Statistical Computing 1, Stat 590 Homework 10 Name:____________________
Fall 2015 Prof. Erik B. Erhardt

Part I. (105 points) Do all calculations in LaTeX + R + knitr. Insert computer text output and graphics to support what you are saying. For this assignment, all R code should well commented and be visible (`echo=TRUE`) in the document where you have written it.

(105$^{\text{pts}}$) **1. Hardy-Weinberg Equilibrium**

If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur in a population with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, according to the Hardy-Weinberg Law. In a sample from the Chinese population of Hong Kong in 1937, blood types occurred with the following frequencies, where M and N are erythrocyte antigens:

	Blood type			
	M	MN	N	Total
Frequency	342	500	187	1029
Category label	1	2	3	

If we let $\underset{\sim}{Y} = [Y_1, Y_2, Y_3]^\top$ be the counts in n categories 1, 2, and 3 based on a sample size m then the joint distribution of the cell counts is multinomial

$$\Pr[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3|\theta] = \frac{m!}{\prod_{i=1}^n y_i!}\prod_{i=1}^n p_i(\theta)^{y_i}$$

where

$$p_1(\theta) = (1-\theta)^2, \quad p_2(\theta) = 2\theta(1-\theta), \quad \text{and} \quad p_3(\theta) = \theta^2$$

are cell probabilities and $0 < \theta < 1$.

(a) (10 pts) **Derive the Muiltinomial log-likelihood, $\ell(\theta)$.** Show that the log-likelihood for an arbitrary sample is

$$\ell(\theta) = \log(m!) - \sum_{i=1}^n \log(y_i!) + (2y_1 + y_2)\log(1-\theta) + (2y_3 + y_2)\log(\theta) + (y_2)\log(2).$$

(b) (15 pts) **Derive the derivatives of the log-likelihood and the expected information.** Derive general expressions for the first and second derivatives of $\ell(\theta)$ ($\dot{\ell}(\theta)$ and $\ddot{\ell}(\theta)$) and the expected information $\mathbf{I}(\theta)$.

(c) (10 pts) **Likelihood equations, $\mathbf{L}(\theta|\underset{\sim}{y})$.** Write down the likelihood equation for an arbitrary sample (that is, $\dot{\ell}(\theta) = 0$) and solve for the MLE.

(d) (10 pts) **Functions to evaluate functions in parts (a) and (b).** Write functions to evaluate $\ell(\theta)$, $\dot{\ell}(\theta)$, $\ddot{\ell}(\theta)$, and $\mathbf{I}(\theta)$ for and arbitrary sample $\underset{\sim}{Y} = [Y_1, Y_2, Y_3]^\top$. Ignore the constant terms in $\ell(\theta)$.

(e) (10 pts) **Plot $\ell(\theta)$, guess at MLE.** For the observed data, plot $\ell(\theta)$. What is a good guess for the MLE based on the graph?

(f) (10 pts) **Newton-Raphson for MLE.** Write an NR procedure to numerically evaluate the MLE of θ for an arbitrary sample.

(g) (10 pts) **Fisher scoring for MLE.** Write a Fisher's Scoring procedure to numerically evaluate the MLE of θ for an arbitrary sample.

(h) (10 pts) **Compare NR and Fisher scoring methods.** For the given data, use each algorithm to evaluate the MLE of θ. Compute estimates of the standard deviation of the MLE $\hat{\theta}$ using both the observed and expected information. Also, compute two approximate 95% CIs for θ of the form $\theta \pm 1.96\text{SE}[\hat{\theta}]$, where the SE is based on the (1) observed and (2) expected information. Discuss the results: what are the similarities/dissimilarities between routines, SEs, CIs, etc., and does the numerically evaluated MLE agree with the answer in Part (c)?

(i) (10 pts) **Sensitivity analysis.** *More open-ended:* By choosing different starting values, examine the sensitivity of NR and Fisher Scoring to the choice of the initial guess of the MLE. Do any graphical summaries help to understand the sensitivity (or lack of sensitivity)? Discuss.

(j) (10 pts) **Summary.** Summarize your results.