STAT 345 Final Exam

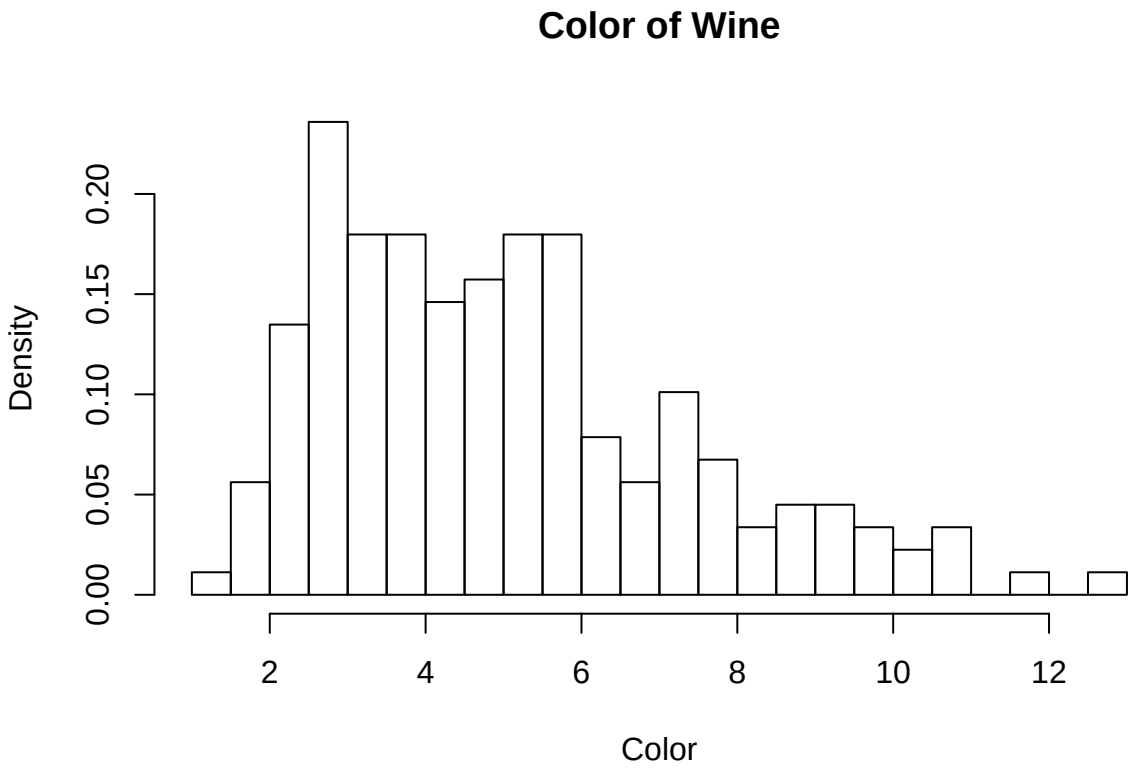
Nicholas Livingstone

Dec 13th, 2019

Data Exploration

1. Data Distribution

```
hist(color, main = "Color of Wine", xlab = "Color", breaks=30, freq=F)
```



The data looks to be right skewed. It appears 2-6 units of intensity is the most common which represents almost half the available range.

2. Statistics of Data

```
xbar <- mean(color)
s2 <- var(color)
s <- sd(color)
skew <- (sum(((color - xbar)/s)^3))/100
```

Sample Mean: 5.06
Sample Variance: 5.37

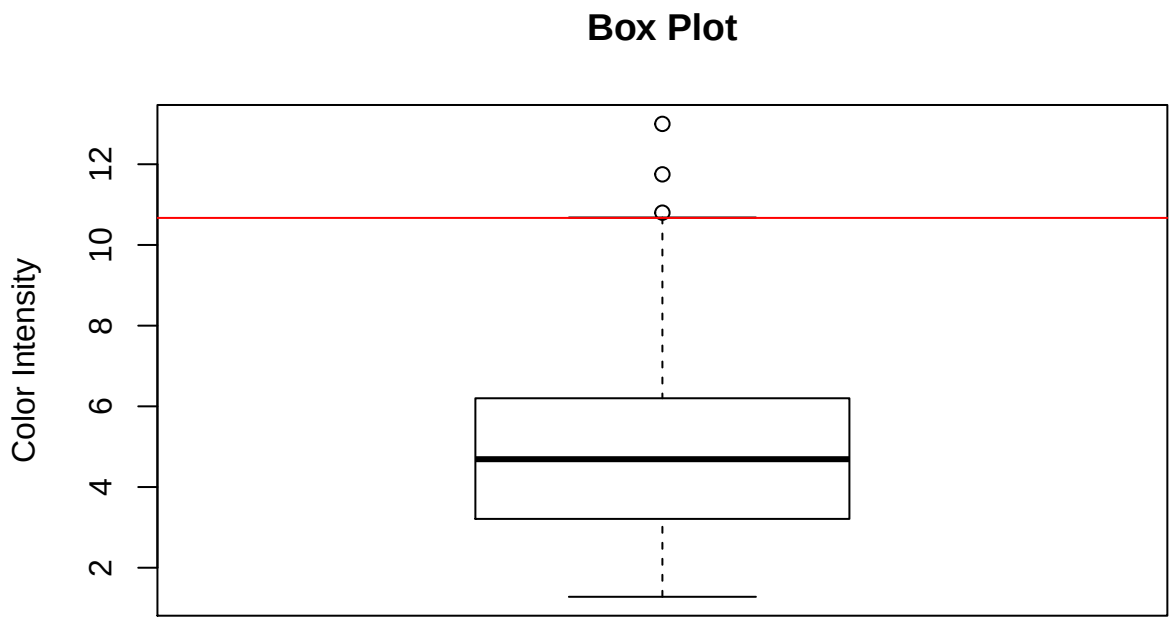
Sample Standard Deviation: 2.32
Skew: 1.52

3. 5 Number Summary(Inline r markdown expressions used here)

Minimum: 1.28
Maximum: 13
Median: 4.69
First Quartile: 3.22
Third Quartile: 6.2

4. Boxplot and Outlier identification

```
boxplot(color, ylab = 'Color Intensity', main = 'Box Plot')
UF <- quantile(color)[[4]] + 1.5 * IQR(color)
LF <- quantile(color)[[2]] - 1.5 * IQR(color)
abline(h = c(UF, LF), col = "red")
```

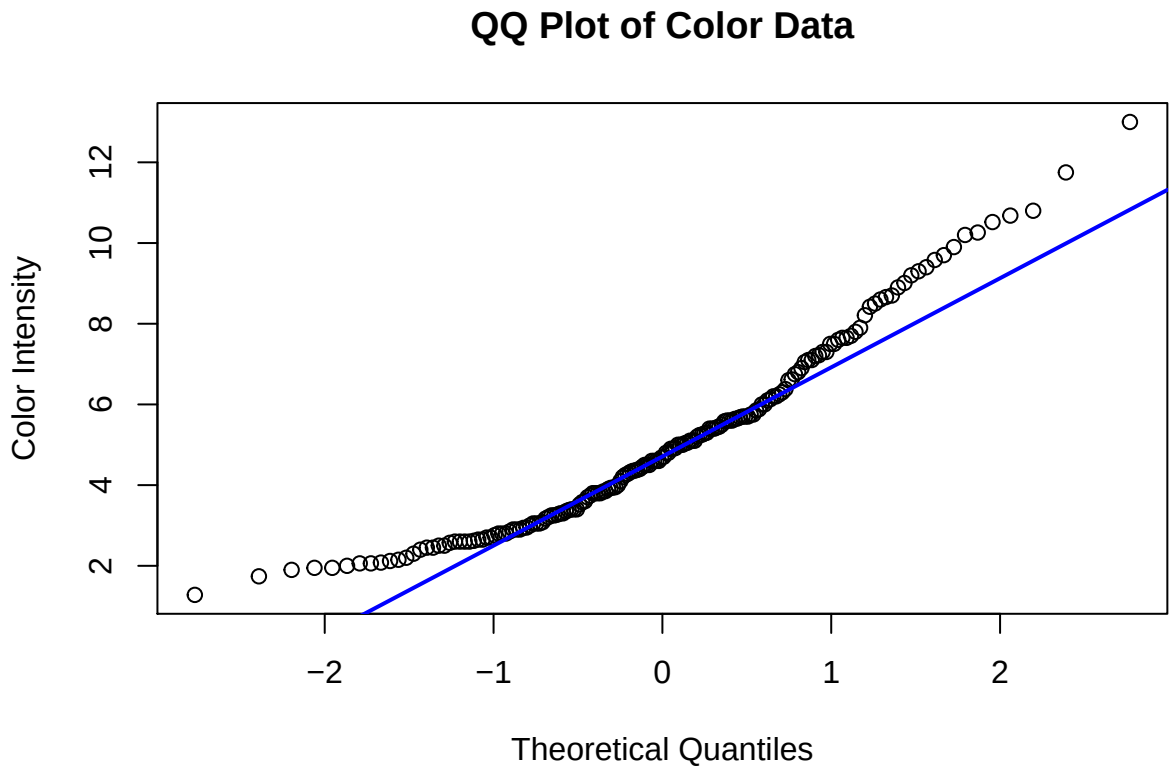



```
outliers <- color[which(UF < color)]
```

Visual inspection suggests there's ~3 outliers. When comparing the IQR * 1.5 to the values we find 4 outliers: 10.8, 13, 11.75, 10.68. All of these outliers are on the upper range of the data, using this method no lower outliers as the Lower Fence went beyond the range of the data.

5. QQ-Plot

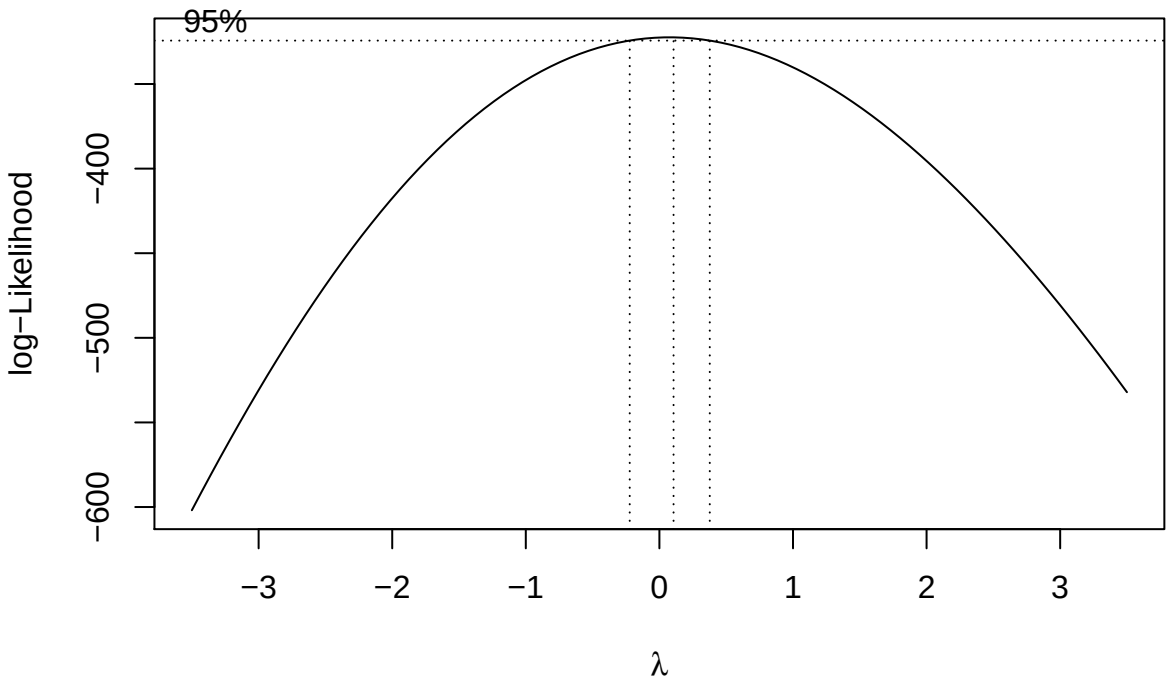
```
qqnorm(color, main = "QQ Plot of Color Data", ylab = "Color Intensity")
qqline(color, col = "blue", lwd = 2)
```



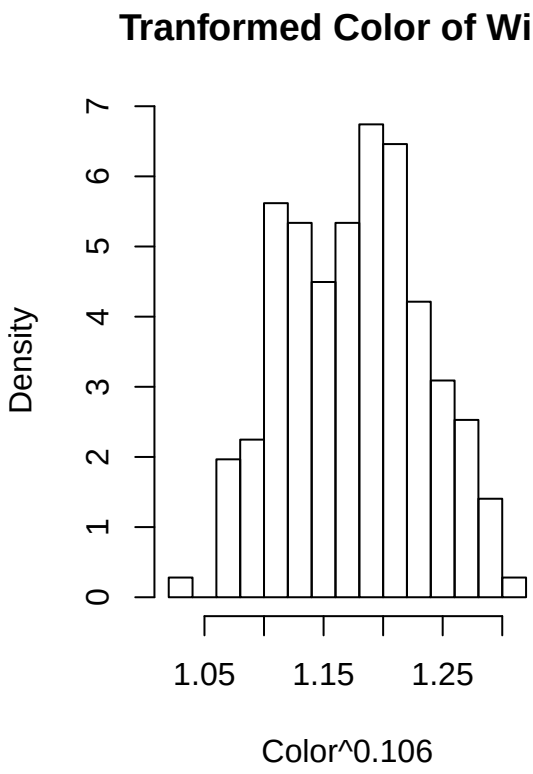
I would not call the data normally distributed. The concave up nature relative to the qqline suggests it's right skewed which is reflected in the original histogram.

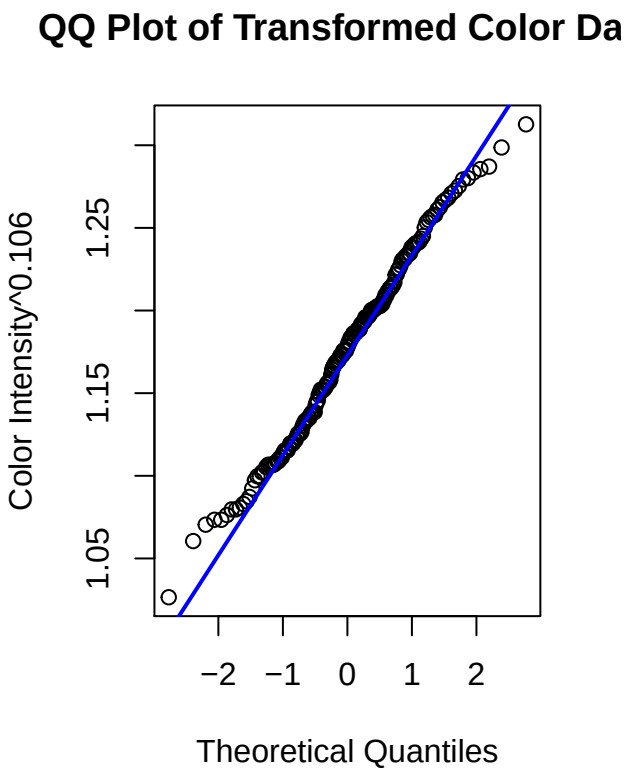
6. Boxcox Plot

```
bc <- boxcox(lm(color~1), lambda=seq(-3.5, 3.5, by=.1))
```




```
best_lambda <- bc$x[which.max(bc$y)]
bc_color <- color^best_lambda
par(mfrow=c(1, 2))
hist(bc_color, main = "Tranformed Color of Wine", xlab = "Color^0.106", breaks=12, freq=F)
qqnorm(bc_color, main = "QQ Plot of Transformed Color Data", ylab = "Color Intensity^0.106")
qqline(bc_color, col = "blue", lwd = 2)
```

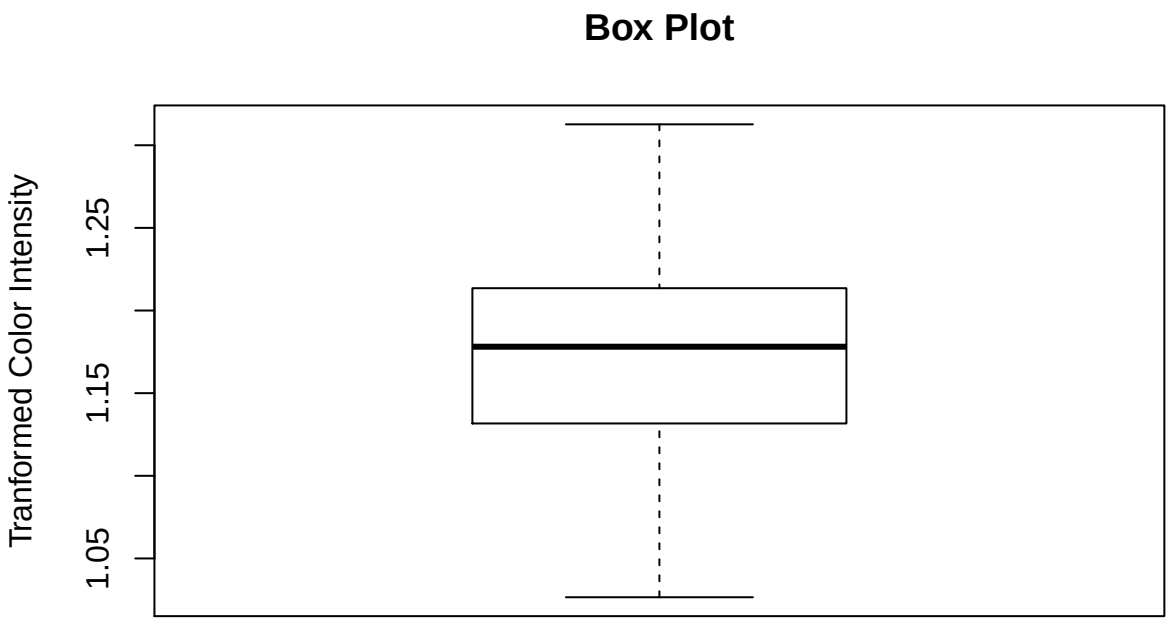




The data appears much more normalized but it's certainly not a perfect image of a normal distr.

7. 1.5×IQR on Transformed Data

```
boxplot(bc_color, ylab = 'Tranformed Color Intensity', main = 'Box Plot')
UF <- quantile(bc_color)[[4]] + 1.5 * IQR(bc_color)
LF <- quantile(bc_color)[[2]] - 1.5 * IQR(bc_color)
abline(h = c(UF, LF), col = "red")
```




```
outliers <- color[which(UF < bc_color)]
```

The tranformed data lacks any outliers as opposed to the previous inspection.

Point Estimation and Distribution Fitting

8. Mean & Variance

$$E(X) = 0.886\theta$$
$$Var(X) = 0.215\theta^2$$

9. Method of Moments Estiamtor for θ

$$E(X) = \bar{X} = 0.886\theta$$
$$\theta = \frac{\bar{X}}{0.886}$$

10. MSE(θ)

$$\begin{aligned} B(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ &= E(\frac{\bar{X}}{0.886}) - \theta \\ &= \frac{1}{0.886} E(\bar{X}) - \theta \\ &= \frac{1}{0.886} \times 0.886\theta - \theta \\ &= \theta - \theta \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(\hat{\theta}) &= Var(\frac{\bar{X}}{0.886}) \\ &= (\frac{1}{0.886})^2 Var(\bar{X}) \\ &= \frac{1}{0.785} \times \frac{0.215\theta^2}{n} \\ &= 0.145\frac{\theta^2}{n} \end{aligned}$$

Because the bias is zero...

$$MSE(\hat{\theta}) = Var(\hat{\theta}) = 0.145\frac{\theta^2}{n}$$

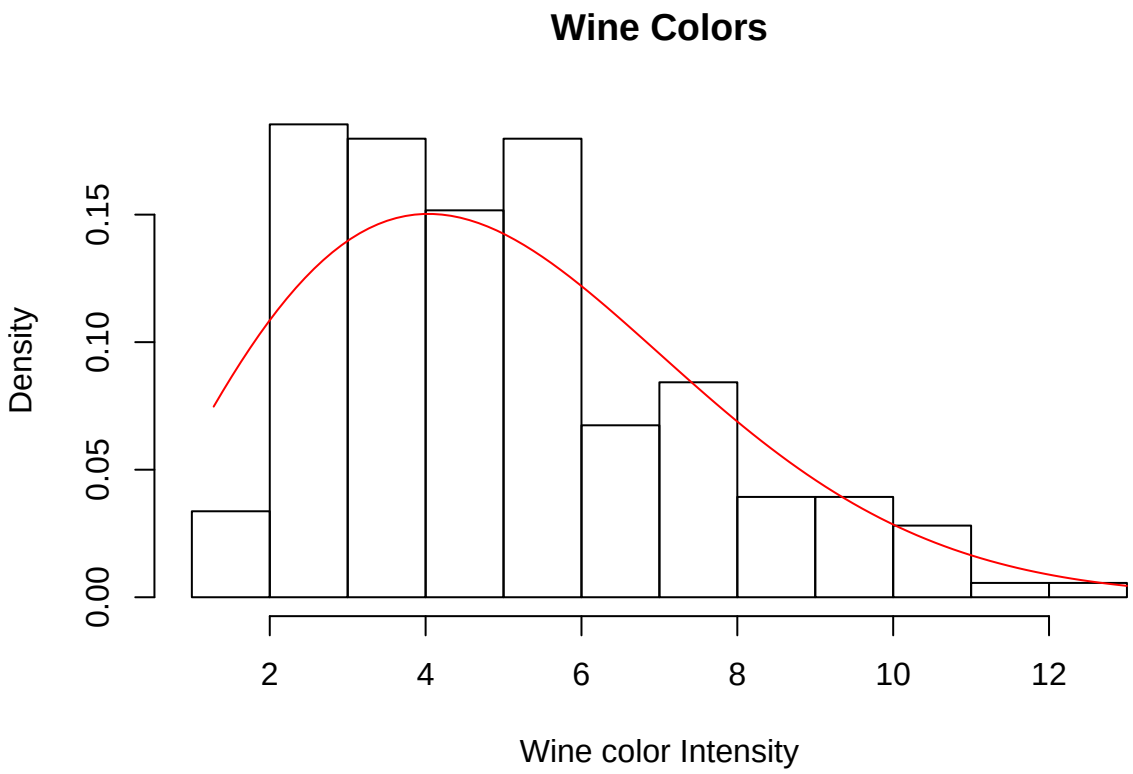
Additionally, the estimator is consistent

$$\lim_{n\to\infty} 0.145\frac{\theta^2}{n} = 0$$

11. MOM Estimate using wine data

$$\hat{\theta} = \frac{\bar{X}}{0.886} = \frac{5.058}{0.886} = 5.707$$

```
hist(color, main='Wine Colors', xlab='Wine color Intensity', freq = F, breaks = 12)
k <- 2
theta <- mean(color)/gamma(3/2)
curve(k * theta^(-k) * x^(k-1) * exp(-(x/theta)^k), from = min(color), to=max(color), add=T, col='red')
```



12. 99th percentile of the Half brain Distribution

$$\int_0^{\tau} \frac{2}{\theta^2} X_i e^{-\frac{X_i}{\theta}^2} dx = 0.99$$
$$\int_0^{\tau} \frac{2}{\theta^2} X_i e^{-\frac{X_i^2}{\theta^2}} dx = 0.99$$

let $u = X_i^2 \qquad du = 2X_i dx$

$$\int_0^{\tau} \frac{2}{\theta^2} X_i e^{-\frac{U}{\theta^2}} \frac{1}{2X_i} du = 0.99$$
$$\theta \int_0^{\tau} \frac{1}{\theta^2} e^{-\frac{U}{\theta^2}} du = 0.99$$

let $w = -\frac{U}{\theta^2} \qquad dW = -\frac{dU}{\theta^2}$

$$
\begin{aligned}
\int_0^\tau \frac{1}{\theta^2} e^W - \theta^2 dW &= 0.99 \\
-e^W \Big|_0^\tau &= 0.99 \\
e^{-\frac{U}{\theta^2}} \Big|_0^\tau &= 0.99 \\
e^{-\frac{x_i^2}{\theta^2}} \Big|_0^\tau &= 0.99 \\
-e^{-\frac{\tau^2}{\theta^2}} + e^{-\frac{0^2}{\theta^2}} &= 0.99 \\
-e^{-\frac{\tau^2}{\theta^2}} + 1 &= 0.99 \\
-e^{-\frac{\tau^2}{\theta^2}} &= 0.99 - 1 \\
e^{-\frac{\tau^2}{\theta^2}} &= 0.01 \\
\ln\left(e^{-\frac{\tau^2}{\theta^2}}\right) &= \ln(0.01) \\
-\frac{\tau^2}{\theta^2} &= -4.605 \\
\tau^2 &= 4.605\theta^2 \\
\tau &= 2.146\theta
\end{aligned}
$$

13. T estimate

```
tau <- theta * 2.146
nonpara <- quantile(color, 0.99)
```

Utilizing the distribution produced a higher τ value than using the quantile function:

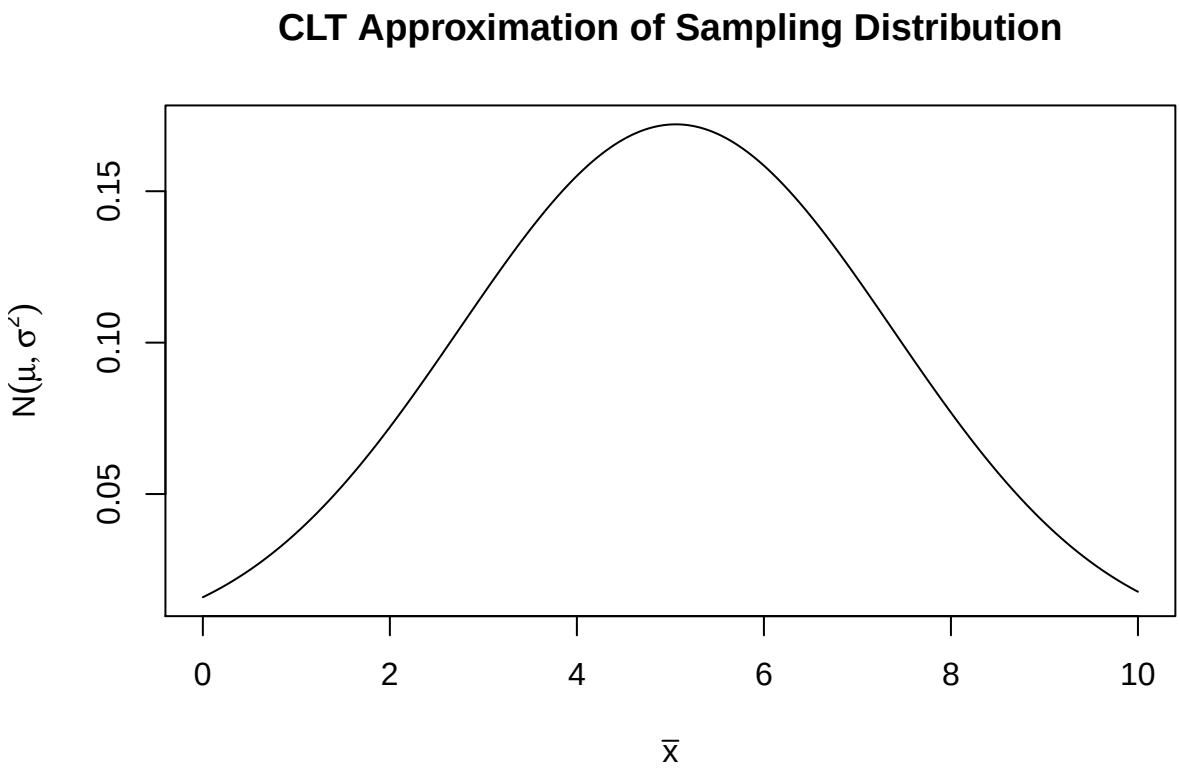
Quantile: 11.02
Distribution: 12.25

This suggests that the actual data is below Halfbrain's distribution. This means if we continued to use the distribution, one could hypothesize that we are more likely to over estimate than under estimate data.

Statistical Inference

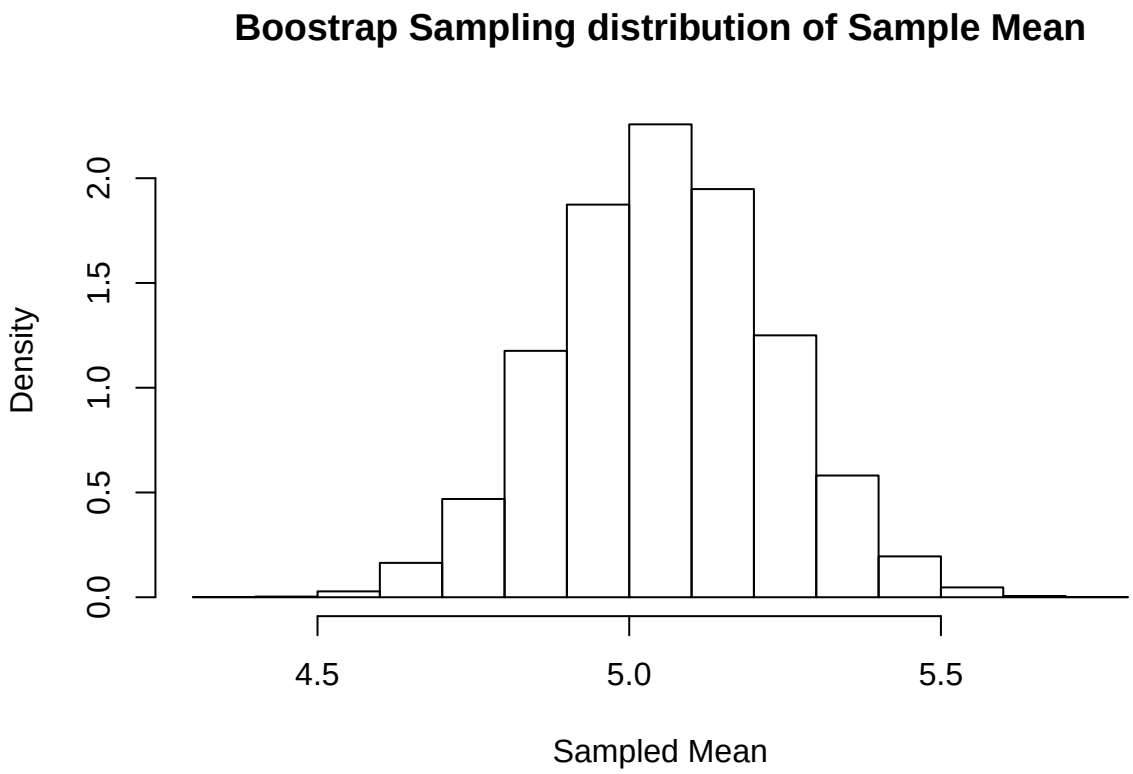
14. CLT Sampling distribution

```
x <- seq(0, 10, by=0.01)
plot(x, dnorm(x, mean(color), sd=sd(color)), type="l", main="CLT Approximation of Sampling Distribution
```



15. Boostrap Algorithm

```
B <- 10000
boot_samples <- rep(NA, B)
for(i in 1:B){
  boot_samples[i] <- mean(sample(color, 178, replace=TRUE))
}
hist(boot_samples, B, freq = F, breaks = 12, main = "Boostrap Sampling distribution of Sample Mean", xla
```



Yes, the CLT appears to apply as the sampled data looks approximately normal.

16. T-Procedures to produce a condfidence interval for μ

```
n <- length(color)
alpha <- 0.03
t <- qt(1-alpha/2, n - 1)
l <- xbar - t * (s/sqrt(n))
u <- xbar + t * (s/sqrt(n))
```

We are 97% certain that the true mean of wine color intensities is between 4.68 and 5.44 color intensity units.

```
boot_ci <- quantile(boot_samples, c(alpha/2, 1-alpha/2))
```

Utilizing the bootstrap algorithm, we achieved very similar results with a lower bound of 4.68 and 5.43 color intensity units.

17. T-Procedure for theta

$$\bar{X} = \hat{\theta}0.886$$

By replacing $\bar{X}$ with $\hat{\theta}$ we can be 97% condifident that the true value of theta is between 5.28 and 6.14.

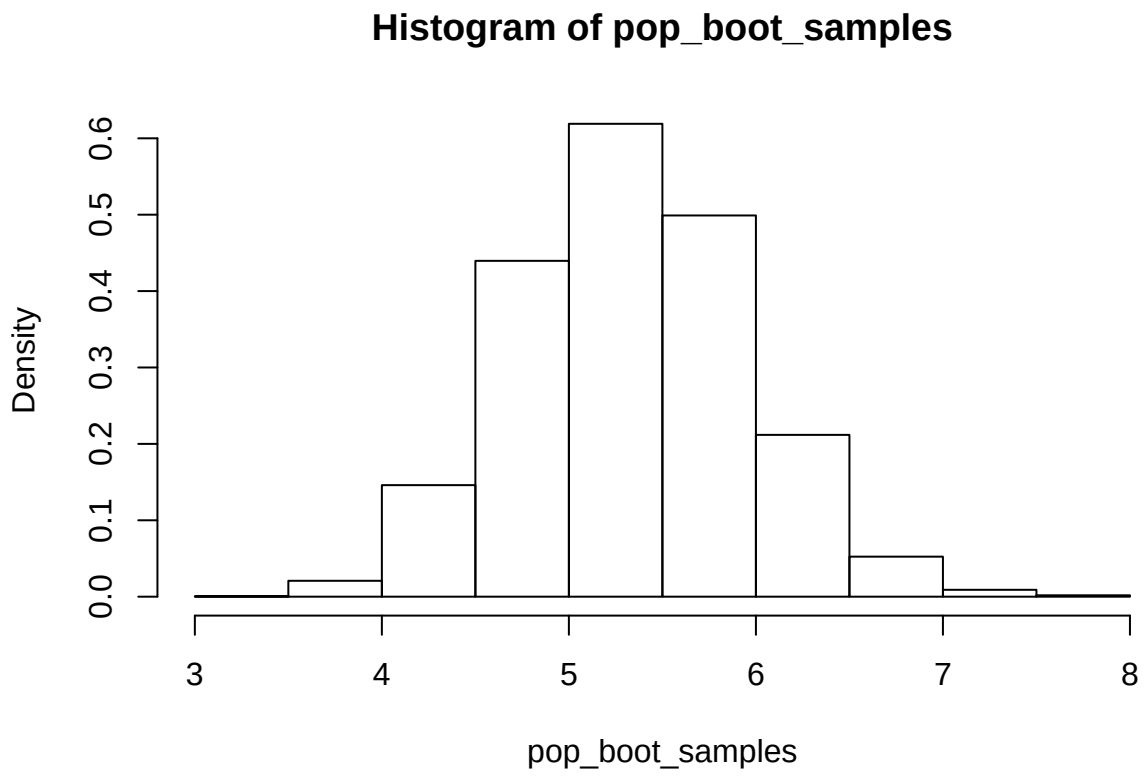
18. 95% Confidence interval for population variance

```
alpha <- 0.05
l <- ((n - 1) * s2)/qchisq(alpha/2, n - 1)
u <- ((n - 1) * s2)/qchisq(1 - alpha/2, n - 1)
```

We are 95% confident the true population variance is between 6.7 and 4.41. Our assumptions have been met as it's surrounding the calculated population variance of 5.37.

19. Bootstrap - Population Variance

```
B <- 10000
pop_boot_samples <- rep(NA, B)
for(i in 1:B){
  pop_boot_samples[i] <- var(sample(color, 178, replace=TRUE))
}
hist(pop_boot_samples, B, freq = F, breaks = 12)
```




```
pop_boot_ci <- quantile(pop_boot_samples, c(0.5/2, 1-0.5/2))
```

Although the bootstrap confidence interval produces a similar lower and upper interval of 4.91 and 5.74, one differernce is that the Chi Distribution produced an upper bound that was less than the lower, the bootstrap algorithm produced the opposite.

20. Wine Cultivar

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

Grignolino

Mean: 3.09

Variance: 0.86

Barolo

Mean: 5.53

Variance: 1.53

Barbera

Mean: 7.4

Variance: 5.34

21. Wine Comparisons

Barolo T-Procedure

```
barolo_xbar <- mean(x2)
barolo_s <- sd(x2)
n <- length(x2)
alpha <- 0.01
t <- qt(1-alpha/2, n - 1)
barolo_l <- barolo_xbar - t * (barolo_s/sqrt(n))
barolo_u <- barolo_xbar + t * (barolo_s/sqrt(n))
```

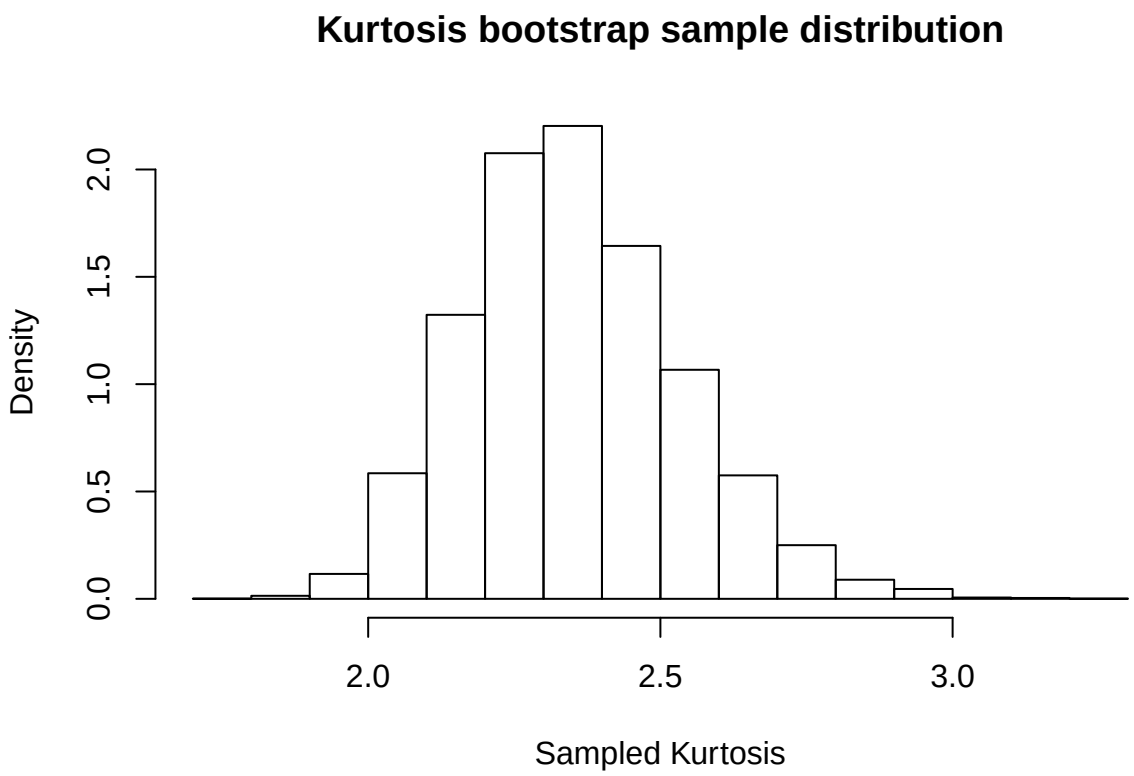
Barbera T-Procedure

```
barbera_xbar <- mean(x3)
barbera_s <- sd(x3)
n <- length(x3)
barbera_l <- barbera_xbar - t * (barbera_s/sqrt(n))
barbera_u <- barbera_xbar + t * (barbera_s/sqrt(n))

wine_u <- barbera_u - barolo_u
wine_l <- barbera_l - barolo_l
```

Using two T-Tests, we can say with 99% confidence that the true difference in means between the two wines is between 1.41 and 2.33. It can then be condluded that the Barbera wines are on average more intense in color than the Barolo.

22. Kurtosis

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

B <- 10000
kurt_boot_samples <- rep(NA, B)
for(i in 1:B){
  kurt_boot_samples[i] <- kurtosis(sample(bc_color, length(bc_color), replace=TRUE))
}
hist(kurt_boot_samples, B, freq = F, breaks = 12, main = "Kurtosis bootstrap sample distribution", xlab
```




```r
kurt_boot_ci <- quantile(kurt_boot_samples, c(0.04/2, 1-0.04/2))
```

Using the bootstrap algorithm, we are 96% confident that the true value of the population kurtosis is between 2.02 and 2.77. Since true value of kurtosis appears to be less than three, it suggests that the data is not normal as the tails have less extreme data than a traditional normal distrubtion. Furthermore, this implies much of the wine color intensities don't have extreme variances and one can generall expect for them to stay in a certain range.