FIRST YEAR EXAM

May 5, 2004

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. Draw a line through work you do not want graded; you do not need to erase.
6. You should attempt to work as many parts as feasible.
7. You have 3 hours to finish.
8. This is a closed book exam. No notes are permitted.
9. The Takehome Exam is due at 12 noon May 7th and should be handed in to Krista Moyle in Room 223 Old Chemistry.

1. Three random quantities X, Y and ϕ have a joint distribution with density $p(x, y \mid \phi)p(\phi)$. Conditional on ϕ, X and Y are bivariate normal with $\mathsf{E}(X \mid \phi) = \mathsf{E}(Y \mid \phi) = 0$, variances $\mathsf{Var}(X \mid \phi) = \mathsf{Var}(Y \mid \phi) = 1/\phi \equiv \sigma^2$ and correlation r, with density

$$p(x, y \mid \phi) = \frac{1}{2\pi} \frac{\phi}{\sqrt{1 - r^2}} \exp\left\{ -\frac{\phi}{2(1 - r^2)} \left(x^2 - 2rxy + y^2 \right) \right\}.$$

The precision ϕ has a Gamma distribution, $\mathsf{Ga}(1/2, 1/2)$, with density

$$p(\phi) \propto \phi^{-1/2} \exp(-\phi/2) \text{ for } \phi > 0.$$

Derive answers to the following:

(a) Find the density function $p(y)$. What is the name of the distribution of Y?

(b) What is the distribution of $\phi \mid Y$?

(c) State the distribution of $X \mid Y, \phi$.

(d) Using (b) and (c) above, derive the density function $p(x \mid y)$ (up to a normalizing constant that you do not need to fill in). What is the name of the distribution of $x \mid y$?

(e) Suppose now that $r = 0$ so that X and Y are uncorrelated conditional on ϕ. In this case, either prove or just directly state reasons for your answers to the following: Under the joint distribution $p(x, y)$ here,

i. are X and Y uncorrelated?

ii. are X and Y independent?

2. Suppose that we have $n+m$ observations on a manufacturing process (Y_i, X_i), for $i = 1, \ldots, n+m$. In the initial stage of the process for $i = 1, \ldots, n$, it is known that the model is

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i \tag{1}$$

where ϵ_i are independent, identically distributed (*i.i.d.*) normal errors with mean 0 and variance σ^2. Between X_n and X_{n+1}, the process under goes a phase change so that model for the second stage is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

for $i = n+1, \ldots, n+m$, with ϵ_i *i.i.d.* normal with the mean 0 and the same variance σ^2 as in stage one. The errors in the first and second phases are also assumed to be mutually independent.

(a) Write the likelihood function for α_0, α_1, β_0, β_1, and σ^2.

(b) Write maximum likelihood estimates (MLEs) $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ in terms of simple linear regression summary statistics (not matrix results). *You may quote any well-known results.*

(c) Let γ denote the value of X at which the phase change occurs, where the two regression lines in (1) and (2) intersect. Find the MLE of γ.

(d) Find the distribution of $(\hat{\alpha}_0 + \hat{\alpha}_1\gamma) - (\hat{\beta}_0 + \hat{\beta}_1\gamma)$ conditional on γ and σ^2.

(e) Sketch the construction of a $(1-\alpha)100\%$ confidence interval for γ by constructing a pivotal quantity that has a $F(1, n+m-4)$ distribution based on $\left[(\hat{\alpha}_0 + \hat{\alpha}_1\gamma) - (\hat{\beta}_0 + \hat{\beta}_1\gamma)\right]^2$. Recall that if t has a student t with ν degrees of freedom, that t^2 has a $F(1, \nu)$ distribution. *You do not need to simplify the expressions completely*

(f) Does such an interval exist?

Useful Information: Recall the simple linear regression model $W_i = \delta_0 + \delta_1 Z_i + \epsilon_i$ with *i.i.d.* errors with mean 0 and variance σ^2. Let

$$S_{zz} = \sum_i (Z_i - \bar{Z})^2 \qquad S_{zw} = \sum_i (Z_i - \bar{Z})(W_i - \bar{W})$$

denote the sum of squares of Z and sum of cross products of W and Z. The least squares estimates are

$$\hat{\delta}_1 = S_{zw}/S_{zz} \qquad \hat{\delta}_0 = \bar{W} - \hat{\delta}_1 \bar{Z}$$

with

$$\begin{aligned} \mathsf{Var}(\hat{\delta}_0) &= \sigma^2 \sum Z_i^2/(nS_{zz}) \\ \mathsf{Var}(\hat{\delta}_1) &= \sigma^2/S_{zz} \\ \mathsf{Cov}(\hat{\delta}_0, \hat{\delta}_1) &= -\sigma^2 \bar{Z}/S_{zz} \end{aligned}$$

3. Suppose we have a sample $(X_1, ..., X_n)$ with $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. We are interested in the two-sided test of $H_0 : \sigma = \sigma_o$ versus $H_1 : \sigma \neq \sigma_o$.

   (a) Find the maximum likelihood estmators of the unknown parameters under H_1 and under H_0.

   (b) Find the likelihood ratio test statistic, Λ, for testing H_0 versus H_1.

   (c) Show that the size α likelihood ratio test region for testing H_0 versus H_1 is equivalent to a region defined in terms of a statistic based on $T \equiv \sum_{i=1}^{n}(X_i - \bar{X})^2$.

   (d) How would you find critical values for a size α test of H_0 versus H_1 based on T?

4. Suppose a simple random sample of size n is taken from a finite population of N units, with $n < N$. Here each unit is equally likely to be selected. Let $Y = (Y_1, Y_2, \ldots, Y_N)$ be the values of the survey variable associated with each unit in the population. Before collecting the data, we model our beliefs about the unknown values of the survey variable using a normal distribution, $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Interest is in the population mean $\bar{Y} = \sum_{i=1}^{N} Y_i/N$.

   Let $\mathcal{S}$ denote the indices of units selected in the simple random sample, and let $\mathbf{Y}_\mathcal{S} = \{Y_i, i \in \mathcal{S}\}$ denote the realized values of the survey variable on the selected units. Since unit labels are not relevant for this problem, you may consider $\mathcal{S} \equiv \{1, \ldots, n\}$ and $\mathbf{Y}_\mathcal{S} \equiv (Y_1, \ldots Y_n)$.

   (a) Using a non-informative prior distribution, $\pi(\mu, \phi) = 1/\phi$, where $\phi \equiv 1/\sigma^2$, find the distribution of μ given collected data $\mathbf{Y}_\mathcal{S}$ and ϕ

   (b) Find the distribution of ϕ given $\mathbf{Y}_\mathcal{S}$, using the prior in (a).

   (c) Find the predictive distribution for any Y_i that was not in the sample (for $i \notin \mathcal{S}$) given the sampled $\mathbf{Y}_\mathcal{S}$ and ϕ.

   (d) Show that the predictive distribution for the population average, $\bar{Y}$, given $\mathbf{Y}_\mathcal{S}$ and ϕ (σ^2) is

$$N\left(\bar{Y}_\mathcal{S}, \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}\right)$$

   where $\bar{Y}_\mathcal{S} = \sum_{i \in \mathcal{S}} Y_i/n$. Hint: Write

$$\bar{Y} = \frac{n\bar{Y}_\mathcal{S} + (N-n)\bar{Y}_{\mathcal{S}^c}}{N}$$

   and obtain the distribution of the average of the unobserved values, $\bar{Y}_{\mathcal{S}^c} = \sum_{i \notin \mathcal{S}} Y_i/(N-n)$, given $\mathbf{Y}_\mathcal{S}$.

   (e) For n large, give an argument that $\bar{Y}$ given observed $\mathbf{Y}_\mathcal{S}$ is approximately normal,

$$N\left(\bar{Y}_\mathcal{S}, \left(1 - \frac{n}{N}\right)\frac{s^2}{n}\right)$$

   where

$$s^2 = \frac{\sum_{i \in \mathcal{S}}(Y_i - \bar{Y}_\mathcal{S})^2}{n-1}$$

   .

5. Let $U_1, \ldots, U_n$ be independent random variables with the uniform distribution on $(0, \theta]$ and consider three possible sequences of estimators of θ, each depending on the first n observations

$$R_n(\vec{U}) \equiv \max(U_1, ..., U_n) \qquad S_n(\vec{U}) \equiv \frac{2}{n} \sum_{i=1}^{n} U_i \qquad T_n(\vec{U}) \equiv \frac{n+1}{n} \max(U_1, ..., U_n)$$

(a) Derive the maximum likelihood estimator (MLE) for this problem. Is it one of R_n, S_n, or T_n?.

(b) Derive the method of moments (MOM) estimator for this problem. Is it one of R_n, S_n, or T_n?.

(c) Derive the Bayesian posterior mean for a Pareto prior distribution with $\pi(\theta > t) = t^{-1}$ for $t \geq 1$. Is it one of R_n, S_n, or T_n?.

(d) Which (if any) of these estimators (MLE, MOM, Bayes) is unbiased?

(e) Prove that each of (MLE, MOM, Bayes) is *consistent* in the sense that it converges almost surely to θ.

6. Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{Bin}(n, p)$ where both n and p are unknown parameters.

   (a) Find the likelihood function $L(n, p)$.

   (b) Find $\hat{p}(n)$, the maximum likelihood estimate of p for a given value of n.

   (c) Define the profile likelihood

   $$L^*(n) = \max_{0 \le p \le 1} L(n, p) = L(n, \hat{p}(n))$$

   where $\hat{p}(n)$ is the MLE of p from above. Find $L^*(n)$.

   (d) Find

   $$L^*(\infty) \equiv \lim_{n \to \infty} L^*(n)$$

   Hint: use the Poisson approximation to the Binomial. Suppose $X \sim \text{Bin}(n, p)$. Define $\lambda = np$ with $Y \sim \text{Poi}(\lambda)$. Then, if n is large and p is small, $\Pr[X = x] \approx \Pr[Y = x]$.

   (e) Fix a prior density π_0 for (n, p) and a small positive number $\epsilon > 0$. Define a class of priors Γ,

   $$\Gamma \equiv \{\pi_k(n, p) = (1 - \epsilon)\pi_0(n, p) + \epsilon\ \delta_{k,\hat{p}}(n, p) : k = 0, 1, \ldots\}$$

   where $\delta_x(y)$ is a distribution degenerate at x, i.e., $\delta_{k,\hat{p}(k)}(n, p) = 1$ if $n = k, p = \hat{p}(k)$, and is zero otherwise. (Yes, this prior depends on the data, but do not worry.) Γ is supposed to be a class of priors "close" to the fixed prior π_0.

   For a prior π_k in Γ, the posterior for n (marginally) is a mixture

   $$\pi_k(n \mid X_1, \ldots, X_m) = (1 - \alpha)\pi_0(n \mid X_1, \ldots, X_m) + \alpha\delta_k(n)$$

   where $\pi_0(n \mid X_1, \ldots, X_m)$ is the posterior distribution for n under prior π_0. Give an expression for α or $\alpha/(1 - \alpha)$ in terms of ϵ and the marginal distributions of the data under $\pi_0(n, p)$ and $\delta_{k,\hat{p}(k)}(n, p)$.

   (f) Find the posterior expectation of n under prior π_k and then show that

   $$\sup_{\pi \in \Gamma} \mathsf{E}_\pi[n | X_1, \ldots, X_m] = \infty.$$

   Hint: Show that $\alpha > 0$ and consider the limit as $k \to \infty$. Part (d) will be useful here. This problem illustrates the difficulty in estimating an unknown Binomial parameter n.

First Year Exam - Takehome

Turn into Krista in Room 223 by Noon May 9, 2003

Scientists are interested in the Earth's temperature change since the last glacial maximum, about 20,000 years ago. The first study to estimate the temperature change was published in 1980, and estimated a change of -1.5 degrees C, ± 1.2 degrees C in tropical sea surface temperatures. The negative value means that the Earth was colder then than now. Since 1980 there have been many other studies. Different studies use different measurement techniques, or proxies. Some proxies can be used over land, others over water. The proxies are

proxy	code
"Mg/Ca"	1
"alkenone"	2
"Faunal"	3
"Sr/Ca"	4
"del 180"	5
"Ice Core"	6
"Pollen"	7
"Noble Gas"	8

Each study reports an estimate `deltaT`, a standard deviation of that estimate `sdev`, the `proxy` used (coded 1 to 8), whether it was a terrestrial or marine study (`T/M`), which is coded as 0 for Terrestrial, 1 for Marine, and the latitude at which data were collected (`latitude`). Data are in the file `http://www.isds.duke.edu/info/FYE/temp`

Using data from 63 studies, conduct an appropriate statistical analysis of the data addressing questions below and summarize your findings in a typed two page report and an executive summary (the summary should be suitable for President Bush's science advisor). You may include a supplemental appendix of no more than five pages with any other key figures, output or more technical expressions to support your analysis that are not included in the main text. All figures and computer output should be clearly labeled and annotated. Any results in the appendix should be referenced in the body of the report. You should be sure to address the following issues, but they should not be considered exclusively.

1. Do estimates vary systematically by proxy?
2. Do terrestrial estimates differ systematically from marine estimates?
3. Do estimates vary systematically by latitude?
4. Can we combine the studies to get a better estimate of the overall temperature change? You must figure out what "overall" means.
5. Are temperatures changing?