Ph.D. Qualifying Exam
Friday–Saturday, January 5–6, 2018

- **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.**
- **Pages 6 and 7 give formulas and tables for possible use in the 5198 problems.**

Problem 1. (5106) Let H be an $n \times n$ householder matrix given by

$$H = I_n - 2\frac{vv^T}{v^T v},$$

for any non-zero n-length column vector v. Prove that H is a symmetric, orthogonal, and reflection matrix. That is, show H satisfies:

$$\text{i) } H = H^T, \quad \text{ii) } HH^T = I_n\,, \quad \text{and} \quad \text{iii) } \det(H) = -1.$$

Problem 2. (5106) Let Y be a continuous random variable with probability density function:

$$Y \sim \alpha_1 f_1(y; \mu_1, \sigma_1^2) + \alpha_2 f_2(y; \mu_2, \sigma_2^2) + \alpha_3 f_3(y; \mu_3, \sigma_3^2),$$

where f_1, f_2 and f_3 are three Gaussian density functions with means μ_1, μ_2, μ_3 and variances σ_1^2, σ_2^2, σ_3^2, respectively. Also, $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$, such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Given n *i.i.d.* observations $\{Y_i\}_{i=1}^n$, our goal is to find the maximum likelihood estimate of

$$\theta = (\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2, \alpha_3, \mu_3, \sigma_3).$$

(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 3. (5166) Three treatments, A, B, and C, are compared in an experimental 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) For a given i, is the process $\{X_i(t),\ t = 1, 2, \ldots\}$ stationary?

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

(c) 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 under-estimates σ^2.

Problem 4. (5166) Consider the following two 2^{8-3} fractional factorial designs:

d_1:	with	6=1234,	7=1235	8=2345
d_2:	with	6=123,	7=145	8 =345

(a) What are the advantages and disadvantages of a fractional factorial design compared with a full factorial design?

(b) What is the defining relation for each of the two designs? What is the resolution of each design? For each design, write out the confounding pattern of the main effects $\{1, 2, 3, 4, 5, 6, 7, 8\}$ with other effects.

(c) Write out the word-length pattern for each design. Which design is better? Please justify your choice.

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

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n,$$

where $\{e_i,\, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma^2)$ variables. Suppose that the predictor x_i in the model is replaced by $z_i = 2x_i + 1$. The simple linear regression model becomes

$$y_i = \gamma_0 + \gamma_1 z_i + e_i = \gamma_0 + \gamma_1(2x_i + 1) + e_i, \quad i = 1, \ldots, n.$$

Let $\{\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2\}$ denote the least-squares estimates of the unknown parameters $\{\gamma_0, \gamma_1, \sigma^2\}$.

(a) Find the expressions for $\hat{\gamma}_0$ and $\hat{\gamma}_1$ in terms of the original predictor values $\{x_i,\, i = 1, \ldots, n.\}$

(b) Find the expression for $\hat{\sigma}^2$ in terms of the original predictor values $\{x_i,\, i = 1, \ldots, n.\}$

(c) Describe how to test the null hypothesis $H_0 :\, \gamma_1 = 0\,.$

Problem 6. (5167) Consider comparing the following two regression models:

$$y_i = \beta_0 + e_i\,, \quad i = 1, \ldots, n, \tag{1}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i\,, \quad i = 1, \ldots, n, \tag{2}$$

where $\{e_i,\, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma^2)$ variables.

(a) Give the expression of R^2 for fitting model (2).

(b) Give the expression of the overall F-test for comparing the two models.

(c) Show that

$$F = \frac{n - p'}{p} \frac{R^2}{1 - R^2}$$

where $p' = p + 1$. Thus the F-statistic is just a transformation of R^2.

Problem 7. (5198) A study was performed to assess the association between a mother's exposure to significant fiber dust during pregnancy and acute lymphoblastic leukemia in the child. 150 children with this leukemia were identified in a specific geographical region, and each was matched with a control child by year of birth, sex and municipality. It was found that the mothers of both the case and control child were exposed to dust for 5 of the pairs; the mother of the case child, but not the control child, was exposed to dust for 10 of the pairs; the mother of the control child, but not the case child, was exposed to dust for 7 of the pairs; and the remaining 128 pairs were such that neither mother had been exposed to significant fiber dust.

(a) Formally test the association between fiber dust and this leukemia. Clearly state the null and alternative hypotheses and your conclusion.

(b) Find the estimated odds ratio for exposure vs nonexposure and give a corresponding 95% confidence interval. Interpret in the context of this study.

(c) Does the matching by child's year of birth, sex and municipality in this study enhance or detract from the strength of conclusions drawn from this study? Explain.

Problem 8. (5198) A case-control study investigated the association between blood levels of vitamin D and colorectal cancer. All subjects diagnosed with colorectal cancer in 2000-2004 in a large county were enrolled. Control subjects were also recruited from the same county and were confirmed free of colorectal cancer (via colonoscopy within the past year). The level of vitamin D was determined by a blood test and dichotomized as low or high.

(a) The estimated odds ratio for colorectal cancer (low vitamin D relative to high vitamin D) was 1.3 with a 95% confidence interval of (1.05, 1.61). Interpret these values in the context of the study.

(b) Subjects were also asked about their smoking habits and categorized as nonsmokers (those who never smoked or had quit smoking more than 5 years ago) or smokers. They found an estimated odds ratio among the smokers of $\widehat{\text{OR}}_S = 0.90$, 95% CI $= (0.6, 1.2)$. The estimated odds ratio among the nonsmokers was $\widehat{\text{OR}}_{NS} = 1.4$, 95% CI $= (1.2, 1.8)$. The Breslow-Day test was used to assess homogeneity of the odds ratios for the smokers and nonsmokers. The test statistic was $X^2_{BD} = 4.3$ on 1 degree of freedom.

State the null and alternative hypotheses for this test and your conclusion.

(c) To adjust for smoking status, the researchers computed the Mantel-Haenszel estimate of the odds ratio as $\widehat{\text{OR}}_{MH} = 0.88$ with 95% CI $= (0.5, 1.5)$. Interpret these values in light of all information in this problem, being clear about any needed assumptions.

(d) Comment on the effect of smoking on the association between vitamin D and colorectal cancer. Is there evidence that smoking is a confounder? Is there evidence that smoking and vitamin D interact?

(e) The vitamin D level was also discretized into 5 categories according to quintiles of the observed distribution. The estimated risks for colorectal cancer within each of the 5 levels of vitamin D were 0.20, 0.14, 0.03, 0.05, 0.08, ordered from the lowest level of vitamin D to the highest level. A χ^2 test for the association between these 5 levels of vitamin D and colorectal cancer resulted in the statistic $X^2 = 10.7$ on 4 df. A test for linear trend among these risks gives the statistic $X^2_{Tr} = 3.2$ on 1 df.

Interpret these results and comment on whether there is evidence of *any* trend in risk with increasing vitamin D levels.

Formulas and Tables for Potential Use in the 5198 Problems

Some known formulae based on the usual 2×2 table

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

$\mathrm{RR} =$ relative risk

$\mathrm{OR} =$ odds ratio

$$\mathrm{AR} = \text{attributable risk} = \frac{R - R_{\overline{E}}}{R}, \text{ where } R \text{ is the overall risk} \tag{1}$$

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

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

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:

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

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

- 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 $V = n_D n_{\bar{D}}(nT_3 - T_2^2)/[n^2(n-1)]$

- 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	9.2	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) Suppose X has density

$$f_X(x) = \frac{1}{\alpha} e^{-x/\alpha}\,, \qquad 0 < x < \infty\,, \quad \alpha > 0\,,$$

and the conditional density of Y given X is

$$f_{Y|X}(y|x) = \frac{\beta x^{\beta}}{y^{\beta+1}}\,, \qquad x < y < \infty, \quad \beta > 2\,.$$

(a) Find EY.

(b) Find $\mathrm{Var}(Y)$.

Problem 10. (5326) Suppose n people play a **modified** version of Russian roulette. Each person has **two** guns: one gun fires with probability α when the trigger is pulled, and the other fires with probability β. (Assume all the guns are independent of each other and successive shots of the same gun are independent.) A round of play consists of everyone who is still alive choosing one of their guns at random (with equal probability) and raising this gun to their temple, and then all firing simultaneously. Play continues until everyone is dead. Answer the following. Simplify your answers as much as possible.

(a) What is the probability that **two** or more people are still alive after k rounds of play?

(b) The **first** person (or persons) to die receives a prize (flowers on the grave). What is the probability this prize goes to only one person?

Problem 11. (5327) Consider the family of bivariate densities $f(\underset{\sim}{x}|\theta)$ with $\underset{\sim}{x} = (x_1, x_2)$ and $\theta = (\theta_1, \theta_2)$ which is defined by

$$f(\underset{\sim}{x}|\theta) = c(\theta)h(\underset{\sim}{x}) \exp(\theta_1 x_1 + \theta_2 x_2)$$

where $h(\underset{\sim}{x}) = \exp\left(-\sqrt{x_1^2 + x_2^2}\right)$, $c(\theta) = \dfrac{1}{2\pi}\left(1 - \theta_1^2 - \theta_2^2\right)^{3/2}$, and $\theta_1^2 + \theta_2^2 < 1$.

(a) State a general formula for the Fisher information matrix $I(\theta)$ and use it to calculate the Fisher information for a single observation $\underset{\sim}{X} = (X_1, X_2)$ from $f(\underset{\sim}{x}|\theta)$ defined above.

(b) Let $\hat{\theta}_n$ denote the maximum likelihood estimator (MLE) for θ based on $\underset{\sim}{X}_1$, $\underset{\sim}{X}_2$, ..., $\underset{\sim}{X}_n$ which are iid from $f(\underset{\sim}{x}|\theta)$ defined above. State a general result for the asymptotic distribution of the MLE for regular families, and use it to obtain an approximate distribution for the MLE $\hat{\theta}_n$ when n is large. (Your answer should be given as explicitly as possible.)

Problem 12. (5327) Suppose we observe $X_1, X_2, \ldots, X_n$ which are iid with the density

$$f(x|\alpha, \beta) = 1 + 6\alpha(2x - 1) + 4\beta(3x^2 - 1) \quad \text{for } 0 \le x \le 1$$

where both α and β are unknown.

Note: This density is well-defined for (α, β) in a parameter space Θ which is somewhat hard to describe, but you do NOT need to know anything about Θ to do the problems below. (Θ is a certain closed, bounded, convex set containing an open ball about the origin $(0, 0)$.)

(a) Find the method of moments (MOM) estimator for (α, β). Your solution should include a general description of the method of moments procedure when there are two parameters.

(b) If the maximum likelihood estimator (MLE) of (α, β) is in the interior of the parameter space (that is, not on the boundary), it will be a solution of a certain set of equations. What are these equations? (Write them out explicitly in this situation.)

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

$\mu((a,b]) = e^{-a} - e^{-b}$ for $0 \leq a < b < \infty$

$\Omega = [0, \infty)$

$\mathcal{F} = \mathcal{B}(\Omega)$, the Borel σ-field on $[0, \infty)$

For $n = 1, 2, ...$, define a sequence of random variables

$$X_n(\omega) = \begin{cases} 0, & \omega \in [0, n), \\ 1, & \omega \in [n, \infty). \end{cases}$$

(a) Find $\sigma(X_1)$, the σ-field generated by X_1.

(b) Find $\sigma(X_1, X_2)$, the σ-field generated by X_1 and X_2.

(c) Prove or disprove: X_1 and X_2 independent.

(d) Let $S_n = \sum_{i=1}^{n} X_i$. Find S such that $S_n \xrightarrow{P} S$. Prove the convergence.

Problem 14. (6346) Let $(\Omega = \mathbb{R}, \mathcal{F} = \mathcal{B}, \mu)$ be the measure space on the reals with the Borel σ-field, Lebesgue measure μ, and $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & x < 0 \\ 1/2^k, & 2k \leq x < 2k+1, \quad k = 0, 1, ... \\ -(1/3^k), & 2k+1 \leq x < 2k+2, \quad k = 0, 1, ... \end{cases}$$

Find $\int f(x) d\mu(x)$ using Lebesgue integral methods. (Note that f is defined for all k simultaneously and is non-zero on $[0, \infty)$.)