our argument above, but the order of the sum does not matter.

In general, the coupon collector problem with n distinct coupons (each equally likely to be chosen) has an answer that is $\mathcal{O}(n \log n)$.

(c) As we saw above, the last few different Pokémon are the hardest to get. In fact, just those last *ten* account for almost 40% of the overall answer! But with the assumption that all Pokémon are equally preferred, no *particular* Pokémon are inherently hard to get.

If we change to e.g. a power-law distribution, though, we would expect the rarest Pokémon (the ones that are 900, 901, ..., 905 times less frequent than the most popular one) to heavily determine the amount of people we need to talk to. We will see 905 of the single most popular Pokémon (say, Charizard) for every one of the least popular Pokémon (say, Mr. Mime). In fact, out of every $\sum_{i=1}^{905} i = 409965$ Pokémon, we would expect only *one* of them to be Mr. Mime! So the expected number of people we need to talk to should be at least 409965, and probably greater, because we may well have failed to see some of the other very-unloved Pokémon. I wrote some code, and the average of 10000 trials was around 515000. The smallest result was 56573, and the largest result was 4135136.

In the even more extreme case that one Pokémon is *much* rarer than the others, that one might almost singlehandedly determine the final answer, i.e., we may find one or more full sets of the other 904 before we find even one of that one. A good approximation of the answer, then, might be $\frac{1}{p}$, where p is the probability of finding the least-liked Pokémon.



Chansey is probably the most CS109 of all Pokémon: the name suggests randomness, ***and*** *it has type Normal.*

III. Just another exciting Friday night

Problem 3. Suppose that we just bought a box of 1000 loose quarters at the bank. We are hoping to find one of the quarters minted in 2019 or 2020 at West Point – they have a “W” mintmark, and not many were made, so they are potentially worth 10 or 20 bucks to collectors. Suppose, somewhat optimistically, that about 1 in 10000 quarters in circulation has a “W” mintmark.

This is one of 10 kinds of “W” quarters in circulation. Ian has found two of these kinds so far.

(a) What is an expression for the exact probability $P(X = x)$ that we will find *exactly* X “W” quarters in the box?

(b) Find and evaluate a Poisson approximation to $P(X \geq 1)$, i.e., the probability that we find *at least* one “W” quarter in the box.

(c) Find and evaluate a normal approximation to $P(X \geq 1)$.

(d) Without knowing the actual answer, which of the two approximations would you trust more? Why?

(e) Which of the approximations required a continuity correction? (Neither? One? Both?)

(f) Suppose we instead wanted the probability of finding *exactly* one “W” quarter in the box. You don’t need to recalculate this, but does this change your answer to (e)?

Solutions to Problem 3.

(a) The distribution here is a binomial: $P(X = 1) = \binom{1000}{1}(0.0001)^1(1 - 0.0001)^{999}$. This is $\boxed{\approx 0.0905}$.

(b) We set the Poisson's λ to be the mean of the binomial distribution, which is $np = 1000 \cdot \frac{1}{10000} = \frac{1}{10}$. Then to find $P(X \geq 1)$, we take $1 - P(X = 0) = 1 - \frac{e^{-0.1} \cdot (0.1)^0}{0!} = 1 - e^{-0.1}$, which is $\boxed{\approx 0.0952}$.

(c) We set the normal's μ to be the mean of the binomial distribution, 0.1, and the variance to be the binomial's variance of $np(1-p) = 0.09999$. Then to find $P(X \geq 1)$, we need to use the translation of the "1" bucket in discrete-land to [0.5, 1.5] in continuous-land. We want the probability of being in at least the 1 bucket (or higher), so, using the continuity correction, we take $1 - \Phi(\frac{0.5-0.1}{\sqrt{0.09999}})$. Using the CS109 CDF calculator, this comes out to $\approx 1 - 0.8971 \boxed{\approx 0.1029}$.

(d) The Poisson approximation works well when p is very small and n is large.[4] Both of those things are true here!

The normal approximation, however, doesn't always do so well when the individual distributions that are being added have a very skewed shape. Each one is a Bernoulli with $p. = 0.0001$, i.e., the PMF has $P(0) = 0.9999, P(1) = 0.0001$. If you think about the distribution resulting from adding together a small number of these, it's clear that it still looks nothing like a Gaussian. Will adding 10000 of them together be enough for the awesome power of the Central Limit Theorem to kick in?

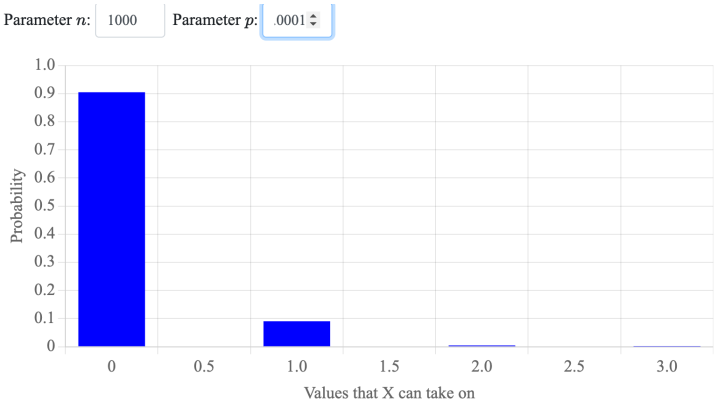
No, as it turns out! The real answer, which you can get as

$$\sum_{i=1}^{1000} \binom{1000}{i} (0.0001)^i (0.9999)^{1000-i}$$

is ≈ 0.0952. So the Poisson approximation is spot on and the normal approximation is quite a ways off!

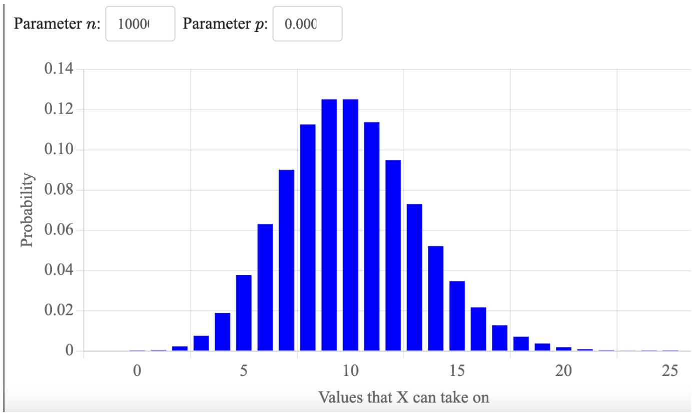
[4] See problem 2 in the Week 6 notes to review why this is.

If we look at the shape of the real binomial distribution, it's no wonder a normal curve can't handle it very well:



Parameter n: 1000 Parameter p: .0001

The point of this problem is to demonstrate that the Central Limit Theorem, as amazing as it is, does **not** mean that the normal distribution is the only one we ever need again! Depending on the distribution in question, it may take a *very* large number of them indeed for the sum to start to look Gaussian. If we use $n = 100000$ rather than $n = 1000$ for the above problem, then we actually do get something kinda Gaussian-looking:



Parameter n: 1000 Parameter p: 0.000

(e) The Poisson distribution is discrete-valued, so a Poisson approximation of a binomial goes from discrete-land to discrete-land, and no continuity correction is needed. But the normal distribution is continuous-valued, so a normal approximation of a binomial goes from discrete-land to continuous-land, which is why we needed the continuity correction in part (c) above.

(f) No. In this case, to get the area under the curve in continuous-land corresponding to the "1" bucket in discrete-land, we would find $\Phi(\frac{1.5-0.1}{\sqrt{0.09999}}) - \Phi(\frac{0.5-0.1}{\sqrt{0.09999}})$. We are still using continuity corrections.

IV. 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