Chris Piech Fall 2024
CS109

CS109 Final Exam

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. For example, describe the distributions and parameter values you used, where appropriate. It is fine for your answers to include summations, **integrals**, products, factorials, exponentials, and combinations.

You can leave your answer in terms of Φ (the CDF of the standard normal) or Φ^{-1} (the inverse CDF). For example $\Phi\left(\frac{3}{4}\right)$ is an acceptable final answer. Recall that the exam is going to be "curved" according to the difficulty of the questions and as such hard questions will not translate to lower grades.

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

I acknowledge and accept the letter and spirit of the honor code. I pledge to write more neatly than I have in my entire life:

Signature: ____

Family Name (print): ____

Given Name (print): ____

Email (preferably your gradescope email): ____

1. Short Answer [17 points]

a. (5 points) Let $X \sim \text{Exp}(\lambda = 2.5)$. What is $P(X > 5)$?

We can use the CDF of the Exponential:

$$\begin{aligned} P(X > 5) &= 1 - F(5) \\ &= 1 - \left(1 - e^{-5\lambda}\right) = e^{-12.5} \end{aligned}$$

b. (6 points) A binary classification machine learning model always outputs 0.6 for the probability that $Y = 1$, regardless of the features given. 60% of the test dataset has $Y = 1$.
i) Would this model be considered accurate? Answer Yes or No and give one sentence of explanation:

For 60% of datapoints, the model would correctly predict that $Y = 1$ (because outputting $P(Y = 1) = 0.6$ would mean predicting a 1). For the remaining 40% of datapoints, $Y = 0$, but the model would still predict 1. Thus, the model makes the correct prediction only 60% of the time, which is not very accurate.

ii) Would it be considered calibrated? Answer Yes or No and give one sentence of explanation:

To test calibration, we divide up model predictions into bins and see if the fraction of datapoints in that bin where $Y = 1$ matches the average prediction for that bin. In this case, we would only have one bin, and all the datapoints would fall in it. In that bin, $Y = 1$ for 60% of the datapoints, so the model will appear perfectly calibrated!

c. (6 points) As servers age, their probability of crashing increases. Here is a table of expected crashes per year for servers of different ages:

Server Age	Expected Crashes Per Year
1-2 years	0.5
3-4 years	3.2
5+ years	9.1

At a large computing center, 30% of servers are 1-2 years old, 50% are 3-4 years old, and the rest are at least 5 years old. A server is chosen at random to be assigned to a user. What are the expected number of crashes next year for this user's server?

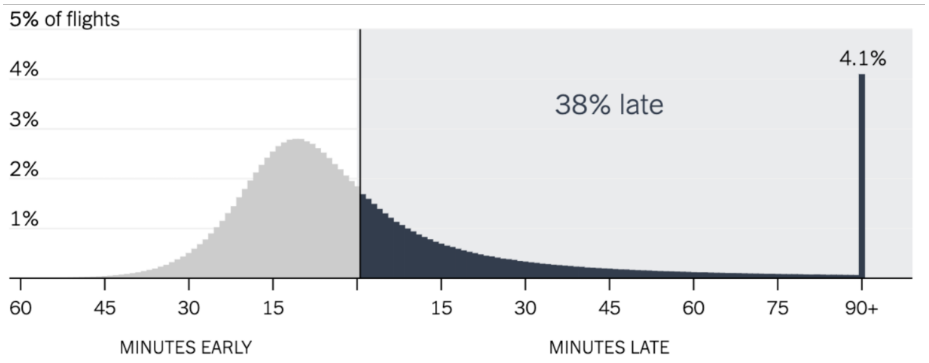
Let X be the number of crashes next year for this server. $E[X]$ depends on the age of the server, and the age is unknown, BUT we have a probability distribution for it. This is an ideal case for applying the Law of Total Expectation.

Let Y be a categorical random variable for which age group this server is in (1-2 years, 3-4 years, or 5+ years). Then LOTE tells us:

$$\begin{aligned} E[X] &= \sum_y E[X|Y = y]P(Y = y) \\ &= E[X|Y = \text{1-2 years}] \cdot 0.3 + E[X|Y = \text{3-4 years}] \cdot 0.5 + E[X|Y = \text{5+ years}] \cdot (1 - 0.3 - 0.5) \\ &= 0.5 \cdot 0.3 + 3.2 \cdot 0.5 + 9.1 \cdot 0.2 \end{aligned}$$

2. Airline Flights [16 points]

38% of all flights in 2023 were late by one minute or more.
4.1% of all flights in 2023 were more than 90 minutes late.
Here is the full distribution of delays for flights in 2023 from the Bureau of Transportation Stats:



Assume all flights are independent: the number of minutes late of any one flight is independent of any other.

a. (5 points) If you take 3 flights in 2023, what is the probability that none are late?

From the problem, $P(\text{one flight is late}) = 0.38$. we can break down a probability of AND between 3 flights using independence:

$$\begin{aligned} P(3 \text{ flights not late} &= P(\text{flight 1 not late AND flight 2 not late AND flight 3 not late}) \\ &= P(\text{one flight not late})^3 \\ &= P(\text{one flight late}^C)^3 = (1 - 0.38)^3 \end{aligned}$$

b. (5 points) If you take 2 flights in 2023, what is the probability that either flight one or flight two is late?

There are two possible interpretations to the wording of this question. For both, let X be the number of late flights. Because flights are independent and are late with a consistent probability, $X \sim \text{Bin}(n = 2, p = 0.38)$.

Interpretation 1: The problem is asking if exactly one of two flights is late (*not* including if both are late). For this assumption, we solve for $P(X = 1) = \binom{2}{1}0.38^1(1 - 0.38)^1 = 2 \cdot 0.38 \cdot 0.62$.

Interpretation 2: The problem is asking if at least one of two flights is late (including if both are late). For this assumption, we solve for $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{2}{0}0.38^0(1 - 0.38)^2 = 1 - 0.62^2$.

c. (6 points) Given that a flight in 2023 is late, what is the probability that it is delayed by more than 90 minutes?

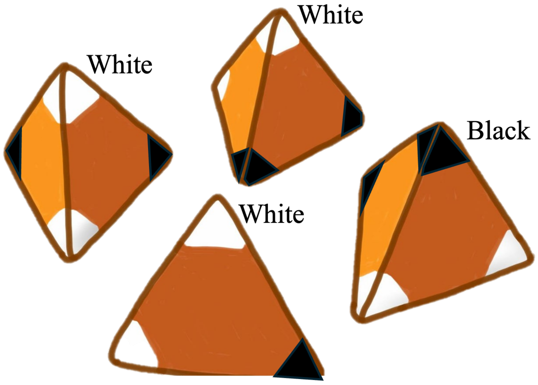
We can use the Definition of Conditional Probability:

$$P(\text{delayed more than 90 min} \mid \text{late}) = \frac{P(\text{delayed more than 90 min} \cap \text{late})}{P(\text{late})} = \frac{0.041}{0.38}$$

We simplify the numerator by recognizing that if a flight is delayed more than 90 min, then it is definitely late, so $P(\text{delayed more than 90 min} \cap \text{late}) = P(\text{delayed more than 90 min})$.

3. Board Games [16 points]

a. (8 points) The Royal Game of Ur is the oldest known board game. This game has four “ur dice.” Here is a picture of the result of rolling the four ur dice where only one dice rolled a “black”:



An “ur dice” is a four-sided pyramid with its four corners painted – two corners are white and two corners are black. A roll of an ur dice is “black” if the corner pointing upward is black. Each of the four corners is equally likely to be the corner pointing upward.

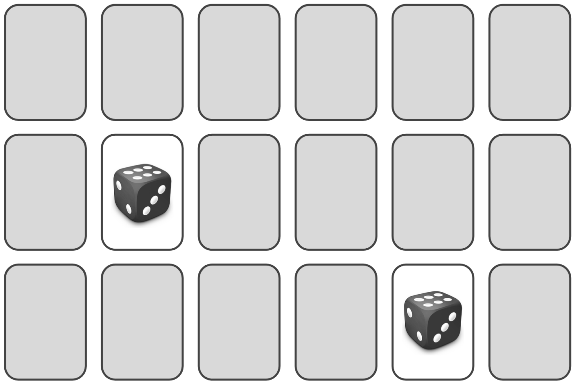
It is your turn and you will win the game if you either roll exactly four blacks or exactly two blacks. What is the probability that you win the game?

We are rolling a fixed number of dice (4), where each dice has 2 possible outcomes (white or black), and those outcomes have a consistent probability. Thus, we can use the Binomial.

Let X be the number of dice that are black after we roll. $X \sim \text{Bin}(n = 4, p = 0.5)$. The event that we roll 4 blacks corresponds to $X = 4$, and the event that we roll 2 blacks corresponds to $X = 2$. These events are mutually exclusive, so the probability of either of them happening is the sum of the individual probabilities:

$$\begin{aligned} P(\text{win}) &= P(X = 2 \text{ or } X = 4) \\ &= P(X = 2) + P(X = 4) \\ &= \binom{4}{2} 0.5^2 (1 - 0.5)^2 + \binom{4}{4} 0.5^4 (1 - 0.5)^0 \\ &= \binom{4}{2} 0.5^4 + 0.5^4 \end{aligned}$$

b. (8 points) In a game of Memory, 18 cards containing 9 distinct matching pairs are placed face down. If you randomly select two cards to turn over, what is the chance that the two cards match?



We can calculate this probability using counting: $P(E) = \frac{|E|}{|S|}$.

Our sample space is all possible ways to choose 2 cards from 18: $|S| = \binom{18}{2}$. Here we count unordered because the most straightforward way to count the event space is unordered, and the two must be counted consistently.

Our event space is all possible distinct ways to choose 2 of the same card. Since there are 9 distinct pairs, there are 9 different possible ways to pick a pair. $|E| = 9$.

Putting these together, $P(\text{match}) = \frac{9}{\binom{18}{2}} = \frac{9}{\frac{18 \cdot 17}{2}} = \frac{1}{17}$.

This solution can also be arrived at by reasoning that after you pick one card, out of 17 remaining cards, only 1 is a match to the card you've picked.

4. Song of the Quarter [25 points]

This quarter in CS109 there were 167 songs that were voted on. For each song, we have a list of votes where each vote is an integer in the set $\{1, 2, 3, 4, 5\}$. We assume all votes for a song are IID samples from the "true" distribution of CS109 opinion on the song.

For each song i we have m_i votes stored in a list **`votes[i]`** = $[x_1, x_2, \ldots, x_{m_i}]$. We have already calculated:

$$\mu_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_j \qquad \text{using } \texttt{np.mean(votes[i])}$$

$$\text{var}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}(x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i])}$$

$$\text{svar}_i = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i], ddof=1)}$$

a. (7 points) Song 1 has $m_1 = 45$ votes. We have calculated:

$$\mu_1 = 3.82 \qquad \text{var}_1 = 1.4 \qquad \text{svar}_1 = 1.5$$

Estimate the probability that the true average rating for song 1 is less than 3.

We can model the average rating of a song using the Central Limit Theorem, since each vote is IID. Let $\bar{X}_1$ be the average rating for song 1 from our collected votes.

$$\bar{X}_1 \sim N(\mu, \frac{S^2}{n})$$

In this case, our best estimate of the mean is $\mu_1 = 3.82$, our sample variance is $S^2 = \text{svar}_1 = 1.5$, and $n = m_1 = 45$. Then, using the CDF of the Normal, the probability we want is:

$$P(\bar{X}_1 < 3) = \phi\left(\frac{3 - 3.82}{\sqrt{\frac{1.5}{45}}}\right)$$

b. (8 points) Song 1 has $m_1 = 45$ votes. Song 2 has $m_2 = 36$ votes. We have calculated:

Song 1:	$\mu_1 = 3.82$	$\text{var}_1 = 1.4$	$\text{svar}_1 = 1.5$
Song 2:	$\mu_2 = 3.79$	$\text{var}_2 = 1.7$	$\text{svar}_2 = 1.8$

What is the probability that the true average of Song 1 is greater than the true average for Song 2?

We again use the CLT to determine distributions of sample means:

$$\bar{X}_1 \sim N(\mu_1, \frac{\text{svar}_1}{m_1})$$

$$\bar{X}_2 \sim N(\mu_2, \frac{\text{svar}_2}{m_2})$$

Then, to determine if the average for song 1 is greater than the average for song 2, we can recall that $a > b$ is equivalent to $a - b > 0$ and create a distribution for the difference in averages between the two songs.

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\text{svar}_1}{m_1} + \frac{\text{svar}_2}{m_2})$$

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 > 0) &= 1 - P(\bar{X}_1 - \bar{X}_2 < 0) \\ &= 1 - \phi\left(\frac{0 - (3.82 - 3.79)}{\sqrt{\left(\frac{1.5}{45} + \frac{1.8}{36}\right)}}\right) \end{aligned}$$

c. (10 points) Write pseudo-code to calculate a p-value for the significance of the difference between the average of song 1 and song 2. That is, find the probability that the votes for both songs are samples from the same universal distribution and the observed difference in averages $|\mu_1 - \mu_2|$ is purely due to random chance.

Let **votes[1]** be the list of votes for song 1.
Let **votes[2]** be the list of votes for song 2.

```
# find the absolute difference of means between the two songs' votes
observed_diff = np.abs(np.mean(votes[1]) - np.mean(votes[2]))

m1 = len(votes[1])
m2 = len(votes[2])

# assume the null hypothesis: all votes came from same dist
pooled_samps = votes[1] + votes[2]

gt_obs_diff_counts = 0

for i in range(10000):
    # repeat the original experiment under the null hypothesis
    resample1 = np.random.sample(pooled_samps, size=m1, replace=True)
    resample2 = np.random.sample(pooled_samps, size=m2, replace=True)

    # count how often the difference in means is more extreme
    # than what we originally observed
    diff = np.abs(np.mean(resample1) - np.mean(resample2))

    if diff >= observed_diff:
        gt_obs_diff_counts += 1

# p-value = fraction of times we see something more extreme
# if we assume the null hypothesis
print(gt_obs_diff_counts / 10000)
```

5. What Name Doesn't Give Away Age? [20 points]

In the Name to Age problem in class, we came up with the following probability distribution that someone was born in year b given that their name is n:

$$P(B = b|N = n) = \frac{\text{count}(b, n)}{\sum_{y \in \text{years}} \text{count}(y, n)}$$

Write pseudo-code to choose the name that leaks the least information about age: specifically, the name where the distribution of the year they were born has the highest entropy. You can use the following variables and functions:

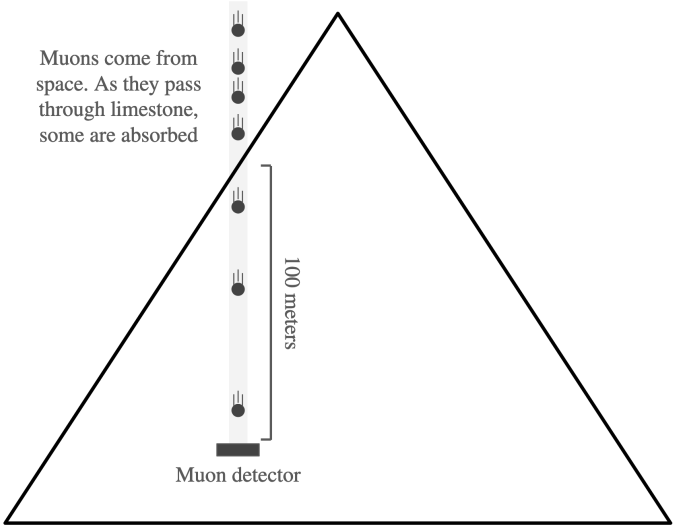
all_names, a list of all possible names to consider.
all_years, a list of all possible years to consider.
count(year, name), returns the number of people born in the specified year with the specified name.

```
# helper functions
def calc_entropy(pmf_dict):
    entropy = 0
    for prob in pmf_dict.values():
        entropy -= prob * math.log2(prob)   # note the minus equals
    return entropy

def make_pmf_for_name(name):
    # implement the equation given for P(B = b | N = name), for all b

    counts_per_year = {year : count(year, name) for year in all_years}
    total = sum(counts_per_year.values())

    pmf = {year : counts_per_year(year) / total for year in all_years}
    return pmf


def find_best_name():
    # loop through all names, keeping track of the name
    # with the highest entropy seen so far

    best_name_so_far = ""
    best_entropy_so_far = 0

    for name in all_names:
        pmf = make_pmf_for_name(name)

        entropy = calc_entropy(pmf)

        if entropy > best_entropy_so_far:
            best_entropy_so_far = entropy
            best_name_so_far = name

    return best_name_so_far
```

6. Hidden Chambers in the Great Pyramid [20 points]

We are going to build a tool to predict the existence of hidden chambers in the Great Pyramid of Giza using muons and probability. A muon is a special type of particle that originates in outer-space and arrives at earth as a Poisson process. When the muons hit limestone, they do not change direction, but as they travel through the limestone, they sometimes get absorbed.

A muon detector is positioned inside a chamber in the Great Pyramid. It is 100 meters below the edge of the pyramid and will only detect muons traveling straight down. Our goal is to estimate: how many meters of limestone are above our detector? This number will help us detect any hidden chambers!


Muons come from space. As they pass through limestone, some are absorbed
100 meters
Muon detector

If you knew how many meters of limestone each muon was passing through, you could calculate the rate of muons arriving per month. If x is the meters of limestone, then the rate is $100 \cdot e^{-x/40}$ muons per month.

a. (6 points) Imagine the entire 100 meter path is limestone. In that case, the rate of muons arriving per month on the detection plate is $100 \cdot e^{-100/40} = 8.2$. What is the probability that in one month you would observe 12 muons?

The problem tells us that muons follow a Poisson process; so, let M be the number of muons that we observe in one month. $M \sim \text{Poi}(\lambda = 8.2)$. Using the PMF of the Poisson:

$$P(M = 12) = \frac{8.2^{12} e^{-8.2}}{12!}$$

b. (14 points) Let X be your belief in the meters of limestone above the detection plate. Your prior belief is that any number of meters from 0 to 100 is equally likely: $X \sim \text{Uni}(0, 100)$. After one month, your detection plate has been hit by 12 muons. What is your updated belief in X?
Recall: You may leave your answer with integrals or sums. You don't need to simplify for full credit.

This is an inference problem, since we want to update our belief in a random variable, X, given an observation. In this case, since X is continuous, we will be updating a PDF.

Let M be the number of muons we detect again. Using Bayes' Theorem:

$$f(X = x|M = 12) = \frac{P(M = 12|X = x)f(X = x)}{P(M = 12)}$$

$P(M = 12|X = x)$ follows a Poisson, where the values of λ is $100 \cdot e^{-x/40}$:

$$P(M = 12|X = x) = \frac{\left(100 \cdot e^{-x/40}\right)^{12} e^{-\left(100 \cdot e^{-x/40}\right)}}{12!}$$

Since X is Uniformly distributed with a range of length 100, $f(X = x) = \frac{1}{100}$.

With inference on a discrete random variable, we expand the denominator using the Law of Total Probability with a summation. For a continuous RV, the summation becomes an integral:

$$P(M = m) = \int_0^{100} P(M = m|X = x)f(X = x)dx$$

Putting it all together:

$$f(X = x|M = 12) = \frac{P(M = 12|X = x)f(X = x)}{P(M = 12)}$$

$$= \frac{\frac{\left(100 \cdot e^{-x/40}\right)^{12} e^{-\left(100 \cdot e^{-x/40}\right)}}{12!} \cdot \frac{1}{100}}{\int_0^{100} \frac{\left(100 \cdot e^{-x/40}\right)^{12} e^{-\left(100 \cdot e^{-x/40}\right)}}{12!} \cdot \frac{1}{100} dx}$$

$$= \frac{e^{-\frac{12x}{40} - 100 \cdot e^{-x/40}}}{\int_0^{100} e^{-\frac{12x}{40} - 100 \cdot e^{-x/40}} dx}$$

Some of the constants cancel out, and the exponents can be combined to simplify a little, but simplifying wasn't required.

7. Sorted Random Values [23 points]

We want to reason about the values produced by the following python code, which creates 10 random uniform values in the range [0, 1] and then sorts them from low to high:

```python
# generate 10 random uniform values
values = []
for i in range(10):
   value_i = random_uniform(0,1)  # sample from standard uniform
   values.append(value_i)

# sort all of the values ascending from low to high
sorted_values = sorted(values)
print(sorted_values)
```

Here is what the list could look like when printed. For clarity each value is rounded to two decimal places:

```
sorted_values = [0.03, 0.13, 0.45, 0.51, 0.52, 0.63, 0.69, 0.82, 0.88, 0.91]
        index :   0     1     2     3     4     5     6     7     8     9
```

a. (10 points) The first value produced by **`random_uniform`** is 0.4. What is the probability that it will end up at index 4 in the sorted list?
In other words: What is the probability that exactly 5 out of the 9 other values are greater than 0.4?

Let $X \sim \text{Uni}(0, 1)$ be any value produced by **`random_uniform`**. $P(X > 0.4) = 1 - P(X < 0.4) = 0.6$, by the CDF of the Uniform (intuitively, it is the fraction of the range that satisfies the event - for a range from 0 to 1, 60% of that range is greater than 0.4).

Then, we need to determine the probability that out of 9 values, exactly 5 are greater than 0.4, and we know that each of the 9 values has a consistent probability of satisfying this event. This fits a Binomial.

Let Y be the number of values greater than 0.4. $Y \sim \text{Bin}(n = 9, p = 0.6)$.

$$P(Y = 5) = \binom{9}{5} 0.6^5 (1 - 0.6)^4$$

b. (10 points) Let X_4 be the value at index 4 in the sorted list. What is the probability density of $X_4 = x$? *In other words: What is the probability density that $X_4 = x$ given that you know exactly 5 out of 9 numbers are greater than x? Recall: You may leave your answer with integrals or sums.*

We can set up this problem using inference, since we want a probability distribution, given that we have an observation to condition on. Using Bayes' Theorem:

$$f(X = x|Y = 5) = \frac{P(Y = 5|X = x)f(X = x)}{P(Y = 5)}$$

Let Y again be the number of values greater than x. We can generalize part A to write a general expression for $P(Y = 5|X = x)$.

We also already know that $f(X = x) = 1$ since $X \sim \text{Uni}(0, 1)$.

If we recall that the denominator will be constant that does not change for different values of X, we can just represent it as $\frac{1}{K}$ and know that we would need to normalize the distribution to sum to 1 later.

Plugging everything in:

$$f(X = x|Y = 5) = K \cdot \binom{9}{5}(1 - x)^5 x^4 \cdot 1 = K \cdot (1 - x)^5 x^4$$

This is recognizable as the PDF of the Beta distribution.

c. (3 points) What is the variance of X_4? Hint: can you recognize X_4 as a random variable we have studied in class?

Since $X_4 \sim \text{Beta}(a = 5, b = 6)$, we can determine the variance using the formula for the variance of any Beta:

$$Var(X_4) = \frac{a \cdot b}{(a + b)^2(a + b + 1)} = \frac{5 \cdot 6}{(5 + 6)^2(5 + 6 + 1)} = \frac{5}{242}$$

8. MLE of Negative Binomial [23 points]

You are trying to estimate the p parameter of a Negative Binomial. You already know that $r = 5$. Recall that if $X \sim \text{NegBin}(r, p)$:

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

You have n samples of X: [7, 12, 9, 12, 8, 12, . . .]. Let k_i be the i^{th} value in this dataset.
Assume that the samples are IID from the same Negative Binomial and that $r = 5$.

a. (5 points) If $p = 0.7$ and $r = 5$ what is the likelihood of seeing the first sample, $X = 7$?

Using the PMF of the Negative Binomial:

$$P(X = 7) = \binom{6}{4} 0.7^5 (1 - 0.7)^2$$

b. (18 points) Explain how you would choose parameter p and provide any necessary derivatives.

We can use MLE!

$$L(p) = \prod_{i=1}^{n} \binom{k_i - 1}{4} p^5 (1-p)^{k_i - 5}$$

$$LL(p) = \sum_{i=1}^{n} \log \binom{k_i - 1}{4} + 5 \log p + (k_i - 5) \log(1-p)$$

We want to find the argmax of $LL(p)$, so we'll take the derivative:

$$\frac{\partial LL(p)}{\partial p} = \sum_{i=1}^{n} \frac{5}{p} - \frac{k_i - 5}{1 - p}$$

To use this derivative to find the best estimate of p, we could use gradient ascent, which allows us to initialize a starting value for p and then by incrementally update it, which each step updating p to a value that improves the likelihood of the data. This will eventually converge to a locally optimal p.

Alternatively, we can set the derivative equal to 0 and solve for p:

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{5}{p} - \frac{k_i - 5}{1 - p} &= 0 \\
\frac{5n}{p} - \frac{1}{1 - p} \sum_{i=1}^{n} (k_i - 5) &= 0 \\
\frac{5n}{p} &= \frac{1}{1 - p} \sum_{i=1}^{n} (k_i - 5) \\
5n(1 - p) &= p \sum_{i=1}^{n} (k_i - 5) \\
5n &= p \left(5n + \sum_{i=1}^{n} (k_i - 5) \right) \\
p &= \frac{5n}{\left(5n + \sum_{i=1}^{n} (k_i - 5) \right)} \\
p &= \frac{5n}{\sum_{i=1}^{n} k_i}
\end{aligned}
$$

9. Recalibrating an Uncalibrated Model [20 points]

You have an uncalibrated binary classification model that outputs values $\hat{p} \in [0, 1]$. These outputs are meant to be the probability that $Y = 1$. However, the outputs from this model are **not** well-calibrated. For instance, among all examples where $\hat{p} \approx 0.9$, it was the case that Y was 1 only 70% of the time. To recalibrate the model's outputs you decide to use Platt Recalibration, where the corrected probability that $Y = 1$ is:

$$P(Y = 1 \mid \hat{p}) = \sigma(a \cdot \hat{p} - 0.5)$$

$\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function and a is the parameter of the recalibration model. Here is the partial derivative of the Platt Recalibration model with respect to a:

$$\frac{\partial}{\partial a}\sigma(a \cdot \hat{p} - 0.5) = \sigma(a \cdot \hat{p} - 0.5) \cdot [1 - \sigma(a \cdot \hat{p} - 0.5)] \cdot \hat{p}$$

a. (4 points) For a new datapoint the uncalibrated model outputs $\hat{p}$ of 0.9. If you use Platt Recalibration with $a = 2$ what is the recalibrated probability that $Y = 1$?

$$P(Y = 1 \mid \hat{p} = 0.9, a = 2) = \sigma(2 \cdot 0.9 - 0.5) = \sigma(1.3)$$

b. (16 points) You are given a training dataset with n outputs from the uncalibrated model $(\hat{p}^{(i)}, y^{(i)})$ where $\hat{p}^{(i)}$ is the uncalibrated output and $y^{(i)} \in \{0, 1\}$ is the true binary outcome. Explain how you could estimate the value of a that makes the $y^{(i)}$ values as likely as possible. Solve for any and all partial derivatives required by your answer.

This problem is related to logistic regression. In both, we can use MLE to estimate parameters, and in both, the likelihood comes from the continuous PMF of the Bernoulli, since here we are still doing binary classification (Y is either 0 or 1):

$$L(a) = \prod_{i=1}^{n} P(Y = 1 \mid \hat{p}^{(i)})^{y^{(i)}} (1 - P(Y = 1 \mid \hat{p}^{(i)}))^{1-y^{(i)}}$$

$$LL(a) = \sum_{i=1}^{n} y^{(i)} \log P(Y = 1 \mid \hat{p}^{(i)}) + (1 - y^{(i)}) \log(1 - P(Y = 1 \mid \hat{p}^{(i)}))$$

To find the value of a that maximizes $LL(a)$, we can take the derivative using the chain rule:

$$\frac{\partial LL(a)}{\partial a} = \frac{\partial LL(a)}{\partial P(Y = 1 \mid \hat{p}^{(i)})} \frac{\partial P(Y = 1 \mid \hat{p}^{(i)})}{\partial a}$$

The second component is given to us in the problem (it is equivalent to $\frac{\partial}{\partial a}\sigma(a \cdot \hat{p}^{(i)} - 0.5)$. The first term looks the same as in logistic regression:

$$\frac{\partial LL(a)}{\partial P(Y = 1 \mid \hat{p}^{(i)})} = \sum_{i=1}^{n} \frac{y^{(i)}}{P(Y = 1 \mid \hat{p}^{(i)})} + \frac{1 - y^{(i)}}{1 - P(Y = 1 \mid \hat{p}^{(i)})}$$

Using this derivative, we would find the best estimate for a using gradient ascent.

That’s all folks! Thank you for the lovely quarter. You were a wonderful class. Airline data is real for US domestic flights. The Royal Game of Ur originated in ancient Mesopotamia around 4,600 years ago. Archeologists have found old boards, dice, and cuneiform tablets with rules. In March 2023 ScanPyramids team discovered a new void in the Great Pyramid using muon tomography and claim they are 99.9999% confident. The inventor of this technology won a Nobel Prize. In the real-life Song of the Quarter I used my confidence in a song averages to select which songs needed more votes. Platt Recalibration (which typically also has a learnable intercept term) is often the best method to fix calibration issues.