Ph.D. Qualifying Exam
Wednesday–Thursday, August 23–24, 2023

- **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 $X_i \in \mathbb{R}^d$, the value $Y_i \in \mathbb{Z}^+$ follows a Poisson distribution with parameter $\lambda_{X_i} = e^{\beta^T X_i}$, where the parameter $\beta \in \mathbb{R}^d$. That is,

$$Y_i \,|\, X_i, \beta \sim \text{Poisson}(\lambda_{X_i})$$

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 w.r.t. the log-likelihood function is semi-negative definite.

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

Problem 2. (5106) Let X be a random variable which follows a Gamma distribution with shape α and rate β. For a constant $a > 0$, our goal is to use the tilted sampling to estimate the tail probability:

$$\theta = P\{X > a\} = \int_a^\infty f(x; \alpha, \beta) dx,$$

where the Gamma density function is $f(x; \alpha, \beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$. We simplify the calculation by assuming $\alpha = 3$ in this problem.

(a) Compute $M(t) = \int_0^\infty e^{tx} f(x; 3, \beta)\, dx$.

(b) Derive an expression to find the optimal amount of tilt t to estimate θ for a given a. (Hint: the mean of X is $\frac{\alpha}{\beta}$.)

(c) Based on the optimal t, write out the sampling method to estimate θ.

Problem 3. (5166) Consider the following linear model for a randomized block design:

$$y_{ti} = \mu + \tau_t + \beta_i + \epsilon_{ti}\,, \quad t = 1, \ldots, k\,; \quad i = 1, \ldots, n,$$

where μ is an overall mean, τ_t is the effect of tth treatment, β_i is the effect of ith block, and $\{\epsilon_{ti} : t = 1, \ldots, k; i = 1, \ldots, n\}$ are assumed to be iid $N(0, \sigma^2)$. Define

$$Y = (Y_{11}, Y_{12}, \ldots, Y_{1n}, Y_{21}, Y_{22}, \ldots, Y_{2n}, \ldots, Y_{k1}, Y_{k2}, \ldots, Y_{kn})'.$$

(a) Find the mean vector $\boldsymbol{\mu} = E(Y)$ and the variance-covariance matrix $V = \mathrm{Cov}(Y)$. What is the distribution of Y?

(b) Define

- S_D: Total Variation of the observations,
- S_T: Sum of Squares for Treatments,
- S_B: Sum of Squares for Blocks,
- S_R: Sum of Squares for Experimental Errors.

Write an ANOVA table for this experiment, including the Sources of Variation, Degrees of Freedom (df), Sums of Squares, and Mean Squares. Derive the expectations for the Mean Squares.

Problem 4. (5166) Consider a 2^{5-2} fractional factorial design with generators **I=1234** and **I=135**. After analyzing the results from this design the researcher decides to perform a second 2^{5-2} design exactly the same as the first but switching the signs of column 2 from "+" to "−" and "−" to "+" (a fold-over design).

(a) What is the defining relation of the first design? What is the resolution of the first design? List the aliases of the main effects in this design.

(b) What is the defining relation of the second design? What is the resolution of the second design? List the aliases of the two-term interactions of factors **1, 2,** and **3** in this design.

(c) What is the defining relation of the combined design (designs 1 and 2 together)? What is the resolution of the combined design? (Hints: combine the defining relations of the two designs together and remove all the defining words that have both"+" and "−" signs.)

(d) Can you modify the first and second designs such that the combined design has a higher resolution?

Problem 5. (5167) Consider independent observations $\{Y_i, X_i\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}$.

(a) Assume that $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ and $\text{Var}(Y_i \mid X_i) = \sigma^2$, where $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$ are all unknown. Derive the OLS estimator for (β_0, β_1).

(b) Suppose we instead assume that $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ and $\text{Var}(Y_i \mid X_i) = \sigma^2/d_i$, where $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$ are all unknown, but $d_i > 0$ is known for $i = 1, \ldots, n$. Describe how you would like to estimate (β_0, β_1).

(c) If we are in doubt of the model assumption $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$, how can we test it?

Problem 6. (5167) Suppose X is a continuous predictor, and F is a factor with two levels $\{+1, -1\}$, represented by two dummy variables $U_+ = I(F = +1)$ and $U_- = I(F = -1)$ where $I(\cdot)$ is the indicator function that takes values 0 and 1. The response Y is a continuous variable. Consider the model

$$E(Y \mid X, F) = \alpha + \beta X + \gamma U_+.$$

(a) Sketch the fitted regression lines (on a plot of Y versus X) for the two levels of F, and determine the value of $E(Y \mid X = 1, F = +1) - E(Y \mid X = 1, F = -1)$.

(b) Interpret γ.

(c) Can we fit a model under the assumption $E(Y \mid X, F) = \alpha + \beta X + \gamma_+ U_+ + \gamma_- U_-$? Why or why not?

Problem 7. (5198) A study of the association between concussions in youth and mental health problems identified in children aged 5-18 years who presented to a medical facility with either a concussion (the exposure) or an orthopedic injury (the comparison group). Children seen for reasons associated with mental health in the year prior to injury were excluded, and all children were covered by the same national health insurance plan. There were 152,231 children who presented with concussion; each was matched by age and sex to multiple children with orthopedic injury. Among the 152,321 children who presented with concussion, 53,863 experienced mental health problems during the subsequent follow up of approximately 10 years. Of the 296,482 children with orthopedic injury, 80,076 showed mental health problems during the follow up.

(a) Display these data in the usual 2×2 format as shown:

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

(b) In this study design, are relative risk (RR) and odds ratio (OR) both appropriate measures for the association between concussion and mental health? Explain.

(c) Using your selected measure of association (RR or OR), use these data to estimate the association and provide an approximate 95% confidence interval. Interpret.

(d) The investigators excluded children with medical visits for reasons associated with mental health in the year prior to injury. Is this a strength or a weakness of the study? Explain.

(e) What advantage, if any, is the comparison group of children with orthopedic injury compared to a comparison group of randomly selected children without concussion?

Problem 8. (5198) A case-control study investigated the association between vaccination against influenza A and B viruses ("flu vaccine") and multiple sclerosis (MS). The investigators identified 110 subjects with MS and matched each to a control subject according to age, sex and nationality. There were 50 pairs in which both the MS case subject and paired control subject had received the flu vaccine; 10 pairs in which the MS case subject had the vaccine and the control subject did not; 25 pairs in which the control subject but not the paired case subject had received the vaccine; and 25 pairs in which neither subject in the pair had received the vaccine.

(a) Formally test the association between flu vaccination and MS. Clearly state the null and alternative hypotheses and your conclusion.

(b) Find the estimated odds ratio for vaccine versus no vaccine and give a corresponding (approximate) 95% confidence interval. Interpret in the context of this study.

(c) The investigators noted that 30% of the MS case subjects had a family history of MS whereas only 12% of control subjects had this family history. Explain how to adjust for family history as a confounder when assessing the association between flu vaccine and MS in this study. Describe how to perform your recommended analysis.

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	9.3	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
7	12.0	14.1	16.0	18.5
8	13.4	15.5	17.5	20.1
9	14.7	16.9	19.0	21.7
10	16.0	18.3	20.5	23.2
11	17.3	19.7	21.9	24.7
12	18.5	21.0	23.3	26.2

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

Problem 9. (5326) A Polya urn initially contains 1 red ball and G green balls. An infinite sequence of balls is drawn from this urn. (Since this is a Polya urn, after each ball is drawn, it is returned to the urn and another ball of the same color is added to the urn.)

Answer the following questions. Justify your answers. Simplify your answers when possible. In these questions let A_i be the event that the i-th ball drawn from the urn is red, and let X be the random number of draws needed to get the first red ball.

(a) Find $P(A_2)$.

(b) Find $P(A_1 \mid A_2)$.

(c) Find $P(X = k)$ where k is an arbitrary positive integer.

(d) Find EX.

(e) Find the probability that the ball drawn immediately after the first red ball is also red. (Note: this is **not** a conditional probability.)

The following facts about the bivariate normal distribution are **not** needed to solve the next problem, but might be used to reduce the work in some parts. You may freely use any of these facts without proof.

Let (X, Y) have a bivariate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, ρ.

The joint density (pdf) of (X, Y) is defined for all $(x, y) \in \mathbb{R}^2$ by

$$g(x,y) = \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \times \exp\left(\frac{-1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

The marginal distributions are $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$.

The conditional distributions are:

$$Y \,|\, X = x \sim N\left(\mu_Y + \rho\sigma_Y\left(\frac{x-\mu_X}{\sigma_X}\right),\ \sigma_Y^2(1-\rho^2)\right),$$
$$X \,|\, Y = y \sim N\left(\mu_X + \rho\sigma_X\left(\frac{y-\mu_Y}{\sigma_Y}\right),\ \sigma_X^2(1-\rho^2)\right).$$

Problem 10. (5326) Let (X, Y) have a joint density which is defined for all $(x, y) \in \mathbb{R}^2$ by

$$f(x,y) = \begin{cases} C\exp\left(-\left(x^2 - xy + \dfrac{y^2}{2}\right)\right) & \text{for } y \geq \sqrt{2}, \\ 0 & \text{for } y < \sqrt{2}. \end{cases}$$

For $y \geq \sqrt{2}$, this density is proportional to a bivariate normal density $g(x, y)$ with $\mu_X = \mu_Y = 0$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 2$, and $\rho = 1/\sqrt{2}$.

Answer the following. Justify your answers. Simplify your answers and give explicit formulas for all answers. If your answer is valid only in a certain range, that should be part of your answer. Some answers may involve Φ, the $N(0, 1)$ cdf defined by $\Phi(z) = \displaystyle\int_{-\infty}^{z} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du$.

(a) Find the value of C. (If you cannot find this value, just leave it as C when doing the later parts.)

(b) Find the marginal density of Y.

(c) Find the marginal density of X.

(d) Find $E(X|Y = y)$ for $y > \sqrt{2}$.

(e) Find $E(Y|X = x)$ for $-\infty < x < \infty$.

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

Name	**Notation**	$f(x)$	E(X)	Var(X)
Normal	$\text{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2
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) We consider $X_1, \ldots, X_n$ i.i.d. from $N(\mu, \sigma^2)$, where the parameter space is $\Theta = \{(\mu, \sigma^2) : \sigma^2 > 0, |\mu| \leq \sigma/10\}$.

(a) Find the method of moments (MOM) estimator for μ. Call it $\hat{\mu}$.

(b) Find the mean squared error (MSE) for the MOM estimator $\hat{\mu}$. Recall that $\text{MSE}_\mu(\hat{\mu}) = E(\hat{\mu} - \mu)^2$.

(c) Consider the estimator

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{1+\lambda},$$

where $\lambda \geq 0$ is a constant. For what values of λ do we have

$$\text{MSE}_\mu(\tilde{\mu}) \leq \text{MSE}_\mu(\hat{\mu})$$

for all $|\mu| \leq \sigma/10$, where $\hat{\mu}$ is the MOM estimator?

Problem 12. (5327) We consider $X_1, \ldots, X_n$ i.i.d. from $N(\mu, \sigma^2)$, where the parameter space is $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$.

(a) Assume that μ 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 μ is unknown, and find the MLE for (μ, σ^2). In finding the MLE in this part, you can use the fact that the stationary point of this likelihood function is the global maximizer.

Problem 13. (6346) Let $(\Omega, \mathcal{B}, P)$ be a probability space and let $\mathcal{G}$ be a sub-sigma-field of $\mathcal{B}$. Suppose X and Y are L_2-integrable random variables on $(\Omega, \mathcal{B}, P)$; that is, $E(|X|^2)$ and $E(|Y|^2)$ are both finite. Also, assume that Y is $\mathcal{G}$-measurable. Let $V(\cdot)$ denote the variance; that is, $V(X) = E\left[\{X - E(X)\}^2\right]$ and $V(X|\mathcal{G}) = E\left[\{X - E(X|\mathcal{G})\}^2 |\mathcal{G}\right]$.

(a) Prove that $V(X|\mathcal{G}) \leq E\left\{(X-Y)^2|\mathcal{G}\right\}$ almost surely.

(b) Use Part (a) and a particular specification of Y to prove that $E[V(X|\mathcal{G})] \leq V(X)$.

(c) Define f to be a contraction on $\mathbb{R}$. That is, $f : \mathbb{R} \to \mathbb{R}$ with

$$|f(x) - f(y)| \leq |x - y| \tag{11}$$

for all $x, y \in \mathbb{R}$ with strict inequality for some pair x, y. Use Part (a) and Equation (11) to prove that $V[f(X)|\mathcal{G}] \leq V(X|\mathcal{G})$ almost surely.

Problem 14. (6346) Let $Y_1, \ldots, Y_n$ be a random sample from F. Suppose you are interested in performing inferences under a different distribution, G. Let F and G have densities with respect to Lebesgue measure, namely f and g. Define the function $w(y) = g(y)/f(y)$, and assume that $f(y) > 0$ whenever $g(y) > 0$.

(a) Define the estimator

$$\widehat{\theta} \equiv \frac{1}{n}\sum_{i=1}^{n} w(Y_i)h(Y_i). \tag{12}$$

Prove that $\widehat{\theta}$ converges almost surely to $E_G\left\{h(Y)\right\}$ as $n \to \infty$, provided the expectation exists. Here, E_G is the expectation operator with respect to G.

(b) Provide the formal statement for the classical Central Limit Theorem (CLT) for independent and identically distributed (iid) random variables. What condition do we require for $\widehat{\theta}$ in (12) to converge asymptotically to a normal distribution? What is this asymptotic distribution for $\widehat{\theta}$?

(c) Let F be the standard Cauchy distribution with median zero and scale one, and let G be a standard t-distribution with 2 degrees of freedom. Let $h(y) = y$. Does $\widehat{\theta}$ follow a CLT? Use Part (b) to justify your answer. Note $f(y) = \frac{1}{\pi(1+y^2)}$ and $g(y) = (2+y^2)^{-3/2}$.