FIRST YEAR EXAM

May 7, 2003

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 for ISDS students; MS students should skip Problem 5.
7. You have 3 hours to finish.
8. The Takehome Exam is due at 12 noon May 9th and should be handed in to Krista Moyle in Room 223 Old Chemistry.

1. Suppose $X_1, X_2, ...$ are independent and identically distributed non-negative random variables with

$$\begin{aligned} \mathsf{E}[X_n] &= a \\ \mathsf{Var}[X_n] &= b^2 \end{aligned}$$

Let N be a non-negative integer-valued random variable which is independent of $X_1, X_2, ...$, with finite mean and variance,

$$\begin{aligned} \mathsf{E}[N] &= c \\ \mathsf{Var}[N] &= d^2. \end{aligned}$$

Let $S_0 = 0, S_1 = X_1, S_2 = X_1 + X_2, ...$, and let $Y = S_N = \sum_{i=1}^{N} X_i$.

(a) Find $\mathsf{E}[Y|N]$ and $\mathsf{E}[Y^2|N]$.

(b) Find $\mathsf{E}[Y]$, $\mathsf{E}[Y^2]$, and $\mathsf{Var}[Y]$.

(c) Show that for any $\alpha \in \mathbb{R}$:

$$\mathsf{E}[e^{-\alpha Y}] = G(F(\alpha))$$

where

$$\begin{aligned} F(\alpha) &= \mathsf{E}[e^{-\alpha X_n}] \\ G(\beta) &= \mathsf{E}[\beta^N], \quad \beta \in [0,1]. \end{aligned}$$

2. When researchers need to ask sensitive questions that they fear respondents may not answer truthfully, they sometimes use a technique called randomized response.

   Suppose you randomly sample $n = 100$ students. Instead of asking them directly, "Have you violated the honor code of Duke?", which students may not want to answer truthfully, use the following process. Give each student a fair coin, and ask him or her to flip it without showing you the result. Instruct the student to do the following based on their flip of the coin:

   - When the coin comes up heads, answer honestly the question, "Have you ever violated the honor code of Duke?"
   - When the coin comes up tails, answer honestly the question, "Is your mother's birth month June, July, or August?"

   With this scheme, the interviewer cannot know whether a "yes" answer means the respondent violated the honor code or means the respondent's mother has a birthday in the summer.

   Let p be the population proportion of Duke honor code violators, and let r be the marginal probability of responding "yes". Assume that respondent's mothers were born in each month with equal probability (ignore differences in length or season of month), and that the coin has a 50-50 chance of coming up heads. Ignore finite population considerations. Let Y dnote the total number of "yes" respondents out of the 100 students.

   (a) Derive a formula for r as a function of p under the assumptions described above.

   (b) Specify the distribution for Y, the number of "yes" answers, and find the maximum likelihood estimator of p.

   (c) Derive an unbiased estimator of p as a function of Y and compute its variance. Will this estimator always take on values in the parameter space?

3. Consider the simple linear regression model with one predictor,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, \ldots, n \tag{1}$$

$$\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_{ee}) \tag{2}$$

with independent, identically distributed normal errors with mean 0 and variance σ_{ee}. Suppose that we are unable to observe X_i directly, but instead observe W_i, a noisy version of X_i,

$$W_i = X_i + u_i \tag{3}$$

where $u_i \overset{i.i.d.}{\sim} N(0, \sigma_{uu})$ represents measurement error in X_i. For $i = 1, \ldots, n$, assume that the vector

$$\begin{pmatrix} X_i \\ \epsilon_i \\ u_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{ee} & 0 \\ 0 & 0 & \sigma_{uu} \end{bmatrix} \right). \tag{4}$$

(a) Find the joint distribution of (Y_i, W_i) given parameters $(\mu_x, \beta_0, \beta_1, \sigma_{ee}, \sigma_{xx}, \sigma_{uu})'$ using the specifications given by equations (1 – 4).

(b) Let $\hat{\gamma}_1$ denote the ordinary least squares regression coefficient computed from the regression of Y on W,

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (W_i - \bar{W})(Y_i - \bar{Y})}{\sum_{i=1}^n (W_i - \bar{W})^2}.$$

Find $\mathsf{E}[\hat{\gamma}_1]$ and show that as an estimator of β_1, the regression coefficient in (1), that $\hat{\gamma}_1$ is biased toward zero.

Useful Information: Recall that if $Z = (Z_1, Z_2)'$ has a normal distribution, with partitioned mean and covariance

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

that $Z_1 | Z_2$ is normal with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \mu_2)$ and covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

4. Suppose $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with density

$$f(x|\theta) = h(x) \exp\{\theta\ T(x) - \chi(\theta)\}, \quad \theta \in (\underline{\theta}, \overline{\theta}).$$

In what follows, assume any needed regularity conditions.

(a) Show that $\mathsf{E}[T(X)|\theta] = \chi'(\theta)$

(b) What is the maximum likelihood estimate of θ based on a sample of size n?

(c) Obtain the uniformly most powerful level α test of $H_0 :\ \theta \leq \theta_0$ versus the alternative $H_1 :\ \theta > \theta_0$

(d) If we take a prior distribution for θ of the form

$$f(\theta|n_0, \mu_0) = c(n_0, \mu_0) \exp\{n_0\mu_0\ \theta - n_0\chi(\theta)\}$$

what is the $\mathsf{E}[\chi'(\theta)]$? For convenience, you may assume that $\lim_{\theta\to\underline{\theta}} f(\theta|n_0, \mu_0) = \lim_{\theta\to\overline{\theta}} f(\theta|n_0, \mu_0) = 0$.

(e) Obtain the posterior mean of $\chi'(\theta)$, $\mathsf{E}[\chi'(\theta)|X_1, \ldots, X_n]$ for a sample of size n.

5. Let U_n be independent random variables with the uniform distribution on $(0, 1]$. Which of the following sequences of random variables converge almost-surely as $N \to \infty$? Why? Cite any relevant theorems etc.

(a) $\displaystyle\sum_{n=1}^{N} \mathbf{1}_{(0,1]}\left(n\sqrt{U_n}\right)$

(b) $\displaystyle\frac{1}{N}\sum_{n=1}^{N} \log U_n$

(c) $\displaystyle\frac{1}{N}\sum_{n=1}^{N} (-1)^n/U_n$

(d) $\displaystyle\prod_{n=1}^{N} (U_n + 1/2)$

6. Biologists want to know the rate λ, measured in number per million years, at which mutations occur. Panama rose from the ocean about three million years ago, separating the Atlantic and Pacific oceans and affording a unique opportunity to study λ. Let S be a species that existed in the ocean three million years ago. The rise of Panama divided the population of S into two subpopulations, one in the Atlantic and one in the Pacific. Each subpopulation followed its own evolutionary path. Gradually the two subpopulations diverged to become two different species today, say S^A and S^P. The number of genetic differences between S^A and S^P can be used to estimate λ.

   Suppose that the number of mutations in each subpopulation in any million year period follows a Poisson(λ) distribution and that mutations arise independently of each other. For simplicity, assume that no more than 1 mutation occurs at a given location in the DNA and that locations of mutations in the two species, S^A and S^P, are different from each other. Let N^A and N^P be the number of mutations in the two subpopulations since the rise of Panama. While we cannot observe N^A and N^P individually, we can observe the number of locations in the DNA at which S^A and S^P differ. I.e., we observe $N = N^A + N^P$.

   (a) What is the distribution of N?

   (b) Suppose there were k species $S_1, \ldots, S_k$ that split into two species, yielding mutation numbers N_i for $i = 1, \ldots, k$. Write down a model for $N_1, \ldots, N_k$ given λ and find the maximum likelihood estimate of λ.

   (c) Call the theory above H_1. An alternate theory, H_2, holds that there were two events that split species into subpopulations. One event, the rise of Panama, occurred three million years ago; the other event occurred two million years ago. Some species were split by the first event, others by the second. In fact, we don't know which species were split at which time. For simplicity, suppose $k = 2$. Show, by adding parameters to your model, that H_1 is nested within H_2, and find the likelihood ratio statistic for comparing H_1 to H_2.

   (d) A well known theorem says that likelihood ratio statistics are asymptotically distributed as χ^2_ν where ν is the number of extra parameters in the more general theory. Does that theorem apply to this example? Explain.

First Year Exam - Takehome

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

Subjects from the general population of Central Prison, Raleigh, NC, volunteered for an experiment involving an "isolation" experience. (Those convicted of felon offenses are housed in this facility.) The experimental treatment exposed inmates to combined sensory restriction and suggestion. The intent was to reduce the psychopathic deviant T scores (Pd T), Scale 4 of the Minnesota Multiphasic Personality Inventory (MMPI) test. Briefly, the three treatments consisted of:

1. Four hours of sensory restriction plus a 15 minute "therapeutic" tape advising that professional help is available.

2. Four hours of sensory restriction plus a 15 minute "emotionally neutral" tape on training hunting dogs.

3. Four hours of sensory restriction but no taped message.

Forty-two subjects were assigned to one of the three treatment groups (for a total of 14 in each treatment group). For each subject the MMPI was administered before and after the experimental treatment. Pre-treatment and post-treatment values of Pd T scores are given below for the 42 individuals.

Pre-Trt1	Post-Trt1	Pre-Trt2	Post-Trt2	Pre-Trt3	Post-Trt3
67	74	88	79	86	90
86	50	79	81	53	53
64	64	67	83	81	102
69	76	83	74	69	67
67	64	79	76	81	76
79	81	76	69	76	81
67	74	71	71	74	69
67	50	67	75	60	60
69	60	69	64	67	69
57	57	67	64	86	83
76	62	67	64	86	107
90	76	74	71	74	71
71	71	81	74	71	71
93	76	81	64	71	81

Conduct an appropriate statistical analysis of the data and summarize your findings on the effectiveness of the experimental treatments in a typed two page report. 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. 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.

- Are the groups the same before treatment?
- Were the treatments effective?
- Do any post-treatment groups differ?
- Is treatment 1 the best?
- MMPI scores follow a bell-shaped distribution and mean of 50 and standard deviation of 10 for the general population. Is this the case in this sub-population? Consider what impact this might have on your analysis.
- Pd T scores are often higher for blacks than whites; is the lack of race information an issue?

This is an open book exam. You may refer to your notes, and published texts and references as needed, however, all work must be your own.

Some Background on the MMPI Pd T Scores:

The Pd T scale was originally developed to identify patients diagnosed as psychopathic personality, asocial or amoral type. General social maladjustment and the absence of strongly pleasant experiences are assessed by the 50 items included in Scale 4. Scores on Scale 4 tend to be related to age, with adolescents and college students often scoring in a T-score range of 55 to 60. Black respondents have also been reported to score higher than white persons on Scale 4. Scale 4 can be thought of as a measure of rebelliousness, with higher scores indicating rebellion and lower scores indicating an acceptance of authority and the status quo. High scorers are very likely to be diagnosed as having some form of personality disorder, but are unlikely to receive a psychotic diagnosis. Low scorers are generally described as conventional, conforming, and submissive.