Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part II
August 27, 2015
1:00-4:00pm, Room 133 SMI

- There are a total of FOUR (4) problems in this exam. Please do a total of TWO (2) problems.
- Each problem must be done in a separate exam book.
- Please turn in TWO (2) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. Throughout, X, Y denote random vectors and t, s denote deterministic vectors with appropriate dimensions, and t' denotes the transpose of t. Let $F(x,y)$ be the joint cumulative distribution function (cdf) of (X,Y), $F_X(x)$ be the cdf of X, and $F_Y(y)$ be the cdf of Y. For almost all y, $F(x|y) = P(X \leq x|Y = y)$ is a cdf, where $X \leq x$ means all components of $x - X$ are nonnegative.

   Definition A. X and Y are said to be independent if and only if $F(x,y) = F_X(x)F_Y(y)$ for all possible x and y.

   (a) Show that, for any Borel function $g(x)$ with $E|g(X)| < \infty$,

   $$E[g(X)|Y = y] = \int g(x)dF(x|y) \quad \text{almost all } y$$

   (b) Show that X and Y are independent according to Definition A if and only if $F(x|y) = F_X(x)$ for almost all y and all x.

   (c) Let $\phi_{X|Y}(t|y) = E(e^{it'X}|Y = y)$ a.s., $i = \sqrt{-1}$. Show that X and Y are independent according to Definition A if and only if $\phi_{X|Y}(t|y)$ does not depend on y a.s.

   (d) Suppose that the probability measure of X is absolutely continuous with respect to a σ-finite measure λ. Show that, for almost all y, there is a nonnegative Borel function $f(x|y)$ such that

   $$F(x|y) = \int_{u \leq x} f(u|y)d\lambda$$

   (e) Assume the condition in part (d). Show that, for any Borel B in the range of (X,Y),

   $$P((X,Y) \in B) = \int_{(x,y)\in B} f(x|y)d\lambda dF_Y(y)$$

2. Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables. This question considers different conditions for almost sure convergence.

   (a) Suppose that $\{X_n\}_{n\geq 1}$ is a monotone sequence with $X_n \xrightarrow{P} X$. Does X_n converge almost surely? Prove or Disprove.

   (b) Prove that $X_n \xrightarrow{as} X$ if and only if

$$\sup_{k\geq n} |X_k - X| \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty.$$

   (c) Prove that $\{X_n\}_{n\geq 1}$ converges almost surely if and only if

$$\text{for all } \epsilon > 0, \quad P\left(\sup_{k\geq 0} |X_{n+k} - X_n| \geq \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty.$$

3. A hospital is interested in studying the variability in serum cholesterol measurements over runs of its spectrophotometer. The study team selects five patients for this study and prepares eight blood sample tubes from each patient. These samples are then analyzed in four spectrophotometer runs: each run includes two samples from each patient for a total of 10 samples per run. Few lines of the data file are as follows:

```
patient run y
1  1  167.3
1  1  166.7
1  2  179.6
1  2  175.3
1  3  169.4
1  3  165.9
1  4  177.7
1  4  177.1
2  1  186.7
2  1  184.2
...
5  4  156.1
5  4  151.0
```

Consider analyzing these data by treating both the patient and run as random factors. Let Y_{ijk} denote the measurement for patient i, run j, sample k.

(a) Write an appropriate linear model for Y_{ijk}, and state the main assumptions.

Consider the following R output for the rest of the questions.

```
chol <- read.table("cholesterol.txt", header=T)
y <- chol$y
patient <- factor(chol$patient)
contrasts(patient) <- contr.sum
run <- factor(chol$run)
contrasts(run) <- contr.sum

result <- lm( y ~ patient*run )

> summary(result)

Call:
lm(formula = y ~ patient * run)

Residuals:
Min      1Q Median      3Q     Max
-4.250 -0.925  0.000  0.925  4.250
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   167.4350     0.4065 411.931  < 2e-16 ***
...
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.571 on 20 degrees of freedom
Multiple R-squared:  0.9975,Adjusted R-squared:  0.9952
F-statistic: 428.6 on 19 and 20 DF,  p-value: < 2.2e-16

tmp <- anova(result)
MS.A <- tmp$"Mean Sq"[1]
MS.B <- tmp$"Mean Sq"[2]
MS.AB <- tmp$"Mean Sq"[3]
MS.Err <- tmp$"Mean Sq"[4]

> c(MS.A, MS.B, MS.AB, MS.Err)
[1] 13152.025875   364.457000     9.123875     6.608500
```

(b) Construct an ANOVA table for your model using the above output.

(c) Do the effects of patient and run appear to be additive? Justify your answer.

(d) Estimate the variance components in your model. Which variance components are significantly different from zero?

(e) The company that manufactures this equipment claims that results from this spectrophotometer are highly consistent from one run to the next, and that measurements from different runs will be no more variable than measurements from the same run. Using your model from part (a), explain under what conditions on the variance components this claim would hold.

(f) Does the data support the claim in part (e)? Justify your answer.

4. Consider a vector $\mathbf{Y}$ of 6 independent random normal variables with known scalar variance σ^2. Suppose that $E(Y)$ is a linear function of: x_1, a three-level factor with values *red*, *yellow*, *blue*, *red*, *yellow*, *blue*, respectively; x_2, a binary predictor with values I, I, I, II, II, II, respectively; and x_3, a continuous variable with values 1, 2, 3, 4, 5, 6, respectively.

   (a) Consider the model omitting x_3 (equivalently, set its coefficient equal to zero). Generate and show an orthogonal matrix $\mathbf{X}$ for use in the model

   $$E(\mathbf{Y}) = \mathbf{X}\beta,$$

   including a constant term, whose coefficient can be referred to if needed as β_0. $\mathbf{X}$ need not be orthonormal.

   (b) For the model in Part (a), give the variance of the following: (the difference in estimated expectations between red and blue observations, other predictors being equal) minus (half the difference in estimated expectations between observations with values "II" and "I", other predictors being equal).

   (c) For the model in Part (a), give the approximate variance of the following: the product of (the difference in estimated expectations between red and blue observations, other predictors being equal) times (half the difference in estimated expectations between observations with values "II" and "I", other predictors being equal), assuming both elements of the product are positive.

   (d) Suppose we now include x_3 as well as an interaction term between x_2 and x_3; denote the interaction term's coefficient as β_{Int}. Give a practical interpretation of $\hat{\beta}_{\text{Int}}$, making sure to define all notation you use.

   (e) For this section assume that we have the predictors and $\mathbf{X}$ matrix of Part (a) but instead of Y being normal let it be binary with values 0 and 1. Also assume the generalized linear model

   $$\eta[E(\mathbf{Y})] = \mathbf{X}\beta$$

   for a continuous monotonic link function $\eta(.)$. Under what circumstances would the estimated covariance matrix of $\hat{\beta}$ be diagonal? Could changing $\eta(.)$ result in different estimates of $E(\mathbf{Y})$? Why or why not?