

**GROUND RULES:**

- This exam contains 8 questions; each question is worth 10 points. Therefore, this exam is worth 80 points.
- 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  $Y_1, Y_2, \dots, Y_n$  are independent (not iid) random variables satisfying

$$Y_i \sim \text{Poisson}(\theta x_i),$$

that is, for each  $i = 1, 2, \dots, n$ , the random variable  $Y_i$  is distributed as Poisson with mean equal to  $\theta x_i$ . The parameter  $\theta > 0$  is unknown and is treated as fixed (not random). The  $x_i$ 's are fixed (not random) and are also observed.

(a) Why aren't the  $Y_i$ 's necessarily identically distributed?

(b) Show that the likelihood function is given by

$$L(\theta) = c_0 \theta^{y_+} \exp(-\theta x_+),$$

where  $c_0$  is a quantity free of  $\theta$ ,  $y_+ = \sum_{i=1}^n y_i$ , and  $x_+ = \sum_{i=1}^n x_i$ .

(c) Suppose we would like to test  $H_0 : \theta = 1$  versus  $H_0 : \theta \neq 1$ . Show that the likelihood ratio test statistic  $\lambda$  is given by

$$\lambda = \left( \frac{x_+}{y_+} \right)^{y_+} \exp(y_+ - x_+).$$

You do not need to perform the test.

2. Consider the same statistical model as in Problem #1, that is,  $Y_1, Y_2, \dots, Y_n$  are independent random variables satisfying

$$Y_i \sim \text{Poisson}(\theta x_i).$$

We will continue to treat the  $x_i$ 's as fixed constants. However, unlike Problem #1, we will now regard  $\theta$  as a random variable with prior distribution

$$g(\theta) = b^{-1} \exp(-\theta/b),$$

where  $b$  is known.

(a) Show that the posterior distribution of  $\theta$  is gamma with parameters  $\alpha = y_+ + 1$  and  $\beta = (x_+ + b^{-1})^{-1}$ .

(b) Find the posterior mode estimator of  $\theta$ .

3. Suppose that  $Y_1, Y_2, \dots, Y_n$  is an iid sample of  $\mathcal{N}(\mu, \sigma^2)$  random variables. The parameter  $\mu$  is best regarded as fixed and is unknown. The Oracle has told us that  $\sigma^2 = 16$ . We would like to test

$$\begin{aligned} H_0 : \mu &= 0 \\ &\text{versus} \\ H_a : \mu &> 0. \end{aligned}$$

(a) Using a rejection region of the form

$$\text{RR} = \{\bar{y} : \bar{y} > k\},$$

we would like to maintain the following criteria:

- Our test should have  $P(\text{Type I Error}) = 0.10$ .
- Our test should have power equal to 0.80 when  $\mu = 1$ .

Determine the values of  $n$  and  $k$  that will meet the outlined criteria.

(b) Sketch a graph of what you expect the power function of the test in part (a) to look like. You do not have to do a lot of calculation here; just a rough sketch will suffice. However, please include key features of this graph, including axes labels and the numerical values possible on both axes.

(c) The rejection region above is the **uniformly most powerful** level 0.10 rejection region when testing  $H_0$  versus  $H_a$ . In a few sentences, describe/explain what this means.

4. Consider the simple linear regression model  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ , for  $i = 1, 2, \dots, n$ , where  $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$  and  $\sigma^2 > 0$  is fixed and is unknown. Recall that in this model the  $x_i$ 's are treated as fixed constants (i.e., they are not random). Let  $\hat{\beta}_1$  denote the least squares estimator of  $\beta_1$ . In class, we proved that  $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2)$ , where

$$c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

We also stated, without proof, that

$$W = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$

where  $\hat{\sigma}^2 = (n-2)^{-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  is the mean squared error. For  $i \neq j$ , define

$$\theta = E(Y_i - Y_j).$$

(a) Show that the parameter  $\theta$  is algebraically equal to

$$\theta = \beta_1(x_i - x_j).$$

(b) Derive a  $100(1 - \alpha)$  percent confidence interval for  $\theta$ . Define any notation used.

(c) For what values of  $x_i$  and  $x_j$  will the length of your confidence interval in part (b) be minimized?

5. Consider the multiple linear regression model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\boldsymbol{\epsilon} \sim \mathcal{N}_4(\mathbf{0}, \sigma^2\mathbf{I})$ . The design matrix and response vector are, respectively,

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

- Calculate the least squares estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ .
- Under the model assumption  $\boldsymbol{\epsilon} \sim \mathcal{N}_4(\mathbf{0}, \sigma^2\mathbf{I})$ , what is the distribution of  $\hat{\boldsymbol{\beta}}$ ?
- Calculate the uncorrected model sum of squares  $\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$ .
- Calculate an unbiased estimate of  $\sigma^2$ .

6. Suppose that  $Y_1, Y_2, \dots, Y_n$  is an iid sample from a negative binomial distribution with known waiting parameter  $r$  and unknown success probability  $\theta$ , where  $0 < \theta < 1$ . 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

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

and  $f_Y(y; \theta)$  denotes the conditional pmf of  $Y$ , given  $\theta$ .

(a) Show that

$$g(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}}.$$

(b) Find the posterior distribution  $g(\theta|\mathbf{y})$ .

(c) Describe how to find a  $100(1 - \alpha)$  percent credible interval for  $\theta$ . A description in words or a properly-labeled picture will suffice.

7. The lifetime random variable  $T$  has a Weibull distribution with pdf

$$f_T(t) = \begin{cases} \beta^{-1} \alpha t^{\alpha-1} \exp(-t^\alpha/\beta), & t > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where  $\alpha > 0$  and  $\beta > 0$  are unknown.

(a) Show that the survivor function of  $T$  is

$$S_T(t) = \begin{cases} 1, & t \leq 0 \\ \exp(-t^\alpha/\beta), & t > 0. \end{cases}$$

(b) Find an expression for  $\phi_{0.5}$ , the median survival time.

(c) When  $\alpha = 1$ , the Weibull distribution reduces to what commonly known distribution?

What is the shape of the corresponding hazard function when  $\alpha = 1$ ?

8. Laryngeal cancer is a disease in which malignant cells form in the tissues of the larynx (voice box). In a recent study, 90 male subjects were followed. Let  $T$  denote the time to death (in years). Not all subjects died during the study; that is, some of the subjects death times were censored. In Figure 1, I have constructed the Kaplan-Meier estimate of the survival function using the data from these 90 subjects.

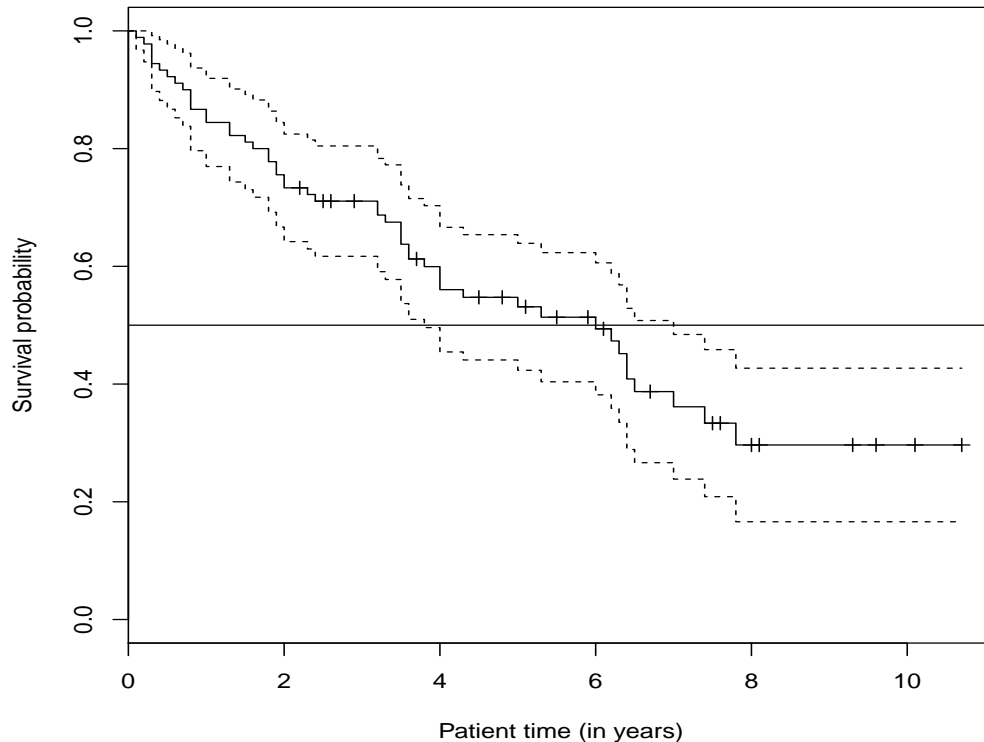


Figure 1: *Kaplan-Meier survival function estimate of time to death from laryngeal cancer among male subjects. A horizontal line at the 0.5 survival probability has been added.*

(a) The Kaplan-Meier estimate is a **nonparametric** estimate of the survival function  $S_T(t)$ . In a few sentences, describe what this means. Are there advantages to using a nonparametric estimate? Disadvantages?

(b) Use the graph above to provide

- (i) a point estimate of the median survival time
- (ii) an approximate 95 percent confidence interval of the median survival time.

(c) In addition to the 90 male subjects, suppose that we had an additional sample of 90 female subjects and their associated death/censoring times. Which nonparametric statistical test could be used to compare the male and female survival functions?

This is an extra page for Problem #8. Use it if you wish.