Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part II
September 3, 2009
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.

4. **Definition.** If $\hat{\theta}$ is the MLE of θ in a parametric model, the Wald type test statistic for testing $H_o : \theta = 0$ vs. $H_a : \theta \neq 0$ is $W = \hat{\theta}/\hat{\sigma}$, where $\hat{\sigma}^2$ is the MLE of $\sigma^2 = \text{Var}_{H_o}(\hat{\theta})$ and H_o is rejected for large $|W|$.

   Suppose that $X_1, \cdots, X_n$ are i.i.d. as $X \sim F$ and $Y_1, \cdots, Y_n$ are i.i.d. as $Y \sim G$, and assume that the X's and Y's are independent.

   (a) Suppose that $f(x) = F'(x) = \lambda_1^{-1} \exp(-x/\lambda_1)$ and $g(u) = G'(y) = \lambda_2^{-1} \exp(-y/\lambda_2)$. Let $\theta = \lambda_2 - \lambda_1$.

      i. Find the MLE $\hat{\theta}$ of θ.

      ii. Find the Wald type test statistic W and use it to construct a test with asymptotic level of significance α for testing $H_o : \theta = 0$ vs. $H_1 : \theta \neq 0$.

      iii. Find the likelihood ratio test for testing $H_o : \theta = 0$ vs. $H_1 : \theta \neq 0$. What critical value will give asymptotic level α?

      iv. Use the test statistic $T = \bar{Y}/\bar{X}$ to construct a test that has exactly level α.

   (b) Suppose that F and G are arbitrary with $E(X) = \mu_X$, $0 < \text{Var}(X) = \sigma_X^2 < \infty$ and $E(Y) = \mu_Y$, $0 < \text{Var}(Y) = \sigma_Y^2 < \infty$.

      i. Find the asymptotic level of the Wald type test in (a)(ii) for testing $H_o : F = G$ vs. $H_1 : F \neq G$ in terms of μ_X, μ_Y, σ_X^2, and σ_Y^2.

      ii. Modify the Wald type test in (a)(ii) so that it has asymptotic level α.

1. Let $\mathcal{L}(A|B)$ denote the conditional distribution of A given B.

Definition. A sequence of random vectors $Z_1, ..., Z_n$ is a Markov chain if

$$\mathcal{L}(Z_t|Z_1, ..., Z_{t-1}) = \mathcal{L}(Z_t|Z_{t-1}), \quad t = 1, ..., n.$$

Let $(x_1, y_1), ..., (x_n, y_n)$ be a sequence of bivariate random variables.

(a) Assume that

$$\mathcal{L}(x_t|y_1, ..., y_n, x_1, ..., x_{t-1}, x_{t+1}, ..., x_n) = \mathcal{L}(x_t|y_{t-1}), \quad t = 2, ..., n. \tag{1}$$

Show that, for $t = 2, ..., n$,

i. $\mathcal{L}(y_t|y_r, x_r, x_{r+1}, x_{r+2}, ..., x_t) = \mathcal{L}(y_t|y_r, x_r, x_{r+2}, ..., x_t), \quad r \leq t-1;$

ii. $\mathcal{L}(y_t|y_1, ..., y_{t-1}, x_1, ..., x_{t-1}) = \mathcal{L}(y_t|y_1, ..., y_{t-1}, x_1, ..., x_{t-2});$

iii. $\mathcal{L}(y_t|y_1, ..., y_{t-1}, x_1, ..., x_{t-1}) = \mathcal{L}(y_t|y_1, ..., y_{t-1}).$

(b) In addition to (1), assume further that

$$\{y_1, ..., y_n\} \quad \text{is a Markov chain,} \tag{2}$$

and show that

i. $\mathcal{L}(y_t|y_1, ..., y_{t-1}, x_1, ..., x_{t-1}) = \mathcal{L}(y_t|y_{t-1}), \quad t = 2, ..., n;$

ii. $\mathcal{L}(y_t|y_{t-1}, x_{t-1}) = \mathcal{L}(y_t|y_{t-1}), \quad t = 2, ..., n;$

iii. $\{(x_1, y_1), ..., (x_n, y_n)\}$ is a Markov chain.

(c) Under the above assumptions (1) and (2), show that

$$\mathcal{L}(y_t|y_{r+1}, x_{r+1}, ..., x_t) = \mathcal{L}(y_t|y_r, y_{r+1}, x_r, ..., x_t),$$

for any $r < t-1$ and $t = 2, ..., n$.

2. A Dirichlet process can be characterized in multiple ways. We will use here the following definition. The sequence $(X_1, X_2, \dots, X_n, \dots)$ is a Dirichlet process with base distribution G_0 and concentration parameter $\alpha_0 > 0$ if G_0 is a probability distribution on R and if:

   - $X_1 \sim G_0$,
   - Conditional on $X_1, X_2, \dots, X_n$, the distribution of X_{n+1} is $\alpha_0 G_0 + \sum_{i=1}^{n} \delta_{X_i}$, appropriately normalized. Recall that δ_x denotes the Dirac measure with probability 1 on singleton $\{x\}$.

   In this problem, let $(X_n)_{n\geq 1}$ be a Dirichlet process with base distribution G_0 and concentration parameter α_0. Assume that G_0 has a first moment μ. In particular, X_1 is integrable and $E(X_1) = \mu$.

   (a) Determine the conditional expectation $E(X_2|X_1)$.

   (b) Determine the distribution of X_2.

   (c) Prove that $(X_2, X_1, X_3, X_4, X_5, \dots)$ forms a Dirichlet process.

   (d) Prove that $(X_n)_{n\geq 3}$ are identically distributed random variables.

   (e) Prove that, conditional on $X_1, X_2, \dots, X_k$, the process $(X_{k+n})_{n\geq 1}$ is a Dirichlet process. Determine its concentration parameter and its base distribution.

   (f) Let $Y_i = 1_{X_i > 0}$. In other words, $Y_i = 1$ if $X_i > 0$ and $Y_i = 0$ if $X_i \leq 0$. Prove that $(Y_n)_{n\geq 1}$ forms a Dirichlet process. Determine its concentration parameter and its base distribution.

   (g) Let Θ be a random variable on $(0, 1)$ with the Beta(α, β) distribution, i.e. with density

   $$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ on } \theta \in (0,1).$$

   Recall that this beta distribution has mean $E(\Theta) = \alpha/(\alpha+\beta)$. Now let X_1, X_2, X_3 be i.i.d. Bernoulli $\mathcal{B}(1, \Theta)$ random variables, conditional on Θ. Prove that (X_1, X_2, X_3) form the beginning of a Dirichlet process. Determine its base distribution and concentration parameter.

3. At a shipping office, three parcels are weighed singly, in pairs, and all together. All the weighings are independent. Denote the weights by

$$\mathbf{y} = (y_{100},\ y_{010},\ y_{001},\ y_{110},\ y_{101},\ y_{011},\ y_{111})'$$

where the suffix 1 indicates the presence of a particular parcel and the suffix 0 indicates its absence. The observations are

$$\mathbf{y} = (50,\ 24,\ 32,\ 69,\ 80,\ 52,\ 100)'.$$

(a) Suppose each weighing has the same variance, i.e.,

$$\text{var}(y_{100}) = \text{var}(y_{010}) = \ldots = \text{var}(y_{111}) = \sigma^2.$$

i. Obtain the best linear unbiased estimators (BLUEs) of the weights of the parcels.

ii. Obtain the variance-covariance matrix of the BLUEs.

(b) Now suppose that the variance of a weighing increases with the number of parcels weighed, such that

$$\begin{aligned} \text{var}(y_{100}) = \text{var}(y_{010}) = \text{var}(y_{010}) &= \sigma^2 \\ \text{var}(y_{110}) = \text{var}(y_{101}) = \text{var}(y_{011}) &= 2\sigma^2 \\ \text{var}(y_{111}) &= 3\sigma^2. \end{aligned}$$

i. Give an alternative design (with seven or fewer weighings) that yields more accurate estimates of the parcel weights.

ii. Show that your design is better.

4. A chemist considered an experiment on the absorption and accumulation of salts by potato cells. He collected data on the rate of uptake (named as absorption) of rubidium (Rb) and bromide (Br) for various numbers of hours. The uptake was measured in the number of microgramme equivalents per 1000 g of water in the potato tissue. All the measurements were done independent of each other.

Duration (Time of immersion (hours))	Absorption (mg. equivalents per 1000 g of water in tissue)	
	Rb	Br
21.7	7.2	6.7
46.0	11.4	12.4
67.0	14.2	15.9
90.2	19.1	18.8
95.5	20.0	21.8

The chemist wants to model absorption as a function of time with a linear model for both of the elements. He wishes to determine (i) whether, at given durations, absorptions of Rb and Br tend to be the same; (ii) if they are not, he wishes to explore whether at least the amount of absorption out of an additional hour are the same for the two elements.

(a) Write out a single multiple linear regression model that would allow you to answer both of the chemist's questions. Explicitly state your assumptions and specify the design matrix.

(b) Specify the hypothesis tests that would answer questions (i) and (ii). State null and alternative hypotheses, corresponding test statistics, along with their distributions.

(c) Use the following computer output to answer the chemist's questions (i) and (ii). In this output, variable `Ions` denote the type of the element: "R" for rubidium and "B" for "bromide".

```
> summary(mR)

Call:
lm(formula = Absorption[Ions == "R"] ~ Duration[Ions == "R"])

Residuals:
      1       2       3       4       5
 0.1511  0.1476 -0.6851  0.2016  0.1848

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.295104   0.501284   6.573 0.007163
Duration[Ions == "R"] 0.172985   0.007186  24.073 0.000157
---
```

```
> summary(mB)

Call:
lm(formula = Absorption[Ions == "B"] ~ Duration[Ions == "B"])

Residuals:
     1       2       3       4       5
-0.4556  0.6777  0.2313 -1.2287  0.7753

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.07762    1.09921    2.80  0.06786
Duration[Ions == "B"]   0.18793    0.01576   11.93  0.00127
---

> summary(m1)

Call:
lm(formula = Absorption ~ Duration + Ions)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0335 -0.5427  0.1298  0.4215  1.0101

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5564     0.6376   5.578 0.000835
Duration      0.1805     0.0085  21.231 1.29e-07
IonsR        -0.7400     0.4687  -1.579 0.158376
---

> summary(m2)

Call:
lm(formula = Absorption ~ Duration * Ions)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2287 -0.3048  0.1680  0.2238  0.7753

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.07762    0.85427   3.603   0.0113
Duration       0.18793    0.01225  15.346 4.84e-06
IonsR          0.21749    1.20811   0.180   0.8631
Duration:IonsR -0.01494    0.01732  -0.863   0.4214
---
```

```
> summary(m0)

Call:
lm(formula = Absorption ~ Duration)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0769 -0.4155 -0.2254  0.4917  1.3801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.186361   0.645883   4.933  0.00114
Duration    0.180456   0.009259  19.491 4.99e-08
---
```

(d) Generalize the above testing strategy for question (i) to K elements, i.e., test whether K different elements have the same linear relationship between absorption and duration. Specify the general form of the linear regression model, hypothesis test, and the test statistic.

(e) The chemist ultimately decides to use the following model (`m0` in the above `R` output)

```
Absorption ~ Duration
```

and considers adding a new variable denoted by `X` into his model. The output from fits with and without an interaction term between `Duration` and `X` are as follows.

```
> summary(m3)

Call:
lm(formula = Absorption ~ Duration + X)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09608 -0.48663 -0.05363  0.30134  1.52821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1613     4.0955  -0.039    0.970
Duration      0.5397     0.4339   1.244    0.254
X             3.6228     4.3744   0.828    0.435

Residual standard error: 0.8236 on 7 degrees of freedom
Multiple R-Squared: 0.9812,     Adjusted R-squared: 0.9758
F-statistic: 182.8 on 2 and 7 DF,  p-value: 9.084e-07
```

```
> summary(m4)

Call:
lm(formula = Absorption ~ Duration * X)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9522 -0.5058 -0.1090  0.3086  1.4575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.455095   4.444333  -0.102    0.922
Duration     0.609946   0.500353   1.219    0.269
X            4.558157   5.305496   0.859    0.423
Duration:X  -0.002055   0.005528  -0.372    0.723

Residual standard error: 0.8795 on 6 degrees of freedom
Multiple R-Squared: 0.9816,     Adjusted R-squared: 0.9725
F-statistic: 106.9 on 3 and 6 DF,  p-value: 1.345e-05
```

Based on these computer outputs, the chemist concludes that the new variable `X` does not have a significant effect on the absorption level. Do you agree with him? Why, why not? Discuss in detail.

(f) After some careful thinking, the chemist is alarmed by the fact that standard error of the coefficient estimate for `X` is quite large in the above model fit `m3` . Argue algebraically how such a large standard error might arise in this model fit. (*Hint:* You may find it more convenient to work with the standardized version of the variables. Let Y denote absorption, Z duration and consider

$$Y_i^* = \beta_0^* + \beta_1^* Z_i^* + \beta_2^* X_i^* + \epsilon_i^*, \quad i = 1, \cdots, n$$

where n denotes the total number of measurements,

$$\begin{aligned} Y_i^* &= \frac{1}{\sqrt{n-1}} \frac{(Y_i - \bar{Y})}{s_Y}, \\ Z_i^* &= \frac{1}{\sqrt{n-1}} \frac{(Z_i - \bar{Z})}{s_Z}, \\ X_i^* &= \frac{1}{\sqrt{n-1}} \frac{(X_i - \bar{X})}{s_X}, \end{aligned}$$

and s_Y, s_Z, and s_X denote respective standard errors of Y, Z, and X.)

(g) Test the hypothesis that, after controlling for duration time, the variations in the amounts of absorption of Rb and Br are the same.

(h) Test whether a model quadratic in duration fits better than the linear one given in model `m0` above.

(j) Construct a 95% prediction interval for the amount of absorption of Rb after a duration of 20 hours based on model `m0`.