Tutorial 2

Lucy Li

5/4/2020

Tutorial 2

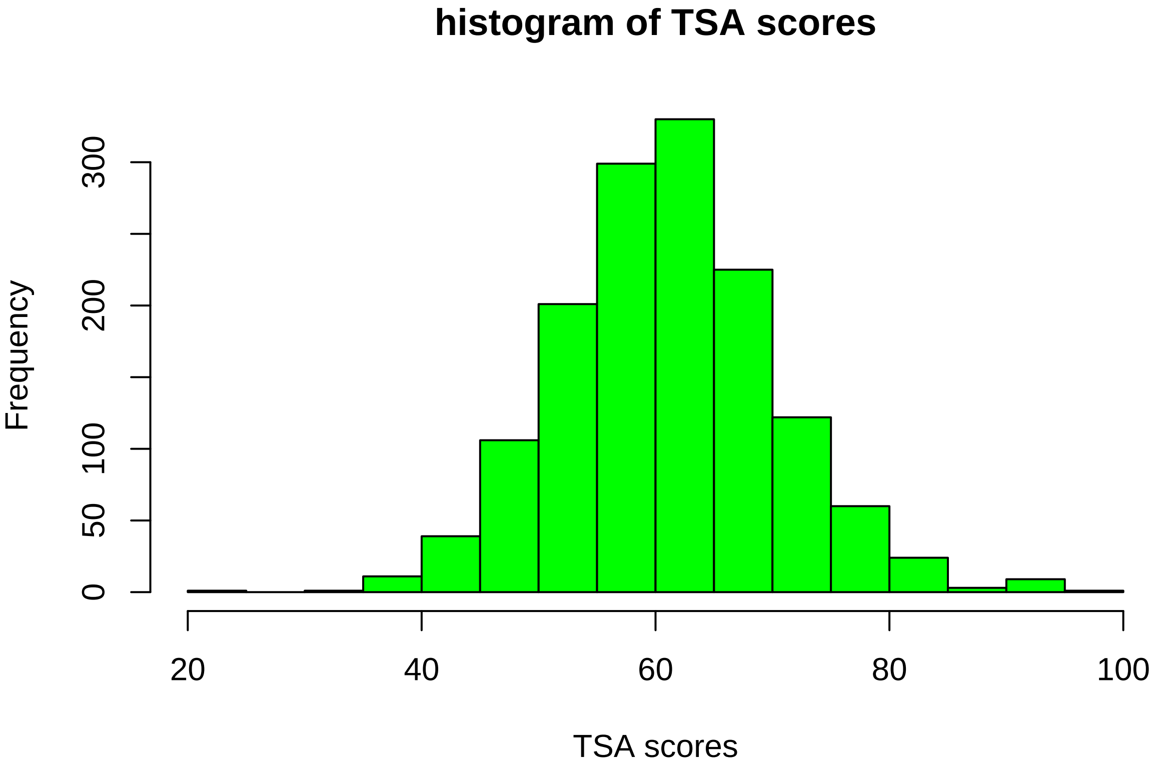
(a) Plot a histogram of the TSA score.

```
library(readr)
tsadata <- read_csv("tsadata.csv")
```

```
## Warning: Missing column names filled in: 'X5' [5], 'X6' [6], 'X7' [7]
```

```
## Parsed with column specification:
## cols(
##   TSA = col_double(),
##   Gender = col_character(),
##   SchoolType = col_character(),
##   Admit = col_double(),
##   X5 = col_logical(),
##   X6 = col_logical(),
##   X7 = col_logical()
## )
```

```
hist(tsadata$TSA, xlab="TSA scores", main="histogram of TSA scores",col="green",border="black")
```



(b) Estimate the mean and the standard deviation of the distribution of TSA scores.

```
mean(tsadata$TSA)
```

```
## [1] 61.05055
```

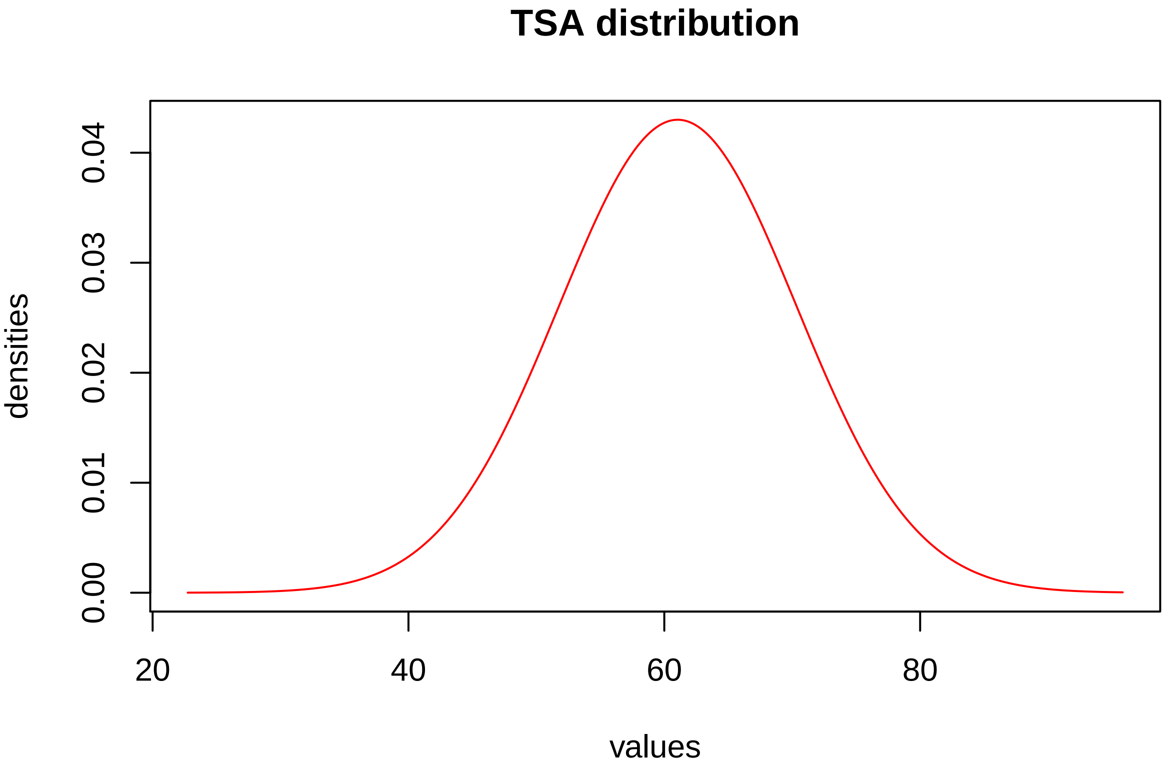
```
sd(tsadata$TSA)
```

```
## [1] 9.278309
```

(c) Using your estimates of the mean and standard deviation, plot the probability density function for a fitted normal distribution and compare it to the histogram which you obtained previously. Does the normal density seem to have captured the distribution of the data adequately?

Ans: The mean and the standard deviation is captured adequately, however, the actual distribution is more weighted to the right tha the normal distribution: if we compare the left intervals with their counterparts on the right and starting from the middle and iteratively increment/decrement (interval 60-65 has more frequency than interval 55-60 interval,interval 65-70 has more frequency than interval 50-55 interval, and so on).

```
x<-seq(min(tsadata$TSA),max(tsadata$TSA),by=.1)
y<-dnorm(x,mean(tsadata$TSA),sd(tsadata$TSA))
matplot(x=x,
        y=cbind(y),
        type="l",
        lty=1,
        col=c("red"),
        xlab="values",
        ylab="densities",
        main="TSA distribution")
```



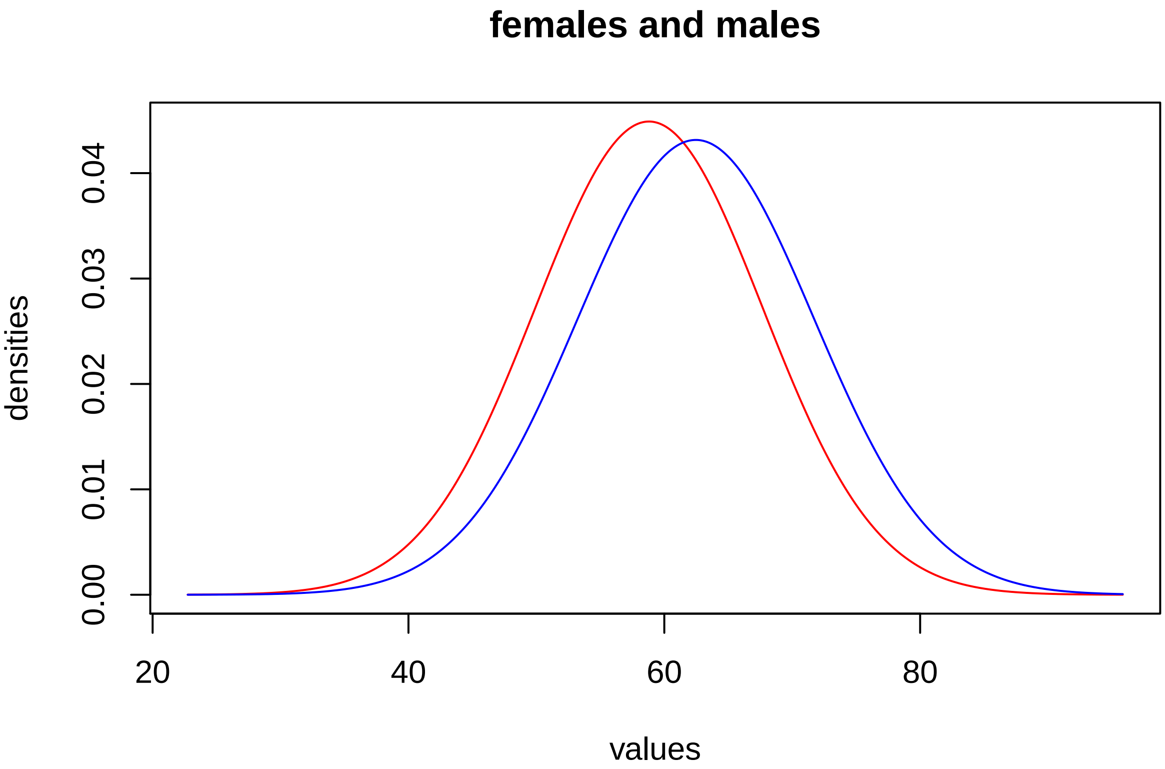
(d) Plot the fitted normal densities for the following on a single set of axes and comment briefly on the differences you see for:

i. Females and males

ii. The three different school types
iii. Those admitted and not admitted

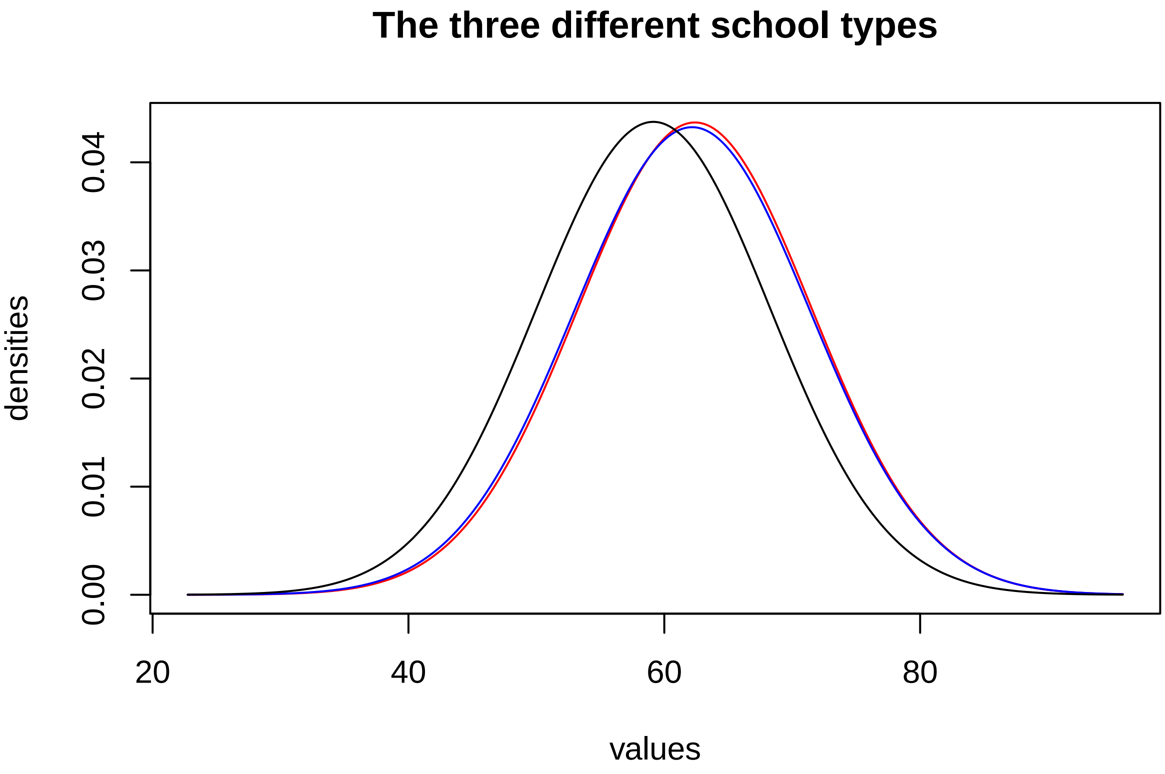
(i) Mean TSA for female (red) (approx 60) is lower than male (blue) (approx 64) , sd lower for female (more female's TSA scores distributed near the mean than male)

```r
x<-seq(min(tsadata$TSA),max(tsadata$TSA),by=.1)
y<-dnorm(x,mean(tsadata$TSA[tsadata$Gender=="F"]),sd(tsadata$TSA[tsadata$Gender=="F"]))
z<-dnorm(x,mean(tsadata$TSA[tsadata$Gender=="M"]),sd(tsadata$TSA[tsadata$Gender=="M"]))
matplot(x=x,
        y=cbind(y,z),
        type="l",
        lty=1,
        col=c("red","blue"),
        xlab="values",
        ylab="densities",
        main="females and males")
```



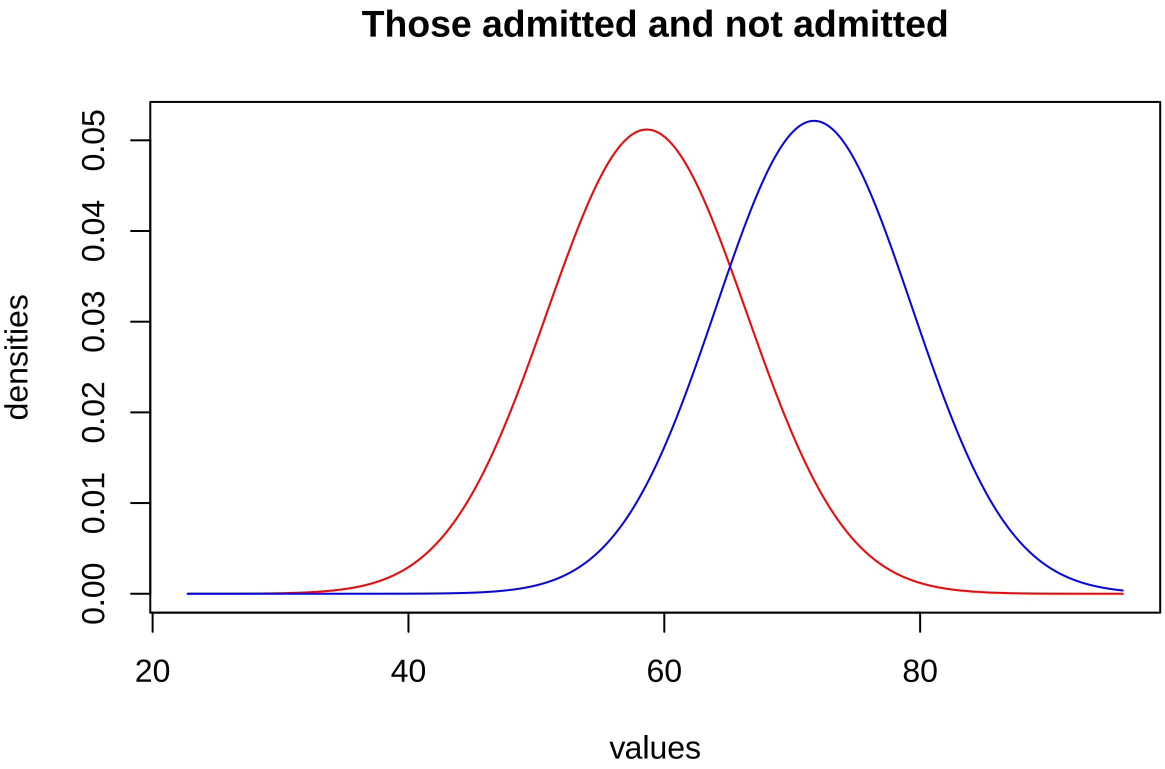
(ii) Very small discernable difference between state and independent school's TSA's normal pdf. Overseas school (black) has lower mean (approx. 60) than state (blue) and independent (red) school and state school's TSA (approx. 62) and approx. same s.d.

```r
x<-seq(min(tsadata$TSA),max(tsadata$TSA),by=.1)
#x<-seq(20,100,by=.1)
y<-dnorm(x,mean(tsadata$TSA[tsadata$SchoolType=="I"]),sd(tsadata$TSA[tsadata$SchoolType=="I"]))
z<-dnorm(x,mean(tsadata$TSA[tsadata$SchoolType=="S"]),sd(tsadata$TSA[tsadata$SchoolType=="S"]))
u<-dnorm(x,mean(tsadata$TSA[tsadata$SchoolType=="O"]),sd(tsadata$TSA[tsadata$SchoolType=="O"]))

matplot(x=x,
        y=cbind(y,z,u),
        type="l",
        lty=1,
        col=c("red","blue","black"),
        xlab="values",
        ylab="densities",
        main="The three different school types")
```



(iii) accepted candidates' mean TSA scores (red) (approx 72) is a lot higher than those not accepted (blue) (approx 58), s.d. of accepted candidates is a tiny bit smaller than those not accepted.

```
x<-seq(min(tsadata$TSA),max(tsadata$TSA),by=.1)
#x<-seq(20,100,by=.1)
y<-dnorm(x,mean(tsadata$TSA[tsadata$Admit=="0"]),sd(tsadata$TSA[tsadata$Admit=="0"]))
z<-dnorm(x,mean(tsadata$TSA[tsadata$Admit=="1"]),sd(tsadata$TSA[tsadata$Admit=="1"]))

matplot(x=x,
        y=cbind(y,z),
        type="l",
        lty=1,
        col=c("red","blue"),
        xlab="values",
        ylab="densities",
        main="Those admitted and not admitted")
```



(e) Test the hypotheses that

i. Females perform worse than males on the TSA test
ii. Applicants from Independent Schools perform better than applicants from State Schools
iii. There is no difference between the TSA test scores of those admitted and those not admitted

---

ANS i. p value is smaller than 1 percent, result is significant at 1 percent level. Reject null hypothesis and say that females perform worse than males on the TSA test

```
t.test(tsadata$TSA[tsadata$Gender=="F"],tsadata$TSA[tsadata$Gender=="M"],alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  tsadata$TSA[tsadata$Gender == "F"] and tsadata$TSA[tsadata$Gender == "M"]
## t = -7.4803, df = 1210.4, p-value = 7.103e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2.857435
## sample estimates:
## mean of x mean of y
##  58.80425  62.46792
```

ii. p value is greater than 10 percent, result is not significant at 10 percent level. Accept null hypothesis and say that there is no evidence to support the fact that applicants from Independent Schools perform better than applicants from State Schools

```
t.test(tsadata$TSA[tsadata$SchoolType=="I"],tsadata$TSA[tsadata$SchoolType=="S"],alternative="greater")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  tsadata$TSA[tsadata$SchoolType == "I"] and tsadata$TSA[tsadata$SchoolType == "S"]
## t = 0.34202, df = 865.75, p-value = 0.3662
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.8118529        Inf
## sample estimates:
## mean of x mean of y
##  62.38967  62.17683
```

iii. p value is much less than 1 percent, result is highly significant at 1 percent level. Reject null hypothesis and say that there is strong evidence to support the fact there is a difference between the TSA test scores of those admitted and those not admitted and that the TSA of admitted students is higher than rejected students

```
t.test(tsadata$TSA[tsadata$Admit==0],tsadata$TSA[tsadata$Admit==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  tsadata$TSA[tsadata$Admit == 0] and tsadata$TSA[tsadata$Admit == 1]
## t = -25.031, df = 398.06, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.10337 -12.04937
## sample estimates:
## mean of x mean of y
##  58.63069  71.70706
```