Student ID

THE AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS

Mid-Semester Exam
Semester 1, 2016

STATXXXX Statistical Inference

Study Time: 15 minutes
Writing Time: 1 $\frac{1}{2}$ hours

Permitted materials:
A4 pages (Two sheets) with handwritten notes on both sides
Paper-based Dictionary, no approval required (must be clear of ALL annotations)
Calculator (Any - programmable or not)

Marks

Question 1	Question 2	Question 3	Question 4	Total

INSTRUCTIONS:

1.) This exam paper comprises a total of 22 pages. Please ensure your paper has the correct number of pages.

2.) The exam includes a total of 4 questions.

3.) After each question there are four blank pages to write your solutions. You may use both sides of each page to write your solutions.

4.) Each question appears on the following pages [marks are indicated]:

- Question 1 is on page 3 [**10 marks**].
- Question 2 is on page 8 [**30 marks**].
- Question 3 is on page 13 [**30 marks**].
- Question 4 is on page 18 [**30 marks**].

5.) Include all workings for each question, as marks will not be awarded for answers that do not include workings.

6.) Draw a $\boxed{\text{box}}$ around each final answer.

7.) Ensure you include your student number on this exam book.

8.) A table of probability distributions is provided with the exam.

Total Marks = 100

This exam is a redeemable exam. It will be worth either 20% or 0% of your final grade based on your final exam mark.

Question 1 [**10 marks**]: A researcher from the College of Medicine states: "I just fit a least-squares model to determine the effects of age and gender on blood pressure." Clearly discuss the appropriateness of this statement.

Sol: Based on what we have learned, there is no least-squares model. For a particular model, we might use least-squares as a way to estimate parameters in the model, thus the model and estimation are separate components. Here we have a particular model, in this case likely a linear regression model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{normal}(0, \sigma^2) \end{aligned}$$

Where:

- Y_i is blood pressure for individual i.
- $x_{1,i} = 1$ if individual i is female and 0 otherwise.
- $x_{2,i}$ is the age of individual i.

To estimate β_0, β_1, and β_2 we may consider using least-squares:

$$min_{\beta_0,\beta_1,\beta_2} \left(Y_i - \beta_0 - \beta_1 x_{1,i} - \beta_2 x_{2,i}\right)^2.$$

Question 2 [**30 marks**]: Let X and Y be independent random variables, where $X \sim \text{gamma}(\alpha = r, \beta = 1)$ and $Y \sim \text{gamma}(\alpha = s, \beta = 1)$. Consider the following random variables based on X and Y:

$$\begin{aligned} Z_1 &= X + Y \\ Z_2 &= \frac{X}{X+Y} \end{aligned}$$

- $Note: E[X] = r; \quad V[X] = r; \quad E[Y] = s; \quad V[Y] = s.$

a. [**10 marks**] Determine the distributions of Z_1 and Z_2.

Sol: Let's solve for two functions, one for X and one for Y in terms of Z_1 and Z_2:

$$\begin{aligned} X &= Z_1 Z_2 = g_1^{-1}(z_1, z_2) \\ Y &= Z_1 - Z_1 Z_2 = g_2^{-1}(z_1, z_2) \end{aligned}$$

Now let's get the determinant of the Jacobian and take the absolute value:

$$\begin{aligned} |J| &= \begin{vmatrix} \frac{\partial X}{\partial Z_1} & \frac{\partial X}{\partial Z_2} \\ \\ \frac{\partial Y}{\partial Z_1} & \frac{\partial Y}{\partial Z_2} \end{vmatrix} \\ &= \begin{vmatrix} Z_2 & Z_1 \\ \\ 1 - Z_2 & -Z_1 \end{vmatrix} \\ &= |-Z_2 Z_1 - (1 - Z_2) Z_1| = |-Z_2 Z_1 - Z_1 + Z_1 Z_2| = |-Z_1| \end{aligned}$$

Now let's get the joint density for Z_1, Z_2, recall that X and Y are independent:

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_X(g_1^{-1}(z_1, z_2), g_2^{-1}(z_1, z_2)) \; f_Y(g_1^{-1}(z_1, z_2), g_2^{-1}(z_1, z_2)) \; |J| \\ &= \frac{1}{\Gamma(r)} (z_1 z_2)^{r-1} \, exp\left(-z_1 z_2\right) \; \frac{1}{\Gamma(s)} (z_1 - z_1 z_2)^{s-1} \, exp\left(-z_1 + z_1 z_2\right) \; |-z_1| \\ &= \frac{1}{\Gamma(r)\Gamma(s)} (z_1)^{r-1} (z_2)^{r-1} \left(z_1(1 - z_2)\right)^{s-1} exp\left(-z_1 + z_1 z_2 - z_1 z_2\right) \; |-z_1| \\ &= \frac{1}{\Gamma(r)\Gamma(s)} (z_1)^{r-1} (z_2)^{r-1} (z_1)^{s-1} (1 - z_2)^{s-1} exp\left(-z_1\right) \; |-z_1| \\ &= (z_1)^{r+s-1} exp\left(-z_1\right) \frac{1}{\Gamma(r)\Gamma(s)} (z_2)^{r-1} (1 - z_2)^{s-1} \\ &= \frac{\Gamma(r+s)}{\Gamma(r+s)} (z_1)^{r+s-1} exp\left(-z_1\right) \frac{1}{\Gamma(r)\Gamma(s)} (z_2)^{r-1} (1 - z_2)^{s-1} \\ &= \underbrace{\frac{1}{\Gamma(r+s)} (z_1)^{r+s-1} exp\left(-z_1\right)}_{f_{Z_1}(z_1)} \underbrace{\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} (z_2)^{r-1} (1 - z_2)^{s-1}}_{f_{Z_2}(z_2)} \end{aligned}$$

From the table we can see:

$$Z_1 \sim \text{gamma}(r+s, 1), \;\; Z_2 \sim \text{beta}(r, s)$$

b. [**5 marks**] Show that Z_1 and Z_2 are independent.

Sol: As Z_1 and Z_2 are not linear combinations of **normally distributed** random variables, then we can not use the fact that the $Cov(Z_1, Z_2) = 0$ to say that Z_1 and Z_2 are independent. To show independence we need to show that we can partition the joint pdf of Z_1,Z_2 into the pdf of Z_1 times the pdf of Z_2.

$$f_{Z_1,Z_2}(z_1, z_2) = f_{Z_1}(z_1) \times f_{Z_2}(z_2)$$

We did this in part (a). Thus Z_1 and Z_2 are independent.

c. [**5 marks**] Determine the means and variances for Z_1 and Z_2.

Sol: From the table we can determine the means and variances:

$$E(Z_1) = r+s, \quad V(Z_1) = r+s, \quad E(Z_2) = \frac{r}{r+s}, \quad V(Z_2) = \frac{rs}{(r+s)^2(r+s+1)}.$$

d. [**10 marks**] Write pseudo-code to determine a direct or indirect computational method to generate random samples of Z_1 and Z_2. You may assume that you are able to generate standard uniform random variables [i.e. $U \sim$ uniform$(0, 1)$]. Additionally, you may assume that r and s are positive integers.

Sol: The simplest way to do this is generate an X and generate a Y and then compute Z_1 and Z_2. If $U \sim$ uniform$(0, 1)$ then $A = -log(U) \sim exponential(1)$:

Let's solve for $U \Rightarrow U = exp(-A) \quad \Rightarrow \frac{d}{dA} = -1exp(-A) = g^{-1}(A)$.

$$\begin{aligned} f_A(a) &= f_u(g^{-1}(a))\,|-1exp(-a)| \\ &= 1\;\,|exp(-a)| \\ &= exp(-a) \end{aligned}$$

Now let $B = \sum_i^k A_i$. Let's use moment generating functions (you could get the mgf from the table):

$$
\begin{aligned}
M_A(t) &= E\left(exp(ta)\right) = \int_0^\infty exp(ta)exp(-a)db \\
&= \int_0^\infty exp(ta-a)da \\
&= \int_0^\infty exp(-a(1-t))da \\
&= \frac{1}{1-t}\int_0^\infty (1-t)exp(-a(1-t))da \\
&= \frac{1}{1-t} \times 1 = \frac{1}{1-t}
\end{aligned}
$$

Now let's get the MGF of B:

$$
M_{\sum B_i}(t) = \left[\frac{1}{1-t}\right]^k
$$

From the table we see that $B \sim \text{gamma}(k, 1)$.

Algorithm 1 Generate Samples for Z_1 and Z_2

Let N be the number of samples we wish to generate
Let $out.z1$ be a vector of length N
Let $out.z2$ be a vector of length N
for n in 1:N **do**
    for j in 1:r **do**
        Generate U_j from a uniform (0,1)
    Let $X = -\sum_{j=1}^{r} \log(U_j)$
    for k in 1:s **do**
        Generate U_k from a uniform (0,1)
    Let $X = -\sum_{k=1}^{s} \log(U_k)$
    Compute $Z_1 = X + Y$
    Compute $Z_2 = X/(X + Y)$
    Store Z_1 in $out.z1$
    Store Z_2 in $out.z2$
return $out.z1, out.z2$

Question 3 [**30 marks**]: An original method for generating random standard normal variables based on random uniform variables was through the following transformation:

$$\begin{aligned} X &= \sum_{i=1}^{12} U_i - 6 \\ U_i &\overset{\text{iid}}{\sim} \text{uniform}(0,1) \end{aligned}$$

a. [**7 marks**] What is the moment generating function (mgf) for X? Use the mgf to determine the $E[X]$.

Sol: Let's get the moment generating function:

$$\begin{aligned} M_{\sum_{i=1}^{12} U_i - 6}(t) &= E\left[exp\left(\sum_{i=1}^{12} U_i - 6)t\right)\right] = E\left[exp(-6t)exp\left(\sum_{i=1}^{12} U_i t\right)\right] \\ &= exp(-6t)E\left[exp\left(\sum_{i=1}^{12} U_i t\right)\right] = exp(-6t)E\left[exp\left(U_1 t\right) \times \cdots \times exp\left(U_{12} t\right)\right] \\ &= exp(-6t)E\left[exp\left(U_1 t\right)\right] \times \cdots \times E\left[exp\left(U_{12} t\right)\right] \\ &= exp(-6t)E\left[exp\left(U_i t\right)\right]^{12} = exp(-6t)\left[M_{U_i}(t)\right]^{12} \end{aligned}$$

Now (you could use the table to get this):

$$M_{U_i}(t) = E\left[exp\left(U_i t\right)\right] = \int_0^1 exp(tx)1dx = \frac{1}{t}exp(tx)\Big|_0^1 = \frac{exp(t)-1}{t}$$

This leads to:

$$M_{\sum_{i=1}^{12} U_i - 6}(t) = exp(-6t)\left[M_{U_i}(t)\right]^{12} = exp(-6t)\left[\frac{exp(t)-1}{t}\right]^{12}$$

Now we need to differentiate this:

$$\begin{aligned} M'_{\sum_{i=1}^{12} U_i - 6}(t) &= -6\ exp(-6t)\ \left[\frac{exp(t)-1}{t}\right]^{12} \\ &\quad + exp(-6t)\ 12\ \left[\frac{exp(t)-1}{t}\right]^{11}\left[\frac{exp(t)t - exp(t)+1}{t^2}\right] \end{aligned}$$

We can't just set $t = 0$, so let's use a Taylor's series expansion around 0 for the exponential functions (or you can use L'Hopital's rule):

$$exp(t) = 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \cdots$$

$$exp(at) = 1 + at + \frac{a^2t^2}{2} + \frac{a^3t^3}{6} + \cdots$$

$$\begin{aligned}
M'_{\sum_{i=1}^{12} U_i - 6}(t) &= -6\left[1 - 6t + 18t^2 - 36t^3 + \cdots\right]\ \left[1 + t/2 + t^2/6 + \cdots\right]^{12} \\
&\quad +12\left[1 - 6t + 18t^2 - 36t^3 + \cdots\right]\left[1 + t/2 + t^2/6 + \cdots\right]^{11}\left[\frac{\frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{8} + \cdots}{t^2}\right] \\
\text{Now when } t = 0 \Rightarrow &= -6(1)(1) \\
&\quad +12(1)(1)\left[\frac{1}{2} + \frac{t}{3} + \frac{t^2}{8} + \cdots\right] \\
&= -6 + 12(1/2) = -6 + 6 = 0.
\end{aligned}$$

b. **[3 marks]** Let $Z \sim$ normal($\mu = 0, \sigma^2 = 1$). Compare the first two moments of X and Z.

Sol: We know the $E[Z] = 0$ and $V[Z] = 1$ from the table. Also, $E[Z^2] = V(Z) - [E[Z]]^2$ so $E[Z^2] = 1$. We have already found the first moment of X so let's find the variance:

$$\begin{aligned}
V(X) &= V\left(\sum_{i=1}^{12} U_i - 6\right) = V\left(\sum_{i=1}^{12} U_i\right) - V(6) \\
&= V\left(\sum_{i=1}^{12} U_i\right) - 0 \\
&= V(U_1) + \cdots + V(U_{12}) = 12\left(\frac{1}{12}\right) = 1
\end{aligned}$$

So the $E[X^2] = 1$. We note that the first two moments for X and Z are the same.

c. **[10 marks]** Justify that X may be considered **approximately** normal($\mu = 0, \sigma^2 = 1$).

Sol: We want to have some sort of a mean for the CLT!

$$
\begin{aligned}
X &= \sum_{i=1}^{12} U_i - 6 = 12\bar{U} - 6 \\
&= \sqrt{12}\left(\frac{\bar{U} - 1/2}{1/\sqrt{12}}\right) \\
&= \left(\frac{\bar{U} - \mu_{\bar{U}}}{\sqrt{V(\bar{U})}}\right) = \left(\frac{\bar{U} - \mu_U}{\sigma_U/\sqrt{n}}\right) \\
&= \sqrt{n}\left(\frac{\bar{U} - \mu_U}{\sigma_U}\right) \approx \text{normal}(0,1)
\end{aligned}
$$

Note: $E[\bar{U}] = 1/2$ and $V[\bar{U}] = \frac{1}{n}V(U) = \frac{1}{12}\frac{1}{12} = \frac{1}{144}$

d. **[10 marks]** Can you think of any obvious ways in which the approximation fails?

Sol: Here the sample size is fixed, so we can't increase or decrease n. But we see that the range of X only goes from $[-6, 6]$, while a standard normal distribution goes from $(-\infty, \infty)$.

Question 4 [**30 marks**]: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta) = \theta x^{\theta-1}$, where $0 \leq x \leq 1$ and $0 < \theta < \infty$.

a. [**8 marks**] Find the maximum likelihood estimator (MLE) of θ.

Sol: Let's get the likelihood:

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

Now let's take the log:

$$\ell(\theta|\boldsymbol{x}) = nlog(\theta) + (\theta-1)\sum_{i=1}^{n} log(x_i) = nlog(\theta) + \theta \sum_{i=1}^{n} log(x_i) - \sum_{i=1}^{n} log(x_i)$$

Now differentiate:

$$\ell'(\theta|\boldsymbol{x}) = \frac{n}{\theta} + \sum_{i=1}^{n} log(x_i) \ \Rightarrow \ \text{set equal to zero and solve}$$

$$\frac{n}{\theta} + \sum_{i=1}^{n} log(x_i) = 0$$

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} log(x_i)}$$

Let's check that the second derivative is negative, so we ensure a maximum:

$$\ell''(\theta|\boldsymbol{x}) = -\frac{n}{\theta^2} \leq 0$$

b. [**14 marks**] What are the mean and variance of the MLE? What happens to the variance as $n \to \infty$.

Sol: Let's determine the distribution for $Y_i = -log(X_i)$ (you can use your results from Question 1):

$$\Rightarrow \ X = exp(-Y) \ \Rightarrow \frac{dx}{dy} = -1exp(-Y)$$

$$
\begin{aligned}
f_Y(y) &= \theta exp(-y)^{(\theta-1)} \big| -1 exp(-y) \big| \\
&= \theta exp(-\theta y)
\end{aligned}
$$

From the table we recognise that this is an exponential distribution. So $Y \sim$ exponential$(1/\theta)$.
Now let's determine $Z = \sum_{i=1}^{n} Y_i$:
The easiest way is through the mgf (you can get the MGF of Y from the table):

$$
\begin{aligned}
M_Y(t) &= E\left(exp(ty)\right) = \int_0^\infty exp(ty)\theta exp(-\theta y)dy \\
&= \theta \int_0^\infty exp(ty - \theta y)dy \\
&= \theta \int_0^\infty exp(-y(\theta - t))dy \\
&= \frac{\theta}{\theta - t} \int_0^\infty (\theta - t) exp(-y(\theta - t))dy \\
&= \frac{\theta}{\theta - t} \times 1 = \frac{\theta}{\theta - t}
\end{aligned}
$$

Now let's get the MGF of Z:

$$
\begin{aligned}
M_{\sum Y_i}(t) &= \left[\frac{\theta}{\theta - t}\right]^n \\
&= \left[\frac{\theta/\theta}{(\theta - t)/\theta}\right]^n = \left[\frac{1}{1 - t/\theta}\right]^n
\end{aligned}
$$

From the table we see that this is the MGF for a gamma distribution. $Z \sim$ gamma$(n, 1/\theta)$.
Finally, we can determine the mean and the variance for the MLE. Let $A = -\sum_{i=1}^{n} log(X_i)$, so the mle is $\hat{\theta} = \frac{n}{A}$.

$$
\begin{aligned}
E\left(\frac{n}{A}\right) &= n\frac{\theta^n}{\Gamma(n)} \int_0^\infty \frac{1}{A} A^{n-1} exp(-\theta A)dA \\
&= n\frac{\theta^n}{\Gamma(n)} \frac{\Gamma(n-1)}{\theta^{n-1}} \int_0^\infty \frac{\theta^{n-1}}{\Gamma(n-1)} A^{(n-1)-1} exp(-\theta A)dA \\
&= n\frac{\theta^n}{\Gamma(n)} \frac{\Gamma(n-1)}{\theta^{n-1}} \times 1 \\
&= \frac{n}{n-1}\theta
\end{aligned}
$$

Thus $\hat{\theta}$ is biased!

$$\begin{aligned}
E\left(\frac{n^2}{A^2}\right) &= n^2\frac{\theta^n}{\Gamma(n)}\int_0^\infty \frac{1}{A^2}A^{n-1}exp(-\theta A)dA \\
&= n^2\frac{\theta^n}{\Gamma(n)}\frac{\Gamma(n-2)}{\theta^{n-2}}\int_0^\infty \frac{\theta^{n-2}}{\Gamma(n-2)}A^{(n-2)-1}exp(-\theta A)dA \\
&= n^2\frac{\theta^n}{\Gamma(n)}\frac{\Gamma(n-2)}{\theta^{n-2}}\times 1 \\
&= \frac{\theta^2 n^2}{(n-1)(n-2)}
\end{aligned}$$

This leads to: $V(\hat{\theta}) = \frac{\theta^2 n^2}{(n-2)(n-1)^2}$. We see $\frac{n^2}{(n-1)^2} \to 1$ as $n \to \infty$:

$$V(\hat{\theta}) = \theta^2\frac{n^2}{(n-1)^2}\frac{1}{(n-2)} \to 0 \text{ as } n \to \infty.$$

c. **[8 marks]** While a closed form solution for the MLE exists, write pseudo-code to perform a Newton-Raphson algorithm to find the MLE. For this particular problem, a friend states that you should have used Fisher scoring. Is your friend correct?

Sol:

Algorithm 2 Newton-Raphson algorithm

Let $\theta = 5$
Let $U = \frac{n}{\theta} + \sum_{i=1}^{n} log(x_i)$
Let $H = -\frac{n}{\theta^2}$
while $check \geq 1e-08$ **do**
    Let $\theta^* = \theta - U/H$
    Let $check = \left|\theta^* - \theta\right|$
    Let $\theta = \theta^*$
return θ

For Fisher Scoring We need to determine:
$I(\theta) = -E[\ell''(\theta)] = -E[-\frac{n}{\theta^2}] = \frac{n}{\theta^2}$. Then we update via:

$$\theta^* = \theta + U/I(\theta)$$

In this case there is **no difference** between the two methods. Your friend didn't go through the calculation! Note there were no data in the second derivative, everything was fixed and not random.

End Of Examination