Written Comprehensive Examination - Theory

Department of Statistics, UC Irvine

Friday, June 16, 2023, 9:00 am to 1:00 pm

- There are 7 questions on the examination. Select any 5 of them to solve. If you attempt to solve more than 5 questions, you are only to turn in the 5 you want graded. If you turn in partial solutions to more than 5 questions, only 5 will be graded.
- Each of the 5 problems you attempt to solve will be worth equal credit, with each accounting for 20% of your final score on this examination.
- Your solutions to each problem should be written on separate sheets of paper **one side only with the backside left blank.** Label each sheet with your identifier number **emailed to you by the Department Manager Laura Swendson**, the problem number, and the page number of that solution written in the upper right hand corner. For example, the labeling on a page may be:

  ID# 8267

  Problem 2, page 3
- You have 4 hours to complete your solution. Please be prepared to turn in your exam at 1:00pm.

1. Let X and Y be independent random variables with $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$. Define random variables Z and W such that $Z = \min\{X, Y\}$ and $W = 1$ if $X \leq Y$ and $W = 0$ if $X > Y$.

   (a) Compute $\Pr(W = 1)$ by using double integrals.

   (b) Compute $\Pr(Z \leq z)$ where $z > 0$ is a constant. Identify the marginal distribution of Z by name and parameter.

   (c) Compute the joint probability $\Pr(Z \leq z, W = 1)$ where $z > 0$ is a constant. Do the same for $\Pr(Z \leq z, W = 0)$.

   (d) Show that W and Z are independent.

(End of Problem 1)

2. The joint pdf of (X, Y) is

$$f(x,y) = \frac{\lambda}{\sqrt{2\pi}} \exp\left\{-\frac{(y-\theta x)^2}{2} - \lambda x\right\}, \quad x > 0,\ -\infty < y < \infty.$$

where $\lambda > 0$ and $\theta \in (-\infty, \infty)$ are parameters.

(a) Find the marginal pdf of X and identify this well known distribution by name and parameter.

(b) Find the conditional mean and variance of Y given X.

(c) Suppose we observe $(X_j, Y_j),\ j = 1, \ldots, n$, iid (independent and identically distributed) with this joint distribution. Identify a three-dimensional sufficient statistic for (θ, λ).

(d) In the setting of part (c) derive the maximum likelihood estimator of (θ, λ).

(End of Problem 2)

3. Suppose $X_j \sim \text{Exp}(\lambda)$ and $Y_j|X_j \sim \text{Poisson}(\theta X_j)$, independently for $j = 1, \ldots, n$. The parameters λ and θ are both positive.

   (a) Compute Fisher's expected information for (λ, θ) based on the n bivariate samples $(X_j, Y_j), j = 1, \ldots, n$. (Hint: this is a 2×2 matrix.)

   (b) Show that the estimator $\tilde{\theta} \equiv \sum_{j=1}^{n} Y_j / \sum_{j=1}^{n} X_j$ is consistent for θ.

   (c) Is the estimator $\tilde{\theta}$ in part (b) unbiased for θ? Explain.

   (d) Is the estimator $\tilde{\theta}$ in part (b) asymptotically normal for θ? Justify.

   (e) Describe how you may obtain an approximate 95% confidence interval for θ, valid for large n.

(End of Problem 3)

4. Suppose $Y_j|(\mu_0, \sigma^2) \sim \mathrm{N}(\mu_0, \sigma^2)$ independently for $j = 1, \ldots, n$. Assume μ_0 is known and $\sigma^2 > 0$ is the only unknown parameter.

   (a) Derive the likelihood ratio test for testing $H_0 : \ \sigma^2 = \sigma_0^2$ vs $H_1 : \ \sigma^2 \neq \sigma_0^2$ where $\sigma_0^2 > 0$ is a constant. Explain how you determine the rejection region to achieve level $\alpha = 0.05$ based on the *exact* distribution of your test statistic.

   (b) Derive a 95% confidence interval for σ^2 based on the statistic $T \equiv \sum_{j=1}^{n}(Y_j - \mu_0)^2$.

   (c) Consider a Bayesian approach under the prior $p(\sigma^2) \propto 1/\sigma^2$. Derive the posterior density of σ^2 and that of $\delta \equiv 1/\sigma^2$. Express the (unnormalized) posterior density of δ in the form of $\delta^{U-1}e^{-\delta V}$ and specify the statistics U and V.

   (d) In the setting of part (c), describe how you may find a 95% highest probability density (HPD) posterior credible interval for δ. (You may use the form of the posterior density of δ in part (c) even if you did not complete part (c).)

   (e) In the setting of part (c), find the Bayes estimator of $\delta \equiv 1/\sigma^2$ under squared error loss. (You may use the form of the posterior density of δ in part (c) even if you did not complete part (c).)

(End of Problem 4)

5. Consider the two-way model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where $i = 1, 2; j = 1, 2$ and $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma^2)$. Note that there is only one observation per cell.

(a) Express the model in the matrix form

$$\mathbf{Y} = \mathbf{X} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \epsilon$$

Note that $rank(X) = 3$ (you do not need to show this part).

(b) Show that μ is NOT estimable.

(c) Consider $m\mu + a_1\alpha_1 + a_2\alpha_2 + b_1\beta_1 + b_2\beta_2$. Assuming that this linear function is estimable, prove that m, the $a_i's$, and the $b_j's$ must satisfy $m = a_1 + a_2 = b_1 + b_2$.

(d) State the Gauss-Markov theorem.

(e) Suppose that we are interested in estimating $2\mu + \alpha_1 + \alpha_2 + 1.5\beta_1 + 0.5\beta_2$. To use Gauss-Markov theorem to find its BLUE, we need to find least-square estimates (LSE) of the parameters. Follow the instructions below to find LSE of $(\mu, \alpha_1, \alpha_2, \beta_1, \beta_2)$. Then find the BLUE of $2\mu + \alpha_1 + \alpha_2 + 1.5\beta_1 + 0.5\beta_2$.

i. Parameterize the model using parameters $\mu, \alpha_1, \alpha_2, \beta_1 - \beta_2, \beta_1 + \beta_2$ then find the corresponding design matrix, which is denoted by $\tilde{X}$.

ii. Note that the rank of the design matrix $\tilde{X}$ is also 3. To find an LSE of $(\mu, \alpha_1, \alpha_2, \beta_1 - \beta_2, \beta_1 + \beta_2)$, we can use a "subset" method by deleting the first and the last columns. In other words, we set $\hat{\mu} = 0, \hat{\beta}_1 + \hat{\beta}_2 = 0$ and find LSE of $\alpha_1, \alpha_2, \beta_1 - \beta_2$. This sounds complicated but you will notice that the three remaining columns are orthogonal with each other.

(End of Problem 5)

6. Segmented regression is a special type of regression in which an explanatory variable is partitioned into intervals and a separate line segment is fitted to each interval. Consider a two-segment linear model. This can be formulated as a linear model with a known "breakpoint", denoted by x_0.

$$Y_i = \begin{cases} \alpha_2 + \beta_2 x_i + \epsilon_i & \text{if } x_i > x_0 \\ \alpha_1 + \beta_1 x_i + \epsilon_i & \text{if } x_i \leq x_0 \end{cases}$$

where $i = 1, \cdots, n$ and $\epsilon_1, \cdots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Without loss of generality (also for convenience), we assume that $x_1 \leq x_0, \cdots, x_m \leq x_0, x_{m+1} > x_0, \cdots, x_n > x_0$.

(a) Rewrite the model in the form $Y = X\theta + \epsilon$, where X is the $n \times 4$ design matrix and $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2)^T$. Specify the design matrix as explicitly as possible.

(b) Examine the design matrix X and the matrix $X^T X$. Explain why the least squares estimate of θ is equivalent to fitting two separate regression lines for the first m and the remaining $n - m$ points separately.

(c) Let RSS_{full} denote the residual sum of squares for the two-segment linear regression. Derive the distribution of RSS_{full}/σ^2.

(d) Based on the model above, $E(Y|x_0) = \alpha_1 + \beta_1 x_0$ but $E(Y|x_0^+) = \alpha_2 + \beta_2 x_0$ where $E(Y|x_0^+)$ is the limit of $E(Y|x)$ as x approaches x_0 from the right. Therefore, the mean response curve $E(Y|x)$ is not necessarily continuous at x_0. We would like to test whether this curve is continuous at x_0. Find a pair of A and c such that the null hypothesis of continuity can be expressed as $A\theta = c$.

(e) Construct an F test to test the hypothesis of a continuous mean response curve. This can be done by either (1) using the distribution of $\hat{\theta}$ or (2) comparing the residual sum of squares (RSS) of a reduced model to RSS_{full}. If you choose (1), describe the distribution of $\hat{\theta}$ and the null distribution of $A\hat{\theta} - c$. If you choose (2), describe the reduced model using parameters α_1, β_1, and β_2. In both cases, provide the null distribution of your F-statistic.

(End of Problem 6)

7. Assume that $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) with density

$$f(x) = \frac{e^{-x/\mu}}{\mu}$$

for $x > 0$, where $\mu > 0$ is unknown (exponential distribution with mean μ). Recall that the variance of this distribution is μ^2.

(a) What is the maximum likelihood estimator (MLE) δ_n of μ^2 based on the first n observations $X_1, X_2, \ldots, X_n$? How do you know it is the MLE?

(b) Prove that δ_n is a consistent sequence of estimators for μ^2, i.e., δ_n converges in probability to μ^2 as $n \to \infty$.

(c) Find the limiting distribution as $n \to \infty$ of $\sqrt{n}\left(\delta_n - \mu^2\right)$.

Table 1: Common distributions and densities.

Distribution	Notation	Density
Bernoulli	$\text{Bern}(\theta)$	$f(y\|\theta) = \theta^y(1-\theta)^{1-y}$
Binomial	$\text{Bin}(n,\theta)$	$f(y\|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$
Multinomial	$\text{Multi}(n;\theta_1,\theta_2,\ldots,\theta_K)$	$f(y\|\theta) = \frac{n!}{y_1!y_2!\ldots y_K!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_K^{y_K}$
Beta	$\text{Beta}(a,b)$	$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)$
Uniform	$U(a,b)$	$p(\theta) = \frac{I_{(a,b)}(\theta)}{b-a}$
Poisson	$\text{Pois}(\theta)$	$f(y\|\theta) = \theta^y e^{-\theta}/y!$
Exponential	$\text{Exp}(\theta)$	$f(y\|\theta) = \theta e^{-\theta y}I_{(0,\infty)}(y)$
Gamma	$\text{Gamma}(a,b)$	$p(\theta) = [b^a/\Gamma(a)]\theta^{a-1}e^{-b\theta}I_{(0,\infty)}(\theta)$
Chi-squared	$\chi^2(n)$	Same as $\text{Gamma}(n/2, 1/2)$
Weibull	$\text{Weib}(\alpha,\theta)$	$f(y\|\theta) = \theta\alpha y^{\alpha-1}\exp(-\theta y^\alpha)\, I_{(0,\infty)}(\theta)$
Normal	$N(\theta, 1/\tau)$	$f(y\|\theta,\tau) = (\sqrt{\tau/2\pi})\exp[-\tau(y-\theta)^2/2]$
Student's t	$t(n,\theta,\sigma)$	$f(y\|\theta) = [1+(y-\theta)^2/n\sigma^2]^{(n+1)/2}$
		$\times\, \Gamma[(n+1)/2]/\Gamma(n/2)\sigma\sqrt{n\pi}$
Cauchy	$\text{Cauchy}(\theta)$	same as $t(1,\theta,1)$
Dirichlet	$\text{Dirichlet}(a_1,a_2,a_3)$	$p(\theta) = \Gamma(a_1+a_2+a_3)/\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)$
		$\times\, \theta_1^{a_1-1}\theta_2^{a_2-1}(1-\theta_1-\theta_2)^{a_3-1}$
		$\times\, I_{(0,1)}(\theta_1)I_{(0,1)}(\theta_2)I_{(0,1)}(1-\theta_1-\theta_2)$

Table 2: Means, Modes, and Variances.

Distribution	Mean	Mode	Variance
Bern(θ)	θ	0 if $\theta < .5$ 1 if $\theta > .5$	$\theta(1-\theta)$
Bin(n, θ)	$n\theta$	integer closest to $n\theta$	$n\theta(1-\theta)$
Beta(a, b)	$a/(a+b)$	$(a-1)/(a+b-2)$ if $a > 1, b \geq 1$	$ab/(a+b)^2(a+b+1)$
$U(a, b)$	$.5(a+b)$	everything a to b	$(b-a)^2/12$
Pois(θ)	θ	integer closest to θ	θ
Exp(θ)	$1/\theta$	0	$1/\theta^2$
Gamma(a, b)	a/b	$(a-1)/b$ if $a > 1$	a/b^2
$\chi^2(n)$	n	$n-2$ if $n > 2$	$2n$
Weib(α, θ)	$\Gamma[(\alpha+1)/\alpha]/\theta$	$[(\alpha-1)/\alpha]^{1/\alpha}/\theta$	$\Gamma[(\alpha+2)/\alpha] - \mu^2$
$N(\theta, 1/\tau)$	θ	θ	$1/\tau$
$t(n, \theta, \sigma)$	θ if $n \geq 2$	θ	$\sigma^2 n/(n-2)$ if $n \geq 3$
Cauchy(θ)	Undefined	θ	Undefined