Name: ____

Biostatistics 1st year Comprehensive Examination: Theory

June 6th, 2016: 9am to 5pm

Instructions:

1. There are six questions and 6 pages (not including the cover sheet).
2. Answer each question to the best of your ability. Be as specific as possible and write as clearly as possible.
3. Put your name and problem number on every sheet of paper; **only use one side** of the paper (the exams will be scanned electronically).
4. This is an in-class examination; do not discuss any part of this exam with anyone while you are taking the exam. **NO BOOKS, NO NOTES, NO FRIENDS, NO PETS, NO INTERNET DEVICES, and NO OUTSIDE ASSISTANCE.**
5. You may leave the examination room to use the restroom or to step out into the hallway for a short break. **HOWEVER, YOU MUST LEAVE YOUR CELL PHONE AND ALL EXAM MATERIALS IN THE EXAMINATION ROOM.** If there is an emergency, please discuss this with the exam proctor.
6. Vanderbilt's academic honor code applies; *adhere to the spirit of this code.*

Question	Points	Score	Comments
1	50		
2	50		
3	50		
4	50		
5	50		
6	50		
Total	**300**		

1. Let $X_1, \dots, X_n \overset{iid}{\sim} f_X(x)$ be location-shifted exponential random variables with

$$f_X(x) = \theta e^{-\theta(x-2)} \quad \text{for} \quad x > 2, \;\; \theta > 0$$

   a. Find the expectation of X_i and variance of X_i?

   b. What is the minimal sufficient statistic for θ?

   c. Is this distribution part of the exponential family? Justify your answer.

   d. Derive a methods of moments estimator (MME) for θ.

   e. Is the MME an unbiased estimator of θ? Justify your answer.

   f. Construct a 95% confidence interval for θ. Be sure to specify any necessary notion of supporting conditions.

Now suppose θ is a random variable with density

$$f_\theta(\theta) = \beta e^{-\beta\theta} \quad \text{where} \quad \infty > \beta > 0$$

   g. Does the expectation of X_i and variance of X_i change? If so, find it.

   h. Find the posterior distribution of θ given $x_1, \dots, x_n$ and suggest a Bayes estimator of θ.

   i. Show that the posterior mean and methods of moments estimator (MME) are approximately equal in large samples.

2. Let X and Y be independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$.

   a. Show that the pdf of $Z = Y + X$ is given by the convolution formula

   $$f_Z(z) = \int f_X(w) f_Y(z - w) dw$$

   b. Derive the analogous "multiplicative" convolution formula for $Z = XY$. For this problem, only consider the case when X and Y are positive random variables.

   Let $M_Z(t)$ represent the moment generating function for a random variable Z.

   c. Let Z be a continuous random variable. Define $M_Z(t)$ and show how a moment generating function is used to find the n^{th} moment of Z.

   Now suppose that $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, \sigma^2)$. The moment generating function for $X \sim N(\mu, \sigma^2)$ is $e^{\mu t + \sigma^2 t^2/2}$.

   d. Use the convolution formula to find the distribution of $Z = Y + X$.

   e. Use the moment generating function to find the distribution of $Z = Y + X$.

   Now let random variables X and Y be *discrete* with probability mass functions $f_X(x) = (3 - x)/15$ for $x \in \{-2, -1, 0, 1, 2\}$ and $f_Y(y) = (1 - y)^2/6$ for $y \in \{-1, 0, 2\}$. Also, X and Y are independent.

   f. Find the pmf of $Z = XY$ (i.e., find $P(Z = z)$ for all z).

   g. Find $P(Z > Y)$.

3. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$ be independent of $Y \sim Ber(1/n)$ where $P(Y = 1) = 1/n$. Consider the following estimator of μ:

$$\hat{\mu}_n = (1 - Y)\,\bar{X}_n + Y\,n^2$$

a. Is $\hat{\mu}_n$ an unbiased estimator of μ? If yes, show it. If not, determine the bias.

b. Show that the limiting distribution of $W_n = \sqrt{n}(\hat{\mu}_n - \mu)$, call it W, is normal. Hint: Consider conditioning on Y.

c. Define the limiting bias of an estimator and find it for $\hat{\mu}_n$.

d. Define the asymptotic bias of an estimator and find it for $\hat{\mu}_n$.

e. Is $\hat{\mu}_n$ a consistent estimator of μ? Prove it or show why not.

f. What does this problem illustrate about the relationship between convergence in distribution and convergence in moments?

g. What is the limiting distribution of $\frac{1+\sqrt{5}}{2}(\bar{X}_n)^{\pi/2}$? Justify your answer.

4. We observe pairs $(Y_1, x_1), \dots, (Y_n, x_n)$. A simple linear regression model for n observations is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for} \quad i = 1, \dots, n$$

Assume that $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$ for all i, and that ε_i is independent of ε_j for all $i \neq j$. Throughout this problem, the x_i are considered fixed and known. All summations are over $i = 1, \dots, n$ unless otherwise specified.

Note: You may use matrix notation to solve this problem. Just be sure to clearly define *all* of your notation. Using matrix notation is not required.

a. Show that the least squares estimators of β_0 and β_1 are:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{\sum (x_i - \overline{x}) y_i}{\sum (x_i - \overline{x})^2}$$

b. Show that $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased estimators of (β_0, β_1).

Now assume that $\varepsilon_i \sim N(0, \sigma^2)$.

c. Find the likelihood function for (β_0, β_1) and derive the MLEs for β_0 and β_1. How do they compare to the least squares estimators?

d. Show that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum (x_i - \overline{x})^2)$.

e. Explain why replacing σ^2 with its estimate, $s^2 = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 / (n-2)$, leads to a t-distribution for $\hat{\beta}_1$. Identify the degrees of freedom. A formal proof is not necessary.

f. Given a fixed sample size n where we are allowed to sample any x_i, how should we select $x_1, \dots, x_n$ to obtain the most efficient estimator of β_1?

5. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f(Y;\theta)$ where $f(\cdot)$ is a member of the exponential family with

$$f(Y;\theta) = \exp\{\theta t(Y) - \kappa(\theta) + c(Y)\}$$

Here $t(\cdot)$ and $c(\cdot)$ are functions of the data Y, and $\kappa(\cdot)$ is a function of the parameter θ. Assume any necessary derivatives or inverses are well defined, e.g. $\kappa'(\theta)$ and κ'^{-1} exist.

a. Find the MLE of θ.

b. Find an estimate of the expected fisher information in the sample $y_1, \ldots, y_n$.

c. Let S_i be the score function for the i^{th} observation. Find a large sample approximation to the distribution of the average score function $\overline{S} = \frac{1}{n}\sum_i S_i$. Hint: consider $\sqrt{n}\,\overline{S}$.

Now suppose that $Y_1, \ldots, Y_n \overset{iid}{\sim} g(Y)$ where $E_g[Y]$ exists and $\kappa(\theta) = \theta^2/2$.

d. When the working model fails, what quantity does the MLE estimate?

e. What is the limiting distribution of $\sqrt{n}\big(\hat{\theta}_n - \theta\big)$ under model failure? Here $\hat{\theta}_n$ is the MLE.

f. Find a robust large-sample $100(1-\alpha)\%$ CI for $E_g[t(Y)]$. Define your notation and identify key assumptions.

6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Ber(\theta)$ where $n \geq 2$. Consider $W(X_1, \ldots, X_n) = X_1 X_2$ such that $E[W] = P(W = 1) = P(X_1 = 1 \text{ and } X_2 = 1)$.

    a. Find $E[W]$.

    b. Show that MVUE for θ^2 is $\overline{X}\left(\overline{X} - \frac{1}{n}\right)\frac{n}{n-1}$.

    c. Let V be the variance of the MVUE for θ^2. Find V or suggest a reasonable approximation.

    d. Find the Cramer-Rao Lower Bound on the variance of an unbiased estimator of θ^2.

    e. Compare the variance you obtained in part (b) to the CRLB from part (c). Note any useful conclusions.

    f. What is the minimum achievable length of an approximate large-sample CI for θ^2 when the CI is based on an unbiased estimator of θ^2?