Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part II
August 28, 2014
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. Let X, Y, and Z be random variables on a probability space. For any random vector W, let $\sigma(W)$ be the σ-field generated by W and $P(A|W)$ be the conditional probability of event A given W. Suppose that

$$P(B|X,Z) = P(B|Z) \quad \text{a.s. for any } B \in \sigma(Y).$$

(a) Show that, for any $A \in \sigma(X)$ and $B \in \sigma(Y)$,

$$P(A \cap B|Z) = P(A|Z)P(B|Z) \quad \text{a.s.}$$

(b) Show that

$$P(A|Y,Z) = P(A|Z) \quad \text{a.s. for any } A \in \sigma(X).$$

(c) Suppose that (X, Y, Z) has a joint probability density $f(x, y, z)$ with respect to the Lebesgue measure. Show that, for any integrable $g(X)$,

$$E[g(X)|Y,Z] = \frac{\int\int g(x)f(x,y,Z)dxdy}{\int\int f(x,y,Z)dxdy} \quad \text{a.s.}$$

You may use (without a proof) the following properties of conditional expectations: $E[E(X|\mathcal{A})|\mathcal{A}_0] = E(X|\mathcal{A}_0) = E[E(X|\mathcal{A}_0)|\mathcal{A}]$ a.s., where $\mathcal{A}_0$ is a sub-σ-field of $\mathcal{A}$. If $\sigma(Y) \subset \mathcal{A}$ and $E|XY| < \infty$, then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$ a.s.

2. Let $\{T_1, \ldots, T_n\}$ be independent and identically distributed exponential random variables with rate λ. Define the sequence

$$M_n = \max_{1 \leq i \leq n} T_i.$$

(a) Find non-random numbers a_n so that

$$\limsup_{n \to \infty} \frac{T_n}{a_n} = 1 \quad \text{almost surely.}$$

Justify your answer.

(b) Find non-random numbers a_n so that

$$\frac{M_n}{a_n} \xrightarrow{a.s.} 1.$$

Justify your answer.

(c) Find non-random numbers a_n, b_n and a non-degenerate random variable M_∞ such that $b_n^{-1}(M_n - a_n)$ converges in distribution to M_∞. Justify your answer.

3. The following data are a portion of the yields collected during an interlaboratory study involving three laboratories. Each laboratory was sent two samples from four different materials. The four materials were labeled A, B, C, and D.

Lab	Material	Yields	
1	A	12.20	12.16
1	B	15.51	15.29
1	C	18.14	18.21
1	D	18.54	18.45
2	A	12.59	12.67
2	B	14.98	15.22
2	C	18.54	18.60
2	D	19.21	18.69
3	A	12.72	12.66
3	B	15.33	15.24
3	C	18.00	17.93
3	D	18.88	18.03

Answer these questions assuming that there is no laboratory by material interaction.

(a) From the design, is the interest in this experiment on the comparison among laboratories or on the comparison among materials?

(b) Write the model equation and give assumptions needed for the analysis. Be sure to define all notations.

(c) Should the effects of laboratories be considered fixed or random?

(d) Construct the appropriate ANOVA table and perform the necessary significance tests. What do you conclude?

(e) The yields were actually the protein contents of the materials sent to the laboratories. The materials were different samples of grain of two species using two cultivation methods. The A samples were from three fields of winter wheat, all from fields that had been producing wheat for at least the two previous years. The B samples were from three fields of winter wheat, all from fields that were planted with soybeans in the previous season. The C samples were from fields of triticale (a hybrid of wheat and rye) that had been planted with wheat for at least the two previous years. The D samples were from fields of triticale that were planted with soybeans in the previous season.

   (i) What set of orthogonal contrasts would be sensible for comparisons among the four materials?

   (ii) Estimate the contrasts and their standard errors.

   (iii) Modify your ANOVA table to include tests of the contrasts and summarize the results of your analysis.

4. A study on anorexia consists of weight data for 72 young women. The women were divided into three groups: *control group with standard treatment* (Cont), *cognitive behaviour treatment group* (CBT), where the participants met with a therapist, and *family therapy group* (FT), where anorexia was interrupted by women's parents when they observed it. Figure 1 provides two exploratory plots of the data. The numbers of women in each of the three categories were 29, 26, and 17, respectively.

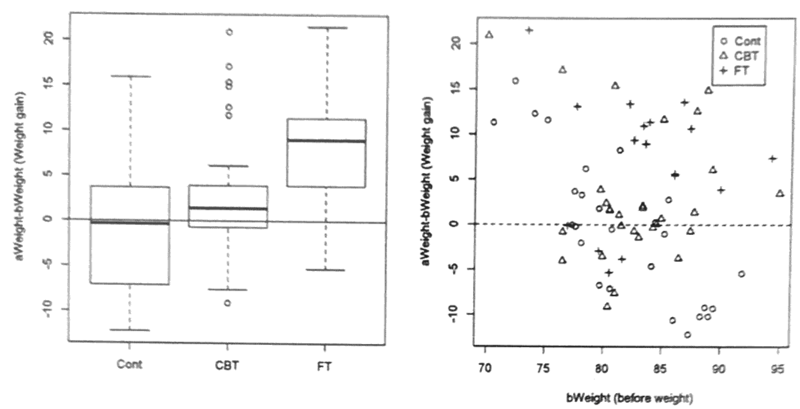

Figure 1: Anorexia Study.

Some data summaries on the before (`bWeight`), after weights (`aWeight`), and weight gains (`aWeight-bWeight`) are as follows.

Group	Mean before weight (lbs)	Mean after weight (lbs)	Mean weight gain (lbs)
Cont	81.56	81.11	-0.45
CBT	82.69	85.70	3.01
FT	83.23	90.49	7.26

Let $y_{i,j}$ denote the weight gain (`aWeight-bWeight`) for the j-th women in i-th treatment group. The total and error sum of squares are given by

$$\begin{aligned} SS_T &= \sum_{i \in \{Cont, CBT, FT\}} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{..})^2 \\ &= 4525.39 \\ SS_E &= \sum_{i \in \{Cont, CBT, FT\}} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i.})^2 \\ &= 3910.74, \end{aligned}$$

where $\bar{y}_{..}$ denotes the grand mean across all women, $\bar{y}_{i.}$ denotes the mean of the i-th treatment group, and n_i is the number of observations in the i-th treatment group.

Consider the following model

$$y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}, \quad i \in \{\text{Cont, CBT, FT}\} \quad j = 1, \cdots, n_i, \tag{1}$$

where $\epsilon_{i,j}$ are independent random variables from $\mathcal{N}(0, \sigma_\epsilon^2)$.

(a) Find the least squares estimators of μ, α_{Cont}, α_{CBT}, and α_{FT}.

(b) Let $\beta = (\mu, \alpha_{Cont}, \alpha_{CBT}, \alpha_{FT})$ and denote its least squares estimator from part (a) with $\hat{\beta}$. Estimate $\text{cov}(\hat{\beta})$.

(c) Is there a significant difference in weight gain between the three treatments based on model (1)? If so, identify which pairs are significantly different at level 0.05.

(d) Consider modeling after weight (`aWeight`) as a function of treatment (`trt`) and before weight (`bWeight`). The following table provides residual sum of squares (RSS) for six models that were fitted to the data. All the models were of the type `aWeight ~ ...`, where the right hand side is specified in the second column of the following table. Use appropriate tests to determine which of these models is most appropriate for these data.

	Model	RSS
2	bWeight	4077.54
3	trt	3665.06
4	trt + bWeight	3311.26
5	trt:bWeight	3245.32
6	trt*bWeight	2844.78

(e) Is it possible to compare all pairs of models in the above table using an F-test. If not, list the pairs for which F-test is not possible.

For parts (f), (g), and (h), consider the following model:

```
Call:
lm(formula = aWeight ~ trt * bWeight, data = anorexia.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8125  -3.8501  -0.9153   4.0010  15.9640

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      92.0515    18.8085   4.894 6.67e-06 ***
trtCBT          -76.4742    28.3470  -2.698  0.00885 **
trtFT           -77.2317    33.1328  -2.331  0.02282 *
bWeight          -0.1342     0.2301  -0.583  0.56173
trtCBT:bWeight    0.9822     0.3442   2.853  0.00578 **
trtFT:bWeight     1.0434     0.4000   2.609  0.01123 *
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.565 on 66 degrees of freedom
Multiple R-squared:  0.3794,Adjusted R-squared:  0.3324
F-statistic:  8.07 on 5 and 66 DF,  p-value: 5.5e-06

> vcov(m6)
                 (Intercept)      trtCBT       trtFT    bWeight trtCBT:bWeight trtFT:bWeight
(Intercept)       353.761280 -353.761280 -353.761280 -4.3172320      4.3172320     4.3172320
trtCBT           -353.761280  803.552512  353.761280  4.3172320     -9.7387679    -4.3172320
trtFT            -353.761280  353.761280 1097.784647  4.3172320     -4.3172320   -13.2261969
bWeight            -4.317232    4.317232    4.317232  0.0529347     -0.0529347    -0.0529347
trtCBT:bWeight      4.317232   -9.738768   -4.317232 -0.0529347      0.1184996     0.0529347
trtFT:bWeight       4.317232   -4.317232  -13.226197 -0.0529347      0.0529347     0.1599758
```

(f) Find a 95% confidence interval for the slope of the line for the standard treatment (Cont) and interpret this confidence interval in terms of the relationship between the before and after weights of women assigned to the standard treatment.

(g) Consider a patient with a before weight of 100 lbs. Find a 95% confidence interval for the difference in her expected after weights on treatments FT and Cont.

(h) Consider the following one-way ANOVA analysis with the partially filled ANOVA table:

```
> anova(lm(bWeight~trt, data=anorexia.dat))
Analysis of Variance Table

Response: bWeight
           Df  Sum Sq Mean Sq F value Pr(>F)
trt           32.57
Residuals     1874.35
```

What could be the purpose of this particular one-way ANOVA for evaluating the design of this study? What do you conclude?

(i) The current design uses "before weights" as a covariate. Describe an alternative design that uses "before weights" in a different way. What form of analysis would be appropriate for your alternative design?