Chris Piech
CS109

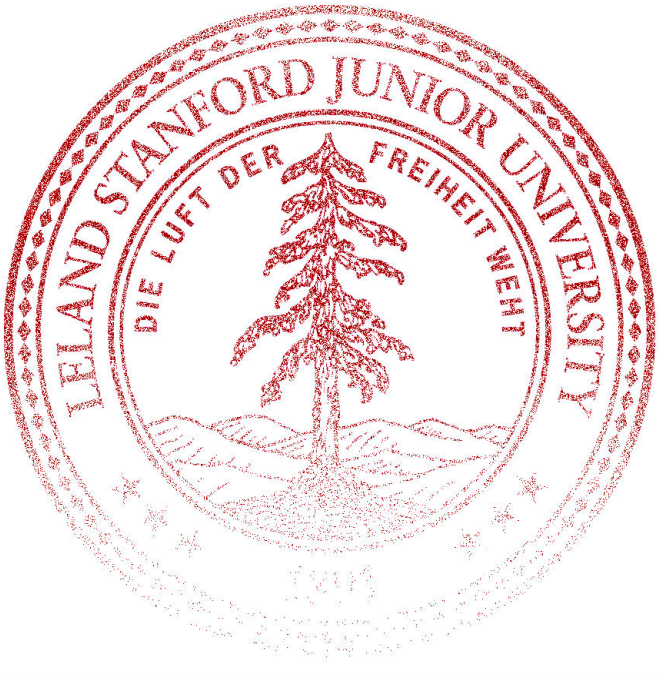
Fall 2022
Nov 2nd

CS109 Midterm Exam (Wednesday)

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam. You have 2 hours (120 minutes) to take the exam. The exam is 120 points, meant to roughly correspond to one point per minute of the exam. You may want to use the point allocation for each problem as an indicator for pacing yourself on 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). For example $\Phi(3/4)$ 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): _____

Stanford Email (@stanford.edu): _____

1 Weather Prediction [20 points]

We are building a weather prediction service for Cloud City. We have access to the full history of weather stored in a list called `history`. `history[i]` is `True` if it rained on day `i`. For example the history list for the last week at Stanford would look like: `history = [False, False, False, False, False, False, True]` if the only time it rained was the last day of the week. The `history` list for Cloud City has over 10,000 elements and represents the daily weather for the last 30 years.

a. (7 points) Write pseudo-code to estimate the probability of rain on a random day.

b. (7 points) Write pseudo-code to estimate the probability of rain tomorrow given that it was sunny today.

c. (6 points) How would you test if rain on a given day is independent of rain on the previous day, using the values calculated in part a and b?

2 Down to Lunch [20 points]

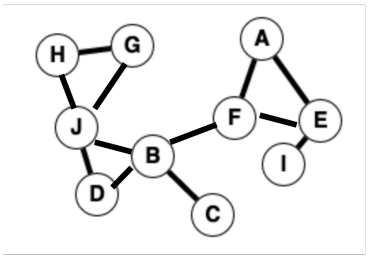
You are writing an app called Down to Lunch which invites people to share a meal. You want to minimize spam while maximizing the chance of creating a good sized group.

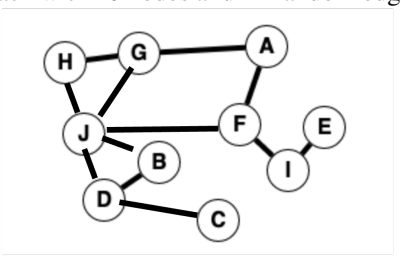
a. (8 points) Assume that each person checks their phones an average of 262 times a day (or 0.182 times a minute), this rate is uniform over the day and events are independent. What is the probability that a person checks their phone in the next 5 minutes?

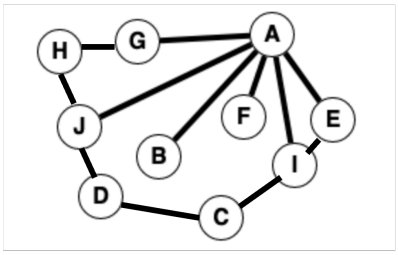
b. (12 points) Let p_a be your answer to part (a). Each person you invite to lunch will come if they check their phone in the next five minutes and they accept the invite. The probability that someone accepts an invitation, given that they check their phone in the next 5 minutes is $1/2$. If you invite 7 people, what is the probability of having a good sized lunch? A good sized lunch group has 1 to 4 guests accept an invite.

3 Random Edges [25 points]

Here are three different networks each with 10 nodes and 12 random edges:







a. (4 points) Consider a network with 10 nodes. Count the number of possible locations for undirected edges. An undirected edge from node A to node B is not distinct from an edge from node B to node A. You can assume that an edge does not connect a node to itself.

b. (6 points) Assume the same network (with 10 nodes) has 12 random edges. If each pair of nodes is equally likely to have an edge, how many unique ways could we chose the location of the 6 edges? Let k be your answer to part a.

c. (10 points) In the same network with 10 nodes and 12 edges, select a node uniformly at random. Let X be the number of edges connected to the node. What is the probability that $X = i$ for $0 \leq i \leq 9$. Recall that there are only 9 nodes to connect to since there are 10 nodes.

d. (5 points) Let $\mathrm{P}(X = i)$ be the answer to the previous part. What is $\mathrm{Var}(X)$?

4 Expressed Trait [15 points]

We would like to predict the probability that a trait is "on". This particular trait is turned on for a cell if either gene_1 or gene_2 are "active". A gene is considered active if at least one RNA copy of the gene was created in the last 10 seconds.

For gene_1, on average 30 RNA copies are created every 10 seconds,
For gene_2 on average 20 RNA copies are created every 10 seconds.

You may assume that the process of creating an RNA copy for a gene occurs independently of the last time an RNA copy was created, that RNA copies are created at a constant average rate, and that genes create RNA copies independently of each other.

What is the probability that the trait is on?

5 Mixing up Gaussians [20 points]

Consider the following function:

```
def sample():
    X = rand_bern(0.3)                        # bernoulli sample
    if X == 0:
       return rand_gauss(mu = -1, std = 2) # gaussian sample
    return rand_gauss(mu = 1, std = 1)      # gaussian sample
```

a. (10 points) What is the probability that the function returns a value less than 0?

b. (10 points) The function returns back the value 0.1. What is the probability that X was 1?

6 Bayesian Viral Load Test [20 points]

We are going to build a Bayesian Viral Load Test which updates a belief distribution regarding a patient's viral load. Though viral load is continuous, in our test we represent it by discretizing the quantity into whole numbers between 0 and 99, inclusive. The units of viral load are the number of viral instances per million samples.

a. (3 points) If a person has a viral load of 9 (in other words, 9 viruses out of every 1 million samples) what is the probability that a random sample from the person is a virus?

b. (7 points) We test 100,000 samples from one person for the virus. If the person's true viral load is 9, what is the probability that exactly 1 of our 100,000 samples is a virus? Use a computationally efficient approximation to compute your answer. Your approximation should respect that there is 0 probability of getting negative virus samples.

c. (10 points) Based on what we know about a patient (their symptoms and personal history) we have encoded a prior belief in a list `prior` where `prior[i]` is the probability that the viral load equals i. `prior` is of length 100 and has keys 0 through 99.

Write an equation for the updated probability that the true viral load is i given that we observe a count of 1 virus sample out of 100,000 tested. Recall that $0 \leq i \leq 99$. You may use approximations.

7 Ranked Evaluation [1 point]

Bonus point: please rank the questions from this exam in order of how confident you feel in the correctness of your answer. For example: writing $3 < 1 < 5$ would mean you are most confident in your answer for question 5, second most confident in your answer for question 1, and least confident in your answer for question 3. Please rank whole questions, not subparts.