UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2016 Final Examination
STA442H1F/2101H1F
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

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

Last/Family Name (Print): _____

First/Given Name (Print): _____

Student Number: _____

Signature: _____

Qn. #	Value	Score
1	10	
2	15	
3	10	
4	15	
5	10	
6	5	
7	10	
8	15	
9	10	
Total = 100 Points		

1. (*10 points*) A randomization test based on 1,000 re-arrangements of the data yields a p-value of $p = 0.034$. Of course this is just an estimate of the p-value of the corresponding permutation test. Give a 95% confidence interval for the permutation test p-value.

   The final answer is a pair of numbers, a lower confidence limit and an upper confidence limit. Show a little work. **Circle your final answer.** Maybe the number 1.96 should have been on the formula sheet.

continued on page 3

2. (*15 points*) Let $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is symmetric and positive definite. Show that $\mathbf{w}^\top \boldsymbol{\Sigma}^{-1} \mathbf{w}$ has a non-central chi-squared distribution with degrees of freedom p and non-centrality parameter $\lambda = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$. You may use anything on the formula sheet except what you are proving.

continued on page 4

3. (*10 points*) When customers call Pizza Gigantico to order pizza, usually the call is answered right away, with zero delay. But if all the operators are taking calls, the customer is put on hold until an operator is available. During one peak period, the wait times for 100 customers are recorded. The data are times on hold for each customer in milliseconds. There are quite a few zeros, but *no ties at any value other than zero.*

   (a) State a reasonable model for these data. *You do not have to justify your model* (though you may make a few remarks if you wish), but the data have certain features that must be reflected in your model for full marks. Please don't get too fancy; the exact distribution you suggest does not matter as long as it's not impossible.

   (b) What is the parameter space for your model?

continued on page 5

4. (*15 points*) In a simple regression through the origin, the response variable Y_i is related to a true explanatory variable T_i, but T_i is latent and unobservable. All we can see is X_i, which is T_i contaminated with measurement error. The model is

$$\begin{aligned} Y_i &= \beta T_i + \epsilon_i \\ X_i &= T_i + \delta_i \end{aligned}$$

where all expected values are zero, $Var(\epsilon_i) = \sigma^2_\epsilon$, $Var(\delta_i) = \sigma^2_\delta$, $Var(T_i) = \sigma^2_t$, and all the T_i, ϵ_i and δ_i variables are independent for $i = 1, \ldots, n$.

Naturally, the data analyst uses the data that are observable, and calculates the least squares estimate

$$\widehat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

Is $\widehat{\beta}_n$ a consistent estimator of β? Answer Yes or No and prove your answer. *Circle the word* ***Yes*** *or the word* ***No****.*

continued on page 6

5. (*10 points*) In a study conducted by a candy company, children whose parents let them volunteer were asked to wait in the waiting room. On the table was a bowl of candy, and they were told to have as much as they liked while they waited. Actually, this was the experiment. Children were randomly assigned to one of three different recipes (labelled A, B and C), and the amount they ate was weighed after they left. Child's age was a covariate; bigger kids tend to eat more.

   (a) Write a regression equation for this problem ($Y_i = \ldots$ and so on). There is a distinct possibility that what recipe is liked most on average might depend on the child's age, and your model should allow for this. You do not need to say exactly how your dummy variables are defined yet. You will do that in the next part.

   (b) Make a table with three rows, one for each experimental treatment. Make columns showing how you would set up the dummy variables. Add a last column showing $E(Y_i|\mathbf{x}_i)$.

   (c) Does difference in preference between recipe A and recipe B (just those two) depend on the child's age? In terms of the β values of your model, what null hypothesis would you test to answer this question?

continued on page 7

6. (*5 points*) Let the $p \times 1$ random vector $\mathbf{w}$ have expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{A}$ be an $m \times p$ matrix of constants, and let $\mathbf{B}$ be an $n \times p$ matrix of constants. Derive a nice simple expression for the $m \times n$ matrix of covariances $C(\mathbf{Aw}, \mathbf{Bw})$. Show your work. This will be useful in the next question.

7. (*10 points*) To establish the independence of the residuals and estimated regression coefficients under normality, calculate $C(\widehat{\boldsymbol{\beta}}, \mathbf{e})$ and simplify. Show your work. **Circle your final answer.**

continued on page 8

8. (*15 points*) Here is R output for the Pig Weight Study. The questions are on the following page.

```
> # Pig weight data
> pigs = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/pigweight.data.txt")
> head(pigs); attach(pigs)
  Drug Momweight Dadweight Pigweight
1    1    133.55    172.97     71.99
2    1    143.65    183.32     76.76
3    1    130.27    186.53     72.22

> # Make dummy variables and fit the model
> n = length(Drug)
> d1=d2=d3 = numeric(n)
> d1[Drug==1] = 1; d2[Drug==2] = 1; d3[Drug==3] = 1
> fullmodel = lm(Pigweight ~ d1+d3 + Momweight + Dadweight); summary(fullmodel)

Call:
lm(formula = Pigweight ~ d1 + d3 + Momweight + Dadweight)

Residuals:
   Min     1Q Median     3Q    Max
-3.905 -1.174  0.187  1.351  3.657

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)  5.87607    9.19763   0.639     0.52499
d1           1.60557    0.52788   3.042     0.00331 **
d3           0.90077    0.53916   1.671     0.09925 .
Momweight    0.26363    0.04727   5.578 0.000000428 ***
Dadweight    0.17442    0.03465   5.034 0.000003580 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.855 on 70 degrees of freedom
Multiple R-squared:  0.4561,Adjusted R-squared:  0.425
F-statistic: 14.67 on 4 and 70 DF,  p-value: 0.000000009393

> # Do a few tests
> L1 = rbind(c(0,1,0,0,0),
+            c(0,0,1,0,0))
> ftest(fullmodel,L=L1)
          F         df1         df2     p-value
 4.63559465  2.00000000 70.00000000  0.01286451
>
> L2 = rbind(c(0,0,0,1,0),
+            c(0,0,0,0,1))
> ftest(fullmodel,L=L2)
                 F                df1                df2            p-value
25.902260245782465  2.000000000000000 70.000000000000000  0.000000003803752
>
> L3 = rbind(c(0, 1, 0, 0, 0)); ftest(fullmodel,L=L3)
          F          df1          df2      p-value
 9.251054028  1.000000000 70.000000000  0.003311102
>
> L4 = rbind(c(0, 0, 1, 0, 0)); ftest(fullmodel,L=L4)
          F         df1         df2     p-value
 2.79120433  1.00000000 70.00000000  0.09924863
>
> L5 = rbind(c(0, 1,-1, 0, 0)); ftest(fullmodel,L=L5)
         F        df1        df2    p-value
 1.7770226  1.0000000 70.0000000  0.1868378
>
> L6 = rbind(c(0, 0, 0, 1,-1)); ftest(fullmodel,L=L6)
         F        df1        df2    p-value
 2.5354089  1.0000000 70.0000000  0.1158241
>
> qt(0.975,df=70) # Critical value
[1] 1.994437
```

continued on page 9

(a) Give a 95% confidence interval for the difference in expected weight between drug treatments 2 and 3 for a pig if the mother weighs 140 pounds and the father weighs 185 pounds. The answer is a pair of numbers, a lower confidence limit and an upper confidence limit. **Circle the numbers**.

(b) For each of the questions below, give the value of the t or F statistic (a number from the printout), and indicate whether or not you reject the null hypothesis.

i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

ii. Allowing for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. Correcting for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iv. Holding mother's weight and father's weight constant, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

(c) We can assume that farmers want their pigs to weigh a lot. In plain, non-statistical language, can you offer some advice to a farmer based on these data? Remember, the farmer must be able to understand your answer or it is worthless.

continued on page 10

9. (*10 points*) In the Tooth Growth data, guinea pigs were randomly assigned to a combination of Supplement Type (Vitamin C or orange juice), and dosage level. The response variable was tooth length after a specified time. Please answer questions based on the following R output. The questions appear after the printout.

```
> head(ToothGrowth); attach(ToothGrowth)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5

> dose = factor(dose)
> # Look at table of means
> meanz = aggregate(len ~ supp+dose, FUN=mean)
> meantable = meanz[,3]; dim(meantable) = c(2,3)
> dimnames(meantable) = list(c("OJ","VC"),c("0.5","1.0","2.0"))
> meantable
     0.5   1.0   2.0
OJ 13.23 22.70 26.06
VC  7.98 16.77 26.14
>
> # Tests: Use effect coding
> contrasts(supp)=contr.sum(2); contrasts(dose)=contr.sum(3)
> contrasts(dose)
    [,1] [,2]
0.5    1    0
1      0    1
2     -1   -1
> full1 = lm(len ~ supp + dose + supp:dose); summary(full1)

Call:
lm(formula = len ~ supp + dose + supp:dose)

Residuals:
   Min     1Q Median     3Q    Max
 -8.20  -2.72  -0.27   2.65   8.27

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.8133     0.4688  40.130  < 2e-16 ***
supp1         1.8500     0.4688   3.946 0.000231 ***
dose1        -8.2083     0.6630 -12.381  < 2e-16 ***
dose2         0.9217     0.6630   1.390 0.170190
supp1:dose1   0.7750     0.6630   1.169 0.247568
supp1:dose2   1.1150     0.6630   1.682 0.098394 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.631 on 54 degrees of freedom
Multiple R-squared:  0.7937,Adjusted R-squared:  0.7746
F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

continued on page 11

```
> anova(full1) # Design is balanced so this works.
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq F value    Pr(>F)
supp       1  205.35  205.35  15.572 0.0002312 ***
dose       2 2426.43 1213.22  92.000 < 2.2e-16 ***
supp:dose  2  108.32   54.16   4.107 0.0218603 *
Residuals 54  712.11   13.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) Write the regression equation used in the R printout.

(b) We are interested in whether the effect of Supplement Type depends on dosage level.

i. In terms of β values from the regression equation, what null hypothesis would you test?

ii. Give the value of the test statistic (t or F). The answer is a number from the printout.

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

iv. Do you conclude that the effect of Supplement Type depends on dosage level? Answer Yes or No.

v. What proportion of the remaining variation does this effect explain, after alowing for other effects in the model? The answer is a number. **Circle the number**.