Yazhen

Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part II
Thursday, September 1, 2011
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. Suppose that X_n, ε_n, $n = 0, \pm 1, \pm 2, \cdots$, are random variables with finite variance on a probability space $(\Omega, \mathcal{F}, P)$, and ε_n are independent and identically distributed random variables with mean zero and finite variance. Assume that X_n and ε_n obey

$$X_n = \theta_1 X_{n-1} + \cdots + \theta_p X_{n-p} + \varepsilon_n,$$

where p is a positive integer and $\theta_1, \cdots, \theta_p$ are real numbers.

Let $\mathcal{F}_n = \sigma\{X_k : k \leq n\}$ be a sigma-field generated by $X_n, X_{n-1}, X_{n-2}, \cdots$, and denote by $\mathcal{B}$ the Borel sigma-field on real line.

(a) Show that the statement:
For any $m > n$, and $A \in \mathcal{B}$,

$$P(X_m \in A | \mathcal{F}_n) = P(X_m \in A | X_n) \tag{1}$$

is equivalent to

$$\theta_2 = \cdots = \theta_p = 0. \tag{2}$$

(b) Show that Statement (1) implies the statement:
For any $m > n$, $k < n$, and $A, B \in \mathcal{B}$,

$$P(X_m \in A, X_k \in B | X_n) = P(X_m \in A | X_n)\, P(X_k \in B | X_n). \tag{3}$$

(c) Show that Statement (3) implies the statement:
For any $m > n$, $k < n$, and all bounded measurable functions f and g,

$$E[f(X_m)\, g(X_k) | X_n] = E[f(X_m) | X_n] E[g(X_k) | X_n].$$

(d) Show that Statement (3) is equivalent to the statement:
For any $m > n$, $k < n$, and $A \in \mathcal{B}$,

$$P(X_m \in A | X_k, X_n) = P(X_m \in A | X_n).$$

2. This question asks you to prove a result regarding the almost sure convergence of order statistics generated from independent and identically distributed (iid) standard uniform random variables. A natural way to prove this result is to utilize the connection between uniform order statistics and standard exponential random variables.

Notation and Setup. We start by setting the notation.

(i) The basic random variables. Consider two independent collections of random variables $\{U_i, 1 \leq i \leq n\}$ and $\{W_i, 1 \leq i \leq n+1\}$, where the U_i are iid uniform random variables distributed on the interval $(0, 1)$ and W_i are iid exponential random variables with mean 1. Also, define the partial sum of the W_i's

$$S_k \equiv \sum_{i=1}^{k} W_i, \qquad k = 1, \cdots, n+1.$$

(ii) Order statistics (extended to the edge of the interval $(0, 1)$). For the random variables $\{U_i, 1 \leq i \leq n\}$, let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ be the order statistics and define $U_{(0)} \equiv 0$ and $U_{(n+1)} \equiv 1$, that is,

$$0 \equiv U_{(0)} < U_{(1)} < U_{(2)} < \cdots < U_{(n)} < U_{(n+1)} \equiv 1.$$

(iii) Partial average notation. We use the $\bar{W}_k$ notation in the standard way, $\bar{W}_k \equiv \frac{1}{k}\sum_{i=1}^{k} W_i$, $k = 1, \cdots, n+1$. Additionally, we use $\bar{W}(i, j)$ for the following

$$\bar{W}(i, j) \equiv \frac{1}{j-i}\left(W_{i+1} + W_{i+2} + \cdots + W_j\right), \qquad 0 \leq i < j \leq n+1.$$

(iv) The range of the indices. Let m_n be a positive integer satisfying $m_n < n/2$ and

$$\frac{m_n}{\log n} \to \infty \qquad \text{as } n \to \infty. \tag{4}$$

Now, let

$$\max_{\star} \text{ and } \sum_{\star}$$

denote the max and sum, respectively, taken over all pairs (i, j) satisfying (A) $0 \leq i < j \leq n+1$ and (B) $j - i \geq m_n$.

(v) The random variables of interest. Define the sequence of random variables

$$T_n \equiv \max_{\star} \left| (n+1)\left(\frac{U_{(j)} - U_{(i)}}{j-i}\right) - 1 \right|.$$

Question. Prove the following parts.

(a) Let $\overset{\text{d}}{=}$ denote equality in distribution. Establish the following

$$\left(U_{(i)} - U_{(i-1)}\right)_{1 \leq i \leq n+1} \overset{\text{d}}{=} \left(\frac{W_i}{S_{n+1}}\right)_{1 \leq i \leq n+1}. \tag{5}$$

Now use (5) to prove

$$T_n \overset{\mathrm{d}}{=} \max_{\star} \left| \frac{\bar{W}(i,j)}{\bar{W}_{n+1}} - 1 \right| . \tag{6}$$

(b) Show that for any $\epsilon \in (0, 1/2)$, we have

$$\sum_n \mathrm{P}\left(T_n \geq \epsilon\right) \leq \sum_n \mathrm{P}\left(\left|\bar{W}_{n+1} - 1\right| \geq \epsilon/3\right) + \sum_n \sum_{\star} \mathrm{P}\left(\left|\bar{W}(i,j) - 1\right| \geq \epsilon/3\right) \tag{7}$$

(c) Show that for some constant c (involving ϵ), we have

$$\mathrm{P}\left(\left|\bar{W}_k - 1\right| \geq \epsilon\right) \leq 2\exp\left(-ck\right) \tag{8}$$

(d) Using the above, and further calculations, prove as $n \to \infty$,

$$T_n \xrightarrow{a.s.} 0.$$

3. Let 2^k denote a *full factorial design* at two levels (+1 and -1) consisting of all level combinations of k factors. For example, a 2^8 has 256 runs. For run size economy, fractional factorial designs are often used in practice. Let 2^{k-p} denote a *fractional factorial design* at two levels (+1 and -1) in k factors, which is a 2^{-p} fraction of a 2^k. The fraction is determined by p *defining words*. For illustration, Table 1 gives a 2^{5-2}, where the column for D equals the product of the columns for A and B, denoted by $A \times B$, and the column for E equals the product of the columns for A and C. The defining words for this design are $D = A \times B$ and $E = A \times C$, or equivalently

$$I = A \times B \times D \text{ and } I = A \times C \times E, \tag{9}$$

where I denotes a column of all +1's. Note that $I = I \times I = A \times A = B \times B = C \times C = D \times D = E \times E$. The two words in (9) and their product can be written together as $I =$

$$A \times B \times D = A \times C \times E = B \times C \times D \times E. \tag{10}$$

The shortest wordlength (i.e., the number of letters) among the three elements in (10) is three and the design is said to have resolution III (three). Generally, for a fractional factorial design 2^{k-p}, p defining words together with their $2^p - p - 1$ products form a set with $2^p - 1$ elements and the *resolution* of the design is the shortest wordlength among the $2^p - 1$ elements. For $p = 2$, $2^p - 1 = 3$ and (10) has three elements.

A scientist studies the impact of five factors A, B, C, D and E on a chemical process using a fractional factorial design in Table 2 (with one replicate per level combination).

(a) Find a set of defining words for this design. What's its resolution?

(b) Suppose the scientist has conducted an experiment using the design in Table 2, with $y_1, \ldots, y_8$ denoting the response values of Runs $1, \ldots, 8$, respectively. He believes that only two factors in Table 2, denoted by W_1 and W_2, can have significant effects on the response and considers a two-factor model

$$y_i = \beta_0 + \beta_1 w_{1i} + \beta_2 w_{2i} + \beta_{12} w_{1i} w_{2i} + \epsilon_i, \text{ for } i = 1, \ldots, 8, \tag{11}$$

where the ϵ_i are independent Normal random variables with mean zero and an unknown variance σ^2, and w_{1i} and w_{2i} represent the entries in the ith rows of W_1 and W_2 in Table 2. For a given pair of W_1 and W_2, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{12})$ and let Σ denote the covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. There are 20 different ways to choose W_1 and W_2 to be two distinct factors from Table 2. Compute the average value of Σ (in terms of a 4 by 4 matrix) over these 20 possibilities of W_1 and W_2. Explain your reasoning clearly.

(c) Suppose the scientist realizes that in addition to A, B, C, D and E, another factor F should also be included in this study. Use the design in Table 2 to construct a fractional factorial design 2^{6-2} with resolution IV (four).

(d) In general, is it possible to construct a fractional factorial design 2^{5-2} with resolution IV (four)? Construct one or prove such a design does not exist.

Table 1: A 2^{5-2} with resolution III

A	B	C	D	E
−1	−1	−1	+1	+1
−1	−1	+1	+1	−1
−1	+1	−1	−1	+1
−1	+1	+1	−1	−1
+1	−1	−1	−1	−1
+1	−1	+1	−1	+1
+1	+1	−1	+1	−1
+1	+1	+1	+1	+1

Table 2: A fractional factorial design of eight runs and five columns

Run #	A	B	C	D	E
1	+1	−1	−1	−1	−1
2	−1	−1	−1	+1	+1
3	+1	+1	+1	−1	+1
4	−1	+1	+1	+1	−1
5	+1	+1	−1	+1	−1
6	−1	+1	−1	−1	+1
7	+1	−1	+1	+1	+1
8	−1	−1	+1	−1	−1

4. Hoping to reduce his heating bills, a Madison homeowner spent $700 to add insulation to his attic in October 2010. The house uses natural gas for heating and for hot water. Electricity is used for everything else. Table 3 shows the monthly heating bill for the house since January 2008. The thermostat in the home is kept at 68 degrees from November through April. *HDD* (heating degree day) is a unit of measurement designed to reflect the demand for energy needed to heat a home or business. It is derived from measurements of outside air temperature. A *therm* is a unit of heat energy, approximately the equivalent of burning 100 cubic feet of natural gas. The current cost of natural gas is $1.12 per therm. Table 4 gives some summary statistics and Table 5 gives historical monthly average data for Madison.

   (a) Determine if the insulation led to reduced energy use, after differences in HDD are accounted for. Test the appropriate hypothesis at the 0.05 level.

   (b) Find a 95% confidence interval for the average monthly amount of heat used for hot water.

   (c) Estimate the mean annual total savings (in dollars) after insulation for the six months from November through April. How long will it take to recover the cost of the insulation?

Table 3: **Monthly utility records**

	2008		2009		2010		2011	
Month	HDD	Therms	HDD	Therms	HDD	Therms	HDD	Therms
Jan	1415	166	1669	188	1301	153	1395	139
Feb	1461	167	1171	139	1167	127	1193	116
Mar	1020	122	866	112	802	94	998	103
Apr	514	80	569	74	385	56	591	68
May	317	42	210	33	211	37	281	31
Jun	16	22	63	20	15	9		
Jul	2	19	34	24	0	15		
Aug	8	19	42	21	4	15		
Sep	97	22	94	22	147	17		
Oct	518	40	622	69	393	29		
Nov	813	90	652	70	787	70		
Dec	1526	176	1503	177	1491	144		

Table 4: **Summary statistics for the data in Table 3**

	HDD		Therms	
	Mean	SD	Mean	SD
Before Oct 2010	604.735	561.657	76.265	59.748
After Oct 2010	962.286	438.561	95.857	41.406

Table 5: **Historical monthly averages for Madison**

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Avg max temp	24.8	30.1	41.5	56.7	68.9	78.2	82.4	79.6	71.5	59.9	44.0	29.8
Avg min temp	7.2	11.1	23.0	34.1	44.2	54.2	59.5	56.9	48.2	37.7	26.7	13.5
HDD	1519	1243	1014	588	294	68.0	12.0	38.0	168	499	888	1342