Assignment 1

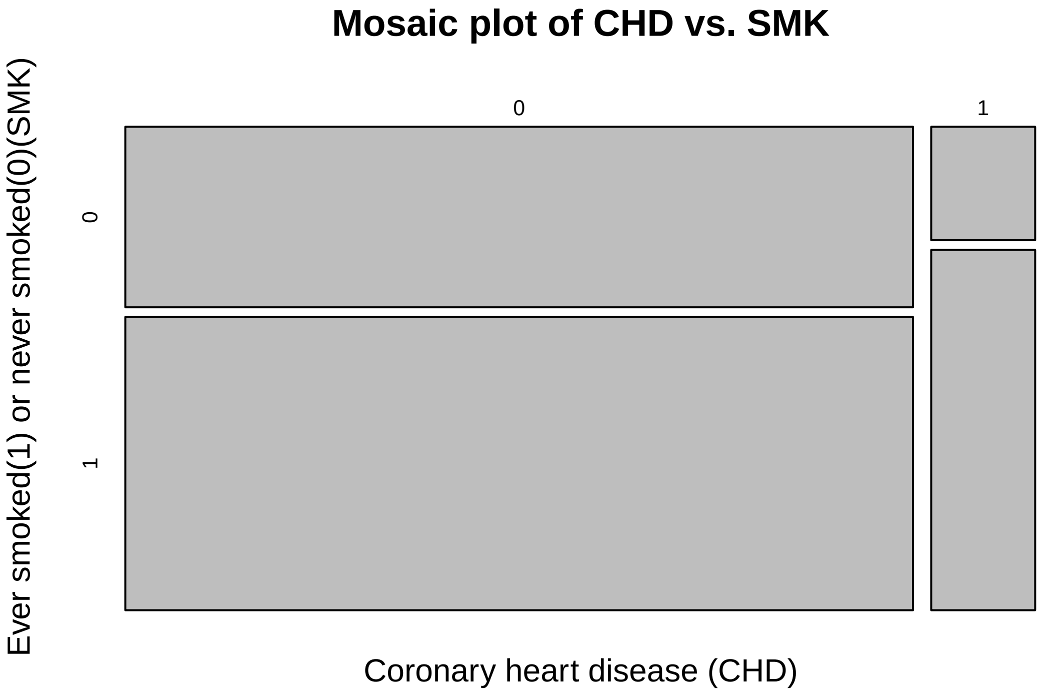
Anant Vishwakarma

4/13/2022

Contingency table and mosaic plot for CHD and smoking

The data set used for this analysis are the results of an epidemiological cohort study in which 609 subjects were followed for 7 years, with coronary heart disease as the outcome of interest. The first task was to investigate the relationship between coronary heart disease and smoking. This was done using a contingency table. A mosaic plot was then made using this contingency table.

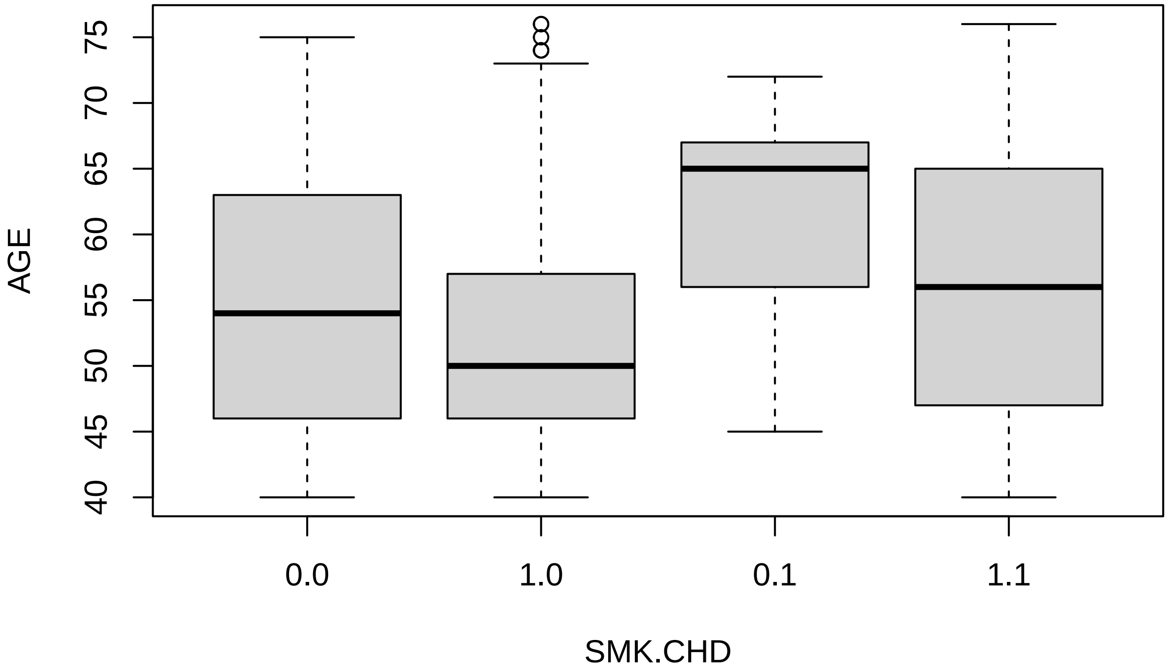
```
##                              Ever smoked(1) or never smoked(0)(SMK)
## Coronary heart disease (CHD)   0   1
##                             0 205 333
##                             1  17  54
```



From the mosiac plot we can see that a large proportion of people who had smoked in their lives but did not have coronary heart disease.

Boxplots for age with respect to CHD and smoking

The next task was to see how this relationship would change when Age was factored in. Since CHD and SMK are categorical variables a contingency table along with a mosaic plot was used, however age is a quantitative variable so a box plot was chosen to be the best option.



The ages of the subjects can be compared in terms of center, spread and shape in decreasing order of importance, The groups with SMK=0 and CHD=0 and with SMK=1 and CHD=1 do not appear to differ much in terms of center. The group with SMK=1 and CHD=0 appears to have a relatively lower center and the group with SMK=0 and CHD=1 appears to have a higher center. Furthermore the group with SMK=0 and CHD=1 appears to have lower spread among its age levels whereas the rest of the groups appear to be similar in their spreads.

Mean and standard deviation of age with respect to CHD and smoking

These results can be further quantified using means and standard deviations.

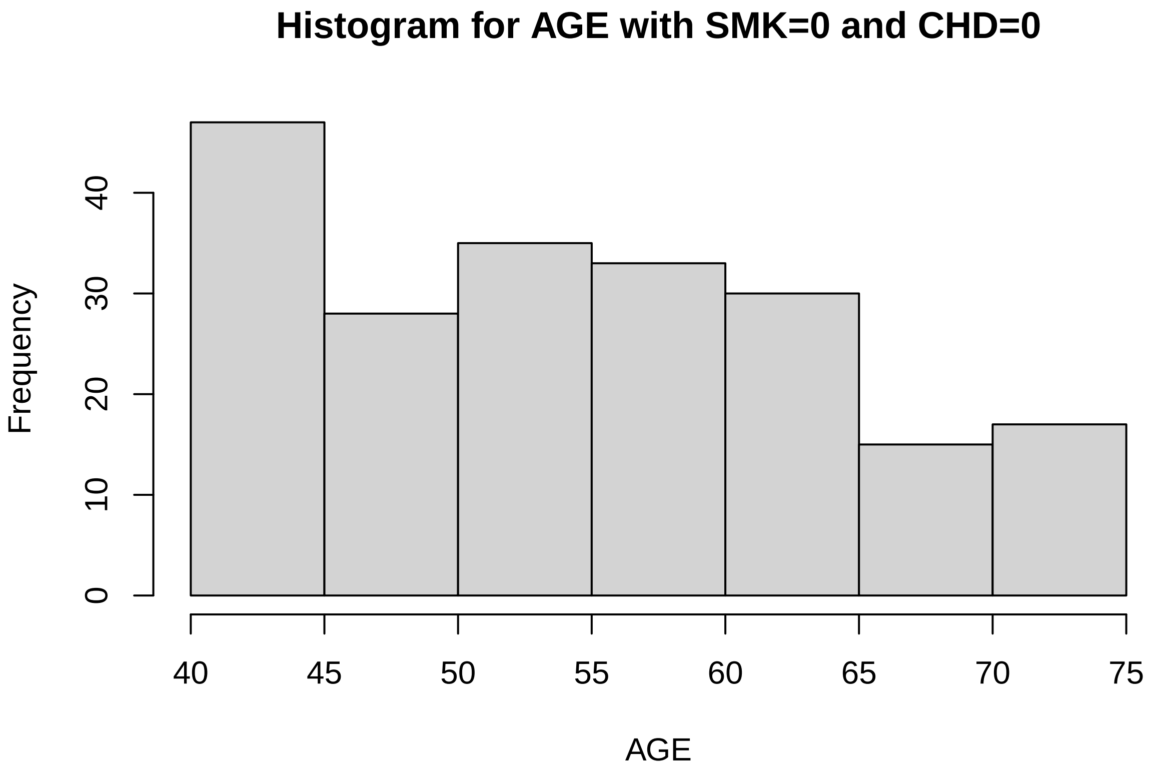
```
##   SMK CHD      AGE
## 1   0   0 54.88780
## 2   1   0 52.22222
## 3   0   1 61.64706
## 4   1   1 55.87037
```

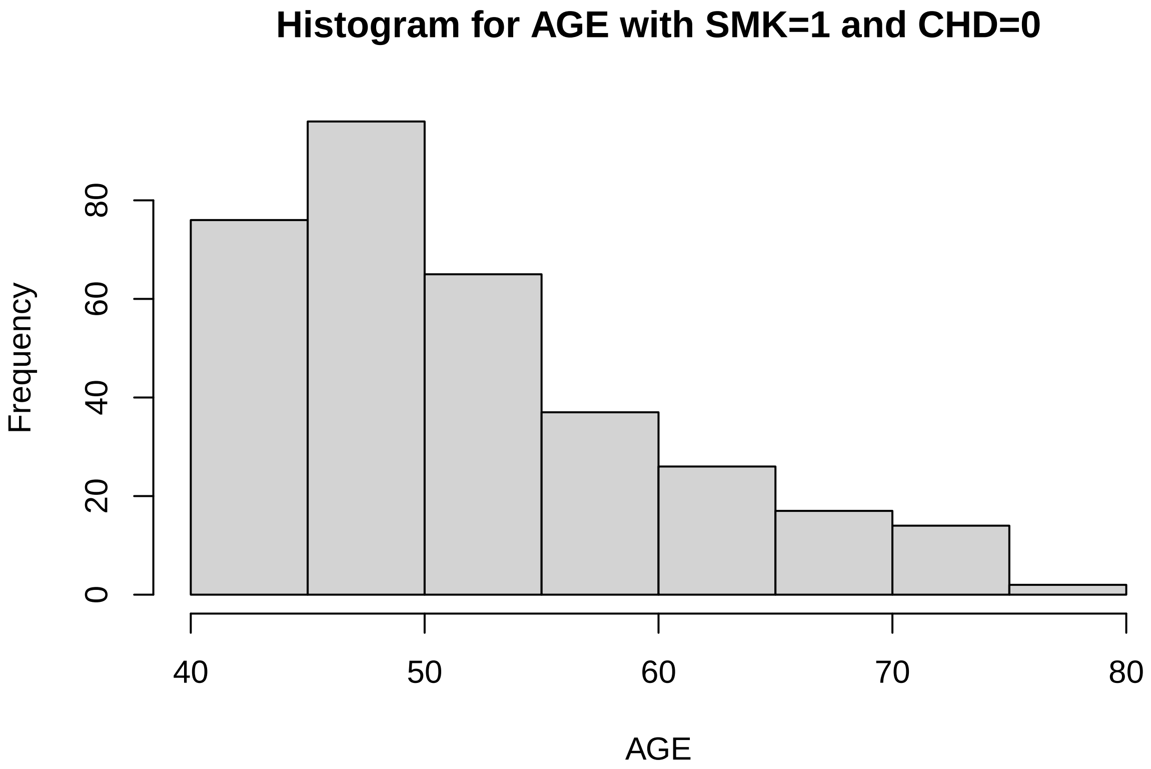
```
##   SMK CHD       AGE
## 1   0   0  9.672181
## 2   1   0  8.491944
## 3   0   1  8.146634
## 4   1   1 10.356014
```

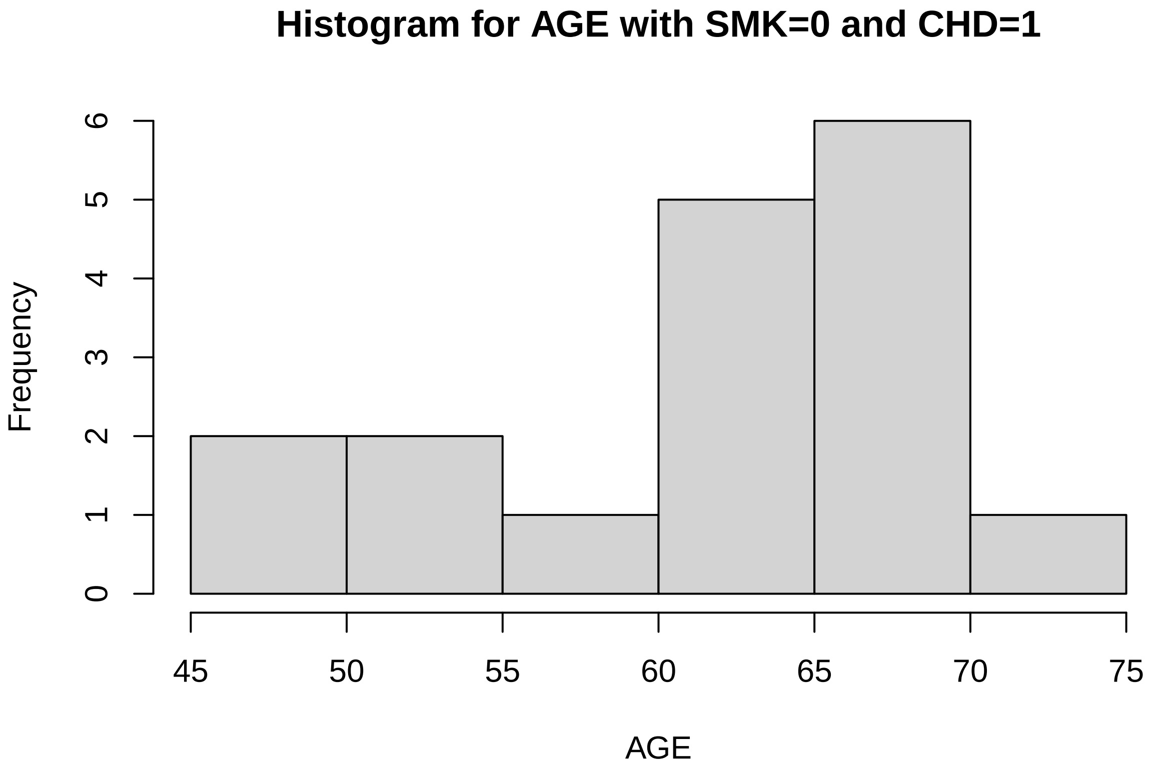
If we look at the standard deviations, the SDs of the group with (SMK=1;CHD=0) and the group with (SMK=0;CHD=1) are closer than they appear to be in the boxplot, other observations all seem to correspond correctly with the spreads visible in the boxplot.

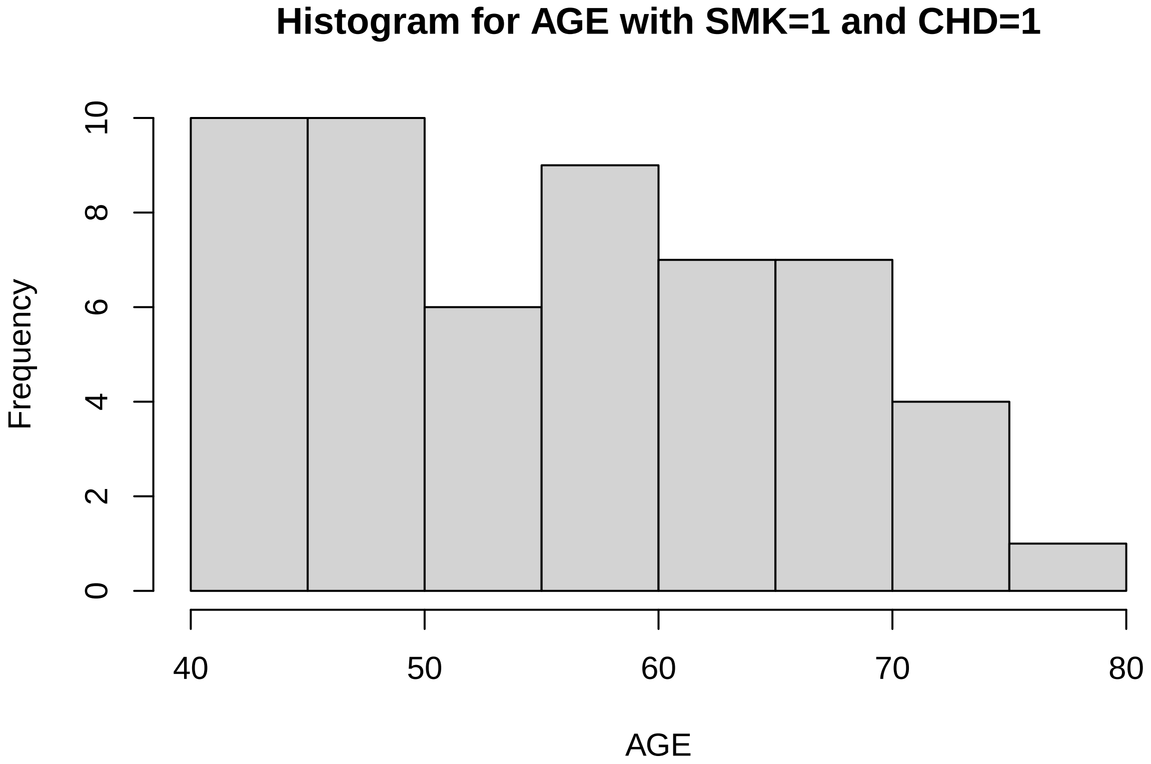
Histograms of age with respect to CHD and smoking

To get a better view of the data we can look at the histograms of the distributions of each group.









For the histograms we can see, the group with SMK=0 and CHD=0 is right skewed with the bulk being to the left, relatively high frequencies of age throughout the histogram and the highest frequency on the left. The group with SMK=0 and CHD=1 is also right skewed with a typical tail and head trend. The group with SMK=0 and CHD=1 has higher frequencies in the 60 to 70 age range. Lastly the group with SMK=1 and CHD=1 has the highest frequencies in the 40 to 50 age range and the lowest in 70 to 80 range.

Discussion of results

The results from these observations show the age range to 40 to 50 seems to have a higher distribution of people who have smoked at least once. Using the mosaic plot and the histogram it can be seen that there was a large number of people who smoked but did not have coronary heart disease. Among the people who smoked, the number of people who had CHD is higher but that isn't enough to establish a connection between SMK and CHD. Therefore, coronary heart disease seems to have more of a correlation with age rather than smoking.

Appendix

```r
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
vars <- c("ID","CHD","CAT","AGE","CHL","SMK","ECG","DBP","SBP","HPT","CH","CC")

evans <- read.table("http://www.stat.ucdavis.edu/~affarris/evans.dat",
                    header=FALSE,col.names=vars)
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
ChdSmkTable <-table("Coronary heart disease (CHD)"=evans$CHD,
                    "Ever smoked(1) or never smoked(0)(SMK)"=evans$SMK)
ChdSmkTable
mosaicplot(ChdSmkTable,main="Mosaic plot of CHD vs. SMK")
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
boxplot(AGE~SMK+CHD,
        data=evans,
        xlab="SMK.CHD")
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
aggregate(AGE~SMK+CHD,data=evans,mean)
aggregate(AGE~SMK+CHD,data=evans,sd)
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center')
hist(evans$AGE[evans$SMK==0&evans$CHD==0],
     xlab = "AGE",
     ylab = "Frequency",
     main = "Histogram for AGE with SMK=0 and CHD=0")
hist(evans$AGE[evans$SMK==1&evans$CHD==0],
     xlab = "AGE",
     ylab = "Frequency",
     main = "Histogram for AGE with SMK=1 and CHD=0")
hist(evans$AGE[evans$SMK==0&evans$CHD==1],
     xlab = "AGE",
     ylab = "Frequency",
     main = "Histogram for AGE with SMK=0 and CHD=1")
hist(evans$AGE[evans$SMK==1&evans$CHD==1],
     xlab = "AGE",
     ylab = "Frequency",
     main = "Histogram for AGE with SMK=1 and CHD=1")
```