Ph.D. Qualifying Exam
Wednesday–Thursday, August 21–22, 2019

- **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) 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 2. (5106) Find the maximum likelihood estimate of θ where θ is the parameter in the multinomial distribution:

$$(x_1, x_2, x_3, x_4) \sim M(n; 0.1(1-\theta), 0.4(2-3\theta), 0.1(1+3\theta), \theta)$$

(a) Choose a variable for the missing data and use the EM algorithm for iteratively estimating θ. Let $\theta^{(m)}$ be the current value of the unknown. Derive the mathematical formula to update for $\theta^{(m+1)}$.

(b) Let $L(\theta)$ denote the likelihood with parameter θ. Prove that

$$L(\theta^{(m+1)}) \geq L(\theta^{(m)}).$$

Problem 3. (5166) Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample (iid) from the $N(\mu, \sigma^2)$ distribution.

(a) Show that $\sum_{i=1}^n (X_i - \bar{X})^2$ can be expressed in the form: $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=2}^n Y_i^2$ where $\{Y_2, Y_3, \ldots, Y_n\}$ are iid random variables with the $N(0, \sigma^2)$ distribution.

(b) Show that $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$.

(c) Let $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. What is the limit distribution of $\sqrt{n-1}(s_n^2 - \sigma^2)$ when n increases to ∞?

Problem 4. (5166) The values in the following table are the burning times (rounded to the nearest tenth of a minute) of random samples of two kinds of emergency flares:

Observation ID	1	2	3	4	5	6	7	8	9	10
Brand A: y_1	15.2	13.1	15.3	28.1	2.5	13.7	15.8	17.2	22.3	12.5
Brand B: y_2	16.6	18.1	14.8	18.6	10.0	26.5	11.3	13.5	12.9	42.5

(a) Assume that the two random samples $\{y_{1i}, i = 1, \ldots, 10\}$ and $\{y_{2i}, i = 1, \ldots, 10\}$ are independent, normally distributed, and with means μ_1 and μ_2 and the same variance σ^2. State your null and alternative hypotheses, and perform a test on the means of the two populations. Test using $\alpha = 0.05$.

(b) Identify any potential outliers in the two samples and give your justifications.

(c) State your null and alternative hypotheses, and test the two distributions using the Wilcoxon Rank-Sum test that is robust against outliers. Test using $\alpha = 0.05$.

Problem 5. (5167) Consider the following simple linear regression model of Y on X_1:

$$Y = \alpha + \beta X_1 + e.$$

Suppose $n = 45$, $\bar{X}_1 = -3.3$, $\bar{Y} = 5.5$, $SXX = 100.0$, $SYY = 123.4$, and $SXY = 88.8$.

(a) Find the following: RSS, R^2, $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2 \equiv \widehat{var}(e)$.

(b) Test the hypothesis (at level 0.05) that $\beta = 1$ versus $\beta \neq 1$.

(c) What assumptions are required for the test in part (b)?

(d) Find the 95% prediction interval for a new observation $X_1^\star = -2$.

(e) Consider the following regression model,

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\text{var}(\epsilon_i) = \sigma^2$ for all i, and $\text{Corr}(\epsilon_i, \epsilon_j) = \rho > 0$ for $i \neq j$. How would you estimate μ? What is the variance of your estimator; and does it go to zero when $n \to \infty$?

Some R output that might be helpful.

```
> qt(c(0.95,0.975,0.99,0.995),42)
[1] 1.681952 2.018082 2.418470 2.698066
> qt(c(0.95,0.975,0.99,0.995),43)
[1] 1.681071 2.016692 2.416250 2.695102
> qt(c(0.95,0.975,0.99,0.995),44)
[1] 1.680230 2.015368 2.414134 2.692278
```

Problem 6. (5167) Suppose we have a univariate response Y and a multivariate predictor $X \in \mathbb{R}^p$. We are interested in the following regression model,

$$Y = \beta_0 + \beta^T X + \varepsilon = \widetilde{\beta}^T \widetilde{X} + \varepsilon,$$

where $\widetilde{\beta} = (\beta_0, \beta^T)^T \in \mathbb{R}^{p+1}$ and $\widetilde{X} = (1, X^T)^T \in \mathbb{R}^{p+1}$. Let $Y_n \in \mathbb{R}^{n\times 1}$, $X_n \in \mathbb{R}^{n\times p}$ and $\widetilde{X}_n = (1_n, X_n) \in \mathbb{R}^{n\times(p+1)}$ be the data matrices, where 1_n is a vector of ones. You may use this result to answer the questions: Suppose $M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ is a $m \times m$ symmetric positive definite matrix, then its inversion can be written as

$$M^{-1} = \begin{pmatrix} A^{-1} - A^{-1}BDB^TA^{-1} & -A^{-1}BD \\ -DB^TA^{-1} & D \end{pmatrix},$$

where $D = (C - B^TA^{-1}B)^{-1}$.

(a) Write down the ordinary least squares estimates for $\widetilde{\beta}$, β_0 and β using the given notation.

(b) Let $\overline{X}$ and S be the sample mean vector and sample covariance matrix of X. Show that

$$(\widetilde{X}_n^T\widetilde{X}_n)^{-1} = \begin{pmatrix} n^{-1} + (n-1)^{-1}\overline{X}^TS^{-1}\overline{X} & -(n-1)^{-1}\overline{X}^TS^{-1} \\ -(n-1)^{-1}S^{-1}\overline{X} & (n-1)^{-1}S^{-1} \end{pmatrix}$$

(c) Suppose we want to estimate the inverse of the covariance matrix denoted as $\Theta \equiv \{\mathrm{Cov}(X)\}^{-1} \in \mathbb{R}^{p\times p}$. When p is very large, directly computing the inversion of the sample covariance matrix S is not practical. When Θ is sparse, the graphical lasso method uses a regression "trick" to obtain the sparse estimator of the inverse covariance. The algorithm updates one column at a time for Θ. To see this, we decompose Θ as

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix},$$

where $\theta_{12} \in \mathbb{R}^{(p-1)\times 1}$. Suppose $\widehat{\Theta} = S^{-1}$, show that $\widehat{\theta_{12}}/\widehat{\theta_{22}}$ is the least squares estimator of the regression coefficient vector for X_p regressed on $(X_1, \ldots, X_{p-1})$. Discuss how to use lasso regression to obtain $\widehat{\Theta}$.

Problem 7. (5198) Imagine we conducted a study to investigate whether waist-to-hip ratio (WHR) is causally related to an elevation in serum concentrations of C-reactive protein (CRP), a marker of inflammation. We recruited 150 subjects with elevated markers of inflammation from the five primary-care clinics in a small town outside Tallahassee. We then selected controls using a random sample of 150 subjects from the community (all were confirmed not to have elevated concentrations of C-reactive protein).

The gold standard for measuring WHR is in-home assessment by trained clinical personnel. Using this gold standard we could determine the "true" distribution of high and low waist-to-hip ratio in cases and controls to be as follows:

	CRP	No CRP
High WHR	67	32
Low WHR	83	118
Total	150	150

(a) Specify the study design's name. Calculate an appropriate measure of the "true" association using these data, derive its 95% confidence interval, and interpret the result.

(b) In reality, we could not use the gold standard for our study (it was impractical and too expensive). Instead we had to rely on self-reports of waist and hip measurements to estimate waist-to-hip ratio. This method was found to have 75% sensitivity and 95% specificity in both cases and controls, relative to the gold standard. Use these values to answer the questions in the following parts.

Complete a table having the format given below showing how the cases and controls would be reclassified using self-reports, given the sensitivity and specificity of that method (round to whole numbers). Show the derivations.

	CRP	No CRP
High WHR	a =	b =
Low WHR	c =	d =
Total		

(c) Calculate the appropriate measure of association we would have observed if we used the self-report method. Derive the 95% confidence interval, and interpret the result.

(d) Describe how the observed value is different from the true value.

Problem 8. (5198) The standardized mean difference (SMD) is frequently used for contrasting two groups on continuous outcomes. Specifically, suppose that μ_0 and μ_1 are the population means of certain continuous outcome measures in two groups (say, control and treatment) in a study, and σ is the true standard deviation of the measures, which is common for both groups. Then, the SMD is calculated as $(\mu_1 - \mu_0)/\sigma$. In practice, it can be estimated by replacing μ_0 and μ_1 with the sample means $\bar{y}_0$ and $\bar{y}_1$ and replacing σ with the pooled sample standard deviation s_{p}.

(a) Suppose we observe the following data displaying the BMI (body mass index) for some patients in two groups:

Group	BMI									
Control	32.1	29.6	29.1	25.4	29.5	31.4	33.2	32.9		
Treatment	37.8	26.2	31.5	31.7	30.4	27.3	35.8	17.0	26.6	29.7

Use these data to estimate the SMD.

(b) The most widely accepted definition of obesity is a BMI of 30 or higher. Calculate the odds ratio (OR) for obesity using the data in (a).

(c) Most clinicians find SMDs difficult to understand, while ORs are easy to interpret. Transforming SMDs to ORs often requires certain cut-offs of the continuous measures (like BMI at 30) to define the binary outcomes (like obesity). One strategy of making such a transformation independent of cut-offs is to assume that the continuous measures follow logistic distributions (rather than normal distributions), whose cumulative distribution functions are $\frac{1}{1+e^{-(x-\mu_0)/s}}$ and $\frac{1}{1+e^{-(x-\mu_1)/s}}$ in the two groups. Here, s is a scale parameter. Derive the means and variances of these two logistic distributions. You may directly use the result that $\int_{-\infty}^{\infty} \frac{x^2 e^{-x}}{(1+e^{-x})^2} dx = \frac{\pi^2}{3}$.

(d) Assuming that the continuous measures follow logistic distributions, show that $\log(\mathrm{OR}) = \frac{\pi}{\sqrt{3}}$ SMD for any cut-off point.

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

Problem 9. (5326) Answer the following. The parts are **not** related.

(a) Suppose the random variable X has density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty.$$

Find $M_X(t)$, the moment generating function (mgf) of X. Give a careful description of $M_X(t)$ for all real values t. Prove your answer.

(b) For each of the two functions given below, answer the following: Does a distribution exist whose mgf $M(t)$ is the given function? If yes, specify this distribution. If no, prove it.

$$\frac{1}{3}\left(e^{3t} + e^{5t} + e^{7t}\right)$$

$$\frac{1}{3}\left(2e^{3t} + 2e^{5t} - e^{7t}\right)$$

(c) Suppose $Y \sim \text{Geometric}(\beta)$ and $Z \,|\, Y \sim \text{Binomial}(Y, p)$. Find $M_Z(t)$, the mgf of Z. Simplify your answer. (Hint: one approach is to use iterated expectations.)

Note: The pmf of Y is $f_Y(y) = (1-\beta)^{y-1}\beta$ for $y = 1, 2, 3, \ldots$ where $0 < \beta < 1$.

Problem 10. (5326) Suppose that n balls are placed at random into n cells. (The balls are independent of each other, and the cells are equally likely.) Let X be the number of empty cells remaining after the balls are placed.

(a) Find $P(X = 0)$.

(b) Find $P(X = 1)$.

(c) Find EX.

(d) Find $\text{Var}(X)$.

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

Name	**Notation**	$f(x)$	$\mu = E(X)$	Var(X)
Poisson	Poisson(λ)	$\frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.$	λ	λ
Exponential	Exp(β)	$\frac{1}{\beta}e^{-x/\beta}, \quad x > 0$	β	β^2
Normal	N(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2
Chi squared	$\chi^2(p)$	$\frac{1}{\Gamma(p/2)2^{p/2}}x^{p/2-1}e^{-x/2}, \quad x > 0$	p	$2p$

Problem 11. (5327) Consider independent observations $X_1, \ldots, X_n \sim$ Poisson(λ) and $Y_1, \ldots, Y_m \sim$ Exp(λ) .

(a) Find a two-dimensional sufficient statistic for λ.

(b) Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^m Y_i$. Consider estimators of the form $\hat{\lambda}(w) = w\bar{X} + (1-w)\bar{Y}, 0 \leq w \leq 1$. Derive a formula for the $MSE_\lambda(w) = E(\hat{\lambda}(w) - \lambda)^2$.

(c) Is it possible to find a w^* such that $MSE_\lambda(w^*) \leq MSE_\lambda(w)$ for all $\lambda > 0$ and $0 \leq w \leq 1$? If so, find this w^*. If not, explain why.

Problem 12. (5327) Consider independent pairs of observations $(X_i, Y_i), i = 1, \ldots, n$, drawn according to the following model:

$$Y_i = \beta X_i + \epsilon_i,$$

where $\beta \in \mathbb{R}, X_i \sim N(0, \sigma^2), \epsilon_i \sim N(0, \sigma^2)$, and ϵ_i is independent of X_i.

(a) Assume that β is known, and find the MLE for σ^2. Call this estimator $\hat{\sigma}_1^2$.

(b) Find the asymptotic distribution of $\log \hat{\sigma}_1^2$, where $\hat{\sigma}_1^2$ is found in part (a).

(c) Now assume that β is unknown, and find the MLE for (β, σ^2). In finding the MLE for this question, you can use the fact that the stationary point of this likelihood function is the global maximizer.

Problem 13. (6346) Let $(\Omega, \mathcal{F})$ be a measurable space.

(a) Give the definition of a measure μ.

(b) Give the definition of a probability measure μ.

(c) Suppose $\mu_i, i = 1, 2, \ldots$, are measures on $(\Omega, \mathcal{F})$. Let $a_1, a_2, \ldots$ be a sequence of finite positive numbers. For $A \in \mathcal{F}$, define

$$\nu(A) = \sum_{i=1}^{\infty} a_i \mu_i(A)$$

Prove or give a counterexample: ν is a measure on $(\Omega, \mathcal{F})$.

(d) Suppose the measures μ_i in part (b) above are **probability** measures on $(\Omega, \mathcal{F})$. Define ν as above. Prove or give a counterexample: ν is a **probability** measure on $(\Omega, \mathcal{F})$.

Problem 14. (6346) An urn initially contains one red and one green ball. Conduct a series of draws. At each draw, select a ball at random, note its color, and return that selected ball and an additional ball of the same color to the urn. Let X_i be the fraction of red balls in the urn after i draws.

(a) Give the definition of a discrete-time martingale.

(b) Prove or disprove: The sequence $X_1, X_2, X_3, \ldots$ is a martingale.

(c) Find $E(X_k)$.