Lab Assignment 2

STA 100 | A. Farris | Spring 2021

A pdf copy of your assignment is due at 5pm on Monday, April 26. Submission of the pdf will be through Gradescope. Please put in the effort to make it look reasonably professional – you're encouraged to use R Markdown. Note that specific tasks for you are highlighted.

Introduction to the Evans data

In this assignment, we will investigate 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. This is an opportunity for us to explore the use of some summary statistics, and to illustrate useful computational tools like subsetting. We begin by obtaining the data:

```
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)
```

This data contains much information about the subjects:

Variable	Description
ID	Subject ID, one observation per subject
CHD	Coronary heart disease (1) or not (0)
CAT	High catecholamine level (1) or not (0)
AGE	Age in years
CHL	Cholesterol level
SMK	Ever smoked (1) or never smoked (0)
ECG	ECG abnormality (1) or not (0)
DBP	Diastolic blood pressure
SBP	Systolic blood pressure
HPT	$= 1$ if DBP ≥ 90 or SBP ≥ 160, otherwise $= 0$
CH	CAT*HPT
CC	CAT*CHL

We can see what the first six subjects look like using the `head` function:

```
head(evans)
```

```
  ID CHD CAT AGE CHL SMK ECG DBP SBP HPT CH  CC
1 21   0   0  56 270   0   0  80 138   0  0   0
2 31   0   0  43 159   1   0  74 128   0  0   0
3 51   1   1  56 201   1   1 112 164   1  1 201
4 71   0   1  64 179   1   0 100 200   1  1 179
5 74   0   0  49 243   1   0  82 145   0  0   0
6 91   0   0  46 252   1   0  88 142   0  0   0
```

We can easily obtain simple summary statistics, for example using `mean` and `sd`:

```
mean(evans$AGE)
```

```
[1] 53.70608
```

```r
sd(evans$AGE)
```

```
[1] 9.258388
```

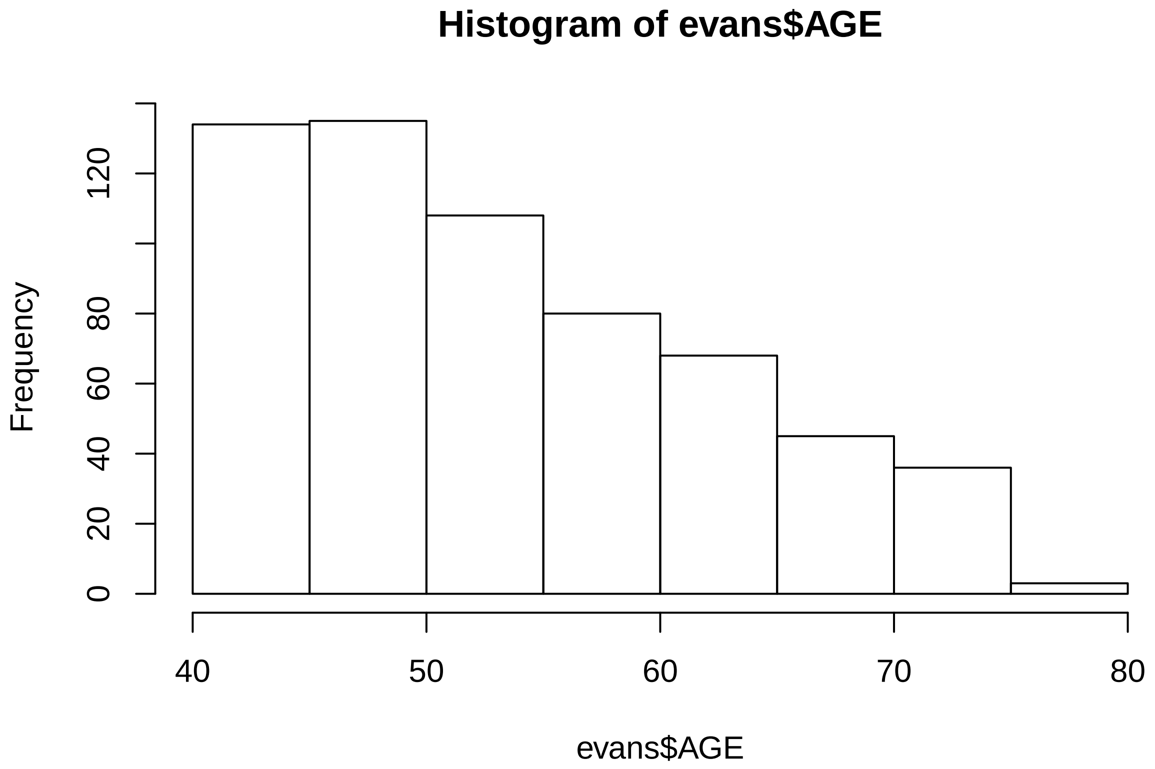
We can get the minimum, maximum, and first three quartiles (along with the mean) using `summary`:

```r
summary(evans$AGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  40.00   46.00   52.00   53.71   60.00   76.00 
```

It is tempting to think, because the mean is greater than the median, that the distribution of ages in this study might be right skewed. We can check this using a histogram :

```r
hist(evans$AGE)
```



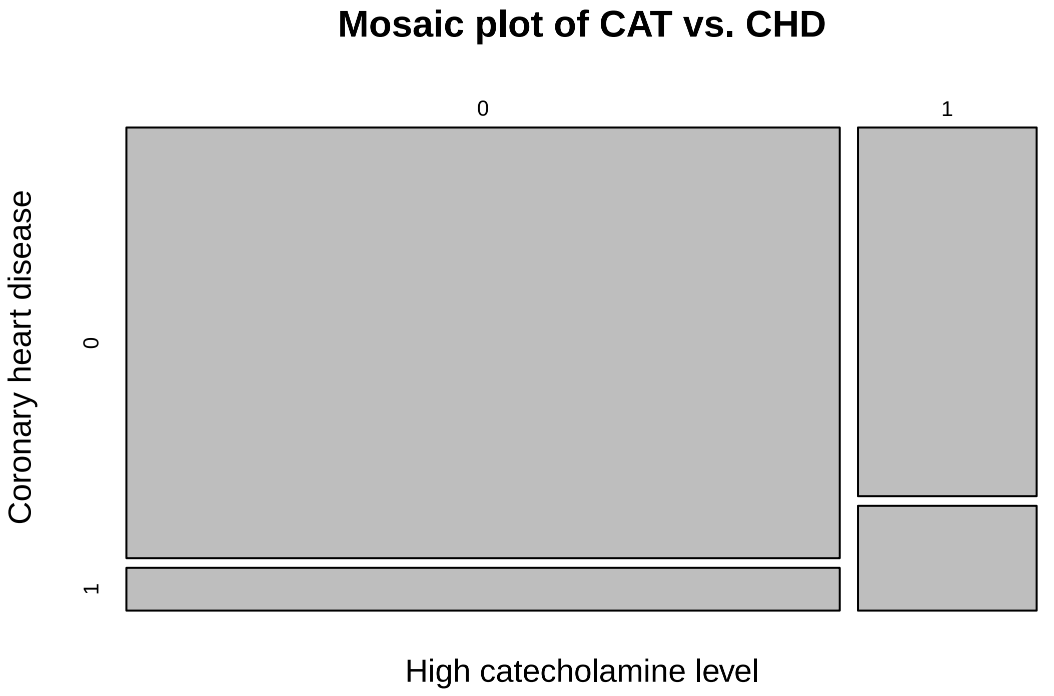
Let's investigate the relationship between coronary heart disease and catecholamine level in the subjects. Because both are recorded as categorical data, we can use a table with the frequency distribution (contingency table):

```r
CatChdTable <- table("High catecholamine level"=evans$CAT,
                  "Coronary heart disease"=evans$CHD)
CatChdTable
```

```
                        Coronary heart disease
High catecholamine level   0   1
                       0 443  44
                       1  95  27
```

and a corresponding mosaic plot:

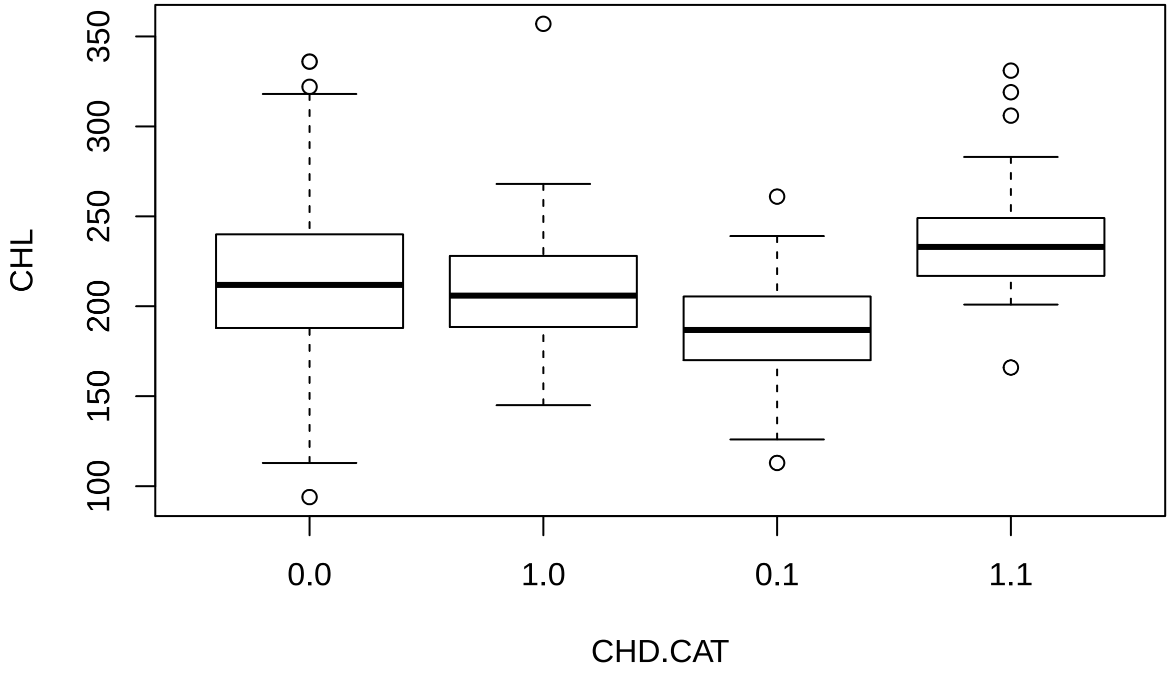
```r
mosaicplot(CatChdTable,
           main="Mosaic plot of CAT vs. CHD")
```



From these we can see that a disproportionate number of the subjects with high catecholamine levels also had coronary heart disease.

Let's investigate further. If we factor in cholesterol levels as well, what can we say? Since cholesterol is a quantitative variable, while CAT and CHD are both categorical (grouping) variables, we could use aligned boxplots:

```r
boxplot(CHL~CHD+CAT,
        data=evans,
        xlab="CHD.CAT")
```



We can compare the cholesterol level distributions for these groups in terms of center, spread, and shape, in decreasing order of importance. The groups with CHD=0 and CAT=0, and with CHD=1 and CAT=0 (respectively marked as 0.0 and 1.0) appear not to differ much in terms of center. Those with CHD=0 and CAT=1 have relatively lower center, though, and those with CHD=1 and CAT=1 have higher center. Furthermore, the group with CHD=0 and CAT=0 seem to have higher spread among their cholesterol levels, while the other three groups' spreads seem similar.

We can further quantify these using means

```
aggregate(CHL~CHD+CAT,
          data=evans,
          mean)
```

```
  CHD CAT      CHL
1   0   0 215.2325
2   1   0 210.6364
3   0   1 187.8211
4   1   1 240.3704
```

and (sample) standard deviations

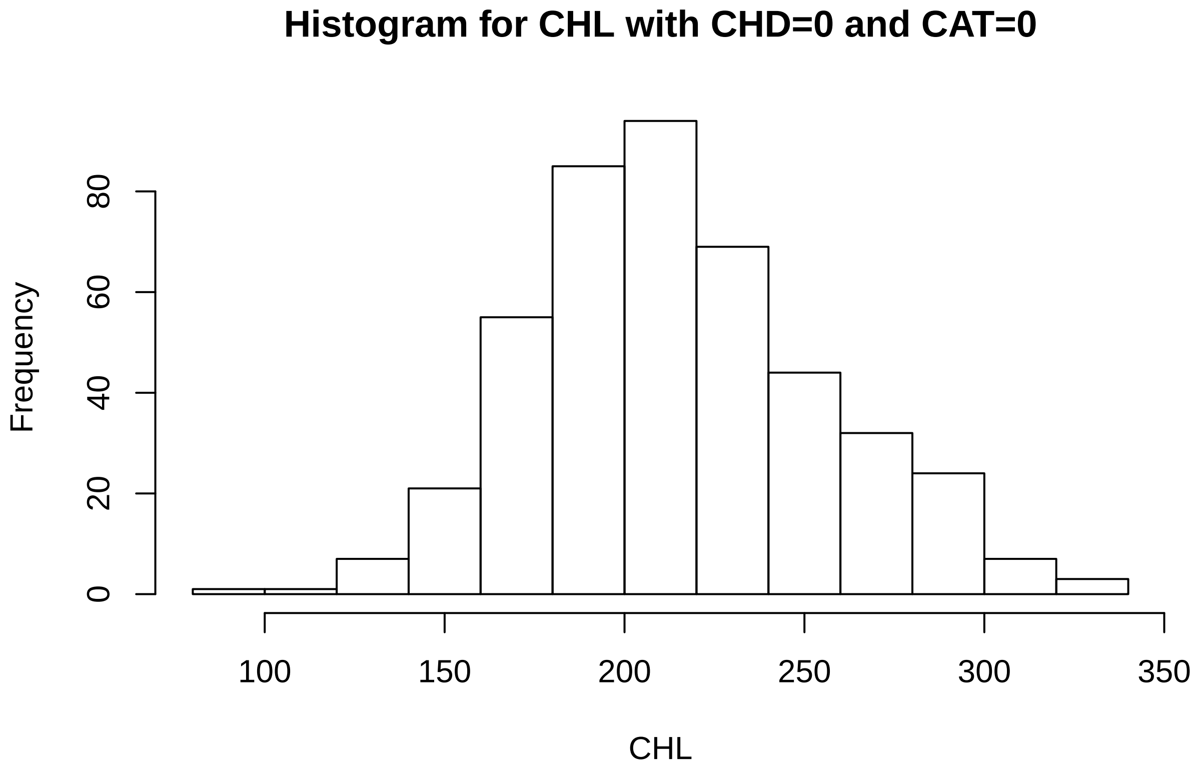
```
aggregate(CHL~CHD+CAT,
          data=evans,
          sd)
```

```
  CHD CAT      CHL
1   0   0 40.34958
2   1   0 37.49238
3   0   1 26.70365
4   1   1 36.88355
```

. Notice that the standard deviations show more variation between spreads in the latter three groups than is apparent in the boxplots. The standard deviations for the CHD=0 and CAT=0 and CHD=1 and CAT=0 groups are closer than the spreads appear in the boxplots, which may be due to the presence of a single, very extreme value in the latter group that is visible in the boxplot (remember that SDs, like means, are attracted to extreme values).

If we need to, we can look at the shapes of the distributions for each group using histograms. For example, for the group with CHD=0 and CAT=0, we see a distribution that is roughly bell shaped, with a slight skew in the direction of higher cholesterol levels:

```
hist(evans$CHL[evans$CHD==0 & evans$CAT==0],
     xlab = "CHL",
     ylab = "Frequency",
     main = "Histogram for CHL with CHD=0 and CAT=0")
```



The differences in centers of the distributions of cholesterol between these groups is the most clearly interpretable result here, with which we can summarize the relationship between catecholamine level, cholesterol levels, and coronary heart disease. We can interpret these differences to tell us that subjects with low catecholamine levels have similar average levels of cholesterol whether they have coronary heart disease or not. However, subjects with high catecholamine levels have seemingly higher levels of cholesterol on average when they have coronary heart disease than when they do not.

Assignment

Using a contingency table, investigate the relationship between coronary heart disease and smoking.

Also, investigate the relationship between coronary heart disease, smoking, and age. How do the distributions of age differ between the four groups obtained by heart disease and smoking status? Use aligned boxplots, means, standard deviations, and if necessary, histograms to compare these. What do these tell you about the relationship between smoking and coronary heart disease?

Appendix: R Script

```r
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)
head(evans)
mean(evans$AGE)
sd(evans$AGE)
summary(evans$AGE)
hist(evans$AGE)
CatChdTable <- table("High catecholamine level"=evans$CAT,
                 "Coronary heart disease"=evans$CHD)
CatChdTable
mosaicplot(CatChdTable,
           main="Mosaic plot of CAT vs. CHD")
boxplot(CHL~CHD+CAT,
        data=evans,
        xlab="CHD.CAT")
aggregate(CHL~CHD+CAT,
          data=evans,
          mean)
aggregate(CHL~CHD+CAT,
          data=evans,
          sd)
hist(evans$CHL[evans$CHD==0 & evans$CAT==0],
     xlab = "CHL",
     ylab = "Frequency",
     main = "Histogram for CHL with CHD=0 and CAT=0")
```