ST 793 Final Exam – Spring 2018

Turn off and put away all cell phones and electronic devices. You may not use a calculator. Put your answers on the sheets of paper handed out.

1. One of the "big" ideas of the class has been to (i) assume a statistical model in order to answer the main scientific question using the primary part, but then (ii) worry about secondary parts of the model when making inference via the asymptotic distributions. In the simplest setting, we have $Y_1, \ldots, Y_n$ iid from some unknown density g and distribution function G. We assume a parametric family density $f(y; \boldsymbol{\theta})$ in order to define a maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ and possibly test statistics about $\boldsymbol{\theta}$.

a. If the true density g is not of the form $f(y; \boldsymbol{\theta})$, then the first thing to worry about is whether $\widehat{\boldsymbol{\theta}}$ converges in probability to a meaningful parameter value, call it $\boldsymbol{\theta}_0$. Give an equation to show how this limit value is defined when $g(y) \neq f(y; \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$.

b. Assuming that the limit in a. is meaningful, let us think next about estimating asymptotic variances.

(i) If the parametric model is correct, $g(y) = f(y; \boldsymbol{\theta_0})$, give two classical estimators (available before 1975, say) for the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$.

(ii) If the parametric model is incorrect, then consider the jackknife and nonparametric bootstrap estimators. I don't think I ever gave the form for these estimators when $\widehat{\boldsymbol{\theta}}$ is a vector. So think about how sample variances generalize to sample covariances and then give the jackknife and nonparametric bootstrap estimators of the asymptotic covariance matrix.

2. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid pairs with $\mathrm{E}\, X_1 = \mu_1$, $\mathrm{E}\, Y_1 = \mu_2$, $\mathrm{Var}\, X_1 = \sigma_1^2$, $\mathrm{Var}\, Y_1 = \sigma_2^2$, and $\mathrm{Cov}\,(X_1, Y_1) = \sigma_{12}$. In the following parts give justifications for the results.

a. Give the asymptotic distribution of $(\overline{X}, \overline{Y})^T$.

b. Suppose that $\mu_1 = \mu_2 = 0$ and let $T = (\overline{X})(\overline{Y})$. Show that

$$nT \xrightarrow{d} Q \qquad \text{as } n \to \infty,$$

and describe the random variable Q.

c. Suppose that $\mu_1 = 0, \mu_2 \neq 0$ and let $T = (\overline{X})(\overline{Y})$. Show that

$$\sqrt{n}T \xrightarrow{d} R \qquad \text{as } n \to \infty,$$

and describe the random variable R.

3. Suppose that $X_1, \ldots, X_n$ are iid with distribution function F.

a. Give the asymptotic joint distribution of the sample quartiles $(\widehat{\eta}_{1/4}, \widehat{\eta}_{3/4})$, stating any assumptions necessary and why the result holds.

b. Give the approximating h_T function for the interquartile range $T = \widehat{\eta}_{3/4} - \widehat{\eta}_{1/4}$.

c. Give the asymptotic variance for $T = \widehat{\eta}_{3/4} - \widehat{\eta}_{1/4}$.

(Hint: check out the theorems on the last page.)

4. Four patients were available to test whether a new treatment was better than the standard one. Two patients were randomly assigned to receive the new treatment, and the remaining two patients received the standard treatment. The new treatent patients obtained scores of 32 and 25 (lower is better), and the two patients receiving the standard treatment obtained scores of 33 and 45.

a. Compute the Wilcoxon Rank Sum statistic, W = sum of the new treatment ranks, for H_0 : treatments are identical versus the new treatment is better than the standard treatment.

b. Give the permutation distribution of W under H_0.

c. Give the permutation p-value = probability that W would be less than or equal to the observed W under the permutation distribution.

5. A small subset of a table from a Monte Carlo study might look like the following:

Distribution	$\widehat{\text{Bias}}(\widehat{\theta})$	$\widehat{\text{Var}}(\widehat{\theta})$	$\widehat{\text{E}}\{\widehat{\text{Var}}(\widehat{\theta})\}$
Normal	0.02	1.47	1.36
Laplace	0.05	1.37	1.25
t_5	0.03	1.28	1.17

These estimates were calculated from $N = 1,000$ Monte Carlo samples. Envision 1,000 independent rows of data, where in each row we have the basic estimator $\widehat{\theta}$ and an estimate of its variance, $\widehat{\text{Var}}(\theta)$, calculated from one Monte Carlo sample. In the above table, $\widehat{\text{Bias}}(\widehat{\theta})$ is just the average of the 1,000 values of $\widehat{\theta}$ minus the true parameter value, say θ. The next column $\widehat{\text{Var}}(\widehat{\theta})$ is simple a sample variance based on the 1,000 values of $\widehat{\theta}$. The last column is the average fo the 1,000 variance estimates.

a. Give an expression for the Monte Carlo standard error of each type of estimate. That is, give an expression for computing a standard error for each entry in columns 2-4 (column 1 is the distribution type). Note, although in practice I like to suggest the use of jackknife or bootstrap standard errors, for this question I want explicit expressions based on exact or asymptotic variances.

b. Usually we want to know if the entries in the 3rd column are close to the entries in the 4th column. However, that is made difficult because the entries in those columns are correlated. Suggest a simple way to combine those estimates to get another column that is easy to use for that purpose.

Theorem 5.19. *Suppose that* $\widehat{\boldsymbol{\theta}}$ *is* $AN_k(\boldsymbol{\theta}, b_n^2\boldsymbol{\Sigma})$ *with* $b_n \to 0$ *and that* g *is a real-valued function with partial derivatives existing in a neighborhood of* $\boldsymbol{\theta}$ *and continuous at* $\boldsymbol{\theta}$ *with* $g'(\boldsymbol{\theta}) = \partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ *not identically zero. Then as* $n \to \infty$

$$g(\widehat{\boldsymbol{\theta}}) \text{ is } AN\left[g(\boldsymbol{\theta}), b_n^2 g'(\boldsymbol{\theta})\boldsymbol{\Sigma}g'(\boldsymbol{\theta})^T\right].$$

Theorem 5.25. *Suppose that* $X_1, \ldots, X_n$ *are iid with distribution function* F *and* $F'(\eta_p)$ *exists and is positive. Then*

$$\widehat{\eta}_p - \eta_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p - I(X_i \le \eta_p)}{F'(\eta_p)}\right] + R_n,$$

where $\sqrt{n}R_n \xrightarrow{p} 0$ *as* $n \to \infty$. **Theorem 5.27.** *For an iid sample* $X_1, \ldots, X_n$*, suppose that each component of* $\widehat{\boldsymbol{\theta}}$ *has the approximation by averages representation given by*

$$\widehat{\theta}_j - \theta_j = \frac{1}{n}\sum_{i=1}^{n} h_j(X_i) + R_{nj}, \quad j = 1, \ldots, k, \tag{1}$$

where $\sqrt{n}R_{nj} \xrightarrow{p} 0$ *as* $n \to \infty$*,* $E\{h_j(X_1)\} = 0$*, and* $Var\{h_j(X_1)\}$ *is finite,* $j = 1, \ldots, k$*. Also assume that* g *is a real-valued function with partial derivatives existing in a neighborhood of the true value* $\boldsymbol{\theta}$ *and continuous at* $\boldsymbol{\theta}$*. Then*

$$g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{k} g_j'(\boldsymbol{\theta})h_j(X_i)\right] + R_n, \tag{2}$$

where $\sqrt{n}R_n \xrightarrow{p} 0$ *as* $n \to \infty$.

Theorem 5.28. *For an iid sample* $X_1, \ldots, X_n$*, suppose that each component of* $\widehat{\boldsymbol{\theta}}$ *has the approximation by averages representation given by*

$$\widehat{\theta}_j - \theta_j = \frac{1}{n}\sum_{i=1}^{n} h_j(X_i) + R_{nj}, \quad j = 1, \ldots, b, \tag{3}$$

where $\sqrt{n}R_{nj} \xrightarrow{p} 0$ *as* $n \to \infty$*,* $E\{h_j(X_1)\} = 0$*, and* $Var\{h_j(X_1)\}$ *is finite,* $j = 1, \ldots, b$*. Also assume that the real-valued function* $q(X_i, \boldsymbol{\theta})$ *has two partial derivatives with respect to* $\boldsymbol{\theta}$*, and*

1. $Var\{q(X_1, \boldsymbol{\theta})\}$ *and* $E\{q'(X_1, \boldsymbol{\theta})\}$ *are finite;*
2. *there exists a function* $M(x)$ *such that for all* $\boldsymbol{\theta}^*$ *in a neighborhood of the true value* $\boldsymbol{\theta}$ *and all* $j, k \in \{1, \ldots, b\}$*,* $|q''(x, \boldsymbol{\theta}^*)_{jk}| \le M(x)$*, where* $E\{M(X_1)\} < \infty$*.*

Then,

$$\frac{1}{n}\sum_{i=1}^{n} q(X_i, \widehat{\boldsymbol{\theta}}) - E\{q(X_1, \boldsymbol{\theta})\} = \frac{1}{n}\sum_{i=1}^{n} h_{\scriptscriptstyle T}(X_i) + R_n,$$

where

$$h_{\scriptscriptstyle T}(X_i) = q(X_i, \boldsymbol{\theta}) - E\{q(X_1, \boldsymbol{\theta})\} + \left[E\{q'(X_1, \boldsymbol{\theta})\}\right]\boldsymbol{h}(X_i),$$

and $\sqrt{n}R_n \xrightarrow{p} 0$ *as* $n \to \infty$.