STAT 345 Fall 2018
Final Exam Name: ______________________________

This exam is intended to combine theory and application. You are free to use your in-class notes, any textbook and online resources (but be careful, other sources may define things differently). You may **NOT** consult with fellow classmates, but are welcome to come to my office hours to ask questions. You may use any software you like, but I highly recommend R since there are many resources available to you on the course webpage.

For this problem, we will consider the CDI dataset (again) which contains demographic information for 440 US counties. In particular, we will focus on the percent of the population with a Bachelors degree. This data can be read into R as follows. *Note: Sometimes copy and paste drops the ˜. If you get an error message from R, make sure the tilde is present.*

```
data <- read.csv(’http://math.unm.edu/~knrumsey/classes/
                 spring17/MiniProjects/data.csv’)
x <- data$PercentWithBachelors
```

Your answers should be given in the correct order. Don’t forget to include graphs and answer all parts of each question. Your report should be typed as much as possible. Leave room in your document to fill in answers that require calculations. **You will turn in a hard copy to me by 11:00am (SMLC 348) on Friday December 12.**

Data Exploration (35 points)

1. Create a Histogram and a Boxplot of the data. Comment on the shape of the distribution (skew, peaks, outliers, etc)
2. Calculate the mean, variance, standard deviation and five number summary of the data.
3. **Fill in the blank:** Approximately 75% of US counties have at least ___________ percent of the population with a Bachelors degree.
4. Use the $1.5 \times IQR$ rule to identify potential outliers. Do these seem like outliers based on visual inspection?
5. Create a QQ-plot of the data. Do you think it is reasonable to assume that this data is Normally distributed?
6. Create a Boxcox plot of the data. Use this plot to suggest a transformation of the data $y_i = g(x_i)$ so that the transformed data is approximately Normally distributed. Create a Histogram and QQ-plot of the transformed data. Is the Normality assumption reasonable now?
7. Use the $1.5 \times IQR$ rule to identify potential outliers, using the transformed data. Are there any outliers now? How does this compare to your visual inspection of outliers?

Point Estimation and Distribution Fitting (25 points)

8. Consider the following distribution, which we will call the *Halfbrain Distribution*, with a single parameter $\theta > 0$.
$$f(x_i|\theta) = 2\theta x_i e^{-\theta x_i^2}, \quad x_i > 0$$
Find the expected value of X for this distribution. *Hint: Use a u-substitution and the gamma function. If you are stuck, set the integral up and use wolframalpha for partial credit.*

9. Assume (momentarily) that $X_1, X_2, \cdots X_n \overset{iid}{\sim} Halfbrain(\theta)$. Find the Method of Moments estimator for θ.

10. Find the CDF of the Halfbrain distribution. Dont forget to define it for all values of x. Consider a new parameter $\tau = P(X > 45)$. Find τ analytically, as a function of θ.

11. Consider two cases:

    - Halfbrain distribution using Method of Moments estimators.
    - Gamma distribution using Method of Moments estimators. (Either calculate these, or get them from your in class notes)

    For each case: **1)** calculate and report the parameter estimates based on the data and **2)** create a Histogram of the data and overlay a density curve for each distribution. Which distribution fits the data best based on visual inspection?

12. Use your estimate for θ (Halfbrain dsitribution) to derive an estimator for τ (problem 10). Also estimate τ using your estimates for the Gamma distribution (you need to do this numerically). Compare both estimates to the nonparametric estimate `mean(x > 45)`. How do your estimates compare to the nonparametric estimate?

13. **Challenge/Optional** Repeat problems 11 and 12 for at least two of the following distributions: Weibull, Lognormal, Beta. *Hint: If you try Beta, don't forget to divide the data by* 100 *before you estimate the parameters.*

Statistical Inference (40 points)

14. Use the Bootstrap algorithm to approximate the sampling distribution of the sample mean $\bar{X}$ setting $B = 10,000$. Create a histogram. What does this say about the validity of the CLT here?

15. Use t-procedures to give a 99.9% lower confidence bound on the true average percent of population with a Bachelors degree. Interpret this bound.

16. Use t-procedures to give a 95% confidence interval for the true mean μ. Interpret your interval.

17. Pivot this interval to give a 95% confidence interval for θ of the Halfbrain distribution.

18. Consider a binary version of the data,

$$Y_i = \begin{cases} 1 & \text{if } X_i > 15 \\ 0 & \text{otherwise} \end{cases}$$

That is, $Y_i \sim \mathit{Bernoulli}(p)$ where Y_i is one if the US county has at least 15% of the population with a Bachelors degree. Find $\sum_{i=1}^{440} Y_i$ using R and state the distribution of this sum (assuming independence). *Hint: in R,* `y = as.numeric(x > 15)`

19. Test the Hypothesis

$$H_0 : p = 0.75 \qquad H_a : p > 0.75$$

at the 10% significance level. Explain what these hypotheses mean in terms of the problem. State your conclusion.

20. Using $\bar{y}$ as a guess for p, what sample size is required so that the Null hypothesis in the test above can be rejected? (Or equivalently, so that a 90% CI for p does not contain 0.75.)

21. Calculate the sample skewness of the data. Is it positive or negative? Does this agree with your answer to question 1? Use the Bootstrap algorithm to approximate the sampling distribution of the sample skewness. Construct a 95% Bootstrapped confidence interval for the population skewness γ. Do you have strong evidence that your data is skewed?

22. **Challenge/Optional**: Kurtosis is a measure of how often a distribution produces outliers. The higher the Kurtosis, the more dispersed the data is likely to be. For any normal distribution, the population Kurtosis (denoted κ) is equal to 3. Consider your transformed data from problem 6. We can construct a hypothesis test for the Normality of this data using Kurtosis. Consider the Hypothesis test:

$$H_o : \kappa = 3 \qquad vs \qquad H_a : \kappa < 3$$

Use the Bootstrap Algorithm to approximate a p-value. Report your conclusion using a significance level of your choice. What does this say about the Normality of your transformed data? *Hint: The Kurtosis can be estimated with "Sample Kurtosis" given below:*

$$\hat{\kappa} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)^4$$