GROUND RULES:

- This exam contains 10 questions; each question is worth 10 points. The maximum number of points on this exam is 100.
- Print your name **at the top of this page in the upper right hand corner**.
- This is a closed-book and closed-notes exam. You may use a calculator if you wish, but SHOW ALL OF YOUR WORK AND EXPLAIN ALL OF YOUR REASONING!!!
- Any discussion or otherwise inappropriate communication between examinees, as well as the appearance of any unnecessary material, will be dealt with severely.
- You have 3 hours to complete this exam. GOOD LUCK!

HONOR PLEDGE FOR THIS EXAM:

After you have finished the exam, please read the following statement and sign your name below it.

> *I promise that I did not discuss any aspect of this exam with anyone other than the instructor, that I neither gave nor received any unauthorized assistance on this exam, and that the work presented herein is entirely my own.*

1. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y;\theta) = \begin{cases} \theta^2 y e^{-\theta y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that the uniformly most powerful (UMP) level α test of

$$H_0 : \theta = 1$$
$$\text{versus}$$
$$H_a : \theta > 1$$

has a rejection region of the form $\text{RR} = \{\boldsymbol{y} : \sum_{i=1}^n y_i < c\}$. Find an expression for c.

2. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson(θ) observations. Show that the level α likelihood ratio test (LRT) of

$$H_0 : \theta = 1$$
versus
$$H_a : \theta \neq 1$$

has a rejection region of the form RR $= \{\boldsymbol{y} : \overline{y} < c_1 \text{ or } \overline{y} > c_2\}$. Don't worry about finding expressions for c_1 and c_2.

3. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. The x_i's are fixed constants. In the notes, we proved that

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

where $c_{11} = 1/\sum_{i=1}^{n}(x_i - \overline{x})^2$, and that

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

is an unbiased estimator of σ^2. We also stated the following two facts (without proof):

- The random variable
$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$
- The statistic $\widehat{\sigma}^2$ is independent of $\widehat{\beta}_1$.

Using this information, derive a $100(1-\alpha)$ percent confidence interval for β_1. Don't just state the answer; your derivation should be thorough. *Hint:* Start by letting Z denote the standardized value of $\widehat{\beta}_1$. Then, find a function of Z and W that has a $t(n-2)$ distribution. Argue that this function is a pivot and then you are off to the races.

4. Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the (fixed) $n \times p$ design matrix and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Let $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. Note that $\mathbf{M}$ is a fixed matrix (i.e., it is not random). Recall that the vector of fitted values is $\widehat{\mathbf{Y}} = \mathbf{M}\mathbf{Y}$.

(a) Show that $E(\widehat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}$.
(b) Show that $V(\widehat{\mathbf{Y}}) = \sigma^2\mathbf{M}$.
(c) Argue that $\widehat{\mathbf{Y}}$ has a multivariate normal distribution. What is the dimension of this distribution?

5. A researcher was studying the relationship between hours of sleep per day and brain weight, body weight, and gestation time. The sample included $n = 14$ subjects. The variables of interest were Y, x_1, x_2, and x_3, where

$$\begin{aligned} Y &= \text{hours of sleep per day } (\texttt{SLEEP}) \\ x_1 &= \text{gestation time } (\texttt{GEST}) \\ x_2 &= \text{brain weight } (\texttt{BRAIN}) \\ x_3 &= \text{body weight } (\texttt{BODY}). \end{aligned}$$

Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_1 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 14$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Fitting this model to the observed data produced the following information:

General Linear Model Procedure

Source	DF	Sum of Squares
Model (Corrected)	3	7.52
Error	10	6.36
Total (Corrected)	13	13.88

(a) Perform a hypothesis test to determine if at least one of the independent variables is important in describing SLEEP. Use $\alpha = 0.05$. State your conclusion.
(b) Your colleague believes that BRAIN (x_2) and BODY (x_3) are not important in the regression. She therefore fits the simple linear regression model

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \epsilon_i,$$

for $i = 1, 2, ..., 14$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Her ANOVA table produces a (corrected) model sum of squares SSR = 5.77. Test whether the simple linear regression model fits the data as well as the multiple linear regression model. Use $\alpha = 0.05$.

6. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(0, \theta)$ distribution, where $\theta > 0$. We would like to do a Bayesian analysis with a Jeffreys prior. Recall that Jeffreys' Principle says to choose the prior $g(\theta) \propto [J(\theta)]^{1/2}$, where

$$J(\theta) = -E\left[\frac{\partial^2 \ln f_Y(Y;\theta)}{\partial \theta^2}\right]$$

is the Fisher information and $f_Y(y;\theta)$ denotes the conditional pdf of Y, given θ.

(a) Find Jeffreys' prior $g(\theta)$. It suffices to say what $g(\theta)$ is proportional to.
(b) Is your prior in part (a) proper or improper? Explain.

7. *Should marijuana be legalized?* Suppose that you ask this question to $n = 100$ USC incoming freshmen. For convenience, treat their (yes/no) responses $Y_1, Y_2, ..., Y_{100}$ as an iid Bernoulli(θ) sample, where θ is the (population) proportion of USC incoming freshmen who are in favor of legalizing marijuana. The parameter θ is best regarded as random and is taken to have a beta(α, β) prior distribution.

(a) Pick a specific prior distribution for θ; that is, specify values of α and β. Defend your choices (even if you go noninformative). Do not let your prior choice be influenced by the information in part (b).
(b) Suppose that a total of 61 students you asked are in favor of legalizing marijuana. With your prior choice in part (a), derive the posterior distribution of θ.
(c) Find the posterior mean $\widehat{\theta}_B = E(\theta|\boldsymbol{Y} = \boldsymbol{y})$.

8. Conditional on $\theta > 0$, suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta$(\theta, 1)$ distribution so that the common pdf is

$$f_{Y|\theta}(y|\theta) = \begin{cases} \theta y^{\theta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

In turn, suppose that θ varies according to a gamma(a, b) prior distribution; i.e.,

$$\theta \sim g(\theta) = \frac{1}{\Gamma(a)b^a}\theta^{a-1}e^{-\theta/b},$$

for $\theta > 0$, where a and b are both known constants larger than zero.

(a) Show that the posterior distribution $g(\theta|\boldsymbol{y})$ is a gamma pdf with shape parameter $\alpha^* = n + a$ and scale parameter $\beta^* = b/(1 - bu)$, where $u = \sum_{i=1}^n \ln y_i$.
(b) Show that a $100(1 - \alpha)$ percent credible set for θ is

$$\left(\frac{b\chi^2_{2(n+a),1-\alpha/2}}{2(1 - bu)}, \frac{b\chi^2_{2(n+a),\alpha/2}}{2(1 - bu)} \right),$$

where $\chi^2_{\nu,\gamma}$ denotes the upper γ quantile of the χ^2 distribution with ν degrees of freedom. *Note:* If you can not show this explicitly, explain to me (in words or using equations) how a credible set would be constructed.

9. For a critical jet engine part, an engineer models (parametrically) the time to failure T using a Rayleigh distribution with pdf

$$f_T(t) = \begin{cases} \frac{2t}{\theta} e^{-t^2/\theta}, & t > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where the parameter $\theta > 0$.

(a) Show that the survivor function of T is

$$S_T(t) = \begin{cases} 1, & t \leq 0 \\ e^{-t^2/\theta}, & t > 0. \end{cases}$$

(b) Show that the hazard rate (function) is given by

$$\lambda_T(t) = \frac{2t}{\theta}.$$

What does this suggest about the rate of failure as a function of time?
(c) The engineer observes an iid sample of n parts and monitors each of them until failure (that is, there are no censored observations). He then estimates $S_T(t)$ in two ways:

- He computes $\widehat{\theta}$, the MLE of θ under the Rayleigh model assumption, and then estimates $S_T(t)$ using $\widehat{S}_T(t) = e^{-t^2/\widehat{\theta}}$.
- He uses the Kaplan-Meier estimate of $S_T(t)$.

The engineer asks for your advice on which estimate to use. What would you tell him?

10. In class, we discussed a clinical trial involving $n = 64$ cancer patients with severe aplastic anemia. Patients were assigned to one of two treatment groups:

- Treatment 1: Cyclosporine and methotrexate (CSP+MTX)
- Treatment 2: Methotrexate only (MTX).

There were 32 patients assigned to each group. The endpoint of interest was

$$T = \text{ time to diagnosis of acute graft versus host disease (AGVHD),}$$

although, as you may remember, some individuals were censored. In class, we computed Kaplan-Meier estimates of the survival functions (see Figure 1 on the next page). For this exam, I used R to produce the logrank test output (see `fit.2` output on the next page).

(a) In words, describe what it means for a patient to have a censored observation.
(b) For the CSP+MTX group, how many patients had observed (not censored) values of T larger than 600 days? Explain how you arrived at this answer.
(c) For the MTX group, how many patients had censored times larger than 600 days? Explain how you arrived at this answer.
(d) Denote by $S_1(t)$ and $S_2(t)$ the population survivor functions for treatment groups 1 and 2, respectively. Using the `fit.2` output, find the logrank test statistic T_{LR} to test

$$H_0 : S_1(t) = S_2(t)$$
$$\text{versus}$$
$$H_a : S_1(t) > S_2(t).$$

State the level $\alpha = 0.10$ rejection region and your conclusion (in plain English; no statistical jargon).

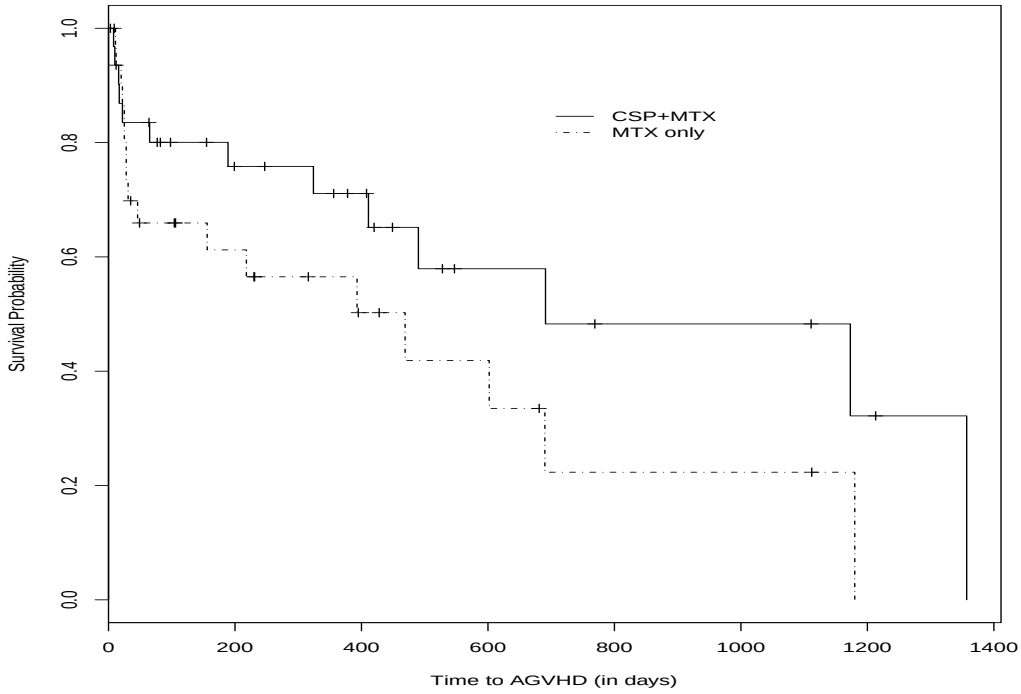

Figure 1: *Kaplan-Meier survival function estimates of the time to diagnosis of AGVHD for two treatment groups.*

```
> fit.2
Call: survdiff(formula = Surv(agvhd.times, cens.agvhd.times) ~ treat)

         N Observed Expected (O-E)^2/E (O-E)^2/V
treat=1 32       13     17.4      1.09      2.74
treat=2 32       17     12.6      1.50      2.74
Chisq = 2.7  on 1 degrees of freedom, p = 0.098
```