18.655 Midterm Exam 2, Spring 2016
Mathematical Statistics
Due Date: 5/12/2016

Answer 4 questions for full credit, and additional question for extra credit.

1. **Estimation for Poisson Model**

   Let $X_1, \ldots, X_n$ be iid $Poisson(\theta)$, where $E[X_i \mid \theta] = \theta$.

   (a). Find $\hat{\theta}_{MLE}$, the maximum likelihood estimate for θ.

   (b). Determine the explicit distribution for $\hat{\theta}_{MLE}$.

   (c). Compute the mean-squared-prediction error of $\hat{\theta}_{MLE}$.

   (d). In a Bayesian framework, suppose

   - π is the prior distribution for θ with probability density function $\pi(\theta)$, $0 < \theta < \infty$.
   - Loss function: $L_k(\theta, a) = \dfrac{(\theta - a)^2}{\theta^k}$, for some fixed $k \geq 0$.

   Give an explicit expression for the Bayes estimate of θ given $\boldsymbol{x} = (x_1, \ldots, x_n)$.

   (e). In (d), suppose the prior distribution is $\pi = Gamma(a, b)$.

   - Is this a conjugate prior distribution?
   - Give an explicit formula for the Bayes estimate; if necessary, condition the values of k for the loss and/or (a, b) the specification of the prior.
   - Comment on the sensitivity of the Bayes estimate to increases/decreases in k, the choice of loss function.

2. **Model-Based Survey Sampling**

Consider the following setup for estimating population parameters with survey sampling:

- The population is finite of size N, for example a census unit.
- We are interested in estimating the average value of a variable, X_i, say current family income.

  The values of the variable for the population are:

  $x_1, x_2, \ldots, x_N$

  and the parameter is

  $\theta = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$
- Suppose that the family income values for the last census are known:

  $u_1, u_2, \ldots, u_N.$
- Ignoring difficulties such as families moving, consider a sample of n families drawn at random without replacement, let

  $X_1, X_2, \ldots, X_n$ denote the incomes of the n families.
- The probability model for the sample is given by

  $$P_{\boldsymbol{x}}[X_1 = a_1, \ldots, X_n = a_n] = \begin{cases} \frac{1}{\binom{N}{n}} & , if \quad \{a_1, \ldots, a_n\} \subset \{x_1, \ldots, x_N\} \\ 0 & , \quad otherwise. \end{cases}$$

  where $\boldsymbol{x} = (x_1, \ldots, x_N)$, is the distribution parameter.
- Consider the sample estimate:

  $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

  of the population parameter $\overline{x}$.

(a). Compute the expectation of $\overline{X}$ and determine whether it is an unbiased estimate for $\overline{x}$.

(b). Verify or correct the following formula for the mean-squared error of $\overline{X}$

$$MSE(\overline{X}) = \sigma_{\boldsymbol{x}}^2 \left(1 - \frac{n-1}{N-1}\right),$$

where $\sigma_{\boldsymbol{x}}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$.

(c). Consider using information contained in $\{u_1, \ldots, u_N\}$ and its probable correlation to $\{x_1, \ldots, x_N\}$, and define the regression estimate:

$$\widehat{\overline{X}_R} \equiv \overline{X} - b(\overline{U} - \overline{u})$$

where

- b is a prespecified positive constant
- U_i is the last census income corresponding to X_i
- $\overline{u} = \frac{1}{N} \quad {}_1^N \, u_i$.
- $\overline{U} = \frac{1}{n} \quad {}_1^n \, U_i$.

Prove that $\widehat{\overline{X}_R}$ is unbiased for any $b > 0$.

(d). In (c), prove that $\widehat{\overline{X}_R}$ has smaller variance than $\overline{X}$ if

$$b < 2Cov(\overline{U}, \overline{X})/Var(\overline{U}).$$

and that the best choice of b is

$$b_{opt} = cov(\overline{U}, \overline{X})/Var(\overline{X}).$$

(e). Show that if $\frac{n}{N} \to \lambda$ as $N \to \infty$, with $0 < \lambda < 1$, and if $E[X_1^2] < \infty$ then

$$\sqrt{n}(\overline{X} - \overline{x}) \xrightarrow{\mathcal{L}} N(0, \tau^2(1 - \lambda)),$$

where $\tau^2 = Var(X_1)$.

(f). Under the same conditions as (e), suppose that the probability model P_θ for

$$\{T_i = (X_i, U_i), i = 1, 2, \ldots, n\}$$

is such that

$X_i = bU_i + \epsilon_i$, $i = 1, \ldots, N$, where the $\{\epsilon_i\}$ are iid and independent of the $\{U_i\}$, with $E[\epsilon_i] = 0$, and $Var(\epsilon_i) = \sigma^2 < \infty$, and $Var(U_i) > 0$.

Show that:

$$\sqrt{n}(\overline{X}_R - \overline{x}) \xrightarrow{\mathcal{L}} N(0, (1 - \lambda)\sigma^2), \text{ with } \sigma^2 < \tau^2.$$

where $\overline{X}_R = \overline{X} - \hat{b}_{opt}(\overline{U} - \overline{u})$,

and

$$\hat{b}_{opt} = \frac{\frac{1}{n} \quad {}_{i=1}^{n}(X_i - \overline{X})(U_i - \overline{U})}{\frac{1}{n} \quad {}_{j=1}^{n}(U_j - \overline{U})^2}$$

(See Problems 3.4.19 and 5.3.11).

3. **Asymptotic Distribution of Correlation Coefficient**

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid as (X, Y) where:

- $0 < E[X^4] < \infty$ and $0 < E[Y^4] < \infty$
- $\sigma_1^2 = Var(X)$, and $\sigma_2^2 = Var(Y)$.
- $\rho^2 = Cov^2(X, Y)/\sigma_1^2\sigma^2$.

Consider estimates:

- $\hat{\sigma}_1^2 = \frac{1}{n}\sum_1^n (X_i - \overline{X})^2$.
- $\hat{\sigma}_2^2 = \frac{1}{n}\sum_1^n (Y_i - \overline{Y})^2$.
- $r^2 = \hat{C}^2/\hat{\sigma}_1^2\hat{\sigma}_2^2$
  where $\hat{C} = \frac{1}{n}\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})$

(a). Write $r^2 = g(\hat{C}, \hat{\sigma}_1^2, \hat{\sigma}_2^2) : R^3 \to R$, where $g(u_1, u_2, u_3) = u_1^2/u_2u_3$.

With focus on ρ and its estimate r, by location and scale invariance, we can use the transformations $\tilde{X}_i = (X_i - \mu_1)/\sigma_1$ and $\tilde{Y}_i = (Y_i - \mu_2)/\sigma_2$, and conclude that we may assume:

$$\mu_1 = \mu_2 = 0, \text{ and } \sigma_1^2 = \sigma_2^2 = 1, \text{ and } \rho = E[XY].$$

Under these assumptions, compute the first order differential of $g()$: $g^{(1)}(u_1, u_2, u_3)$.

(b). If $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, then show that

$$\sqrt{n}[\hat{C} - \rho, \hat{\sigma}_1^2 - 1, \hat{\sigma}_2^2 - 1]^T$$

has the same asymptotic distribution as

$$n^{1/2}[\frac{1}{n}\sum_1^n X_iY_i - \rho, \frac{1}{n}\sum_1^n X_i^2 - 1, \frac{1}{n}\sum_1^n Y_i^2 - 1]^T$$

(c). If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then

$$\sqrt{n}(r^2 - \rho^2) \to N(0, 4\rho^2(1 - \rho^2)^2).$$

and if $\rho \neq 0$, then

$$\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} N(0, (1 - \rho^2)^2).$$

(d). If $\rho = 0$, then

$$\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} N(0, 1).$$

(See Problem 5.3.9)

4. **Bounding Errors in Expectation Approximations**

   Suppose that

   - $X_1, \ldots, X_n$ are iid from a population with distribution P on $\mathcal{X} = R$.
   - $\mu_j = E[X_1^j], j = 1, 2, 3, 4$
   - $\mu_4 = E[X_1^4] < \infty$
   - $h : R \to R$ has derivatives of order k: $h^{(k)}$, $k = 1, 2, 3, 4$ and
     $$|h^{(4)}(x)| \leq M$$
     for all x and some constant $M < \infty$.

   Show that
   $$E[h(\overline{X})] = h(\mu) + \tfrac{1}{2}h^{(2)}(\mu)\tfrac{\sigma^2}{n} + R_n$$
   where
   $$|R_n| \leq h^{(3)}(\mu)|\mu_3|/6n^2 + M(\mu_4 + 3\sigma^2)/24n^2.$$
   (See Problem 5.3.23)

5. **Linear Model with Stochastic Covariates**

   Let $X_i = (Z_i^T, Y_i)^T, i = 1, 2, \ldots, n$ be iid as $X = (X^T, Y)^T$, where Z is a $p \times 1$ vector of explanatory varialbes and Y is the response variable of interest. Assume that

   - $Y = \alpha + Z^T\beta + \epsilon$, where
     $\epsilon \sim N(0, \sigma^2)$, independent of Z and $E[Z] = 0$. It follows that
     $$Y \mid Z = z \sim N(\alpha + z^T\beta, \sigma^2).$$
   - The stochastic covariates have distribution $Z \sim H_0$ with density h_0 and $E[ZZ^T]$ is nonsingular.

   (a). Show that the MLE of β exists (with probability 1 for sufficiently large n) and is given by
   $$\hat{\beta} = [\tilde{Z}_{(n)}^T\tilde{Z}_{(n)}]^{-1}\tilde{Z}_{(n)}^T\boldsymbol{Y}$$
   where $\tilde{Z}_{(n)}$ is the $n \times p$ matrix $||Z_{ij} - \overline{Z}_{\cdot j}||$, with $\overline{Z}_{\cdot j} = \frac{1}{n}\sum_{i=1}^n Z_{ij}$.

   (b). Show that the MLEs of α and σ^2 are
   $$\hat{\alpha} = \overline{Y} - \textstyle\sum_{j=1}^p \overline{Z}_{\cdot j}\hat{\beta}_j$$
   $$\hat{\sigma}^2 = \tfrac{1}{n}|\boldsymbol{Y} - (\hat{\alpha} + Z_{(n)}\hat{\beta}))|^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^n(\boldsymbol{Y}_i - (\hat{\alpha} + Z_i^T\hat{\beta}))^2$$

where $Z_{(n)}$ is the $n \times p$ matrix $||Z_{ij}||$.

(c). Prove that the asymptotic distribution of the mle's satisfy

$$(\sqrt{n}(\hat{\alpha}-\alpha, \hat{\beta}-\beta, \hat{\sigma}^2-\sigma^2) \xrightarrow{\mathcal{L}} N(\mathbf{0}, diag(\sigma^2, \sigma^2[E(ZZ^T)]^{-1}, 2\sigma^4)).$$

18.655 Mathematical Statistics
Spring 2016