UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2013 Final Examination
STA442H1F/2101HF
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

Aids: Calculator Model(s): Any calculator without wireless capability is okay.
Formula sheet supplied.

Last/Surname (Print): ______________________________

First/Given Name (Print): ______________________________

Student Number: ______________________________

Signature: ______________________________

Qn. #	Value	Score
1	10	
2	18	
3	10	
4	15	
5	19	
6	23	
7	5	
Total = 100 Points		

1. (*10 points*) In a taste test, a sample of n consumers is asked to indicate which of two coffees they like more. The coffees are labelled A and B.

   We'll model this by saying that $Y_1, \ldots, Y_n$ is a random sample from a Bernoulli distribution with parameter $\theta \in (0, 1)$, where θ is the probability that Coffee A is preferred.

   (a) We want to test for difference in preference for the two coffees. What is the null hypothesis, in symbols?

   (b) Suppose $n = 100$ consumers participate in the study, and 60 of them choose Coffee A. What is the likelihood ratio test statistic G^2? The answer is a number. Show some work. You do not have to derive the MLEs; just write them down if you know them. **Circle your final answer**.

continued on page 3

(c) Continuing with the taste test study of Question 1, what is the critical value at $\alpha = 0.05$? The answer is a number.

(d) Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

(e) In plain, non-statistical language, what do you conclude?

2. (*18 points*) Here is a regression model with a single random explanatory variable, in which the slope equals the intercept. Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta + \beta X_i + \epsilon_i,$$

where $E(X_i) = \mu_x$, $E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i.

(a) Propose an estimator of β that will be consistent. Call it $\widehat{\beta}_n$. You don't have to derive it; just guess (perhaps based on homework), and **give a formula for** $\widehat{\beta}_n$. Don't prove consistency yet; you'll do that in the next part.

continued on page 4

(b) Continuing with Question 2, show that your estimator $\widehat{\beta}_n$ is consistent.

continued on page 5

(c) Still continuing with Question 2, is your estimator $\widehat{\beta}_n$ unbiased? Answer Yes or No and prove your answer. Remember that X_i is *random*, and the the expected value of a ratio is *not* the ratio of expected values.

continued on page 6

3. (*10 points*) For this question you may use the following facts without proof; they are not directly on the formula sheet. Let $D_1, \ldots, D_n$ be a random sample (i.i.d.) from a $N(\mu, \sigma^2)$ distribution. Then

- $\frac{\sum_{i=1}^n (D_i - \overline{D})^2}{\sigma^2} \sim \chi^2(n-1)$
- $\sum_{i=1}^n (D_i - \overline{D})^2$ and $\overline{D}$ are independent.

The question is about the 2-sample t-test. Let $X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$ be *independent* random samples. Notice the variance σ^2 is the same. The usual t statistic for testing $H_0 : \mu_1 = \mu_2$ is

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}}, \text{ where } S_p^2 = \frac{\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2}{n_1 - n_2 - 2}.$$

(a) What is the distribution of $\overline{X} - \overline{Y}$? Just write down the answer.

(b) Now standardize your answer to obtain a standard normal random variable. Call it Z.

(c) Next, you need a "well chosen" chi-squared random variable W to use in the denominator. Propose a random variable W and show it has a chi-squared distribution. What are the degrees of freedom?

continued on page 7

(d) How do you know Z and W are independent? My answer consists of four short statements.

(e) Finally, write $T = \frac{Z}{\sqrt{W/\nu}}$, where ν refers to the degrees of freedom. Simplify, obtaining the usual formula for the test statistic given at the beginning of this question.

continued on page 8

4. (*15 points*) In a study of the effects of combining two blood pressure drugs, patients with high blood pressure are randomly assigned to one of four treatment conditions. They get either Drug A or a placebo, and they get either Drug B or a placebo. So each patient takes two pills a day for 6 weeks. Their blood pressure after 6 weeks is the response variable. Age is a covariate. Use this regression model for the problem:

$$E[Y|\mathbf{X}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \text{ where}$$

- $x_1 = 1$ if the patient got both drugs. Otherwise, $x_1 = 0$.
- $x_2 = 1$ if the patient got Drug A but not B. Otherwise, $x_2 = 0$.
- $x_3 = 1$ if the patient got Drug B but not A. Otherwise, $x_3 = 0$.
- $x_4 = 1$ if the patient got neither drug. Otherwise, $x_4 = 0$.
- $x_5 =$ the patient's age.

(a) Write $E[Y|\mathbf{X}]$ for each treatment combination in the table below.

	Drug B	
Drug A	*Yes*	*No*
Yes		
No		

(b) Give the null hypothesis you would test to answer each question below. The answers are in terms of the β parameters of the regression model. Some of the answers are the same. You may assume that each question begins with "Controlling for the patient's age, ..."

Question	*Null Hypothesis*
Averaging across Drug A versus Placebo, does Drug B affect blood pressure?	
Does the effect of Drug A on blood pressure depend on whether the patient is also taking Drug B?	
Is it better to take both drugs, or neither drug?	
Is it better to take just Drug A, or just Drug B?	
Is Drug A alone better than nothing?	
Is Drug B alone better than nothing?	
Is there a main effect for Drug A?	
Is there a main effect for Drug B?	
Is the average response to just Drug A and just Drug B different from the response to both drugs at once?	
Is there a statistically significant interaction?	
Test both main effects and the interaction, all at the same time.	

continued on page 9

5. (*19 points*) The general mixed linear model is

$$\mathbf{Y} \;=\; \mathbf{X}\boldsymbol{\beta} \;+\; \mathbf{Z}\mathbf{b} \;+\; \boldsymbol{\epsilon}, \text{ where}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.
- $\mathbf{b}$ and $\boldsymbol{\epsilon}$ are independent.

(a) What is the distribution of $\mathbf{Y}$ under this model? Show only as much work as you need to. You may be able to just write down the answer.

(b) Suppose you used the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}$ on the formula sheet to estimate $\boldsymbol{\beta}$.

i. What is the distribution of $\widehat{\boldsymbol{\beta}}$ under the mixed model of this question? Show a bit of work.

ii. Is $\widehat{\boldsymbol{\beta}}$ unbiased under the mixed model of this question? Just answer Yes or No.

iii. Suppose you give a confidence interval for a single regression coefficient β_j using the usual formulas for a fixed effects model, but in fact the true model is the mixed model including the $\mathbf{Zb}$ part. Do you think your confidence interval would be too wide, too narrow, or about right? Why?

continued on page 10

(c) Continuing with Question 5, a one-factor random effects analysis of variance model may be written

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij},$$

where

μ is an unknown constant parameter.

$\tau_j \sim N(0, \sigma^2_\tau)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, with σ^2_τ and σ^2 both unknown.

τ_j and ϵ_{ij} are all independent, $i = 1, \ldots n$ and $j = 1, \ldots, k$. Note that the total number of observations is nk.

Suppose you have $k = 2$ treatments with $n = 3$ observations per treatment, for a total of siz observations. Write the mixed model for this problem in matrix form, showing the *complete* matrices. As a start, $\mathbf{Y}$ is given below.

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{pmatrix} =$$

continued on page 11

6. (*23 points*) The training to be an astronaut is very demanding; most candidates who enter the programme do not finish successfully. Trainers at the Space Agency were able to pre-test a large sample of candidates and then observe whether they completed the training programme successfully. They combined the assessments into a single number called `Pretest` in the output below.

```
> spacemodel = glm(Success ~ Pretest, family=binomial)
> summary(spacemodel)

Call:
glm(formula = Success ~ Pretest, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.6823  -1.0544  -0.5418   1.0624   1.9458  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.74279    1.42083  -4.042 5.30e-05 ***
Pretest      0.05733    0.01420   4.037 5.42e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.84  on 144  degrees of freedom
Residual deviance: 181.16  on 143  degrees of freedom
AIC: 185.16

Number of Fisher Scoring iterations: 4

> Vn = vcov(spacemodel); Vn # Asymptotic covariance matrix
            (Intercept)       Pretest
(Intercept)  2.01875785 -0.0200200107
Pretest     -0.02002001  0.0002017034
```

(a) An extra ten points on the Pretest (ten points, not one) multiplies the estimated odds of Success by ... **Circle your answer below.** The answer is a number.

(b) For a candidate who scores 100 on the Pretest, the estimated probability (not odds) of success is ... **Circle your answer below.** The answer is a number.

continued on page 12

(c) Continuing with Question 6, training potential astronauts is expensive, especially since so many drop out. The Space Agency wants to train only those with at least a 90% chance of success. So the question is, what score on the pre-test would yield a 0.90 probability of success? Denote that point by $x_{.90}$. Suppose you knew the true values of the parameters β_0 and β_1. Give a formula for $x_{.90}$. Show your work. **Circle your answer.**

(d) Give an *estimate* of $x_{.90}$ based on the R output. Your answer is a number. **Circle your answer.**

(e) The estimator of $x_{.90}$ is definitely a random variable. Call it $\widehat{X}_{.90}$. How do you know that the distribution of $\widehat{X}_{.90}$ is asymptotically normal? A rough, informal answer is okay; no formal proof is required.

continued on page 13

(f) Still continuing with Question 6, calculate an estimate of the asymptotic variance of $\widehat{X}_{.90}$. Your answer is a single number. Show your work. **Circle your final answer.**

continued on page 14

(g) Still continuing with Question 6, give a 95% margin of error for your estimate of $x_{.90}$. The confidence interval is the estimate plus or minus the margin of error, but you are only being asked for one number. In case you forgot, the critical value you need is 1.96. **Circle your answer.**

7. (*5 points*) An observational medical study (no random assignment) finds that the more organic food a person eats (as a percent of total calories), the better the person's overall health on average. Briefly discuss at least one omitted variable that could have produced this result.

Total Marks = 100 points