Ph.D. Qualifying Exam
Monday–Tuesday, January 3–4, 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 Y be a discrete random variable with nonnegative integer values. For any $y \in \{0, 1, 2, \cdots\}$,

$$\text{Prob}(Y = y) = \alpha_1 P_1(y; \lambda_1) + \alpha_2 P_2(y; \lambda_2),$$

where P_1 and P_2 are two Poisson probability mass functions with means λ_1 and λ_2, respectively. Also, $0 \leq \alpha_1, \alpha_2 \leq 1$, such that $\alpha_1 + \alpha_2 = 1$. Given n *i.i.d.* observations $\{Y_i\}_{i=1}^n$, our goal is to find the maximum likelihood estimate of

$$\theta = (\alpha_1, \lambda_1, \alpha_2, \lambda_2).$$

Use the EM algorithm for iteratively estimating θ. Let $\theta^{(m)}$ be the estimated value at the m^{th} iteration. Derive the mathematical formula to update it for $\theta^{(m+1)}$.

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) The pine engraver beetle is an insect that attacks stressed trees. The beetle is, therefore, an important pest in recently clear-cut forest areas. An experiment was conducted to determine the effects of host material and type of lure on trap catches of pine engraver. A factorial unbalanced experimental design consisting of three levels of factor *lure* type and three levels of factor *host* material was used.

The three levels for factor host material were red pine (R), white pine (W), and white spruce (S). The three levels for the factor lure type were chemical (A), a group of 20 male insects (B), and a control with no lure (C). Six-inch bolts of red pine, white pine, or white spruce were placed in a metal cage with wire screening, next to a Lindgren Funnel Trap. For lure treatment A, 0.1 mg of a 100:1 mixture of Ipsdienol:Lanierone was placed inside the cage. For lure treatment B, 20 live males were placed inside the cage. For treatment C, no lure was placed inside the cage.

Treatment combinations of the two factors above were randomly assigned to 63 trapping cages. After 48 hours, the total number of beetles (males and females) caught in the Lindgren trap was counted. One trap was attacked by raccoons, and so the trap's contents could not be counted.

Assume significance level 0.05 for all tests.

(a) Write the factor-effects model for this experiment, including both predictors and their interaction. State all model assumptions.

(b) Fill in the blanks in the tables for Models 1 and 2 below. Briefly explain each answer.

Model 1

Source	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
factor(host)		1861.9			1.124e-10
factor(lure)		7535.3			<2.2e-16
Residuals		1508.8		–	–

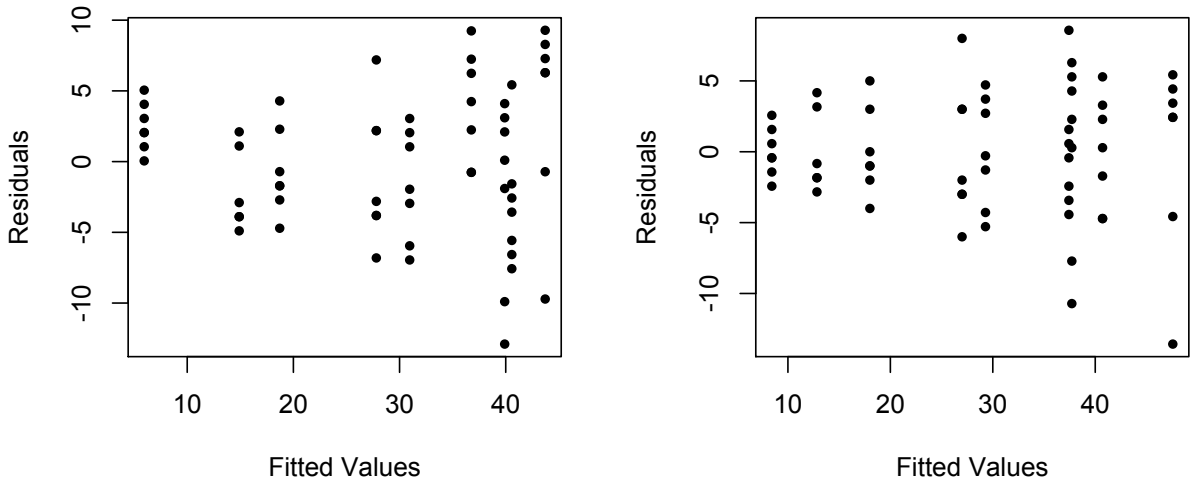
Model 2

Source	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
factor(host)		1861.9			3.763e-12
factor(lure)		7535.3			<2.2e-16
factor(host):factor(lure)		412.5			0.00174
Residuals				–	–

(c) Which model is better? Why?

(d) If the two predictors, host and lure, enter the model in reverse order, will their sums of squares be different from those shown in the tables? Justify your answer.

(e) The plots in the figure below show the residuals versus the fitted values for Models 1 and 2. These plots are useful for diagnosing the fits with respect to which model assumption? Does that assumption appear to be violated for either model? If so, suggest a remedy.



Residuals versus fitted values for Model 1 (left) and Model 2 (right).

Problem 4. (5166) Suppose that an experimenter performs a 2^{5-2} fractional factorial design for factors **1**, **2**, **3**, **4**, and **5**. The experimenter chose generators **4 = 123** and **5 = 23**. After analyzing the results from this design, the experimenter decided to perform a second 2^{5-2} design exactly the same as the first but with signs changed in column **3** of the design matrix. Answer the following questions, and give your justifications.

(a) How many runs does the first design contain?

(b) Specify the confounding pattern of the first design.

(c) Give the sets of generators for the second design.

(d) What is the resolution of the second design?

(e) What are the defining relation and the resolution of the combined design?

(f) Given the sign switching for factor **3**, can you find a set of generators for the first design such that the resolution of the resulting combined design is higher than that derived in **(e)**? If no, give the reason; if yes, specify such generators, and explain why they satisfy the condition.

Problem 5. (5167) Suppose we fit a regression of Y on (X, Z) with the true mean function $\mathrm{E}(Y \mid X = x, Z = z) = 2 + 3x + 4z$. Further suppose that (X, Z) is bivariate normal, with the five parameters $(\mu_x, \mu_z, \sigma_x^2, \sigma_z^2, \rho_{xz})$.

(a) What are the true regression parameters of X on Z? (Hint: $X = \beta_0 + \beta_1 Z + \varepsilon$, $\mathrm{var}(\varepsilon) = \sigma^2$.)

(b) Provide conditions under which the mean function for $\mathrm{E}(Y \mid X)$ is linear but has a negative coefficient for X.

Problem 6. (5167) Based on the following R output, answer the questions.

```
> pairs(cbind(Y,X1,X2))
> cor(cbind(Y,X1,X2))
           Y        X1        X2
Y  1.0000000 0.8906967 0.8943581
X1 0.8906967 1.0000000 0.9965905
X2 0.8943581 0.9965905 1.0000000
> m<-lm(Y~X1+X2)
> summary(m)

Call:
```

```
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-14900.1  -2386.1   -192.5   1888.4  30253.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6168.1959   642.6794  -9.598  < 2e-16 ***
X1             -0.7658     2.8748  -0.266  0.79017
X2              8.4317     2.8922   2.915  0.00387 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4543 on 258 degrees of freedom
Multiple R-squared:  0.7999,Adjusted R-squared:  0.7984
F-statistic: 515.8 on 2 and 258 DF,  p-value: < 2.2e-16

> m2<-lm(Y~X2)
> summary(m2)

Call:
lm(formula = Y ~ X2)

Residuals:
     Min       1Q   Median       3Q      Max
-14934.5  -2423.2   -181.8   1839.7  30205.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6167.6009   641.5219  -9.614   <2e-16 ***
X2              7.6639     0.2382  32.175   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4535 on 259 degrees of freedom
Multiple R-squared:  ?.???,Adjusted R-squared:  ?.???
F-statistic:  1035 on 1 and 259 DF,  p-value: < 2.2e-16
> qt(c(0.95,0.975,0.99,0.995),258)
[1] 1.650781 1.969201 2.340888 2.595019
> qt(c(0.95,0.975,0.99,0.995),259)
[1] 1.650758 1.969166 2.340831 2.594945
> qt(c(0.95,0.975,0.99,0.995),260)
[1] 1.650735 1.969130 2.340775 2.594870
> qnorm(c(0.95,0.975,0.99,0.995))
[1] 1.644854 1.959964 2.326348 2.575829
```

(a) What does the first line of R script "**pairs(cbind(Y,X1,X2))**" create?

(b) Write down the fitted regression model of Y on X_2. What are the assumptions of this simple linear model? Is there any evidence against those model assumptions?

(c) What are the R^2 values for models m and $m2$?

(d) If we use an F-test to compare models m and $m2$, what are the null and alternative hypotheses? What would be the p-value?

(e) For a new observations $(X_1^*, X_2^*) = (200, 2000)$ what are the predicted values from model m and from $m2$? Which prediction do you think is more accurate, and why?

Problem 7. (5198) An investigation of the association between fruit and vegetables in the diet and risk for adenomatous polyps of the colon recruited individuals who had recently undergone sigmoidoscopy of the colon. Individuals for whom adenomatous polyps were found were matched to individuals for whom no adenomatous polyps were found by clinic, time of screening, age and sex. Low fruit and vegetable consumption was defined as two or fewer servings per day (on average) as reported by the subject. In 45 pairs, the person with adenomatous polyps reported low consumption, and the polyp-free individual did not. In 24 pairs, the person with adenomatous polyps did not report low consumption and the polyp-free individual did. There were 11 pairs in which both the person with and the person without adenomatous polyps reported low consumption, and the remaining 415 people with adenomatous polyps together with their matched polyp-free subjects both reported high consumption.

(a) Do these data provide evidence for an association between low fruit and vegetable consumption and elevated risk for adenomatous polyps? Explain clearly your chosen association measure, its estimate and uncertainty, and provide a clearly stated conclusion. If you perform hypothesis testing, state all hypotheses explicitly and clearly state your conclusion.

(b) Does your conclusion change if the matching performed in this study is ignored in the analysis? Can you explain why there is (or isn't) a difference in the results of the matched and unmatched analyses?

Problem 8. (5198) A study of the association between a history of playing soccer and dementia evaluated 7676 former soccer players and found evidence of neurodegenerative disease in 386. Evaluation of 23,028 non-soccer players (chosen to be similar to the soccer players in other aspects) found evidence of neurodegenerative disease in 366.

(a) Estimate the relative risk of neurodegenerative disease among former soccer players compared to non-soccer players. Provide a 95% confidence interval and interpret.

Problem 8 continued: The attributable risk we defined in class is given by $(R - R_{\overline{E}})/R$, where R is the overall risk. However, some investigators use a different measure they also call "attributable risk," given by

$$\phi = \frac{R_E - R_{\overline{E}}}{R_E} = \frac{RR - 1}{RR} = 1 - \frac{1}{RR}.$$

Here ϕ is the fraction of risk among the exposed that exceeds the risk in the unexposed. Given the estimate $\widehat{\text{RR}}$ of RR, ϕ is estimated using $\hat{\phi} = 1 - 1/\widehat{\text{RR}}$.

(b) Notice that $\ln(1 - \hat{\phi}) = -\ln(\widehat{\text{RR}})$. Use this relationship to give an expression for an approximate 95% confidence interval for $\ln(1 - \phi)$.

(c) If the confidence interval for $\ln(1-\phi)$ is (L, U), what is the corresponding confidence interval for ϕ?

(d) Give an approximate 95% confidence interval for ϕ for the data relating a history of soccer playing to neurodegenerative disease. Interpret.

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

$$\begin{aligned}
&\text{RR} = \text{relative risk} \\
&\text{OR} = \text{odds ratio} \\
&\text{AR} = \text{attributable risk} = \frac{R - R_{\overline{E}}}{R}, \text{ where } R \text{ is the overall risk} \qquad (1)
\end{aligned}$$

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

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

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

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

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:

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

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

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

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

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

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

where now $\mathrm{E}(a_i)$ and $\mathrm{Var}(a_i)$ 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 $V = n_D n_{\bar{D}} (n T_3 - T_2^2)/[n^2(n-1)]$

- 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			
$\nu = \mathrm{df}$	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. (The parts are **not** related.)

(a) Let $M(t)$ be the moment generating function (mgf) of X. Assume that $M(t)$ is well-defined and finite for all values of t in a neighborhood $(-\varepsilon, \varepsilon)$ of zero. Define $\psi(t) = \log M(t)$. Show that

$$\frac{d}{dt}\psi(t)\bigg|_{t=0} = EX \quad \text{and} \quad \frac{d^2}{dt^2}\psi(t)\bigg|_{t=0} = \text{Var}(X)\,.$$

Clearly state all the facts used in your argument.

(b) Let Z have a Binomial(n, p) distribution. Find $E(Ze^{tZ})$ where t is an arbitrary real number.

Problem 10. (5326) Let X, Y, Z be independent Poisson random variables with means α, β, and λ, respectively.

(a) Find the distribution of $X \,|\, X+Y$. That is, find $P(X = j \,|\, X+Y = k)$ for arbitrary integers j, k satisfying $0 \le j \le k$.

(b) Find the distribution of $X+Y+Z \,|\, X+Y$. That is, find $P(X+Y+Z = j \,|\, X+Y = k)$ for arbitrary integers j, k satisfying $j \ge k \ge 0$.

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

Name	**Notation**	$f(x)$	E(X)	Var(X)
Poisson	Poisson(λ)	$\dfrac{\lambda^x e^{-\lambda}}{x!}\,, \quad x = 0, 1, 2, \ldots$	λ	λ
Normal	N(μ, σ^2)	$\dfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2

Problem 11. (5327) We consider $X_1, \ldots, X_n$ i.i.d. from Poisson(λ), where $\lambda > 0$ is the unknown parameter.

(a) Consider the estimator $\hat{\lambda} = \sum_{i=1}^n a_i X_i$, where $a_i \geq 0, \sum_{i=1}^n a_i = 1$ are weights. What a_i should we use to minimize $E(\hat{\lambda} - \lambda)^2$?

(b) For your choice of a_i in part (a), is the resulting $\hat{\lambda}$ a sufficient statistic for λ? Why or why not?

(c) For your choice of a_i in part (a), find the asymptotic distribution of $\log \hat{\lambda}$.

Problem 12. (5327) We consider $X_1, \ldots, X_n$ i.i.d from $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. We denote $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

(a) Show that $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

(b) Show that $\hat{\sigma}^2$ is not the maximum likelihood estimator.

(c) Show that $\hat{\sigma}^2$ is $\sqrt{n}$-consistent and find its asymptotic distribution.

Problem 13. (6346) Let X_n be a martingale with respect to the filtration $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots \subset \mathcal{F}$ where $\mathcal{F}_i = \sigma(X_1, X_2, \ldots, X_i)$.

(a) Give the definition of a stopping time T.

(b) Give the definition of $\mathcal{F}_T$.

(c) If X_n is a martingale, show X_T is measurable with respect to $\mathcal{F}_T$.

(d) If S is also a stopping time, show $\min(S, T)$ is a stopping time.

Problem 14. (6346) Let X_n be a sequence of independent Bernoulli random variables with fixed parameter $p \in (0, 1)$. Let X be a Bernoulli random variable with the same parameter p and independent of the X_n.

(a) Describe the relations between convergence in distribution and convergence in probability.

(b) Prove or disprove: $X_n \xrightarrow{D} X$.

(c) Prove or disprove: $X_n \xrightarrow{P} X$.