UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2012 Final Examination
STA442H1F/2101HF
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

Aids: Calculator Model(s): Any calculator is okay.
Formula sheet supplied.

Last/Surname (Print): ______________________________

First/Given Name (Print): ______________________________

Student Number: ______________________________

Signature: ______________________________

Qn. #	Value	Score
1	8	
2	10	
3	20	
4	4	
5	25	
6	11	
7	6	
8	16	
Total = 100 Points		

1. (*8 points*) Show that if $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $Y = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom. You need a certain matrix to do this, and you may just use it without proving its existence or any of its properties.

continued on page 3

2. (*10 points*) Let $Y_1, \ldots, Y_n$ be a random sample from a Poisson distribution with parameter $\mu > 0$, so that $E(Y_i) = Var(Y_i) = \mu$. It is suggested that for testing $H_0 : \mu = \mu_0$ against $H_0 : \mu \neq \mu_0$, a good test statistic might be $Z_n = 2\sqrt{n}\left(\sqrt{\overline{Y}_n} - \sqrt{\mu_0}\right)$. Find the limiting distribution of Z_n under the assumption that $H_0 : \mu = \mu_0$ is true. Show your work. For full marks, cite material from the formula sheet when you use it.

continued on page 4

3. (*20 points*) In a political poll, a random sample of n registered voters are to indicate which of two candidates they prefer.

    (a) State a reasonable model for these data, in which the population proportion of registered voters favouring Candidate A is denoted by π. Denote the observations $Y_1, \ldots, Y_n$.

    (b) What is the null hypothesis? Give your answer in symbols.

    (c) What is the asymptotic (that is, approximate large-sample) distribution of $\overline{Y}_n$? You don't have to prove anything; just give the distribution and the parameters.

    For the last two parts of this question, my answers use the symbol "$\dot{\sim}$" to mean "distributed approximately as."

    (d) To test which candidate is preferred by a majority of voters, the test statistic will be $Z_n^2 = \frac{n(\overline{Y}_n - 1/2)^2}{\overline{Y}(1-\overline{Y})}$. Under the null hypothesis, what is the asymptotic distribution of Z_n^2? Justify your answer. Your argument need not be completely rigorous, but I have to be able to follow it.

continued on page 5

(e) Under the alternative hypothesis, $Z_n^2 = \frac{n(\overline{Y}_n - 1/2)^2}{\overline{Y}(1-\overline{Y})}$ has an approximate non-central chi-squared distribution. Find the non-centrality parameter λ. Again, your justification need not be completely rigorous, but I have to be able to follow it. Use the formula sheet.

continued on page 6

(f) Suppose that the true proportion of registered voters favouring candidate A is 0.40. What is the smallest sample size required for a power of at least 0.90? The answer is an integer. You will use a calculator, but this R output will help. *You have a lot more room than you need.*

```
> critval = qchisq(0.95,df=1); critval
[1] 3.841459
> # help(uniroot) Says: "The function uniroot searches the interval from lower
> #                      to upper for a root (i.e., zero) of the function f with
> #                      respect to its first argument.
>
> f = function(lambda) # Specific to power of 0.90 for chi-square(1)
+     { f = 1-pchisq(critval,df=1,ncp=lambda) - 0.90; f}
>
> Lambda = uniroot(f,lower=0,upper=20)$root  # Search between lower and upper
> Lambda
[1] 10.50742
> # Now, what is Lambda? This should settle it.
> 1-pchisq(critval,df=1,ncp=Lambda)
[1] 0.9
```

continued on page 7

4. (*4 points*) Naturally, the normal distribution is a member of the natural exponential family. Identify the natural link function by fitting the normal density $f(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(y-\mu)^2}{\sigma^2}}$ into the framework of the natural exponential density. Your objective is to show convincingly what the natural link function is, and that's all.

continued on page 8

5. (*25 points*) This question will be a lot easier if you remember that if $X \sim \chi^2(\nu)$, then $E(X) = \nu$ and $Var(X) = 2\nu$. You don't have to prove this; just use it.

   For the usual linear regression model with normal errors, σ^2 is usually estimated with $MSE = \frac{SSE}{n-p}$.

   (a) Show that MSE is an unbiased estimator of σ^2.

   (b) Show that MSE is a consistent estimator of σ^2.

continued on page 9

(c) Under the usual regression model what is the joint distribution of $\epsilon_1, \ldots, \epsilon_n$?

(d) Let $T_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2$. What is $E(T_n)$? Show a little work.

(e) How do you know that $T_n \stackrel{p}{\rightarrow} \sigma^2$?

(f) Show that $Var(T_n) < Var(MSE)$.

(g) So it would appear that T_n is a better estimator of σ^2 than MSE is, since they are both unbiased and the variance of T_n is lower. Why do you think MSE is used in regression analysis instead of T_n?

continued on page 10

6. (*11 points*) Information about a sample of traveling sales representatives include years of experience, technical knowledge (measured by a test), gender (Female or Male), and how many major contracts they got their clients to sign last month. There are quite a few zeros even for good salespeople, so the data are definitely not normal. *The quantitative covariates have been centered*, so that $x_1 = 0$ actually means that the first covariate equals $\overline{x}_1$, the mean of x_1 for the entire sample.

   (a) State a reasonable model for these data, by giving

      i. A probability distribution for the response variable.

      ii. A link function and a linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$.

   You don't need to give any justification for your model. Please make it a proportional means model for now.

   (b) Make a table with two rows, showing how you would set up indicator dummy variables for sex, with Males as the reference category. Add another column showing the expected number of sales (signed contracts).

   (c) Controlling for experience and technical knowledge, are male and female sales representatives equally effective? What null hypothesis would you test to answer this question?

   (d) Allowing for experience and technical knowledge, the expected number of signed contracts for male sales representatives is ___________ as great as the expected number for female sales representatives. Answer in terms of the parameters in your model. Be careful not to get the answer backwards!

   (e) You want to estimate the number of sales for a female sales representative of average experience and technical knowledge (average for the whole sales force, that is). Give your answer in terms of $\widehat{\beta}$ values.

   (f) Finally, you want to test whether the proportional means model was a good idea. Maybe we should have done this first. Give the linear predictor (regression equation) for the full model, and *state the null hypothesis in symbols.*

continued on page 11

7. (*6 points*) Arsenic is a powerful poison, which is why it has been used on farms for many years to kill insects. Even in very small amounts, arsenic can cause cancer, and recently it has been found that rice and foods made from rice tend to be very high in arsenic. Brown rice is worse, by the way.

   In a controlled experiment, pots of rice were prepared by either washing the rice first or not, and by cooking the rice in either a low, a medium or a high amount of water. The response variable is amount of arsenic in the cooked rice. The table below shows cell means.

	Amount of Water		
	Low	Medium	High
Washed	μ_{11}	μ_{12}	μ_{13}
Unwashed	μ_{21}	μ_{22}	μ_{23}

   Just write your answers to the questions below. You don't have to show any work.

   (a) If you wanted to test whether the effect of washing the rice depended on how much water you cook it in, what is the null hypothesis? Give your answer in terms of μ_{ij} values.

   (b) If you wanted to test whether the effect of the amount of water used to cook the rice depends on whether you wash it, what is the null hypothesis? Give your answer in terms of μ_{ij} values.

8. (*16 points*) This question is based upon the following R output. Remember that there were three courses, Catch-up, Elite and Mainstream.

```
> head(math)
  hsgpa hsengl hscalc   course passed outcome
1  78.0     80    Yes Mainstrm     No  Failed
2  66.0     75    Yes Mainstrm    Yes  Passed
3  80.2     70    Yes Mainstrm    Yes  Passed
4  81.7     67    Yes Mainstrm    Yes  Passed
5  86.8     80    Yes Mainstrm    Yes  Passed
6  76.7     75    Yes Mainstrm    Yes  Passed
> attach(math)
> model9 = glm(passed ~ hsgpa + course, family=binomial)
> model10 = glm(passed ~ hsgpa + course + hsgpa:course, family=binomial)
> anova(model9,model10,test='Chisq')
Analysis of Deviance Table

Model 1: passed ~ hsgpa + course
Model 2: passed ~ hsgpa + course + hsgpa:course
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       390     428.90
2       388     428.45  2  0.44679   0.7998
```

continued on page 12

```
> summary(model9)

Call:
glm(formula = passed ~ hsgpa + course, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5030  -0.9590   0.4413   0.8826   2.0971

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -16.02245    2.06714  -7.751 9.12e-15 ***
hsgpa            0.19229    0.02607   7.376 1.63e-13 ***
courseElite      2.21871    0.63711   3.482 0.000497 ***
courseMainstrm   1.28489    0.45300   2.836 0.004562 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 530.66  on 393  degrees of freedom
Residual deviance: 428.90  on 390  degrees of freedom
AIC: 436.9

Number of Fisher Scoring iterations: 4

> anova(model9,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: passed

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     393     530.66
hsgpa   1   87.221       392     443.43 < 2.2e-16 ***
course  2   14.539       390     428.90 0.0006964 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> vcov(model9)
               (Intercept)         hsgpa  courseElite courseMainstrm
(Intercept)     4.27307885 -0.0526881470 -0.230629028  -0.1359942135
hsgpa          -0.05268815  0.0006796534  0.000542363  -0.0006783835
courseElite    -0.23062903  0.0005423630  0.405914412   0.1880425694
courseMainstrm -0.13599421 -0.0006783835  0.188042569   0.2052058666
```

continued on page 13

(a) Consider the first `anova` command.

i. Write the linear predictor (regression model for the log odds) *for the full model.* You do not need to say how the dummy variables are defined, even though you probably know.

ii. In terms of the β values of your linear predictor, what is the null hypothesis?

iii. Does this result make you doubt the proportional odds model? Just answer Yes or No.

(b) Based on the output for Model 9, estimate the probability of passing the course for a student in the Mainstream course with a High School Grade Point Average of 90%. The answer is a number between zero and one. Show some work. **Circle your final answer.**

(c) Controlling for High School Grade Point Average, the odds of passing the coure are estimated to be ___________ times as great for a student in the Elite course as for a student in the Mainstream course.

continued on page 14

(d) Suppose we want to *test* whether, controlling for High School Grade Point Average, the chances of passing the course are different for students in the Elite and Mainstream courses.

i. Calculate a test statistic. The answer is a number. Personally, I think Z is easier than χ^2.

ii. In plain, non-statistical language, what do you conclude from this test?

Total Marks = 100 points