FIRST YEAR EXAM

May 8, 2002

Notes:

1. This is a closed book exam. No notes are permitted.
2. Please write your solutions on the paper provided. Put the problem number and your 3 character code on the top of **each page** that you turn in.
3. Write only on **one side** of the page.
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; draw a line through work you do not want graded.
6. All 6 problems will be graded.
7. You have 3 hours to finish.
8. The Takehome Portion will be given out at 9am May 9th, and due at 12 noon May 10th.

1. Let $X_1, ..., X_n \sim f(x|\theta)$ be *i.i.d.* random variables, each with probability density function

$$f(x|\theta) = \begin{cases} e^{\theta - x} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

for some uncertain $\theta \in \Theta = \mathbb{R}$. Our goal is to estimate θ.

(a) Find the mean $\mathsf{E}[X_1 \mid \theta]$ and variance $\mathsf{V}[X_1 \mid \theta]$.

(b) Find the maximum likelihood estimator (MLE) $\hat{\theta} = \hat{\theta}(\vec{x})$ for a random sample $\vec{x} = \{X_1, ..., X_n\}$. Show your work.

(c) Is the MLE sufficient for θ? Why or why not?

(d) Find the probability density function for the MLE $\hat{\theta}$ (it will depend on θ).

(e) Find the p-value for a test of the hypothesis $H_0 : \theta = 0$ *vs.* $H_1 : \theta > 0$ with the data set $\vec{x} = \{2.5, 1.0, 1.5, 4.0, 6.0\}$. Would you accept or reject H_0 at level $\alpha = 0.01$?

2. Landrigan et al. published the article, "Neuropsychological Dysfunction in Children with Chronic Low-Level Lead Absorption," in the medical journal, Lancet, more than 25 years ago. The study investigated the relationship between lead absorption and neuropsychological function. Blind evaluations were undertaken in 46 children aged 3-15 years with high blood-lead concentrations and in 78 ethnically and socioeconomically similar children with low blood-lead levels. All children lived within 5 miles of a large lead-emitting smelter, and in many cases residence there had been lifelong. Intelligence (IQ) scales, neurological tests, and medical histories were compared between the high and low lead groups. (These groups will be referred to as the lead-absorption and control groups).

   A table in the paper lists the following IQ means (standard deviations) for the lead-absorption and control groups. There are 3 subscales of the WPPSI test: verbal, performance, and full-scale.

WPPSI Test	Lead-absorption Group	Control group
Verbal IQ	84.92 (11.60)	84.47 (12.39)
Performance IQ	90.67 (11.64)	100.27 (11.29)
Full-scale IQ	86.17 (11.17)	91.20 (11.88)

   (a) Calculate a 95% confidence or credible interval for the difference in group means on the WPPSI Performance IQ. Of children taking the WPPSI performance test, 12 were in the lead-absorption group and 15 were in the control group. Be sure to write down ALL numbers that go into your calculation and state any assumptions that you make.

   (b) Does your interval support rejecting the null hypothesis of equal means for the two groups as the paper claims? Why or why not?

   (c) Another table gives the IQ results after deleting children with PICA (craving for nonfood items). Now the difference in WPPSI performance IQ averages is no longer statistically significant at the .05 level [92.22 (sd =11.77) versus 99.93 (11.64)]. Which of the following rationales could account for this? (Select all that apply)

      i. The children with PICA had bigger differences in performance IQ between the lead-absorption and control groups than did the children without PICA.

      ii. Removing the PICA children decreased the sample size.

      iii. Removing the PICA children increased the standard deviation of the estimate of the difference in average performance IQ between groups.

      iv. There was more variability in the PICA group than in the non-PICA group.

      v. The power of the test increased.

   (d) In the study there were at least 200 tests of significance carried out, although many were not reported. About 1 dozen or so are statistically significant at the 0.05 level. If all 200 null hypotheses were true, and the tests of significance were independent, how many would you expect to reject at the 0.05 level?

3. A scientist wants to measure a physical constant θ. There are two measuring devices available, **A** and **B**. The accuracy of **A** is known; the accuracy of **B** is not. Consider the following two scenarios:

   1 Measurements $x_1, \ldots, x_{n_A}$ are taken on device **A**; where

   $$x_1, \ldots, x_{n_A} \overset{i.i.d.}{\sim} N(\theta, 1)$$

   .

   2 Measurements $y_1, \ldots, y_{n_B}$ are taken on device **B**; where

   $$y_1, \ldots, y_{n_B} \overset{i.i.d.}{\sim} N(\theta, 1/\tau)$$

   and τ is an unknown precision parameter.

   (a) Find a minimal sufficient statistic for θ under Scenario 1.

   (b) Approximately how well can θ be estimated under Scenario 1?

   (c) Find a minimal sufficient statistic for (θ, τ) under Scenario 2.

   (d) If τ has the prior density

   $$p(\tau) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \tau^{\alpha-1} e^{-\tau/\beta} \qquad \tau > 0,$$

   and $p(\theta|\tau) \propto 1$, find the posterior density of (θ, τ) and the marginal posterior density of τ under Scenario 2.

   (e) Now suppose you are told that $\tau = 0.1$. Measurements on **A** cost \$1000 each and measurements on **B** cost \$100 each. The scientist has \$1000 to spend on measurement. Which is better: one measurement on **A** or 10 measurements on **B**?

4. Suppose that the conditional distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ has mean

$$\mathsf{E}[\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2] = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \tag{1}$$

and constant variance,

$$\mathsf{V}[\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2] = \sigma^2 I_n \tag{2}$$

where I_n is a $n \times n$ identity matrix, $\mathbf{X}_1$ is a $n \times p$ matrix and $\mathbf{X}_2$ is a $n \times q$ matrix of predictor variables, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are p and q dimensional vectors of regression coefficients, respectively.

(a) Suppose the mean function is given by (1), but we fit the mean as a function of $\mathbf{X}_1$ alone. Find $\mathsf{E}[\mathbf{Y}|\mathbf{X}_1]$ assuming that $\mathbf{X}_2|\mathbf{X}_1$ has a distribution with finite mean, $\mathsf{E}[\mathbf{X}_2|\mathbf{X}_1]$, and variance, $\mathsf{V}[\mathbf{X}_2|\mathbf{X}_1]$.

(b) Find $\mathsf{V}[\mathbf{Y}|\mathbf{X}_1]$. Give sufficient conditions for this to be constant for all values of $\mathbf{X}_1$.

(c) If we fit a regression of $\mathbf{Y}$ on $\mathbf{X}_1$ of the form $\mathbf{X}_1\boldsymbol{\alpha}$ using ordinary least squares (OLS), what is the expected value of $\hat{\boldsymbol{\alpha}}$ (the OLS estimator), when the mean is actually given by $\mathsf{E}[\mathbf{Y}|\mathbf{X}_1]$?

(d) If the distibution of $\mathbf{Y}$ given $\mathbf{X}_1$ and $\mathbf{X}_2$ has mean and variance given by (1) and (2), respectively, is the OLS estimator $\hat{\boldsymbol{\alpha}}$ from above an unbiased estimator of $\boldsymbol{\beta}_1$? Justify.

(e) Many studies that estimate health effects of EPA criteria pollutants, such as NO_x and particulate matter (PM), do regression analyses using one pollutant at a time. If particulate matter has a mean that is a linear in NO_x and has variance independent of NO_x, what are possible consequences for policy making regarding Particulate Matter (PM), if NO_x is not included in the regression model? (you may assume that PM and NO_x are positively correlated and the sample sizes are large).

5. Brain tumour samples from n patients are allocated into two tumour types: benign tumour (0) or aggressive tumour (1). There are n_0 benign and n_1 aggressive tumours.

   A specific gene is recorded as being active ($g_i = 1$) or inactive ($g_i = 0$) for each tumour. It is of interest to explore whether or not the active/inactive status of the gene relates to whether or not the tumour is benign/aggressive. This is addressed in a model under which the g_i represent a sequence of independent Bernoulli trials with $Pr(g_i = 1) = \pi_0$ for benign tumours and $Pr(g_i = 1) = \pi_1$ for aggressive tumours.

   Write H for the assumption (hypothesis) that $\pi_0 = \pi_1$ and $\bar{H}$ for the alternative, more general hypothesis $\pi_0 \neq \pi_1$. Write $G = \{g_1, \ldots, g_n\}$ and suppose that the data indicate x_0 of the n_0 benign tumours have the gene active, while x_1 of the n_1 aggressive tumours have the gene inactive; write $x = x_0 + x_1$ and $n = n_0 + n_1$.

   The Bernoulli sampling models under $\bar{H}$ and H respectively are

$$\begin{aligned} \text{Under} \quad \bar{H}: \quad & p(G|\pi_0, \pi_1) = \pi_0^{x_0}(1-\pi_0)^{n_0-x_0} \times \pi_1^{x_1}(1-\pi_1)^{n_1-x_1} \\ \text{Under} \quad H: \quad & p(G|\pi) = \pi^x(1-\pi)^{n-x} \end{aligned}$$

   (a) Evaluate the marginal (or prior predictive) density $p(G|H) = \int_0^1 p(G|\pi)p(\pi)d\pi$ as a function of (n, x), assuming the prior $\pi \sim U(0, 1)$.

   (b) Evaluate the related marginal density $p(G|\bar{H})$ as a function of (n_1, x_1) and (n_0, x_0), assuming the priors $\pi_0 \sim U(0, 1)$ and $\pi_1 \sim U(0, 1)$, independently.

   (c) Suppose that you assign $Pr(H) = 0.5$. Give an expression for the posterior probability $Pr(H|G)$ in terms of $p(G|H)$ and $p(G|\bar{H})$.

   (d) The clinical research protocol chooses $n_0 = 34$ benign and $n_1 = 40$ aggressive tumours, and the analysis reports $x_0 = 5$ and $x_1 = 17$. This results in the Bayes' factor $B = p(G|H)/p(G|\bar{H}) \approx 0.13$. What is the implied posterior probability $Pr(H|G)$ in this case? Is this evidence for or against an association between the gene activity and the tumour type?

   (e) Based on this specific data set, and assuming $\bar{H}$, what are the posterior distributions for π_0 and π_1? What are the posterior means of π_0 and π_1?

   It is generally agreed that the incidence rate of aggressive tumours is about 15% – i.e., for a randomly selected patient, the probability of an aggressive tumour is about 0.15.

   (f) A further patient is assessed, pre-surgery, for activity of the gene; it is determined that the gene is inactive for this individual. Assuming $\bar{H}$, give an expression for the probability, θ, that this patient has an aggressive tumour, in terms of π_0 and π_1.

   (g) With the data values above, what is the MLE of θ?

   (h) Describe how you would compute the approximate posterior mean and a 90% posterior credible interval for θ.

A useful identity: $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ when $a, b > 0$.

6. Gambling Joe is playing a fair game against the house where he has equal probability of winning or losing his bet each time. Suppose Joe begins with $M_0 = \$35$ and quits when he hits \$0 or has at least \$100. Let τ be the time Joe quits, and let M_n be the amount of money Joe has after n games.

   (a) Joe's bet on the $(n+1)$st game is a function of the outcomes on the first n games (under the restriction that the bet is at most M_n; Joe cannot bet more money than he has.) Show that M_n is a martingale with respect to the information in the first n games.

   (b) Show that if Joe always bets at least one dollar, that τ is finite with probability 1.

   (c) If Joe always bets \$1, what is the probability that $M_\tau = 0$ (that Joe loses everything?)

   (d) Now suppose Joe bets anywhere between \$1 and \$10 (but he still can't bet more money than he currently has). Give upper and lower bounds for $P(M_\tau) = 0$.

   (e) Because Joe didn't study probability in school, he switches to an unfair game, where the probabilty of winning is 49/100 and the probability of losing is 51/100. Now he only quits if he is out of money, and he always bets \$2. If τ_2 is the time that Joe runs out of money, find
   $$\mathsf{E}[\tau_2 | M_0 = 20].$$

First Year Exam - Takehome
May 9, 2002
Understanding Sleep

Background

A particular individual sometimes has difficulty falling asleep, and sometimes awakens during the night repeatedly or for significant amounts of time. Upon consulting a sleep doctor and having his sleep monitored, he was told told that his sleep patterns were characteristic of a normal sleeper who spends too much time in bed. The doctor's hypothesis is that a typical human needs a certain amount of sleep per night (which may vary across humans), so if the human is habitually given much more time in bed, they will just sleep less efficiently. The subject was instructed to keep a sleep log for several weeks, find the average amount of actual total sleep time (TST), add about half an hour, and that was the amount of time he should regularly spend in bed trying to sleep.

Other factors may also be involved in sleep patterns and quality. Exercise is supposed to help with sleep. Alcohol is supposed to decrease the amount of time it takes to fall asleep, but also decrease the quality of sleep, leading to more frequent awakenings during the night. The subject's mother read somewhere that calcium-magnesium supplements taken at dinnertime are supposed to help one fall asleep faster and sleep better. Staying up later than typical may decrease the time to fall asleep. The quantity and/or quality of the previous night's sleep could affect the current night.

The Data

In the file `http://www.isds.duke.edu/info/FYE/sleep.txt` you will find information from the subject's sleep log, recorded over the course of several months (excluding trips out of town and the days immediately after returning). In this file, you will find the following variables:

1. Date (recorded in the morning after arising)
2. Bedtime (recorded as a four-digit number where the first two are the number of hours past noon and the last two are the number of minutes past that, e.g., a bedtime of 11:30PM is recorded as 1130, a bedtime of 1:05AM is recorded as 1305)
3. Time To Sleep (TTS, the number of minutes it took to fall asleep)
4. Total Sleep Time (TST, the total number of minutes spent asleep)
5. Total Bed Time (TBT, the total number of minutes spent in bed attempting to sleep)
6. Alc (an indicator for the consumption of alcohol later than the end of dinner)
7. Cal (an indicator for the consumption of a calcium-magnesium supplement with dinner)
8. Run (the number of miles run that day)

Of course, there may be measurement error in the TTS and TST variables, but since it's not clearly biased in either direction, ignore possible measurement errors for the sake of this problem.

Questions to Consider

Summarize your findings on this subject's sleep patterns in a two page report. You may include a supplemental appendix of no more than five pages. You should be sure to address the following issues, but they should not considered exclusively.

1. According to the doctor, how much time (TBT) should the subject be spending in bed trying to sleep?
2. Is the sleep doctor's hypothesis of a fixed amount of required sleep reasonable, given this dataset?
3. Do calcium-magnesium supplements help with either falling asleep or staying asleep during the night?
4. Do any other factors affect sleep patterns?