ISDS First Year Exam
May 4, 1996

Please answer 5 of the following questions. You have four hours to complete the exam. The exam is closed book; you may not use any reference materials (books, notes, friends, enemies, etc). Use of a calculator is o.k.

- Make sure that each answer is on a separate sheet(s) - and not on the exam. Do not use the back of the page.
- Please make up a three letter code and write it on the top of each page. DO NOT put your name on any of the pages.
- Do not turn in more than 5 questions.

Questions will carry substantial credit for correct *partial* answers.

Problem 1

A client of yours is interested in predicting Y *(log algal concentration in lakes)* based on the values of X *(log phosphorus per* cm^3*)* and Z *(log nitrogen per* cm^3). Both X and Z are important. The client has data from three different lakes over time. Lake 1 has data

$$y_{11}....y_{1n}, \quad x_{11}....x_{1n}$$

lake 2 has data

$$y_{21}....y_{2m}, \quad z_{21}....z_{2m}$$

and lake 3 has data

$$x_{31}....x_{3q}, \quad z_{31}....z_{3q}$$

Your goal is to build a model to make predictions for Y using X and Z.

1. You can only ask two very brief questions of your client. What would you ask?
2. Specify a model for the data and explain how to use it to perform the predictions. Write down enough detail that your RA could go off and compute an estimate of $p(y < 2|x = 1, z = .5)$ given the observed data above.
3. What is the most restrictive assumption in the model you specified? How would your expect this assumption to affect the answer that your RA will get?

Problem 2

PART A: IQs of students are distributed approximately normally with mean m and variance τ^2. Let x be your 'true' IQ, so that, assuming you are a random student, $x \sim \mathrm{N}(m, \tau^2)$. A standard IQ test is known(!) to be unbiased with normally distributed errors having a variance σ^2. So, if you take the test and your 'true ' IQ is x, the test measure y of your result has the conditional distribution $(y|x) \sim \mathrm{N}(x, \sigma^2)$.

Solve the following, quoting (NOT proving) standard normal theory results:

1. Find the marginal (= initial or prior predictive) distribution of y. If $m = 115$, $\tau = 10$ and $\sigma = 5$, how surprised will you be if your test result exceeds 137?

2. Your test result is y is observed; what do you now think about your IQ x?

3. If $m = 115$, $\tau = 10$ and $\sigma = 5$, how large will y need to be in order that $P(x > 125|y) \geq 0.95$? Interpret this result.

PART B:

Amongst diseased individuals ($\theta = 1$), measured blood levels x of a biochemical indicator are distributed as $(x|\theta = 1) \sim \mathrm{N}(m_1, \tau^2)$. Amongst healthy individuals ($\theta = 0$), the distribution is $(x|\theta = 0) \sim \mathrm{N}(m_0, \tau^2)$. Also $m_1 > m_0$, levels being higher amongst diseased people. The disease incidence rate is assumed known at 0.01, so that θ has discrete prior mass function $p(\theta)$ given by $p(1) = 0.01$ and $p(0) = 0.99$.

For a randomly chosen individual, use Bayes' theorem to find, as a function of the test result x, the *posterior odds on disease* $O(1|x) = P(\theta = 1|x)/P(\theta = 0|x)$. For a fixed number k ($>> 1$), show that the posterior odds exceed k times the prior odds of 1:99 when

$$x > \bar{m} + \log(k)\tau^2/(m_1 - m_0),$$

where $\bar{m} = .5(m_0 + m_1)$. Interpret this result, and comment on how the result depends on the size of $m_1 - m_0$.

Problem 3

A sociologist is studying sex discrimination in the workplace. Establishments are chosen randomly and a job is chosen randomly within the establishment to obtain a sample of size N. Let n_i and f_i denote the number of workers and women respectively, performing the selected job in establishment i for $i = 1, \ldots N$. We can regard f_i as an observation from a Binomial distribution with parameters n_i and p_i. The sociologist would like to learn about the distribution of p_i's across the country. What do you recommend? Can you suggest a way to address this problem? Please be as specific as you can.

Problem 4

A statistician is unsure whether a sequence of independent observations $\{Y_j\}$ arose from a standard exponential distribution, with density function $f_0(y) = \exp(-y)$, or possibly from a χ^2_1 distribution, with density function $f_1(y) = \exp(-y/2)/\sqrt{2\pi y}$. Both of these distributions have mean one, so she can't expect to tell which might be the right distribution just by looking at the sample mean $\bar{Y}_n$.

For both Bayesian and Frequentist reasons[1], she decides to compute the Bayes factor (or likelihood ratio) $B_n = \prod_{j=1}^{n} \frac{f_0(y_j)}{f_1(y_j)}$. Confident that B_n will converge to either zero or infinity, she plans to continue sampling until either $B_n > e^3 \approx 20$ (when she'll decide H_0 is probably true) or $B_n < e^{-3} \approx .05$ (when she'll reject H_0 in favor of H_1).

1. Her first ten observations had partial sum $\sum_{j=1}^{10} Y_j = 12$ and partial product $\prod_{j=1}^{10} Y_j = 4$; has she satisfied her stopping rule yet?

2. Which of the two distributions do these data appear to favor? With equal prior probabilities $\Pr[Y_j \sim f_0(y)] = \Pr[Y_j \sim f_1(y)] = 1/2$, what are the corresponding posterior probabilities?

3. Find $r_0 > 0$ and $r_1 > 0$ so that $M_n \equiv B_n / r_i^n$ is a martingale if $Y_j \sim f_i(y)$, for $i = 0, 1$. In what sense will M_n converge (almost surely, in L^p, in probability, or in distribution), and to what limit? What is the expectation of M_n? This should give you an idea of how fast B_n will converge.

You may find it useful to recall that the Gamma function $\Gamma(\alpha) \equiv \int_0^\infty t^{\alpha-1} e^{-t}\, dt$ satisfies $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and $\Gamma(1/2) = \sqrt{\pi}$.

[1] A Bayesian would find that the posterior odds of the hypothesis H_0 that the data come from $f_0(x)$ are B_n times the prior odds; by the Neyman-Pearson Lemma, a Frequentist would find that the most powerful test of the hypothesis H_0 will be one that rejects for small values of B_n.

Problem 5

PART A: Arrivals of buses at a certain bus stop is a Poisson process having rate $1/q$, the average number of arrivals in any hour-long period. So, for example, the probability of no buses arriving in any time period of length t is $\exp(-t/q)$. Also, the time between arrivals is exponential with mean q. You would like to estimate q. You learn that the driver of each bus must sign in at the bus stop with the time of the bus's arrival. You plan to arrive at a randomly selected time and check the latest time on the sign-in sheet. Subtracting from the current time gives you time T_1 (in hours) and seems to be an estimate of q. Then you realize that if you were to wait until the next bus arrives, say for time T_2, that time seems also to be an estimate of q. So $(T_1 + T_2)/2$ must be a better estimate than either time separately. But $T_1 + T_2$ is the actual time between two buses, so it too seems to be an estimate of q. Only one of the two, $(T_1 + T_2)/2$ and $T_1 + T_2$, can be an unbiased estimate of q. Giving reasons for your answer, which is it? Is this the maximum likelihood estimate for q? Explain.

PART B: Suppose your prior distribution for θ, the proportion of North Carolinians who support the death penalty, is beta with mean 0.6 and standard deviation 0.3.

1. Determine the parameters α and β of your prior distribution.
2. What is your prior predictive distribution for x, the proportion of North Carolinians in a sample of size n that favour the death penalty? (i.e. the marginal distribution $p(x)$)?
3. A random sample of 1000 North Carolinians is taken, and 20% support the death penalty. What are your posterior mean and variance for θ?
4. Instead of a beta prior we now use a Gamma prior on the odds $\eta = \theta/(1-\theta)$, with parameters chosen to imply still the same prior mean 0.6 and standard deviation 0.3 for θ. Find the mean and variance of the implied prior predictive distribution.

Hint: The p.d.f. of a Beta $Be(\alpha, \beta)$ r.v. is:

$$\begin{aligned} f(x) &= c\, x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \\ c &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned}$$

with expected value $E(X) = \frac{\alpha}{\alpha+\beta}$ and variance $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Problem 6

This question will carry substantial credit for correct *partial* answers. Solutions to parts (2), (3) and (4) rely on the result for $r(x)$ stated in (1), but you can attempt them even if you can't prove (1).

A simple version of signal detection models in observational astronomy hypothesizes Poisson arrival times for nuclear particles generated from astronomical sources, with such signals being obscured by "background" noise. Observations are made to investigate an hypothesis that there exists a source (eg: a supernova) of particles in a particular region of the sky. If such a source exists, the detected radiation from it will be weak and so assessment of whether or not it exists is obscured by the background noise.

Suppose particles arrive at a *known* background rate b and, independently, at an *unknown* rate β from the source. Then the model for the number of particles X recorded is $X \sim Po(b+\beta)$ with density $p(x|\beta) = (b+\beta)^x \mathrm{e}^{-(b+\beta)}/x!$, over $x = 0, 1, \ldots$.

Note that $\beta > 0$ is consistent with source existence; write $H = \{\beta > 0\}$, and suppose

- $P(H) = P(\beta > 0) = w$, the prior probability of source existence; and
- a prior density $p(\beta|H)$ over $0 < \beta$ for the source rate assuming the source exists.

We now observe $X = x$ particles in the unit of time.

1. Show that the posterior probability $P(H|x)$ is given by $P(H|x) = \mathcal{O}(H|x)/(1+\mathcal{O}(H|x))$ where $\mathcal{O}(H|x)$ is the posterior odds ratio

   $$\mathcal{O}(H|x) = \frac{w}{1-w} r(x)$$

   and where $r(x)$ is the generalised likelihood ratio

   $$r(x) = \int_0^\infty (1+\beta/b)^x \mathrm{e}^{-\beta} p(\beta|H) d\beta$$

2. Deduce that $x = 0$ always provides evidence *against* source existence.

3. Show that, whatever the prior $p(\beta|H)$ is, $r(x) \leq (1+\hat{\beta}/b)^x \mathrm{e}^{-\hat{\beta}}$ where $\hat{\beta}$ is the MLE of β under H. *(Hint: "a maximum always exceeds an average").* Comment on this result, and indicate its relevance in connection with the use of maximum likelihood ratio testing of source existence versus Bayesian assessment based on posterior odds ratios.

4. Find $\hat{\beta}$ as a function of x and b. If $b = 0.73$ (a background rate in a recent study of supernova SN1987A); compute the upper bounds on $r(1)$ and $r(2)$. Comment on the potential evidence in favour of source existence if only one or two counts are recorded.

Problem 7

PART A:

Based on 3 independent Bernoulli random observations $X_1, X_2,$ and X_3 with probability of success $p,\ 0 \leq p \leq 1$, it is desired to test

$$H_0 : p = \frac{1}{2} \quad vs. \quad H_1 : p \neq \frac{1}{2}.$$

When observed, the three observations are all equal to 1 (i.e. $X_1 = X_2 = X_3 = 1$).

1. Calculate the P-value (two-sided).
2. If H_0 and H_1 are each given prior probability $1/2$, and the prior mass on H_1 is distributed according to a uniform density on $[\lambda, 1-\lambda],\ 0 \leq \lambda < \frac{1}{2}$, calculate the posterior probability of H_0 as a function of λ.
3. Find the minimum posterior probability of H_0 in part 2 over all possible values of $\lambda,\ 0 \leq \lambda < \frac{1}{2}$.
4. Comment on the results in part 3.

PART B:

College Entrance Test. Because of the role of college aptitude test scores in college entrance decision, there are minicourses that purport to teach students how to take these tests. A particular aptitude test has been found to produce scores that are normally distributed, with mean 500 and standard deviation 60. If the minicourse directed at this test is effective (on average), the mean score θ of students who take the course is larger than 500; otherwise it is not. We want to test

$$H_0 : 480 \leq \theta \leq 520 \quad \text{versus} \quad H_1 : \theta \text{ not in } [480, 520],$$

and our prior for θ is $N(520, 30^2)$.

1. If 25 observations give the mean 510, perform the test and make the decision.
2. Find 95% credible set for θ. Compare the obtained credible set with the frequentist 95% confidence interval for the unknown mean θ. Explain why credible sets tend to be shorter than the corresponding confidence intervals.

Problem 8

PART A: An anthropologist has collected the breadths y_i of $n = 84$ male Etruscan skulls. The average skull breadth is $\bar{y} = 143.8mm$. The anthropologist also computed the *standard error* as $s/\sqrt{n} \approx 0.65$ where

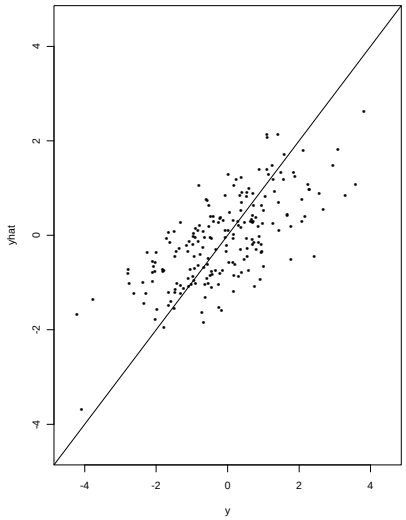
$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{(n-1)}}$$

For modern Italian males the maximum head breadth averages $132.4mm$. The anthropologist noticed that $(143.8 - 132.4)/.65 \approx 17$ and reasoned that the breadth of a randomly selected male Etruscan skull is very unlikely to be less than 132.4, because 132.4 is about 17 standard errors away from 143.8.

Do you agree with this reasoning? Why or why not?

PART B:

A researcher in the School of the Environment is creating a regression model to predict the concentration of chlorophyll a in Lake Okeechobee, Florida. Let y_i be the i'th measurement of chlorophyll a and $\hat{y}_i$ be its prediction from the model. The researcher produced the following plot of $\hat{y}$ versus y,



noticed that most small values of y had larger values of $\hat{y}$, most large values of y had smaller values of $\hat{y}$ and wrote "With an ordinary least squares (OLS) model, the plot of model predictions vs. observations forms a cloud of points about a line with a slope of one, which corresponds to predicted equals observed. ... For the model developed here it is clear that some 'shrinkage' toward the mean has occurred, since the cloud of points is not oriented along the solid line, which has a slope of one.

Can you explain what happened?

Problem 9

Consider the following model for a mixture of two normals.

$$\begin{aligned} y_i &\overset{iid}{\sim} \frac{1}{2} \cdot N(\mu_1, 1) + \frac{1}{2} \cdot N(\mu_2, 1), \quad i = 1, \ldots, n \\ \mu_j &\overset{iid}{\sim} N(0, 1/A^2) \end{aligned}$$

1. Show that an improper prior with $A^2 = 0$, i.e. $p(\mu_1, \mu_2) = const$, leads to an improper posterior.

2. To avoid the improper posterior we change the model to

$$y_i \overset{iid}{\sim} \frac{1}{2} \cdot N(\mu, 1) + \frac{1}{2} \cdot N(\mu + \mu_2, 1), \quad i = 1, \ldots, n \tag{1}$$

$$p(\mu_2|\mu) = N(0, 1), \tag{2}$$

$$p(\mu) \propto 1. \tag{3}$$

   The prior is noninformative in μ, and proper in μ_2 given μ.
   Show that the posterior could be multimodal.

3. To implement a Gibbs sampling scheme we replace the mixture (1) by:

$$y_i|z_i \sim \begin{cases} N(\mu, 1) & \text{if } z_i = 0 \\ N(\mu + \mu_2, 1) & \text{if } z_i = 1 \end{cases} \tag{4}$$

$$Pr(z_i = 1) = \frac{1}{2}. \tag{5}$$

   Show that model (4) and (5) are equivalent to model (1).

4. Propose a Gibbs sampling scheme to simulate from the posterior under model (4), (5) and (2),(3). Give all complete conditionals which are required for resampling:

   (a) $p(\mu|\mu_2, z_1, \ldots, z_n, y)$;

   (b) $p(\mu_2|\mu, z_1, \ldots, z_n, y)$;

   (c) $p(z_i|\mu, \mu_2, z_j, j \neq i, y)$, $i = 1, \ldots, n$.

Problem 10

Suppose you are given a data set of survival times (in weeks) t_{ij}, $i = 1, ..., 15$, $j = 1, 2$, for two groups of animals, each containing 15 animals, suffering from cancer. One group ($j = 1$) is treated with a drug, the other group ($j = 2$) is treated with a placebo. The 13th, 14th, and 15th observations of the treatment group and the 15th observation of the control group are right censored.

Consider a proportional hazard model to analyse the data, with a hazard function $h(t_{ij}, z_j) = h_0(t_{ij}) \cdot \exp(1 + \beta \cdot z_j)$. For $j = 1$, the treatment group, set $z_j = 1$, for $j = 2$, the control group, set $z_j = 0$, and let the parameter β describe the differential effect. Let the survival times t_{ij} follow a Weibull density, i.e.

$$f(t_{ij}) = \rho\mu_j \cdot t_{ij}^{\rho-1} \exp(-\mu_j \cdot t_{ij}^{\rho})$$

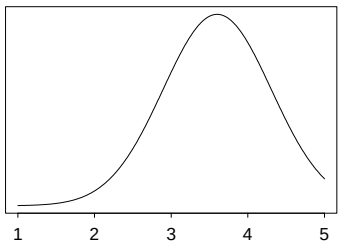
with shape parameter $\rho > 0$ and $\mu_j = \exp(1 + \beta \cdot z_j)$.

1. Determine the likelihood function under this survival model.

2. Give the full conditional distributions for β and ρ up to a normalizing constant assuming a constant prior density for both β and ρ. What kind of Markov Chain Monte Carlo method would you choose for updating the full conditional distributions?

3. Derive a formula for the median survival time m for the treatment group (j=1). How would you calculate m in a Markov Chain Monte Carlo scheme?

4. A data analysis using the model given above led to the following posterior marginal density for the parameter ρ.



Would you consider reanalysing the data under the simplifying assumption of an exponential survival time distribution? Reasoning?