Ph.D. Qualifying Exam
Wednesday–Thursday, August 17–18, 2022

- **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 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a sequence of i.i.d. paired samples, where $X_i \in \mathbf{R}^d, Y_i \in \mathbf{R}, i = 1, \dots, n$. Conditioned on X_i, Y_i follows a Poisson distribution with mean parameter $\lambda_{X_i} = e^{\beta^T X_i}$, where $\beta \in \mathbf{R}^d$ and β^T indicates the transpose of β. 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) Find the expressions for the gradient $L = \dfrac{\partial l(\beta)}{\partial \beta}$ and Hessian matrix $H = \dfrac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}$. Prove that the log-likelihood function $l(\beta)$ is concave. (Hint: show that H is semi-negative definite.)

(c) Write out the Newton-Raphson algorithm to find the root of the gradient L.

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

Problem 3. (5166) Table 1 summarizes the observed data for a study with the outcome being the measurement of some blood biomarkers. Two factors, A and B, are of interest. Each factor indicates the absence (0) or presence (1) of some potential risk factor. Twelve subjects were recruited for the study, three for each of the four combinations of the two factors. Answer the following questions. Conduct all hypothesis tests at the 0.05 significance level.

Table 1: Observed data for the study.

Subject	Outcome	A	B
1	42.4	1	1
2	43.2	1	1
3	42.2	1	1
4	43.6	1	0
5	42.3	1	0
6	41.2	1	0
7	42.5	0	1
8	42.7	0	1
9	42.6	0	1
10	39.7	0	0
11	41.5	0	0
12	40.4	0	0

(a) (3 points) Write the factor-effects model for this study, including the main effects and interaction. Define all quantities that appear in your model specification, and state all model assumptions.

(b) (5 points) Some output results for the model fit are shown in Table 2. Find the values of the missing items [i] to [viii] in the table. Show your derivations and justifications.

Table 2: Analysis of variance table.

```
Source       Df  Sum Sq  Mean Sq  F value   Pr(>F)
A             1  2.5208   2.5208   3.9439  0.08229
B           [i]    [iv]   3.9675   6.2073  0.03744
A:B        [ii]     [v]    [vii]   3.9439  0.08229
Residuals [iii]    [vi]   [viii]
```

(c) (5 points) Based on these results, compute a F-statistic and state the degrees of freedom for a joint test of whether or not the interaction term and the main effect of B are significant. Give the conclusion of this joint test. You may need the information from the critical values of F-distribution in Table 3.

Table 3: Critical values for F-distribution with degrees of freedom (df_1, df_2) and 0.05 tail probability.

	$\text{df}_1 = 1$	2	3	4
$\text{df}_2 = 5$	6.608	5.786	5.409	5.192
6	5.987	5.143	4.757	4.534
7	5.591	4.737	4.347	4.120
8	5.318	4.459	4.066	3.838
9	5.117	4.256	3.863	3.633
10	4.965	4.103	3.708	3.478

(d) (3 points) Compute a F-statistic and state the degrees of freedom for testing whether the overall effects of the two factors are significant. Give the conclusion of this joint test. You may need the information from the critical values of F-distribution in Table 3.

(e) (4 points) Suppose we fit an additive effects model (no interaction) to the data. Will the MSE change? If so, how and why? If the MSE will change, compute the new MSE and the various F statistics.

Problem 4. (5166) Suppose that we aim to determine the effects of 6 factors in an experiment. We refer to these 6 factors as A, B, C, D, E, and F.

(a) (8 points) Construct a 2^{6-2} fractional factorial design such that its resolution is as high as possible. Specify the generators of your design. What is the resolution?

(b) (8 points) What is the defining relation of your design? Give the confounding pattern.

(c) (4 points) Which effects are confounded with the main effect of factor A in your design? Which effects are confounded with the two-factor interaction AB?

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

$$Y = \alpha + \beta X_1 + e\,. \tag{1}$$

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, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2 \equiv \widehat{var}(e)$.

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

(c) What assumptions (in addition to (1)) 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 $i \neq j$. Is the sample mean $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ still the MLE for μ? What is the variance of $\overline{Y}$? Does the variance go to zero when $n \to \infty$ in this case, and why?

Note: Here is 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 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., $(Y_{1i}, Y_{2i}, X_i^T) \sim (Y_1, Y_2, X^T)$, $i = 1, \ldots, n$.

(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) It is widely believed that the influenza vaccine can lower risk for myocardial infarction (MI). The data below come from a study investigating the association between the influenza vaccine and MI among people aged 65 years and older. Electronic health records (EHRs) were used to identify patients who had no record of a previous flu vaccine and received the flu vaccine between Sept. 1, 1996 and January 31, 1997. Unvaccinated patients in the EHRs were designated as controls. This “1997 cohort” was followed for a year, with subjects censored upon death, subsequent flu vaccination (for controls), and departing their clinical practice.

The “1997 Cohort”:

	MI	No MI	Total
Flu Vaccine	80	7,607	7,687
No Flu Vaccine	423	54,534	54,957
			62,644

(a) Estimate the odds ratio for MI associated with vaccination vs no vaccination for these data. Provide a 95% confidence interval and interpret.

(b) The investigators also classified patients according to whether they had comorbidities associated with increased risk for MI. Among the patients with comorbidities, they found $\widehat{\text{OR}}_{co} = 1.50$ with 95% CI (1.16, 1.92); among those without comorbidities, they found $\widehat{\text{OR}}_{noco} = 0.16$, 95% CI (0.04, 0.67). Moreover, when considering presence/absence of comorbidities as strata, they found

- The Mantel-Haenszel estimate of the common OR is 1.23 with 95% CI (0.96, 1.56).
- The Cochran-Mantel-Haenszel test that the common OR is 1.0 yields the test statistic of 2.71 on 1 degree of freedom.
- The Breslow-Day test yields a test statistic of 13.5 on 1 degree of freedom.

Interpret these analyses – what do you conclude about whether cormorbidities *confound* the association between vaccination and MI risk? Do comorbidities *interact* with vaccine status to affect risk for MI? Be sure to clearly state all associated hypotheses.

Problem 8. (5198)

(a) In the usual 2×2 table (see formula sheet), derive the approximate variance for log of the odds ratio estimate

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

Justify every step in your derivation.

(b) What is an approximate variance for the odds ratio estimate (i.e., not the log OR estimate) that follows from (2)?

Formula Sheet for STA 5198 Problems

Some known formulae based on the usual 2 × 2 table $\begin{array}{c|c|c|} & D & \overline{D} \\ \hline E & a & b \\ \hline \overline{E} & c & d \\ \hline \end{array}$.

$$\begin{aligned} \text{RR} &= \text{relative risk} \\ \text{OR} &= \text{odds ratio} \\ \text{AR} &= \text{attributable risk} = \frac{R - R_{\overline{E}}}{R}, \text{ where } R \text{ is the overall risk} \end{aligned} \tag{3}$$

$$\widehat{\text{Var}}\left(\log \widehat{\text{RR}}\right) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d} \tag{4}$$

$$\widehat{\text{Var}}\left(\log \widehat{\text{OR}}\right) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{5}$$

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

$$\frac{(ad-bc)\exp(\pm u)}{nc + (ad-bc)\exp(\pm u)}, \tag{6}$$

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:

$$\left(\sum_i \frac{a_i d_i}{n_i}\right) \Big/ \left(\sum_i \frac{b_i c_i}{n_i}\right) \tag{7}$$

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

$$X^2 = \frac{\left(\sum a_i - \sum E(a_i)\right)^2}{\sum \text{Var}(a_i)} \tag{8}$$

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

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

$$X^2 = \sum \frac{(a_i - \text{E}(a_i))^2}{\text{Var}(a_i)}, \tag{9}$$

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 $\{O_{ij}\}$:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n}n \tag{10}$$

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

$$X^2_{(L)} = \frac{(T_1 - \frac{n_D}{n} T_2)^2}{V}, \tag{11}$$

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) $\{E_{ij}\}$:

$$\kappa = \frac{p_O - p_E}{1 - p_E}, \qquad p_O = \sum O_{ii}/n, \quad p_E = \sum E_{ii}/n \tag{12}$$

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

z	0.84	1.04	1.28	1.64	1.96	2.33
$P(Z \leq z)$	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 $\Pr(W \geq q) = p$.

	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) A monkey types **13** digits at random. (Each keystroke is independent of the others, and all 10 digits $0, 1, 2, \ldots, 9$ are equally likely with probability $1/10$.) Let X_i = the number of digits which appear **exactly** i times in the monkey's typed output. Find the following. Explain your answers.

(a) $P(X_4 = 3)$

(b) EX_3

(c) $P(X_5 = 0)$

Problem 10. (5326) Let X_1 and X_2 be independent $N(0, 1)$ random variables.

(a) Find the density (pdf) of $Y = \dfrac{(X_1 - X_2)^2}{2}$.

(b) Find the density (pdf) of $Z = \dfrac{X_1^2}{X_1^2 + X_2^2}$.

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

Name	**Notation**	$f(x)$	E(X)	Var(X)
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
Bernoulli	Bernoulli(p)	$p^x(1-p)^{(1-x)}, \quad x = 0, 1$	p	$p(1-p)$
Beta	Beta(α, β)	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

where $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.

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 moment (MOM) estimator for μ. Call it $\hat{\mu}$.

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

(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

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

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

Problem 12. (5327) We consider $X_1, \ldots, X_n$ i.i.d from Bernoulli(p), where the parameter space is $\Theta = \{p : 0.25 \leq p \leq 0.75\}$.

(a) Find the maximum likelihood estimator for p.

(b) Find the maximum likelihood estimator for $\log\left(\frac{p}{1-p}\right)$.

(c) The Beta distribution is the conjugate prior for the Bernoulli distribution. If we want to construct a Bayesian estimator for p, would you use the Beta distribution as the prior distribution? Why or why not?

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

(a) Give the definition of a probability measure.

(b) Suppose X is a random variable on $(\Omega, \mathcal{F}, \mu)$ with $X \geq 0$ and $\mathbb{E}X = 1$. Let $\nu(A) = \mathbb{E}(X 1_A)$ for all $A \in \mathcal{F}$. Show ν is a probability measure on $(\Omega, \mathcal{F})$.

(c) Show that $\mu(A) = 0$ implies $\nu(A) = 0$.

(d) If $X > 0$ instead of $X \geq 0$, show that $\mathbb{E}_\nu Y = \mathbb{E}_\mu(XY)$ where $\mathbb{E}_\lambda$ is the expectation with respect to the measure λ.

Problem 14. (6346) Answer the following.

(a) Define convergence in probability.

(b) Let X and Y_n be random variables with $\mathbb{E}Y_n = var(Y_n) = 1/n$. Show $X_n = X + Y_n$ converges to X in probability.

(c) Let W_n be a sequence of random variables. Prove that W_n converges in probability if and only if for every $\varepsilon > 0$,

$$P(|W_n - W_m| > \varepsilon) \to 0$$

as n and m go to infinity.