Recitation Final Practice

Math-UA 185: Probability and Statistics / Spring 2026

Ed Chen

I've put together a few practice problems in light of the final exam coming up. I'll go over these during the final recitation on May 1st. Problems 1 through 4 are decently comprehensive and touch on most of the more challenging concepts in this course, whilst the latter 10 questions are more straightforward standard issue questions.

A fully explained solution set is also on Brightspace. Portions of the last 10 questions are pulled from the Dekking textbook, which is also a good resource for more practice. Partial inspiration for Problems 1 and 2 are taken from EECS126, at Berkeley.

Problem 1. Office Hours

A student arrives to office hours at 1 pm. The TA is particularly lazy today, and instead of being in his office continuously, checks his office according to a Poisson process with rate $\lambda = 5$ times per hour independently of the student's arrival time.

When the TA finally arrives, he gives the student bonus points linearly proportional to the time since his previous visit to the office, as he does not know how long the student has been waiting. Assume this rate is 1 point per hour.

a) Let W be the amount of time, in hours, that the student waits until the TA next appears. What is the distribution of W? What is $\mathbb{E}[W]$?

b) Let A be the amount of time, in hours, since the TA's previous office visit when the student arrives. What is the distribution of A? Are A and W independent?

c) Let B be the number of bonus points. Compute $\mathbb{E}[B]$, and claim the distribution of B.

d) What is the probability that the student receives more than 1 bonus point?

e) While waiting, the student sends emails according to a separate Poisson process at rate 3 emails per hour, independently of the TA's office-checking process. Let N be the number of emails sent before the TA appears. Find $\mathrm{P}(N = 0)$.

f) Now find the distribution of N, the number of emails sent before the TA appears.

Problem 2. Concert Time

After taking the final, you decide to go to a concert. There are only two staff members letting fans into the venue. Staff 1 has service times distributed as Exponential(2), and staff 2, who is more experienced at their job, has service times distributed as Exponential(3). All service times are independent, and measured in minutes.

When you arrive, both staff are busy with one fan each, and there is a single line of $n-1$ fans in front of you waiting for the first available staff member. Let T be the total time, in minutes, from when you enter the line until you are let into the venue.

a) Let $X_1 \sim$ Exponential(2) be the remaining time until staff 1 finishes with their current fan, and let $X_2 \sim$ Exponential(3) be the remaining time until staff 2 finishes with their current fan. Find the distribution and expectation of

$$M = \min\{X_1, X_2\}$$

b) Find the CDF of

$$Z = \max\{X_1, X_2\}$$

Interpret what Z represents in this scenario.

c) After waiting for a while, you are now first in line. Yay! You notice staff 1 is giving out free merch. What is the probability that staff 1 is the first worker to become available? Which staff is more likely to be available first?

d) Find the expected amount of time from arrival until you first reach a staff member.

e) Once you reach a staff member, what is your expected service time?

f) Find $\mathbb{E}[T]$, the expected total time from entering the line until being let into the concert.

g) Imagine now that the line is quite busy. The concert organizers, who are getting flak for only hiring two staff members, are promising a free drink ticket if your wait time is over 60 minutes. Use Markov's inequality to give an upper bound on the probability that your total time is over 60 minutes in terms of the number of people in line, including yourself n.

h) Your friend is quite excited at the prospect of a free drink and demands to know a tighter probability bound after they count a total of $n = 148$ fans in line (they haven't taken this course). Compute $\mathrm{Var}(T)$ and use Chebyshev's inequality to bound

$$\mathrm{P}\left(|T - \mathbb{E}[T]| \geq r\right)$$

What can you say about $\mathrm{P}(T \geq 60)$ using the bound you have found?

Problem 3. Boba Price Ceiling

An eager NYU Data Science graduate has decided to capitalize on a lucrative business opportunity to open a new boba shop near NYU. She is trying to decide how aggressively she can price drinks before students remember they have rent to pay.

She wants to better understand the distribution that describes the willingness-to-pay of students, X. As an avid data scientist, she assumes it follows a bounded rescaled Beta distribution $X \sim \text{Boba}(\alpha, \theta)$, where θ describes the maximum possible price the population would be willing to pay for a drink, and where α describes how clustered others' willingness-to-pay is around that maximum. The probability density is

$$f_{\alpha,\theta}(x) = \begin{cases} \dfrac{\alpha x^{\alpha-1}}{\theta^\alpha}, & 0 < x < \theta, \\ 0, & \text{otherwise} \end{cases}$$

For student i, let X_i be the maximum price, in dollars, that student is willing to pay for one drink. Suppose a poll is taken of n students, collecting samples

$$X_1, \dots, X_n$$

which we assume are i.i.d. with density $X \sim \text{Boba}(\alpha, \theta)$, with unknown $\alpha > 0$ and $\theta > 0$.

a) Confirm that $f_{\alpha,\theta}$ is a valid density for arbitrary α, θ, and find the CDF of X_i.

b) Write the likelihood and log-likelihood functions for α, θ. Recall since θ dictates the support of the distribution, you must use an indicator function in accurately describing the PDF.

c) Assuming a fixed α, find the Maximum Likelihood Estimator of θ, denoted $\widehat{\theta}$.

d) Now using $\widehat{\theta}$, find the Maximum Likelihood Estimator of α, denoted $\widehat{\alpha}$.

e) Compute the bias of the Maximum Likelihood Estimator $\widehat{\theta}$ of θ. You may find it helpful to first attain the PDF/CDF of $\widehat{\theta}$.

f) Assuming a known α, construct an unbiased estimator, denoted $\widehat{\theta}_0$ of θ using $\widehat{\theta}$.

g) Suppose now a helpful friend who also runs a boba shop lets our business owner know that their prior market research has established that the maximum possible willingness-to-pay is

$$\theta = 12$$

That is, we assume no student is willing to pay more than \$12 for one boba. Restate the MLE $\widehat{\alpha}$, compute the Fisher Information $I_n(\alpha)$, and state the Cramer–Rao lower bound for unbiased estimators of α.

h) A related large-sample principle to Cramer–Rao is that under regularity conditions the MLE is approximately normal with variance that meets the lower bound given by Cramer–Rao, ergo

$$\widehat{\alpha} \approx \mathcal{N}\left(\alpha, I_n^{-1}(\alpha)\right)$$

Use this approximation to give an approximate 95% confidence interval for α.

i) With this new information, suppose she sets the price of one drink to be p, where $0 \leq p \leq 12$. A randomly chosen student buys the drink if and only if it is below their willingness-to-pay, that is

$$X_i \geq p$$

Find the expected revenue from one student as a function of p.

j) Eager at this revelation, the polling study is conducted and the MLE $\widehat{\alpha}$ is determined to be 2. Find the price p^* that maximizes the expected revenue $r(p)$.

Problem 4. Not Exam Advice

A student wants to understand how caffeine consumption relates to exam performance. For student i, let

$$x_i = \text{number of caffeinated drinks consumed in the previous 24 hours} \qquad y_i = \text{exam score}$$

The student polls some friends, and fits the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}[\varepsilon_i] = \sigma^2$$

For the full dataset of $n = 8$ students, the following summary statistics are given:

$$\overline{x} = 3, \qquad \overline{y} = 86$$

$$S_{xx} = \sum_{i=1}^{8}(x_i - \overline{x})^2 = 28, \qquad S_{xy} = \sum_{i=1}^{8}(x_i - \overline{x})(y_i - \overline{y}) = 140$$

$$TSS = \sum_{i=1}^{8}(y_i - \overline{y})^2 = 1260$$

Afterwards, the student realizes there is actually another important distinction, and separates the data points into high and low sleepers. Within each group, the summary statistics are:

	n	$\overline{x}$	$\overline{y}$	S_{xx}	S_{xy}
Low sleep	4	1.5	73.5	5	-5
High sleep	4	4.5	98.5	5	-5

a) Compute the least-squares line of the overall data

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

b) Compute the residual sum of squares RSS and R^2 for the overall data.

c) Compute the least-squares slope separately within each sleep group.

d) Compare the overall slope with the within-group slopes. Can you explain what is happening?

e) Suppose the model errors in the overall caffeine-only regression are normally distributed with known standard deviation

$$\sigma = \sqrt{7}$$

Argue that the estimator $\widehat{\beta}_1$ is normally distributed and utilizing the transformation to the standard normal, give an approximate 95% confidence interval for the overall slope β_1.

Problem 5. Menu Combinations

A food truck sells rice bowls. There are 4 bases, 5 proteins, 3 sauces, and 6 toppings. A regular bowl has 1 base, 1 protein, 1 sauce, and 1 topping. A deluxe bowl has 1 base, 2 different proteins, 2 different sauces, and 2 different toppings.

a) How many regular bowls are possible?

b) How many deluxe bowls are possible?

c) Suppose 7 students sit in a row, but two particular students refuse to sit next to each other. How many arrangements are possible?

d) A randomly chosen regular bowl is equally likely among all regular bowls. Let A be the event that base 1 is chosen and B be the event that topping 1 is chosen. Are A and B independent?

Problem 6. Disease Testing

A disease affects 2% of a population. A rapid test is positive with probability 0.96 for someone with the disease. For someone without the disease, the test is falsely positive with probability 0.07. A symptom checker, conditionally independent of the rapid test given disease status, flags someone with probability 0.80 if they have the disease and with probability 0.10 otherwise.

a) Find the probability that both the test and symptom checker are positive.

b) Given that both are positive, find the probability that the person has the disease.

c) Explain why the probability of having the disease given the positive rapid test and the flagged symptom checker result is much larger than the probability from the positive rapid test alone.

Problem 7. Functions of RVs

Let X have probability mass function

x	-2	-1	0	1	2
$p_X(x)$	c	$2c$	$3c$	$2c$	c

a) Find c.

b) Compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and $\mathbb{E}[X^3]$.

c) Let $Y = X^2$. Find the PMF of Y.

d) Are X and $Y = X^2$ independent?

Problem 8. Discrete RVs

A machine produces independent items, each defective with probability $p = 0.08$.

a) Let X be the number of items inspected until the first defective item, counting the defective item. Identify the distribution of X and derive $\mathbb{E}[X]$.

b) Compute $\mathrm{P}(X > 20)$.

c) Given that the first 20 items are not defective, compute the probability that more than 30 items are needed until the first defective item.

d) Let Y be the number of defective items among the first 50 items. Identify the distribution of Y and write $\mathrm{P}(Y \leq 3)$ as a summation.

Problem 9. Triangles

Let X have density

$$f_X(x) = \begin{cases} cx, & 0 \le x \le 1, \\ c(2-x), & 1 < x \le 2, \\ 0, & \text{otherwise} \end{cases}$$

a) Find c.

b) Find the CDF $F_X(x)$.

c) Compute $\mathbb{E}[X]$ without integrating.

d) Now consider random variables X, Y that admit a joint PDF

$$f_{XY} = \begin{cases} 2, & 0 < y < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

Find the marginal densities $f_X(x)$ and $f_Y(y)$.

e) Are X and Y independent?

Problem 10. RV Functions

Let X, Y, Z be independent random variables such that

$$X \sim \text{Uniform}(-1, 1), \qquad Y \sim \text{Uniform}(-1, 1), \qquad \mathbb{E}[Z] = 0, \qquad \text{Var}(Z) = 1$$

Define

$$U = X + Z, \qquad V = Y + 2Z, \qquad R = X^2$$

a) Find the CDF of R.

b) Find the PDF of R.

c) Compute $\mathbb{E}[R]$ and $\text{Var}(R)$.

d) Compute $\text{Cov}(U, V)$.

e) Compute $\text{Var}(U)$, $\text{Var}(V)$, and $\text{Cor}(U, V)$.

f) Are U and V independent?

g) Are R and V independent?

Problem 11. LLN vs CLT

Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = 40$ and $\mathrm{Var}(X_i) = 25$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

a) State the weak law of large numbers for $\overline{X}_n$.

b) Use Chebyshev's inequality to bound $\mathrm{P}(|\overline{X}_{100} - 40| \geq 1)$.

c) Use the CLT to approximate $\mathrm{P}(39 < \overline{X}_{100} < 41)$ in terms of Φ.

d) Explain why the Chebyshev bound and the CLT approximation are not the same kind of statement.

Problem 12. Sample Variance Bias

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Define

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

a) Show that $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$.

b) Give an unbiased estimator of σ^2.

Problem 13. Least Squares

For data (x_i, y_i), fit $y_i \approx \beta_0 + \beta_1 x_i$ by minimizing

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

a) Derive the normal equations.

b) Show that the fitted line passes through $(\overline{x}, \overline{y})$.

c) Give the formula for $\widehat{\beta}_1$.

d) If $\widehat{\beta}_1 = 0$, must $R^2 = 0$?

Problem 14. Cramer Rao

Suppose $X_1, \dots, X_n$ are i.i.d. Exponential(λ) with density $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

a) Find the MLE of λ.

b) What does the LLN imply about $\widehat{\lambda}$?

c) Compute the Fisher information $I_n(\lambda)$.

d) State the Cramer–Rao lower bound for unbiased estimators of λ.

Problem 15. Confidence Interval

A sample of $n = 64$ measurements has sample mean $\overline{x} = 12.4$. The distribution is known to be Gaussian and the population standard deviation is known to be $\sigma = 3.2$.

a) Give a 95% confidence interval for the population mean μ using $z_{0.025} = 1.96$.

b) What changes if σ is unknown and estimated by the sample standard deviation s?

c) Interpret the 95% t-confidence level.