Written Comprehensive Examination - Theory

Department of Statistics, UC Irvine

Friday, June 18, 2021, 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. Label each sheet with your student identification number, 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# 912346378

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

1. A task can be done using any of three different machines.

   - Machine A involves one single step, with the time (X) completing the step following an exponential distribution with the mean time being 2 hours, i.e., $f_X(x) = \frac{1}{2}e^{-x/2}, x > 0$, where $f_X(x)$ denotes the pdf of X.
   - Machine B requires two steps with the time for completing step i follows an exponential distribution with mean 1 hour, i.e., $f_{Y_i}(y_i) = e^{-y_i}, y_i > 0, i = 1, 2$. We assume that Y_1 and Y_2 are independent.
   - Machine C uses two independent engines and the job will be done as long as one engine completes the job. Let Z_i (where $i = 1, 2$) be the time required by the ith engine. It is known that Z_i follows an exponential distribution with mean 4 hours, i.e., $f_{Z_i}(z_i) = \frac{1}{4}e^{-z_i/4}, z_i > 0, i = 1, 2$.

   (a) Let Y be the time required for Machine B to complete a job. Find the mean and variance of Y. Between Machine A and Machine B, which one is more reliable in terms of variance?

   (b) Argue that, from a statistical point of view, there is no difference between Machine A and Machine C.

   (c) Between Machine A and Machine B, which one is more likely to finish a job within 2 hours? Answer this question by computing $Pr(X < 2)$ and $Pr(Y < 2)$.

   (d) Suppose there are 10 jobs on a certain day and the probability that a job is assigned to Machine B is 10%. Let T be the total time needed to finish the 10 jobs. Find $E(T)$ and $Var(T)$.

(End of Problem 1)

2. Let $Y_1, Y_2, \cdots$ be a sequence of independent random variables with $Y_i \sim \text{Binomial}(i, 1/i), i = 1, 2, \cdots$.

   (a) Use moment generating function (mgf) to show that the sequence converges in distribution to X where X follows a Poisson distribution.

   (b) Show that $Var(\bar{Y}_n) < \frac{1}{n}$, where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

   (c) Give the definition of convergence in probability.

   (d) Chebyshev's inequality says

$$P[g(X)) \geq r] \leq \frac{E[g(X)]}{r},$$

   where $g(\cdot)$ is a non-negative function and X is a random variable. Use Chebyshev's inequality to prove that $\bar{Y}_n$ converges in probability to 1.

   (e) Based on (d), what can you say about $\sqrt{\bar{Y}_n}$? Justify your answer.

(End of Problem 2)

3. Let $X_1, \ldots, X_n$ be a random sample from a distribution with the following pdf:

$$f_X(x|\theta) = \frac{x}{\theta} \exp\left(-\frac{x^2}{2\theta}\right),$$

where $x > 0$ and $\theta > 0$. The mean and the variance of the distribution are

$$E(X) = \sqrt{\frac{\pi\theta}{2}}, \quad \text{Var}(X) = \frac{4-\pi}{2}\theta.$$

(a) Find a minimal sufficient statistic for θ.

(b) Find the method of moment estimator for θ based on the first moment, and show if it is an unbiased estimator or not.

(c) Find the maximum likelihood estimator for θ, and show if it is an unbiased estimator or not.

(d) Calculate the Cramer-Rao lower bound for unbiased estimators of θ, and show if the best unbiased estimator of θ attains the Cramer-Rao lower bound. If yes, find the best unbiased estiomator. Otherwise, justfy your answer.

(e) Calculate the Cramer-Rao lower bound for unbiased estimators of θ^2, and show if the best unbiased estimator of θ^2 attains the Cramer-Rao lower bound? If yes, find the best unbiased estimator. Otherwise, justify your answer.

(End of Problem 3)

4. Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population. Consider testing

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

(a) If σ^2 is known, show that the test that rejects H_0 when

$$\bar{X} > \theta_0 + z_\alpha \sqrt{\sigma^2/n}$$

is a test of size α, where $\bar{X}$ is the sample meant and z_α is the upper 100α-th percentile of a $N(0, 1)$ distribution.

(b) Show that the test in (a) can be derived as a likelihood ratio test (LRT).

(c) Show that the test in (a) is a uniformly most powerful (UMP) test.

(d) If σ^2 is unknown, show that the test that rejects H_0 when

$$\bar{X} > \theta_0 + t_{n-1,\alpha} \sqrt{S^2/n}$$

is a test of size α, where S^2 is the sample variance and $t_{n-1,\alpha}$ is the upper 100α-th percentile of a T_{n-1} distribution.

(e) Show that the test in (d) can be derived as an LRT.

(f) Now assume $X_1, \ldots, X_n$ are not necessarily generated from a normal distribution, but from some unknown distribution with mean θ and variance σ^2. Show that the test in (d) is a test of size α asymptotically.

(End of Problem 4)

5. Use one-way ANOVA as an example to describe and explain

   (a) over-parameterization;

   (b) estimable and non-estimable functions;

   (c) identifiability constraints;

   (d) Gauss-Markov theory.

(End of Problem 5)

6. In statistical modeling, one frequently encountered problem is nuisance parameters. For example, in linear models we might have covariates whose coefficients can be considered as nuisance parameters. Although their effects are not of primary interest, failing to adjust for these effects will lead to biased estimates and violates the exchangeability assumption that is required by permutation tests, which are special nonparametric resampling tests. Let's consider a linear model

$$Y = X\beta + Z\gamma + \epsilon, \epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

where $\mathbf{I}_n$ is the identity matrix of size n, Y is a random vector of length n, $Z_{n\times 1}$ and $X_{n\times p}$ are fixed matrices with $rank(X) = p \geq 1$, $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}$ are the coefficients of X and Z, respectively. The parameter of interest is γ.

(a) One intuitively reasonable approach is to first regress Y on X to obtain the residuals, denoted by e, and then regress e on Z to make inference about γ. Show that $e = (\mathbf{I}_n - P)Y$ where $P = X(X^TX)^{-1}X^T$.

(b) Find the variance-covariance of e and show that the n residuals are not independent.

The correlated residuals make the above procedure not attractive. A more interesting approach is to find an $n \times (n-p)$ transformation matrix G such that $GG^T = \mathbf{I}_n - P$ and $G^TG = \mathbf{I}_{n-p}$. In class we showed the existence of such matrix G for a projection matrix. Based on this, answer questions (c)-(e).

(c) Let $\tilde{Y} = G^TY$ and $\tilde{Z} = G^TZ$. Show that $\tilde{Y} \sim N(\tilde{Z}\gamma, \sigma^2\mathbf{I}_{n-p})$.

(d) Let $\tilde{\gamma}$ denote the LSE of γ based on $\tilde{Y}$ and $\tilde{Z}$. It is not difficult to find that the the residual sum of squares is

$$RSS = \tilde{Y}^T(\mathbf{I}_{n-p} - \frac{\tilde{Z}\tilde{Z}^T}{\tilde{Z}^T\tilde{Z}})\tilde{Y}.$$

Show that $RSS/\sigma^2 \sim \chi^2_{n-p-1}$.

(e) Based on the results in (c) and (d), describe how you can construct a $100(1-\alpha)\%$ C.I. for γ. Note that σ^2 is unknown and has to be estimated from $\tilde{Y}$ and $\tilde{Z}$.

(End of Problem 6)

7. Mitochondria are cell organelles that generate energy for cells to perform their functions. Mitochondria have their own DNA, distinct from the rest of the cell DNA that is stored in the cell nucleus. Mitochondrial DNA sequences mutate faster that the nucleic DNA and, as a result, have more information about evolutionary forces that shaped genetic diversity we observe today. Suppose we have aligned mitochondrial DNA sequences from humans and chimpanzees. We store these aligned sequences in a $2 \times L$ matrix:

$$\mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1L} \\ y_{21} & y_{22} & \cdots & y_{2L} \end{pmatrix},$$

where $y_{ij} \in \{A, G, C, T\}$ and L is the sequence alignment length. We would like to estimate divergence time of these two species (time to the most recent common ancestor of the two species) using this DNA alignment. First, to make this problem tractable, we assume a known mutation rate $\alpha = 0.025$, measured in the number of mutations, per genomic location, per million years. One simple model of DNA mutational process, known as a Jukes-Cantor model, says that the sequence alignment can be condensed into a sufficient statistic $N_1 = \sum_{i=1}^{L} 1_{\{y_{1i} \neq y_{2i}\}}$ — the number of variable columns in the sequence alignment. Under the Jukes-Cantor model, columns in the sequence alignment are independent and identically distributed, and

$$Pr(y_{1i} \neq y_{2i}) = \frac{3}{4}\left(1 - e^{-\alpha t}\right),$$

where t is the human/chimpanzees divergence time, in millions of years.

(a) Show that the maximum likelihood estimator (MLE) of the divergence time is

$$\hat{t} = -\frac{1}{\alpha} \ln\left(1 - \frac{4}{3}\frac{N_1}{L}\right).$$

What conditions should N_1 and L satisfy for this MLE to exist?

(b) Use observed Fisher information to derive the asymptotic variance of $\hat{t}$.

(c) Instead of the Fisher information, use asymptotic normality of N_1/L and the delta method to derive the asymptotic variance of $\hat{t}$.

(d) Use $\alpha = 0.025$, $N_1 = 975$, and $L = 9993$ to obtain the MLE and an asymptotic 95% confidence interval for the divergence time t.

(End of Problem 7)

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