STAT 345 Spring 2018
Homework 11 - Confidence Intervals Name: _____

Please adhere to the homework rules as given in the Syllabus.

1. Data Analysis. Recall the CDI dataset which contains demographic information on US counties. A random subset of this data can be found at

`math.unm.edu/~knrumsey/cdi_sample.csv`

The variable `TotalSeriousCrimes` represents a count of all violent crimes recorded in each county during a year.

a) Make a histogram of the crimes variable. Do you think that t-procedures are safe to use for this data?

b) We can attempt to normalize the data by dividing by the Population variable (now we are examining crimes per capita). Make a new histogram (turn these in together). How do you feel about using t-procedures now?

c) Use R, Excel or any software of your choice to construct a 95% confidence interval for μ, the true average number of crimes per capita in large US counties. Show the important calculations below and interpret your interval in words.

2. Significance. Suppose a new drug is in development for reducing Systolic blood pressure. The drug is intended for a population of people with high SBP (say SBP = 135 mmHg). When a member of this population takes this drug, the reduction in SBP is assumed to be Normal with mean μ and sd σ where both parameters are unknown. Researchers have decided that any effective drug should reduce your blood pressure by at least 7 mmHg.

a) Statistically Insignificant, but (maybe) Practically Significant. Suppose than $n = 4$ subjects are given the drug and the reduction in SBP is measured for each subject, giving $\bar{x} = 8.4$ and $S = 4.7$. Calculate a 98% confidence interval for μ and interpret your interval. Mr. S claims that, since zero is in the interval, there is "no effect" and suggests that further research on this drug be abandoned. Explain why Mr. S is jumping to conclusions too quickly.

a) Statistically Significant, but Practically Insignificant. Now we consider an alternative drug. This drug is given to $n = 348$ members of the population, and the reduction in SBP is measured for each subject, with $\bar{x} = 1.8$ and $S = 6.37$. Calculate a 98% confidence interval for μ and interpret your interval. Mr. S claims that, since zero is **not** in this interval, the results are statistically significant and this drug is ready for the market. Explain what is wrong with his logic.

3. Confidence Interval for Population Proportion. In the 2016-2017 season, Stephen Curry shot $n = 790$ three-pointers and made 325 of them. Let p be the true proportion of three-pointers that Curry will make.

a) Create a 80% confidence interval for p using the Normal approximation to the Binomial. Interpret your interval.

c) According to your interval, do you feel comfortable asserting that Stephen Curry makes more than 40 percent of his three-pointers?

d) If you wanted to know the true proportion within 1% (i.e. 0.01), how many Stephen Curry three-pointers do you have to observe?

4. Exponential Distribution. Assume that $X_1, X_2, \cdots X_{15}$ are iid $Exp(\lambda)$, and suppose we calculate $\sum_{i=1}^{15} x_i = 31$ and $\sum_{i=1}^{15} x_i^2 = 130$.

a) Compute $\bar{x}$ and s^2. *Hint: Use the shortcut method to find* s^2.

b) Set $\mu = \frac{1}{\lambda}$ and compute a small sample 95% CI for μ (using t-procedures).

c) Convert your CI for μ to a CI for λ by noting that $P(a < \mu < b) = P(1/b < \lambda < 1/a)$.

5. Bootstrap CI. The skew of a distribution was discussed in an earlier HW (homework 6). Denote the "true" skew of a distribution as γ, and consider the estimator

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

where $\bar{x}$ and s are the sample mean and standard deviation respectively.

a) Consider the raw number of total serious crimes (not per capita) from the CDI dataset in problem 1. Calculate the sample skewness and report the value $\hat{\gamma}$ here.

b) Remember from homework 6 that Exponential distributions always have a true skew of $\gamma = 2$. Create a histogram of the TotalSeriousCrimes data (use the option `freq=FALSE`) and add an exponential density curve to the plot using the estimator $\hat{\lambda} = 1/\bar{x}$. Comment on the fit.

c) Use a bootstrap procedure (notes are on the course web page) to create an approximate 95% confidence interval for γ. Interpret this interval in words and turn in a histogram of the bootstrap sampling distribution.

6. *Challenge Problem.* **Logistic Regression.** Consider the problem of predicting whether or not an NFL kicker will make a field goal, or better yet the probability that he makes it. We are able to analyze this data (as in problem 3) by assuming that we know nothing else about a kick so that the probability of each kick is the same. This is silly of course, and we can do much better if we take additional knowledge into account such as weather conditions, elevation of the stadium and **distance**. Distance is the most important factor in a kick (other than the kicker himself, perhaps) so that will be our focus for now. Let x be the distance of a kick and let $Y = 1$ if the kick is good and $Y = 0$ otherwise. The simple logistic regression model is

$$Y \sim Bern(p(x))$$

where

$$p(x) = \frac{1}{1 + \exp(-\alpha - \beta x)}$$

The probability of a successful kick now depends on two parameters α and β and *changes* with the distance x. Consider another parameter, θ which gives the probability that a kick is made from 60 yards. To get θ, we just plug in 60 for x in the equation above. If α and β are estimated with Maximum Likelihood, then the invariance property says that the ML of θ is

$$\hat{\theta}_{ML} = \frac{1}{1 + \exp(-\hat{\alpha}_{ML} - \hat{\beta}_{ML} \cdot 60)}.$$

Although there is no close-form solution for the ML estimators, the estimates can be done easily using R.

A database consisting of $n = 984$ kicks can be found by typing

```
data <- read.table('http://users.stat.ufl.edu/~winner/data/fieldgoal.dat')
```

The first column (V1) gives the distance x and the second column (V2) gives the make/miss indicator Y. We can estimate the parameters (with MLE by default) in R by typing

```
fit = glm(y~x, family=binomial); alpha_hat = fit$coeff[1]; beta_hat = fit$coeff[2]
```

Now the ML estimate for θ can be calculated.

a) Give the ML estimate for α, β and θ using the data. Using a computer, provide a plot of $p(x)$ for $x \in [18, 117]$.

b) Use a bootstrap to give an approximate confidence interval for the true value of θ. Give a histogram of the bootstrap sampling distribution of $\hat{\theta}_{ML}$.