Introductory Econometrics: Chapter 1

Author: J M Woolridge

R Code Compilation by RJ Neel

End of Chapter 1 exercises (Page 39)

Computer Exercises

C1: Use the data in WAGE1 for this exercise

```
## Computer Exercise C1
library(wooldridge) #load the Woolridge Package
wage1
?wage1 #Description of the dataset

## starting httpd help server ... done

head(wage1) #First 6 rows. Easy to view
ncol(wage1) # No of rows
nrow(wage1) #No. of columns
```

(i) Find the average education level in the sample. What are the lowest and highest years of education?

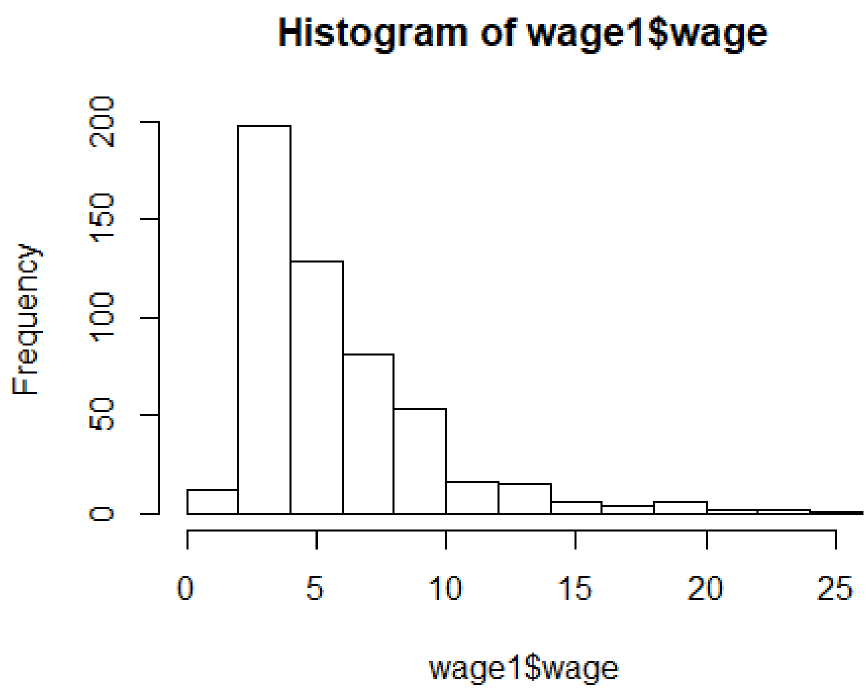
Solution

```
summary(wage1$educ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   12.56   14.00   18.00

#Alternatively
mean(wage1$educ) #avg education level

## [1] 12.56274

min(wage1$educ) #min education leve

## [1] 0

max(wage1$educ) #max

## [1] 18
```

(ii) Find the average hourly wage in the sample. Does it seem high or low?

Solution

```
mean(wage1$wage) #Gives you the average hourly wage

## [1] 5.896103

summary(wage1$wage) #Wage appears to be low

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.530   3.330   4.650   5.896   6.880  24.980

hist(wage1$wage) #Clearly skewed towards right
```



(iii) The wage data are reported in 1976 dollars. Using the Internet or a printed source, find the Consumer Price Index (CPI) for the years 1976 and 2013.

Solution Using Table B-60 in the 2004 Economic Report of the President, the CPI was 56.9 in 1976 and 233 in 2013.

(iv) Use the CPI values from part (iii) to find the average hourly wage in 2013 dollars. Now does the average hourly wage seem reasonable?

Solution To convert 1976 dollars into 2013 dollars, we use the ratio of the CPIs, which is 233/56.9 ≈ 4.09. Therefore, the average hourly wage in 2013 dollars is roughly 4.09($\5.90) ≈ $\24.13, which is a reasonable figure.

(v) How many women are in the sample? How many men?

Solution

```
head(wage1)

##   wage educ exper tenure nonwhite female married numdep smsa northcen sout
h
## 1 3.10   11     2      0        0      1       0      2    1        0
0
## 2 3.24   12    22      2        0      1       1      3    1        0
0
## 3 3.00   11     2      0        0      0       0      2    0        0
0
## 4 6.00    8    44     28        0      0       1      0    1        0
0
## 5 5.30   12     7      2        0      0       1      1    0        0
0
## 6 8.75   16     9      8        0      0       1      0    1        0
0
##   west construc ndurman trcommpu trade services profserv profocc clerocc
## 1    1        0       0        0     0        0        0       0       0
## 2    1        0       0        0     0        1        0       0       0
## 3    1        0       0        0     1        0        0       0       0
## 4    1        0       0        0     0        0        0       0       1
## 5    1        0       0        0     0        0        0       0       0
## 6    1        0       0        0     0        0        1       1       0
##   servocc    lwage expersq tenursq
## 1       0 1.131402       4       0
## 2       1 1.175573     484       4
## 3       0 1.098612       4       0
## 4       0 1.791759    1936     784
## 5       0 1.667707      49       4
## 6       0 2.169054      81      64

#Notice the female column is a binary variable implying 1 for female and 0 fo
r male requring us to proceed with 'dplyr'
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

w=nrow(wage1 %>% group_by(female) %>% filter(female=='1'))
w

## [1] 252
```

```
m=nrow(wage1)-w
m

## [1] 274
```

End of Computer Exercise 1

C2: Use the data in BWGHT to answer this question

(i) How many women are in the sample, and how many report smoking during pregnancy?

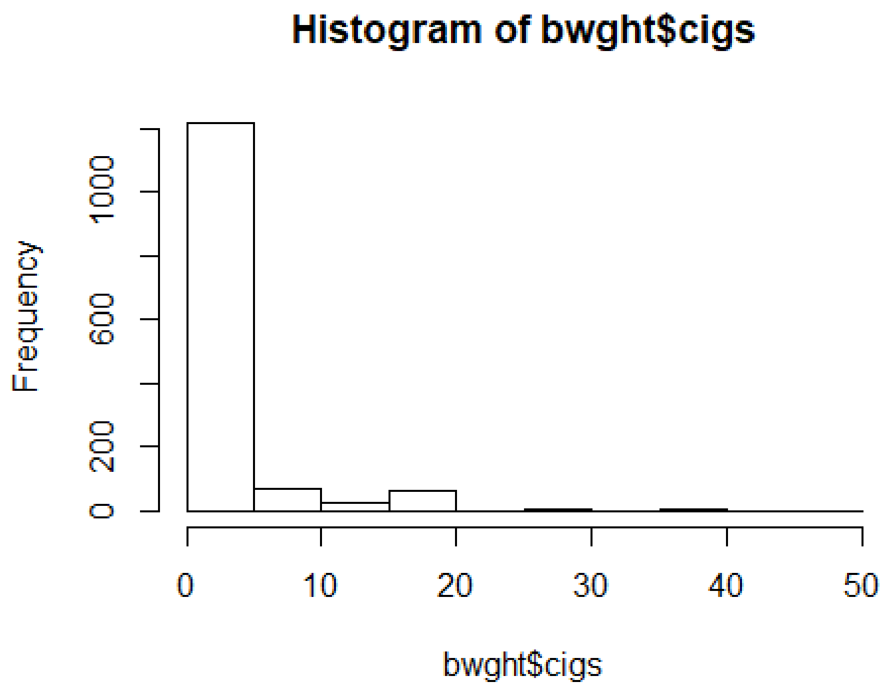
Solution

```
#Note: This data set contains all women
nrow(bwght) #No of women smoking

## [1] 1388
```

(ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.

Solution

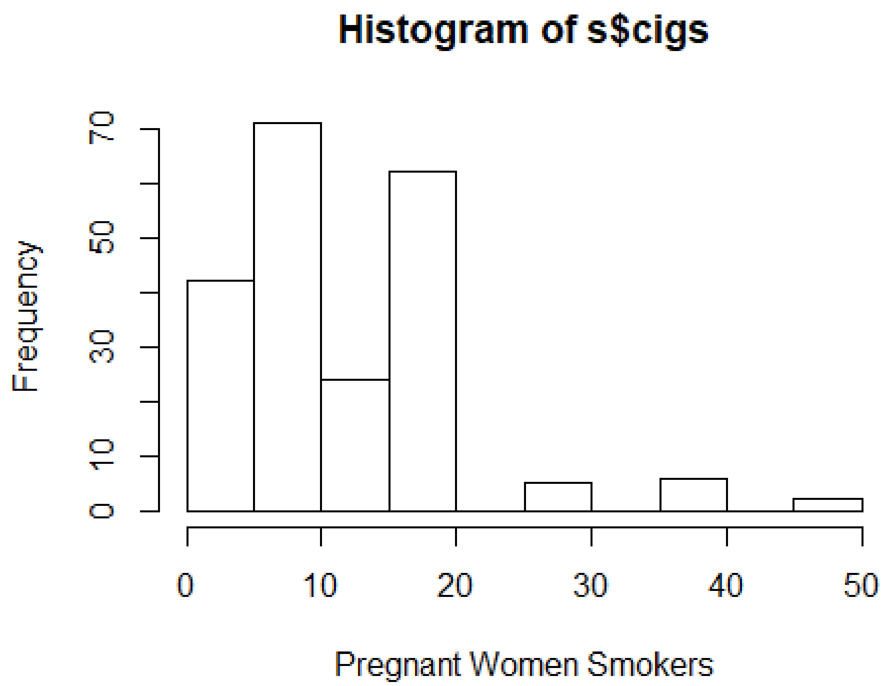
```
mean(bwght$cigs)

## [1] 2.087176

summary(bwght$cigs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.087   0.000  50.000

hist(bwght$cigs)
```




```
#Based on the histogram and range it appears to be a good measure.
```

(iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?

```
avg_all=mean(bwght$cigs)
avg_all

## [1] 2.087176

library(dplyr)
nrow(bwght %>% group_by(cigs) %>% filter(cigs=='0'))

## [1] 1176

s=(bwght %>% group_by(cigs) %>% filter(cigs>'0'))
avg_s=mean(s$cigs)
avg_s

## [1] 13.66509

hist(s$cigs, xlab='Pregnant Women Smokers')
```




```
#It markedly differs from the privious average by about 11 units.
```

C3 The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

```
head(meap01)

##   dcode bcode math4 read4 lunch enroll  expend    exppp  lenroll  lexpend
## 1  1010  4937  83.3  77.8 40.60    468 2747475 5870.673 6.148468 14.82619
## 2  2070   597  90.3  82.3 27.10    679 1505772 2217.632 6.520621 14.22482
## 3  2080  4860  61.9  71.4 41.75    400 2121871 5304.678 5.991465 14.56781
## 4  3010   790  85.7  60.0 12.75    251 1211034 4824.836 5.525453 14.00698
## 5  3010  1403  77.3  59.1 17.08    439 1913501 4358.772 6.084499 14.46445
## 6  3010  4056  85.2  67.0 23.17    561 2637483 4701.396 6.329721 14.78534
##     lexppp
## 1 8.677725
## 2 7.704195
## 3 8.576344
## 4 8.481532
## 5 8.379946
## 6 8.455615
```

(i) Find the largest and smallest values of math4. Does the range make sense? Explain.

Solution

```
head(meap01)
```

```
##   dcode bcode math4 read4 lunch enroll  expend    exppp  lenroll  lexpend
## 1  1010  4937  83.3  77.8 40.60    468 2747475 5870.673 6.148468 14.82619
## 2  2070   597  90.3  82.3 27.10    679 1505772 2217.632 6.520621 14.22482
## 3  2080  4860  61.9  71.4 41.75    400 2121871 5304.678 5.991465 14.56781
## 4  3010   790  85.7  60.0 12.75    251 1211034 4824.836 5.525453 14.00698
## 5  3010  1403  77.3  59.1 17.08    439 1913501 4358.772 6.084499 14.46445
## 6  3010  4056  85.2  67.0 23.17    561 2637483 4701.396 6.329721 14.78534
##     lexppp
## 1 8.677725
## 2 7.704195
## 3 8.576344
## 4 8.481532
## 5 8.379946
## 6 8.455615
```

```
summary(meap01$math4) # It makes sense as percentage is between 0 and 100
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   61.60   76.40   71.91   87.00  100.00
```

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

Solution

```
library(dplyr)
passrate100=nrow(meap01 %>% group_by(math4) %>% filter(math4=='100'))
passrate100
```

```
## [1] 38
```

```
samplesize=nrow(meap01)
samplesize
```

```
## [1] 1823
```

```
percent_passrate=round((passrate100/samplesize)*100,2)
percent_passrate
```

```
## [1] 2.08
```

(iii) How many schools have math pass rates of exactly 50%?

Solution

```
library(dplyr)

nrow(meap01 %>% group_by(math4) %>% filter(math4=='50'))
```

```
## [1] 17
```

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

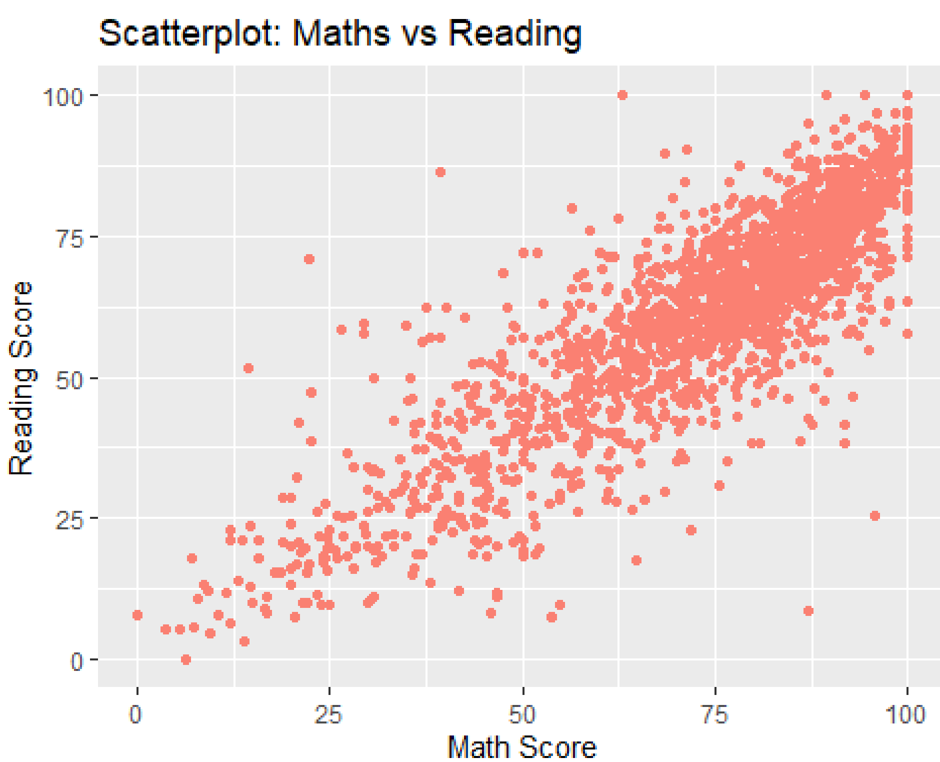
Solution

```
pass_m=mean(meap01$math4)
pass_m

## [1] 71.909

pass_r=mean(meap01$read4)
pass_r

## [1] 60.06188

#Cleary Reading is much more difficult to pass
```

(v) Find the correlation between math4 and read4. What do you conclude?

Solution

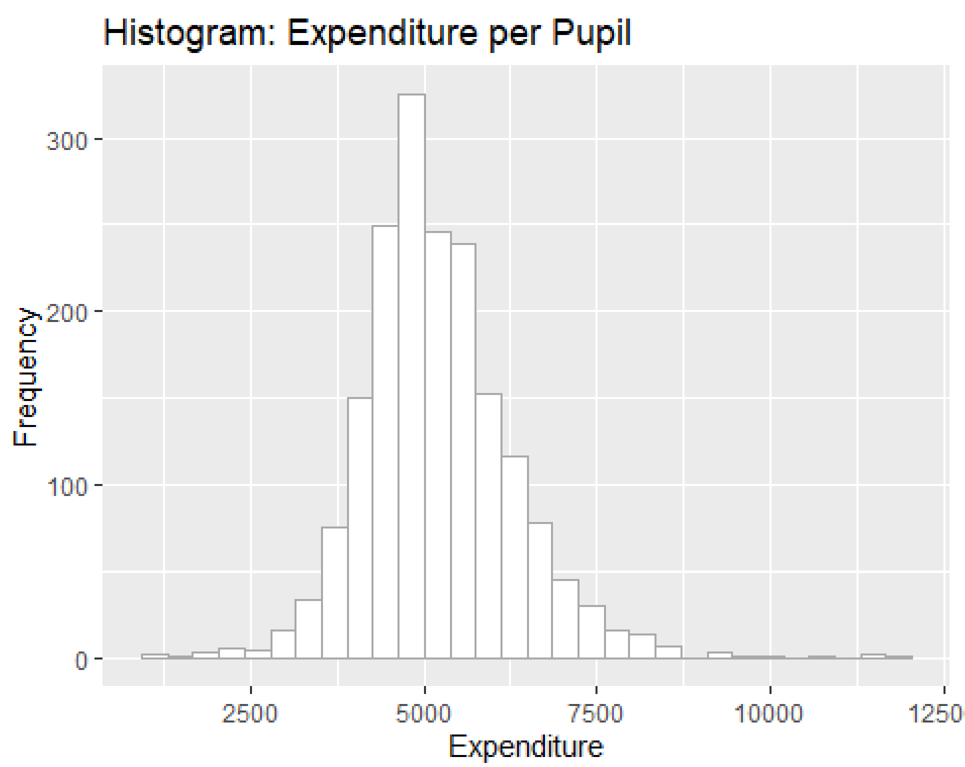
```
cor(meap01$math4,meap01$read4)

## [1] 0.8427281

library(ggplot2)
ggplot(data=meap01,aes(x=meap01$math4,y=meap01$read4))+geom_point(col='salmon
')+ggtitle("Scatterplot: Maths vs Reading")+xlab("Math Score")+ylab("Reading
Score")
```




```
#It is strongly positive
```

(vi) The variable exppp is expenditure per pupil. Find the average of exppp along with its standard deviation. Would you say there is wide variation in per pupil spending?

```
mean(meap01$exppp)

## [1] 5194.865

sd(meap01$exppp)

## [1] 1091.89

summary(meap01$exppp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1207    4502    5078    5195    5767   11958

library(ggplot2)
ggplot(data=meap01,aes(x=meap01$exppp))+geom_histogram(col='dark grey',fill='
white')+ggtitle("Histogram: Expenditure per Pupil")+xlab("Expenditure")+ylab(
"Frequency")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Considering Min=1207 and Max=11958, It is significantly wide
```

(vii) Suppose School A spends $6,000 per student and School B spends $5,500 per student. By what percentage does School A's spending exceed School B's? Compare

this to 100 · [log(6,000) – log(5,500)], which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Appendix A.)

Solution

```
round((log(6000)-log(5500))*100,2) # Gives the Percentage

## [1] 8.7

round(((6000-5500)/ 5500)*100,2)

## [1] 9.09
```

C4: The data in JTRAIN2 come from a job training experiment conducted for low-income men during 1976–1977; see Lalonde (1986).

```
head(jtrain2)

##   train age educ black hisp married nodegree mosinex re74 re75     re78 un
em74
## 1     1  37   11     1    0       1        1      13    0    0  9.93005
1
## 2     1  22    9     0    1       0        1      13    0    0  3.59589
1
## 3     1  30   12     1    0       0        0      13    0    0 24.90950
1
## 4     1  27   11     1    0       0        1      13    0    0  7.50615
1
## 5     1  33    8     1    0       0        1      13    0    0  0.28979
1
## 6     1  22    9     1    0       0        1      13    0    0  4.05649
1
##   unem75 unem78 lre74 lre75      lre78 agesq mostrn
## 1      1      0     0     0  2.295566  1369     13
## 2      1      0     0     0  1.279792   484     13
## 3      1      0     0     0  3.215249   900     13
## 4      1      0     0     0  2.015723   729     13
## 5      1      0     0     0 -1.238599  1089     13
## 6      1      0     0     0  1.400318   484     13

?jtrain2
```

(i) Use the indicator variable train to determine the fraction of men receiving job training.

Solution

```
tr=nrow(jtrain2 %>% group_by(train) %>% filter(train=='1'))
tr

## [1] 185
```

```
total=nrow(jtrain2)
total
```

```
## [1] 445
```

```
percent_trained=round((tr/total)*100,2)
percent_trained
```

```
## [1] 41.57
```

(ii) The variable re78 is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of re78 for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

Solution

```
train78=jtrain2 %>% select(re78,train) %>% group_by(re78) %>% filter(train=='
1')
train78
```

```
## # A tibble: 185 x 2
## # Groups:   re78 [141]
##       re78 train
##      <dbl> <int>
##  1  9.93      1
##  2  3.60      1
##  3 24.9       1
##  4  7.51      1
##  5  0.290     1
##  6  4.06      1
##  7  0         1
##  8  8.47      1
##  9  2.16      1
## 10 12.4       1
## # ... with 175 more rows
```

```
mean(train78$re78)
```

```
## [1] 6.349145
```

```
untrain78=jtrain2 %>% select(re78,train) %>% group_by(re78) %>% filter(train=
='0')
untrain78
```

```
## # A tibble: 260 x 2
## # Groups:   re78 [169]
##     re78 train
##    <dbl> <int>
##  1  0        0
##  2 12.4      0
##  3  0        0
##  4 10.7      0
```

```
##   5 11.8       0
##   6  9.23      0
##   7 10.6       0
##   8  6.04      0
##   9  3.88      0
## 10  0         0
## # ... with 250 more rows
```

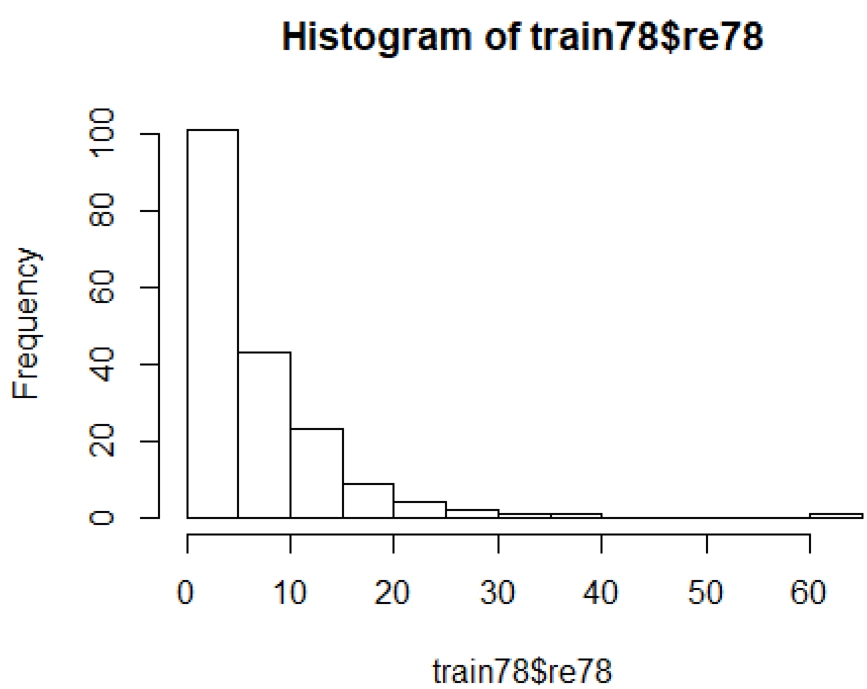
```
mean(untrain78$re78)
```

```
## [1] 4.554802
```

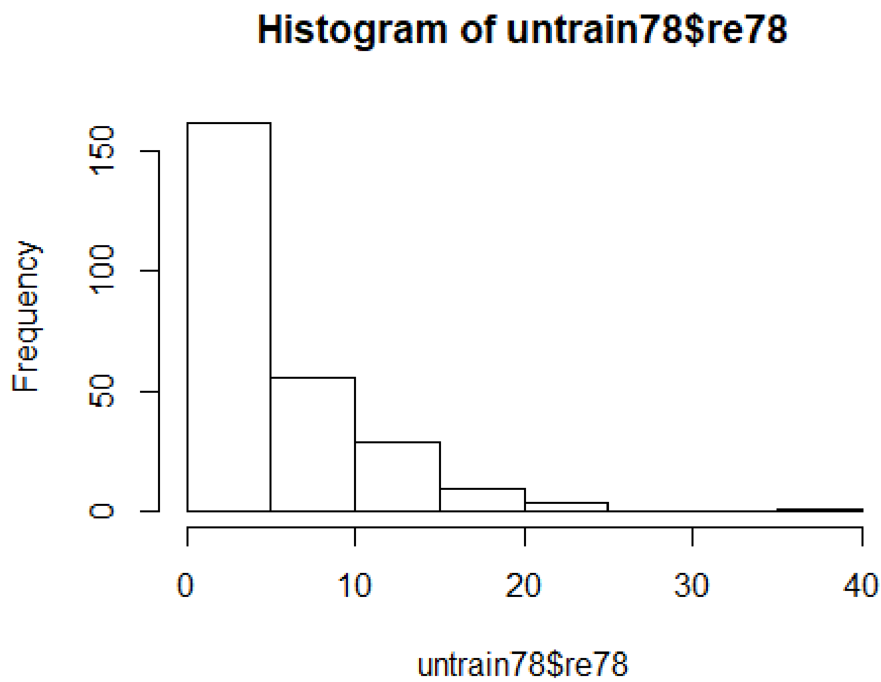
```
library(sm)
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

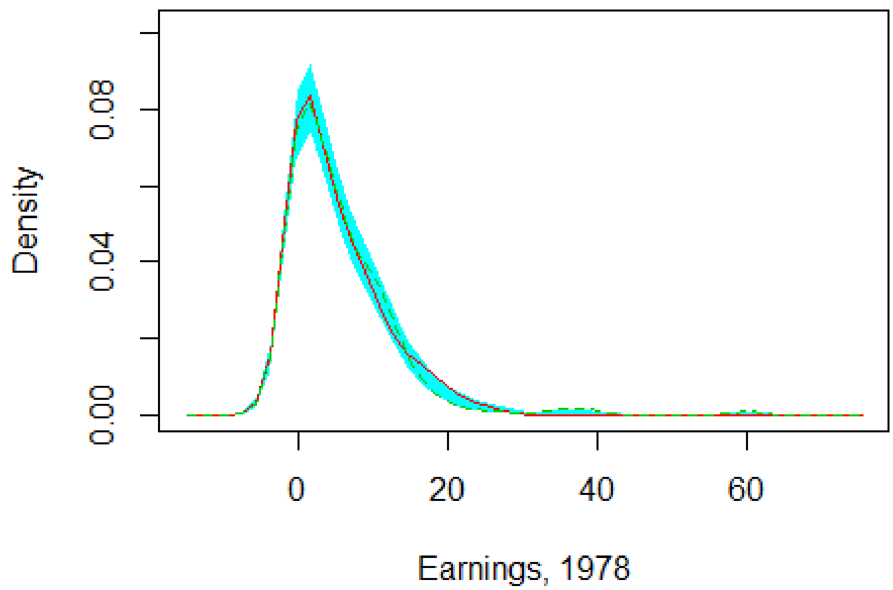
```
ab=c(train78$re78,untrain78$re78)
hist(train78$re78)
```




```
hist(untrain78$re78)
```




```
g=rep(1:2, rep(100,2))
sm.density.compare(ab,g, xlab="Earnings, 1978",model="equal") # plot densitie
s

## Warning in cbind(X, group): number of rows of result is not a multiple of
vector
## length (arg 2)

##
## Test of equal densities:  p-value =  0.95
```




```
#Mean for Trained worker is 6.35 and Untrained is 4.55. Significant Differenc
e
```

(iii) The variable unem78 is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.

Solution

```
train_uemp=jtrain2 %>% select(unem78,train) %>% group_by(train) %>% filter(tr
ain=='1' & unem78=='1')
tu=nrow(train_uemp)
tu

## [1] 45

tot_train=nrow(jtrain2 %>% select(unem78,train) %>% group_by(train) %>% filte
r(train=='1'))
tot_train

## [1] 185

round((tu/tot_train)*100,2) #Percentage Unemployed with Training

## [1] 24.32

untrain_unemp=jtrain2 %>% select(unem78,train) %>% group_by(train) %>% filter
(train=='0' & unem78=='1')
```

```
utue=nrow(untrain_unemp)
utue

## [1] 92

tot_untrain=nrow(jtrain2 %>% select(unem78,train) %>% group_by(train) %>% fil
ter(train=='0'))

round((utue/tot_untrain)*100,2) #Percentage Unemployed without Training

## [1] 35.38

#Unemployment figure among untrained individuals is much larger than trained
```

(iv) From parts (ii) and (iii), does it appear that the job training program was effective? What would make our conclusions more convincing?

Solution The differences in earnings and unemployment rates suggest the training program had strong, positive effects. Our conclusions about economic significance would be stronger if we could also establish statistical significance

End of C4

C5 The data in FERTIL2 were collected on women living in the Republic of Botswana in 1988. The variable children refers to the number of living children. The variable electric is a binary indicator equal to one if the woman's home has electricity, and zero if not.

```
head(fertil2)

##   mnthborn yearborn age electric radio tv bicycle educ ceb agefbrth childr
en
## 1        5       64  24        1     1  1       1   12   0        NA      
0
## 2        1       56  32        1     1  1       1   13   3        25      
3
## 3        7       58  30        1     0  0       0    5   1        27      
1
## 4       11       45  42        1     0  1       0    4   3        17      
2
## 5        5       45  43        1     1  1       1   11   2        24      
2
## 6        8       52  36        1     0  0       0    7   1        26      
1
##   knowmeth usemeth monthfm yearfm agefm idlnchld heduc agesq urban urb_edu
c
## 1        1       0      NA     NA    NA        2    NA   576     1       1
2
## 2        1       1      11     80    24        3    12  1024     1       1
3
## 3        1       0       6     83    24        5     7   900     1
```

```
5
## 4         1         0         1      61     15          3     11   1764       1
4
## 5         1         1         3      66     20          2     14   1849       1         1
1
## 6         1         1        11      76     24          4      9   1296       1
7
##   spirit protest catholic frsthalf educ0 evermarr
## 1      0       0        0        1     0        0
## 2      0       0        0        1     0        1
## 3      1       0        0        0     0        1
## 4      0       0        0        0     0        1
## 5      0       1        0        1     0        1
## 6      0       0        0        0     0        1

?fertil2
```

(i) Find the smallest and largest values of children in the sample. What is the average of children?

Solution

```
summary(fertil2$children)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   2.268   4.000  13.000
```

(ii) What percentage of women have electricity in the home?

Solution

```
library(dplyr)
tot_women=nrow(fertil2)
tot_women

## [1] 4361

elec_women=nrow(fertil2 %>% filter(electric=="1"))
elec_women

## [1] 611

perc_with_elect=round((elec_women/tot_women)*100,2)
perc_with_elect

## [1] 14.01
```

(iii) Compute the average of children for those without electricity and do the same for those with electricity. Comment on what you find.

Solution

```
library(dplyr)
elect_child=fertil2 %>% select(children,electric) %>% filter(electric==1)

avg_chil_with_elec=mean(elect_child$children)

avg_chil_with_elec #Average Children in Electricy households

## [1] 1.898527

non_elect_child=fertil2 %>% select(children,electric) %>% filter(electric==0)

avg_chil_with_no_elec=mean(non_elect_child$children)

avg_chil_with_no_elec #Average Children in Non Electricy households

## [1] 2.327729

#Clearly the electricity households have fewer children
```

(iv) From part (iii), can you infer that having electricity “causes” women to have fewer children? Explain.

Solution Not directly byt may be influency of electricity devices like TV and radio are more likely to promote fewer children and more liberal thoughts.

C6 Use the data in COUNTYMURDERS to answer this question. Use only the year 1996. The variable murders is the number of murders reported in the county. The variable execs is the number of executions that took place of people sentenced to death in the given county. Most states in the United States have the death penalty, but several do not.

```
head(countymurders)

##   arrests countyid density popul perc1019 perc2029 percblack percmale
## 1       2     1001   54.05 32216    20.63    15.28     22.33    40.25
## 2       3     1001   53.66 31984    20.19    15.55     22.07    40.36
## 3       2     1001   53.75 32036    19.66    15.73     21.80    40.42
## 4       7     1001   53.78 32056    19.10    15.88     21.53    40.47
## 5       3     1001   53.91 32128    18.54    15.92     21.26    40.51
## 6       1     1001   54.11 32248    18.06    15.87     20.96    40.45
##   rpcincmaint rpcpersinc rpcunemins year murders  murdrate arrestrate stat
efips
## 1      167.67    8780.80      29.16 1980       2 0.6208096  0.6208095
1
## 2      167.99    8232.80      43.92 1981       1 0.3126563  0.9379690
1
## 3      166.63    8327.61      71.41 1982       3 0.9364465  0.6242977
1
## 4      176.53    8545.55      72.22 1983       7 2.1836790  2.1836790
1
## 5      166.25    8965.16      40.36 1984       2 0.6225100  0.9337650
```

```
1
## 6      153.12    9254.02      44.54 1985       2 0.6201935  0.3100968
1
##   countyfips execs   lpopul execrate
## 1          1     0 10.38022        0
## 2          1     0 10.37299        0
## 3          1     0 10.37462        0
## 4          1     0 10.37524        0
## 5          1     0 10.37748        0
## 6          1     0 10.38121        0

?countymurders
#This is a Panel dataset
```

(i) How many counties are there in the data set? Of these, how many have zero murders? What percentage of counties have zero executions? (Remember, use only the 1996 data.)

Solution

```
library(dplyr)
county96=countymurders %>% select(countyid,year,murdrate,execs) %>% filter(ye
ar=="1996")
nrow(county96)

## [1] 2197

county96_0murd=countymurders %>% select(countyid,year,murdrate,execs) %>% fil
ter(year=="1996" & murdrate=="0")
nrow(county96_0murd) #Counties with Zero murders

## [1] 1051

county96_0exec=countymurders %>% select(countyid,year,murdrate,execs) %>% fil
ter(year=="1996" & execs=="0")
nrow(county96_0exec) #Counties with Zero executions

## [1] 2166

perc_county_0exe=round(nrow(county96_0exec)/nrow(county96),2)*100
perc_county_0exe #Percentage of Counties with zero exexutions

## [1] 99
```

(ii) What is the largest number of murders? What is the largest number of executions? Why is the average number of executions so small?

Solution

```
library(dplyr)
county96 %>% filter(murdrate==max(county96$murdrate))
```

```
##   countyid year murdrate execs
## 1    48033 1996 12.39157     0
```

```r
max(county96$murdrate) #Maximum Murder rate
```

```
## [1] 12.39157
```

```r
max(county96$execs) #Maximum Executions
```

```
## [1] 3
```

```r
mean(county96$execs) #The mean of execution is so small because many counties
do not practice death penalty and many have no executions
```

```
## [1] 0.01593081
```

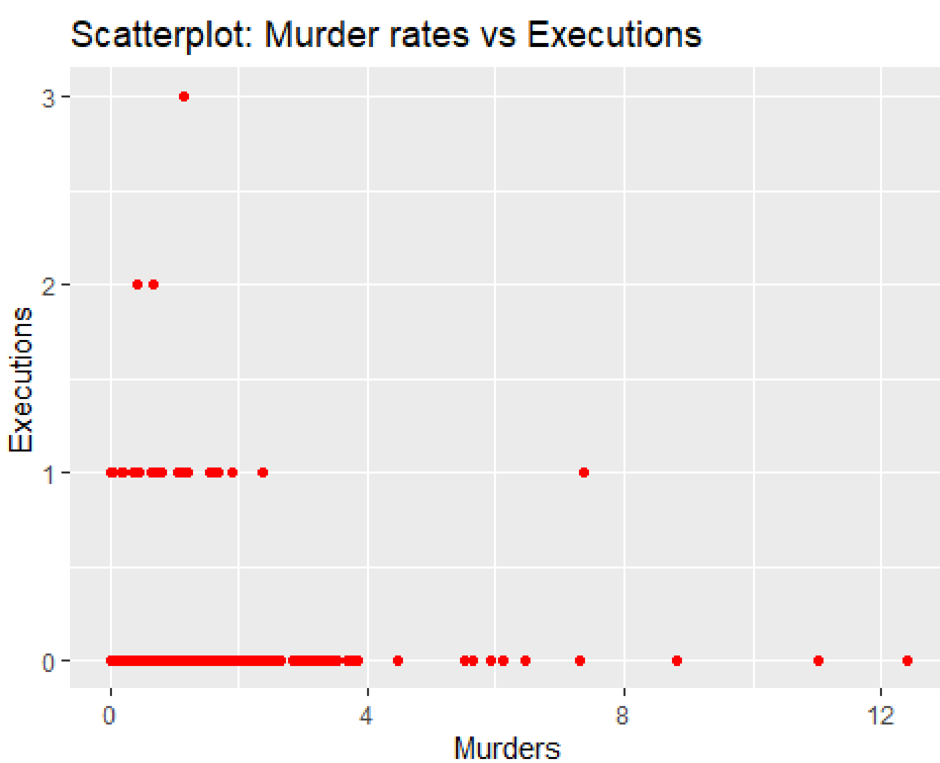
(iii) Compute the correlation coefficient between murders and execs and describe what you find.

Solution

```r
cor(county96$murdrate,county96$execs)
```

```
## [1] 0.08567281
```

```r
library(ggplot2)
ggplot(data=county96,aes(x=murdrate,y=execs))+geom_point(col='red')+ggtitle("
Scatterplot: Murder rates vs Executions")+xlab("Murders")+ylab("Executions")
```



(iv) You should have computed a positive correlation in part (iii). Do you think that more executions cause more murders to occur? What might explain the positive correlation?

Solution It is a very weak positive correlation. We would expect the two to have negative relationship. Here the data is influenced by the fact that many states have outlawed death penalty

End of C6

C7 The data set in ALCOHOL contains information on a sample of men in the United States. Two key variables are self-reported employment status and alcohol abuse (along with many other variables). The variables employ and abuse are both binary, or indicator, variables: they take on only the values zero and one.

```
head(alcohol)

##   abuse status unemrate age educ married famsize white exhealth vghealth
## 1     1      1      4.0  50    4       1       1     1        0        0
## 2     0      3      4.0  37   12       1       5     1        0        0
## 3     0      3      4.0  53    9       1       3     1        1        0
## 4     0      3      3.3  59   11       1       1     1        1        0
## 5     0      3      3.3  43   10       1       1     1        1        0
## 6     0      3      3.3  38   10       1       1     1        1        0
##   goodhealth fairhealth northeast midwest south centcity outercity qrt1 qr
t2
## 1          0          0         0       1     0        0         0    1
0
## 2          1          0         0       1     0        0         0    1
0
## 3          0          0         0       1     0        0         0    1
0
## 4          0          0         1       0     0        1         0    1
0
## 5          0          0         1       0     0        1         0    1
0
## 6          0          0         1       0     0        1         0    1
0
##   qrt3 beertax cigtax ethanol mothalc fathalc livealc inwf employ agesq
## 1    0   0.334     38 2.03946       0       0       0    0      0  2500
## 2    0   0.334     38 2.03946       0       0       0    1      1  1369
## 3    0   0.334     38 2.03946       0       0       0    1      1  2809
## 4    0   0.240     26 2.44998       0       0       0    1      1  3481
## 5    0   0.240     26 2.44998       0       1       1    1      1  1849
## 6    0   0.240     26 2.44998       0       0       0    1      1  1444
##   beertaxsq cigtaxsq ethanolsq educsq
## 1  0.111556     1444  4.159397     16
## 2  0.111556     1444  4.159397    144
## 3  0.111556     1444  4.159397     81
## 4  0.057600      676  6.002402    121
```

```
## 5  0.057600       676  6.002402    100
## 6  0.057600       676  6.002402    100

?alcohol
#Cross sectional data set
```

(i) What is percentage of the men in the sample report abusing alcohol? What is the employment rate?

Solution

```
ab=nrow(alcohol %>% filter(abuse=="1"))
tot=nrow(alcohol)

p_ab=(round((ab/tot)*100,2))
p_ab #Percentage of men reporting alcohol abuse

## [1] 9.92

employed=nrow(alcohol %>% filter(status=="3"))
employed

## [1] 8822

employment_rate=(round((employed/tot)*100,2))
employment_rate #Employment rate in the sample

## [1] 89.82
```

(ii) Consider the group of men who abuse alcohol. What is the employment rate?

Solution

```
ab_tot=nrow(alcohol %>% filter(abuse=="1"))
ab_tot

## [1] 974

ab_employed=nrow(alcohol %>% filter(abuse=='1' & status=="3"))
ab_employed

## [1] 850

ab_employment_rate=(round((ab_employed/ab_tot)*100,2))
ab_employment_rate #Employment rate among alcohol abuses in the sample

## [1] 87.27
```

(iii) What is the employment rate for the group of men who do not abuse alcohol?

```
no_ab_tot=nrow(alcohol %>% filter(abuse=="0"))
no_ab_tot

## [1] 8848
```

```r
no_ab_employed=nrow(alcohol %>% filter(abuse=='0' & status=="3"))
no_ab_employed

## [1] 7972

no_ab_employment_rate=(round((no_ab_employed/no_ab_tot)*100,2))
no_ab_employment_rate #Employment rate among alcohol abuses in the sample

## [1] 90.1
```

(iv) Discuss the difference in your answers to parts (ii) and (iii). Does this allow you to conclude that alcohol abuse causes unemployment?

Solution There is hardly a difference of around 3%. Not much significant to make any conclusion with reasonabale certainty.