Department of Statistical Science Duke University

FIRST YEAR EXAM - SPRING 2013

Monday, May 6th 2013

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Put the problem number and your assigned code on the top of **each page**.
3. Write only on **one side** of the page (solutions on the reverse side of the page will be ignored).
4. Start each problem on a new page.
5. It is to your advantage to show your work and explain your answers.
   Do not erase anything– just draw a line through work you do not want graded.
6. You have 3 hours to finish the written exam: Questions 1-6 inclusive. Attempt all questions; note that credit is not necessarily equally allocated across questions.
7. This is a closed book exam. No notes are permitted.

1. A random string of digits $X = X_0X_1X_2\ldots$ is created sequentially as follows: $X_0 = 0$ and for $i \geq 1$,

$$X_i = \begin{cases} X_{i-1} & \text{with probability } 0.1 \\ X_{i-1} \oplus 1 & \text{with probability } 0.9, \end{cases}$$

where $\oplus$ denotes "addition modulo 10", i.e., $k \oplus 1 = k + 1$ for $k = 0, 1, \cdots, 8$, but $9 \oplus 1 = 0$.

(a) Let Z_1 denote the first $i \geq 1$ with $X_i = 0$. For example, with

$$X = 01234566788901234567890123456 7\ldots$$

we will count $Z_1 = 12$. Show that $\mathsf{E}(Z_1) = 10$.

(b) Suppose the sequential construction is carried on until '0' appears for the 41st time (i.e., 40th reappearance after the initial '0'). Prove that with probability one the string terminates with a finite length.

(c) Under the stopping rule of part (b), let $X = X_0X_1\ldots X_M$ be the full string, with X_M the 41st '0'. Clearly, M is a random integer and the smallest value it can realize is 40 (when X is a string of 41 '0's). Is $P(M > 440) > 0.5$? Justify.

2. Skyler selects a random sample of zartylblats and tests each for strength by dropping it on the floor to see if it breaks. Denote by θ the probability that a dropped zartylblat will break; and assume independence for the sample.

   In the sample of size $n = 10$, Skyler observed $X = 8$ broken and 2 unbroken zartylblats.

   (a) What is the likelihood function for θ in this problem?

   (b) Skyler announces that her **posterior** distribution for θ has density function

   $$\pi(\theta \mid X = 8) = c\theta^{10}(1-\theta)^6, \qquad 0 < \theta < 1$$

   where $c = \Gamma(18)/\{\Gamma(11)\Gamma(7)\} = 136136$. What was Skyler's *prior* distribution? Give the answer as either a density function (correctly normalized if possible) or give the name and the value(s) of any parameter(s).

   (c) Blake has a different prior distribution— she believes that θ can take only one of two different values, $1/3$ and $2/3$, each with prior probability $1/2$. With the same data, what is Blake's *posterior* distribution for θ?

   (d) Alex wants to find the (frequentist) P-value for a test of the hypothesis $H_0 : \theta = 1/2$ against the one-sided alternative $H_1 : \theta < 1/2$, with the same data.

   Which of the following *is* that P-value?

   a. $\mathsf{P}[X \geq 8 \mid \theta = 1/2]$ b. $\mathsf{P}[|X-5| \geq 3 \mid \theta = 1/2]$ c. $\mathsf{P}[X \leq 8 \mid \theta = 1/2]$
   d. $\mathsf{P}[\theta < 1/2 \mid X = 8]$ e. $\mathsf{P}[\theta \neq 1/2 \mid X = 8]$ f. $\mathsf{P}[X = 8 \mid \theta < 1/2]$

   One of the following is the correct numerical value of P. Report the correct value and tell how you knew (or how you eliminated the others). If you're unsure, explain. You should not need a calculator or table.

   a. $P = 0.000$ b. $P = 0.011$ c. $P = 0.500$
   d. $P = 0.800$ e. $P = 0.989$ f. $P = 1.000$

3. Toxicologists want to assess whether two treatment groups differ in the expected number of tumors per animal. There are n_1 animals in treatment group 1 and $n_2 = n_1$ animals in treatment group 2. Animals are exposed at the beginning of the experiment and at the end the number of tumors is counted so that a count $y_i \in \{0, 1, 2, \ldots\}$ is recorded for each animal. Based on previous studies, scientists claim that the Poisson assumption is warranted.

   Question: Let the null hypothesis correspond to equivalence between the two groups and the alternative to any difference in the expected tumor counts per animal. Calculate the Bayes factor in favor of the alternative hypothesis under conjugate $Ga(1, 1)$ priors, expressed only in terms of the number of animals $n = n_1 + n_2$ and the total number of tumors on all animals.

4. (a) Two scalar random quantities x, z have a joint distribution with complete conditionals $(x|z) \sim N(x|\phi z, v)$ and $(z|x) \sim N(z|\phi x, v)$ for some known, positive parameter $\phi \in (-1, 1)$ and where $v = s(1 - \phi^2)$ for some $s > 0$.

   i. What are the margins $p(x)$ and $p(z)$? Show your reasoning.

   ii. What is the joint distribution of (x, z)

   iii. What is the precision matrix of the joint distribution of (x, z)?

   (b) A third random quantity y is conditionally distributed as $(y|x) \sim N(y|x, w)$ for some known $w > 0$. Also, given x, y is conditionally independent of z.

   i. What is the joint distribution of (y, x, z)

   ii. What is $E(x|y)$?

   iii. What is $E(z|y, x)$?

   iv. What is $E(z|y)$?

5. Suppose we have observations Y_{pc} for $p = 1, \ldots, P$, a sequence of peptide locations, and $c = 1, 2$ corresponding to two experimental conditions. To assess existence of a change point at a given location p^* (one of the P locations), the following model is assumed on these observations:

$$\begin{aligned} Y_{p1} &\sim N(\mu_p, \sigma^2), \;\; p = 1, \ldots, P \\ Y_{p2} &\sim N(\mu_p, \sigma^2), \;\; p = 1, \ldots, P; \;\; p \neq p^* \\ Y_{p^*2} &\sim N(\mu^*, \sigma^2), \end{aligned}$$

with the observations being conditionally independent within and across treatments given model parameters $\mu_1, \ldots, \mu_P$, μ^* and σ^2. Let $\Delta = (\mu^* - \mu_{p^*})/2$, $\bar{\mu}_{p^*} = (\mu^* + \mu_{p^*})/2$ and $\bar{\mu}_p = \mu_p$ for $p \neq p^*$.

(a) Find the distributions of $D_p = (Y_{p1} - Y_{p2})/2$ and $M_p = (Y_{p1} + Y_{p2})/2$ for $p = 1, \ldots, P$.

(b) Are the vectors $D = (D_1, \ldots, D_P)^T$ and $M = (M_1, \ldots, M_P)^T$ independent of each other given Δ, $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_P)^T$ and σ^2?

(c) Find the MLE of σ^2 and simplify it to a function of D alone.

(d) Find the MLE $\hat{\Delta}$ of Δ and the standard error $SE_{\hat{\Delta}}$ of $\hat{\Delta}$.

(e) What is the distribution of $(\hat{\Delta} - \Delta)/SE_{\hat{\Delta}}$?

6. Suppose $X_i|n_i, p \sim Bin(n_i, p), i = 1, 2, ..., I$, conditionally independent with the n_i and p unknown. Suppose we adopt a $Be(\alpha, \beta)$ prior for p. Suppose, a priori, the n_i are independent and consider two improper priors for them:

   (I) $\pi(n_i) \propto 1$, a discrete uniform

   (II) $\pi(n_i) \propto 1/n_i$, a "scale" or geometric prior

   We can immediately write down the joint posterior for $\{n_i\}, p$ up to proportionality and, for each of these priors, we can marginalize over the n_i to obtain the marginal posterior distribution for p.

   (a) Show that, under prior (I), the posterior distribution for p is improper if $\alpha < I$.

   (b) Show that, under prior (II), if all of the $x_i > 0$, the posterior distribution on p is the same as the prior.

   (c) Without doing any calculations, speculate on what the story would be if we considered the prior, $\pi(n_i) \propto 1/n_i^2$.

FYE 2013
May 6th to 8th 2013

Department of Statistical Science
Duke University

Take Home Data Analysis Problem

Right Heart Catheterization (RHC) is a procedure for directly measuring how well the heart is pumping blood to the lungs. RHC is often applied to critically ill patients for directing immediate and subsequent treatment. However, administering RHC may cause serious complications, though the risks are usually small. There is some debate whether the use of RHC actually leads to improved treatment.

The data set "`rhc_study`" (more information below) contains data on 3824 hospitalized adult patients at five medical centers in the U.S. The variable `rhc` (column 1) indicates whether RHC was applied within 24 hours of admission (TRUE/FALSE). Each patient was followed up with some treatment procedures that may have been influenced by the RHC result if it was performed on the patient. The outcome variable is `surv30` (column 54) which is a prognosis score describing the probability of survival at 30 days after completion of treatment. The prognosis score derivation is a standardized procedure and is calculated by following the same protocol for all patients at all centers. Based on information from a panel of experts, a set of 52 variables were identified that are potentially related to both the decision to use RHC and the outcome `surv30`.

An accompanying data set "`rhc_new`" contains information on 1911 "new" patients admitted to the same five centers for serious health complications similar to the patients in `rhc_study`. These new patients are to be treated upon, with or without RHC being used to determine the choice of treatment. The data set `rhc_new` has the same 54 columns as `rhc_study`, but its column 1 and column 54 contain only `NA`'s.

Present a three page (maximum) report addressing the question: *Should RHC be performed to assist treatment choices for new patients in order to maximize their individual prognosis scores?* Your report should discuss all relevant aspects of your analysis (exploratory, modeling and validation) with graphical and numerical summaries that are important for communicating results. The report should be written so that doctors could understand and apply the findings while making RHC recommendation for future care. While you may include code and other plots in the supplemental appendix, you should not assume that graders will read beyond the main report; all relevant material should be within the three page limit.

You can access the two data sets in either tab delimited format (readable to `R`) or comma delimited 'csv' format (readable to `Excel` and `R`) from `http://www.stat.duke.edu/~st118/fye13/takehome/`. Details on all 54 variables are given on the next page.

Submit your report electronically to Karen Herndon by email (`karen@stat.duke.edu`). Your report file should be named `fye13_codename.pdf`. Your report should not contain your name or any other identifier. It MUST include your assigned code name.

rhc	RHC applied (TRUE / FALSE)
age	Age (years)
sex	Male/Female
race	black/white/other
edu	Education (years)
income	> $50k/ $11-$25k/ $25-$50k/ Under $11k
ninsclas	Insurance (Medicaid/Medicare Medicare & Medicaid/No insurance/Private/Private & Medicare)
cat1	Primary disease category (ARF/CHF/Cirrhosis/Colon Cancer/Coma/COPD/Lung Cancer /MOSF+Malignancy/MOSF+Sepsis)
cat2	Secondary disease category (Cirrhosis/Colon Cancer/Coma/Lung Cancer/MOSF+Malignancy/ MOSF+Sepsis/ None)
resp	Respiratory diagnosis (Yes/No)
card	Cardiovascular diagnosis (Yes/No)
neuro	Neurological diagnosis (Yes/No)
gastr	Gastrointestinal diagnosis (Yes/No)
renal	Renal diagnosis (Yes/No)
meta	Metabolic diagnosis (Yes/No)
hema	Hematological diagnosis (Yes/No)
seps	Sepsis diagnosis (Yes/No)
trauma	Trauma diagnosis (Yes/No)
ortho	Orthopedic diagnosis (Yes/No)
das2d3pc	DASI – Duke Activity Status Index
dnr1	Do Not Resuscitate status on day 1 (Yes/No)
ca	Metastatic cancer (Yes/No)
surv2md1	Estimate of prob. of surviving 2 months
aps1	APACHE score
scoma1	Glasgow coma score
wtkilo1	Weight (Kg)
temp1	Temperature (Celsius)
meanbp1	Mean Blood Pressure
resp1	Respiratory Rate
hrt1	Heart Rate
pafi1	PaO2=FI02 ratio
paco21	PaCO2
ph1	PH
wblc1	WBC
hema1	Hematocrit
sod1	Sodium
pot1	Potassium
crea1	Creatinine
bili1	Bilirubin
alb1	Albumin
cardiohx	Cardiovascular symptoms (TRUE/FALSE)
chfhx	Congestive Heart Failure (TRUE/FALSE)
dementhx	Dementia, stroke or cerebral infarct, Parkinsons disease (TRUE/FALSE)
psychhx	Psychiatric history, active psychosis or severe depression (TRUE/FALSE)
chrpulhx	Chronic pulmonary disease, severe pulmonary disease renalhx (TRUE/FALSE)
renalhx	Chronic renal disease, chronic hemodialysis or peritoneal dialysis (TRUE/FALSE)
liverhx	Cirrhosis, hepatic failure (TRUE/FALSE)
gibledhx	Upper GI bleeding (TRUE/FALSE)
malighx	Solid tumor, metastatic disease, chronic leukemia=myeloma, acute leukemia, lymphoma (TRUE/FALSE)
immunhx	Immunosuppression, organ transplant, HIV, Diabetes Mellitus, Connective Tissue Disease (TRUE/FALSE)
transhx	Transfer (> 24 hours) from another hospital (TRUE/FALSE)
amihx	Definite myocardial infarction (TRUE/FALSE)
wt0	Missing weight recorded as 0 (TRUE/FALSE)
surv30	Outcome: prognosis of chance of surviving more than 30 days

Take-home Applied Exam

- Keep your answer concise and to the point.
- Present your results in a three page (maximum) report addressing the primary questions posed.
- Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results.
- You may include code and other plots in a supplemental appendix; BUT, you should not assume that graders will read beyond the main report; all relevant material should be within the three page limit.
- You may use all notes, books, software etc from courses and studies to date, and build on your cumulated experience in applied modeling and data analysis.
- BUT– you are also bound by this honor pledge and must sign below to confirm this:
  - I confirm that this Take-home Exam submission is my work alone.
  - I have not consulted at all with any other students, whether they are taking the exam or not.
  - I have not copied nor adapted the work of others, nor provided help or advice to others on this exam.
  - I have not sought out or used any external sources (past student projects, publications, web sites, etc) that explicitly address any aspects of the specific data set and applied problem here. In particular, I have not used web searches to find previous references to the data and earlier analyses of this specific data set and problem, of any kind.
- Sign below and hand this in with your solution.

Name:

Signature:

Date: May 8th 2013

Distribution	Notation	$f(x) =$ **pdf (pmf)**	Support	Mean	Variance	
Beta	$Be(a,b)$	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$x \in (0,1)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	
Bernoulli	$Bern(p)$	$f(x) = p^x q^{(1-x)}$	$x \in \{0,1\}$	p	$p\,q$	$(q = 1-p)$
Binomial	$Bin(n,p)$	$f(x) = \binom{n}{x} p^x q^{(n-x)}$	$x \in \{0,\cdots,n\}$	$n\,p$	$n\,p\,q$	$(q = 1-p)$
Chi-square	$\chi^2(\nu)$	$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}\, e^{-x/2}$	$x \in \mathbb{R}_+$	ν	2ν	
Exponential	$Ex(\lambda)$	$f(x) = \lambda\, e^{-\lambda x}$	$x \in \mathbb{R}_+$	$1/\lambda$	$1/\lambda^2$	
Gamma	$Ga(\nu,\lambda)$	$f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1}\, e^{-\lambda x}$	$x \in \mathbb{R}_+$	ν/λ	ν/λ^2	
Geometric	$Geo(p)$	$f(x) = p\, q^x$	$x \in \mathbb{Z}_+$	q/p	q/p^2	$(q = 1-p)$
		$f(y) = p\, q^{y-1}$	$y \in \{1,\ldots\}$	$1/p$	q/p^2	$(y = x+1)$
HyperGeo.	$HG(n,M,N)$	$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$x \in 0,\cdots,n$	$n\,p$	$n\,p\,(1-p)\frac{N-n}{N-1}$	$(p = \frac{M}{N})$
Logistic	$Lo(\mu,\beta)$	$f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$	$x \in \mathbb{R}$	μ	$\pi^2\beta^2/3$	
Log Normal	$LN(\mu,\sigma^2)$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}_+$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$	
Neg. Binom.	$NB(\alpha,p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha\, q^x$	$x \in \mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$	$(q = 1-p)$
		$f(y) = \binom{y-1}{y-\alpha} p^\alpha\, q^{y-\alpha}$	$y \in \{\alpha,\ldots\}$	α/p	$\alpha q/p^2$	$(y = x+\alpha)$
Normal	$N(\mu,\sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}$	μ	σ^2	
Pareto	$Pa(\alpha,\epsilon)$	$f(x) = \alpha\,\epsilon^\alpha / x^{\alpha+1}$	$x \in (\epsilon,\infty)$	$\frac{\epsilon\,\alpha}{\alpha-1}$	$\frac{\epsilon^2\alpha}{(\alpha-1)^2(\alpha-2)}$	
Poisson	$Poi(\lambda)$	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{Z}_+$	λ	λ	
Snedecor F	$F(\nu_1,\nu_2)$	$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \times x^{\frac{\nu_1-2}{2}} \left[1+\frac{\nu_1}{\nu_2}\,x\right]^{-\frac{\nu_1+\nu_2}{2}}$	$x \in \mathbb{R}_+$	$\frac{\nu_2}{\nu_2-2}$	$\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}$	
Student t	$t(\nu)$	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}[1+x^2/\nu]^{-(\nu+1)/2}$	$x \in \mathbb{R}$	0	$\nu/(\nu-2)$	
Uniform	$U(a,b)$	$f(x) = \frac{1}{b-a}$	$x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Weibull	$Wei(\alpha,\beta)$	$f(x) = \alpha\beta\, x^{\alpha-1}\, e^{-\beta\, x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	