May 2015 MS/PhD Qualifying Examination
Department of Statistics
University of South Carolina
May 18, 2015: 9:00AM–3:00PM

Instructions: You will be given a folder with a number written on it; inside the folder is the Qualifying Exam as well as a slip of paper with the same number that is on the folder. Write your name on the numbered piece of paper and place it inside the envelope provided by the proctor. This envelope will be sealed until after the exams are graded.

This exam consists of six problems; answer all of them. Use separate sheets of paper for each problem, and write only on one side of each page (do **not** write on both sides). Each sheet that you turn in should have the problem number clearly labeled at the top as well as the number on your folder.

You are allowed to use the computers and the software in the examination room. However, you are **not** allowed to use the Internet, except to examine help files of the software and to download data sets that are needed in some of the problems. Provide complete details in your solutions.

You have **six hours** to complete this examination. Good luck!

1. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) Poisson(λ), where $\lambda \in (0, \infty)$.

(a) Find a complete sufficient statistic for this problem and justify your answer.

(b) Show that for any constant B,

$$E(B^{\sum_{i=1}^n X_i}) = e^{n\lambda(B-1)}.$$

(c) Find an UMVUE (Uniformly Minimum Variance Unbiased Estimator) of $P(X_i = 0)$ and justify why it is UMVUE.

(d) Consider the estimator $e^{-\bar{X}_n}$ of $P(X_i = 0)$ where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Is this a consistent estimator? Why or why not?

(e) Find σ^2 such that

$$\sqrt{n}(e^{-\bar{X}_n} - e^{-\lambda}) \xrightarrow{d} N(0, \sigma^2).$$

2. A large car dealership asks you to develop a statistical model to relate the count of cars sold (on a given day) to several independent variables. A random sample of 100 days was obtained, available online at `http://people.stat.sc.edu/hansont/carsales2015.txt`

The dealership president suspects that car sales may depend on whether the day is on a weekend; however, this effect could depend on whether a day is a national holiday (when many potential buyers could be off work). The president also suspects some type of weather effect on car sales, e.g., sales may be related to the daily high temperature. The precise nature of this relationship is uncertain and should be investigated. In addition, the president suspects that because of a monthly sales quota system, the dealership tends to sell more cars later in any given month, on average.

Build a model that will fit these data well and perform hypothesis tests related to each of the president's suspicions. Be sure to justify your model choice, using diagnostics to check model fit, model assumptions, etc. Using at least a couple of paragraphs, write a mini-report summarizing your model choices and your findings. Discuss your substantive conclusions about the president's conjectures.

3. There are three six-sided dice in a bag, all physically balanced but with different labelings on the sides. The dice are labeled as in the figure on the right. You are blindfolded and you randomly select one die from the bag. You roll this die 29 times, and your friend tells you that your rolls produced the following frequency table of results:

Outcome	Frequency
1	5
2	11
3	6
4	7

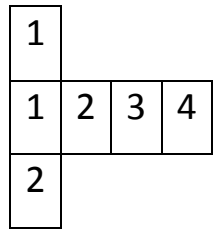
You are not told which die you rolled.

(a) The parameter of interest is $\mathbf{p} = (p_1, p_2, p_3, p_4)$, the probability vector for the die that you rolled. State the parameter space.

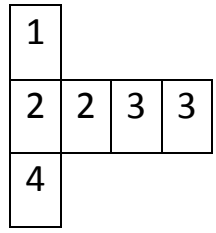
(b) Find the maximum likelihood estimate of $\mathbf{p} = (p_1, p_2, p_3, p_4)$.

(c) Given the observed data, find the probability that you picked Die A.

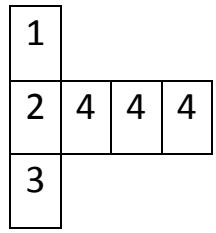
(d) Set up and carry out a hypothesis test of $H_0 : \mathbf{p} = \mathbf{p}_C$ versus $H_a : \mathbf{p} \neq \mathbf{p}_C$ with a size as close to 0.10 as possible. Here, $\mathbf{p}_C$ denotes the probability vector corresponding to Die C. Give a P-value for your test. State whether your test is exact or approximate.



Die A



Die B



Die C

4. When a beam of light is passed through a chemical solution, a fraction of the incident light will be absorbed or reflected and the rest, i.e. the *transmitted* light, will make it through the solution. The intensity of the transmitted light decreases as the concentration of the chemical solution increases.

The data tabled below are from an experiment in which several solutions of known concentrations x_i (milligrams/liter) of a pure chemical were used to measure the amount of transmitted light Y_i (arbitrary units).

i	y_i	x_i	i	y_i	x_i
1	2.86	0.0	2	2.64	0.0
3	1.57	1.0	4	1.24	1.0
5	0.45	2.0	6	1.02	2.0
7	0.65	3.0	8	0.18	3.0
9	0.15	4.0	10	0.01	4.0
11	0.04	5.0	12	0.36	5.0

These data are available online at `http://people.stat.sc.edu/hansont/light.txt`

(a) Fit the model

$$Y_i = \gamma_1 + \gamma_2 e^{\gamma_3 x_i} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

to these data. Report maximum likelihood estimates of all four model parameters with confidence intervals.

(b) Obtain two plots: (i) the fitted mean curve superimposed on a scatterplot of the raw data, and (ii) Pearson residuals $r_i = (y_i - \hat{\gamma}_1 - \hat{\gamma}_2 e^{\hat{\gamma}_3 x_i})/\hat{\sigma}$ vs. concentration x_i. Comment on the fit of the model in light of these plots.

(c) As the concentration $x \to \infty$, what does the mean $\mu(x) = \gamma_1 + \gamma_2 e^{\gamma_3 x}$ converge to when $\gamma_3 < 0$? Obtain an estimate and confidence interval for this limiting value. Test that this limiting value is equal to zero at the 5% level.

(d) Obtain a 95% CI for $\gamma_1 + \gamma_2$, the mean at zero concentration.

(e) Test for constant variance by fitting

$$Y_i = \gamma_1 + \gamma_2 e^{\gamma_3 x_i} + \epsilon_i, \quad \epsilon_i \overset{ind.}{\sim} N(0, e^{\tau_0 + \tau_1 x_i}).$$

Discuss your conclusion.

(f) Redo the analyses in parts (a) and (b) with a quadratic mean

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Also obtain the AIC for both models. Which model is preferred? Why?

5. Suppose U_1 and U_2 are $iid\ Uniform(0,1)$ random variables.

(a) Show that

$$Z_1 = \cos(2\pi U_1)\sqrt{-2\log U_2}, \quad Z_2 = \sin(2\pi U_1)\sqrt{-2\log U_2}$$

are i.i.d $N(0,1)$ random variables.

(b) Consider the usual simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0,\sigma^2),$$

where $i = 1,\ldots,n$. In matrix notation this is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{I}\sigma^2),$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)'$, $\boldsymbol{\beta} = (\beta_0,\beta_1)'$, $\mathbf{X}$ is an $n \times 2$ design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)'$. The least squares (and maximum likelihood) estimators for $\boldsymbol{\beta}$ are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Derive the sampling distribution of $\hat{\boldsymbol{\beta}}$ directly using the properties of multivariate normal distributions.

(c) The Cholesky factorization of a symmetric, positive definite matrix $\boldsymbol{\Sigma}$ gives upper triangular $\mathbf{A}$ such that $\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{A}$. For $\mathbf{Z} = (Z_1, Z_2)'$ in part (a), argue that one can simulate $\mathbf{V} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\mathbf{V} = \boldsymbol{\mu} + \mathbf{A}'\mathbf{Z}$ where $\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{A}$ and $\mathbf{Z} \sim N_2(\mathbf{0}, \mathbf{I})$. In R, the `chol` function returns the upper triangular $\mathbf{A}$.

(d) For $\boldsymbol{\beta} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}$, and $\sigma^2 = 1$, use (a), (b), and (c) directly along with `runif(2)` to simulate 1000 independent copies of $\hat{\boldsymbol{\beta}}$ in R. Provide a scatterplot of simulated values ($\hat{\beta}_0$ along the x-axis and $\hat{\beta}_1$ along the y-axis); include your R code.

(e) Find an exact expression for the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ for the $\mathbf{X}$ from part (d). In general, is this correlation free of $\boldsymbol{\beta}$ and σ^2?

6. Consider data obtained from subjects involved in a randomized, double-blind, parallel-group, multicenter study comparing two oral treatments (terbinafine versus itraconazole) for toenail infection (De Backer et al., 1998). Patients received one of the two treatments (1 or 0) and were evaluated for the degree of onycholysis (degree of separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48 thereafter. The onycholysis outcome variable is binary ($Y_{ij} = 0$ for "none to mild" versus $Y_{ij} = 1$ for "moderate to severe"); $Y_{ij} = 1$ implies greater toenail separation, a bad thing. The binary outcome was evaluated on $n = 294$ patients at 7 visits comprising a total of 1908 measurements. Y_{ij} is the measurement taken at visit j for patient i, where $i = 1, \ldots, 294$ are subjects and $j = 1, \ldots, 7$ correspond to scheduled visits at 0, 4, 8, 12, 24, 36, and 48 weeks. The exact timing of measurements in months m_j is included. The data are online at `http://people.stat.sc.edu/hansont/toenail.txt`

From left to right are recorded the observation number (ignore this), a unique patient id, the response Y_{ij}, the treatment t_i, the month m_j, and the visit j. You will perform a simplified analysis of these data. Consider the random intercept model where the log-odds of greater separation is linear in months:

$$\text{logit}\{P(Y_{ij} = 1)\} = \beta_0 + \beta_1 t_i + \beta_2 m_j + \beta_3 t_i m_j + u_i, \quad u_1, \ldots, u_{294} \overset{iid}{\sim} N(0, \sigma^2).$$

(a) Why is an interaction between time (in months) and treatment included?

(b) Show that, conditional on u, the odds ratio of greater separation comparing $t = 1$ to $t = 0$ at month m is

$$OR(m) = e^{\beta_1 + \beta_3 m}.$$

(c) Fit the model above and report parameter estimates with 95% CI's for all model parameters including σ^2.

(d) Obtain the odds ratios in part (b) with 95% CI for $m = 0, 3, 6, 9, 12$ months; do not worry about multiple comparisons here. At which time (in months) are significant treatment differences observed at the 5% level? Which treatment is better?

(e) Formally test whether there is significant variability among patients in terms of their latent $u_1, \ldots, u_{294}$.

(f) Plot the estimated odds of greater separation (assuming $u = 0$) vs. m for $t = 0$ and $t = 1$. Your plot should have two curves. Do both treatments "work"? Does the plot confirm what you would expect at time $m = 0$?