Ph.D. Qualifying Exam
Friday–Saturday, August 23–24, 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) 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 2. (5106) Let Y be a discrete random variable with probability mass function:

$$Y \sim \alpha_1 f_1(y; \lambda_1) + \alpha_2 f_2(y; \lambda_2) + \alpha_3 f_3(y; \lambda_3),$$

where f_1, f_2 and f_3 are three Poisson mass functions with means λ_1, λ_2, λ_3, 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

$$\theta = (\alpha_1, \lambda_1, \alpha_2, \lambda_2, \alpha_3, \lambda_3).$$

(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)

Assume that two random samples $\{X_i, i = 1, \cdots, n\}$ with $n > 30$ and $\{Y_j, j = 1, \cdots, m\}$ with $m > 30$ are independently drawn from distributions with means μ_1 and μ_2 and the same variance σ^2. Define $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{j=1}^{m} Y_j/m$. Let

$$s_1^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1) \quad \text{and} \quad s_2^2 = \sum_{j=1}^{m} (Y_j - \bar{Y})^2/(m-1)\,.$$

(a) Approximately, what are the distributions of $\bar{X}$ and $\bar{Y}$? Give your justifications.

(b) Show that both s_1^2 and s_2^2 are unbiased estimates for σ^2.

(c) When the two samples are both from normal distributions, what are the distributions of s_1^2 and s_2^2? Why is $s^2 = \dfrac{(n-1)s_1^2 + (m-1)s_2^2}{m+n-2}$ a better estimate for σ^2 than s_1^2 and s_2^2?

Problem 4. (5166)

Two treatments, A and B, are compared in an experimental design. For each treatment, measurements are recorded successively in time. Suppose that the measurements follow the model:

$$X_i(t) = \mu_i + \epsilon_i(t) - \theta\epsilon_i(t-3),$$

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

(a) For a given i, is the process $\{X_i(t), t = 1, 2, \ldots, \}$ stationary? Give your justifications.

(b) What are the distributions of $\bar{X}_i$ for $i = 1, 2$?

(c) Describe how to estimate σ_1^2 and σ_2^2. What are the properties of your estimates? Describe how to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

Problem 5. (5167) Suppose the regression of Y on (X, Z) has 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 the 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 is the R^2 for model m and $m2$?

(d) If we use the 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 observation $(X_1^*, X_2^*) = (200, 2000)$, what are the predicted values from model m and from $m2$? Which prediction do you think is more accurate, and why?

Problem 7. (5198) Answer the following.

(a) In the usual 2×2 table (see later formulas), derive the approximate variance for the log of the relative risk estimate

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

Justify every step in your derivation.

(b) A study recruited individuals at high risk for coronary heart disease and assessed their use of omega-3 fatty acid supplements. These individuals were then followed for 3 years and all cardiovascular disease events were recorded. Using the data reported below with E denoting supplement use and D denoting occurrence of any cardiovascular events in the 3 years of follow up, provide a point estimate and 95% confidence interval for the relative risk of a cardiovascular event for individuals taking supplements to those not. *Interpret your findings.*

	D	$\overline{D}$
E	122	771
$\overline{E}$	147	2342

Problem 8. (5198) Answer the following.

(a) Suppose two investigators are planning case-control studies, and both determine to randomly select 100 cases and 100 controls from their populations of interest. The first investigator believes that exposure probabilities in the cases and controls are about 0.4 and 0.1, respectively (so an odds ratio of about 6 is expected). The second investigator believes that, in her situation, the exposure probabilities in the cases and controls are roughly 0.2 and 0.04, respectively (so that again an odds ratio of about 6 is expected). Which study has the greater power to detect a significant association between the relevant exposure and disease? *Justify.*

(b) 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	$114,601$	0.77	$(0.71, 0.85)$
Women aged < 60	$31,037$	0.67	$(0.56, 0.81)$
Women aged ≥ 60	$83,564$	0.85	$(0.77, 0.94)$

With age considered a stratification factor,

- the Mantel-Haenszel estimate and 95% confidence interval for the odds ratio are 0.81 (0.74, 0.88)
- the Breslow-Day test statistic is $X^2_{BD} = 4.46$ on 1 degree of freedom
- the Cochran-Mantel-Haenszel test statistic is $X^2_{CMH} = 21.92$ on 1 degree of freedom

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

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

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

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

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

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

- 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:

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

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

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

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

  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 $\{O_{ij}\}$:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n} n \qquad (8)$$

- 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			
$\nu =$ 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 gambler is betting on the successive spins of a colored wheel. When the wheel stops spinning, an attached arrow points to one of the colors **A**mber, **B**urgundy, or **C**hartreuse with probabilities A, B, and C, respectively, where $A + B + C = 1$. The gambler **wins** \$1 when the arrow points to **A**mber, **loses** \$1 when it points to **B**urgundy, and gets **nothing** (neither wins nor loses) when it points to **C**hartreuse. The gambler has an initial fortune of z dollars and a goal of g dollars, i.e., starting with z dollars, he continues playing until he reaches g dollars or goes broke (loses all his money). (Assume $0 < z < g$.)

(a) Let $\psi(z)$ denote the probability of reaching the goal g as a function of the initial fortune z. Use the Law of Total Probability to find an equation that $\psi(z)$ must satisfy. (State your argument in detail.)

(b) Let $R = B/A$. Show that the function

$$\psi(z) = \frac{R^z - 1}{R^g - 1}$$

satisfies the equation found in (a) and also the boundary conditions $\psi(0) = 0$ and $\psi(g) = 1$. (Assume that $A \neq B$.)

(c) **Given that the gambler reached the goal**, what is the probability that he lost the first two spins? (Assume $z \geq 2$.)

Problem 10. (5326) Two gamblers (A and B) play the following game. To start off, each of them puts one coin in "the pot". One of these coins is marked with an X. Then the players alternate taking turns starting with A (that is, A, B, A, B, ...). Each turn consists of the following: a player reaches into the pot and pulls out a coin at random. If it is the marked coin, the player wins all the money in the pot and the game is over. If it is not the marked coin, the player puts it back into the pot, and then adds one more coin to the pot and the game continues. Let Y denote the total length of the game, that is, the total number of draws from the pot.

(a) Find $P(Y = k)$ for all $k \geq 1$.

(b) Find EY.

(c) Suppose the game is modified as follows: If a player selects the **marked** coin, then he tosses the coin. If it comes up heads, then the player wins all the money in the pot and the game is over. If it comes up tails, the player puts the marked coin back into the pot, and then adds one more coin to the pot and the game continues. For this modified game find $P(Y = k)$ for all $k \geq 1$. (Assume the marked coin is a fair coin.)

Problem 11. (5327) Suppose X_1, X_2 are i.i.d. with the density

$$f(x \mid \beta) = \frac{1}{\beta} e^{-x/\beta},\ 0 \le x < \infty,\ \beta > 0.$$

Note that $EX_1 = EX_2 = \beta$.

(a) Find a complete sufficient statistic for β.

(b) Find the uniformly minimum variance unbiased estimator (UMVUE) for β.

(c) Show that $X_1 + X_2$ is independent of $\dfrac{X_1}{X_2}$.

Problem 12. (5327) Consider n independent random variables $X_i \sim \text{Gamma}(\alpha_i, \beta), i = 1, \ldots, n$, where $\alpha_i > 0$ are known and $\beta > 0$ is unknown. Note that, for a random variable $X \sim \text{Gamma}(\alpha, \beta)$, its probability density function is

$$f(x \,|\, \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 \le x < \infty,$$

and $EX = \alpha\beta, \text{Var}(X) = \alpha\beta^2$.

(a) Let $\tilde{\beta} = \frac{1}{n}\sum_{i=1}^n \frac{X_i}{\alpha_i}$. Prove that $\tilde{\beta}$ is unbiased and find its variance.

(b) Find the Cramer-Rao lower bound for the variance of an unbiased estimator of β.

(c) Using your answers to (a) and (b) or in any other way, prove that, for any positive constants $z_1, \ldots, z_m$, we have

$$\frac{1}{\sum_{i=1}^m z_i} \le \frac{1}{m^2} \sum_{i=1}^m \frac{1}{z_i}.$$

(d) Find the maximum likelihood estimator for β.

Problem 13. (6346) Answer the following. Be specific about assumptions and use measure-theoretic notation.

(a) State and prove Markov's inequality.

(b) State and prove Chebyshev's inequality.

Problem 14. (6346) Answer the following.

(a) Let X_i be independent and identically distributed uniform $[0,1]$ random variables and $M_n = \max(X_1, X_2, \ldots, X_n)$. Show $M_n \xrightarrow{P} 1$.

(b) Show $M_n \xrightarrow{D} 1$.

(c) Let Y_i, $i = 1, 2, 3, \ldots$ be independent and identically distributed random variables with

$$P(Y_i = k) = 1/10, \quad k = 0, 1, 2, \ldots, 9.$$

Set

$$W_n = \sum_{i=1}^{n} \frac{Y_i}{10^i}$$

Show W_n converges in distribution to a uniform $[0,1]$ random variable.