Department of Statistical Science Duke University

# FIRST YEAR EXAM - SPRING 2011

Monday, May 9th 2011

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 the written exam: Questions 1-6 inclusive. Attempt all questions; note that credit is not necessarily equally allocated across questions.
7. This is a closed book exam. No notes are permitted.

1. Let $X \sim \textit{Ex}(\theta)$ be an exponentially distributed random variable with rate $\theta > 0$, with pdf

$$f(x) = \theta e^{-\theta x}, \qquad x > 0$$

and let $Y \sim \textit{Bin}(n, p)$ be a Binomial random variable with pmf

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, \cdots, n,$$

with $X$ and $Y$ independent.

(a) Briefly describe what it means for the continuous variable $X$ and the discrete variable $Y$ to be independent.

(b) What is the probability of the event $A = [X \geq Y]$? Simplify the resulting expression.

(c) What is the conditional distribution of $Y$ given the event $A$? It is a familiar one, give its name and the value of any parameter(s).

(d) Now consider infinitely-many independent random variables $X_n \sim \textit{Ex}(\theta)$ for $n = 1, 2, \ldots$. How many of the events

$$A_k = [\ X_k > \log k\ ]$$

will occur— finitely-many or infinitely-many? Justify your answer.

2. Let $X$ and $Y$ be two independent random variables with $X \sim Ex(\lambda)$ and $Y \sim Ex(\mu)$ [pdfs: $f_X(x) = \lambda e^{-\lambda x},\ x > 0$; $f_Y(y) = \mu e^{-\mu y}, y > 0$.] Suppose $X$ and $Y$ are not directly observable. Instead we record

$$Z = \min(X, Y), \quad \text{and} \quad W = \begin{cases} 1 & \text{if } Z = X \\ 0 & \text{if } Z = Y \end{cases}.$$

(a) Show that

$$P(Z > z, W = 1) = \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)z} \quad \text{and} \quad P(Z > z, W = 0) = \frac{\mu}{\lambda + \mu} e^{-(\lambda+\mu)z},$$

for every $z > 0$.

(b) Write down the likelihood function in $(\lambda, \mu)$ based on $n$ independent observations $(z_i, w_i)$, $i = 1, ..., n$, of $(Z, W)$ and find the MLE of $\lambda$ and $\mu$ (simplify expessions).

(c) Now suppose even $Z$ is not recorded and we only have $n$ independent observations $w_1, \cdots, w_n$ of $W$. Can we test $H_0 : \lambda \leq \mu$ vs. $H_1 : \lambda > \mu$ based on these observations? Give a brief explanation.

3. A survey of 100 Duke lower-classmen included $n$ Freshmen and $m = 100 - n$ Sophomores. Of these, $X$ Freshmen and $Y$ Sophomores are recorded to be in favor of a recent federal policy. Suppose a $\theta \in (0, 1)$ proportion of all Freshmen and a $\lambda \in (0, 1)$ proportion of all Sophomores at Duke favor the policy. We are interested in testing $H_0 : \theta = \lambda$ vs. $H_1 : \theta \neq \lambda$ at some fixed level $\alpha \in (0, 1)$.

   (a) Under what assumption(s) on the survey, is it OK to model the data as: $X \sim \textit{Bin}(n, \theta)$ independently of $Y \sim \textit{Bin}(m, \lambda)$?

   (b) For the binomial model, describe a level-$\alpha$ test for $H_0 : \theta = \lambda$ vs. $H_1 : \theta \neq \lambda$ based on a normal approximation and the statistics

$$D = \frac{X}{n} - \frac{Y}{m} \quad \text{and} \quad S = \frac{X + Y}{n + m}.$$

   For your test, provide the rule for rejecting $H_0$ and state whether the level condition is met exactly or asymptotically. Justify your answer. A detailed proof is not needed, simply give a sketch of an argument and cite appropriate probability results.

   (c) Under what assumption(s) on the survey, is it OK to view the data as the two-way table

| | In favor | Against | Total |
|---|---|---|---|
| Freshman | $X$ | $n - X$ | $n$ |
| Senior | $Y$ | $m - Y$ | $m$ |
| Total | $X + Y$ | $m + n - (X + Y)$ | $m + n$ |

   with the vector of cell counts having a multinomial distribution?

   (d) For the multinomial model what is the Pearson's $\chi^2$ test statistic for testing $H_0 : \theta = \lambda$ vs. $H_1 : \theta \neq \lambda$? State the degrees of freedom of the associated $\chi^2$ distribution. You don't have to simplify the expression of the statistics, you can leave it in terms of "observed" and "expected" counts, with precise expressions given for these counts.

4. Two small, independent experiments measure the levels of a continuous response $x$ of an experimental treatment compared to a control. The measured data are assumed normal with unit variance, so that the data are independently distributed as $x_i \sim N(\mu_0, 1)$ (under control) and $x_i \sim N(\mu_1, 1)$ (under treatment). Assume reference priors $p(\mu_0) \propto constant$ independently of $p(\mu_1) \propto constant$.

   Experiment A measures 50 observations on the treatment and 10 on the control, while Experiment B measures 20 observations on the treatment and 20 on the control. The two experiments are run under exactly the same conditions so it is perfectly valid to combine the two data sets. The summary statistics (sample mean and sample size in each case) are as follows:

| | Treatment ($\mu_1$) | | Control ($\mu_0$) | |
|---|---|---|---|---|
| *Experiment A:* | $\bar{x}_1 = 10$ | $n_1 = 50$ | $\bar{x}_0 = 10$ | $n_0 = 10$ |
| *Experiment B:* | $\bar{x}_1 = 6$ | $n_1 = 20$ | $\bar{x}_0 = 6$ | $n_0 = 20$ |
| *A & B combined:* | $\bar{x}_1 = 8.86$ | $n_1 = 70$ | $\bar{x}_0 = 7.33$ | $n_0 = 30$ |

   You are to make inferences about the effect of treatment: $\delta = \mu_1 - \mu_0$. Do this by answering (a) and (b) below. *NOTES– 1: Do NOT derive any distribution theory results; simply use standard normal theory and just quote the results used. 2: Give numerical results in part (a). 3: Parts (a) and (b) are equally weighted for credit for this question.*

   (a) (i) What is the posterior for $\delta$ based only on the data from Experiment A?
   (ii) What is the posterior for $\delta$ based only on the data from Experiment B?
   (iii) What is the posterior for $\delta$ based on the data from A & B combined?

   (b) Based on interpreting these posterior inferences, two analysts reasoned as follows:
   - ◇ Analyst Joe: *In each Experiment A and B separately, there is no evidence at all of a non-zero treatment effect; the two experiments agree therefore I conclude there is no effect.*
   - ◇ Analyst Jemima: *Combining the data from A & B indicates a very significant and positive treatment effect, so I conclude there is indeed an effect.*

   Discuss these competing views.

5. Consider the following statistical model for $n$ records $(x_i, Y_i)$, $i = 1, \cdots, n$. For any record $i$, let the random variable $Z_i$ be independently distributed as

$$P(Z_i = 1|x_i) = \eta, \quad P(Z_i = 0|x_i) = 1 - \eta, \quad 0 < \eta < 1,$$

and let $Y_i$ given $(Z_i, x_i)$ (and parameters) be independently distributed with Poisson pmfs:

$$Z_i = 0: \ P(Y_i = y|Z_i = 0, x_i, \lambda_0) = (\lambda_0 x_i)^y \frac{e^{-\lambda_0 x_i}}{y!}, \ \ y = 0, 1, 2, \cdots,$$
$$Z_i = 1: \ P(Y_i = y|Z_i = 1, x_i, \lambda_1) = (\lambda_1 x_i)^y \frac{e^{-\lambda_1 x_i}}{y!}, \ \ y = 0, 1, 2, \cdots.$$

Set $\lambda_1 = \lambda_0 \alpha$ and consider a prior specification that treats $\eta$, $\lambda_0$ and $\alpha$ as independent with

$$\eta \sim \textit{Be}(a, b): \ \pi_\eta(\eta) \propto \eta^{a-1}(1-\eta)^{b-1}, \text{ for some } a > 1, b > 1$$
$$\lambda_0 \sim \textit{Ga}(v, s): \ \pi_{\lambda_0}(\lambda_0) \propto \lambda_0^{v-1} e^{-\lambda_0 s}, \text{ for some } v > 0, s > 0$$
$$\alpha \sim \textit{Ga}(w, t): \ \pi_\alpha(\alpha) \propto \alpha^{w-1} e^{-\alpha t}, \text{ for some } w > 0, t > 0.$$

Let $X$ represent the collection of all $x_i$, and let $Y$ represent the collection of all $y_i$.

(a) Show that the likelihood function of the "parameters" $\lambda_0, \alpha, \eta, Z_1, \cdots, Z_n$ is given by

$$L(\lambda_0, \alpha, \eta, Z_1, \cdots, Z_n) \propto \lambda_0^{Y^+} \alpha^{Y_1^+} \exp\{-\lambda_0 X_0^+ - \alpha \lambda_0 X_1^+\}$$

where $Y^+ = \sum_i y_i$, $Y_1^+ = \sum_{i:Z_i=1} y_i$, $X_0^+ = \sum_{i:Z_i=0} x_i$ and $X_1^+ = \sum_{i:Z_i=1} x_i$.

(b) Find the conditional distribution of $\lambda_0$ given $X, Y, Z_1, \cdots, Z_n, \alpha, \eta$. It is acceptable to derive the conditional density up to a multiplicative constant, but identify the name of the conditional distribution.

(c) Show that, given $X, Y, \alpha, \lambda_0, \eta$, the variables $Z_1, \cdots, Z_n$ are conditionally independent. Identify the conditional distribution (name + expression for parameter(s)) of each $Z_i$.

6. Suppose

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \right), \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix} \right)$$

with $X$ and $Y$ independent. Define

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = A \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \quad V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = A \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}$$

where

$$A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix}.$$

Show that $U_1$ is independent of $V_2$ given $V_1$.

**FYE 2011: Take Home Question** *Department of Statistical Science, Duke University*

## Background

The prevalence of maternal cigarette smoking during pregnancy is about 14% in the United States despite warnings of its effects on the developing fetus. More generally, it is thought that many complex diseases and disorders have origins *in utero*, resulting from adaptations to the gestational environment and/or from exposure to toxicants, such as those that result from maternal smoking during pregnancy. Epigenetics is one mechanism by which *in utero* exposures may influence health outcomes of the offspring. For example, exposures may affect the epigenetic regulation of imprinted genes that are critical to normal growth and development and thereby increase risk of adverse health outcomes.

Imprinted genes are those whose regulation is either maternally or paternally controlled. Normally, one of the two alleles of an imprinted gene is silenced via methylation at differentially methylated regions (DMRs) located on that allele. If the gene is maternally expressed, the paternal allele is silenced; if it is paternally expressed, the maternal allele is silenced. Hence when such a DMR is assayed, a normal individual's methylation fraction (measured using the DNA of many cells from the subject) should be near 50%.

A well-characterized imprinted domain on human chromosome 11p15.5 contains the genes for paternally expressed Insulin-like Growth Factor II (IGF2) and maternally expressed H19. Deregulation of IGF2 expression has been linked to overgrowth disorders, obesity and cancer. Imprinted expression and transcription of IGF2 are regulated in large part through the patterns of differential methylation of at least two regulatory DMRs, one of which is located near the H19 promoter (H19 DMR) and the other upstream of the three IGF2 promoters that are subject to imprinting (IGF2 DMR). Both DMRs have been shown to exhibit altered methylation in cigarette smoking-related malignancies.

The goal of this analysis is to examine the influence of a developing child's *in utero* exposure to maternal cigarette smoking byproducts on percent methylation at the IGF2 DMR using umbilical cord blood samples. The data are from a multi-ethnic birth cohort established to facilitate study of the effects of early exposures on epigenetic profiles and phenotypic outcomes. The IGF2R DMR methylation is measured as the percent of the subject's (offspring's) alleles that are methylated in a sample of their DNA and, hence, takes values between 0 and 100. Methylation at this locus was assayed twice for 294 of the 314 subjects and only once for the remaining 20. The 608 DNA samples were arrayed on 22 96-well plates and each plate was processed separately. Each plate has eight rows, designated 'A' through 'H,' and twelve columns, designated 1 through 12; only a subset of wells were used on each plate.

## Data

The data are in the (tab delimited) file

`http://www.stat.duke.edu/~clyde/epigen.dat`

There are 314 rows where each row corresponds to one subject (offspring and mother) with columns corresponding to:

**age** Maternal age coded 'lt30' (younger than 30 at delivery), '30to39' (30 to 39 years old at delivery), and 'ge40' (more than 40 years old at delivery).

**BMI** Maternal body mass index measured as weight before pregnancy (in kg) divided by height (in meters) squared and coded '0' (=less than 30) or '1' (greater than or equal to 30).

**smoke** Maternal cigarette smoking during pregnancy; coded '0' if did not smoke and '1' if smoked during early pregnancy and stopped *or* if smoked throughout pregnancy.

**gestage** Gestational age of the infant. Coded '1' if less than 37 weeks and '0' if greater than or equal to 37 weeks.

**gender** Infant's gender ('1'=male, '0'=female).

**edu** Mother's education level (coded 'ltHS' for less than high school, 'ltCollege' for high school/GED, and 'geCollege' for at least some college)

**race** Mother's race/ethnicity ('AA'=African American, 'EA'=Caucasian or 'Other').

**methyl1** the first replicate measurement of the subject's child's methylation level.

**methyl2** the second replicate measurement of the subject's child's methylation level. 20 subjects do not have a second measurement so this is missing.

**plate1** (and 'plate2') plates on which the subject's first and second replicate measurements were made, respectively. 'plate2' is missing for 20 subjects who have only a single measurement.

**row1** (and 'row2') row in which the subject's first and second replicate measurements were placed on the plates whose IDs appear in 'plate1' and in 'plate2,' respectively. 'row2' is missing for 20 subjects.

**column1** (and 'column2') column in which the subject's first and second replicate measurements were placed on the plates whose IDs appear in 'plate1' and in 'plate2,' respectively. 'column2' is missing for 20 subjects.

**well1** (and 'well2') well in which the subject's first and second replicate measurements were placed on the plates whose IDs appear in 'plate1' and in 'plate2,' respectively. 'well2' is missing for 20 subjects.

## Objectives

A child's methylation level may depend on gender and race and may be associated with maternal smoking or unmeasured exposures associated with maternal age, maternal BMI, maternal education level, and/or gestational age. Explore and analyze the data, characterizing the relationships between percent methylation in the offspring and maternal smoking status adjusting for any other covariates. It is also of interest to determine if the relationship between smoking and methylation is the same across race and gender.

Present your results in a three page (maximum) report addressing the primary question: Does maternal smoking modify patterns of methylation in the offspring? Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results. While you may include code and other plots in a supplemental appendix, you should not assume that graders will read beyond the main report; all relevant material should be within the three page limit.

| **Distribution** | **Notation** | $f(x) =$ **pdf (pmf)** | **Support** | **Mean** | **Variance** | |
|---|---|---|---|---|---|---|
| **Beta** | $Be(a,b)$ | $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ | $x \in (0,1)$ | $\frac{a}{a+b}$ | $\frac{ab}{(a+b)^2(a+b+1)}$ | |
| **Binomial** | $Bin(n,p)$ | $f(x) = \binom{n}{x} p^x q^{(n-x)}$ | $x \in 0, \cdots, n$ | $n\,p$ | $n\,p\,q$ | $(q = 1-p)$ |
| **Chi-square** | $\chi^2(\nu)$ | $f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}\, e^{-x/2}$ | $x \in \mathbb{R}_+$ | $\nu$ | $2\nu$ | |
| **Exponential** | $Ex(\lambda)$ | $f(x) = \lambda\, e^{-\lambda x}$ | $x \in \mathbb{R}_+$ | $1/\lambda$ | $1/\lambda^2$ | |
| **Gamma** | $Ga(\nu,\lambda)$ | $f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1}\, e^{-\lambda x}$ | $x \in \mathbb{R}_+$ | $\nu/\lambda$ | $\nu/\lambda^2$ | |
| **Geometric** | $Geo(p)$ | $f(x) = p\, q^x$ | $x \in \mathbb{Z}_+$ | $q/p$ | $q/p^2$ | $(q = 1-p)$ |
| | | $f(y) = p\, q^{y-1}$ | $y \in \{1, ...\}$ | $1/p$ | $q/p^2$ | $(y = x+1)$ |
| **HyperGeo.** | $HG(n,M,N)$ | $f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$ | $x \in 0, \cdots, n$ | $n\,p$ | $n\,p\,(1-p)\frac{N-n}{N-1}$ | $(p = \frac{M}{N})$ |
| **Logistic** | $Lo(\mu,\beta)$ | $f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$ | $x \in \mathbb{R}$ | $\mu$ | $\pi^2\beta^2/3$ | |
| **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.** | $NB(\alpha,p)$ | $f(x) = \binom{x+\alpha-1}{x} p^\alpha\, q^x$ | $x \in \mathbb{Z}_+$ | $\alpha q/p$ | $\alpha q/p^2$ | $(q = 1-p)$ |
| | | $f(y) = \binom{y-1}{y-\alpha} p^\alpha\, q^{y-\alpha}$ | $y \in \{\alpha, ...\}$ | $\alpha/p$ | $\alpha q/p^2$ | $(y = x+\alpha)$ |
| **Normal** | $N(\mu,\sigma^2)$ | $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ | $x \in \mathbb{R}$ | $\mu$ | $\sigma^2$ | |
| **Pareto** | $Pa(\alpha,\epsilon)$ | $f(x) = \alpha\, \epsilon^\alpha / x^{\alpha+1}$ | $x \in (\epsilon, \infty)$ | $\frac{\epsilon\,\alpha}{\alpha-1}$ | $\frac{\epsilon^2\alpha}{(\alpha-1)^2(\alpha-2)}$ | |
| **Poisson** | $Poi(\lambda)$ | $f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$ | $x \in \mathbb{Z}_+$ | $\lambda$ | $\lambda$ | |
| **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** | $Wei(\alpha,\beta)$ | $f(x) = \alpha\beta\, x^{\alpha-1}\, e^{-\beta\, x^\alpha}$ | $x \in \mathbb{R}_+$ | $\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$ | $\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$ | |