MATH 476 Statistics

Fred J. Hickernell **Spring 2010**
Final Exam **Tuesday, May 4**

Instructions:

i. *This test consists of* FIVE *questions. Answer all of them.*

ii. *The time allowed for this test is* 2 hours.

iii. *The data and situations portrayed in this test are fictitious, but realistic.*

iv. *This test is closed book, but you may use* four *double-sided letter-size sheets of notes.*

v. *Calculators, even of the programmable variety, are allowed. Computers using JMP or MATLAB, are also allowed. No internet access, web browsing, email, chat, etc. is allowed.*

vi. *Show all your work to justify your answers. Answers without adequate justification will not receive credit.*

1. (15 marks)
The geometric probability mass function is defined as $f(x;\theta) = \theta(1-\theta)^x$, for $x = 0, 1, \ldots$. Here the parameter θ denotes the probability of a success, and the random variable X with this distribution is the number of failures experienced before obtaining the first success. What is the maximum likelihood estimator of θ?

2. (20 marks)
William thinks that his lucky number is 8. Every week a lottery selects 5 numbers at random out of the numbers 1 through 40.

   a) If the lottery numbers are really chosen randomly, what is the probability that the number 8 will be chosen this week.

   b) William notices that during the last 208 weeks, the number 8 has been chosen 30 times. Does this provide compelling evidence that 8 is chosen *more often* than it would be by pure chance?

   c) Does the data mentioned in the previous part provide compelling evidence that 8 is chosen *no more often* than it would be pure chance?

3. (20 marks)
Incoming undergraduate students are polled on how they think that the new iPod Touches that they will be receiving *should be* used in their classes. They are asked to choose their preference from one of four categories. Professors for the Level 100 and 200 classes are asked the same question. Their responses are given below. This sample is taken to represent the views of students and faculty nationwide.

   A. Not at all

   B. Classroom demonstrations but no graded work

   C. Classroom demonstrations and homework, but no tests or exams, and

   D. All aspects of learning, including classroom demonstrations, homework, quizzes, tests and exams.

	Extent of Use			
	A	B	C	D
Students	45	142	186	179
Professors	30	52	28	12

a) Construct a 95% confidence interval for the proportion of incoming students that believe that iPod Touches should be used in all aspects of learning.

b) Does this data provide compelling evidence that faculty and students have different views about the role of iPod Touches for learning.

4. (25 marks)
A linear regression model may be written as $Y = \beta_0 + \beta_1 g_1(\boldsymbol{x}) + \cdots \beta_p g_p(\boldsymbol{x}) + \varepsilon$, where $\boldsymbol{x}$ is the d-dimensional vector of inputs, Y is the random output, $g_1, \ldots, g_p$ are given functions, $\beta_0, \ldots, \beta_p$ are the unknown, but deterministic, regression coefficients, and ε is the random noise. After observing n data points, $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$, one may write the model with data in matrix form, $\boldsymbol{Y} = \mathsf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathsf{G} = \begin{pmatrix} 1 & g_1(\boldsymbol{x}_1) & \cdots & g_p(\boldsymbol{x}_1) \\ \vdots & \vdots & & \vdots \\ 1 & g_1(\boldsymbol{x}_n) & \cdots & g_p(\boldsymbol{x}_n) \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Here it is assumed that the ε_i are i.i.d. Gaussian with mean zero and variance σ^2, i.e., $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \mathsf{I})$. The least squares regression estimate for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathsf{G}^T\mathsf{G})^{-1}\mathsf{G}^T\boldsymbol{Y}$.

a) Is $\hat{\boldsymbol{\beta}}$ random or deterministic, and why?

b) Is $\hat{\boldsymbol{\beta}}$ a biased or unbiased estimator of $\boldsymbol{\beta}$?

c) The regression diagnostics in JMP and other statistical packages produce an ANOVA table and an F-test. What is the null hypothesis for this F-test in mathematical terms?

d) The regression diagnostics in JMP and other statistical packages also produce t-tests for the regression coefficients. If the p-value for the t-test on coefficient β_3 is 0.01, then what are the relevant null and alternative hypotheses, and what is your conclusion at the 5% significance level?

e) For the situation described in the previous part, does the p-value depend on β_3, $\hat{\beta}_3$, and/or σ^2? Answer yes or no for each one.

5. (20 marks)
Consider the following data set for the period of a simple pendulum as a function of the amplitude:

Amplitude	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Period	6.23	6.29	6.09	6.24	6.22	6.34	6.38	6.42	6.78	6.74

Amplitude	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
Period	6.54	6.95	7.03	6.99	7.22	7.37	7.58	7.86	8.06	8.26

Amplitude	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
Period	8.61	8.9	9.4	9.85	10.18	10.79	11.58	12.71	13.9	16.22

Find a regression model that fits this data well.