FIRST YEAR EXAM

Monday May 10, 2010; 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. No problem is associated with any particular course, so 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.
   A page with common p.d.f. and p.m.f. formulas is attached.
9. The Take-Home practical will be available from Karen Herndon in 214 Old Chemistry immediately after dropping of this written exam and is required to be handed in by 5:00 PM on Tuesday May 11 to Karen Herndon in 214 Old Chemistry.

1. Suppose we want to study the effect of eduction (in years) X on people's income Y and collect data (X_i, Y_i) from a simple random sample of size n from general population. It is known that the distribution of income is usually heavily right-skewed, thus we assume the following linear model (assume $Y > 0$ and both X and Y are centered):

$$\log Y_i = X_i\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

   (a) Assuming σ is unknown, what are the MLEs for β and σ^2? Are they unbiased and why?

   (b) Assuming σ is known, is the MLE of β the uniformly minimum variance unbiased estimator for β? Why or why not?

   (c) Assuming σ is known, give an approximate 95% confidence interval for β when n is large. Describe any theorem that you use.

   (d) Assuming from previous studies that we know $\beta \sim N(\mu_0, \tau^2)$, what is the posterior distribution of β given the observed data (assume σ is known)?

2. Suppose $W_l = \mu_l + \epsilon_l, l = 1, 2, ..., L$ where the μ_l are viewed as constants and the ϵ_l are i.i.d. with distribution having cdf F and pdf f on R^+.

   (a) Obtain the form of the density for $V = \max_{l=1,2,...,L} \ W_l$.

   (b) If the μ_l are all equal, what is the probability that $W_j = v$, for any $j = 1, ..., L$?

   (c) If the μ_l are all distinct, obtain an expression for $\Pr(W_j = V)$.

   (d) If the ϵ's have a Gumbel distribution, i.e., have cdf $F(c) = \exp(-\exp(-c))$, show that $\Pr(W_j = V) = \frac{\exp(\mu_j)}{\sum_l \exp(\mu_l)}$.

3. Let $\mathbf{Y}$ be a $n \times 1$ vector that is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance $\sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Under the null model, $\boldsymbol{\mu} = \mathbf{1}_n \alpha$, where $\mathbf{1}_n$ is a vector of ones of length n. The alternative model has mean vector $\boldsymbol{\mu} = \mathbf{1}_n \alpha + \mathbf{X}\boldsymbol{\beta}$ where $\mathbf{X}$ is an $n \times p$ matrix of rank $p < n$ that has been centered so that $\mathbf{X}^T \mathbf{1}_n = \mathbf{0}_p$.

   (a) Using Zellner's g-prior the (marginal) likelihood of g under the alternative model is

   $$l(g) = (1+g)^{(n-1-p)/2}(1+g(1-R^2))^{-(n-1)/2}.$$

   Find the (marginal) maximum likelihood estimator of g, $\hat{g}$.

   (b) If the null model is true, show that

   $$R^2 \equiv \frac{\mathbf{Y}^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}{\|\mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}\|^2}$$

   has a beta distribution and find its expected value. ($\|\mathbf{Y}\|^2 = \mathbf{Y}^T\mathbf{Y}$ and $\bar{\mathbf{Y}}$ is the mean of $\mathbf{Y}$).

   (c) If p is increasing with n, such that $p/(n-1-p) = r$, with $0 < r < \infty$, what happens to R^2 as $n, p \to \infty$ assuming that the null model is true (show)?

   (d) What happens to $\hat{g}$ under the above conditions as n and p go to infinity?

4. Let $X_1, \ldots, X_n$ be an iid sample from a $\text{Be}(\mu, 1)$ and let $Y_1, \ldots, Y_m$ be an iid sample from a $\text{Be}(\nu, 1)$ pdf. Assume these samples are independent.

   (a) Find likelihood ratio test for $H_0 : \mu = \nu$ versus $H_A : \mu \neq \nu$.

   (b) Show that the test depends upon the statistic

$$T = \frac{\sum \ln X_i}{\sum \ln X_i + \sum \ln Y_j}.$$

   (c) Find the distribution of T under the null, and use this to determine a test of size $\alpha = 0.1$. (hint: For two independent random variables $W \sim \text{Ga}(n, 1/\mu)$ and $V \sim \text{Ga}(m, 1/\mu)$, the distribution of $T = W/(W + V)$ is $\text{Be}(m, n)$.)

5. Let $X \sim \mathsf{Bi}(n, p)$ and $Y \sim \mathsf{Ex}(\lambda)$ be independent random variables with the Binomial distribution (with mean np) and the Exponential distribution (with mean $1/\lambda$), respectively.

   (a) Find the probability $\Pr[X \leq Y]$ explicitly, in closed form.

   (b) If $X_n \sim \mathsf{Bi}(n, p_n)$ and $Y_n \sim \mathsf{Ex}(\lambda)$, $n \in \mathbb{N}$, are all independent, with $\lambda = \log 2$ and $p_n = 1/n$, what is the probability of the event

$$E = \{X_n \leq Y_n \text{ for infinitely-many } n\}?$$

   Why? Show your work.

6. (a) Two real-valued random quantities X and Y have distributions $X \sim N(0,1)$ and $Y \sim N(0,1)$, and it is known that they are uncorrelated, $Cor(X,Y) = 0$. Are X and Y independent? Given an argument, or counter-example, supporting your answer.

   (b) Two real-valued random quantities X and Y have a bivariate joint distribution that is a $50:50$ mixture of two bivariate normals, having pdf

$$p(x,y) = 0.5N(\mathbf{0},\mathbf{U}) + 0.5N(\mathbf{0},\mathbf{V})$$

   where

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}.$$

   i. Sketch the contours of $p(x,y)$.

   ii. What is the marginal distribution of X?

   iii. What is the correlation $Cor(X,Y)$?

   iv. Are X and Y independent?

Beta	$Be(\alpha,\beta)$	$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$x\in(0,1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	
Binomial	$Bi(n,p)$	$f(x)=\binom{n}{x}p^x q^{(n-x)}$	$x\in 0,\cdots,n$	$n\,p$	$n\,p\,q$	$(q=1-p)$
Exponential	$Ex(\lambda)$	$f(x)=\lambda\,e^{-\lambda x}$	$x\in\mathbb{R}_+$	$1/\lambda$	$1/\lambda^2$	
Gamma	$Ga(\alpha,\lambda)$	$f(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\,e^{-\lambda x}$	$x\in\mathbb{R}_+$	α/λ	α/λ^2	
Geometric	$Ge(p)$	$f(x)=p\,q^x$	$x\in\mathbb{Z}_+$	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.	$HG(n,A,B)$	$f(x)=\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}$	$x\in 0,\cdots,n$	$n\,P$	$n\,P\,(1-P)\,\frac{N-n}{N-1}$	$(P=\frac{A}{A+B})$
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	$LN(\mu,\sigma^2)$	$f(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\log x-\mu)^2/2\sigma^2}$	$x\in\mathbb{R}_+$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$	
Neg. Binom.	$NB(\alpha,p)$	$f(x)=\binom{x+\alpha-1}{x}p^\alpha\,q^x$	$x\in\mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$	$(q=1-p)$
		$f(y)=\binom{y-1}{y-\alpha}p^\alpha\,q^{y-\alpha}$	$y\in\{\alpha,...\}$	α/p	$\alpha q/p^2$	$(y=x+\alpha)$
Normal	$N(\mu,\sigma^2)$	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$	$x\in\mathbb{R}$	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	$Po(\lambda)$	$f(x)=\frac{\lambda^x}{x!}e^{-\lambda}$	$x\in\mathbb{Z}_+$	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	$U(a,b)$	$f(x)=\frac{1}{b-a}$	$x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Weibull	$We(\alpha,\beta)$	$f(x)=\alpha\beta\,x^{\alpha-1}\,e^{-\beta\,x^\alpha}$	$x\in\mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	

Take Home Question

A paper (citation omitted) reported on seven published studies on the question of whether chlorinated water is a risk factor for bladder cancer. Each of the primary studies is summarized below in Table 1 and (over the page) Table 2. Each study estimated an adjusted odds ratio that contrasted people who were exposed to chlorinated drinking water with people who drank mostly unchlorinated water. Each study reported an estimated odds ratio and 95% confidence interval adjusted for one or more confounding variables.

Explore and analyze the data. Present your analysis results addressing the primary question: *What is the evidence that chlorinated water is a risk factor for bladder cancer?*

Your report should discuss all aspects of your analysis – exploratory and modeling – with relevant graphical and numerical summaries.

Table 1. Studies of Chlorinated water and bladder cancer

Principal Author	Year	Adjusted OR	LCL	UCL	Method	Quality[c]
Cantor	1987	1.19	1.07	1.32	Logistic	78
Zierler	1988	1.40	1.20	2.10	M-H[a]	71
Wilkins	1986	2.20	0.71	6.82	Logistic	61
Gottlieb	1982	1.18	0.95	1.45	Adj[b]	49
Brenniman	1980	0.98	0.77	1.25	Adj	46
Young	1981	1.15	0.70	1.89	Logistic	45
Alvanja	1978	1.69	1.07	2.67	adj	43

a. M-H is the Mantel-Haenszel method, which produces and approximate logistic regression estimate
b. The odds ratio was adjusted by some method other than logistic regression
c. Each paper was rated for quality on the basis of selection of subjects, measurement of and adjustment for confounding variables, exposure assessment, and statistical analysis. Interpret the score as the percentage of quality criteria that were met in each study.

Table 2. Study Characteristics

Investigator	Population	Cases	Controls	Exposure ascertainment	Adjusted for[a]
Cantor	White residents of 10 US regions	2962 newly diagnosed cases of bladder cancer	Community controls matched 2:1 for age, sex, and region by random dialing and medicare files	History of residence and beverage consumption combined with sampling of water utilities.	Age, sex, smoking, occupation, size of place of longest residence
Zierler	MA residents	51,645 kidney, bladder, lung, breast, pancreas, GI cancer deaths	cardiovasular, cerebro-vascular, or pulmonary deaths or lymphatic cancer deaths.	Chlorination / chloramination at address on death certificate.	Age, sex surface / groundwater, size, poverty of county of residence.
Wilkins	31,000 residents of Washington Co, MD	Diagnosis of cancer in 12 years after initial survey.	Unexposed: Deep well users.	Exposed: Users of chlorinated surface water in Hagerstown, MD	Age, sex, smoking, marital status, education.
Gottlieb	Louisiana residents	10,205 kidney, bladder, stomach, liver, colorectal cancer deaths.	Other cancer deaths and noncancer deaths matched 1:1 for age, sex, race, and county.	Chlorinated / unchlorinated water ad address on death cert. and at birthplace.	surface / groundwater, cardiovascular disease, age, sex, race, county
Brenniman	White residents of IL	3208 GI and urinary tract cancer deaths	Noncancer deaths matched 14:1 for age, sex, and county.	Chlorinated / unchlorinated water at address on death cert.	Urbanicity, population densit, age, sex, county.
Young	White, Female WI residents	8029 cancer deaths	Noncancer deaths matched 1:1 for age, sex, race, and county	Chlorinated / unchlorinated water at address on death cert.	Pop density, occupation, marital status, rural runoff in water, age, sex, race, county
Alvanja	NY State residents	3446 GI and urinary tract cancer deaths	Noncancer and lung cancer deaths matched 1:1 for age, sex, county.	Chlorinated / unchlorinated water at address on death cert.	Urbanicity, occupation, age, sex, county

a. In addition to variables used to match cases and controls.