Chris Piech Fall 2021
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, 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.

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 True of False (36 points)

Answer True or False for each of the following questions. You must give a brief justification for your answer.

a. (6 points) If two random variables X and Y are independent, then it is still possible for some assignments $X = x$ and $Y = y$ to be dependent.

b. (6 points) The log likelihood function that is used to estimate p of a Bernoulli, for a set of observations, must always be 0 or smaller.

c. (6 points) You sampled two sets of numbers from unknown distributions: $(a_1 ... a_n)$ and $(b_1 ... b_n)$. You observe a difference between their modes, and you decide to use bootstrapping to ascertain whether the modes of their respective distributions are truly different. If your algorithm estimates a p-value of 0.99, this implies that the modes of the underlying distributions are not meaningfully different.

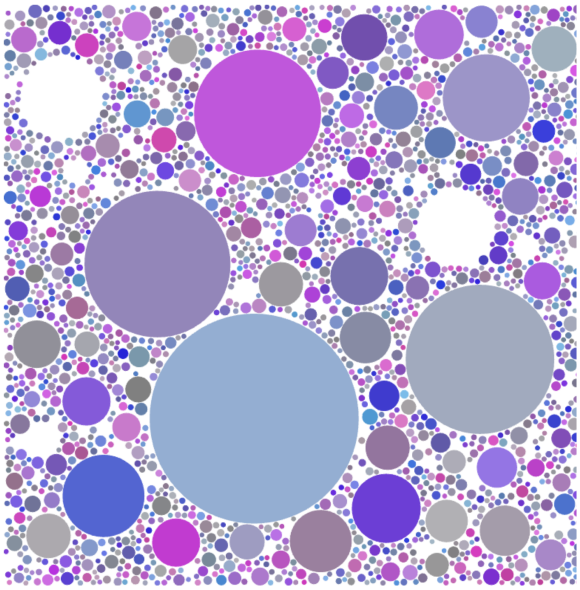
d. (6 points) Let $X_1 \ldots X_n$ be i.i.d. Then $(\frac{1}{n} \sum_{i=1}^{n} X_i)$ is a new random variable which tends toward a normal distribution with mean 0 and variance 0 as n tends toward infinity.

e. (6 points) Let $X \sim \text{Poi}(\lambda = 5)$ and let $Y \sim \text{Exp}(\lambda = 5)$. $P(X = 0)$ is equal to $P(Y > 1)$

f. (6 points) Let $X_1, X_2, ..., X_n$ be (not necessarily independent) random variables each with mean 1. It must be the case that $P(X_1 + ... + X_n >= n)$ is greater than 0.

2 Algorithmic Art (32 points)

We want to generate probabilistic artwork, efficiently. We are going to use random variables to make a picture filled with non-overlapping circles:



In our art, the circles are different sizes. Specifically, each circle's **radius** is drawn from a Pareto distribution (which is described below). The placement algorithm is greedy: we sample 1000 circle sizes. Sort them by size, largest to smallest. Loop over the circle sizes and place circles one by one.

To place a circle on the canvas, we sample the location of the center of the circle. Both the x and y coordinates are uniformly distributed over the dimensions of the canvas. Once we have selected a prospective location we then check if there would be a collision with a circle that has already been placed. If there is a collision we keep trying new locations until you find one that has no collisions.

Pareto Distribution
Notation: $X \sim \text{Pareto}(\alpha)$
Parameters: α, the shape parameter
Support: 1 to ∞
PDF: $f(x) = \frac{\alpha}{x^{\alpha+1}}$
CDF: $F(x) = 1 - \frac{1}{x^{\alpha}}$

a. (6 points) You sample a single radius from a Pareto distribution with $\alpha = 2$. What is the probability that the radius is 300 or greater?

b. (6 points) If you sample 1000 radii (radii is plural of radius) from the same distribution, Pareto($\alpha = 2$), what is the probability that there are at most two circles whose radii are 300 or greater? Provide an equation that you could use to compute the exact answer. Let p be your answer from part (a).

c. (5 points) You are trying to place a circle and have not been able to find a place without collisions yet. We are going to estimate p, the probability of finding a space for your current circle, as a Beta (which has the Uniform prior). After 100 tries with zero successes, how confident are you that the true probability of success is < 0.01? You may leave your answer in terms of $\text{betaCdf}(x, a, b)$, a function that returns the CDF of a beta random variable with parameters a and b at value x.

d. (15 points) Your artwork is inspired by the size of sand particles which also follow a Pareto distribution. You would like the alpha in your artwork to match that of sand in your local beach. You go to the beach and collect 100 particles of sand and measure their size. Call the measured radii $x_1 \ldots x_{100}$. Derive a formula for the MLE estimate of α.

3 How Lucky! (15 points)

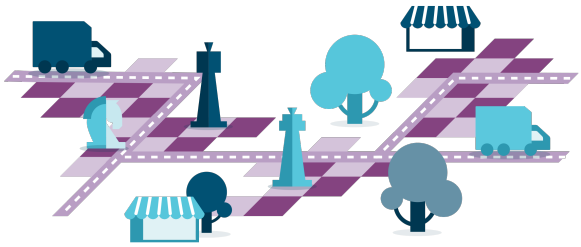
Craps is a game where you roll two dice (and look at the sum). In the game, you don't want to roll a 7. If you roll a 7, you are out. The current world record for the most rolls **without a 7** is from Patricia Demauro who rolled 154 pairs of dice before she rolled her first 7.

a. (5 points) What is the probability of getting a run consisting of 154 rolls without any 7s followed by a 7 on the 155th?

b. (10 points) There have been (estimated) 10^9 games of craps played in the last decade. What is the probability that exactly two games out of 10^9 will have a run of length 154 without any 7s (followed by a 7 on the 155th)? Use an approximation that could be used to efficiently compute the answer. Let p be the answer to part (a).

4 Probabilistic Supply Chains (25 points)

You are managing inventory of giant bolts in a company that produces airplanes (airplanes need giant bolts).



You have developed a probabilistic model of the joint probability between supply of bolts, S (how many bolts you will have in the next week) and demand for bolts, D (how many bolts you will need in the next week while constructing airplanes).

You can use the helper function `joint_pr(s, d)` which returns the joint probability $P(S = s, D = d)$. You may assume that both supply and demand are **discrete, integer** values in the range 0 to 100. You can't have negative bolts, and neither supply nor demand is ever more than 100. You can assume that all parameters passed to your functions are in the range 0 to 100. `joint_pr(s, d)` is defined for all s, d in range.

a. (5 points) Write pseudocode for a function `supply_pr(s)` that returns $P(S = s)$, the marginal distribution for supply. The function takes in s, a value of S.

b. (5 points) Write pseudocode for a function `demand_conditional_ex(s)` that returns $E[D|S = s]$, the conditional expectation of demand. The function takes in s, a value of S.

c. (7 points) Write pseudocode for a function `undersupply_pr()` that returns $P(D > S)$, the probability that demand > supply,

d. (8 points) The cost associated with having too few bolts is $5000 because it means you will delay the construction of an aircraft. There is an additional cost of $100 per bolt in your supply regardless of whether or not it is used. Write pseudocode for a function `expected_cost()` which returns your expected cost.

5 Chess.com Puzzles (10 points)

Chess.com is a website for playing chess. They are trying to estimate how well a player can solve chess puzzles (puzzle ability) as a random variable, A, which can take on **integer** values in the range 0 to 2000 inclusive. Higher abilities mean the player is better at chess puzzles. Note that ability is **discrete**.

Assume that the probability that a player gets a particular puzzle correct, conditioned on their ability being equal to a, is:

$$p_{\text{correct}} = 0.1 + 0.9 \cdot \sigma(a - 1200)$$

$\sigma(x)$ is the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Our user gets the puzzle correct. Write an expression to calculate the posterior belief that their ability equals a. In your calculation you should use the prior belief that chess.com had about their ability (their belief in the player's ability before they saw this puzzle result). Let prior(i) be a function which returns the prior belief that $P(A = i)$.

6 P-Hacking (36 points)

It turns out that science has a bug! If you test many hypotheses but only report the one with the lowest p-value you are more likely to get a spurious result (one resulting from chance, not a real pattern).

Recall p-values: A p-value was meant to represent the probability of a spurious result. It is the chance of seeing a difference in means (or in whichever statistic you are measuring) at least as large as the one observed in the dataset if the two populations were actually identical. A p-value < 0.05 is considered "statistically significant". In class we compared sample means of two populations and calculated p-values. What if we had 5 populations and searched for pairs with a significant p-value?

To explore this idea, we are going to look for patterns in a dataset which is totally random – every value is Uniform(0,1) and independent of every other value. There is clearly no significance in any difference in means in this toy dataset. However, we might find a result which looks statistically significant just by chance. Here is an example of a simulated dataset with 5 random populations, each of which has 20 samples:

	Pop 1	Pop 2	Pop 3	Pop 4	Pop 5
1	0.330	0.272	0.959	0.985	0.175
2	0.386	0.353	0.929	0.575	0.386
3	0.232	0.839	0.009	0.229	0.899
	0.836		0.003		
			...		
		0.833	0.333		
	0.649	0.723	0.565	0.061	0.479
20	0.726	0.158	0.678	0.498	0.645
Sample mean	0.534	0.579	0.474	0.437	0.545

The numbers in the table above are just for demonstration purposes. You should not base your answer off of them. We call each population a random population to emphasize that there is no pattern.

a. (2 points) How many ways can you choose a pair of two populations from a set of five to compare? The values of elements within the population do not matter nor does the order of the pair.

b. (3 points) What is the variance of a Uniform(0, 1)?

c. (5 points) What is an approximation for the distribution of the mean of 20 samples from Uniform(0,1)? If necessary, you can leave the parameters of the distribution in terms of v, the answer to (b).

d. (6 points) What is an approximation for the distribution of the mean from one population minus the mean from another population? Note: this value may be negative if the first population has a smaller mean than the second.

e. (8 points) What is the smallest difference in means, k, that would look statistically significant if there were only two populations? In other words, the probability of seeing a difference in means of k or greater is < 0.05.

f. (5 points) Give an expression for the probability that the smallest sample mean among 5 random populations is less than 0.2.

g. (7 points) Use the following functions to write code that estimates the probability that among 5 populations you find a difference of means which would be considered significant (using the bootstrapping method designed to compare 2 populations). Run at least 10,000 simulations to estimate your answer. You may use the following helper functions. Write pseudocode:

```
# the smallest difference in means that would look statistically significant
k = calculate_k()

# create a matrix with n_rows by n_cols elements, each of which is Uni(0, 1)
matrix = random_matrix(n_rows, n_cols)

# from the matrix, return the column (as a list) which has the smallest mean
min_mean_col = get_min_mean_col(matrix)

# from the matrix, return the row (as a list) which has the largest mean
max_mean_col = get_max_mean_col(matrix)

# calculate the p-value between two lists using bootstrapping (like in pset5)
p_value = bootstrap(list1, list2)
```

7 Calibration of ML Algorithms (26 points)

The heart disease model that you built for HW6 is going to be used by medical doctors! For their purposes the output probability, P(Y=1|X=x), of the logistic regression model is much more useful than the class prediction. Imagine two patients which are both predicted to be class Y=1. For one patient the logistic regression model says the probability that Y=1 is 0.99 and for the other it is only 0.60. In class we measured the accuracy of the predictions. Can we measure if the output probabilities are accurate as well?

To do so we look at all the output probabilities for our model on patients in the test set. We bucket all patients into five groups based on the output probability and then observe how many patients in that group actually had Y=0 and Y=1. For example the patient with a 0.99 output probability would go into the "very high p" group because the output probability is very high. Here is information for all five groups:

Group name	Average Logistic Regression Output Prob. for Group	Count(Y = 0) for Group	Count(Y = 1) for Group
Very low p	0.1	70	30
Low p	0.3	30	20
Borderline	0.5	20	20
High p	0.7	20	30
Very High p	0.9	30	70

Logistic Regression Output Probability is the average output probability for all patients in the group. Count(Y=0) is the number of patients in the group who had a true label 0.

a. (6 points) What is the accuracy of the model on the 100 patients in the "very high p" group?

b. (6 points) What is the probability of seeing 70 patients with $Y = 1$ and 30 patients with $Y = 0$ assuming the true probability of $Y = 1$ was actually 0.9?

c. (4 points) Based on your analysis of this group do you think that the probabilities output by the model are too high or too low? Explain in a sentence or so.

d. (10 points) We need to fix our probabilities. One solution is to "calibrate" the model on our test set. To do so, we are going to use a method call Platt Scaling. We will train a separate model which takes the logistic regression output probability and turns it into a better probability.

$$p_{\text{better}} = \sigma(\theta \cdot p_{\text{output}})$$

It only has one parameter, θ. Chose the value of θ that optimizes the likelihood of the **class labels** in the **test** set. Explain briefly how to learn this parameter θ. Include any derivations which would be necessary if one were to implement your strategy.

That's all folks. Algorithmic Art is based off a painting by Tyler Hobbs and our inspiration for this problem is thanks to Erin McCoy. The story of craps is true and it is thought to be one of the most improbably runs in betting history. One of the issues with our current supply chain is that folks were not reasoning about uncertainty. Chess.com is working with the piech lab to understand how people learn. P-hacking is a real problem – people find a spurious result and never publish the other hypotheses that they checked. Calibration is a great analysis to use for any machine learning model. Thank you all for the wonderful quarter and we hope you have a fantastic break. This was a great class and the teaching team really appreciated how positive, curious and intelligent you all were. All the best.