Name: _________________________________

Biostatistics 1st year Comprehensive Examination: Theory

May 26th, 2015: 9am to 5pm

Instructions:
1. There are six questions and 6 pages. Answer each question to the best of your ability.
2. Be as specific as possible and write as clearly as possible.
3. 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, NO OUTSIDE ASSISTANCE.
4. You may leave the examination room to use the restroom or to step out into the hallway for a short breather. 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.
5. Vanderbilt’s academic honor code applies.

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Score</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>300</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ random variables.

   a. Show that the moment generating function (MGF) for $X_i$ is $e^{\lambda(e^t - 1)}$.

   b. Show that the conditional distribution of $X_1$ given $X_1 + X_2$ is Binomial.

   c. Show that $S_n/n \xrightarrow{p} \lambda$ as $n \to \infty$.

   d. Find the asymptotic distribution of $\sqrt{S_n/n}$. Be sure to justify all steps. Explain why this result can be useful in practice.

   e. Find a large-sample $100(1 - \alpha)$% CI for $\lambda$ using part (e). What are the three key statistical properties that justify this interval?

   f. Suppose the investigator is only willing to assume that the first moment is $\lambda$ and the second moment is constant. Provide a large-sample $100(1 - \alpha)$% CI for $\lambda$ under these circumstances. Justify your answer.

   g. Find the minimum variance unbiased estimator (MVUE) of $\lambda^2$. 
2. A standard Gumbel distribution has cdf \( F(X) = \exp\{-e^{-(X-\mu)/\sigma}\} \) for 
\(-\infty < X < \infty, -\infty < \mu < \infty, \) and \( \sigma > 0. \)

a. What is the pdf of \( X? \)

b. What is the median of \( X? \)

c. What is the mode of \( X? \)

d. How would you generate data from the location-scale transformed Gumbel using only a uniform(0,1) distribution?

e. The moment generating function (mgf) of \( X \) is \( \Gamma(1 - \sigma t)e^{\mu t}. \) What is the mean of \( X? \) [Hint: Euler’s constant is \( \gamma = -\Gamma'(1). \)]

f. Prove that if \( Y \sim Exp(1) \) then \( -\log(Y) \sim Gumbel(0,1). \)

g. Let \( X_i \overset{iid}{\sim} Gumbel(\mu, 1) \) for \( i = 1, ..., n. \) What is the MLE of \( \mu? \)
3. Let \((X_1, Y_1), \ldots, (X_n, Y_n) \sim iid\) Bivariate Normal with mean \(E(X_i) = 0, E(Y_i) = 0\) and covariances \(Var(X_i) = 1, Var(Y_i) = 1, Cov(X_i, Y_i) = \rho \in (-1, 1)\) for any \(1 \leq i \leq n\) and \(-\infty < x_i, y_i < \infty\). The MGF is

\[ M_{XY}(t_1, t_2) = e^{\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)} \]

and density is

\[ f(x_i, y_i) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp \left\{ -\frac{x_i^2 - 2\rho x_i y_i + y_i^2}{2(1 - \rho^2)} \right\} \]

Let \(T_X = \sum_{i=1}^{n} x_i^2\) and \(T_Y = \sum_{i=1}^{n} y_i^2\) be statistics.

a. Find a minimal sufficient statistic for \(\rho\).

b. Show that both \(T_X\) and \(T_Y\) are ancillary statistics for \(\rho\).

c. Explain why the joint statistic \((T_X, T_Y)\) cannot ancillary for \(\rho\).

Now let \((X_i, Y_i)\) be paired observations such that \(X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2), Y_1, \ldots, Y_n \sim N(\mu_y, \sigma_y^2)\) and \(\rho = \text{corr}(x_i, y_i)\) for \(i = 1, \ldots, n\) with \(\rho = \text{corr}(x_i, y_j) = 0\) for \(i \neq j\).

d. What is the standard \(\alpha\)-level test for testing the null hypothesis \(H_0: \mu_x = \mu_y\) when the observations are assumed unpaired, independent, and with equal variance? [You do not have to derive this test; just state the test statistic, its distribution and assumptions.]

e. What happens to the Type I Error of this test when \(\rho = 0\) and the variances are not equal? What happens when \(0 < |\rho| < 1\) and the variances are not equal?

f. Under these circumstances, which is more important to get right: the independence structure or the variance structure? Justify your answer. Suggest an alternative approach that could avoid these problems.
4. Let $X_i, i = 1, ..., n$ be iid $Uniform(0, \theta)$ random variables.
   
   a. Show that the MLE of $\theta$ is the maximum order statistic $X_{(n)}$.
   
   b. What is the density function (pdf) of $X_{(n)}$?
   
   c. What is the expectation of $X_{(n)}$?
   
   d. Is $X_{(n)}$ a biased estimator of $\theta$? Why or why not?
   
   e. Is $X_{(n)}$ a consistent estimator of $\theta$? Why or why not?
   
   f. Prove that $n(\theta - X_{(n)})$ converges in distribution to an $Exponential(\theta)$ distribution with CDF

   \[ F(t) = 1 - e^{-t/\theta} \]

   g. Use (f) to provide a $(1 - \alpha)100\%$ level lower (one-sided) confidence interval for $\theta$. 
5. Consider observations \( X_1, \ldots, X_n \overset{iid}{\sim} f(X) \) and hypotheses \( H_0: f(X) = f_0(X) \) and \( H_1: f(X) = f_1(X) \). For this problem, assume the densities are everywhere positive so that \( f(X) > 0 \) for all \( X \in S \), that derivatives and inverses exist, and that hypothesized densities are distinct, i.e. they are not equal at one or more points in the sample space.

a. Show that the Kullback-Leibler divergence between the two hypothesized densities is always positive. That is, show that
\[
E_{f_0} \left[ \log \frac{f_0(X)}{f_1(X)} \right] > 0
\]

b. Write down the likelihood ratio for \( H_1 \) vs. \( H_0 \) based on observations \( X_1, \ldots, X_n \). Show that it converges to \( \infty \) under \( H_1 \), and 0 under \( H_0 \), as \( n \to \infty \).

c. Write down the probability of observing a large likelihood ratio in support of \( H_1 \) when \( H_0 \) is true (where 'large' is defined a likelihood ratio of \( k > 1 \) or greater). Show that this probability converges to 0 as \( n \to \infty \).

d. Suggest a way of holding the probability in part (c) constant as the sample size grows. Name a statistical technique where this is done.

Let \( T = T(X_1, \ldots, X_n) \) be a test statistic with distribution function \( G(T) \). Then \( Y = 1 - G_0(T) \) is the \textit{p-value}.

e. Show that the distribution function for the \textit{p-value}, \( F_{Y}(y) \), is
\[
F_{Y}(y) = P(Y \leq y) = 1 - G\{G_0^{-1}(1 - y)\}
\]
and find the distribution of the \textit{p-value} under the null hypothesis \( (H_0) \).

f. Show that the probability distribution function (pdf) for the \textit{p-value} under the alternative hypothesis \( (H_1) \) is given by the likelihood ratio
\[
f_{Y}(y) = \frac{g_1\{T\}}{g_0\{T\}}
\]

g. Use the definition of a \textit{p-value} for a composite null hypothesis to posit how the result in part (e) might change when testing composite nulls. Then, using part (f), posit why a \textit{p-value} would overstate the strength of evidence against the null hypothesis under a composite alternative. [No proofs; a broad conceptual argument is fine here.]
6. The hazard function, \( h(t) = \lim_{\Delta \to 0} P(t \leq T < t + \Delta | T \geq t) / \Delta \). Assume that \( T \) is a continuous random variable.

a. Show that \( h(t) = f(t)/S(t) = -\frac{\partial}{\partial t} \log S(t) \) where \( f(t) \) is the pdf and \( S(t) = 1 - F(t) \) is the survivor function.

b. Prove that \( S(t) = \exp \left\{- \int_0^t h(u)du \right\} \).

c. Suppose subject \( i \) is followed for time \( t_i \) until an event occurs (\( \delta_i = 1 \)) or the subject is censored (\( \delta_i = 0 \)). Assume \( t_1, ..., t_n \) are independent and argue that the likelihood function is

\[
\prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}
\]

d. Let \( h(t) = h_0(t) \exp(\beta x) \) and \( H_0(t) = \int_0^t h(u)du \). Show that the likelihood can be written as

\[
\prod_{i=1}^n \left[ \mu_i^{\delta_i} e^{-\mu_i} \right] \left[ \frac{h_0(t_i)}{H_0(t_i)} \right]^{\delta_i}
\]

where \( \mu_i = H_0(t_i)e^{\beta x_i} \)

e. The ‘kernal’ of this likelihood is similar to the kernal of the likelihood for what other distribution? What other options for estimating/computing \( \beta \) does this inspire?

f. What is \( h_0(t_i) \) when \( T|X = x_i \sim \text{Exp}(e^{\beta x_i}) \)? [Here \( y \sim \text{Exp}(\lambda) \) implies that \( f(y; \lambda) = \lambda e^{-\lambda y} \).]

g. Suppose \( n = 100, x = \{0,1\}, \sum \delta_i x_i = 25 \) and \( \sum t_i x_i = 100 \). What is the MLE if data are assumed exponential as in (f)?