Chris Gregg
CS109

Summer 2026
August 3rd

CS109 Midterm 2 Practice Problems

These three problems are written to be similar in *scope and style* to what you will see on Midterm 2. They are practice, not a preview: the topics here are drawn from across the second half of the course, and the exam draws from the same pool but is not built from these problems. Do not read anything into which topics do or do not appear below.

You are welcome to use the course reader for this practice, but for the midterm you will be limited to five back/front handwritten pages of notes.

As on the exam, it is fine for your answers to include summations, products, factorials, exponentials, logarithms, and combinations. You may leave answers in terms of Φ (the CDF of the standard normal) or Φ^{-1}. Do not use a calculator — the numeric values in the solutions are there so you can check yourself afterwards, not because you need to produce them.

In the event of an incorrect answer, an explanation of how you got there can earn partial credit. Naming the distribution and its parameters is almost always worth writing down.

Topics covered: Poisson and Multinomial random variables, conditioning and the law of total expectation, maximum likelihood estimation, the Central Limit Theorem, p-values and resampling, and entropy and information gain.

1 Cache Tiers

A web service receives read requests independently at a rate of 12 per minute. Each arriving request is served, independently of every other request, by one of three tiers: the **L1** in-memory cache with probability $1/2$, the **L2** shared cache with probability $1/3$, and, if both caches miss, the **disk** with probability $1/6$.

You may not use code to answer any part of this problem.

(a) Suppose you are told that exactly 9 requests arrived during one particular minute. Given that, what is the probability that exactly 5 were served by L1, exactly 3 by L2, and exactly 1 by disk?

(b) Now stop conditioning on the total. Let N be the number of requests in a one-minute window and let D be the number of those requests served by disk. State the distribution of N (with parameters) and compute $E[D]$ using the law of total expectation.

(c) It is a fact (which you may use without proof) that $D \sim \text{Poi}(2)$. What is the probability that at least 2 requests go to disk in a given minute?

(d) An L1 hit takes 1 ms, an L2 hit takes 5 ms, and a disk read takes 40 ms. Let T be the service time of a single request. Compute $E[T]$ and $\text{Var}(T)$.

2 Does the New Ranker Help?

For a single user session on your site, the time spent reading has mean $\mu = 12$ minutes and standard deviation $\sigma = 9$ minutes. Sessions are independent of one another. You do *not* know the shape of the per-session distribution.

(a) You collect $n = 400$ independent sessions and compute the sample mean $\bar{X}$. Approximate $P(\bar{X} > 13)$. Justify why the approximation is available to you even though the per-session distribution is unknown.

(b) You now run an experiment. 400 sessions use the old ranker and 400 different sessions use a new ranker, and the new group's mean is 0.8 minutes higher. State the null hypothesis you would test, and define, in one or two sentences, what the p-value for this experiment means.

(c) The code below estimates that p-value with a permutation test. Line (i) is written for you; fill in the three remaining blanks. `np.concatenate([a, b])` joins two lists; `np.random.shuffle(a)` randomly reorders `a` in place; `np.mean(a)` returns the average of `a`.

```
import numpy as np
NUM_TRIALS = 10000

def p_value(old, new):
    obs_diff = np.mean(new) - np.mean(old)
    pooled   = np.concatenate([old, new])   # (i)
    n_new    = len(new)
    count = 0
    for i in range(NUM_TRIALS):
        np.random.shuffle(pooled)
        sim_new  = pooled[:n_new]
        sim_old  = pooled[n_new:]


        sim_diff = ______________________________________  # (ii)


        if _____________________________________________:  # (iii)

            count += 1


    return _____________________________________________   # (iv)
```

(d) The function returns 0.031. State what you conclude at a significance threshold of $\alpha = 0.05$, and give one reason why this conclusion could still be wrong.

3 One Word of Evidence

An email is spam ($Y = 1$) with probability 0.25 and legitimate ($Y = 0$) otherwise. Let $W = 1$ if the email contains the word "urgent" and $W = 0$ otherwise, with

$$P(W = 1 \mid Y = 1) = 0.6, \qquad P(W = 1 \mid Y = 0) = 0.1.$$

Use $\log_2$ throughout parts (a)–(c), so that entropies are in bits. It is fine to leave answers as expressions involving logarithms.

(a) Compute $H(Y)$.

(b) Compute $P(W = 1)$, and then $P(Y = 1 \mid W = 1)$ and $P(Y = 1 \mid W = 0)$.

(c) Compute the conditional entropy $H(Y \mid W)$ and the information gain $H(Y) - H(Y \mid W)$. Interpret the information gain in one sentence.