Ph.D. Qualifying Exam
Thursday–Friday, August 20–21, 2020

- **Begin your solution to each problem on a new sheet of paper.**
- **Statistics PhD students should do the 5106 problems.**
- **Biostatistics PhD students should do the 5198 problems.**
- **All students should do the 5166 and 5167 problems.**

Problem 1. (5106) Let $x = (x_1, \ldots, x_n)$ be a given binary, Markovian sequence. In particular,

$$P(x_1 = 0) = 0.5\,, \quad P(x_1 = 1) = 0.5\,,$$

and

$$P(x_i = x_{i-1}) = p\,, \quad P(x_i = 1 - x_{i-1}) = 1 - p\,, \quad i = 2, \ldots, n\,.$$

Let $y = (y_1, \ldots, y_n)$ be a noisy observation of x. That is,

$$y_i = x_i + e_i,$$

with $e_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$.

Assuming p and σ^2 known, we can use the Maximum A Posteriori method to reconstruct x from y in the following form:

$$\hat{x} = \text{argmax}_{\{x_i\}} \sum_{i=1}^{n} (-\frac{(y_i - x_i)^2}{2\sigma^2}) + \sum_{i=2}^{n} \log(1_{x_i = x_{i-1}} p + 1_{x_i \neq x_{i-1}}(1-p))$$

Write out a pseudocode for a Dynamic Programming procedure to compute $\hat{x}$ in the computational order of n.

Problem 2. (5106) Consider a data set of observations $\{x_n\}$ where $n = 1, \ldots, N$, and x_n is a Euclidean variable with dimensionality D. Our goal is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data. Prove that the optimal linear projection for which the variance of the projected data is maximized is defined by the M eigenvectors $U_1, \ldots, U_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, \ldots, \lambda_M$.

Problem 3. (5166) In a one-way layout experiment, consider the following linear model:

$$Y_{ti} \;=\; \mu + \tau_t + \epsilon_{ti}, \quad t = 1, \ldots, k; \;\; i = 1, \ldots, n_t,$$

where μ is the overall mean and $\{\tau_t, t = 1, \ldots, k\}$ are the effects of the treatments. The random noise $\{\epsilon_{ti}, t = 1, \ldots, k; i = 1, \ldots, n_t\}$ follows the model:

$$\epsilon_{ti} \;=\; a_{ti} + \theta a_{t,i-1},$$

where the a_{ti}'s are independent and identically distributed random variables with mean zero and variance σ^2.

(a) Let $N = \sum_{t=1}^{k} n_t$, $\bar{Y}_{t\cdot} = n_t^{-1} \sum_{i=1}^{n_t} Y_{ti}$, and $\bar{Y} = N^{-1} \sum_{t=1}^{k} \sum_{i=1}^{n_t} Y_{ti}$. Calculate the expected values and variances of $\bar{Y}_{t\cdot}$ and $\bar{Y}$.

(b) Find $\mathrm{Cov}(\bar{Y}_{t\cdot}, \bar{Y})$.

(c) Let $S_T = \sum_{t=1}^{k} n_t (\bar{Y}_{t\cdot} - \bar{Y})^2$, $S_R = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (Y_{ti} - \bar{Y}_{t\cdot})^2$, and $S_D = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (Y_{ti} - \bar{Y})^2$. Show that $S_D = S_T + S_R$. Find the expected value of S_T.

Problem 4. (5166) A study was conducted to determine whether the age of customers is related to the type of movie he or she watches. A researcher went to a movie store and observed all of the 450 movie rentals that took place during one day. These movie rentals were categorized by customer age and movie type, and the results are displayed in the contingency table given below. Assume that the cell frequencies in this table are independent Poisson variables.

Age	Documentary	Comedy	Mystery
12–20	13	49	38
21–40	34	86	60
41 and over	73	50	47

(a) Conditional on the total sample size being $n = 450$, what is the distribution of the nine categories? Show your justifications. What are the mean and variance of each cell frequency?

(b) Run a chi-square test of independence and draw your conclusion. Use $\alpha = 0.05$.

(c) Run a chi-square test of "Comedy" versus "Mystery". Use $\alpha = 0.05$.

Note: A small chi-square table is given on page 7.

Problem 5. (5167) Suppose we fit a regression of Y on (X, Z) with the true mean function $\mathrm{E}(Y \mid X = x, Z = z) = 2 + 3x - 4z$.

(a) If X and Z are independent, what are the expected results from fitting a simple linear regression of Y on X? Assume that $\mathrm{Var}(Y \mid X = x, Z = z) = \sigma^2$ is constant. (Hint: see how $E(Y|X)$ and $\mathrm{Var}(Y|X)$ depend on Z.)

(b) Provide conditions under which the mean function for $\mathrm{E}(Y \mid X)$ is linear but has a negative coefficient for X.

(c) Assume that (X, Z) is bivariate normal, with the five parameters $(\mu_x, \mu_z, \sigma_x^2, \sigma_z^2, \rho_{xz}) = (0, 1, 1, 2^2, 0.8)$. Derive the regression model parameters of X on Z.

(d) Suppose we collected data $(X_i, Y_i, Z_i = 1)$, $i = 1, \ldots, n$, and $(X_i, Y_i, Z_i = 2)$, $i = n+1, \ldots, 2n$, and wanted to estimate the mean function $E(Y \mid X = x, Z = z)$. If we assume that $\mathrm{Var}(Y \mid X = x, Z = z) = 4z^2$, what would be the RSS of weighted least squares and the solution from minimizing it?

Problem 6. (5167) Consider a heteroscedastic linear model with data consisting of n independent copies of $(Y, \mathbf{X}, W)$, where $Y \in \mathbb{R}^1$, $\mathbf{X} \in \mathbb{R}^p$, and $W > 0$ is a weight with expectation one:

$$Y = \mu + \boldsymbol{\beta}^T \mathbf{X} + \epsilon / \sqrt{W},$$

where ϵ is independent of (X, W), and $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$.

(a) Discuss why we can assume $\mathrm{E}(W) = 1$ without loss of generality.

(b) Describe the estimation procedure for the parameters μ, $\boldsymbol{\beta}$ and σ^2.

(c) Provide a likelihood justification for the previous step. (Hint: by assuming $\epsilon \sim$ Normal)

(d) If $\mathbf{X} \mid W \sim \mathrm{N}(\boldsymbol{\mu}_x, W^{-1}\boldsymbol{\Sigma})$, what would be a good estimator for $\boldsymbol{\Sigma}$?

Begin your solution to each problem on a new sheet of paper.

Problem 9. (5326) Let (X, Y) have a uniform distribution inside the circle of radius 2 about the origin, that is, (X, Y) has joint density

$$f(x,y) = \begin{cases} C & \text{if } x^2 + y^2 \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

for some constant C. Answer the following. Justfy your answers.

(a) Find $f_X(x)$, the marginal density of X.

(b) Find $E(Y \mid X = x)$ for $-2 < x < 2$.

(c) Are X and Y independent?

(d) Find $\text{Corr}(X, Y)$, the correlation between X and Y.

(e) Find $\text{Corr}(X^2, Y^2)$, the correlation between X^2 and Y^2.
(Hint: For non-negative integers r, s, the quantity $EX^{2r}Y^{2s}$ may be expressed (after some manipulations) as a Beta integral. Another possibility is to use polar coordinates.)

Problem 10. (5326) An urn contains R red balls, G green balls, and B blue balls. A player randomly selects **9** balls, doing this one by one and with**OUT** replacement. The player wins a dollar every time he selects three red balls in a row. (Assume $R \geq 3$ and $R + G + B \geq 9$.) To be more precise, a player receives a dollar after the i-th draw, if he selected red balls on draws $i - 2$, $i - 1$, and i. Note that under these rules, a player drawing 4 red balls in a row receives a total of 2 dollars; a player drawing 5 red balls in a row receives a total of 3 dollars, etc. Let X be the player's total winnings.

(a) Find EX.

(b) Find EX^2.

The following table will be useful for solving the STA 5327 problems.

Name	**Notation**	$f(x)$	$E(X)$	Var(X)
Normal	$\mathrm{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2

Problem 11. (5327) Consider the probability density function:

$$f_\nu(x; \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})\sigma} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \quad \text{for } x \in \mathbb{R},$$

where $\nu > 2$ is a known constant, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters, and $\Gamma(\cdot)$ is the Gamma function. We observe i.i.d. observations $X_1, \ldots, X_n \sim f_\nu(x; \mu, \sigma^2)$.

(a) Find an equation system that the maximum likelihood estimator of (μ, σ^2) must satisfy. (But you do not need to solve the equation system.)

(b) Let $(\hat{\mu}, \hat{\sigma}^2)$ be the maximum likelihood estimator of (μ, σ^2). Determine whether $\hat{\mu}$ and $\hat{\sigma}^2$ are asymptotically independent.

Problem 12. (5327) Consider $X_1, \ldots, X_n$ i.i.d from $\frac{1}{2}N(\mu, 1) + \frac{1}{2}N(-\mu, 1)$, where $\mu \geq 0$.

(a) Find the method of moments estimator of μ. Discuss when it exists.

(b) Design a test for the following hypotheses with Type I error at 0.05:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu > 0.$$

Problem 13. (6346) Let X_n be a martingale with respect to $\mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n)$. Suppose $X_n \in L_2$ and S and T are bounded stopping times with $S \leq T$.

(a) Give the definition of a martingale.

(b) Give the definitions of a stopping time and a bounded stopping time.

(c) Show X_S and X_T are in L_2.

(d) Show $E[(X_T - X_S)^2 | \mathcal{F}_S] = E[X_T^2 - X_S^2 | \mathcal{F}_S]$.

Problem 14. (6346) For the problems below, you may assume that X_n and X are random variables with all moments finite.

(a) Define convergence in probability.

(b) Define convergence in L_p.

(c) Prove or give a counterexample: $X_n \xrightarrow{L_p} X \Rightarrow X_n \xrightarrow{P} X$ for finite p.

(d) Prove or give a counterexample: $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L_p} X$ for finite p.