Draft 1

STA442H1F/2101H1F Midterm Test

Methods of Applied Statistics
Jerry Brunner
Duration - 2 hours

Aids Allowed: Any calculator without wireless capability.
Formula sheet supplied.

Last/Family Name (Print): ______________________

First/Given Name (Print): Jerry

Student Number: ______________________

Signature: ______________________

Qn. #	Value	Score
1		
2		
3		
4		
5		
6		
Total = 100 Points		

1. (15 *points*) Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance μ^2.

3pts (a) What is the ~~asymptotic~~ distribution of $\overline{X}_n$? You do not need to show any work. Just write down the answer.

$$\overline{X}_n \sim N\left(\mu, \frac{\mu^2}{n}\right)$$

12 pts (b) Find a variance-stabilizing transformation. That is, find a function $g(x)$ such that the limiting distribution of

$$Y_n = \sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right)$$

does not depend on μ. Show your work.

By delta method, need $g'(\mu)^2 \cdot \mu^2 = 1$

So seek $g(\mu)$ with $g'(\mu) = \frac{1}{\mu}$, so the function is $g(x) = \ln(x)$

students could get this by separation of variables, but they don't have to

$$\frac{dg}{d\mu} = \frac{1}{\mu} \Rightarrow \int dg = \int \frac{1}{\mu}\, d\mu = \log(\mu)$$

5

continued on page 3

2. (15 *points*) A linear null hypothesis like $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ may be written in infinitely many equivalent ways. The formula sheet has a formula for the Wald test statistic W_n, which may be used to test such a linear null hypothesis. Let $\mathbf{A}$ be an $r \times r$ matrix with an inverse. Clearly, $H_0 : \mathbf{AL}\boldsymbol{\theta} = \mathbf{Ah}$ is true if and only if $\mathbf{L}\boldsymbol{\theta} = \mathbf{h}$, so $H_0 : \mathbf{AL}\boldsymbol{\theta} = \mathbf{Ah}$ is just another way of writing the same null hypothesis. How does this way of re-expressing H_0 affect W_n? Show your work. *After the calculation, answer the question.* How does this way of re-expressing H_0 affect W_n?

For testing $H_0 : AL\theta = Ah$,

$$
\begin{aligned}
W_n &= (AL\hat{\theta}_n - Ah)^T (AL\hat{V}_n(AL)^T)^{-1}(AL\hat{\theta}_n - Ah) \\
&= [A(L\hat{\theta}_n - h)]^T (AL\hat{V}_n L^T A^T)^{-1} A(L\hat{\theta}_n - h) \\
&= (L\hat{\theta}_n - h)^T \underbrace{A^T (A^T)^{-1}}_{I} (L\hat{V}_n L^T)^{-1} \underbrace{A^{-1}A}_{I} (L\hat{\theta}_n - h) \\
&= (L\hat{\theta}_n - h)^T (L\hat{V}_n L^T)^{-1} (L\hat{\theta}_n - h)
\end{aligned}
$$

which is the test statistic for $H_0 : L\theta = h$.

The transformation has no effect on W_n.

5.30 or so, say 6

continued on page 4

15

3. (*points*) Independently for $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2_\epsilon$, and ϵ_i is independent of X_i. Let

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

*!

Is $\widehat{\beta}_n$ a consistent estimator of β_1? Answer Yes or No and prove your answer.

$$E(X_i Y_i) = E(X_i(\beta_0 + \beta_1 X_i + \varepsilon_i)) =$$

$$= \beta_0 E(X_i) + \beta_1 E(X_i^2) + E(X_i)E(\varepsilon_i)$$

$$= \beta_0 \mu + \beta_1(\sigma^2 + \mu^2) + 0$$

And $E(X_i^2) = \sigma^2 + \mu^2$, so

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} \xrightarrow{as} \frac{E(X_i Y_i)}{E(X_i^2)}$$ By LLN and continuous mapping

$$= \frac{\beta_0 \mu + \beta_1(\sigma^2 + \mu^2)}{\sigma^2 + \mu^2} = \beta_1 + \frac{\beta_0 \mu}{\sigma^2 + \mu^2}$$

No, not consistent

15/30

continued on page 5

4. (26 *points*) Pigs are routinely given large doses of antibiotics even when they show no signs of illness, to protect their health under unsanitary conditions. Pigs were randomly assigned to one of three antibiotic drugs. Dressed weight (weight of the pig after slaughter and removal of head, intestines and skin) was the response variable. Explanatory variables are Drug type, Mother's live adult weight and Father's live adult weight.

2 pts (a) Write the regression equation for the full model, including ϵ_i. Let x_1 = mother's weight and x_2 = father's weight. There are no product terms.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 d_1 + \beta_4 d_2 + \varepsilon$$

8 pts (b) Make a table with one row for every drug, with columns showing how the dummy variables were defined. Make another column giving $E(y|\mathbf{x})$ for each drug.

Drug	d_1	d_2	$E(y)$
1	1	0	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3$
2	0	1	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4$
3	0	0	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$

4 pts (c) What is the expected dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds? Your answer is a formula involving some β values.

$$E(y) = \beta_0 + 140\beta_1 + 185\beta_2 + \beta_4$$

(d) In symbols, give the null hypotheses you would test to answer the following questions. Your answers are statements involving the β values from your regression equation.

3 i. Allowing for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig? $H_0: \beta_3 = \beta_4 = 0$

3 ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2? $H_0: \beta_3 = \beta_4$

3 iii. Correcting for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3? $H_0: \beta_3 = 0$

3 iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3? $H_0: \beta_4 = 0$

continued on page 6

5. (15 *points*) In the geometric distribution, the random variable X is like the number of tosses of a coin required to get the first head. If $P(\text{Head}) = p$, then $P(X = x) = (1-p)^{x-1}p$ for $x = 1, 2, \ldots$. This might not be a bad model for the day on which some disaster happens.

Accordingly, let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with parameter p, and let the prior distribution of p be uniform on the interval from zero to one. Find the posterior density. Show your work. The answer must be a density that would integrate to one, so it includes the normalizing constant. (You do not need to actually carry out the integration.) **Circle your final answer**.

$$\pi(p|x) \propto f(x|p)\pi(p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p \cdot 1$$

$$= (1-p)^{\Sigma x_i - n} p^n = p^{(n+1)-1} (1-p)^{(\Sigma x + 1 - n) - 1}$$

From the formula sheet, this is Beta with $\alpha = n+1$ and $\beta = \sum_{i=1}^{n} x_i - n + 1$, so

$$\pi(p|x) = \frac{\Gamma(\Sigma x_i + 2)}{\Gamma(n+1)\Gamma(\Sigma x - n + 1)} p^{(n+1)-1} (1-p)^{(\Sigma x_i - n + 1) - 1}$$

27

continued on page 7

6. (14 *points*) The following R session shows input and output for the infamous Mystery data: $f(x) = \frac{\theta e^{\theta(x-\mu)}}{(1+e^{\theta(x-\mu)})^2}$.

```
> # Logistic mystery
> rm(list=ls()); options(scipen=999)
> mystery = scan("http://www.utstat.toronto.edu/~brunner/data/legal/mystery.data.txt")
Read 50 items
> c(mean(mystery), sd(mystery))
[1] 3.380600 2.086256
> mloglike = function(theta,x)
+       {
+       mu = theta[1]; T = theta[2]
+       n = length(x); xbar = mean(x)
+       value =  n * (log(T) + T*(xbar-mu)) -
+               2 * sum(log(1+exp(T*(x-mu))))
+       value = - value # For MINUS log likelihood
+       return(value)
+       } # End function loglike
> # Start at mu = mean and theta = standard deviation
> begin = c(3.381,2.09)
> search = nlm(mloglike,begin,hessian=TRUE,x=mystery); search
$minimum
[1] 106.7216

$estimate
[1] 3.3392008 0.8760226

$gradient
[1] -0.000007826352 -0.000014949819

$hessian
           [,1]       [,2]
[1,] 12.8499586  0.3085511
[2,]  0.3085511 92.3628136

$code
[1] 1

$iterations
[1] 10

> solve(search$hessian) # Yields the inverse
              [,1]          [,2]
[1,]  0.0778275056 -0.0002599938
[2,] -0.0002599938  0.0108277366
>
```

The questions appear on the final page.

continued on page 8

7 pts (a) Carry out a large-sample Z test for $H_0 : \mu = 0$ at $\alpha = 0.05$. If you do not recall the critical value, it's 1.96. Circle the value of the test statistic (a number), and state whether you reject H_0.

$$Z = \frac{3.34 - 0}{\sqrt{0.0778}} = 11.97$$

Don't deduct for rounding error

> 1.96 so Reject H_0

7 pts (b) Give a large-sample 95% confidence interval for θ. You do *not* need to derive it. The answer is a pair of numbers, a lower confidence limit and an upper confidence limit.

$$0.876 \pm 1.96\sqrt{.010828} = 0.876 \pm 1.96(0.10406)$$

$= 0.876 \pm 0.204$, so CI is

$(0.636, 1.08)$