FIRST YEAR EXAM

Monday May 7, 2007; 9:30 – 12:30

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 Exam is due at 10:00 AM on Wednesday May 9th.
   It should be handed in to Krista Moyle in Room 223 Old Chemistry.

1. Let $Y_i \overset{\text{iid}}{\sim} \mathsf{No}(\mu, 1)$, with $\mu \in \{0, 1\}$ known to be zero or one. Upon observing a sample $\mathbf{y} = \{y_i\}$ of size n we must choose one of three possible actions, labeled 1, 2, 3. If μ is the population mean and if we make decision d, we incur a loss $L(\mu, d)$ which we would like to be as small as possible. The loss function is given by

$L(\mu, d)$	$d = 1$	$d = 2$	$d = 3$
$\mu = 0$	0	1	5
$\mu = 1$	5	1	0

Evidently for large value of $\mathbf{y}$ (suggesting that $\mu = 1$) we should make choice $D(\mathbf{y}) = 3$, while for small values of $\mathbf{y}$ (suggesting that $\mu = 0$) we should make choice $D(\mathbf{y}) = 1$, and for intermittant values perhaps $D(\mathbf{y}) = 2$ is best, suggesting a decision function of the form

$$D(\mathbf{y}) = \begin{cases} 1 & \text{if} \quad \bar{y} \le b \\ 2 & \text{if } b < \bar{y} \le c \\ 3 & \text{if } c < \bar{y} \end{cases}$$

for some numbers $-\infty < b \le c < \infty$. For prior distribution

$$\pi(\mu = 0) = \frac{1}{2} = \pi(\mu = 1),$$

find the values of b and c for which the decision procedure $D(\mathbf{y})$ above minimizes the expected loss. Simplify your answers as much as possible.

2. If $\{X_j\} \overset{\text{iid}}{\sim} \mathsf{Po}(\lambda)$ are independent Poisson-distributed random variables, then the Central Limit Theorem asserts that the partial sum $S_n := \sum_{j=1}^{n} X_j$ and sample average $\bar{X}_n := S_n/n$ are asymptotically normally distributed, with means and variances you can compute easily (they'll depend on λ and n).

   (a) For any smooth (say, twice continuously differentiable) function g with non-vanishing derivative $g'(x) \neq 0$, the random variables

   $$Y_n := g\big(\bar{X}_n\big)$$

   are also approximately normally-distributed for large n (you don't have to prove that). Use a Taylor expansion of $g(\cdot)$ to find the approximate mean and variance of Y_n, in terms of n, λ, and g and its derivatives. Try to make your answers **correct to order** $1/n$.

   (b) Find a power $p \neq 0$ for which the function $g(x) := |x|^p$ leads to random variables $Y_n := g(\bar{X}_n)$ whose (approximate) variance doesn't depend on λ. Such a "variance stabilizing transformation" is sometimes used to make regression models behave better. Give the approximate mean and variance of $Y_n = |\bar{X}_n|^p$.

3. Suppose the following linear model relates responses y_i and (non-random) predictors x_{ij}:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathsf{No}(0, \sigma^2) \tag{1}$$

for $1 \leq i \leq n$. Let $\mathbf{y} = \{y_i\}$ denote the $n \times 1$ vector of responses, $\mathbf{x}_1 := \{x_{i1}\}$ denote the $n \times 1$ vector containing the values of the first predictor, and $\mathbf{x}_2 := \{x_{i1}\}$ denote the $n \times 1$ vector containing all values of the second predictor. You may assume that $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{1}_n$ (the $n \times 1$ vector of ones) are linearly independent, so the $n \times 3$ *design matrix*

$$\mathbf{X} := [\mathbf{1}_n \ \mathbf{x}_1 \ \mathbf{x}_2]$$

has rank three.

(a) What is the distribution of the least-squares estimate $\hat{\beta}$ of $\boldsymbol{\beta} := (\beta_0, \beta_1, \beta_2)'$?

(b) Construct centered response and predictor vectors by subtracting their means— *i.e.*, set $\mathbf{y}_c := \mathbf{y} - \bar{y}\mathbf{1}_n$ and $\mathbf{x}_{cj} := \mathbf{x}_j - \bar{x}_j\mathbf{1}_n$. True or false: There exist numbers $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ such that

$$\mathsf{E}(\mathbf{y}_c) = \alpha_0 + \mathbf{x}_{c1}\alpha_1 + \mathbf{x}_{c2}\alpha_2.$$

If true, find $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$ in terms of $\boldsymbol{\beta}$; if false, show why.

(c) Now transform the response by squaring it, so that $\mathbf{y}^2$ is the vector containing the squared elements of $\mathbf{y}$. Transform the covariate vectors in the same way, yielding $\mathbf{x}_1^2$ and $\mathbf{x}_2^2$. You wish to fit a linear model for $\mathbf{y}^2$, in terms of $\mathbf{x}_1^2$ and $\mathbf{x}_2^2$. True or false: There exist numbers $\gamma_0, \gamma_1, \gamma_2 \in \mathbb{R}$ such that

$$\mathsf{E}(\mathbf{y}^2) = \gamma_0 + \mathbf{x}_1^2\gamma_1 + \mathbf{x}_2^2\gamma_2.$$

If true, find $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)'$ in terms of $\boldsymbol{\beta}$; if false, show why.

4. Infectious diseases are sometimes modelled with a so called *SIR* model (the letters stand for Susceptible, Infected, and Recovered). People begin in class S, then possibly migrate to class I (*i.e.*, become infected), and then to class R (*i.e.*, recover); no other transitions are possible. In a simple version of the model, the i^{th} individual begins in class S, waits a random amount of time $T_i \sim \mathsf{Ex}(\lambda)$ before migrating to class I, then waits another random amount of time $U_i \sim \mathsf{Ex}(\mu)$ before migrating to class R, with all the exponentially-distributed random variables $\{T_i, U_i\}$ independent.

   (a) For each $t > 0$, find the C.D.F. $\mathsf{P}[T_i \leq t]$.

   (b) Let N denote the number of Susceptibles at time 0 and let X_t be the number of these who become infected by time t. Find the probability distribution of X_t.

   (c) Let W_1 be the length of time until the *first* of those people becomes infected, *i.e.*, $W_1 \equiv \min\{T_1, \ldots, T_N\}$. Find the probability distribution for W_1.

   (d) Let W_N be the length of time until the *last* of those people becomes infected, *i.e.*, $W_N \equiv \max\{T_1, \ldots, T_N\}$. Find the probability density function for W_N.

   (e) Let $Y_i = T_i + U_i$ be the total amount of time the i^{th} Susceptible waits before joining class R. Find the probability distribution of Y_i under the (simplifying) assumption $\mu = \lambda$.

5. The random variables $X_i \overset{\text{iid}}{\sim} \mathsf{Un}(0, \theta)$ are independent, all uniformly distributed on some range $0 < x < \theta$, with partial products

$$Y_n := \prod_{i=1}^{n} X_i = X_1 \times X_2 \times \cdots \times X_n.$$

(a) What is the expectation $\mathsf{E}[Y_n]$?

(b) Find the limit $\lim_{n\to\infty} Y_n$, as a function of $\theta > 0$. For which values of θ is it finite?

(c) Show that Lebesgue's dominated convergence theorem applies to $\{Y_n\}$ if $\theta = 1$, but prove that it does *not* apply if $\theta = 2.5$.

(XC) Does Lebesgue's dominated convergence theorem apply to $\{Y_n\}$ if $\theta = 1.5$? Why?

6. The random variable X takes one of five possible values, with probability mass function $f(x) = \mathsf{P}[X = x]$ given by either $f = f_0$ or $f = f_1$ from the following table:

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
$f_0(x) =$	0.50	0.05	0.20	0.10	0.15
$f_1(x) =$	0.10	0.25	0.10	0.40	0.15

We take a single observation $X = x$ and wish to consider the hypotheses $f = f_\theta$ with

$$H_0 : \theta = 0 \qquad \textit{vs.} \qquad H_1 : \theta = 1.$$

(a) Find the P-value for the most powerful test of H_0 *vs.* H_1, for each possible value of x.

(b) Find the power of the most powerful test of H_0 *vs.* H_1 of size $\alpha = 0.15$.

(c) For a prior distribution giving equal probabilities to H_0 and H_1, evaluate the posterior probability $\pi[\theta = 0 \mid X = x]$ for each possible value of x.

FYE '07 Take-Home Problem

Turn in solution to Krista Moyle in Room 223 Old Chemistry by 10am on May 9, 2007

The dataset for this exercise consists of the results of 987 screening mammograms administered at the Group Health Cooperative in the state of Washington during the year 2002. Five radiologists were selected at random from those who read lots of mammograms in this cooperative. For each of these radiologists, approximately 200 of the mammograms they read were selected at random. Recorded for each mammogram is a numeric code (1–999) identifying the radiologist who read it, along with two outcomes.

One outcome is an indicator of whether or not there was a breast cancer diagnosis within 12 months following the screening mammogram (1=Yes, 0=No); the second is an indicator of whether or not the subject is recalled for further diagnostic testing (1= recall for further diagnostic testing, 0="normal"). In addition, several risk factors identified in previous studies, are provided; *referent* values for a "typical female" are indicated by asterisks:

AGE 40–49*, 50–59, 60–69, 70 and older

FAMILY HISTORY OF BREAST CANCER 0=No*, 1=Yes

HISTORY OF BREAST BIOPSY/SURGERY 0=No*, 1=Yes

BREAST CANCER SYMPTOMS 0=No*, 1=Yes

MENOPAUSE/HORMONE THERAPY STATUS Pre-menopausal, Post-menopausal & no HT, Post-menopausal & HT*, Post-menopausal & unknown HT

PREVIOUS MAMMOGRAM 0=No*, 1=Yes

BREAST DENSITY CLASSIFICATION 1=Almost entirely fatty, 2=Scattered fibroglandular tissue*, 3=Heterogenously dense, 4=Extremely dense

---

Please address the five problems below. If you do not have enough time to complete the computing you might wish to do for certain parts, please indicate what you *would* do if you had more time.

(a) Develop a descriptive table that provides marginal cancer and recall rates for the risk factors.

(b) Given a set of risk factor levels and a particular radiologist, there is a conceptual 2×2 table of *screening test outcome* by *cancer outcome*. To learn about the probabilities associated with any such table we propose to model the chance of recall given risk factor levels and radiologist and, then, the chance of cancer given screening outcome, risk factor levels and radiologist. Fit these models for the given data. Interpret and comment upon the results.

(c) For a "typical" female (no history of breast biopsy or surgery or family history of breast cancer, age between 40 to 49, post-menopausal and using hormone replacement therapy, has density breast classification 2, and has no reported symptoms), estimate (point and interval, if you can) the chance of the joint event of no recall and no cancer. (Assume an "average" radiologist.)

(d) For a "typical" female (as above), estimate (point and interval, if you can) the chance of a false positive, *i.e.*, the chance of recall given no cancer within 12 months following the screening mammogram. (Again, assume an "average" radiologist.)

(e) Since cancers are rare, someone might choose all of the cancer cases seen by a selected radiologist and a random sample of his/her non-cancer cases. Does this affect the proposed analysis above? If so, explain why.

The data set is available at URL

`http://www.stat.duke.edu/programs/grad/fye/brca.txt`

All entries are numeric. For risk factors with just two levels, the referent level is represented by zero and the alternative by one. For risk factors having more than two levels, the referent level is specified and columns are presented only for incidence of the nonreferent levels. Thus in all cases the referent level would have all columns 0's.

Column coding is as follows:

```
Col  1 -- radiologist ID (1--999)
Col  2 -- cancer outcome (1=Cancer, 0=Normal)
Col  3 -- recall outcome (1=Y  0=N)
Col  4 -- intercept (always 1)
Col  5 -- patient age 50-59 (the referent is patient age is 40-49)
Col  6 -- patient age 60-69
Col  7 -- patient age 70+
Col  8 -- family history of breast cancer  (1=Y 0=N)
Col  9 -- history of breast biopsy/surgery (1=Y 0=N)
Col 10 -- breast cancer symptoms (lump/nipple discharge; 1=Y 0=N)
Col 11 -- pre-menopause (referent is post menopause with hormone treatment)
Col 12 -- post-menopause, no hormone treatment
Col 13 -- post-menopause, unknown hormone treatment
Col 14 -- patient had a previous mammogram
Col 15 -- breast density 1 (breast density 2 is the referent)
Col 16 -- breast density 3
Col 17 -- breast density 4
```

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