**STATISTICS Ph. D. COMPREHENSIVE EXAM**
August, 2015

General Instructions: Write your ID number on your answer sheets. Do not put your name on any of your answer sheets. All solutions should be rigorously explained. Show all work. If you cannot solve a problem, at least explain what the problem is about and what approach may lead to its solution.

Reminder: Unless expressly stated otherwise, all assertions should be rigorously proved.

1. *Sandwich Estimation.* Assume the independent, heteroscedastic regression model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{Cov}(e) = V = D(\sigma_i^2).$$

The generalized least squares estimates, the BLUEs and, under normality, the MLEs, and the UMVU estimates all satisfy

$$X\hat{\beta}_G = X\left[X'V^{-1}X\right]^{-1} X'V^{-1}Y,$$

except that this is not an estimate because we do not know the variances involved. One option is to simply use least squares.

(a) Find the covariance matrix of the least squares estimates, say $X\hat{\beta}$.

(b) Find $\mathrm{E}\left[D^2(Y - X\beta)\right]$, where $D^2$ applied to a diagonal matrix $D$ is just $DD$.

(c) What is the obvious estimate of $\mathrm{Cov}(X\hat{\beta})$? (This should be a "sandwich estimate.")

(d) Is the previous estimate unbiased? Is the previous estimate any good? (This last question is the entire point of the exercise!)

2. Let $X = (x_1, x_2, \ldots, x_n)$ be a random sample from the exponential distribution with density

$$f\left(x\,|\lambda\right) = \lambda \exp\left(-\lambda x\right); x \geq 0.$$

Consider testing $\mathrm{H}_0 : \lambda = \lambda_0 = 1$ vs. $\mathrm{H}_1 : \lambda = \lambda_1 = 2$.

(a) Show that, under $\mathrm{H}_0$, $Z(X) = \sqrt{n}\left(\bar{X}_n - 1\right)$ converges in distribution to a $N(0,1)$.

(b) To test $\mathrm{H}_0$ against $\mathrm{H}_1$ should we use a critical region of the form $Z(X) > \gamma$ or $Z(X) < \gamma$? Why?

(c) Show that, under $\mathrm{H}_1$, $Z(X)$ is approximately a $N\left(-\frac{\sqrt{n}}{2}, \frac{1}{4}\right)$.

(d) Find an expression for the power function of the test that rejects $\mathrm{H}_0$ when $Z(X) < \gamma$ in terms of $\Phi(\cdot)$, the cumulative distribution of a $N(0,1)$.

3. Let $X \sim N(\mu, 1)$ and $Y \sim N(\nu, 1)$ be independent, and suppose we wish to construct a confidence interval for $\theta = \mu/\nu$. Show that $(X - \theta Y)/(1 + \theta^2)^{1/2}$ is a pivotal quantity, and that the confidence set obtained from it could consist of the whole real line, a single interval or two disjoint intervals. Is the confidence set ever empty? Hint: Find the roots of a quadratic equation in $\theta$.

4. Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed with joint density $f(y;\theta)$ with $\theta$ a scalar parameter. Suppose a one-to-one differentiable transformation is made from $\theta$ to $\phi = \phi(\theta)$.

   (a) What is the relationship between the Fisher information $i(\phi)$ about $\phi$ contained in the sample and $i(\theta)$, the Fisher information about $\theta$?

   (b) In Bayesian inference for $\theta$, it was suggested by Jeffreys that a uniform prior should be assumed for that function $\phi$ of $\theta$ for which the Fisher information $i(\phi)$ is constant. By considering the relationship between the prior density of $\theta$ and that of $\phi$, show that the suggestion leads to a prior density for $\theta$ itself which is proportional to $i(\theta)^{1/2}$.

   (c) Let $X$ be a single observations from a Binomial$(k, \theta)$ distribution. What is the *Jeffreys' prior* for $\theta$?

5. Let $X_1, X_2, \ldots$ be independent with $X_n$ taking the values $\sqrt{n-1}$, 1, $-1$, and $-\sqrt{n-1}$ each with probability 1/4. Find the asymptotic distribution of $\bar{X}_n$.

6. *Machine Learning.* A client has data $(x_i, y_i)$, $i = 1, \ldots, n$ that she expects to fall (roughly) on one half of a circle (**not** the half disk). She wants to estimate the center and the radius of the circle.

   (a) Would $(\bar{x}, \bar{y})$ make a good estimate for the center of the circle? Explain.

   (b) Without worrying about the statistical properties of the estimates (which is what makes this a machine learning problem), find a way to use a regression program to estimate the center and the radius. Use the notation

   $$(x_i - \eta)^2 + (y_i - \gamma)^2 \doteq r^2,$$

   where the center is $(\eta, \gamma)$ and the radius is $r$. Hint: Write an observable quantity as a linear combination of unknown parameters and other observable quantities.

   (c) Explain how your regression estimates give the desired estimates.

   (d) What would it take to make your estimate of the radius nonsensical?