Ph.D. Qualifying Exam
Wednesday–Thursday, August 18–19, 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 Y be a continuous random variable with probability density function:

$$Y \sim \alpha_1 f_1(y; \mu_1, \sigma_1^2) + \alpha_2 f_2(y; \mu_2, \sigma_2^2),$$

where f_1 and f_2 are two Gaussian density functions with means μ_1, μ_2 and variances σ_1^2, σ_2^2, 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

$$\theta = (\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2).$$

(a) 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. Prove that

$$L(\theta^{(m+1)}) \geq L(\theta^{(m)}).$$

Problem 2. (5106) Let $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be a sequence of i.i.d. paired samples, where $X_i \in \mathbf{R}^d, Y_i \in \mathbf{R}, i = 1, \cdots, n$. 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

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

(b) Find the expressions for the gradient $L = \frac{\partial l(\beta)}{\partial \beta}$ and Hessian matrix $H = \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}$. 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 3. (5166) Consider the following linear model for a one-factor randomized design:

$$Y_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i,$$

where for each i, the random errors $\{\epsilon_{ij}, j = 1, \ldots, n_i\}$ are assumed to be i.i.d. $N(0, \sigma_i^2)$. The sample variances for the k treatments are denoted by $s_i^2, \;\; i = 1, \ldots, k$.

(a) It is well known that

$$(n_i - 1)s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \sim \sigma_i^2 \chi^2_{n_i - 1}.$$

Based on this result, show that the distribution of $\log(s_i^2)$ can be approximated by $N(\log(\sigma_i^2), 2(n_i - 1)^{-1})$.

(b) In a one-way layout design with $k = 10$ and $n_1 = \cdots = n_{10} = 6$, the s_i^2's and $\log(s_i^2)$'s are given in the following table:

s_i^2	0.56	0.82	0.45	1.02	0.34	0.77	0.52	0.88	0.31	1.20
$\log(s_i^2)$	−0.58	−0.20	−0.80	0.02	−1.08	−0.26	−0.65	−0.13	−1.17	0.18

Based on the result in part (a), perform a test on "$H_0 : \sigma_1^2 = \cdots = \sigma_{10}^2$" against the alternative hypothesis "$H_a :$ some of the variances are not equal" using $\alpha = 0.05$.

Problem 4. (5166) Suppose a researcher wants to design a 2^{7-2} experiment with factors A, B, C, D, E, F, G. Upon further consideration, it was determined that the last two factors, F and G, must be assigned to interactions involving the first five factors.

(a) Suppose the factor assignment of F = ABC and G = BCD is made. Ignoring three-factor interactions and higher, list all of the two-factor interaction aliases.

(b) What is the resolution of the design in (a) and explain how it can be determined.

(c) Instead, suppose the factor assignment of F = ABCD and G = ABDE was made. Ignoring three-factor interactions and higher, list all of the two factor interaction aliases.

(d) In general, would either of the designs of (a) or (c) be preferred over the other? Explain why or why not.

Problem 5. (5167) Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i\,, \quad i = 1, \ldots, n, \tag{1}$$

where $e_i \sim N(0, \sigma^2)$ are iid and independent of X_i. Suppose the predictor is transformed as $Z_i = 2X_i + 1$ and we fit a linear regression model of Y_i on Z_i.

(a) Write down the simple linear regression model of Y on Z (with specific model assumptions).

(b) What are the least-squares estimates of the unknown parameters in (a). (Hint: please include the estimate of σ^2.)

(c) Express part (b)'s answer in terms of the original data (Y_i, X_i).

(d) Describe the hypothesis test for $\beta_0 = 0$ in model (1) using the transformed data (Y_i, Z_i) and the estimates in part (b).

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

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

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) Suppose we are interested in forward stepwise variable selection. What is the Bayesian information criterion in this model? Describe how to use it in the forward stepwise variable selection.

Problem 7. (5198) An epidemiological study identified and enrolled a sample of 100 individuals. On January 1, 2018, the study start date, 3 of the 100 individuals were found to have the chronic disease under study. During the one-year study, 7 new cases of disease were found. The diagram below shows the follow-up for the 10 cases. Subjects 1, 2 and 3 contracted disease prior to the study start on January 1, 2018. The lines depict the time that the subject lived with the disease, with the × marking time of death and ○ indicating the subject is still alive. Assume the remaining 90 subjects did not become ill or die and were observed for the full year.

Estimate the following quantities:

(a) The *prevalence* of disease on January 1, 2018.

(b) The *prevalence* of disease on September 30, 2018.

(c) The cumulative *incidence* of *disease* during 2018 (i.e., during the study period).

(d) The cumulative *incidence* of *death* during 2018 (i.e., during the study period).

(e) Give an approximate 95% confidence interval for the cumulative *incidence* of *disease* during the study.

(f) If a new treatment were developed that *cured* people of this disease within one month of their diagnosis, how would this affect the *prevalence* rate in future studies? How would this affect the *incidence* rate in future studies?

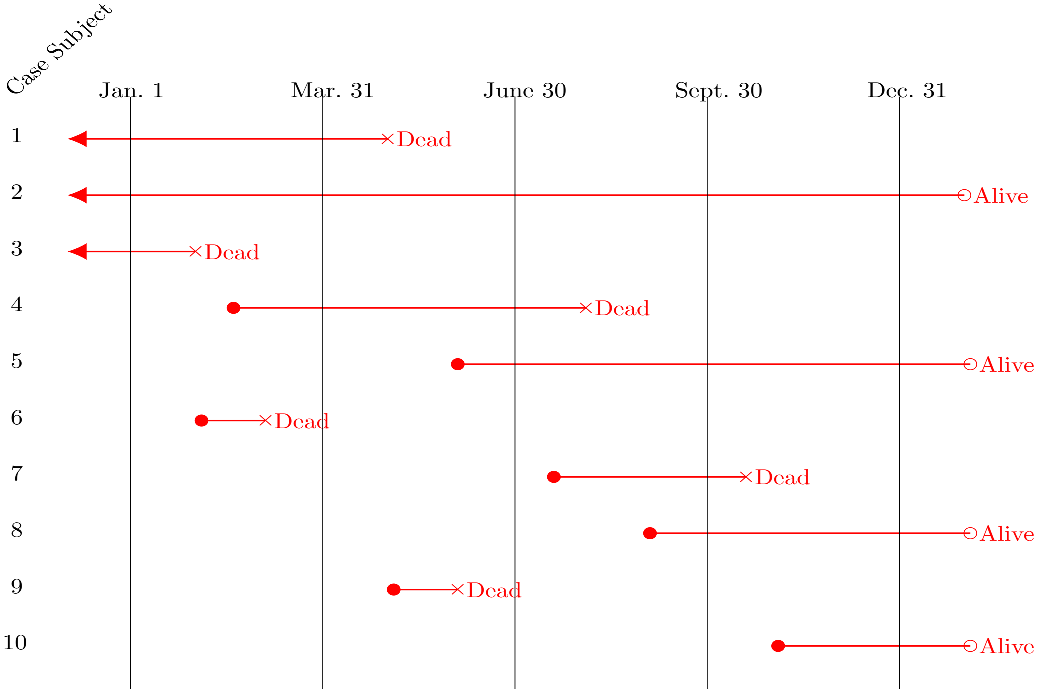

Figure 1: Follow up for the 10 cases in the epidemiological study. ← pre-existing disease; ● disease onset; × death; ○ still alive.

Problem 8. (5198) A recent study recruited women aged 50-71 years with no cancer diagnosis and assessed their history of oral contraceptive use using a baseline survey. The women were then prospectively followed until their first diagnosis of any cancer or their death. Of particular interest is the odds ratio for endometrial cancer among oral contraceptive (OC) users relative to non-users.

Using the results below, *discuss whether OC use is associated with endometrial cancer diagnosis and whether age confounds or interacts with the effect of OC use on this cancer.*

Analysis set	n	Crude $\widehat{OR}$	95% confidence interval
All women	$114,601$	0.89	$(0.82, 0.97)$
Women aged < 60	$30,613$	1.23	$(1.02, 1.47)$
Women aged ≥ 60	$83,988$	0.80	$(0.72, 0.88)$

With age considered a stratification factor,

- the Mantel-Haenszel estimate and 95% confidence interval for the odds ratio are 0.88 (0.81, 0.96)
- the Breslow-Day test statistic is $X^2_{BD} = 16.2$ on 1 degree of freedom
- the Cochran-Mantel-Haenszel test statistic is $X^2_{CMH} = 8.3$ on 1 degree of freedom

Formula Sheet for STA 5198 Problems

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

$$\begin{aligned}
\text{RR} &= \text{relative risk} \\
\text{OR} &= \text{odds ratio} \\
\text{AR} &= \text{attributable risk} = \frac{R - R_{\overline{E}}}{R}, \text{ where } R \text{ is the overall risk}
\end{aligned} \tag{2}$$

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

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

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

$$\frac{(ad - bc)\exp(\pm u)}{nc + (ad - bc)\exp(\pm u)}, \tag{5}$$

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:

$$\left(\sum_i \frac{a_i d_i}{n_i}\right) \Big/ \left(\sum_i \frac{b_i c_i}{n_i}\right) \tag{6}$$

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

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

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

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

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

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 $\{O_{ij}\}$:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{O_{i\cdot}}{n} \frac{O_{\cdot j}}{n} n \tag{9}$$

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

$$X^2_{(L)} = \frac{(T_1 - \frac{n_D}{n} T_2)^2}{V}, \tag{10}$$

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

$$\kappa = \frac{p_O - p_E}{1 - p_E}, \qquad p_O = \sum O_{ii}/n, \quad p_E = \sum E_{ii}/n \tag{11}$$

- 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	9.3	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 T be the triangle in the (x, y)-plane bounded by the lines $x = 0$, $y = 0$, and $x + y = 1$; the three vertices of this triangle are the points $(0, 0)$, $(1, 0)$, and $(0, 1)$.

(a) Let U be a random variable with density (pdf) given by

$$f_U(u) = 2u\,, \quad 0 < u < 1\,.$$

The line $x + y = U$ intersects the triangle T. Let S be the part of T which lies below this line; S is the triangle with vertices $(0, 0)$, $(U, 0)$, and $(0, U)$. Let W be the area of S. Find the density of the random variable W.

(b) Suppose n random points are distributed uniformly and independently on the triangle T. (The points are also independent of U.) Let J be the number of these points which lie inside the smaller triangle S. Find EJ.

(c) Find $\mathrm{Var}(J)$.

(d) Find $P(J = i)$ for $i \in \{0, 1, \ldots, n\}$.

Note: The answers in (b), (c), and (d) should **NOT** be conditional on U or W.

Problem 10. (5326) Suppose that tosses of a coin are independent with probability p of heads. Answer the following. Simplify your answers if possible.

(a) Two people each toss this coin until they get their first head (and then stop). What is the probability they both toss the coin exactly the same number of times?

(b) Three people each toss this coin until they get their first head. What is the probability that all three of them have exactly the same number of tosses?

(c) Three people each toss this coin until they get their first head. What is the probability that all three have a **different** number of tosses?

Hint: One approach to part (c) is to use the principle of inclusion-exclusion.

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

Name	**Notation**	density $f(x)$	$E(X)$	Var(X)
Normal	$\mathrm{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2

For both STA 5327 problems, we consider $X_1, \ldots, X_n$ i.i.d from $N(\theta, 1)$. We denote $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Problem 11. (5327) Consider the parameter space $\Theta = \{\theta : \theta \in \mathbb{R}\}$.

(a) Construct a Bayesian estimator for θ by considering the prior $\theta \sim N(0, \lambda)$, where $\lambda > 0$.

(b) Find the mean squared error (MSE) for the Bayesian estimator you constructed.

(c) When is the Bayesian estimator better than $\bar{X}$ in terms of MSE?

Problem 12. (5327) Consider the parameter space $\Theta^* = \{\theta : 0 \le \theta < \infty\}$,

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

(b) Show that, when $\theta = 0$, the MLE is closer to the truth than $\bar{X}$ with a probability of 1/2.

(c) Is the MLE unbiased?

(d) For the hypotheses

$$H_0 : \theta \le 1, \theta \in \Theta^* \quad \text{vs.} \quad H_1 : \theta > 1, \theta \in \Theta^*$$

show that the likelihood ratio test rejects H_0 if $\bar{X} > c$ for some constant c.

Problem 13. (6346) Suppose $(\Omega, \mathcal{F}_0, \mu)$ is a probability space, $\mathcal{F}$ is a σ-field with $\mathcal{F} \subset \mathcal{F}_0$, $X : \Omega \to \mathbb{R}$ is measurable with respect to $\mathcal{F}_0$, and $\mathbb{E}|X| < \infty$.

(a) Give the definition of conditional expectation, $\mathbb{E}(X|\mathcal{F})$.

(b) Let $Y = \mathbb{E}(X|\mathcal{F})$. Show that $\mathbb{E}|Y| \leq \mathbb{E}|X|$.

(c) Suppose $\mathcal{G}$ is a σ-field with $\mathcal{G} \subset \mathcal{F}$. Show $\mathbb{E}\left[\mathbb{E}(X|\mathcal{F})|\mathcal{G}\right] = \mathbb{E}(X|\mathcal{G})$.

Problem 14. (6346) Suppose $f_i : \Omega \to \mathbb{R}$ are measurable with respect to $(\Omega = \mathbb{R}, \mathcal{F} = \mathbb{B})$, $i = 1, 2, ...$, where $\mathbb{B}$ is the Borel σ-field.

(a) Give the definition of a measurable function.

(b) Show that $\sup_i f_i$ and $\inf_i f_i$ are measurable.

(c) Give the definitions of $\limsup_i f_i$ and $\liminf_i f_i$.

(d) Show that $\limsup_i f_i$ and $\liminf_i f_i$ are measurable.