Department of Statistical Science Duke University

FIRST YEAR EXAM - SPRING 2017

Monday May 8th 2017, 9:00 AM – 1:00 PM

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Please use **black pen/ink** (no pencils) to complete your final solutions.
3. Put the problem number and your assigned code on the top of **each page**.
4. Write only on **one side** of the page (solutions on the reverse side of the page will be ignored).
5. Start each problem on a new page.
6. 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.
7. You have 4 hours to finish the written exam.
8. **You may choose to complete five out of the six questions. In case you attempt all six questions, only five will be graded and please clearly indicate which five you choose to be graded.**
9. All five graded questions will carry *equal* weight.
10. This is a closed book exam. No notes are permitted.

1. In a random sequence of independent binary trials, $x_i = 0$ or 1 $(i = 1, \ldots, n)$, let π be the success probability for each trial and write $x_{1:n} = \{x_1, \ldots, x_n\}$.

   (a) What is the distribution of $t = \sum_{i=1}^{n} x_i$? What is the expected value of t given π, n?

   (b) Derive the information function $I(x_{1:n}|\pi) = -\frac{\delta^2}{\delta\pi^2} log(p(x_{1:n}|\pi))$, and deduce the formula for the Fisher information function $I(\pi) = \mathsf{E}(I(x_{1:n}|\pi))$.

   (c) Under the reference prior $p(\pi) \propto I(\pi)^{1/2}$, what is the reference posterior $p(\pi|x_{1:n})$?

   (d) Under this reference prior, what is the predictive probability of success at the next trial, $\Pr(x_{n+1} = 1|x_{1:n})$? Comment on the form of this as a function of $x_{1:n}$, referring particularly to cases when the observed success total t is close to or equal to either 0 or n.

   (e) Find $\hat{\pi}$, the MLE of π. Identify the asymptotic normal approximation to $p(\pi|x_{1:n})$, namely $\mathsf{N}(\hat{\pi}, \sigma_n^2)$, and give a formula for σ_n^2. Show that, whatever the data may be, $\sigma_n \leq (4n)^{-1/2}$.

   (f) In sample survey reports, estimates of population percentages are typically presented as, for example, *"39% with a 3% margin of error";* typical sample sizes are around $n = 1{,}000$. With relevant discussion, show how such statements may be explicitly interpreted in terms of approximate 95% highest density regions for π.

   (g) An analyst is interested in inference on the log-odds ratio $\mu = \log(\pi/(1-\pi))$. How would you advise her to summarize inferences on μ? Give detailed comments and support them theoretically. Be explicit about how such inferences will be computed. Include comments on this when n is large as well as for cases when t is close to either 0 or 1.

2. Let $X \sim \text{Uniform}(0, \theta)$ and consider estimating θ under squared error loss.

   (a) Suppose we place an $\text{Exp}(\lambda)$ prior on θ

   $$\pi(\theta) = \lambda e^{-\lambda\theta} \quad \text{for } \theta > 0.$$

   What is the posterior of θ? What is the Bayes estimator? What is the Bayes risk of this estimator?

   (b) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from $\text{Uniform}(0, \theta)$, let $\delta = \max(X_1, X_2 \ldots, X_n) + 1$. Is δ an admissible estimator for θ? Why or why not?

3. A random variable Y is normally distributed: $f(Y) = \mathsf{N}(\mu, \sigma^2)$. We want to estimate μ and σ^2. Suppose we try to collect a sample of n independent observations, but we observe only $0 < N_1 \leq n$ of them. Here, N_1 is a random variable; that is, before you take the sample, you don't know how many values will be missing.

   Let R be a random variable so that $R = 1$ when an observation is observed and $R = 0$ when it is missing. Each observation is observed with probability p and missing with probability $(1-p)$, independently of whether other observations are missing or observed. Let $(R_1, \ldots, R_n)$ be the vector of zeros and ones corresponding to whether each observation is observed or missing.

   Suppose that observations are missing for reasons unrelated to the values of Y. Thus, $f(Y \mid R = 0) = f(Y \mid R = 1) = \mathsf{N}(\mu, \sigma^2)$.

   (a) Suppose for the moment that you know $(R_1, \ldots, R_n)$ and $\sum_i R_i = n_1$, but you do not yet know the values of any Y_i. Find expressions for $\mathsf{E}(\sum_{i:R_i=1} Y_i/n_1 \mid \mu, \sigma^2)$ and $\mathsf{Var}(\sum_{i:R_i=1} Y_i/n_1 \mid \mu, \sigma^2)$. Here, $\sum_{i:R_i=1} Y_i/n_1$ is the mean of the observed values of Y.

   (b) Now go back to the case where you don't know $(R_1, \ldots, R_n)$, which is the case for the remainder of the problem. Someone proposes to use the estimator, $\bar{Y}_{obs} = \sum_{i:R_i=1} y_i/N_1$. This is the sample mean of the observed cases. Prove that $\mathsf{E}(\bar{Y}_{obs} \mid \mu, \sigma^2) = \mu$.

   (c) Show that $\mathsf{Var}(\bar{Y}_{obs} \mid \mu, \sigma^2) = \sigma^2 \mathsf{E}(1/N_1)$.

   (d) Someone proposes the following procedure to estimate μ and σ^2. First, set all the $n - N_1$ values of Y equal to the sample mean of the observed cases, i.e., equal to $\bar{Y}_{obs}$. We write the resulting data as $(Y_1^*, \ldots, Y_n^*)$, where $Y_i^* = Y_i$ when $R_i = 1$ and $Y_i^* = \bar{Y}_{obs}$ when $R_i = 0$. Second, estimate μ with $\bar{Y}^* = \sum_{i=1}^n Y_i^*/n$. Third, estimate σ^2 with $s^{2*} = \sum_{i=1}^n (Y_i^* - \bar{Y}^*)^2/(n-1)$.

      i. Is $\bar{Y}^*$ unbiased for μ? Give an intuitive explanation or proof justifying your answer.

      ii. Is s^{2*} unbiased for σ^2? Give an intuitive explanation or proof justifying your answer.

4. Consider the sequence of random variables $X_1, X_2, ...$, where the pdf of X_n is equal to

$$f_n(x) = \begin{cases} (n-1)/2 & -1/n < x < 1/n \\ 1/n & n < x < n+1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is $\mathsf{E}[X_n]$?

(b) What is the variance of X_n? What is the variance of X_n as $n \to \infty$?

(c) Is there convergence in probability for X_n or stated in another way what is the probability that X_n is less than ε from zero?

5. Suppose that we have random variables $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\theta, 1)$, but some observations are censored at a, that is, we observe $\min(Y_i, a)$ for these censored observations. Order the observations so that $(y_1, \ldots, y_m)$ are uncensored and $(y_{m+1}, \ldots, y_n)$ are censored (and so equal a). We wish to estimate θ.

   (a) Give the likelihood function for all the observed (censored and uncensored) data.

   (b) Introduce missing data $(z_{m+1}, \ldots, z_n)$ and give the density of the missing data.

   (c) Write the complete data likelihood.

   (d) Give the expected complete data log-likelihood and derive the EM algorithm for finding the MLE of θ.

6. The entropy of a random variable measures our inability to predict it. For a continuous random variable with density $p(y)$, the entropy is given by

$$-\int \ln p(y)p(y)dy,$$

which is the negative of the "average height" of the log density.

(a) Using the fact that $\ln x \leq x - 1 \ \forall x \geq 0$ show that

$$-q \ln q \leq -q \ln p + (p - q) \text{ for all } p \geq 0, q \geq 0.$$

(b) Let $p_\theta(y) = \exp\{t(y) \cdot \theta - a(\theta)\}$ be a member of a continuous K-parameter regular exponential family with $\theta \in \Theta$, the natural parameter space. Let $q(y)$ be any other probability density such that $\mathrm{E}_q[t(y)] = \mathrm{E}_\theta[t(y)] \equiv \int t(y)p_\theta(y)\ dy$. Show that the entropy of $p_\theta(y)$ is at least as great as that of $q(y)$, i.e.

$$-\int q(y) \ln q(y)dy \leq -\int p_\theta(y) \ln p_\theta(y)dy.$$

(c) Let $q(y)$ be a probability density, let $t(y) : \mathcal{Y} \to \mathbb{R}^K$ and let ψ_0 be the q-expectation of $t(y)$, i.e.

$$\int t(y)q(y)\ dy = \psi_0.$$

Now suppose we are to sample data from q but model it as having come from a member of the exponential family $\{p_\theta(y) = \exp\{t(y) \cdot \theta - a(\theta\} : \theta \in \Theta\}$, of which q is not necessarily a member.

Let $Y_1, \ldots, Y_n \sim$ i.i.d. q and let $l(\theta) = \frac{1}{n} \sum \ln p_\theta(y_i)$.

i. For a given θ, what will $l(\theta)$ converge to as $n \to \infty$? Give your reasoning.

ii. Let $\hat{\theta}$ be the maximizer of $l(\theta)$ and $\hat{\psi} = \mathrm{E}_{\hat{\theta}}[t(Y)]$. What is $\hat{\psi}$ converging to?

FYE 2017
May 8, 1pm to May 11, 12noon, 2017

Department of Statistical Science
Duke University

Take Home Data Analysis Problem

A question that radiologists examine is how well can one predict survival outcomes of patients from images of a tumor. In this problem you are given clinical data for almost 100 brain tumors, Glioblastomas, and statistical summaries capturing the morphology of MRI images of these tumors. The morphological summaries were specified by collaborations between radiologists and computer scientists,

There are two data files at `www.stat.duke.edu/~sayan/FYE_17`

(1) The clinical variables are in `TCGA_Clinical_Data.csv`

(2) The morphological summaries are in `Morphometric_Features.txt`

The two main clinical variables of interest are disease free survival (`Disease Free`) and overall survival (`Survival`) with both in `TCGA_Clinical_Data.csv`. Your goal is to provide some quantification of how the morphological features perform as covariates in predicting each of the the two clinical variables or in explaining variation of the two clinical variables. Analysis of each of the clinical outcomes as well as a joint analysis of the outcomes is of interest.

Write a report (maximum 3 pages) based on your analysis describing your findings. Be thorough in your exploratory analyses and exploratory use of models; applied work that overly emphasizes complicated modeling to begin is often less valuable than careful, incisive evaluation of data through simpler, exploratory models– at least to begin.

Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results.

Take-home Applied Exam

- Present your results in a three page (maximum) report addressing the primary questions posed. Keep your answers concise and to the point.
- 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.
- You may freely use other resources– code, literature, etc– from whatever source you like, so long as you do not violate the condition 4 below.
- **To Confirm**– you are also bound by this honor pledge and must sign below to confirm this:
    1. I confirm that this Take-home Exam submission is my work alone.
    2. I have not consulted at all with any other students, whether they are taking the exam or not.
    3. I have not copied nor adapted the work of others, nor provided help or advice to others on this exam.
    4. 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 before or at 12noon, May 11 (Thursday) 2017 to Lori Rauch at Room 214, Old Chem.

Name:

Signature:

Date: May 11, 2017

DISTRIBUTION	NOTATION	$f(x) =$ PDF(PMF)	RANGE	MEAN	VARIANCE	
Beta	$Be(\alpha, \beta, a, b)$	$f(x) = \frac{1}{b-a}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{x-a}{b-a}\right)^{\alpha-1}\left(\frac{b-x}{b-a}\right)^{\beta-1}$	$x \in (a, b)$	$a + (b-a)\frac{\alpha}{\alpha+\beta}$	$\frac{(b-a)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	
Binomial	$Bi(n, p)$	$f(x) = \binom{n}{x} p^x q^{(n-x)}$	$x \in 0, \cdots, n$	$n\,p$	$n\,p\,q$	$(q = 1-p)$
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	$Ga(\alpha, \beta)$	$f(x) = \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}\,e^{-x/\beta}$	$x \in \mathbb{R}_+$	$\alpha\beta$	$\alpha\beta^2$	
Geometric	$Ge(p)$	$f(x) = p\,q^x$	$x \in \mathbb{Z}_+$	q/p	q/p^2	$(q = 1-p)$
		$f(y) = p\,q^{y-1}$	$y \in \{1, ...\}$	$1/p$	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	$\mathsf{Lo}(\mu, \beta)$	$f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$	$x \in \mathbb{R}$	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.	$NB(\alpha, p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha\,q^x$	$x \in \mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$	$(q = 1-p)$
		$f(y) = \binom{y-1}{y-\alpha} p^\alpha\,q^{y-\alpha}$	$y \in \{\alpha, ...\}$	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	$\mathsf{Pa}(\alpha, \epsilon)$	$f(x) = \alpha\,\epsilon^\alpha/x^{\alpha+1}$	$x \in (\epsilon, \infty)$	$\frac{\epsilon\,\alpha}{\alpha-1}$	$\frac{\epsilon^2\alpha}{(\alpha-1)^2(\alpha-2)}$	
Poisson	$Po(\lambda)$	$f(x) = \frac{\lambda^x}{x!}e^{-\lambda}$	$x \in \mathbb{Z}_+$	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	$We(\alpha, \beta)$	$f(x) = \alpha\beta\,x^{\alpha-1}\,e^{-\beta\,x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	