Ph.D. Qualifying Exam
Wednesday–Thursday, January 4–5, 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 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 $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. We simplify the calculation by assuming $\alpha = 2$ in this problem.

(a) Derive an expression to find the optimal amount of tilt t to estimate θ for a given a.

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

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

$$(x_1, x_2, x_3, x_4) \sim M(n; 0.1(2-\theta), 0.3(2-\theta), 0.2(1+\theta), 0.2\theta)$$

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

(b) Let $L(\theta)$ denote the likelihood with parameter θ and assume θ^* is the true, unique maximum likelihood point. Use a graphing method to illustrate how the EM algorithm searches for θ^*. That is, draw a graph to show

$$\theta^{(m)} \to \theta^* \quad (m \to \infty)$$

Problem 3. (5166) A study was conducted to determine whether the age of customers is related to the type of movie he or she watches. A sample is shown in the following table.

Age	Documentary	Comedy	Mystery
12-20	33	29	28
21-40	45	46	69
41 and over	43	60	57

(a) Given that the total sample size $n = 410$ is fixed, what is the distribution of the nine categories? What are the mean value and variance of each cell frequency? Run a chi-square test of independence and draw your conclusion. Use $\alpha = 0.05$.

Note: A small chi-square table may be found on page 6.

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

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

$$X_i(t) = \mu_i + \epsilon_i(t) + \theta\epsilon_i(t-2),$$

where $\{\epsilon_i(t) : i = 1, 2, 3; t = -1, 0, 1, \ldots, n\}$ are independent and identically distributed random variables with mean zero and variance σ^2. Let $\bar{X}_i = \sum_{t=1}^{n} X_i(t)/n$ and $\bar{X} = \sum_{i=1}^{3}\sum_{t=1}^{n} X_i(t)/(3n)$.

(a) Calculate the means and variances of $\bar{X}_i$ and $\bar{X}$.

(b) Let $s^2 = \sum_{i=1}^{3}\sum_{t=1}^{n}(X_i(t) - \bar{X}_i)^2/[3(n-1)]$. Calculate the mean value $\mathrm{E}(s^2)$. Find the range of θ for which s^2 over-estimates σ^2.

Problem 5. (5167) For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$, answer the following.

(a) Write down the residual from OLS fitting in terms of $\mathbf{Y}$ and the n-by-n projection matrix onto $\mathbf{X}$.

(b) Derive an unbiased estimator for σ^2 and prove its unbiasedness.

(c) How does the answer in part (b) change with the rank of the matrix $\mathbf{X}$?

Problem 6. (5167) Suppose we have two response variables $Y_1, Y_2 \in \mathbb{R}$ and a multivariate predictor $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$. We are interested in the following regression model:

$$Y_1 = \beta_0 + \beta^T X + \varepsilon,$$

$$Y_2 = \gamma_0 + \gamma^T X + \delta,$$

where (ε, δ) is bivariate normal with mean zero and $\text{corr}(\varepsilon, \delta) = \rho > 0$. Suppose the data are i.i.d. and $(Y_{1i}, Y_{2i}, X_i^T) \sim (Y_1, Y_2, X^T)$, $i = 1, \ldots, n$.

(a) Write down the ordinary least squares estimates for $\beta_0, \beta, \gamma_0, \gamma$ and prove that these are the maximum likelihood estimators.

(b) Derive an estimator (e.g., the maximum likelihood estimator) for ρ.

(c) Discuss how $\rho = 0$ would affect the estimates in parts (a) and (b).

(d) If ρ is given, how can you improve the estimates in parts (a) and (b)?

Problem 7. (5198) A study performed in Norway investigated the association between prenatal benzodiazepines/z-hypnotics (B/Z, exposure) and childhood attention deficit/hyperactivity disorder (ADHD). Women who had participated in a population-based study and who had provided self-report information about their use of B/Z during pregnancy were included. Their children were then followed to a mean age of 11 years for development of ADHD. Our focus is the effect of B/Z use during early pregnancy (gestational age 0-16 weeks). Among the 82,201 mother-child pairs in the study, 435 mothers reported use of B/Z during early pregnancy and, among these, 19 children developed ADHD. Of the remaining 81,766 mothers who did not report use of B/Z, 2236 children developed ADHD.

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

	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 B/Z use and ADHD? 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 noted that the women who took B/Z may be systematically different from women who did not take B/Z (as B/Z was not randomly assigned). Among the covariates available for the mothers was smoking during pregnancy and lifetime history of major depression, both of which might conceivably affect development of ADHD in a child. Indeed, the proportions of smokers were 16.4% and 7.5% in the exposed and unexposed groups, and the proportions with a history of depression were 19.7% and 5.7%, respectively. Describe *two ways* to adjust for these potential confounders when assessing the B/Z-ADHD association.

Problem 8. (5198) Investigators performed a cross sectional study associating diversity of the gut microbiome (low or high) with cognitive statue (impaired or normal) in middle-aged adults in the U.S.. They found an estimated relative risk for impaired status (low diversity compared to high diversity) of 1.5 with 95% confidence interval (1.2, 2.0). In analyses of males and females separately, they reported

$$\text{for males:} \quad \widehat{\text{RR}}_M = 1.2, \ \text{se}(\log \widehat{\text{RR}}_M) = 0.20$$
$$\text{for females:} \quad \widehat{\text{RR}}_F = 2.1, \ \text{se}(\log \widehat{\text{RR}}_F) = 0.10$$

Computing the estimated relative risk ratio (females to males) as $\text{RRr} = 2.1/1.2 = 1.75$, they claimed that low diversity of the gut microbiome is riskier for females than for males with respect to cognitive health. Evaluate the statistical evidence for their claim.

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

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

- 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) Let X have the folded Cauchy density given by

$$f(x) = \frac{2}{\pi(1+x^2)} \quad \text{for } x > 0 \text{ and } f(x) = 0 \text{ otherwise}.$$

Answer the following. The different parts are unrelated.

Note: In one part it may be useful to know that $\dfrac{d}{dx}\tan^{-1}(x) = \dfrac{1}{1+x^2}$.

(a) Let $M_X(t)$ be the moment generating function of X. Determine the set of values t for which $M_X(t)$ is finite.

(b) Find a transformation g such $U = g(X)$ has an exponential distribution with density $f_U(u) = e^{-u}$ for $u > 0$. Give a closed form expression (i.e., an explicit formula) for this transformation.

(c) Let X and Y be independent where X is folded Cauchy with the density given earlier and $Y \sim \text{Uniform}(0,1)$. Find the density of $Z = Y/X$.

Problem 10. (5326) Suppose you have a fair coin with the sides labeled $+1$ and -1 whose tosses are independent of each other. Toss this coin n times and let X_i, $i = 1, 2, \ldots, n$, be the value observed on the i-th toss. Define $X_{n+1} = \prod_{i=1}^{n} X_i$. Then define $Y_i = I_{\{X_i=1\}}$ for $i = 1, 2, \ldots, n+1$, that is, $Y_i = 1$ if $X_i = 1$, and $Y_i = 0$ otherwise. Finally, define $Z = \sum_{i=1}^{n+1} Y_i$. Answer the following **and fully justify your answers.** Assume $n \geq 2$.

[Note that $Y_i = (X_i + 1)/2$ for all i. This fact might be useful depending on how you do the parts.]

(a) Are $Y_1, Y_2, \ldots, Y_{n+1}$ mutually independent random variables?

(b) Find EZ.

(c) Find $\text{Var}(Z)$.

(d) Find the probability mass function (pmf) of Z. (To save a little work, consider only the case where n is even.)

Problem 11. (5327) Consider n independent random variables $X_i \sim N(\mu, 1), i = 1, \ldots, n$, where μ is the unknown parameter. We do not know the exact values of X_i, but only observe the events $X_i > 0$ or $X_i \leq 0$. In answering the following questions, you can use $\Phi(x)$ to denote the cumulative distribution function for a standard normal random variable, Φ^{-1} to denote the inverse of Φ, and ϕ to denote the probability density function for a standard normal random variable.

(a) Find the maximum likelihood estimator (MLE) for μ.

(b) Find an estimate for the variance of the MLE for μ.

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 $E\left(X_{(n)} \mid \dfrac{\bar{X} - X_{(1)}}{X_{(n)} - \bar{X}}\right)$, where $X_{(1)} = \min_{i=1,\ldots,n} X_{(1)}$, $X_{(n)} = \max_{i=1,\ldots,n} X_{(n)}$ and $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Problem 13. (6346) Let $\{X_n, n \geq 1\}$ be i.i.d. $N(0,1)$. The goal of this problem will be to find the limiting distribution of

$$\frac{n\left(X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}\right)^2}{\left(\sum_{i=1}^{2n} X_i^2\right)^2} \tag{11}$$

as $n \to \infty$. To do this answer the following:

(a) Slutsky's Theorem: Suppose $\{Z_n, Z, Y_n, n \geq 1\}$ are random variables. If Z_n converges in distribution to Z and Y_n converges in probability to a non-zero constant c. Then does Z_n/Y_n converge in distribution? If so, what does it converge to?

(b) Does $\left(\frac{\sum_{i=1}^{2n} X_i^2}{2n}\right)^2$ converge in probability? If so to what? Prove your answers.

(c) Does $\sqrt{n}\left(\frac{X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}}{n} - 0\right)$ converge in distribution? If so, what is the limiting distribution? Prove your answers.

(d) Does $\frac{\left(X_1X_2 + X_3X_4 + \ldots + X_{2n-1}X_{2n}\right)^2}{n}$ converge in distribution? If so, what is the limiting distribution? Prove your answers.

(e) Use parts $(a), (b)$, and (d) to find the limiting distribution of (11).

Problem 14. (6346) Let $(\Omega, \mathcal{B})$ and $(\Omega', \mathcal{B}')$ be two measurable spaces.

(a) When is the map $X : \Omega \to \Omega'$ referred to as a "random variable"?

(b) Let $\mathcal{B}' = \sigma(\mathcal{C}')$ for some generating set $\mathcal{C}'$. Suppose $X^{-1}(\mathcal{C}') \subset \mathcal{B}$. Is X measurable? Prove your claim.

(c) Let $\{X_n, n \geq 1\}$ be random variables on the probability space $(\Omega, \mathcal{B}, P)$ and let $S_0 = 0$ and $S_n = \sum_{i=1}^{n} X_i$. Let $\tau \equiv \inf\{n > 0 : S_n > 0\}$. Prove that τ is a random variable.

(d) If $\tau(\omega) < \infty$ for all $\omega \in \Omega$, prove that S_τ is a random variable.