Name: ____

Biostatistics 1st year Comprehensive Examination: Theory

May 29th, 2018: 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 and Y be random variables from the joint distribution given by pdf

$$f(x, y) = 15x^2y \quad \text{for} \quad 0 < x < y < 1$$

   a. Find the marginal distribution of Y. Be sure to show your answer is indeed a valid pdf.

   b. Find $E[Y]$ and $Var[Y]$.

   c. Find the median of Y.

   d. Describe an algorithm for generating $Y_1, \ldots, Y_{100} \overset{iid}{\sim} f(Y)$ from 100 uniform random variables $U_1, \ldots, U_{100} \overset{iid}{\sim} U(0,1)$.

   e. Find the pdf of $Z = Y^2$?

   f. Find the conditional distribution of X given Y.

   g. Are X and Y independent random variables? Justify your answer.

   h. What is $Cov(2X, Y)$?

   i. What is $P(X + Y < 1)$?

2. Let $X_1, \ldots, X_n \overset{iid}{\sim} f(X;\theta)$ where f has a $Beta(\theta, 1)$ distribution

$$f(X;\theta) = \theta X^{\theta-1} \quad \text{for} \quad 0 < X < 1 \;\; \text{and} \;\; \theta > 0$$

with $E[X] = \dfrac{\theta}{\theta+1}$, $E[X^2] = \dfrac{\theta}{(\theta+1)^2(\theta+2)}$, $E[\log X_i] = \dfrac{1}{\theta}$, $E[(\log X_i)^2] = \dfrac{2}{\theta^2}$

a. Find the minimal sufficient statistic for θ.

b. Find the Maximum Likelihood Estimator of $1/\theta$.

c. Find the Cramer-Rao Lower Bound for an unbiased estimator of $1/\theta$.

d. Find the minimum variance unbiased estimator of $1/\theta$.

e. Under model failure, what quantity is the MLE consistent for (i.e., what is the object of inference)?

f. Find the robust variance estimator for the MLE under model failure.

g. Consider the estimator $\hat{\phi} = \left(\sqrt{n}\pi \; - \sum_i \log x_i\right)/n$. Is $\hat{\phi}$ is a consistent estimator of $1/\theta$? If yes, prove it. If not, explain why.

3. Let $Y|X,\beta,\sigma^2 \sim N(\beta X,\sigma^2)$. Assume $\beta|\sigma^2 \sim N(\mu,\sigma^2)$, and $\sigma^2 > 0$ has an inverse-gamma distribution $IV(a,b)$, where

$$f(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-\frac{b}{\sigma^2}}$$

where $a,b > 0$ and $\Gamma(w) = (w-1)!$ These prior distributions are conjugate.

a. Assume $a > 1$. Find $E[\sigma^2]$.

b. Find $E[Y|X]$.

c. Find $Var[Y|X]$.

d. Show the posterior distribution, $f(\beta,\sigma^2|Y=y,X=x)$, is proportional to

$$\frac{1}{\sigma}e^{-\frac{1}{2\sigma^2}\left(\beta-\frac{xy+\mu}{x^2+1}\right)^2(x^2+1)}(\sigma^2)^{a-\frac{1}{2}-1}\ e^{-\frac{1}{2\sigma^2}\left(2b+\mu^2+y^2-\frac{(xy+\mu)^2}{x^2+1}\right)}$$

e. What is the posterior mean of β?

The model described above will be considered our first model, M_1, but we are also considering a second model, M_2, where $Y|X,\gamma,\sigma^2 \sim N(\gamma,\sigma^2)$ and $\gamma|\sigma^2 \sim N(\mu,\sigma^2)$, and $\sigma^2 > 0$ follows and inverse-gamma distribution $IV(a,b)$. Models M_1 and M_2 are equally likely a priori.

f. Find the Bayes factor comparing models M_1 and M_2 and provide an interpretation for it. You do not need to reduce this expression.

4. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Exp(\phi_y)$ and $X_1, \ldots, X_m \overset{iid}{\sim} Exp(\phi_x)$ be two independent samples where the exponential distributions are parameterized as

$$f(y; \phi_y) = \frac{1}{\phi_y} e^{-y/\phi_y}$$

for $\phi_y > 0, y > 0$ so that $E[Y_i] = \phi_y$ and $Var[Y_i] = \phi_y^2$.

a. What is the generalized likelihood ratio test statistic for testing $H_0: \phi_y = \phi_x$ vs. the alternative hypothesis $H_1: \phi_y \neq \phi_x$? Denote this statistic by $\Lambda_{n,m}$.

b. What is the large-sample distribution of $\Lambda_{n,m}$?

c. Propose an alternative test of $H_0: \phi_y = \phi_x$. Be sure to provide enough details (test statistic, sampling distribution under the null, type of test, etc.) to establish the test as a valid test.

Now suppose the null hypothesis is $H_0: \phi_y = \phi_x \ and \ \phi_y, \ \phi_x < c_0$ while the alternative hypothesis is $H_1: \phi_y \neq \phi_x \ and \ \phi_y, \ \phi_x > c_1$ for constants $c_1 > c_0$.

d. Define the generalized likelihood ratio test statistic that would be used in this case, denote it by $\Lambda^*_{n,m}$, and make its form as simple as possible. Hint: Use a picture to illustrate the situation and motivate the analytical expression.

e. What is the large-sample distribution of $\Lambda^*_{n,m}$? Either justify an analytical solution or provide a numerical algorithm that would lead to this distribution. Be specific.

5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(X_i;\theta)$ where f is a shifted exponential distribution

$$F(X_i;\theta) = \begin{cases} 1 - e^{-(X_i-\theta)} & X_i \geq \theta \\ 0 & X_i < \theta \end{cases}$$

Consider the two hypotheses: $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$ where $\theta_1 > \theta_0$. Denote the smallest observation by $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and the data vector by $\underline{X} = (X_1, \ldots, X_n)$.

a. Find the likelihood function for θ, call it $L\left(\theta|\underline{X}\right)$.

b. Show that $P\left(X_{(1)} > c\right) = e^{-n(c-\theta)}$ for $c > \theta$. [Hint: $X_{(1)} > c \Longrightarrow X_i > c \ \forall\, i$].

c. Consider the rejection region $\left\{\underline{X}: X_{(1)} > c\right\}$ and find c that gives a one-sided α-level test of H_0.

d. What is the *p-value* for the set of observations $z_1, \ldots, z_k$?

e. Find the power function for the test specified in part (c).

f. What is the set of null hypotheses that do not reject from the test in part (c)?

g. Use the Karlin-Rubin theorem to show that the test from part (c) is a most powerful test of size α.

h. Use the Neyman-Pearson Lemma to find a test that is different from (c) but also a most powerful test of size α.

6. *Causal Inference.* Let Y_z denote the potential outcome of Y if a subject is assigned treatment $Z = z$, where $z = 0$ or 1. Y_z is a random variable. Let X denote a covariate. An individual's 'causal effect' is said to be $ICE = Y_1 - Y_0$. But without strong assumptions it is impossible to identify ICE because no subject can be assigned to both $Z = 1$ or $Z = 0$ simultaneously. So researchers focus on the average causal effect, $ACE = E[Y_1 - Y_0]$, which can be identified under weaker assumptions. Two assumptions that identify ACE are

   (i) "Consistency": $Y_z = Y$ if $Z = z$. For example, $f(Y_0|Z = 0) \equiv f(Y|Z = 0)$

   (ii) "Ignorable treatment assignment (i.e., randomization)": $(Y_1, Y_0) \perp Z$

   a. Prove that under assumptions (i) and (ii), $ACE = E[Y|Z = 1] - E[Y|Z = 0]$

   b. Suppose that 100 subjects are randomized to treatment ($Z = 1$) and 100 subjects are randomized to control ($Z = 0$). Use the result from part (a) to provide an unbiased estimator of ACE and find its variance.

   c. Provide and justify an approximate 95% confidence interval for ACE.

   When using observational data, assumption (ii) often does not hold. A weaker assumption that might be reasonable is the following:

   (iii) "Conditional ignorable treatment assignment (i.e., no unmeasured confounding)": $\{(Y_1, Y_0) \perp Z\}|X$.

   d. The propensity score is defined as $e(x) \equiv P(Z = 1|X = x) = E[Z|X = x]$. Prove that

$$ACE = E\left[\frac{YZ}{e(X)}\right] - E\left[\frac{Y(1 - Z)}{1 - e(X)}\right]$$

   assuming that (i) and (iii) hold and that $0 < e(x) < 1 \ \forall x$ (this constraint on the propensity score is called 'positivity'). This is called an inverse probability of treatment weighting scheme. Hint: Make use of iterative expectations.