Additional final practice problems for CS109

Disclaimer: You should prioritize official CS109 review materials (past final and midterm exams, homeworks, sections) first! These problems are not guaranteed to match actual CS109 exam problems in style, scope, or difficulty (mine probably skew harder, and some of them involve calculations that would be too onerous by hand). Hence "final practice problems" rather than "practice final problems".

These are unofficial and were written for Ian's CS109A class for Winter 2022. Anyone is welcome to share them around, but please direct any questions about the problems to `itullis@stanford.edu`.

Star ratings: *, ** = bread and butter, *** = extra thought required, **** = stretch

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 When expectations don't meet expectations

If we roll a six-sided die six times, we would *expect* (in the sense of "expectation") to see every number once. But remember that expectations sometimes don't match our expectations! Should we actually be surprised if we don't see every number once? If we see the same number come up twice, should we be suspicious?

In the following problem, we will use 4-sided dice to make the math easier.[1] Suppose we roll a fair 4-sided die four times. Let C_1, C_2, C_3, C_4 be the counts of ones, twos, threes, and fours that we see (so $C_1 + C_2 + C_3 + C_4 = 0$). Let C_{max} be $\max(C_1, C_2, C_3, C_4)$, i.e., the largest number of instances of any one roll that we see. If we roll two 3s and two 4s, for example, then $C_{max} = 2$. Or if we roll one 1 and three 2s, then $C_{max} = 3$.

(***) What is the complete probability distribution of C_{max}? (Hint: Find the values one at a time, in whatever order makes it easiest.)

3 Let's pretend that industry cares about grades

Suppose that a company is building a model that tries to predict success in industry (somewhat unrealistically using a binary variable S, where $S = 1$ means success) based on grades in n Stanford CS courses $X_1, X_2, ..., X_n$ (using one multinomial variable per course, ignoring +/-, so A = 4, B = 3, C = 2, D = 1, NP = 0). The company uses complete data (grades + success) from 1000 students.

(a) (**) Suppose that the model is like Naive Bayes but *without* the independence assumption, and estimates each individual parameter of the form $P(X_1 = x_1, ..., X_n = n|S = y)$ or $P(S = y)$ directly from the data. How many of these individual parameters would need to be estimated? Give your answer in terms of n.

(b) (*) What is the major problem with the above approach, apart from having to estimate so many parameters? (Hint: Unless n is very small, what will most of those estimated probabilities end up being?)

(c) (**) Now suppose that the model *does* use the Naive Bayes assumption. How many individual parameters of the form $P(X_i = x_i|S = y)$ or $P(S = y)$ would need to be estimated? Give your answer in terms of n.

(d) (*) Give one "real-world" reason why the Naive Bayes assumption might be unrealistic in this particular scenario.

(e) (**) Show how you would estimate the parameter $P(X_1 = 3|Y = 1)$ from the data, *including Laplace smoothing.*

[1] Believe me, this problem is *hot garbage* to solve with 6-sided dice, and I say this as a combinatorics fan...

Now, for the remaining parts, instead assume that the company uses a logistic regression model. Before training, an extra "intercept" feature is added; it has the value 1 for every student.

(f) (**) What undesirable consequences might arise if the model did not include this intercept feature? (How would its absence limit the expressiveness of the model?)

(g) (**) Why would such an intercept feature be unnecessary/useless in a Naive Bayes model?

(h) (**) Suppose that $n = 2$, and the vector of weights found by the logistic regression model is $[-2.1, 0.4, 0.5]$, with the weights corresponding to the intercept term, course 1, and course 2, respectively. Would the model predict success for a student with an A in course 1 and a B in course 2? (Justify your answer mathematically.)

(i) (***) Suppose that S is truly independent of e.g. X_7. What would you expect the corresponding weight for X_7 to be?

(j) (***) Suppose that e.g. X_{10} is accidentally an exact copy of X_9. Explain why in this case, a gradient-based method might give different sets of weights depending on the learning rate.

4 Algorithmic fairness

Suppose that a certain protected demographic D makes up 10% of the population. Within demographic D (i.e. $D = 1$), 5% have a certain health condition ($H = 1$). In the remaining 90% of the population, only 1% have that health condition.

(The setup, and parts (a) and (b), are the same as in the Week 2 notes. But then the problem goes on to some new places!)

(a) (*) For a randomly selected person, what is $P(D = 1|H = 1)$?

(b) (*) Suppose we are trying to predict H. Consider a stupid model that, when given any new person, invariably predicts that $H = 0$. When applied to a representative sample of the population, how often will the stupid model be correct? (The takeaway is that impressive-looking accuracy numbers may mask serious problems!)

(c) (**) Does the stupid model in (b) achieve **parity**? What about **calibration**? Explain.

(d) (***) Now suppose that we try to incorporate additional features, including age. Suppose that the true distribution of age is $\mathcal{N}(\mu = 50, \sigma = 10)$ both within and outside of demographic D. So, for example, we would expect the mean age of a sample of 18 people (regardless of their demographic status) to be 50, with a variance of

$(\frac{1}{18})^2(18\sigma^2) = \frac{\sigma^2}{18}$, and therefore a standard deviation of $\frac{\sigma}{\sqrt{18}} \approx 2.35$. (For this problem, do not worry about the $n-1$ bias correction.)

Suppose that a sample of 10 people has only 1 from demographic D. To try to fight bias, the researchers create 8 more copies of that individual. (They copy the data point, not the actual person!) The mean age of this sample is still 50, but now what would you expect the standard deviation of that mean to be? (This illustrates one issue with trying to correct for bias in this way!)

5 Miscellaneous shorties

(a) (**) You are playing an intense game of Monopoly[2], and you suspect that one of the six-sided dice might be unfair. You start with a uniform prior and then record 3 1s, 2 2s, 5 3s, 4 4s, no 5s, and 1 6. What is the form of your Dirichlet posterior? (Answer as $Dir()$ with some number of parameters; you do not need to calculate the actual PDF.)

(b) (**) Suppose we are doing Naive Bayes (without Laplace smoothing) on a dataset with two (binary) features and a binary output. Even if the Naive Bayes assumption is correct, why is the following statement incorrect in general?

$$P(Y=0|X_1=1, X_2=0) = 1 - P(X_1=1|Y=1) \cdot P(X_2=0|Y=1) \cdot P(Y=1)$$

(c) (***) Let $W \sim Bin(10, 0.1)$, $X \sim Bin(10, 0.1)$, and $Y \sim Bin(20, 0.2)$ be random variables, with X and Y independent but W and X not necessarily independent. Let $A = W + 2X$ and $B = X + 2Y$. For each of the following, say whether or not it can be calculated exactly, and give the exact answer (or an expression for it) if so: $\mathbf{E}[A]$, $\mathbf{Var}[A]$, $\mathbf{E}[B]$, $\mathbf{Var}[B]$.

6 Quad workout

Suppose that we try to define a *quad distribution*[3] as a continuous distribution supported on $[0, 1]$. It has real nonnegative parameters a and b, and the PDF

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

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).

[2] In all honesty, I recommend choosing literally any other board game instead.

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

(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.

(g) (***) Now suppose that we believe some real phenomenon is modeled by a random variable $X \sim Quad(a, b)$, with a and b unknown. Suppose that we observe two values $X_1 = 0$ and $X_2 = 1$. Explain how you would find maximum likelihood values for a and b, given this data. Include any necessary equations.

BEST OF LUCK ON THE EXAM!

Solutions begin on the next page.

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.
- The only way to get $C_{max} = 1$ is for every value from 1 to 4 to come up exactly once. We can solve this using sample and event spaces: the sample space is of size 4^4 (all possible outcomes, treating the dice as distinct) and the event space is of size 4! (the number of ways for 1, 2, 3, 4 to come up in some order on the four rolls). So $P(C_{max} = 1) = \frac{4!}{4^4} = \frac{24}{256}$.
- The only way to get $C_{max} = 4$ is for all the dice to come up the same value. We can view this as the probability that the second through fourth dice all match whatever came up on the first die. This probability is $(\frac{1}{4})^5$. Just for ease of comparison with the other values, let's call it $\frac{4}{256}$.
- To get the size of the event space for $P(C_{max} = 3)$, we can start by picking the value to be repeated (there are 4 choices) and the other value (there are 3 choices left over), then use the binomial coefficient $\binom{4}{3} = 4$ to get the number of ways to distribute these relative to one another. Therefore the size of the event space is $4 \cdot 3 \cdot 4 = 48$, so the overall probability is $\frac{48}{256}$.
- Now we can note that 1, 2, 3, and 4 are mutually exclusive and exhaustive outcomes, so $P(C_{max} = 2) = \frac{256-24-4-48}{256} = \frac{180}{256}$. But it's good to review how to get this value directly, so let's do that.

  We can do something like what we did for $P(C_{max} = 3)$, but now there are two possibilities for the leftover two: either they are the same (Scenario 1) or different (Scenario 2).

- In Scenario 1, we have one 2×-repeated value and two (different) singleton values. The multinomial coefficient gives us the number of ways to order these relative to each other: $\binom{4}{2,1,1} = \frac{4!}{2!1!1!} = 12$. There are $4 \cdot \binom{3}{2} = 12$ to choose the three values while avoiding double-counting. (See problem 8 in the 109A midterm practice problems for why it's $\binom{3}{2}$ here and not $3 \cdot 2$. Yes, it is annoying and I also have to double-check and think through it every time!) So the total number of ways here is 144.
- In Scenario 2, we have we have two 2×-repeated values. There are $\binom{4}{2} = 6$ ways to pick these while avoiding double-counting. Then there are $\binom{4}{2} = 6$ ways to order these relative to each other. So we get $6 \cdot 6 = 36$.

Since Scenario 1 and Scenario 2 are mutually exclusive, the total event space size is $144 + 36 = 180$, which is what we wanted. Yay!

(If you get flustered over situations like Scenario 2, it helps to just write out some values. For example, writing out 1122, 1212, 1221, 2112, 2121, 2211, then 1133, 1313... makes it much clearer that the number of ways for Scenario 2 to happen is 36.)

Therefore the must common outcome, *by far*, is to see at most two of one number. So even though we really do expect (in the mathematical sense) to see each value 1 time in the 4 rolls, we should not confuse ourselves and expect to see this happen often on any given set of 4 rolls. The mathematical expectation is what happens on average *over multiple sets of rolls*, and is often not super useful for thinking about what happens on any particular roll.

3. (a) We need $P(S = 0)$ and $P(S = 1)$. Once we have one of those from the data, we know the other, but then again, it's hard to imagine a way of processing the data where we couldn't have just as easily kept track of both the 0s and 1s at the same time.

For each of the n courses, there are 5 possible grade values, and then each of those comes in two flavors: conditioned on $S = 1$, and conditioned on $S = 0$. So there are $2 \cdot 5^n$ such parameters. However, notice that if we know $5^n - 1$ out of the 5^n possible parameters conditioned on $S = 1$, for instance, we know the other one as well, because their values must sum to $P(S = 1)$. So we technically only need to estimate $5^n - 1$ parameters from each of the two $|S = 0$ and $|S = 1$ categories, but again, this trick is only of theoretical interest since we need to look at all the data.

In summary, there are $2 \cdot 5^n + 2$ parameters total (which is the important part), but we could find the right set of only $2 \cdot 5^n - 1$ of them and then derive the other 3 (which is less important).

(b) There are only 1000 students, so there will be many grade/success combinations

that are not represented by any data point, resulting in a probability estimate of 0. This also means that estimates for new students with these previously-unknown grade/success combinations would come out as 0, which kinda defeats the purpose of making the model. Laplace smoothing would get rid of the 0s but would not address the model's poor ability to generalize.

(c) We'll leave out the discussion from (a) of which probabilities can be derived from which others. As before, we need $P(S = 0)$ and $P(S = 1)$. But now we only need parameters corresponding to grades in individual courses, not combinations of grades in all courses. Conditioning on $S = 1$, for example, we need $P(X_1 = 4|S = 1), P(X_1 = 3|S = 1), P(X_1 = 2|S = 1), P(X_1 = 1|S = 1), P(X_1 = 0|S = 1)$. (Again, once we have gotten four of these from the data, we technically know the fifth for free.)

So here there are $2 \cdot 5 \cdot n + 2 = 10n + 2$ probabilities in total.

(d) Naive Bayes assumes that – conditioning on success, for example – a student's grade in one class is independent of their grades in other classes. This doesn't seem very plausible. That is, suppose the model learns that

- for class X_1, $P(X_1 = 4|S = 1) = 0.5$ and $P(X_1 = 3|S = 1) = 0.3$
- for class X_2, $P(X_1 = 4|S = 1) = 0.6$ and $P(X_1 = 3|S = 1) = 0.2$

But it's likely that in general (and even within the $S = 1$ cohort), people who get an A in the first class tend to be much more likely to get an A in the second class as well. This would violate the independence assumption. (Remember that this doesn't mean that Naive Bayes can't be used! We know it's naive – it's right in the name...)

(e) This is just like what you did in Problem 2 of Homework 6, except now there are 5 possible values that X_1 can take on, and the Laplace smoothing adds one "bonus" instance to each of them. So the denominator has a +5 instead of a +1.

$$P(X_1 = 3|Y = 1) = \frac{(\text{count of data points with } X_1 = 3, Y = 1) + 1}{(\text{count of data points with } Y = 1) + \mathbf{5}}$$

(f) A logistic model with no intercept term produces a decision boundary that *has* to go through the origin. This is usually an undesired restriction.

E.g., in the case of this model, suppose there are two courses. We want to draw a decision boundary (line, in this case) that separates the $S = 1$ points from the $S = 0$ points. But we are hampered in our ability to do so if that line must go through the origin, especially since most of the points from both categories will be far from the origin.

(g) If we added an extra feature X_0 to Naive Bayes that was 1 for everyone, it would just create a multiplicative $P(X_0 = 1|S = 1)$ or $P(X_0 = 1|S = 0)$ term that would always be 1, and so it would essentially not be there.

(h) The dot product of the weights vector and the feature vector is $[-2.1, 0.4, 0.5] \cdot [1, 4, 3] = 1.0$. Plugging this into the sigmoid, the predicted probability is

$\boxed{\frac{1}{1+e^{-1}}}$. This is greater than 0.5, so the model predicts success.

(i) If S and X_7 are truly independent within the dataset, the model can derive no useful predictive information from X_7 and should in theory give it a weight of 0.

Unfortunately this isn't always the case in a more complex model, because of interactions with other features. Consider these data points:

S	X_7	X_8
1	0	0
1	1	1
0	0	0
0	1	0

By inspection of the first two columns, S and X_7 are independent. (Specifically, $P(S = s, X_7 = x) = P(S = s)P(X_7 = x)$ for all possible (s, x) pairs.) However, when I use our HW 6 code or sites to fit this, I get nonzero weights for the X_7 term.

(j) Suppose that our fitting function assigns weights w_9 and w_{10} to X_9 and X_{10}. Then observe that because X_{10} is the same as X_9, one of many other equally good sets of weights would be $w_9 + w_{10}$ and 0 for X_9 and X_{10}. Depending on how our learning rate pushes us along the landscape, we might end up at different equally good optima. (In this case, the landscape is still convex, but not *strictly* convex.)

4. (a) By Bayes' Rule, $P(D = 1|H = 1) = \frac{P(H=1|D=1)P(D=1)}{P(H=1)}$. We are given that $P(H = 1|D = 1) = 0.05$ and $P(D = 1) = 0.1$. Using the Law of Total Probability for the denominator, we have $P(H = 1) = P(H = 1|D = 1)P(D = 1) + P(H = 1|D = 0)P(D = 0)$, and we are told that $P(H = 1|D = 0) = 0.01$ and $P(D = 0) = 0.9$. Putting this all together, we get $\frac{(0.05)(0.1)}{(0.05)(0.1)+(0.01)(0.9)} = \boxed{\frac{5}{14}}$.

Here's another less formal way to solve problems like this: say there are 1000 people in the population. Then 100 of them have $D = 1$, and 5 of those have $H = 1$. The remaining 900 people have $D = 0$, and 9 of those have $H = 1$. So there are 14 people with $H = 1$, and 5 of them have $D = 1$.

(b) The total fraction of people with $H = 1$ is $(0.1)(0.05) + (0.9)(0.01) = 0.014$. Therefore the total fraction of people with $H = 0$ is $1 - 0.014 = 0.986$. So the model will be correct 98.6% of the time, even though it is not even trying to do the thing it is supposed to do, and even though it is producing worse results for the demographic that presumably needs this disease detection the most.

(c) The stupid model achieves parity because the probability of a positive prediction ($H = 1$) is the same even when conditioned on each group: $P(H = 1|D = 1) = P(H = 1|D = 0) = 0.986$. However, it does not achieve calibration, because its probability of a correct response (T) is different when conditioned on each group: $P(H = T|D = 1) = 0.95$, and $P(H = T|D = 0) = 0.99$.

(d) Let X_1 be a random variable corresponding to the copied person, and X_2 through X_{10} be random variables corresponding to the other people. Each of these random variables has mean $\mu = 50$, $\sigma = 10$, and $\sigma^2 = 100$. Then the mean of the new artificial sample is distributed as $\frac{1}{18}(9X_1 + X_2 + ... + X_{10})$. (The $9X_1$ is *not* the same as $X_1 + ... + X_1$ nine times... make sure you understand why!) The parenthetical part has variance $81\sigma^2 + 9\sigma^2 = 90\sigma^2$, and then the variance of $\frac{1}{18}$ of that is $(\frac{1}{18})^2 90\sigma^2 = \frac{5}{18}\sigma^2 = \frac{500}{18}$. Accordingly, $\sigma = \sqrt{\frac{500}{18}} = \boxed{\frac{5\sqrt{10}}{3}} \approx 5.27$, which is much larger!

This is another one I didn't fully believe until I wrote the code:

```
from scipy.stats import norm
import numpy as np

vals = []
TRIALS = 1000000
for i in range(TRIALS):
    r = list(norm.rvs(loc=50, scale=10, size=18))
    vals.append(sum(r) / 18)
print("Original:", np.std(vals))

vals = []
for i in range(TRIALS):
    r = list(norm.rvs(loc=50, scale=10, size=10))
    r.extend([r[0]]*8)
    vals.append(sum(r) / 18)
print("With copied person:", np.std(vals))
```

5. (a) The beta is a special case of the Dirichlet, which is just like a beta where there can be multiple outcomes (e.g., 1, 2, 3, 4, 5, or 6) instead of just two (e.g., heads or tails). Just as a uniform prior for a (necessarily two-outcome) beta is $Beta(1, 1)$, a uniform prior for a (six-outcome) Dirichlet is $Dir(1, 1, 1, 1, 1, 1)$. Then, just as seeing heads and tails increments the first and second parameters of the beta (as we do Bayesian updates), our Dirichlet will be updated to $Dir(4, 3, 6, 5, 1, 2)$.

(b) The statement is incorrect because we are trying to apply the Law of Total Probability here, but these are likelihoods, not probabilities. Likelihoods come from

products of PDFs (and a denominator that we often ignore), and they are not restricted to being between 0 and 1. So you can't use a 1 minus trick – you have to calculate the likelihoods for the two options ($Y = 1$ and $Y = 0$) separately.

(c) The means and variances of *independent* random variables add, *even if they are not identically distributed*. This is just a consequence of how the mean and variance are defined. So $\mathbf{E}[B] = \mathbf{E}[X + 2Y] = \mathbf{E}[X] + \mathbf{E}[2Y] = \mathbf{E}[X] + 2\mathbf{E}[Y] = (10)(0.1) + (40)(0.2) = 9$, and $\mathbf{Var}[B] = \mathbf{Var}[X + 2Y] = \mathbf{Var}[X] + \mathbf{Var}[2Y] = \mathbf{Var}[X]+2^2\mathbf{Var}[Y] = (10)(0.1)(1-0.1)+(4)(20)(0.2)(1-0.2) = 0.9+12.6 = 13.7$.

The means of random variables add even if they are neither independent nor identically distributed. So $\mathbf{E}[A] = \mathbf{E}[W+X] = \mathbf{E}[W]+\mathbf{E}[X] = (10)(0.1)+(10)(0.1) = 2$. However, there is no way to find the variance in this case. It depends on the joint distributions of W and X – for instance, if W acts like an exact copy of X (e.g., maybe the two variables are defined on the same draw from $Bin(10, 0.1)$), then their combined variance will be as large as possible.

6. (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 $\boxed{1-\Phi(\frac{-9}{\sqrt{13}})}$.

(g) The likelihood of seeing these two data points is proportional to the product of the $f(x)$ evaluated at each point, i.e.,

$$L(a,b) \propto f(0)\cdot f(1) = \frac{2a(0)+2b}{a+2b}\cdot\frac{2a(1)+2b}{a+2b} = \frac{4b(a+b)}{(a+2b)^2}$$

Now we find the partial derivatives of $L(a, b)$ with respect to a and b. (Notice that there is no reason to work with log likelihoods here, because there are no exponential terms to eliminate.) Using the quotient rule for derivatives:

$$\frac{\partial L(a,b)}{\partial a} = \frac{(a+2b)^2(4b) - (4b(a+b))(2(a+2b)(1)}{(a+2b)^4} = -\frac{4ab}{(a+2b)^3}$$

$$\frac{\partial L(a,b)}{\partial b} = \frac{(a+2b)^2(4a+8b) - (4b(a+b))(2(a+2b)(2)}{(a+2b)^4} = \frac{4a^2}{(a+2b)^3}$$

At this point, unless we find some clever means of seeing an exact solution, we need to use gradient ascent. This entails finding these gradients at our starting location (say, (1, 1)), using those to take a small step (with some learning rate that is not too large to overshoot, but large enough to make progress), finding gradients at our new location, etc. An optimal solution (that obeys the constraints in (a)) turns out to be $a = 0, b = 1$.