Statistics 104 Spring 2020 Midterm

Lars Lien Ankile

03.09.2020

Problem Scoring		
Problem	Point Value	Points Scored
1	46	
2	38	
3	44	
4	72	
Total	200	

I confirm that I have worked independently on this take-home exam, except for any assistance I may have received from the teaching staff with technical issues. All the work is my own, and I have not collaborated in any way with fellow students.

Signature:

PROBLEM 1: SHORT ANSWER

a)

i. There are several reasons why this study and conclusion might be misleading, first, (1) the study was conducted over a very short period of time and the athletes might require much more time before one has fully adapted to the new diet, and the observed effect might be only temporary. (2) The people in the study wasn't randomly allocated into the study and control group, but was allowed to decide themselves, which might introduce spurious causes into the study it is hard to control for. (3) The people on the ketogenic diet only showd signs "indicative" of impaired bone health, deteriorating bones weren't detected directly, and it is unknown whether a ketogenic diet could highten these biomarkers for other reasons than deteriorating bone. (4) It could also be that people require a different amount of calories when they're one a ketogenic diet than when they're not and that the biomarkers could be observed because the participants weren't ingesting enough calories because they requried larger amounts on the ketogenic diet. (5) Lastly it is uknown how big the study population was, which makes it hard to know if the results are statistically significant or just chance.

ii. Two reasons: (1) His subjective perception of the diet and how he feels should weigh heavily in my opinion. Since he's feeling positive effects of the diet, I think that indicates that it could be beneficial for him to continue. Also, the study lasted for only a month, while our friend is three months into the diet, which might make the results not applicable. That is because it's possible that the indicators of deteriorating bone health only occurs in the first period and fades once you're accustomed to the diet. (2) The study was conducted on people who self-selected into the different groups, and it doesn't say anything about what kind of poeple they were. It is highly uncertain if the effects would apply to our friend. E.g. it is not certain that a ketogenic diet will affect the bone health of an old man the same ways it would for a young woman.

b)

i. Since FBG levels are normally distributed, we can use the `qnorm` function to find the level that would place us in the upper 1%, i.e. the 99th percentile. When `qnorm` is used with the approriate arguements, we get that a FBG level of 127.9 will place you in the 99th percentile. See below for R-code.

```
# Find 99th percentile for FBG levels
qnorm(p = 0.99, mean = 100, sd = 12)
```

```
## [1] 127.9162
```

ii. When using a mean of 100 mg/dL and SD of 12 mg/dL with the function `pnorm` we can get the probability of someone having a FBG level between 100 and 125mg/dL by first getting the probability of them having below 125, and then subtracting away the probability of them having below 100. Then we can raise this number to the second power to get our final answer, since we're assuming these measurements are independent. When performing this in R below, I get that the probability of a non-diabetic being classified as pre-diabetic is 0.2317, or 23.2%.

```
# Probability that a non-diabetic will be classified as pre-diabetic
p_fbg_100_125 <- (pnorm(q = 125, mean = 100, sd = 12)
                  - pnorm(q = 100, mean = 100, sd = 12))
```

```
p_fbg_100_125^2
```

```
## [1] 0.2317359
```

c) The probability of the number of rolls exceeding 3 is the same as 1 minus the probability of the number of rolls being 3 or less, i.e. $P(N > 3) = 1 - P(n \leq 3)$. The probability of failing on the first roll is $P(R_1 \leq 1) = \frac{1}{6}$, the probability of faling on the second roll is $P(R_2 \leq 2) = \frac{2}{6} = \frac{1}{3}$, and the probability of failing on the third roll is $P(R_3 \leq 3) = \frac{3}{6} = \frac{1}{2}$. Putting it all together we get following equations,

$$\begin{aligned} P(N > 3) &= 1 - P(N \leq 3) \\ &= P(\text{Fail on 1st roll}) \qquad + P(\text{Fail on 2nd roll}) + P(\text{Fail on 3rd roll}) \\ &= \frac{1}{6} \qquad + \frac{5}{6} \cdot \frac{1}{3} + \frac{5}{6} \cdot \frac{2}{3} \cdot \frac{1}{2} \\ &= 0.2777778, \end{aligned}$$

and get that the probability of rolling the die more than 3 times is 0.2778, or 27.8%

```
# Probability that N is bigger than 3
1 - (1/6 + 5/6*1/3 + 5/6*2/3*1/2)
```

```
## [1] 0.2777778
```

d)

```
# Defining some variables that will be useful in the following problems
inf_rate_covid <- 20 / 1000
inf_rate_flu <- 30 / 1000
num_undergrads <- 6800
```

i. We know there's a 3% chance of getting the flu in any given month, and 2% for COVID. Therefore, for one person the probability of not getting either in a given month is $0.97 \cdot 0.98 = 0.9506$. Since we want to know the probability of three people avoiding both diseases, assuming infections are independent between the people, we can raise the individual probability to the 3rd power, and we get $0.9506^3 = 0.859$, or 85.9%, probability of none of the roommates getting sick from any of the diseases.

```
# Infections from each disease is independent of each other
# Probability of one person not contracting flu in one month
p_no_flu <- 1 - inf_rate_flu

# Probability of one person not contracting covid-20 in one month
p_no_covid <- 1 - inf_rate_covid

# Probability of three people not contracting either disease in one month
(p_no_flu * p_no_covid)^3
```

```
## [1] 0.8590005
```

ii. The assumptions required is that (1) infections from the different diseases are independent and (2) infections between rommates are independent. Assumption (1) is not realistic, first, because once one is infected by let's say the flu, one's immune system is likely weakened, which makes one more susceptible to new infections. At the same time, once one's infected with the flu or COVID, one's probably going to stay home or isolated which might make it less likely to get infected by other diseases. These two effects work in opposite directions, and it's hard to know which has the stronger effect. Assumption (2) is not realistic because if someone you live with gets an infection, your chance of getting infected too probably rises dramatically.

iii. The first month 2% out of 6800 will be infected. The month after, 2% out of the ones who weren't infected last month will get it, and same the month after. From the calculation below, we see that in expectation, almost 400 people will get infected with COVID-20.

```
# Amount of undergrads who we expect will get COVID-20
# Assuming that one can only be infected once
covid_month_1 <- inf_rate_covid * num_undergrads
covid_month_2 <- inf_rate_covid * (num_undergrads - covid_month_1)
covid_month_3 <- inf_rate_covid * (num_undergrads - covid_month_1 - covid_month_2)

covid_month_1 + covid_month_2 + covid_month_3
```

```
## [1] 399.8944
```

iv. We can consider this a binomial process where the probability of "success" (in quotes because an COVID-infection isn't *really* a success) is 0.02, and the number of trials is 6800, and the amount of "successes" we want to see more than 150. We can use the `pbinom` function in R, with the `lower.tail`-argument set to `FALSE` to get $P(N > 150)$. Doing this gives a chance of the school going into shut-down mode of 0.1059, or 10.6%.

```
# Probability that Harvard will go into shut-down mode in March
shut_down_threshold <- 150
pbinom(q = shut_down_threshold,
       size = num_undergrads,
       prob = inf_rate_covid,
       lower.tail = FALSE)
```

```
## [1] 0.1058926
```

PROBLEM 2: RETIREMENT PLANNING

a) Below is my implementation of the simulation. I chose to write it as a function that takes in the amount percentage that are supposed to be in stocks as a parameter because the code in for a), b), and c) is practically the same, except for the proportion of money in stocks and bonds. For my code, I start with getting the file, and getting out the relevant data into new variables for stock returns and bond returns for easier handling later. Then I define some variables for use in the simulation. After that I create the vector that is going to hold the resuls of the replicates of the simulation, and therefore it has the length of the variable `replicates`. Then I set the seed for the pseudorandom number generator, and start the outer for-loop. This for-loop simulates 40 years of investing a 1000 times. The inner for loop simulates one year of investing. It first updates the account balance based on the start contribution, current year and the contribution growth rate. Then it samples a random year, and applies the weighted average of the returns from the stock and bond market to the account, and then repeats. Every time the 40 years of investing is done, the final account value is saved in the result vector. This vector is returned in the end.

```
# Here I've chosen to define a function to be able to reuse some code
retirement_saving_sim <- function(stocks_perc = 0.8) {
  # Load data
  stocksbonds <- read.csv('stocksbonds.csv')
  sp500 <- stocksbonds$sp500
  bonds <- stocksbonds$bonds

  # Set parameters
  # The number of years we're investing for
  num_years <- 40
  # The initial contribution to the account
  add_funds <- 2000
  # The yearly growth of the contribution
  yearly_inc <- 0.03

  # The amount of times to run the simulation
  replicates <- 1000

  # Create storage vector
  account_values <- vector(mode = 'numeric', length = replicates)

  # Set seed
  set.seed(2020)

  # Simulate returns
  # Run the simulation a thousand times
  for (k in 1:replicates) {

    # For each simulation, first set the account value to 0
    account_value <- 0
    # Simulate each year with one round of this loop
```

```
    for (year in 1:num_years) {
      # First, at the start of each year, add funds to the account
      # This value increases by 3% each year, and therefore it must be raised to
      # power of the year less 1 (theres no increase in added funds first year)
      # Account value at the start of the year
      account_value <- account_value + add_funds * (1 + yearly_inc)^(year - 1)

      # Get the returns for both sp500 and bonds for the year
      # This is done by getting a random return from the list of yearly returns
      # for both security classes
      sample_year <- sample(1:nrow(stocksbonds), size = 1, replace = TRUE)
      sp_return <- sp500[sample_year]
      bond_return <- bonds[sample_year]

      # Calculate weighted average over returns from bonds and stocks
      avg_return <- stocks_perc*sp_return + (1-stocks_perc)*bond_return

      # Then add the value of the returns that year to the account value
      # Account value at the end of the year after returns on stocks and bonds
      account_value <- account_value * (1 + avg_return)

    }
    # Add the final account value after 40 years of
    # investing to the result vector
    account_values[k] <- account_value
  }
  return (account_values)
}
```

b) When running the simulation above, I get that the 90th percentile account value after 40 years of investing is roughly $2.5 million.

```
# Compute 90th percentile
# Here I first run the simulation with a 80/20 split in stocks/bonds
account_values_b <- retirement_saving_sim(stocks_perc = 0.8)
# Here I define some quantiles I'm interested in, most importantly the 90th
quants <- c(0.1, 0.25, 0.5, 0.75, 0.9)
# Here I ask R to give me values for the quantiles
quantile(account_values_b, probs = quants)
```

```
##       10%       25%       50%       75%       90%
##  433874.7  656048.7 1059225.4 1668828.3 2503401.9
```

c) Once you change the portfolio to be 100% invested in stocks the 90th percentile account value after 40 years of investing is roughly $3.85 million. This is a lot more. However, we see that the lowest 10th percentile account value is lower for the portfolio in 100% stocks, so it's more risky as well.

```r
# Here I run the simulation with 100% invested in stocks
account_values_c <- retirement_saving_sim(stocks_perc = 1.0)
# Here I get the same quantiles as above, just for the new results
quantile(account_values_c, probs = quants)
```

```
##       10%       25%       50%       75%       90%
##  415959.2  690864.1 1269657.5 2247894.7 3850718.8
```

d) When running the simulation for the modified version of investing one's age below, I get that the 90th percentil return is $1.4 million.

```r
# Here I choose to do the whole simulation again instead of using
# the function from above because the internals are suffieciently different
# Load data
stocksbonds <- read.csv('stocksbonds.csv')
sp500 <- stocksbonds$sp500
bonds <- stocksbonds$bonds

# Set parameters
# All parameters, vectors, and seeds are the same as before
num_years <- 40
add_funds <- 2000
yearly_inc <- 0.03

replicates <- 1000


# Create storage vector
account_values_d <- vector(mode = 'numeric', length = replicates)

# Set seed
set.seed(2020)

# Simulate returns
# Run the simulation a thousand times
for (k in 1:replicates) {

  # For each simulation, first set the account value to 0
  account_value <- 0
  # For each simulation, reset stocks percentage to 100%
  stocks_perc <- 1.0
  # Simulate each year with one round of this loop
  for (year in 1:num_years) {
    # First, at the start of each year, add funds to the account
    # This value increases by 3% each year, and therefore it must be raised to
    # power of the year less 1 (theres no increase in added funds first year)
    # Account value at the start of the year
    account_value <- account_value + add_funds * (1 + yearly_inc)^(year - 1)
```

```
    # Get the returns for both sp500 and bonds for the year
    # This is done by getting a random return from the list of yearly returns
    # for both security classes
    sample_year <- sample(1:nrow(stocksbonds), size = 1, replace = TRUE)
    sp_return <- sp500[sample_year]
    bond_return <- bonds[sample_year]

    # Calculate weighted average over returns from bonds and stocks
    avg_return <- stocks_perc*sp_return + (1-stocks_perc)*bond_return

    # Then add the value of the returns that year to the account value
    # Account value at the end of the year after returns on stocks and bonds
    account_value <- account_value * (1 + avg_return)

    # Decrease the percentage of money to put in
    # stocks by 2 percentage points each year
    stocks_perc <- stocks_perc - 0.02

  }
  # Add the final account value after 40 years of investing to the result vector
  account_values_d[k] <- account_value
}

# Do the actual calculation of quantiles, using the same quantiles as before
quantile(account_values_d, probs = quants)
```

```
##       10%       25%       50%       75%       90%
##  428739.6  568943.3  745466.2 1033727.7 1409085.2
```

e) There are mainly two things to consider when saving for retirment, or saving in general: (1) expected returns and (2) risk. We see that having 100% of the portfolio in stocks for the whole period yields the by far hightest 90th percentile gain (3.85 mill for c) to 2.5 and 1.4 mill for b) and c), respectively). However, we also see that the portfolio in 100% stocks also has the worst returns in the 10th percentile, i.e it is also the riskiest portfolio. But, seeing that the risky portfolio outperforms the other schemes for the 25th percentile and all above and the difference in return is small in the 10th percentile, I'd say that the reward outweighs the risk (to me at least), and one should opt for having 100% of one's savings in stocks.

f) In the following I've chosen to use my function from a) to calculate the expected returns for being 100% invested in the stock market a thousand times. Then I calculate how much the retirement schema might cost me, and subtract that sum from every one of the thousand entries in the returns vector. Then, I calculate the proportion of those entries went below zero which gives me the chance of not having enough money for retirement. The result was 0.525, or 52.5%, chance of not having enough money for retirement.

```
# Initialize a variable to hold the total amount of expenses
total_expenses <- 0
# Loop over the number of years we're spending money
```

```r
for (year in 1:20) {
  # Every year I'll add to the total expenses yearly
  # expenses times the growth of the expenses
  total_expenses <- total_expenses + 50000 * (1 + 0.03)^(year - 1)
}

# Here I get the expected returns from the stock market,
# a thousand simulations of 40 years investing
account_values_f <- retirement_saving_sim(stocks_perc = 1.0)

# Calcualate how much money there'd be left in each of the thosand cases
money_left <- account_values_f - total_expenses

# Find the number of years that the balance went to zero or below
years_below_0 <- money_left[money_left < 0]

# Find the actual percentage of cases where the balance went below zero
# to where it didn't
length(years_below_0) / length(account_values_f)
```

```
## [1] 0.525
```

PROBLEM 3: GENETICS OF AUSTRALIAN CATTLE DOGS

a) Let B be the event that an Australian cattle dog has a black coat and M be the event that an Australian cattle dog has any mask (either unilateral or bilateral). Then, the probability that a dog has a black coat and a mask is $P(B \cap M) = P(B) \cdot P(M|B) = 0.4 \cdot (0.35 + 0.25) = 0.24$, or 24%.

b) Let R be the event that an Australian cattle dog has a red coat. To find what percentage of Australian cattle dogs have bilateral masks, we can add the probability that a black-coated dog has on with the probability that a red-coated has one. I.e. $P(M_2) = P(M_2|B) \cdot P(B) + P(M_2|R) \cdot P(R) = 0.35 \cdot 0.4 + 0.1 \cdot (1 - 0.4) = 0.2$, or 20% of all Australian cattle dogs have bilateral masks.

c) To find out what probability any given dog with a bilateral mask is of being a Red Heeler we can apply Bayes' theorem with the same events as defined in a) and b).

$$\begin{aligned} P(R|M_2) &= \frac{P(R)P(M_2|R)}{P(M_2)} \\ &= \frac{0.6 \cdot 0.1}{0.2} \\ &= 0.3 \end{aligned}$$

From the above we see that the probability of a dog with a bilateral mask being a Red Heeler is 0.3, or 30%.

d)

i. Let R be the event that a Australian cattle dog is a Red Heeler, D_0 be a dog without hearing deficits, D be a dog with any hearing deficit, and M_2 be a dog with a bilateral mask. Then, the probability that any given dog is a Red Heeler with a bilateral mask and no hearing deficits is

$$\begin{aligned} P(M_2 \cap D_0 \cap R) &= P(D_0|M_2, R)P(M_2|R)P(R) \\ &= (1 - P(D|M_2, R))P(M_2|R)P(R) \\ &= P(M_2|R)P(R) - P(M_2 \cap D \cap R) \\ &= 0.1 \cdot 0.6 - (0.045 + 0.012) \\ &= 0.003, \end{aligned}$$

or 0.3%.

ii. We know that out of all dogs, $0.012 + 0.045 = 0.057$, or 5.7%, of them will be Blue Heelers with a bilateral mask and some degree of hearing deficit. If we divide that by 0.14, since 14% of all dogs are Blue Heelers with bilateral masks, we get that $\frac{0.057}{0.14} = 0.4071$ out of Blue Heelers with bilateral masks have some degree of hearing deficit. The complement of that, $1 - 0.4071 = 0.5929$, or 59.3%, out of the black coated dogs with bilateral masks have no hearing deficits.

iii. Using the same events as in the problem formulation and in my previous answers, the prevalence of deafness in Reed Heelers is given by

$$\begin{aligned}
P(D|R) &= P(D \cap M_0|R) + P(D \cap M_1|R) + P(D \cap M_2|R) \\
&= P(D|M_0, R)P(M_0|R) + P(D|M_1, R)P(M_1|R) + \frac{P(D \cap M_2 \cap R)}{P(R)} \\
&= 0.4 \cdot 0.5 + 0.4 \cdot 0.4 + \frac{0.012 + 0.045}{0.6} \\
&= 0.455
\end{aligned}$$

I.e. 45.5% of Red Heelers have some hearing deficit.

The prevalence of deafness in Blue Heelers is given by

$$\begin{aligned}
P(D|B) &= P(D_1|B) + P(D_2|B) \\
&= P(D_1 \cap (M_0 \cup M_1)|B) + P(D_2 \cap (M_0 \cup M_1)|B) + P(D \cap M_2|B) \\
&= P(D_1|M_0 \cup M_1, B)(P(M_0|B)P(M_1|B)) + P(D_2|M_0 \cup M_1, B)(P(M_0|B)P(M_1|B)) \\
&\quad + \frac{P(M_2 \cap D \cap B)}{P(B)} \\
&= 0.05 \cdot (0.4 + 0.25) + 0.01 \cdot (0.4 + 0.25) + \frac{0.045 + 0.012}{0.4} \\
&= 0.1815.
\end{aligned}$$

I.e. 18.2% of Blue Heelers suffer from deafness.

In short, 45.5% of Red Heelers, as compared to 18.2% of Blue Heelers have hearing deficits, i.e. a much higher proportion of Red Heelers suffer from deafness.

iv. To find the probability that a dog with no hearing deficits is a Blue Heeler we can again use Bayes' theorem. Let D_0 be the event that a dog is not deaf and B be the event that a dog is a Blue Heeler. Then the probability $P(B|D_0)$ is

$$\begin{aligned}
P(B|D_0) &= \frac{P(B)P(D_0|B)}{P(D_0)} \\
&= \frac{0.4 \cdot 0.8185}{0.4 \cdot 0.8185 + 0.6 \cdot 0.545} \\
&= 0.5003,
\end{aligned}$$

or 50% that any given dog with no hearing deficits is a Blue Heeler.

PROBLEM 4: SELF-MANAGEMENT

a) The first thing I notice which seems natural is that the median age is pretty high, at 70, and the oldest is 92 years and the youngest is 28. Most of the people are between the ages of 50 and 90, which makes sense seeing that old people are much more at risk for most chronic diseases. However, the data is slightly left-skewed with some cases among younger people. Furthermore we see that there are some more men than women in the study, with roughly 60% being men and 40% women, which is probably caused by men having a higher rate of these chronic diseases. Lastly, we see from both the last barchart above, and also the two last tables in the numerical summaries, that the most common diseas is diabetes type-II, with roughly 37% percent of participants having it, the next most common disease is chronic obstructive pulmonary disease, which roughly 25% of the study participants have, and Chronic heart failure and chronic renal disease come in at last place with around 19% percent of study participants each.

```
#load the data
load("self_manage.Rdata")

#numerical summaries
cat("Statistical summary of age:\n")
```

```
## Statistical summary of age:
```

```
summary(self.manage$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   28.00   63.00   70.00   69.64   78.00   92.00       3
```

```
cat("\nSummary of sex (count):")
```

```
##
## Summary of sex (count):
```

```
addmargins(table(self.manage$sex))
```

```
##
##   male female    Sum
##    694    458   1152
```

```
cat("\nSummary of sex (ratio):")
```

```
##
## Summary of sex (ratio):
```

```
prop.table(table(self.manage$sex))
```

```
##
##      male    female
## 0.6024306 0.3975694
```

```
cat("\nSummary of diseases (count):")
```

```
##
```

```
## Summary of diseases (count):
```

```
addmargins(table(self.manage$disease))
```

```
##
## DM-II  COPD    HF   CRD   Sum
##   422   290   223   219  1154
```

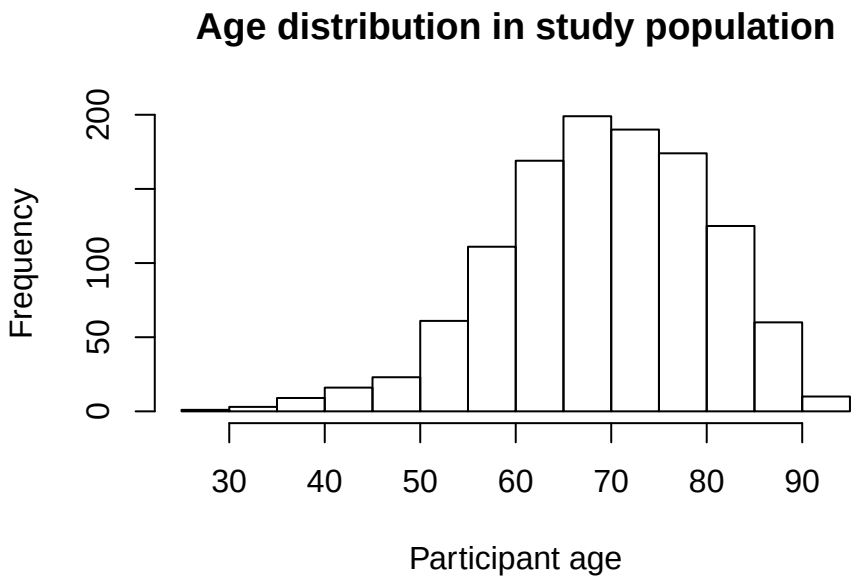
```
cat("\nSummary of diseases (ratio):")
```

```
##
## Summary of diseases (ratio):
```

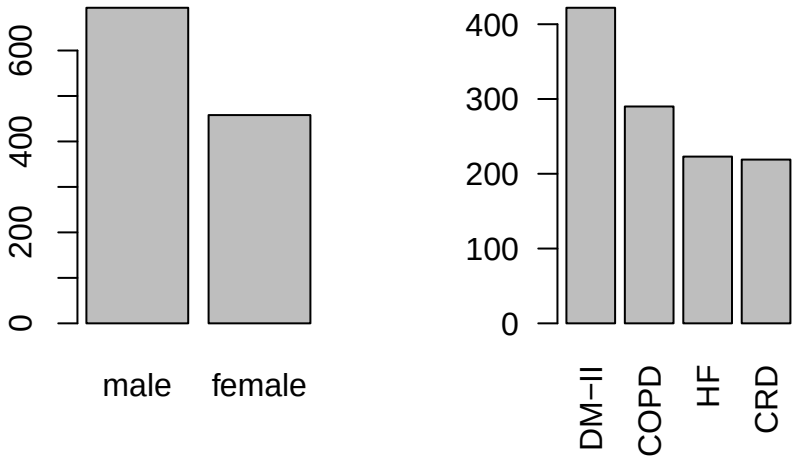
```
prop.table(table(self.manage$disease))
```

```
##
##     DM-II      COPD        HF       CRD
## 0.3656846 0.2512998 0.1932409 0.1897747
```

```
#graphical summaries
hist(self.manage$age,
     main = "Age distribution in study population",
     xlab = "Participant age")
```




```
par(mfrow = c(1, 2))
barplot(table(self.manage$sex))
barplot(table(self.manage$disease), las = 2)
```



b) From the below summary and histogram over the pam scores, we see that the data i slightly right-skewed, with a median around 53, i.e. slightly above the middle of the score range for that parameter. One peculiar thing about the histogram is that we can observe is that especially 2 bins are very large: the bin from 45-50 and 55-60. One reason might be because being above 47 in rating will barely place you in category 2, while a score of 55 will barely place you in category 3, so it can seem like participants have been cognizant of this while rating themselves, or it might be randomness, hard to know. When it comes to categories, we see from the above barchart that category 3 is the most common, with 33% of participants being in that category. There are almost as many in category 1 and 2, with 27% and 28% percent of participants being in those categories, respectively. The by far smallest category is category 4 with only 12% of participants being able to successfully keep new behaviors over time.

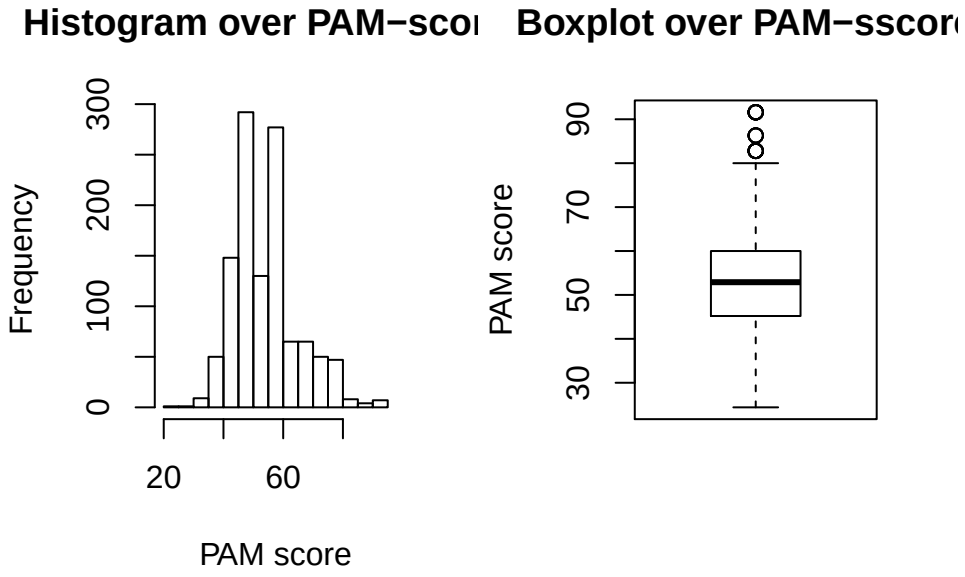
```
#PAM-13 score
cat("Statistical summary of pam score:\n")
```

```
## Statistical summary of pam score:
```

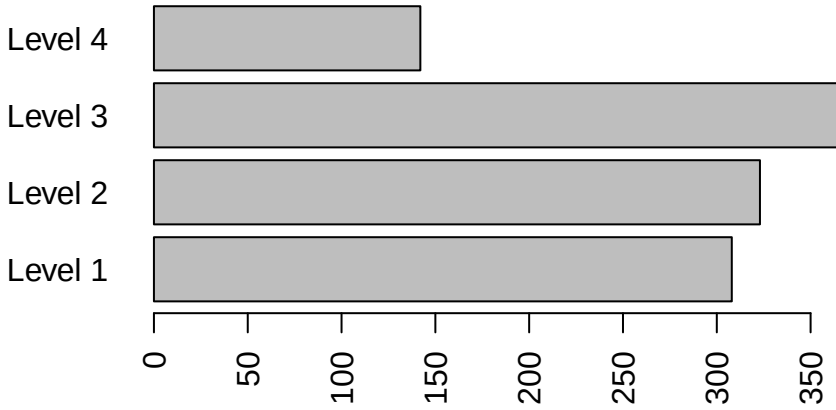
```
summary(self.manage$pam.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    24.4    45.2    52.9    54.1    60.0    91.6
```

```
par(mfrow = c(1, 2))
hist(self.manage$pam.score,
     main = "Histogram over PAM-scores",
     xlab = "PAM score")
boxplot(self.manage$pam.score,
        main = 'Boxplot over PAM-sscores',
        ylab = "PAM score")
```




```
#PAM level
par(mfrow = c(1, 1))
barplot(table(self.manage$pam.cat), las = 2, horiz = TRUE)
```




```
cat("\nOverview of pam categories (ratio):")
```

```
##
## Overview of pam categories (ratio):
```

```
prop.table(table(self.manage$pam.cat))
```

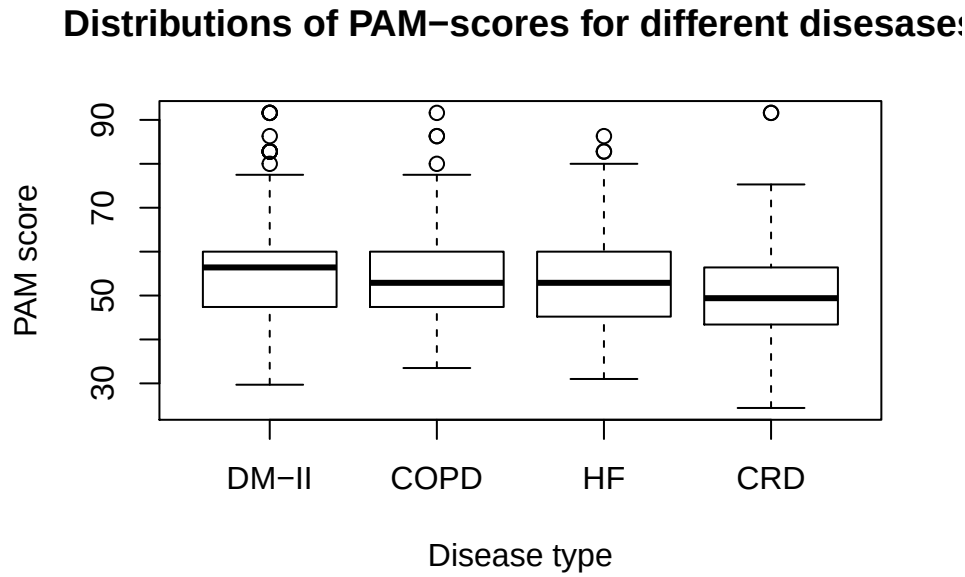
```
##
##   Level 1   Level 2   Level 3   Level 4
## 0.2668977 0.2798960 0.3301560 0.1230503
```

c)

i. From the boxplot below, wee can see that there are some differences between the PAM-scores for people with the different diseases. We see that middle 50% of the population is roughly the same for DM-II, COPD, and HF, except that DM-II has a slightly higher median. People who suffer from CRD appears to have the lowest median score and also the fewest outliers, but there's one outlier that is very high. There are no outliers in the lower range, only in the

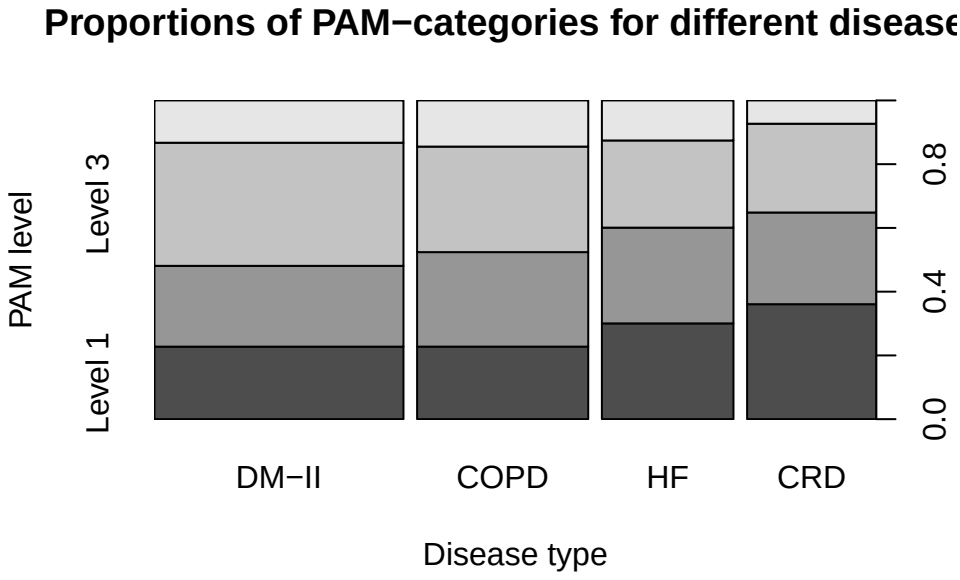
higher range. In short, DM-II seems to have the best scores, CRD seems to have the worst, while COPD and HF are somewhere in the middle.

```
#disease type and PAM-13
plot(self.manage$pam.score ~ self.manage$disease,
     main = 'Distributions of PAM-scores for different disesases',
     xlab = "Disease type",
     ylab = "PAM score")
```



ii. The below stacked barchart shows as a proportion of the total number within each disease class are in the different PAM-categories. The categories are sorted from Level 1 at the bottom of the stack to Level 4 at the top. This view confirms the conclusion from i). Here we see that DM-II has the highest proportion of people in category 3 and 4, while CRD has the highest proportion of people in category 1 and 2, which indicates that people suffering from DM-II are the most adept at self-management.

```
#disease type and PAM level
plot(self.manage$pam.cat ~ self.manage$disease,
     main = 'Proportions of PAM-categories for different diseases',
     xlab = "Disease type",
     ylab = "PAM level")
```

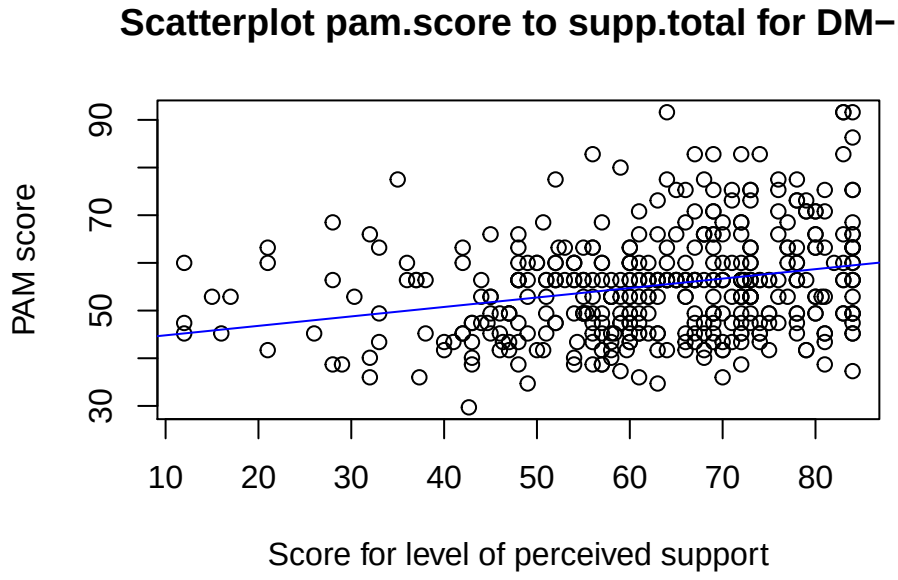


iii. I like the visualization from i) because it clearly indicates some key factors like how spread out the distributions are and where we find the median and most of the population. It is also the visualization that contains the most information. However, I also like the stacked bars from ii) which more intuitively shows how many people are in the different categories for to the different diseases. In short, the summary from i) carries more information, but the one from part ii) is however easier to understand and more intuitive, so I find that the most informative for understanding the relationship.

d)

i. There seems to be a weak positive correlation between perceived social support and PAM-score within all diseases. However, the correlation is slightly stronger within DM-II and HF, while it appears to be nearly non-existant within COPD and CRD, from looking at the scatterplots and the correlation values.

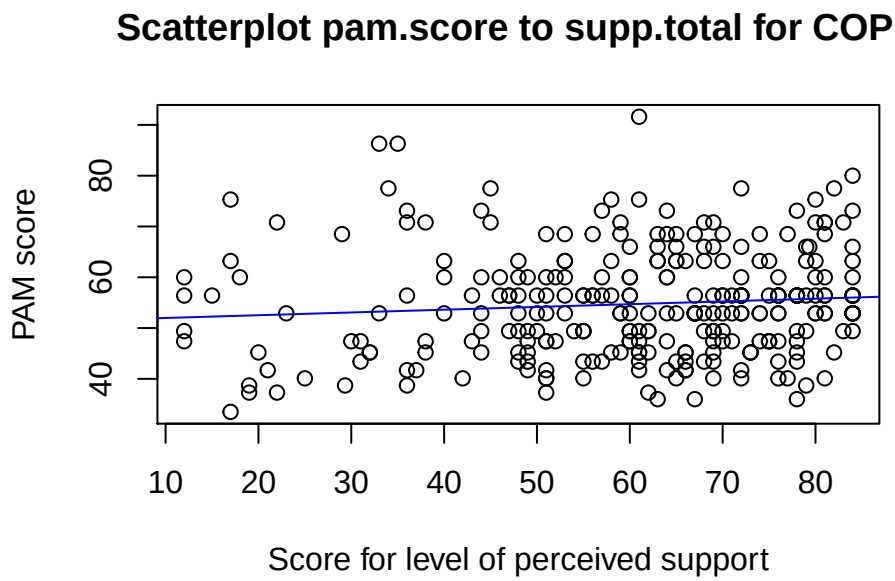
```
#PAM-13 and social support
dmii <- self.manage[self.manage$disease == 'DM-II', ]
copd <- self.manage[self.manage$disease == 'COPD', ]
hf <- self.manage[self.manage$disease == 'HF', ]
crd <- self.manage[self.manage$disease == 'CRD', ]

# Scatterplot regression-line and correlation for DM-II
plot(dmii$pam.score ~ dmii$supp.total,
     main = 'Scatterplot pam.score to supp.total for DM-II',
     xlab = "Score for level of perceived support",
     ylab = "PAM score")
abline(lm(dmii$pam.score ~ dmii$supp.total, data = mtcars), col = "blue")
```




```
cat("Correlation between social support and PAM-score for DM-II: ",
    cor(dmii$supp.total, dmii$pam.score, use = "complete.obs"))
```

```
## Correlation between social support and PAM-score for DM-II:  0.2711669
```

```
# Scatterplot regression-line and correlation for COPD
plot(copd$pam.score ~ copd$supp.total,
     main = 'Scatterplot pam.score to supp.total for COPD',
     xlab = "Score for level of perceived support",
     ylab = "PAM score")
abline(lm(copd$pam.score ~ copd$supp.total, data = mtcars), col = "blue")
```

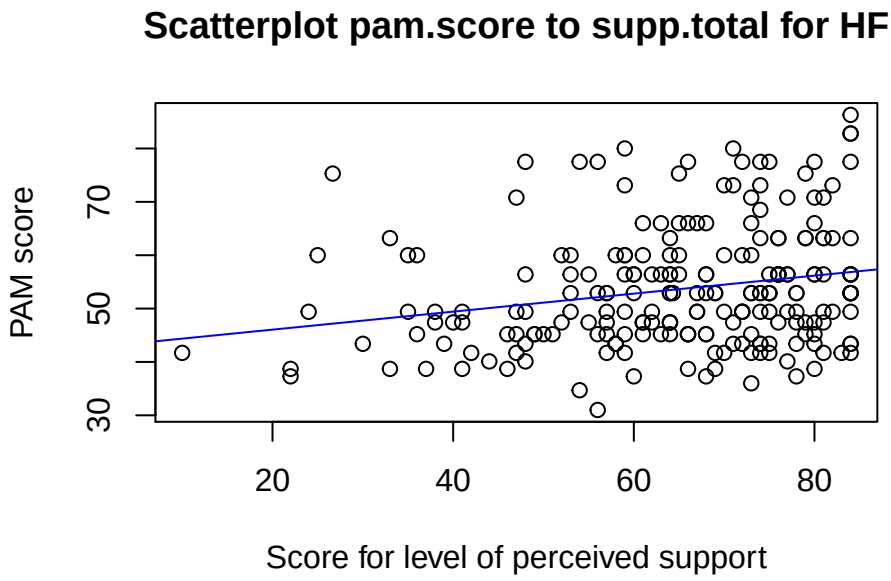



```
cat("\nCorrelation between social support and PAM-score for COPD: ",
    cor(copd$supp.total, copd$pam.score, use = "complete.obs"))
```

```
##
```

```
## Correlation between social support and PAM-score for COPD:  0.0905129
```

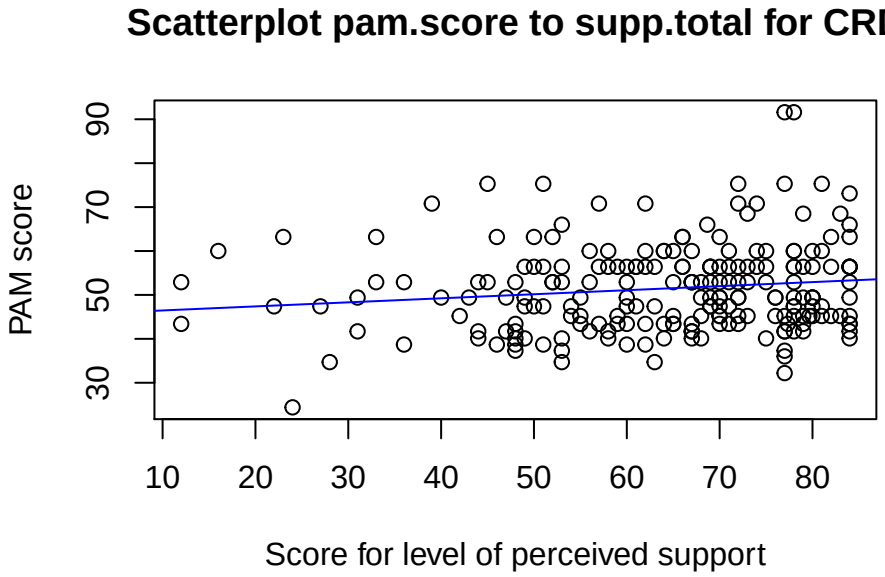
```
# Scatterplot regression-line and correlation for HF
plot(hf$pam.score ~ hf$supp.total,
     main = 'Scatterplot pam.score to supp.total for HF',
     xlab = "Score for level of perceived support",
     ylab = "PAM score")
abline(lm(hf$pam.score ~ hf$supp.total, data = mtcars), col = "blue")
```




```
cat("\nCorrelation between social support and PAM-score for nHF: ",
    cor(hf$supp.total, hf$pam.score, use = "complete.obs"))
```

```
##
## Correlation between social support and PAM-score for nHF:  0.2252179
```

```
# Scatterplot regression-line and correlation for CRD
plot(crd$pam.score ~ crd$supp.total,
     main = 'Scatterplot pam.score to supp.total for CRD',
     xlab = "Score for level of perceived support",
     ylab = "PAM score")
abline(lm(crd$pam.score ~ crd$supp.total, data = mtcars), col = "blue")
```




```
cat("\nCorrelation between social support and PAM-score for CRD: ",
    cor(crd$supp.total, crd$pam.score, use = "complete.obs"))
```

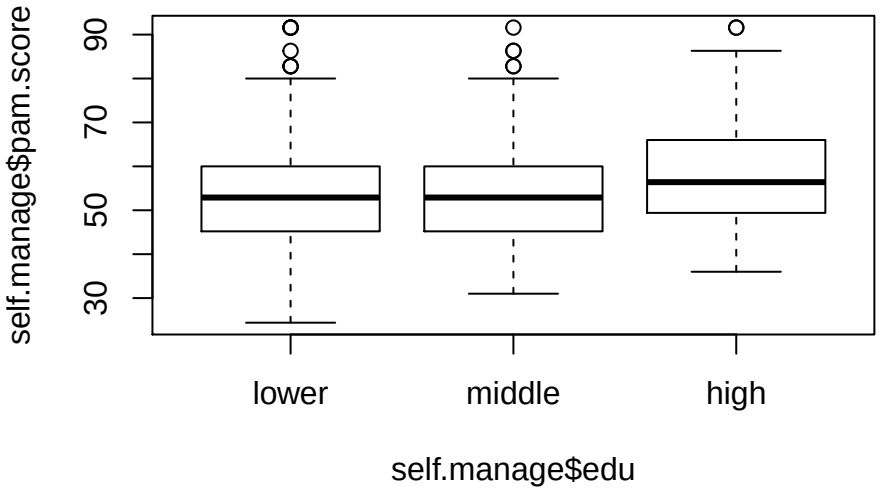
```
##
## Correlation between social support and PAM-score for CRD:  0.1408107
```

ii. Even though we see a (weak) positive correlation between social support and PAM-score, I don't think we know enough to conclude that social support causes higher PAM-scores, even though that might sound reasonable. Another explanation might for example be that people who tend to take better care of themselves also happens to make sure to have better relationships with people around them, and not the other way around. We must not confuse correlation with causation.

e)

i. From the below boxplots we see that there appears to be a relationship between education level and PAM-score, where higher education tends to mean higher PAM-scores. The effect isn't too pronounced, though, with the median among people with high education being 56.4 and 52.9 for people with either lower or middle level of education. Both the 25th and 75th percentiles are higher for the high education group than for the others.

```
#PAM-13 and educational level
plot(self.manage$pam.score ~ self.manage$edu)
```




```
cat("Median for Lower education level:",
    median(self.manage[self.manage$edu == 'lower', ]$pam.score, na.rm = TRUE))
```

```
## Median for Lower education level: 52.9
```

```
cat("\nMedian for Middle education level:",
    median(self.manage[self.manage$edu == 'middle', ]$pam.score, na.rm = TRUE))
```

```
##
## Median for Middle education level: 52.9
```

```
cat("\nMedian for High education level:",
    median(self.manage[self.manage$edu == 'high', ]$pam.score, na.rm = TRUE))
```

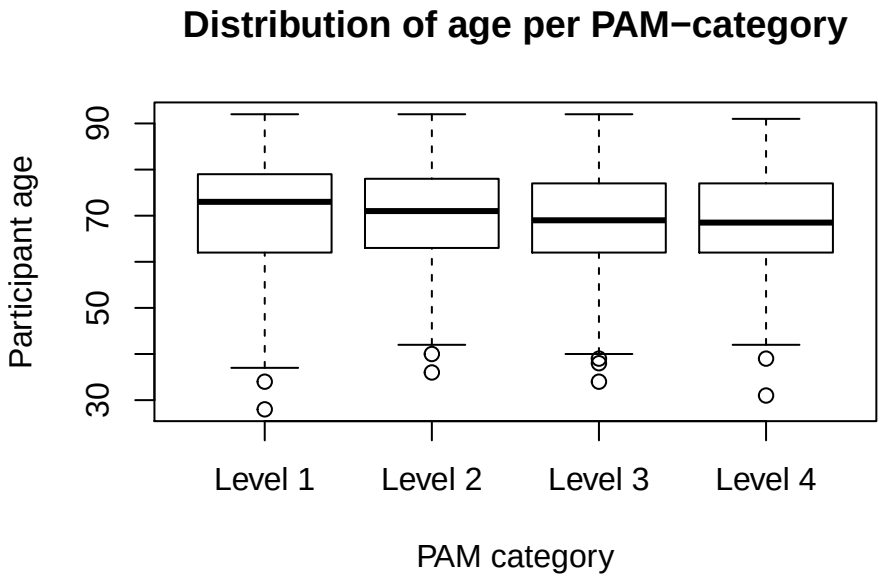
```
##
## Median for High education level: 56.4
```

ii. The PAM-score measures the *"self-reported knowledge, skills, and confidence for self-management."* Therefore, it could be the case that people with higher education just have higher confidence in their own abilites and knowledge. This they could have e.g. because they're actually better or because their education gives them the belief that they are in general more knowledgeable and confident in their own abilities without them actually being better.

f)

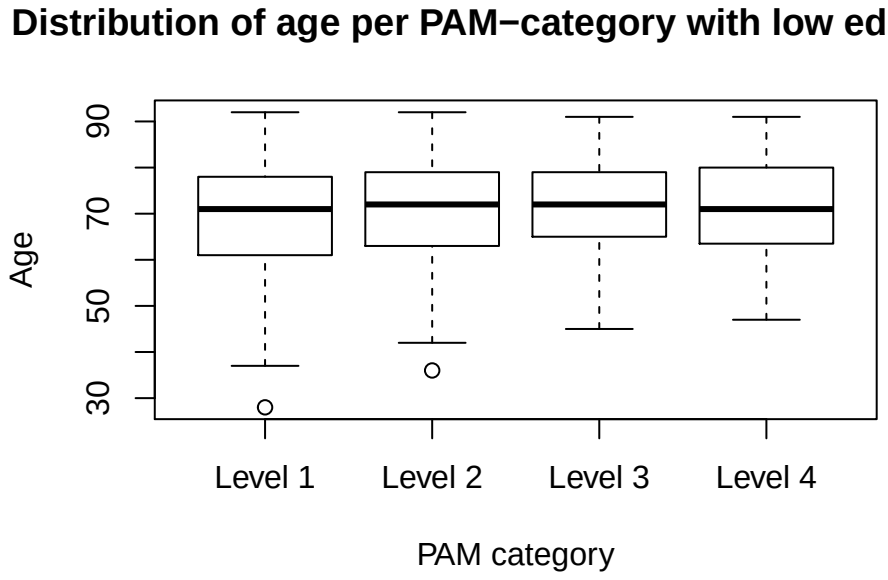
i. From the below boxplots it appears that lower age is slightly correlated with higher PAM-level. The median age is slightly decreasing for each higher PAM-level. However, we see that some of the youngest study participants can be seen as outliers in Level 1.

```
#age and PAM level
plot(self.manage$age ~ self.manage$pam.cat,
     main = "Distribution of age per PAM-category",
     xlab = "PAM category",
     ylab = "Participant age")
```

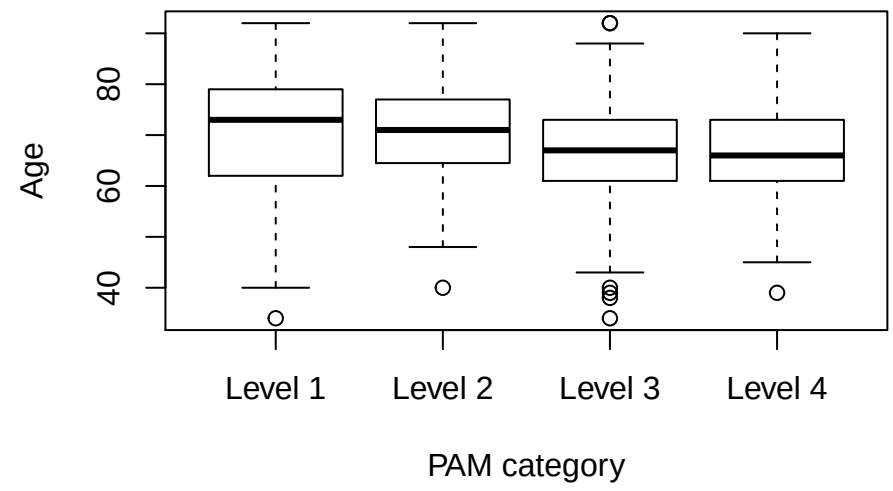


ii. Within the group of people with low educational level, the age effect seems to disappear completely, i.e. the median age is almost the same. What we do see, however, is that there's a larger spread in ages for the people in the lower levels as compared to the higher levels. For people with middle or high levels of education the effect seems to be back again, with higher PAM-level correlating with lower median age. Interestingly, within the group of people with high education levels, the median age goes down with increasing PAM-levels, except for level 4, where median age suddenly increases a little.

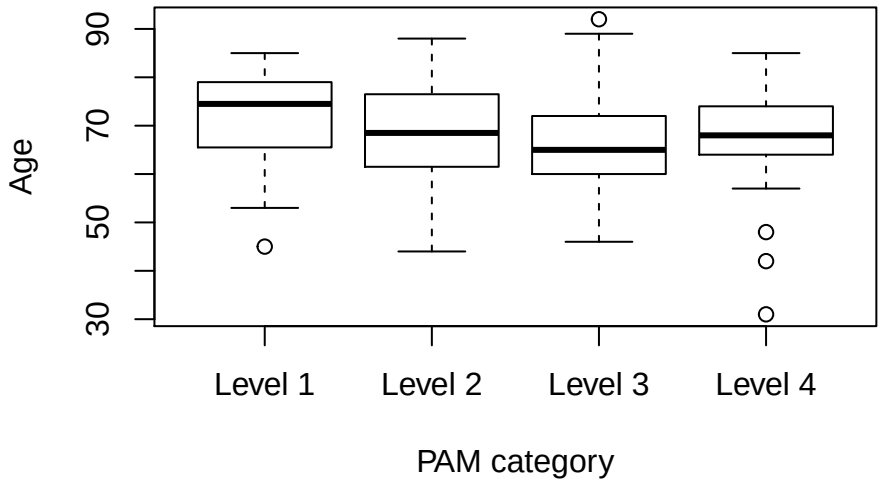
```
#age and PAM level, educational level
lower <- self.manage[self.manage$edu == 'lower', ]
middle <- self.manage[self.manage$edu == 'middle', ]
high <- self.manage[self.manage$edu == 'high', ]

plot(lower$age ~ lower$pam.cat,
     main = "Distribution of age per PAM-category with low education",
     xlab = "PAM category",
     ylab = "Age")
```




```
plot(middle$age ~ middle$pam.cat,
     main = "Distribution of age per PAM-category with middle education",
     xlab = "PAM category",
     ylab = "Age")
```




```
plot(high$age ~ high$pam.cat,
     main = "Distribution of age per PAM-category with high education",
     xlab = "PAM category",
     ylab = "Age")
```



iii. Overall, there seems to be a very weak relationship between lower age and increasing PAM-level since we see that the median age is slightly lower for people in PAM-level 4 compared to level 1. This effect goes away if we look at people with low levels of education, though. If we look at people with middle or high levels of education, we see weak correlation. Still, we see that the median age for people with high education is higher for people in PAM-level 4 than in PAM-level 3, which might suggest that this relationship isn't particularly strong. In short, there's not sufficient data to suggest that there's a relationship between age and PAM-category at this time, in my opinion.

g)

i. From the below calculations we see that the risk of being in PAM-category 1 if you're suffering from depressive disorder is 0.195, or 19.5%, while the number is 0.028 or 2.8% if you're not depressive. This indicates that the risk of being in PAM-category 1 is much larger if you're depressive.

```
#calculate risks
depressive <- na.omit(self.manage$pam.cat[self.manage$hads.depress >= 11])
non_depressive <- na.omit(self.manage$pam.cat[self.manage$hads.depress < 11])

# Probability of being in category 1 if you're suffering from depression
length(depressive[depressive == "Level 1"]) / length(depressive)
```

```
## [1] 0.5
```

```
# Probability of being in category 1 if you're not suffering from depression
length(non_depressive[non_depressive == "Level 1"]) / length(non_depressive)
```

```
## [1] 0.239801
```

ii. Depressed people have a mean PAM-score of 49.6 while the non-depressed have a PAM-score of 54.6, which suggests that people with depression tend on average to view themselves as less knowledgeable and confident about self-management than people who are not suffering from depression.

```r
# difference in mean score
depressive_score <- na.omit(self.manage$pam.score[self.manage$hads.depress >= 11])
non_depressive_score <- na.omit(self.manage$pam.score[self.manage$hads.depress < 11])
# Mean score for people suffering from depressive disorders
mean(depressive_score)
```

```
## [1] 49.62333
```

```r
# Mean score for people suffering from depressive disorders
mean(non_depressive_score)
```

```
## [1] 54.59313
```

iii. I think it's likely that people that do not suffer from depression have a likelyhood of not having responded to the HADS questionaire. This is because people who are suffering from depression will most likely seek help for it, and part of that process, I imagine, is taking this HADS questionaire. People who do not suffer will with higher likelyhood think that the questionaire is a waste of time and don't answer it. That makes it likely that th results could be skewed by the existence of NA's. At the same time, as we can se from the below calculation, the amount of missing responses is only roughly 2.5% of responses so I doubt that it would bias the outcome very much.

```r
# Calculate the percentage of NA's among the participants
nas <- length(self.manage$hads.depress[is.na(self.manage$hads.depress)])
all_rows <- length(self.manage$hads.depress)
nas; all_rows
```

```
## [1] 29
```

```
## [1] 1154
```

```r
nas / all_rows
```

```
## [1] 0.02512998
```