Ph.D. Qualifying Exam
Thursday–Friday, January 3–4, 2019

- **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 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be a sequence of i.i.d. paired samples. Conditioned on $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{Z}^+$ follows a Poisson distribution with parameter $\lambda_{X_i} = e^{\beta^T X_i}$, where the parameter $\beta \in \mathbb{R}^d$. That is,

$$Y_i \mid X_i, \beta \sim \text{Poisson}(\lambda_{X_i})$$

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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, l(\beta).$$

(b) Prove that the Hessian matrix of the log-likelihood function is semi-negative definite.

(c) Write out the Newton-Raphson algorithm to find $\hat{\beta}$.

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) A pharmaceutical company would like to examine the potency of a liquid medication mixed in large vats. To do this, a random sample of three vats from a month's production was obtained and five separate samples were selected from each vat. The results are listed in the following table:

						Mean
Vat 1	4.0	3.4	4.6	3.2	3.8	3.8
Vat 2	3.0	3.5	2.6	2.4	3.5	3.0
Vat 3	2.8	3.0	2.2	2.6	2.4	2.6

(a) Write a random-effect model for this experiment and state your assumptions. What are the differences between a random-effect model and a fixed-effect model?

(b) Let $SSA = 5 \times \sum_{i=1}^{3} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$ and MSA be the mean square for factor Vat. Calculate the expected value $\mathrm{E}(MSA)$.

(c) State your null and alternative hypotheses. Given $SSW = \sum_{i=1}^{3} \sum_{j=1}^{5} (y_{ij} - \bar{y}_{i\cdot})^2 = 2.62$, conduct an analysis of variance based on the sample data using $\alpha = 0.05$. Estimate the variances in your model.

Problem 4. (5166) Let X denote the repair time in days required for a certain component in an airplane. We wish to test whether X has a Poisson distribution with mean μ. The repair times for 60 components were recorded, with the results shown in the following table. In some cases the component could be repaired immediately on-site, which is recorded as zero days.

Repair Time (Days)	0	1	2	3	4	5	≥ 6
Number of Components	2	5	12	15	16	7	3

(a) The maximum likelihood estimate of $\mu = E(X)$ is the average repair time of the 60 components. Calculate this value $\hat{\mu}$. (Treat ≥ 6 as 6.)

(b) Now suppose that $X \sim \text{Poisson}(\hat{\mu})$. Calculate the cell probabilities $P(X = k)$ for $0 \leq k \leq 5$ and $P(X \geq 6)$ as well as the expected cell frequencies under this assumption.

(c) Combine cells so that expected frequencies are not less than 5 in any cell. Make a chi-square test of agreement of the observed with the expected cell frequencies using $\alpha = 0.05$.

Problem 5. (5167) Suppose the regression of Y on (X, Z) has the true mean function $\mathrm{E}(Y \mid X = x, Z = z) = 4 + 3x + 2z$. 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 Z on X? (Hint: $Z = \beta_0 + \beta_1 X + \varepsilon$, $\mathrm{var}(\varepsilon) = \sigma^2$.)

(b) If $\rho_{xz} = 0$, what are the true regression parameters of Y on X?

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

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) What is the R^2 for model m and $m2$?

(c) Can we use the t-test to compare models m and $m2$? If so, what are the null and alternative hypotheses and the p-value? If not, what test should we use (also, give the null and alternative hypotheses and p-value)?

(d) Describe how to evaluate the influence of each observation $(Y_i, X_{1i}, X_{2i}),\ i = 1, \ldots, n$.

(e) Suppose we have $p = 20$ predictors instead of just two. Give at least two procedures for model selection and compare the procedures.

Problem 7. (5198) The Department of Defense recognized after the 1991 Gulf War that there was a need to collect more information about the long-term health of service members. The Millennium Cohort Study was designed to address that critical need, and the study was launched in 2001. Since 2001, a total of 77,040 military service members have been enrolled and followed until 2012.

Based on a recent report, the Millennium Cohort investigators were concerned about a potential link between psychiatric disorders and leptospirosis. The investigators decided to have a 10% random sample of stored bloods that was collected at enrollment pulled and tested for the presence of leptospirosis. Blood samples were linked to outcomes using a computer database. A total of 2500 samples tested positive for leptospirosis, of which 38 were from service members who were subsequently diagnosed with a psychiatric disorder. Among the remaining negative samples, 15 were from service members subsequently diagnosed with a psychiatric disorder.

(a) Write down the exposure and disease, and summarize these data in a 2 × 2 table having the format given below.

	D	$\overline{D}$	Total
E			
$\overline{E}$			
Total			

(b) Using the data in the 2 × 2 table, calculate the cumulative incidence ratio and its 95% confidence interval. Interpret the cumulative incidence ratio.

(c) What is the risk of psychiatric illness among all those whose blood samples were tested that is attributable to infection with leptospirosis?

(d) Assuming the results from this study can be generalized to the 20 million people living in Florida and that 5% have been infected with leptospirosis, what percent of psychiatric illness in the Florida population is attributable to infection with leptospirosis?

Problem 8. (5198) Answer the following questions about systematic error.

(a) The following table shows the agreement between six consultants and initial readers on grading chest X-ray film quality.

Initial	Six Consultant Readers			Total
Readers	Good	Acceptable	Poor	Readings
Good	771	922	220	1913
Acceptable	176	387	111	674
Poor	37	113	59	209
Total	984	1422	390	2796

Calculate the κ statistic. What does the κ statistic tell you about these data?

(b) Consider a case–control study with the observed data in the following 2×2 table:

	Case	Control
E	a	b
$\overline{E}$	c	d
Total	N_1	N_0

Denote π_1 and P_1 as the true and observed exposure probabilities for cases, respectively, and π_0 and P_0 as those for controls. Consequently, we may assume $a \sim Bin(N_1, P_1)$ and $b \sim Bin(N_0, P_0)$. Also, suppose that the exposure is misclassified, while the disease condition is not. Let Se_1 and Sp_1 be the sensitivity and specificity for cases, and Se_0 and Sp_0 be those for controls; they are fixed known values and satisfy $Se_i + Sp_i \geq 1$ $(i = 0, 1)$. Show that

$$P_i = \pi_i Se_i + (1 - \pi_i)(1 - Sp_i) \text{ for } i = 0, 1.$$

(c) Based on (b), the true (misclassification-bias-corrected) odds ratio (OR) is OR $= [\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$. Show that it can be estimated as

$$\widehat{\text{OR}} = \frac{(N_1 Sp_1 - c)(N_0 Se_0 - b)}{(N_0 Sp_0 - d)(N_1 Se_1 - a)}.$$

(Hint: express π_i in terms of P_i and the sensitivity and specificity, and use an appropriate estimate for P_i.)

(d) Based on (b) and (c), show that

$$\widehat{\text{Var}}[\log(\widehat{\text{OR}})] \approx \frac{N_1 ac(Se_1 + Sp_1 - 1)^2}{(N_1 Se_1 - a)^2 (N_1 Sp_1 - c)^2} + \frac{N_0 bd(Se_0 + Sp_0 - 1)^2}{(N_0 Se_0 - b)^2 (N_0 Sp_0 - d)^2}.$$

(Hint: use the delta method.)

Begin your solution to each problem on a new sheet of paper.

Problem 9. (5326) Suppose X and Y are i.i.d. Geometric(p) with $0 < p < 1$. Answer the parts given below.

Here are some possibly useful facts you may use without proof:

- Geometric(p) pmf: $f(x) = p(1-p)^{x-1}$, $x = 1, 2, \ldots$
- Geometric(p) mgf: $M(t) = \dfrac{pe^t}{1-(1-p)e^t}$ for $t < -\log(1-p)$
- Infinite series: $\displaystyle\sum_{k=1}^{\infty} \frac{u^k}{k} = -\log(1-u)$ for $|u| < 1$.

(a) Find $P\left(\dfrac{X}{X+1} < 0.77\right)$.

(b) Find $\operatorname{Cov}\left(\dfrac{X}{Y}, Y\right)$.

(c) Find $P\left(\dfrac{X}{Y} = \dfrac{1}{3}\right)$.

Problem 10. (5326) There are **four** snipers firing at the enemy from hidden positions. The enemy is constantly searching for the snipers' locations, and when a sniper's location is discovered, she is killed in a hail of bullets. Assume that the snipers begin their work at the same time and that the snipers' "lifetimes" (the time until they are discovered and killed) are i.i.d. exponential random variables with a mean of β hours.

(a) Find the mean and variance of the time until the last sniper dies.

(b) The snipers are killed off one by one. Assume that, while k snipers remain alive, they kill the enemy soldiers at an average rate of ck^2 deaths per hour. (c is an arbitrary positive value.) What is the expected value of the total number of enemy soldiers killed by the four snipers?

Problem 11. (5327) Consider n independent random variables $X_i \sim N(\mu, 1), i = 1, \ldots, n$, where μ is the unknown parameter. We do not know the exact values of X_i, but only observe the events $X_i > 0$ or $X_i \leq 0$. In answering the following questions, you can use $\Phi(x)$ and $\phi(x)$ to denote the cumulative distribution function and probability density function, respectively, for a $N(0, 1)$ random variable, and Φ^{-1} to denote the inverse of Φ.

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

(b) Find an estimate for the variance of the MLE for μ.

Problem 12. (5327) Let $X_1, \ldots, X_n$ be i.i.d. observations from Uniform$(1, a+1)$, where $a > 0$ is the unknown parameter. Answer the following questions.

(a) Find a complete sufficient statistic for a.

(b) Find the best unbiased estimate for a (i.e, the uniformly minimum variance unbiased estimator).

(c) Find $E\left(X_{(n)} \,\middle|\, \frac{\bar{X} - X_{(1)}}{X_{(n)} - \bar{X}}\right)$, where $X_{(1)} = \min\limits_{1 \le i \le n} X_i$, $X_{(n)} = \max\limits_{1 \le i \le n} X_i$ and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Problem 13. (6346) Suppose X_i are i.i.d. with mean μ and finite variance σ^2. The sample variance is given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 .$$

(a) Give the definition of convergence in probability.

(b) Give the definition of a consistent estimator.

(c) Show S^2 is consistent, OR, find conditions on S^2 that make it a consistent estimator for σ^2.

Problem 14. (6346) Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, $X : \Omega \to \mathcal{O}$, $g : \mathcal{O} \to \mathbb{R}$, and μ_X be the measure induced by X from μ. Prove

$$E(g(X)) = \int g(u) d\mu_X(u).$$