PhD Qualifying Examination

Department of Statistics
University of South Carolina
May 24, 2019 - 9:00AM–3:00PM

READ FIRST THESE INSTRUCTIONS

1. **Procedure for assigning examinee codename:** Each examinee will draw a piece of paper from a box. The piece of paper from this box will contain a codename. Write your real name and codename on a sheet of white paper provided in the examination packet. Place this sheet of paper in an envelope provided in the examination packet. Seal the envelope and give the envelope to the proctor. The proctor will keep these envelopes sealed until all exams are graded and the Examination Committee has decided on the results. The Examination Committee will then open these envelopes to determine the examinees' names and their results.

2. DO NOT write your name on any of your answer sheets. Instead, write the codename that you drew from Step 1.

3. There are six (6) problems on this examination. Write your solutions on the yellow sheets of paper provided in your examination packet. **Write only on one side of each sheet of paper.**

4. The data set lowbirth.txt referred to in Problem # 4 is saved in the pin drive that is included in your examination packet.

5. You may use a computer in the examination room, R and SAS OnDemand on the computer, and/or a calculator. However, you are **not allowed** to use search engines during the examination. But you may use the HELP manuals of the statistical packages that you use. Please adhere to the HONOR CODE in this instance. Any violation of the HONOR CODE (such as using search engines) will lead to a zero for the exam.

6. SAS OnDemand can be accessed through the following link https://odamid.oda.sas.com.

7. You have six hours for this examination. All six problems will be graded and are of equal weight.

The Problems

1. Suppose A, B, and C are events in a sample space S with $P(A) > 0$. We say that the events B and C are *conditionally independent* (CI) given A if

$$P(B \cap C|A) = P(B|A)P(C|A).$$

Consider the following statements and additionally assume $P(B) > 0$, $P(C) > 0$ and $P(A \cap B) > 0$. If the statement is true, prove it. Otherwise, give a counterexample.

(a) If B and C are CI given A, then $P(C|A \cap B) = P(C|A)$.

(b) If B and C are independent, then they are CI given A.

(c) If B and C are CI given A, then B and C are independent.

2. For $i = 1, ..., n; j = 1, 2$, let

$$\begin{aligned} Y_{ij} &= \alpha + \beta X_j + \epsilon_{ij} \\ \epsilon_{ij} &\sim \quad iid \quad N(0, \sigma^2) \\ X_1 &= -1 \\ X_2 &= +1 \end{aligned}$$

(a) Let $(\widehat{\alpha}, \widehat{\beta})$ be the least squares estimator of (α, β). Show that regardless of whether X is included in the model,

$$\widehat{\alpha} = \overline{\overline{Y}} = \frac{\sum_{j=1}^{2} \sum_{i=1}^{n} Y_{ij}}{2n}.$$

In addition, show that if X is included,

$$\widehat{\beta} = \frac{\overline{Y_2} - \overline{Y_1}}{2},$$

where $\overline{Y_j} = n^{-1} \sum_{i=1}^{n} Y_{ij}$.

(b) For cases when X is/(is not) included in the model, compute $\widehat{Y_{ij}}$, seperately.

(c) Derive the test statistic for testing the null hypothesis $H_0 : \beta = 0$.

(d) Consider the prediction model that uses $a\widehat{\beta}$ rather than $\widehat{\beta}$. That is, $\widehat{Y_{ij}} = \widehat{\alpha} + a\widehat{\beta} X_j$. Show that the MSE of $\widehat{\alpha} + a\widehat{\beta}$ for estimating $(\alpha + \beta)$ (since $|X| = 1$ for predicting the Y_{ij}) is:

$$MSE(\widehat{\alpha} + a\widehat{\beta}) = E[(\widehat{\alpha} + a\widehat{\beta}) - (\alpha + \beta)]^2 = (1 + a^2)\frac{\sigma^2}{2n} + (1 - a)^2\beta^2.$$

(e) If you are choosing between not including $X (a = 0)$ or including $X (a = 1)$ show that you should use $a = 1$ if and only if $\beta^2 > \sigma^2/(2n)$ based on minimizing MSE.

(f) More generally, show that the optimal a for minimizing MSE is:

$$\begin{aligned} a(\sigma^2, \beta, n) &= \frac{t^2}{1 + t^2}, \text{ where} \\ t^2 &= \frac{2n\beta^2}{\sigma^2}. \end{aligned}$$

Comment on the relationship between t obtained here and the test statistic obtained in part (c).

3. Consider two independent random variables, $X_1 \sim N(\cos\mu, \kappa^{-1})$ and $X_2 \sim N(\sin\mu, \kappa^{-1})$, which involve two parameters, $\mu \in [0,\, 2\pi)$ and $\kappa > 0$, with κ^{-1} being the variance. Define $\mathbf{X} = (X_1,\, X_2)^{\mathrm{T}}$. Viewing $\mathbf{X}$ as a random vector, we are interested in its direction, specified by an angle $\theta \in [0,\, 2\pi)$. Denote by r the Euclidean norm of $\mathbf{X}$, that is, $r = \sqrt{X_1^2 + X_2^2}$, and thus $X_1 = r\cos\theta$ and $X_2 = r\sin\theta$.

   (a) Show that the joint probability density function (pdf) of (r, θ) is given by

$$f_{r,\theta}(r,\theta) = \frac{\kappa r}{2\pi}\exp\left[-\frac{\kappa}{2}\left\{r^2 - 2r\cos(\theta-\mu)+1\right\}\right],\ \text{for } r \geq 0 \text{ and } \theta \in [0,\, 2\pi).$$

   (b) Show that the conditional density of θ given r is given by

$$f_{\theta|r}(\theta|r) = \frac{1}{2\pi I_0(r\kappa)}\exp\{r\kappa\cos(\theta-\mu)\},\ \text{for } \theta \in [0,\, 2\pi), \tag{1}$$

   where

$$I_0(t) = \frac{1}{2\pi}\int_0^{2\pi}\exp\{t\cos(\xi-\mu)\}\,d\xi$$

   is the modified Bessel function of order zero. One can show that (you do not need to show this)

$$I_0(t) = \frac{1}{\pi}\int_0^{\pi}\exp(t\cos\xi)d\xi.$$

   (c) Setting $r = 1$ in (1) yields the following density function,

$$f_\theta(\theta) = \frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(\theta-\mu)\}, \tag{2}$$

   which is the pdf of θ as the angle of a bivariate normal vector on the unit circle. This distribution for θ is known as the *circular normal distribution*, with μ as the mean direction and κ as the concentration parameter. Is the family of circular normal distributions an exponential family? Explain.

   (d) Let $(\theta_1, \ldots, \theta_n)$ be an iid sample from the circular normal distribution specified by (2). Find a bivariate sufficient statistic for $(\mu,\, \kappa)$. Is the statistic you find a complete statistic? Explain.

   (e) Suppose it is known that $\mu = 0$. Provide the uniformly most powerful level α test based on a random sample $(\theta_1, \ldots, \theta_n)$ from a mean-zero circular normal distribution for testing the null hypothesis $H_0 : \kappa = 0$ versus the alternative $H_1 : \kappa = 1$. When the null distribution of a test statistic is not one of the named distributions, you may use the following lemma to approximate a percentile of the null distribution.

   Lemma C: If $(\theta_1, \ldots, \theta_n)$ is a random sample from uniform$(0,\, 2\pi)$, then, for a large n, $n^{-1}\sum_{i=1}^n \cos\theta_i$ follows $N(0,\, 1/(2n))$ approximately.

Some trigonometric results that may be helpful for this problem:

$$\begin{aligned}
&\cos(a-b) = \cos a \cos b + \sin a \sin b, \\
&\sin(a+b) = \sin a \cos b + \cos a \sin b, \\
&\sin(-a) = -\sin a,\ \cos(-a) = \cos a, \\
&\frac{d}{da}\sin a = \cos a,\ \frac{d}{da}\cos a = -\sin a, \\
&\sin^2 a + \cos^2 a = 1.
\end{aligned}$$

Probability density functions (pdf) associated with distributions relevant to this problem:

- For $X \sim N(\mu, \sigma^2)$, the pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

- For $X \sim \text{uniform}(a, b)$, the pdf is given by

$$f(x) = \frac{1}{b-a} I(a \leq x \leq b).$$

4. **Low Birth Weight**

The goal of this analysis is to understand the factors associated with birth weight (BWT). Data were collected on 189 women including several variables which were thought to be of importance. Babies that are born weighing less than 2500 grams are considered "low birth weight" (low=1 if birth weight < 2500, 0 otherwise) and are at increase risk of morbidity and mortality. The variables collected in the study are:

- low: Low Birth Weight (0= Birth Weight ≥ 2500 grams, 1=Birth Weight < 2500 grams)
- age: Age of the Mother in Years
- lwt: Mom's Weight in Pounds at the Last Menstrual Period
- race: Race (1 = White, 2 = Black, 3 = Other)
- smoke: Smoking Status During Pregnancy (1 = Yes, 0 = No)
- ptl: History of Premature Labor (0 = None 1 = One, etc.)
- ht: History of Hypertension (1 = Yes, 0 = No)
- ui: Presence of Uterine Irritability (1 = Yes, 0 = No)
- ftv: Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)
- bwt: Babies' Birth Weight in Grams

First 5 observations:

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	4	< 2500 g	28	120	Other	Yes	1	No	Yes	0	709
2	10	< 2500 g	29	130	White	No	0	No	Yes	2	1021
3	11	< 2500 g	34	187	Black	Yes	0	Yes	No	0	1135
4	13	< 2500 g	25	105	Other	No	1	Yes	No	0	1330
5	15	< 2500 g	25	85	Other	No	0	No	Yes	0	1474

Use the low birth weight data to answer the following questions:

(a) Build a good logistic regression model for predicting low birth weight. This should include model selection and interpreting the regression coefficient estimates. Remember to write down your final model.

(b) Given the following characteristics of a pregnant woman: age=30, lwt=120, race=White, ptl=1, ht=0, ui=0, ftv=1, what is the estimated probability of having a low birthweight baby if she does not smoke? What is the estimated probability of having a low birthweight baby if she smokes? Please also report corresponding 95% bootstrap confidence intervals for these two true probabilities.

(c) Report the expected quality of prediction from this model. Compare your final model with one that only has mothers' smoking status using cross-validated area under the ROC curve.

(d) Summarize the results of the analysis in a way that is clear to a subject-matter expert.

5. Consider $Y \sim$ Bernoulli(p) and $X|Y = y \sim N(y\mu_1 + (1-y)\mu_0, 1)$, where $p \in (0,1)$ and $\mu_1, \mu_2 \in \mathbb{R}$. Denote $\theta = (p, \mu_0, \mu_1)'$. Let $(X_1, Y_1), \ldots, (X_K, Y_K)$ be an iid sample from the joint distribution of (X, Y).

   (a) Based on the three expectations $E(Y)$, $E(XY)$, and $E(X(1-Y))$, construct a method of moments (MoM) estimator of θ.

   (b) Find the maximum likelihood estimator (MLE) of θ. You do not need to check second-order conditions. Compare the MoM estimator with the MLE. Which one do you think is better? (*Justify your answer*).

   Now consider that K is an even number; i.e., $K = 2n$ for a positive integer n. Define $W_i = (X_{2i-1} + X_{2i})/2$ and $Z_i = \max(Y_{2i-1}, Y_{2i})$ for $i = 1, \ldots, n$ (*This type of data often arises in the applications that use specimen pooling*). Suppose $\{(X_1, Y_1), \ldots, (X_K, Y_K)\}$ is latent and we only observe $\{(W_1, Z_1), \ldots, (W_n, Z_n)\}$.

   (c) Write down the likelihood function of θ using $\{(W_1, Z_1), \ldots, (W_n, Z_n)\}$ (*Hint: start from the distributions of* $W_i|Z_i = 0$ *and* $W_i|Z_i = 1$. *Note that* $Z_i = 0$ *is equivalent to* $Y_{2i-1} = Y_{2i} = 0$).

6. Let $X_1, \ldots, X_n$ be an independent and identically distributed (iid) sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ an iid sample from $N(\mu_2, \sigma_2^2)$. The two samples are mutually independent. Denote $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.

   (a) Discuss the limiting distribution of $\sqrt{n}(\bar{X}_n \bar{Y}_n - \mu_1 \mu_2)$ as n goes to $+\infty$. When $\mu_1 = \mu_2 = 0$, find the exact distribution of $n(\bar{X}_n \bar{Y}_n - \mu_1 \mu_2)$.

   (b) Assuming $\sigma_1 = \sigma_2 = 1$, develop a size α likelihood ratio test for $H_0 : \mu_1 \times \mu_2 = 0$ versus $H_1 : \mu_1 \times \mu_2 \neq 0$ (*Note that H_0 is true if* ***at least one*** *of μ_1 and μ_2 is zero. This null hypothesis often arises in genetics applications. Recall that a size α test is a test of which the* ***supremum*** *of the probability of Type I error is* ***equal*** *to α. In your solution, you must show that your test is a size α test.*)

   (c) Prove that the power function of the likelihood ratio test derived in part (b) approaches 1 as n goes to infinity, if $\mu_1 \times \mu_2 \neq 0$.