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 a dataset consisting of measurements from 178 red wines (three different types or ”grape cultivars”) from a specific region in Italy. In particular, we will focus on the ”color intensity” of each wine. The data can be read using R using the following code:

```
install.packages(’candisc’)
library(’candisc’)
data(Wine)
color <- Wine$Color
```

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 of the data. Comment on the shape of the distribution (skew, peaks, outliers, etc)

2. Calculate the sample mean, variance, standard deviation and skew of the data.

3. Calculate the 5 number summary and interpret the first and third quartile.

4. Create a Boxplot of the data and 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 (30 points)

8. Consider the following distribution, which we will call the *Halfbrain Distribution*. The distribution has two parameters, θ and k.

$$f(x_i|\theta) = k\theta^{-k}x_i^{k-1}e^{-(x_i/\theta)^{k-1}}, \quad x_i > 0$$

The expected value and variance of this distribution is given by

$$E(X) = \theta\Gamma(1 + 1/k)$$

$$Var(X) = \theta^2\left(\Gamma(1 + 2/k) - \Gamma(1 + 1/k)^2\right)$$

What is the mean and variance of $X \sim Halfbrain(\theta, k)$ when $k = 2$?

9. If it is known that $k = 2$, what is the Method of Moments **estimator** for θ?

10. Find the MSE of the method of moments estimator. Is the estimator consistent?

11. What is the Method of moments **estimate** for θ using the wine color intensity data? Create a histogram of the data (use freq=FALSE) and add the Halfbrain density curve using this estimate.

12. Assume again that $k = 2$. Let τ be a parameter denoting the 99^{th} percentile of the Halfbrain distribution (i.e. $P(X \leq \tau) = 0.99$). Find τ in terms of θ.

13. Use your answers to questions 11 and 12 to provide an estimate for τ using the wine color intensity data. How well does this match the nonparametric estimate given in R by `quantile(color, 0.99)`?

Statistical Inference (40 points)

14. According to the CLT, what is the (approximate) sampling distribution of $(\bar{X})$, the sample average wine color intensity?

15. Use the Bootstrap algorithm to approximate the sampling distribution of $\bar{X}$ (setting $B = 10,000$). Does the CLT appear to be valid here?

16. Use t-procedures to give a 97% confidence interval for μ, the true mean color intensity of the wines in this population. Interpret the interval and compare it to the 97% CI using the Bootstrap distribution.

17. Assuming that $k = 2$, pivot the t-procedure interval interval to give a 97% confidence interval for θ of the Halfbrain distribution.

18. Construct a 95% CI for the population variance (σ^2) of the wine color intensity data. Interpret this interval and discuss whether or not the assumptions have been met.

19. Compare your 95% interval in the previous problem to a 95% CI constructed using the Bootstrap. If the intervals differ drastically, explain why you think they are different.

20. There are three different types of red wine in this dataset, each from a different grape cultivar. The color intensity of each group can be extracted in R by typing:

```
x1 <- color[Wine$Cultivar == 'grignolino']
x2 <- color[Wine$Cultivar == 'barolo']
x3 <- color[Wine$Cultivar == 'barbera']
```

Calculate the mean and standard deviation of color intensity for each wine cultivar.

21. Create a 99% CI for $\mu_{barbera} - \mu_{barolo}$ using two-sample t-procedures. Based on this interval, do you feel confident asserting that the mean color intensity of barbera wines is greater than the mean color intensity of barolo wines?

22. 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. Use the Bootstrap algorithm to construct a 96% upper confidence bound for the kurtosis of this data. Interpret this confidence bound and explain what it says about whether or not the transformed data is normally distributed. *Hint: The Kurtosis can be estimated with "Sample Kurtosis" given below:*

```
kurtosis <- function(x){
  z <- (x-mean(x))/sd(x)
  mean(z^4)
}
```