

FIRST YEAR EXAM

Monday May 5, 2008; 9:00 – 12:00am

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.
Do not erase anything— just draw a line through work you do not want graded.
6. You have 3 hours to finish.
7. This is a closed book exam. No notes are permitted.
A page with common p.d.f. and p.m.f. formulas is attached.
8. The Take-Home practical can be picked up from the front of the room immediately after dropping of this written exam and is required to be handed in by 5:00 PM on Tuesday May 6th, 2008 to Karen Herndon in 223C Old Chemistry.

1. Let X_1, \dots, X_n be IID with density

$$f(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} \mathbf{1}(\nu \leq x) \quad \theta > 0, \nu > 0,$$

where $\mathbf{1}(\cdot)$ is the indicator function.

- (a) Find a 2-dimensional sufficient statistic for the model.
- (b) Suppose θ is a known constant. Find the MLE for ν .
- (c) Now suppose that θ is unknown but $\nu = 1$. Find the score function $S(\theta)$, the derivative of the log-likelihood, and determine its asymptotic distribution at the true value θ_0 . Carefully state any theorems that you use.
- (d) Suppose $\nu = 1$. Find the MLE for θ and determine its asymptotic distribution. Carefully state any theorems that you use.
- (e) Suppose $\nu = 1$. Find the asymptotic distribution of the MLE estimator of $\exp[-\theta]$.

2. Let $Z_i \stackrel{IID}{\sim} N(0, 1)$ and $\xi \sim \text{Bi}(1, \frac{1}{2})$ be independent standard Normal random variables for $i = 1, 2$ (with CDF $\Phi(z) = P[Z_i \leq z]$) and a Bernoulli random variable.

Define

$$X = \frac{\sqrt{2}}{2}(Z_1 + Z_2) \quad Y = (\xi) Z_1 + (1 - \xi) Z_2$$

(so Y is Z_1 with probability $\frac{1}{2}$ and Z_2 with probability $\frac{1}{2}$).

- (a) Find the indicated CDFs (distribution functions).

$$F_X(x) = P[X \leq x] =$$

$$F_Y(y) = P[Y \leq y] =$$

- (b) For $z \in \mathbb{R}$, find the indicated conditional probabilities:

$$P[X \leq x \mid Z_1 = z] =$$

- (c) For $z \in \mathbb{R}$, find the indicated conditional expectations:

$$E[X \mid Z_1 = z] =$$

$$E[Y \mid Z_1 = z] =$$

- (d) Find:

$$E[Z_1^2 \mid Y = y] =$$

- (e) Let $\{X_n\}$ be a sequence of random variables that converge almost surely to the random variable $X_\infty \equiv 0$. Show that $\cos(X_n) \rightarrow 1$ in L_p for every $1 \leq p < \infty$, but show that it does *not* follow that $\cos(X_n) \rightarrow 1$ in L_∞ .

3. If a person is suspected of having an undesirably high blood alcohol level in his blood a rapid test is carried out. This test has only a 75% chance of being correct: i.e. of giving a positive result conditioned on the alcohol level is high, or of giving a negative result conditioned on the level being low. If the rapid test gives a positive result the person undergoes a second, more reliable test. If the alcohol content was actually high at the time of the first test, the second has a 90% chance of detecting it. If the content was low at the time of the first test then the second test never gives a false result. These results on the second test are independent of those on the first.

Suppose 20% of all suspects actually have high levels of blood alcohol. Answer the following questions.

- (a) What proportion of suspects will have a second test that does not detect high levels of alcohol?
- (b) What is the posterior probability that such a person does in fact have a high levels?
- (c) What proportion of tested persons will not have a second test?
- (d) If a suspect does not have the second test, what is the posterior probability that his blood alcohol level is in fact high?

4. Recall the definition of stochastic order, i.e., Y is stochastically larger than X if $F_Y(c) \leq F_X(c), \forall c$. A family of distributions $F_\theta(\cdot)$ is stochastically increasing (decreasing) in θ if $F_\theta(c) \downarrow (\uparrow)$ in θ for every c .
- (a) Suppose that a strictly positive random variable Y has a scale parameter distribution, i.e., for $\theta > 0$, $P_\theta(Y \leq y) = F(y/\theta)$ where F is any cdf. Show that this family is stochastically increasing in θ .
- (b) Suppose F is any cdf. Consider the family, $F^\theta(\cdot)$ where $\theta > 0$. What can you say about stochastic order for this family?
- (c) Recall the association inequality, i.e., if $X \sim f$, f a pdf over \mathcal{X} a subset of R^1 and g_1, g_2 are non-decreasing functions, then $E_f(g_1(X)g_2(X)) \geq E_f(g_1(X))E_f(g_2(X))$. So, suppose f is a density and g is any non-negative increasing function over \mathcal{X} . Create the new density $h(x) = f(x)g(x) / \int_{\mathcal{X}} f(x)g(x)dx$. (Assume the denominator integral exists.) Use the inequality to show that h is stochastically larger than f .
- (d) Suppose F and G are two cdf's such that F is stochastically larger than G . Let $w(x)$ be any non-decreasing function on R^1 . Assuming the expectations exist, show that $E_F(w(X)) \geq E_G(w(X))$. (Hint: One way to show this is to prove it for suitable indicator functions.)

5. Suppose the following linear model is true:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

where $i = 1, \dots, n$ and the $\{\epsilon_i\}$ are independent. Let \mathbf{y} denote the $n \times 1$ vector of responses, \mathbf{x}_1 denote the $n \times 1$ vector containing the values of the first predictor, and \mathbf{x}_2 denote the $n \times 1$ vector containing all values of the second predictor.

- a. What is the distribution of the least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$?
- b. Suppose the response and predictors are centered by subtracting the mean. $\mathbf{y}_c = \mathbf{y} - \bar{y}$, and \mathbf{x}_{jc} is an $n \times 1$ vector of centered predictors, where $j = 1, 2$. Consider the following assumption:

$$E(\mathbf{y}_c) = \alpha_1 \mathbf{x}_{1c} + \alpha_2 \mathbf{x}_{2c}$$

Is the assumption true? Describe the relationship between $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ and $\boldsymbol{\beta}$.

- c. Suppose you transform the response by squaring it, so that \mathbf{y}^2 is the vector containing the squared elements of \mathbf{y} . You transform the covariate vectors in the same way, yielding \mathbf{x}_1^2 and \mathbf{x}_2^2 . You plan to fit the following linear model for \mathbf{y}^2 :

$$\mathbf{y}^2 = \gamma_0 + \mathbf{x}_1^2 \gamma_1 + \mathbf{x}_2^2 \gamma_2 + \boldsymbol{\epsilon}^*$$

Are you correct if you assume that $\boldsymbol{\epsilon}^*$, the vector of random errors, has mean 0? Explain.

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

- (a) Find the observed information

$$\mathcal{I}(x_{1:n}|\pi) = -\frac{\partial^2}{\partial \pi^2} \log(p(x_{1:n}|\pi)),$$

and deduce the expected (Fisher) information function $I(\pi) = E(\mathcal{I}(x_{1:n}|\pi))$.

- (b) Use the resulting reference prior $p(\pi) \propto I(\pi)^{1/2}$ to deduce the (reference) posterior $p(\pi|x_{1:n})$.
- (c) What is the predictive probability of success at the next trial, $Pr(x_{n+1} = 1|x_{1:n})$?
- (d) Find $\hat{\pi}$, the MLE of π . Using part (6a), show that the asymptotic normal approximation to $p(\pi|x_{1:n})$ is $N(\hat{\pi}, s^2)$ where $s^2 = \hat{\pi}(1 - \hat{\pi})/n$.
- (e) Show that, whatever the data may be, $s \leq (4n)^{-1/2}$.
- (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 $n = 1000$. Using the results in (d), show how such statements may be interpreted in terms of approximate 95% highest posterior density regions for π .

7. FYE '08 Take-Home Problem

Turn in solution to Karen Herndon in Room 223 Old Chemistry by 5pm on Tuesday May 6.

What do Barbie dolls, food wrap, edamame, and spermicides have in common? And what do they have to do with low sperm counts, precocious puberty, and breast cancer? "Everything," say those who support the notion that hormone mimics are disrupting everything from fish gender to human fertility. "Nothing," counter others who regard the connection as trumped up, alarmist chemophobia. The controversy swirls around the significance of a number of substances that behave like estrogens and appear to be practically everywhere—from plastic toys to topical sunscreens.

Estrogens are a group of hormones produced in both the female ovaries and male testes, with larger amounts made in females than in males. They are particularly influential during puberty, menstruation, and pregnancy, but they also help regulate the growth of bones, skin, and other organs and tissues. In general, they have a strong effect of endocrine function by disrupting these functions.

Over the past 10 years, many synthetic compounds and plant products present in the environment have been found to affect hormonal functions in various ways. Those that have estrogenic activity have been labeled as environmental estrogens. There is increasing concern that chemicals in the environment referred to as environmental estrogens may be causing adverse effects through endocrine disruption.

Hence, there is a need for new approaches for screening chemicals for endocrine disrupting effects. The rat uterotrophic bioassay provides one approach for identifying agonists or antagonists of estrogen. An estrogen antagonist is a compound that blocks the binding of estrogen and so blocks the action of estrogen. An estrogen agonist is a compound that enhances the action of estrogen.

Rats in this study are either immature or have their ovaries removed and therefore do not produce estrogen. The point of the study is to use the rats as an assay to test the effect of estrogen agonists and antagonists on a particular hormonal response, the weight of the uterus. This is done by varying the amount of the agonist or antagonist given to the rat. The response is the weight of the uterus, with uterus weight expected to exhibit an increasing dose response trend for chemicals acting as estrogen agonists and with estrogen antagonists acting to block such estrogen effects. It is expected that the uterus gets heavier with the increase of estrogen agonist dose.

The basic design randomizes female rats to treatment groups, with groups consisting of a control group and several groups having increasing doses of the test agent. An international multilaboratory study was conducted to compare the results of the rat uterotrophic bioassay using a known estrogen agonist (EE) and a known estrogen antagonist (ZM). The main goal of the study was to assess whether the results were consistent across the laboratories.

Variables

Protocol	A = immature female rats dosed by oral gavage (3 days) B = immature female rats dosed by injection (3 days) C = adult ovariectomized female rats dosed by injection (3 days) D = adult ovariectomized female rats dosed by injection (7 days)
Uterus	Uterus weight (mg)
Weight	Body weight of rat (g)
EE	Dose of estrogen agonist, EE in mg/kg/day
ZM	Dose of estrogen antagonist, ZM in mg/kg/day
Lab	Laboratory at which assay was conducted
Group	Lab replicate group (6 rats were used per group)

Questions:

- Select an appropriate model for the data and justify your choice.
- Is the uterotrophic bioassay successful at identifying estrogenic effects of EE and anti-estrogenic effects of ZM? Do some labs fail to detect such effects? At what dose level of EE is there a change relative to the control and does this level vary across labs?
- Does the dose response vary across labs? If so, are there certain labs that are outliers?
- Do the protocols differ in their sensitivity to detecting estrogenic and anti-estrogenic effects? If so, is there one protocol that can be recommended?

Please be rigorous in providing a full justification for each of your answers, including all relevant statistical details, calculations and results. Report the results in a manner interpretable by a toxicologist interested in the study conclusions.

The data can be found at <http://www.stat.duke.edu/~sayan/bioassay.txt>

Beta	$\text{Be}(\alpha, \beta)$	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$x \in (0, 1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Binomial	$\text{Bi}(n, p)$	$f(x) = \binom{n}{x} p^x q^{(n-x)}$	$x \in 0, \dots, n$	np	npq ($q = 1 - p$)
Exponential	$\text{Ex}(\lambda)$	$f(x) = \lambda e^{-\lambda x}$	$x \in \mathbb{R}_+$	$1/\lambda$	$1/\lambda^2$
Gamma	$\text{Ga}(\alpha, \lambda)$	$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$x \in \mathbb{R}_+$	α/λ	α/λ^2
Geometric	$\text{Ge}(p)$	$f(x) = p q^x$	$x \in \mathbb{Z}_+$	q/p	q/p^2 ($q = 1 - p$)
		$f(y) = p q^{y-1}$	$y \in \{1, \dots\}$	$1/p$	q/p^2 ($y = x + 1$)
HyperGeo.	$\text{HG}(n, A, B)$	$f(x) = \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}}$	$x \in 0, \dots, n$	nP	$nP(1-P) \frac{N-n}{N-1}$ ($P = \frac{A}{A+B}$)
Logistic	$\text{Lo}(\mu, \beta)$	$f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$	$x \in \mathbb{R}$	μ	$\pi^2 \beta^2 / 3$
Log Normal	$\text{LN}(\mu, \sigma^2)$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log x - \mu)^2 / 2\sigma^2}$	$x \in \mathbb{R}_+$	$e^{\mu + \sigma^2 / 2}$	$e^{2\mu + \sigma^2} (e^{\sigma^2 - 1})$
Neg. Binom.	$\text{NB}(\alpha, p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha q^x$	$x \in \mathbb{Z}_+$	$\alpha q / p$	$\alpha q / p^2$ ($q = 1 - p$)
		$f(y) = \binom{y-1}{y-\alpha} p^\alpha q^{y-\alpha}$	$y \in \{\alpha, \dots\}$	α / p	$\alpha q / p^2$ ($y = x + \alpha$)
Normal	$\text{No}(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2 / 2\sigma^2}$	$x \in \mathbb{R}$	μ	σ^2
Pareto	$\text{Pa}(\alpha, \epsilon)$	$f(x) = \alpha \epsilon^\alpha / x^{\alpha+1}$	$x \in (\epsilon, \infty)$	$\frac{\epsilon \alpha}{\alpha-1}$	$\frac{\epsilon^2 \alpha}{(\alpha-1)^2 (\alpha-2)}$
Poisson	$\text{Po}(\lambda)$	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{Z}_+$	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	$\text{Un}(a, b)$	$f(x) = \frac{1}{b-a}$	$x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Weibull	$\text{We}(\alpha, \beta)$	$f(x) = \alpha\beta x^{\alpha-1} e^{-\beta x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$