CS109A Week 7 Notes

Ian Tullis

May 10, 2022

This week we'll practice for the upcoming Quiz 2!

I. Bayesian Networks

We have been cooped up inside all day coding, and we are about to leave Stanford to head into San Francisco on 101. We check the traffic on Google Maps. Oh no, 101 is bright red! Why is the traffic so much worse than usual? We come up with a couple of theories:

- There might be a Warriors game tonight.
- It might be raining, and as we all know, in general, people from the Bay Area can't drive in any kind of weather.[1]

We also know that on days when there is a Warriors game, people are more likely to wear Warriors jerseys.

Of course, traffic is sometimes bad for other reasons, and it might just happen that more people wear Warriors jerseys on some days. So we can't rule out the effects of chance.

Klay Thompson famously ended up as part of a random street interview about scaffolding in New York City. I wonder what his thoughts are on Bayesian networks?

[1] I include myself in this statement!

Problem 1. Let:

- W be the event that there is a Warriors game,
- R be the event that it is raining,
- T be the event that there is unusually bad traffic, and
- J be the event that more people than usual are wearing Warriors jerseys.

(a) Translate the information given on this page into a Bayesian network with circles for W, R, T, and J, and arrows between them as appropriate. (Assume, rather unrealistically, that there are no other specific factors involved besides chance. Note that this model assumes that R and W are independent, which makes sense for an indoor sport!)

(b) In each of the following situations, assume that the person operates under the belief system above, but unless otherwise stated, they start off knowing nothing – i.e., they do not know whether there is a Warriors game tonight, whether it is raining, whether there is unusually bad traffic, or whether more people than usual are wearing Warriors jerseys. or whether there is a Warriors game tonight. For this part, try to think intuitively and not in terms of specific probabilities. Remember that events T and J could each happen, with at least *some* probability, by chance; that is, it is possible that more people than usual are wearing Warriors jerseys, even though there is no Warriors game.

(i) Klay already knows that traffic is unusually bad. He checks the weather and sees that it is not raining. Does this new information change his belief about whether there is a Warriors game tonight?

(ii) Draymond sees that more people than usual are wearing Warriors jerseys. Does this change his belief about whether traffic is unusually bad? Does this change his belief about whether it is raining?

(iii) Steph already knows that there is no Warriors game tonight. Then he sees that more people than usual are wearing Warriors jerseys. Does this change his belief about whether traffic is unusually bad?

(iv) Ayesha already knows there is a Warriors game tonight. Then she sees that traffic is unusually bad. Does this change her belief about whether it is raining?

(c) Now suppose that the actual underlying model is:

- $P(R) = 0.2$. (Hey, we can dream it's that high.)
- $P(W) = 0.1$.
- $P(T|R \cap W) = 0.9$; $P(T|R^c \cap W) = 0.8$; $P(T|R \cap W^c) = 0.7$; $P(T|R^c \cap W^c) = 0.1$.
- $P(J|W) = 0.6$; $P(J|W^c) = 0.1$.

(i) What are $P(R \cap W), P(R^c \cap W), P(R \cap W^c)$, and $P(R^c \cap W^c)$?

(ii) What are $P(T), P(T|W)$, and $P(W|T)$? Does your answer for $P(W|T)$ agree with your reasoning in (b)(i)?

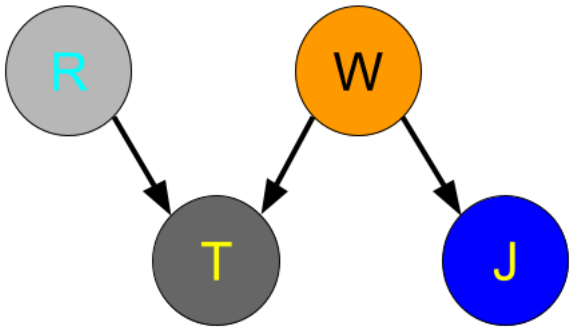
(iii) What is $P(W|T \cap R)$? Comparing this to $P(W|T)$, does this fit your answer to (b)(i)?

(iv) Time permitting, try to check some of the other statements as well. It may be easiest to make a table with the probabilities of all 16 of the possible scenarios, although the whole point of these Bayesian networks is to avoid having to do this explicitly. Make sure you know how to do this yourself, but the results are provided here. (Note: you should do the previous parts without directly using this table.)

W	R	T	J	Prob.
0	0	0	0	0.5832
0	0	0	1	0.0648
0	0	1	0	0.0648
0	0	1	1	0.0072
0	1	0	0	0.0486
0	1	0	1	0.0054
0	1	1	0	0.1124
0	1	1	1	0.0136
1	0	0	0	0.0064
1	0	0	1	0.0096
1	0	1	0	0.0256
1	0	1	1	0.0384
1	1	0	0	0.0008
1	1	0	1	0.0012
1	1	1	0	0.0072
1	1	1	1	0.0108

Solutions to Problem 1.

(a) The Bayesian network looks like this:



(b) (i) Yes. If rain and the Warriors game are the two possible explanations (other than chance) for the bad traffic, and rain is ruled out, then it is much more likely that a Warriors game is to blame.

(ii) Yes; No. Given that more people than usual are wearing jerseys, Draymond's belief that there is a Warriors game is strengthened. Because this is known to result in bad traffic, his belief that there is bad traffic is also strengthened.

However, intuitively, this all has nothing to do with whether it is raining. Even though rain might make already bad traffic worse, Draymond knows that whether or not it rains is independent of whether or not there is a Warriors game, so feeling more confident that there is a Warriors game shouldn't tell him anything about rain.

We might fool ourselves with an argument like "well, we think traffic is bad, and bad traffic is associated with rain", but our new belief about the traffic is already fully explained by our new belief about the Warriors game. We really haven't learned anything about the rain.

(iii) No. First of all, since Steph knows that there is no Warriors game, and that is the only factor (besides chance) that influences whether more people than usual are wearing Warriors jerseys, he can safely conclude that the latter is due to chance.

Now, usually, the only value in knowing whether more people are wearing Warriors jerseys is that it makes Steph more likely to believe there is a game, but given that he already knows there isn't one, he learns nothing.

(iv) $\boxed{\text{Yes.}}$ This one is tricky! At first it might seem that Ayesha's knowledge of the Warriors game fully explains the bad traffic. But that would only be true if a Warriors game *surely* resulted in bad traffic. Otherwise, there is some chance that the bad traffic is *not* due to the Warriors game, and in that case, as usual, bad traffic increases our suspicion that rain may be involved.

There is a set of rules, involving something called "d-separation", to make this type of analysis easier. This is not in scope for CS109, but you might enjoy learning more: `https://bayes.cs.ucla.edu/BOOK-2K/d-sep.html`. I personally always have trouble remembering these rules, and I prefer to think through example situations.

(c) (i) $P(R \cap W) = P(R)P(W) = 0.2 \cdot 0.1 = 0.02$ because the events are independent (and we have no other information). Similarly, $P(R^c \cap W) = P(R^c)P(W) = (1 - 0.2)(0.1) = 0.08$. $P(R \cap W^c) = 0.18$, and $P(R^c \cap W^c) = 0.72$.

(ii) By the Law of Total Probability,

$$\begin{aligned} P(T) = &P(T|W \cap R)P(W \cap R) + P(T|W \cap R^c)P(W \cap R^c) \\ &+P(T|W^c \cap R)P(W^c \cap R) + P(T|W^c \cap R^c)P(W^c \cap R^c) \end{aligned}$$

That is, we're splitting up the traffic across these four mutually exclusive and exhaustive cases. Using our results from (a), we get

$$P(T) = 0.9 \cdot 0.02 + 0.8 \cdot 0.08 + 0.7 \cdot 0.18 + 0.1 \cdot 0.72 = \boxed{0.28}$$

We can calculate $P(T|W)$ in a similar way, by splitting across the mutually exclusive and exhaustive cases of rain and no rain:

$$P(T|W) = P(T|W \cap R)P(R) + P(T|W \cap R^c)P(R^c)$$

This is $0.9 \cdot 0.2 + 0.8 \cdot 0.8 = \boxed{0.82}$.

Finally, by Bayes' Rule,

$$P(W|T) = \frac{P(T|W)P(W)}{P(T)} = \frac{0.82 \cdot 0.1}{0.28} \approx \boxed{0.2929}$$

This is larger than $P(W) = 0.1$, the estimate that we would make in the absence of knowledge of bad traffic. That makes sense! A Warriors game is one possible explanation for the bad traffic.

(iii) $P(W|T \cap R) = \frac{P(W \cap T \cap R)}{P(T \cap R)}$.

Let's rewrite the numerator as $P(R)P(W|R)P(T|R \cap W)$. We know that W is independent of R (basketball games go on rain or shine!), so $P(W|R) = P(W)$. Therefore we have $0.2 \cdot 0.1 \cdot 0.9 = 0.018$.

What about the denominator? We can rewrite it as $P(R)P(T|R)$. Now, using the Law of Total Probability as in the previous part, $P(T|R) = P(T|R \cap W)P(W) + P(T|R \cap W^c)P(W^c) = 0.9 \cdot 0.1 + 0.7 \cdot 0.9 = 0.72$. So $P(TR) = 0.2 \cdot 0.72 = 0.144$.

Therefore $P(W|T \cap R) = \frac{0.018}{0.144} = \boxed{0.125}$.

Comparing this with $P(W|T)$, it is quite a bit lower. That is, the rain provides a potential explanation for the traffic, which is consistent with what we said in (b)(i). However, 0.125 is still larger than $P(W) = 0.1$. That is, we still have some reason to suspect a Warriors game; the rain does not completely explain away the traffic.

II. Important engagement with real-world issues

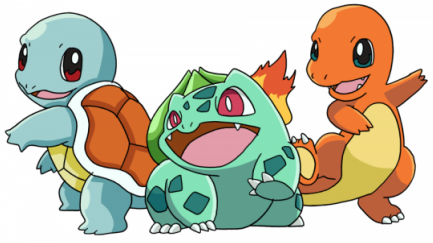
Problem 2.

(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[2]. 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?



[2] 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!

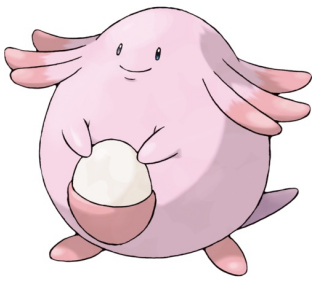
Therefore the answer is $\boxed{\sum_{i=1}^{905} \frac{905}{i}}$, which means we need to talk to ≈ 6684 people (in expectation) to catch 'em all. Notice that the summation looks like it goes in the opposite order of our argument above, but the order of the sum does not matter.

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.[3] 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?

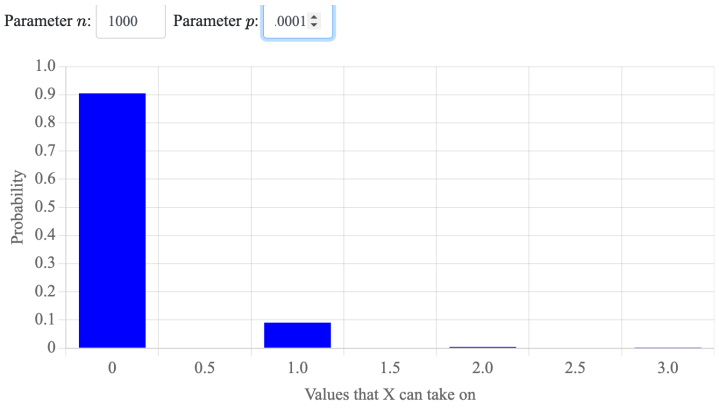
No, as it turns out! The real answer, which you can get as

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

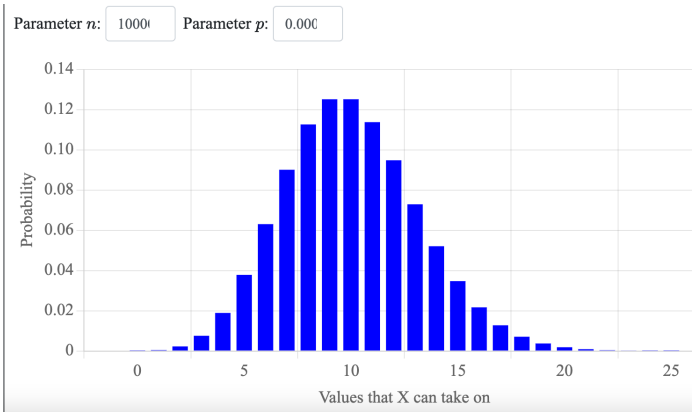
is ≈ 0.0952. So the Poisson approximation is spot on and the normal approximation is quite a ways off!

[3] See problem 5 in the Week 5 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:



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:



(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. Additional practice problems

Solutions appear at the end of this section.

1 Jane Street

You are trying to walk from the Oval to your next class in the Main Quad, but you still have to cross Jane Stanford Way, which has a seemingly endless flow of cyclists!

On average, one cyclist goes by (i.e. passes along right in front of you) every 2 seconds. Assume that the cyclists behave independently; also, their directions of travel don't matter, and they go by instantaneously.

Parts (a)-(c) are warm-up, and parts (d)-(e) are about crossing the street.

(a) (**) What is the probability that the next cyclist you see will appear between 1 and 4 seconds from now?

(b) (**) In part (a), you might have been uncomfortable that the problem didn't say when the most recent cyclist went by. Why does this not matter?

(c) (**) What is the probability that you will see exactly 1 cyclist in the interval between 1 and 4 seconds from now?

(d) (**) You know it takes you 6 seconds to cross the street, and you can't stop in the middle, so you need to wait for an interval of at least 6 seconds between successive cyclists before you start to cross. (You can see the bikes coming, so you know when it is safe to go.) What is the probability that you can start crossing *right now*?

(e) (****) What is the expected amount of time you will need to wait before you can start to cross?

Here is an integral that may prove useful as you solve this part, and that you would not be asked to do on an exam: for a constant a, $\int axe^{-ax}dx = \frac{e^{-ax}(ax+1)}{a}$.

2 Central "Express" way, except when it isn't!

As usual, I am running late, so I leave home 25 minutes before my next class starts. Suppose that there are 16 traffic lights on my drive from home to campus. Also suppose that if I hit only green lights, it would take me 20 minutes to reach campus. However:

- Each light has (independently) a probability of 0.3 of being red.
- For each light (independently), the amount of additional time I have to wait at that light is normally distributed with mean 60 seconds and standard deviation 15 seconds.

(a) (*) What is the expected amount of time it will take me to reach campus?

(b) (***) What is the **exact** probability that I make it to campus in time to teach class? (Assume, somewhat unrealistically, that I teleport instantaneously from my parking spot on campus to the classroom.) Your answer can involve a summation and Φ.

You can use the fact (which I think is out of scope for the exam) that the sum of n independent $\mathcal{N}(\mu, \sigma^2)$ is itself a normal distribution $\mathcal{N}(n\mu, n\sigma^2)$. So if I hit 3 lights, for instance, my total waiting time is distributed as $\mathcal{N}(3 \cdot 60, 3 \cdot 15^2)$. However, this question *does* test something that could be on the exam! What distribution can you use for the number of lights I hit?

(c) (***) Suppose that the traffic lights are not independent, in some way that guarantees that I will never hit two red lights in a row, but that the probability of hitting each one is still 0.3 overall. (As one example, suppose that with probability 0.5, each odd-numbered light is red with probability 0.6 and even-numbered lights are never red, and with probability 0.5, each even-numbered light is red with probability 0.6 and odd-numbered lights are never red.) Would the answer to (a) to be larger, smaller, or the same, or would it depend on the specific way in which these conditions hold?

3 Quad workout

Suppose that we try to define a *quad distribution*[4] as a continuous distribution supported on $[0, 1]$. It has real nonnegative parameters a and b, and the PDF

$$f(x) = \begin{cases} k(ax+b) & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

k is a positive real constant (in terms of a and b) that makes $f(x)$ a valid PDF; you will calculate it in part (b).

(a) (**) What restrictions are there on a and/or b such that (for an appropriate choice of positive real k) $f(x)$ is a valid probability distribution?

(b) (**) What is k, expressed in terms of (validly chosen) a and b, such that $f(x)$ is a valid probability distribution? What is the form of $f(x)$ with k replaced by what you just found?

(c) (**) What is the mean of a random variable $X \sim Quad(a, b)$ (i.e. distributed according to a quad distribution with parameters a and b), expressed in terms of a and b?

(d) (**) What is the CDF $F(x)$ of $f(x)$, expressed in terms of x, a, and b?

(e) (**) Set up – but do not evaluate – an expression for the median of a random variable with the distribution $Quad(a, b)$.

(f) (**) Approximate the probability that a sum of 162 independent draws from the same $Quad(a = 1, b = 1)$ distribution is greater than 81.

[4] Not an actual name, as far as I know!

4 The obligatory CS109 medical problems

(a) (*) Assume that people who are suffering from a certain illness (which is not COVID because goodness knows we've had to think about that enough) have a temperature distributed as $\mathcal{N}(\mu = 101, \sigma = 1)$. However, people who don't have the illness have a temperature distributed as $\mathcal{N}(\mu = 98, \sigma = 1)$.

Suppose that we observe a person with a temperature of 101 or higher. Why shouldn't we necessarily conclude that the person has the illness, even though 101 clearly has a larger probability density in the "illness" distribution than in the "non-illness" distribution? (I.e., what other piece of information do we need before making our decision? The intended answer is *not* that there are other diseases in the world too, although that is a reasonable point... but here, suppose that we have somehow narrowed it down to the person either having that illness or having no illness.)

(b) (**) In Homework 3, Problem 3, when testing for measles (in a population with a 5% rate of measles), we tested pooled samples of 6 people at a time, then tested them all individually if there was a positive pooled test. Suppose that we replaced 6 with N. What is the smallest value of N (greater than 1, of course) for which this strategy becomes *worse* (in terms of expected total number of tests) than just testing everyone individually at the outset? (Use e.g. Wolfram Alpha for calculations.)

5 Not quite six sigma

(a) (*) In a normally distributed population, what fraction of the population do you equal or exceed if you are 1 standard deviation (σ) above the mean? 2σ? deviations? 3σ? (This wouldn't be a midterm question, because it can't be hand-solved, but: do you know how to find this info? I do think it's useful (for life) to memorize these numbers...)

(b) (*) Tests like the SAT have often been scaled based on an assumed normal distribution with a mean of 500 and a standard deviation of 100. However, scores are reported rounded to the nearest 10-point increment. What fraction of test-takers would you expect to earn a score of 670? (This isn't a trick question where the answer is 0... what range of the distribution does this actually represent?)[5] You can leave your answer in terms of Φ expressions.

[5] This question should not be taken as an endorsement of standardized testing, which has its problems, to say the least. Nor do one's test scores determine one's worth in any way. That said, you should still try to do your best on the midterm!

V. Solutions To Additional Practice Problems

(1) (a) Because these events (the cyclists passing by) are independent, it is reasonable to use an exponential distribution to model the time until the next cyclist. If we see one cyclist every 2 seconds on average, then the exponential parameter λ is $\frac{1}{2}$. (Recall that the mean of the exponential distribution is $\frac{1}{\lambda}$, just like how the mean of the exponential distribution's discrete analog, the geometric distribution, is $\frac{1}{p}$.)

Then the probability that the next cyclist appears between 1 and 4 seconds from now is given by integrating the exponential PDF:

$$\int_{x=1}^{4} \lambda e^{-\lambda x} = \int_{x=1}^{4} \frac{1}{2} e^{-\frac{x}{2}} = [-e^{-\frac{x}{2}}]_1^4 = \boxed{e^{-\frac{1}{2}} - e^{-2}}$$

.

(b) An exponential distribution (or Poisson process) is "memoryless". Remember Beepworld from the Week 5 notes... the expected time to the next beep was *always* 10 seconds, *no matter when we started listening*. In our case here, it makes no difference whether a cyclist just went by at the start of the problem, or no cyclists have gone by for a whole minute!

(c) In a situation where the exponential distribution models the time to the next event, the Poisson distribution models the number of events per time interval. Every time interval of a given size **behaves exactly the same way**, so asking about the interval from 1 to 4 seconds from now is the same as asking about any 3-second interval. (It doesn't matter whether other cyclists might go by between 0 and 1 seconds!)

Since there is 1 cyclist on average every 2 seconds, there are $\frac{3}{2}$ on average every 3 seconds, so $\lambda = \frac{3}{2}$ here. (Remember – we have to choose λ to fit the time interval the problem is asking about! It is the same rate as we used in part (a), but here we're looking at events per time interval of 3 seconds, and *not* using the second as the basic unit of time, as we did in part (a).)

Therefore the probability of seeing exactly one cyclist in that 3-second interval seconds is $P(X = 1) = \frac{e^{-\lambda}\lambda^1}{1!} = \boxed{\frac{3e^{-3}}{2}}$.

(d) To find the probability that we will be able to cross the street right now, we can use either the exponential distribution or the Poisson distribution.

- Exponential: We can cross now only if the time to the next cyclist is 6 seconds or more. Using the same $\lambda = \frac{1}{2}$ from part

(a), the probability of this is $1 - \int_{x=0}^{6} e^{-\frac{x}{2}} = 1 - [-e^{-\frac{x}{2}}]_0^6 = 1 - (-e^{-3} + 1) = \boxed{e^{-3}}$.

- Poisson: We can cross now only if there are 0 cyclists in the next 6-second interval. For this interval, $\lambda = 3$, since we expect 3 cyclists in 6 seconds. So the answer is $P(X = 0) = \frac{e^{-3}3^0}{0!} = \boxed{e^{-3}}$.

(e) One way to approach this challenging problem is to observe that we are sampling from the exponential distribution (intervals between successive cyclists) until we happen to draw a sample of 6 seconds or larger. The probability of any one draw meeting these criteria is e^{-3}, so – using the expectation of a geometric distribution – we will need to wait for an expected $\frac{1}{e^{-3}} = e^3$ draws before we get our first time interval long enough for us to cross in. That is, in expectation, we will see $e^3 - 1$ "bad" intervals (less than 6 seconds) before our one "good" interval (6 seconds or more). Once the good interval hits, we can start crossing immediately, without waiting for that interval to elapse. So we only need to consider the total time of the "bad" intervals.

Let X be a random variable for the number of bad intervals we will encounter, and let Y be a random variable for the duration of an arbitrary bad interval. Since the durations of these bad intervals are independent (of each other and of the total number of bad intervals), the expected total wait time is $P(X = 0)\cdot 0 + P(X = 1)\cdot\mathbb{E}[Y] + P(X = 2)\cdot 2\mathbb{E}[Y] + ...$, i.e., $\sum_{x=0}^{\infty} P(X = x)x\mathbb{E}[Y]$. Since $\mathbb{E}[Y]$ is a constant, we pull it out: $\mathbb{E}[Y]\sum_{x=0}^{\infty} P(X = x)x$. Now the summation is just an expression for $\mathbb{E}[X]$, so we have $\mathbb{E}[Y]\cdot\mathbb{E}[X]$.

(We had to be careful here because – unlike with linearity of expectation – the expectation of a product of random variables is only equal to the product of their expectations if those two random variables are independent.)

The expected wait time for each cyclist, $\mathbb{E}[Y]$, should be 2 seconds, right? So the answer is $2(e^3 - 1)$, right?

Unfortunately, this is wrong because the expectation of a bad interval is not 2. We have truncated the exponential distribution by conditioning on the bad intervals being less than 6 seconds. So we need to find the average of the exponential distribution between 0 and 6...

Except if we want to find the mean over just that range, now all the probabilities need to be readjusted to make the distribution between 0 and 6 behave like a PDF (i.e. have the integral be 1). We already know (from part (d)) that the probability of an interval being bad is

$1 - e^{-3}$, so we can just correct the probabilities by a factor of $\frac{1}{1-e^{-3}}$. So we get $\int_0^6 \frac{1}{2}x(e^{-\frac{x}{2}})(\frac{1}{1-e^{-3}})dx = \frac{1}{1-e^{-3}}[-(x+2)e^{-\frac{x}{2}}]_0^6 = \frac{2-8e^{-3}}{1-e^{-3}}$.

Therefore the true answer is $(e^3 - 1)(\frac{2-8e^{-3}}{1-e^{-3}})$, which turns out to simplify to $= \boxed{2e^3 - 8}$, which is about 32.17.

This situation is an example of a *Poisson process*, if you want to learn more about that!

This was another of those problems that was worth checking with code. The result was 32.17.

```
from scipy import stats

def trial():
    t = 0
    while True:
        wait = stats.expon.rvs(scale=2)
        if wait >= 6:
            return t
        else:
            t += wait

NUM_TRIALS = 1000000
tot = 0
for i in range(NUM_TRIALS):
    tot += trial()

print(tot / NUM_TRIALS)
```

(2) (a) Each traffic light has an 0.3 probability of adding an expected 60 seconds of delay, so it adds 18 seconds in expectation. By linearity of expectation, the sum for all 16 lights is 288 seconds = 4 minutes, 48 seconds. So I will on average take $\boxed{24:48}$ to reach campus, which is cutting it close!

(b) Suppose that I hit n lights. Then the total delay is the sum of n independent and identically distributed $\mathcal{N}(\mu = 60, \sigma^2 = 225)$ random variables. Per the rules for adding independent normal distributions, as given in the problem, the total delay is $\mathcal{N}(\mu = 60n, \sigma^2 = 225n)$. Then the probability that I make it to campus on time (i.e. that the delays are 300 seconds or less) is $\Phi(\frac{300-60n}{\sqrt{15n}})$. Since this delay time is continuous, there is no need for a continuity correction.

However, the number of lights I hit is binomially distributed: $P(N = n) = \binom{16}{n}(0.3)^n(0.7)^{16-n}$. The distribution of the total delay depends on this value, so it is:

$$\boxed{0.7^{16} + \sum_{n=1}^{16} \binom{16}{n} (0.3)^n (0.7)^{16-n} \Phi(\frac{300 - 60n}{\sqrt{15n}})}.$$

The separate 0.7^{16} term is there because we can't have the summation start at 0, or we would have a division by 0 in the Φ term. In that case, the Φ term wouldn't even make sense, because there would be no distributions involved! If I hit no red lights, then I always make it on time.

Observe that earlier in the summation (for small values of n), the Φ part will be very close to 1 (since there are so few lights that they can't possibly cause enough of a delay), and later in the summation, the Φ part will be very close to 0 (since there are so many lights that I can't possibly avoid being too delayed) *and* the binomial part will be close to 0 (since it is unlikely to see so many red lights, given that they are individually uncommon). If we do the math, it turns out that within the summation, only the $n = 1$ through $n = 4$ terms really matter.

(c) Linearity of expectation holds regardless of the independence or non-independence of the individual traffic lights; notice that the argument in (a) never invoked independence. So the answer is exactly the same regardless of the exact way the non-independence manifests.

(3) (a) We need a PDF to always be nonnegative. Since k is positive, we must also have $ax + b$ positive everywhere over $[0, 1]$. Now, for $x = 0$ this becomes b and for $x = 1$ this becomes $a + b$, so we need $b \geq 0$ and $a > -b$.

(b) We need $f(x)$ to integrate to 1 over its supported range of $[0, 1]$, so

$$\int_{x=0}^{1} k(ax + b)dx$$

Then

$$k[\frac{ax^2}{2} + bx]_0^1 = k(\frac{a}{2} + b) = 1$$

and

$$k = \frac{1}{\frac{a}{2} + b} = \boxed{\frac{2}{a + 2b}}$$

.

So

$$f(x) = \frac{2ax + 2b}{a + 2b}$$

.

(c) The mean of a PDF is given by $\int xf(x)dx$, so here we have

$$\int_{x=0}^{1} \frac{(2ax+2b)x}{a+2b} = \frac{1}{a+2b}\int_{x=0}^{1}(2ax^2+2bx)dx = \frac{1}{a+2b}[\frac{2ax^3}{3}+bx^2]_0^1$$

$$= \frac{1}{a+2b}(\frac{2a}{3}+b) = \boxed{\frac{2a+3b}{3(a+2b)}}$$

(d) The CDF is found by integrating $f(x)$ up to a certain point y:

$$\int_0^y \frac{2ax+2b}{a+2b}dx = \frac{1}{a+2b}[ax^2+2bx]_0^x$$

Expressing this as $F(x)$, this is

$$\boxed{F(x) = \frac{ax^2+2bx}{a+2b}}$$

(e) We can find the median of a PDF by setting its CDF to $\frac{1}{2}$ and solving for x:

$$\frac{ax^2+2bx}{a+2b} = \frac{1}{2}$$

This can be evaluated using the quadratic formula, but the answer is gross.

(f) We should use the Central Limit Theorem here, but in order to do so, we need the variance of $Quad(a,b)$. Let's find it as $\mathbf{E}[X^2]-(\mathbf{E}[X])^2$. We have $\mathbf{E}[X] = \frac{2a+3b}{3(a+2b)}$ from part (c); using $a=1, b=1$ in this part of the problem, this is $\frac{5}{9}$. But we still need $\mathbf{E}[X^2]$; we can find this as in part (c), but with x^2 in place of x:

$$\int_{x=0}^{1} \frac{(2ax+2b)x^2}{a+2b} = \frac{1}{a+2b}\int_{x=0}^{1}(2ax^3+2bx^2)dx = \frac{1}{a+2b}[\frac{ax^4}{2}+\frac{bx^3}{3}]_0^1$$

$$= \frac{1}{a+2b}(\frac{ax^4}{2}+\frac{2bx^3}{3}) = \frac{1}{3}(\frac{1}{2}+\frac{2}{3}) = \frac{7}{18}$$

Then the variance is this minus the square of the mean:

$$\frac{7}{18}-(\frac{5}{9})^2 = \frac{7}{18}-\frac{25}{81} = \frac{63-50}{162} = \frac{13}{162}$$

Let Y be the sum of 162 independent draws from $Quad(1,1)$. Then, per the Central Limit Theorem, the mean of Y is $162(\frac{5}{9}) = 90$, and the variance is $162(\frac{13}{162}) = 13$. Then, using the standard normal CDF, the probability that Y is less than or equal to 81 is $\Phi(\frac{81-90}{\sqrt{13}})$, and so the probability that it is greater than 81 is $\boxed{1-\Phi(\frac{-9}{\sqrt{13}})}$.

(4) (a) The issue is that we do not know how common the illness is overall. Suppose that the illness is extremely rare, occurring in, e.g., 0.1% of the population. But we would expect about 0.13% of the non-illness population to have this temperature or higher! (This comes from evaluating the normal CDF with $\mu = 98$, $\sigma = 1$, $x = 101$, and then subtracting the result from 1 because we want the area to the right of that point, under the long tail of the curve.) So in this case, the fraction of the population with the illness is pretty similar to the fraction of the population that doesn't have the illness but just happens to have that high of a temperature. Therefore, it's hard to say which of those two groups the person is in!

The takeaway point is that we need to know **the overall frequency of the illness** to make a sensible prediction. Notice that the section problem had to give you this piece of information (it's 20% in that case). A similar situation arose in that problem, where even though 100 "looks" much more like a temperature from the group with the flu, the proportion of people with the flu also factors into the calculation, and so the answer ends up actually being not far off from 50%.

(b) On the homework, we found that the expected number of tests was $1(0.95^6)+(1+6)(1-0.95^6)$. Here, we replace 6 with the more general N: $1(0.95^N)+(1+N)(1-0.95^N)$. The pooled-test strategy becomes worse once that quantity exceeds N (which is the number of tests we would need if we skipped the pooled test and just tested everyone up front). So, to get the threshold N above which the pooled-test strategy is worse, we solve

$$1(0.95^N) + (1 + N)(1 - 0.95^N) = N$$

$$0.95^N + 1 - 0.95^N + N - N \cdot 0.95^N = N$$

$$1 = N \cdot 0.95^N$$

Using Wolfram Alpha, we find that $N \approx 1.05$ or $N \approx 87.08$. Both of these are thresholds, but only one is the kind we want (where the pooled test goes from useful to not useful).

Let's check for $N = 87$: $1(0.95^{87})+(1+87)(1-0.95^{87}) \approx 86.997 < 87$ - still good!

And for $N = 88$: $1(0.95^{88}) + (1 + 88)(1 - 0.95^{88}) \approx 88.036 > 88$ - no longer good!

So the answer is $\boxed{88}$.

(5) (a) We can get these values from the course reader's Gaussian CDF calculator: they are $\Phi(1), \Phi(2), \Phi(3)$, which are $\boxed{0.8413, 0.9772, 0.9987}$.

(b) 670 would correspond to anything between 665 and 675; anything just lower than 665 would be rounded to 660, and anything just higher than 675 would be rounded to 680. So we want the integral from 665 to 675 of a normal distribution with mean 500 and standard deviation 100. As usual, we don't actually do this integral, but instead frame it as the difference between two evaluations of the CDF: $\boxed{\Phi(\frac{675-500}{100}) - \Phi(\frac{665-500}{100})}$, which is $\approx 0.9599 - 0.9505 \approx 0.0094$.