Assignment2

Question 4

a)

```
non_white <- scan("wage_nonwhite.txt")
white <- scan("wage_white.txt")
av_difference <- mean(non_white) - mean(white);av_difference
```

```
## [1] -1.442187
```

```
mdn_difference <- abs(median(non_white) - median(white));mdn_difference
```

```
## [1] 1.135
```

b)

```
ks.test(non_white, white)
```

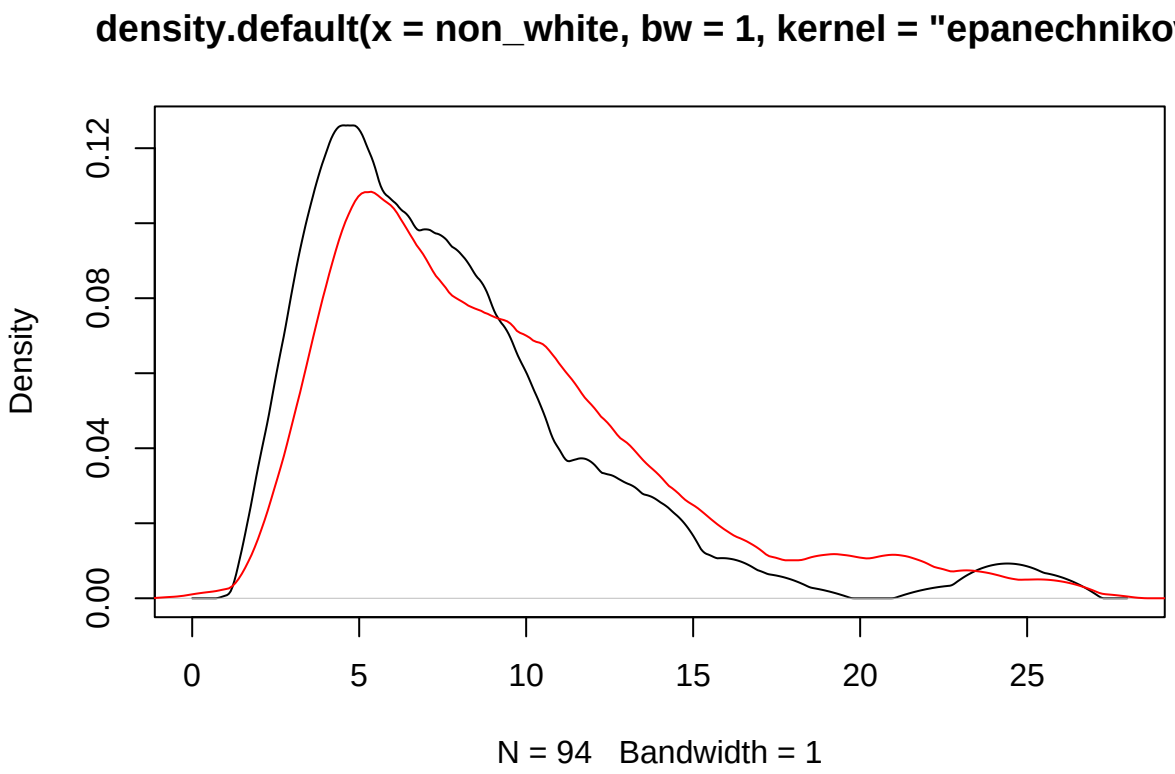
```
## Warning in ks.test(non_white, white): p-value will be approximate in the
## presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  non_white and white
## D = 0.17123, p-value = 0.02131
## alternative hypothesis: two-sided
```

The P-value is smaller than 0.05 therefore we can reject the null hypothesis and assume on a 95% confidence interval the samples come from different distributions.

c)

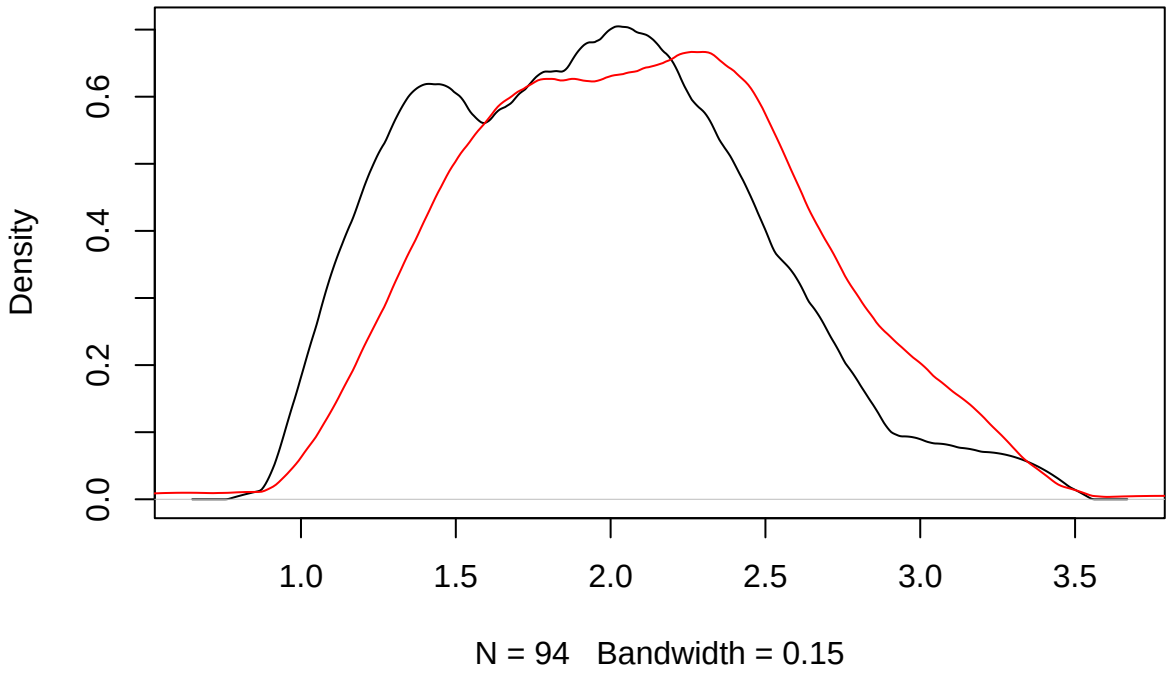
```
plot(density(non_white, bw=1, kernel="epanechnikov"))
lines(density(white, bw=1, kernel="epanechnikov"), col='red')
```



I used a bandwidth of 1 as a smaller bandwith causes the data to be over fitted, a larger bandwidth would cause the data to be over smoothed. I used an epanechnikov kernel to balance the bias-variance trade off. # d)

```
log_non <- log(non_white)
log_white <- log(white)

plot(density(log_non, bw= 0.15, kernel="epanechnikov"))
lines(density(log_white, bw=0.15, kernel="epanechnikov"), col='Red')
```



I used a bandwidth of 1 as a smaller bandwith causes the data to be over fitted, a larger bandwidth would cause the data to be over smoothed. I used an epanechnikov kernel to balance the bias-variance trade off. # e)

```
muw <- mean(log_white)
sigw <- sd(log_white)
mun <- mean(log_non)
sign <- sd(log_non)

ks.test(log_non,pnorm(mun,sign))
```

```
## Warning in ks.test(log_non, pnorm(mun, sign)): cannot compute exact p-value with
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  log_non and pnorm(mun, sign)
## D = 1, p-value = 0.2757
## alternative hypothesis: two-sided
```

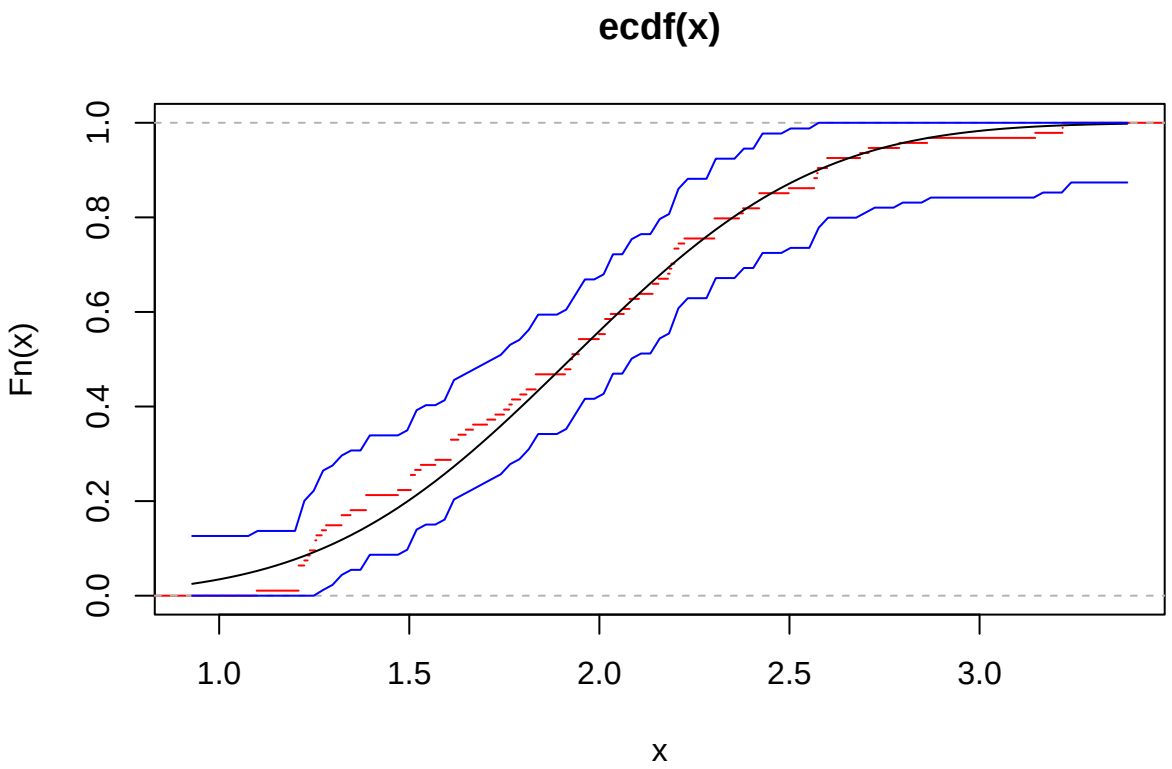
```
ks.test(log_white, pnorm( muw, sigw))
```

```
## Warning in ks.test(log_white, pnorm(muw, sigw)): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  log_white and pnorm(muw, sigw)
## D = 0.99318, p-value = 0.2786
## alternative hypothesis: two-sided
```

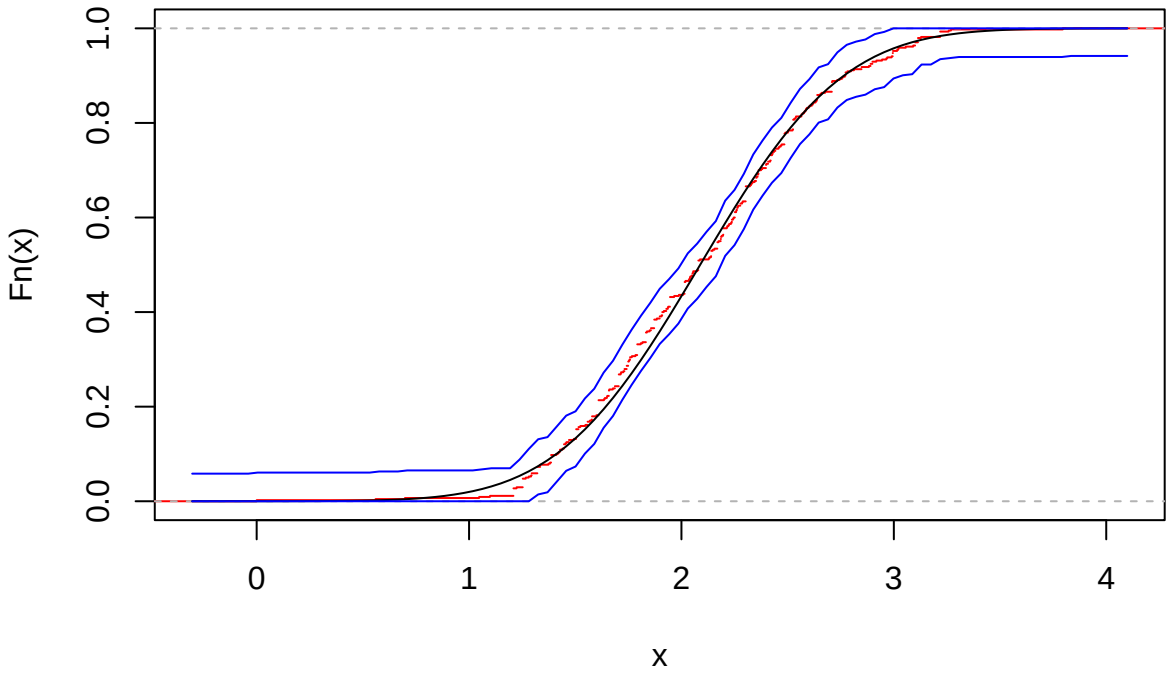
```
n = length(log_non)
alpha = 0.1
x = log_non
epsilon = sqrt(1/(2*n)*log(2/alpha))

est = ecdf(x)
plot(est,do.points=FALSE,col='red')
curve(pnorm(x, mun, sign),add=TRUE)
curve(pmax(est(x)-epsilon,0),col='blue',add=TRUE)
curve(pmin(est(x)+epsilon,1),col='blue',add=TRUE)
```




```
n = length(log_white)
alpha = 0.1
x = log_white
epsilon = sqrt(1/(2*n)*log(2/alpha))

est = ecdf(x)
plot(est,do.points=FALSE,col='red')
curve(pnorm(x, muw, sigw),add=TRUE)
curve(pmax(est(x)-epsilon,0),col='blue',add=TRUE)
curve(pmin(est(x)+epsilon,1),col='blue',add=TRUE)
```



a) White worker make an average of $1.44 more than non-white workers The difference in the median wage values is $1.14

b) The data sets come from different distributions. This means the spreads and shapes of the data will be different for non-white and white workers

c) non-white workers have a sharp increase and peaks at $4. This means most non-white workers have a wage of around $4.
the density of non-white wages then slowly decreases as wages increase. This means as wages increase the number of non-white workers earning that amount decreases.

white workers have a sharp increase and peaks at $6. This means most white workers have a wage of around $6.
the density of non-white wages then slowly decreases as wages increase. This means as wages increase the number of non-white workers earning that amount decreases.

the density of white workers seems to decrease slower than the density of non-white workers meaning there is a higher proportion of white workers

d) The white wages curve is shifted right when compared to the non-white wage curve. This suggests white workers have a higher average wage compared to non-white workers.

e) both p-values are greater than 0.1 in the ks test. This means we fail to reject our null hypothesis and therefore can conclude with 90% confidence that both log-wage samples come from a normal distribution. The ECDF's are not violated which confirms we can conclude with 90% confidence log wages are normally distributed.