CS109A Week 9 Notes

Ian Tullis

March 1, 2022

I. That's a hard pass

One of the hardest tests I'm aware of is Level 1 of the Kanji Kentei, the standard test of knowledge of Chinese characters as they are used in Japanese. You have to know *everything* about ≈6000 characters[1] **and** their compounds, most of which are very obscure at best, and there is no standard study list. The test is pass/fail, but can be attempted as many times as one likes until one passes.

Suppose that the exam board decides that too many people are passing – all it takes is getting lucky once! They propose to prevent this by requiring that the test be passed *twice in a row*. That is, passing once gives you a "partial pass", and then you must pass the test on your next attempt to get an "overall pass". Otherwise, you lose your partial pass status and have to start over.

Problem 1. Suppose we want to get an overall pass. We have a 10% chance of passing on each attempt, independently across attempts.

(a) What is the expected number of times X we will need to take the test before getting our first **partial** pass (i.e. passing the test once)?

(b) Now suppose we want to know the expected number of times X' we will need to take the test before getting an **overall** pass. At first it may seem like taking the expectation of a negative binomial distribution (with success probability p, and $r = 2$) works for this, but it doesn't. Why not?

(c) So what *is* the expected number of times X' we will need to take the test before getting an overall pass? (Hint: I recommend thinking about this in terms of Chris's "algorithmic analysis" lecture – that is, try to write an expression for $E[X']$ that depends on one or more other instances of $E[X']$. Think about the various scenarios that can happen – e.g., we can fail right away, or pass once and then fail, or...)

(d) Why doesn't it work to say that the probability of two passes in a row is $0.1^2 = 0.01$, then use the geometric distribution to get $E[X] = \frac{1}{0.01} = 100$?

[1] Even this is mercifully fewer than the total number of characters (including obscure ones) in Chinese.

Solutions to Problem 1. We will use p to represent the success probability, just to keep things general, even though the problem tells us that $p = 0.1$.

(a) Setting aside the weird rules about getting an overall pass, this is an instance of a geometric distribution: we fail the test until we succeed once, at which point we have a partial pass. So the answer is the usual expectation for a geometric distribution: $\boxed{\frac{1}{p} = 10}$.

(b) The negative binomial distribution counts the number of trials needed to get a total of r successes. But that's not how the new rules of this test work. For example, suppose that we pass, then fail, then pass. The negative binomial distribution with $r = 2$ would say we are done, but this is not an overall pass: we got a partial pass, then lost it, then got another partial pass.

(c) Suppose we are starting this whole process and about to take our first test. Let $E[X']$ be the expected number of tests until we get an overall pass. Let's break down what can happen:

- With probability $1 - p$, we fail. Then we have taken one test, but *we are still right back where we started.* So in this scenario, the total expected number of tests is 1 (the attempt we just made) plus $E[X']$ (since it's like we're starting over).
- With probability $(p)(1-p)$, we pass once (getting a provisional pass), then fail. Then we have taken two tests, but we are right back where we started. So in this scenario, the total expected number of tests is 2 (the two attempts we just made) plus $E[X']$.
- Otherwise, with probability p^2, we pass twice and are done (we get our overall pass). In this scenario, the total number of tests is 2.

After convincing ourselves that these situations are mutually exclusive and exhaustive, we can write the following:

$$E[X'] = (1-p)(1 + E[X']) + p(1-p)(2 + E[X']) + p^2(2)$$

Simplifying:

$$E[X'] = 1 - p + E[X'] - pE[X'] + 2p + pE[X'] - 2p^2 - p^2E[X'] + 2p^2$$

$p^2E[X'] = 1 + p$. (Now THAT's what I call cancellation!)

$$E[X'] = \boxed{\frac{1+p}{p^2} = 110}$$

(d) It is true that (on any two consecutive tests) the probability of two passes in a row is 0.01, and the answer of 100 is pretty close, but the method is invalid. The geometric model assumes that we keep making attempts, failing zero or more times, until we succeed. But here a "success" entails passing twice, and consists of two rounds, whereas failures can be either one round (immediate fail) or two rounds (pass then fail). The geometric distribution can't account for this difference in what counts as an attempt.

II. A break from bootstrapping

Although it's not covered in CS109, it's worth knowing that the t-test is one alternative to bootstrapping[2] when we want to argue that a difference between the means of two samples is (or isn't) likely to be due to chance. It's another frequentist method that produces a p-value that tells us how often we expect to see as extreme a difference in means as the actual difference just by chance – i.e., in the null hypothesis world in which there is no real underlying distinction between the two samples.

Unlike bootstrapping, the t-test requires us to assume that our populations are normally distributed. But what if we have a good reason to think that our samples *aren't* normally distributed? And what if we also don't want to bootstrap? There is no magic bullet, of course, but there are other statistical methods that do not assume a particular distribution for the data. Today we'll look at one of them that is based only on *rankings* within the data.[3] This test is also not covered in CS109, but it uses ideas from the class, and it will be a good excuse to review some combinatorics!

We'll walk through it beginning on the next page...

You may have heard of "Student's t-test", but Student is actually this man, William Gosset. He discovered the idea while working as a Guinness brewer, and had to hide the company's (and his) identity when publishing, hence the weird pseudonym.

[2] It's really the other way around. The t-test predates bootstrapping by about a century.
[3] If you want to read more about this, check out the Mann-Whitney U-test.

Problem 2. Suppose that our friend from Cal (from last week) has woken up and wants to THROW DOWN with another claim: Cal students drink more milk tea than Stanford students. As before, she surveyed some students about how many milk teas they have per month, and she got the following results.[4]

- Cal students: 67, 5, 15, 30, 18
- Stanford students: 0, 17, 4, 2, 1

The values do *not* seem to be normally distributed, which makes sense; some people drink lots of milk tea, whereas others have it rarely or not at all. Let's walk through the method, which doesn't care about the non-normality:

(a) Treating the two samples as one combined pool (just like in bootstrapping!), assign ranks from 1 to 10 to each of these values. The largest value gets 1, and the lowest value gets 10.[5]

Then find the sum S_C of the ranks of the Cal students.

(b) Suppose that we had ignored the data and assigned the ten ranks from 1 to 10 to the two samples at random, keeping the sample sizes the same at 5 each. Treating the students within each sample as indistinct (i.e. looking at just the sets of ranks), how many ways are there to do this? For example, one possibility would be: Cal gets the set of ranks $\{3, 4, 7, 8, 9\}$, and Stanford gets the set of remaining ranks $\{1, 2, 5, 6, 10\}$.

(c) Now let's step into the null hypothesis world in which there is no difference between the two groups. Suppose that you performed the operation in part (b), but then also found the rank sum S'_C, as in part (a). What is the probability – across the randomness of how the ranks get assigned – of seeing an S'_C *at least* as small as your answer to (a)?

Notice that you don't need to simulate anything here; this can be calculated exactly. What sets of 5 distinct ranks can you write down that add up to less than or equal to your answer to (a)? It turns out there are very few of them.

(d) The value you computed in (c) is a p-value. Based on the result, do you believe your friend's assertion?

(e) Although this rank-based method does not make explicit assumptions about the data, can you think of any shortcomings, or situations in which it might not work well?

[4]The sample sizes are equal here for ease of explanation, but the method generalizes to samples of different sizes.

[5]The values are all different here, also for ease of explanation, but the method can handle ties.

Solutions to Problem 2.

(a) The ranks are:

- Cal: 1, 6, 5, 2, 3
- Stanford: 10, 4, 7, 8, 9

So $S_C = 1 + 6 + 5 + 2 + 3 = \boxed{17}$.

(b) We can view this as choosing 5 out of the 10 ranks to assign to Cal (and then giving the remaining 5 to Stanford), and the number of ways to do this is just $\boxed{\binom{10}{5} = 252}$.

(If you want a bit of extra review for the final: what would have changed about this if we had had a tie? e.g. ranks 1, 2, 3, 4, 5.5, 5.5, 7, 8, 9, 10)

(c) Just by inspection / trial and error, the sets of 5 ranks that sum to 17 or less are:

- {1, 2, 3, 4, 5}
- {1, 2, 3, 4, 6}
- {1, 2, 3, 4, 7}
- {1, 2, 3, 5, 6}

(Notice that if we had 1, 2, 4 as our smallest values, then the remaining two would sum to at least 5 + 6, which is too big. So all the sets have to include 1, 2, 3.)[6]

Therefore there are only 4 ways for this to happen, out of a possible 252 (as found in part (b)), so the probability is $\frac{4}{252} = \boxed{\frac{1}{63}}$.

(d) Because $p < 0.05$, we might provisionally believe our friend this time. That is, it is very unlikely that just by chance, so many of the highest-ranked values would end up on the Cal side.[7]

(e) Using ranks is great for dealing with outliers (like the 67 in this data set), but not so great in other situations in which many similar values are clustered together. Suppose that we had had a bunch of data points like 22, 23, 24, 25, 26 here... the differences between these could be just noise, but the method would give the difference between 23 and 24 (or 22) just as much weight as the difference between 30 and 67 from the original dataset!

[6] If you liked this part, you might consider checking out Kakuro / Cross Sums puzzles! But... maybe after the quarter is over...

[7] You might wonder: shouldn't we also account for the possibility of seeing this lopsided a difference, but in favor of Stanford instead of Cal? This is the difference between "one-tailed" and "two-tailed" statistical tests, if you want to read more about that.

So this kind of rank-based test is most appropriate when the values are pretty well separated. If we feed the test a bunch of ranks based on extremely similar values, where the differences are within the margin of the noise inherent in measurement... then, garbage in, garbage out.

III. MLE: Half-Life Confirmed

For a moment let's put on our experimental physicist hats. We have just created three atoms of the same new element, and we are watching each one to see when it decays. Suppose we find that these decay times are 40, 70, and 85 nanoseconds, respectively. (We have a *very* keen eye and are *very* fast with a stopwatch!)

There is a good scientific reason to believe that the time taken for each atom to decay follows an exponential distribution. But we don't know the parameter λ. How can we estimate it? In this situation it's not obvious what we should do! When we flip a coin 100 times and see 57 heads, it feels kinda obvious that the best estimate of the coin's true probability of coming up heads is 0.57. But we might have less intuition when it comes to scary exponential functions!

Problem 3.

(a) Assume that the three atoms decayed independently of one another, but according to the same parameter λ. What is the joint probability of seeing these particular three decay times, given λ? (Use the exponential PDF. I'd list it here, but I want you to remember it or look it up, for practice!)

(b) In part (a) you found the *likelihood function*. Notice that we use the word "likelihood" rather than "probability". Why is it not correct to call that expression a probability?

(c) We want to find the value of λ that maximizes the likelihood function. To do this, we can take the derivative (with respect to λ) and set it to 0, then solve for λ. But sometimes it's easier to do this to the log of the likelihood function. What's the (easier-to-work-with) log of your answer to (a)?

(d) Find the derivative of the log likelihood function in (c), set it to 0, and solve for the maximum. What value of λ do you get?

(e) Why can we be confident that this gave us the same answer that we would have gotten using the derivative of the (non-logged) likelihood function?

(f) How can we be sure that the value in (e) is a maximum rather than a minimum? (What else would we need to do to satisfy the math police, even though we will often skip it in CS109?)

(g) What is the expected value of the exponential distribution with this λ? Does that seem to fit the data?

Solutions to Problem 3.

(a) The exponential PDF is $f(X = x) = \lambda e^{-\lambda x}$. Let X_1, X_2, X_3 be the independent decay times for three atoms. We can say

$$f(X_1 = 40, X_2 = 70, X_3 = 85|\lambda) = (\lambda e^{-40\lambda})(\lambda e^{-70\lambda})(\lambda e^{-85\lambda}) = \lambda^3 e^{-(40+70+85)\lambda}$$

$$= \boxed{\lambda^3 e^{-195\lambda}}$$

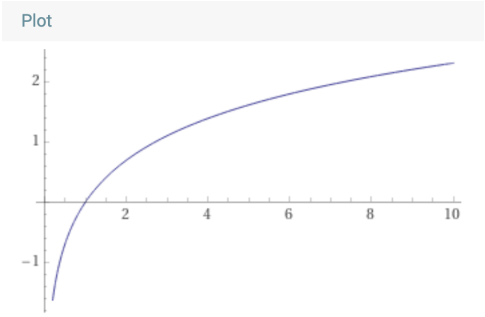
(b) The expression in part (a) is a joint probability density function. Unlike PMFs, these give probability densities, not probabilities.

(c) The log likelihood is

$$\log\left(\lambda^3 e^{-195\lambda}\right) = \log(\lambda^3) + \log(e^{-195\lambda}) = \boxed{3\log(\lambda) - 195\lambda}$$

(d) $\frac{d}{d\lambda}(3\log(\lambda) - 195\lambda) = \frac{3}{\lambda} - 195$. Setting this to 0: $\lambda = \frac{3}{195} = \boxed{\frac{1}{65}}$.

(e) This works because the log function is "everywhere increasing", i.e., for any two values x and y, $\log x > \log y$ if and only if $x > y$. This is even true in the weird part of the function's range, $(0, 1]$.



(However, we should note that if the likelihood equals 0, then the log likelihood is nonsensical, since 0 is not in the domain of log. But we are not likely to try to use MLE in a situation where the likelihood would be 0...)

(f) Technically, we should check that this is a minimum by confirming that the second derivative is "concave down" at $\lambda = \frac{1}{65}$: $\frac{d}{d\lambda}\frac{3}{\lambda} - 195 = -\frac{3}{\lambda^2}$, which is certainly negative because $\lambda = \frac{1}{65}$ is positive. Yay!

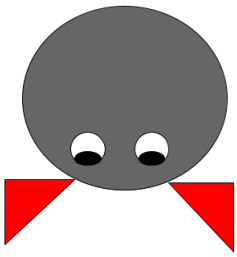
A mnemonic I just made up: $\log x$ is not convex! (But you have to remember the "log" / "not" part of the rhyme as well.)

(g) The expected value of an exponential is $\frac{1}{\lambda}$, so, $\boxed{65}$ here. This is the mean of our values of 40, 70, and 85, so it seems sensible. But it was not obvious, without doing MLE, that we should have taken the average.

IV. MAP and the slacker robot

Both MLE and MAP try to find the value of an unknown parameter, but MAP does so in a Bayesian way, by incorporating a prior. Let's practice by diving into another example, and this time we'll be trying to estimate *two* unknown parameters...

Problem 4: This robot is not trying very hard. Our friend created a bot to play Rock Paper Scissors.



beep boop I am a robot thrown together in 1 minute in Google Drawings. It's unclear how I would use my triangular appendages to play Rock Paper Scissors

It is not very sophisticated; we know that on each turn, independently of its choices on all other turns, it picks Rock with probability r_{true}, Paper with probability p_{true}, and Scissors with probability $1 - r_{\text{true}} - p_{\text{true}}$. We know that r_{true} and p_{true} are constants, but we do not know their values. This sounds like a job for MLE or MAP, and this time we'll use the latter.

We wish to model our uncertainty about these values r_{true} and p_{true} as a joint probability distribution / likelihood function $f(R = r, P = p)$.

(a) Suppose that our prior belief is that the joint distribution of r and p is uniform. That is, $f(R = r, P = p) = k$ for some constant k. Show that $k = 2$. (Hint: we want f to assign the same probability to every valid pair (r, p) with $r \geq 0, p \geq 0, r + p \leq 1$. What does this region look like? What would be the area under a region of the same shape floating k units above the xy-plane? If this part is confusing, don't worry about it, leave it as some constant K, and move on, since AFAICT, Chris is not emphasizing this type of multivariable calculation this quarter. It'll all still work out.)

For the remaining parts, suppose that after establishing our prior, we observed three rounds and saw the robot play Paper twice and Rock once.

(b) What is $P(\text{Paper twice, Rock once} \mid R = r, P = p)$? Your answer should be in terms of r and p.

Oh no, there's more on the next page...

(c) Using Bayes' Rule, write an expression for

$f(R = r, P = p \mid \text{Paper twice, Rock once})$.

Note that the two terms in the numerator are your answers from (b) and (a).

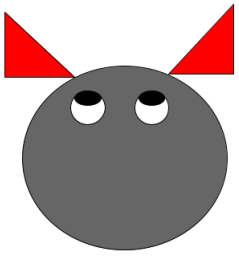
(d) Evaluate the denominator using an analogue of the Law of Total Probability, then simplify your answer to (c) in terms of r and p. (Hint: the expression should be a double integral over r and p, and the upper limit of the inner integral should not be 1... the result is a constant. If you don't want to do a double integral, feel free to skip this part and look up the answer, since – again – it does not seem to be crucial for this quarter's CS109.)

(e) Using your answer from (d), how much more likely is it that $p_{\text{true}} = \frac{2}{3}, r_{\text{true}} = \frac{1}{3}$, compared with $p_{\text{true}} = \frac{1}{3}, r_{\text{true}} = \frac{1}{3}$?

(f) Show that $p = \frac{2}{3}, r = \frac{1}{3}$ maximizes $f(p, r)$. You can assume without proof that the maximum occurs when $p + r = 1$. (Hint: reduce the expression to one variable, then use calculus to find a maximum.)

(g) Congratulations, you just found a new and super-useful distribution![8] It was mentioned in lecture on Monday... do you happen to remember its name?

(h) Notice that because we used a uniform prior, we got the same result in (f) as if we had done MLE. What prior would we have instead used if we were incorporating Laplace smoothing? Roughly how would this change the answer to (f)?

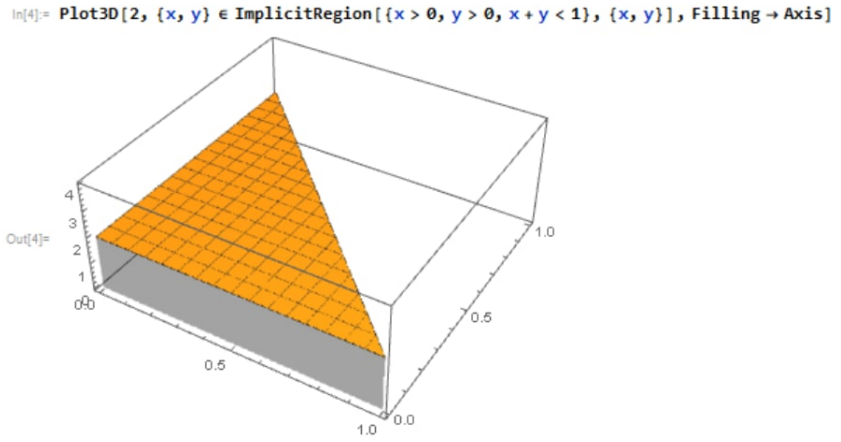


boop beep I have rotated 180 degrees to cheer you on while solving this difficult problem

[8] You will see a lot of it if you take CS238: Decision Making Under Uncertainty.

Solutions to Problem 4.

(a) The region corresponding to all possible (r, p) pairs is the triangle bounded by the r-axis, the p-axis, and the line $p = 1 - r$. The corresponding uniform $f(r, p)$ is an identically-shaped triangle floating at a height of k above our region. The integral of this needs to be 1 for it to be a probability distribution; that integral is the volume of a prism, given by $\frac{1}{2}(1)(1)(k) = \frac{k}{2}$. Setting this equal to 1, we get the desired $k = 2$.



Mmm, pie!

If you dislike pleasantly simple formulas for volumes of shapes but enjoy multivariable calculus, a correct double integral is

$$\int_0^1 \int_0^{1-r} k \; dpdr = \int_0^1 [kp]_0^{1-r} dr = \int_0^1 k(1-r)dr = k[r - \tfrac{r^2}{2}]_0^1 = \tfrac{k}{2}$$

and then we have $\frac{k}{2} = 1$ and $k = 2$ as before.

(b) $P(\text{Paper twice, Rock once} \mid R = r, P = p)$ has a multinomial distribution:

$\binom{3}{2,1,0} r^1 p^2 (1 - r - p)^0$. This simplifies to just $\boxed{3rp^2}$.

(c) $f(R = r, P = p \mid \text{Paper twice, Rock once}) =$

$$\boxed{\frac{P(\text{Paper twice, Rock once} \mid R = r, P = p) f(R = r, P = p)}{P(\text{ Paper twice, Rock once })}}$$

(d) The denominator looks like the numerator, but integrated over all possible pairs of (r, p). The limits are the same as in the double integral from (a):

$$\int_0^1 \int_0^{1-r} (3rp^2)(2) \; dpdr = 2 \int_0^1 [rp^3]_0^{1-r} dr = 2 \int_0^1 r(1-r)^3 dr$$

$= 2[\frac{r}{2} - r^2 + \frac{3r^3}{4} - \frac{r^4}{5}]_0^1 = 2(\frac{1}{2} - 1 + \frac{3}{4} - \frac{1}{5}) = \frac{1}{10}$.

Therefore, using this and the information from (a) and (b),

$f(R = r, P = p \mid \text{Paper twice, Rock once}) = \frac{(3rp^2)(2)}{\frac{1}{10}} = \boxed{60rp^2}$, or $6Krp^2$ if you left the constant.

(e) $f(R = \frac{1}{3}, P = \frac{2}{3}) = 60(\frac{1}{3})(\frac{2}{3})^2$.

$f(R = \frac{1}{3}, P = \frac{1}{3}) = 60(\frac{1}{3})(\frac{1}{3})^2$.

Taking the ratio of these (which makes the constant K go away if you are using K), we see that the former is $\boxed{4}$ times more likely.

(f) If we assume that $p+r = 1$ (i.e. that we can't do better by assigning probability to Scissors, which seems reasonable since we have not observed a Scissors play yet), then $p = 1 - r$ and $60pr^2 = 60(1 - r)r^2 = 60r^2 - 60r^3$. The first derivative of this with respect to r is $120r - 180r^2$. Setting this to 0, we find that maxima/minima occur at $r = 0$ and $r = \frac{2}{3}$. (Same answer if using K.)

We could further find the second derivative with respect to r: $120 - 360r$. Then we have a positive value for $r = 0$ and a negative value for $r = \frac{2}{3}$, so the latter is indeed a maximum.

(g) We have performed MAP to estimate multinomial parameters. Per slide 75 of lecture 22, the conjugate distribution is the Dirichlet, which is a multidimensional analogue of the beta distribution. Our uniform prior distribution is $Dir(1, 1, 1)$ (just as $Beta(1, 1)$ is uniform), and we ended up with the posterior $Dir(3, 2, 1)$.

(h) If we had used Laplace smoothing instead, we would have started with $Dir(2, 2, 2)$. We would have ended up with $Dir(4, 3, 2)$; we didn't spend much time going into how to get an answer out of this (you can do the same math as above to find forms for $Dir(2, 2, 2)$ $(120rp(1 - r - p))$ and $Dir(4, 3, 2)$ $(3360r^2p^3(1 - r - p))$), but the best values for p and r end up being $\frac{1}{2}$ and $\frac{1}{3}$. Note the $p : r : 1 - p - r$ ratio of 3:2:1, which is the same as taking 4, 3, 2 and subtracting one from each; similarly, our $Dir(3, 2, 1)$ gave us a ratio of 2:1:0.