Data 102 Fall 2026 **Midterm 1** **A. Strang**

NAME (FIRST LAST): _____ **SID:** _____

Room: ◯ Pimentel 1, ◯ Wheeler 20, ◯ Mulford 106, ◯ Mulford 107

Seat Number: _____

TIME: 110 minutes

Conditions: closed book/internet; no reference sheet; no calculator/computer

GRADING

- There are 9 questions and 60 points total.
- The exam will be graded for accuracy. Partial credit will be granted.
- Please mark all selection bubbles clearly. Fill in the bubble fully. Do not use x's or check marks.

YOUR WRITTEN WORK

- There is a space for your full name and SID number on each page. **You must fill this in.** It will ensure that we can identify your work during the scanning process.
- If you need scratch paper, please use the backs of the pages of the exam. Note: **only the front pages will be scanned and graded.**
- When working on short-answer problems, clearly indicate your answers, and justify your answers. We will give partial credit based on demonstrated knowledge. Make sure your work is organized and legible. We will not give partial credit for work we cannot follow.
- When working on short-answer problems from practice problems you must justify your answers as when solving the problems from scratch. Correct answers without justification will not receive full credit.
- Numerical answers may be left as mathematical expressions (need not be fully simplified) unless we explicitly ask for simplification. For example, $(6 \times 10 + 15 \times 3)/3$.

HONOR CODE

Exams are individual work. By signing on the line below you verify that all answers are your own and that you will not share the content of exams until answers are released. Unsigned exams may not be graded. **Please abide by UC Berkeley's Honor Code:**

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Your signature: _____

Full Name and SID: _____

Part 1: *Answer all parts of questions 1 - 5.*

Question 1 [6 points]. Consider a binary decision making problem. Let $R \in \{0, 1\}$ denote the possible states of reality and $D \in \{0, 1\}$ denote the decision returned by a decision making tool.

1. Define each error rate listed below as the conditional probability of an event:

   (a) The false positive rate equals Pr(_____|_____).

   (b) The false negative rate equals Pr(_____|_____).

   (c) The false discovery proportion equals Pr(_____|_____).

2. Select all true statements below:

   ☐ a decision procedure is sensitive if it has a false positive rate near zero

   ☐ a decision procedure is specific if it has a true negative rate near one

   ☐ a decision procedure with a small false positive rate must also have a small false discovery proportion

Question 2 [10 points]. In the space below, explain how to apply a null hypothesis statistical test. Let x represent the observed data, and X hypothetical data. Use H_0 to denote the null hypothesis and H_1 the alternative. Let $s(x)$ denote the observed value of the test statistic and $S = s(X)$ denote hypothetical values. Let $p(s)$ denote the p-value given the observed value of the test statistic, s. Let α denote the level of the test.
Your explanation must:

- clearly define the p value as a probability
- clearly explain the conditions under which we reject the null, or fail to reject the null.

Procedure:

Guarantee: This procedure controls the (◯ *false negative rate*, ◯ *false positive rate*, ◯ *false discovery proportion*) in the sense:

Pr(_____) ≤ _____.

Full Name and SID: _____

Question 3 [5 points]. Consider two simple hypotheses:

$$H_0 : X \sim f_0, \qquad H_1 : X \sim f_1$$

where f_0 and f_1 are different probability density functions.
In the space below, state the Neyman-Pearson lemma. Use the same notation from question 2 for the observed data, hypothetical data, test statistic, etc. Your statement must:

- Clearly define the test statistic suggested by the lemma in terms of f_0 and f_1.
- Clearly describe the rationale for using this statistic (e.g. what can we guarantee about the error rates?)

Answer: Any test that adopts the likelihood ratio (or a monotonic injective transformation of the likelihood ratio),

Question 4 [4 points]. Explain each error rate listed below in plain English, then suggest a multiple testing procedure that would control it. If you use a symbol in your definition, you must define it (e.g. $n_{0,0}$ is the number of true negative decisions).

- **False Discovery Rate:**

  To control the FDR, we should adopt the (◯ *Bonferroni*, ◯ *Benjamini-Hochberg*) procedure.

- **Family-Wise Error Rate:**

  To control the FWER, we should adopt the (◯ *Bonferroni*, ◯ *Benjamini-Hochberg*) procedure.

Full Name and SID: _____

Question 5 [5 points]. Let Θ equal a randomly sampled unknown, X randomly sampled data, and $d(\cdot)$ a decision function that accepts data and returns a decision. Let $\ell(d(x), \theta)$ denote the loss associated with the decision $d(x)$ if $\Theta = \theta$. Then, define each notion of risk below:

- **Frequentist Risk:** $R_F(_____) =$ _____
- **Posterior Risk:** $R_p(_____) =$ _____
- **Bayes Risk:** $R_B =$ _____

Then provide two equations, one relating Bayes risk to the Frequentist risk, and one relating the Bayes risk to the Posterior risk.

$R_B =$ _____ $=$ _____.

Full Name and SID: _____

Part 2: *Complete 2 of the 4 problems. Indicate which problems you want graded below.*

Please Grade: ◯ Question 6, ◯ Question 7, ◯ Question 8, ◯ Question 9

If you fill in more than two bubbles, we will grade the first two problems you mark for grading.

Question 6 [15 points]. Suppose that $T \sim \text{Exponential}(\lambda)$ and wish to test

$$H_0 : \lambda = c \quad \text{vs} \quad H_1 : \lambda = b \times c$$

for some $b > 1$. Recall that, if $Y \sim \text{Exponential}(\mu)$ then $Y \geq 0$ with PDF and CDF:

$$f_Y(y) = \mu e^{-\mu y}, \quad F_Y(y) = 1 - e^{-\mu y}$$

(a) Compute the likelihood ratio, $\text{LR}(t)$, explicitly in terms of b and c.

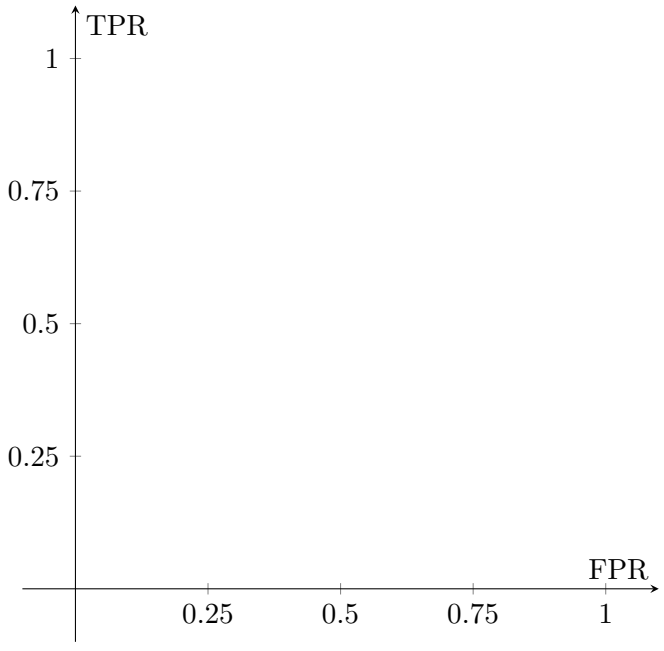
(b) Let $S = s(T) = \text{LR}(T)$. Let α be our false positive rate ($0 < \alpha < 1$). Compute the value of the threshold $s_*(\alpha)$ so that the FPR of the test is equal to α if we reject the null when $S \geq s_*(\alpha)$.

Full Name and SID: _____

(c) What is the true positive rate of this test? Your answer should be expressed in terms of b, c, and α.

(d) On the axes below, carefully draw a plot showing the TPR of the test as a function of desired FPR, α, for $b = 1$, $b = 2$, and $b = 10$. Clearly label each curve with the corresponding value of b. To receive full credit, it must be possible to correctly answer each question listed below using your plot (you do not need to answer the questions in writing, just provide sufficient illustration):

1. Is the TPR a monotonic function of the FPR? If so, is it increasing or decreasing in the FPR?
2. Is the TPR a concave or convex function of the FPR?
3. What is the TPR when $\alpha = 0$ and when $\alpha = 1$?
4. Holding α fixed, how does the TPR depend on b (increasing, decreasing, constant)?



Full Name and SID: _____

Question 7 [15 points]. Suppose that we can observe $\{X_i\}_{i=1}^n$, drawn i.i.d. from either f_0 or f_1. To determine which distribution produced our data, we will run a hypothesis test. Treat the hypothesis, $X_i \sim f_0$ as the null.

(a) Show that the log-likelihood ratio is: $\log(\text{LR}(\vec{x})) = \sum_{i=1}^n \log(\text{LR}(x_i))$

(b) Consider the test statistic, $s(\vec{x}) = \frac{1}{n}\log(\text{LR}(\vec{x}))$. Is large s evidence in favor of, or against, the null? Justify your answer.

(c) Show that a test that decides for or against the null by comparing s to a threshold is maximally sensitive for any controlled significance.

Full Name and SID: _____

(d) Let $R_i = \log(\text{LR}(X_i))$, $\mathbb{E}[R_i|H_0] = \mu_0$, and $\text{SD}[R_i|H_0] = \sigma_0$. Assume that all of the expected values and standard deviations are finite. Then, show that, under the null, $S = s(\vec{X})$ converges in probability to μ_0, with deviations on the scale $\sigma_0/\sqrt{n}$, and is asymptotically Gaussian, in the limit as $n \to \infty$.

(e) Let $s_*(\alpha)$ denote the decision threshold we compare S to when testing the null hypothesis f_0 at significance level α. Using part (d), identify the limiting value of $s_*(\alpha)$ and whether $s_*(\alpha)$ approaches its limit from above or below. Justify your answer.

Full Name and SID: _____

Question 8 [15 points]. Suppose that you have census data from a random sample of the US adult population. Your data includes two numerical fields: `Age` and `Hours worked per week`. It also includes five categorical fields: Gender, Education (post high school, post college), Marriage status and whether the person's income is greater than $50,000. We will use this dataset to test the hypotheses of whether each of the categorical fields have any effect on the expectation of the numerical fields.

(a) Consider the following hypothesis:

H_0 : There is no difference in the average value of `numerical_col` between the two groups specified in `binary_col`.

H_1 : The average value of `numerical_col` is different for the two groups specified in `binary_col`.

Explain, in pseudocode, how you would estimate p-values under hypothesis H_0 for any pair of numerical and categorical fields.

Full Name and SID: _____

(b) Suppose that, after 10,000 simulations, you receive a table of ten p-values, one for each possible combination of categorical and numerical columns:

		1	2
		Age	Hours per week
a	Post HS?	0.7	0.9
b	Post College?	0.4	0.15
c	Ever Married?	0	0.016
d	Gender	0.06	0.018
e	50K?	0.001	0.08

In your following answers, you may use H_{a1} to mean the "Post HS?" and "Age" hypothesis, H_{a2} to mean the "Post HS?" and "Hours per week" hypothesis, etc.

(i) Suppose we want to guarantee a Family-wise Error Rate (FWER) of 0.05. Given the p-values provided above, for which tests do we reject the null hypothesis?

(ii) Suppose we want to guarantee a False Discovery Rate (FDR) of 0.05. Given the p-values provided above, for which tests do we reject the null hypothesis?

(c) How do the results from (b.i) and (b.ii) compare? Explain how and why these results are different.

Full Name and SID: _____

Question 9 [15 points]. Consider a medical diagnostic decision. For each person that undergoes testing, the null hypothesis is that they don't have the virus, and the alternative is that they do.

When making medical decisions, wrong diagnoses carry different costs. Deciding that a patient does not have the virus when in reality they do can have severe consequences: lack of treatment, risk of infecting others, and even premature death.

On the other hand, deciding that a patient has the virus when in reality they don't has consequences of that include distress, unnecessary treatment and costs of subsequent testing. These errors are typically considered less severe.

(a) Let's assume that we estimate the cost of making a false negative mistake to be k-times larger than the cost of a false positive. We can express that via the loss function shown below:

$$\begin{cases} \ell(D = 0 \mid R = 0) = \ell(D = 1 | R = 1) = 0 \\ \ell(D = 1 \mid R = 0) = 1 \\ \ell(D = 0 \mid R = 1) = _____ \end{cases}$$

Fill in the missing value above.

(b) Suppose that you run the test on 100 patients and observe the following outcomes:

	$D = 0$	$D = 1$
$R = 0$	84	12
$R = 1$	1	3

Compute the average loss incurred over these 100 patients.

(c) Suppose that you can implement a test that, for a false positive rate α, achieves a true positive rate $\beta = 1 - (\alpha - 1)^2$. Suppose that the prevalence of the disease is 0.04. In the space below, express the Bayes risk as a function of $\alpha \in [0, 1]$ and k.

(d) Let $\alpha_*(k)$ denote the choice of α that minimizes the Bayes risk as a function of k. (Note: you should be able to answer each question below without solving for α_* explicitly. If you want to solve for α_*, use the back of this page.)

(i) If k increases, then $\alpha_*(k)$ (◯ *increases*, ◯ *decreases*)

(ii) If the prevalence of the disease increases, then $\alpha_*(k)$ (◯ *increases*, ◯ *decreases*)

Full Name and SID: _____

(e) Given a prevalence of 0.04, write, find, as a function of α, the posterior probability that a patient has the disease if they test positive, $\Pr(R = 1|D = 1)$.

(f) Given that the test returns a positive result, a doctor can decide to start treatment ($T = 1$), or not ($T = 0$). Assume that $\ell(T|R) = \ell(D|R)$. Find the posterior risk of treating, or not treating, the patient, given that the test was positive:

- Posterior risk if untreated:

- Posterior risk if treated:

(g) Suppose that $\alpha = 0.5$. Given that a patient tests positive, should the doctor recommend treatment?