Stat 5102 Final Exam **May 14, 2015**

Name ______________________________ Student ID ____________________

The exam is closed book and closed notes. You may use three $8\frac{1}{2} \times 11$ sheets of paper with formulas, etc. You may also use the handouts on “brand name distributions” and Greek letters. You may use a calculator.

Put all of your work on this test form (use the back if necessary). Show your work or give an explanation of your answer. No credit for numbers or formulas with no indication of where they came from. Leave no undone integrals or derivatives in your answers, but other than that requirement there is no unique “correct” simplification. Any correct answer gets full credit, except as explicitly stated in questions.

Abbreviations used: distribution function (DF), independent and identically distributed (IID), maximum likelihood estimate (MLE), probability density function (PDF), and probability mass function (PMF).

The points for the questions total to 200. There are 12 pages and 8 problems.

Questions begin on the following page.

1. [25 pts.] Suppose (X_i, Y_i), $i = 1, \ldots, n$, are IID from the distribution with PDF

$$f_\theta(x, y) = e^{-x\theta - y/\theta}, \qquad 0 < x < \infty, 0 < y < \infty.$$

where $\theta > 0$ is an unknown parameter. (Hint: What are the marginal distributions of X and Y?)

(a) Find the Jeffreys prior for this distribution.

(b) Say whether it is proper or improper.

2. [25 pts.] The following Rweb output fits a linear model. The covariate `x` is numeric and the covariate `color` is categorical (what R calls a factor).

```
Rweb:>  out1 <- lm(y ~ x + color)
Rweb:>  summary(out1)

Call:
lm(formula = y ~ x + color)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6101 -3.7974  0.5263  2.8854 10.4138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.35847    1.94831   2.237 0.030398 *
x            0.20438    0.04896   4.174 0.000139 ***
color2       3.35899    2.22408   1.510 0.138120
color3       3.99087    2.22570   1.793 0.079835 .
color4       2.97893    2.22839   1.337 0.188157
color5       9.66366    2.23215   4.329 8.51e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.972 on 44 degrees of freedom
Multiple R-squared:  0.4809,    Adjusted R-squared:  0.4219
F-statistic: 8.153 on 5 and 44 DF,  p-value: 1.639e-05
```

(a) Find a 95% confidence interval for the true unknown regression coefficient for the predictor `x`. (Hint: The 0.95 quantile of the standard normal distribution is 1.645, the 0.975 quantile of the standard normal distribution is 1.960, the 0.95 quantile of the t distribution on 44 degrees of freedom is 1.680, and the 0.975 quantile of the t distribution on 44 degrees of freedom is 2.015.)

(b) Perform a hypothesis test of whether the regression coefficient for the predictor `x` is zero (the null hypothesis) versus nonzero (the alternative hypothesis). State the value of the test statistic, the distribution of the test statistic under the null hypothesis, whether this distribution is exact or approximate (asymptotic, large n), and the P-value. Interpret the P-value. What does it say about what predictor variable or variables should or should not be in the model?

The Rweb output continues.

```
Rweb:>  out0 <- lm(y ~ x)
Rweb:>  anova(out0, out1)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + color
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     48 1579.6
2     44 1087.7  4    491.86 4.9742 0.00214 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c) What is this test about? State the null and alternative hypotheses, the value of the test statistic, the distribution of the test statistic under the null hypothesis, whether this distribution is exact or approximate (asymptotic, large n), and the P-value. Interpret the P-value. What does it say about what predictor variable or variables

3. [25 pts.] Suppose $X_1, \ldots, X_n$ are IID from the NegBin(r, p) distribution, where r is considered known so p is the only unknown parameter. Show that there is a one-dimensional sufficient statistic, and identify this statistic.

4. [25 pts.] Suppose (X_i, Y_i), $i = 1, \ldots, n$ are IID (two-dimensional) random vectors, the marginal distribution of X_i is $\text{Poi}(\mu)$, and the conditional distribution of Y_i given X_i is $\text{Bin}(X_i, p)$, where μ and p are unknown parameters satisfying $0 < \mu < \infty$ and $0 < p < 1$. Show that this statistical model is a two-dimensional exponential family (two natural statistics and two natural parameters). Identify the natural statistics and natural parameters.

   Clarification: Since X_i can be zero, we need to know what the binomial distribution with sample size zero means. Setting sample size to zero in the usual definition of the binomial distribution from this brand name distributions handout, the sample space is all of the integers from zero to zero. Thus the distribution is concentrated at zero and $\Pr(Y_i = 0 \mid X_i = 0) = 1$. The usual formula for the PMF

$$f(y_i \mid x_i) = \binom{x_i}{y_i} p^{y_i} (1 - p)^{x_i - y_i}$$

   still works when $x_i = y_i = 0$, giving

$$\binom{0}{0} p^0 (1 - p)^0 = \frac{0!}{0!\,0!} p^0 (1 - p)^0 = 1$$

   so we do not have to fuss about this special case (the general formula for the PMF works in this case too).

5. [25 pts.] Suppose $X_1, \ldots, X_n$ are IID from the distribution with PDF

$$f_\theta(x) = \begin{cases} 3\theta/[(1+\theta)(1+|x|)^4], & -\infty < x < 0 \\ 3\theta/[(1+\theta)(1+\theta|x|)^4], & 0 < x < \infty \end{cases}$$

where θ is an unknown parameter satisfying $0 < \theta < \infty$. This distribution has mean and variance

$$E_\theta(X) = \frac{1-\theta}{2\theta}$$

$$\mathrm{var}_\theta(X) = \frac{3 - 2\theta + 3\theta^2}{4\theta^2}$$

(you do not have to prove these).

(a) Find a method of moments estimator of θ.

(b) Find the asymptotic normal distribution of your estimator $\hat{\theta}_n$.

6. [25 pts.] The function

$$F_\theta(x) = \begin{cases} x^2/(1+\theta x + x^2), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

is a DF, where the parameter θ satisfies $0 < \theta < \infty$.

Hint: The roots of the quadratic polynomial $ax^2 + bx + c$ are

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

(a) Find the median of this distribution.

(b) Find the PDF corresponding to this DF.

(c) Find the asymptotic variance of the sample median for an IID sample $X_1, \ldots, X_n$ from this distribution.

7. [25 pts.] Suppose $X_1, \ldots, X_n$ are IID from the distribution with PDF

$$f_\theta(x) = \theta^2 x e^{-\theta x}, \qquad 0 < x < \infty,$$

where θ is an unknown parameter satisfying $0 < \theta < \infty$.

(a) Find the maximum likelihood estimator of θ.

(b) Show that your solution to part (a) is the unique global maximizer of the likelihood if it is. If you cannot show that it is the unique global maximizer, show that it is a local maximizer.

(c) Calculate expected Fisher information for sample size n.

(d) Give an approximate, large sample 95% confidence interval for θ. (Hint: The 0.95 quantile of the standard normal distribution is 1.645, and the 0.975 quantile of the standard normal distribution is 1.96.)

8. [25 pts.] The *Rayleigh distribution* (not a distribution we have met before) has PDF

$$f_\sigma(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \qquad 0 < x < \infty,$$

where $\sigma > 0$ is an unknown parameter. Suppose we have IID data X_1, $\ldots$, X_n having this distribution and want to do a Bayesian analysis with a $\text{Gam}(\alpha, \lambda)$ prior distribution on the parameter $\theta = 1/\sigma^2$. Find the posterior distribution for θ. (Hint: the posterior is a brand name distribution.)