Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

AMS 206: Quiz 3 *[45 points]*
(required for graduate students, extra credit for undergraduates)

Name: _____

In this problem, and all subsequent assignments involving computer-based calculations and plots, please put Your code (in whatever statistical computing environment You use) in an Appendix at the end of Your solutions; this will enable the grader to give You more accurate partial credit if Your solution is not fully correct.

(Neyman-style frequentist inference for the variance and standard deviation in the Gaussian model with known mean) People in Las Vegas who are experts on the National Football League (NFL) provide a prediction in the form of a *point spread* for every football game before it occurs, as a measure of the difference in ability between the two teams (and taking account of where the game will be played and other variables, such as injuries to key players). For example, if Denver is a 3.5–point favorite to defeat San Francisco, the implication is that betting on whether Denver's final score minus 3.5 points exceeds or falls short of San Francisco's final score is an even-money (50/50) proposition. The data set `actual-minus-predicted.txt` on the course web page (based on data from Gelman et al. (2014)) records the differences y = (actual outcome – point spread) for a collection of $n = 672$ professional football games in 1981, 1982 and 1984. Thinking of this data set as like a random sample from a population $\mathcal{P}$ of similar NFL games in other years, the Gelman et al. authors propose the sampling distribution $(Y_i \,|\, \sigma\, \mathcal{G}\, \mathcal{B}) \stackrel{\text{IID}}{\sim} N(0, \sigma^2)$ $(i = 1, \ldots, n)$ for the observed differences y_i, in which $\mathcal{G}$ stands for the Gaussian sampling distribution assumption (which Gelman et al. made after looking at the data, and which is therefore not part of $\mathcal{B}$). (Note: if this distribution didn't have a mean that's close to 0, the experts would be *uncalibrated* and you could make money by betting against them.)

(1) Read the data set into `R` (or Your favorite alternative statistical computing environment); in `R` this can be done by changing directory to where You've downloaded the data file from the web page and using the command `y <- scan( 'actual-minus-predicted.txt' )` . Compute the sample mean $\bar{y} = \frac{1}{n} \sum_i^n y_i$ and SD $s = \sqrt{\frac{1}{n-1} \sum_i^n (y_i - \bar{y})^2}$ of the y_i values, and make a histogram of them on the density scale with about 30 bars, superimposing the Normal density curve on this plot with mean 0 and SD s . Make a Normal quantile-quantile plot of the y_i, and superimpose the target straight line implied by a mean of 0 and an SD of s. (`R` code for making plots similar to these, and saving them as PDF for incorporation into documents, may be found in the files called *R and Maple code for Case Study 1* and *R code for the Central Limit Theorem simulation in Case Study 2* on the course web page.) Looking at Your histogram and Normal qqplot, do You agree with Gelman et al. that the Gaussian sampling model with mean 0 is reasonable for this data set? Explain briefly. *[20 points]* (You can either insert (if You're using text-processing software to prepare Your answers) or attach Your plots (if not) at the end of the quiz.)

For the rest of the problem, however You answered part (1), let's assume the Gaussian sampling distribution $(Y_i \,|\, \sigma\, \mathcal{G}\, \mathcal{B}) \overset{\text{IID}}{\sim} N(0, \sigma^2)$ $(i = 1, \ldots, n)$ for the observed y_i in the NFL data set.

(2) The usual frequentist estimate of the Gaussian variance σ^2 when the mean is known to be 0 is $s_*^2 = \frac{1}{n} \sum_i^n y_i^2$. It can be shown (You're not asked to show this) that the repeated-sampling distribution of s_*^2 in the IID $N(0, \sigma^2)$ sampling model is implied by the relationship

$$\frac{n\, s_*^2}{\sigma^2} \sim \chi_n^2 \,, \tag{1}$$

in which χ_n^2 is the *chi-squared* distribution with n degrees of freedom, having density function

$$\theta \sim \chi_n^2 \quad \text{iff} \quad p(\theta) = \left\{ \begin{array}{cc} \frac{1}{2^{\frac{n}{2}}\, \Gamma\left(\frac{n}{2}\right)} \theta^{\frac{n}{2}-1} e^{-\frac{\theta}{2}} & \text{if} \quad \theta > 0 \\ 0 & \text{otherwise} \end{array} \right\} . \tag{2}$$

The `R` built-in function `qchisq` computes inverse cumulative distribution function (CDF) values (quantiles) for this distribution. For example, for a given $0 < \alpha < 1$, calling this function with the `R` command `A <- qchisq( alpha / 2, n )` computes

> the place on the horizontal axis where the χ^2 distribution with n degrees of freedom has $\frac{\alpha}{2}$ of its probability to the left of that place

and assigns that value to the numeric object A; this is equivalent to saying that $A = F_{\chi_n^2}^{-1}\left(\frac{\alpha}{2}\right)$, in which $F_{\chi_n^2}(\theta)$ is the CDF of the χ_n^2 density. In a similar manner `B <- qchisq( 1 - alpha / 2, n )` computes $B = F_{\chi_n^2}^{-1}\left(1 - \frac{\alpha}{2}\right)$, the $\left(1 - \frac{\alpha}{2}\right)$ quantile of the χ_n^2 distribution, so that

$$\text{if} \quad \theta \sim \chi_n^2 \quad \text{then} \quad P_F(A < \theta < B) = 1 - \alpha \,, \tag{3}$$

in which $P_F(\cdot)$ is the frequentist (repeated-sampling) version of probability. But now, in view of equation (1) above, this implies that, under the IID $N(0, \sigma^2)$ sampling model,

$$P_F\left(A < \frac{n\, s_*^2}{\sigma^2} < B\right) = 1 - \alpha \,. \tag{4}$$

(a) Rearrange the inequality inside $P_F(\cdot)$ in (4) to have σ^2 in the middle, thereby showing that equation (4) implies that

$$P_F\left(\frac{n\, s_*^2}{B} < \sigma^2 < \frac{n\, s_*^2}{A}\right) = 1 - \alpha \,. \tag{5}$$

You have just performed Mr. Neyman's *confidence trick*, demonstrating that $\left(\frac{n\, s_*^2}{B}, \frac{n\, s_*^2}{A}\right)$ is a $100\,(1 - \alpha)\%$ confidence interval (CI) for σ^2 in the IID $N(0, \sigma^2)$ sampling model. Use this result to show that $\left(s_* \sqrt{\frac{n}{B}}, s_* \sqrt{\frac{n}{A}}\right)$ is a $100\,(1 - \alpha)\%$ CI for σ in this model. *[10 points]*

(b) Use the results You derived above to compute 95% CIs for σ^2 and σ with the NFL data. If someone said to You, "NFL football games are hard to predict: the people who post the point spreads are usually wrong by about two touchdowns (14 points)," would You say that the data set examined here supports this claim? In other words, is the difference between s_* (the best frequentist estimate of σ in the $N(0, \sigma^2)$ sampling model) and this person's stated value of $\sigma_{claim} = 14$ statistically significant? Is that difference practically significant? Explain briefly. *[15 points]*