Ph.D. Qualifying Exam
Monday–Tuesday, January 4–5, 2021

- **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 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be a sequence of i.i.d. paired samples. Conditioned on $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{Z}^+$ follows a Poisson distribution with mean $\lambda_{X_i} = e^{\beta^T X_i}$, where the parameter $\beta \in \mathbb{R}^d$. That is,

$$Y_i \,|\, X_i, \beta \sim \text{Poisson}(\lambda_{X_i}), \quad i = 1, \cdots, n.$$

Our goal is to find the maximum likelihood estimate (MLE) of β with the observations $\{(X_i, Y_i)\}_{i=1}^n$.

(a) Derive an expression for the log-likelihood function

$$l(\beta) = \sum_{i=1}^{n} \log(f(Y_i|X_i, \beta)),$$

such that the MLE is given by

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\ l(\beta).$$

(b) Prove that the Hessian matrix of the log-likelihood function w.r.t. β is semi-negative definite.

(c) Write out the Newton-Raphson algorithm to find $\hat{\beta}$.

Problem 2. (5106) Consider a data set of observations $\{x_n\}$ where $n = 1, \cdots, N$, 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 $U_1, \cdots, U_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, \cdots, \lambda_M$.

Problem 3. (5166) Twenty mice were divided at random into two groups of size ten and fed different diets y_1 and y_2. At four weeks the weights of the mice were:

y_1:	12.5	18.0	12.0	13.2	15.0	14.8	14.6	16.2	17.5	16.2
y_2:	16.2	16.0	17.2	15.3	17.0	16.5	16.8	18.2	15.5	18.5

(a) Test the correlation between y_1 and y_2. What assumptions on the observations have you made for the testing?

(b) Do the mean weights for diets y_1 and y_2 differ at level $\alpha = 0.05$? (State your null and alternative hypotheses and carry out a test.)

(c) Test a hypothesis about the two populations using the Wilcoxon Rank-Sum Test with a normal approximation and $\alpha = 0.05$. State your hypotheses clearly.

Problem 4. (5166) Two treatments, A and B, are compared in an experiment design. For each treatment, measurements are recorded successively in time. Suppose that the measurements follow the model:

$$X_1(t) = \mu_1 + \epsilon_1(t) - \theta_1\epsilon_1(t-1) + \theta_3\epsilon_1(t-3),$$

$$X_2(t) = \mu_2 + \epsilon_2(t) + \alpha_1\epsilon_2(t-1) + \alpha_2\epsilon_2(t-2),$$

where $\{\epsilon_i(t) : i = 1, 2;\ t = \ldots, -1, 0, 1, \ldots\}$ are independent normal random variables with mean zero and variance σ_i^2. Let $\bar{X}_i = \sum_{t=1}^{n} X_i(t)/n$ for $i = 1, 2$.

(a) Calculate $\mathrm{Cov}(X_1(t),\, X_1(t+k))$ for $k \geq 0$.

(b) For a given i, is the process $\{X_i(t), t = 1, 2, \ldots, \}$ stationary?
Please give your justifications.

(c) Calculate the mean and variance of $\bar{X}_1$. What is the distribution of $\bar{X}_1$?

Problem 5. (5167) Consider the multiple linear regression with i.i.d. samples $(Y_i, \mathbf{X}_i) \sim (Y, \mathbf{X})$, $i = 1, \ldots, n$.

(a) Write down the hat matrix $\mathbf{H}$.

(b) Show that $\mathbf{I} - \mathbf{H}$ and $\mathbf{H}$ are orthogonal. Use this result to show that the slope of the regression of fitted residual $\widehat{e}$ on fitted response $\widehat{Y}$ is 0.

(c) Provide conditions under which the mean function for $\mathrm{E}(Y \mid \mathbf{X})$ is linear but has non-negative coefficients for $\mathbf{X}$.

(d) What are the intercept and slope of the regression of $\widehat{e}$ on Y? and why?

(e) What are the intercept and slope of the regression of Y on $\widehat{Y}$? and why?

(f) What are the intercept and slope of the regression of Y on $\widehat{e}$? and why?

Problem 6. (5167) Consider a heteroscedastic linear model with data consisting of n independent copies of $(Y, \mathbf{X}, W)$, with

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

where $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$, $W > 0$ with $\mathrm{E}(W) = 1$, and $\epsilon \in \mathbb{R}$ is independent of $(\mathbf{X}, W)$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2$.

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

(b) Provide a likelihood justification for part (a). (Hint: by assuming $\epsilon \sim$ Normal)

(c) If $\mathbf{X} \mid W \sim \mathrm{N}(\boldsymbol{\mu}_x, W^{-1}\boldsymbol{\Sigma})$, describe how to estimate $\boldsymbol{\Sigma}$.

(d) Discuss how the results in part (a) will change if we can assume $\mathrm{E}(W) = \omega$ for an arbitrary positive constant ω instead of assuming $\mathrm{E}(W) = 1$.

Problem 7. (5198) A recent study examined the association between autoimmune diseases among pregnant women and subsequent diagnoses of autism in their children. Cases ($n = 407$) were children with a diagnosis of autism spectrum disorder (ASD) in the Kaiser Permanente clinical database. Controls ($n = 2095$) were children with no autism diagnosis drawn from the same birth hospitals as the cases. Of particular interest is the odds ratio associating maternal psoriasis and child's diagnosis of ASD.

Using the results below, *discuss whether the occurrence of maternal psoriasis during pregnancy is associated with subsequent ASD diagnosis and whether maternal age at delivery confounds or interacts with the effect of psoriasis on offspring ASD.* Be sure to

- Include 95% confidence intervals for any odds ratio estimates you cite, together with your interpretation of these intervals.
- Explicitly state hypotheses for test statistics and clearly state your conclusions from tests.

Analysis set	n	Crude $\widehat{OR}$	se($\log \widehat{OR}$)
All women	2502	2.9	0.38
Women aged < 35	1872	2.0	0.59
Women aged ≥ 35	630	4.1	0.51

With age considered a stratification factor,

- the Mantel-Haenszel estimate and 95% confidence interval for the odds ratio are 2.98 (1.41, 6.28)
- the Breslow-Day test statistic is $X^2_{BD} = 0.80$ on 1 degree of freedom
- the Cochran-Mantel-Haenszel test statistic is $X^2_{CMH} = 9.02$ on 1 degree of freedom

---

Problem 8. (5198) A study was performed to assess the association between tonsillectomy and Hodgkin's lymphoma (a type of blood cancer). 85 cases of Hodgkin's disease were identified and each was matched to a healthy sibling of the same sex and age within 5 years. It was found that 26 of the cases and their matched sibling controls both had had tonsillectomy; in 15 pairs the case had had tonsillectomy but not the control; in 7 pairs the control, but not the case, had had tonsillectomy; and there were 37 instances in which neither the case nor control siblings had had tonsillectomy.

(a) Formally test the association between history of tonsillectomy and Hodgkin's lymphoma. Clearly state the null and alternative hypotheses and your conclusion.

(b) Find the estimated odds ratio associated with history of tonsillectomy and give a corresponding 95% confidence interval. Interpret in the context of this study.

(c) Does the matching by siblingship, sex and age in this study enhance or detract from the strength of conclusions? How would you expect the odds ratio estimate to change if the matching were ignored? Explain.

Formula Sheet for STA 5198 Problems

Some known formulae based on the usual 2×2 table

	D	$\overline{D}$
E	a	b
$\overline{E}$	c	d

RR = relative risk

OR = odds ratio

$$\text{AR} = \text{attributable risk} = \frac{R - R_{\overline{E}}}{R}, \text{ where } R \text{ is the overall risk} \tag{1}$$

$$\widehat{\text{Var}}\left(\log \widehat{\text{RR}}\right) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d} \tag{2}$$

$$\widehat{\text{Var}}\left(\log \widehat{\text{OR}}\right) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{3}$$

Approximate $100(1-\alpha)\%$ confidence interval for AR is given by

$$\frac{(ad-bc)\exp(\pm u)}{nc + (ad-bc)\exp(\pm u)}, \tag{4}$$

where

$$u = \frac{z_{\alpha/2}(a+c)(c+d)}{ad-bc}\sqrt{\frac{ad(n-c)+c^2b}{nc(a+c)(c+d)}}.$$

- Mantel-Haenszel estimate of the odds ratio:

$$\left(\sum_i \frac{a_i d_i}{n_i}\right) / \left(\sum_i \frac{b_i c_i}{n_i}\right) \tag{5}$$

- Cochran-Mantel-Haenszel test for significance of the E-D association adjusted for confounder:

$$X^2 = \frac{\left(\sum a_i - \sum E(a_i)\right)^2}{\sum \text{Var}(a_i)} \tag{6}$$

$$\text{E}(a_i) = \frac{D_i E_i}{n_i}, \qquad \text{Var}(a_i) = \frac{D_i \overline{D}_i E_i \overline{E}_i}{n_i^2 (n_i - 1)}$$

- Breslow-Day test for homogeneity of the relative risks (or odds ratios) across strata:

$$X^2 = \sum \frac{(a_i - \text{E}(a_i))^2}{\text{Var}(a_i)}, \tag{7}$$

  where now $\text{E}(a_i)$ and $\text{Var}(a_i)$ are computed using the Mantel-Haenszel estimate of the common relative risk across strata.

- χ^2 test for association in 2-way table with observed counts $\{O_{ij}\}$:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n} n \tag{8}$$

- χ^2 test for trend in $\ell \times 2$ table with exposure scores $x_1, x_2, \ldots, x_\ell$:

$$X^2_{(L)} = \frac{(T_1 - \frac{n_D}{n} T_2)^2}{V}, \tag{9}$$

where $T_1 = \sum a_i x_i$, $T_2 = \sum m_i x_i$, $T_3 = \sum m_i x_i^2$ and $V = n_D n_{\bar{D}}(nT_3 - T_2^2)/[n^2(n-1)]$

- κ statistic for agreement in square 2-way table with observed counts $\{O_{ij}\}$ and expected counts (assuming independence of rows and columns) $\{E_{ij}\}$:

$$\kappa = \frac{p_O - p_E}{1 - p_E}, \qquad p_O = \sum O_{ii}/n, \quad p_E = \sum E_{ii}/n \tag{10}$$

- Some normal quantiles. Here $Z \sim N(0,1)$.

z	0.84	1.04	1.28	1.64	1.96	2.33
$P(Z \le z)$	0.80	0.85	0.90	0.95	0.975	0.99

- Critical values of the χ^2_ν distribution. For $W \sim \chi^2_\nu$, the entries are q such that $\Pr(W \ge q) = p$.

	Probability p			
$\nu =$ df	0.10	0.05	0.025	0.01
1	2.7	3.8	5.0	6.6
2	4.6	6.0	7.4	9.2
3	6.3	7.8	9.2	11.3
4	7.8	9.5	11.1	13.3
5	9.2	11.1	12.8	15.1
6	10.6	12.6	14.4	16.8

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

Problem 9. (5326) Answer the following.

(a) Suppose the random variable X has density (pdf) given by

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

for $\lambda > 0$. Find $M_X(t)$, the moment generating function (mgf) of X. (Specify $M_X(t)$ for all $t \in \mathbb{R}$.)

(b) Suppose the random variable Y has mgf

$$M_Y(t) = \frac{e^{\beta t^2}}{1-t} \quad \text{for } t < 1$$

for some value $\beta > 0$. Find the mean and variance of Y.

(c) Suppose the sequence of random variables $Z_1, Z_2, Z_3, \ldots$ have mgf's given by

$$M_{Z_n}(t) = \frac{e^{t^2/n}}{1-t} \quad \text{for } t < 1\,.$$

For $z > 0$, find $\lim_{n\to\infty} P(Z_n \leq z)$.

Problem 10. (5326) The hazard function $h_T(t)$ of a random variable T is defined by

$$h_T(t) = \lim_{\delta \downarrow 0} \frac{P(t \leq T < t+\delta \,|\, T \geq t)}{\delta}\,.$$

(Note: $\delta \downarrow 0$ is the same as $\delta \to 0^+$.)

(a) Show that if T is a continuous random variable with density $f_T(t)$ and cumulative distribution function (cdf) $F_T(t)$, then

$$h_T(t) = \frac{f_T(t)}{1-F_T(t)} = -\frac{d}{dt}\log(1-F_T(t))\,.$$

(b) Let $X_1, X_2, \ldots, X_n$ be independent random variables with cdf's $F_1, F_2, \ldots, F_n$. Define $Y = \min_i X_i$. State and prove a formula for the cdf of Y.

(c) If $X_1, X_2, \ldots, X_n$ in part (b) are all continuous random variables, prove that

$$h_Y(t) = h_{X_1}(t) + h_{X_2}(t) + \cdots + h_{X_n}(t)\,.$$

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

Name	**Notation**	$f(x)$	$E(X)$	Var(X)
Uniform	Unif(a, b)	$\frac{1}{b-a}$, $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
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 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). (Hint: If $Z \sim N(0, 1)$, then $Z^2 \sim \chi^2(1)$.)

(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 without proof the fact that the stationary point of this likelihood function is the global maximizer.

Problem 12. (5327) Let $X_1, \ldots, X_n$ be iid observations from Uniform$(1, a + 1)$, where $a > 0$ is the unknown parameter. Answer the following questions.

(a) Find the complete sufficient statistic for a.

(b) Find the best unbiased estimate for a (i.e, the uniformly minimum variance unbiased estimator).

(c) Find $E\left(X_{(n)} \;\middle|\; \frac{\bar{X} - X_{(1)}}{X_{(n)} - \bar{X}}\right)$, where $X_{(1)} = \min\limits_{i=1,\ldots,n} X_i$, $X_{(n)} = \max\limits_{i=1,\ldots,n} X_i$ and $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$.

Problem 13. (6346) Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \cdots \subset \mathcal{F}$ be a sequence of σ-fields.

(a) Give the definition of a martingale.

(b) Suppose Y is a random variable with $E|Y| < \infty$. Let $X_n = E(Y|\mathcal{F}_n)$. Show X_n is a martingale.

Problem 14. (6346) Answer the following.

(a) State and prove Markov's inequality.

(b) Suppose $X_i, i = 1, 2, \ldots$ are random variables with finite mean μ and

$$\lim_{i\to\infty} \text{var}(X_i) = 0.$$

Show X_i converges to μ in probability.

(c) Prove or give a counterexample to the converse of part (b): Suppose X_i are random variables with finite mean μ and X_i converges to μ in probability. Then

$$\lim_{i\to\infty} \text{var}(X_i) = 0.$$