Statistical Methods for Engineers

Laboratory activities

Mikel Molina

2023-02-18

1st PART

Consider the following two games.

- **In Game A you flip a fair coin. If the coin comes up Heads you get two dollars, whereas if it comes up Tails you get one dollar.**
- **In Game B you roll a fair die. If the six-spot comes up, you win twenty-five dollars. If you get 2, 3, 4, or 5, nothing happens. If the one-spot comes up, you lose fifteen dollars.**

In this exercise you need to answer two questions:

1. **If you could choose to play either Game A or Game B *just once*, which game would you prefer to play, and why? (Use only your intuition)**

   I would rather play game A, since in game B, as opposed to winning it, there would be bigger chances to not win anything at all or to even lose money.

2. **Suppose that you can choose either to play Game A ten thousand times, or Game B ten thousand times. Which choice would you prefer, and why? (Use only your intuition)**

   This time I would choose Game B, because the reward is bigger and I think I would end up with more money rather than with less.

Answers to the first question will vary from person to person, depending on circumstances and personal taste. On the other hand, astute consideration of expected values leads most people to answer the second question in the same way.

1st activity

Which random variable we have to define for representing the first game? And the second game? Find the probability mass functions of both random variables and plot them.

Let's start by getting the probability mass function (p.m.f) of the first game. We'll define x as:

$$X \equiv \text{"Money won in game A"} = \{1, 2\}$$

And then we find the values of the p.m.f for each value:

$$P(X = 2) = P(heads) = 0.5$$
$$P(X = 1) = P(tail) = 0.5$$

Furthermore, we know that the previous variable X is a discrete random variable and that the possible values of X are consecutive integers from 1 to 2, and that each value of X has an equal probability. So we have that:

a = 1 and b = 2, then $f(x) = \frac{1}{b-a+1} \Rightarrow f(x) = \frac{1}{2}$, for $1 \leq x \leq 2$ (For discrete numbers, which means that it's just value

This is the p.m.f for the first game. Now for the game B we will do pretty much the same, with some little differences. This time we define X as:

$$X \equiv \{Money\ won\ in\ game\ B\} = \{-15, 0, 25\}$$

And then we find the values of the p.m.f for each value:

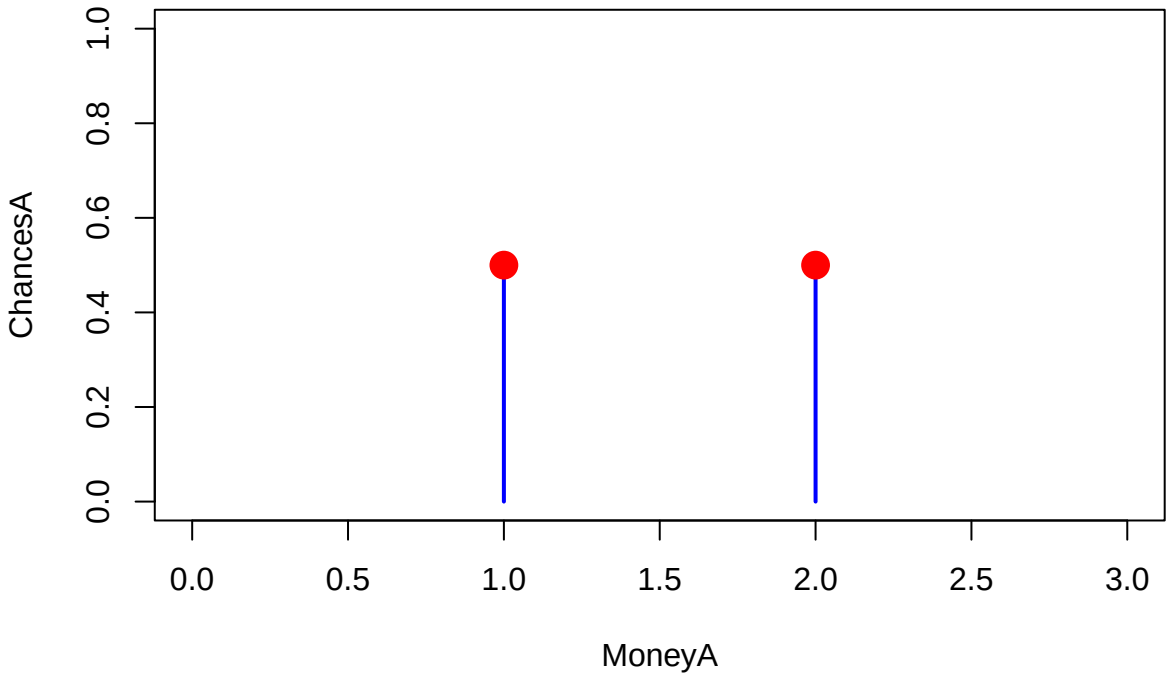
$$P(X = 25) = P(6) = \frac{1}{6}$$
$$P(X = -15) = P(1) = \frac{1}{6}$$
$$P(X = 0) = P(2) + P(3) + P(4) + P(5) = \frac{4}{6}$$

This time though, we won't be able to use the same "formula" for the function we used before, so we will define the p.m.f like so:

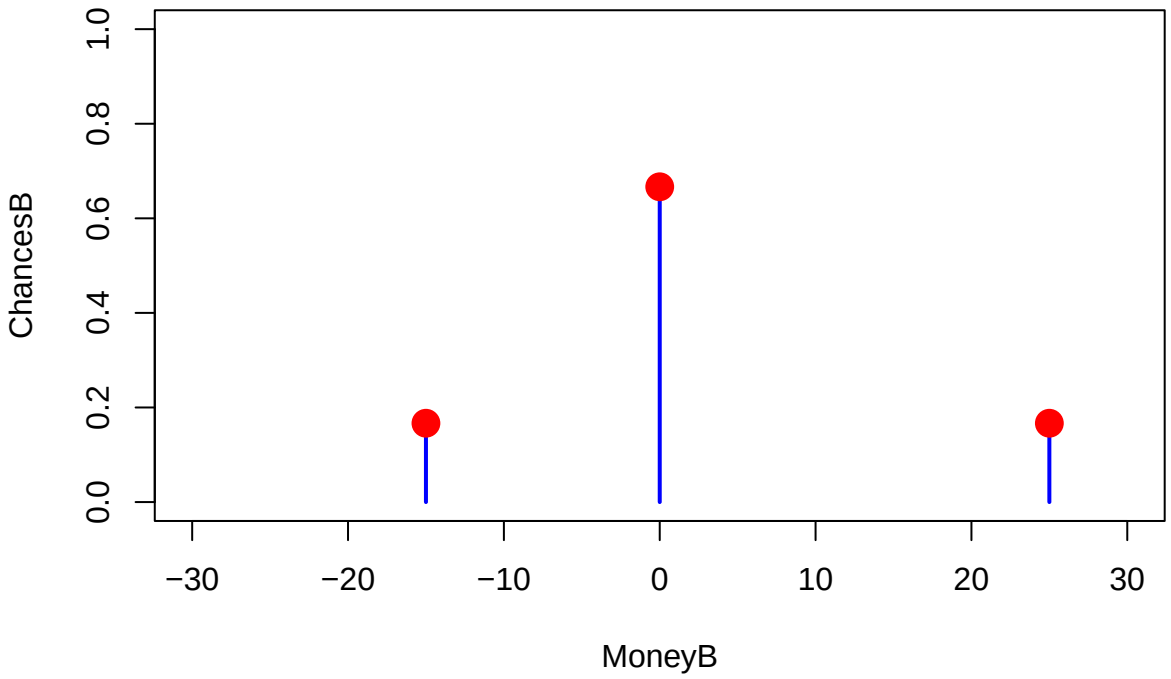
$$f(x) = P(X = x) = \begin{cases} \frac{1}{6}, & x = 25 \\ \frac{1}{6}, & x = -15 \\ \frac{4}{6}, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

Now that we have the p.m.f of both games, let's plot them.

```
# Plotting function A
 MoneyA <- c(1,2) #Money that can be won in game A.
 ChancesA <- c(1/2,1/2) #Chances to get the money in game  A, respectively.
 plot(MoneyA, ChancesA, type = "h", xlim = c(0,3), ylim=c(0,1), lwd = 2, col ="blue", ylab= "ChancesA"
 points(MoneyA,ChancesA,pch=16,cex=2,col="red")
```




```r
#Plotting function B
 MoneyB <- c(-15,0,25) #Money that can be won or lost in game B.
 ChancesB <- c(1/6,4/6,1/6) #Chances to get the money in game B, respectively.
 plot(MoneyB, ChancesB, type = "h", xlim = c(-30,30), ylim = c(0,1), lwd = 2, col = "blue", ylab = "Cha
 points(MoneyB, ChancesB, pch = 16, cex = 2, col = "red")
```



2nd activity

Find the expected values and the variances of previously defined random variables. Which are the meaning of these values?

We can compute the expected value and variance of the game A using the formulas from the Discrete Uniform Distribution, since we are using the same function:

Game A

$$\text{Expected value: } \mu = E[X] = \frac{a+b}{2} = \frac{3}{2} = 1.5$$

$$\text{Variance: } \sigma^2 = E[(X-\mu)^2] = E[X^2] - (E[X])^2 = \frac{(b-a+2)(b-a)}{12} = \frac{3}{12} = \frac{1}{4} = 0.25$$

As for game B, we cannot use the same formulas used for game A, so we will have to compute it ourselves:

Game B

$$\text{Expected value: } \mu = E[X] = \sum_x x \cdot f(x) = -15 \cdot \frac{1}{6} + 0 \cdot \frac{4}{6} + 25 \cdot \frac{1}{6} = \frac{10}{6} \approx 1.66$$

$$\text{Variance: } \sigma^2 = E[(X-\mu)^2] = E[X^2] - (E[X])^2 \overset{(1)}{=} \frac{850}{6} - \frac{10^2}{6^2} = \frac{1250}{9} \approx 138.88$$

$$(1): E[X^2] = \sum_x x^2 \cdot f(x) = (-15)^2 \cdot \frac{1}{6} + 0^2 \cdot \frac{4}{6} + 25^2 \cdot \frac{1}{6} = \frac{850}{6}$$

Now, let's compute this values to check whether we got them right.

```
ExpectedA <- sum(MoneyA*ChancesA) #Expected value of the game A.
ExpectedASquared <- ExpectedA^2 #Expected value of the game A squared, representing (E[X])^2.
MoneyASquared <- MoneyA^2 #X^2 in the argument of E[X^2].
VarianceA <- sum(MoneyASquared * ChancesA) - ExpectedASquared #Variance of game A.
ExpectedB <-sum(MoneyB*ChancesB) #Expected value of the game B.
xminusmusquared <- (MoneyB - ExpectedB)^2
VarianceB <- sum(xminusmusquared*ChancesB) # Variance of game B
sprintf("Expected value of game A: %.1f", ExpectedA)
```

```
## [1] "Expected value of game A: 1.5"
```

```
sprintf("Variance of game A: %.2f", VarianceA)
```

```
## [1] "Variance of game A: 0.25"
```

```
sprintf("Expected value of game B: %.2f", ExpectedB)
```

```
## [1] "Expected value of game B: 1.67"
```

```
sprintf("Variance of game B: %.3f", VarianceB)
```

```
## [1] "Variance of game B: 138.889"
```

All the values are according to what we calculated.

3rd activity

Simulate the results from both games when the variable n is the number of times we play (check the `sample()` function help). Repeat the simulation for each value of n five times and fill the tables:

```
#Sample with n = 100
GameASample <- sample(x = MoneyA, size = 100, replace = TRUE, prob = ChancesA)
sprintf("Mean with n = 100 for game A: %.2f", mean(GameASample))
```

```
## [1] "Mean with n = 100 for game A: 1.52"
```

```
GameBSample <- sample(x = MoneyB, size = 100, replace = TRUE, prob = ChancesB)
sprintf("Mean with n = 100 for game B: %.2f", mean(GameBSample))
```

```
## [1] "Mean with n = 100 for game B: 1.40"
```

```
#Sample with n = 1000
GameASample <- sample(x = MoneyA, size = 1000, replace = TRUE, prob = ChancesA)
sprintf("Mean with n = 1000 for game A: %.3f", mean(GameASample))
```

```
## [1] "Mean with n = 1000 for game A: 1.502"
```

```
GameBSample <- sample(x = MoneyB, size = 1000, replace = TRUE, prob = ChancesB)
sprintf("Mean with n = 1000 for game B: %.3f", mean(GameBSample))
```

```
## [1] "Mean with n = 1000 for game B: 2.165"
```

```
#Sample with n = 10000
GameASample <- sample(x = MoneyA, size = 10000, replace = TRUE, prob = ChancesA)
sprintf("Mean with n = 10000 for game A: %.4f", mean(GameASample))
```

```
## [1] "Mean with n = 10000 for game A: 1.5045"
```

```
GameBSample <- sample(x = MoneyB, size = 10000, replace = TRUE, prob = ChancesB)
sumB <- sum(GameBSample)
sprintf("Mean with n = 10000 for game B: %.4f", mean(GameBSample))
```

```
## [1] "Mean with n = 10000 for game B: 1.7455"
```

- $n = 100$

Repeat	**1**	**2**	**3**	**4**	**5**
$\overline{x}$	Game A: 1.45 Game B: 2.15	Game A: 1.52 Game B: 4.75	Game A: 1.56 Game B: 2.85	Game A: 1.53 Game B: 3.2	Game A: 1.47 Game B: 2.2

- $n = 1,000$

Repeat	**1**	**2**	**3**	**4**	**5**
$\overline{x}$	Game A: 1.521 Game B: 1.325	Game A: 1.481 Game B: 1.405	Game A: 1.508 Game B: 1.515	Game A: 1.528 Game B: 1.015	Game A: 1.462 Game B: 1.395

- $n = 10,000$

Repeat	**1**	**2**	**3**	**4**	**5**
$\overline{x}$	Game A: 1.5065 Game B: 1.8115	Game A: 1.5134 Game B: 1.6505	Game A: 1.5007 Game B: 1.6685	Game A: 1.4932 Game B: 1.8385	Game A: 1.4948 Game B: 1.5485

4th activity

Which is the probability of winning $5,000$ dollars or more in with the first game if we play it $10,000$ times? And with the second game?

Game A

In game A it is guaranteed that we will win more than 5000 if we were to play the game 10000 times. Even in the worst-case scenario, where we would only get one dollar for every throw in all the 10000 trials, we would still be getting 10000 $, which would be the double.

Game B

As for game B, things get a little bit more complicated; in this game we can lose 15 dollars. Furthermore, the outcome with the highest chances of happening for each trial is getting 0 $, so there is the possibility of not getting our desired 5000 dollars, or more. However, let us compute how likely it is for us to get more than the desired amount:

We want to find the probability that the sum of the 10000 values is more than 5000. Since the sample size is large enough ($10000 > 30$), our distribution will follow that of the Normal Distribution, we know this due to the Central Limit Theorem (C.L.T.). We already have everything necessary to calculate what we want:

In the previous exercises we calculated the mean and the variance of Game B:

$$\begin{cases} \mu = E[X_i] = \frac{10}{6} \\ \sigma^2 = VAR[X_i] = \frac{1250}{9} \\ n = 10000 > 30, \text{We can use the Central Limit Theorem} \end{cases}$$

$X \equiv$ "Money won in game B"

$$S_{10000} = X_1 + X_2 + \ ... \ + X_{10000} \overset{C.L.T.}{\sim} N(n\mu, \sqrt{n}\sigma) = N\left(10000 \cdot \frac{10}{6}, \sqrt{10000} \cdot \sqrt{\frac{1250}{9}}\right)$$

$$P(S_{10000} > 5000) \overset{(1)}{=} P\left(\frac{S_{10000} - n\mu}{\sqrt{n}\sigma} > \frac{5000 - n\mu}{\sqrt{n}\sigma}\right) = P\left(\frac{S_{10000} - 10000 \cdot \frac{10}{6}}{\sqrt{10000} \cdot \sqrt{\frac{1250}{9}}} > \frac{5000 - 10000 \cdot \frac{10}{6}}{\sqrt{10000} \cdot \sqrt{\frac{1250}{9}}}\right) =$$

$$= P(Z > -9.899) \overset{(2)}{=} 1 - P(Z \leq -9.899) \approx 1$$

$$(1) : Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0, 1)$$

$(2) : P(Z \leq -9.899)$ Is such a small value that we might as well approximate it to zero.

So the chances of getting more than 5000 dollars in game B are, approximately, 100 %. We are guaranteed to win money in game B as well, with absolutely no risks.

Now we are going to compute to check whether we got it right:

```
Z_ScoreB <- ((5000 - 10000*(10/6))/sqrt(10000*(1250/9))) # We make use of the z-score.
Pb <- pnorm(Z_ScoreB, lower.tail = FALSE) #Chances of winning more than 5000 $ in game B.
sprintf("P(Sb > 5000) = %f. As we can see, theoretically, it would be impossible to win less than 5000 d
```

```
## [1] "P(Sb > 5000) = 1.000000. As we can see, theoretically, it would be impossible to win less than 5
```

2nd PART

The adult data set is a famous dataset from the UCI - machine learning repository.

The idea is to predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset. Extraction was done by Barry Becker from the 1994 Census database.

Read a little bit about how this dataset is built. Which variables made up the dataset and what are their meanings? What is the character of each variable (qualitative, quantitative, ...). You can also use the `summary()` function to get some information.

In this dataset there are in total 14 variables, those are, respectively from left to right:

Age, quantitative continuous variable. The workclass, qualitative variable with 8 different categories. Fnlwgt, or final weight, the number of total people with the same features as that person, quantitative discrete variable. Education , qualitative variable with 16 categories. Education number, quantitative discrete variable. Marital-status, qualitative variable with 7 different categories. Occupation, qualitative variable with 14 categories. Relationship, qualitative variable with 6 different categories. Race, qualitative variable with 5 categories. Sex, qualitative variable with 2 categories. Capital-gain, the amount of money won by selling an asset that has increased its value by short-term or long-term, quantitative discrete variable. Capital-loss, the amount of money lost by selling an asset that has decreased its value by short-term or long-term, quantitative discrete variable. Hours per week, the amount of hours a person works per week, quantitative continuous variable. Native country, where the person has been born.

```
myData <- read.csv("adult.csv")
str(myData)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native.country: chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ year.income   : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

```
ages.data <- myData$age #All the ages of the dataset stored in the array list.
summary(myData)
```

```
##       age         workclass            fnlwgt         education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education.num   marital.status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex            capital.gain    capital.loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
```

```
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country     year.income
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

1st activity

Choose a variable from your data set and describe it in depth. You should use graphs and statistics for description. To do this, work on the following sections:

We are going to choose the age variable.

- **Which type of variable is it?**

  A quantitative discrete variable, the age.

- **How many individuals are in the sample?**

```
ages.sorted <- sort(myData$age)
ages.elements <- length(ages.sorted)
cat("There are" ,ages.elements,"observations. \n")
```

```
## There are 32561 observations.
```

- **Which graph do you think will show properly the values that the variable takes?**

  A histogram or a barplot would be the best way to show the values the variable takes.

- **Compute central tendency statistics and interpret them.**

  $$\text{Mean} = \frac{1}{N}\sum_{i=1}^{32561} x_i = 38.58165$$

  With this, we gather that most of the people in the data are around 38 years old. Then again, the mean is sensitive to extreme values, so depending on the amount of people that are a lot of years old or very young it may be possible tha the mean doesn't fully represent the age. Let us compute the median.

  Since the number of elements is odd: $\text{Median} = \frac{N+1}{2} = 37$

  As we can see, the mean and the median are pretty close, so the data should have a fairly symmetrical distribution. Lastly, as calculated before, the mode is 36, once again, pretty close to the mean and the median, so the graph will ha neither negative nor positive skewness.

```
ages.mean <- mean(ages.data) #Mean of the ages.
cat("The mean of the data is",ages.mean,".\n")
```

```
## The mean of the data is 38.58165 .
```

```r
ages.median <- median(ages.data) #Median of ages
median.p <- ((length(ages.data)+1)/2) # position of the median
cat("The median of the data is",ages.median,". And its position is:", median.p,"\n")
```

```
## The median of the data is 37 . And its position is: 16281
```

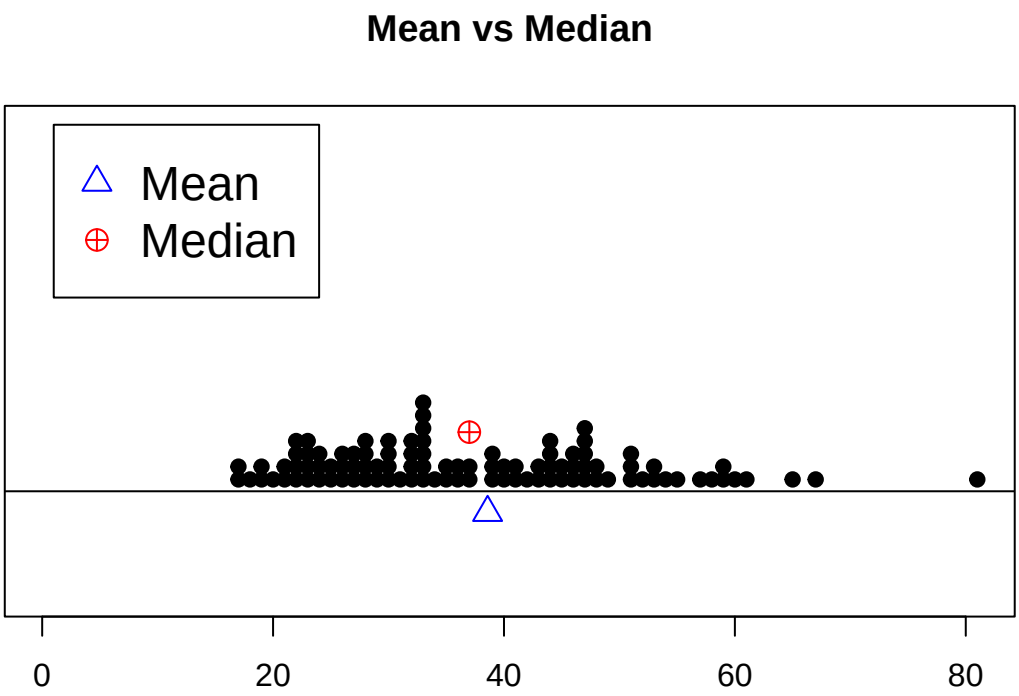
```r
Table <- table(ages.data); Table
```

```
## ages.data
##  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
## 395 550 712 753 720 765 877 798 841 785 835 867 813 861 888 828 875 886 876 898
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
## 858 827 816 794 808 780 770 724 734 737 708 543 577 602 595 478 464 415 419 366
##  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
## 358 366 355 312 300 258 230 208 178 150 151 120 108  89  72  67  64  51  45  46
##  77  78  79  80  81  82  83  84  85  86  87  88  90
##  29  23  22  22  20  12   6  10   3   1   1   3  43
```

```r
cat("The value of the mode is 36, is in the same position and appears 898 times.")
```

```
## The value of the mode is 36, is in the same position and appears 898 times.
```

```r
legend <- c("Mean", "Median")
Ages.sample <- sample(x = ages.sorted, size = 100) #Taking a sample so as to avoid the plot from being
stripchart(Ages.sample, method = "stack",
           xlim = c(0,max(Ages.sample)),
           axes = TRUE, main = "Mean vs Median",
           pch = 19,at = 0.5)
abline(h=0.45)
points(ages.mean, 0.36, col = "blue",pch = 2, cex = 1.5) # Mean of ages.
points(ages.median, 0.7, col = "red",pch = 10, cex =1.5) # Median of ages.
legend(1,2,legend, pch = c(2,10), col = c("blue", "red"), cex = 1.5)
```



The median is slightly to the left of the median, so we may deduce that the distribution is fairly symmetrical

as we have already said before.

- **Find the maximum, minimum and quartiles.**

```
ages.max <- max(ages.data) #Maximum age in the data.
cat("Maximum of the data is:",ages.max,"\n")
```

```
## Maximum of the data is: 90
```

```
ages.min <- min(ages.data) #Minimum age in the data.
cat("Minimum of the data is:",ages.min,"\n")
```

```
## Minimum of the data is: 17
```

```
quantile(ages.data)
```

```
##   0%  25%  50%  75% 100%
##   17   28   37   48   90
```

- **What would you say about dispersion?**

```
ages.sd <- sqrt(var(ages.data)) # Standard deviation of ages.
ages.sd
```

```
## [1] 13.64043
```

```
ages.range <- ages.max - ages.min # The range of the ages.
ages.range
```

```
## [1] 73
```

The standard deviation, as computed above, is 13.64. This means that most values fall around 13.64 from the mean.

- **Is it fine to use Is it okay to use Chebyshev's inequality with the distribution of this variable? Why? If it is possible, use it and explain the result you obtain.**

  Yes, as Chebyshev's rule applies to every and any data distribution. Let us calculate the intervals with the values k = 2 and 3.

  According to Chebysev's rule, the proportion of observations within k standard deviations of the mean is:

  $f(|x_i - \bar{x}| \leq ks) \geq 1 - \frac{1}{k^2}, k > 1,$ i.e.,

  With k = 2: $1 - \frac{1}{2^2} = 0.75$

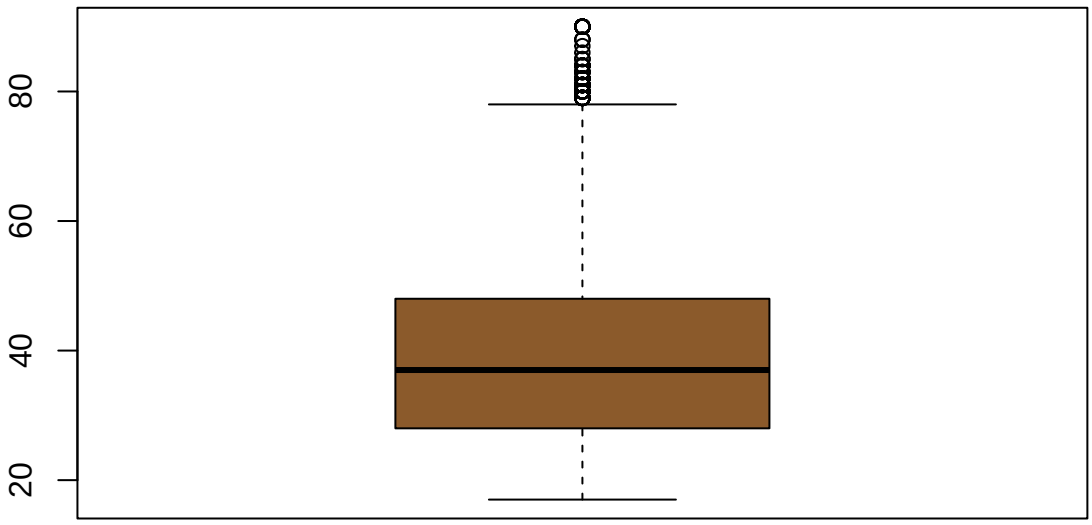
  With k = 3: $1 - \frac{1}{3^2} = 0.89$

  The proportion of observations that is between 2 standard deviations of the mean, $(\bar{x} - 2s, \bar{x} + 2s) = (11.3, 65.86)$ is e

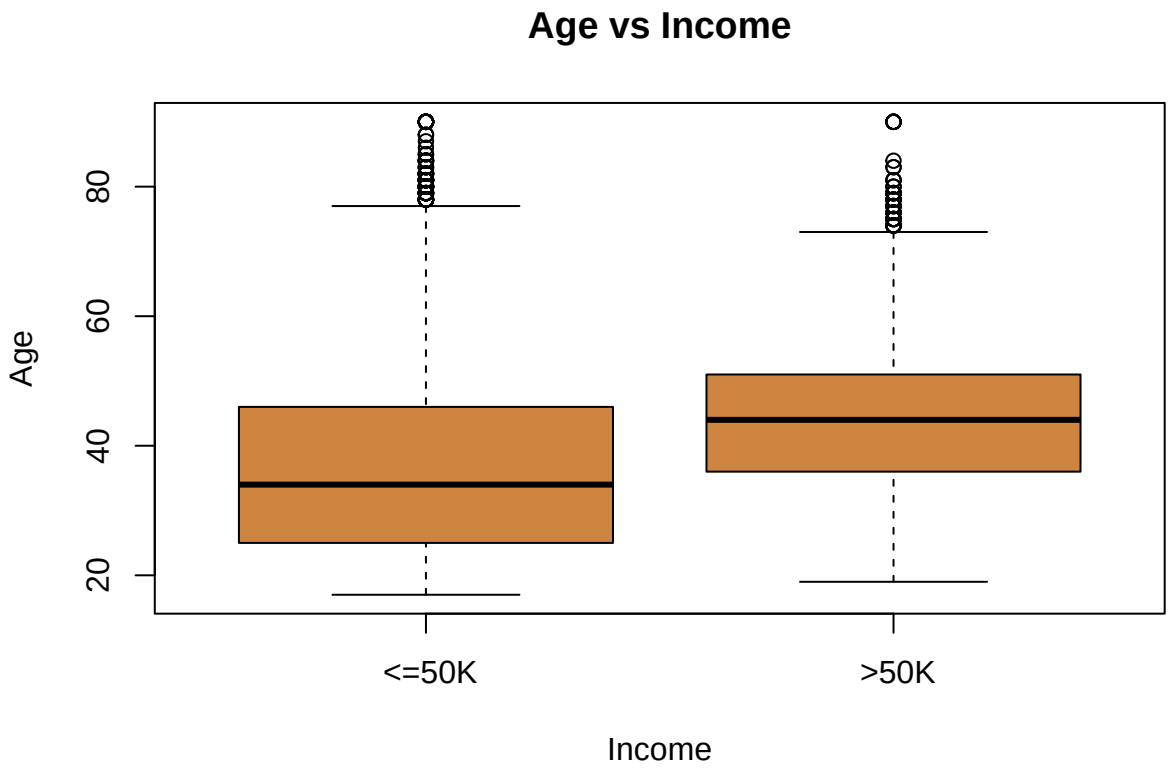
  The proportion of observations that is between 2 standard deviations of the mean, $(\bar{x} - 3s, \bar{x} + 3s) = (-2.34, 79.5)$ is e

  To finish, let's draw a boxplot to visualize the values we've calculated, also, let's do some graphics of the age variable.

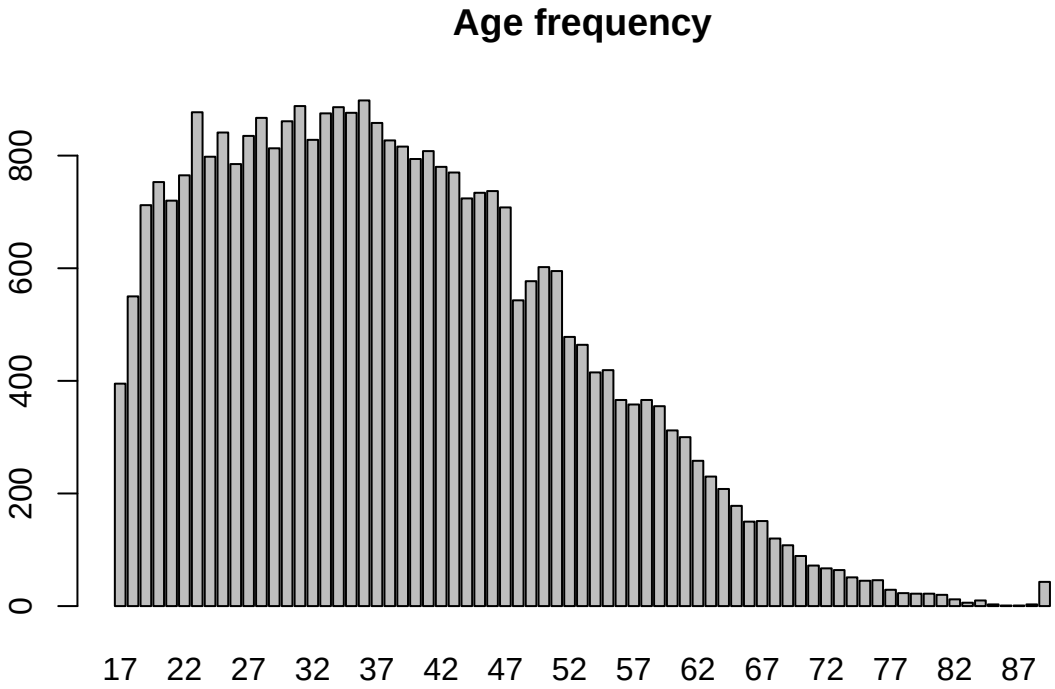
```
boxplot(ages.data, col = "tan4")
```




```
boxplot(age ~ year.income, data = myData, main = "Age vs Income", xlab = "Income", ylab = "Age", col =
```



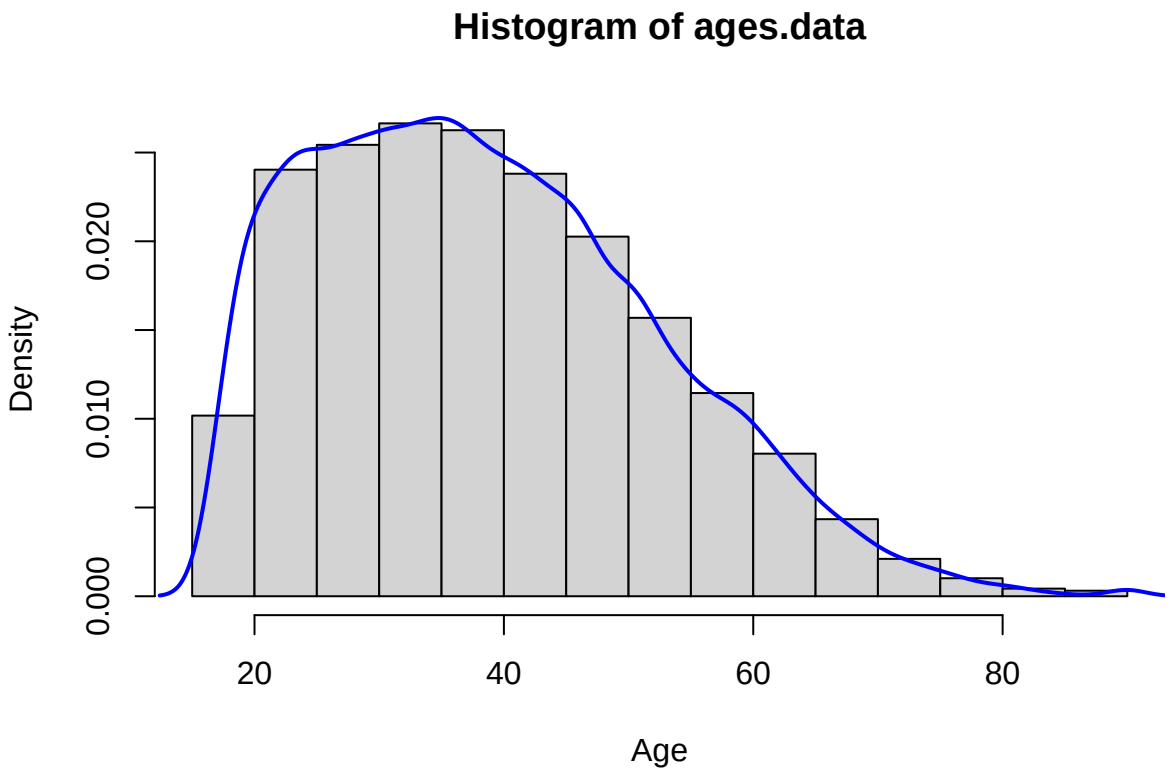

```
xlimit <- c(ages.min, ages.max)
ylimit <- c(0, 1600)
barplot(table(ages.data), main = "Age frequency")
```



In this boxplot, the values between 28 and 48 are inside the box-plot. And inside the two lines out of the box, m1 and m2, are the values between -2 and 78. The rest are outliers, which makes sense since there aren't any people in the data negative ages old, given that the minimum is 17. However, there are some people that are more than 78 years old, so it makes sense that there are outliers with bigger value than m2, as the maximum is 90. Also, the line in the middle of the box represents the median, which in this case is 37.

We can see in the boxplot between age and income, that interestingly enough, the older the people get the less money they make per year, whereas the people who earn the most are in their mid career.

```
hist(ages.data, include.lowest = TRUE, freq = FALSE, right = FALSE, xlab = "Age")
lines(density(ages.sorted), lwd = 2, col = 'blue')
```



Before, we predicted that the distribution would be symmetrical. As a matter of fact, the mean and the median were pretty close in the graph shown before, so if the amount of data had been smaller, perhaps we would have been able to see a more symmetric (or normal) distribution.

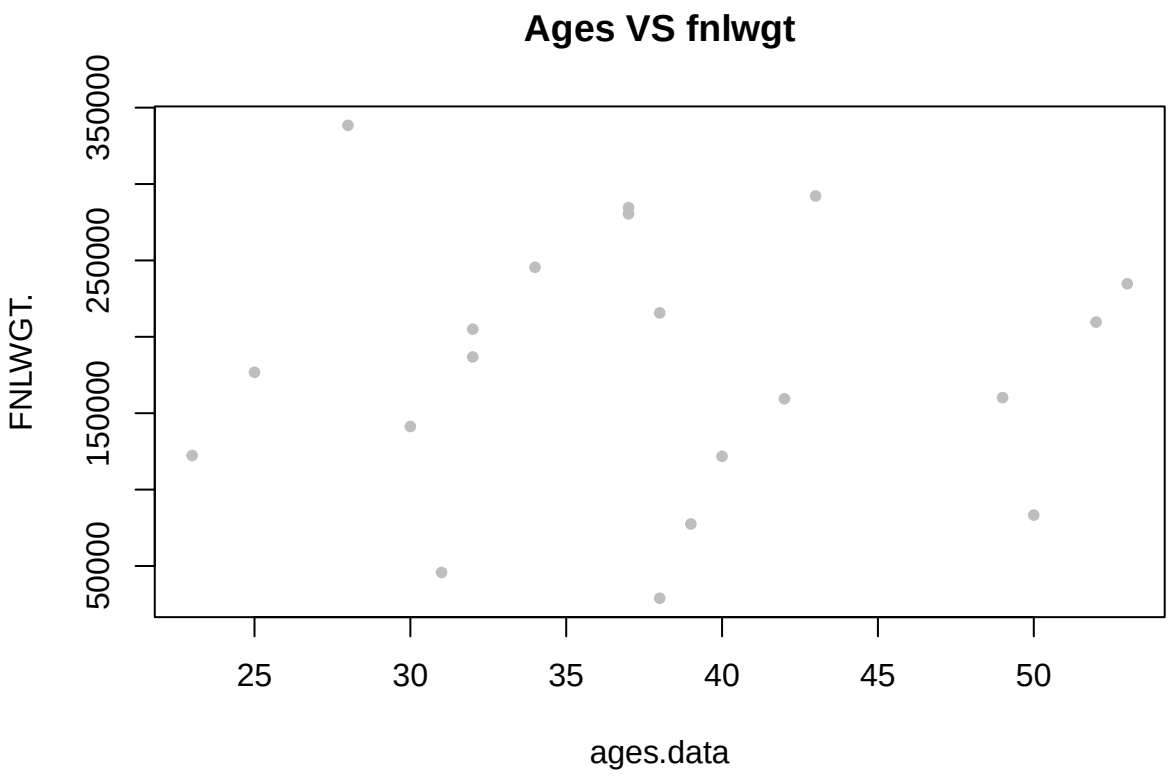
Another reason behind the fact that this histogram isn't symmetrical is because there are more people who are older, as the boxplot that we have shown before, in which whereas there are outliers above the boxplot, we cannot see any behind it, thus making the plot a little skewed.

2nd activity

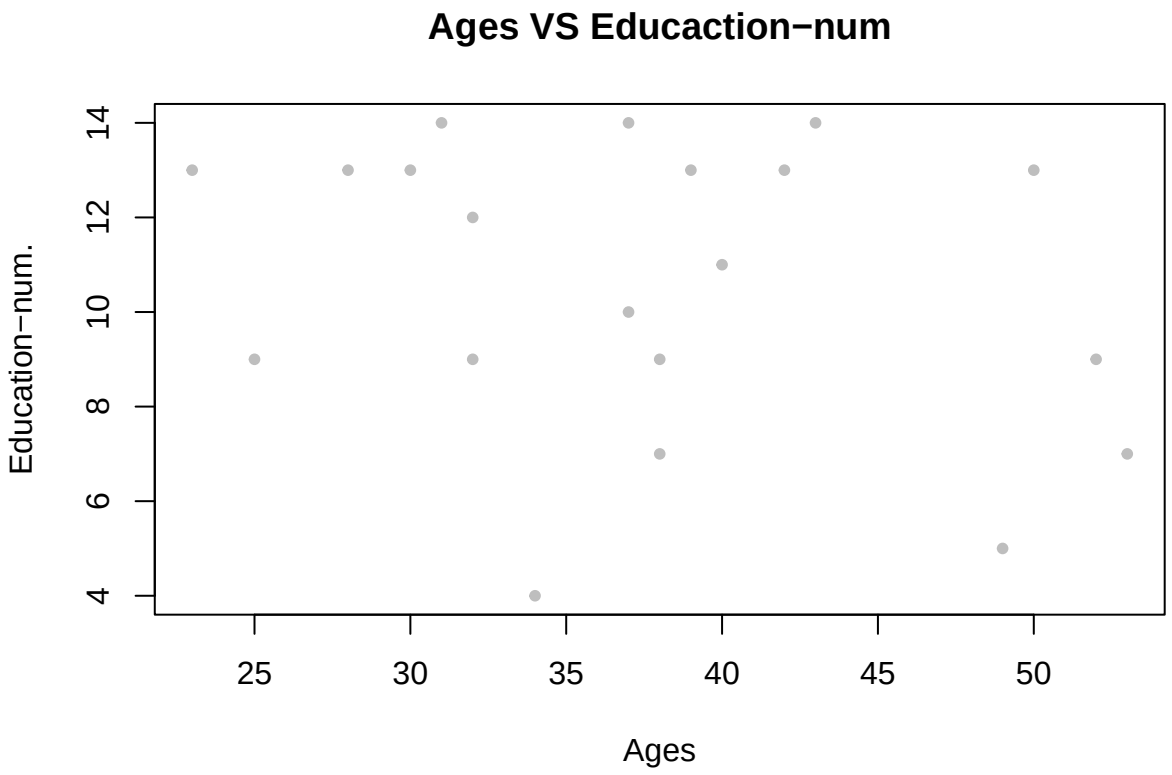
How does the variable you chose in the previous section relate to the other variables? Are the variables in the dataset correlated? Follow these instructions:

- **Make appropriate graphs to see if there is a relationship between your chosen variable and other variables.**

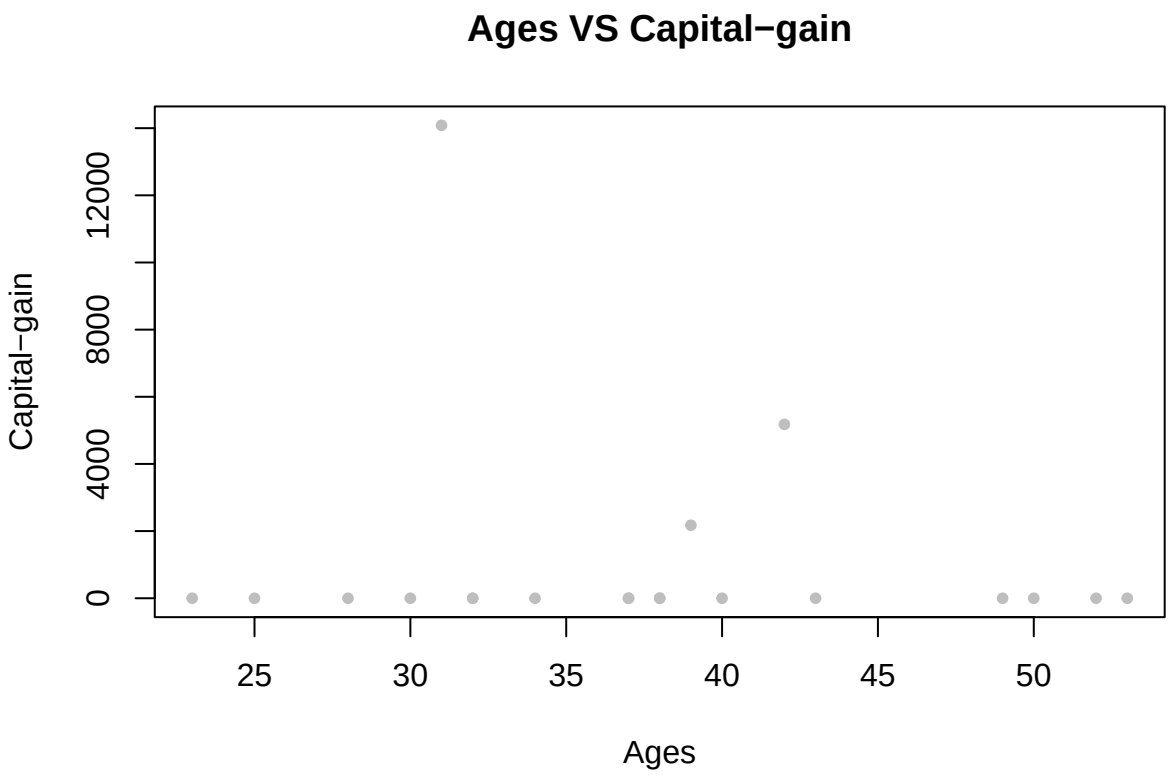
```
# Plot with respect to fnlwgt
plot(fnlwgt[1:20] ~ ages.data[1:20], ylab = "FNLWGT.", xlab = "ages.data", main= "Ages VS fnlwgt", data
```



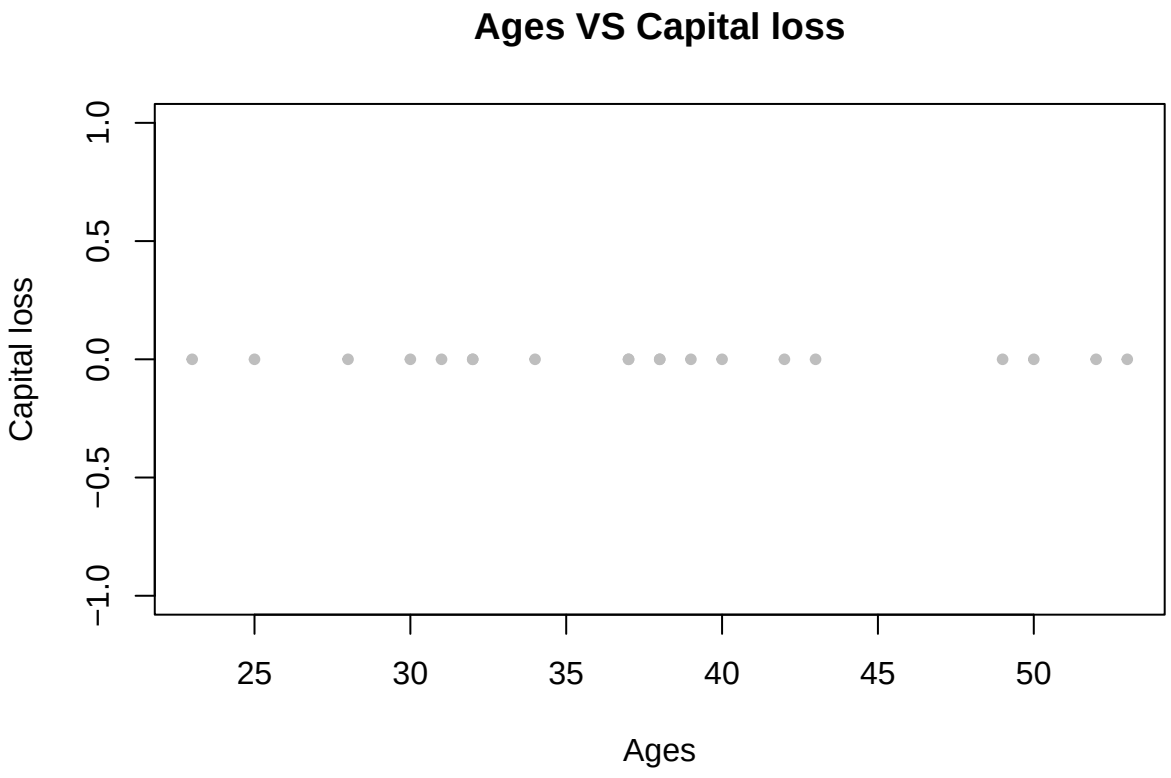

```
# Plot with respect to the education number.
plot(education.num[1:20] ~ ages.data[1:20], ylab = "Education-num.", xlab = "Ages", main= "Ages VS Educa
```



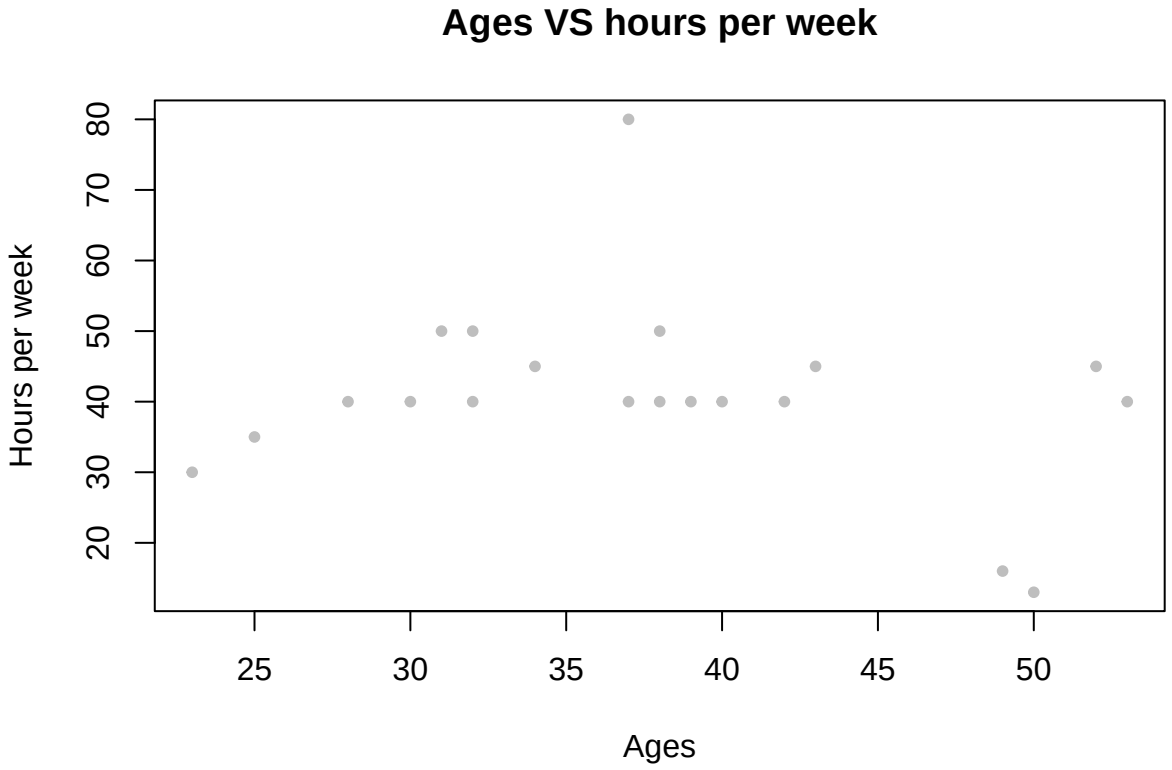

```
# Plot with respect to the capital-gain.
plot(capital.gain[1:20] ~ ages.data[1:20], ylab ="Capital-gain", xlab = "Ages", main= "Ages VS Capital-g
```




```
#Plot with respect to the capital loss.
plot(capital.loss[1:20] ~ ages.data[1:20], ylab = "Capital loss", xlab = "Ages", main= "Ages VS Capital
```




```
#Plot with respect to the hours per week.
plot(hours.per.week[1:20] ~ ages.data[1:20], ylab = "Hours per week", xlab = "Ages", main= "Ages VS hou
```



Note that we have only taken the first 20 elements of every variable, this is so as to avoid the plot from being full with dots, making impossible to see any kind of relationship; however, all the calculations will be done using all the available data.

In all these scatterplots we have taken the Ages as the **predictor**. All the other variables are the **response**. Let's analyze all the results.

In the fist plot, there isn't any kind of relationship between the fnlwgt and the ages, as the final weight just indicates the number of people that have the same characteristics as one person.

The second plot does not show any kind of relationship between the education number and the ages.

From plot four and five the isn't much to say since most of the people's incomes come from their work's salary and not from investment sources.

That leaves the last plot, were the ages of a person and their hours per week are plotted, in this plot we can see some relationship between those two variables, so probably the strongest correlation coefficient is going to be this one's.

- **Calculate the correlation coefficient between the chosen variable and each of the other variables. With which do you think the chosen variable has the strongest relationship? How is it? And the weakest?**

```
cor.fnlwgt <- cor(myData$age, myData$fnlwgt)
cat("The correlation coefficient with respect to the fnlwgt is:", cor.fnlwgt,"\n")
```

```
## The correlation coefficient with respect to the fnlwgt is: -0.07664587
```

```
cor.num <- cor(myData$age, myData$education.num)
cat("The correlation coefficient with respect to the education num is:", cor.num,"\n")
```

```
## The correlation coefficient with respect to the education num is: 0.03652719
```

```r
cor.gain <- cor(myData$age, myData$capital.gain)
cat("The correlation coefficient with respect to the capital gain is:", cor.gain,"\n")
```

```
## The correlation coefficient with respect to the capital gain is: 0.0776745
```

```r
cor.loss <- cor(myData$age, myData$capital.loss)
cat("The correlation coefficient with respect to the capital loss is:", cor.loss,"\n")
```

```
## The correlation coefficient with respect to the capital loss is: 0.05777454
```

```r
cor.hours <- cor(myData$age, myData$hours.per.week)
cat("The correlation coefficient with respect to the hours per week is:", cor.hours,"\n")
```

```
## The correlation coefficient with respect to the hours per week is: 0.06875571
```

```r
cat("The strongest relationship is with respect to the hours per week. (Aside from the capital gain, bu
```

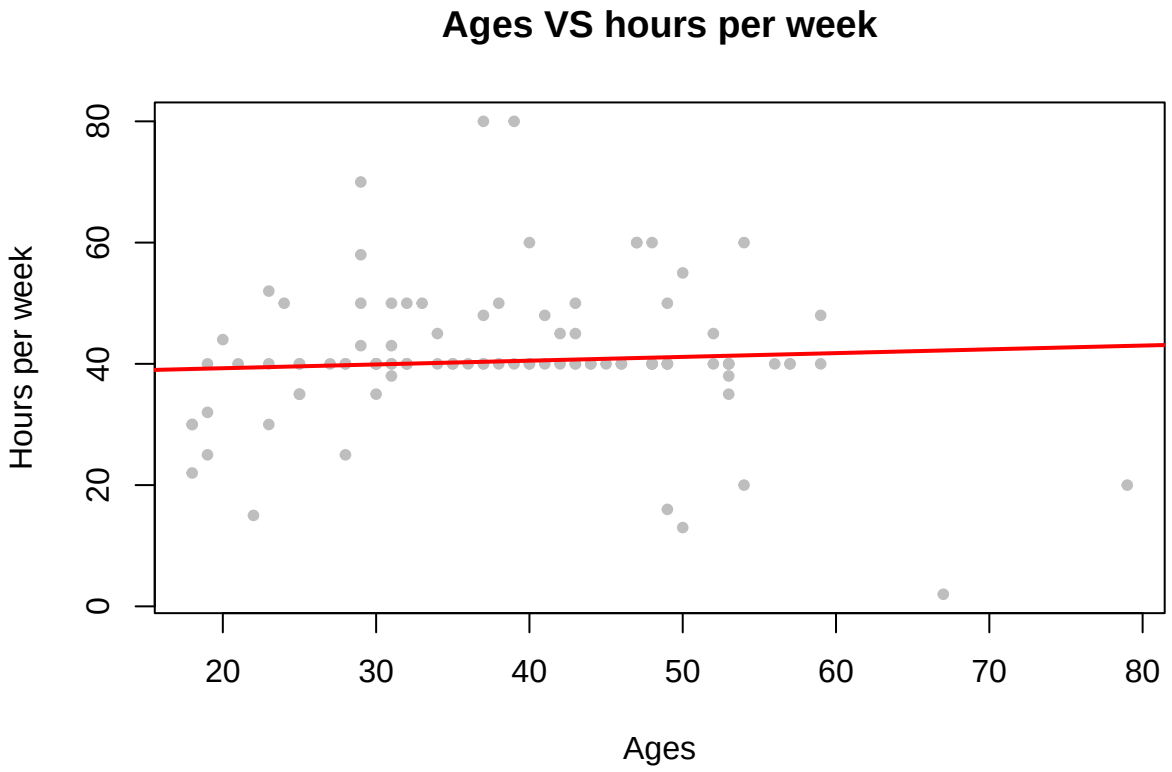
```
## The strongest relationship is with respect to the hours per week. (Aside from the capital gain, but s
```

- **Find two variables (we'll call them X and Y here) and calculate the regression line. Give the value predicted by the model for a value of Y that does not appear in the data set.**

```r
ageslm <- lm(hours.per.week ~ age, data = myData)
summary(ageslm)
```

```
##
## Call:
## lm(formula = hours.per.week ~ age, data = myData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.077  -1.957  -0.215   4.474  59.781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.036203   0.204804  185.72   <2e-16 ***
## age          0.062238   0.005005   12.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.32 on 32559 degrees of freedom
## Multiple R-squared:  0.004727,   Adjusted R-squared:  0.004697
## F-statistic: 154.6 on 1 and 32559 DF,  p-value: < 2.2e-16
```

```r
plot(hours.per.week[1:100] ~ ages.data[1:100], ylab = "Hours per week", xlab = "Ages", main= "Ages VS h
abline(ageslm, lwd = 2, col = 'red')
```




```
# We are going to predict the hours per week for a person 16 years old.
pred <- coef(ageslm)[1] + 16*coef(ageslm)[2]
cat("The hours our model predicts that a person 16 years old is going to work is:", pred)
```

```
## The hours our model predicts that a person 16 years old is going to work is: 39.03201
```

3rd activity

In the context set by data and reasoning about what you are doing:

- **Calculate a 95% confidence interval for a variable that you decide is useful.**

  Since we have been working with the parameter of ages all this time, let us construct a 95 % confidence interval for the ages variable.

  Just like before, since the sample size is large enough, we can apply the Central Limit Theory.

  It is important to note that since we don't know the variance, we will be using the estimate of the variance, s^2 . Because of this, the distribution followed will be that of the Student's t distribution.

$X \equiv$ "Ages of the individual."

$$\begin{cases} \hat{\mu} = \overline{X} = 38.58 \\ \hat{\sigma} = s = 13.64 \\ n = 32561 > 30, \text{We can use the Central Limit Theorem} \end{cases}$$

$$\frac{X_1 + ... + X_{32561}}{32561} \overset{C.L.T}{\sim} N(\mu, s/\sqrt{n}) \text{ With } n - 1 = 32560 \text{ degrees of freedom.}$$

$$P\left(\frac{|\overline{x} - \mu|}{s/\sqrt{n}} \leq t_{\alpha/2}\right) = 0.95$$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow t_{0.05/2} = t_{0.025} = \pm 1.96$$

So we have that our interval with 95% confidence is:

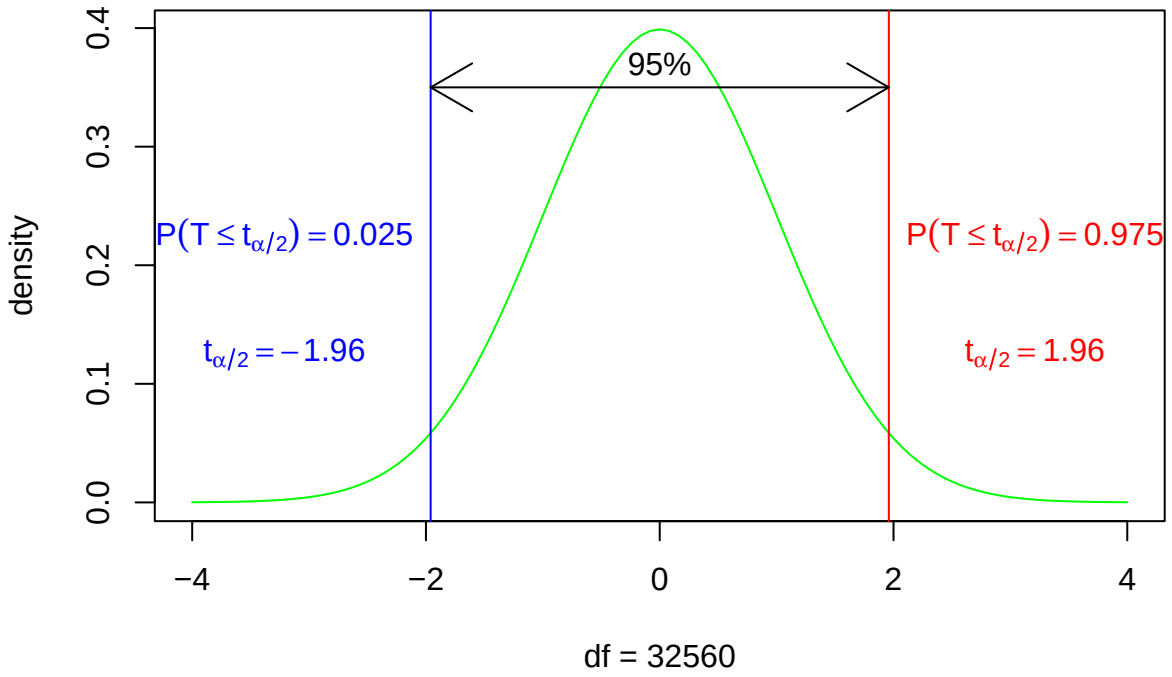
$$I_{\mu}^{0.95} = \left(\overline{X} - t_{0.025} \cdot \frac{s}{\sqrt{n}}, \overline{X} + t_{0.025} \cdot \frac{s}{\sqrt{n}}\right) = (38.43, 38.72)$$

Now, let's program it so as to get a more precise result and check whether our interval is accurate:

```
c1 <- 0.95 #Confidence interval we want to know.
ages.df <- ages.elements - 1 #Degrees of freedom.
ages.alpha <- (1-c1)/2 #Alpha/2, probability of P(t <= -t)
ages.t <- qt(ages.alpha, df = ages.df ) # t value, -1.96
ages.t <- abs(ages.t)# +/- 1.96.
standardError <- ages.sd/sqrt(ages.elements)
error <- (ages.t * standardError)
e1 <- ages.mean - error
e2 <- ages.mean + error
cat("So our interval is from,", e1, "to", e2 )
```

```
## So our interval is from, 38.43348 to 38.72981
```

```
  #Plotting the t student distribution and its confidence interval
curve(dt(x, df = 32560), from = -4, to = 4, type = "l", ylab = "density", xlab = "df = 32560", col = "g
text(ages.t + 1.25, ages.alpha + 0.2, expression(P(T<=t[alpha/2])==0.975), col = 'red')
text(-ages.t - 1.25, ages.alpha + 0.2, expression(P(T<=t[alpha/2])==0.025), col = 'blue')
text(ages.t + 1.25, ages.alpha + 0.1, expression(t[alpha/2]==1.96), col = 'red')
text(-ages.t - 1.25, ages.alpha + 0.1, expression(t[alpha/2]==-1.96), col = 'blue')
abline(v = -1.96, col = 'blue')
abline(v = 1.96, col = 'red')
arrows(x0 = -1.96, y0 = 0.35, x1 = 1.96, y1 = 0.35, code = 3)
text(x=0, y=0.37,labels='95%')
```



As we can see, we have gotten the same result, so the interval checks out.

Pose a hypothesis testing for a variable that you decide is useful. What decision have you made?

Let's make a hypothesis to see whether men or women had to work more in 1996, first:

```
# First we divide the male and female into two different lists and compute the mean and sd of their hou

male_hours <- subset(myData, sex == " Male", select = c("hours.per.week"))
female_hours <- subset(myData, sex == " Female", select = c("hours.per.week"))
female.size <- length(female_hours$hours.per.week) # Number of females
male.size <- length(male_hours$hours.per.week) #Number of males.
male.mean <- mean(male_hours$hours.per.week) # Mean of hours per week for males
male.sd <- sd(male_hours$hours.per.week) # Sd of hours per week for males
female.mean <- mean(female_hours$hours.per.week) # Mean of hours per week for females.
female.sd <- sd(female_hours$hours.per.week)# Sd of hours per week for females.
```

Male simple size is $n_M = 21790$ with average and sd: $\overline{X}_M = 42.42, S_A = 12.11$

Female simple size is $n_F = 10771$ with average and sd: $\overline{X}_M = 36.41, S_A = 11.81$

Just by looking at this data, we can gather that male back in 1998 had to work more hours per week, so we are going define

$$\begin{cases} H_0 : \mu_M - \mu_F = 0 \\ Ha : \mu_M - \mu_F \neq 0 \end{cases}$$ With a significance level $\alpha = 0.05$

The standard deviation is:

$$s^2 = \frac{(n_M - 1)S_M^2 + (n_F - 1)S_F^2}{n_M + n_F - 2}$$

Since both sample sizes are much bigger than 30, we may make use of the Central Limit Theory, so they will follow a Normal Distribution. However, since the variance is unknown, we will use its estimate, which will

change our distribution to that of the Student's t distribution.

$$\overline{X}_M - \overline{X}_F \sim N(\mu_M - \mu_F, \sigma\sqrt{(2/n)})$$
$$t = \left(\frac{(\overline{X}_M - \overline{X}_F) - (\mu_M - \mu_F)}{s\sqrt{(2/n)}} \right) \sim t_{n+n-2}$$

Now that we know which distribution we will be using, all that's left is to compute the p-value taking the null hypothesis into account:

$$p - value = P(|T| > t|\mu_M - \mu_F = 0)$$

And the t in this case is:

```
mfsd <- (((male.size-1)*male.sd^2) + ((female.size-1)*female.sd^2))/(male.size+female.size-2) ##Standar
t_score <- (male.mean-female.mean)/(mfsd*sqrt(2/32561))
t_score # The t-score of the distribution male female
```

```
## [1] 5.315676
```

```
p_value <- pt(t_score, 32559, lower.tail = FALSE)# Lastly, we compute the p-value.
p_value #Which is practically 0.
```

```
## [1] 5.348043e-08
```

$$P(|T| > 5.315676) \approx 0 < 0.025 = \frac{\alpha}{2}$$

So since the p-value is smaller than the α value, we can reject the null hypothesis, thus we conclude that male or female did not work the same amount of hours per week back in 1998.

However, we do have to note that in this data set the women are underrepresented; that is, there are less women than men in this data set, which may be due to some biased census. Perhaps that's why we have obtained this result, so we have to keep in mind that even though our results point to a conclusion, the real answer might be a totally different one.