Data 102, Spring 2025
Midterm 1

- You have **110 minutes** to complete this exam. There are **7 questions**, totaling **50 points**.
- You may use **one** 8.5×11 sheet of handwritten notes (front and back), and the provided reference sheet. No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down something so we can give you partial credit.
- We have provided a blank page of scratch paper at the **beginning** of the exam. No work on this page will be graded.
- You may, without proof, use theorems and facts given in the discussions or lectures, **but please cite them**.
- We don't answer questions individually. If you believe something is unclear, bring your question to us and if we find your question valid we will make a note to the whole class.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Berkeley email	
Name of person to your left	
Name of person to your right	

Honor Code [1 pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ____

This page has been intentionally left blank. No work on this page will be graded.

1 True or False [5 Pts]

For each of the following, determine whether the statement is true or false. For this question, no work will be graded and no partial credit will be assigned.

(a) [1 Pt] True or False: Sampling methods like Markov Chain Monte Carlo (MCMC) can be used to approximate any of the following: posterior means, posterior modes, and posterior probabilities.

○ True ○ False

(b) [1 Pt] True or False: When using Gibbs sampling, the distribution of the samples is guaranteed to converge to the true prior distribution as the number of iterations becomes very large.

○ True ○ False

(c) [1 Pt] True or False: MCMC methods (as discussed in Data 102) are usually more efficient than rejection sampling because we keep more samples.

○ True ○ False

(d) [1 Pt] True or False: When computing the frequentist risk, we average over all possible values of the observed data by using the likelihood.

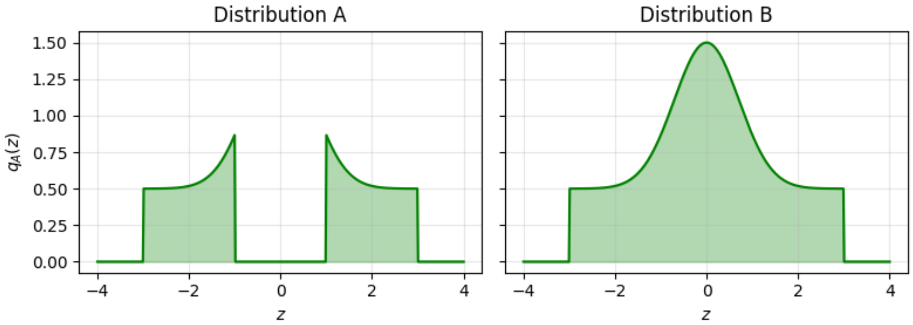
○ True ○ False

(e) [1 Pt] In a Beta-binomial hierarchical model (like the kidney cancer example from lecture), a Bayesian approach is most helpful when some of the groups are relatively small, because the prior helps reduce the variance of the estimates for small groups.

○ True ○ False

2 Sampling [4 Pts]

We use rejection sampling to approximate two distributions over z shown below. Assume that any values of z not shown in the plot(s) have zero density.



For all the questions below, assume we chose any parameters optimally to get the largest possible proportion of accepted samples.

(a) [2 Pts] Which of the following are valid proposal distributions when using rejection sampling to approximate Distribution A (left)? Select all answers that apply **by filling in the square next to each correct answer**.

□ $z \sim \text{Uniform}(-4, 4)$

□ $z \sim \text{Uniform}(-3, 3)$

□ $z \sim \text{Uniform}(-2, 2)$

□ $z \sim N(0, 5)$

(b) [2 Pts] Suppose we use rejection sampling to approximate each distribution (A and B), choosing a correct proposal distribution from the choices above, and using the same proposal distribution for both A and B.

For which distribution will we obtain a higher proportion of accepted samples? Choose the single best answer **by filling in the circle next to it.** Explain your answer in two sentences or less. **You must explain your answer to receive credit.**

○ A

○ B

○ They will be the same

Explanation:

3 Binary Decisions Potpourri [8 Pts]

(a) [3 Pts] Fill in the blanks in the statements below. Each blank should contain one term from the following list: **specificity, sensitivity, precision, FDP**.

Hint: Precision = 1 - FDP.

(i) The ____________of a test is determined by its p-value threshold (and by nothing else).

(ii) When using Neyman-Pearson, we fix a desired level of ____________, and use a particular test statistic to achieve the best possible ____________.

(iii) If the row-wise rates are fixed and known (and not equal to 0 or 1), then an increase in prevalence will increase the ____________.

(b) [2 Pts] Below is an incomplete expression that is supposed to compute the **Bayesian posterior risk** under the ℓ_2 (squared error) loss for a decision $\delta(x)$ (computed from data x) and an unknown variable θ. Fill in the four blanks below to complete the expression:

$$\int \left(\delta(x)\underline{\qquad}\right)^{\underline{\qquad}} p\left(\underline{\qquad}\right) d\underline{\qquad}$$

(c) [3 Pts] Consider a set of n unique p-values, where the smallest is 0 and the largest is 1. Using these p-values, define the following quantities:

- N_1 is the number of discoveries made using naive thresholding with significance α
- N_2 is the number of discoveries made using Bonferroni correction with FWER α
- N_3 is the number of discoveries made using Benjamini-Hochberg with FDR α

For each of the blanks below, use one of $=, \leq, <, \geq, >$ to describe the relationship. Choose the single best answer **by filling in the circle next to it.** If multiple options are possible, choose "Cannot be determined".

(i) N_1__N_2

○ $=$ ○ $<$ ○ $\leq$ ○ $>$ ○ $\geq$ ○ Cannot be determined

(ii) N_2__N_3

○ $=$ ○ $<$ ○ $\leq$ ○ $>$ ○ $\geq$ ○ Cannot be determined

(iii) N_1__N_3

○ $=$ ○ $<$ ○ $\leq$ ○ $>$ ○ $\geq$ ○ Cannot be determined

4 Did ChatGPT Write This Question? (10 pts)

A professor is grading student essays for her class of $n = 80$ students, and wants to determine whether they were written by ChatGPT. She identifies ten words that are used much more often in ChatGPT responses than in typical student essays and calls these "LLM words". She counts the number of times LLM words occur in each student's essay, w_i for $i = 1, \ldots, n$. She assumes that $w_i \sim \text{Poisson}(\lambda)$, and defines two hypotheses for each essay:

- H_0: This essay was written by the student ($\lambda = 20$).
- H_1: This essay was written by ChatGPT ($\lambda > 20$).

(a) [2 Pts] The first student's essay has a p-value of 0.04. If the professor uses a p-value threshold of 0.1, which of the following must be true? Select all answers that apply **by filling in the square next to each correct answer**.

□ Given that H_0 is true, the conditional probability that the first student's essay contains w_1 or more "LLM words" is 0.04.

□ The probability of the professor's test correctly classifying any student-written essay is 0.9.

□ The first student's essay must have been written by the student.

□ The first student's essay must have been written by ChatGPT.

(b) [2 Pts] Suppose for this part only that 50% of students turn in essays written by ChatGPT, and that the false omission proportion (FOP) is 0.25. The professor continues to use a p-value threshold of 0.1.

Fill in the entries of the confusion matrix below with the *expected* number of FP, TP, FN, and TN events, assuming her hypotheses are correctly specified and the test is well-designed to rule out other possibilities. If you do not have enough information to fill in an entry, draw an X through it.

	D=0	D=1
R=0		
R=1		

(c) [2 Pts] For this part only, the professor decides to control for family-wise error rate (FWER). Which of the following facts, if true, would support this decision? Select all answers that apply **by filling in the square next to each correct answer**.

- □ If the professor rejects the null hypothesis for any student, she immediately gives that student an F in the course for academic misconduct.
- □ The professor is willing to accept a relatively low power for her test.
- □ The number of students doubles to $n = 160$.

(d) [2 Pts] The professor wants to calculate the power of her test for any particular threshold value for w_i that she chooses. Rewrite either the null or the alternative hypothesis so that this is possible, or explain in one sentence why neither needs to be rewritten.

- ○ Rewrite null
- ○ Rewrite alternative
- ○ Rewrite neither

New hypothesis (or explanation):

(e) [2 Pts] Suppose for this part only that $n = 5$, and the professor obtains the following p-values for the five students: 0.04, 0.01, 0.7, 0.0002, 0.027. She uses the Benjamini-Hochberg procedure with FDR α, and rejects the null for two out of the five days. What values of α are possible for her to have used? Specify an interval or intervals, e.g., "$\alpha \in [0, 0.1)$ or $\alpha \in [0.7, 0.9)$", or if this is impossible, explain why. **You must show your work to receive credit.**

5 The One With A New Beetle (7 pts)

Cindy travels to Greenland and discovers a new species of beetle. She isn't sure whether it belongs to a more common genus, where each beetle's antennae are the same size, or a less common genus, where the left one is larger. So, she collects n beetles and for each one, measures the difference in length between the left and right antennae in mm. She calls this measurement x_i, so that her data are $x_1, \dots, x_n$.

She formulates two hypotheses:

- H_0: the beetles belong to the more common genus, with $x_i \sim \mathcal{N}(0, 1)$
- H_1: the beetles belong to the less common genus, with $x_i \sim \mathcal{N}(1, 1)$

(a) [2 Pts] Show that the likelihood ratio is

$$LR = \exp\left(\sum_{i=1}^{n} x_i - \frac{n}{2}\right).$$

For this question, you should follow the convention used in lecture, where the null likelihood is in the denominator.

For the remainder of the question, you should assume that:

- $n = 9$
- Cindy calculates the sample average $t = \sum_i \frac{x_i}{n}$ and uses this as her test statistic.
- Cindy's decision rule is: she will reject the null hypothesis if the sample average t is greater than or equal to 0.5 mm.

(b) [2 Pts] Cindy's decision rule corresponds to rejecting the null hypothesis if the alternative hypothesis is k or more time(s) more likely than the null hypothesis. What is the value of k? You must show your work to earn credit.

Hint: This question can be answered with very little computation.

(c) [3 Pts] What is the power of the test with Cindy's decision rule? You should express your answer either as an integral (you do not need to solve it), or in terms of the standard normal CDF Φ, where $\Phi(z)$ represents the probability that a standard normal random variable (i.e., $\mathcal{N}(0, 1)$) is less than or equal to z.

Hint: Start by determining the appropriate distribution of t.

Any work below this line will not be graded.

6 Community Television [10 Pts]

Troy watches N old TV shows, but he has a short attention span. For any show i, let T_i be the number of episodes he watches before stopping. He assumes T_i is geometrically distributed: $T_i \overset{iid}{\sim} \text{Geometric}(q)$, where q is the probability that, after watching any episode, he stops watching that show.

$$\mathbb{P}(T_i = t) = (1-q)^{t-1}q, \qquad t = 1, 2, \ldots$$

You may assume that the shows are long enough (and q is large enough) that Troy never reaches the end of any show he watches.

You may use (without proof) the fact that the Beta distribution is the conjugate prior for the Geometric likelihood: in other words, if $q \sim \text{Beta}(a, b)$, then

$$q|T_1, \ldots, T_N \sim \text{Beta}\left(a + N, b - N + \sum_{i=1}^{N} T_i\right)$$

(a) [2 Pts] For each of the following descriptions, determine whether it applies to the MAP, MLE, or neither. Choose the single best answer **by filling in the circle next to it.**

(i) The most likely estimate given the observed dataset

○ MAP ○ MLE ○ Neither

(ii) The estimate that makes the observed dataset most likely

○ MAP ○ MLE ○ Neither

(iv) The estimate that minimizes the average 0-1 loss, where the average is taken over all possible q given how many episodes he watched of each show.

○ MAP ○ MLE ○ Neither

(b) [2 Pts] For each of the following probability, determine which distribution should be used to calculate it. Choose the single best answer **by filling in the circle next to it.**

(i) The probability of watching only one episode of each show if Troy has a 90% chance of giving up on a show after any episode

○ Likelihood ○ Prior ○ Posterior

(ii) Given that Troy stopped after one episode for 10 different shows, the conditional probability that q is less than 0.1

○ Likelihood ○ Prior ○ Posterior

(iii) If Troy tries six shows, the probability that he watches the exact same number of episodes for all six, as a function of q

○ Likelihood ○ Prior ○ Posterior

(iv) Troy's best guess for his average stopping probability, before starting any shows

○ Likelihood ○ Prior ○ Posterior

For the rest of this question, assume that Troy watches four shows. For the first three shows, he records $T_1 = T_2 = 4$ and $T_3 = 7$. He decides to use a Beta$(2, 4)$ prior for q.

(c) [3 Pts] For this part only, he finds that using MAP and MLE estimation both return the same result. How many episodes of the fourth show did he watch? Provide a number or explain why this is impossible. You must show your work to earn credit.

Hint: You may use (without proof) the fact that, given i.i.d. observations $x_1, \ldots, x_m$ *of a geometric random variable with parameter q, the MLE for* q *is* $\frac{m}{\sum_{i=1}^{m} x_i}$.

(d) [1 Pt] For this part only, suppose $T_4 = 7$. Given all four observations and the prior above, what is the probability that q is greater than 0.5? Express your answer as an integral: you do not need to simplify.

(e) [2 Pts] Given the information above, which of the following scenarios is guaranteed to increase Troy's MAP estimate for q? Select all answers that apply **by filling in the square next to each correct answer**.

□ Watching more episodes of the fourth show

□ Using a stronger prior with the same mean

□ Using a Beta$(4, 2)$ prior instead of a Beta$(2, 4)$ prior

□ Watching five additional shows (for a total of nine shows), and stopping after only one episode for each of the last five shows

7 Berkeley Landlord [6 Pts]

Laura manages apartment buildings in Berkeley, and wants to better understand how to manage maintenance requests from her tenants. She gathers data from each tenant across all her buildings for last year, and defines the following variables:

- r_{ij} is the number of maintenance requests last year for tenant i in building j.
- λ_j is the average number of requests per tenant last year in building j.
- α/β is the average number of requests for all tenants across all her buildings last year.

She assumes that $r_{ij}|\lambda_j \sim \mathrm{Poisson}(\lambda_j)$.

(a) [2 Pts] For this part only, assume she has two buildings with three tenants each. Draw a graphical model for this model, **assuming that the λ_j are fixed**. Your model should include all relevant variables. **If you shade in any nodes, please only shade the border of the node so that we can read what's inside while grading.**

For the remainder of this question, she decides to treat each λ_j as a random variable, and uses a Gamma prior: $\lambda_j|\beta \sim \mathrm{Gamma}(\alpha, \beta)$

(b) [2 Pts] *Questioning assumptions:* In two sentences or less of plain English, describe one assumption Laura is making by using this model for r and λ, and why the assumption might be wrong.

(c) [2 Pts] Suppose that building 1 has five tenants. Compute the posterior distribution for λ_1 in terms of α, β, and the data $r_{11}, \dots, r_{51}$, or explain why we need to use approximate inference. If you can compute the posterior distribution, you should express your answer as a known distribution. For example, you might answer $\lambda_1 \sim \mathcal{N}\left(\sum_i r_{i1}\ ,\ \alpha/\beta\right)$ (note that this is not the correct answer).

Any work below this line will not be graded.

8 Congratulations [0 Pts]

Congratulations! You have completed Midterm 1.

- **Make sure that you have written your student ID number on *every other page* of the exam.** You may lose points on pages where you have not done so.
- Also ensure that you have **signed the Honor Code** on the cover page of the exam for 1 point.
- If more than 10 minutes remain in the exam period, you may hand in your paper and leave. If $\leq$ 10 minutes remain, please **sit quietly** until the exam concludes.

[Optional, 0 pts] Draw a picture or cartoon that's related to your favorite thing you've learned in Data 102 so far.

Midterm 1 Reference Sheet

Useful Distributions:

Distribution	Support	PDF/PMF	Mean	Variance	Mode
$X \sim \text{Poisson}(\lambda)$	$x = 0, 1, 2, \ldots$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$\lfloor \lambda \rfloor$
$X \sim \text{Binomial}(n, p)$	$x \in \{0, 1, \ldots, n\}$	$\binom{n}{x} p^x (1-p)^{1-x}$	np	$np(1-p)$	$\lfloor (n+1)p \rfloor$
$X \sim \text{Beta}(\alpha, \beta)$	$0 \le x \le 1$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}$	$\frac{\alpha-1}{\alpha+\beta-2}$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \ge 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2	μ
$X \sim \text{Exponential}(\lambda)$	$x \ge 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0
$X \sim \text{InverseGamma}(\alpha, \beta)$	$x \ge 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta}{\alpha+1}$

Conjugate Priors: For observations x_i, $i = 1, \ldots, n$:

Likelihood	Prior	Posterior
$x_i \vert \theta \sim \text{Bernoulli}(\theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \vert x_{1:n} \sim \text{Beta}\left(\alpha + \sum_i x_i, \beta + \sum_i (1 - x_i)\right)$
$x_i \vert \mu \sim \mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(\mu_0, 1)$	$\mu \vert x_{1:n} \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+n}\left(\mu_0 + \frac{1}{\sigma^2}\sum_i x_i\right), \frac{\sigma^2}{\sigma^2+n}\right)$
$x_i \vert \lambda \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \vert x_{1:n} \sim \text{Gamma}\left(\alpha + n, \beta + \sum_i x_i\right)$
$x_i \vert \lambda \sim \text{Poisson}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \vert x_{1:n} \sim \text{Gamma}\left(\alpha + \sum_i x_i, \beta + n\right)$
$x_i \vert \lambda \sim \mathcal{N}(\mu, \sigma^2)$	$\sigma \sim \text{InverseGamma}(\alpha, \beta)$	$\sigma \vert x_{1:n} \sim \text{InverseGamma}\left(\alpha + n/2, \beta + \left(\sum_{i=1}^n (x_i - \mu)^2\right)/2\right)$

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Logistic	sigmoid	Bernoulli
Poisson	exponential	Poisson
Negative binomial	exponential	Negative binomial

Some powers of e:

x	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$y = e^x$	1.05	1.11	1.22	1.35	1.49	1.65	1.82	2.01	2.23	2.46	2.72